Bandwidth-Efficient Digital Modulation with Application to Deep-Space Communications

Marvin K. Simon

MONOGRAPH 3
DEEP–SPACE COMMUNICATIONS AND NAVIGATION SERIES

Bandwidth-Efficient Digital Modulation with Application to Deep-Space Communications

DEEP-SPACE COMMUNICATIONS AND NAVIGATION SERIES

Issued by the Deep-Space Communications and Navigation Systems
Center of Excellence
Jet Propulsion Laboratory
California Institute of Technology

Joseph H. Yuen, Editor-in-Chief

Previously Published Monographs in this Series

1. *Radiometric Tracking Techniques for Deep-Space Navigation*
   C. L. Thornton and J. S. Border

2. *Formulation for Observed and Computed Values of Deep Space Network Data Types for Navigation*
   Theodore D. Moyer

Bandwidth-Efficient Digital Modulation with Application to Deep-Space Communications

Marvin K. Simon

Jet Propulsion Laboratory
California Institute of Technology

With Technical Contributions by

Dennis Lee
Warren L. Martin
Haiping Tsou
Tsun-Yee Yan

of the Jet Propulsion Laboratory

MONOGRAPH 3
DEEP–SPACE COMMUNICATIONS AND NAVIGATION SERIES

Bandwidth-Efficient Digital Modulation with
Application to Deep-Space Communications
(JPL Publication 00-17)

June 2001

The research described in this publication was carried out at the
Jet Propulsion Laboratory, California Institute of Technology,
under a contract with the National Aeronautics and Space Administration.

Table of Contents

Foreword vii

Preface ix

Chapter 1: Introduction 1

Chapter 2: Constant Envelope Modulations 3

2.1 The Need for Constant Envelope 3

2.2 Quadriphase-Shift-Keying and Offset (Staggered) Quadriphase-Shift-Keying 4

2.3 Differentially Encoded QPSK and Offset (Staggered) QPSK 8

2.4 π/4-QPSK: A Variation of Differentially Encoded QPSK with Instantaneous Amplitude Fluctuation Halfway between That of QPSK and OQPSK 9

2.5 Power Spectral Density Considerations 12

2.6 Ideal Receiver Performance 12

2.7 Performance in the Presence of Nonideal Transmitters 12

2.7.1 Modulator Imbalance and Amplifier Nonlinearity ... 12

2.7.2 Data Imbalance 26

2.8 Continuous Phase Modulation 26

2.8.1 Full Response—MSK and SFSK 27

2.8.2 Partial Response—Gaussian MSK 57

2.9 Simulation Performance 113

References 116

Chapter 3: Quasi-Constant Envelope Modulations 125

3.1 Brief Review of IJF-QPSK and SQORC and their Relation to FQPSK 129

3.2 A Symbol-by-Symbol Cross-Correlator Mapping for FQPSK 136

3.3 Enhanced FQPSK 143

3.4 Interpretation of FQPSK as a Trellis-Coded Modulation 146
3.5 Optimum Detection 147
3.6 Suboptimum Detection 152
3.6.1 Symbol-by-Symbol Detection 152
3.6.2 Average Bit-Error Probability Performance 159
3.6.3 Further Receiver Simplifications and FQPSK-B Performance 161
3.7 Cross-Correlated Trellis-Coded Quadrature Modulation 166
3.7.1 Description of the Transmitter 168
3.7.2 Specific Embodiments 172
3.8 Other Techniques 177
3.8.1 Shaped Offset QPSK 177
References 184
Chapter 4: Bandwidth-Efficient Modulations with More Envelope Fluctuation 187
4.1 Bandwidth-Efficient TCM with Prescribed Decoding Delay—Equal Signal Energies 190
4.1.1 ISI-Based Transmitter Implementation 190
4.1.2 Evaluation of the Power Spectral Density 195
4.1.3 Optimizing the Bandwidth Efficiency 204
4.2 Bandwidth-Efficient TCM with Prescribed Decoding Delay—Unequal Signal Energies 212
References 218
Chapter 5: Strictly Bandlimited Modulations with Large Envelope Fluctuation (Nyquist Signaling) 219
5.1 Binary Nyquist Signaling 219
5.2 Multilevel and Quadrature Nyquist Signaling 223
References 223
Chapter 6: Summary 225
6.1 Throughput Performance Comparisons 225
References 226

Foreword

The Deep Space Communications and Navigation Systems Center of Excellence (DESCANSO) was recently established for the National Aeronautics and Space Administration (NASA) at the California Institute of Technology's Jet Propulsion Laboratory (JPL). DESCANSO is chartered to harness and promote excellence and innovation to meet the communications and navigation needs of future deep-space exploration.

DESCANSO's vision is to achieve continuous communications and precise navigation—any time, anywhere. In support of that vision, DESCANSO aims to seek out and advocate new concepts, systems, and technologies; foster key scientific and technical talents; and sponsor seminars, workshops, and symposia to facilitate interaction and idea exchange.

The Deep Space Communications and Navigation Series, authored by scientists and engineers with many years of experience in their respective fields, lays a foundation for innovation by communicating state-of-the-art knowledge in key technologies. The series also captures fundamental principles and practices developed during decades of deep-space exploration at JPL. In addition, it celebrates successes and imparts lessons learned. Finally, the series will serve to guide a new generation of scientists and engineers.

Joseph H. Yuen
DESCANSO Leader

Preface

Traditional modulation methods adopted by space agencies for transmitting telecommand and telemetry data have incorporated subcarriers as a simple means of separating different data types as well ensuring no overlap between the radio frequency (RF) carrier and the modulated data's frequency spectra. Unfortunately, subcarrier modulation suffers from a number of disadvantages, namely, greater spacecraft complexity, additional losses in the modulation/demodulation process, and most important, at least from the standpoint of this monograph, a large, occupied bandwidth. One effort to mitigate the latter was to replace the more traditional square-wave subcarriers with sine-wave carriers, but this was not considered to be an acceptable solution for all space-exploration missions.

In the early digital communication years (i.e., 1960s and 1970s), bandwidth occupancy was really not an issue because of low data rates and the requirement for only a few data channels (subcarriers). Consequently, other attempts at limiting bandwidth occupancy were not considered at that time. As missions became more complex, however, the RF spectrum became more congested, and data rates continued to grow, thus requiring an attendant increase in subcarrier frequencies (equivalently, occupied bandwidth) and along with that, an increased susceptibility to interference from different spacecraft. A point came at which it was no longer feasible to use subcarrier-based modulation methods. Fortunately, during this same period, improved bandwidth-efficient modulation methods that directly modulated the carrier were being developed, which, along with improved data formatting methods (e.g., packet transfer frame telemetry) to handle the multiple channel separation problem, eliminated the need for subcarriers. Combining the packet telemetry format with any of the direct modulation methods and applying

additional spectral pulse shaping to the latter now made it possible to transmit messages at a high data rate while using a comparatively small bandwidth.

The purpose of this monograph is to define, describe, and then give the performance (power and bandwidth) of digital communication systems that incorporate a large variety of the bandwidth-efficient modulations referred to above. In addition to considering the ideal behavior of such systems, we shall also cover their performance in the presence of a number of practical (non-ideal) transmitter and receiver characteristics such as modulator and phase imbalance, imperfect carrier synchronization, and transmitter nonlinearity. With regard to the latter, the requirement of operating the transmitter at a high power efficiency, i.e., running the power amplifier in a saturated or near-saturated condition, implies that one employ a constant envelope modulation. This constraint restricts the type of modulations that can be considered, which in turn restricts the amount of spectral occupancy and power efficiency that can be achieved. Relaxing the constant envelope condition (which then allows for a more linear but less efficient transmitter power amplifier operation) potentially eases the restrictions on power and bandwidth efficiency to the extreme limit of Nyquist-type signaling, which, in theory, is strictly bandlimited and capable of achieving the maximum power efficiency. Because of this inherent trade-off between envelope (or more correctly, instantaneous amplitude) fluctuation of the modulation and the degree of power and bandwidth efficiency attainable, we have chosen to structure this monograph in a way that clearly reflects this issue. In particular, we start by discussing strictly constant envelope modulations and then, moving in the direction of more and more envelope fluctuation, end with a review of strictly bandlimited (Nyquist-type) signaling. Along the way, we consider a number of quasi-constant envelope modulations that have gained considerable notoriety in recent years and represent a good balance among the above-mentioned power and bandwidth trade-off considerations.

Finally, it should be mentioned that although the monograph attempts to cover a large body of the published literature in this area, the real focus is on the research and the results obtained at the Jet Propulsion Laboratory (JPL). As such, we do not offer this document to the readership as an all-inclusive treatise on the subject of bandwidth-efficient modulations but rather one that, as the title reflects, highlights the many technical contributions performed under NASA-funded tasks pertaining to the development and design of deep-space communications systems. When taken in this context, we hope that, in addition to being informative, this document will serve as an inspiration to future engineers to continue the fine work that was initiated at JPL and has been reported on herein.

Marvin K. Simon
June 2001

Chapter 1
Introduction

The United States Budget Reconciliation Act of 1993 mandates reallocation of a minimum of 200 MHz of spectrum below 5 GHz for licensing to nonfederal users. One of the objectives is to promote and encourage novel spectrum-inspired technology developments and wireless applications. Many user organizations and communications companies have been developing advanced modulation techniques in order to more efficiently use the spectrum.

In 1998, the international Space Frequency Coordination Group (SFCG) adopted a spectral mask that precludes the use of a number of classical modulation schemes for missions launched after 2002. The SFCG has recommended several advanced modulations that potentially could reduce spectrum congestion. No one technique solves every intended application. Many trade-offs must be made in selecting a particular technique, the trade-offs being defined by the communications environment, data integrity requirements, data latency requirements, user access, traffic loading, and other constraints. These new modulation techniques have been known in theory for many years, but have become feasible only because of recent advances in digital signal processing and microprocessor technologies.

This monograph focuses on the most recent advances in spectrum-efficient modulation techniques considered for government and commercial applications. Starting with basic, well-known digital modulations, the discussion will evolve to more sophisticated techniques that take on the form of constant envelope modulations, quasi-constant envelope modulations, nonconstant envelope modulations, and finally Nyquist-rate modulations. Included in the discussion will be a unified treatment based on recently developed cross-correlated trellis-coded quadrature modulation (XTCQM), which captures a number of state-of-the-art spectrally efficient modulation schemes. Performance analysis, computer simulation results, and their hardware implications will be addressed. Comparisons of

different modulation schemes recommended by the Consultative Committee for Space Data Systems (CCSDS), an international organization for cross support among space agencies, for SFCG will be discussed.

Chapter 2
Constant Envelope Modulations

2.1 The Need for Constant Envelope

Digital communication systems operate in the presence of path loss and atmospheric-induced fading. In order to maintain sufficient received power at the destination, it is required that a device for generating adequate transmitter output power based on fixed- but-limited available power be employed, examples of which are traveling-wave tube amplifiers (TWTAs) and solid-state power amplifiers (SSPAs) operated in full- saturation mode to maximize conversion efficiency. Unfortunately, this requirement introduces amplitude modulation-amplitude modulation (AM-AM) and amplitude modulation-phase modulation (AM-PM) conversions into the transmitted signal. Because of this, modulations that transmit information via their amplitude, e.g., quadrature amplitude modulation (QAM), and therefore need a linear amplifying characteristic, are not suitable for use on channels operated in the above maximum transmitter power efficiency requirement.[1] Another consideration regarding radio frequency (RF) amplifier devices that operate in a nonlinear mode at or near saturation is the spectral spreading that they reintroduce due to the nonlinearity subsequent to bandlimiting the modulation prior to amplification. Because of the need for the transmitted power spectrum to fall under a specified mask imposed by regulating agencies such as the FCC or International Telecommunications Union (ITU), the modulation must be designed to keep this spectral spreading to a minimum. This constraint necessitates limiting the amount of instantaneous amplitude fluctuation in the transmitted waveform in addition to imposing the requirement for constant envelope.

[1] An approach whereby it might be possible to generate QAM-type modulations using separate nonlinearly operated high-power amplifiers on the inphase (I) and quadrature (Q) channels is currently under investigation by the author.

Because of the above considerations regarding the need for high transmitter power efficiency, it is clearly desirable to consider modulations that achieve their bandwidth efficiency by means other than resorting to multilevel amplitude modulation. Such constant envelope modulations are the subject of discussion in the first part of this monograph. Because of the large number of possible candidates, to keep within the confines of a reasonable size book, we shall restrict our attention to only those that have some form of inphase-quadrature phase (I-Q) representation and as such an I-Q form of receiver.

2.2 Quadriphase-Shift-Keying and Offset (Staggered) Quadriphase-Shift-Keying

M-ary phase-shift-keying (M-PSK) produces a constant envelope signal that is mathematically modeled in complex form[2] as

$$\tilde{s}\,(t) = \sqrt{2P}e^{j(2\pi f_c t+\theta(t)+\theta_c)} = \tilde{S}\,(t)\,e^{j(2\pi f_c t+\theta_c)} \tag{2.2-1}$$

where P is the transmitted power, f_c is the carrier frequency in hertz, θ_c is the carrier phase, and $\theta(t)$ is the data phase that takes on equiprobable values $\beta_i = (2i-1)\,\pi/M,\ i = 1, 2, \cdots, M$, in each symbol interval, T_s. As such, $\theta(t)$ is modeled as a random pulse stream, that is,

$$\theta(t) = \sum_{n=-\infty}^{\infty} \theta_n p\,(t - nT_s) \tag{2.2-2}$$

where θ_n is the information phase in the nth symbol interval, $nT_s < t \leq (n+1)T_s$, ranging over the set of M possible values β_i as above, and $p(t)$ is a unit amplitude rectangular pulse of duration T_s seconds. The symbol time, T_s, is related to the bit time, T_b, by $T_s = T_b \log_2 M$ and, thus, the nominal gain in bandwidth efficiency relative to binary phase-shift-keying (BPSK), i.e., $M = 2$, is a factor of $\log_2 M$. The signal constellation is a unit circle with points uniformly spaced by $2\pi/M$ rad. Thus, the complex signal transmitted in the nth symbol interval is

$$\tilde{s}(t) = \sqrt{2P}e^{j(2\pi f_c t+\theta_n+\theta_c)}, \quad nT_s < t \leq (n+1)T_s, \quad n = -\infty, \cdots, \infty \tag{2.2-3}$$

[2] The actual (real) transmitted signal is $s\,(t) = \mathrm{Re}\,\{\tilde{s}(t)\} = \sqrt{2P}\cos\left(2\pi f_c t + \theta(t) + \theta_c\right)$.

Note that because of the assumed rectangular pulse shape, the complex baseband signal $\tilde{S}(t) = \sqrt{2P}e^{j\theta_n}$ is constant in this same interval and has envelope $\left|\tilde{S}(t)\right| = \sqrt{2P}$.

A special case of M-PSK that has an I-Q representation is quadriphase-shift-keying (QPSK), and corresponds to $M = 4$. Here it is conventional to assume that the phase set $\{\beta_i\}$ takes on values $\pi/4,\ 3\pi/4, 5\pi/4, 7\pi/4$. Projecting these information phases on the quadrature amplitude axes, we can equivalently write QPSK in the nth symbol interval in the complex I-Q form[3]

$$\tilde{s}(t) = \sqrt{P}\left(a_{In} + ja_{Qn}\right)e^{j(2\pi f_c t + \theta_c)}, \quad nT_s < t \leq (n+1)T_s \tag{2.2-4}$$

where the information amplitudes a_{In} and a_{Qn} range independently over the equiprobable values ± 1. Here again, because of the assumed rectangular pulse shape, the complex baseband signal $\tilde{S}(t) = \sqrt{P}\left(a_{In} + ja_{Qn}\right)$ is constant in this same interval. The real transmitted signal corresponding to (2.2-4) has the form

$$s(t) = \sqrt{P}m_I(t)\cos(2\pi f_c t + \theta_c) - \sqrt{P}m_Q(t)\sin(2\pi f_c t + \theta_c),$$

$$m_I(t) = \sum_{n=-\infty}^{\infty} a_{In}p(t - nT_s), \quad m_Q(t) = \sum_{n=-\infty}^{\infty} a_{Qn}p(t - nT_s) \tag{2.2-5}$$

If one examines the form of (2.2-4) it becomes apparent that a large fluctuation of the instantaneous amplitude between symbols corresponding to a 180-deg phase reversal can occur when both a_{In} and a_{Qn} change polarity at the same time. As mentioned in Sec. 2.1, it is desirable to limit the degree of such fluctuation to reduce spectral regrowth brought about by the transmit amplifier nonlinearity, i.e., the smaller the fluctuation, the smaller the sidelobe regeneration and vice versa. By offsetting (staggering) the I and Q modulations by $T_s/2$ s, one guarantees the fact that a_{In} and a_{Qn} cannot change polarity at the same time. Thus, the maximum fluctuation in instantaneous amplitude is now limited to that corresponding to a 90-deg phase reversal (i.e., either a_{In} or a_{Qn}, but not both, change polarity). The resulting modulation, called offset (staggered) QPSK (OQPSK), has a signal of the form

---

[3] One can think of the complex carrier as being modulated now by a complex random pulse stream, namely, $\tilde{a}(t) = \sum_{n=-\infty}^{\infty}\left(a_{In} + ja_{Qn}\right)p(t - nT_s)$.

$$s(t) = \sqrt{P} m_I(t) \cos(2\pi f_c t + \theta_c) - \sqrt{P} m_Q(t) \sin(2\pi f_c t + \theta_c),$$

$$m_I(t) = \sum_{n=-\infty}^{\infty} a_{In} p(t - nT_s), \quad m_Q(t) = \sum_{n=-\infty}^{\infty} a_{Qn} p\left(t - \left(n + \frac{1}{2}\right) T_s\right) \tag{2.2-6}$$

While it is true that for M-PSK with $M = 2^m$ and m an arbitrary integer, the information phases can be projected on the I and Q coordinates and as such obtain, in principle, an I-Q transmitter representation, it should be noted that the number of possible I-Q amplitude pairs obtained from these projections exceeds M. Consequently, decisions on the resulting I and Q multilevel amplitude signals at the receiver are not independent in that each pair of amplitude decisions does not necessarily render one of the transmitted phases. Therefore, for $M \geq 8$ it is not practical to view M-PSK in an I-Q form.

The detection of an information phase can be obtained by combining the detections on the I and Q components of this phase. The receiver for QPSK is illustrated in Fig. 2-1(a) while the analogous receiver for OQPSK is illustrated in Fig. 2-1(b). The decision variables that are input to the hard-limiting threshold devices are

$$\left.\begin{aligned} y_{In} &= a_{In}\sqrt{P}T_s + N_{In} \\ y_{Qn} &= a_{Qn}\sqrt{P}T_s + N_{Qn} \end{aligned}\right\} \tag{2.2-7}$$

where for QPSK

$$\left.\begin{aligned} N_{In} &= \mathrm{Re}\left\{\int_{nT_s}^{(n+1)T_s} \tilde{N}(t)\, dt\right\} \\ N_{Qn} &= \mathrm{Im}\left\{\int_{nT_s}^{(n+1)T_s} \tilde{N}(t)\, dt\right\} \end{aligned}\right\} \tag{2.2-8}$$

whereas for OQPSK

$$\left.\begin{aligned} N_{In} &= \mathrm{Re}\left\{\int_{nT_s}^{(n+1)T_s} \tilde{N}(t)\,dt\right\} \\ N_{Qn} &= \mathrm{Im}\left\{\int_{(n+1/2)T_s}^{(n+3/2)T_s} \tilde{N}(t)\,dt\right\} \end{aligned}\right\} \tag{2.2-9}$$

In either case, N_{In}, N_{Qn} are zero mean Gaussian random variables (RVs) with variance $\sigma_N^2 = N_0 T_s/2$ and thus conditioned on the data symbols, y_{In}, y_{Qn} are also Gaussian RVs with the same variance.

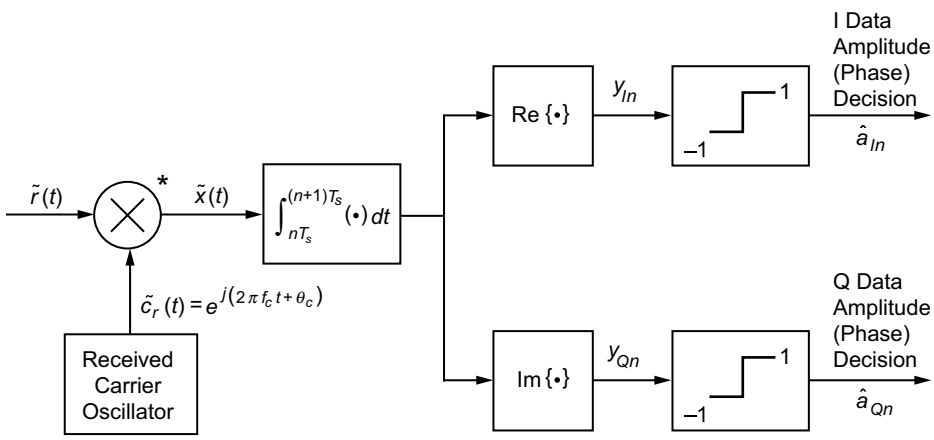


Fig. 2-1(a). Complex form of optimum receiver for ideal coherent detection of QPSK over the AWGN.

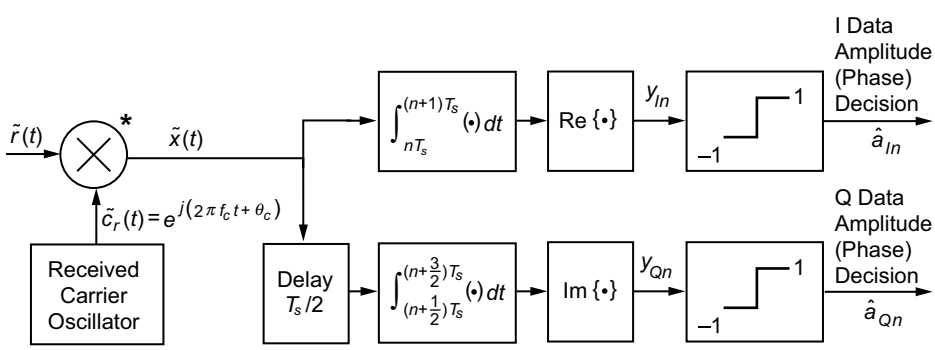


Fig. 2-1(b). Complex form of optimum receiver for ideal coherent detection of OQPSK over the AWGN.

2.3 Differentially Encoded QPSK and Offset (Staggered) QPSK

In an actual coherent communication system transmitting M-PSK modulation, means must be provided at the receiver for establishing the local demodulation carrier reference signal. This means is traditionally accomplished with the aid of a suppressed carrier-tracking loop [1, Chap. 2]. Such a loop for M-PSK modulation exhibits an M-fold phase ambiguity in that it can lock with equal probability at the transmitted carrier phase plus any of the M information phase values. Hence, the carrier phase used for demodulation can take on any of these same M phase values, namely, $\theta_c + \beta_i = \theta_c + 2i\pi/M,\ i = 0, 1, 2, \cdots, M-1$. Coherent detection cannot be successful unless this M-fold phase ambiguity is resolved.

One means for resolving this ambiguity is to employ differential phase encoding (most often simply called differential encoding) at the transmitter and differential phase decoding (most often simply called differential decoding) at the receiver following coherent detection. That is, the information phase to be communicated is modulated on the carrier as the difference between two adjacent transmitted phases, and the receiver takes the difference of two adjacent phase decisions to arrive at the decision on the information phase.[4] In mathematical terms, if $\Delta\theta_n$ were the information phase to be communicated in the nth transmission interval, the transmitter would first form $\theta_n = \theta_{n-1} + \Delta\theta_n$ modulo 2π (the differential encoder) and then modulate θ_n on the carrier.[5] At the receiver, successive decisions on θ_{n-1} and θ_n would be made and then differenced modulo 2π (the differential decoder) to give the decision on $\Delta\theta_n$. Since the decision on the true information phase is obtained from the difference of two adjacent phase decisions, a performance penalty is associated with the inclusion of differential encoding/decoding in the system.

For QPSK or OQPSK, the differential encoding/decoding process can be performed on each of the I and Q channels independently. A block diagram of a receiver for differentially encoded QPSK (or OQPSK) would be identical to that shown in Fig. 2-1(a) [or Fig. 2-1(b)], with the inclusion of a binary differential decoder in each of the I and Q arms following the hard-decision devices [see

---

[4] Note that this receiver (i.e., the one that makes optimum coherent decisions on two successive symbol phases and then differences these to arrive at the decision on the information phase) is suboptimum when $M > 2$ [2]. However, this receiver structure, which is the one classically used for coherent detection of differentially encoded M-PSK, can be arrived at by a suitable approximation of the likelihood function used to derive the true optimum receiver, and at high signal-to-noise ratio (SNR), the difference between the two becomes mute.

[5] Note that we have shifted our notation here insofar as the information phases are concerned so as to keep the same notation for the phases actually transmitted.

Figs. 2-2(a) and 2-2(b)].[6] Inclusion of differentially encoded OQPSK in our discussion is important since, as we shall see later on, other forms of modulation, e.g., minimum-shift-keying (MSK), have an I-Q representation in the form of pulse-shaped, differentially encoded OQPSK.

2.4 π/4-QPSK: A Variation of Differentially Encoded QPSK with Instantaneous Amplitude Fluctuation Halfway between That of QPSK and OQPSK

Depending on the set of phases, $\{\Delta\beta_i\}$, used to represent the information phase, $\Delta\theta_n$, in the nth transmission interval, the actual transmitted phase, θ_n, in this same transmission interval can range either over the same set, $\{\beta_i\} = \{\Delta\beta_i\}$, or over another phase set. If for QPSK, we choose the set $\Delta\beta_i = 0, \pi/2,\ \pi, 3\pi/2$ to represent the information phases, then starting with an initial transmitted phase chosen from the set $\pi/4,\ 3\pi/4, 5\pi/4, 7\pi/4$, the subsequent transmitted phases, $\{\theta_n\}$, will also range over the set $\pi/4,\ 3\pi/4, 5\pi/4, 7\pi/4$ in every transmission interval. This is the conventional form of differentially encoded QPSK, as discussed in the previous section. Now suppose instead that the set $\Delta\beta_i = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$ is used to represent the information phases, $\{\Delta\theta_n\}$. Then, starting, for example, with an initial phase chosen from the set $\pi/4, 3\pi/4, 5\pi/4, 7\pi/4$, the transmitted phase in the next interval will range over the set $0, \pi/2, \pi, 3\pi/2$. In the following interval, the transmitted phase will range over the set $\pi/4, 3\pi/4, 5\pi/4, 7\pi/4$, and in the interval following that one, the transmitted phase will once again range over the set $0, \pi/2, \pi, 3\pi/2$. Thus, we see that for this choice of phase set corresponding to the information phases, $\{\Delta\theta_n\}$, the transmitted phases, $\{\theta_n\}$, will alternatively range over the sets $0, \pi/2, \pi, 3\pi/2$ and $\pi/4, 3\pi/4, 5\pi/4, 7\pi/4$. Such a modulation scheme, referred to as $\pi/4$-QPSK [3], has an advantage relative to conventional differentially encoded QPSK in that the maximum change in phase from transmission to transmission is 135 deg, which is halfway between the 90-deg maximum phase change of OQPSK and 180-deg maximum phase change of QPSK.

In summary, on a linear additive white Gaussian noise (AWGN) channel with ideal coherent detection, all three types of differentially encoded QPSK, i.e., conventional (nonoffset), offset, and $\pi/4$ perform identically. The differences among the three types on a linear AWGN channel occur when the carrier demodulation phase reference is not perfect, which corresponds to nonideal coherent detection.

---

[6] Since the introduction of a 180-deg phase shift to a binary phase sequence is equivalent to a reversal of the polarity of the binary data bits, a binary differential encoder is characterized by $a_n = a_{n-1}b_n$ and the corresponding binary differential decoder is characterized by $b_n = a_{n-1}a_n$ where $\{b_n\}$ are now the information bits and $\{a_n\}$ are the actual transmitted bits on each channel.

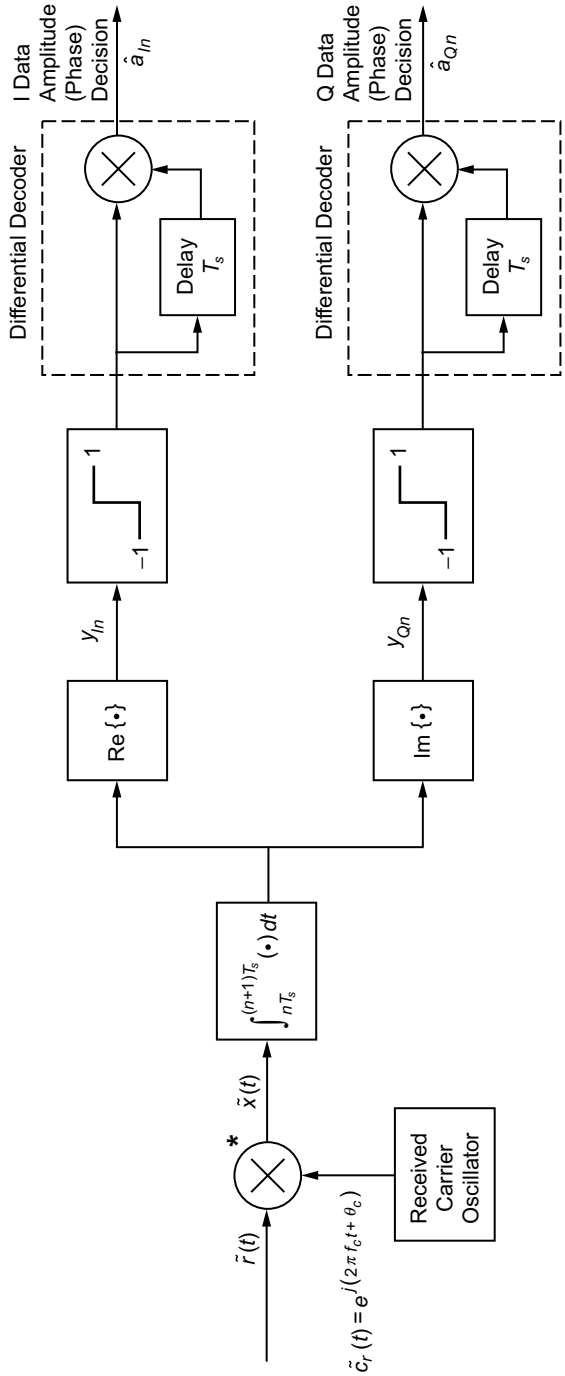


Fig. 2-2(a). Complex form of optimum receiver for ideal coherent detection of differentially encoded QPSK over the AWGN.

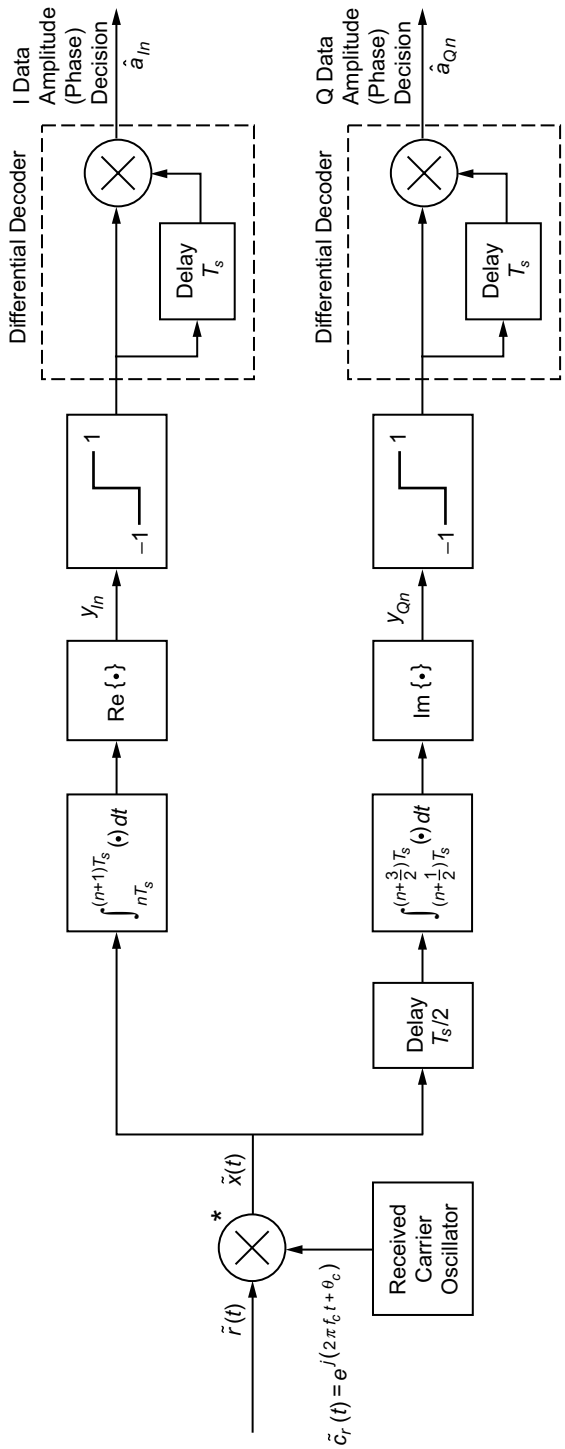


Fig. 2-2(b). Complex form of optimum receiver for ideal coherent detection of differentially encoded OQPSK over the AWGN.

2.5 Power Spectral Density Considerations

The power spectral densities (PSD) of QPSK, OQPSK, and the differentially encoded versions of these are all identical and are given by

$$S(f) = PT_s \left(\frac{\sin \pi f T_s}{\pi f T_s} \right)^2 \tag{2.5-1}$$

We see that the asymptotic (large f) rate of rolloff of the PSD varies as f^{-2}, and a first null (width of the main lobe) occurs at $f = 1/T_s = 1/2T_b$. Furthermore, when compared with BPSK, QPSK is exactly twice as bandwidth efficient.

2.6 Ideal Receiver Performance

Based upon the decision variables in (2.2-7) the receiver for QPSK or OQPSK makes its I and Q data decisions from

$$\left.\begin{aligned} \hat{a}_{In} &= \operatorname{sgn} y_{In} \\ \hat{a}_{Qn} &= \operatorname{sgn} y_{Qn} \end{aligned}\right\} \tag{2.6-1}$$

which results in an average bit-error probability (BEP) given by

$$P_b(E) = \frac{1}{2} \operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}} \right), \quad E_b = PT_b \tag{2.6-2}$$

and is identical to that of BPSK. Thus, we conclude that ideally BPSK, QPSK, and OQPSK have the identical BEP performance although the latter two occupy half the bandwidth.

2.7 Performance in the Presence of Nonideal Transmitters

2.7.1 Modulator Imbalance and Amplifier Nonlinearity

The deleterious effect on receiver performance of modulator phase and amplitude imbalance and amplifier nonlinearity has been studied by several researchers [3–10]. With regard to modulator imbalances, the primary source of degradation comes about because of the effect of the imbalance on the steady-state lock point of the carrier tracking loop, which has a direct impact on the determination of

accurate average BEP performance. Here, we summarize some of these results for QPSK and OQPSK, starting with modulator imbalance acting alone and then later on in combination with amplifier nonlinearity. We begin our discussion with a description of an imbalance model associated with a modulator for generating these signals.

2.7.1.1 Modulator Imbalance Model. QPSK can be implemented with two balanced modulators, one on each of the I and Q channels, as illustrated in Fig. 2-3. Each of these modulators is composed of two AM modulators with inputs equal to the input nonreturn-to-zero (NRZ) data stream and its inverse (bit polarities inverted). The difference of the outputs of the two AM modulators serves as the BPSK transmitted signal on each channel. A mathematical description of the I and Q channel signals in the presence of amplitude and phase imbalances introduced by the AM modulators is[7]

$$s_I(t) = \frac{\sqrt{P}}{2} m_I(t) \left[\cos(2\pi f_c t + \theta_{cI}) + \Gamma_I \cos(2\pi f_c t + \theta_{cI} + \Delta\theta_{cI})\right]$$

$$+ \frac{\sqrt{P}}{2} \left[\cos(2\pi f_c t + \theta_{cI}) - \Gamma_I \cos(2\pi f_c t + \theta_{cI} + \Delta\theta_{cI})\right] \qquad (2.7\text{-}1a)$$

$$s_Q(t) = \frac{\sqrt{P}}{2} m_Q(t) \left[\sin(2\pi f_c t + \theta_{cQ}) + \Gamma_Q \sin(2\pi f_c t + \theta_{cQ} + \Delta\theta_{cQ})\right]$$

$$+ \frac{\sqrt{P}}{2} \left[\sin(2\pi f_c t + \theta_{cQ}) - \Gamma_Q \sin(2\pi f_c t + \theta_{cQ} + \Delta\theta_{cQ})\right] \qquad (2.7\text{-}1b)$$

$$s(t) = s_I(t) + s_Q(t)$$

where θ_{cI}, θ_{cQ} are the local oscillator carrier phases associated with the I and Q balanced modulators, Γ_I, Γ_Q (both assumed to be less than unity) are the relative amplitude imbalances of these same modulators, and $\Delta\theta_{cI}, \Delta\theta_{cQ}$ are the phase imbalances between the two AM modulators in each of the I and Q

[7] To be consistent with the usage in Ref. 8, we define the transmitted signal as the sum of the I and Q signals, i.e., $s(t) = s_I(t) + s_Q(t)$ rather than their difference as in the more traditional usage of (2.2-5). This minor switch in notation is of no consequence to the results that follow.

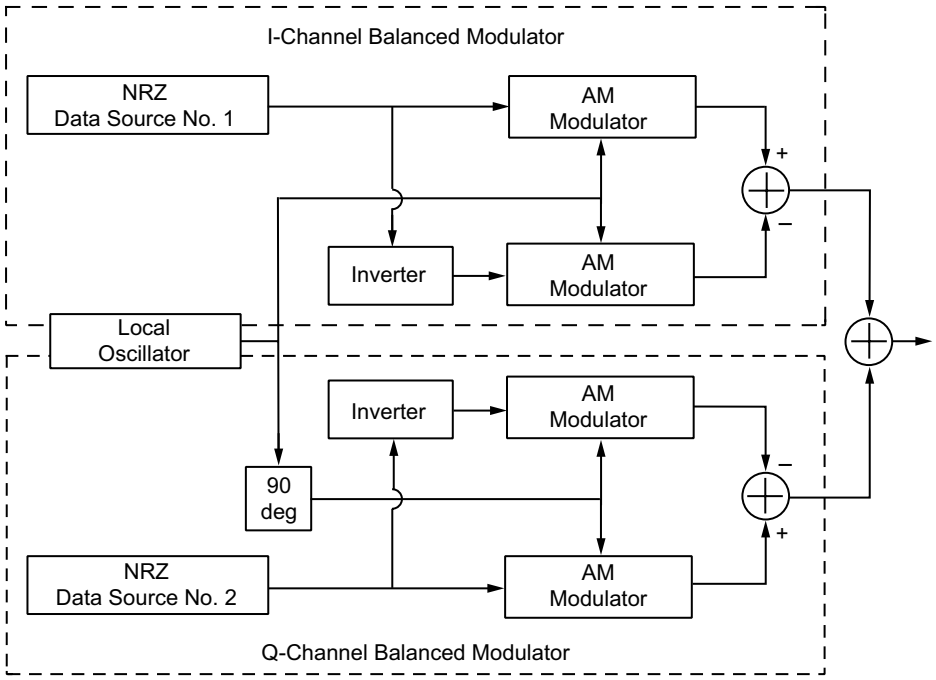


Fig. 2-3. Balanced QPSK modulator implementation.

balanced modulators, respectively. Note that by virtue of the fact that we have introduced separate notation for the I and Q local oscillator phases, i.e., θ_{cI} and θ_{cQ}, we are also allowing for other than a perfect 90-deg phase shift between I and Q channels. Alternatively, the model includes the possibility of an interchannel phase imbalance, $\Delta\theta_c = \theta_{cI} - \theta_{cQ}$. Since we will be interested only in the difference $\Delta\theta_c$, without loss of generality we shall assume $\theta_{cQ} = 0$, in which case $\theta_{cI} = \Delta\theta_c$. Finally, note that if $\Gamma_I = \Gamma_Q = 1$, $\Delta\theta_{cI} = \Delta\theta_{cQ} = 0$, and $\theta_{cI} = \theta_{cQ} = \theta_c$, then we obtain balanced QPSK as characterized by (2.2-5).

As shown in Ref. 8, the transmitted signal of (2.7-1a) and (2.7-1b) can, after some trigonometric manipulation, be written in the form

$$s(t) = \sqrt{P}\Big\{\big[\alpha_I + \beta_I m_I(t) - \gamma_Q\big(1 - m_Q(t)\big)\big]\cos 2\pi f_c t$$

$$+\big[\alpha_Q + \beta_Q m_Q(t) + \delta_I - \gamma_I m_I(t)\big]\sin 2\pi f_c t\Big\} \qquad (2.7\text{-}2)$$

where

$$\left.\begin{aligned}
\alpha_I &= \frac{(1-\Gamma_I\cos\Delta\theta_{cI})\cos\Delta\theta_c+\Gamma_I\sin\Delta\theta_{cI}\sin\Delta\theta_c}{2},\\
\alpha_Q &= \frac{1-\Gamma_Q\cos\Delta\theta_{cQ}}{2}\\
\beta_I &= \frac{(1+\Gamma_I\cos\Delta\theta_{cI})\cos\Delta\theta_c-\Gamma_I\sin\Delta\theta_{cI}\sin\Delta\theta_c}{2},\\
\beta_Q &= \frac{1+\Gamma_Q\cos\Delta\theta_{cQ}}{2}\\
\gamma_I &= \frac{(1+\Gamma_I\cos\Delta\theta_{cI})\sin\Delta\theta_c+\Gamma_I\sin\Delta\theta_{cI}\cos\Delta\theta_c}{2},\\
\gamma_Q &= \frac{\Gamma_Q\sin\Delta\theta_{cQ}}{2}\\
\delta_I &= \frac{-(1-\Gamma_I\cos\Delta\theta_{cI})\sin\Delta\theta_c+\Gamma_I\sin\Delta\theta_{cI}\cos\Delta\theta_c}{2}
\end{aligned}\right\} \tag{2.7-3}$$

The form of the transmitted signal in (2.7-2) clearly identifies the crosstalk introduced by the modulator imbalances, i.e., the dependence of the I channel signal on the Q channel modulation and vice versa, as well as the lack of perfect quadrature between I and Q channels. Note the presence of a spurious carrier component in (2.7-3), i.e., a discrete (unmodulated) carrier component that is not present in the balanced case. Note that for perfect quadrature between the I and Q channels, i.e., $\Delta\theta_c = 0$, we have $\gamma_I = \delta_I = (1/2)\Gamma_I \sin\Delta\theta_{cI}$, and (2.7-2) becomes the symmetric form

$$\begin{aligned}
s(t) = \sqrt{P}\Big\{&\big[\alpha_I+\beta_I m_I(t)-\gamma_Q\big(1-m_Q(t)\big)\big]\cos 2\pi f_c t\\
&+\big[\alpha_Q+\beta_Q m_Q(t)+\gamma_I\big(1-m_I(t)\big)\big]\sin 2\pi f_c t\Big\}
\end{aligned} \tag{2.7-4}$$

which corresponds to the case of modulator imbalance alone. If now the phase imbalance is removed, i.e., $\Delta\theta_{cI} = \Delta\theta_{cQ} = 0$, then $\gamma_I = \gamma_Q = 0$, and the crosstalk in the transmitted signal disappears, i.e., modulator amplitude imbalance alone does not cause crosstalk. It is important to note, however, that the lack of crosstalk in the transmitted signal does not guarantee the absence

of crosstalk at the receiver, which affects the system error probability performance. Finally, note that for the perfectly balanced case, $\beta_I = \beta_Q = 1$ and $\alpha_I = \alpha_I = 0$, $\gamma_I = \gamma_Q = 0$, and (2.7-4) results in (2.2-5) with the exception of the minus sign discussed in Footnote 7.

2.7.1.2 Effect on Carrier Tracking Loop Steady-State Lock Point. When a Costas-type loop is used to track a QPSK signal, it forms its error signal from $IQ\left(I^2 - Q^2\right)$, where the letters I and Q now refer to signals that are synonymous with the outputs of the inphase and quadrature integrate-and-dump (I&D) filters, y_{In} and y_{Qn}, shown in Fig. 2-2(a). In the presence of modulator imbalance and imperfect I and Q quadrature, the evaluation of the steady-state lock point of the loop was considered in Ref. 8 and, in the most general case, was determined numerically. For the special case of identically imbalanced I and Q modulators and no quadrature imperfection, i.e., $\Gamma_I = \Gamma_Q = \Gamma$, $\Delta\theta_{cI} = \Delta\theta_{cQ} = \Delta\theta_u$ and $\Delta\theta_c = 0$, a closed-form result for the steady-state lock point is possible and is given by

$$\phi_0 = -\frac{1}{4}\tan^{-1}\frac{6\Gamma^2\sin 2\Delta\theta_u + \Gamma^4\sin 4\Delta\theta_u}{1 + 6\Gamma^2\cos 2\Delta\theta_u + \Gamma^4\cos 4\Delta\theta_u} \tag{2.7-5}$$

Note that for perfect modulator amplitude balance ($\Gamma = 1$), we obtain $\phi_0 = -\Delta\theta_u/2$, as expected. This shift in the lock point exists independently of the loop SNR and thus can be referred to as an irreducible carrier phase error.

2.7.1.3 Effect on Average BEP. Assuming that the phase error is constant over the bit time (equivalently, the loop bandwidth is small compared to the data rate) and that the 90-deg phase ambiguity associated with the QPSK Costas loop can be perfectly resolved (e.g., by differential encoding), the average BEP can be evaluated by averaging the conditional (on the phase error, ϕ) BEP over the probability density function (PDF) of the phase error, i.e.,

$$\left.\begin{aligned} P_{bI}(E) &= \int_{\phi_0-\pi/4}^{\phi_0-\pi/4} P_{bI}(E;\phi)\, p_\phi(\phi)\, d\phi \\ P_{bQ}(E) &= \int_{\phi_0-\pi/4}^{\phi_0-\pi/4} P_{bQ}(E;\phi)\, p_\phi(\phi)\, d\phi \end{aligned}\right\} \tag{2.7-6}$$

where

$$p_\phi(\phi) = 4\frac{\exp\Big(\rho_{4\phi}\cos\left(4\left(\phi-\phi_0\right)\right)\Big)}{2\pi I_0\left(\rho_{4\phi}\right)}, \quad |\phi-\phi_0| \le \frac{\pi}{4} \tag{2.7-7}$$

is the usual Tikhonov model assumed for the phase error PDF [11] with ϕ_0 determined from (2.7-5). The parameter $\rho_{4\phi}$ is the loop SNR of the four times phase error process (which is what the loop tracks) and $I_0(\cdot)$ is the modified first-order Bessel function of the first kind. Based on the hard decisions made on y_{In} and y_{Qn} in Fig. 2-2(a), the conditional BEPs on the I and Q channels in the presence of imbalance are given, respectively, in Ref. 8, Eqs. (11a) and (11b):

$$\begin{aligned} P_{bI}(E;\phi) = {} & \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos(\phi+\Delta\theta_c)+\sin\phi\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos(\phi+\Delta\theta_c)-\Gamma_Q\sin(\phi+\Delta\theta_{cQ})\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_I\cos(\phi+\Delta\theta_{c_I}+\Delta\theta_c)-\sin\phi\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_I\cos(\phi+\Delta\theta_{c_I}+\Delta\theta_c)+\Gamma_Q\sin(\phi+\Delta\theta_{cQ})\right]\right) \end{aligned} \tag{2.7-8a}$$

and

$$\begin{aligned} P_{bQ}(E;\phi) = {} & \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos\phi-\sin(\phi+\Delta\theta_c)\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos\phi+\Gamma_I\sin(\phi+\Delta\theta_{cI}+\Delta\theta_c)\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_Q\cos(\phi+\Delta\theta_{cQ})+\sin(\phi+\Delta\theta_c)\right]\right) \\ & + \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_Q\cos(\phi+\Delta\theta_{cQ})-\Gamma_I\sin(\phi+\Delta\theta_{cI}+\Delta\theta_c)\right]\right) \end{aligned} \tag{2.7-8b}$$

Substituting (2.7-7) together with (2.7-8a) and (2.7-8b) in (2.7-6) gives the desired average BEP of the I and Q channels for any degree of modulator imbalance. Note that, in general, the error probability performances of the I and Q channels are not identical.

For a maximum amplitude imbalance (Γ_I or Γ_Q) of 0.2 dB, a maximum phase imbalance ($\Delta\theta_{cI}$ or $\Delta\theta_{cQ}$) of +2 deg, and a maximum I-Q quadrature imbalance ($\Delta\theta_c$) of +2 deg (the values recommended by the CCSDS), Figs. 2-4(a) and 2-4(b) plot the I and Q average BEPs as computed from (2.7-6) for the best and worst combinations of imbalance conditions. In these plots, the loop SNR, $\rho_{4\phi}$, is assumed to have infinite value ("perfect" carrier synchronization), and, consequently, the degradation corresponds only to the shift in the lock point. The case of perfectly balanced QPSK is also included in these plots for comparison purposes. We observe that the best imbalance condition gives a performance virtually identical to that of balanced QPSK, whereas the worst imbalance condition results in an E_b/N_0 loss of 0.33 dB at an average BEP of 10^{-2}.

The extension of the above results to the case of OQPSK is presented in Ref. 9. The same modulator imbalance model as that illustrated in Fig. 2-3 is considered, with the exception that the Q channel data stream is now offset with respect to the I channel data stream, requiring a half-symbol delay between the NRZ data source 2 and AM modulator. Also, the amplitude imbalance, Γ, between the I and Q channels, is now *explicitly* included as an additional independent parameter. Therefore, analogous to (2.7-1b), the Q component of the transmitted OQPSK signal becomes [the I component is still given by (2.7-1a)]

$$s_Q(t) = \Gamma\frac{\sqrt{P}}{2}m_Q\left(t-\frac{T_s}{2}\right)\left[\sin\left(2\pi f_c t+\theta_{cQ}\right)+\Gamma_Q\sin\left(2\pi f_c t+\theta_{cQ}+\Delta\theta_{cQ}\right)\right]$$

$$+\Gamma\frac{\sqrt{P}}{2}\left[\sin\left(2\pi f_c t+\theta_{cQ}\right)-\Gamma_Q\sin\left(2\pi f_c t+\theta_{cQ}+\Delta\theta_{cQ}\right)\right] \qquad (2.7\text{-}9)$$

Using similar trigonometric manipulations for arriving at (2.7-2), the transmitted signal $(s_I(t)+s_Q(t))$ can now be written as

$$s(t) = \sqrt{P}\left\{\left[\alpha_I+\beta_I m_I(t)-\gamma_Q\left(1-m_Q\left(t-\frac{T_s}{2}\right)\right)\right]\cos 2\pi f_c t\right.$$

$$\left.+\left[\alpha_Q+\beta_Q m_Q\left(t-\frac{T_s}{2}\right)+\delta_I-\gamma_I m_I(t)\right]\sin 2\pi f_c t\right\} \qquad (2.7\text{-}10)$$

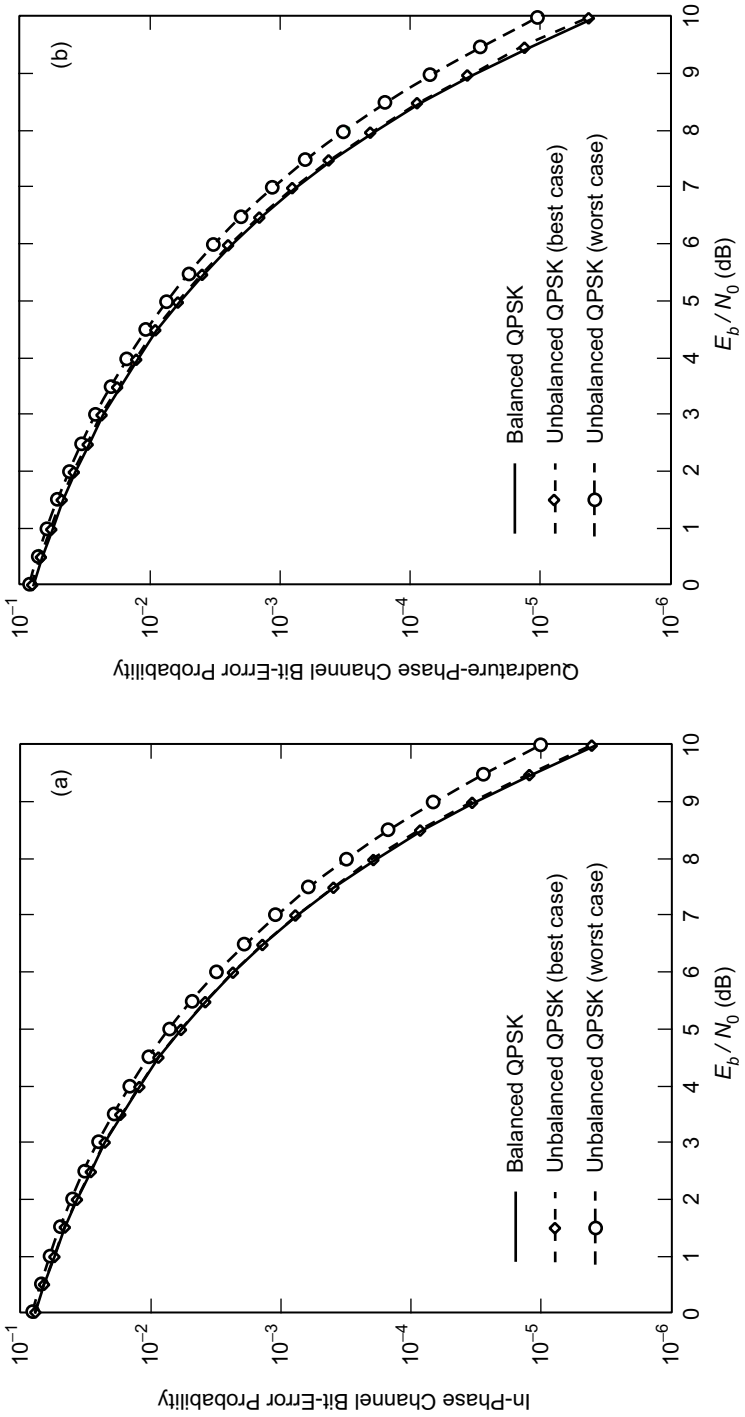


Fig. 2-4. Bit-error performance of imbalanced QPSK signals: (a) in-phase channel and (b) quadrature channel.

where the only changes in the parameters of (2.7-3) are that α_Q, β_Q, and γ_Q are now each multiplied by the I-Q amplitude imbalance parameter, Γ.

The carrier-tracking loop assumed in Ref. 9 is a slightly modified version of that used for QPSK, in which a half-symbol delay is added to its I arm so that the symbols on both arms are aligned in forming the $IQ\left(Q^2 - I^2\right)$ error signal. This loop as well as the optimum (based on maximum a posteriori (MAP) estimation) OQPSK loop, which exhibits only a 180-deg phase ambiguity, are discussed in Ref. 12. The evaluation of the steady-state lock point of the loop was considered in Ref. 9 and was determined numerically. The average BEP is still determined from (2.7-6) (again assuming perfect 90-deg phase ambiguity resolution), but the conditional I and Q BEPs are now specified by

$$
\begin{aligned}
&P_{bI}\left(E;\phi\right) = \\
&\frac{1}{16}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos\left(\phi+\Delta\theta_c\right)+\sin\phi\right]\right) \\
&+\frac{1}{16}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos\left(\phi+\Delta\theta_c\right)-\Gamma\Gamma_Q\sin\left(\phi+\Delta\theta_{cQ}\right)\right]\right) \\
&+\frac{1}{16}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_I\cos\left(\phi+\Delta\theta_{c_I}+\Delta\theta_c\right)-\Gamma\sin\phi\right]\right) \\
&+\frac{1}{16}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_I\cos\left(\phi+\Delta\theta_{c_I}+\Delta\theta_c\right)+\Gamma\Gamma_Q\sin\left(\phi+\Delta\theta_{cQ}\right)\right]\right) \\
&+\frac{1}{8}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\cos\left(\phi+\Delta\theta_c\right)-\frac{\Gamma\Gamma_Q}{2}\sin\left(\phi+\Delta\theta_{cQ}\right)+\frac{\Gamma}{2}\sin\phi\right]\right) \\
&+\frac{1}{8}\operatorname{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma_I\cos\left(\phi+\Delta\theta_{cI}+\Delta\theta_c\right)+\frac{\Gamma\Gamma_Q}{2}\sin\left(\phi+\Delta\theta_{cQ}\right)-\frac{\Gamma}{2}\sin\phi\right]\right)
\end{aligned}
\tag{2.7-11a}
$$

and

$$P_{bQ}\left(E;\phi\right) =$$

$$\frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\cos\phi - \sin\left(\phi+\Delta\theta_c\right)\right]\right)$$

$$+\,\frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\cos\phi + \Gamma_I\sin\left(\phi+\Delta\theta_{cI}+\Delta\theta_c\right)\right]\right)$$

$$+\,\frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\Gamma_Q\cos\left(\phi+\Delta\theta_{cQ}\right) + \sin\left(\phi+\Delta\theta_c\right)\right]\right)$$

$$+\,\frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\Gamma_Q\cos\left(\phi+\Delta\theta_{cQ}\right) - \Gamma_I\sin\left(\phi+\Delta\theta_{cI}+\Delta\theta_c\right)\right]\right)$$

$$+\,\frac{1}{8}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\cos\phi + \frac{\Gamma_I}{2}\sin\left(\phi+\Delta\theta_{cI}+\Delta\theta_c\right) + \frac{1}{2}\sin\left(\phi+\Delta\theta_c\right)\right]\right)$$

$$+\,\frac{1}{8}\,\mathrm{erfc}\left(\sqrt{\frac{E_b}{N_0}}\left[\Gamma\Gamma_Q\cos\left(\phi+\Delta\theta_{cQ}\right) - \frac{\Gamma_I}{2}\sin\left(\phi+\Delta\theta_{cI}+\Delta\theta_c\right)\right.\right.$$

$$\left.\left.+\frac{1}{2}\sin\left(\phi+\Delta\theta_c\right)\right]\right) \tag{2.7-11b}$$

Substituting (2.7-7) together with (2.7-11a) and (2.7-11b) in (2.7-6) gives the desired average BEP of the I and Q channels for any degree of modulator imbalance. Note again that, in general, the error probability performances of the I and Q channels are not identical.

For the same maximum amplitude imbalance, maximum phase imbalance, and maximum I-Q quadrature imbalances as for the QPSK case and in addition an I-Q amplitude imbalance (Γ) of -0.2 dB (corresponding to an actual Q-channel power that is 0.4 dB less than that in the I channel), Figs. 2-5(a) and 2-5(b) plot the I and Q average BEPs as computed from (2.7-6) for the best and worst combinations of imbalance conditions. These results also include the effect of a finite loop SNR of the ϕ process, $\rho_\phi = \rho_{4\phi}/16$, which was chosen equal to 22 dB and held constant along the curves. The case of perfectly balanced QPSK is included in these plots for comparison purposes. The curve labeled

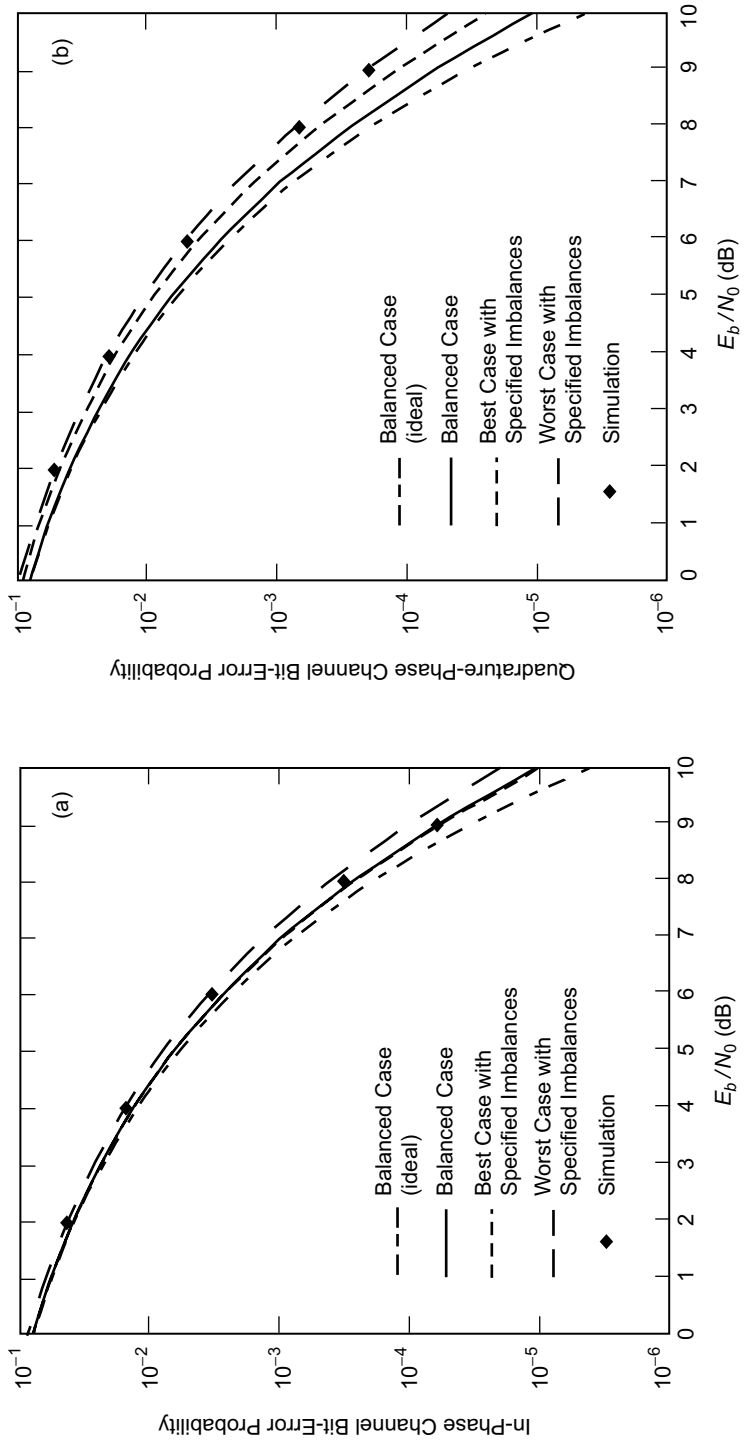


Fig. 2-5. Bit-error performance of OQPSK signals under imperfect carrier synchronization: (a) in-phase channel and (b) quadrature-phase channel.

balanced QPSK (ideal) refers to the case where the loop SNR is assumed infinite, as was the case shown in Figs. 2-4(a) and 2-4(b). Finally, simulation points that agree with the analytical results are also included in Figs. 2-5(a) and 2-5(b). We observe from these figures that the worst imbalance condition results in an E_b/N_0 loss of 0.61 dB for the I channel and 1.08 dB for the Q-channel at an average BEP of 10^{-4}, the larger loss for the Q channel coming as a result of its 0.4-dB power deficiency caused by the I-Q amplitude imbalance. When the I and Q results are averaged, the overall E_b/N_0 degradation becomes 0.86 dB. If perfect carrier synchronization had been assumed, then as shown in Ref. 9, these worst-case losses would be reduced to 0.34 dB for the I channel and 0.75 dB for the Q channel, which translates to a 0.58-dB average performance degradation.

Aside from intrachannel and interchannel amplitude and phase imbalances, the inclusion of a fully saturated RF amplifier modeled by a bandpass hard limiter in the analytical model causes additional degradation in system performance. The performance of OQPSK on such a nonlinear channel was studied in Ref. 10, using the same modulator imbalance model as previously discussed above. The results are summarized as follows.

The transmitter is the same as that illustrated in Fig. 2-3 (with the inclusion of the half-symbol delay in the Q channel as previously discussed), the output of which is now passed through a nonlinear amplifier composed of the cascade of a hard limiter and a bandpass filter (a bandpass hard limiter [13]). The hard limiter clips its input signal at levels $\pm\sqrt{2P_1}\,(\pi/4)$, and the bandpass (zonal) filter removes all the harmonics except for the one at the carrier frequency. The resulting bandpass hard-limited OQPSK signal is a constant envelope signal that has the form

$$\hat{s}\,(t) = \sqrt{2P_1}\cos\left(2\pi f_c t + \theta_d\,(t)\right) \tag{2.7-12}$$

where $P_1 = P\left(\beta_I^2 + \gamma_I^2\right)$ with β_I, γ_I as defined in (2.7-3) and[8]

$$\theta_d\,(t) = \tan^{-1}\frac{\gamma_I}{\beta_I} - \tan^{-1}\left(\frac{Gm_Q\left(t - \dfrac{T_s}{2}\right)\cos\Delta\theta + A\cos\psi}{m_I\,(t) + Gm_Q\left(t - \dfrac{T_s}{2}\right)\sin\Delta\theta + A\sin\psi}\right) \tag{2.7-13}$$

with

[8] The arctangents in (2.7-13) are taken in their principal value sense. Thus, adding π to some of these values is required to place $\theta_d\,(t)$ into its appropriate quadrant.

$$\left.\begin{aligned} G &= \sqrt{\frac{\beta_Q^2+\gamma_Q^2}{\beta_I^2+\gamma_I^2}} \\ A &= \sqrt{\frac{(\alpha_I-\gamma_Q)^2+(\alpha_Q+\delta_I)^2}{\beta_I^2+\gamma_I^2}} \\ \Delta\theta &= \tan^{-1}\frac{\gamma_Q}{\beta_Q}-\tan^{-1}\frac{\gamma_I}{\beta_I} \\ \psi &= \tan^{-1}\frac{\alpha_I-\gamma_Q}{\alpha_Q+\delta_I}-\tan^{-1}\frac{\gamma_I}{\beta_I} \end{aligned}\right\} \tag{2.7-14}$$

Since in any half symbol interval, $m_I(t)$ and $m_Q(t-[T_s/2])$ only take on values ± 1, then in that same interval, $\theta_d(t)$ takes on only one of four equiprobable values, namely, $\theta_{1,1}, \theta_{-1,1}, \theta_{1,-1}, \theta_{-1,-1}$, where the subscripts correspond, respectively, to the values of the above two modulations.

The average BEP is again computed from (2.7-6) together with (2.7-7), where the conditional BEPs are now given by [10, Eqs. (10a) and (10b)]

$$\begin{aligned} P_{bI}(E;\phi) &= \frac{1}{2}\overline{\mathrm{erfc}\left(\sqrt{\frac{2E_b'}{N_0}}\cos\left(\frac{\theta_d^{(1)}-\theta_d^{(2)}}{2}\right)\left|\cos\left(\frac{\theta_d^{(1)}+\theta_d^{(2)}}{2}+\phi\right)\right|\right)} \\ P_{bQ}(E;\phi) &= \frac{1}{2}\overline{\mathrm{erfc}\left(\sqrt{\frac{2E_b'}{N_0}}\cos\left(\frac{\theta_d^{(2)}-\theta_d^{(3)}}{2}\right)\left|\cos\left(\frac{\theta_d^{(2)}+\theta_d^{(3)}}{2}+\phi\right)\right|\right)} \end{aligned} \tag{2.7-15}$$

where $\theta_d^{(j)}$ is the value of the symbol phase $\theta_d(t)$ in the interval $(j-1)T_s/2 \le t \le jT_s/2$, the overbar denotes the statistical average over these symbol phases, and $E_b' = P_1T/2T_s = \left(\beta_I^2+\gamma_I^2\right)PT_s/2 = \left(\beta_I^2+\gamma_I^2\right)E_b$ is the actual I-channel bit energy. Using now the steady-state lock point (irreducible carrier phase error) found numerically in Ref. 10 for this scenario, the average overall and I and Q BEPs are illustrated in Figs. 2-6(a), 2-6(b), and 2-6(c) using parameters identical to those used in arriving at Figs. 2-5(a) and 2-5(b). The final result is that, in the presence of modulator imbalance, the nonlinear amplifier tends to produce a more balanced signal constellation, and thus, the relative BEP performance

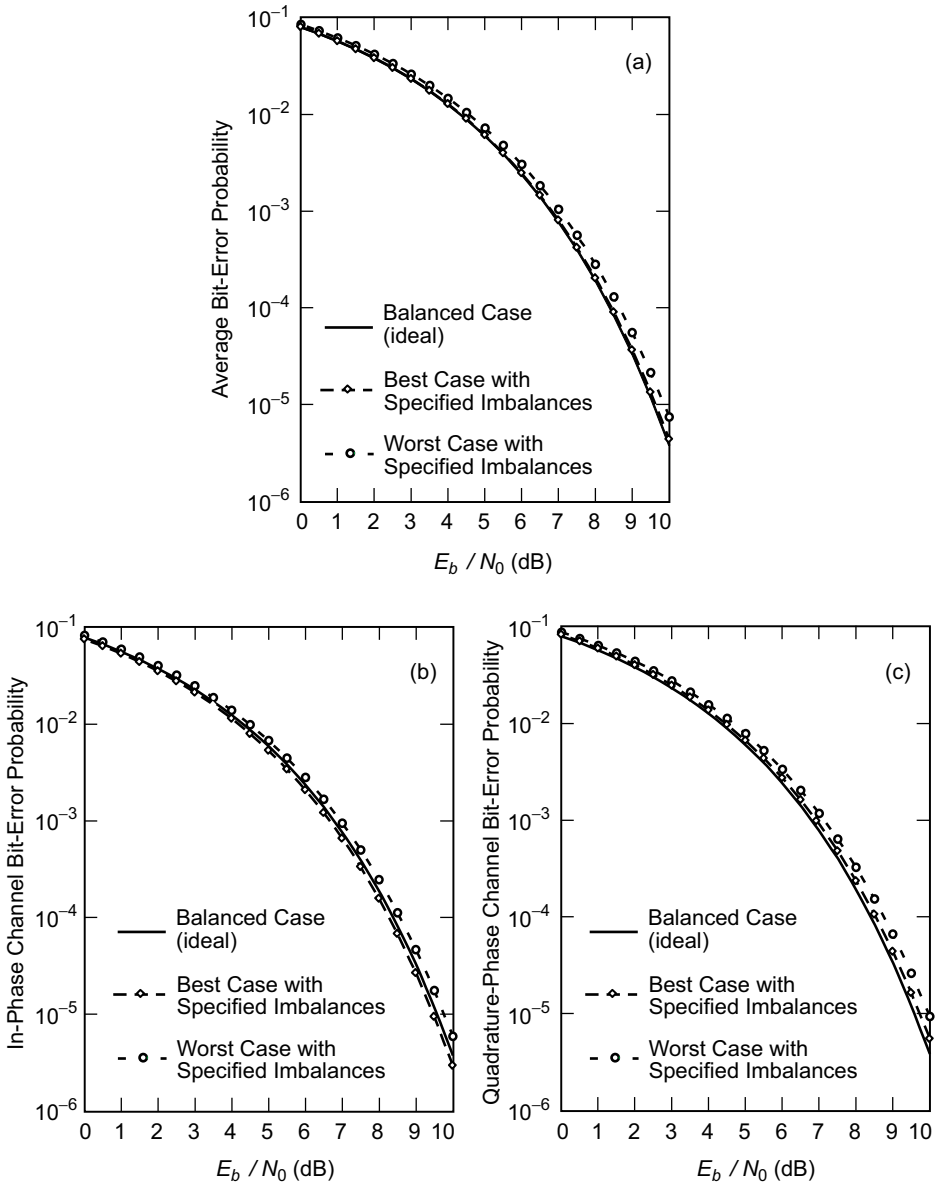


Fig. 2-6. Bit-error performance of nonlinear OQPSK links with imperfect carrier synchronization (i.e., with a carrier-tracking loop SNR fixed at 22 dB): (a) overall channel, (b) in-phase channel, and (c) quadrature-phase channel.

between the I and Q channels is itself more balanced. Furthermore, the average BEPs themselves are much closer to that of a perfectly balanced OQPSK system than those found for the linear channel.

2.7.2 Data Imbalance

The presence of data imbalance (positive and negative bits have different a priori probabilities of occurrence) in the transmitted waveform results in the addition of a discrete spectral component at dc to the continuous PSD component described by (2.5-1). Specifically, if p denotes the probability of a mark (+1), then the total PSD is given by [11, Eq. (1-19)]

$$S(f) = PT_s \left[\frac{1}{T_s} (1-2p)^2 \delta(f) + 4p(1-p) \frac{\sin^2 \pi f T_s}{(\pi f T_s)^2} \right] \tag{2.7-16}$$

Clearly, for the balanced data case, i.e., $p = 1/2$, (2.7-16) reduces to (2.5-1). Since the total power in the transmitted signal is now split between an unmodulated tone at the carrier frequency and a data-bearing component, the carrier tracking process at the receiver (which is designed to act only on the latter) becomes affected even with perfect modulator balance. The degrading effects of a residual carrier on the Costas loop performance for binary PSK are discussed in Ref. 14. The extension to QPSK and OQPSK modulations is straightforward and not pursued here.

Further on in this monograph in our discussion of simulation models and performance, we shall talk about various types of filtered QPSK (which would then no longer be constant envelope). At that time, we shall observe that the combination of data imbalance and filtering produces additional discrete spectral harmonics occurring at integer multiples of the symbol rate.

2.8 Continuous Phase Modulation

Continuing with our discussion of strictly constant envelope modulations, we now turn our attention to the class of schemes referred to as continuous phase frequency modulation (CPFM) or more simply continuous phase modulation (CPM). The properties and performance (bandwidth/power) characteristics of this class of modulations are sufficiently voluminous to fill a textbook of their own [15]. Thus, for the sake of brevity, we shall only investigate certain special cases of CPM that have gained popularity in the literature and have also been put to practice.

CPM schemes are classified as being full response or partial response, depending, respectively, on whether the modulating frequency pulse is of a single bit duration or longer. Within the class of full response CPMs, the subclass of schemes having modulation index 0.5 but arbitrary frequency pulse shape results in a form of generalized MSK [16].[9] Included as popular special cases are MSK, originally invented by Doelz and Heald, as disclosed in a 1961 U.S. patent [19], having a rectangular frequency pulse shape, and Amoroso's sinusoidal frequency-shift-keying (SFSK) [20], possessing a sinusoidal (raised cosine) frequency pulse shape. The subclass of full-response schemes with rectangular frequency pulse but arbitrary modulation index is referred to as continuous phase frequency-shift-keying (CPFSK) [21], which, for all practical purposes, served as the precursor to what later became known as CPM itself. Within the class of partial-response CPMs, undoubtedly the most popular scheme is that of Gaussian minimum-shift-keying (GMSK) which, because of its excellent bandwidth efficiency, has been adopted as a European standard for personal communication systems (PCSs). In simple terms, GMSK is a partial-response CPM scheme obtained by filtering the rectangular frequency pulses characteristic of MSK with a filter having a Gaussian impulse response prior to frequency modulation of the carrier.

In view of the above considerations, in what follows, we shall focus our CPM discussion only on MSK, SFSK, and GMSK, in each case presenting results for their spectral and power efficiency behaviors. Various representations of the transmitter, including the all-important equivalent I-Q one, will be discussed as well as receiver performance, both for ideal and nonideal (modulator imbalance) conditions.

2.8.1 Full Response—MSK and SFSK

While the primary intent of this section of the monograph is to focus specifically on the properties and performance of MSK and SFSK in the form they are most commonly known, the reader should bear in mind that many of these very same characteristics, e.g., transmitter/receiver implementations, equivalent I-Q signal representations, spectral and error probability analysis, apply equally well to generalized MSK. Whenever convenient, we shall draw attention to these analogies so as to alert the reader to the generality of our discussions. We begin the mathematical treatment by portraying MSK as a special case of the more general CPM signal, whose characterization is given in the next section.

[9] Several other authors [17,18] coined the phrase "generalized MSK" to represent generalizations of MSK other than by pulse shaping.

2.8.1.1 Continuous Phase Frequency Modulation Representation. A binary single-mode (one modulation index for all transmission intervals) CPM signal is a constant envelope waveform that has the generic form (see the implementation in Fig. 2-7)

$$s(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \phi\left(t, \boldsymbol{\alpha}\right) + \phi_0\right), \quad nT_b \le t \le (n+1)T_b \tag{2.8-1}$$

where, as before, E_b and T_b respectively denote the energy and duration of a bit ($P = E_b/T_b$ is the signal power), and f_c is the carrier frequency. In addition, $\phi\left(t, \boldsymbol{\alpha}\right)$ is the phase modulation process that is expressable in the form

$$\phi\left(t, \boldsymbol{\alpha}\right) = 2\pi \sum_{i \le n} \alpha_i h q\left(t - iT_b\right) \tag{2.8-2}$$

where $\boldsymbol{\alpha} = (\cdots, \alpha_{-2}, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \cdots)$ is an independent, identically distributed (i.i.d.) binary data sequence, with each element taking on equiprobable values ± 1, $h = 2\Delta f T_b$ is the modulation index (Δf is the peak frequency deviation of the carrier), and $q(t)$ is the normalized phase-smoothing response that defines how the underlying phase, $2\pi\alpha_i h$, evolves with time during the associated bit interval. Without loss of generality, the arbitrary phase constant, ϕ_0, can be set to zero.

For our discussion here it is convenient to identify the derivative of $q(t)$, namely,

$$g(t) = \frac{dq(t)}{dt} \tag{2.8-3}$$

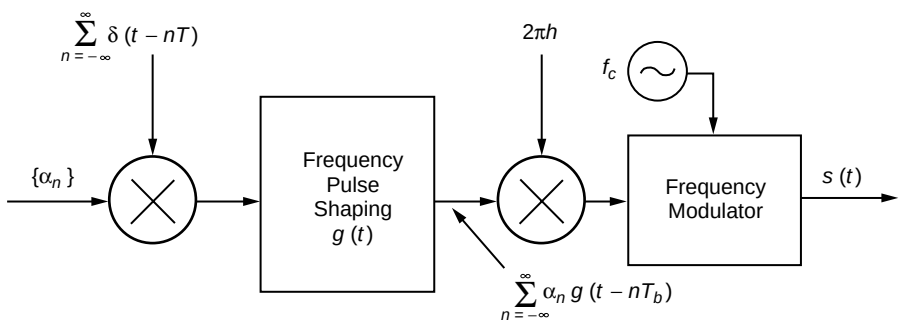


Fig. 2-7. CPM transmitter.

which represents the instantaneous frequency pulse (relative to the nominal carrier frequency, f_c) in the zeroth signaling interval. In view of (2.8-3), the phase smoothing response is given by

$$q(t) = \int_{-\infty}^{t} g(\tau)\, d\tau \tag{2.8-4}$$

which, in general, extends over infinite time. For full response CPM schemes, as will be the case of interest here, $q(t)$ satisfies the following:

$$q(t) = \begin{cases} 0, & t \leq 0 \\ \dfrac{1}{2}, & t \geq T_b \end{cases} \tag{2.8-5}$$

and, thus, the frequency pulse, $g(t)$, is nonzero only over the bit interval, $0 \leq t \leq T_b$. In view of (2.8-5), we see that the ith data symbol, α_i, contributes a phase change of $\pi\alpha_i h$ rad to the total phase for all time after T_b seconds of its introduction, and, therefore, this fixed phase contribution extends over all future symbol intervals. Because of this overlap of the phase smoothing responses, the total phase in any signaling interval is a function of the present data symbol as well as all of the past symbols, and accounts for the memory associated with this form of modulation. Consequently, in general, optimum detection of CPM schemes must be performed by a maximum-likelihood sequence estimator (MLSE) form of receiver [1] as opposed to bit-by-bit detection, which is optimum for memoryless modulations such as conventional binary FSK with discontinuous phase.

As previously mentioned, MSK is a full-response CPM scheme with a modulation index $h = 0.5$ and a rectangular frequency pulse mathematically described by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}, & 0 \leq t \leq T_b \\ 0, & \text{otherwise} \end{cases} \tag{2.8-6}$$

For SFSK, one of the generalized MSK schemes mentioned in the introduction, $g(t)$, would be a raised cosine pulse given by

$$g(t) = \begin{cases} \dfrac{1}{2T_b}\left[1 - \cos\left(\dfrac{2\pi t}{T_b}\right)\right], & 0 \leq t \leq T_b \\ 0, & \text{otherwise} \end{cases} \tag{2.8-7}$$

The associated phase pulses defined by (2.8-4) are

$$q(t) = \begin{cases} \dfrac{t}{2T_b}, & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{2.8-8}$$

for MSK and

$$q(t) = \begin{cases} \dfrac{1}{2T_b}\left[t - \dfrac{\sin 2\pi t/T_b}{2\pi/T_b}\right], & 0 \le t \le T_b \\ \dfrac{1}{2}, & t \ge T_b \end{cases} \tag{2.8-9}$$

for SFSK.

Finally, substituting $h = 0.5$ and $g(t)$ of (2.8-6) in (2.8-1) combined with (2.8-2) gives the CPM representations of MSK and SFSK, respectively, as

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b}\sum_{i \le n} \alpha_i \left(t - iT_b\right)\right), \quad nT_b \le t \le (n+1)T_b \tag{2.8-10}$$

and

$$s_{\mathrm{SFSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \cos\left(2\pi f_c t + \frac{\pi}{2T_b}\sum_{i \le n} \alpha_i \left[t - iT_b - \frac{\sin 2\pi\left(t - iT_b\right)/T_b}{2\pi/T_b}\right]\right),$$

$$nT_b \le t \le (n+1)T_b \tag{2.8-11}$$

both of which are implemented as in Fig. 2-7, using $g(t)$ of (2.8-6) or (2.8-7) as appropriate.

Associated with MSK (or SFSK) is a phase trellis that illustrates the evolution of the phase process with time, corresponding to all possible transmitted sequences. For MSK, the phase variation with time is linear [see (2.8-10)], and, thus, paths in the phase trellis are straight lines with a slope of $\pm\pi/2T_b$. Figure 2-8 is an illustration of the MSK phase trellis where the branches are labeled with the data bits that produce the corresponding phase transition. Note that

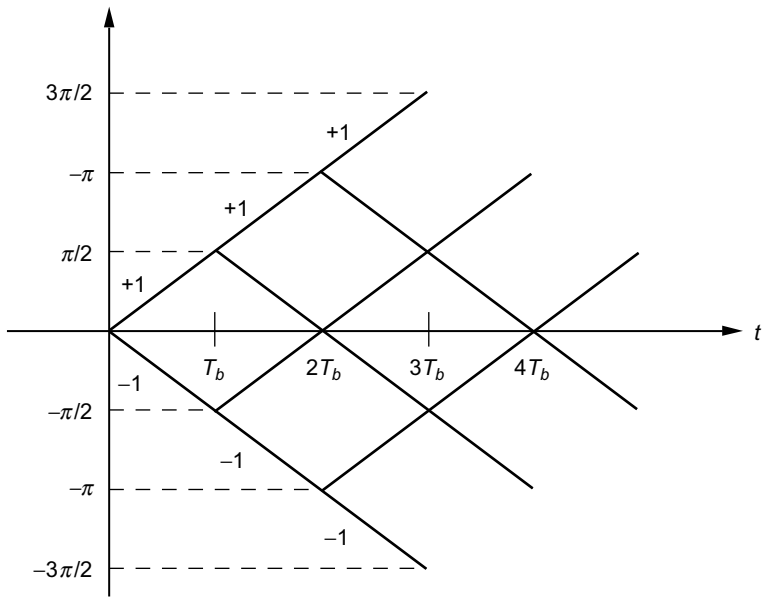


Fig. 2-8. Phase trellis (time-varying) for conventional MSK. Phase states (mod 2π) are $(0, \pi)$ for n even and $(\pi/2, 3\pi/2)$ for n odd.

the change in phase over a single bit time is either $\pi/2$ or $-\pi/2$, depending on the polarity of the data bit, α_i, corresponding to that bit time. Also note that the trellis is time-varying in that the phase states (modulo 2π) alternate between 0 and π at even multiples of the bit time and $\pi/2$ and $3\pi/2$ at odd multiples of the bit time. For SFSK, the phase trellis would appear as in Fig. 2-8 with, however, a sinusoidal variation in phase superimposed over the straight line paths. Here again the change in phase over a single bit time would be either $\pi/2$ or $-\pi/2$, depending on the polarity of the data bit, α_i, corresponding to that bit time.

2.8.1.2 Equivalent I-Q Representation of MSK. Although, as stated above, CPM schemes, because of their inherent memory, require a memory-type of detection, e.g., MLSE, full-response modulations with $h = 0.5$ such as MSK and SFSK can in fact be detected using a memoryless I-Q form of receiver. The reason for this is that for these modulations the transmitter can be implemented in an I-Q form analogous to that of OQPSK. To see this mathematically, we first rewrite the excess phase in the nth transmission interval of the MSK signal in (2.8-10) as

$$\phi(t,\boldsymbol{\alpha}) = \frac{\pi}{2T_b}\sum_{i\le n}\alpha_i\,(t-iT_b) = \alpha_n\frac{\pi}{2T_b}\,(t-nT_b) + \frac{\pi}{2}\sum_{i\le n-1}\alpha_i = \alpha_n\frac{\pi}{2T_b}t + x_n,$$

$$nT_b \le t \le (n+1)T_b \qquad (2.8\text{-}12)$$

where $(\pi/2)\sum_{i\le n-1}\alpha_i$ is the accumulated phase at the beginning of the nth transmission interval that is equal to an odd integer (positive or negative) multiple of $\pi/2$ when n is odd and an even integer (positive or negative) multiple of $\pi/2$ when n is even, and x_n is a phase constant required to keep the phase continuous at the data transition points $t = nT_b$ and $t = (n+1)\,T_b$. Note also that x_n represents the y-intercept (when reduced modulo 2π) of the path in the phase trellis that represents $\phi(t,\boldsymbol{\alpha})$. In the previous transmission interval, the excess phase is given by

$$\phi(t,\boldsymbol{\alpha}) = \alpha_n\frac{\pi}{2T_b}\big(t-(n-1)\,T_b\big) + \frac{\pi}{2}\sum_{i\le n-2}\alpha_i = \alpha_{n-1}\frac{\pi}{2T_b}t + x_{n-1},$$

$$(n-1)\,T_b \le t \le nT_b \qquad (2.8\text{-}13)$$

For phase continuity at $t = nT_b$, we require that

$$\alpha_n\frac{\pi}{2T_b}\,(nT_b) + x_n = \alpha_{n-1}\frac{\pi}{2T_b}\,(nT_b) + x_{n-1} \qquad (2.8\text{-}14)$$

or equivalently

$$x_n = x_{n-1} + \frac{\pi n}{2}\,(\alpha_{n-1}-\alpha_n) \qquad (2.8\text{-}15)$$

Equation (2.8-15) is a recursive relation that allows x_n to be determined in any transmission interval given an initial condition, x_0.

We observe that $(\alpha_{n-1}-\alpha_n)\,/2$ is a ternary random variable (RV) taking on values $0, +1, -1$, with probabilities $1/2, 1/4, 1/4$, respectively. Therefore, from (2.8-15), when $\alpha_{n-1} = \alpha_n$, $x_n = x_{n-1}$, whereas when $\alpha_{n-1} \ne \alpha_n$, $x_n = x_{n-1} \pm \pi n$. If we arbitrarily choose the initial condition $x_0 = 0$, then we see that x_n takes on values of 0 or π (when reduced modulo 2π). Using this fact in (2.8-12) and applying simple trigonometry to (2.8-10), we obtain

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\cos\phi\left(t,\alpha\right)\cos 2\pi f_c t - \sin\phi\left(t,\alpha\right)\sin 2\pi f_c t\right],$$

$$nT_b \le t \le (n+1)T_b \qquad (2.8\text{-}16)$$

where

$$\left.\begin{aligned}
\cos\phi\left(t,\alpha\right) &= \cos\left(\alpha_n \frac{\pi}{2T_b}t + x_n\right) = a_n \cos\frac{\pi}{2T_b}t, \quad a_n = \cos x_n = \pm 1 \\
\sin\phi\left(t,\alpha\right) &= \sin\left(\alpha_n \frac{\pi}{2T_b}t + x_n\right) = \alpha_n a_n \sin\frac{\pi}{2T_b}t = b_n \sin\frac{\pi}{2T_b}t, \\
& \qquad\qquad\qquad\qquad\qquad\qquad b_n = \alpha_n \cos x_n = \pm 1
\end{aligned}\right\} \qquad (2.8\text{-}17)$$

Finally, substituting (2.8-17) in (2.8-16) gives the I-Q representation of MSK as

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[a_n C\left(t\right)\cos 2\pi f_c t - b_n S\left(t\right)\sin 2\pi f_c t\right], \quad nT_b \le t \le (n+1)T_b \qquad (2.8\text{-}18)$$

where

$$\left.\begin{aligned}
C\left(t\right) &= \cos\frac{\pi t}{2T_b} \\
S\left(t\right) &= \sin\frac{\pi t}{2T_b}
\end{aligned}\right\} \qquad (2.8\text{-}19)$$

are the effective I and Q pulse shapes, and $\{a_n\}, \{b_n\}$, as defined in (2.8-17), are the effective I and Q binary data sequences.

For SFSK, the representation of (2.8-18) would still be valid with a_n, b_n as defined in (2.8-17), but now the effective I and Q pulse shapes become

$$\left.\begin{aligned}
C\left(t\right) &= \cos\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right] \\
S\left(t\right) &= \sin\left[\frac{\pi}{2T_b}\left(t - \frac{\sin 2\pi t/T_b}{2\pi/T_b}\right)\right]
\end{aligned}\right\} \qquad (2.8\text{-}20)$$

To tie the representation of (2.8-18) back to that of FSK, we observe that

$$\left.\begin{aligned} C(t)\cos 2\pi f_c t &= \frac{1}{2}\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] \\ &\quad + \frac{1}{2}\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right] \\ S(t)\sin 2\pi f_c t &= -\frac{1}{2}\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] \\ &\quad + \frac{1}{2}\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right] \end{aligned}\right\} \tag{2.8-21}$$

Substituting (2.8-21) in (2.8-18) gives

$$\begin{aligned} s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\Bigg[&\left(\frac{a_n+b_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] \\ &+\left(\frac{a_n-b_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\Bigg], \quad nT_b \le t \le (n+1)T_b \end{aligned} \tag{2.8-22}$$

Thus, when $a_n = b_n$ ($\alpha_n = 1$), we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right] \tag{2.8-23}$$

whereas when $a_n \neq b_n$ ($\alpha_n = -1$) we have

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right] \tag{2.8-24}$$

which establishes the desired connection.

Note from (2.8-19), that since $C(t)$ and $S(t)$ are offset from each other by a time shift of T_b seconds, it might appear that $s_{\mathrm{MSK}}(t)$ of (2.8-18) is in the form of OQPSK with half-sinusoidal pulse shaping.[10] To justify that this is indeed the case, we must examine more carefully the effective I and Q data sequences $\{a_n\}, \{b_n\}$ in so far as their relationship to the input data sequence $\{\alpha_i\}$ and the rate at which they can change. Since the input α_n data bit can change every bit time, it might appear that the effective I and Q data bits, a_n and b_n, can also change every bit time. To the contrary, it can be shown that as a result of the phase continuity constraint of (2.8-15), $a_n = \cos x_n$ can change only at the zero crossings of $C(t)$, whereas $b_n = \alpha_n \cos x_n$ can change only at the zero crossings of $S(t)$. Since the zero crossings of $C(t)$ and $S(t)$ are each spaced $2T_b$ seconds apart, then a_n and b_n are constant over $2T_b$-second intervals (see Fig. 2-9 for an illustrative example). Further noting that the continuous waveforms $C(t)$ and $S(t)$ alternate in sign every $2T_b$ seconds, we can incorporate this sign change into the I and Q data sequences themselves and deal with a fixed, positive, time-limited pulse shape on each of the I and Q channels. Specifically, defining the pulse shape

$$p(t) = \begin{cases} \sin \dfrac{\pi t}{2T_b}, & 0 \le t \le 2T_b \\ 0, & \text{otherwise} \end{cases} \tag{2.8-25}$$

then the I-Q representation of MSK can be rewritten in the form

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[d_c(t)\cos 2\pi f_c t - d_s(t)\sin 2\pi f_c t\right] \tag{2.8-26}$$

where

$$\left.\begin{aligned} d_c(t) &= \sum_n c_n p\big(t - (2n-1)T_b\big) \\ d_s(t) &= \sum_n d_n p\left(t - 2nT_b\right) \end{aligned}\right\} \tag{2.8-27}$$

with

[10] A similar statement can be made for SFSK, where the pulse shaping is now described by (2.8-20).

n	α_n	x_n (mod 2π)	a_n	b_n	Time Interval
0	1	0	1	1	$0 \le t \le T_b$
1	–1	π	–1	1	$T_b \le t \le 2T_b$
2	–1	π	–1	1	$2T_b \le t \le 3T_b$
3	1	0	1	1	$3T_b \le t \le 4T_b$
4	1	0	1	1	$4T_b \le t \le 5T_b$
5	1	0	1	1	$5T_b \le t \le 6T_b$
6	–1	0	1	–1	$6T_b \le t \le 7T_b$
7	1	π	–1	–1	$7T_b \le t \le 8T_b$
8	–1	π	–1	1	$8T_b \le t \le 9T_b$

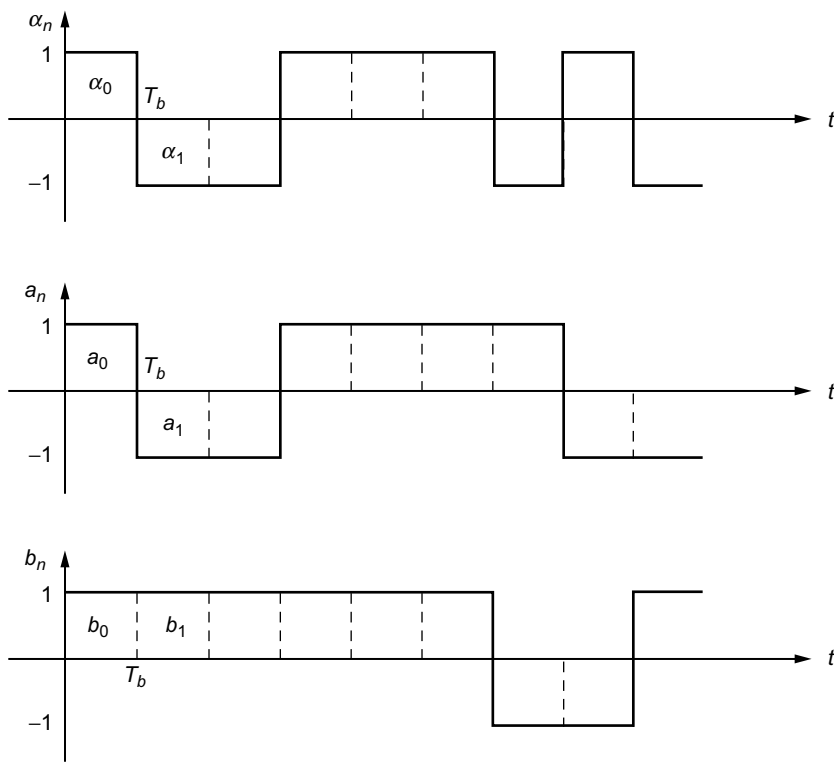


Fig. 2-9. An example of the equivalent I and Q data sequences represented as rectangular pulse streams. Redrawn from [1].

$$\left.\begin{aligned} c_n &= (-1)^n a_{2n-1} \\ d_n &= (-1)^n b_{2n} \end{aligned}\right\} \tag{2.8-28}$$

To complete the analogy between MSK and sinusoidally pulse shaped OQPSK, we must examine the manner in which the equivalent I and Q data sequences needed in (2.8-28) are obtained from the input data sequence $\{\alpha_n\}$. Without going into great mathematical detail, we can say that it can be shown that the sequences $\{a_{2n-1}\}$ and $\{b_{2n}\}$ are the odd/even split of a sequence, $\{v_n\}$, which is the differentially encoded version of $\{\alpha_n\}$, i.e., $v_n = \alpha_n v_{n-1}$ (see Fig. 2-10 for an illustrative example). Finally, the I-Q implementation of MSK as described by (2.8-26)–(2.8-28) is illustrated in Fig. 2-11. As anticipated, we observe that this figure resembles a transmitter for OQPSK except that here, the pulse shaping is half-sinusoidal (of symbol duration $T_s = 2T_b$) rather than rectangular; in addition, we see that a differential encoder is applied to the input data sequence prior to splitting it into even and odd sequences, each at a rate $1/T_s$. The interpretation of MSK as a special case of OQPSK with sinusoidal pulse shaping along with trade-offs and comparisons between the two modulations is further discussed in Refs. 22 and 23.

Before concluding this section, we note that the alternative representation of MSK as in (2.8-22) can be also expressed in terms of the differentially encoded bits, v_n. In particular,

For n odd

$$\begin{aligned} s_{\text{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \Bigg[&\left(\frac{v_{n-1}+v_n}{2}\right) \cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \\ &- \left(\frac{v_{n-1}-v_n}{2}\right) \cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]\Bigg], \\ &nT_b \le t \le (n+1)T_b \end{aligned} \tag{2.8-29a}$$

For n even

$$\begin{aligned} s_{\text{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \Bigg[&\left(\frac{v_{n-1}+v_n}{2}\right) \cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \\ &+ \left(\frac{v_{n-1}-v_n}{2}\right) \cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right]\Bigg], \\ &nT_b \le t \le (n+1)T_b \end{aligned} \tag{2.8-29b}$$

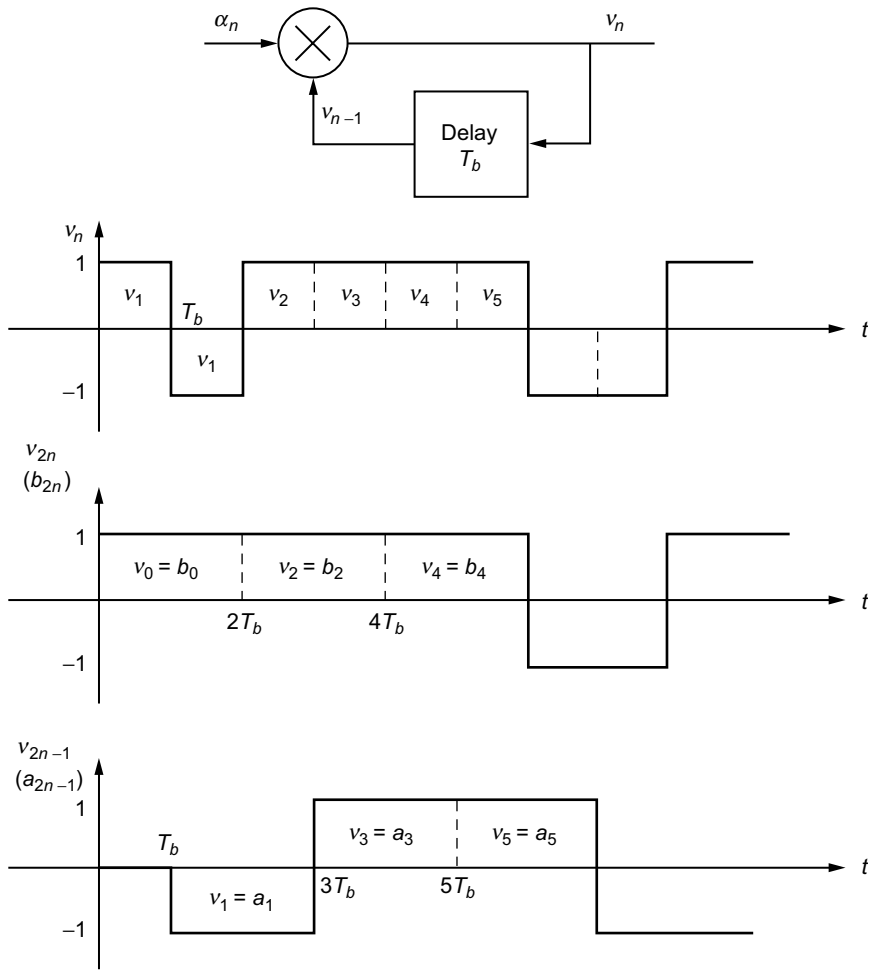


Fig. 2-10. An example of the equivalence between differentially encoded inputs bits and effective I and Q bits. Redrawn from [1].

Combining these two results we get

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\left(\frac{v_{n-1}+v_n}{2}\right)\cos\left[2\pi\left(f_c+\frac{1}{4T_b}\right)t\right]\right.$$

$$\left.+(-1)^n\left(\frac{v_{n-1}-v_n}{2}\right)\cos\left[2\pi\left(f_c-\frac{1}{4T_b}\right)t\right]\right],$$

$$nT_b \le t \le (n+1)T_b \qquad (2.8\text{-}30)$$

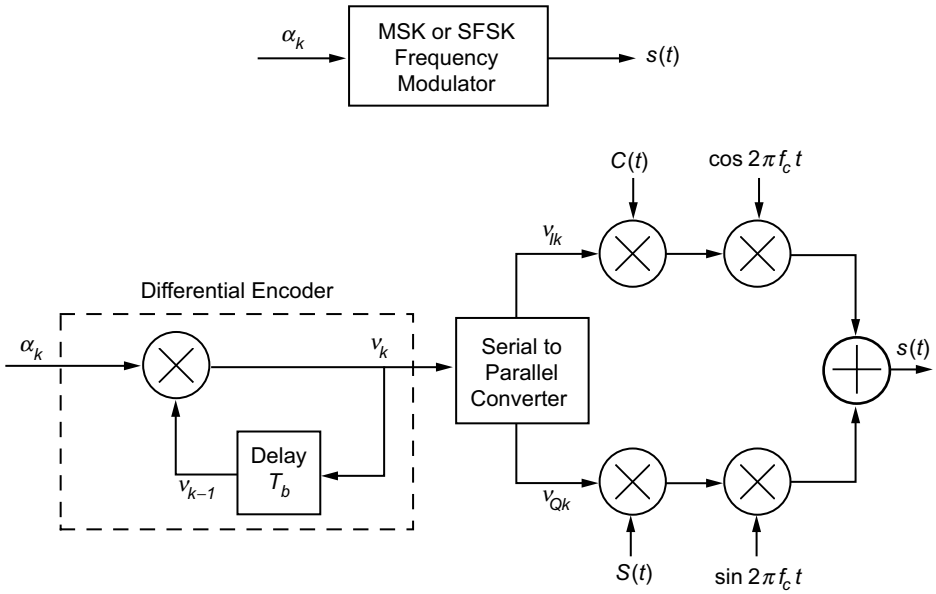


Fig. 2-11. CPM and equivalent I-Q implementations of MSK or SFSK.

2.8.1.3 Precoded MSK. The differential encoder that precedes the I-Q portion of the transmitter in Fig. 2-11 requires a compensating differential decoder at the receiver following I-Q demodulation and detection (see Fig. 2-12). Such a combination of differential encoding at the transmitter and differential decoding at the receiver results in a loss in power performance relative to that obtained by conventional OQPSK (this will be discussed in more detail later on in the chapter). It is possible to modify MSK to avoid such a loss by first recognizing that the CPM form of modulator in Fig. 2-7 for implementing MSK can be preceded by the cascade of a differential encoder and a differential decoder without affecting its output (Fig. 2-13). That is, the cascade of a differential encoder and a differential decoder produces unity transmission—input = output. Thus, comparing Fig. 2-13 with Fig. 2-11, we observe that precoding the CPM form of MSK modulator with a differential decoder, resulting in what is referred to as precoded MSK [1, Chap. 10] will be equivalent to the I-Q implementation of the latter without the differential encoder at its input (see Fig. 2-14), and thus the receiver for precoded MSK is that of Fig. 2-12 without the differential decoder at its output. A similar precoding applied to SFSK would also allow for dispensing with the differential decoder at the output of its I-Q receiver. Finally, we note that both MSK (or SFSK) and its precoded version have identical spectral characteristics and, consequently, for all practical purposes, the improvement in power performance provided by the latter comes at no expense.

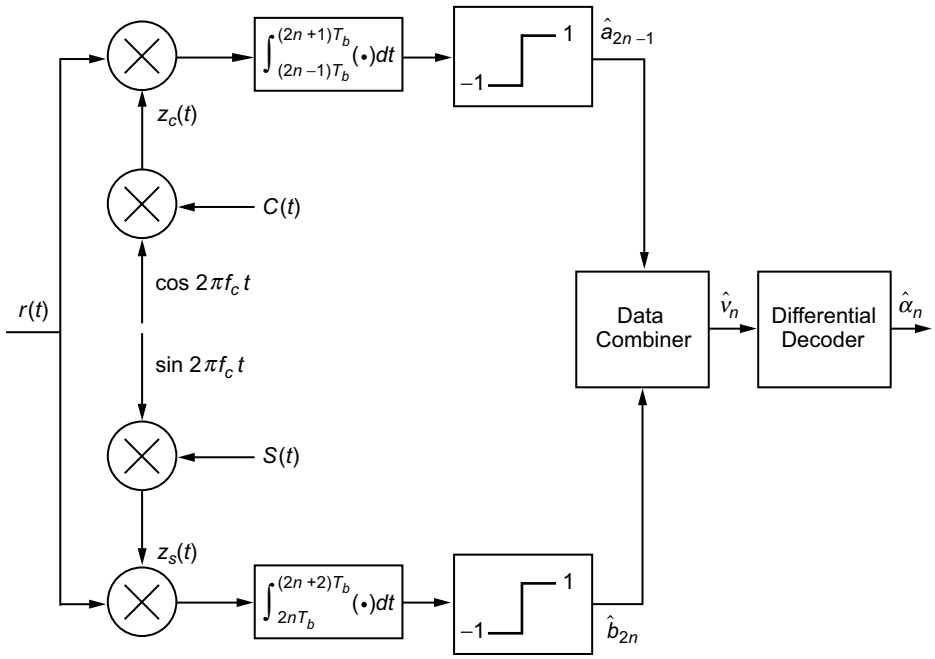


Fig. 2-12. An I-Q receiver implementation of MSK.

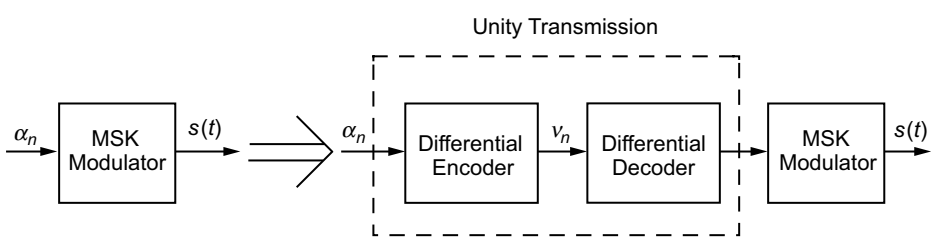


Fig. 2-13. Two equivalent MSK transmitters.

2.8.1.4 Spectral Characteristics. The ability to express MSK in the offset I-Q form of (2.8-18) allows for simple evaluation of its PSD. In particular, for a generic offset I-Q modulation formed by impressing two lowpass modulations (random pulse trains of rate $1/2T_b$) of equal power and pulse shape on inphase and quadrature carriers, i.e.,

$$s(t) = Am_I(t)\cos 2\pi f_c t - Am_Q(t)\sin 2\pi f_c t,$$

$$m_I(t) = \sum_n a_n p(t - 2nT_b), \quad m_Q(t) = \sum_n b_n p\big(t - (2n-1)T_b\big) \quad (2.8\text{-}31)$$

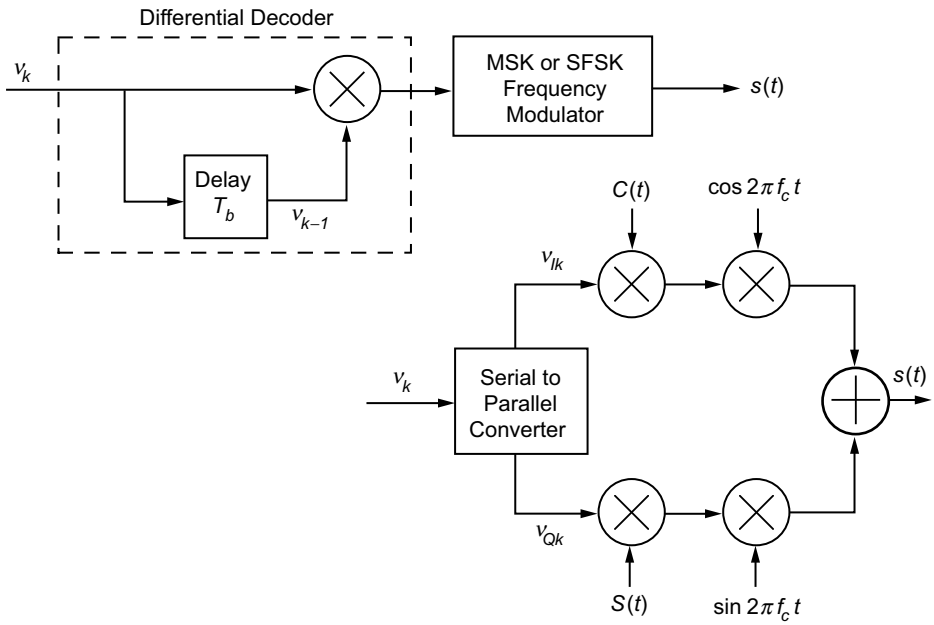


Fig. 2-14. CPM and equivalent I-Q implementations of precoded MSK or SFSK.

the PSD is given by [1, Chap. 2]

$$S_s(f) = \frac{1}{4}\left[G(f - f_c) + G(f + f_c)\right] \tag{2.8-32}$$

where $G(f)$ is the equivalent baseband PSD and is related to the PSD, $S_m(f)$, of $m_I(t)$ or $m_Q(t)$ by

$$G(f) = 2A^2 S_m(f), \quad S_m(f) = \frac{1}{2T_b}|P(f)|^2 \tag{2.8-33}$$

with $P(f)$ denoting the Fourier transform of the pulse shape $p(t)$. For MSK, we would have $A = \sqrt{2E_b/T_b}$ and $p(t)$ given by (2.8-25) with Fourier transform

$$P(f) = \frac{4T_b}{\pi} e^{-j2\pi f T_b} \frac{\cos 2\pi f T_b}{1 - 16 f^2 T_b^2} \tag{2.8-34}$$

Substituting (2.8-34) in (2.8-33) gives the equivalent baseband PSD of MSK as

$$G(f) = \frac{32E_b}{\pi^2} \frac{\cos^2 2\pi f T_b}{\left(1 - 16 f^2 T_b^2\right)^2} \tag{2.8-35}$$

and the corresponding bandpass PSD as [1, Chap. 2]

$$S_s(f) = \frac{8E_b}{\pi^2} \left[\frac{\cos^2 2\pi (f - f_c) T_b}{\left(1 - 16 (f - f_c)^2 T_b^2\right)^2} + \frac{\cos^2 2\pi (f + f_c) T_b}{\left(1 - 16 (f + f_c)^2 T_b^2\right)^2} \right] \tag{2.8-36}$$

We observe from (2.8-35) that the main lobe of the lowpass PSD has its first null at $f = 3/4T_b$. Also, asymptotically for large f, the spectral sidelobes roll off at a rate f^{-4}. By comparison, the equivalent PSD of OQPSK wherein $A = \sqrt{E_b/T_b}$ and $p(t)$ is a unit amplitude rectangular pulse of duration $2T_b$, is given by

$$G(f) = 4E_b \frac{\sin^2 2\pi f T_b}{\left(2\pi f T_b\right)^2} \tag{2.8-37}$$

whose main lobe has its first null at $f = 1/2T_b$ and whose spectral sidelobes asymptotically roll off at a rate f^{-2}. Thus, we observe that while MSK (or precoded MSK) has a wider main lobe than OQPSK(or QPSK) by a factor of 3/2, its spectral sidelobes roll off at a rate two orders of magnitude faster. Figure 2-15 is an illustration of the normalized lowpass PSDs, $G(f)/2E_b$, of MSK and OQPSK obtained from (2.8-35) and (2.8-37), respectively, as well as that of SFSK, which is given by [1, Chap. 2]

$$G(f) =$$

$$2E_b \left[J_0\left(\frac{1}{4}\right) A_0(f) + 2\sum_{n=1}^{\infty} J_{2n}\left(\frac{1}{4}\right) B_{2n}(f) + 2\sum_{n=1}^{\infty} J_{2n-1}\left(\frac{1}{4}\right) B_{2n-1}(f) \right]^2,$$

$$A(f) = 2\frac{\sin 2\pi f T_b}{2\pi f T_b},$$

$$A_0(f) = \frac{1}{2} A\left(f + \frac{1}{4T_b}\right) + \frac{1}{2} A\left(f - \frac{1}{4T_b}\right) = \frac{4}{\pi} \frac{\cos 2\pi f T_b}{1 - 16 f^2 T_b^2},$$

$$A_{2n}(f) = \frac{1}{2}A\left(f + \frac{2n}{T_b}\right) + \frac{1}{2}A\left(f - \frac{2n}{T_b}\right),$$

$$A_{2n-1}(f) = \frac{1}{2}A\left(f + \frac{2n-1}{T_b}\right) - \frac{1}{2}A\left(f - \frac{2n-1}{T_b}\right),$$

$$B_{2n}(f) = \frac{1}{2}A_{2n}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n}\left(f - \frac{1}{4T_b}\right),$$

$$B_{2n-1}(f) = -\frac{1}{2}A_{2n-1}\left(f + \frac{1}{4T_b}\right) + \frac{1}{2}A_{2n-1}\left(f - \frac{1}{4T_b}\right),$$

$$J_n(x) = n\text{th order Bessel function of the first kind} \tag{2.8-38}$$

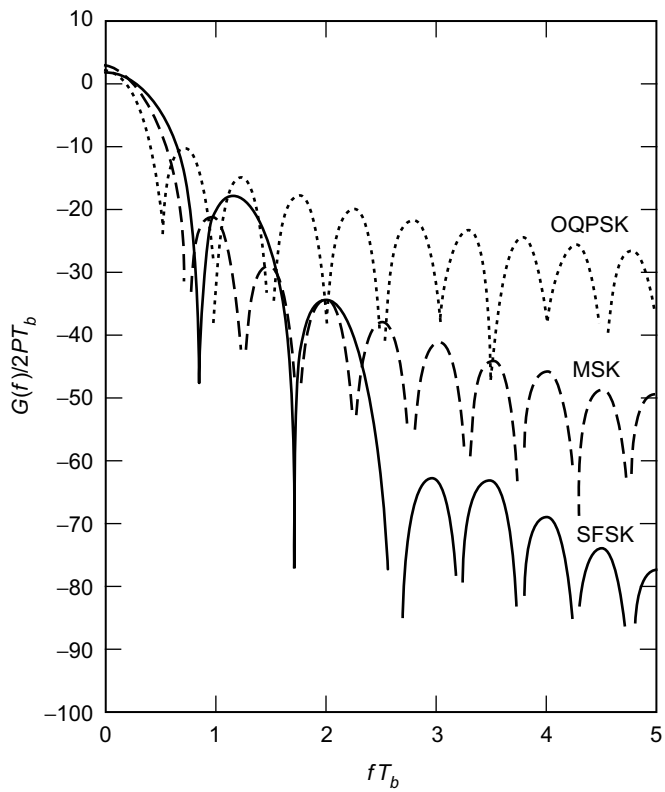


Fig. 2-15. A comparison of the equivalent baseband PSDs of MSK, OQPSK, and SFSK. Redrawn from [16].

whose main lobe is wider than that of MSK, but whose spectral sidelobes asymptotically roll off four orders of magnitude faster, i.e., at a rate f^{-8}. In fact, for the class of generalized MSK schemes, we can conclude that the smoother we make the shape of the frequency pulse, i.e., the more derivatives that go to zero at the endpoints $t = 0$ and $t = 2T_b$, the wider will be the main lobe but the faster the sidelobes will roll off.

Another way of interpreting the improved bandwidth efficiency that accompanies the equivalent I and Q pulse shaping is in terms of the fractional out-of-band power, defined as the fraction of the total power that lies outside a given bandwidth, i.e.,

$$\eta = 1 - \frac{\int_{-B/2}^{B/2} G(f)\, df}{\int_{-\infty}^{\infty} G(f)\, df} \tag{2.8-39}$$

Figure 2-16 is a plot of the fractional out-of-band power (in dB) versus BT_b for MSK, OQPSK, and SFSK, using the appropriate expression for $G(f)$ as determined from (2.8-35), (2.8-37), and (2.8-38), respectively.

2.8.1.5 Other Transmitter Representations.

a. Cross-Coupled I-Q Transmitter. A variation of the I-Q transmitter discussed in Sec. 2.8.1.2 is illustrated in Fig. 2-17 [24,25,26]. An modulated carrier at frequency f_c is multiplied by a lowpass sinusoidal signal at frequency $1/4T_b$ to produce a pair of unmodulated tones (carriers) at $f_2 = f_c + 1/4T_b$ and $f_1 = f_c - 1/4T_b$. These tones are separately extracted by narrow bandpass filters whose outputs, $s_1(t)$ and $s_2(t)$, are then summed and differenced to produce

$$\begin{aligned} z_c(t) = s_1(t) + s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] + \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \\ &= \cos\left(\frac{\pi t}{2T_b}\right)\cos 2\pi f_c t \end{aligned} \tag{2.8-40}$$

$$\begin{aligned} z_s(t) = s_1(t) - s_2(t) &= \frac{1}{2}\cos\left[2\pi\left(f_c - \frac{1}{4T_b}\right)t\right] - \frac{1}{2}\cos\left[2\pi\left(f_c + \frac{1}{4T_b}\right)t\right] \\ &= \sin\left(\frac{\pi t}{2T_b}\right)\sin 2\pi f_c t \end{aligned}$$

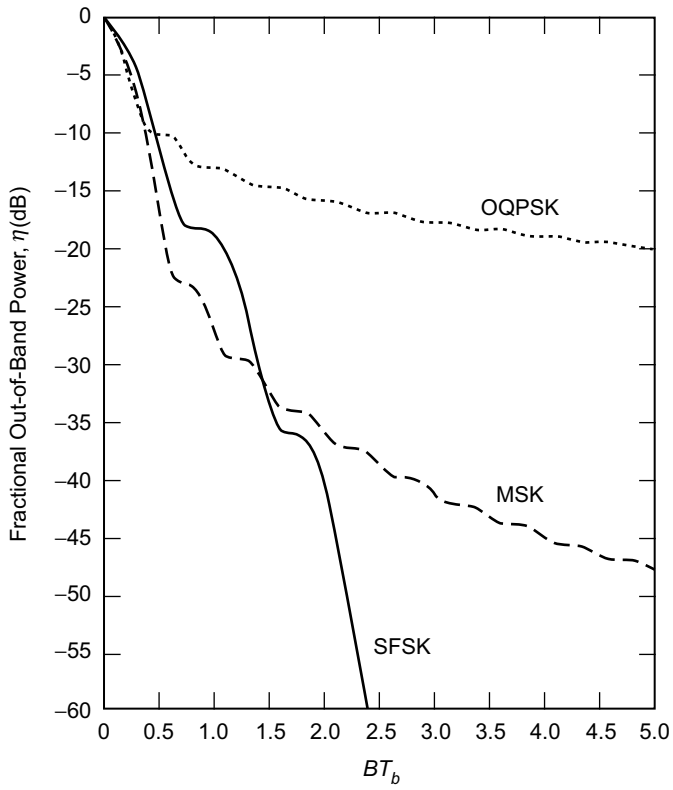


Fig. 2-16. A comparison of the fractional out-of-band power performance of MSK, OQPSK, and SFSK. Redrawn from [16].

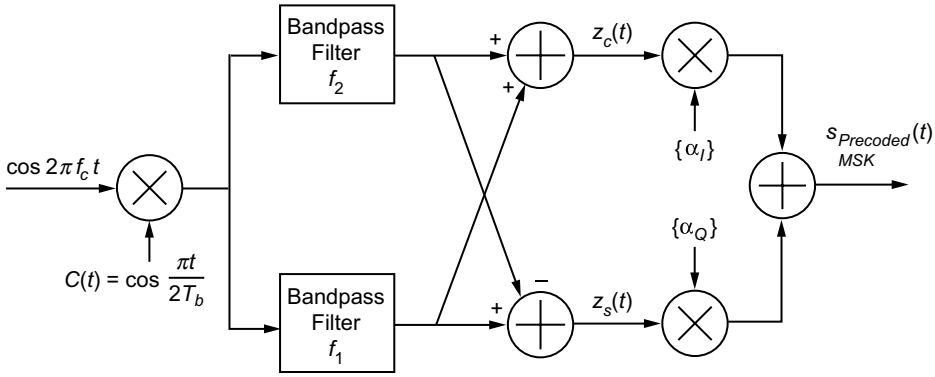


Fig. 2-17. Cross-coupled implementation of precoded MSK.

The signals $z_c(t)$ and $z_s(t)$ are respectively multiplied by I and Q data sequences $\{\alpha_I\}$ and $\{\alpha_Q\}$, each at a rate of $1/2T_b$ (and offset from each other by T_b s), and then differenced to produce the MSK (actually precoded MSK) output. The advantage of the implementation of Fig. 2-17 is that the signal coherence and the frequency deviation ratio are largely unaffected by variations in the data rate [25].

b. Rimoldi's Representation. As previously stated, the conventional CPM implementation of MSK produces a phase trellis that is symmetric about the horizontal axis, but that is time varying in that the possible phase states (reduced modulo 2π) alternate between $(0, \pi)$ and $(\pi/2, 3\pi/2)$ every T_b seconds. To remove this time-variation of the trellis, Rimoldi [27] demonstrated that CPM with a rational modulation index could be decomposed into the cascade of a memory encoder (finite-state machine) and a memoryless demodulator (signal waveform mapper). For the specific case of MSK, Rimoldi's transmitter is illustrated in Fig. 2-18. Imbalanced (0's and 1's) binary bits, $U_n = (1 - \alpha_n)/2$, are input to a memory one encoder. The current bit and the differentially encoded version of the previous bit (the encoder state) are used to define, via a binary-coded decimal (BCD) mapping, a pair of baseband signals (each chosen from a set of four possible waveforms) to be modulated onto I and Q carriers for transmission over the channel. Because of the imbalance of the data, the phase trellis is tilted as shown in Fig. 2-19, but on the other hand, it is now time invariant, i.e., the phase states (reduced modulo 2π) at all time instants (integer multiples of the bit time) are $(0, \pi)$. This transmitter implementation suggests the use of a simple two-state trellis decoder, which will be discussed in the next section

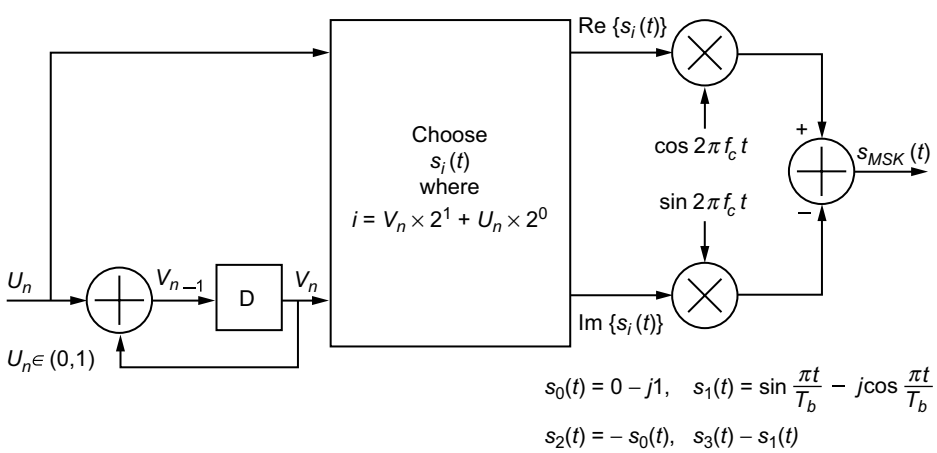


Fig. 2-18. MSK transmitter based on Rimoldi decomposition of CPM.

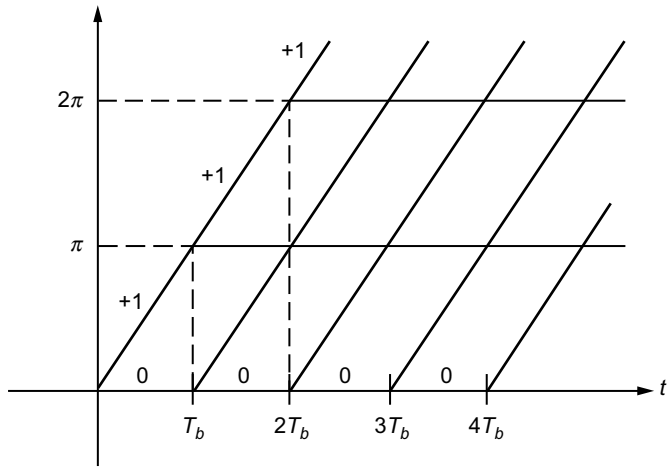


Fig. 2-19. Tilted (time-invariant) phase trellis for Rimoldi's MSK representation. Phase states (mod 2π) are $(0, \pi)$ for all n.

dealing with memory receiver structures. Also, later on in Chap. 4, we shall use Rimoldi's representation as the basis for developing bandwidth-efficient MSK-type modulations with memory greater than one under the constraint of finite decoding delay. Such modulations are not constrained to be constant envelope (rather, the transmitted signals are constrained to have equal energy) and thus, we defer our discussion of these schemes until that time.

Rimoldi's representation can also be used to implement precoded MSK. The appropriate transmitter is illustrated in Fig. 2-20.

2.8.1.6 Receiver Performance—Coherent Detection. Depending on the particular form used to represent the MSK signal, e.g., CPM, parallel I-Q, serial, etc., many different forms of receivers have been suggested in the literature for performing coherent detection. These various forms fall into two classes: structures based on a memoryless transmitter representation and structures based on a memory transmitter representation. As we shall see, all of these structures, however, are, themselves, memoryless.

a. Structures Based on a Memoryless Transmitter Representation. The most popular structure for coherent reception of MSK that is based on a memoryless transmitter representation corresponds to a parallel I-Q representation and has already been illustrated in Fig. 2-12. Here, the received signal plus noise is

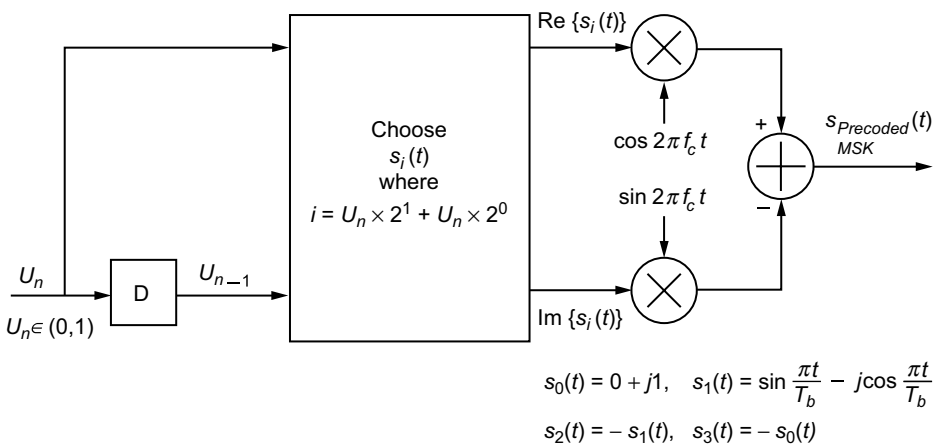


Fig. 2-20. Precoded MSK transmitter based on Rimoldi decomposition of CPM.

multiplied by the I and Q "carriers,"[11] $z_c(t)$ and $z_s(t)$, respectively, followed by integrate-and-dump (I&D) circuits of duration $2T_b$ seconds that are timed to match the zero crossings of the I and Q symbol waveforms. The multiplier-integrator combination constitutes a matched filter that, in the case of AWGN and no intersymbol interference (ISI), results in optimum detection. Means for producing the I and Q demodulation signals $z_c(t)$ and $z_s(t)$ will be discussed in the section on synchronization techniques.

b. Structures Based on a Memory Transmitter Representation. As noted in Sec. 2.8.1.5*b*, MSK (or precoded MSK) can be viewed as a cascade of a memory one encoder and a memoryless modulator. As such, a receiver can be implemented based on MLSE detection. For precoded MSK, the trellis diagram that appropriately represents the transitions between states is given in Fig. 2-21. Each branch of the trellis is labeled with the input bit (0 or 1) that causes a transition and the corresponding waveform (complex) that is transmitted as a result of that transition. The decision metrics based on a two-symbol observation that result in the surviving paths illustrated in Fig. 2-21 are

$$\int_{nT_b}^{(n+1)T_b} r(t)\, s_1(t)\, dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_0(t)\, dt > \int_{nT_b}^{(n+1)T_b} r(t)\, s_3(t)\, dt$$

$$+ \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_1(t)\, dt \tag{2.8-41a}$$

[11] The word "carrier" here is used to denote the combination (product) of the true carrier and the symbol waveform (clock).

and

$$\int_{nT_b}^{(n+1)T_b} r(t)\, s_1(t)\, dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_2(t)\, dt > \int_{nT_b}^{(n+1)T_b} r(t)\, s_3(t)\, dt$$

$$+ \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_3(t)\, dt \tag{2.8-41b}$$

Noting from Fig. 2-20 that $s_3(t) = -s_0(t)$ and $s_2(t) = -s_1(t)$, (2.8-41a) and (2.8-41b) can be rewritten as

$$\int_{nT_b}^{(n+1)T_b} r(t)\, s_0(t)\, dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_0(t)\, dt > -\int_{nT_b}^{(n+1)T_b} r(t)\, s_1(t)\, dt$$

$$+ \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_1(t)\, dt \tag{2.8-42a}$$

and

$$\int_{nT_b}^{(n+1)T_b} r(t)\, s_0(t)\, dt + \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_0(t)\, dt > -\int_{nT_b}^{(n+1)T_b} r(t)\, s_1(t)\, dt$$

$$+ \int_{(n+1)T_b}^{(n+2)T_b} r(t)\, s_1(t)\, dt \tag{2.8-42b}$$

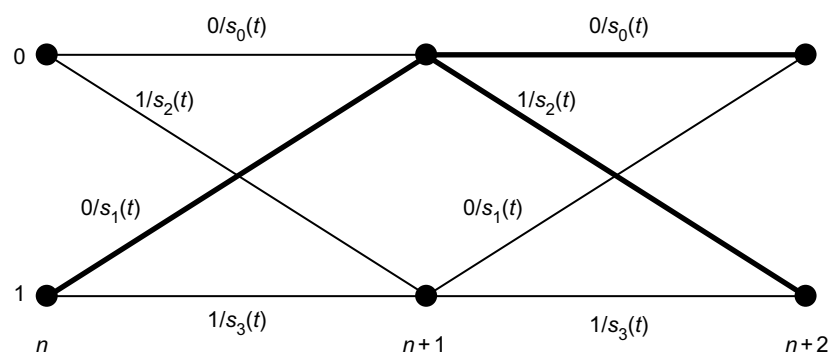


Fig. 2-21. A complex baseband trellis. Surviving paths for decoding $\hat{U}_n = 0$ in the interval $(n, n+1)$ assuming state "1" at time n are indicated by heavy lines.

which are identical and suggest the memoryless receiver illustrated in Fig. 2-22 [27].[12] Thus, we conclude that MSK (or precoded MSK) is a memory one type of trellis-coded modulation (TCM) that can be decoded with a finite (one bit) decoding delay, i.e., the decision on the nth bit can be made at the conclusion of observing the received signal for the n+1st transmission interval.

Massey [28] suggests an alternative representation of MSK (or precoded MSK) in the form of a single-input, two-output sequential transducer followed by an RF selector switch (Fig. 2-23). Analogous to the representation in (2.8-30), for precoded MSK, the sequential transducer implements the ternary sequences $\alpha_k^+ = (1/2)(\alpha_{k-1} + \alpha_k)$ and $\alpha_k^- = (-1)^k (1/2)(\alpha_{k-1} - \alpha_k)$. Note as before that α_k^+ is nonzero only when α_k^- is zero and vice versa. The function of the RF selector switch is to select one of the carriers for the signal to be transmitted in each bit interval according to the rule:

$$s(t) = \begin{cases} r_2(t) & \text{if } \alpha_k^+ = 1 \\ -r_2(t) & \text{if } \alpha_k^+ = -1 \\ r_1(t) & \text{if } \alpha_k^- = 1 \\ -r_1(t) & \text{if } \alpha_k^- = -1 \end{cases}, \quad r_i(t) = \sqrt{\frac{2E_b}{T_b}} \cos 2\pi f_i t, \quad i = 1, 2 \qquad (2.8\text{-}43)$$

which represents four mutually exclusive possibilities. This form of modulator has the practical advantage of not requiring addition of RF signals or RF filtering since there is no actual mixing of the carriers with the modulating signals.

Massey shows that, analogous to what is shown in Fig. 2-21, the output of the modulator can be represented by a trellis (Fig. 2-24), where again each branch is labeled with the input bit and the signal transmitted. Note that the trellis is time varying (the branch labels alternate with a period of two). In view of the trellis representation in Fig. 2-24, the optimum receiver is again an MLSE that has the identical structure as that in Fig. 2-22, where the complex demodulation signals $s_0(t - (n+1)T_b)$ and $s_1(t - (n+1)T_b)$ are replaced by the real carriers $r_1(t)\, r_1(t)$ and $r_2(t)$ of (2.8-43), the real part of the comparator (difference) output is omitted, and the decision device outputs balanced $+1, -1$ data rather than $0, 1$ data.

Regardless of the particular receiver implementation employed, the BEP performance of ideal coherent detection[13] of MSK is given by

---

[12] It can be shown that the surviving paths corresponding to being in state "0" at time n leads to the identical decision metric as that in (2.8-41a) or (2.8-41b).

[13] By "ideal coherent detection," we mean a scenario wherein the local supplied carrier reference is perfectly phase (and frequency) synchronous with the received signal carrier. Later on, we explore the practical implications of imperfect carrier synchronization.

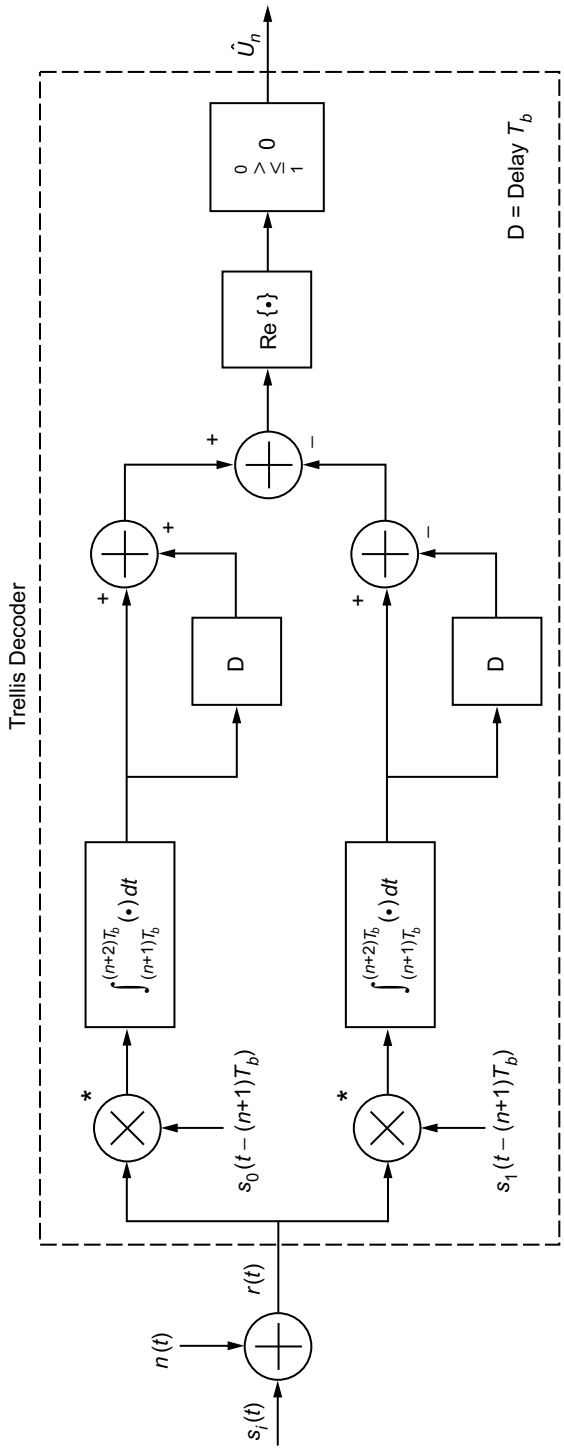


Fig. 2-22. Complex MLSE receiver.

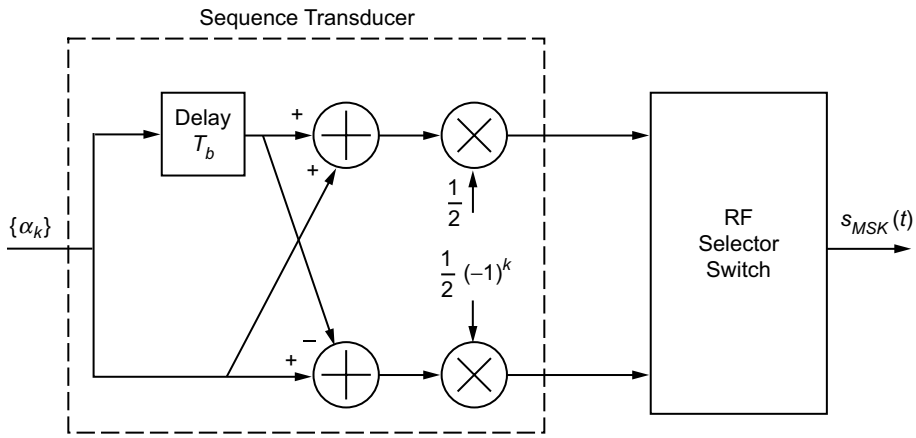


Fig. 2-23. Massey's precoded MSK transmitter.

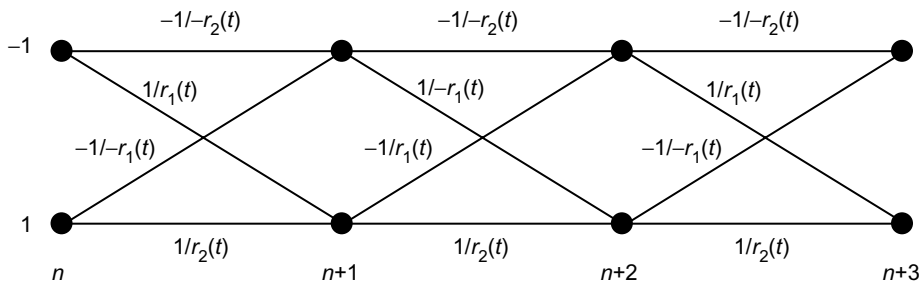


Fig. 2-24. Transmitter output trellis diagram.

$$P_b(E) = \operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\left(1 - \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}}\right) \tag{2.8-44}$$

whereas the equivalent performance of precoded MSK is

$$P_b(E) = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E_b}{N_0}} \tag{2.8-45}$$

which is identical to that of ideal coherent detection of BPSK, QPSK, or OQPSK [see (2.6-2)]. Comparing (2.8-44) with (2.8-45), we observe that the former can be written in terms of the latter as

$$P_b\left(E\right)|_{\text{MSK}} = 2P_b\left(E\right)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\left(1 - P_b\left(E\right)\Big|_{\substack{\text{precoded}\\ \text{MSK}}}\right) \tag{2.8-46}$$

which reflects the penalty associated with the differential encoding/decoding operation inherent in MSK but not in precoded MSK as previously discussed. At a BEP of 10^{-5}, this amounts to a penalty of approximately a factor of two in error probability or equivalently a loss of 0.75 dB in E_b/N_0.

2.8.1.7 Receiver Performance—Differentially Coherent Detection. In addition to coherent detection, MSK can be differentially detected [29], as illustrated in Fig. 2-25. The MSK signal plus noise is multiplied by itself delayed one bit and phase shifted 90 deg. The resulting product is passed through a lowpass zonal filter that removes second harmonics of the carrier frequency terms. Also assumed is that the carrier frequency and data rate are integer related, i.e., $f_cT_b = k$, with k integer. Assuming that the MSK signal input to the receiver is in the form of (2.8-1) combined with (2.8-12), i.e.,

$$s(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left(2\pi f_c t + \alpha_n \frac{\pi}{2T_b}t + x_n\right) = \sqrt{\frac{2E_b}{T_b}}\cos\Phi\left(t, \boldsymbol{\alpha}\right),$$

$$nT_b \le t \le (n+1)T_b \tag{2.8-47}$$

then the differential phase $\Delta\Phi \triangleq \Phi\left(t, \boldsymbol{\alpha}\right) - \Phi\left(t - T_b, \boldsymbol{\alpha}\right)$ is given by

$$\Delta\Phi \triangleq -\left(\alpha_{n-1} - \alpha_n\right)\frac{\pi}{2}\left(\frac{t}{T_b} - k\right) + \alpha_{n-1}\frac{\pi}{2} \tag{2.8-48}$$

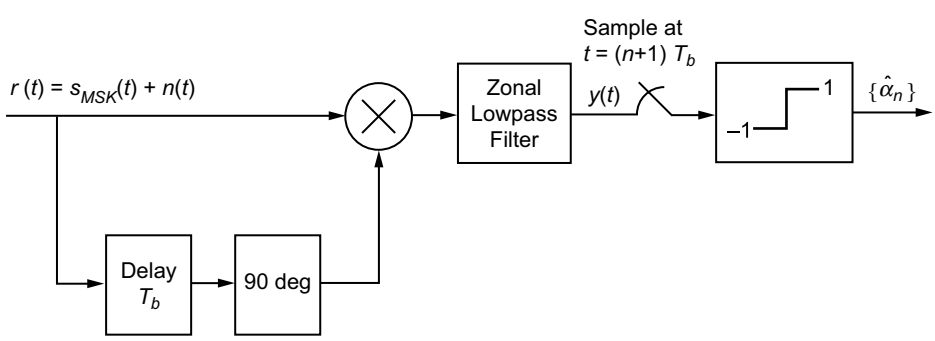


Fig. 2-25. Differentially coherent MSK receiver.

where we have made use of the phase continuity relation in (2.8-15) in arriving at (2.8-48). The mean of the lowpass zonal filter output can be shown to be given by

$$\overline{y(t)} = s(t)\, s_{90}(t) = \frac{E_b/T_b}{2} \sin \Delta\Phi \tag{2.8-49}$$

where the "90" subscript denotes a phase shift of 90 deg in the corresponding signal. Combining (2.8-48) and (2.8-49), the sampled mean of the lowpass zonal filter output at time $t = (n+1)\, T_b$ becomes

$$\overline{y((k+1)\, T_b)} = \frac{E_b/T_b}{2} \sin\left(\alpha_k \frac{\pi}{2}\right) = \alpha_k \frac{E_b/T_b}{2} \tag{2.8-50}$$

which clearly indicates the appropriateness of a hard limiter detector in the presence of noise. Figure 2-26 is an illustration of the various waveforms present in the differentially coherent receiver of Fig. 2-25 for a typical input data sequence.

2.8.1.8 Synchronization Techniques. In our discussion of coherent reception in Sec. 2.8.1.6, we implicitly assumed that a means was provided in the receiver for synchronizing the phase of the local demodulation reference(s) with that of the received signal carrier and also for time synchronizing the I&D circuits. Here we discuss several options for implementing such means.

One form of combined carrier and clock recovery that is synergistic with the transmitter form in Fig. 2-17 was originally proposed by DeBuda [30,31].[14] With reference to Fig. 2-27, the received MSK signal is first squared to produce an FSK signal at twice the carrier frequency and with twice the modulation index, i.e., $h = 1$, which is known as Sunde's FSK [32]. Whereas the MSK signal has no discrete (line) spectral components, after being squared, it has strong spectral components at $2f_1$ and $2f_2$, which can be used for synchronization. In fact, Sunde's FSK has 50 percent of its total power in these two line components (the other 50 percent of the total power is in a discrete line component at dc). To demonstrate this transformation from continuous to discrete spectrum, we square the MSK signal form in (2.8-30), which gives

[14] DeBuda also referred to MSK, in conjunction with his self-synchronizing circuit, as "fast FSK (FFSK)," which at the time was the more popular terminology in Canada.

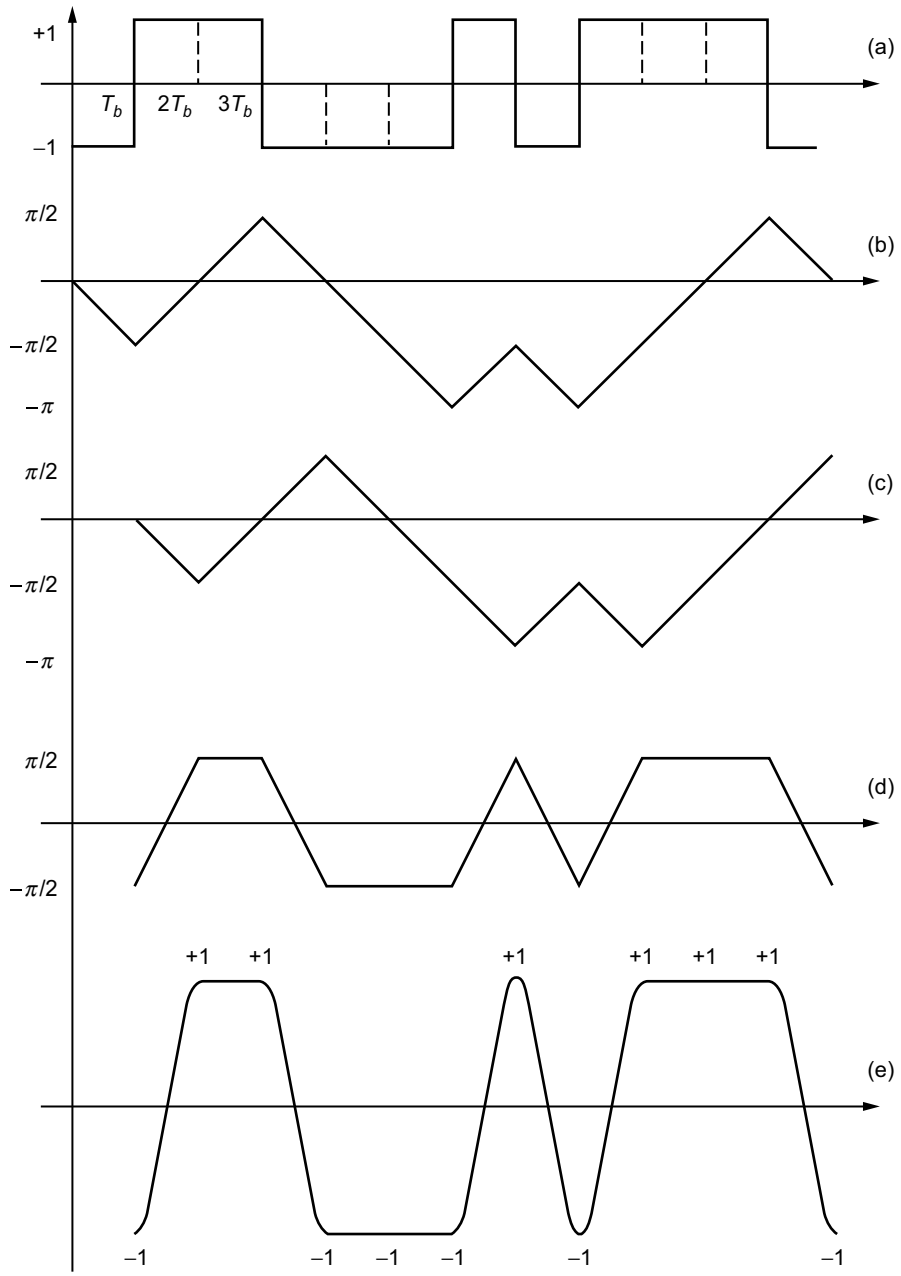


Fig. 2-26. Various waveforms present in the differentially coherent receiver shown in Fig. 2-25: (a) transmitted bit sequence, (b) transmitted phase, (c) transmitted phase delayed, (d) difference phase, and (e) multiplier output (sine of difference phase).

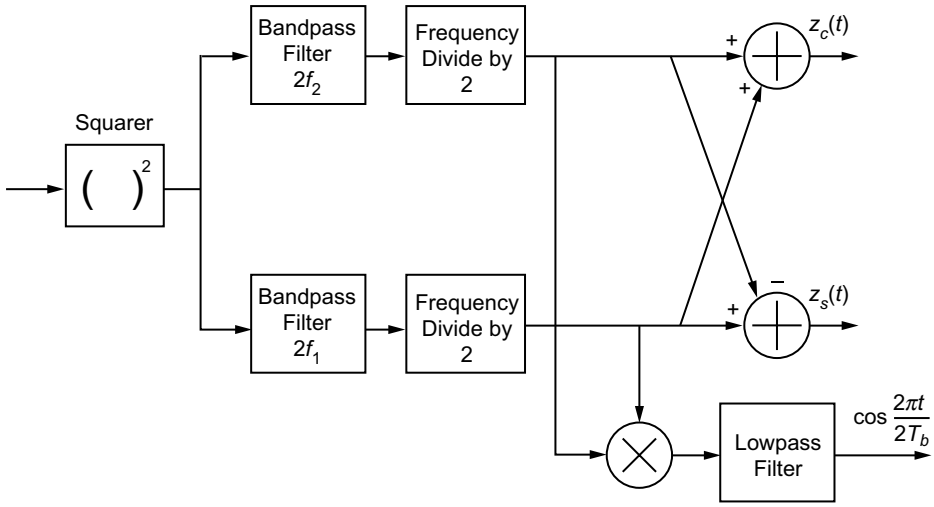


Fig. 2-27. DeBuda's carrier and symbol synchronization scheme.

$$s^2_{\text{MSK}}(t) = \frac{2E_b}{T_b}\left[\left(v_n^+\right)^2\cos^2 2\pi f_2 t + \left(v_n^-\right)^2\cos^2 2\pi f_1 t + 2v_n^+ v_n^- \cos 2\pi f_2 t \cos 2\pi f_1\right]$$

$$= \frac{2E_b}{T_b}\left[\frac{1}{2} + \frac{1}{2}\left(v_n^+\right)^2\cos 4\pi f_2 t + \frac{1}{2}\left(v_n^-\right)^2\cos 4\pi f_1 t\right],$$

$$v_n^+ = \frac{v_{n-1}+v_n}{2}, \quad v_n^- = (-1)^n\left(\frac{v_{n-1}-v_n}{2}\right) \qquad (2.8\text{-}51)$$

where we have made use of the fact that since either v_n^+ or v_n^- is always equal to zero, then $v_n^+ v_n^- = 0$. Also, either $\left(v_n^+\right)^2 = 1$ and $\left(v_n^-\right)^2 = 0$ or vice versa, which establishes (2.8-51) as a signal with only discrete line components. The components at $2f_1$ and $2f_2$ are extracted by bandpass filters (in practice, phase-locked loops) and then frequency divided to produce $s_1(t) = (1/2)\cos 2\pi f_1 t$ and $s_2(t) = (1/2)\cos 2\pi f_2 t$. The sum and difference of these two signals produce the reference "carriers" $z_c(t) = C(t)\cos 2\pi f_c t$ and $z_s(t) = S(t)\sin 2\pi f_c t$, respectively, needed in Fig. 2-12. Finally, multiplying $s_1(t)$ and $s_2(t)$ and low-pass filtering the result produces $(1/8)\cos 2\pi t/2T_b$ (a signal at 1/2 the bit rate), which provides the desired timing information for the I&Ds in Fig. 2-12.

Another joint carrier and timing synchronization scheme for MSK was derived by Booth [33] in the form of a closed loop motivated by the maximum a posteriori (MAP) estimation of carrier phase and symbol timing. The resulting

structure [Fig. 2-28(a)] is an overlay of two MAP estimation I-Q closed loops—one typical of a carrier synchronization loop, assuming known symbol timing [Fig. 2-28(b)] and one typical of a symbol timing loop, assuming known carrier phase [Fig. 2-28(c)]. In fact, the carrier synchronization component loop is identical to what would be obtained for sinusoidally pulse-shaped OQPSK.

Finally, many other synchronization structures have been developed for MSK and conventional (single modulation index) binary CPM, which, by definition, would also be suited to MSK. A sampling of these is given in Refs. 34–40. In the interest of brevity, however, we do not discuss these here. Instead, the interested reader is referred to the cited references for the details.

2.8.2 Partial Response—Gaussian MSK

GMSK was first introduced by Murota, Kinoshita, and Hirada [41] in 1981 as a highly bandwidth-efficient constant envelope modulation scheme for communication in the 900-MHz land mobile radio environment (see [42,43] for field experimental results of performance in this frequency band). In simple terms, GMSK is an $h = 0.5$ partial-response CPM scheme obtained by filtering the rectangular frequency pulses characteristic of MSK with a filter having a Gaussian impulse response prior to frequency modulation of the carrier.[15] As such, the GMSK frequency pulse is the difference of two time-displaced (by T_b seconds) Gaussian probability integrals (Q-functions), i.e.,[16]

$$g(t) = \frac{1}{2T_b}\left[Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\left(\frac{t}{T_b} - 1\right)\right) - Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\frac{t}{T_b}\right)\right],$$

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy, \quad -\infty \le t \le \infty \qquad (2.8\text{-}52)$$

[15] It is important to emphasize that although the acronym GMSK was assigned to the term *Gaussian-filtered MSK* in [41], the modulation actually described in this reference applies the Gaussian filtering at baseband, i.e., prior to modulation onto the carrier, and, hence, it does not destroy the constant envelope property of the resulting modulation. Perhaps because of this poor usage of the term *Gaussian-filtered MSK*, occasionally there appears in the literature [44, p. 519] a misleading statement alluding to the fact that GMSK is an "MSK modulated signal passed through a Gaussian filter ...," which would imply Gaussian filtering at RF, thereby destroying the constant envelope nature of the signal. This interpretation is not in keeping with the original description of GMSK in [41] and the large number of references that followed; thus, we caution the reader against adopting this usage.

[16] We assume here a frequency pulse shape, $g(t)$, that results from excitation of the Gaussian filter (arbitrarily assumed to have zero group delay) with the unit rectangular pulse $p(t) = 1, 0 \le t \le T_b$.

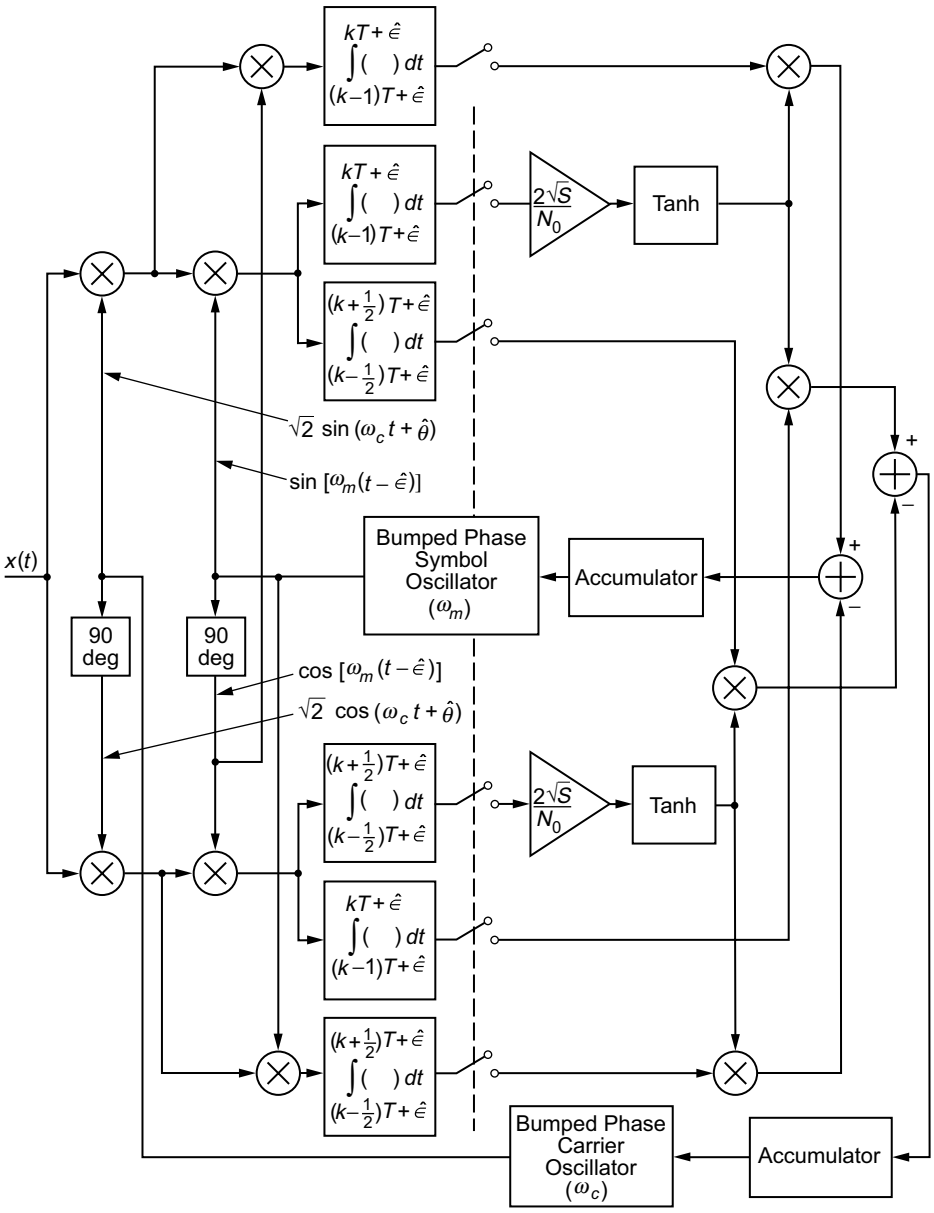


Fig. 2-28(a). Joint carrier and symbol MAP estimation loop for MSK modulation.

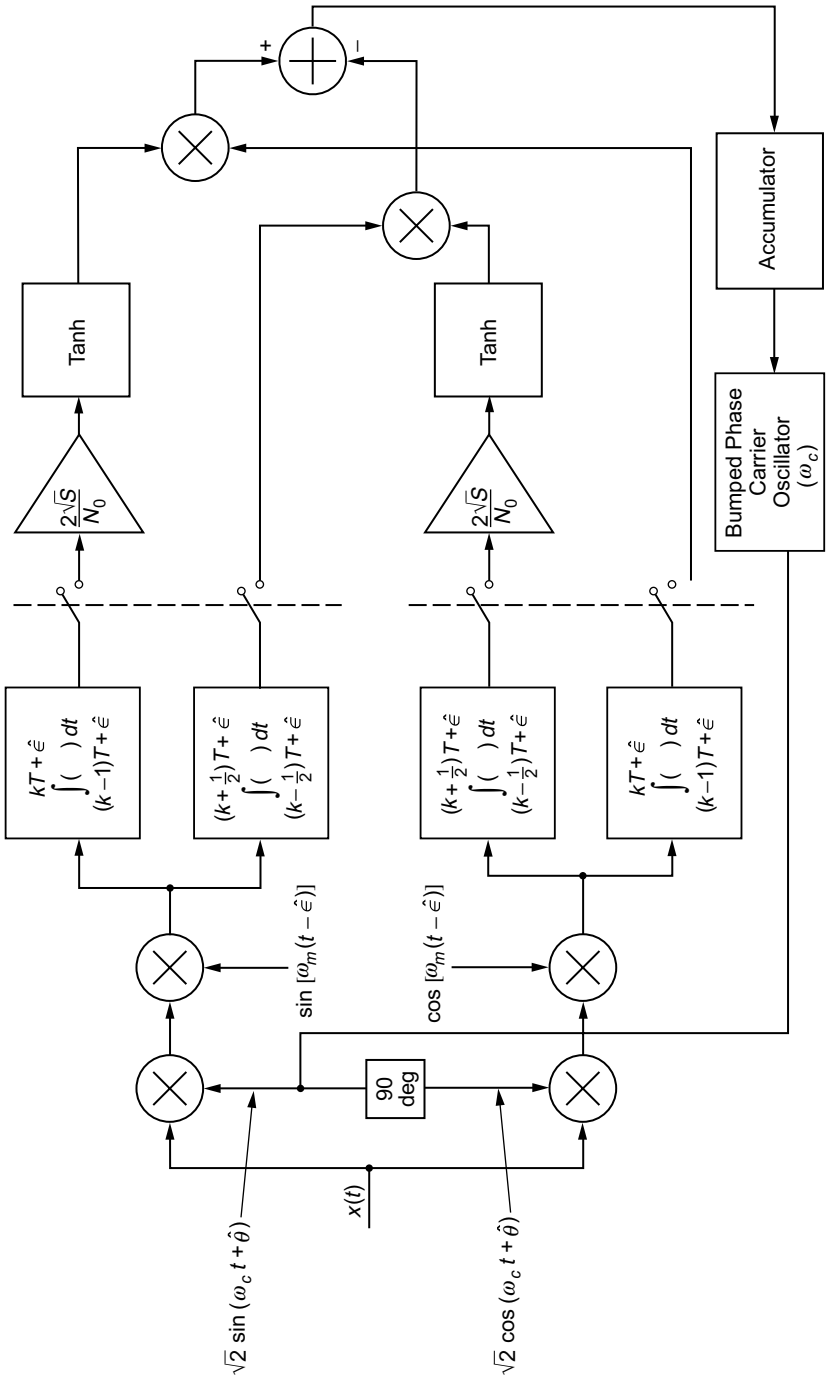


Fig. 2-28(b). Joint carrier and symbol MAP estimation loop for MSK modulation (carrier synchronization component).

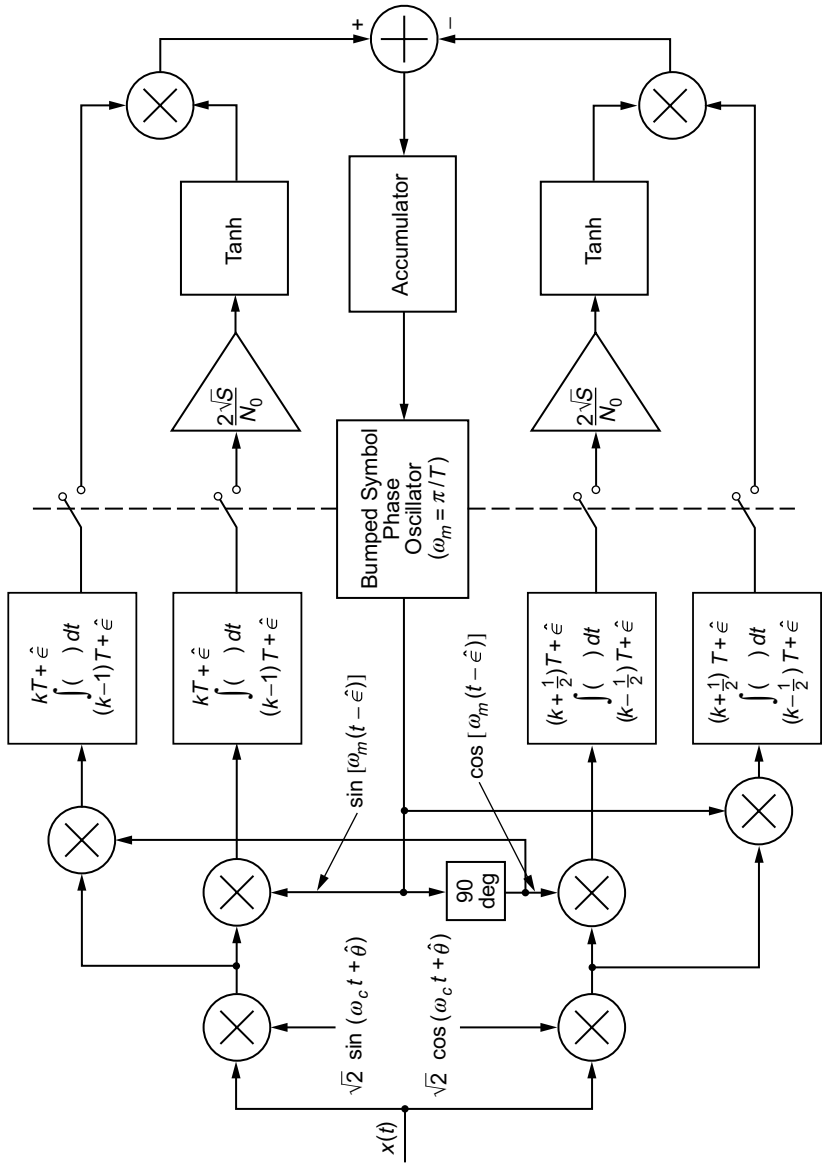


Fig. 2-28(c). Joint carrier and symbol MAP estimation loop for MSK modulation (symbol synchronization component).

where B is the 3-dB bandwidth of the lowpass Gaussian filter and is related to the noise bandwidth, B_N, of this filter by [45, Eq. (14)]

$$\frac{B}{B_N} = 2\sqrt{\frac{\ln 2}{\pi}} = 0.93944 \tag{2.8-53}$$

Smaller values of BT_b lead to a more compact spectrum but also introduce more ISI and, therefore, a degraded error probability performance. Thus, for a given application, the value of BT_b is selected as a compromise between spectral efficiency and BEP performance.

Since the Gaussian Q-function is doubly infinite in extent, it is common practice to time-truncate the GMSK frequency pulse so as to deal with finite ISI. For $BT_b = 0.25$, truncating $g(t)$ of (2.8-52) to four bit intervals is appropriate [46] whereas for $BT_b = 0.3$, the value used in the Global System for Mobile (GSM) application [47], considering ISI only from adjacent bits (i.e., time truncation to three bit intervals) has been shown to be sufficient [48]. Thus, in practical GMSK implementations, one employs the approximation (see Fig. 2-29)

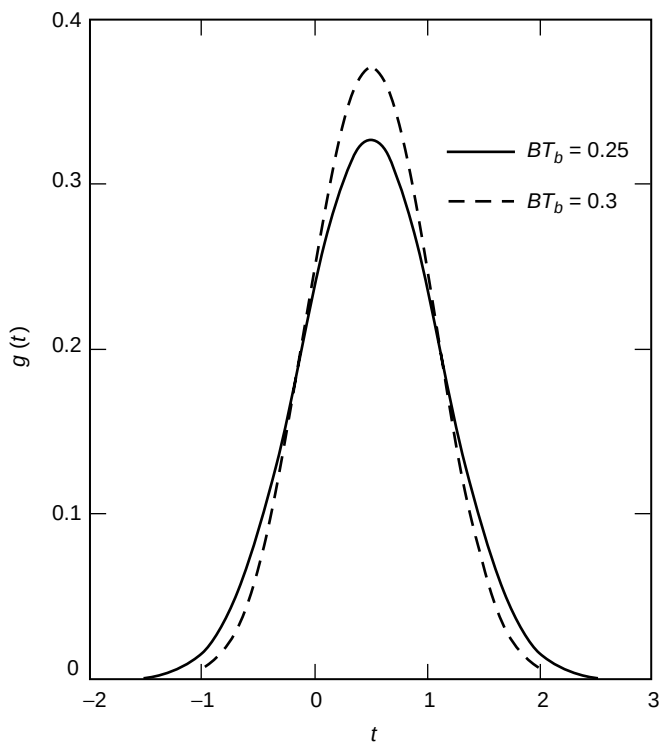


Fig. 2-29. GMSK frequency pulse.

$$g(t) = \begin{cases} \dfrac{1}{2T_b}\left[Q\left(\dfrac{2\pi BT_b}{\sqrt{\ln 2}}\left(\dfrac{t}{T_b}-1\right)\right) - Q\left(\dfrac{2\pi BT_b}{\sqrt{\ln 2}}\dfrac{t}{T_b}\right)\right], & -(L-1)\,T_b/2 \le t \\ & \le (L+1)\,T_b/2 \\ 0, & \text{otherwise} \end{cases} \tag{2.8-54}$$

where L is chosen as above in accordance with the value of BT_b.[17] Also, although $g(t)$ of (2.8-54) appears to have a "Gaussian-looking" shape, we emphasize that the word *Gaussian* in GMSK refers to the *impulse* response of the filter through which the input rectangular pulse train is passed and not the shape of the resulting frequency pulse.

2.8.2.1 Continuous Phase Modulation Representation. Based on the above, the CPM representation of GMSK is, analogous to (2.8-10),

$$s_{\text{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\cos\left(2\pi f_c t + \frac{\pi}{2T_b}\sum_i \alpha_i \int \left[Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\left(\frac{\tau}{T_b}-(i+1)\right)\right)\right.\right.$$

$$\left.\left. -Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\left(\frac{\tau}{T_b}-i\right)\right)\right] d\tau\right), \quad nT_b \le t \le (n+1)\,T_b \tag{2.8-55}$$

which is implemented, analogous to Fig. 2-7, in Fig. 2-30(a). Equivalently, if the input is represented by its equivalent NRZ data stream (i.e., the frequency pulse stream that would ordinarily be inputted to the FM modulator in MSK), then the filter impulse response, $h(t)$, becomes Gaussian, as implied by the GMSK acronym, i.e.,

$$h(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \sigma^2 = \frac{\ln 2}{(2\pi B)^2} \tag{2.8-56}$$

(appropriately time-truncated as discussed above), and the implementation appears as in Fig. 2-30(b).

---

[17] Technically speaking, $g(t)$ of (2.8-53) should be scaled by a constant C so as to satisfy a condition analogous to (2.8-5), namely,

$$q(t) = \int_{-\infty}^{t} g(\tau)d\tau = \begin{cases} 0, & t \le -(L-1)\,T_b/2 \\ 1/2, & t \ge (L+1)\,T_b/2 \end{cases}$$

However, for the values of BT_b of practical interest, i.e., $BT_b \ge 0.25$, the scaling constant is ignored, i.e., C is nominally taken as unity.

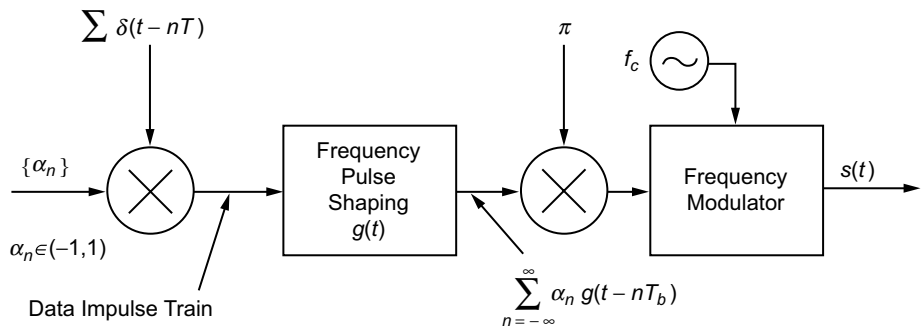


Fig. 2-30(a). GMSK transmitter (CPM representation).

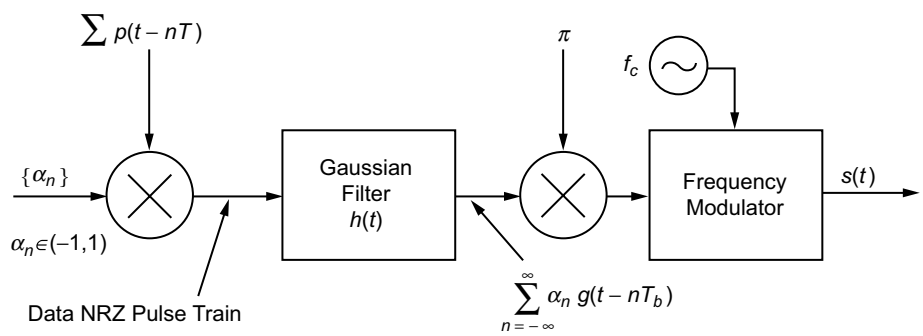


Fig. 2-30(b). Equivalent GMSK transmitter (CPM representation).

The frequency modulator in Fig. 2-30(a) or 2-30(b) is typically implemented with a phase-locked loop (PLL) synthesizer whose voltage-controlled oscillator (VCO) input is the point at which the modulation is injected. When long strings of zeros or ones are present in the data, the spectrum of the modulation extends to dc, which presents a problem, since PLL frequency synthesizers implemented as above do not respond to this low-frequency signal due to their inherent high-pass filter characteristic. As such, the VCO output (the location of the modulated signal) would not contain the low-frequency content of the information (modulating) signal. By contrast, if the modulation were to be injected at the input of the master oscillator preceding the PLL (the oscillator must be capable of being modulated by a voltage signal), then since this oscillator is not in the loop, the VCO output would contain the low-frequency content of the modulation (i.e., that within the loop filter bandwidth) but not its high-frequency content. Clearly then, a combination of these two approaches would yield the desirable result of constant modulation sensitivity, irrespective of the loop bandwidth. Such an FM scheme is referred to as two-point modulation [49] and corresponds to a

dc-coupled GMSK modulator wherein the Gaussian filtered input signal is split sending one portion to the VCO modulation input and the other to the PLL master oscillator input.

2.8.2.2 Equivalent I-Q Representations. For high carrier frequencies, direct synthesis of the GMSK signal as in Fig. 2-7, using a digital approach is impractical since maintaining an adequate sampling rate requires an extremely high operating frequency. Instead, one can resort to a quadrature implementation where lowpass I and Q signals containing the phase information are generated that vary much slower than the phase of the modulated carrier, thus making it feasible to implement them digitally. Applying the simple trigonometric rule for the cosine of the sum of two angles to (2.8-55), we obtain

$$s_{\mathrm{MSK}}(t) = \sqrt{\frac{2E_b}{T_b}}\left[\cos\phi(t,\boldsymbol{\alpha})\cos 2\pi f_c t - \sin\phi(t,\boldsymbol{\alpha})\sin 2\pi f_c t\right] \tag{2.8-57}$$

where

$$\phi(t,\boldsymbol{\alpha}) =$$

$$\frac{\pi}{2T_b}\sum_i \alpha_i \int \left\{ Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\left(\frac{\tau}{T_b} - (i+1)\right)\right) - Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\left(\frac{\tau}{T_b} - i\right)\right)\right\} d\tau \tag{2.8-58}$$

Conceptually then, an I-Q receiver for GMSK is one that performs the following sequence of steps: first, the Gaussian-filtered NRZ data stream is generated. Next, integration is performed to produce the instantaneous phase of (2.8-58). Finally, the integrator output is passed through sine and cosine read-only memories (ROMs) whose outputs are applied to I and Q carriers (see Fig. 2-31). Such a scheme has also been referred to as quadrature cross-correlated GMSK (see [50, Fig. 4.3.20] for an illustration similar to Fig. 2-31). Several commercial vendors and industrial organizations, e.g., Alcatel and Aerospace, have digitally implemented this generic approach in the transmitter design of their GMSK modems. In these implementations, the block labeled "Gaussian filter" is either an actual filter that approximates the Gaussian impulse response as per (2.8-54) or, more efficiently, a ROM table lookup, whereas the block labeled "integrator" is typically performed by a "phase accumulator."[18]

[18] Without loss in generality, the Gaussian filter and integrator blocks can be switched as is the case in some of the implementations.

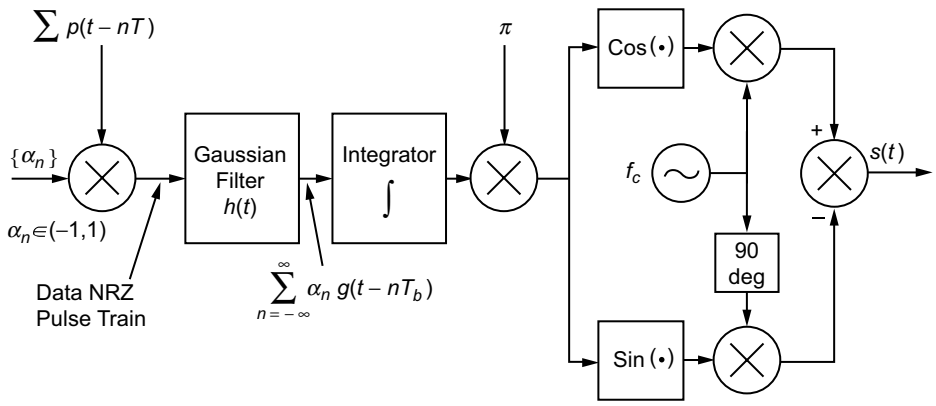


Fig. 2-31. GMSK transmitter (I-Q representation).

In [45], an efficient I-Q implementation of a GMSK modulator is presented that skips the above sequence of steps and instead generates the I and Q baseband signals directly from the binary data, thereby eliminating the errors in filtering, phase truncation, and sine/cosine computation inherent in the conventional architecture. A brief description of this method is as follows, based on the assumption of ISI only from adjacent symbols, i.e., $L = 3$.

Consider the GMSK frequency response (pulse train) that generates the phase of (2.8-58). If we impose the condition that this response in the mth bit interval, $mT_b \le |t| \le (m+1)\,T_b$, be dependent only on the bit of interest, α_m, and its two nearest neighbors, α_{m-1} and α_{m+1}, i.e., only adjacent ISI, then it can be shown [45, Eqs. (28), (29)] that it is sufficient to require

$$\left.\begin{aligned} Q\left(\frac{2\pi BT_b}{\sqrt{\ln 2}}\right) &\cong 0 \\ Q\left(-\frac{2\pi BT_b}{\sqrt{\ln 2}}\right) &\cong 1 \end{aligned}\right\} \tag{2.8-59}$$

Assuming (2.8-59) is true, then since by superposition the response to a train of NRZ pulses varying from -1 to 1 is the equivalent to the response to a rectangular pulse train varying from 0 to 2 minus a constant of value 1, the normalized frequency response in the above interval can be expressed as

$$g_m(t) \triangleq \sum_{i=m-1,m,m+1} (\alpha_i + 1) \left[Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - (i+1) \right) \right) \right.$$

$$\left. - Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - i \right) \right) \right] dt - 1$$

$$\cong (\alpha_{m-1} + 1)\, Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - m \right) \right)$$

$$+ (\alpha_m + 1) \left[Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - (m+1) \right) \right) - Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - m \right) \right) \right]$$

$$+ (\alpha_{m+1} + 1) \left[1 - Q\left(\frac{2\pi B T_b}{\sqrt{\ln 2}} \left(\frac{t}{T_b} - (m+1) \right) \right) \right] - 1 \qquad (2.8\text{-}60)$$

Alternatively, since the Gaussian Q-function can be expressed in terms of the error function by $Q(x) = (1/2)\left[1 + \operatorname{erf}\left(x/\sqrt{2}\right)\right]$, then letting $\alpha_i' = (1/2)(\alpha_i + 1)$ denote the (0,1) equivalent of the $(-1, 1)$ α_i's, and introducing the constant $\beta \triangleq \pi B \sqrt{2/\ln 2}$, as in Eq. (19) of Ref. 45, (2.8-60) can be rewritten as

$$g_m(t) \cong \alpha_{m-1}' \left[1 - \operatorname{erf}\big(\beta (t - m T_b)\big) \right]$$

$$+ \alpha_m' \left[\operatorname{erf}\big(\beta (t - m T_b)\big) - \operatorname{erf}\big(\beta (t - (m+1) T_b)\big) \right]$$

$$+ \alpha_{m+1}' \left[1 + \operatorname{erf}\big(\beta (t - (m+1) T_b)\big) \right] - 1 \qquad (2.8\text{-}61)$$

Corresponding to the values (0,1) for each of the three α_i's in (2.8-61), there are eight possible waveforms $f_i(t - m T_b), i = 0, 1, 2, \cdots, 7$ that characterize the frequency response in the mth bit interval. These are given in Table 2-1 assuming $m = 0$ for simplicity.

Table 2-1. Possible frequency responses in the interval $0 \leq t \leq T_b$.

$\alpha'_{-1}, \alpha'_0, \alpha'_1$	i	$f_i(t)$
000	0	-1
001	1	$\operatorname{erf}\big(\beta(t-T_b)\big)$
010	2	$\operatorname{erf}(\beta t) - \operatorname{erf}\big(\beta(t-T_b)\big) - 1$
011	3	$\operatorname{erf}(\beta t)$
100	4	$-\operatorname{erf}(\beta t)$
101	5	$-\operatorname{erf}(\beta t) + \operatorname{erf}\big(\beta(t-T_b)\big) + 1$
110	6	$-\operatorname{erf}\big(\beta(t-T_b)\big)$
111	7	1

We observe from this table that there are only three independent frequency response waveforms, i.e., $f_2(t), f_3(t), f_7(t)$, in that the remaining five can be obtained from these three by means of simple operations, namely,

$$\left.\begin{aligned}
f_0(t) &= -f_7(t) \\
f_1(t) &= f_3(t) - f_2(t) - f_7(t) = f_3(t - T_b) \\
f_4(t) &= -f_3(t) \\
f_5(t) &= -f_2(t) \\
f_6(t) &= -f_1(t)
\end{aligned}\right\} \tag{2.8-62}$$

In view of the above, the frequency modulating signal corresponding to the phase modulating signal of (2.8-58) can be expressed in the form of a data-dependent pulse train as

$$f(t, \alpha) = \frac{1}{2\pi}\frac{d}{dt}\phi(t, \alpha) = \frac{1}{4T_b}\sum_i f_{l(i)}(t - iT_b)\, p(t - iT_b) \tag{2.8-63}$$

where as before, $p(t)$ is a unit amplitude rectangular pulse in the interval $0 \leq t \leq T_b$ and the index $l(i) = 4a_{i-1} + 2a_i + a_{i+1}$ is the decimal equivalent of

the 3-bit binary sequence influencing the ith bit interval and determines the particular frequency waveform for that interval in accordance with Table 2-1. The corresponding complex phase modulating signal can be written in the form

$$\exp\{\phi(t,\alpha)\} = \exp\left\{\sum_i \phi_{l(i)}(t - iT_b)\, p(t - iT_b)\right\},$$

$$\phi_i(t) = \frac{\pi}{2T_b}\int_0^t f_i(\tau)\, d\tau + \phi_i(0) \tag{2.8-64}$$

where $\phi_i(0)$ is the initial phase value that depends on the past history of the data sequence. Analogous to Table 2-1, there are eight possible phase responses in any given bit interval. These are evaluated in Ref. 45, using the approximation of (2.8-59) (reformulated in terms of the error function as $\mathrm{erf}(\beta T_b) \cong 1$, $\mathrm{erf}(-\beta T_b) \cong -1$), along with appropriate asymptotic expansions of the error function. Once again, there are only three independent phase response waveforms, e.g., $\phi_2(t), \phi_3(t), \phi_7(t)$, and the remaining five can be obtained from these three by means of simple operations.

The phase responses are used to determine phase trellises, keeping in mind that the sequences of possible phase trajectories generated by the 3-bit data sequences in each bit interval are constrained by the fact that only one of the 3 bits changes from interval to interval. Thus, for example, 010 can only be followed by 100 or 101. Furthermore, since we are interested only in the sine and cosine of the phase, it is sufficient to consider the phase trajectories modulo 2π. Using these trellises, it is shown in Ref. 45 that only four curves of T_b-s duration are needed to generate the I (from the cosine of the phase) and Q (from the sine of the phase) signals directly from the input data sequence for that bit interval. This is accomplished with a table lookup ROM that stores these four basic curves.

Finally, the practical trade-offs in terms of recent digital integrated circuit (IC) technology between the FM/VCO and I-Q transceiver architectures are discussed in Ref. 51.

2.8.2.3 Other GMSK Representations—The Laurent Expansion. A decade and a half ago, Laurent [52] described a representation for CPM in the form of a superposition of phase-shifted amplitude-modulation pulse (AMP) streams, the number of such being dependent on the amount of partial response in the modulation, as described by the duration (in bits) of the frequency pulse. A full-response scheme such as MSK required only a single pulse stream (with complex symbols). The primary focus of this work was on binary modulation

because of its relative simplicity of implementation.[19] The motivation for presenting such a representation was twofold. First, it allowed for easier evaluation of the autocorrelation and PSD of such modulations; in particular, simple results were specifically obtained for half-integer index modulations, i.e., ones whose frequency modulation index was of the form $h = n + 1/2$, n integer. Second, it allowed for approximation (with good accuracy) of CPM by a single pulse stream with one optimized pulse shape (called the "main pulse") and as such offered a synthesis means no more complicated than MSK.

Three years later, Kaleh [46] exploited Laurent's representation of CPM to allow for simple implementation of coherent receivers of such modulations, in particular, for the case of GMSK. Two forms of such receivers were considered, namely, a simplification of the optimum MLSE receiver and a linear MSK-type receiver, both of which yielded small degradation relative to the true optimum MLSE receiver.

In this section, we summarize the key results of these papers in so far as the transmitter implementation is concerned, devoting more time to the interpretations of the results than to the details of the derivations.

a. Exact AMP Representation of GMSK. In what follows, it will be convenient to deal with the complex envelope of the signal $s(t)$, i.e., the complex baseband signal, $\tilde{S}(t)$, defined by the relation

$$s(t) = \mathrm{Re}\left\{\tilde{S}(t)e^{j2\pi f_c t}\right\} \tag{2.8-65}$$

Thus, from (2.8-1), we have for binary CPM

$$\tilde{S}(t) = \sqrt{\frac{2E_b}{T_b}} \exp\left\{j\phi\left(t, \boldsymbol{\alpha}\right)\right\}, \quad nT_b \le t \le (n+1)T_b \tag{2.8-66}$$

For $h = 0.5$ partial-response CPM, where the frequency pulse has duration LT_b (remember from our previous discussion that the value of L used to approximate GMSK is a function of the value of BT_b of interest), Laurent showed after much manipulation that the complex envelope in (2.8-66) could be expressed as[20]

---

[19] The work was later extended to the M-ary case by Mengali and Morelli [53].

[20] For observation of the signal in the Nth transmission interval, $(N-1)T_b \le t \le NT_b$, the upper limit on n in (2.8-67) and (2.8-68) can be changed from ∞ to $N-1$ since the signal does not depend on future data bits. Furthermore, for a finite data sequence of length N, i.e., $\alpha_0, \alpha_1, \cdots, \alpha_{N-1}$, the lower limit on n in (2.8-67) and (2.8-68) can be changed from $-\infty$ to 0.

$$\tilde{S}(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{K=0}^{2^{L-1}-1} \left[\sum_{n=-\infty}^{\infty} e^{j\frac{\pi}{2}A_{K,n}} C_K\left(t - nT_b\right) \right]$$

$$\triangleq \sqrt{\frac{2E_b}{T_b}} \sum_{K=0}^{2^{L-1}-1} \left[\sum_{n=-\infty}^{\infty} \tilde{a}_{K,n} C_K\left(t - nT_b\right) \right] \tag{2.8-67}$$

which results in the real CPM signal

$$s(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{K=0}^{2^{L-1}-1} \left[\sum_{n=-\infty}^{\infty} C_K\left(t - nT_b\right) \cos\left(2\pi f_c t + \frac{\pi}{2} A_{K,n}\right) \right] \tag{2.8-68}$$

i.e., a superposition of 2^{L-1} amplitude-/phase-modulated pulse streams. In (2.8-68), $C_K\left(t\right)$ is the equivalent pulse shape for the kth AMP stream and is determined as follows:

First, define the generalized phase pulse function by

$$\Psi(t) = \begin{cases} \pi q(t), & 0 \le t \le LT_b \\ \dfrac{\pi}{2}\left[1 - 2q\left(t - LT_b\right)\right], & LT_b \le t \end{cases} \tag{2.8-69}$$

which is obtained by taking the nonconstant part of the phase pulse, $q(t)$, that exists in the interval $0 \le t \le LT_b$ and reflecting it about the $t = LT_b$ axis.[21] Therefore, in view of (2.8-69), $\Psi(t)$ is a waveform that is nonzero in the interval $0 \le t \le 2LT_b$ and symmetric around $t = LT_b$. The symmetry around $t = LT_b$ assumes that the frequency pulse, $g\left(t\right)$, is even symmetric around $t = LT_b/2$ and, thus, the phase pulse $q\left(t\right)$ is odd symmetric around the value $\pi/4$ at $t = LT_b/2$. Next define

$$\left.\begin{aligned} S_0(t) &\triangleq \sin\Psi(t) \\ S_n(t) &\triangleq S_0(t + nT) = \sin\Psi(t + nT) \end{aligned}\right\} \tag{2.8-70}$$

Finally,

[21] For the Laurent representation, it is convenient to shift the frequency pulse of (2.8-54) to the interval $0 \le t \le LT_b$ before integrating it to get the phase pulse, $q(t)$.

$$C_K(t) = S_0(t) \prod_{i=1}^{L-1} S_{i+L\beta_{K,i}}(t), \quad 0 \le K \le 2^{L-1} - 1, \quad 0 \le t \le T_{bK},$$

$$T_{bK} = T_b \times \min_{i=1,2,\cdots,L-1} \left[L\left(2 - \beta_{K,i}\right) - i \right] \quad (2.8\text{-}71)$$

where $\beta_{K,i}, i = 1, 2, \cdots, L-1$ are the coefficients in the binary representation of the integer K, i.e.,

$$K = \sum_{i=1}^{L-1} 2^{i-1} \beta_{K,i} \quad (2.8\text{-}72)$$

Note from (2.8-71) that each of the equivalent pulse waveforms, $C_K(t)$, in general have different durations, and, consequently, the pulse streams in (2.8-68) consist of overlapping pulses.

The complex phase coefficient $\tilde{a}_{K,n} \triangleq e^{j(\pi/2)A_{K,n}}$ associated with the nth T-s translate of this Kth pulse shape, namely $C_K(t - nT)$, is also expressible in terms of the binary representation of the integer K as given in (2.8-72). In particular,

$$\left.\begin{aligned} A_{K,n} &= \sum_{i=-\infty}^{n} \alpha_i - \sum_{i=1}^{L-1} \alpha_{n-i}\beta_{K,i} = A_{0,n} - \sum_{i=1}^{L-1} \alpha_{n-i}\beta_{K,i} \\ A_{0,n} &= \alpha_n + A_{0,n-1} \end{aligned}\right\} \quad (2.8\text{-}73)$$

and thus,

$$\tilde{a}_{K,n} \triangleq e^{j(\pi/2)A_{K,n}} = \exp\left[j\frac{\pi}{2} \left(A_{0,n-L} + \sum_{i=0}^{L-1} \alpha_{n-i} - \sum_{i=1}^{L-1} \alpha_{n-i}\beta_{K,i} \right) \right]$$

$$= \tilde{a}_{0,n-L} e^{j(\pi/2)\alpha_n} \prod_{i=1}^{L-1} e^{j(\pi/2)\alpha_{n-i}[1-\beta_{K,i}]} \quad (2.8\text{-}74)$$

Before proceeding further, we present an example corresponding to a particular value of L to illustrate the above description of the representation. Consider the case of $L = 4$, which as previously mentioned, is adequate to represent GMSK with $BT_b \ge 0.25$. Therefore, from (2.8-71), there are $2^{L-1} = 8$ different $C_K(t)$'s,

i.e., $C_0(t), C_1(t), \cdots, C_7(t)$, each of which is a product of the basic generalized pulse shape $S_0(t)$ and $L-1=3$ other $S_i(t)$'s with the particular ones being chosen according to the coefficients in the binary representation of the index, K. For example, for $K=3$, we would have

$$K = 3 = 2^0 \times 1 + 2^1 \times 1 + 2^2 \times 0 \quad \Rightarrow \quad \beta_{3,1} = 1, \beta_{3,2} = 1, \beta_{3,3} = 0 \tag{2.8-75}$$

and thus,

$$C_3(t) = S_0(t) \prod_{i=1}^{3} S_{i+4\beta_{3,i}}(t) = S_0(t)S_5(t)S_6(t)S_3(t), \quad 0 \le t \le T_{b3} = 2T_b \tag{2.8-76}$$

In summary,

$$\left.\begin{aligned}
C_0(t) &= S_0(t)S_1(t)S_2(t)S_3(t), \quad 0 \le t \le 5T_b \\
C_1(t) &= S_0(t)S_2(t)S_3(t)S_5(t), \quad 0 \le t \le 3T_b \\
C_2(t) &= S_0(t)S_1(t)S_3(t)S_6(t), \quad 0 \le t \le 2T_b \\
C_3(t) &= S_0(t)S_3(t)S_5(t)S_6(t), \quad 0 \le t \le 2T_b \\
C_4(t) &= S_0(t)S_1(t)S_2(t)S_7(t), \quad 0 \le t \le T_b \\
C_5(t) &= S_0(t)S_2(t)S_5(t)S_7(t), \quad 0 \le t \le T_b \\
C_6(t) &= S_0(t)S_1(t)S_6(t)S_7(t), \quad 0 \le t \le T_b \\
C_7(t) &= S_0(t)S_5(t)S_6(t)S_7(t), \quad 0 \le t \le T_b
\end{aligned}\right\} \tag{2.8-77}$$

From (2.8-74) the set of complex phase coefficients for the third pulse train corresponding to $C_3(t)$ of (2.8-76) would be

$$\tilde{a}_{3,n} = \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} \prod_{i=1}^{3} e^{j(\pi/2)\alpha_{n-i}[1-\beta_{3,i}]} = \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-3}} \tag{2.8-78}$$

The complete group of phase coefficient sets for all eight pulse trains is given by (also see [46, Eq. (A.19)])

$$\left.\begin{aligned}
\tilde{a}_{0,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-1}} e^{j(\pi/2)\alpha_{n-2}} e^{j(\pi/2)\alpha_{n-3}} \\
&= j^{\alpha_n+\alpha_{n-1}+\alpha_{n-2}+\alpha_{n-3}} \tilde{a}_{0,n-4} = j^{\alpha_n} \tilde{a}_{0,n-1} \\
\tilde{a}_{1,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-2}} e^{j(\pi/2)\alpha_{n-3}} \\
&= j^{\alpha_n+\alpha_{n-2}+\alpha_{n-3}} \tilde{a}_{0,n-4} = j^{\alpha_n} \tilde{a}_{0,n-2} \\
\tilde{a}_{2,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-1}} e^{j(\pi/2)\alpha_{n-3}} \\
&= j^{\alpha_n+\alpha_{n-1}+\alpha_{n-3}} \tilde{a}_{0,n-4} = j^{\alpha_n+\alpha_{n-1}} \tilde{a}_{0,n-3} \\
\tilde{a}_{3,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-3}} = j^{\alpha_n+\alpha_{n-3}} \tilde{a}_{0,n-4} = j^{\alpha_n} \tilde{a}_{0,n-3} \\
\tilde{a}_{4,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-1}} e^{j(\pi/2)\alpha_{n-2}} = j^{\alpha_n+\alpha_{n-1}+\alpha_{n-2}} \tilde{a}_{0,n-4} \\
\tilde{a}_{5,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-2}} = j^{\alpha_n+\alpha_{n-2}} \tilde{a}_{0,n-4} \\
\tilde{a}_{6,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} e^{j(\pi/2)\alpha_{n-1}} = j^{\alpha_n+\alpha_{n-1}} \tilde{a}_{0,n-4} \\
\tilde{a}_{7,n} &= \tilde{a}_{0,n-4} e^{j(\pi/2)\alpha_n} = j^{\alpha_n} \tilde{a}_{0,n-4}
\end{aligned}\right\} \quad (2.8\text{-}79)$$

It is to be emphasized that to the extent that GMSK can be approximated by a partial-response CPM with finite L, the AMP representation is *exact*. For the case of $L = 4$, Ref. 46 states that, based on computer simulations, the first AMP component corresponding to the pulse stream $\{C_0(t - nT)\}$ contains the fraction 0.991944 of the total signal energy, and the second component corresponding to the pulse stream $\{C_1(t - nT)\}$ contains the fraction 0.00803 of the total energy. Thus, the remaining six components contain only the fraction 2.63×10^{-5} of the total signal energy, and thus, to a good approximation can be ignored. Hence, we conclude that for values of BT_b, where $L = 4$ is an appropriate value for the truncation of the frequency pulse, *a two pulse stream AMP representation corresponding to $K = 1$ and $K = 2$ is sufficient to approximate GMSK*, i.e.,

$$\tilde{S}_{\text{GMSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \left[\sum_{n=-\infty}^{\infty} \tilde{a}_{0,n} C_0\left(t - nT_b\right) + \sum_{n=-\infty}^{\infty} \tilde{a}_{1,n} C_1\left(t - nT_b\right) \right] \quad (2.8\text{-}80)$$

where $C_0(t)$ and $C_1(t)$ are determined from the first two equations in (2.8-77) (see Fig. 2-32 for an illustration of these two waveforms) and, likewise, $\tilde{a}_{0,n}$ and $\tilde{a}_{1,n}$ are determined from the first two equations of (2.8-79). Since the actual data symbols, $\{\alpha_n\}$, range over the values ± 1, then the even and odd complex symbols for each of the two pulse streams take on values

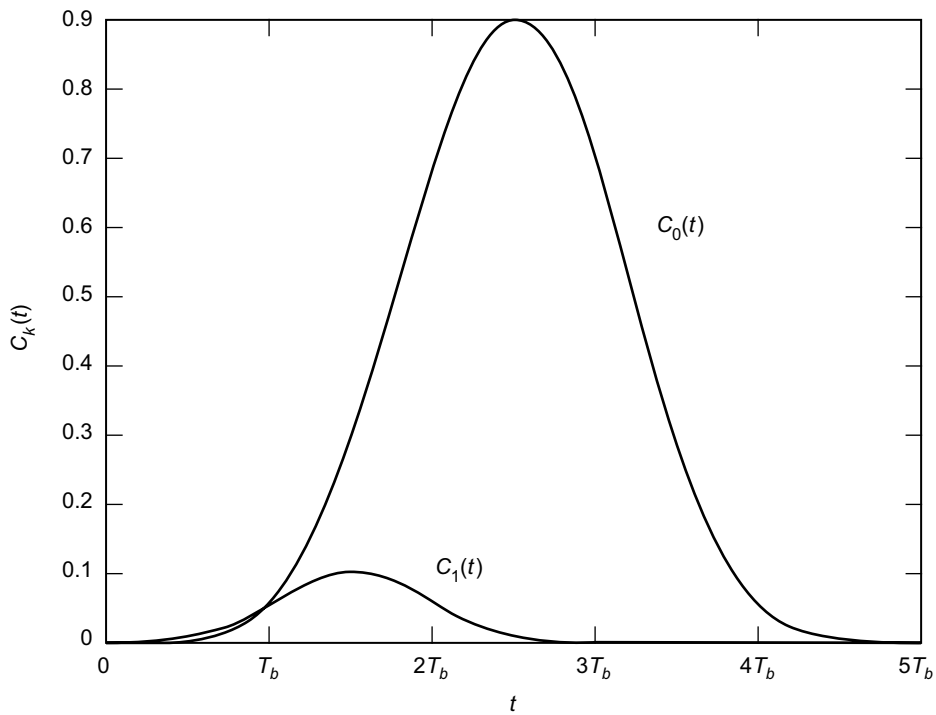


Fig. 2-32. Pulse shapes for first and second AMP streams.

$$\left.\begin{array}{ll} \{\tilde{a}_{0,2n}\} \in \{j,-j\}\,, & \{\tilde{a}_{0,2n+1}\} \in \{1,-1\} \\ \{\tilde{a}_{1,2n}\} \in \{1,-1\}\,, & \{\tilde{a}_{1,2n+1}\} \in \{j,-j\} \end{array}\right\} \tag{2.8-81}$$

which clearly indicate a representation composed of the superposition of two I-Q signals. Note that the sequences $\{\tilde{a}_{0,n}\}$ and $\{\tilde{a}_{1,n}\}$ are themselves uncorrelated as well as being mutually uncorrelated, viz.,

$$\begin{aligned} E\left\{\tilde{a}_{0,k}\tilde{a}^*_{1,k-m}\right\} &= E\left\{j\alpha_k j\alpha_{k-1}\cdots j\alpha_{k-m-1}\tilde{a}_{0,k-m-2} \times -j\alpha_{k-m}\tilde{a}^*_{0,k-m-2}\right\} \\ &= \pm E\left\{\alpha_k\alpha_{k-1}\cdots\alpha_{k-m-1}\alpha_{k-m}\right\} = 0, \quad m > 0 \end{aligned} \tag{2.8-82}$$

Furthermore, since for binary ± 1 data, $j^{\alpha_n} = j\alpha_n$, then the first two equations of (2.8-79) become

$$\left.\begin{aligned} \tilde{a}_{0,n} &= j\alpha_n \tilde{a}_{0,n-1} \\ \tilde{a}_{1,n} &= j\alpha_n \tilde{a}_{0,n-2} \end{aligned}\right\} \tag{2.8-83}$$

which clearly identifies the fact that the complex symbols for the two pulse streams are obtained from a *differentially encoded* version of the input data. Finally, the corresponding real (± 1) symbols on the I and Q channels for the two pulse streams are

$$\left.\begin{aligned} a_{0,n}^{I} &= \tilde{a}_{0,2n+1} = \mathrm{Re}\left\{\tilde{a}_{0,2n+1}\right\} \\ a_{0,n}^{Q} &= -j\tilde{a}_{0,2n} = \mathrm{Im}\left\{\tilde{a}_{0,2n}\right\} \\ a_{1,n}^{Q} &= -j\tilde{a}_{1,2n+1} = \mathrm{Im}\left\{\tilde{a}_{1,2n+1}\right\} \\ a_{1,n}^{I} &= \tilde{a}_{1,2n} = \mathrm{Re}\left\{\tilde{a}_{1,2n}\right\} \end{aligned}\right\} \tag{2.8-84}$$

and, hence, the real GMSK two pulse stream approximation corresponding to (2.8-80) is

$$\begin{aligned} s_{\mathrm{GMSK}}(t) = \sqrt{\frac{2E_b}{T_b}} \Bigg[& \sum_{n=-\infty}^{\infty} a_{0,n}^{I} C_0\left(t-(2n+1)T_b\right)\cos 2\pi f_c t \\ & - \sum_{n=-\infty}^{\infty} a_{0,n}^{Q} C_0\left(t-2nT_b\right)\sin 2\pi f_c t \\ & + \sum_{n=-\infty}^{\infty} a_{1,n}^{I} C_1\left(t-2nT_b\right)\cos 2\pi f_c t \\ & - \sum_{n=-\infty}^{\infty} a_{1,n}^{Q} C_1\left(t-(2n+1)T_b\right)\sin 2\pi f_c t \Bigg] \end{aligned} \tag{2.8-85}$$

which has the implementation of Fig. 2-33. Note that each of the pulse streams is in the form of a pulse-shaped OQPSK modulation with overlapping pulses and effective symbol rate $T_s = 2T_b$ on each of the quadrature channels. Also, the encoding of the first pulse stream is a conventional differential encoder whereas the second pulse stream is generated from a product of the input data stream and a delayed version of the differentially encoded output of the first stream. Therefore, from a data encoding standpoint, the first pulse stream resembles MSK whereas the second does not.

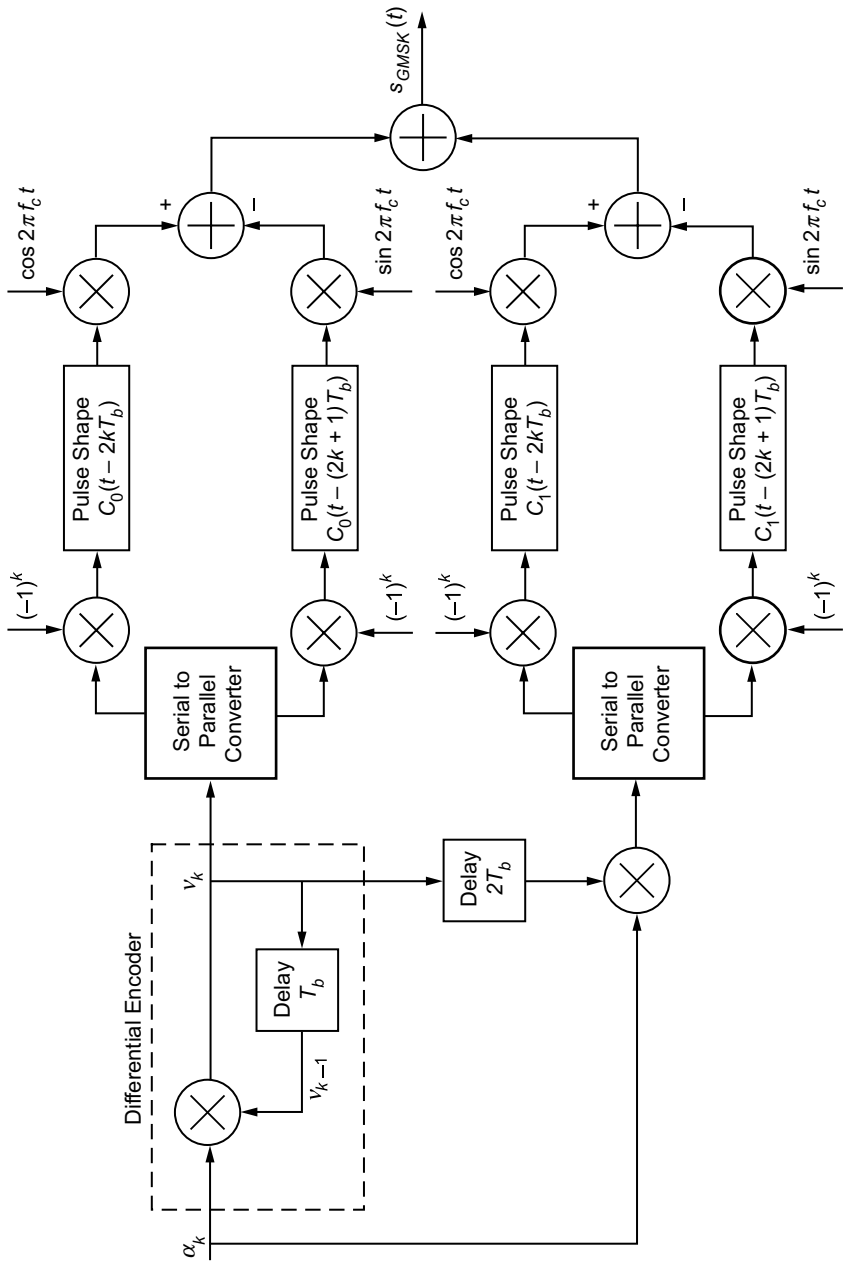


Fig. 2-33. Two-pulse stream I-Q implementation of GMSK.

b. Precoded GMSK. Because of the different encodings for the first and second pulse streams, we can only fully compensate for one of the two with a precoding operation. Thus, proceeding as in the MSK case, if we precede the GMSK modulator with a differential decoder [see Fig. 2-34(a)], then, as was true for MSK, the first pulse stream of the equivalent I-Q representation would no longer have a differential encoder at its front end. The effect of the precoding on the second pulse stream of the equivalent I-Q representation is to produce a particular feed-forward type of encoding [see Fig. 2-34(b)] that can be shown to be equivalent to a two-stage differential decoder (see Fig. 2-35). Such precoded GMSK has been considered by several authors in the literature [54–57] as a means of improving receiver performance.

2.8.2.4 Power Spectral Density Considerations. As mentioned above, one advantage of the Laurent representation is that it provides a simple means of computing the PSD. In particular, since the various pulse-stream equivalent data sequences are both self- and mutually uncorrelated, for the GMSK signal with complex form as in (2.8-67), the PSD is simply

$$S(f) = E_b \sum_{k=0}^{2^{L-1}-1} \frac{1}{T_b} \left|\mathcal{C}_k(f)\right|^2, \quad \mathcal{C}_k(f) = \mathcal{F}\{C_k(t)\} \tag{2.8-86}$$

or for $L = 4$ and the two pulse stream approximation of (2.8-80),

$$S(f) = \frac{E_b}{T_b}\left[\left|\mathcal{C}_0(f)\right|^2 + \left|\mathcal{C}_1(f)\right|^2\right] \tag{2.8-87}$$

Figure 2-36 is a plot of the normalized (all curves start at zero decibels at zero frequency) GMSK PSD as computed from (2.8-86), with frequency pulse length in bits, L, as a parameter. The values of BT_b have been chosen equal to the reciprocal of L. Thus, for example, a value of $L = 4$ results in a curve for $BT_b = 0.25$ that corresponds to the special case previously considered. Comparing

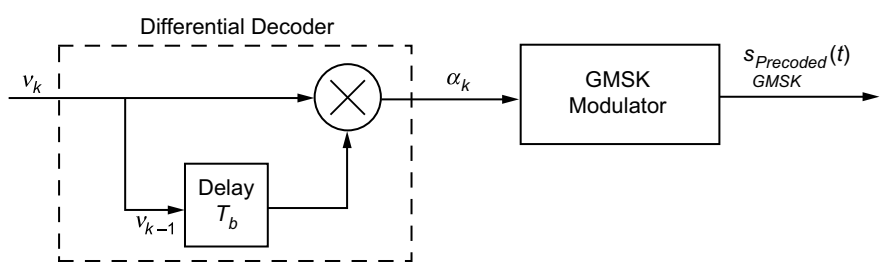


Fig. 2-34(a). Precoded GMSK transmitter.

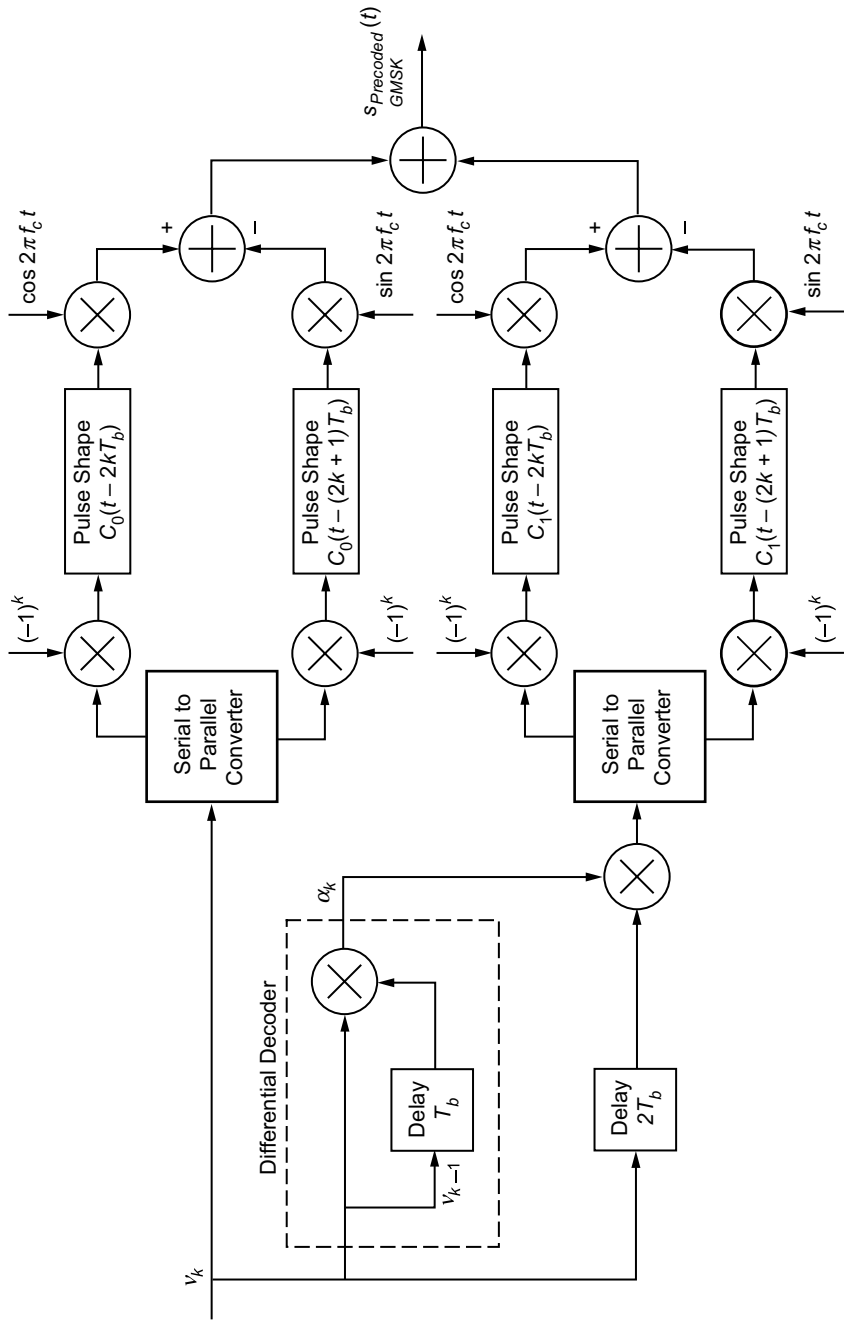


Fig. 2-34(b). Two-pulse stream I-Q implementation of precoded GMSK.

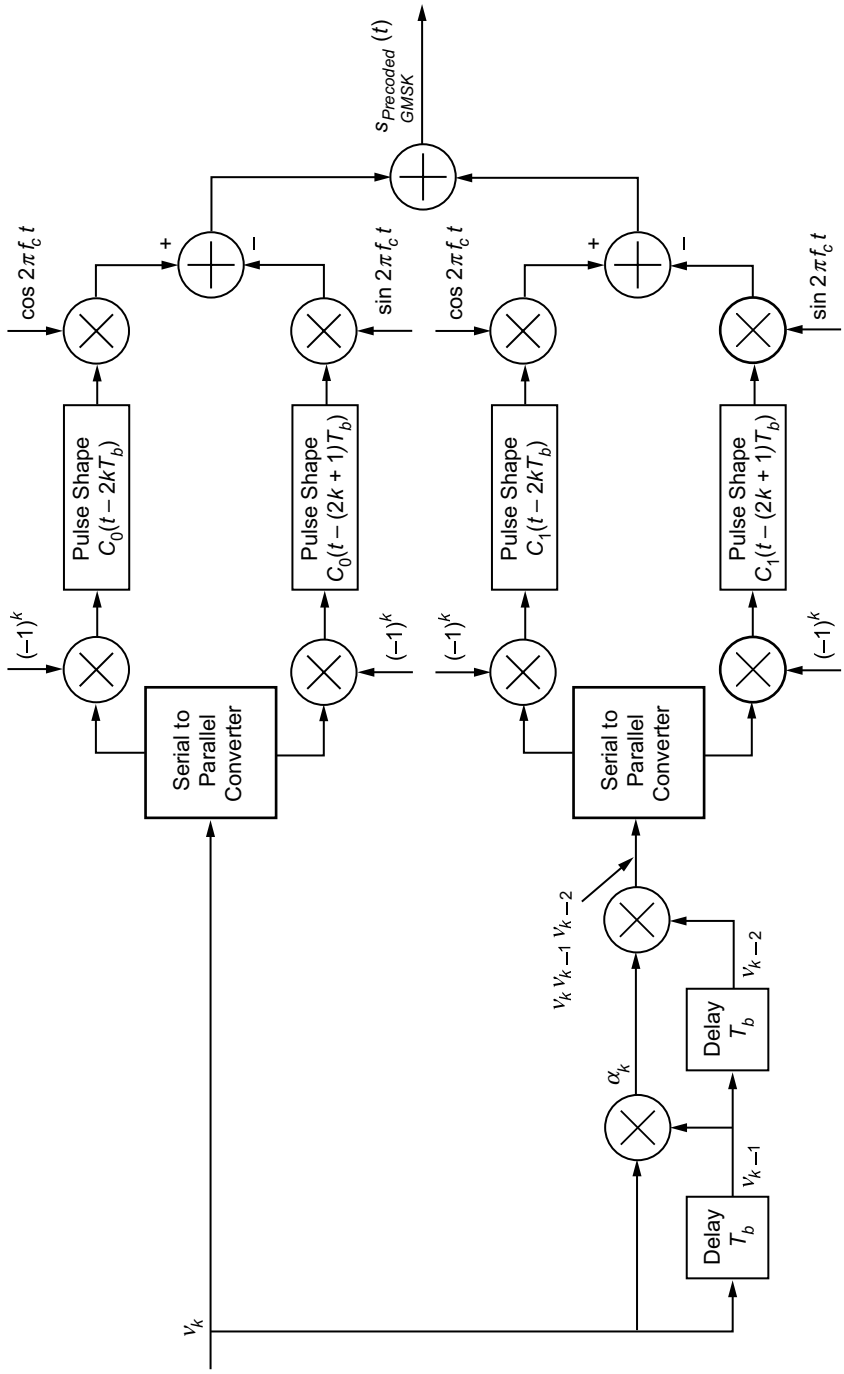


Fig. 2-35. Alternative two-pulse stream I-Q implementation of precoded GMSK.

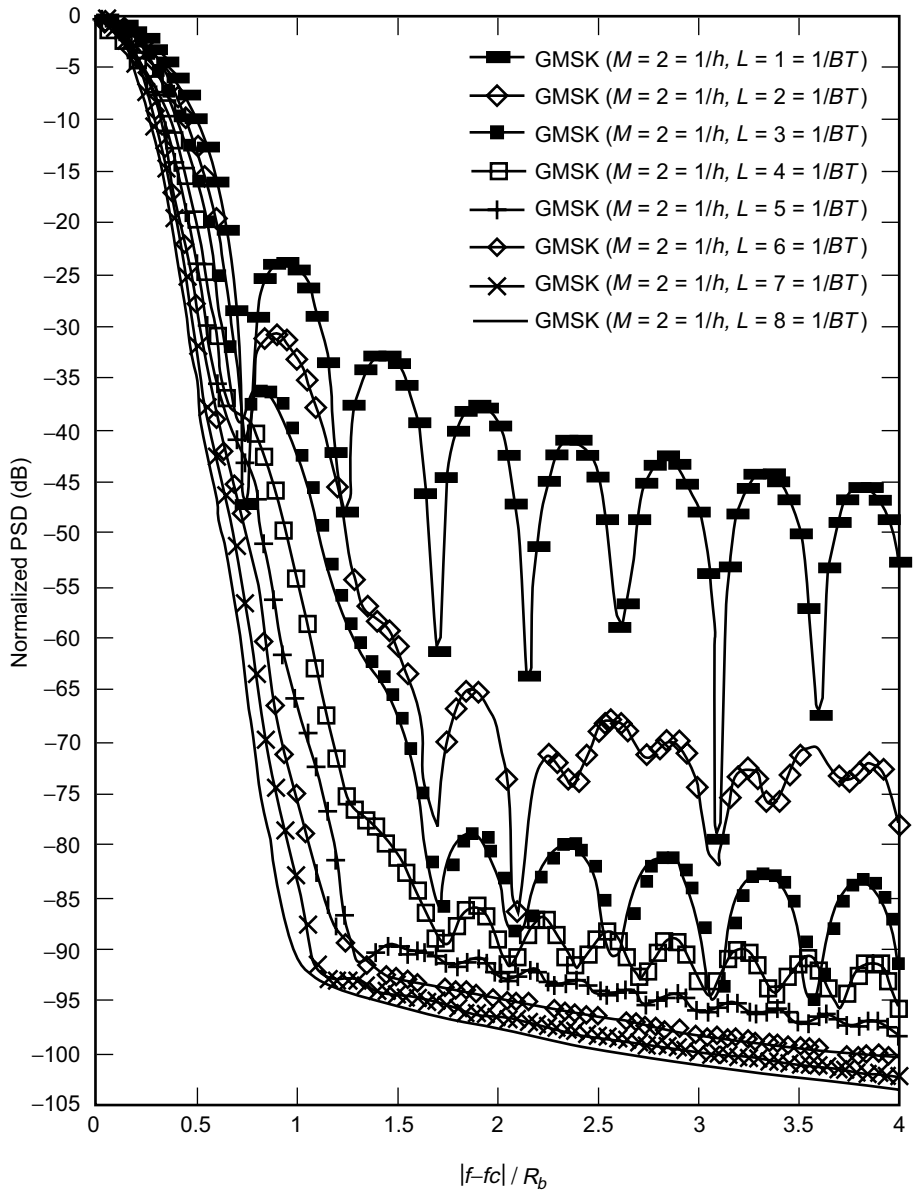


Fig. 2-36. Power spectral density of GMSK with BT_b as a parameter. Redrawn from [54].

Fig. 2-36 with Fig. 2-15, we observe the potentially significant improvement in spectral efficiency of the partial-response CPM modulation (GMSK) over the full-response CPM modulation (MSK), when the value of BT_b is sufficiently less than unity. Finally, we note that the PSD of precoded GMSK is identical to that of GMSK.

2.8.2.5 Approximate AMP Representation of GMSK Based on a Single Pulse Stream. Before moving on to a discussion of the various types of receivers that have been proposed for GMSK, it is instructive to further approximate (simplify) the AMP representation, since the structure of one of these receivers stems from this approximation. In the AMP representation of (2.8-67) or (2.8-68), the dominant term is the pulse stream corresponding to $C_0(t)$ (for full-response CPM, i.e., $L = 1$, it would be the only one), since its duration is the longest (at least $2T_b$ longer than any other pulse component) and it also conveys the most significant part of the total energy of the signal. (Although Laurent never proves the latter mathematically, it appears to be the case for all practical CPM scenarios.) Thus, approximating the AMP representation with a single pulse stream, which is an exact representation for MSK, is a reasonable thing to do. As such, we propose an approximation of (2.8-67), where the pulse shape, $P(t)$ (called the "main pulse" in Ref. 52), used for the single-stream AMP representation should have the same phase shift as that associated with $C_0(t)$ and must satisfy some optimization criterion in the sense of being the best approximation of the signal, viz.,

$$\hat{\tilde{S}}(t) = \sum_{n=-\infty}^{\infty} e^{j(\pi/2)A_{0,n}} P(t - nT_b) = \sum_{n=-\infty}^{\infty} e^{j(\pi/2)\sum_{i=0}^{n}\alpha_i} P(t - nT_b) \quad (2.8\text{-}88)$$

The optimization criterion selected by Laurent consisted of minimizing the average energy of the difference between the complete signal and its approximation. Two methods are proposed in Ref. 52 for solving this optimization problem in the general case of CPM with modulation index, h, the second of which is preferred because it illustrates the important properties of the main pulse. In particular, $P(t)$ is expressed as a weighted superposition of time-shifted versions of the finite duration components, $C_i(t)$. It is further shown in Ref. 52, that regardless of the value of L, for $h = 0.5$ (as is the case for GMSK), the main pulse is simply given by $C_0(t)$. Hence, we conclude that using only the first AMP component of the signal is the best—and naturally the simplest—possible approximation in the above mean-square energy sense.

2.8.2.6 Coherent GMSK Receivers and Their Performance. A variety of different types of receivers [46,54,56,57] have been proposed for coherent

detection of GMSK; most of them are based on the Laurent representation and employ the Viterbi Algorithm (VA) [58]. To start the discussion, we consider first the optimum receiver based upon an L-bit duration GMSK frequency pulse.

Because of the memory inherent in CPM, regardless of its mathematical representation, the optimum receiver (from the standpoint of minimizing the message error probability) has the form of an MLSE which is typically implemented using the VA. This receiver employs $m = 2^{L-1} - 1$ filters matched to each of the m pulse shapes in the complex baseband AMP representation of (2.8-67). These filters act on the received complex signal plus noise, and their outputs are inputted to a VA whose decision metric is based upon the equivalent data stream encodings of (2.8-79) (see Fig. 2-37). The number of states in the trellis diagram characterizing the receiver is equal to 2^L, e.g., $L = 4$ would require a 16-state trellis.

a. Optimum Receiver. When the GMSK signal of (2.8-68) is transmitted over an AWGN channel, the received signal is given by

$$z(t) = s(t) + n(t) \tag{2.8-89}$$

where $n(t)$ is as before a zero mean Gaussian process, independent of the signal, with single-sided PSD equal to N_0 watts/hertz. Since for a length N data

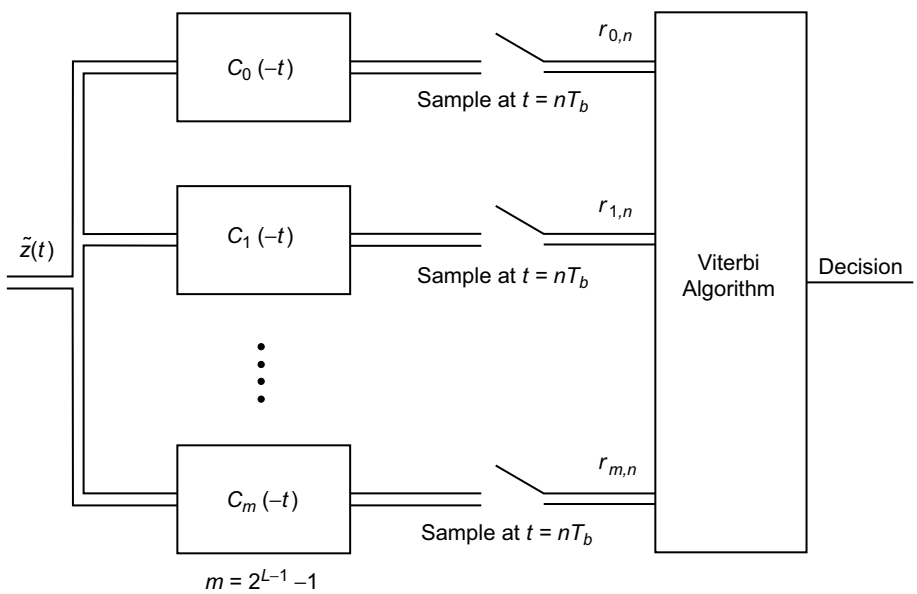


Fig. 2-37. Optimum GMSK receiver.

sequence, all of the possible 2^N transmitted signals have equal energy and are equally likely, the optimum receiver that minimizes the message (sequence) error probability chooses that message, i, that maximizes the metric

$$\Lambda_i = 2\int_{-\infty}^{\infty} z(t)s_i(t)dt = \mathrm{Re}\left\{\int_{-\infty}^{\infty} \tilde{Z}(t)\tilde{S}_i^*(t)dt\right\} \tag{2.8-90}$$

where the second equality ignores the second harmonic carrier term, $s_i(t)$ is the signal corresponding to the ith data sequence with complex envelope $\tilde{S}_i(t)$, and analogous to (2.8-65)

$$z(t) = \mathrm{Re}\left\{\tilde{Z}(t)e^{j2\pi f_c t}\right\} \tag{2.8-91}$$

Substituting (2.8-67) in (2.8-90) yields an additive form for the metric, namely,

$$\Lambda_i = \sqrt{\frac{2E_b}{T_b}}\sum_{n=0}^{N-1}\lambda_i(n) \tag{2.8-92}$$

where $\lambda_i(n)$ is the trellis branch metric given by

$$\begin{aligned}\lambda_i(n) &= \mathrm{Re}\left\{\sum_{K=0}^{2^{L-1}-1} \tilde{a}_{K,n}^{i}{}^{*}\int_{-\infty}^{\infty}\tilde{Z}(t)C_K\left(t-nT_b\right)dt\right\}\\ &\triangleq \mathrm{Re}\left\{\sum_{K=0}^{2^{L-1}-1} \tilde{a}_{K,n}^{i}{}^{*} r_{K,n}\right\}\end{aligned} \tag{2.8-93}$$

The superscript "i" on the equivalent complex data symbols denotes the ith data sequence, i.e., these are the N symbols corresponding to the signal, $s_i(t)$. The correlation values

$$r_{K,n} = \int_{-\infty}^{\infty}\tilde{Z}(t)C_K\left(t-nT_b\right)dt, \quad 0 \le K \le 2^{L-1}-1, \quad 0 \le n \le N \tag{2.8-94}$$

are thus sufficient statistics for making the message decision and can be obtained for any fixed n by sampling the outputs of the bank of 2^{L-1} matched filters in Fig. 2-37 at time $t = nT_b$.

Computation of an nth branch metric requires knowledge of the equivalent complex data sequence $\{\tilde{a}_{K,n}\}$. This in turn can be found from the current data symbol α_n and a state defined by the vector $(a_{0,n-L}, \alpha_{n-L+1}, \alpha_{n-L+2}, \cdots, \alpha_{n-2}, \alpha_{n-1})$. Therefore, the decision rule can be implemented by inputting the set of matched filter output samples to a VA using the above state definition and current symbol to define the trellis states and transitions between them. The complexity of the VA is proportional to the number of states, which as previously mentioned, is equal to 2^L.

b. Simplified (Suboptimum) GMSK Receivers. Using the approximate AMP representation discussed in Sec. 2.8.2.5, Kaleh [46] first derived a reduced-complexity Viterbi detector that achieved near-optimal performance. By "reduced-complexity," we mean that the number of matched filters and VA states is appreciably smaller than would be required for the truly optimum receiver. In particular, a receiver consisting of only two matched filters and a four-state VA resulted in a performance degradation of less than 0.24 dB relative to the optimum and much more complex receiver. In addition to the simplification of the optimum receiver based on an error probability criterion, Kaleh also considered an optimum coherent linear receiver based on a minimum mean-square error (MMSE) criterion. Such a receiver has a generic form analogous to that used for detection of MSK (as such it was referred to in Ref. 46 as an MSK-type receiver) except for the fact that the receive filter is composed now of a combination of matched and Wiener-type filters. In what follows, we explore these two receiver options.

The complexity of the optimum MLSE receiver can be reduced by approximating the AMP representation with a smaller number of pulse streams, as previously discussed. In particular, consider replacing the 2^{L-1} pulse streams in (2.8-67) by the first $\hat{K}$ of them, where $\hat{K}$ is chosen such that the remaining components cumulatively have very small energy [(2.8-80) is a particular example of this where $\hat{K} = 2$]. As such, we can write the transmitted signal in the approximate (simplified) complex baseband form

$$\hat{\tilde{S}}(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{K=0}^{\hat{K}-1} \left[\sum_{n=-\infty}^{\infty} e^{j\pi h A_{K,n}} C_K\left(t - nT_b\right) \right]$$

$$\triangleq \sqrt{\frac{2E_b}{T_b}} \sum_{K=0}^{\hat{K}-1} \left[\sum_{n=-\infty}^{\infty} \tilde{a}_{K,n} C_K\left(t - nT_b\right) \right] \tag{2.8-95}$$

where the "hat" is used to denote approximation. Then, in accorandance with (2.8-92) the simplified receiver computes the approximate metric

$$\hat{\Lambda}_i = \sqrt{\frac{2E_b}{T_b}} \sum_{n=0}^{N-1} \hat{\lambda}_i(n) \tag{2.8-96}$$

where

$$\hat{\lambda}_i(n) = \mathrm{Re}\left\{ \sum_{K=0}^{\hat{K}-1} \tilde{a}_{K,n}^{i}{}^{*} \int_{-\infty}^{\infty} \hat{\tilde{Z}}(t) C_K\left(t - nT_b\right) dt \right\} \triangleq \mathrm{Re}\left\{ \sum_{K=0}^{\hat{K}-1} \tilde{a}_{K,n}^{i}{}^{*} r_{K,n} \right\} \tag{2.8-97}$$

and $\hat{\tilde{Z}}(t)$ corresponds to the received version of $\hat{\tilde{S}}(t)$. Since $r_{K,n}, K = \hat{K}, \hat{K} + 1, \cdots, 2^{L-1} - 1$ are considered as irrelevant, the number of matched filters needed in Fig. 2-37 is reduced from 2^{L-1} to $\hat{K}$. Also, a great reduction in the complexity of the VA is achieved, since the number of states can accordingly be reduced from 2^L to $2\hat{K}$.

Pursuing now in detail the case where $\hat{K} = 2$, the even branch metrics in (2.8-97) are given by

$$\begin{aligned}
\hat{\lambda}_i(2n) &= \mathrm{Re}\left\{ \tilde{a}_{0,2n}^{i}{}^{*} r_{0,2n} + \tilde{a}_{1,2n}^{i}{}^{*} r_{1,2n} \right\} \\
&= \mathrm{Re}\left\{ a_{0,2n}^{i}{}^{*} r_{0,2n} \right\} + \mathrm{Re}\left\{ -j\alpha_{2n}^{i} \tilde{a}_{0,2n-2}^{i}{}^{*} r_{1,2n} \right\} \\
&= a_{0,2n}^{i} \mathrm{Re}\left\{ r_{0,2n} \right\} + \mathrm{Re}\left\{ -\frac{\tilde{a}_{0,2n}^{i}}{\tilde{a}_{0,2n-1}^{i}} \tilde{a}_{0,2n-2}^{i}{}^{*} r_{1,2n} \right\} \\
&= a_{0,2n}^{i} \mathrm{Re}\left\{ r_{0,2n} \right\} + \mathrm{Re}\left\{ -a_{0,2n}^{i} j a_{0,2n-1}^{i} a_{0,2n-2}^{i} r_{1,2n} \right\} \\
&= \overbrace{a_{0,2n}^{i}}^{\text{current bit}} \mathrm{Re}\left\{ r_{0,2n} \right\} + \overbrace{a_{0,2n}^{i}}^{\text{current bit}} \overbrace{a_{0,2n-1}^{i} a_{0,2n-2}^{i}}^{\text{2 previous bits}} \mathrm{Im}\left\{ r_{1,2n} \right\}
\end{aligned} \tag{2.8-98a}$$

whereas the odd branch metrics are given by

$$\hat{\lambda}_i(2n-1) = \overbrace{a_{0,2n-1}^{i}}^{\text{current bit}} \mathrm{Im}\left\{ r_{0,2n-1} \right\} - \overbrace{a_{0,2n-1}^{i}}^{\text{current bit}} \overbrace{a_{0,2n-2}^{i} a_{0,2n-3}^{i}}^{\text{2 previous bits}} \mathrm{Re}\left\{ r_{1,2n-1} \right\} \tag{2.8-98b}$$

Consequently, at any time, nT_b, we see that the state vector is defined by $a^i_{0,n-1}a^i_{0,n-2}$, i.e., the two previous equivalent real bits, which results in the four-state trellis illustrated in Fig. 2-38. The VA makes decisions, $\hat{a}_n$, on the real equivalent bits, a_n, from which the decisions on the actual transmitted bits are obtained from the differential decoding operation

$$\left.\begin{aligned} \hat{\alpha}_{2n} &= -\hat{a}_{2n}\hat{a}_{2n-1} \\ \hat{\alpha}_{2n+1} &= \hat{a}_{2n+1}\hat{a}_{2n} \end{aligned}\right\} \tag{2.8-99}$$

The performance of the simplified Viterbi receiver was computed in Ref. 46, based on an upper bound obtained from the minimum Euclidean distance of the signaling set. In particular, it is well known that for modulations characterized by a trellis-type decoding algorithm, the bit error probability is upper bounded by

$$P_b(E) \leq CQ\left(\frac{d_{\min}}{\sqrt{2N_0}}\right) \tag{2.8-100}$$

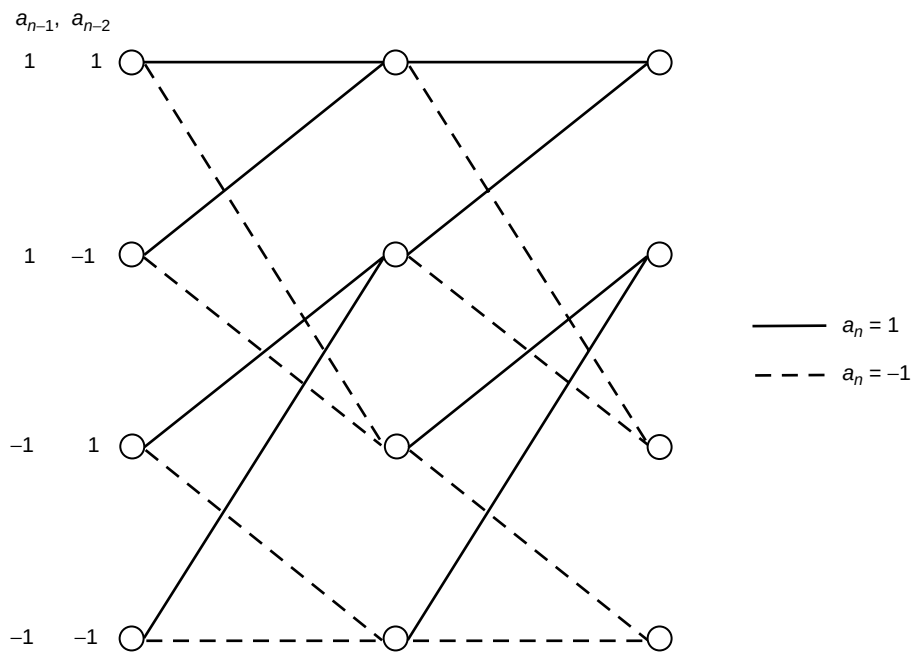


Fig. 2-38. Trellis diagram for a simplified Viterbi receiver for GMSK; $BT_b = 0.25$.

where C is a constant that depends on the number of nearest neighbors in the constellation to the transmitted signal and $d_{\min}$ is the minimum Euclidean distance between transmitted signals. Using this measure of performance, it was shown in Ref. 46 that *the simplified Viterbi receiver that uses two matched filters and a 4-state VA has a degradation of less than 0.24 dB when compared to the optimum Viterbi receiver that requires 8 matched filters and a 16-state VA.*

Next, we consider the implementation of simple MSK-type linear receivers for GMSK (see Fig. 2-39). Such receivers are memoryless and make bit-by-bit decisions on the transmitted data. In the case of true MSK, the receiver operates in the absence of ISI and, thus, the receive filter is merely the matched filter to the transmitted amplitude pulse shape, i.e., $C_0(t) = S_0(t) = \sin \pi t/2T_b$. Even in the case of generalized MSK with $h = 0.5$, the receive filter still operates in the absence of ISI with a matched receive filter in accordance with $C_0(t)$, which from (2.8-69) and (2.8-70) would be now given by

$$C_0(t) = S_0(t) = \begin{cases} \sin\big(\pi q(t)\big), & 0 \le t \le T_b \\ \sin\left(\dfrac{\pi}{2}\big[1 - 2q(t - T_b)\big]\right), & T_b \le t \le 2T_b \end{cases} \tag{2.8-101}$$

Before showing how such a receiver must be modified for GMSK, we first review its application to MSK.

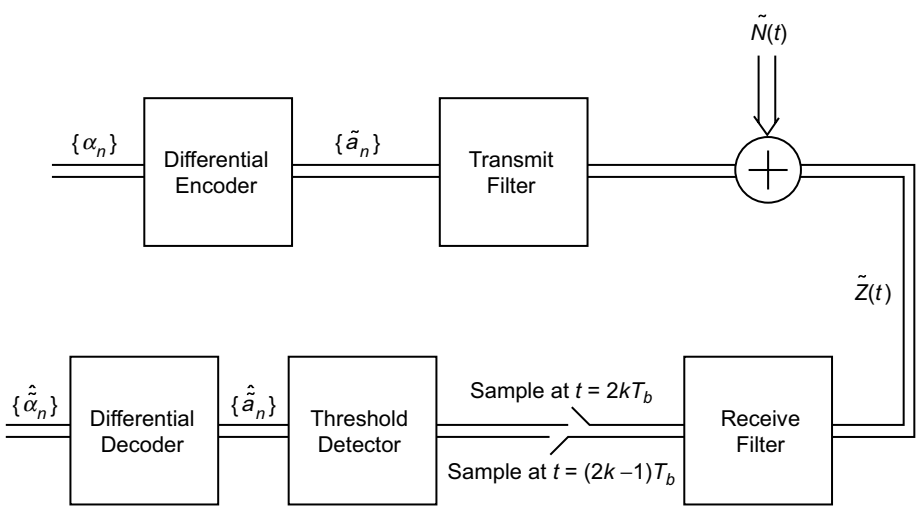


Fig. 2-39. Baseband model of an MSK-type system for GMSK.

For MSK with $L = 1$, the branch metric of (2.8-93) would simplify to

$$\lambda_i(n) = \mathrm{Re}\left\{\tilde{a}_{0,n}^{i}{}^{*} r_{0,n}\right\} = \mathrm{Re}\left\{\tilde{a}_{0,n}^{i}\right\}\mathrm{Re}\left\{r_{0,n}\right\} + \mathrm{Im}\left\{\tilde{a}_{0,n}^{i}\right\}\mathrm{Im}\left\{r_{0,n}\right\}$$

$$= \begin{cases} a_{0,n}\mathrm{Re}\left\{r_{0,n}\right\}, & n \text{ even} \\ a_{0,n}\mathrm{Im}\left\{r_{0,n}\right\}, & n \text{ odd} \end{cases} \tag{2.8-102}$$

where

$$r_{0,n} = \int_{-\infty}^{\infty} \tilde{Z}(t) C_0\left(t - nT_b\right) dt, \quad 0 \le n \le N \tag{2.8-103}$$

which is the output of a single filter matched to $C_0(t)$ and sampled at time $t = nT_b$. Since the branch metric only depends on the current symbol, a memoryless receiver is appropriate for making decisions on the equivalent real data bits, $\{a_{0,n}\}$. Therefore, if we alternately sample the real and imaginary parts of the matched filter output at intervals of T_b s, we obtain ISI-free decisions on these bits. The true data bits are still obtained by following these decisions with the differential decoding operation of (2.8-99).

For GMSK, a superimposed I-Q representation is still possible. However, because the equivalent pulse shapes now spread over many symbol intervals and because more than one AMP component is present, the receive filter in Fig. 2-39 must be more complex than just a simple matched filter in order to account for the ISI inherent in the signal and the interference produced by the other AMP component(s). The nature of the modification of the receive filter required to accommodate these additional degradations is discussed below.

Consider then the transmitted signal of (2.8-95), where we explicitly substitute $\hat{K} = 2$ so as to correspond to the approximation discussed above for GMSK. Omitting herein the "hat" on $\tilde{S}(t)$ to simplify the notation, we have [see (2.8-80)]

$$\tilde{S}(t) = \sqrt{\frac{2E_b}{T_b}} \sum_{n=-\infty}^{\infty} \tilde{a}_{0,n} C_0\left(t - nT_b\right) + \sqrt{\frac{2E_b}{T_b}} \sum_{n=-\infty}^{\infty} \tilde{a}_{1,n} C_1\left(t - nT_b\right) \tag{2.8-104}$$

with corresponding received signal

$$\tilde{Z}(t) = \tilde{S}(t) + \tilde{N}(t) \tag{2.8-105}$$

The second term in (2.8-104) may be viewed as an interference term. Since we have restricted ourselves to a linear receiver of the type in Fig. 2-39, we shall

at first ignore this interference term and design the receive filter to match only the first of the two AMP components in (2.8-104). Hence, in view of (2.8-102), we form the output statistics $\mathrm{Re}\{r_{0,2k}\}$ and $\mathrm{Im}\{r_{0,2k+1}\}$, which are obtained by alternately sampling the real and imaginary components of the output of a matched filter (impulse response $h(t) = C_0(t)$) at T_b-s intervals, where $C_0(t)$ is now defined as in (2.8-77). Substituting (2.8-105) together with (2.8-104) into (2.8-103) and simplifying gives

$$\begin{aligned}
\mathrm{Re}\{r_{0,2k}\} &= \sqrt{\frac{2E_b}{T_b}}\left[\sum_m \tilde{a}_{0,2k-2m} p_{00}(2mT_b) + \sum_m \tilde{a}_{1,2k-2m+1} p_{10}(2mT_b)\right] \\
&\quad + \mathrm{Re}\{w_{2k}\} \\
\mathrm{Im}\{r_{0,2k+1}\} &= \sqrt{\frac{2E_b}{T_b}}\left[\sum_m \mathrm{Im}\{\tilde{a}_{0,2k-2m+1}\} p_{00}(2mT_b)\right. \\
&\quad \left. + \sum_m \mathrm{Im}\{\tilde{a}_{1,2k-2m+2}\} p_{10}(2mT_b)\right] + \mathrm{Im}\{w_{2k+1}\}
\end{aligned} \tag{2.8-106}$$

where

$$\left.\begin{aligned}
p_{00}(t) &\triangleq \int C_0(\tau) C_0(\tau - t)\, d\tau \\
p_{10}(t) &\triangleq \int C_1(\tau) C_0(\tau - t)\, d\tau \\
w_k &\triangleq \int \tilde{N}(t) C_0(t - kT)\, dt
\end{aligned}\right\} \tag{2.8-107}$$

Notice that even if we ignore the interference term in (2.8-104), i.e., assume the $\tilde{a}_{1,k}$'s are all equal to zero, the metric components in (2.8-106) still contain ISI terms due to the $\tilde{a}_{0,k}$ symbols in that $p_{00}(2mT_b) \neq 0$ for $m \neq 0$. Thus, bit-by-bit decisions based on the $r_{0,k}$'s are not optimum. Furthermore, when the interference term in (2.8-104) is accounted for, then bit-by-bit decisions based on the $r_{0,k}$'s are even more suboptimum. In Ref. 46, it is shown that by applying an MMSE criterion, the performance of the linear receiver can be improved by inserting between the matched filter and the threshold decision device a Weiner

estimator, which takes the form of a finite impulse response (FIR) filter. This should not be too surprising, since it is well-known that such a filter combination is optimum (in the mean-square error sense) for any binary pulse stream that contains ISI and is transmitted over an AWGN. The input-output sample characteristic of the Wiener filter with real coefficients $\{c_k, -N \leq k \leq N\}$ has the mathematical form

$$y_n = \sum_{k=-N}^{N} c_k r_{0,n-2k} \tag{2.8-108}$$

Equivalently, the transfer function of this filter is given by

$$C\left(e^{j2\pi f(2T_b)}\right) = \sum_{k=-N}^{N} c_k e^{j2\pi fk(2T_b)} \tag{2.8-109}$$

Consequently, bit-by-bit decisions are made using $\mathrm{Re}\{y_{2k}\}$ and $\mathrm{Im}\{y_{2k+1}\}$ in place of $\mathrm{Re}\{r_{0,2k}\}$ and $\mathrm{Im}\{r_{0,2k+1}\}$ in (2.8-102). The evaluation of the coefficients $\{c_k\}$ is performed in Ref. 46 as the solution to a set of Wiener-Hopf (linear) equations involving samples of $p_{00}(t)$ and $p_{10}(t)$, namely,

$$\left.\begin{aligned}
\sum_{k=-N}^{N} \Psi_{ik} c_k &= \left(\sqrt{\frac{2E_b}{T_b}}\right)^{-1} p_{00}\left(-2iT_b\right), \quad -N \leq i \leq N \\
\Psi_{ik} &= \sum_m p_{00}\left(2mT_b\right) p_{00}\big(2\left(m+k-i\right)T_b\big) \\
&\quad + \sum_m p_{10}\big(\left(2m-1\right)T_b\big) p_{10}\big(2\left(m+k-i-1\right)T_b\big) \\
&\quad + \frac{N_0 T_b}{2E_b} p_{00}\left(2\left(k-i\right)T_b\right)
\end{aligned}\right\} \tag{2.8-110}$$

Instead of implementing two separate filters, the matched and Wiener filters can be combined into a single optimum filter with impulse response

$$h_o(t) = \mathcal{F}^{-1}\left\{\mathcal{F}\{C_0(-t)\} C\left(e^{j2\pi f(2T_b)}\right)\right\} = \sum_{k=-\infty}^{\infty} c_k C_0\left(-t+2kT_b\right) \tag{2.8-111}$$

Alternating samples of the real and imaginary parts of the output of $h_o(t)$ in (2.8-111) taken at T_b-s intervals produces the $\mathrm{Re}\{y_{2k}\}$ and $\mathrm{Im}\{y_{2k+1}\}$ values needed for decisions on $\{a_{0,n}\}$. A comparison of the impulse response of the optimum receive filter as given by (2.8-111), with just the matched filter portion, i.e, $h(t) = C_0(-t)$, is illustrated in Fig. 2-40. The eye diagram of the signal at the output of the optimum receive filter is illustrated in Fig. 2-41 for the case of $BT_b = 0.25$.

Upper and lower bounds on the error probability of the linear MSK-type receiver for GMSK are derived in Ref. 46 in the form of the sum of two Gaussian probability integrals with appropriate arguments. These bounds were evaluated for the case of $BT_b = 0.25$ and $N = 11$ FIR filter coefficients. While it is true that the four-state VA receiver performs better than the linear receiver because in the former, the second AMP component [see (2.8-104)] is considered as relevant whereas in the latter, it is treated as interference, the difference in performance between the two is quite small. The reason that the performance difference is small in the GMSK case is because the second AMP component has small energy. For CPM schemes with rational modulation index other than 0.5, one might expect a larger improvement from the simplified VA scheme.

2.8.2.7 Spectral Considerations in the Presence of Data Imbalance. Analogous to the discussion in Sec. 2.7.2, we consider here the spectral behavior of MSK, GMSK, and precoded GMSK in the presence of data imbalance. For linear modulations produced by amplitude modulation of a binary pulse stream on a carrier, e.g., QPSK and OQPSK, the effect of data imbalance on the PSD is well documented, e.g., Chap. 2 of Ref. 1, manifesting itself in the addition of a discrete spectral component to the overall PSD with no effect on the shape of

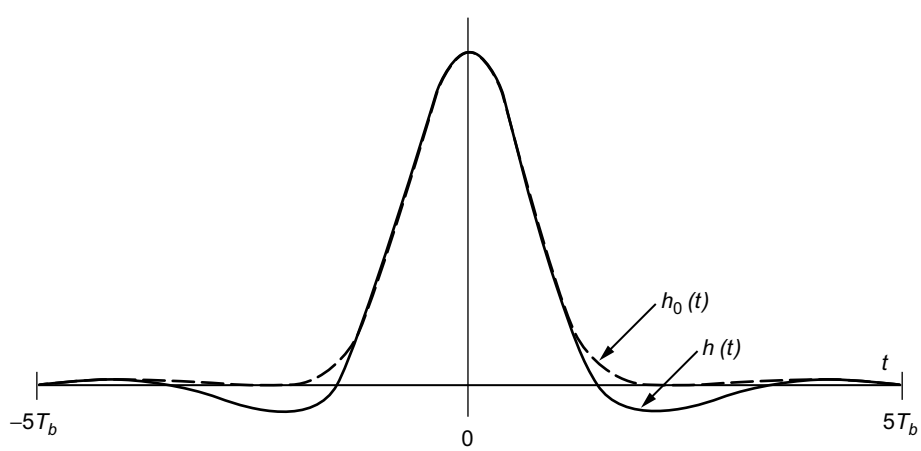


Fig. 2-40. Comparison of the impulse responses of the optimum and matched receive filters. Vertical scaling is normalized. Redrawn from [46].

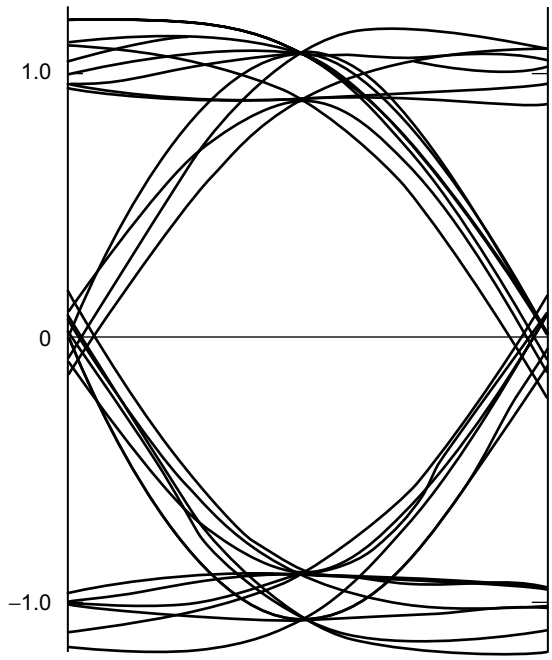


Fig. 2-41. Eye diagram at the output of the optimum receiver filter. Scaling is normalized. Redrawn from [46].

the continuous component. For phase (or frequency) modulation, the evaluation of the PSD is considerably more complex, and the effect of data imbalance is quite different in terms of its impact on both the discrete and continuous spectral components of the modulator output. Because of these important differences and their significance in relation to the specification on the tolerable amount of data imbalance, the presentation will devote more attention to detail, considering first the more generic MSK-type modulations and then including GMSK as a specific case.

Of the many techniques available for evaluating the PSD of CPM schemes [1,15,27,52,59], the one deemed most convenient by the author, particularly for MSK-type modulations with data imbalance, is that which results from the Laurent representation. As previously noted, when the input binary data is random and balanced, the complex data sequences that characterize each of the 2^{L-1} AMP components are themselves uncorrelated, and, furthermore, are uncorrelated with each other. As such, the PSD of the composite CPM waveform is equal to the sum of the PSDs of the AMP components, each of which is computed by conventional PSD evaluation techniques for binary amplitude (unit magnitude)

modulation of a carrier with a complex i.i.d. data sequence. The resulting PSD under these ideal circumstances was presented in Sec. 2.8.2.4.

Here we expand upon the PSD evaluation found in Laurent [52] to include the case of input data imbalance. Specifically, we shall show that because of the presence of data imbalance, the effective complex data sequences that typify each AMP pulse stream are now themselves correlated and, in addition, are correlated with each other. The correlation properties of each of these sequences resemble those of a first-order Markov process, and, hence, the PSD for each contains a factor due to the pulse shape as well as a factor due to the sequence correlation. Likewise, the cross-correlation properties of the sequences contain pulse shape and correlation factors.

We begin by presenting the generic result for the PSD of a modulation composed of a group of correlated data pulse trains, each of which contains its own real pulse shape and complex data stream. Next, we apply this generic PSD formula to first MSK and then GMSK. Since MSK is a full-response scheme, its Laurent representation has only a single pulse stream, and, thus, the PSD has no cross-correlation components. In line with our previous discussions of approximate AMP representations of GMSK, in evaluating the PSD of GMSK, we shall employ the two pulse stream approximation discussed in Sec. 2.8.2.3*a* and characterized by (2.8-80) and later on in (2.8-104). The results that follow are taken primarily from Ref. 60.

a. A Generic Expression for the PSD of a Sum of Random Pulse Trains with Complex Data Symbols. Consider finding the PSD of a complex signal, $\tilde{S}(t)$, of the form in (2.8-67). The traditional method of evaluating such a PSD is to first find the autocorrelation function of $\tilde{S}(t)$, namely, $R_{\tilde{S}}(t, t+\tau) = E\left\{\tilde{S}(t)\,\tilde{S}^*(t+\tau)\right\}$, then time-average to remove the cyclostationary property, and, finally, take the Fourier transform of the result, i.e.,

$$S_{\tilde{S}}(f) = \mathcal{F}\left\{\left\langle R_{\tilde{S}}(t, t+\tau)\right\rangle\right\} \tag{2.8-112}$$

By a straightforward extension of the results in Chap. 2 of Ref. 1, the following result can be obtained:

$$S_{\tilde{S}}(f) = \sum_{i=0}^{2^{L-1}-1} S_{ii}(f) + \sum_{i=0}^{2^{L-1}-1} \sum_{\substack{j=0 \\ i<j}}^{2^{L-1}-1} S_{ji}(f) \tag{2.8-113}$$

where

$$S_{ii}(f) = S_{\tilde{a}_i}(f)\, S_{p_i}(f) \tag{2.8-114}$$

with

$$\left.\begin{aligned} S_{\tilde{a}_i}(f) &= \sum_{l=-\infty}^{\infty} R_{\tilde{a}_i}(l)\, e^{-j2\pi f l T_b}, \quad R_{\tilde{a}_i}(l) = E\left\{\tilde{a}_{i,k}\tilde{a}_{i,k+l}^*\right\} \\ S_{p_i}(f) &= \frac{1}{T_b}\left|P_i(f)\right|^2, \quad P_i(f) \triangleq \mathcal{F}\left\{C_i(t)\right\} \end{aligned}\right\} \tag{2.8-115}$$

and

$$S_{ji}(f) = 2\ \mathrm{Re}\left\{S_{\tilde{a}_{ji}}(f)\, S_{p_{ji}}(f)\right\} \tag{2.8-116}$$

with

$$\left.\begin{aligned} S_{\tilde{a}_{ji}}(f) &= \sum_{l=-\infty}^{\infty} R_{\tilde{a}_{ji}}(l)\, e^{-j2\pi f l T_b}, \quad R_{\tilde{a}_{ji}}(l) = E\left\{\tilde{a}_{j,k}\tilde{a}_{i,k+l}^*\right\} \\ S_{p_{ji}}(f) &= \frac{1}{T_b} P_i(f)\, P_j^*(f), \quad P_i(f) \triangleq \mathcal{F}\left\{C_i(t)\right\} \end{aligned}\right\} \tag{2.8-117}$$

Clearly then, the evaluation of the PSD involves finding the Fourier transform of the various pulse shapes in the AMP representation and both the autocorrelation and cross-correlation of the equivalent complex data sequences.

b. Cross-Correlation Properties of the Equivalent Complex Data Symbols and Evaluation of the PSD. For MSK, the equivalent complex data symbols, $\{\tilde{a}_{0,n}\}$, are defined in terms of the actual input data symbols $\{\alpha_n\}$ by the iterative (complex differential encoding) relation

$$\tilde{a}_{0,n} \triangleq e^{j(\pi/2)A_{0,n}} = j\alpha_n \tilde{a}_{0,n-1} \quad \Rightarrow \quad \tilde{a}_{0,2n} \in \{j, -j\}, \quad \tilde{a}_{0,2n+1} \in \{1, -1\} \tag{2.8-118}$$

Suppose now that $\{\alpha_n\}$ characterizes a random i.i.d. imbalanced source, i.e., one where $\Pr\{\alpha_n = 1\} = 1 - p, \quad \Pr\{\alpha_n = -1\} = p$ with $0 \le p \le 1$. Then, it is straightforward to show that $\{\tilde{a}_{0,n}\}$ is a first-order Markov source and as such, it is balanced, i.e.,

$$\left.\begin{array}{lll} \Pr\{\tilde{a}_{0,n}=j\}=\dfrac{1}{2}, & \Pr\{\tilde{a}_{0,n}=-j\}=\dfrac{1}{2} & \text{for } n \text{ even} \\ \Pr\{\tilde{a}_{0,n}=1\}=\dfrac{1}{2}, & \Pr\{\tilde{a}_{0,n}=-1\}=\dfrac{1}{2} & \text{for } n \text{ odd} \end{array}\right\} \tag{2.8-119}$$

and, thus, $E\{\tilde{a}_{0,n}\}=0$. However, while the differential encoding operation converts the imbalanced random i.i.d. source to a balanced source,[22] the symbols of the latter are now correlated. Using the defining relation for $\{\tilde{a}_{0,n}\}$, it is straightforward to show that

$$R_{\tilde{a}_0}(l) \triangleq E\left\{\tilde{a}_{0,n}\tilde{a}^*_{0,n+l}\right\} = \left[-j\left(1-2p\right)\right]^l, \quad l \text{ integer}, \quad R_{\tilde{a}_0}(-l) = R^*_{\tilde{a}_0}(l) \tag{2.8-120}$$

i.e., $\{\tilde{a}_{0,n}\}$, behaves analogously to a first-order Markov source having a transition probability equal to p. The discrete Fourier transform of (2.8-120) as needed in (2.8-115) is obtained as

$$\begin{aligned} S_{\tilde{a}_0}(f) &= \sum_{l=-\infty}^{\infty} R_{\tilde{a}_0}(l)\, e^{-j2\pi f l T_b} = \sum_{l=-\infty}^{\infty} \left[-j\left(1-2p\right)\right]^l e^{-j2\pi f l T_b} \\ &= 1 + 2\sum_{l=-\infty}^{\infty} \left(1-2p\right)^l e^{-j2\pi l(fT_b+[1/4])} \end{aligned} \tag{2.8-121}$$

Using a well-known result [61] for the series in (2.8-121), namely,

$$\sum_{k=1}^{\infty} a^k \cos k\theta = \frac{a\cos\theta - a^2}{1-2a\cos\theta + a^2} \tag{2.8-122}$$

we obtain the closed-form result

$$S_{\tilde{a}_0}(f) = \frac{4p\left(1-p\right)}{2\left(1-2p\right)\left(1+\sin 2\pi f T_b\right)+4p^2} \tag{2.8-123}$$

Finally, taking the Fourier transform of the pulse shape $C_0(t) = \sin \pi t/2T_b$ and substituting its squared magnitude in (2.8-115), the complex baseband PSD of MSK with imbalanced data input becomes

[22] The implication of a balanced equivalent complex symbol stream for the AMP representation of MSK is that no discrete spectrum will be generated.

$$S_{\tilde{m}}(f;p) = T_b \frac{16}{\pi^2} \frac{\cos 2\pi f T_b}{\left(1 - 16 f^2 T_b^2\right)^2} \left[\frac{4p(1-p)}{2(1-2p)(1+\sin 2\pi f T_b) + 4p^2} \right] \quad (2.8\text{-}124)$$

Note that because of the presence of the term $\sin 2\pi f T_b$ in the denominator of (2.8-123), the equivalent baseband spectrum of (2.8-124) is not symmetric around $f = 0$. Since the PSD of the true MSK signal is related to the equivalent baseband PSD by

$$S_s(f;p) = \frac{1}{4}\left[S_{\tilde{m}}(f + f_c; p) + S_{\tilde{m}}(-f + f_c; p)\right] \quad (2.8\text{-}125)$$

then, equivalently, the PSD of (2.8-125) will have a tilt around the carrier. Also, since in addition from (2.8-124) we have

$$S_{\tilde{m}}(f; 1-p) = S_{\tilde{m}}(-f; p) \quad (2.8\text{-}126)$$

the tilt of the PSD of (2.8-125) reverses when the probability distribution of the input data is reversed. Finally, for $p = 1/2$, i.e., balanced random data input, the factor in brackets in (2.8-124) becomes equal to unity, and one obtains the well-known two-sided PSD of conventional MSK [see (2.8-35)], which is symmetrical around the origin.

For GMSK, the equivalent complex data symbols, $\{\tilde{a}_{0,n}\}$, are defined in terms of the actual input data symbols $\{\alpha_n\}$ by the iterative relations in the first two equations of (2.8-79). Suppose that $\{\alpha_n\}$ again characterizes a random i.i.d. imbalanced source; then, the autocorrelation function of the first equivalent symbol stream is given by (2.8-120) and its associated discrete Fourier transform by (2.8-123). Thus, the PSD of the first component of the AMP representation of GMSK is

$$S_{00}(f;p) = \frac{1}{T_b} |P_0(f)|^2 \left[\frac{2p(1-p)}{(1-2p)(1+\sin 2\pi f T_b) + 2p^2} \right], \quad P_0(f) \triangleq \mathcal{F}\{C_0(t)\} \quad (2.8\text{-}127)$$

with $C_0(t)$ defined in (2.8-77) and evaluated from (2.8-69) and (2.8-70), using the GMSK phase pulse.

Following a similar procedure as that used to derive (2.8-20), it can be shown that the autocorrelation function of the second equivalent symbol stream (which is also balanced and therefore has zero mean) is given by [60, Appendix]

$$R_{\tilde{a}_1}(l) \triangleq E\left\{\tilde{a}_{1,n}\tilde{a}^*_{1,n+l}\right\} = \begin{cases} 1, & l=0 \\ -j\left(1-2p\right)^3, & l=1, \\ \left[-j\left(1-2p\right)\right]^l, & l \geq 2 \end{cases} \quad R_{\tilde{a}_1}(-l) = R^*_{\tilde{a}_1}(l) \tag{2.8-128}$$

with discrete Fourier transform

$$S_{\tilde{a}_1}(f) = \sum_{l=-\infty}^{\infty} R_{\tilde{a}_1}(l)\, e^{-j2\pi f l T_b} = S_{\tilde{a}_0}(f) + 8p\left(1-2p\right)\left(1-p\right)\sin 2\pi f T_b \tag{2.8-129}$$

Therefore, the PSD of the second component of the AMP representation of GMSK is

$$S_{11}(f;p) = \frac{1}{T_b}\left|P_1(f)\right|^2 4p\left(1-p\right)$$

$$\times\left[\frac{1}{2\left(1-2p\right)\left(1+\sin 2\pi f T_b\right)+4p^2} + 2\left(1-2p\right)\sin 2\pi f T_b\right],$$

$$P_1(f) \triangleq \mathcal{F}\left\{C_1(t)\right\} \tag{2.8-130}$$

Note again that because of the presence of the term $\sin 2\pi f T_b$ in the denominator of (2.8-130), the equivalent baseband spectrum is not symmetric around $f=0$.

What remains is to compute the cross-correlation function of the two equivalent complex symbol streams. Following the same procedure as for obtaining the autocorrelation function of the individual pulse streams, we obtain [60, Appendix]

$$R_{\tilde{a}_{10}}(l) \triangleq E\left\{\tilde{a}_{1,n}\tilde{a}^*_{0,n+l}\right\} = \begin{cases} \left[-j(1-2p)\right]^{l+1}, & l \geq 0 \\ (1-2p)^2, & l=-1\ , \\ \left[j(1-2p)\right]^{-(l+1)}, & l \leq -2 \end{cases} \quad R_{\tilde{a}_{01}}(-l) = R^*_{\tilde{a}_{10}}(l) \tag{2.8-131}$$

with discrete Fourier transform

$$S_{\tilde{a}_{10}}(f) = \sum_{l=-\infty}^{\infty} R_{\tilde{a}_{10}}(l)\, e^{-j2\pi f l T_b} = e^{j2\pi f T_b}\left[S_{\tilde{a}_0}(f) - 4p(1-p)\right] \tag{2.8-132}$$

Thus, the cross-spectrum of $\tilde{m}(t)$ is from (2.8-116)

$$S_{10}(f;p) =$$

$$8p(1-p)\,\mathrm{Re}\left\{\left[\frac{1}{2(1-2p)(1+\sin 2\pi f T_b)+4p^2} - 1\right] e^{j2\pi f T_b}\frac{1}{T_b}P_0(f)\,P_1^*(f)\right\} \tag{2.8-133}$$

which is also not symmetric around $f = 0$. Finally, the complex baseband PSD of GMSK (based on the two pulse stream AMP approximation) with imbalanced data input becomes

$$S_{\tilde{m}}(f;p) = S_{00}(f;p) + S_{11}(f;p) + S_{10}(f;p) \tag{2.8-134}$$

where $S_{00}(f;p)$, $S_{11}(f;p)$, and $S_{10}(f;p)$ are defined in (2.8-127), (2.8-130), and (2.8-133), respectively.

Before proceeding, we point out that with some additional computation (which would be warranted if one were interested in very low PSD levels), the PSD evaluation procedure discussed above can be extended to include more than just the first two (dominant) AMP pulse streams. In fact, the results of Sec. 2.8.2.7*a* are quite general and, analogous to (2.8-120), (2.8-128), and (2.8-131), all one needs to compute are the autocorrelation and cross-correlation functions of the remainder of the equivalent data symbol streams, e.g., see (2.8-79) for $L = 4$.

c. Precoded MSK and GMSK. As previously discussed in Secs. 2.8.1.3 and 2.8.2.3*b*, conventional I-Q-type receivers for MSK and GMSK modulations suffer a small performance penalty due to the differential encoding operation inherent in these modulations and, thus, the need for differential decoding at the receiver. Precoding the input data with a differential decoder removes the need for differential decoding at the receiver and, thus, eliminates this penalty. From a spectral standpoint, this precoding operation has no effect on the PSD of the transmitted signal when the input data are balanced. However, when the input data are imbalanced, as is the case of interest in this section, the precoder has a definite effect on the transmitted signal PSD. To see how this comes about, we shall first consider the simpler case of MSK.

Suppose that prior to phase modulation of the carrier the input data stream $\boldsymbol{\alpha} = (\cdots, \alpha_{-2}, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \cdots)$ is first converted to a complex data stream via

$$\alpha'_n = \alpha_n j^n \tag{2.8-135}$$

and then passed through a differential decoder that satisfies the recursion relation

$$\beta_n = -j\alpha'_n \left(\alpha'_{n-1}\right)^* \tag{2.8-136}$$

where β_n denotes the complex binary output of the decoder (input to the MSK modulator) in the nth bit interval. Substituting (2.8-135) into (2.8-136), we see that

$$\beta_n = -j\left(\alpha_n j^n\right)\left(\alpha_{n-1}\left(-j\right)^{n-1}\right) = \alpha_n \alpha_{n-1} \tag{2.8-137}$$

which is a conventional differential decoding of the true input data bits. Since the cascade of the MSK differential encoding relationship $\tilde{a}_{0,n} = j\alpha_n \tilde{a}_{0,n-1}$ and the differential decoder of (2.8-136) produces a unity gain transmission path, i.e.,

$$\beta_n = -j\tilde{a}_{0,n}\tilde{a}^*_{0,n-1} = -j\left(j\alpha_n \tilde{a}_{0,n-1}\right)\tilde{a}^*_{0,n-1} = \alpha_n \left|\tilde{a}_{0,n-1}\right|^2 = \alpha_n \tag{2.8-138}$$

one can deduce that for an input binary complex i.i.d. bit sequence, $\boldsymbol{\alpha}' = \left(\cdots, \alpha'_{-2}, \alpha'_{-1}, \alpha'_0, \alpha'_1, \alpha'_2, \cdots\right)$, as in (2.8-135), precoded MSK using the precoder (differential decoder) in (2.8-136) is exactly the same as a Laurent representation of MSK (a single, complex symbol pulse stream with half-sinusoidal pulse shape) with the same input data sequence, i.e., $\{\tilde{a}_{0,n}\} = \boldsymbol{\alpha}'$.

The consequence of the above equivalence is that since the conversion from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha}'$ does not change the statistical (correlation) properties of the sequence, then based on the Laurent AMP representation, we conclude that the PSD of precoded MSK is that of a linear modulation with an i.i.d. uncorrelated complex imbalanced data source and as such, has a continuous component given by

$$S_{\text{P-MSK}}\left(f\right)|_{cont.} = 4p\left(1-p\right)T_b \frac{16}{\pi^2}\frac{\cos^2 2\pi f T_b}{\left(1-16f^2T_b^2\right)^2} \tag{2.8-139}$$

and a discrete component derived analogously to the results in Chap. 2 of Ref. 1 as

$$S_{\text{P-MSK}}(f)\,|_{discr.} = (1-2p)^2 \sum_{k=-\infty}^{\infty} \frac{4}{\pi^2} \frac{1}{(1-4k^2)^2} \delta\left(f - \frac{k}{2T_b}\right) \tag{2.8-140}$$

where P-MSK denotes precoded MSK. Thus, in summary, the addition of a precoder to the input of an MSK modulator with imbalanced data input removes the tilt of the MSK spectrum due to the imbalance and replaces it with a discrete spectral component, as is typical of linear modulations.

For precoded GMSK approximated by the first two AMP components, the PSD in the presence of data imbalance was derived in Ref. 62. Without going into great detail, the resulting expressions for the continuous and discrete PSD components are given below:

$$\begin{aligned}
&S_{\text{P-GMSK}}(f)\,|_{cont.} = \\
&4p(1-p)\frac{1}{T_b}|P_0(f)|^2 + \Big\{1-(1-2p)^6 + 2\left[(1-2p)^2-(1-2p)^6\right]\cos 2\pi f T_b \\
&+ 2\left[(1-2p)^4-(1-2p)^6\right]\cos 4\pi f T_b\Big\}\frac{1}{T_b}|P_1(f)|^2 \\
&+ 2\,\mathrm{Re}\left\{\left[(1-2p)^2-(1-2p)^4\right]\left[1+\exp(-2\pi f T_b)+\exp(-4\pi f T_b)\right]\right\} \\
&\times \frac{1}{T_b}P_0(f)\,P_1^*(f)
\end{aligned} \tag{2.8-141a}$$

$$\begin{aligned}
&S_{\text{P-GMSK}}(f)\,|_{discr.} = \\
&\left[(1-2p)^2\sum_{k=-\infty}^{\infty}\left(\frac{1}{2T_b}\right)^2\left|P_0\left(\frac{k}{2T_b}\right)\right|^2 + (1-2p)^6\sum_{k=-\infty}^{\infty}\left(\frac{1}{2T_b}\right)^2\left|P_1\left(\frac{k}{2T_b}\right)\right|^2\right. \\
&\left.+2(1-2p)^2\sum_{k=-\infty}^{\infty}\left(\frac{1}{2T_b}\right)^2\mathrm{Re}\left\{P_0\left(\frac{k}{2T_b}\right)P_1^*\left(\frac{k}{2T_b}\right)\right\}\right]\delta\left(f-\frac{k}{2T_b}\right)
\end{aligned} \tag{2.8-141b}$$

where, as before, $P_0(f)$ and $P_1(f)$ are the Fourier transforms of the AMP pulse shapes $C_0(t)$ and $C_1(t)$. As was true for precoded MSK, applying a precoder to

a GMSK modulator with imbalanced data input removes the tilt of the GMSK spectrum due to the imbalance and replaces it with a discrete spectral component. Figures 2-42 and 2-43 illustrate the PSD of precoded GMSK as computed from the sum of (2.8-141a) and (2.8-141b) for $BT_b = 0.25$ with 10 percent data imbalance and 60 percent data imbalance ($p = 0.2$), respectively. Included with the theoretical results are numerical results obtained from computer simulation [62]. As can be seen from these illustrations, the theory matches very closely the simulation results.

2.8.2.8 Synchronization Techniques. In addition to the previously discussed advantages of the AMP representation in so far as spectrum evaluation and ideal receiver implementation, there is yet another advantage having to do with carrier synchronization of the receiver. Mengali and Andrea [63] discuss the use of the Laurent representation for CPM primarily in the context of the single pulse stream approximation and, as such, arrive at decision-directed phase estimation structures that are analogous to those used for MSK. Similar decision-directed (data-aided) methods of obtaining symbol time and carrier phase tracking estimates for precoded CPM (in particular, GMSK) were also considered in Ref. 55.

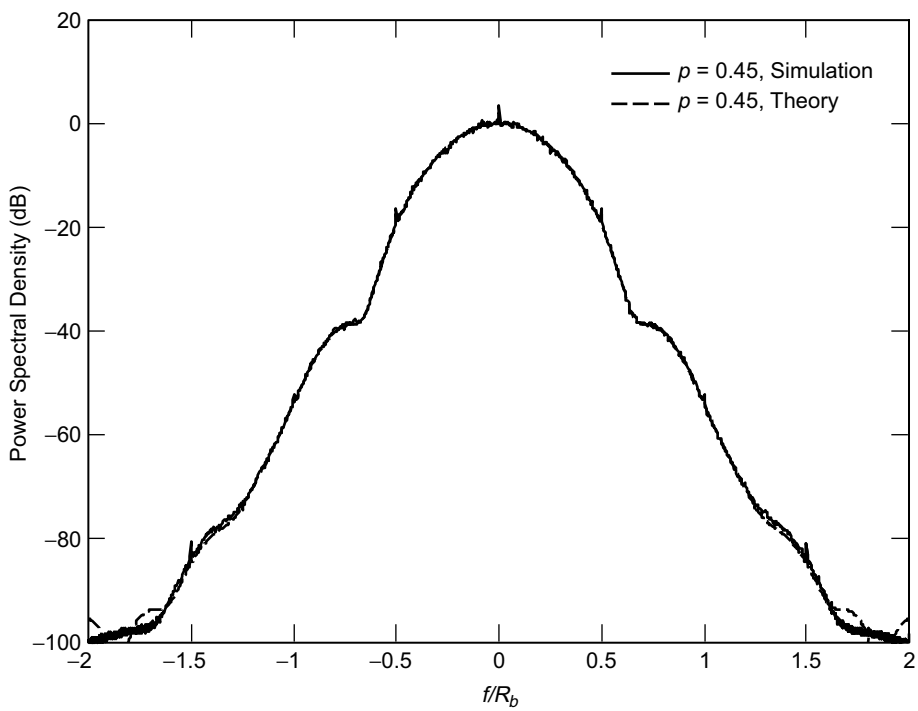


Fig. 2-42. GMSK spectrum with precoding and 10 percent data imbalance.

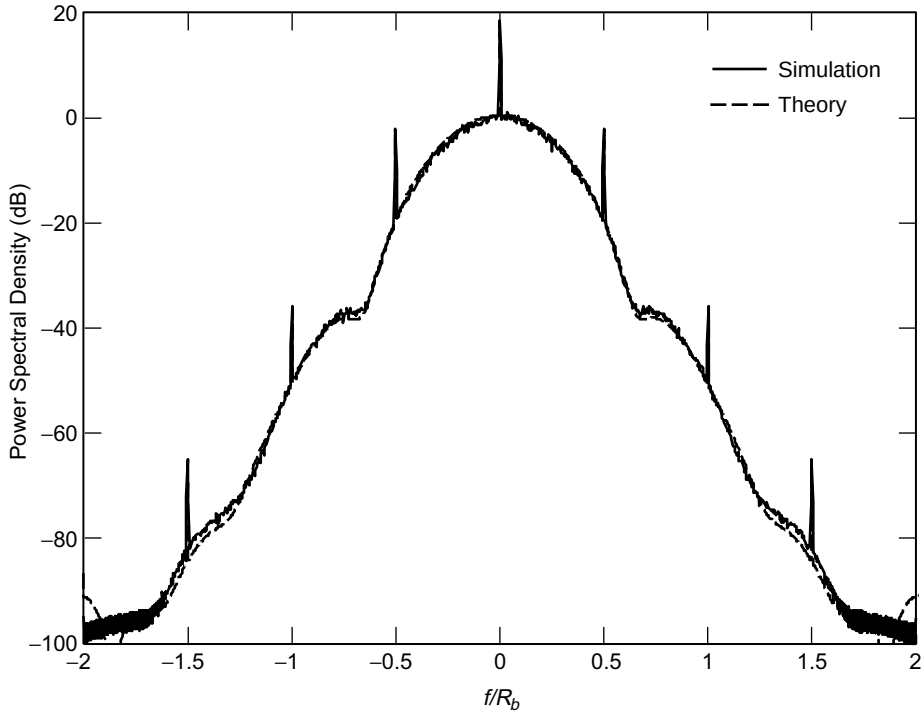


Fig. 2-43. GMSK spectrum with precoding and 60 percent data imbalance.

In this section of the monograph, we extend the carrier synchronization problem two steps further, with the goal of achieving a better solution. First, we consider the two pulse stream approximation suggested by Kaleh [46] rather than the single (main) pulse approximation. Second, using the MAP approach for carrier phase estimation as applied to pulse stream modulations with ISI [64,65], we arrive at an optimum[23] closed-loop structure that is not limited to a decision-directed form and, moreover, accounts for the ISI directly within its implementation. Finally, the tracking performance of this optimum structure is evaluated in terms of its mean-square phase error. Some of the results to be presented here are extracted from Ref. 66.

a. MAP Estimation of Carrier Phase. Consider the received signal, $y(t)$, composed of the sum of $s(t;\theta)$ and AWGN, $n(t)$ (with single-sided PSD, N_0 watts/hertz), where $s(t;\theta)$ is given by (2.8-85) with the addition of a

[23] By optimum we mean that closed-loop structure whose error signal is motivated by the derivative of the log-likelihood ratio associated with the MAP estimation of carrier phase.

uniformly distributed phase, θ, included in each carrier component. Based on an observation of $y(t)$ over the interval $0 \le t \le T_0$, where we arbitrarily assume that T_0 is an even integer multiple, say K_b, of the bit time, T_b ($K_s = K_b/2$ is then an integer multiple of the symbol time, $T_s = 2T_b$), we wish to find the MAP estimate of θ, i.e., the value of θ that maximizes the a posteriori probability, $p\left(\theta \,|y(t)\,\right)$, or since θ is assumed to be uniformly distributed, the value of θ that maximizes the conditional probability $p\left(y(t)\,|\theta\,\right)$. For an AWGN channel with single-sided noise power spectral density N_0 watts/hertz, $p\left(y(t)\,|\theta\,\right)$ has the form

$$p\left(y(t)\,\middle|\mathbf{a}_0^I, \mathbf{a}_0^Q, \mathbf{a}_1^I, \mathbf{a}_1^Q, \theta\right) = C \exp\left(-\frac{1}{N_0}\int_0^{T_0} \left(y(t) - s\left(t;\theta\right)\right)^2 dt\right) \quad (2.8\text{-}142)$$

where C is a normalization constant, and we have added to the conditioning notation the implicit dependence of $s(t;\theta)$ on the i.i.d. I and Q data sequences of the two pulse streams. For a constant envelope (energy) signal such as GMSK, it is sufficient to consider only the term involving the correlation of $y(t)$ and $s(t;\theta)$ and lump the remaining terms into the normalization constant.[24] Thus, we rewrite (2.8-142) as

$$p\left(y(t)\,\middle|\mathbf{a}_0^I, \mathbf{a}_0^Q, \mathbf{a}_1^I, \mathbf{a}_1^Q, \theta\right) = C \exp\left(\frac{2}{N_0}\int_0^{T_0} y(t) s\left(t;\theta\right) dt\right) \quad (2.8\text{-}143)$$

where for convenience, we still use C to denote the normalization constant.

Evaluation of (2.8-143) for $s(t;\theta)$ corresponding to a single binary pulse stream, e.g., BPSK, with ISI was considered in Refs. 64 and 65. Extension of the result to an $s(t;\theta)$ corresponding to a single pair of quadrature binary pulse streams (such as QPSK) with identical ISI on the I and Q channels is straightforward and was somewhat discussed in Ref. 64. What we have for the AMP representation of GMSK in (2.8-85) is two pairs of offset quadrature binary pulse streams, each pair having different amounts of ISI. (Recall that $C_0(t)$ is a pulse of width $5T_b$, and $C_1(t)$ is a completely different pulse of width $3T_b$.)

---

[24] We note that for the general ISI problem as treated in Refs. 64 and 65, the energy-dependent exponential term $\exp\{-(1/N_0)\int_0^{T_0} s^2(t;\theta)dt\}$ is not constant and, in fact, depends on the data sequence. However, for the "exact" representation of GMSK by the two pulse stream AMP form, we can make the constant envelope assumption and hence ignore the energy-dependent term.

Without belaboring the details, following substitution of (2.8-85) into (2.8-143) and averaging over the four i.i.d. component data sequences $\mathbf{a}_0^I, \mathbf{a}_0^Q, \mathbf{a}_1^I, \mathbf{a}_1^Q$, we obtain [66]

$$p\left(y(t)\left|\theta\right.\right) = C \prod_{\substack{k=-3 \\ k \text{ odd}}}^{K_b-1} \cosh\left\{I_c\left(k,0,\theta\right)\right\} \prod_{\substack{k=-4 \\ k \text{ even}}}^{K_b-2} \cosh\left\{I_s\left(k,0,\theta\right)\right\}$$

$$\times \prod_{\substack{k=-2 \\ k \text{ even}}}^{K_b-2} \cosh\left\{I_c\left(k,1,\theta\right)\right\} \prod_{\substack{k=-1 \\ k \text{ odd}}}^{K_b-1} \cosh\left\{I_s\left(k,1,\theta\right)\right\} \tag{2.8-144}$$

where

$$\left.\begin{aligned} I_c\left(k,l,\theta\right) &\triangleq \frac{2\sqrt{2E_b/T_b}}{N_0} \int_0^{K_bT_b} r(t)\cos\left(\omega_c t+\theta\right) C_l\left(t-kT_b\right) dt \\ I_s\left(k,l,\theta\right) &\triangleq \frac{2\sqrt{2E_b/T_b}}{N_0} \int_0^{K_bT_b} r(t)\sin\left(\omega_c t+\theta\right) C_l\left(t-kT_b\right) dt \end{aligned}\right\} \tag{2.8-145}$$

Note that because of the presence of ISI in each of the component pulse streams, the arguments of the hyperbolic cosine terms involve integration over the entire observation interval $0 \le t \le K_bT_b$ rather than just integration over a single bit interval, as is customary in such problems when ISI is absent. (Actually the finite duration of $C_0\left(t-kT_b\right)$ and $C_1\left(t-kT_b\right)$ will truncate these integrations to an interval (depending on the value of k) smaller than the observation time interval but still larger than the bit interval.) Finally, the MAP estimate of θ, i.e., θ_{MAP}, is the value of θ that maximizes (2.8-144).

b. Closed-Loop Carrier Synchronization of GMSK. As has been done many times in the past to arrive at closed-loop carrier synchronizers based on open-loop MAP estimates, one takes the natural logarithm of the likelihood ratio, differentiates it with respect to θ, and then uses this as the error signal, $e\left(\theta\right)$, in a closed-loop configuration. The reasoning behind this approach is that $e\left(\theta\right)$ will be equal to zero when $\theta = \theta_{\text{MAP}}$ and, thus, the closed loop will null at the point corresponding to the open MAP phase estimate. Proceeding in this fashion, we obtain

$$e(\theta) \triangleq \frac{d}{d\theta} \ln p\left(y(t)\,|\theta\right)$$

$$= \sum_{\substack{k=-3 \\ k \text{ odd}}}^{K_b-1} I_s(k,0,\theta) \tanh\left\{I_c(k,0,\theta)\right\} - \sum_{\substack{k=-4 \\ k \text{ even}}}^{K_b-2} I_c(k,0,\theta) \tanh\left\{I_s(k,0,\theta)\right\}$$

$$+ \sum_{\substack{k=-2 \\ k \text{ even}}}^{K_b-2} I_s(k,1,\theta) \tanh\left\{I_c(k,1,\theta)\right\} - \sum_{\substack{k=-1 \\ k \text{ odd}}}^{K_b-1} I_c(k,1,\theta) \tanh\left\{I_s(k,1,\theta)\right\}$$

$$\triangleq e_0(\theta) + e_1(\theta) \tag{2.8-146}$$

where we have made use of the fact that from (2.8-145) $I_c(k,l,\theta)$ and $I_s(k,l,\theta)$ are derivatives of each other.

The result in (2.8-146) suggests a superposition of two loops, each contributing a component to the error signal corresponding to associated pulse stream in the two pulse stream AMP representation of GMSK. Figures 2-44(a) and 2-44(b) illustrate the two loop components that must be superimposed to arrive at the closed-loop GMSK carrier synchronizer suggested by the error signal in (2.8-146).[25] We offer this scheme as the "optimum" (in the sense of being MAP motivated) GMSK carrier synchronizer. As is customary, the tanh nonlinearity can be approximated by a linear or hard limiter device for low and high SNR applications, respectively. The rate at which the loop updates its carrier phase estimate can vary from every T_b to every K_bT_b seconds. In the case of the latter extreme, the observation intervals used for each carrier phase estimate do not overlap and, as such, the loop represents a sequential block-by-block implementation of the MAP open-loop estimator. In the case of the former extreme, the observation intervals used for each carrier phase estimate overlap by $(K_b-1)\,T_b$ s and, as such, the loop represents a sliding window MAP phase estimator.

c. Performance of the GMSK Loop Based on a Single Pulse AMP Representation. In this section, we consider the mean-square error performance of the previously derived closed loop, using just a single pulse stream for the AMP representation of GMSK. As such, the error signal is described by only the first two out of the four terms in (2.8-146), which leads to the implementation in Fig. 2-44(a), i.e., there is no contribution to the error signal from Fig. 2-44(b).

[25] A value of $K_b = 6$ (for any larger value, the noise-free (signal) components of $I_c(k,l,\theta)$ and $I_s(k,l,\theta)$ would not change due to the truncation of the integral caused by the time limitation of $C_0(t-kT_b)$, and $C_1(t-kT_b)$) is no doubt sufficient for these figures.

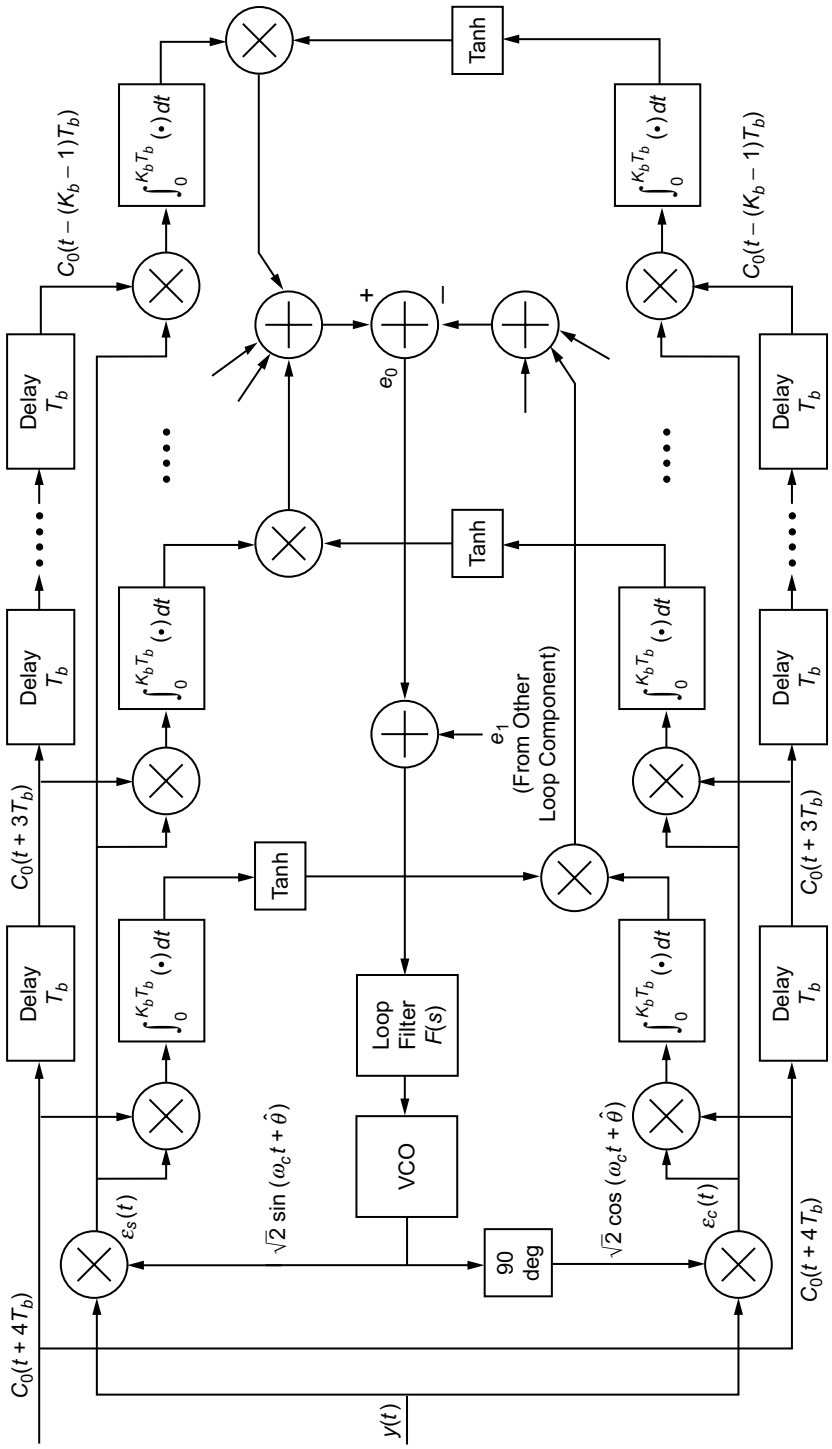


Fig. 2-44(a). Block diagram of a suboptimum ISI-compensated MAP estimation loop for GMSK (first signal component).

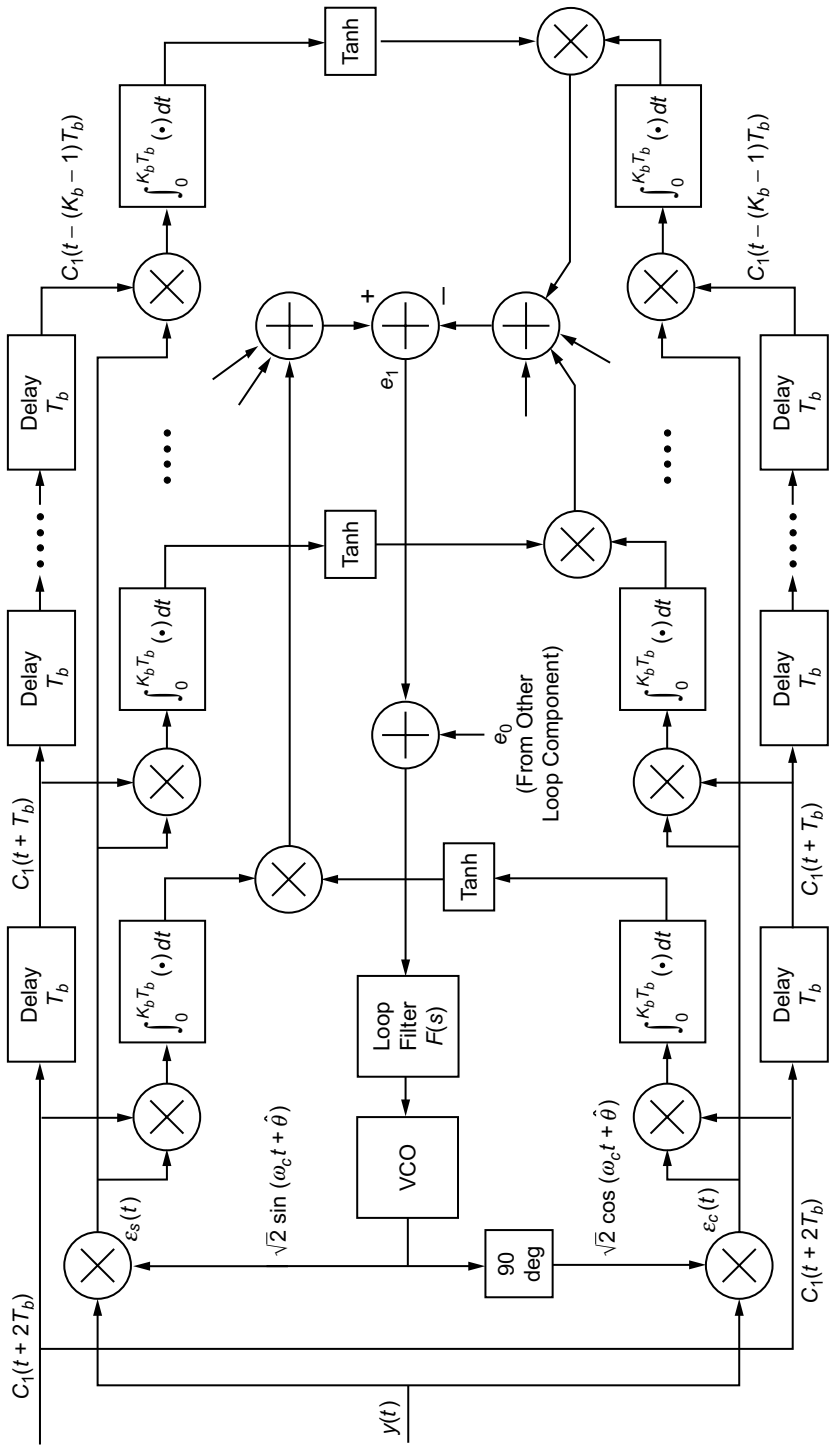


Fig. 2-44(b). Block diagram of a suboptimum ISI-compensated MAP estimation loop for GMSK (second signal component).

In evaluating the performance, we shall consider the linear loop model, wherein the tanh nonlinearity is replaced by a linear function.

To obtain the mean-square phase error performance, we follow the approach taken in [11,67], resulting in the expression

$$\sigma_{2\phi}^2 = \frac{N_E B_L}{K_g^2} \tag{2.8-147}$$

where B_L denotes the loop bandwidth, N_E is the flat single-sided PSD of the equivalent noise process perturbing the loop, and K_g is the slope (with respect to 2ϕ) of the loop S-curve at the origin. Without going into great detail, K_g is obtained as

$$K_g = PT_s^2 \sum_{i=-K_s+1}^{2} \sum_{j=-K_s+1}^{2} \left[-I_{i-(1/2),j}^2 + I_{i-(1/2),j-(1/2)}^2 + I_{i,j}^2 - I_{i,j-(1/2)}^2 \right] \tag{2.8-148}$$

where $P = E_b/T_b$ is the signal power, and the $I_{i,j}$'s are ISI parameters defined by

$$I_{i,j} \triangleq \frac{1}{T_s} \int_0^{K_s T_s} C_0(t + iT_s) C_0(t + jT_s) dt = I_{j,i} \tag{2.8-149}$$

Furthermore, N_E is evaluated as

$$N_E = 2N_0^2 T_s^3 \beta + 4PN_0 T_s^4 \alpha \tag{2.8-150}$$

where $T_s = 2T_b$ is again the effective symbol rate in each of the quadrature channels and the coefficients α and β are given as follows:

$$\begin{aligned} \alpha = & \sum_{i=-K_s+1}^{2} \sum_{l=-K_s+1}^{2} \left[I_{i-(1/2),l-(1/2)}^2 + I_{i,l}^2 - 2I_{i,l-(1/2)}^2 \right] \\ & + 2 \sum_{n=1}^{K_s-1} \sum_{i=-K_s+1}^{2} \sum_{l=-K_s+1}^{2} \\ & \times \left[J_{i-(1/2),l-(1/2)}^2(n) + J_{i,l}^2(n) - J_{i,l-(1/2)}^2(n) - J_{i-(1/2),l}^2(n) \right] \end{aligned} \tag{2.8-151}$$

and

$$\beta = \sum_{i=-K_s+1}^{2} \sum_{l=-(K_s+n)+1}^{2-n} \Bigg[I_{i-(1/2),l-(1/2)}$$

$$\times \sum_{j=-K_s+1}^{2} \left(I_{i-(1/2),j-(1/2)} I_{l-(1/2),j-(1/2)} + I_{i-(1/2),j} I_{l-(1/2),j} \right)$$

$$- I_{i-(1/2),l} \sum_{j=-K_s+1}^{2} \left(I_{i-(1/2),j-(1/2)} I_{l,j-(1/2)} + I_{i-(1/2),j} I_{l,j} \right)$$

$$+ I_{i,l} \sum_{j=-K_s+1}^{2} \left(I_{i,j} I_{l,j} + I_{i,j-(1/2)} I_{l,j-(1/2)} \right)$$

$$- I_{i,l-(1/2)} \sum_{j=-K_s+1}^{2} \left(I_{i,j} I_{l-(1/2),j} + I_{i,j-(1/2)} I_{l-(1/2),j-(1/2)} \right) \Bigg]$$

$$+ 2 \sum_{n=1}^{K_s-1} \sum_{i=-K_s+1}^{2} \sum_{l=-(K_s+n)+1}^{2-n} \Bigg[J_{i-(1/2),l-(1/2)}(n)$$

$$\times \sum_{j=-K_s+1}^{2} \left(I_{i-(1/2),j-(1/2)} I_{l-(1/2),j-(1/2)}(n) + I_{i-(1/2),j} I_{l-(1/2),j}(n) \right)$$

$$- J_{i-(1/2),l}(n) \sum_{j=-K_s+1}^{2} \left(I_{i-(1/2),j-(1/2)} I_{l,j-(1/2)}(n) + I_{i-(1/2),j} I_{l,j}(n) \right)$$

$$+ J_{i,l}(n) \sum_{j=-K_s+1}^{2} \left(I_{i,j} I_{l,j}(n) + I_{i,j-(1/2)} I_{l,j-(1/2)}(n) \right)$$

$$- J_{i,l-(1/2)}(n) \sum_{j=-K_s+1}^{2} \left(I_{i,j} I_{l-(1/2),j}(n) + I_{i,j-(1/2)} I_{l-(1/2),j-(1/2)}(n) \right) \Bigg] \tag{2.8-152}$$

where the additional ISI parameters are defined by

$$I_{i,j}(k) \triangleq \frac{1}{T_s}\int_{kT_s}^{(k+K_s)T_s} C_0(t+iT_s)C_0(t+jT_s)dt = I_{j,i}(k) \tag{2.8-153}$$

and

$$J_{i,j}(k) \triangleq \frac{1}{T_s}\int_{kT_s}^{K_sT_s} C_0(t+iT_s)C_0(t+jT_s)dt = J_{j,i}(k) \tag{2.8-154}$$

which differs from $I_{i,j}(k)$ of (2.8-153) only in that the upper limit is kept fixed at $K_sT_s = K_bT_b$, independent of k. Note also that $J_{i,j}(0) = I_{i,j}(0) = I_{i,j}$, as defined in (2.8-149).

It is customary to rewrite (2.8-147) in the form[26]

$$\sigma_{2\phi}^2 = \frac{4N_0B_L}{PS_L} = \frac{4}{\rho_{\text{PLL}}\mathcal{S}_L} \tag{2.8-155}$$

where $\rho_{\text{PLL}} = P/N_0B_L$ denotes the loop SNR for a phase-locked loop (PLL) and $\mathcal{S}_L$ denotes the so-called squaring loss, which represents the additional degradation in loop SNR caused by the presence of $S\times S$, $S\times N$, and $N\times N$ components in the error signal. Combining (2.8-147) and (2.8-153), we obtain the following expression for the squaring loss:

$$\mathcal{S}_L = \frac{\left\{\sum_{i=-K_s+1}^{2}\sum_{j=-K_s+1}^{2}\left[-I_{i-(1/2),j}^2 + I_{i-(1/2),j-(1/2)}^2 + I_{i,j}^2 - I_{i,j-(1/2)}^2\right]\right\}^2}{\dfrac{N_E}{4PN_0T_s^4}} \tag{2.8-156}$$

From (2.8-150), it is possible to write the equivalent normalized flat noise spectral density as

$$\frac{N_E}{4PN_0T_s^4} = \alpha + \frac{\beta}{2PT_s/N_0} = \alpha + \frac{\beta}{2E_s/N_0} \tag{2.8-157}$$

[26] The factor of "4" in (2.8-155) comes from the fact that we are characterizing the variance of the 2ϕ process rather than the ϕ process.

where E_s/N_0 is the symbol energy-to-noise ratio. Finally, using (2.8-157) in (2.8-156), we get the desired result for the squaring loss, namely,

$$\mathcal{S}_L = \frac{\left\{ \sum_{i=-K_s+1}^{2} \sum_{j=-K_s+1}^{2} \left[-I^2_{i-(1/2),j} + I^2_{i-(1/2),j-(1/2)} + I^2_{i,j} - I^2_{i,j-(1/2)} \right] \right\}^2}{\alpha + \dfrac{\beta}{2E_s/N_0}} \tag{2.8-158}$$

which is expressed entirely in terms of the symbol energy-to-noise ratio and the ISI parameters defined in (2.8-149), (2.8-153), and (2.8-154), all of which can easily be computed from knowledge of the main pulse shape, $C_0(t)$.

For GMSK with $BT_b = 0.25$ (equivalently $L = 4$) corresponding to the pulse shape $C_0(t)$ shown in Fig. 2-32, Fig. 2-45 illustrates the squaring loss (in dB) as computed from (2.8-158) versus $E_b/N_0 = (1/2)E_s/N_0$ (in dB) with K_s, the number of symbols in the observation interval, as a parameter. Because the dominant pulse in the AMP representation of GMSK is 5 bits (2.5 symbols) long, it appears that extending the observation beyond the duration of the pulse (i.e., values of $K_s > 3$) provides no further improvement in performance. In fact, the results for $K_s = 2$ and $K_s = 3$ are virtually indistinguishable from

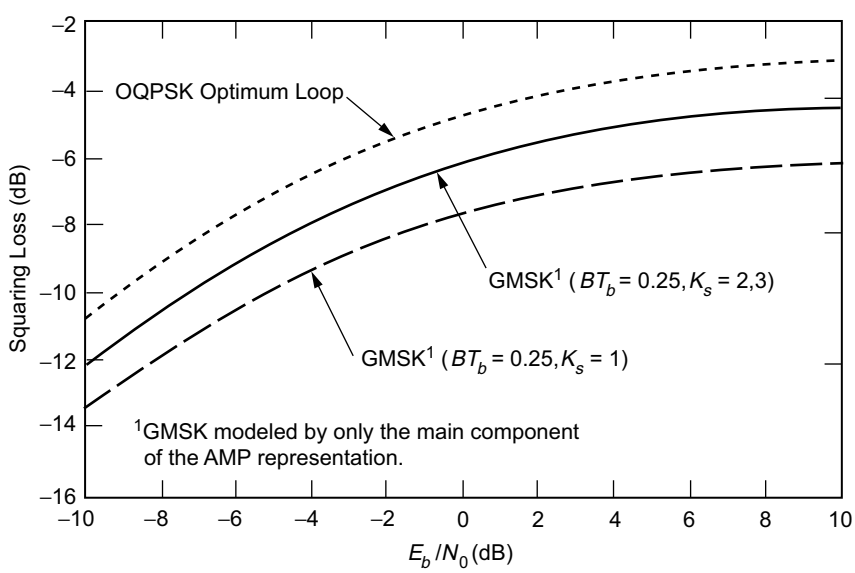


Fig. 2-45. The squaring loss performance of OQPSK and GMSK loops.

each other. Thus, for the chosen value of BT_b, implementing the loop based on a value of $K_s = 2$ is sufficient, thereby reducing the implementation complexity, which increases with the value of K_s. Note, however, that there is a significant improvement in performance by building the structure with $K_s = 2\,(K_b = 4)$ rather than $K_s = 1\,(K_b = 2)$.

Also shown in Fig. 2-45 for purpose of comparison is the performance of the corresponding MAP-motivated (optimum) OQPSK carrier synchronization loop, as obtained from the results in Ref. 12, which employs square pulses of duration T_s and so does not suffer from ISI. Although the OQPSK loop outperforms that of GMSK, we see that the difference between the two (in terms of squaring loss or, equivalently, in terms of equivalent loop SNR) is only a little more than 1 dB. This difference appears to be constant across a large range of E_b/N_0 values (-10 dB to 10 dB) and is a small price to pay for the large improvement in bandwidth efficiency that GMSK affords over OQPSK. Of particular importance is that the loop will, in fact, acquire and track a GMSK modulation at very low E_b/N_0 values, which is important in applications where high-power, efficient, error-correction coding, e.g., convolutional or turbo coding, is added to the system.

Since squaring loss is not a physical quantity that can be determined from computer simulation, to demonstrate the excellent agreement between simulation and analysis, Fig. 2-46 directly plots the equivalent linear loop SNR [i.e.,

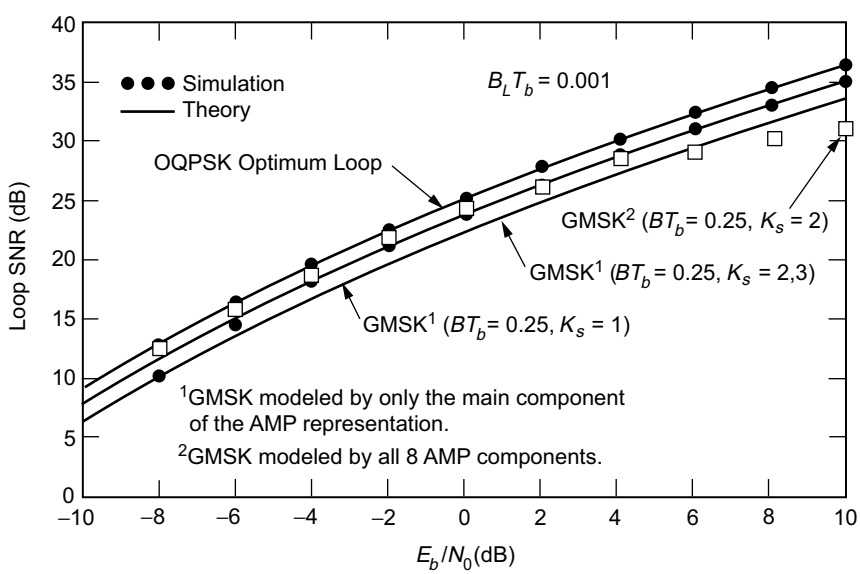


Fig. 2-46. Loop SNR performance of OQPSK and GMSK loops.

the reciprocal of the mean-square phase error as computed from (2.8-155)] versus E_b/N_0 (in dB) for the same parameter values as in Fig. 2-45 and a loop bandwidth-bit time product, $B_L T_b = 0.001$. Here, several different GMSK options were investigated. All of the analytically computed results assumed a carrier loop implementation based on an AMP approximation of the transmitted GMSK corresponding to one (the main) pulse stream. For this case, we see virtually perfect agreement between simulation and analytically computed results. For the computer simulation, another option was explored wherein the true GMSK (which requires eight AMP components to fully represent the transmitted waveform) was transmitted. Here, we have a bit of a mismatch between the receiver and the transmitter because the carrier loop is matched to only one of the eight AMP components that compose the GMSK modulation. Thus, at high SNR (where the signal dominates the noise), the simulation reveals a bit of performance degradation. This performance degradation can be diminished by implementing the receiver with a second layer corresponding to the second AMP component and adding the two components to produce the resulting error signal, as was previously suggested. Although this requires additional implementation complexity, in some applications, it may be warranted.

2.9 Simulation Performance

Aside from supporting analysis, simulations are especially useful in providing results in situations where analysis is either unavailable or, because of the complexity of the system model, would be too difficult to perform. In this section, we present some of these simulation results obtained from modeling the various systems on a Signal Processing WorkSystem (SPW) workstation.

Figure 2-47 is a block diagram of the simulation used to model precoded GMSK with concatenated block [(255,233) Reed-Solomon] and convolutional (rate 1/2, constraint length $K = 7$) error correction coding. The uncoded portion of the receiver is based on the suboptimum scheme proposed by Kaleh [46] (and discussed in Sec. 2.8.2.6b), which incorporates a Wiener filter following the matched filter prior to the decision device. The idealized (no data imbalance) BEP performance obtained from running this simulation is illustrated in Fig. 2-48, corresponding to values of $BT_b = 0.25$ and $BT_b = 0.50$. Also included for comparison is the performance of BPSK with the same error correction coding. We observe in this figure that whereas coded GMSK with $BT_b = 0.25$ suffers a small E_b/N_0 penalty (relative to coded BPSK) of something less than 0.2 dB, coded GMSK with $BT_b = 0.50$ has virtually identical performance to coded BPSK. This is a rather striking result when one considers the significant improvement in bandwidth efficiency offered by the former relative to the

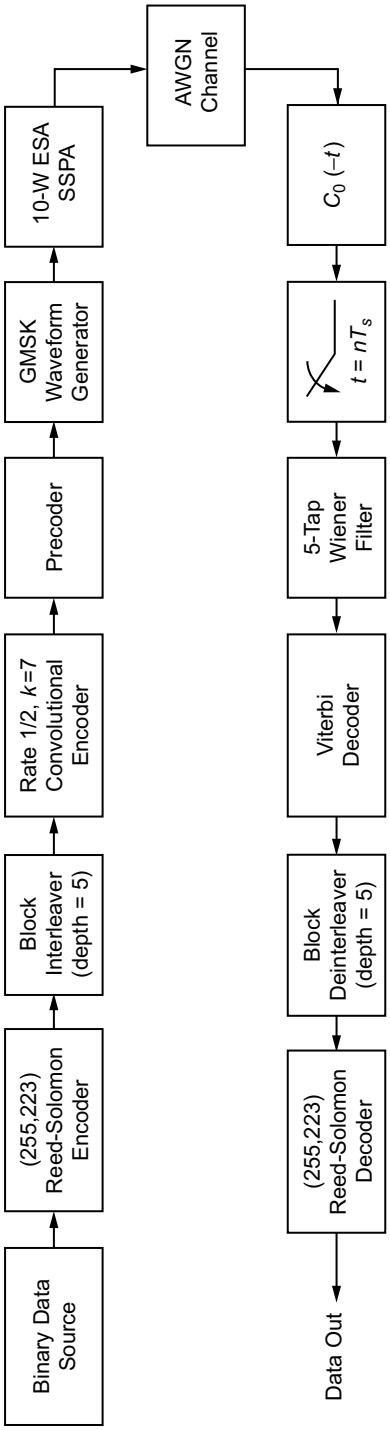


Fig. 2-47. Block diagram of a simulation for precoded GMSK with concatenated Reed-Solomon/Viterbi error-correction coding.

latter. If one now eliminates the error correction coding from the simulation, then equivalent performance results are illustrated in Fig. 2-49. Here again, we observe that GMSK with $BT_b = 0.50$ has virtually identical performance to BPSK. Finally, the performance of uncoded GMSK in the presence of data imbalance is illustrated in Fig. 2-50 for the case of $BT_b = 0.25$. Surprisingly, even with 60 percent data imbalance, the degradation in E_b/N_0 is rather small (on the order of 0.25 dB). If one increases the value of BT_b to 0.5, then even at this rather large data imbalance, the degradation becomes virtually nil. The apparent conclusion to be drawn from what is illustrated in these figures is that while data imbalance has a pronounced effect on the PSD of GMSK, its effect on BEP is quite insignificant.

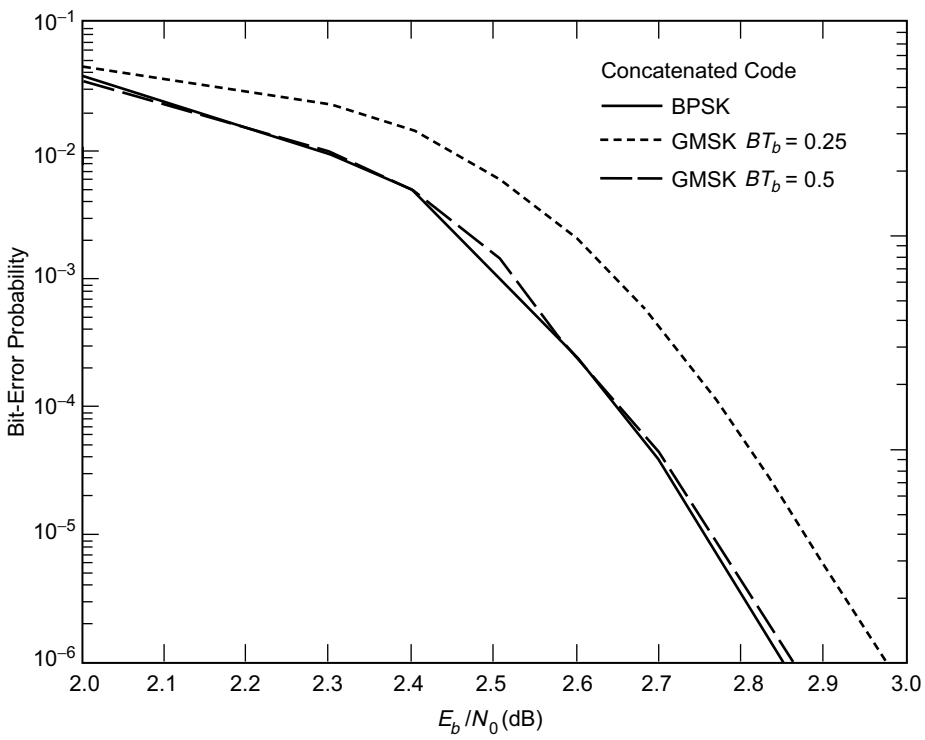


Fig. 2-48. Bit-error probability performance of precoded GMSK with concatenated (Reed-Solomon/convolutional) error-correction coding.

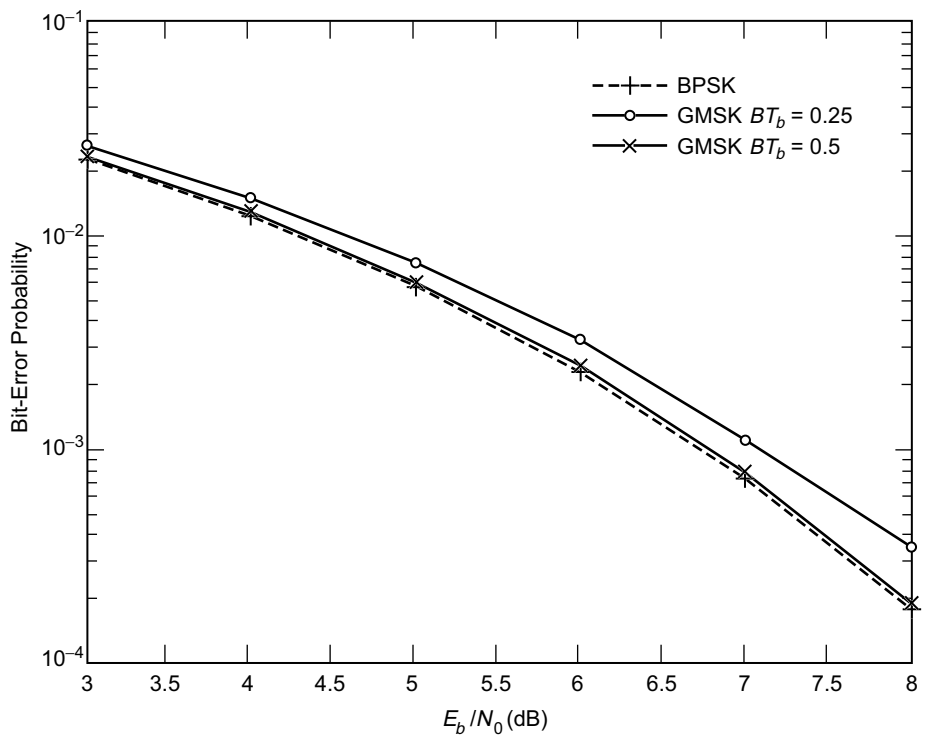


Fig. 2-49. Bit-error probability performance of uncoded GMSK with a suboptimum (Wiener filter-type) receiver.

References

[1] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Upper Saddle River, New Jersey: Prentice Hall, 1995.

[2] M. K. Simon and D. Divsalar, "On the optimality of classical coherent receivers of differentially encoded M-PSK," *IEEE Communications Letters*, vol. 1, no. 3, pp. 67–70, May 1997.

[3] T. M. Nguyen, "On the effects of a spacecraft subcarrier unbalanced modulator," *IEEE Journal of Digital and Analog Communication Systems*, vol. 6, pp. 183–192, 1993.

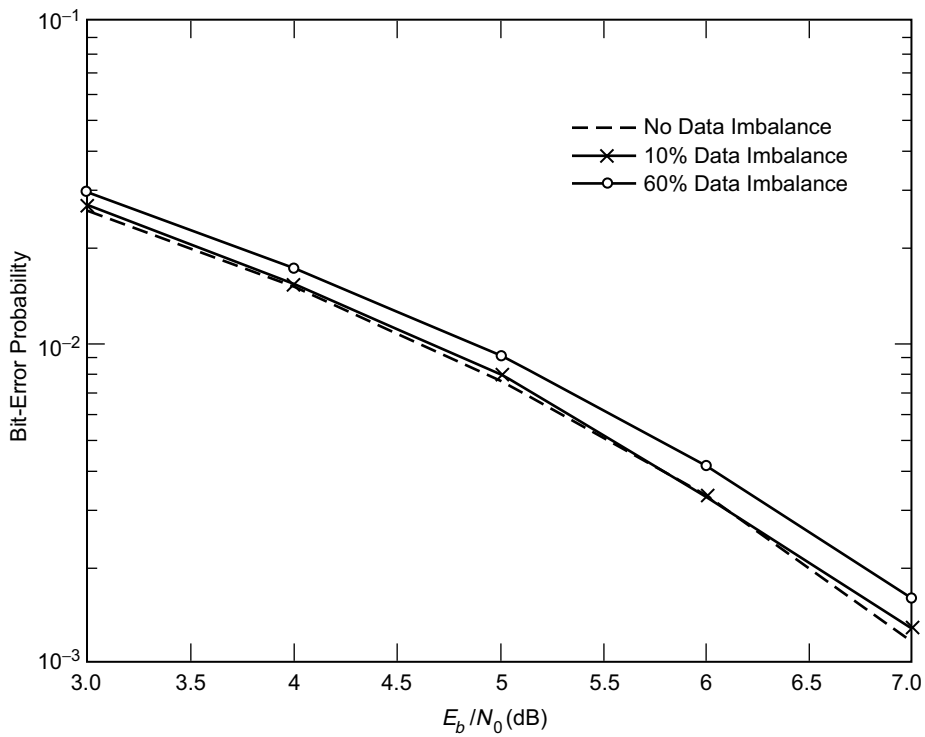


Fig. 2-50. Bit-error probability performance of uncoded GMSK in the presence of data imbalance.

[4] J.-L. Gerner, "A position paper on the effect of phase unbalanced modulator on the performance of PSK modulation schemes for category A missions," Consultative Committee for Space Data Systems, *Proceedings of the CCSDS RF and Modulation Subpanel 1E Meeting*, CCSDS 421.0-G-1, Green Book, pp. 287–301, September 1989.

[5] T. M. Nguyen and Y. Owens, "Cross-talk in QPSK communication systems," Consultative Committee for Space Data Systems, *Proceedings of the CCSDS RF and Modulation Subpanel 1E Meeting*, CCSDS B20.0-Y-1, Yellow Book, pp. 85–93, September 1993.

[6] T. M. Nguyen and A. Anabtawi, "Cross-talk due to phase imbalance between the channels in QPSK communication systems," paper presented at the Consultative Committee for Space Data Systems RF and Modulation Subpanel 1E Meeting, Pasadena, California, June 1994.

[7] M. K. Simon, "The effect of modulator unbalance on QPSK performance," paper presented at the Consultative Committee for Space Data Systems RF and Modulation Subpanel 1E Meeting, Pasadena, California, May 1996.

[8] H. Tsou, "The effect of phase and amplitude imbalance on the performance of BPSK/QPSK communication systems," *Telecommunications and Data Acquisition Progress Report 42-130*, vol. April–June 1997, August 15, 1997, http://tmo.jpl.nasa.gov/progress_report/issues.html Accessed March 2, 2001.

[9] H. Tsou, "The effect of phase and amplitude imbalance on the performance of offset quadrature phase-shift-keyed (OQPSK) communication systems," *Telecommunications and Mission Operations Progress Report 42-135*, vol. July–September 1998, November 15, 1998, http://tmo.jpl.nasa.gov/progress_report/issues.html Accessed March 2, 2001.

[10] H. Tsou, "The combined effect of modulator imbalances and amplifier nonlinearity on the performance of offset quadrature phase-shift-keyed (OQPSK) communication systems," *Telecommunications and Mission Operations Progress Report 42-137*, vol. January–March 1999, May 15, 1999, http://tmo.jpl.nasa.gov/progress_report/issues.html Accessed March 2, 2001.

[11] W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Upper Saddle River, New Jersey: PTR Prentice Hall, 1973.

[12] M. K. Simon, "Carrier synchronization of offset quadrature phase-shift keying," *Telecommunications and Mission Operations Progress Report 42-133*, vol. January–March 1998, May 15, 1998, http://tmo.jpl.nasa.gov/progress_report/issues.html Accessed March 2, 2001.

[13] W. B. Davenport, Jr. and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.

[14] M. K. Simon, "The effects of residual carrier on Costas loop performance as applied to the Shuttle S-band Uplink," *IEEE Transactions on Communications (Special issue on Space Shuttle Communications and Tracking)*, vol. COM-26, no. 11, pp. 1542–1548, November l978.

[15] J. B. Anderson, T. Aulin, and C.-E. Sundberg, *Digital Phase Modulation*, New York: Plenum Press, 1986.

[16] M. K. Simon, "A generalization of MSK-Type signaling based upon input data symbol pulse shaping," *IEEE Transactions on Communications*, vol. COM-24, no. 8, pp. 845–856, August l976.

[17] M. L. Doelz and E. T. Heald, "Minimum-shift data communication system," U.S. patent no. 2,977,417, March 28, 1961.

[18] P. Galko and S. Pasupathy, "Generalized MSK," *Proceedings of the IEEE International Electrical & Electronics Conference & Exposition*, Toronto, Ontario, Canada, October 5–7, 1981.

[19] I. Korn, "Generalized MSK," *IEEE Transactions on Information Theory*, vol. IT-26, no. 2, pp. 234–238, March 1980.

[20] F. Amoroso, "Pulse and spectrum manipulation in the minimum (frequency) shift keying (MSK) format," *IEEE Transactions on Communications*, vol. COM-24, no. 3, pp. 381–384, March 1976.

[21] M. G. Pelchat, R. C. Davis, and M. B. Luntz, "Coherent demodulation of continuous phase binary FSK signals," *Proceedings of the International Telemetry Conference*, Washington, D.C., 1971.

[22] H. R. Mathwich, J. F. Balcewicz, and M. Hecht, "The effect of tandem band and amplitude limiting on the E_b/N_0 performance of minimum (frequency) shift keying (MSK)," *IEEE Transactions on Communications*, vol. COM-22, no. 10, pp. 1525–1540, October 1974.

[23] S. A. Gronemeyer and A. L. McBride, "MSK and offset QPSK modulation," *IEEE Transactions on Communications*, vol. COM-24, no. 8, pp. 809–820, August 1976.

[24] D. M. Brady, "A constant envelope digital modulation technique for millimeter-wave satellite system," *International Conference on Communications*, Minneapolis, Minnesota, June 17–19, 1974, p. 36C-1.

[25] D. P. Taylor, "A high speed digital modem for experimental work on the communications technology satellite," *Canadian Electrical Engineering Journal*, vol. 2, no. 1, pp. 21–30, 1977.

[26] R. M. Fielding, H. L. Berger, and D. L. Lochhead, "Performance characterization of a high data rate MSK and QPSK channel," *International Conference on Communications*, Chicago, Illinois, pp. 3.2.42–3.2.46, June 12–15, 1977.

[27] B. E. Rimoldi, "A decomposition approach to CPM," *IEEE Transactions on Information Theory*, vol. IT-34, no. 2, pp. 260–270, May 1988.

[28] J. L. Massey, "A generalized formulation of minimum shift keying modulation," *International Conference on Communications*, vol. 2, Seattle, Washington, pp. 26.5.1–26.5.5, June 1980.

[29] T. Masamura, S. Samejima, Y. Morihiro, and H. Fuketa, "Differential detection of MSK with nonredundant error correction," *IEEE Transactions on Communications*, vol. COM-27, no. 6, pp. 912–918, June 1979.

[30] R. DeBuda, "The Fast FSK modulation system," *International Conference on Communications*, Montreal, Canada, pp. 41-25–45-27, June 14–16, 1971.

[31] R. DeBuda, "Coherent demodulation of frequency-shift-keying with low deviation ratio," *IEEE Transactions on Communications*, vol. COM-20, no. 3, pp. 429–435, June 1972.

[32] W. R. Bennett and S. O. Rice, "Spectral density and autocorrelation functions associated with binary frequency shift keying," *Bell System Technical Journal*, vol. 42, no. 5, pp. 2355–2385, September 1963.

[33] R. W. Booth, "An illustration of the MAP estimation method for deriving closed-loop phase tracking topologies: the MSK signal structure," *IEEE Transactions on Communications*, vol. COM-28, no. 8, pp. 1137–1142, August 1980.

[34] S. J. Simmons and P. J. McLane, "Low-complexity carrier tracking decoders for continuous phase modulations," *IEEE Transactions on Communications*, vol. COM-33, no. 12, pp. 1285–1290, December 1985.

[35] J. Huber and W. Liu, "Data-aided synchronization of coherent CPM receivers," *IEEE Transactions on Communications*, vol. 40, no. 1, pp. 178–189, January 1992.

[36] M. Moeneclaey and I. Bruyland, "The joint carrier and symbol synchonizability of continuous phase modulated waveforms," *International Conference on Communications*, vol. 2, Toronto, Canada, pp. 31.5.1–31.5.5, June 1986.

[37] A. N. D'Andrea, U. Mengali, and R. Reggiannini, "A digital approach to clock recovery in generalized minimum shift keying," *IEEE Transactions on Vehicular Technology*, vol. 39, no. 3, pp. 227–234, August 1990.

[38] A. N. D'Andrea, U. Mengali, and M. Morelli, "Multiple phase synchronization in continuous phase modulation," in *Digital Signal Processing 3*, New York: Academic Press, pp. 188–198, 1993.

[39] U. Lambrette and H. Meyr, "Two timing recovery algorithms for MSK," *International Conference on Communications*, New Orleans, Louisiana, vol. 2, pp. 1155–1159, May 1–5, 1994.

[40] A. N. D'Andrea, U. Mengali, and M. Morelli, "Symbol timing estimation with CPM modulation," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1362–1371, October 1996.

[41] K. Murota, K. Kinoshita, and K. Hirade, "Spectrum efficiency of GMSK land mobile radio," *International Conference on Communications*, vol. 2, pp. 23.8.1–23.8.5, June 14–20, 1981.

[42] K. Hirade, K. Murota, and M. Hata, "GMSK transmission performance in land mobile radio," *Global Communications Conference*, pp. B3.4.1-B3.4.6.

[43] K. Daikoku, K. Murota, and K. Momma, "High-speed digital transmission experiments in 920 MHz urban and suburban mobile radio channels," *IEEE Transactions on Vehicular Technology*, vol. VT-31, no. 2, pp. 70–75, May 1982.

[44] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Upper Saddle River, New Jersey: Prentice-Hall, 1996.

[45] A. Linz and A. Hendrickson, "Efficient implementation of an I-Q GMSK modulator," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 43, no. 1, pp. 14–23, January 1996.

[46] G. K. Kaleh, "Simple coherent receivers for partial response continuous phase modulation," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 9, pp. 1427–1436, December 1989.

[47] M. R. L. Hodges, "The GSM radio interface," *British Telecom Technological Journal*, vol. 8, no. 2, January 1990.

[48] J. Haspeslagh et al., "A 270 Kb/s 35-mW modulation IC for GSM cellular radio hand held terminals," *IEEE Journal on Solid State Circuits*, vol. 25, no. 12, pp. 1450–1457, December 1990.

[49] R. Hunter and F. Kostedt, "Enhance GMSK performance with two-point modulation," *Microwaves & RF*, vol. 39, no. 4, pp. 59–69, April 2000.

[50] K. Feher, *Wireless Digital Communications*, Upper Saddle River, New Jersey: Prentice Hall, 1995.

[51] F. Wellesplein, "Trends in silicon radio large scale integration," *Microwave Engineering Europe*, pp. 37–45, May 2000.

[52] P. A. Laurent, "Exact and approximate construction of digital phase modulations by superposition of amplitude modulated pulses," *IEEE Transactions on Communications*, vol. COM-34, no. 2, pp. 150–160, February 1986.

[53] U. Mengali and M. Morelli, "Decomposition of M-ary CPM signals into PAM waveforms," *IEEE Transactions on Information Theory*, vol. 41, no. 5, pp. 1265–1275, September 1995.

[54] K. Tsai and G. L. Lui, "Binary GMSK: Characteristics and performance," 99-G1-2, *International Telemetering Conference*, Las Vegas, Nevada, October 25–28, 1999.

[55] G. L. Lui and K. Tsai, "Data-aided symbol time and carrier phase tracking for pre-coded CPM signals," 99-G1-4, *International Telemetering Conference*, Las Vegas, Nevada, October 25–28, 1999.

[56] G. L. Lui and K. Tsai, "Viterbi and serial demodulators for pre-coded binary GMSK," 99-G1-3, *International Telemetering Conference*, Las Vegas, Nevada, October 25–28, 1999.

[57] G. L. Lui, "Threshold detection performance of GMSK signal with BT=0.5," *MILCOM' 98 Conference Proceedings*, vol. 2, pp. 515-519, October 19–21, 1998.

[58] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 260–269, April 1967.

[59] J. Proakis, *Digital Communications*, 3rd edition, New York: McGraw-Hill, 1995.

[60] M. K. Simon, P. Arabshahi, L. Lam, and T.-Y. Yan, "Power spectrum of MSK-Type Modulations in the Presence of Data Imbalance," *Telecommunications and Data Acquisition Mission Operations Progress Report 42-134*, vol. April–June 1998, August 15, 1998, http://tmo.jpl.nasa.gov/progress_report/issues.html Accessed March 2, 2001.

[61] L. B. W. Jolley, *Summation of Series*, New York: Dover Publications, 1961.

[62] D. Lee, "Occupied bandwidth of MSK and GMSK in the presence of data imbalance," Consultative Committee for Space Data Systems, *Proceedings of the CCSDS RF and Modulation Subpanel 1E Meeting*, European Space Research and Technology Centre (ESTEC), Noordwiik, The Netherlands, October 18–22, 1999.

[63] U. Mengali and A. N. D'Andrea, *Synchronization Techniques for Digital Receivers*, New York: Plenum Press, 1997.

[64] S. M. Hinedi, "Carrier Synchronization in Bandlimited Channels," Ph.D. dissertation, University of Southern California, 1987.

[65] M. K. Simon and S. Hinedi, "Suppressed carrier synchronizers for ISI channels," CD-ROM, *Global Telecommunications Conference*, London, England, November 18–22, 1996.

[66] M. K. Simon, "MAP-motivated carrier synchronization of GMSK based on the Laurent AMP representation," CD-ROM, *Global Telecommunications Conference*, Sydney, Australia, November 8–12, 1998.

[67] W. C. Lindsey and M. K. Simon, "Optimum performance of suppressed carrier receivers with Costas loop tracking," *IEEE Transactions on Communications*, vol. COM-25, no. 2, pp. 215–227, February 1977.

Chapter 3
Quasi-Constant Envelope Modulations

In Chap. 2, we restricted our consideration to bandwidth-efficient modulations that were *strictly* constant envelope, thus rendering themselves maximally power efficient when transmitted over a nonlinear channel operating in saturation. As a compromise between the two extremes of constant and nonconstant envelope, we turn our attention to modulations that deviate slightly rom the former but make up for the attendent small loss in power efficiency by offering a more significant improvement in bandwidth efficiency. The most promising modulation in this category is Feher-patented quadriphase-shift-keying (FQPSK) [1], whose generic form finds its roots in cross-correlated PSK (XPSK), introduced in 1983 [2], and which has recently been given a more insightful interpretation [3], thereby allowing further enhancements [3,4]. Since the basic form of FQPSK and its predecessor, XPSK, are well documented in several of Feher's textbooks and papers [5-8], our focus here will be on the recent advancements [3,4] that allow additional improvements in power and bandwidth efficiencies. Nevertheless, we shall present a brief review of FQPSK in its originally conceived form, since it provides insight into the new interpretation and enhancements that followed. Before proceeding with the technical details, we present a brief historical perspective as well as the current state of the art regarding the practical application of FQPSK in government- and commercially developed hardware.

FQPSK was invented by Kamilo Feher (Digcom, Inc. and the University of California, Davis). It is covered by a number of U.S. and Canadian patents [1], and is exclusively licensed by Digcom, Inc. It has been adopted by the U.S. Department of Defense Joint Services Advanced Range Telemetry (ARTM) program as their Tier 1 modulation for missile, aircraft, and range applications to replace existing pulse code modulation/frequency modulation (PCM/FM) systems. FQPSK modems operating at a data rate of 20 Mb/s are currently

available as an off-the-shelf product from several commercial vendors. The suitability of FQPSK for high-speed application has been demonstrated under a joint program between Goddard Space Flight Center (GSFC) and the Jet Propulsion Laboratory (JPL), with the development of a 300-Mb/s modem based on the enhanced architecture suggested in Ref. 3. Actually, the receiver, which employs all-digital parallel processing, can operate at 600 Mb/s and although originally designed for BPSK/QPSK, can accommodate FQPSK simply by having one reprogram the detection filter coefficients. Furthermore, the GSFC development has demonstrated that the same synchronization (carrier, bit, etc.) and detection techniques used for QPSK can be used for FQPSK without hardware modification to achieve good (not necessarily optimal) performance. Finally, at the CCSDS/SFCG meeting held in October 1999, it was recommended "that either GMSK or FQPSK-B[1] be used for high data-rate transmissions whenever practicable, and, in any case, for operating at frequencies where the available bandwidth is limited."

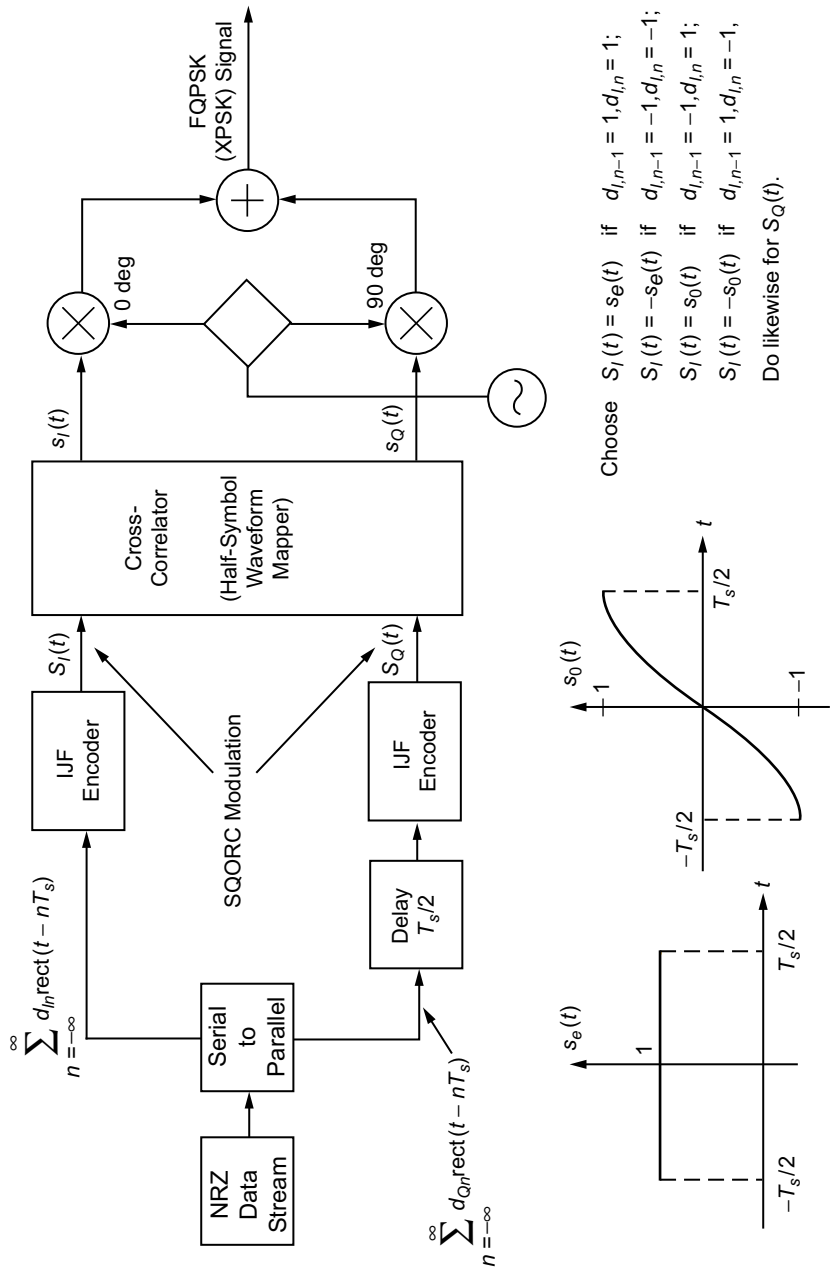
As implied above, in its generic (unfiltered) form, FQPSK is conceptually the same as XPSK, introduced in 1983 by Kato and Feher [2].[2] This technique was in turn a modification of the previously introduced (by Feher et al. [10]) interference- and jitter-free QPSK (IJF-QPSK), with the express purpose of reducing the 3-dB envelope fluctuation characteristic of IJF-QPSK to 0 dB (or as close to zero as possible), thus making it appear constant envelope.[3] It is further noted that using a constant waveshape for the even pulse and a sinusoidal waveshape for the odd pulse, which was the case considered in [2], IJF-QPSK becomes identical to the staggered quadrature overlapped raised cosine (SQORC) scheme introduced by Austin and Chang [11]. (We shall demonstrate this shortly.) The means by which Kato and Feher achieved their 3-dB envelope reduction was the introduction of an intentional but controlled amount of cross-correlation between the I and Q channels. This cross-correlation operation was applied to the IJF-QPSK (SQORC) baseband signal prior to its modulation onto the I and Q carriers (see Fig. 3-1). Specifically, this operation constituted mapping in each *half*-symbol the 16 possible combinations of I and Q channel waveforms

---

[1] The acronym FQPSK-B refers to Butterworth-filtered FQPSK. The exact filter type and optimal value of bandwidth-symbol time product, BT_s, are proprietary.

[2] More recent versions of FQPSK, referred to as FQPSK-B [9], include proprietary designed filtering for additional spectrum containment. At the moment, such filtering is not germane to our discussions although it should be mentioned now that the enhancements to be discussed are also applicable to FQPSK-B and similarly provide improved performance.

[3] The reduction of the envelope fluctuation from 3 dB to 0 dB occurs only at the uniform sampling instants on the inphase (I) and quadrature (Q) channels. It is for this reason that XPSK is referred to as being pseudo- or quasi-constant envelope, i.e., its envelope has a small amount of fluctuation between the uniform sampling instants.



Choose $S_I(t) = s_e(t)$ if $d_{I,n-1} = 1, d_{I,n} = 1$;
$S_I(t) = -s_e(t)$ if $d_{I,n-1} = -1, d_{I,n} = -1$;
$S_I(t) = s_0(t)$ if $d_{I,n-1} = -1, d_{I,n} = 1$;
$S_I(t) = -s_0(t)$ if $d_{I,n-1} = 1, d_{I,n} = -1$,
Do likewise for $S_Q(t)$.

Fig. 3-1. Conceptual block diagram of FQPSK (XPSK). "IJF" (intersymbol interference/jitter-free encoder) is a waveform mapping function without any error-correcting capability.

present in the SQORC signal into a new[4] set of 16 waveform combinations chosen in such a way that the cross-correlator output is time continuous and has unit (normalized) envelope[5] at all I and Q uniform sampling instants. Because the cross-correlation mapping is based on a half-symbol characterization of the SQORC signal, there is no guarantee that the *slope* of the cross-correlator output waveform is continuous at the half-symbol transition points. In fact, it can be shown [3] that for a random-data input sequence, such a discontinuity in slope occurs one quarter of the time.

It is a well-known fact that the rate at which the sidelobes of a modulation's PSD roll off with frequency is related to the smoothness of the underlying waveforms that generate it. That is, the more derivatives of a waveform that are continuous, the faster its Fourier transform decays with frequency. Thus, since the first derivative of the FQPSK waveform is discontinuous (at half-symbol transition instants) on the average one quarter of the time, one can anticipate that an improvement in PSD rolloff could be had if the FQPSK cross-correlation mapping could be modified so that the first derivative is always continuous. By restructuring the cross-correlation mapping into a symbol-by-symbol representation, the slope discontinuity referred to above is placed in evidence and is particularly helpful in suggesting a means to eliminate it. This representation also has the advantage that it can be described directly in terms of the data transitions on the I and Q channels, and, thus, the combination of IJF encoder and cross-correlator can be replaced simply by a single modified cross-correlator. The replacement of the conventional FQPSK cross-correlator by this modified cross-correlator that eliminates the slope discontinuity leads to what is referred to as enhanced FQPSK [3]. Not only does enhanced FQPSK have a better PSD (in the sense of reduced out-of-band energy) than conventional FQPSK but from a modulation symmetry standpoint, it is a more logical choice.

A further and more important advantage of the reformulation as a symbol-by-symbol mapping is the ability to design a receiver for FQPSK or enhanced FQPSK that specifically exploits the correlation introduced into the modulation scheme to significantly improve power efficiency or, equivalently, error-probability performance. Such a receiver, which takes a form analogous to those used for trellis-coded modulations, will yield significant performance improvement over receivers that employ symbol-by-symbol detection, thus ignoring the inherent memory of the modulation.

---

[4] Of the 16 possible cross-correlator output combinations, only 12 of them are in fact new, i.e., for 4 of the input I and Q combinations, the cross-correlator outputs the identical combination.

[5] Actually, Kato and Feher allow (through the introduction of a transition parameter $k = 1 - A$ to be defined shortly) for a controlled amount of envelope fluctuation. For quasi-constant envelope, one should choose $A = 1/\sqrt{2}$.

3.1 Brief Review of IJF-QPSK and SQORC and Their Relation to FQPSK

The IJF-QPSK scheme (alternately called FQPSK-1) is based on defining waveforms, $s_o(t)$ and $s_e(t)$, which are respectively odd and even functions of time over the symbol interval $-T_s/2 \leq t \leq T_s/2$, and then using these and their negatives, $-s_o(t), -s_e(t)$, as a 4-ary signal set for transmission in accordance with the values of successive pairs of data symbols in each of the I and Q arms. Specifically, if d_{In} denotes the I channel data symbols in the interval $(n-(1/2))\,T_s \leq t \leq (n+(1/2))\,T_s$, then the transmitted waveform, $x_I\,(t)$, in this same interval would be determined as follows:

$$\left.\begin{array}{ll} x_I(t) = s_e\,(t-nT_s) \triangleq s_0\,(t-nT_s) & \text{if } d_{I,n-1}=1, d_{I,n}=1 \\ x_I\,(t) = -\,s_e\,(t-nT_s) \triangleq s_1\,(t-nT_s) & \text{if } d_{I,n-1}=-1, d_{I,n}=-1 \\ x_I\,(t) = s_o\,(t-nT_s) \triangleq s_2\,(t-nT_s) & \text{if } d_{I,n-1}=-1, d_{I,n}=1 \\ x_I\,(t) = -\,s_o\,(t-nT_s) \triangleq s_3\,(t-nT_s) & \text{if } d_{I,n-1}=1, d_{I,n}=-1 \end{array}\right\} \quad (3.1\text{-}1)$$

The Q channel waveform, $x_Q\,(t)$, would be generated by the same mapping as in (3.1-1), using instead the Q channel data symbols, $\{d_{Qn}\}$, and then delaying the resulting waveform by a half-symbol. If the odd and even waveforms, $s_o(t)$ and $s_e(t)$, are defined by

$$\left.\begin{array}{ll} s_e\,(t) = 1, & -\dfrac{T_s}{2} \leq t \leq \dfrac{T_s}{2} \\ s_o\,(t) = \sin\dfrac{\pi t}{T_s}, & -\dfrac{T_s}{2} \leq t \leq \dfrac{T_s}{2} \end{array}\right\} \quad (3.1\text{-}2)$$

then typical waveforms for the I and Q IJF encoder outputs are illustrated in Figs. 3-2(a) and 3-2(b).

An identical modulation to $x_I\,(t)$ (and likewise for $x_Q\,(t)$) generated from the combination of (3.1-1) and (3.1-2) can be obtained directly from the binary data sequence, $\{d_{In}\}$, itself, without the need for defining a 4-ary mapping based on the transition properties of the sequence. In particular, if we define the two-symbol-wide, raised-cosine pulse shape

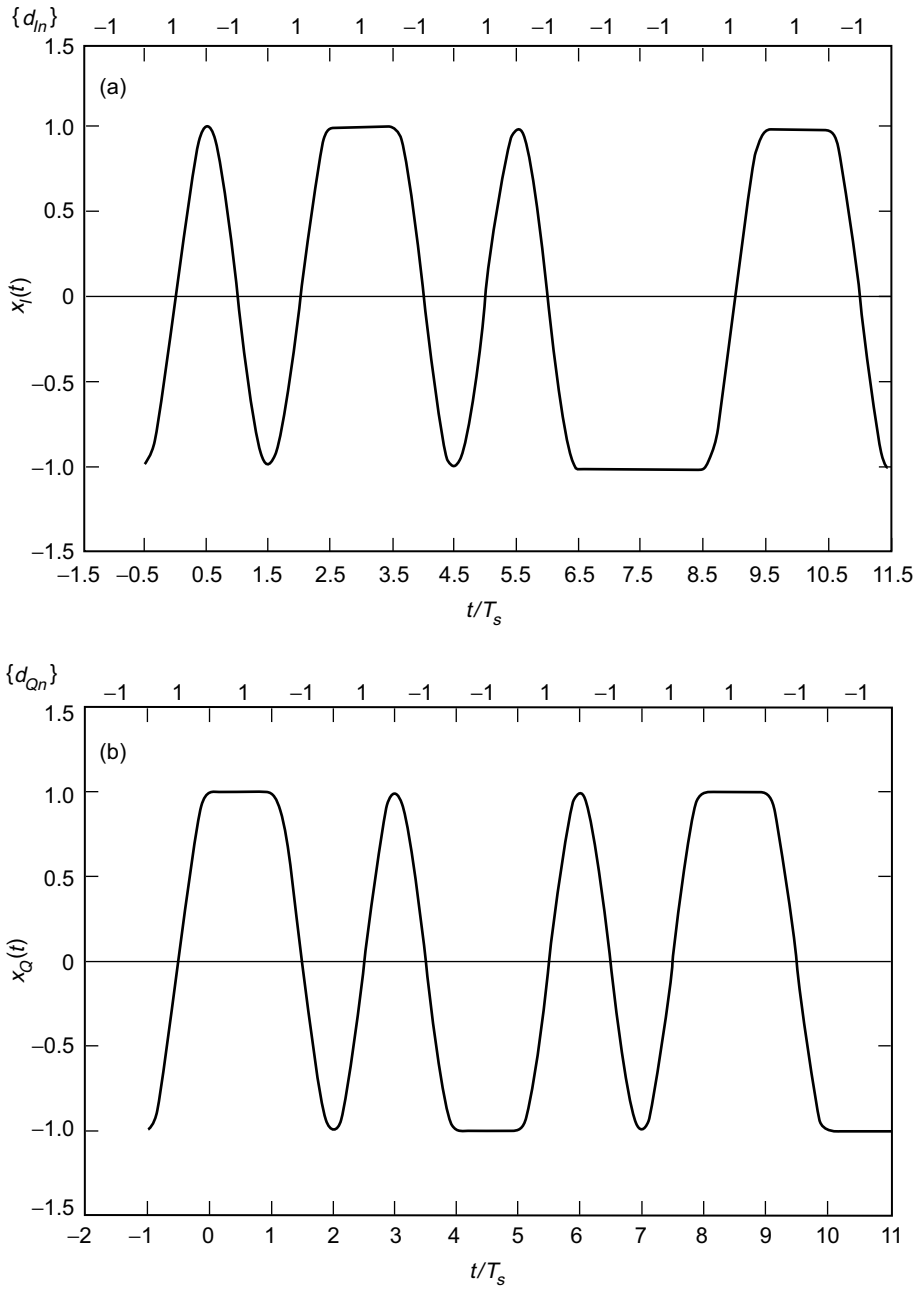


Fig. 3-2. IJF encoder output: (a) in-phase and (b) quadrature phase. Redrawn from [3].

$$p(t) = \sin^2\left(\frac{\pi\left(t+\frac{T_s}{2}\right)}{2T_s}\right), \quad -\frac{T_s}{2} \le t \le \frac{3T_s}{2} \tag{3.1-3}$$

then the I modulation

$$x_I(t) = \sum_{n=-\infty}^{\infty} d_{In} p(t - nT_s) \tag{3.1-4}$$

will be identical to that generated by the above IJF scheme, assuming the choice of odd and even waveforms as in (3.1-2). Similarly,

$$x_Q(t) = \sum_{n=-\infty}^{\infty} d_{Qn} p\left(t - \left(n+\frac{1}{2}\right)T_s\right) \tag{3.1-5}$$

would also be identical to that generated by the above IJF scheme. A quadrature modulation scheme formed from $x_I(t)$ of (3.1-4) and $x_Q(t)$ of (3.1-5) is precisely what Austin and Chang referred to as SQORC modulation [11], namely, independent I and Q staggered modulations with overlapping raised cosine pulses on each channel. The resulting carrier modulated waveform is described by

$$x(t) = x_I(t)\cos\omega_c t + x_Q(t)\sin\omega_c t \tag{3.1-6}$$

and is implementated as shown in Fig. 3-3. Figure 3-4 shows the equivalence of the transmitted SQORC baseband waveforms with the IJF-QPSK even and odd waveforms of (3.1-2) for a pair of consecutive data bits.

Although SQORC exhibits a 3-dB envelope fluctuation, it is nevertheless a highly bandwidth-efficient modulation. In fact, except for a normalization constant, its PSD is the *product* of the PSDs of OQPSK and MSK, i.e.,

$$S_{\text{SQORC}}(f) = \left(\frac{\sin\pi f T_s}{\pi f T_s}\right)^2 \frac{\cos^2 2\pi f T_s}{\left(1-16f^2T_s^2\right)^2} \tag{3.1-7}$$

which asymptotically decays as f^{-6}. It was with this in mind that Feher and Kato sought to tailor the transmitted waveform in such a fashion as to reduce the envelope fluctuation to near 0 dB, yet maintain the high bandwidth efficiency

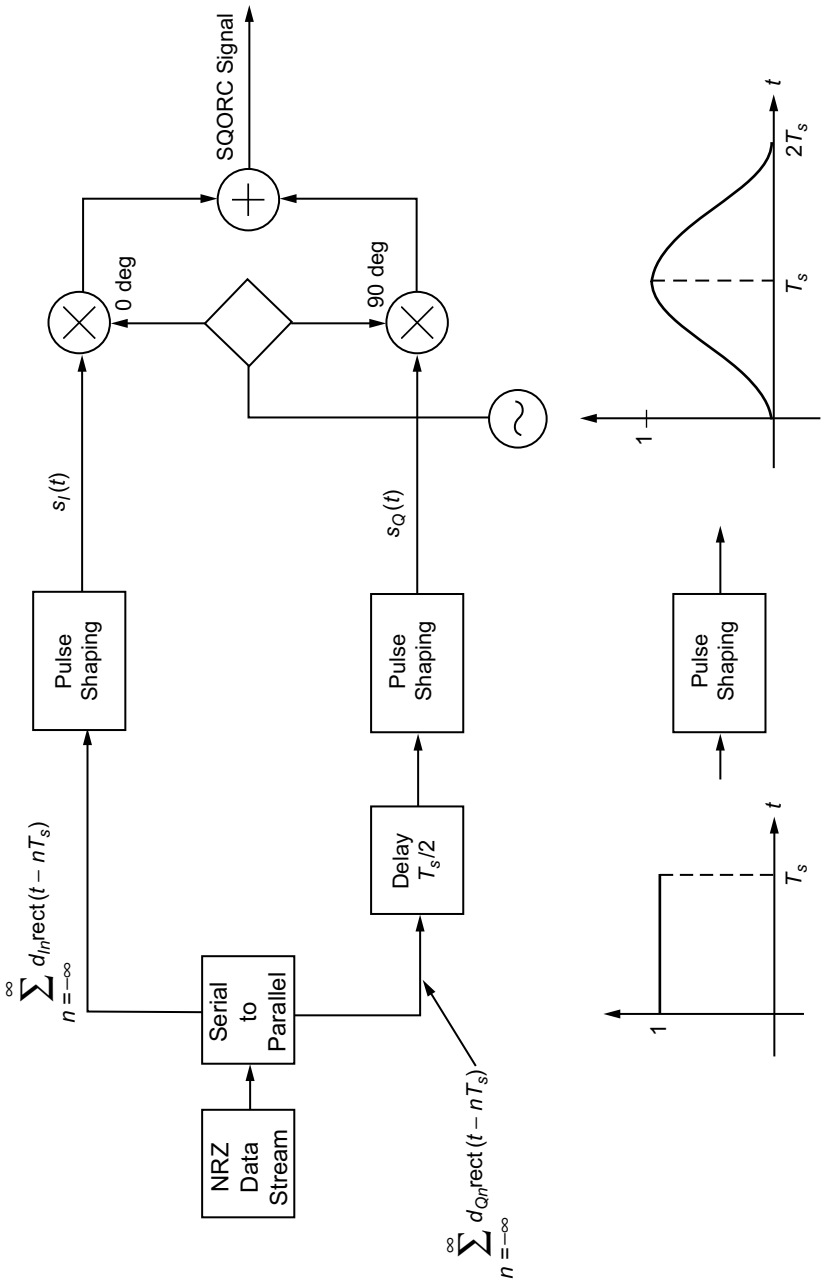


Fig. 3-3. Conceptual block diagram of an SQORC transmitter.

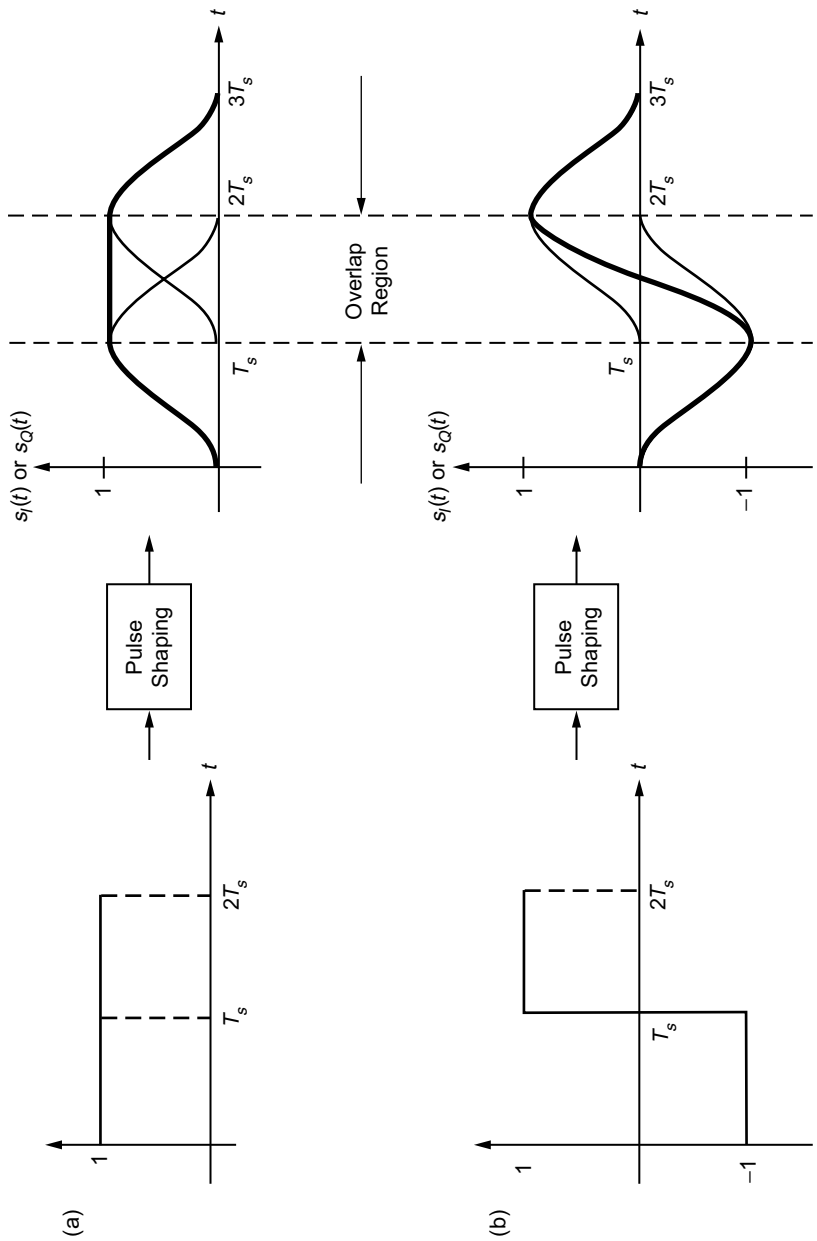


Fig. 3-4. Baseband SQORC waveforms: (a) consecutive symbols of like polarity and (b) consecutive symbols of alternating polarity.

inherent in SQORC. The specifics of how this is accomplished are wrapped up in the block labeled "cross-correlator" in Fig. 3-1 and are described below.

With reference to Fig. 3-1, in any given half-symbol, there are 16 possible combinations of the SQORC signal components, $S_I(t)$ and $S_Q(t)$, at the input of the cross-correlator. These combinations are illustrated in Fig. 3-5 and are composed of specific combinations of the signals $\pm 1, \pm \sin \pi t/T_s, \pm \cos \pi t/T_s$. For each of the I-Q component pairs, $S_I(t), S_Q(t)$, the cross-correlator generates a new I-Q component pair denoted by $s_I(t), s_Q(t)$, whose purpose is to reduce the envelope fluctuation of the resulting I and Q symbol steams. As such, the cross-correlator acts as a half-symbol waveform mapper. A mathematical description of the 16 possible cross-correlated signal combinations is given in Table 3-1.

Table 3-1. I and Q cross-correlated signal combinations.

$s_I(t)$ (or $s_Q(t)$)	$s_Q(t)$ (or $s_I(t)$)	Number of Combinations
$\pm \cos \pi t/T_s$	$\pm \sin \pi t/T_s$	4
$\pm A \cos \pi t/T_s$	$f_1(t)$ or $f_3(t)$	4
$\pm A \sin \pi t/T_s$	$f_2(t)$ or $f_4(t)$	4
$\pm A$	$\pm A$	4

The transition functions $f_i(t), i = 1, 2, 3, 4$ referred to in Table 3-1 are defined in the interval $0 \le t \le T_s/2$ by [2]

$$\left.\begin{aligned} f_1(t) &= 1 - (1-A)\cos^2 \frac{\pi t}{T_s} \\ f_2(t) &= 1 - (1-A)\sin^2 \frac{\pi t}{T_s} \\ f_3(t) &= -1 + (1-A)\cos^2 \frac{\pi t}{T_s} \\ f_4(t) &= -1 + (1-A)\sin^2 \frac{\pi t}{T_s} \end{aligned}\right\} \tag{3.1-8}$$

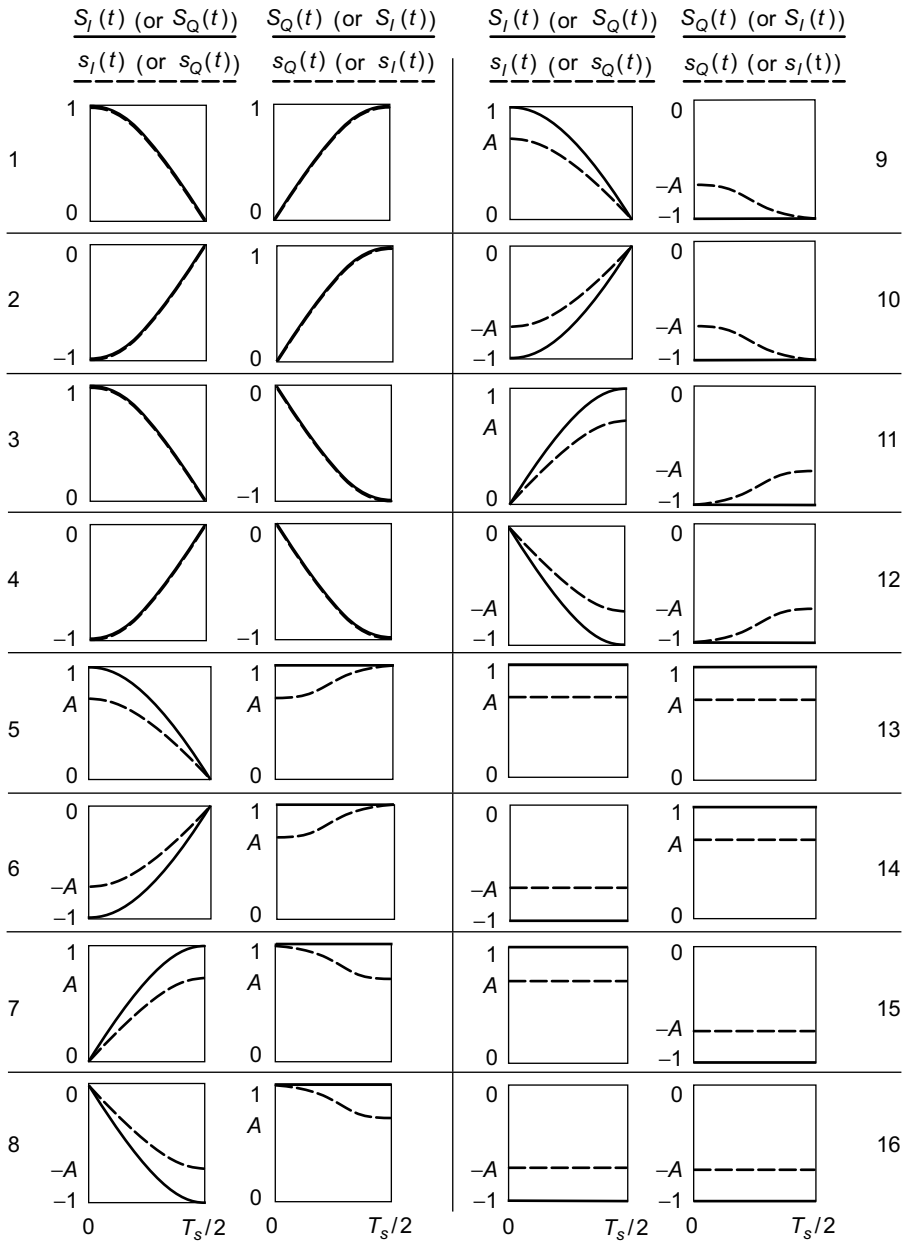


Fig. 3-5. FQPSK half-symbol waveform mapper
($A = 1 / \sqrt{2}$ for "constant" envelope).

where A is a transition parameter that can take on values in the interval $1/\sqrt{2}$ to 1 and is used to trade off between the amount of envelope fluctuation and bandwidth efficiency. The 16 mapped symbol pairs, $s_I(t), s_Q(t)$, that appear at the output of the cross-correlator, corresponding to the combinations in Table 3-1 are superimposed on the corresponding 16 possible input pairs, $S_I(t), S_Q(t)$, shown in Fig. 3-5. Note that the input pairs $S_I(t), S_Q(t)$ for combinations 1, 2, 3, and 4 are already constant envelope, and, thus, remapping of these signals is not necessary. For the same I and Q data sequences that generated the SQORC signal components in Figs. 3-2(a) and 3-2(b) and $A = 1/\sqrt{2}$, Figs. 3-6(a) and 3-6(b) illustrate the corresponding I and Q cross-correlator outputs. One can observe from these figures that at the uniform sample points on the I and Q channels, i.e., $t = nT_s$ and $t = (n + (1/2))T_s$ (n integer), the transmitted baseband signal is now *precisely* constant envelope. At other than the uniform sample points, the maximum fluctuation in the baseband signal envelope has been shown to be equal to 0.18 dB [2]—a small price for the significant bandwidth efficiency that has been afforded by this modulation, as will now be demonstrated.

Illustrated in Fig. 3-7 is the PSD of unfiltered FQPSK (as described above) along with that corresponding to other modulations previously discussed in this monograph. Figure 3-8 illustrates the corresponding plots of out-of-band energy versus normalized bandwidth, $BT_b = B/R_b$, as computed from

$$P_{ob} = 1 - \frac{\int_{-B}^{B} S_m(f)\,df}{\int_{-\infty}^{\infty} S_m(f)\,df} \tag{3.1-9}$$

When compared with OQPSK and MSK, which are both constant envelope, unfiltered FQPSK, which is virtually constant envelope, clearly provides an improvement in spectral efficiency. When compared with constant envelope GMSK, however, filtering must be applied to FQPSK in order to make it comparable in spectral efficiency. The PSD of FQPSK-B is superimposed on Figs. 3-7 and 3-8 and clearly shows a spectral advantage when compared, for example, with $BT_b = 0.5$ GMSK.

3.2 A Symbol-by-Symbol Cross-Correlator Mapping for FQPSK

In Ref. 3, the original characterization of FQPSK in terms of a cross-correlation operation performed on the pair of IJF encoder outputs every half-symbol interval was reformulated into a mapping performed directly on the input I and Q data sequences every full symbol interval. To do this, 16 waveforms, $s_i(t); i = 0, 1, 2, \cdots, 15$, were defined over the interval $-T_s/2 \le t \le T_s/2$, which

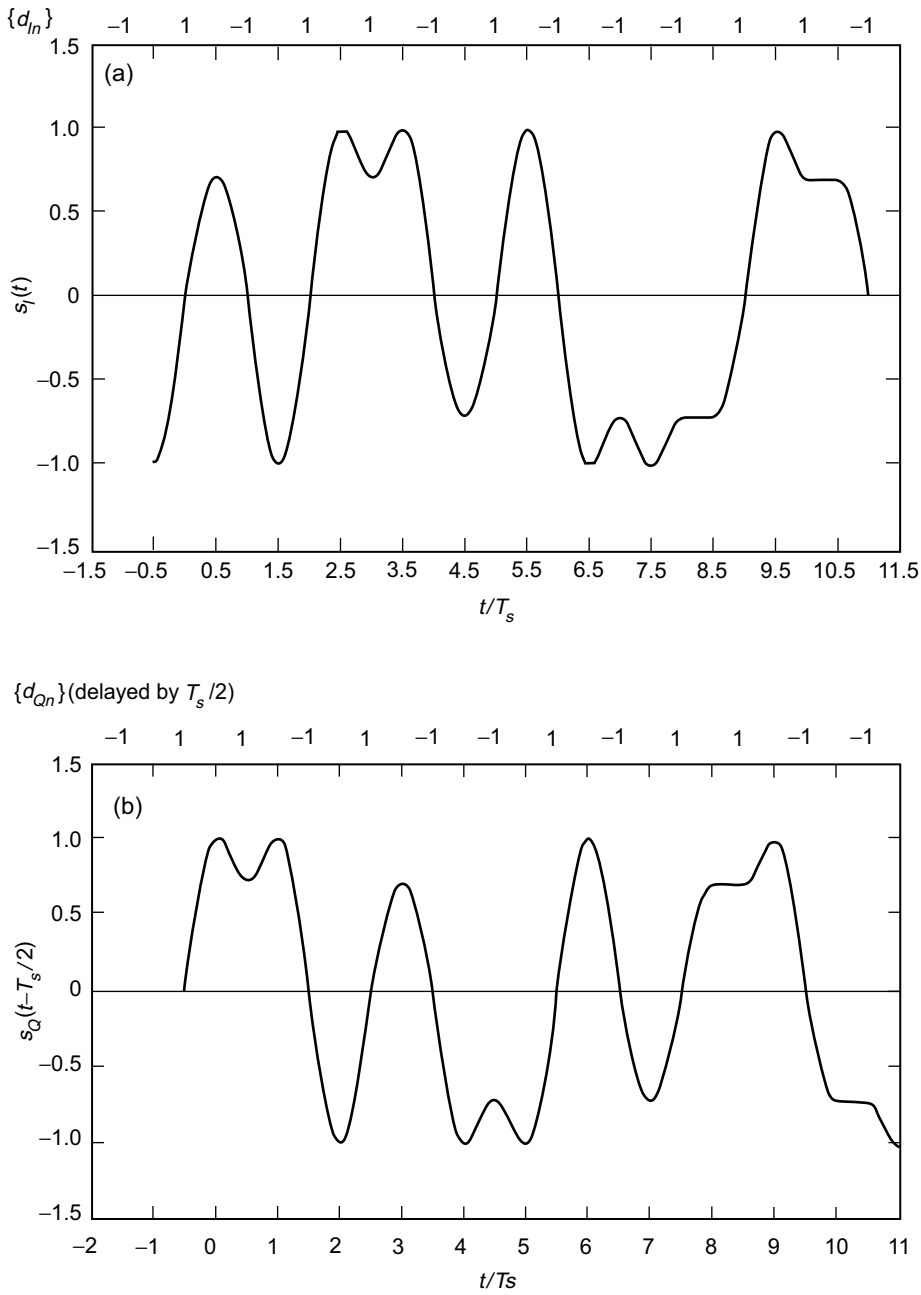


Fig. 3-6. FQPSK (XPSK) output: (a) in-phase and (b) quadrature-phase. Redrawn from [3].

collectively formed a transmitted signaling set for the I and Q channels. The particular I and Q waveforms chosen for any particular T_s-s signaling interval on each channel depended on the most recent data transition on that channel as well as the two most recent successive transitions on the other channel and such identified FQPSK as a modulation with inherent memory. The specific details are presented in Ref. 3 and are summarized as follows: Define (see Fig. 3-9)

$$\left.\begin{array}{lll}
s_0(t) = & A, & -\frac{T_s}{2} \le t \le \frac{T_s}{2}, \quad s_8(t) = -s_0(t) \\[2ex]
s_1(t) = & \begin{cases} A, & -\frac{T_s}{2} \le t \le 0, \\ 1-(1-A)\cos^2\frac{\pi t}{T_s}, & 0 \le t \le \frac{T_s}{2}, \end{cases} & s_9(t) = -s_1(t) \\[2ex]
s_2(t) = & \begin{cases} 1-(1-A)\cos^2\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le 0, \\ A, & 0 \le t \le \frac{T_s}{2}, \end{cases} & s_{10}(t) = -s_2(t) \\[2ex]
s_3(t) = & 1-(1-A)\cos^2\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le \frac{T_s}{2}, \quad s_{11}(t) = -s_3(t) \\[2ex]
s_4(t) = & A\sin\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le \frac{T_s}{2}, \quad s_{12}(t) = -s_4(t) \\[2ex]
s_5(t) = & \begin{cases} A\sin\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le 0, \\ \sin\frac{\pi t}{T_s}, & 0 \le t \le \frac{T_s}{2}, \end{cases} & s_{13}(t) = -s_5(t) \\[2ex]
s_6(t) = & \begin{cases} \sin\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le 0, \\ A\sin\frac{\pi t}{T_s}, & 0 \le t \le \frac{T_s}{2}, \end{cases} & s_{14}(t) = -s_6(t) \\[2ex]
s_7(t) = & \sin\frac{\pi t}{T_s}, & -\frac{T_s}{2} \le t \le \frac{T_s}{2}, \quad s_{15}(t) = -s_7(t)
\end{array}\right\} \tag{3.2-1}$$

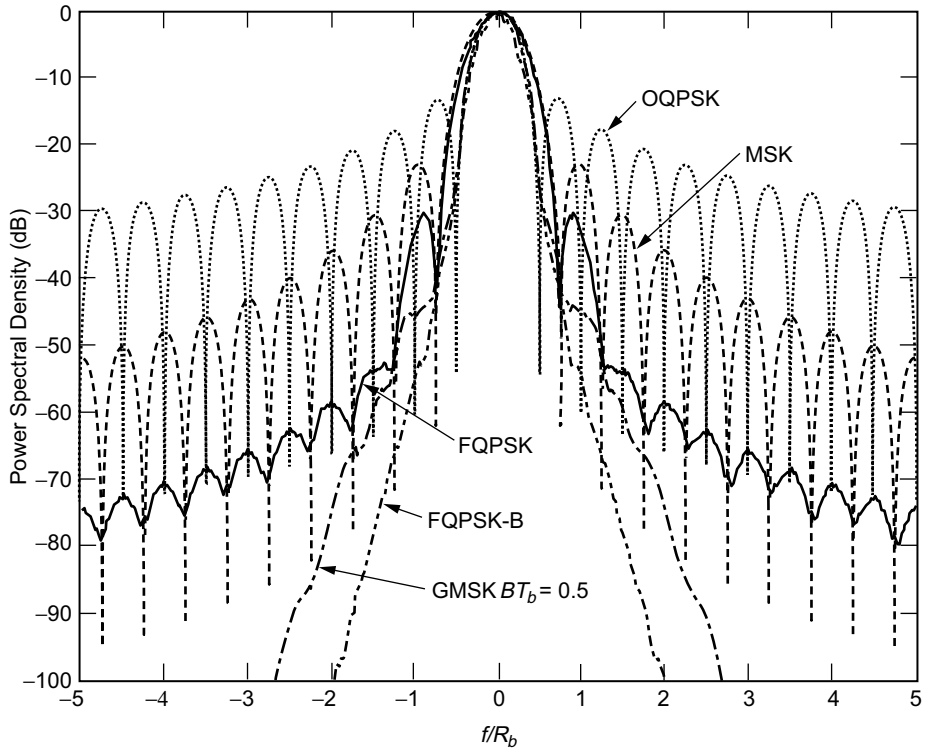


Fig. 3-7. Power spectral density of various modulations.

Note that for any value of A other then unity, e.g., $A = 1/\sqrt{2}$, $s_5(t)$ and $s_6(t)$ as well as their negatives, $s_{13}(t)$ and $s_{14}(t)$, will have a discontinuous slope at their midpoints (i.e., at $t = 0$) whereas the remaining 12 waveforms all have a continuous slope throughout their defining interval. Also, all 16 waveforms have zero slope at their endpoints and thus, concatenation of any pair of these will not result in a slope discontinuity. We will return to the issue of the discontinuous midpoint slope shortly.

The mapping function that assigns the particular baseband I and Q channel waveforms transmitted in the nth signaling interval chosen from the set in (3.2-1) is specified in terms of the transition properties of the I and Q data symbol sequences. For example, if $d_{I,n-1} = 1, d_{I,n} = 1$ (i.e., no transition on the I sequence and both data bits positive), then the transmitted I-channel signal, $y_I(t) = s_I(t)$, in the nth signaling interval $(n - (1/2))\,T_s \le t \le (n + (1/2))\,T_s$ is chosen as follows.

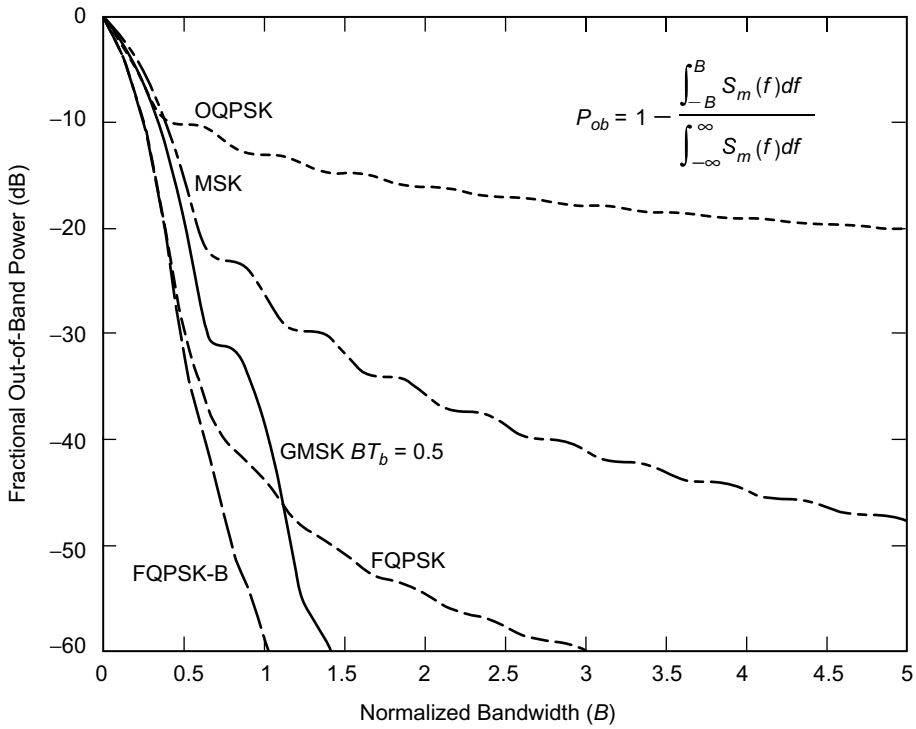


Fig. 3-8. Out-of-band power of various modulations.

A. $y_I(t) = s_0(t - nT_s)$ if $d_{Q,n-2}, d_{Q,n-1}$ results in no transition and $d_{Q,n-1}, d_{Q,n}$ results in no transition.

B. $y_I(t) = s_1(t - nT_s)$ if $d_{Q,n-2}, d_{Q,n-1}$ results in no transition and $d_{Q,n-1}, d_{Q,n}$ results in a transition (positive or negative).

C. $y_I(t) = s_2(t - nT_s)$ if $d_{Q,n-2}, d_{Q,n-1}$ results in a transition (positive or negative) and $d_{Q,n-1}, d_{Q,n}$ results in no transition.

D. $y_I(t) = s_3(t - nT_s)$ if $d_{Q,n-2}, d_{Q,n-1}$ results in a transition (positive or negative) and $d_{Q,n-1}, d_{Q,n}$ results in a transition (positive or negative).

Without belaboring the details, the assignments for the remaining three combinations of $d_{I,n-1}$ and $d_{I,n}$ follow similarly. Finally, making use of the signal properties in (3.2-1), the mapping conditions for the I-channel baseband output can be summarized in a concise form described by Table 3-2(a):

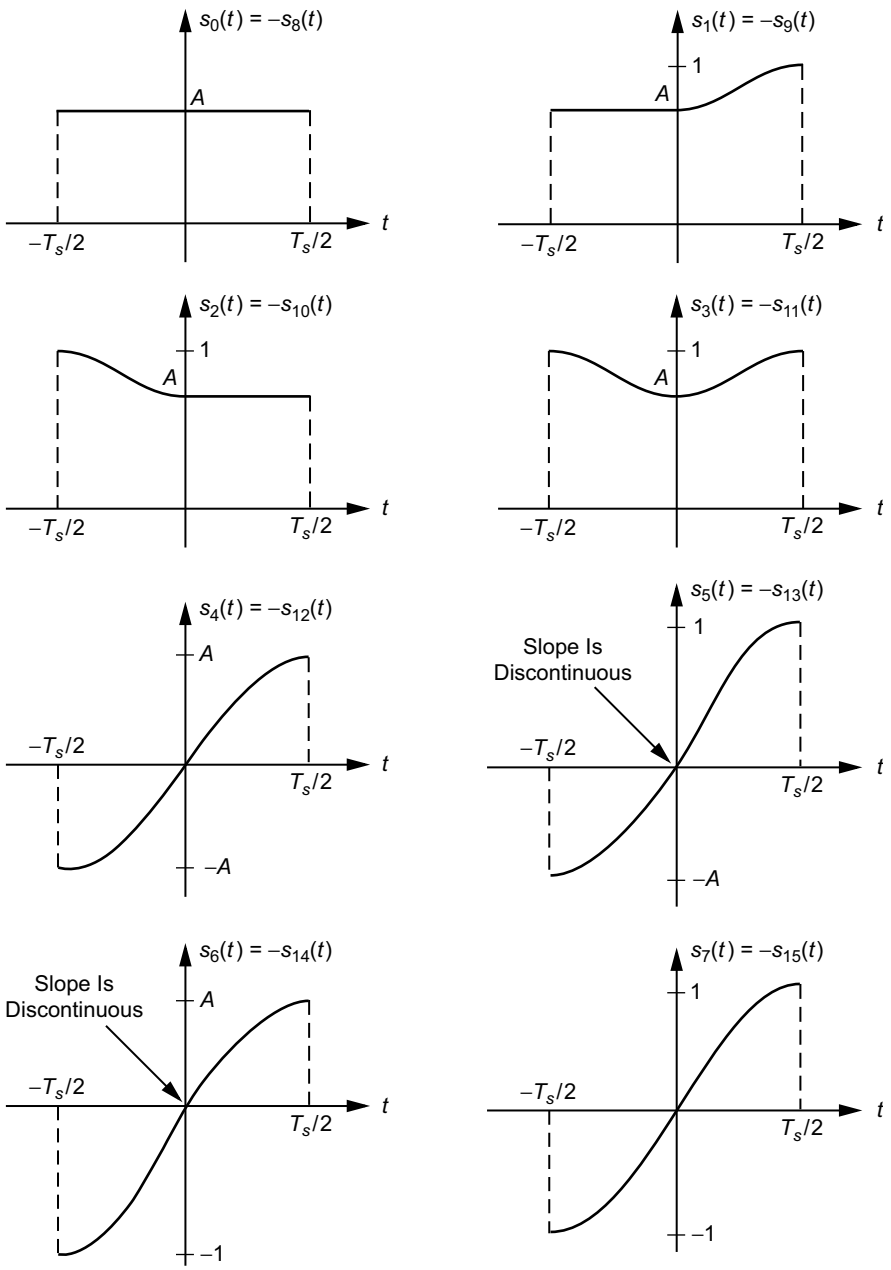


Fig. 3-9. FQPSK full-symbol waveforms
($A = 1 / \sqrt{2}$ for "constant" envelope). Redrawn from [3].

Table 3-2(a). Mapping for inphase (I)-channel baseband signal, $y_I(t)$, in the interval $(n - [1/2])\,T_s \le t \le (n + [1/2])\,T_s$.

$\left\lvert \frac{d_{In} - d_{I,n-1}}{2} \right\rvert$	$\left\lvert \frac{d_{Q,n-1} - d_{Q,n-2}}{2} \right\rvert$	$\left\lvert \frac{d_{Qn} - d_{Q,n-1}}{2} \right\rvert$	$s_I(t)$
0	0	0	$d_{In}s_0\,(t - nT_s)$
0	0	1	$d_{In}s_1\,(t - nT_s)$
0	1	0	$d_{In}s_2\,(t - nT_s)$
0	1	1	$d_{In}s_3\,(t - nT_s)$
1	0	0	$d_{In}s_4\,(t - nT_s)$
1	0	1	$d_{In}s_5\,(t - nT_s)$
1	1	0	$d_{In}s_6\,(t - nT_s)$
1	1	1	$d_{In}s_7\,(t - nT_s)$

A similar construction for the baseband Q-channel transmitted waveform, $y_Q\,(t) = s_Q\,(t - T_s/2)$, in the nth signaling interval, $nT_s \le t \le (n+1)\,T_s$, in terms of the transition properties of the I and Q data symbol sequences, $\{d_{In}\}$ and $\{d_{Qn}\}$, leads to Table 3-2(b):

Table 3-2(b). Mapping for quadrature (Q)-channel baseband signal, $y_Q(t)$, in the interval $nT_s \le t \le (n + 1)\,T_s$.

$\left\lvert \frac{d_{Qn} - d_{Q,n-1}}{2} \right\rvert$	$\left\lvert \frac{d_{I,n} - d_{I,n-1}}{2} \right\rvert$	$\left\lvert \frac{d_{I,n+1} - d_{I,n}}{2} \right\rvert$	$s_Q(t)$
0	0	0	$d_{Qn}s_0\,(t - nT_s)$
0	0	1	$d_{Qn}s_1\,(t - nT_s)$
0	1	0	$d_{Qn}s_2\,(t - nT_s)$
0	1	1	$d_{Qn}s_3\,(t - nT_s)$
1	0	0	$d_{Qn}s_4\,(t - nT_s)$
1	0	1	$d_{Qn}s_5\,(t - nT_s)$
1	1	0	$d_{Qn}s_6\,(t - nT_s)$
1	1	1	$d_{Qn}s_7\,(t - nT_s)$

Note from Tables 3-2(a) and 3-2(b) that subscript i of the transmitted signal $s_i\,(t - nT_s)$ or $s_i\,(t - (n + (1/2))\,T_s)$, as appropriate, is the binary-coded decimal (BCD) equivalent of the three transitions.

Applying the mappings in Tables 3-2(a) and 3-2(b) to the I and Q data sequences of Figs. 3-2(a) and 3-2(b) produces the identical I and Q baseband transmitted signals to those that would be produced by passing the I and Q IJF encoder outputs of this figure through the cross-correlator (half-symbol mapping) of the FQPSK (XPSK) scheme as described in Ref. 2 and illustrated in Figs. 3-6(a) and 3-6(b). Thus, we conclude that *for arbitrary I and Q input sequences, FQPSK can alternatively be generated by the symbol-by-symbol mappings of Tables 3-2(a) and 3-2(b) as applied to these sequences.*

3.3 Enhanced FQPSK

We now return to the issue of the midpoint slope discontinuity associated with four of the waveforms in Fig. 3-9. As discussed in Sec. 3.2, the symbol-by-symbol mapping representation of FQPSK identifies the fact that $s_5(t), s_6(t), s_{13}(t)$ and $s_{14}(t)$ have a slope discontinuity at their midpoint. Consequently, for random I and Q data symbol sequences, on the average, the transmitted FQPSK waveform will likewise have a slope discontinuity at one quarter of the uniform sampling time instants. To prevent this from occurring, we now redefine these four transmitted signals in a manner analogous to $s_1(t), s_2(t), s_9(t), s_{10}(t)$, namely,

$$\left.\begin{array}{ll}
s_5(t) = \begin{cases} \sin\dfrac{\pi t}{T_s} + (1-A)\sin^2\dfrac{\pi t}{T_s}, & -\dfrac{T_s}{2} \le t \le 0 \\ \sin\dfrac{\pi t}{T_s}, & 0 \le t \le \dfrac{T_s}{2} \end{cases} & s_{13}(t) = -s_5(t) \\
s_6(t) = \begin{cases} \sin\dfrac{\pi t}{T_s}, & -\dfrac{T_s}{2} \le t \le 0 \\ \sin\dfrac{\pi t}{T_s} - (1-A)\sin^2\dfrac{\pi t}{T_s}, & 0 \le t \le \dfrac{T_s}{2} \end{cases} & s_{14}(t) = -s_6(t)
\end{array}\right\} \quad (3.3\text{-}1)$$

Note that the signals $s_5(t), s_6(t), s_{13}(t), s_{14}(t)$ as defined in (3.3-1) do not have a slope discontinuity at their midpoint nor, for that matter, anywhere else in the defining interval. Also, the zero slope at their endpoints has been preserved. Therefore, using (3.3-1) in place of the corresponding signals of (3.2-1) will result in a modified FQPSK signal that has no slope discontinuity anywhere in time regardless of the value of A. Figure 3-10 illustrates a comparison of the

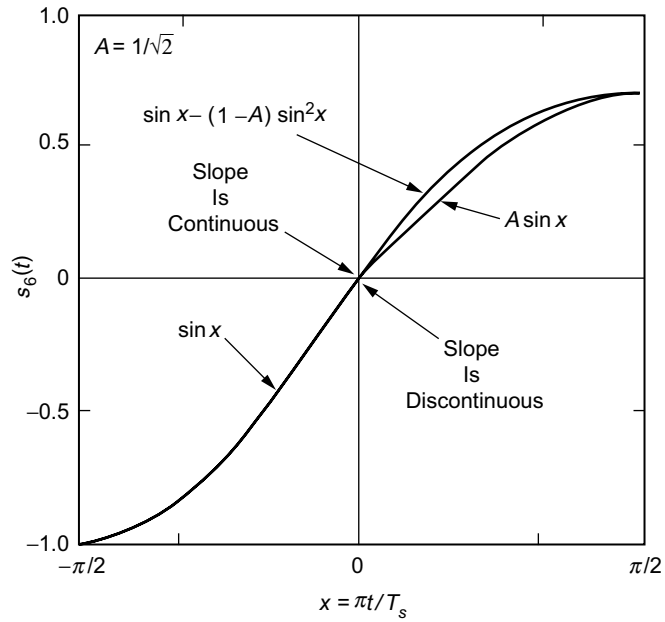


Fig. 3-10. Original and enhanced FQPSK pulse shapes. Redrawn from [3].

signal $s_6(t)$ of (3.3-1) with that of (3.2-1) for a value of $A = 1/\sqrt{2}$. Figure 3-11 illustrates the power spectral density of conventional FQPSK and its enhancement obtained by using the waveforms of (3.3-1) as replacements for those in (3.2-1). The significant improvement in spectral rolloff rate is clear from a comparison of the two.

As it currently stands, the signal set chosen for enhanced FQPSK has a symmetry property for $s_0(t), s_1(t), s_2(t), s_3(t)$ that is not present for $s_4(t), s_5(t), s_6(t), s_7(t)$. In particular, $s_1(t)$ and $s_2(t)$ are each composed of one-half of $s_0(t)$ and one-half of $s_3(t)$, i.e., the portion of $s_1(t)$ from $t = -T_s/2$ to $t = 0$ is the same as that of $s_0(t)$, whereas the portion of $s_1(t)$ from $t = 0$ to $t = T_s/2$ is the same as that of $s_3(t)$ and vice versa for $s_2(t)$. To achieve the same symmetry property for $s_4(t) - s_7(t)$, one would have to reassign $s_4(t)$ as

$$s_4(t) = \begin{cases} \sin\dfrac{\pi t}{T_s} + (1-A)\sin^2\dfrac{\pi t}{T_s}, & -\dfrac{T_s}{2} \le t \le 0 \\ \\ \sin\dfrac{\pi t}{T_s} - (1-A)\sin^2\dfrac{\pi t}{T_s}, & 0 \le t \le \dfrac{T_s}{2} \end{cases} \qquad s_{12}(t) = -s_4(t) \tag{3.3-2}$$

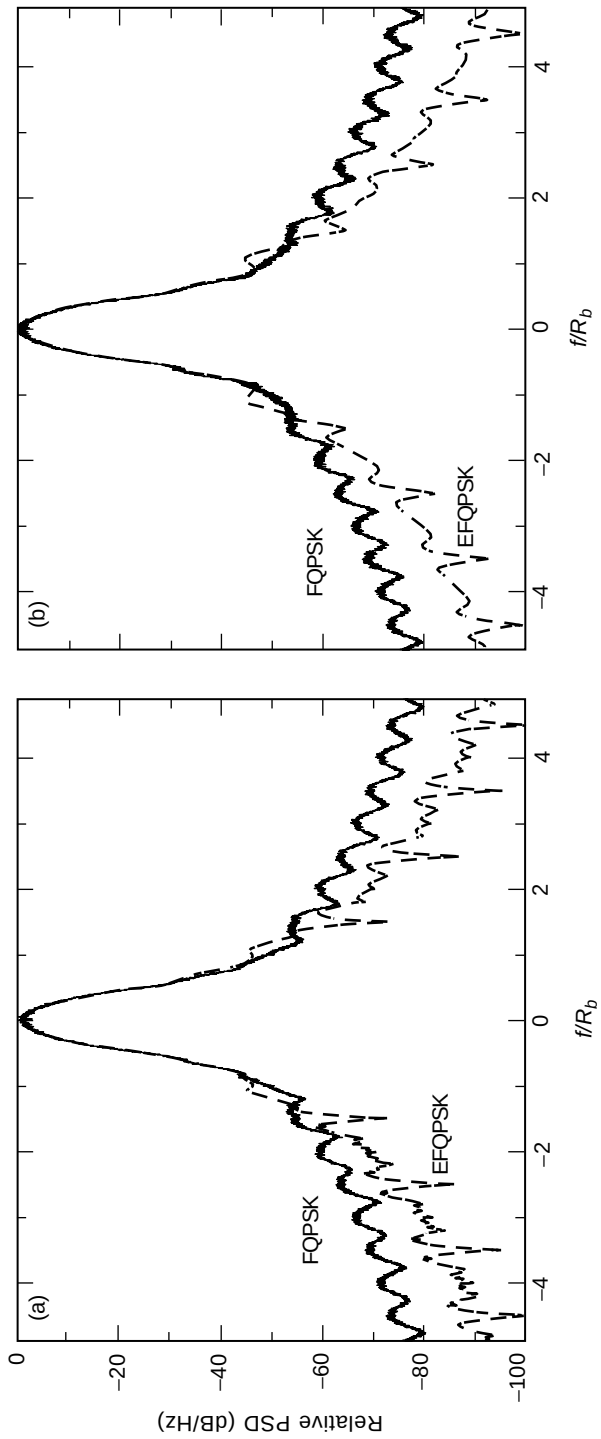


Fig. 3-11. Power spectrum of conventional and enhanced FQPSK: (a) without high-power amplifier and (b) with high-power amplifier. Redrawn from [3].

This minor change, which produces a complete symmetry in the waveform set, has an advantage from the standpoint of hardware implementation and produces a negligible change in spectral properties of the transmitted waveform. Nevertheless, for the remainder of the discussion, we shall ignore this minor change and assume the version of enhanced FQPSK first introduced in this section.

3.4 Interpretation of FQPSK as a Trellis-Coded Modulation

The I and Q mappings given in Tables 3-2a and 3-2b can alternatively be described in terms of the (0,1) representation of the I and Q data

$$\left.\begin{aligned} D_{In} &\triangleq \frac{1-d_{In}}{2} \\ D_{Qn} &\triangleq \frac{1-d_{Qn}}{2} \end{aligned}\right\} \tag{3.4-1}$$

which both range over the set (0,1). Then, defining the BCD representation of the indices i and j by

$$\left.\begin{aligned} i &= I_3 \times 2^3 + I_2 \times 2^2 + I_1 \times 2^1 + I_0 \times 2^0 \\ j &= Q_3 \times 2^3 + Q_2 \times 2^2 + Q_1 \times 2^1 + Q_0 \times 2^0 \end{aligned}\right\} \tag{3.4-2}$$

with

$$\left.\begin{aligned} I_0 &= D_{Qn} \oplus D_{Q,n-1}, & Q_0 &= D_{I,n+1} \oplus D_{In} \\ I_1 &= D_{Q,n-1} \oplus D_{Q,n-2}, & Q_1 &= D_{In} \oplus D_{I,n-1} = I_2 \\ I_2 &= D_{In} \oplus D_{I,n-1}, & Q_2 &= D_{Qn} \oplus D_{Q,n-1} = I_0 \\ I_3 &= D_{In}, & Q_3 &= D_{Qn} \end{aligned}\right\} \tag{3.4-3}$$

we have $y_I(t) = s_i(t - nT_s)$ and $y_Q(t) = s_j(t - (n+1/2))T_s)$. That is, in each symbol interval $((n-(1/2))T_s \le t \le (n+(1/2))T_s$ for $y_I(t)$ and $nT_s \le t \le (n+1)T_s$ for $y_Q(t))$, the I and Q channel baseband signals are each chosen from a set of 16 signals, $s_i(t), i = 0, 1, \cdots, 15$, in accordance with the 4-bit

BCD representations of their indices defined by (3.4-2) together with (3.4-3). A graphical illustration of the implementation of this mapping is given in Fig. 3-12.

Another interpretation of the mapping in Fig. 3-12 is as a 16-state trellis code with 2 binary (0,1) inputs $D_{I,n+1}, D_{Qn}$ and 2 waveform outputs, $s_i(t), s_j(t)$, where the state is defined by the 4-bit sequence, $D_{In}, D_{I,n-1}, D_{Q,n-1}, D_{Q,n-2}$. The trellis is illustrated in Fig. 3-13, and the transition mapping is given in Table 3-3. In this table, the entries in the column labeled "input" correspond to the values of the two input bits, $D_{I,n+1}, D_{Qn}$, that result in the transition, while the entries in the column labeled "output" correspond to the subscripts, i and j, of the pair of symbol waveforms, $s_i(t), s_j(t)$, that are outputted.

3.5 Optimum Detection

In designing receivers for FQPSK, the approach taken in the past has been to disregard the inherent memory of the transmitted modulation (actually the

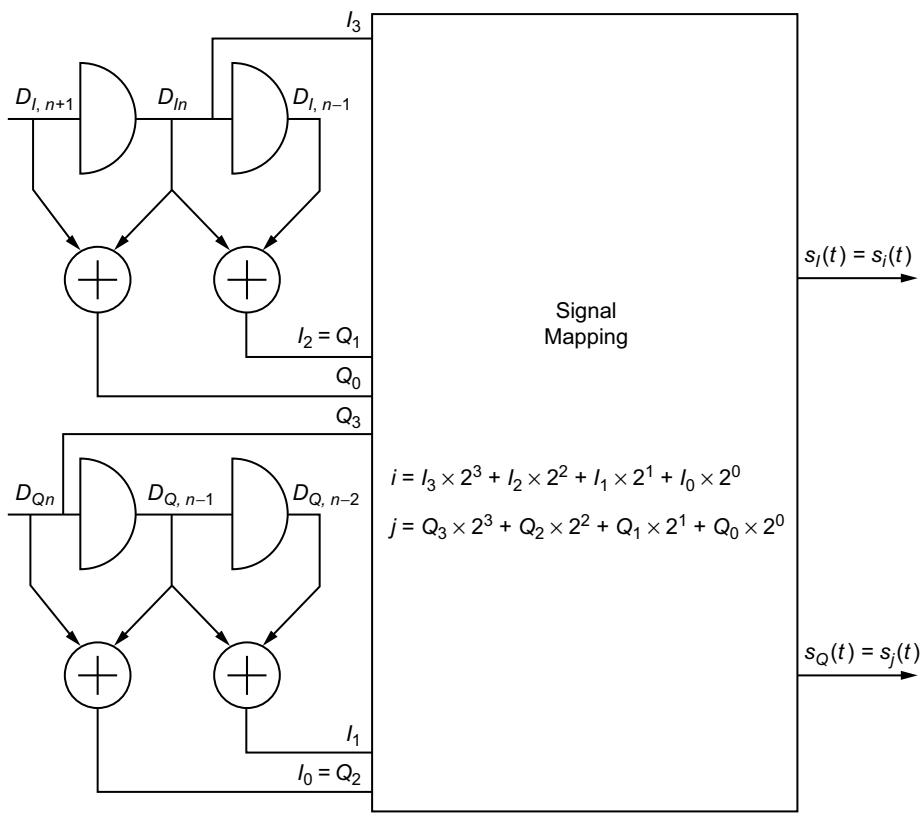


Fig. 3-12. Alternative implementation of FQPSK baseband signals. Redrawn from [3].

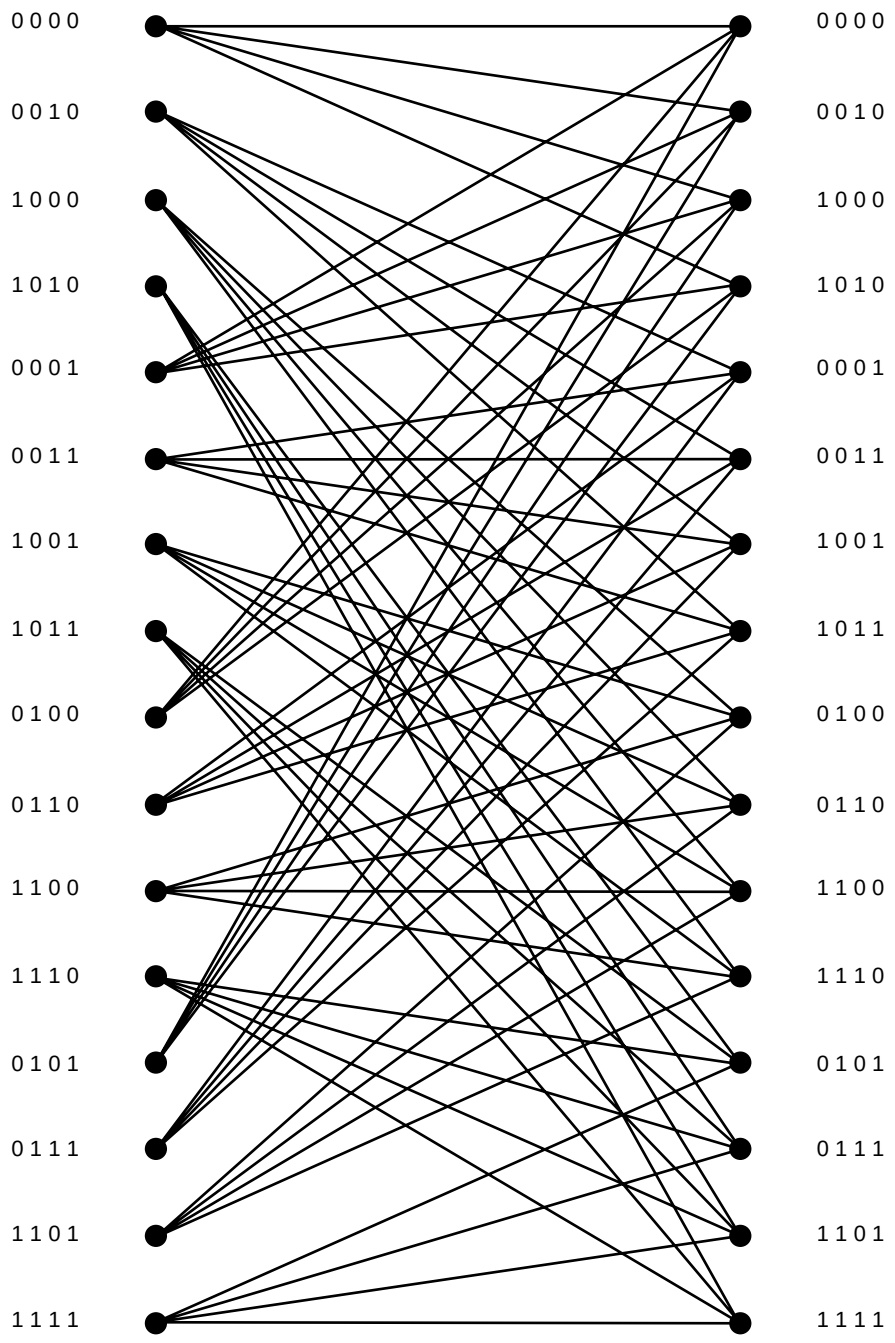


Fig. 3-13. The 16-state trellis diagram for FQPSK. Redrawn from [3].

Table 3-3. Trellis state transitions.

Current State	Input	Output	Next State
0 0 0 0	0 0	0 0	0 0 0 0
0 0 0 0	0 1	1 12	0 0 1 0
0 0 0 0	1 0	0 1	1 0 0 0
0 0 0 0	1 1	1 13	1 0 1 0
0 0 1 0	0 0	3 4	0 0 0 1
0 0 1 0	0 1	2 8	0 0 1 1
0 0 1 0	1 0	3 5	1 0 0 1
0 0 1 0	1 1	2 9	1 0 1 1
1 0 0 0	0 0	12 3	0 1 0 0
1 0 0 0	0 1	13 15	0 1 1 0
1 0 0 0	1 0	12 2	1 1 0 0
1 0 0 0	1 1	13 14	1 1 1 0
1 0 1 0	0 0	15 7	0 1 0 1
1 0 1 0	0 1	14 11	0 1 1 1
1 0 1 0	1 0	15 6	1 1 0 1
1 0 1 0	1 1	14 10	1 1 1 1
0 0 0 1	0 0	2 0	0 0 0 0
0 0 0 1	0 1	3 12	0 0 1 0
0 0 0 1	1 0	2 1	1 0 0 0
0 0 0 1	1 1	3 13	1 0 1 0
0 0 1 1	0 0	1 4	0 0 0 1
0 0 1 1	0 1	0 8	0 0 1 1
0 0 1 1	1 0	1 5	1 0 0 1
0 0 1 1	1 1	0 9	1 0 1 1
1 0 0 1	0 0	14 3	0 1 0 0
1 0 0 1	0 1	15 15	0 1 1 0
0 1 1 0	0 0	7 6	0 0 0 1
0 1 1 0	0 1	6 10	0 0 1 1
0 1 1 0	1 0	7 7	1 0 0 1
0 1 1 0	1 1	6 11	1 0 1 1
1 1 0 0	0 0	8 1	0 1 0 0
1 1 0 0	0 1	9 13	0 1 1 0
1 1 0 0	1 0	8 0	1 1 0 0
1 1 0 0	1 1	9 12	1 1 1 0
1 1 1 0	0 0	11 5	0 1 0 1
1 1 1 0	0 1	10 9	0 1 1 1

Table 3-3 (cont'd). Trellis state transitions.

Current State	Input	Output	Next State
1 1 1 0	1 0	11 4	1 1 0 1
1 1 1 0	1 1	10 8	1 1 1 1
0 1 0 1	0 0	6 2	0 0 0 0
0 1 0 1	0 1	7 14	0 0 1 0
0 1 0 1	1 0	6 3	1 0 0 0
0 1 0 1	1 1	7 15	1 0 1 0
0 1 1 1	0 0	5 6	0 0 0 1
0 1 1 1	0 1	4 10	0 0 1 1
0 1 1 1	1 0	5 7	1 0 0 1
0 1 1 1	1 1	4 11	1 0 1 1
1 1 0 1	0 0	10 1	0 1 0 0
1 1 0 1	0 1	11 13	0 1 1 0
1 1 0 1	1 0	10 0	1 1 0 0
1 1 0 1	1 1	11 12	1 1 1 0
1 1 1 1	0 0	9 5	0 1 0 1
1 1 1 1	0 1	8 9	0 1 1 1
1 1 1 1	1 0	9 4	1 1 0 1
1 1 1 1	1 1	8 8	1 1 1 1

original formulation of FQPSK as the combination of SQORC and a cross-correlator [half-symbol mapper] followed by I-Q carrier modulation did not recognize the existence of this inherent memory) and employ symbol-by-symbol detection. Based on the reformulation of FQPSK as a trellis-coded modulation (TCM) as discussed in Sec. 3.4, an optimum receiver would be one that exploited this characterization. In accordance with the foregoing representation of FQPSK as a TCM with 16 states, the optimum receiver (employing the Viterbi Algorithm) for FQPSK is illustrated in Fig. 3-14. Note that the 16 waveforms of Fig. 3-9 are not all equal in energy, and, thus, the matched filter outputs in this figure must be biased before applying them to the Viterbi decoder. Later on, we shall illustrate average BEP results obtained from a computer simulation of this receiver. For the moment, we shall just compare its asymptotic (limit of infinite energy-to-noise ratio) performance with that of the optimum receiver for conventional uncoded offset OQPSK based on normalized squared Euclidean distance, $d^2_{\min}/2\bar{E}_b$, where $\bar{E}_b$, denotes the average energy per bit. For the latter, $d^2_{\min}/2\bar{E}_b = 2$, which is the same as that for BPSK. For FQPSK, $d^2_{\min}/2\bar{E}_b$ is evaluated in Ref. 3 with the following results:

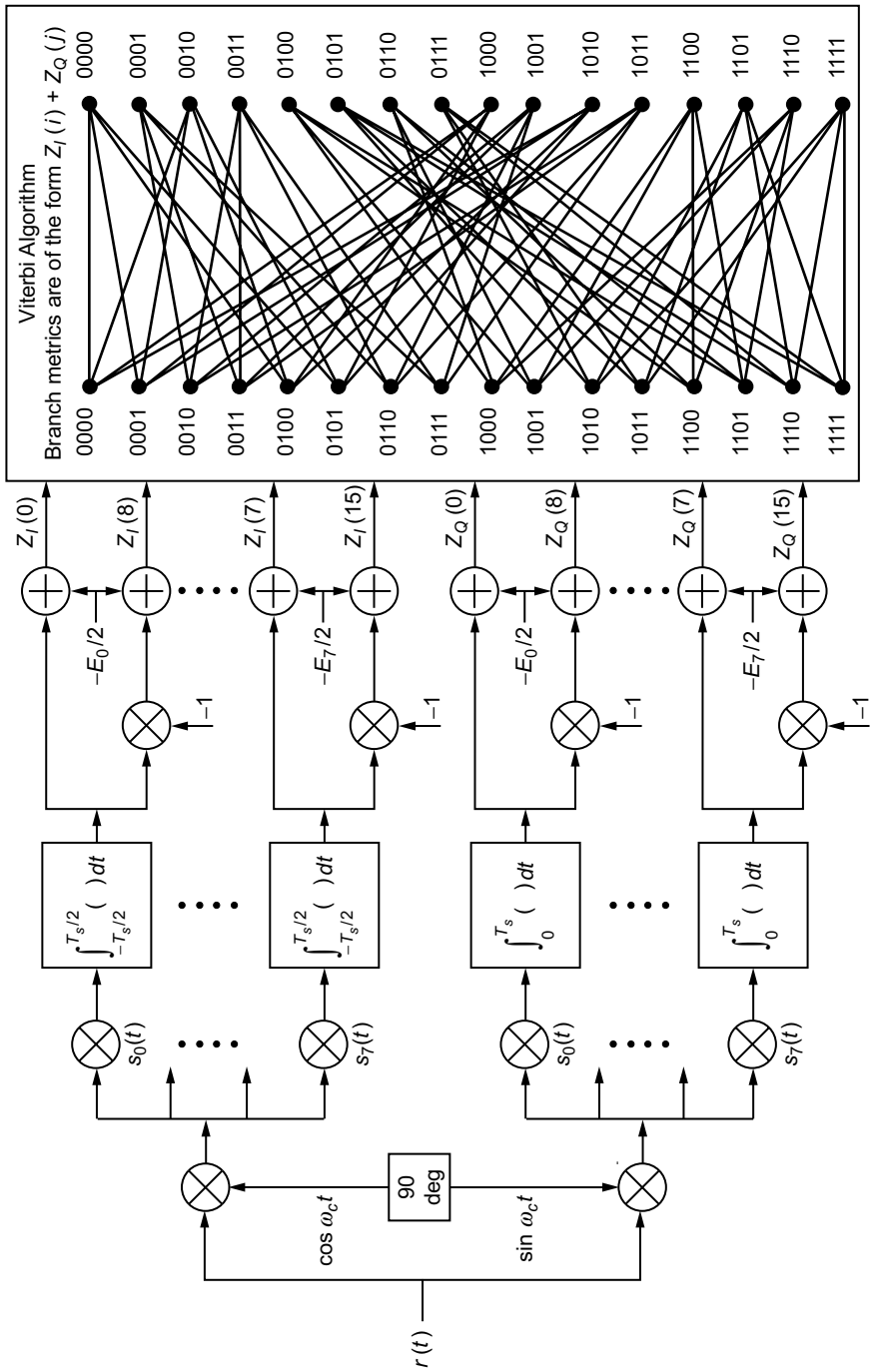


Fig. 3-14. Optimum trellis-coded receiver for FQPSK. Redrawn from [3].

$$\frac{d^2_{\min}}{2\bar{E}_b} = \frac{16\left[\frac{7}{4} - \frac{8}{3\pi} - A\left(\frac{3}{2} + \frac{4}{3\pi}\right) + A^2\left(\frac{11}{4} + \frac{4}{\pi}\right)\right]}{(7 + 2A + 15A^2)} \tag{3.5-1}$$

which for $A = 1/\sqrt{2}$ evaluates to

$$\frac{d^2_{\min}}{2\bar{E}_b} = 1.56 \tag{3.5-2}$$

Similarly, for enhanced FQPSK

$$\frac{d^2_{\min}}{2\bar{E}_b} = \frac{\left(3 - 6A + 15A^2\right)}{\frac{21}{8} - \frac{8}{3\pi} - A\left(\frac{1}{4} - \frac{8}{3\pi}\right) + \frac{29}{8}A^2} \tag{3.5-3}$$

which for $A = 1/\sqrt{2}$ coincidentally evaluates to (3.5-2), i.e., it is identical to that for FQPSK. Thus, the enhancement of FQPSK provided by using the waveforms of (3.3-1) as replacements for their equivalents in (3.2-1) is significantly beneficial from a spectral standpoint, with no penalty in receiver performance.

Finally, we conclude that as a trade against the significantly improved bandwidth efficiency afforded by FQPSK and its enhanced version relative to that of OQPSK, *an asymptotic loss of only 10 log(2/1.56) = 1.07 dB is experienced when optimum reception is employed.*[6]

3.6 Suboptimum Detection

As previously stated, FQPSK receivers traditionally use symbol-by-symbol detection, which is suboptimum and results in a significant loss relative to the performance of ideal OQPSK. In this section, we start by examining the extent of this degradation, following which we consider other suboptimum reception methods that also require less implementation complexity than the optimum receiver discussed in Sec. 3.5.

3.6.1 Symbol-by-Symbol Detection

Here we examine the performance of FQPSK when the detector makes decisions on a symbol-by-symbol basis, i.e., the inherent memory introduced by the trellis coding is ignored at the receiver. In order to understand how this can be accomplished, we will first establish the fact that in any typical transmission interval, there exists a fixed number (in particular, eight) of possible

[6] At smaller (finite) SNRs, the loss between uncoded OQPSK and trellis-decoded FQPSK will be even less.

waveforms (pulse shapes) that represent the FQPSK signal, and each of these occurs with equal probability. As such, from symbol-to-symbol, the FQPSK signal appears as an equiprobable M-ary signaling set (with $M = 8$) and thus can be detected accordingly. With this in mind, we shall investigate two possible simple structures, both of which are suboptimum relative to the trellis-coded receiver previously discussed that exploits the memory inherent in the modulation. The first structure is a standard offset QPSK receiver that employs simple I&Ds as detectors and, as such, ignores the pulse shaping associated with the above-mentioned M-ary symbol-by-symbol representation. The second structure, which shall be referred to as an average matched filter receiver, improves on the first one by replacing the I&Ds with matched filters, where the match is made to the average of the waveshapes in the M-ary signal set representation. Without loss in generality, the following description shall consider the case $n = 0$, corresponding to the I channel interval $-T_s/2 \leq t \leq T_s/2$ and the Q channel interval $0 \leq t \leq T_s$. We shall focus our attention only on the I channel, with the initial goal of defining the eight equally likely waveforms that typify an FQPSK waveform in the interval $0 \leq t \leq T_s$. To avoid confusion with the previously defined signals such as those defined in (3.2-1), we shall use upper-case notation, i.e., $S_i(t)\,,\ i = 0, 1, \cdots, 7$ to describe these new waveforms. As we shall see momentarily, each of these new waveforms is composed of the latter half (i.e., that which occurs in the interval $0 \leq t \leq T_s/2$) of the I channel waveform transmitted in the interval $-T_s/2 \leq t \leq T_s/2$, followed by the first half (i.e., that which occurs in the interval $T_s/2 \leq t \leq T_s$) of the I channel waveform transmitted in the interval $T_s/2 \leq t \leq 3T_s/2$. As stated above, only eight such possible combinations can exist, and all are equiprobable.

3.6.1.1 Signal Representation. In Ref. 3, it is shown that for $d_{I0} = 1$ and $s_I(t) = s_0(t)$ in the interval $-T_s/2 \leq t \leq T_s/2$, the transmitted signal, $S_i(t)$, for the interval $0 \leq t \leq T_s$ is composed of the latter half of $s_0(t)$ followed by the first half of either $s_0(t)\,, s_1(t)\,, s_{12}(t)$ or $s_{13}(t)$. Looking at the definitions of $s_0(t)\,, s_1(t)\,, s_{12}(t)\,, s_{13}(t)$ in (3.2-1), we see that this yields only two distinct possibilities for $S_i(t)$, namely,

$$\left.\begin{aligned} S_0(t) &= A, \quad 0 \leq t \leq T_s \\ S_1(t) &= \begin{cases} A, & 0 \leq t \leq \dfrac{T_s}{2} \\ A \sin \dfrac{\pi t}{T_s}, & \dfrac{T_s}{2} \leq t \leq T_s \end{cases} \end{aligned}\right\} \tag{3.6-1a}$$

both of which are equally likely.

Following a similar procedure (still for $d_{I0} = 1$), it can be shown that for each of the other possible waveforms in $-T_s/2 \leq t \leq T_s/2$, i.e., $s_1(t), s_2(t), s_3(t)$, $s_4(t), s_5(t), s_6(t)$ and $s_7(t)$, there are only two possible distinct waveforms in $0 \leq t \leq T_s$, which are again equally likely. These possibilities are summarized in Table 3-4.

Table 3-4. Possible distinct signal pairs.

Signal in $-T_s/2 \leq t \leq T_s/2$	Signal in $0 \leq t \leq T_s$
$s_1(t)$	$S_2(t), S_3(t)$
$s_2(t)$	$S_0(t), S_1(t)$
$s_3(t)$	$S_2(t), S_3(t)$
$s_4(t)$	$S_4(t), S_5(t)$
$s_5(t)$	$S_6(t), S_7(t)$
$s_6(t)$	$S_4(t), S_5(t)$
$s_7(t)$	$S_6(t), S_7(t)$

where the signals $S_2(t), S_3(t), S_4(t), S_5(t), S_6(t), S_7(t)$ are defined as

$$\left.\begin{aligned}
S_2(t) &= 1-(1-A)\cos^2\frac{\pi t}{T_s}, \quad 0 \leq t \leq T_s \\
S_3(t) &= \begin{cases} 1-(1-A)\cos^2\dfrac{\pi t}{T_s}, & 0 \leq t \leq \dfrac{T_s}{2} \\ \sin\dfrac{\pi t}{T_s}, & \dfrac{T_s}{2} \leq t \leq T_s \end{cases} \\
S_4(t) &= \begin{cases} A\sin\dfrac{\pi t}{T_s}, & 0 \leq t \leq \dfrac{T_s}{2} \\ A, & \dfrac{T_s}{2} \leq t \leq T_s \end{cases} \\
S_5(t) &= A\sin\frac{\pi t}{T_s}, \quad 0 \leq t \leq T_s \\
S_6(t) &= \begin{cases} \sin\dfrac{\pi t}{T_s}, & 0 \leq t \leq \dfrac{T_s}{2} \\ 1-(1-A)\cos^2\dfrac{\pi t}{T_s}, & \dfrac{T_s}{2} \leq t \leq T_s \end{cases} \\
S_7(t) &= \sin\frac{\pi t}{T_s}, \quad 0 \leq t \leq T_s
\end{aligned}\right\} \tag{3.6-1b}$$

In comparing the performances of the suboptimum receivers of FQPSK to that of uncoded OQPSK, we shall reference them all to the same average transmitted power, $\bar{P}$, or, equivalently, the same average energy-per-bit to noise spectral density ratio, $\bar{E}_b/N_0 = \bar{P}T_b/N_0$. In order to do this, we must first compute the energy, $E_i = \int_0^{T_s} S_i^2(t)\,dt$, of each of the waveforms in (3.5-4a) and (3.5-4b) and take their average. The results are summarized below [3]:

$$\left.\begin{aligned}
E_0 &= A^2 T_s \\
E_1 &= \frac{3}{4}A^2 T_s \\
E_2 &= \left(\frac{3}{8} + \frac{1}{4}A + \frac{3}{8}A^2\right) T_s \\
E_3 &= \left(\frac{7}{16} + \frac{1}{8}A + \frac{3}{16}A^2\right) T_s \\
E_4 &= \frac{3}{4}A^2 T_s \\
E_5 &= \frac{1}{2}A^2 T_s \\
E_6 &= \left(\frac{7}{16} + \frac{1}{8}A + \frac{3}{16}A^2\right) T_s \\
E_7 &= \frac{1}{2}T_s
\end{aligned}\right\} \tag{3.6-2}$$

and

$$\bar{E} = \frac{1}{8}\sum_{i=0}^{7} E_i = \left(\frac{7 + 2A + 15A^2}{32}\right) T_s \tag{3.6-3}$$

Since the average power transmitted in the I channel is one-half the total (I+Q) average transmitted power, $\bar{P}$, then we have

$$\frac{\bar{P}}{2} = \frac{\bar{E}}{T_s} = \frac{7 + 2A + 15A^2}{32} \tag{3.6-4}$$

or, equivalently, the average energy-per-symbol is given by

$$\bar{P}T_s \triangleq \bar{E}_s = 2\bar{E}_b = \frac{7 + 2A + 15A^2}{16} T_s \tag{3.6-5}$$

Note that the evaluation of average energy per symbol based on the symbol-by-symbol M-ary representation of FQPSK is identical to that obtained from the representation as a trellis-coded modulation. Also note that for $A = 1$, which corresponds to SQORC modulation, we have $\bar{E}_s = (4/3)T_s$ which is consistent with the original discussions of this modulation in Ref. 11.

In accordance with our discussion at the beginning of Sec. 3.6.1, we shall consider two suboptimum receivers for symbol-by-symbol detection of FQPSK, the difference being the manner in which the detector is matched to the received signal. For the average matched filter case, the detector is implemented as a multiplication of the received signal by $\bar{S}(t) \triangleq \frac{1}{8}\sum_{i=0}^{7} S_i(t)$, followed by an I&D filter and binary hard decision device (see Fig. 3-15). For the OQPSK receiver, the detector is purely an I&D (i.e., matched to a rectangular pulse), which is tantamount to assuming $\bar{S}(t) = 1$. Thus, we can cover both cases at the same time, leaving $\bar{S}(t)$ as an arbitrary premultiplication pulse shape and later substitute the appropriate waveform.

Assuming the M-ary symbol-by-symbol representation of FQPSK just described, then the decision variable Z in Fig. 3-15 is given by

$$Z = \int_0^{T_s} S(t)\,\bar{S}(t)\,dt + \int_0^{T_s} n(t)\,\bar{S}(t)\,dt \triangleq \bar{Z} + N \tag{3.6-6}$$

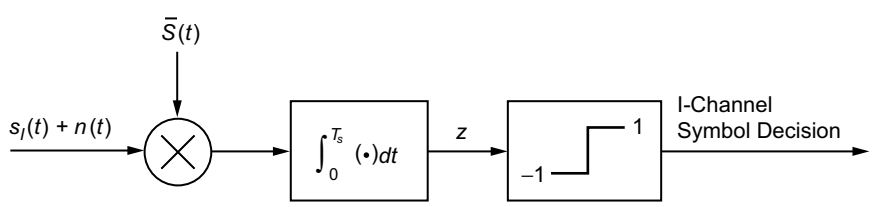


Fig. 3-15. Suboptimum receiver for FQPSK, based on symbol-by-symbol detection. Redrawn from [3].

where $S(t)$ is the transmitted waveform in $0 \leq t \leq T_s$ and ranges over the set of eight waveforms in (3.6-1a) and (3.6-2b) with equal probability. The random variable, N, is zero mean Gaussian with variance $\sigma_N^2 = N_0 E_{\bar{S}}/2$ where $E_{\bar{S}} \triangleq \int_0^{T_s} \bar{S}^2(t)\,dt$. Thus, the I channel symbol error probability (the same as the Q channel symbol error probability) conditioned on the particular $S(t) = S_i(t)$ corresponding to the transmitted symbol $d_{I0} = 1$ is shown to be

$$P_{si}(E) = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{1}{N_0}\frac{\left(\int_0^{T_s} S_i(t)\,\bar{S}(t)\,dt\right)^2}{E_{\bar{S}}}}\right) \tag{3.6-7}$$

and hence the average symbol error probability is given by

$$P_s(E) \triangleq \frac{1}{8}\sum_{i=0}^{7} P_{si}(E) \tag{3.6-8}$$

3.6.1.2 Conventional OQPSK Receiver. For the conventional OQPSK receiver, we set $\bar{S}(t) = 1$, or equivalently, $E_{\bar{S}} = T_s$ in (3.6-7), resulting in

$$\begin{aligned} P_{si}(E) &= \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{T_s}{N_0}\left(\frac{1}{T_s}\int_0^{T_s} S_i(t)\,dt\right)^2}\right) \\ &= \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\left(\frac{32}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}\left(\frac{E_i}{T_s}\right)^2}\right) \end{aligned} \tag{3.6-9}$$

Substituting the average energies from (3.6-2) in (3.6-9) for each signal and then performing the average as in (3.6-8) gives the final desired result for average symbol error probability, namely,

$$
\begin{aligned}
P_{si}(E) = {} & \frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\left(\frac{32A^4}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}}\right) \\
& + \frac{1}{8}\,\mathrm{erfc}\left(\sqrt{\left(\frac{18A^4}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}}\right) \\
& + \frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\left(\frac{\left(3+2A+3A^2\right)^2}{2\left(7+2A+15A^2\right)}\right)\frac{\bar{E}_b}{N_0}}\right) \\
& + \frac{1}{8}\,\mathrm{erfc}\left(\sqrt{\left(\frac{\left(7+2A+3A^2\right)^2}{8\left(7+2A+15A^2\right)}\right)\frac{\bar{E}_b}{N_0}}\right) \\
& + \frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\left(\frac{8A^4}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}}\right) \\
& + \frac{1}{16}\,\mathrm{erfc}\left(\sqrt{\left(\frac{8}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}}\right)
\end{aligned}
\tag{3.6-10}
$$

3.6.1.3 Average Matched Filter Receiver. For the average matched filter, we need to compute the correlations of each of the pulse shapes in (3.6-1a) and (3.6-1b) with the average pulse shape, $\bar{S}(t)$, and also the energy, $E_{\bar{S}}$, of the average pulse shape. Rewriting (3.6-7) in a form analogous to (3.6-9), namely,

$$
P_{si}(E) = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\left(\frac{32}{7+2A+15A^2}\right)\frac{\bar{E}_b}{N_0}\,\frac{\left(\frac{1}{T_s}\int_0^{T_s} S_i(t)\,\bar{S}(t)\,dt\right)^2}{\frac{1}{T_s}E_{\bar{S}}}}\right)
\tag{3.6-11}
$$

then the results necessary to evaluate (3.6-11) are tabulated below:

$$\frac{1}{T_s}\int_0^{T_s} S_0(t)\,\bar{S}(t)\,dt = \frac{A}{4}\left[\frac{1}{2}+\frac{2}{\pi}+A\left(\frac{3}{2}+\frac{2}{\pi}\right)\right]$$

$$\frac{1}{T_s}\int_0^{T_s} S_1(t)\,\bar{S}(t)\,dt = \frac{1}{T_s}\int_0^{T_s} S_4(t)\,\bar{S}(t)\,dt$$

$$= \frac{A}{4}\left[\frac{1}{2}+\frac{5}{3\pi}+A\left(1+\frac{7}{3\pi}\right)\right]$$

$$\frac{1}{T_s}\int_0^{T_s} S_2(t)\,\bar{S}(t)\,dt = \frac{1}{4}\left[\frac{3}{8}+\frac{4}{3\pi}+A\left(\frac{3}{4}+\frac{2}{\pi}\right)+A^2\left(\frac{7}{8}+\frac{2}{3\pi}\right)\right] \tag{3.6-12}$$

$$\frac{1}{T_s}\int_0^{T_s} S_3(t)\,\bar{S}(t)\,dt = \frac{1}{T_s}\int_0^{T_s} S_6(t)\,\bar{S}(t)\,dt$$

$$= \frac{1}{4}\left[\frac{7}{16}+\frac{4}{3\pi}+A\left(\frac{5}{8}+\frac{7}{3\pi}\right)+A^2\left(\frac{7}{16}+\frac{1}{3\pi}\right)\right]$$

$$\frac{1}{T_s}\int_0^{T_s} S_5(t)\,\bar{S}(t)\,dt = \frac{A}{2}\left[\frac{1}{4}+\frac{2}{3\pi}+A\left(\frac{1}{4}+\frac{4}{3\pi}\right)\right]$$

$$\frac{1}{T_s}\int_0^{T_s} S_7(t)\,\bar{S}(t)\,dt = \frac{1}{2}\left[\frac{1}{4}+\frac{2}{3\pi}+A\left(\frac{1}{4}+\frac{4}{3\pi}\right)\right]$$

and

$$\frac{1}{T_s}E_{\bar{S}} = \frac{1}{16}\left[(1+A)^2\left(\frac{3}{2}+\frac{4}{\pi}\right)+\frac{3}{8}(1-A)^2-2\left(1-A^2\right)\left(\frac{1}{2}+\frac{2}{3\pi}\right)\right] \tag{3.6-13}$$

Finally, substituting (3.6-12) and (3.6-13) into (3.6-11) and averaging as in (3.6-8) gives the desired result, which we shall not explicitly write in closed form.

3.6.2 Average Bit-Error Probability Performance

The average BEP of the two suboptimum receivers discussed in Sec. 3.5.2.2 is illustrated in Fig. 3-16 for the case $A = 1/\sqrt{2}$. These results are obtained directly

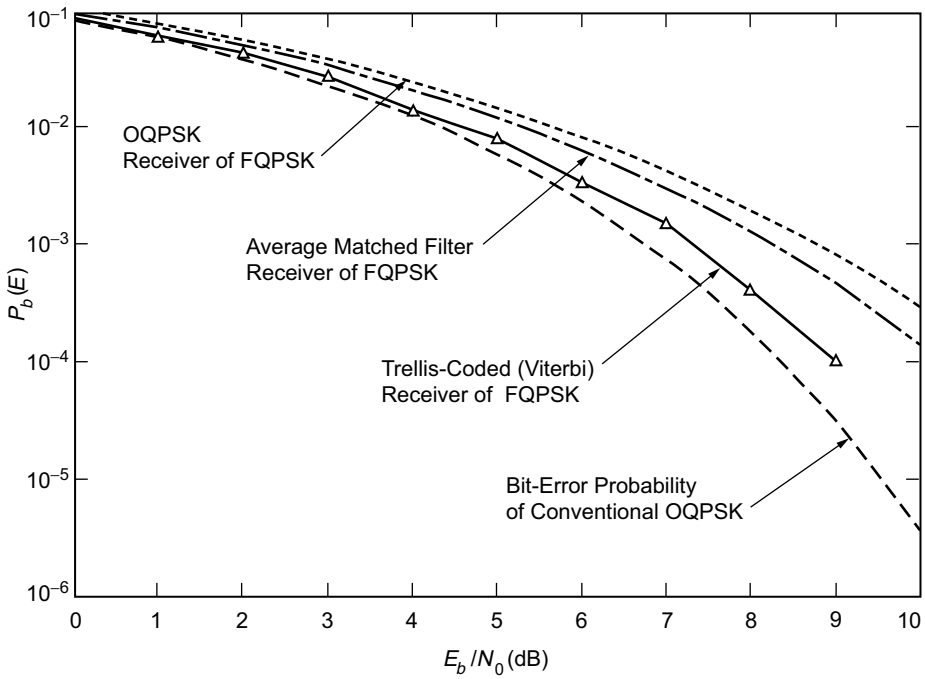


Fig. 3-16. Bit-error probability performance of various receivers of FQPSK modulation (reference curve is bit-error probability of OQPSK). Redrawn from [3].

from (3.6-10) for the OQPSK receiver and from (3.6-8) in combination with (3.6-11)–(3.6-13) for the average matched filter receiver. Also included in this figure is the performance corresponding to the optimum uncoded OQPSK receiver (same performance as for uncoded BPSK), i.e., $P_b(E) = (1/2)\text{erfc}\sqrt{E_b/N_0}$ as well as simulation results obtained for the optimum trellis-coded receiver of Fig. 3-14. We observe, as one might expect, that the average matched filter receiver outperforms the OQPSK receiver, since an attempt to match the transmitted pulse shape (even on an average basis) is better than no attempt at all. We also observe that the trellis-coded receiver at $P_b(E) = 10^{-4}$ is more than 1 dB better than the average matched filter receiver, granted that the latter is considerably simpler in implementation. Finally, for the same average BEP, the trellis-coded receiver of FQPSK is only about 0.6 dB inferior to uncoded OQPSK performance, which is a relatively small penalty paid for the vast improvement in PSD afforded by the former relative to the latter.

3.6.3 Further Receiver Simplifications and FQPSK-B Performance

As discussed above, symbol-by-symbol detection of FQPSK pays a significant penalty relative to optimum detection at the cost of a significant increase in implementation complexity. In Ref. 4, the authors introduce a simplified Viterbi-type receiver that still exploits the memory inherent in the modulation but has a reduced trellis and significantly less complexity (fewer states in the VA), with only a slight BEP degradation compared to that of the optimum (full Viterbi) receiver. The reduction in the number of states of the trellis comes about by grouping signal waveforms (see Fig. 3-9) with similar characteristics and using a single averaged matched filter for each group. In this sense, this simplified receiver acts as a compromise between the very simple averaged matched filter of Sec. 3.6.1.3, which uses a single matched filter equal to the average of all waveforms, and the optimum receiver, which uses a full bank of filters individually matched to each waveform. In discussing this simplified receiver, we shall consider its performance (obtained by simulation) in the context of FQPSK-B since, as shown in Figs. 3-7 and 3-8, FQPSK-B is much more spectrally efficient than unfiltered FQPSK but has ISI introduced by the filtering.

With reference to Fig. 3-14, the optimum Viterbi receiver for FQPSK implements a bank of matched filters to produce the correlations of the received signal with each of the 16 waveforms in Fig. 3-9 (actually only eight matched filters are required for each of the I and Q channels since $s_8(t), s_9(t), \cdots, s_{15}(t)$ are the negatives of $s_0(t), s_1(t), \cdots, s_7(t)\, s_0(t), s_1(t), \cdots, s_7(t)$). The Viterbi receiver then acts on these 32 correlation values (actually the branch metrics are the sum of each of the 16 energy-biased correlations from the corresponding I and Q channels) to produce a joint decision on the I and Q signals transmitted in a given symbol interval. A simplified FQPSK (or FQPSK-B) Viterbi receiver can be formed by observing certain similarities in the waveforms of Fig. 3-9 and thereby separating them into four different groups. For example, for $A = 1$, waveforms $s_0(t), s_1(t), \cdots, s_3(t)$ would become identical (similarly, for waveforms $s_8(t), s_9(t), \cdots, s_{11}(t)$). Thus, it is reasonable for arbitrary A to form group 1 as $s_0(t), s_1(t), \cdots, s_3(t)$ and also group 3 as $s_8(t), s_9(t), \cdots, s_{11}(t)$. Likewise, for $A = 1$, waveforms $s_4(t), s_5(t), \cdots, s_7(t)$ would become identical (similarly, for waveforms $s_{12}(t), s_{12}(t), \cdots, s_{15}(t)$). Consequently, for arbitrary A, it is again reasonable to form group 2 as $s_4(t), s_5(t), \cdots, s_7(t)$ and also group 4 as $s_{12}(t), s_{12}(t), \cdots, s_{15}(t)$. A close examination of the mapping that produced the trellis of Fig. 3-13 reveals that with this grouping, the trellis-coded structure of FQPSK splits into two independent I and Q two-state trellises. By independent, we mean that the I and Q decisions are no longer produced jointly but rather separately by individual VAs acting on the enery-biased correlations derived from the I and Q demodulated signals, respectively. A block diagram of

the simplified FQPSK-B receiver[7] is illustrated in Fig. 3-17. First, the received signal is demodulated and correlated against the arithmetic average of each of the above waveform groups as given by

$$\left.\begin{aligned} q_0(t) &= \frac{1}{4}\sum_{i=0}^{3} s_i(t) \\ q_1(t) &= \frac{1}{4}\sum_{i=4}^{7} s_i(t) \\ q_2(t) &= \frac{1}{4}\sum_{i=8}^{11} s_i(t) = -q_0(t) \\ q_3(t) &= \frac{1}{4}\sum_{i=12}^{15} s_i(t) = -q_1(t) \end{aligned}\right\} \tag{3.6-14}$$

and illustrated in Fig. 3-18. Since $q_2(t)$ and $q_3(t)$ are the negatives of $q_0(t)$ and $q_1(t)$, only two correlators (matched filters) are needed for the I and Q channels. Next, the VA metrics are formed by energy-biasing the matched filter outputs, where the appropriate energies are now those corresponding to the group averaged waveforms in (3.6-14). Figure 3-19 shows the two-state trellis associated with the grouped signals for each of the I and Q channels. The trellis is symmetric and has two transitions to each state. The two VAs could also be combined into a single four-state VA. When compared to the full Viterbi receiver of Fig. 3-14, the simplified Viterbi receiver has 12 fewer correlators and an 8-fold reduction in the number of VA computations performed per decoded bit.

Figure 3-20 illustrates the simulated BEP of the simplified and full FQPSK-B Viterbi receivers and compares them with that of the conventional symbol-by-symbol I&D (also referred to as sample-and-hold (S&H) in Ref. 4) receiver and ideal QPSK (or equivalently OQPSK). The simulated channel includes a nonlinear solid-state power amplifier (SSPA) operating in full saturation, which restores a constant envelope to the transmitted signal. For the full 16-state Viterbi receiver, a truncation path length (number of bits decoding delay before making decisions) of 50 bits was used. Due to the short constraint length nature of the reduced trellis in the simplified receiver, a truncation path length of only

[7] The primary difference between the simplified receiver for unfiltered FQPSK and FQPSK-B is the inclusion of an appropriate bandpass filter at the input to the receiver to match the filtering applied to the modulation at the transmitter.

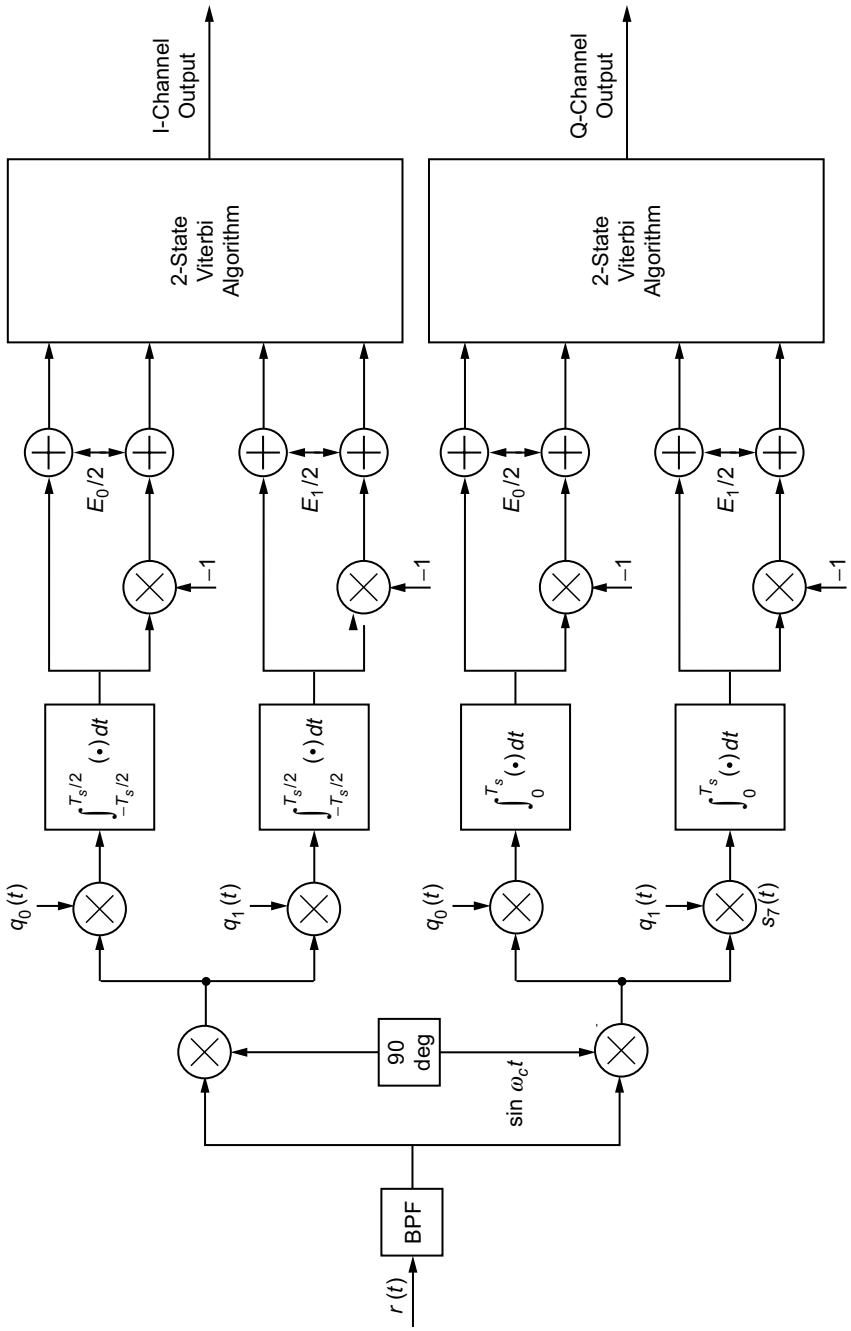


Fig. 3-17. Simplified FQPSK-B Viterbi receiver.

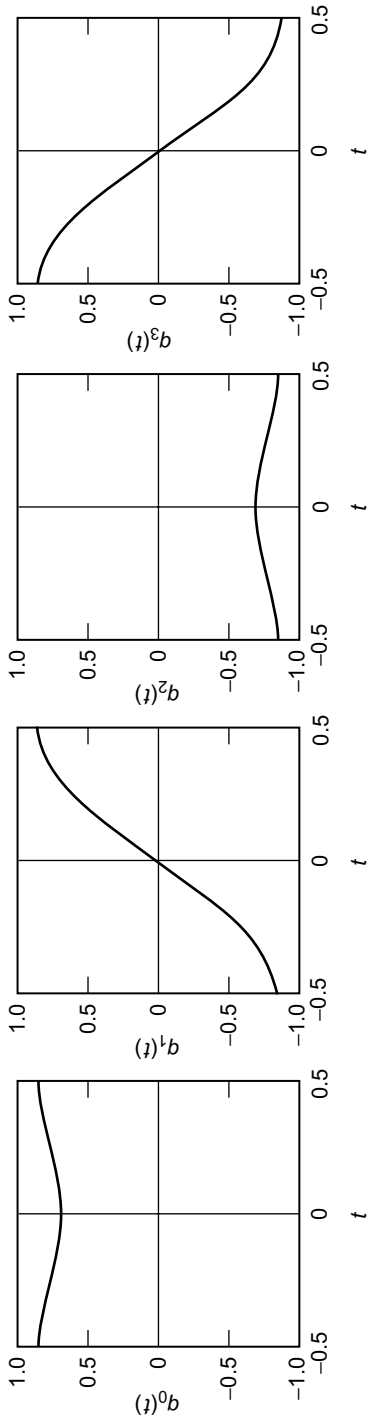


Fig. 3-18. Averaged waveforms for a simplified Viterbi receiver. Redrawn from [4].

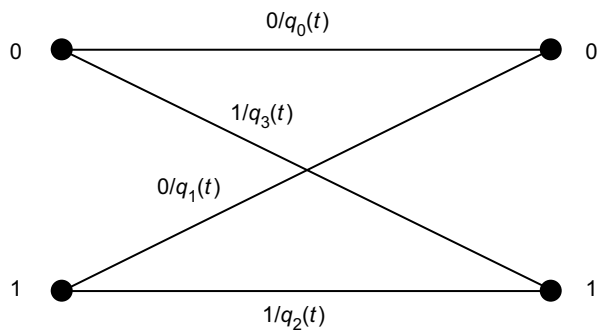


Fig. 3-19. Trellis diagram for a simplified FQPSK-B Viterbi receiver. Redrawn from [4].

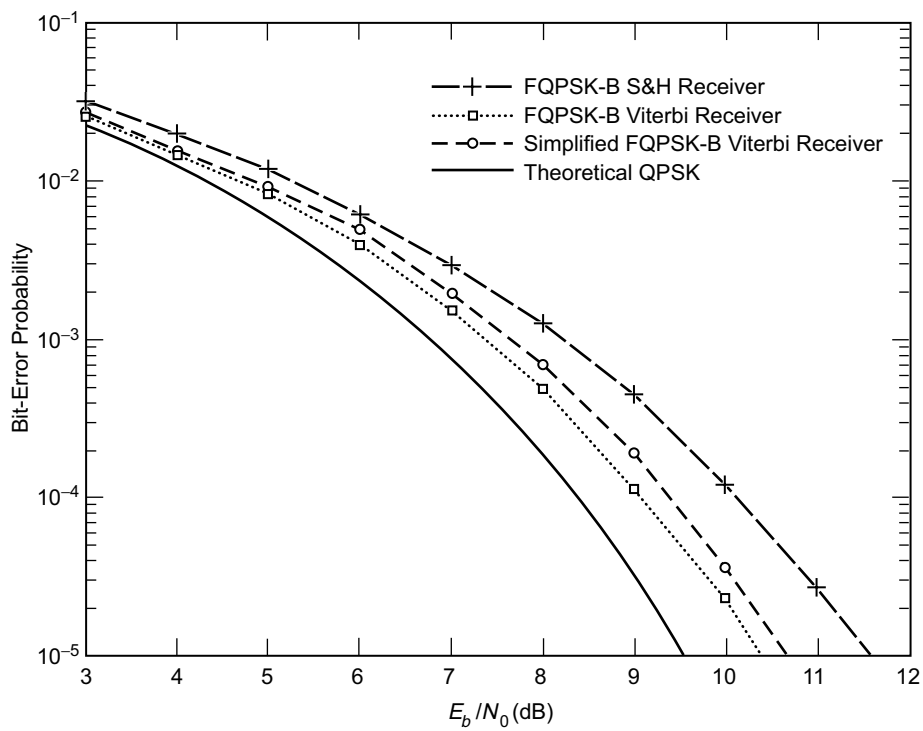


Fig. 3-20. Bit-error probability performance of FQPSK-B S&H and Viterbi receivers, with saturated SSPA. Redrawn from [4].

10 bits was shown to be necessary. Using the results in Fig. 3-20, Table 3-5 summarizes a comparison of the performances of the three FQPSK-B receivers at BEPs of 10^{-3} and 10^{-5}.

Table 3-5. Comparison of FQPSK-B performances.

Receiver	E_b/N_0 (dB) for 10^{-3} BEP	Loss Compared to Ideal QPSK at $P_b(E) = 10^{-3}$	E_b/N_0 (dB) for 10^{-5} BEP	Loss Compared to Ideal QPSK at $P_b(E) = 10^{-5}$
Full Viterbi receiver	7.4	0.6	10.4	0.8
Simplified receiver	7.65	0.85	10.7	1.1
S&H receiver	8.2	1.4	11.6	2.0

We observe from this table that the full Viterbi receiver performs 0.8 dB better than the symbol-by-symbol S&H receiver at a BEP of 10^{-3}, which is comparable to the analogous comparison made in Ref. 3 for unfiltered FQPSK. Thus, we can conclude that the Viterbi receiver works almost as well for the filtered version of FQPSK as it does for the unfiltered version. When compared with the full Viterbi receiver, the simplified FQPSK-B receiver suffers a slight degradation (0.25 dB) but is still better than the S&H receiver at a BEP of 10^{-3}. At a BEP of 10^{-5}, the full and simplified FQPSK-B Viterbi receivers are, respectively, 1.2 and 0.9 dB better than the S&H receiver.

Finally, to allow evaluation of BEP at values sufficiently small as to make simulation impractical, the BEP performance of the symbol-by-symbol S&H receiver for FQPSK-B was derived using superposition arguments in the appendix of Ref. 4 where the channel was modeled as being linear. Because of the ISI inherent in FQPSK-B, there are a large number of terms in the analytical expression for BEP. However, it was shown there that 32 terms is sufficient to give a very close match to simulation results (see Fig. 3-21). Comparing the results for the linear channel, given in Fig. 3-21, with the S&H simulation results for the nonlinear channel in Fig. 3-20 indicates a very small difference between them, the latter being the worse of the two.

3.7 Cross-Correlated Trellis-Coded Quadrature Modulation

Cross-correlated trellis-coded quadrature modulation (XTCQM) [12] is a technique that expands on the notion of combined bandwidth/power efficiency

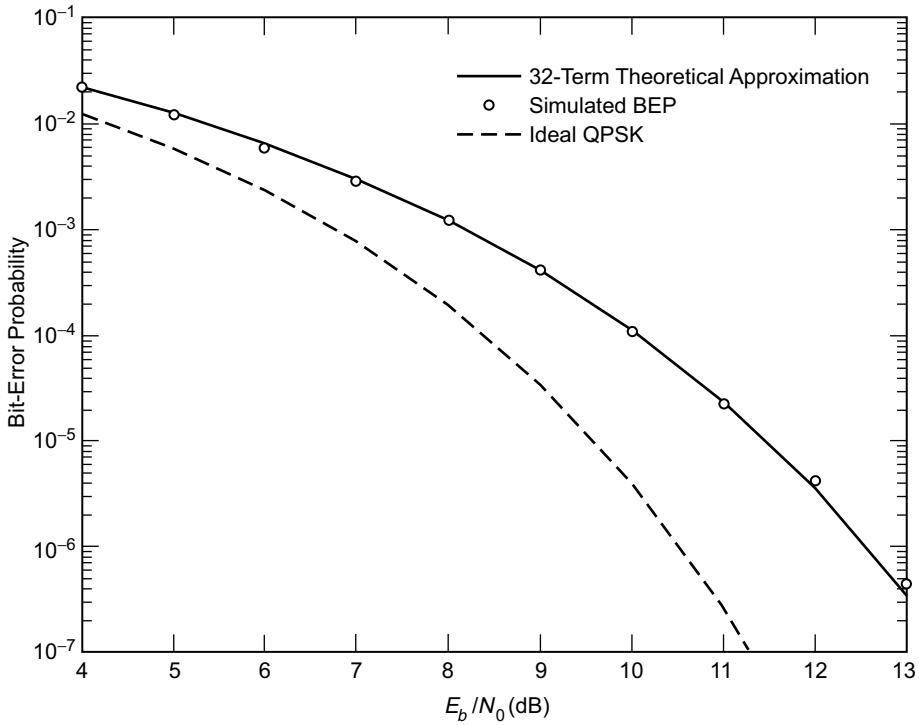


Fig. 3-21. Comparison of theoretical and simulated bit-error probability for an FQPSK-B S&H receiver.

indigenous to TCM, with particular emphasis on the spectral occupancy of the transmitted signal, while at the same time paying careful attention to the desirability of small envelope fluctuation. Whereas TCM combines conventional multilevel or multiphase modulations with error-correction coding through a suitable mapping that simultaneously exploits the desirable properties of these two functions, XTCQM focuses on a quadrature structure with coded I and Q channels that are cross-correlated and whose outputs are mapped into an M-ary waveform modulation. By virtue of its form, XTCQM produces a transmitted waveform with high spectral efficiency and allows for the design of a highly-power-efficient receiver. Its transmitter and receiver make use of standard, currently available devices, e.g., OQPSK modulator, convolutional encoders, matched filters, and Viterbi decoders, for its implementation. As we shall see, specific embodiments of XTCQM manifest themselves as FQPSK and trellis-coded versions of OQPSK and SQORC modulation. However, the generic structure provides considerably more flexibility for trading off between power and spectral efficiencies than these more restrictive embodiments.

3.7.1 Description of the Transmitter

With reference to Fig. 3-22, consider an input binary (± 1) i.i.d. data (information) sequence, $\{d_n\}$, at a bit rate, $R_b = 1/T_b$. This sequence is split into inphase (I) and quadrature (Q) sequences, $\{d_{In}\}$ and $\{d_{Qn}\}$, respectively, which consist of the even and odd bits of the information bit sequence, $\{d_n\}$, occurring at a rate, $R_s = 1/T_s = 1/2T_b$. We assume that the I and Q sequences, $\{d_{In}\}$ and $\{d_{Qn}\}$, are time synchronous and that the bit, d_{In} (or d_{Qn}), occurs during the interval $(n - (1/2))\,T_s \leq t \leq (n + (1/2))\,T_s$. As was the case for the TCM representation of FQPSK illustrated in Fig. 3-12, it is more convenient to work with the $(0, 1)$ equivalents of the I and Q data sequences, namely, $\{D_{In}\}$ and $\{D_{Qn}\}$, as in (3.4-1). The sequences, $\{D_{In}\}$ and $\{D_{Qn}\}$, are applied to I and Q rate $r = 1/N$ convolutional encoders (the two encoders are in general different, i.e., they have different tap connections and different modulo 2 summers but are assumed to have the same code rate). Let $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ and $\left\{E_{Qk}\,\middle|_{k=1}^{N}\right\}$ respectively denote the sets of N $(0, 1)$ output symbols of the I and Q convolutional encoders corresponding to a single-bit input to each. These sets of output symbols will be used to determine a pair of baseband waveforms, $s_I(t), s_Q(t)$, which ultimately modulate I and Q carriers for transmission over the channel. In order to generate an offset form of modulation, the signal, $s_Q(t)$, will be delayed by $T_s/2 = T_b$ s prior to modulation on the quadrature carrier.[8] The mapping of the symbol sets $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ and $\left\{E_{Qk}\,\middle|_{k=1}^{N}\right\}$ into $s_I(t)$ and $s_Q(t)$ and the size and content (waveshapes) of the waveform sets from which the latter are selected are the two most significant constituents of the XTCQM modulation scheme.

3.7.1.1 The Mapping. The mapping of the sets $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ and $\left\{E_{Qk}\,\middle|_{k=1}^{N}\right\}$ into $s_I(t)$ and $s_Q(t)$ is illustrated in Fig. 3-23. Consider that each of these sets of N $(0, 1)$ output symbols is partitioned into three groups as follows: For the first group, let $I_{l_1}, I_{l_2}, \cdots, I_{l_{N_1}}$ be a subset containing N_1 elements of $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ that will be used only in the selection of $s_I(t)$. For the second group, let $Q_{l_1}, Q_{l_2}, \cdots, Q_{l_{N_2}}$ be a subset containing N_2 elements of $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ that will be used only in the selection of $s_Q(t)$. Finally, for the third group, let $I_{l_{N_1+1}}, I_{l_{N_1+2}}, \cdots, I_{l_{N_1+N_3}} = Q_{l_{N_2+1}}, Q_{l_{N_2+2}}, \cdots, Q_{l_{N_2+N_3}}$ be a subset containing N_3 elements of $\left\{E_{Ik}\,\middle|_{k=1}^{N}\right\}$ that will be used both for the selection of $s_I(t)$ and $s_Q(t)$; hence, the term "cross-correlation" in the name of the modulation

[8] Note that delaying the waveform one-half of a symbol at the output of the mapping allows synchronous demodulation and computation of the path metric at the receiver. This is also true for the TCM implementation of FQPSK and, as such, is different than the conventional FQPSK approach, which applies the half-symbol delay to the Q data stream prior to any further processing (see Fig. 3-1).

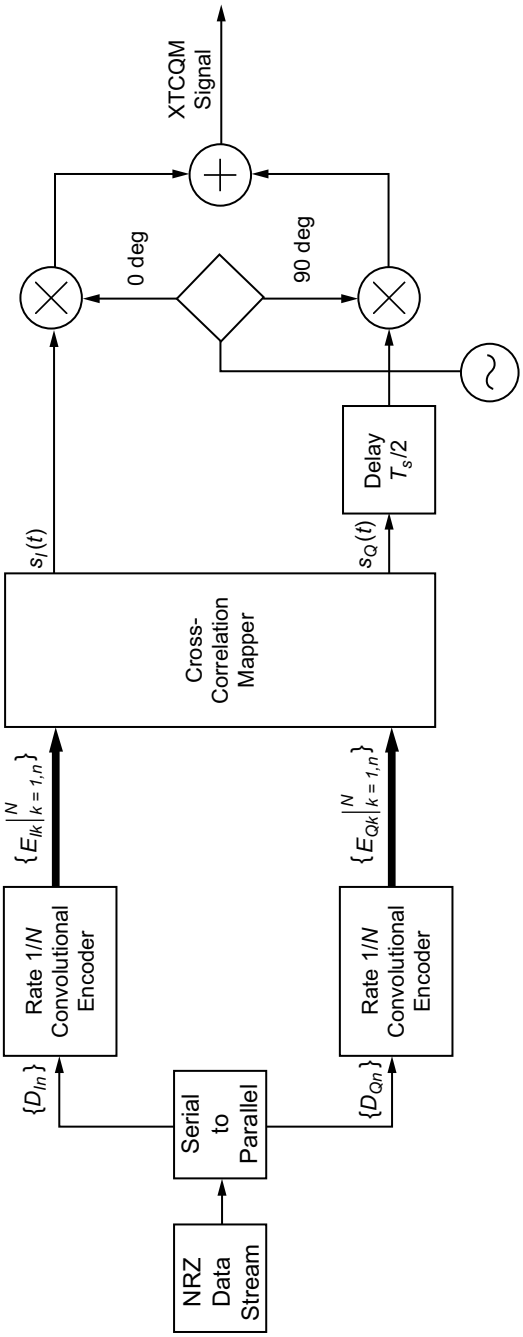


Fig. 3-22. Conceptual block diagram of an XTCQM transmitter.

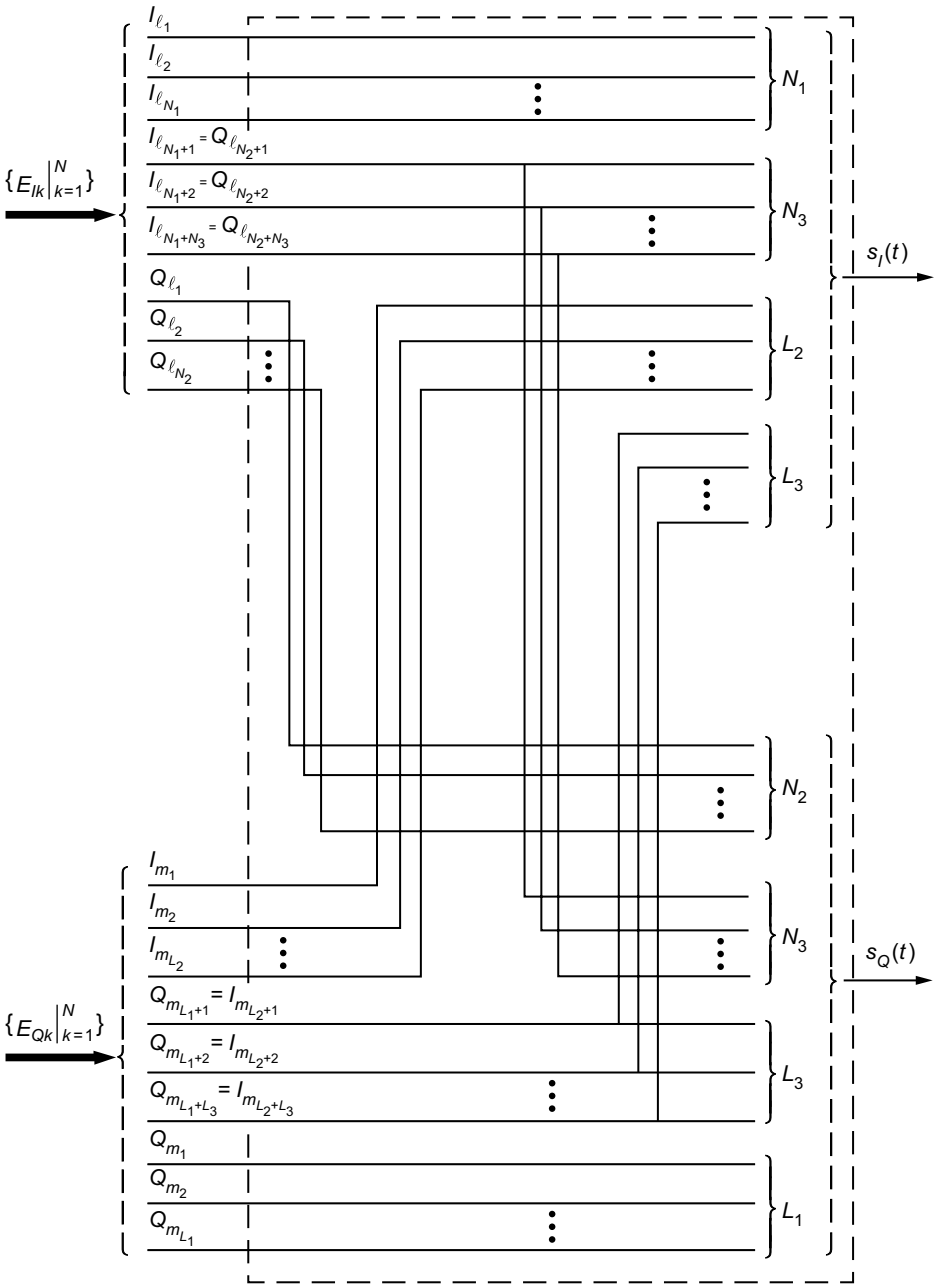


Fig. 3-23. Cross-correlation mapper.

scheme. Since all of the output symbols of the I encoder are used either to select $s_I(t)$ or $s_Q(t)$, or both, then, clearly, we must have $N_1+N_2+N_3=N$. A similar three-part grouping of the Q encoder ouput symbols $\left\{E_{Qk}\left|_{k=1}^{N}\right.\right\}$ is assumed to occur. That is, for the first group, let $Q_{m_1}, Q_{m_2}, \cdots, Q_{m_{L_1}}$ be a subset containing L_1 elements of $\left\{E_{Qk}\left|_{k=1}^{N}\right.\right\}$ that will be used only in the selection of $s_Q(t)$. For the second group, let $I_{m_1}, I_{m_2}, \cdots, I_{m_{L_2}}$ be a subset containing L_2 elements of $\left\{E_{Qk}\left|_{k=1}^{N}\right.\right\}$ that will be used only in the selection of $s_I(t)$. Finally, for the third group let $Q_{m_{L_1+1}}, Q_{m_{L_1+2}}, \cdots, Q_{m_{L_1+L_3}} = I_{m_{L_2+1}}, I_{m_{L_2+2}}, \cdots, I_{m_{L_2+L_3}}$ be a subset containing L_3 elements of $\left\{E_{Qk}\left|_{k=1}^{N}\right.\right\}$ that will be used both for the selection of $s_I(t)$ and $s_Q(t)$. Once again, since all of the output symbols of the Q encoder are used either to select $s_I(t)$ or $s_Q(t)$, or both, then, clearly, we must have $L_1+L_2+L_3=N$. More often than not, because of symmetry properties associated with the resulting modulation, we shall want to choose $L_1=N_1, L_2=N_2$ and $L_3=N_3$; however, the proposed XTCQM scheme is not restricted to this particular selection.

In summary, based on the above, the signal $s_I(t)$ will be determined from symbols $I_{l_1}, I_{l_2}, \cdots, I_{l_{N_1+N_3}}$ from the output of the I encoder and symbols $I_{l_1}, I_{l_2}, \cdots, I_{l_{L_2+L_3}}$ from the output of the Q encoder. Therefore, the size of the signaling alphabet used to select $s_I(t)$ will be $2^{N_1+N_3+L_2+L_3} \triangleq 2^{N_I}$. Similarly, the signal $s_Q(t)$ will be determined from symbols $Q_{l_1}, Q_{l_2}, \cdots, Q_{l_{L_1+L_3}}$ from the output of the Q encoder and symbols $Q_{l_1}, Q_{l_2}, \cdots, Q_{l_{N_2+N_3}}$ from the output of the I encoder. Thus, the size of the signaling alphabet used to select $s_Q(t)$ will be $2^{L_1+L_3+N_2+N_3} \triangleq 2^{N_Q}$. If it is desired that the size of the signaling alphabets for selecting $s_I(t)$ and $s_Q(t)$ be equal (a case of common interest), we need to have $N_I=N_Q$ or, equivalently, $L_1+N_2=N_1+L_2$. This condition is clearly satisfied if the condition $L_1=N_1, L_2=N_2$ is met; however, the former condition is less restrictive and does not require the latter to be true.

Having now assigned the encoder output symbols to either $s_I(t)$ or $s_Q(t)$ or both, the final step in the signal mapping is to form appropriate BCD numbers from these symbols and use these as indices i and j for choosing $s_I(t)=s_i(t)$ and $s_Q(t)=s_j(t)$, where $\left\{s_i(t)\left|_{i=1}^{N_I}\right.\right\}$ and $\left\{s_j(t)\left|_{j=1}^{N_Q}\right.\right\}$ are the signal waveform sets assigned for transmission of the I and Q channel signals. Specifically, let $I_0, I_1, \cdots, I_{N_I}$ be the particular set of symbols (taken from both I and Q encoder outputs) used to select $s_I(t)$, and let $Q_0, Q_1, \cdots, Q_{N_Q}$ be the particular set of symbols (taken from both I and Q encoder outputs) used to select $s_Q(t)$. Then, the BCD indices needed above are $i = I_{N_I-1} \times 2^{N_I-1} + \cdots + I_1 \times 2^1 + \cdots + I_0 \times 2^0$ and $j = Q_{N_Q-1} \times 2^{N_Q-1} + \cdots + Q_1 \times 2^1 + \cdots + Q_0 \times 2^0$.

3.7.1.2 The Signal Sets (Waveforms). While, in principle, any set of N_I waveforms of duration T_s s defined on the interval $-T_s/2 \leq t \leq T_s/2$, can be used for selecting the I channel transmitted signal, $s_I(t)$, and,

likewise, any set of N_Q waveforms of duration T_s s also defined on the interval $-T_s/2 \leq t \leq T_s/2$, can be used for selecting the Q channel transmitted signal, $s_Q(t)$, there are certain properties that should be invoked on these waveforms to make them desirable both from a power and spectral efficient standpoint. For the purpose of this discussion, we shall assume the special case $N_I = N_Q \triangleq N^*$ although, as pointed out previously, this is not a limitation on the invention. First, to achieve maximum distance in the waveform set (i.e., good power efficiency) one should divide the signal set, $\left\{ s_i(t) \big|_{i=1}^{N^*} \right\}$, into two equal parts-the signals in the second part being antipodal to (the negatives of) those in the first part. Mathematically, the signal set would have the composition $s_0(t), s_1(t), \cdots, s_{N^*/2-1}(t), -s_0(t), -s_1(t), \cdots, -s_{N^*/2-1}(t)$. Second, to achieve good spectral efficiency, one should choose the waveforms to be as smooth (i.e., as many continuous derivatives) as possible. Furthermore, to prevent discontinuities at the symbol transition-time instants, the waveforms should have a zero first derivative (slope) at their endpoints, $t = \pm T_s/2$.

In the next section, we explore the FQPSK embodiment as well as several other embodiments corresponding to well-known modulation schemes previously discussed in this monograph.

3.7.2 Specific Embodiments

3.7.2.1 FQPSK. Consider as an example the mapping scheme of Sec. 3.7.1.1, corresponding to $N_1 = N_2 = N_3 = 1$, i.e., rate $r = 1/N = 1/3$ encoders, and $L_1 = L_2 = L_3 = 1$, e.g., of the three output symbols from the I encoder, one is used to choose the I-channel signal, one is used to choose the Q-channel signal, and one is used to choose both the I- and Q-channel signals. Suppose now that the specific symbol assignments for the three partitions of the I encoder output are: I_3 (group 1), Q_0 (group 2), $I_2 = Q_1$ (group 3), and, similarly, the specific symbol assignments for the three partitions of the Q encoder output are: Q_3 (group 1), I_1 (group 2), $I_0 = Q_2$ (group 3). Furthermore, since $N_I = N_Q = 4$, the size of the signaling alphabet from which both $s_I(t)$ and $s_Q(t)$ are to be selected will be composed of $2^4 = 16$ signals. If, then, the rate 1/3 encoders are specifically chosen as in Fig. 3-12, and the 16 waveforms are selected as in Fig. 3-9, it immediately follows that FQPSK becomes a particular embodiment of the XTCQM scheme.

3.7.2.2 Trellis-Coded OQPSK. Consider an XTCQM scheme in which the mapping function is performed identical to that in the FQPSK embodiment (i.e., as in Fig. 3-12) but the waveform assignment is made as follows (see Fig. 3-24):

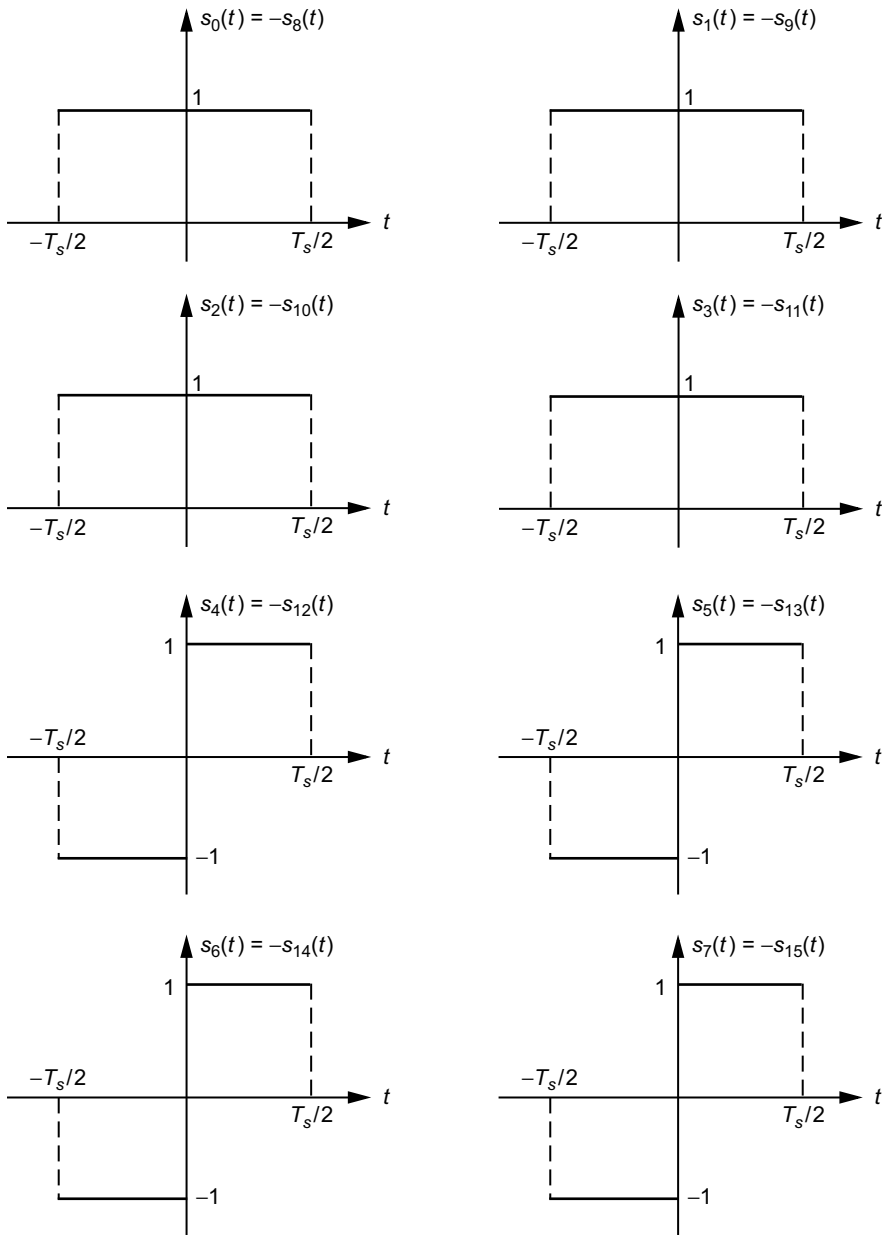


Fig. 3-24. Trellis-coded OQPSK full-symbol waveforms.

$$\left(\begin{array}{l} s_0(t) = s_1(t) = s_2(t) = s_3(t) = 1, \quad -\dfrac{T_s}{2} \le t \le \dfrac{T_s}{2} \\ \\ s_4(t) = s_5(t) = s_6(t) = s_7(t) = \begin{cases} -1, & -\dfrac{T_s}{2} \le t \le 0 \\ 1, & 0 \le t \le \dfrac{T_s}{2} \end{cases} \\ \\ s_i(t) = -s_{i-8}(t), \quad i = 8, 9, \cdots, 15 \end{array}\right) \tag{3.7-1}$$

that is, the first four waveforms are identical (a rectangular unit pulse) as are the second four (a split rectangular unit pulse), and the remaining eight waveforms are the negatives of the first eight. As such, there are only four unique waveforms, which we denote by $c_i(t)\big|_{i=0}^{3}$, where $c_0(t) = s_0(t), c_1(t) = s_4(t), c_2(t) = s_8(t), c_3(t) = s_{12}(t)$. Since in the BCD representations for each group of four identical waveforms, the two least significant bits are irrelevant, i.e., the two most significant bits are sufficient to define the common waveform for each group, we can simplify the mapping scheme by eliminating the need for I_0, I_1 and Q_0, Q_1. With reference to Fig. 3-12, elimination of I_0, I_1 and Q_0, Q_1 accomplishes two purposes. First, each encoder (both of which are now identical) needs only a single shift-register stage, and, second, the correlation between the two encoders insofar as the mapping of either one's output symbols to both $s_I(t)$ and $s_Q(t)$ has been eliminated, which, therefore, results in what might be termed a degenerate form of XTCQM, dubbed trellis-coded OQPSK [13]. The resulting embodiment is illustrated in Fig. 3-25. Since, insofar as the mapping is concerned, the I and Q channels are now decoupled (as indicated by the dashed line in the signal mapping block of Fig. 3-25), it is sufficient to examine the trellis structure and its distance properties for only one of the two channels (I or Q). The trellis diagram for either the I or Q channel of this modulation scheme would simply have two states and is illustrated in Fig. 3-26. The dashed line indicates a transition caused by an input "0," and the solid line indicates a transition caused by an input "1." Also, the branches are labeled with the output signal waveform that results from the transition. An identical trellis diagram would exist for the Q channel.

What is interesting about this embodiment of XTCQM is that, as far as the transmitted signal is concerned, it has a PSD identical to that of uncoded OQPSK (which is the same as for uncoded QPSK). In particular, because of the constraints imposed by the signal mapping, the waveforms $c_1(t) = s_4(t)$ and $c_3(t) = s_{12}(t)$ can never occur twice in succession. Thus, for any input information sequence, the sequence of signals $s_I(t)$ and $s_Q(t)$ cannot transition at a rate

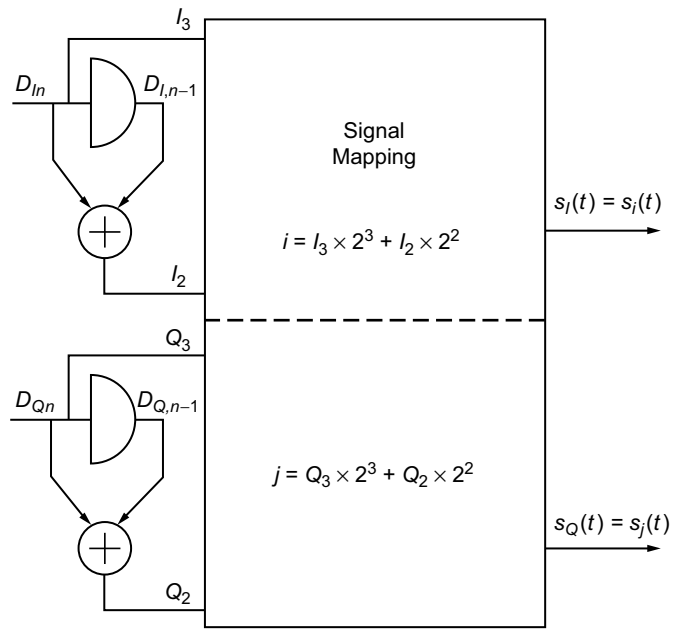


Fig. 3-25. The trellis-coded OQPSK embodiment of an XTCQM transmitter.

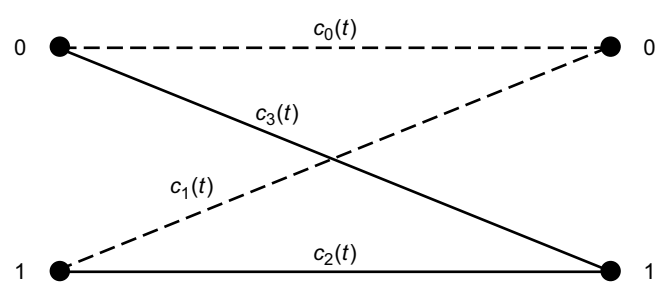


Fig. 3-26. Two-state trellis diagram for OQPSK.

faster than $1/T_s$. Stated another way, there cannot be any segment of sequences of $s_I(t)$ or $s_Q(t)$ that is constant for less than T_s s. This additional spectrum conservation constraint imposed by the signal mapping function of XTCQM will naturally result in a reduction of the coding (power) gain relative to that which could be achieved with another mapping, which does not prevent the successive repetition of $c_1(t)$ and $c_3(t)$. However, the latter occurrence would result in a bandwidth expansion by a factor of two.

3.7.2.3 Trellis-Coded SQORC. If, instead of a split rectangular pulse in (3.7-1), a sinusoidal pulse is used, namely,

$$\left.\begin{aligned} &s_4(t) = s_5(t) = s_6(t) = s_7(t) = \sin\frac{\pi t}{T_s}, \quad -\frac{T_s}{2} \le t \le \frac{T_s}{2} \\ &s_i(t) = -s_{i-8}(t), \quad i = 12, 13, 14, 15 \end{aligned}\right\} \tag{3.7-2}$$

the simplification of the mapping function shown in Fig. 3-25 again occurs (i.e., decoupling of the I and Q channels) and the trellis diagram of Fig. 3-26 is still appropriate for either the I or Q channel [13]. Once again, insofar as the transmitted signal is concerned, it has a PSD identical to that of uncoded SQORC (which is the same as for uncoded QORC).

3.7.2.4 Uncoded OQPSK. If we further simplify the signal assignment and mapping of Fig. 3-12 such that

$$\left.\begin{aligned} &s_0(t) = s_1(t) = \cdots = s_7(t) = 1, \quad -\frac{T_s}{2} \le t \le \frac{T_s}{2} \\ &s_i(t) = -s_{i-8}(t), \quad i = 8, 9, \cdots, 15 \end{aligned}\right\} \tag{3.7-3}$$

then in the BCD representations for each group of eight identical waveforms, the three least significant bits are irrelevant, i.e., only the first significant bit is needed to define the common waveform for each group. Hence, we can simplify the mapping scheme by eliminating the need for I_0, I_1, I_2 and Q_0, Q_1, Q_2. Defining the two unique waveforms, $c_0(t) = s_0(t), c_1(t) = s_8(t)$, we obtain the simplified degenerate mapping of Fig. 3-27, which corresponds to uncoded OQPSK with NRZ data formatting.

Likewise, if instead of the signal assignment in (3.7-3), we were to use

$$\left.\begin{aligned} &s_0(t) = s_1(t) = \cdots = s_7(t) = \begin{cases} -1, & -\frac{T_s}{2} \le t \le 0 \\ 1, & 0 \le t \le \frac{T_s}{2} \end{cases} \\ &s_i(t) = -s_{i-8}(t), \quad i = 8, 9, \cdots, 15 \end{aligned}\right\} \tag{3.7-4}$$

then the mapping of Fig. 3-27 produces uncoded OQPSK with Manchester (biphase) data formatting.

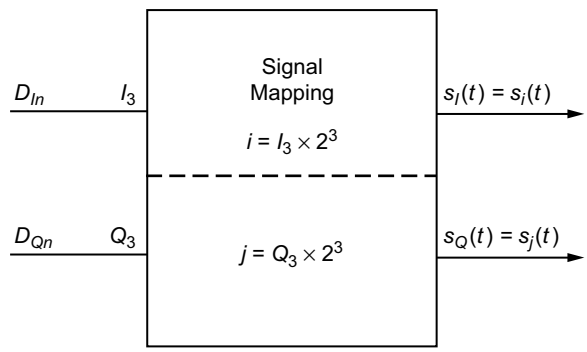


Fig. 3-27. Uncoded OQPSK embodiment of an XTCQM transmitter with NRZ data formatting.

3.8 Other Techniques

In the early 1980s, shaped BPSK (SBPSK) was introduced by Dapper and Hill [14] as a means of bandlimiting a BPSK signal while, at the same time, keeping its envelope constant. Further development of the SBPSK concept led to a variant of this scheme for offset quadrature modulation referred to as shaped offset QPSK (SOQPSK). In 2000, Hill [15] reported on a specific SOQPSK scheme with an enhanced waveform that offers spectral containment and detection efficiency comparable to or better than FQPSK-B, depending on the specifics of the comparison. Since SOPQSK is nonproprietary, whereas FQPSK-B is not, then, in view of the above-mentioned performance similarity, the former should be considered as a potential candidate in bandwidth-efficient modulation applications. In this section, we briefly review the results presented in Ref. 14 placing them in the context and notation of previous sections of this monograph. The material that follows should have been included in Chap. 2 since SOQPSK is truly constant envelope. However, we delayed discussing it there so that we might first present the material on FQPSK, thereby allowing the reader a better understanding of the performance comparison.

3.8.1 Shaped Offset QPSK

As a prelude to understanding the concept behind shaped offset QPSK (SOQPSK), it is instructive to first demonstrate that conventional OQPSK (rectangular pulse shaping implied) can be represented as a special case of CPM modulation. Specifically, OQPSK has the form in (2.8-1) together with (2.8-2), where $h = 1/2$; the frequency pulse, $g(t)$, of (2.8-3) is a delta function, i.e., $g(t) = (1/2)\delta(t)$ [equivalently, the phase pulse, $q(t)$, is a step function, i.e., $q(t) = (1/2)u(t)$]; and the ith element of the effective data se-

quence, α_i, can be shown to be related to the true input data bit sequence $\mathbf{a} = (\cdots, a_{-2}, a_{-1}, a_0, a_1, a_2, \cdots)$ by[9]

$$\alpha_i = (-1)^{i+1} \frac{a_{i-1}(a_i - a_{i-2})}{2} \tag{3.8-1}$$

Since the a_i's take on ± 1 values, then the α_i's come from a ternary (−1,0,+1) alphabet. However, in any given bit (half-symbol) interval, the α_i's can only assume one of two equiprobable values, namely, 0 and +1 or 0 and −1, with the further restriction that a +1 cannot be followed by a −1, or vice versa. Thus, in reality, the modulation scheme is a binary CPM but one whose data alphabet can vary (between two choices) from bit interval to bit interval. Another way of characterizing the variation rule for the data alphabet is as follows: If the previous bit is 0, then the data alphabet for the current bit is switched relative to that available for the previous bit, i.e., if it was (0,+1) for the previous transmission, it becomes (0,−1) for the current transmission, and vice versa. On the other hand, if the previous bit is a +1 or a −1, then the data alphabet for the current bit remains the same as that available for the previous bit, e.g., if it was (0,+1) for the previous transmission, it is again (0,+1) for the current transmission.

Since $h = 1/2$ together with the factor of 1/2 in $g(t)$ corresponds to a phase change of $\pi/2$ rad, then a value of $\alpha_i = 0$ suggests no change in carrier phase (no transition occurs in the I (or Q) data symbol sequence at the midsymbol time instant of the Q (or I) data symbol), whereas a value of $\alpha_i = \pm 1$ suggests a carrier phase change of $\pm\pi/2$ (a transition occurs in the I (or Q) data symbol sequence at the midsymbol time instant of the Q (or I) data symbol). Finally, note that since the duration of the frequency pulse does not exceed the baud (bit) interval, then, in accordance with the discussion in Sec. 2.7, the CPM representation of OQPSK is full response and can be implemented with the cascade of a precoder satisfying (3.8-1) and a conventional CPM modulator such as in Fig. 2-7.

In the early conception of SOQPSK, a rectangular pulse of duration equal to the bit period was used for $g(t)$. In this sense, one might think that SOQPSK resembled MSK; however, we remind the reader that for the latter, the data alphabet was fixed at $-1, +1$ whereas for the former it varies between $0, -1$ and $0, +1$. Thus, whereas in a given bit interval, the phase for MSK is always linearly varying with either a positive or negative slope, the phase for SOQPSK can either vary linearly or remain stationary. As such, the phase trellis for SOQPSK will have plateaus during the bit intervals where $\alpha_i = 0$. Since for OQPSK

[9] Note that the I and Q data symbols a_{In}, a_{Qn} of (2.2-4) are respectively obtained as the even and odd bits of the sequence $\mathbf{a}$. Also note that, whereas the I-Q representation of OQPSK contains I and Q data sequences at the symbol rate $1/T_s$, the effective data sequence for the CPM representation occurs at the half-symbol (bit) rate, $1/(T_s/2) = 1/T_b$.

itself, the phase trellis would only have plateaus (no linear variations), then in this sense, SOQPSK with a square frequency pulse can be viewed as a hybrid of OQPSK and MSK.

In Ref. 14, two variants of SOQPSK, referred to as SOQPSK-A and SOQPSK-B, were considered based upon a frequency pulse shape that is a minor modification of the impulse response corresponding to a spectral raised cosine filter. The modification corresponds to applying a raised cosine (in the time domain) window to the above impulse response, which alone would have doubly infinite extent. Specifically,

$$g(t) = g_1(t)\, g_2(t) \tag{3.8-2}$$

where

$$\left.\begin{aligned} g_1(t) &\triangleq \frac{A\cos \pi\alpha Bt/T_s}{1-4\left(\alpha Bt/T_s\right)^2}\,\frac{\sin \pi Bt/T_s}{\pi Bt/T_s} \\ g_2(t) &\triangleq \begin{cases} 1, & \left|\dfrac{t}{T_s}\right| \le \varepsilon_1 \\ \dfrac{1}{2}+\dfrac{1}{2}\cos\dfrac{\pi\left(|t/T_s|-\varepsilon_1\right)}{\varepsilon_2}, & \varepsilon_1 < \left|\dfrac{t}{T_s}\right| \le \varepsilon_1+\varepsilon_2 \\ 0, & \left|\dfrac{t}{T_s}\right| > \varepsilon_1+\varepsilon_2 \end{cases} \end{aligned}\right\} \tag{3.8-3}$$

In (3.8-3), $g_1(t)$ is the impulse response of the spectral raised cosine filter with amplitude, A, fractional rolloff factor, α, and additional time-scaling factor, B, and $g_2(t)$ is the above-mentioned windowing function that limits the duration of $g(t)$ to $2(\varepsilon_1+\varepsilon_2)T_s$. The values of the parameters in (3.8-3) that define SOQPSK-A and SOQPSK-B are tabulated in Table 3-6.

Table 3-6. Parameter values for SOQPSK-A and SOQPSK-B.

Parameter	SOQPSK-A	SOQPSK-B
α	1.0	0.5
B	1.35	1.45
ε_1	1.4	2.8
ε_2	0.6	1.2

From Table 3-6, we see that SOQPSK-A has a frequency pulse duration of $4T_s$ ($8T_b$) and SOQPSK-B has a frequency pulse duration of $8T_s$ ($16T_b$); thus, both of these schemes correspond to partial-response CPM. Hill [15] clearly indicates that the parameter values chosen to represent SOQPSK-A and SOQPSK-B "are not 'optimum' in any mathematical sense" but are simply representative examples of what can be achieved with the functional form in (3.8-2) together with (3.8-3).

Figure 3-28 illustrates the simulated PSDs of SOQPSK-A and SOQPSK-B along with that of the earlier full-response version of SOQPSK, which uses a rectangular T_b-s frequency pulse (Ref. 14 refers to the latter as military standard (MIL-STD) SOQPSK, since this version was in fact adopted as a military standard). We observe from this figure that the difference between SOQPSK-A and SOQPSK-B down to a level of -40 dB is virtually nil; however, below that level, SOQPSK-A offers a significant spectral improvement over SOQPSK-B. To compare the spectral behavior of SOQPSK with that of FQPSK-B, Hill [15] uses the results illustrated in Fig. 3-28 (which were verified by experimental measurement) along with measured hardware results obtained from an FQPSK-B modem built by RF Networks, Inc. and tested at the ARTM Project facility at Edwards

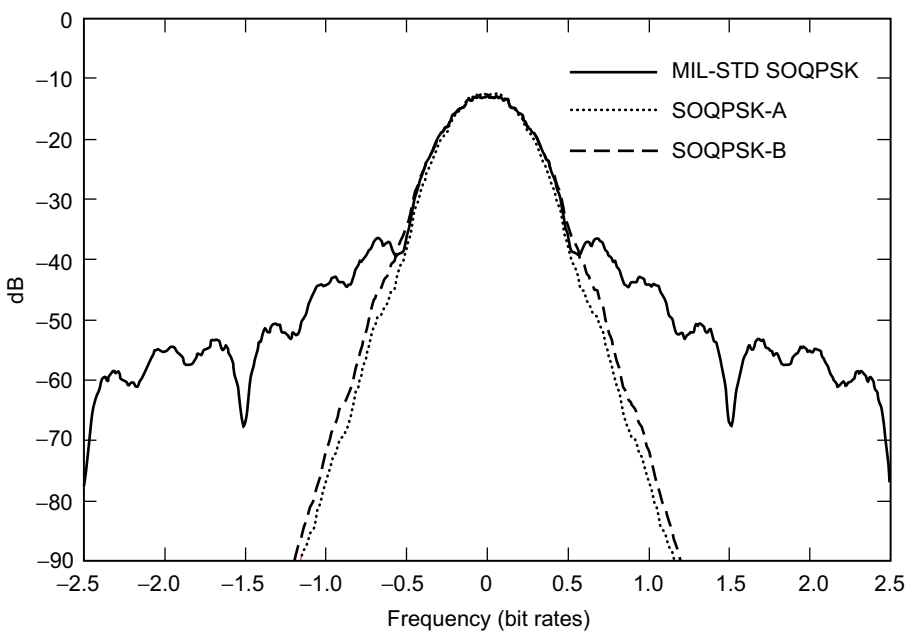


Fig. 3-28. Simulated SOQPSK power spectral densities. Resolution bandwidth = 20 kHz. Redrawn from [15].

Air Force Base. Two sets of comparisons were made. In one case, the PSDs of SOQPSK-A and SOQPSK-B were compared with that of FQPSK-B without nonlinear amplification (as such, FQPSK-B is therefore nonconstant envelope). Figure 3-29 illustrates this comparison, where it can be observed that: (a) down to −25 dB, the three PSDs are virtually indistinguishable from one another, and (b) below −25 dB, FQPSK-B is the most compact, SOQPSK-A is slightly wider, and SOQPSK-B is wider still. The second comparison pits SOQPSK-A against FQPSK-B with nonlinear amplification (to produce a constant envelope modulation). Figure 3-30 illustrates this comparison, where SOQPSK-A now has a narrower PSD than FQPSK-B. (Note that since SOQPSK-A is constant envelope, the nonlinear amplification theoretically has no effect on its PSD. This was also confirmed experimentally, as indicated in Ref. 14).

To complete this discussion, Figs. 3-31(a) and 3-31(b) illustrate the simulated and measured BEP of MIL-STD SOQPSK, SOQPSK-A, SOQPSK-B, and FQPSK-B. The simulated results were obtained using a conventional OQPSK receiver that for all three modulations is suboptimum, since no attempt is made to match the equivalent I and Q pulse shapes. We observe from Fig. 3-31(a) that

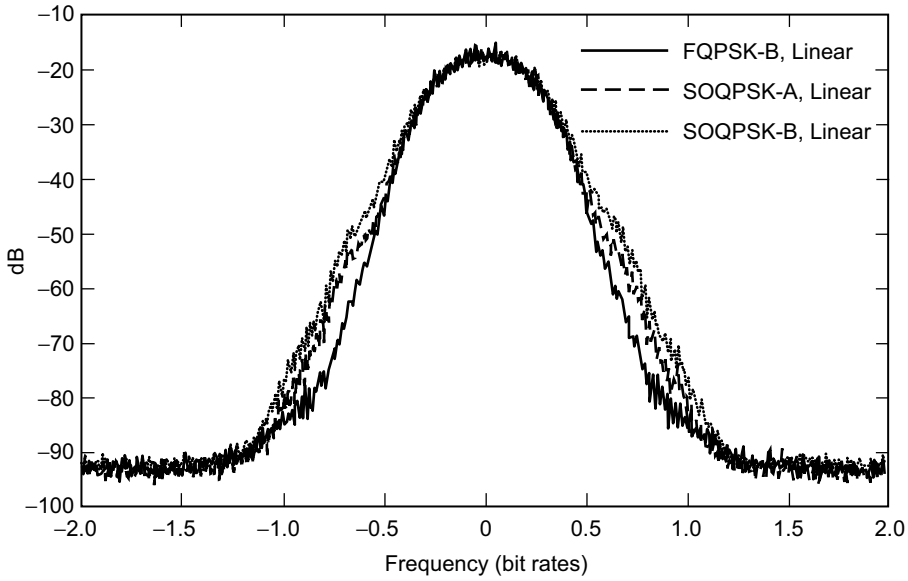


Fig. 3-29. A comparison of the power spectral densities of SOQPSK and FQPSK-B without nonlinear amplification. Resolution bandwidth = 3 kHz, video bandwidth = 10 kHz, and data rate = 1.0 Mb/s. Data from the Advanced Range Telemetry Lab at Edwards Air Force Base, California. Redrawn from [15].

at a BEP of 10^{-5}, MIL-STD SOQPSK and SOQPSK-B produce about the same E_b/N_0 performance penalty, i.e., 2.4 dB, relative to ideal unfiltered OQPSK, whereas SOQPSK-A is only about 0.25 dB worse. The conclusion to be drawn from this fact is that the MIL-STD variant of SOQPSK, which employs a rectangular frequency pulse, can be modified to generate SOQPSK-A or SOQPSK-B, with virtually no penalty in detection (power) efficiency but a considerable improvement in bandwidth efficiency. It seems clear that further optimizing the receiver by including appropriate matched filtering and trellis decoding (recall that SOQPSK-A or SOQPSK-B are memory modulations by virtue of the fact that they are partial-response CPMs) would yield additional improvement in power efficiency. The measured BEP performance curves in Fig. 3-31(b) reveal that at a BEP of 10^{-5}, SOQPSK-A is comparable but about 0.5 dB worse than nonlinearly amplified FQPSK-B, whereas SOQPSK-B is about 0.75 dB better.

As this book was going to press, the author became aware of results [16] describing a trellis detector for SOQPSK-A and SOQPSK-B that, by accounting for the pulse shaping and memory inherent in the modulations, provides superior detection performance as compared to the traditional OQPSK detector considered in Ref. 15.

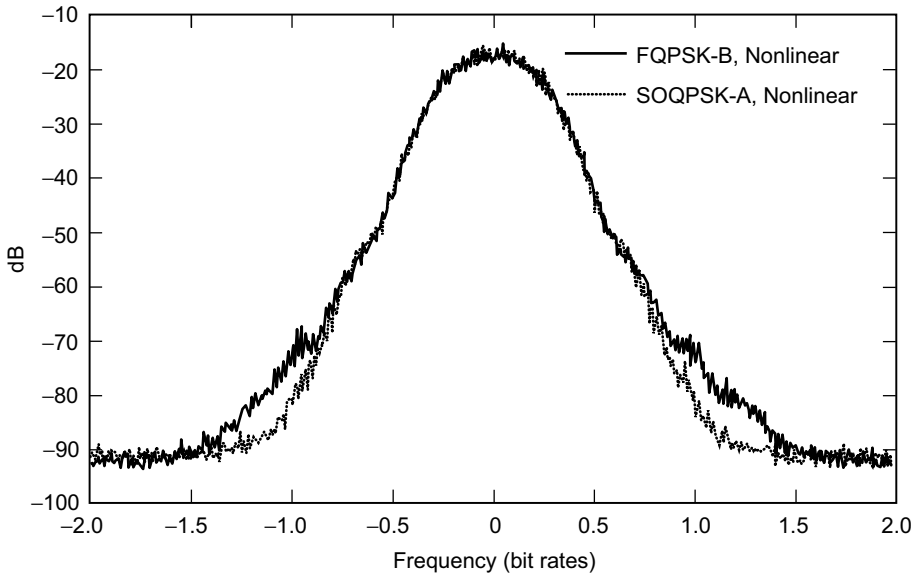


Fig. 3-30. A comparison of the power spectral densities of SOQPSK-A and FQPSK with nonlinear amplification. Resolution bandwidth = 3 kHz, video bandwidth = 10 kHz, and data rate = 1.0 Mb/s. Data from the Advanced Range Telemetry Lab at Edwards Air Force Base, California. Redrawn from [15].

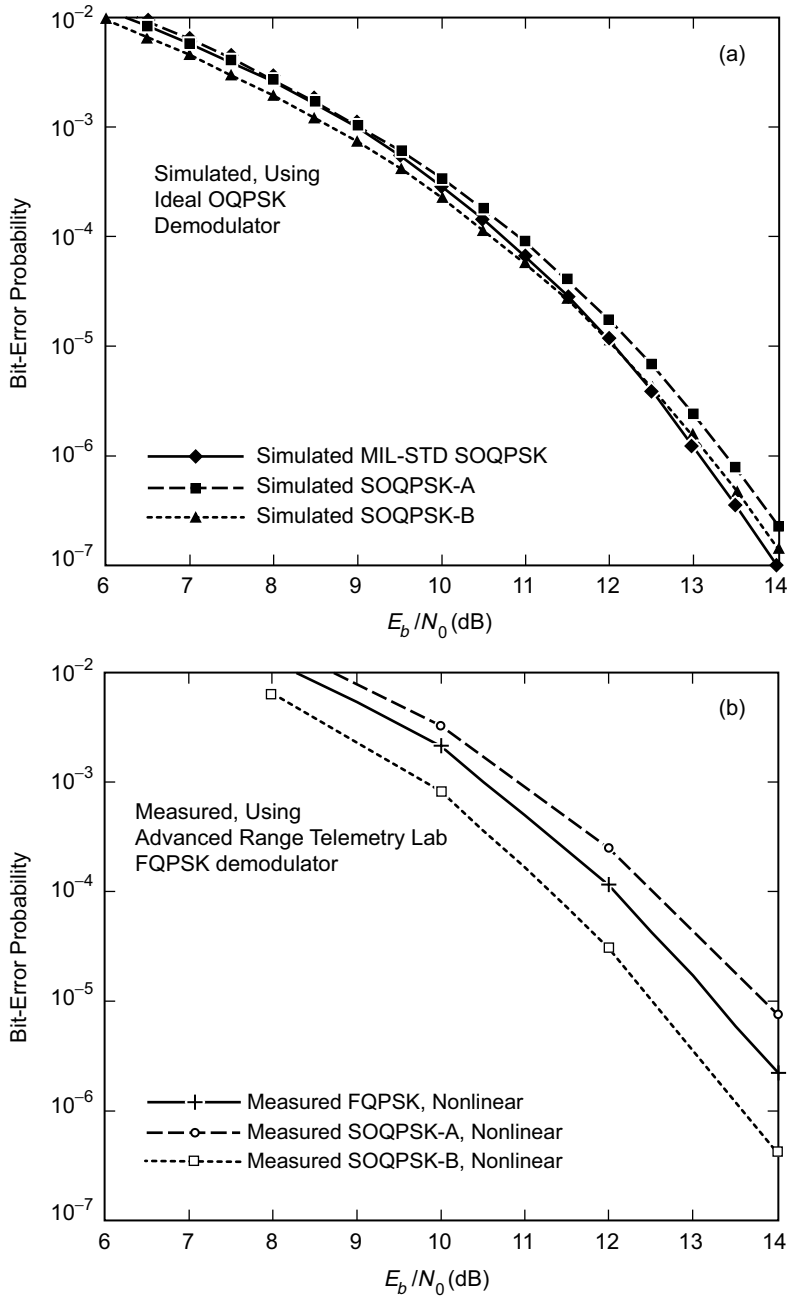


Fig. 3-31. Bit-error probability results for FQPSK-B, SOQPSK-A, and SOQPSK-B: (a) simulated and (b) measured. Redrawn from [15].

References

[1] K. Feher et al., U.S. patents: 4,567,602; 4,339,724; 4,644,565; 5,784,402; 5,491,457. Canadian patents: 1,211,517; 1,130,871; 1,265,851.

[2] S. Kato and K. Feher, "XPSK: A new cross-correlated phase-shift-keying modulation technique," *IEEE Transactions on Communications*, vol. 31, no. 5, pp. 701–707, May 1983.

[3] M. K. Simon and T.-Y. Yan, "Unfiltered Feher-patented quadrature phase-shift-keying (FQPSK): Another interpretation and further enhancements: Parts 1, 2," *Applied Microwave & Wireless Magazine*, pp. 76–96/pp. 100–105, February/March 2000.

[4] D. Lee, M. K. Simon, and T.-Y. Yan, "Enhanced Performance of FQPSK-B Receiver Based on Trellis-Coded Viterbi Demodulation," *International Telemetering Conference*, San Diego, California, October 23–26, 2000.

[5] K. Feher, *Wireless Digital Communications: Modulation and Spread Spectrum Applications*, Upper Saddle River, New Jersey: Prentice Hall, 1995.

[6] K. Feher, *Digital Communications: Satellite/Earth Station Engineering*, Littleton, Colorado: Crestone Engineering, 1996.

[7] K. Feher, "F-QPSK-A superior modulation technique for mobile and personal communications," *IEEE Transactions on Broadcasting*, vol. 39, no. 2, pp. 288–294, June 1993.

[8] K. Feher, "FQPSK transceivers double the spectral efficiency of wireless and telemetry systems," *Applied Microwave & Wireless Magazine*, June 1998. Also presented at *European Telemetry Conference*, Garmish-Patternk, Germany, May 5–8, 1998.

[9] W. L. Martin, T-Y. Yan, and L. V. Lam, "CCSDS-SFCG: Efficient modulation methods study at NASA/JPL, Phase 3: End-to end performance," *Proceedings of the SFGC Meeting*, Galveston, Texas, September 16–25, 1997.

[10] T. Le-Ngoc, K. Feher, and H. Pham Van, "New modulation techniques for low-cost power and bandwidth efficient satellite earth stations," *IEEE Transactions on Communications*, vol. 30, no. 1, pp. 275–283, January 1982.

[11] M. C. Austin and M. V. Chang, "Quadrature overlapped raised-cosine modulation," *IEEE Transactions on Communications*, vol. 29, no. 3, pp. 237–249, March 1981.

[12] M. K. Simon and T.-Y. Yan, "Cross-correlated trellis coded quadrature modulation," patent filed October 5, 1999.

[13] M. K. Simon, P. Arabshahi, and M. Srinivasan, "Trellis-coded quadrature phase shift keying (QPSK) with variable overlapped raised-cosine pulse shaping," *Telecommunications and Mission Operations Progress Report 42-136*, vol. October–December 1998, February 15, 1999.
http://tmo.jpl.nasa.gov/progress_report/issues.html
Accessed March 2, 2001.

[14] M. J. Dapper and T. J. Hill, "SBPSK: A robust bandwidth-efficient modulation for hard-limited channels," *MILCOM Conference Record*, Los Angeles, California, pp. 31.6.1–31.6.6, October 21–24, 1984.

[15] T. J. Hill, "An enhanced, constant envelope, interoperable shaped offset QPSK (SOQPSK) waveform for improved spectral efficiency," *International Telemetering Conference*, San Diego, California, October 23–26, 2000. Also see "A non-proprietary, constant envelope, variant of shaped offset QPSK (SOQPSK) for improved spectral containment and detection efficiency," *MILCOM Conference Record*, vol. 1, Los Angeles, California, pp. 347–352, October 23–26, 2000.

[16] M. Geoghegan, "Implementation and performance results for trellis detection of SOQPSK," to be presented at *International Telemetering Conference 2001*, Las Vegas, Nevada, October 22–25, 2001.

Chapter 4
Bandwidth-Efficient Modulations with More Envelope Fluctuation

Thus far in our discussions, we have focused on constant or quasi-constant envelope modulations many of which, by virtue of their inherent memory, required a trellis decoder (as implemented by the VA [1]) for optimum reception. In theory, the VA can start producing a ML estimate of the transmitted signal only after observing the channel output corresponding to the entire transmitted signal, resulting in an infinite decoding delay. By decoding delay, we mean the amount of time (typically measured in number of bits) after which one begins to decode. Algorithms such as the truncated VA [2] can be used to reduce the decoding delay, but, in general, these lead to suboptimum receiver structures.

In certain applications, achieving a finite and small decoding delay is desirable. The natural question to ask is whether the requirement for finite decoding delay imposes constraints on the modulation/demodulation scheme that would reduce its optimality from a power and bandwidth-efficiency standpoint. Furthermore, to what extent would these constraints compromise the constant envelope nature of the transmitted signal set?

The ultimate goal would be to understand the possible trade-offs among minimum Euclidean distance (or, more generally, distance profile), bandwidth (or, more generally, PSD) and decoding delay. Such a goal is beyond the scope of this monograph. Instead, we consider here a reduced goal that investigates the above trade-offs for a particular structure derived from a generalization of that which implements MSK. The seeds for this investigation were planted in a paper presented at the 1997 International Symposium on Information Theory [3], in which Li and Rimoldi proposed a particular transmitter structure [the combination of

an encoder of memory, ν, and a waveform mapper—see Fig. 4-1(a)][1] for TCMs that, under certain constraints placed on the *differences* of the transmitted waveforms, guaranteed optimum decoding (using a conventional trellis decoder) with a finite (ν-bit duration) delay. Specifically, the encoder was simply a tapped delay line whose ν taps together with the input bit were mapped into a set of $M = 2^{\nu+1}$ waveforms (signals) of one-bit duration in accordance with a BCD relationship. That is, if $U_n \in 0, 1$ denotes the nth input bit and $U_{n-1}, U_{n-2}, \cdots, U_{n-\nu}$ the previous ν bits (the state of the encoder), then the signal transmitted in the interval $nT_b \leq t \leq (n+1)\,T_b$ would be $s_i\,(t)$, where the index, i, is defined in terms of these bits by $i = U_n \times 2^{\nu} + U_{n-1} \times 2^{\nu-1} + \cdots + U_{n-\nu-1} \times 2^1 + U_{n-\nu} \times 2^0$. It was also shown in Ref. 3 that, in addition to the constraints placed on the waveform differences, it was possible to further constrain the signals so as to

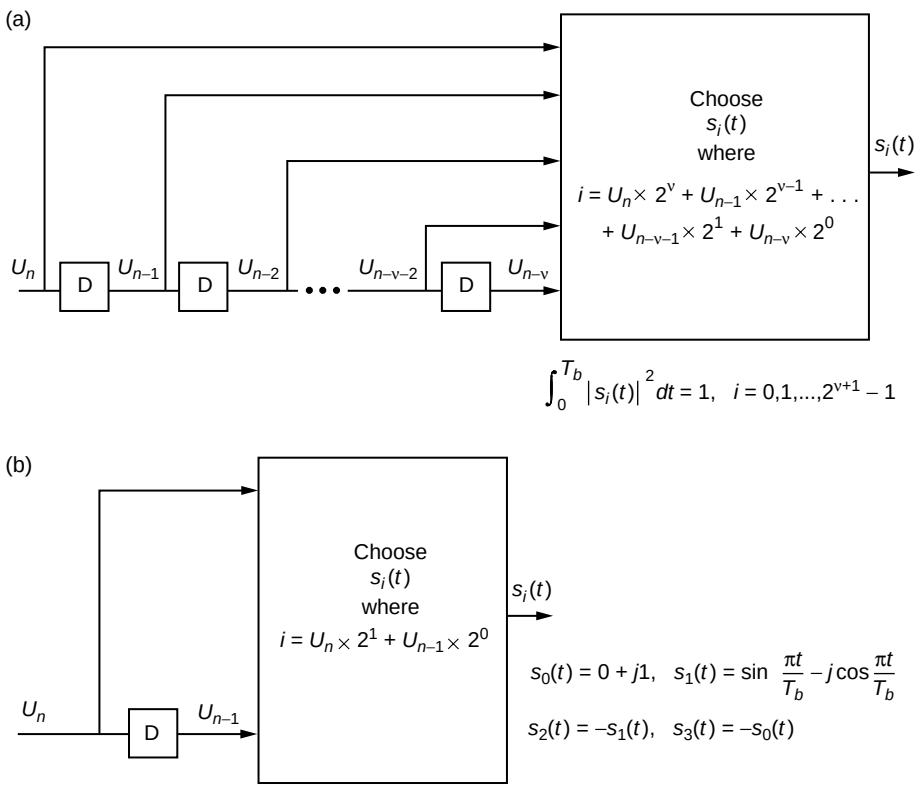


Fig. 4-1: (a) A trellis-coded modulation complex baseband transmitter and (b) the special case of "MSK" (ν = 1).

[1] Note the synergy of this structure with the TCM representation of FQPSK illustrated in Fig. 3-12.

maximize the value of the minimum squared Euclidean distance taken over all pairs of error event paths, namely, $d^2_{\min} = 2$. Such a maximum value of $d^2_{\min}$, which corresponds to a number of binary modulations such as BPSK and MSK, indicates that the receiver is providing optimum reception from a power conservation standpoint. Finally, in the presence of all of the above constraints, Li and Rimoldi [3] showed that it is possible to further optimize the system by selecting a set of waveforms that minimize the bandwidth-bit time product, BT_b.

In this chapter, we investigate an alternative (simpler) representation of the transmitter configuration suggested in Ref. 3 that consists of nothing more than a single filter (with complex impulse response) whose input is the ± 1 equivalent of the input data bits, namely, $\bar{U}_n = 1 - 2U_n$ for all n. This representation is arrived at by viewing the transmitted signal as a random pulse train with a pulse shape that extends beyond a single bit interval, i.e., one that contributes intersymbol interference (ISI) to its neighbors. As we shall see, such a pulse shape of duration $(\nu + 1)\, T_b$ can be constructed by designing its $\nu + 1$ partitions of duration T_b s in terms of the waveform differences that are outputted from Li and Rimoldi's transmitter. Such an ISI-based transmitter representation has the advantage that the PSD, and hence, the bandwidth are readily evaluated using known results for uncoded, random binary complex pulse trains. It also allows applying the insight provided in Forney's classic paper [4] on the VA, in particular, the discussion regarding the use of this algorithm to combat ISI.

One of the requirements placed on the set of possible transmitted waveforms $s_i\,(t)\,, i = 0, 1, \cdots, M$ in Ref. 3 is that they all have equal energy.[2] Following consideration of the alternative representation described above, we discuss the impact of relaxing the equal energy restriction on the power efficiency of the modulation scheme in its ability to achieve the largest value of $d^2_{\min}$. In particular, we propose an additional set of constraints (now on the differences of the *energies* of the signals) that must be satisfied to achieve the same finite decoding delay, using again the optimum sequence receiver, and then demonstrate that such a set of constraints results in a signal design with a maximum value of $d^2_{\min}$ less than two. Allowing the signals to have unequal energy, however, suggests the possibility of additional flexibility in the design of these signals in order to achieve the best bandwidth efficiency. Thus, the reduction in $d^2_{\min}$ caused by the unequal energy requirement can possibly trade off against an additional reduction in signal bandwidth. Additional consideration of this notion warrants investigation.

---

[2] Note that the assumption of equal energy does not imply constant envelope, as was the case for the CPMs studied in Ref. 5, which served as the motivation for the work leading up to the results in Ref. 3. Nevertheless, the envelope fluctuation of the resulting signal designs will be small when compared with Nyquist designs of comparable bandwidth efficiencies, to be discussed later on.

4.1 Bandwidth-Efficient TCM with Prescribed Decoding Delay—Equal Signal Energies

4.1.1 ISI-Based Transmitter Implementation

The decomposition of a memory modulation into a cascade of an encoder and a memoryless modulator was first applied to CPM by Rimoldi [5]. In particular, for MSK (see Sec. 2.8.1.5b of this monograph), the transmitter obtained is illustrated in Fig. 2-18. Comparing Fig. 2-18 with the special case of Fig. 4-1(a), corresponding to $\nu = 1$ and illustrated in Fig. 4-1(b), we note that in the former, the state is represented by the differentially encoded version of the current input bit $V_n = U_n \oplus V_{n-1}$ whereas, in the latter, it would be just the previous input bit, U_{n-1} itself. Furthermore, because of the differential encoding associated with the state in Fig. 2-18, a differential decoder would be required in the receiver following the trellis decoder, which would result in a small loss in BEP performance. We have previously shown in Sec. 2.8.1.3 that precoding true MSK with a differential decoder at the transmitter results in a modulation that is equivalent (spectral and power efficiently) to MSK but without the need for differential decoding at the receiver. It is such precoded MSK that is implemented by the simpler configuration of Fig. 4-1(b) and denoted by the quotation marks around MSK in the caption. In what follows, when referring to MSK in the context of Fig. 4-1(b) or its equivalents, we shall assume that precoded MSK is implied.

Consider an uncoded random binary (± 1) sequence, $\left\{\bar{U}_n\right\}$, that generates a random pulse train

$$s'(t) = \sum_{n=-\infty}^{\infty} \bar{U}_n p(t - nT_b) \tag{4.1-1}$$

where $p(t) \triangleq p_R(t) + jp_I(t)$ is a complex pulse shape defined on the interval $0 \leq t \leq (\nu+1)T_b$. Consider partitioning $p(t)$ into $\nu+1$ adjoint pieces corresponding to its one-bit interval sections. That is, we define the set of T_b-s duration waveforms

$$p_k(t) \triangleq p_{Rk}(t) + jp_{Ik}(t) = \begin{cases} p(t+kT), & 0 \leq t \leq T_b \\ 0, & \text{otherwise} \end{cases}, \quad k = 0, 1, 2, \cdots, \nu \tag{4.1-2}$$

From (4.1-1), in any T_b-s interval, e.g., the nth, the signal $s'(t)$ will be described by one of $M = 2^{\nu+1}$ complex waveforms, i.e., $s'_k(t - nT_b), k = 0, 1, 2, \cdots, 2^{\nu+1} - 1$, which are expressed in terms of $p(t)$ and the data sequence, $\left\{\bar{U}_n\right\}$, by

$$s'_k\left(t-nT_b\right)=\bar{U}_n p_0\left(t-nT_b\right)+\bar{U}_{n-1}p_1\left(t-nT_b\right)+\cdots+\bar{U}_{n-\nu}p_\nu\left(t-nT_b\right),$$

$$k=0,1,2,\cdots,2^{\nu+1}-1 \tag{4.1-3}$$

where the index, k, is the equivalent (0,1) bit sequence $\{U_n, U_{n-1}, \cdots, U_{n-\nu}\}$ expressed in BCD form. As an example, the set of waveforms for memory $\nu = 2$ is given below:

$$\left.\begin{aligned}
s'_0\left(t-nT_b\right)&=p_0\left(t-nT_b\right)+p_1\left(t-nT_b\right)+p_2\left(t-nT_b\right)\\
s'_1\left(t-nT_b\right)&=p_0\left(t-nT_b\right)+p_1\left(t-nT_b\right)-p_2\left(t-nT_b\right)\\
s'_2\left(t-nT_b\right)&=p_0\left(t-nT_b\right)-p_1\left(t-nT_b\right)+p_2\left(t-nT_b\right)\\
s'_3\left(t-nT_b\right)&=p_0\left(t-nT_b\right)-p_1\left(t-nT_b\right)-p_2\left(t-nT_b\right)\\
s'_4\left(t-nT_b\right)&=-p_0\left(t-nT_b\right)+p_1\left(t-nT_b\right)+p_2\left(t-nT_b\right)\\
s'_5\left(t-nT_b\right)&=-p_0\left(t-nT_b\right)+p_1\left(t-nT_b\right)-p_2\left(t-nT_b\right)\\
s'_6\left(t-nT_b\right)&=-p_0\left(t-nT_b\right)-p_1\left(t-nT_b\right)+p_2\left(t-nT_b\right)\\
s'_7\left(t-nT_b\right)&=-p_0\left(t-nT_b\right)-p_1\left(t-nT_b\right)-p_2\left(t-nT_b\right)
\end{aligned}\right\} \tag{4.1-4}$$

We note from (4.1-4) that, because of the BCD construction, the following properties hold for the signal differences:

$$s'_0(t)-s'_1(t)=s'_2(t)-s'_3(t)=s'_4(t)-s'_5(t)=s'_6(t)-s'_7(t)=2p_2(t) \tag{4.1-5a}$$

$$s'_0(t)-s'_2(t)=s'_4(t)-s'_6(t)=2p_1(t) \tag{4.1-5b}$$

Also, an equivalent (at least insofar as the first equality is concerned) condition to (4.1-5b) is

$$s'_0(t)-s'_4(t)=s'_2(t)-s'_6(t)=2p_0(t) \tag{4.1-5c}$$

In the more generic case for arbitrary ν, the conditions corresponding to (4.1-5a) and (4.1-5b) would be summarized as:

$$s_0'(t) - s_{2^m}'(t) = s_{2^{m+1}l}'(t) - s_{2^{m+1}l+2^m}'(t) = 2p_{\nu-m}(t),$$

$$m = 0, 1, 2, \cdots, \nu - 1, \quad l = 1, 2, \cdots, 2^{\nu-m} - 1 \tag{4.1-6}$$

and, furthermore, the generalization of (4.1-5c) becomes

$$s_0'(t) - s_{2^\nu}'(t) = s_{2^{\nu-1}}'(t) - s_{2^\nu+2^{\nu-1}}'(t) = 2p_0(t) \tag{4.1-7}$$

Associating the $2^{\nu+1}$ signals $\{s_k'(t)\}$ expressed as in (4.1-3) with the assumed equal energy, $\{s_k(t)\}$, derived from the implementation in Fig. 4-1(a), we see that the conditions on the signal differences of $s_i'(t)$ given in (4.1-6) are precisely those of Theorem I in Ref. 3, which guarantees a finite decoding delay of ν bits using an optimum trellis-coded receiver.[3] Therefore, since $p(t)$ is entirely specified by its adjoint T_b-s sections, T_b, it would appear that the transmitter of Fig. 4-1(a) can be equivalently implemented [see Fig. 4-2(a)] by passing the input ± 1 data sequence, $\{\bar{U}_n\}$ (modeled as a random impulse train), through a filter with complex impulse response

$$\left.\begin{aligned} p(t) &= \sum_{i=0}^{\nu} p_i(t - iT_b) \\ p_i(t) &= \frac{1}{2}\left[s_0'(t) - s_{2^{\nu-i}}'(t)\right] \end{aligned}\right\} \tag{4.1-8}$$

or equivalently [see Fig. 4-2(b)], the real and imaginary parts of the baseband signal (to be modulated onto quadrature carriers for transmission over the

[3] Li and Rimoldi also note that these conditions guarantee that the Euclidean distance between any pair of paths in the trellis decoder diverging at time n and remerging at time $n + \nu + 1$ is the same. Furthermore, the number of correlators (matched filters) needed to implement the optimum (MLSE) receiver will now vary *linearly* with memory, i.e., $\nu + 1$, as opposed to *exponentially* with memory, i.e., $2^{\nu+1}$, which is the case when no constraints are imposed on the decoding delay.

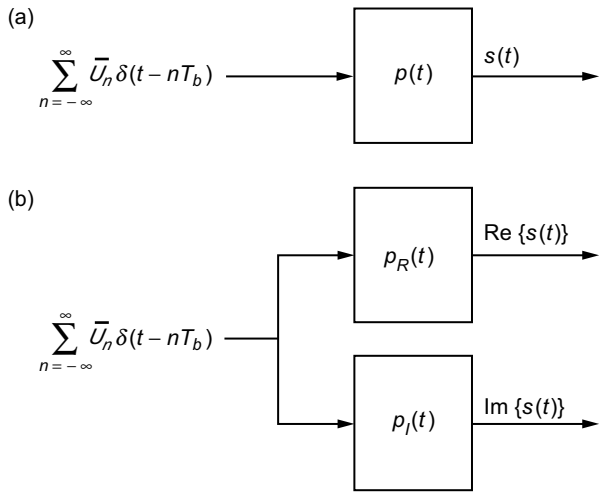


Fig. 4-2: (a) Complex baseband transmitter for MSK equivalent to Fig. 4-1(b) and (b) an I-Q baseband transmitter for MSK equivalent to Fig. 4-1(b).

channel) can be obtained by passing the common input ± 1 data sequence, $\{\bar{U}_n\}$, through a pair of filters with respective impulse responses

$$\left.\begin{aligned} p_{Ri}(t) &= \frac{1}{2}\left[s'_{R0}(t) - s'_{R2^{\nu-i}}(t)\right] \\ p_{Ii}(t) &= \frac{1}{2}\left[s'_{I0}(t) - s'_{I2^{\nu-i}}(t)\right] \end{aligned}\right\} \tag{4.1-9}$$

Unfortunately, the implementation in Fig. 4-2(a) is not always equivalent to that in Fig. 4-1(a), but as we shall see momentarily, for the case of most practical interest, i.e., a signal set $\{s_k(t)\}$ with maximum minimum Euclidean distance between its members, the equivalence between the two implementations is guaranteed, i.e., $\{s'_k(t)\}$ and $\{s_k(t)\}$ are identical. Before showing this, we note that even though $p_R(t)$ and $p_I(t)$ are constructed from the real and imaginary components of a set of equal energy complex signals, $\{s'_k(t), k = 0, 1, 2, \cdots, 2^{\nu+1} - 1\}$, they themselves do not necessarily have equal energy. We shall see that this is true, even for the simple case of MSK.

Note that because of the symmetry of the BCD mapping, the signals in the memory two example of (4.1-4) also satisfy the conditions

$$\left.\begin{aligned} s_0'(t) &= -s_7'(t) \\ s_1'(t) &= -s_6'(t) \\ s_2'(t) &= -s_5'(t) \\ s_3'(t) &= -s_4'(t) \end{aligned}\right\} \tag{4.1-10}$$

which, in the case of arbitrary memory, ν, would become

$$s_m'(t) = -s_{2^{\nu+1}-1-m}'(t), \quad m = 0, 1, \cdots, 2^\nu - 1 \tag{4.1-11}$$

The conditions of (4.1-11), which correspond to an antipodal signaling set, are precisely those given in Ref. 3. They achieve the maximum value of minimum-squared Euclidean distance, namely, $d_{\min}^2 = 2$. Thus, the implementation of Fig. 4-2(a) not only achieves finite decoding delay but also automatically achieves the optimum performance from the standpoint of power efficiency. This result should not be surprising in view of the findings in Ref. 4, which indicate that an MLSE-form of receiver such as the trellis decoder can completely remove ISI and thereby achieve the performance of a zero-ISI (full-response) system. However, since the implementation in Fig. 4-1(a) can produce a set of signals, $\{s_k(t)\}$, that satisfies the difference properties needed for finite decoding delay without requiring them to have maximum minimum Euclidean distance, then the two implementations will be equivalent, i.e., $\{s_k(t)\} = \{s_k'(t)\}$ only when this additional requirement is imposed. A formal proof of this equivalence is presented in Ref. 3. In what follows, we consider only the important practical case of antipodal signal sets and, as such, drop the prime notation on the signals derived from $p(t)$.

What remains is to consider the bandwidth efficiency of signals designed according to the constraints of (4.1-6), (4.1-7), and (4.1-11). This is where the ISI-based representation of Fig. 4-2(a) helps considerably, since the evaluation of the PSD of the transmitted signal can be trivially accomplished using well-known relations [6] for random pulse trains. This is considered in the next section.

4.1.2 Evaluation of the Power Spectral Density

In this section, we compute the PSD of a random complex pulse train, e.g., that in (4.1-1), modulated onto quadrature carriers. That is, if the transmitted bandpass signal is given by[4]

$$\begin{aligned}\tilde{s}(t) &= \mathrm{Re}\left\{s(t)\, e^{j2\pi f_m t}\right\} \\ &= \left(\sum_{n=-\infty}^{\infty} \bar{U}_n p_R(t - nT_b)\right)\cos 2\pi f_m t \\ &\quad - \left(\sum_{n=-\infty}^{\infty} \bar{U}_n p_I(t - nT_b)\right)\sin 2\pi f_m t \end{aligned} \tag{4.1-12}$$

then it is straightforward to show using an extension of the methods in Chap. 2 of Ref. 6 that the PSD of $\tilde{s}(t)$ is given by

$$\begin{aligned} S(f) &= \frac{1}{4T_b}\left|P_R(f - f_m) + jP_I(f - f_m)\right|^2 \\ &\quad + \frac{1}{4T_b}\left|P_R(f + f_m) - jP_I(f + f_m)\right|^2 \\ &= S_u(f) + S_l(f) \end{aligned} \tag{4.1-13}$$

where

$$\left.\begin{aligned} P_R(f) &\triangleq \mathcal{F}\{p_R(t)\} \\ P_R(f) &\triangleq \mathcal{F}\{p_I(t)\} \end{aligned}\right\} \tag{4.1-14}$$

are the Fourier transforms of the real and imaginary pulse shapes which, in general, are complex functions of f, and the u and l subscripts denote upper and lower sideband, respectively. Note that the signal in (4.1-12) differs from the usual QPSK-type of signal in that here, the same data sequence is passed

[4] We use the notation f_m for the actual modulating frequency of the quadrature carriers to distinguish it from the carrier frequency around which the PSD is symmetric, which will be denoted by f_c. More about this shortly.

through both the I and Q filters whereas for QPSK, the two sequences passing through these filters would be different and independent of one another. As such, the PSD in (4.1-13) cannot, in general, be written in the form [6, Eq. (2.131)]

$$S(f) = \frac{1}{4}G(f - f_c) + \frac{1}{4}G(f + f_c) \tag{4.1-15}$$

where $G(f)$ is the equivalent baseband (symmetrical around $f = 0$) PSD and is a real function of f, and f_c is some arbitrary carrier frequency.[5]

To demonstrate the above point, consider the specific case of MSK ($\nu = 1$), for which the four complex signals are given by[6]

$$\left.\begin{aligned} s_0(t) &= 0 + j1 \\ s_1(t) &= \sin\frac{\pi t}{T_b} - j\cos\frac{\pi t}{T_b} = s_0^*(t)\, e^{j\frac{\pi t}{T_b}} \\ s_2(t) &= -s_1(t) \\ s_3(t) &= -s_0(t) \end{aligned}\right\} \tag{4.1-16}$$

In terms of the ISI-based representation, we obtain from (4.1-8) that

$$\left.\begin{aligned} p_0(t) &= \frac{1}{2}\sin\frac{\pi t}{T_b} + j\frac{1}{2}\left[1 - \cos\frac{\pi t}{T_b}\right] \\ p_1(t) &= -\frac{1}{2}\sin\frac{\pi t}{T_b} + j\frac{1}{2}\left[1 + \cos\frac{\pi t}{T_b}\right] \end{aligned}\right\} \tag{4.1-17}$$

[5] What is meant by an "equivalent baseband PSD" is a PSD around zero frequency that is *identical* to the upper or lower sideband of the bandpass PSD, frequency-shifted to the origin. While it is always possible to express (4.1-13) in the form $S(f) = (1/4)G_u(f - f_c) + (1/4)G_l(f + f_c)$ where $G_u(f) = G_l(-f)$, in general, there is no guarantee that $G_u(f)$ [or equivalently, $G_l(f)$] has symmetry about the origin, or for that matter, about any frequency f_c. Stated another way, while demodulating the bandpass signal with a carrier at some frequency f_c (not necessarily equal to the modulating frequency f_m) will always produce a symmetric PSD around the origin, the resulting baseband PSD will, in general, be a combination (sum) of the aliased upper and lower sidebands, and may or may not appear as a simple frequency translation of either of these sidebands.

[6] Note that for the Rimoldi decomposition of MSK illustrated in Fig. 2-18, the signals satisfy the condition $s_0(t) - s_1(t) = -(s_2(t) - s_3(t))$ rather than $s_0(t) - s_1(t) = s_2(t) - s_3(t)$, as required by (4.1-5a), for the signals of (4.1-16) corresponding to precoded MSK.

Thus, using (4.1-17) to define the complex pulse shape of (4.1-8), we obtain

$$p(t) = \frac{1}{2}\sin\frac{\pi t}{T_b} + j\frac{1}{2}\left[1 - \cos\frac{\pi t}{T_b}\right], \quad 0 \le t \le 2T_b \tag{4.1-18}$$

That is, an appropriate implementation for MSK that guarantees a decoding delay of one bit is that of Fig. 4-2(b), with I and Q filters having impulse responses

$$\left.\begin{aligned} p_R(t) &= \frac{1}{2}\sin\frac{\pi t}{T_b}, \quad 0 \le t \le 2T_b \\ p_I(t) &= \frac{1}{2}\left[1 - \cos\frac{\pi t}{T_b}\right], \quad 0 \le t \le 2T_b \end{aligned}\right\} \tag{4.1-19}$$

Taking the Fourier transforms of $p_R(t)$ and $p_I(t)$ of (4.1-8) and using these in (4.1-13), we arrive at the following result for the bandpass PSD:

$$\begin{aligned} S(f) &= \frac{T_b}{4}\frac{\sin^2 2\pi(f - f_m)T_b}{\pi^2}\left[\frac{1}{1 - 2(f - f_m)T_b} + \frac{1}{2(f - f_m)T_b}\right]^2 \\ &\quad + \frac{T_b}{4}\frac{\sin^2 2\pi(f + f_m)T_b}{\pi^2}\left[\frac{1}{1 + 2(f + f_m)T_b} - \frac{1}{2(f + f_m)T_b}\right]^2 \\ &= S_u(f) + S_l(f) \end{aligned} \tag{4.1-20}$$

Note that while $S(f)$ is an even function of f (as it should be for a real signal), its upper and lower sidebands, $S_u(f)$ and $S_l(f)$, are not symmetric around f_m and $-f_m$, respectively. However, there does exist a frequency, $f_c \neq f_m$, around which the upper sideband (and similarly for the lower sideband) is symmetric. To understand why this is so, we remind the reader that according to Rimoldi's decomposition [5], the modulation frequency chosen for the quadrature carriers should be shifted from the carrier frequency f_c, around which the bandpass spectrum is to be symmetric by an amount equal to $1/4T_b$, i.e., $f_m = f_c - 1/4T_b$. This stems from the fact that the specification of the signals as in (4.1-16) results in a tilted trellis where the phase tilt is equal to $\pi/2$ rad. (Note that a frequency shift of $\Delta f = 1/4T_b$ is equal to a phase shift $2\pi\Delta f T = \pi/2$). To demonstrate that this is indeed the case, we evaluate the PSD of MSK, using (4.1-20) with the shifted value of modulating frequency, $f_m = f_c - 1/4T_b$. When this is done, the result in (4.1-15) is obtained with

$$G(f) = \frac{16T_b}{\pi^2} \frac{\cos^2 2\pi f T_b}{\left(1 - 16 f^2 T_b^2\right)^2} \tag{4.1-21}$$

which corresponds (except for a normalization factor) to the well-known PSD of MSK [6, Eq. (2.148)].

The question that comes about now is: For arbitrary memory, ν, and a baseband signal design satisfying (4.1-6), (4.1-7), and (4.1-11), is it possible to find a modulating frequency, f_m, that will produce a symmetric bandpass PSD around some other carrier frequency, f_c? If not, then one cannot find an equivalent baseband PSD, and, hence, the bandwidth (whatever measure is used) of the signal must be determined from the RF waveform.

4.1.2.1 The Memory One Case. To shed some light on the answer to the above question, we consider the simplest case of unit memory, where the complex pulse shape of (4.1-8) is given by

$$\begin{aligned} p(t) &= \frac{1}{2}\left[s_0(t) - s_2(t) + s_0(t - T_b) - s_1(t - T_b)\right] \\ &= \frac{1}{2}\left[s_0(t) + s_0(t - T_b) + s_1(t) + s_2(t - T_b)\right], \quad 0 \le t \le 2T_b \end{aligned} \tag{4.1-22}$$

where, in accordance with (4.1-11), we have used the fact that $s_1(t) = -s_2(t)$ in order to achieve $d_{\min}^2 = 2$. The Fourier transform of $p(t)$ in (4.1-22) is given by

$$\begin{aligned} P(f) = \frac{1}{2}\Bigg[&\int_0^{T_b} s_0(t)\left(1 + e^{-j2\pi f T_b}\right) e^{-j2\pi f t} dt \\ &+ \int_0^{T_b} s_1(t)\, e^{-j2\pi f t} dt + e^{-j2\pi f T_b} \int_0^{T_b} s_2(t)\, e^{-j2\pi f t} dt\Bigg] \end{aligned} \tag{4.1-23}$$

Since from (4.1-13), the upper spectral sideband is $S_u(f) = (1/4T_b)\left|P(f - f_m)\right|^2$, then in order for this to be symmetric around f_c, we must have

$$\left|P(f_c + f - f_m)\right|^2 = \left|P(f_c - f - f_m)\right|^2 \tag{4.1-24}$$

or letting $f_s \triangleq f_c - f_m$ denote the separation between the actual modulation frequency and the bandpass frequency around which symmetry is desired, $s_0(t)$ and $s_1(t)$ must be chosen to satisfy

$$|P(f_s+f)|^2 = |P(f_s-f)|^2 \tag{4.1-25a}$$

or equivalently

$$|P(f_s+f)|^2 = |P^*(f_s-f)|^2 \tag{4.1-25b}$$

for some f_s. In terms of (4.1-23), the spectral equality in (4.1-25b) requires that we have

$$\begin{aligned}
&\left|\int_0^{T_b}\left(s_0(t)\,e^{-j2\pi f_s t}\right)e^{-j2\pi f t}dt + e^{-j2\pi(f_s+f)T_b}\int_0^{T_b}\left(s_0(t)\,e^{-j2\pi f_s t}\right)e^{-j2\pi f t}dt\right.\\
&\left.+\int_0^{T_b}\left(s_1(t)\,e^{-j2\pi f_s t}\right)e^{-j2\pi f t}dt + e^{-j2\pi(f_s+f)T_b}\int_0^{T_b}\left(s_2(t)\,e^{-j2\pi f_s t}\right)e^{-j2\pi f t}dt\right|^2\\
&=\left|\int_0^{T_b}\left(s_0^*(t)\,e^{j2\pi f_s t}\right)e^{-j2\pi f t}dt + e^{j2\pi(f_s-f)T_b}\int_0^{T_b}\left(s_0^*(t)\,e^{j2\pi f_s t}\right)e^{-j2\pi f t}dt\right.\\
&\left.+\int_0^{T_b}\left(s_1^*(t)\,e^{j2\pi f_s t}\right)e^{-j2\pi f t}dt + e^{j2\pi(f_s-f)T_b}\int_0^{T_b}\left(s_2^*(t)\,e^{j2\pi f_s t}\right)e^{-j2\pi f t}dt\right|^2
\end{aligned} \tag{4.1-26}$$

Sufficient conditions on the signals, $\{s_i(t)\}$, for (4.1-26) to be satisfied are

$$\left.\begin{aligned}
s_1(t) &= s_0^*(t)\,e^{j4\pi f_s t}\\
s_2(t) &= e^{j4\pi f_s T_b}s_0^*(t)\,e^{j4\pi f_s t}
\end{aligned}\right\} \tag{4.1-27}$$

However, since in arriving at (4.1-26), we have already assumed that $s_1(t) = -s_2(t)$, then (4.1-27) further requires that $f_s = 1/4T_b$, from which we obtain the complete signal set

$$\left.\begin{aligned} s_1(t) &= s_0^*(t)\, e^{j\pi t/T_b} \\ s_2(t) &= -\, s_0^*(t)\, e^{j\pi t/T_b} \\ s_3(t) &= -\, s_0(t) \end{aligned}\right\} \tag{4.1-28}$$

Note that for memory one it is only necessary to specify $s_0(t)$ in order to arrive at the complete signal set. Also, the signal set of (4.1-28) satisfies the finite decoding delay condition of Ref. 3, namely, $s_0(t) - s_1(t) = s_2(t) - s_3(t)$.

The equivalent lowpass PSD is obtained by first using $s_1(t) = -s_2(t)$ in (4.1-23), resulting in

$$P(f) = \frac{1}{2}\Big[S_0(f) + S_1(f) + e^{-j2\pi f T_b}\big(S_0(f) - S_1(f)\big)\Big] \tag{4.1-29}$$

from which one immediately gets

$$\frac{1}{T_b}\,|P(f)|^2 =$$

$$\frac{1}{2T_b}\left[|S_0(f)|^2 + |S_1(f)|^2 + \mathrm{Re}\Big\{\big(S_0^*(f) + S_1^*(f)\big)\big(S_0(f) - S_1(f)\big)e^{-j2\pi f T_b}\Big\}\right] \tag{4.1-30}$$

In (4.1-29) and (4.1-30), $S_i(f)$ denotes the Fourier transform of $s_i(t)$. Using the first symmetry condition of (4.1-28) in (4.1-30) gives the desired equivalent lowpass PSD, namely,

$$\begin{aligned} \frac{1}{T_b}\left|P\left(f + \frac{1}{4T_b}\right)\right|^2 = &\left|S_0\left(f + \frac{1}{4T_b}\right)\right|^2 [1 - \sin 2\pi f T_b] \\ &+ \left|S_0\left(-f + \frac{1}{4T_b}\right)\right|^2 [1 + \sin 2\pi f T_b] \\ &+ 2\left[\mathrm{Re}\left\{S_0\left(f + \frac{1}{4T_b}\right)\right\}\mathrm{Im}\left\{S_0\left(-f + \frac{1}{4T_b}\right)\right\}\right. \\ &\left. + \mathrm{Re}\left\{S_0\left(-f + \frac{1}{4T_b}\right)\right\}\mathrm{Im}\left\{S_0\left(f + \frac{1}{4T_b}\right)\right\}\right]\cos 2\pi f T_b \end{aligned} \tag{4.1-31}$$

which is clearly an even function of frequency.

Although (4.1-28) is satisfied by the MSK signals of (4.1-16) as should be the case, this condition applies in a more general context since it does not explicitly specify $s_0(t)$ but rather only the *relation between* $s_0(t)$ and $s_1(t)$. This should not be surprising since it has been shown in the past that there exists an entire class of MSK-type signals (referred to in Ref. 7 as generalized MSK) which happen to also be constant envelope (in addition to being equal energy) and achieve $d^2_{\min} = 2$ as well as a decoding delay of one bit interval. In particular, the class of binary full-response CPM signals with modulation index $h = 1/2$ and equivalent phase pulse $q(t)$, which satisfies the conditions of (2.8-5), is appropriate, an example of which is Amoroso's SFSK [8] for which $q(t)$ is given by (2.8-9).

4.1.2.2 The Memory Two Case. For memory two, the pulse shape is given by

$$p(t) = \frac{1}{2}\left[s_0(t) - s_4(t) + s_0(t-T_b) - s_2(t-T_b) + s_0(t-2T_b) - s_1(t-2T_b)\right]$$

$$= \frac{1}{2}\left[s_0(t) + s_0(t-T_b) + s_0(t-2T_b) + s_3(t) - s_2(t-T_b) - s_1(t-2T_b)\right],$$

$$0 \le t \le 3T_b \qquad (4.1\text{-}32)$$

with Fourier transform

$$P(f) = \frac{1}{2}\left[\left(1 + e^{-j2\pi f T_b} + e^{-j4\pi f T_b}\right)\int_0^{T_b} s_0(t)\, e^{-j2\pi f t} dt + \int_0^{T_b} s_3(t)\, e^{-j2\pi f t} dt\right.$$

$$\left. - e^{-j2\pi f T_b}\int_0^{T_b} s_2(t)\, e^{-j2\pi f t} dt - e^{-j4\pi f T_b}\int_0^{T_b} s_1(t)\, e^{-j2\pi f t} dt\right] \qquad (4.1\text{-}33)$$

Applying (4.1-33) to (4.1-25b) and letting $s_3(t) = s_2(t) - s_0(t) + s_1(t)$, in accordance with (4.1-5a), we obtain the bandpass spectral symmetry condition

$$\Bigg| e^{-j2\pi(f_s+f)T_b} \int_0^{T_b} \left(s_0(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt + e^{-j4\pi(f_s+f)T_b}$$

$$\times \int_0^{T_b} \left(s_0(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt$$

$$+ \int_0^{T_b} \left(s_2(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt - e^{-j2\pi(f_s+f)T_b} \int_0^{T_b} \left(s_2(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt$$

$$+ \int_0^{T_b} \left(s_1(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt - e^{-j4\pi(f_s+f)T_b} \int_0^{T_b} \left(s_1(t)\, e^{-j2\pi f_s t}\right) e^{-j2\pi f t} dt \Bigg|^2$$

$$= \Bigg| e^{j2\pi(f_s-f)T_b} \int_0^{T_b} \left(s_0^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt + e^{j4\pi(f_s-f)T_b}$$

$$\times \int_0^{T_b} \left(s_0^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt$$

$$+ \int_0^{T_b} \left(s_2^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt - e^{j2\pi(f_s-f)T_b} \int_0^{T_b} \left(s_2^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt$$

$$+ \int_0^{T_b} \left(s_1^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt - e^{j4\pi(f_s-f)T_b} \int_0^{T_b} \left(s_1^*(t)\, e^{j2\pi f_s t}\right) e^{-j2\pi f t} dt \Bigg|^2 \tag{4.1-34}$$

Analogous with (4.1-27), satisfying (4.1-34) implies the set of conditions

$$s_1(t) + s_2(t) = \left(s_1^*(t) + s_2^*(t)\right) e^{j4\pi f_s t} \tag{4.1-35a}$$

$$s_0(t) - s_2(t) = e^{j4\pi f_s T_b} \left(s_0^*(t) - s_2^*(t)\right) e^{j4\pi f_s t} \tag{4.1-35b}$$

$$s_0(t) - s_1(t) = e^{j8\pi f_s T_b} \left(s_0^*(t) - s_1^*(t)\right) e^{j4\pi f_s t} \tag{4.1-35c}$$

Again letting $f_s = 1/4T_b$ and summing (4.1-35a), (4.1-35b), and (4.1-35c) gives

$$s_1(t) + s_2(t) = \left(s_1^*(t) + s_2^*(t)\right)e^{j\pi t/T_b} \tag{4.1-36a}$$

$$s_0(t) = s_2^*(t)\, e^{j\pi t/T_b} \left(\text{or equivalently } s_2(t) = s_0^*(t)\, e^{j\pi t/T_b}\right) \tag{4.1-36b}$$

$$s_0(t) - s_1(t) = \left(s_0^*(t) - s_1^*(t)\right)e^{j\pi t/T_b} \tag{4.1-36c}$$

Actually, (4.1-36c) is not an independent condition since it can be derived from (4.1-36a) and (4.1-36b). Thus, (4.1-36a) and (4.1-36b) are sufficient to determine the signal design.

Following along the lines of (4.1-29) and (4.1-30), the equivalent PSD of the memory two modulation may be found. In particular, the Fourier transform of the equivalent pulse shape in (4.1-8) is given as

$$P(f) =$$

$$\frac{1}{2}\Big[S_0(f) + S_3(f) + e^{-j2\pi f T_b}\big(S_0(f) - S_2(f)\big) + e^{-j4\pi f T_b}\big(S_0(f) - S_1(f)\big)\Big] \tag{4.1-37}$$

Using the additional relation, $S_3(f) = S_1(f) + S_2(f) - S_0(f)$, to achieve finite decoding delay, one immediately gets the desired equivalent lowpass PSD as

$$\frac{1}{T_b}\left|P\left(f + \frac{1}{4T_b}\right)\right|^2 =$$

$$\frac{1}{4T_b}\Bigg[\left|S_1\left(f + \frac{1}{4T_b}\right) + S_2\left(f + \frac{1}{4T_b}\right)\right|^2$$

$$+ \left|S_0\left(f + \frac{1}{4T_b}\right) - S_2\left(f + \frac{1}{4T_b}\right)\right|^2 + \left|S_0\left(f + \frac{1}{4T_b}\right) - S_1\left(f + \frac{1}{4T_b}\right)\right|^2$$

$$+ 2\mathrm{Re}\Bigg\{\left[S_1^*\left(f + \frac{1}{4T_b}\right) + S_2^*\left(f + \frac{1}{4T_b}\right)\right]$$

$$\times \left[S_0\left(f + \frac{1}{4T_b}\right) - S_2\left(f + \frac{1}{4T_b}\right)\right] e^{-j2\pi(f + [1/4T_b])T_b}\Bigg\}$$

$$+ 2\mathrm{Re}\left\{\left[S_0^*\left(f+\frac{1}{4T_b}\right) - S_2^*\left(f+\frac{1}{4T_b}\right)\right]\right.$$

$$\left.\times\left[S_0\left(f+\frac{1}{4T_b}\right) - S_1\left(f+\frac{1}{4T_b}\right)\right]e^{-2\pi(f+[1/4T_b])T_b}\right\}$$

$$+ 2\mathrm{Re}\left\{\left[S_1^*\left(f+\frac{1}{4T_b}\right) + S_2^*\left(f+\frac{1}{4T_b}\right)\right]\right.$$

$$\left.\left.\times\left[S_0\left(f+\frac{1}{4T_b}\right) - S_1\left(f+\frac{1}{4T_b}\right)\right]e^{-4\pi(f+[1/4T_b])T_b}\right\}\right] \tag{4.1-38}$$

which, when (4.1-36) is used, can be shown to be an even function of frequency, as is necessary.

4.1.3 Optimizing the Bandwidth Efficiency

Having obtained expressions for the equivalent baseband PSD, it is now straightforward to use these to determine the sets of signals that satisfy all of the previous constraints and, in addition, maximize the power within a given bandwidth, B. In mathematical terms, we search for the set of signals that for a given value of B maximizes the fractional in-band power

$$\eta = \frac{\int_{-B/2}^{B/2} G(f)\,df}{\int_{-\infty}^{\infty} G(f)\,df}, \quad G(f) \triangleq \frac{1}{T_b}\left|P\left(f+\frac{1}{4T_b}\right)\right|^2 \tag{4.1-39}$$

subject to the unit power constraint

$$\frac{1}{T_b}\int_0^{T_b} |s_i(t)|^2\,dt = \frac{1}{T_b}\int_{-\infty}^{\infty} |S_i(f)|^2\,dt = 1, \quad i = 0, 1, 2, \cdots, M-1 \tag{4.1-40}$$

4.1.3.1 Memory One Case. For the case of $\nu = 1$, we observed that the entire signal set may be determined from the single complex signal, $s_0(t)$. Thus, optimizing bandwidth efficiency corresponds to substituting the PSD of (4.1-31) (which is entirely specified in terms of the Fourier transform of $s_0(t)$) into (4.1-39) and then maximizing η subject to (4.1-40). Such a procedure would result in an optimum $S_0(f)$ from whose inverse Fourier transform one could determine the optimum signal set. Since $S_0(f)$ exists, in general, over the entire doubly infinite frequency axis, it is perhaps simpler to approach the optimization in the

time domain, since $s_0(t)$ is indeed time limited to the interval $0 \le t \le T_b$. To do this, we need to first rewrite the PSD of (4.1-40) in terms of $s_0(t)$ rather than $S_0(f)$ and then perform the integrations on f required in (4.1-39). After considerable manipulation, and for simplicity of notation normalizing $T_b = 1$ (i.e., $BT_b = B$), it can be shown that

$$\int_{-B/2}^{B/2} G(f)\,df =$$

$$B\int_0^1\int_0^1 s_0(t)\,s_0^*(\tau)\,e^{-j(\pi/2)(t-\tau)}$$

$$\times\left[\text{sinc}\,\pi B(t-\tau) - j\frac{1}{2}\text{sinc}\,\pi B(t-\tau+1) + j\frac{1}{2}\text{sinc}\,\pi B(t-\tau-1)\right]dtd\tau$$

$$+\frac{1}{2}B\,\text{Im}\left\{\int_0^1\int_0^1 s_0(t)\,s_0(\tau)\,e^{-j(\pi/2)(t+\tau)}\right.$$

$$\left.\times\left[\text{sinc}\,\pi B(t-\tau+1) + \text{sinc}\,\pi B(t-\tau-1)\right]dtd\tau\right\} \tag{4.1-41}$$

where $\text{sinc}\,x \triangleq \sin x/x$. Furthermore, it is straightforward to show that

$$\int_{-\infty}^{\infty} G(f)\,df = 1 \tag{4.1-42}$$

and, thus, η is given directly by (4.1-41).

The maximization of (4.1-41) subject to the energy constraint of (4.1-40) has been carried out numerically, using the MATLAB$^{(r)}$ (software application) optimization toolbox function "fminunc" (quasi-Newton method of convergence). In particular, for each value of B (BT_b if $T_b \neq 1$), the optimum complex signal, $s_0(t)$, [represented by N uniformly spaced samples in the interval $(0,1)$], is determined, from which the fractional out-of-band power, $1-\eta$, is calculated using (4.1-41) for η. Because of complexity issues involved in computing the optimum solution, the number of sample points, N, is limited to 64. Furthermore, since the Gaussian integration required to evaluate with high accuracy the double integral of (4.1-41) requires a much higher density of sample values (not necessarily uniformly spaced), then to allow for Fourier interpolation, we assume the signal

to be bandlimited[7] to the Nyquist rate, i.e., 32 ($32/T_b$ if $T_b \neq 1$). Because of this bandlimiting assumption, certain optimum signal waveforms (particularly those at small values of B) that exhibit a sharp discontinuity will have a ringing behavior. This ringing behavior can be minimized by additional interpolation (filtering) but has proven difficult to eliminate completely.

Figures 4-3(a) and 4-3(b) are 3-D plots of the optimum real and imaginary parts of $s_0(t)$ versus t as a function of B in the interval $0 \leq B \leq 3$. Figures 4-4(a)–(h) are a number of cuts of these 3-D plots taken at distinct values of B in the same range. For small values of B, we observe that the real part of $s_0(t)$ has sharp discontinuities at $t = 0$ and $t = 1$ and, thus, exhibits the ringing behavior alluded to above. As B increases, the sharpness of the discontinuity at the edges diminishes, and in the limit of large B, both the real and imaginary parts of $s_0(t)$ approach a sinusoid with unit period. Specifically, $s_0(t)$ tends toward the form $-\alpha_1 \sin 2\pi t + j(\beta_1 + \alpha_2 \cos 2\pi t)$, where $\alpha_1, \alpha_2, \beta_1$ are constants that also must satisfy the unit energy constraint, i.e., $\beta_1^2 + (1/2)\left(\alpha_1^2 + \alpha_2^2\right) = 1$. Figure 4-5 is the corresponding plot of optimum (minimum) fractional out-of-band power versus B. Also shown are corresponding results for MSK and SFSK modulations that can readily be found in Fig. 2.11 of Ref. 6.[8] We observe that by optimizing the signal set at each value of B without loss in $d_{\min}^2$ or finite decoding delay performance, we are able to obtain a significant improvement in bandwidth efficiency. The quantitative amount of this improvement is given in Table 4-1 for the 99 percent and 99.9 percent bandwidths corresponding respectively to the -20 dB and -30 dB out-of-band power levels.

Before concluding this section, we note that the maximization of (4.1-41) subject to the constraint in (4.1-40) can be carried out analytically using the method of calculus of variations. Unfortunately, however, the resulting solution for $s_0(t)$ is in the form of an integral equation that does not lend itself to a

Table 4-1. Bandwidth-efficient performance of TCM with prescribed decoding delay.

Signal	$1/B_{99}T_b$ ([b/s]/Hz)	% Improvement over MSK	$1/B_{99.9}T_b$ ([b/s]/Hz)	% Improvement over MSK
MSK	0.845	—	0.366	—
Optimum ($\nu = 1$)	0.896	6.04	0.659	79.7
Optimum ($\nu = 2$)	1.23	45.6	—	—

[7] Of course, in reality the continuous time-limited signal would have infinite bandwidth.

[8] Note that the definition of bandwidth B in Ref. 1 is one-half of that used in this monograph.

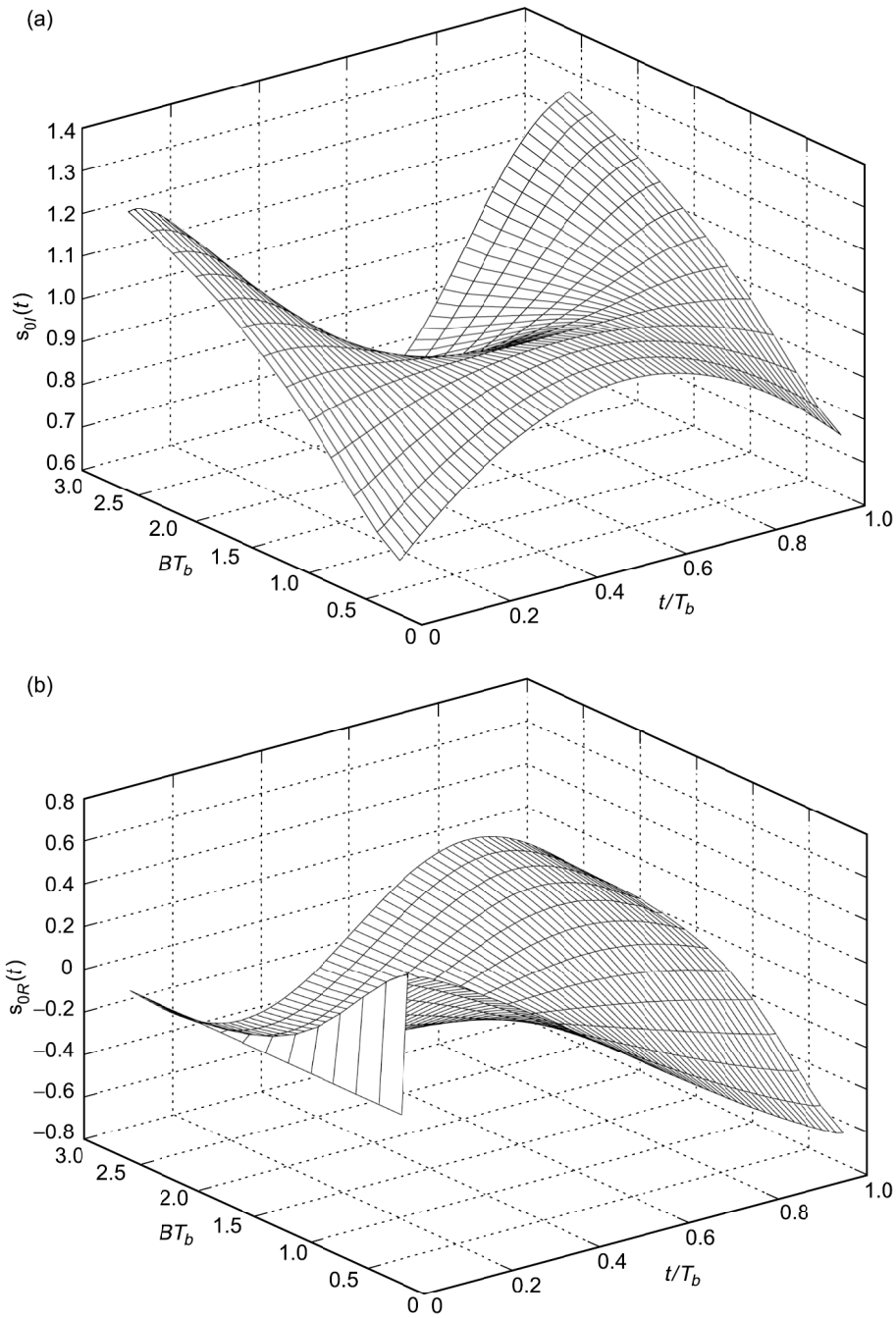


Fig. 4-3. Profiles of the optimum signal as a function of the bandwidth-bit time product: (a) the imaginary part and (b) the real part.

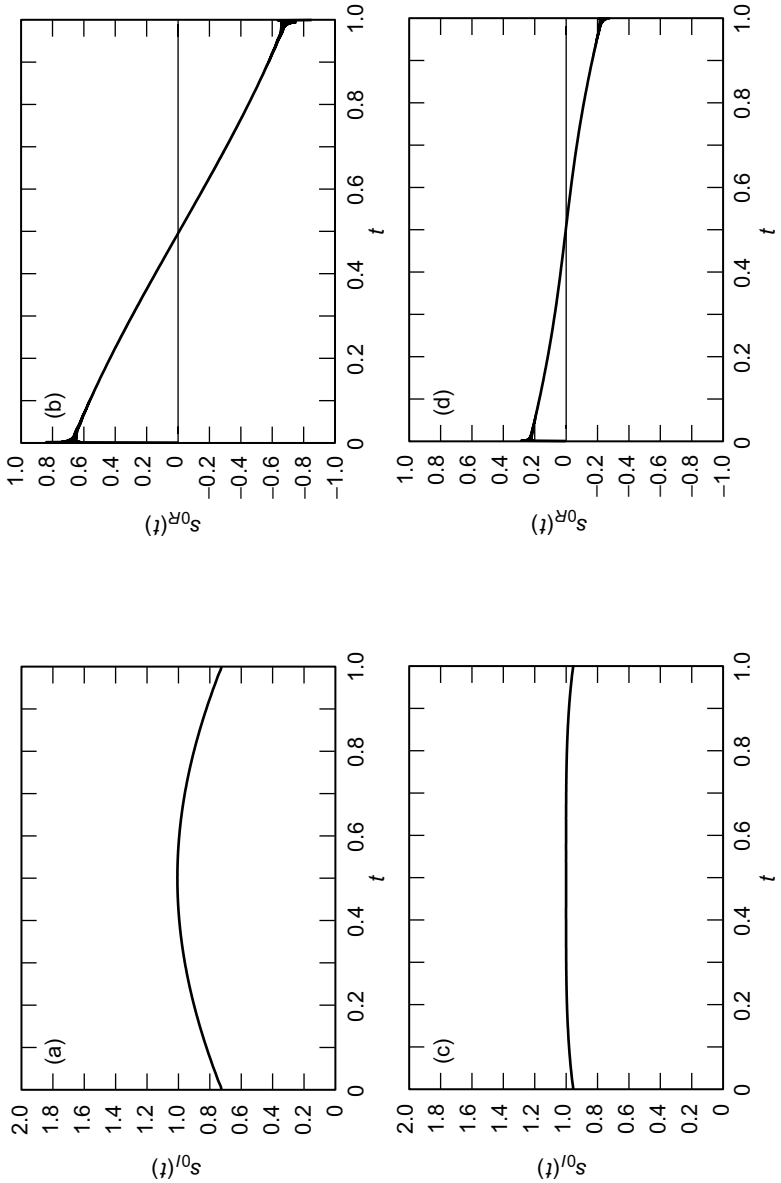


Fig. 4-4. The optimum signal for bandwidth-time product: (a) the imaginary part of the optimum signal for bandwidth-time product = 0.2, (b) the real part of the optimum signal for bandwidth-time product = 0.2, (c) the imaginary part of the optimum signal for bandwidth-time product = 1.0, and (d) the real part of the optimum signal for bandwidth-time product = 1.0.

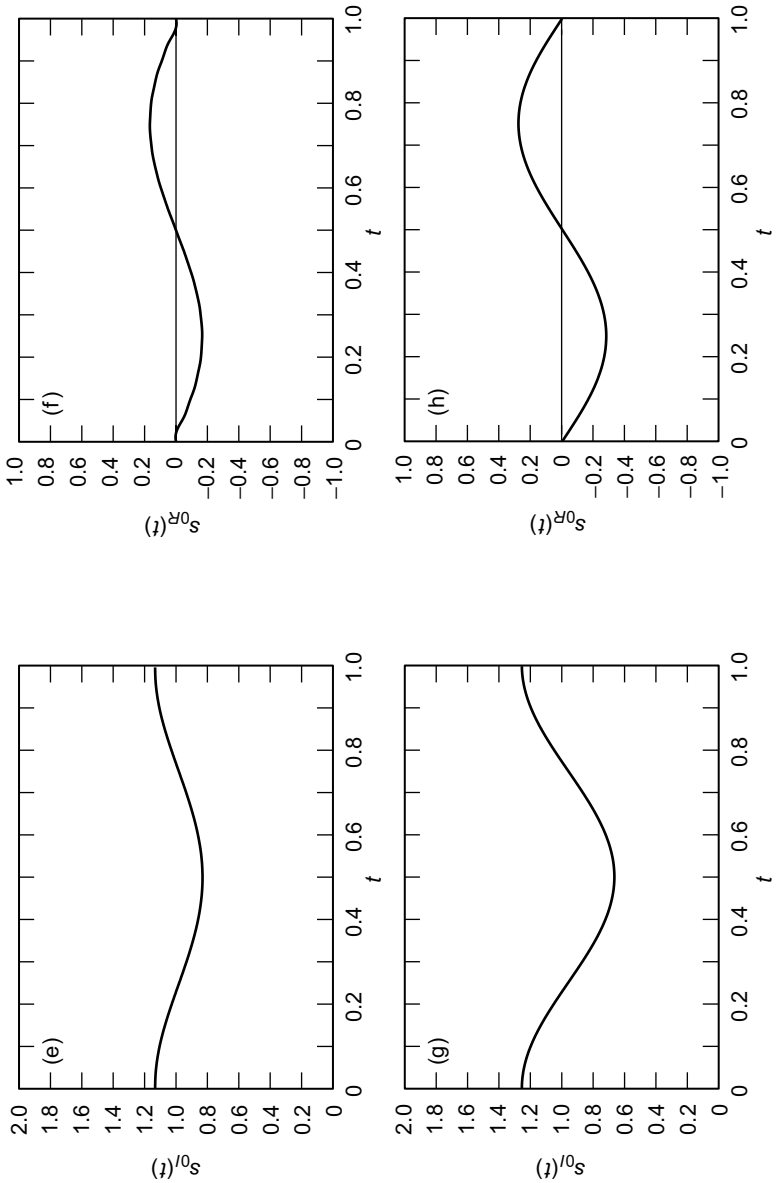


Fig. 4-4 (cont'd). The optimum signal for bandwidth-time product: (e) the imaginary part of the optimum signal for bandwidth-time product = 1.8, (f) the real part of the optimum signal for bandwidth-time product = 1.8, (g) the imaginary part of the optimum signal for bandwidth-time product = 2.6, and (h) the real part of the optimum signal for bandwidth-time product = 2.6.

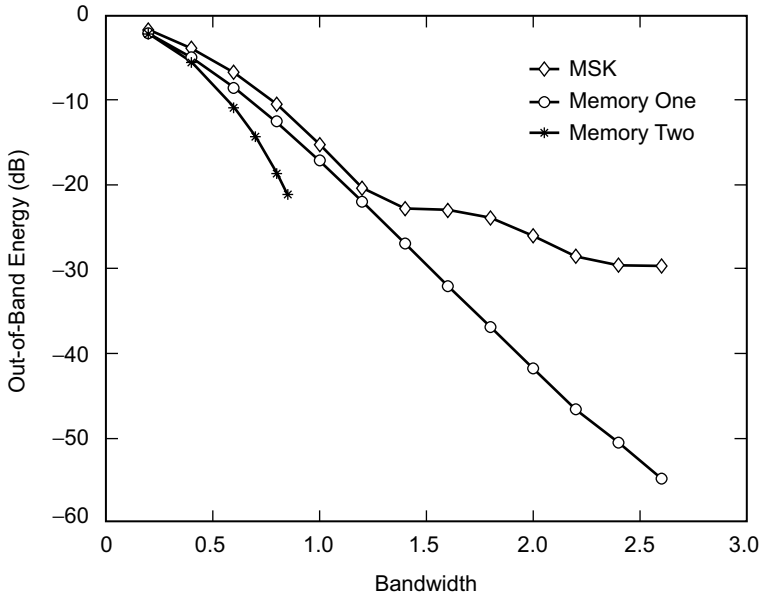


Fig. 4-5. Comparison of fractional out-of-band powers.

closed-form solution. Thus, there is no strong advantage to presenting these results here since we have already obtained a numerical solution as discussed above by direct maximization of (4.1-41). One interesting observation does result from applying the calculus of variations approach: $s_{0R}(t)$ is an odd function around its midpoint (at $t = 1/2$) and $s_{0I}(t)$ is an even function around its same midpoint. Clearly, this observation is justified by the numerical results illustrated in the various parts of Fig. 4-4.

4.1.3.2 Memory Two Case. Analogous to what was done for the memory one case, we need to maximize the fractional in-band power of (4.1-39), using now (4.1-31) for $G(f)$. Expressing the various Fourier transforms of (4.1-31) in terms of their associated signal waveforms and then performing the integration on frequency between $-B/2$ and $B/2$ as required in (4.1-39) produces the following result (again normalizing $T_b = 1$):

$$\int_{-B/2}^{B/2} G(f)\, df = \sum_{i=1}^{6} P_i \tag{4.1-43}$$

where

$$\left.\begin{aligned}
P_1 =& \frac{B}{4}\int_0^1\int_0^1 \left(s_1^{(2)}(t)+s_2^{(2)}(t)\right)\left(s_1^{(2)}(\tau)+s_2^{(2)}(\tau)\right)^* \\
&\times e^{-j(\pi/2)(t-\tau)}\text{sinc}\ \pi B(t-\tau)dtd\tau \\
P_2 =& \frac{B}{4}\int_0^1\int_0^1 \left(s_0^{(2)}(t)-s_2^{(2)}(t)\right)\left(s_0^{(2)}(\tau)-s_2^{(2)}(\tau)\right)^* \\
&\times e^{-j(\pi/2)(t-\tau)}\text{sinc}\ \pi B(t-\tau)dtd\tau \\
P_3 =& \frac{B}{4}\int_0^1\int_0^1 \left(s_0^{(2)}(t)-s_1^{(2)}(t)\right)\left(s_0^{(2)}(\tau)-s_1^{(2)}(\tau)\right)^* \\
&\times e^{-j(\pi/2)(t-\tau)}\text{sinc}\ \pi B(t-\tau)dtd\tau \\
P_4 =& 2\ \text{Re}\left\{\frac{B}{4}\int_0^1\int_0^1 \left(s_0^{(2)}(t)-s_2^{(2)}(t)\right)\left(s_1^{(2)}(\tau)+s_2^{(2)}(\tau)\right)^* \right. \\
&\left.\times e^{-j(\pi/2)(t-\tau+1)}\text{sinc}\ \pi B(t-\tau+1)dtd\tau\right\} \\
P_5 =& 2\ \text{Re}\left\{\frac{B}{4}\int_0^1\int_0^1 \left(s_0^{(2)}(t)-s_1^{(2)}(t)\right)\left(s_0^{(2)}(\tau)-s_2^{(2)}(\tau)\right)^* \right. \\
&\left.\times e^{-j(\pi/2)(t-\tau+1)}\text{sinc}\ \pi B(t-\tau+1)dtd\tau\right\} \\
P_6 =& 2\ \text{Re}\left\{\frac{B}{4}\int_0^1\int_0^1 \left(s_0^{(2)}(t)-s_1^{(2)}(t)\right)\left(s_1^{(2)}(\tau)+s_2^{(2)}(\tau)\right)^* \right. \\
&\left.\times e^{-j(\pi/2)(t-\tau+2)}\text{sinc}\ \pi B(t-\tau+2)dtd\tau\right\}
\end{aligned}\right\} \quad (4.1\text{-}44)$$

From the constraint in (4.1-36b), $s_2^{(2)}(t)$ can be expressed in terms of $s_0^{(2)*}(t)$ and then substituted in (4.1-44). Thus, the optimization problem reduces to finding only two signals, $s_0^{(2)}(t)$ and $s_1^{(2)}(t)$, by joint maximization of (4.1-43) combined with (4.1-44). (Note that $s_3^{(2)}(t)$ can be found from $s_3^{(2)}(t) = s_2^{(2)}(t) - s_0^{(2)}(t) + s_1^{(2)}(t)$, once $s_0^{(2)}(t)$ and $s_1^{(2)}(t)$ are determined.)

Superimposed on Fig. 4-5 are the optimum fractional out-of-band power results for the memory two case. Due to the extremely time-consuming nature of the computer algorithms that perform the joint optimization procedure, particularly at low levels of fractional out-of-band power where extreme accuracy in satsifying the constraints is required, only results corresponding to values of $BT_b < 1$ (or equivalently $B < 1$ for $T_b = 1$) have been obtained thus far. Nevertheless, we are able to extract from these results the bandwidth-efficiency improvement relative to MSK for the 99 percent (-20 dB) out-of-band power level, and this improvement is included in Table 4-1. We observe that there is a significant improvement in out-of-band power performance, with no power efficiency penalty, by going from a memory one (1-bit decoding delay) modulation to one that has memory two (2-bit decoding delay).

4.2 Bandwidth-Efficient TCM with Prescribed Decoding Delay—Unequal Signal Energies

In the introduction to this chapter, we said that a relaxation of the equal energy condition on the signals could be used to potentially trade off between the power and bandwidth efficiency of the system. We now investigate the additional constraints that must be placed on the signals in order that the optimum TCM receiver still achieve a finite decoding delay equal to the memory of the modulation. In order to accomplish this, we first briefly review the received signal plus noise model, branch metric, and accompanying decision rule leading up to the conditions on the signal differences in Theorem I of Ref. 3 [summarized herein in (4.1-6) and (4.1-7)] and then modify them so as to apply to the case of unequal signal energies.

Corresponding to the baseband signal, $s(t)$, of (4.1-1) transmitted over an AWGN channel, the received signal is

$$R(t) = s(t) + N(t) \tag{4.2-1}$$

where $N(t)$ is again a zero-mean complex Gaussian noise process with PSD N_0 watts/hertz. For equal energy signals, the maximum-likelihood (Viterbi) receiver uses as its branch metric in the nth interval

$$\begin{aligned}\lambda_n(s_i) &= \mathrm{Re}\left\{\int_{nT}^{(n+1)T} R^*(t)\, s_i(t-nT)\, dt\right\} \\ &= \mathrm{Re}\left\{\int_0^T R^*(t+nT)\, s_i(t)\, dt\right\}, \quad i \in \left\{0, 1, \cdots, 2^{\nu+1}-1\right\}\end{aligned} \tag{4.2-2}$$

As previously stated, without any constraints on the signal set, for true optimality, the Viterbi receiver *theoretically* needs to observe the entire transmitted sequence (sum over an infinite number of branch metrics), resulting in an infinite decoding delay although in practice one may decode with finite delay using a truncated (but suboptimal) form of VA. If the signal differences are constrained as in (4.1-6) and (4.1-7), then, as previously stated in Theorem I of Ref. 3, the receiver can optimally decode the nth information symbol after ν symbol intervals, according to the decision rule:

$$\text{Choose } U_n = 0 \text{ if } \sum_{i=\nu}^{n+\nu} \lambda_i \left(s_0 - s_{2^{n+\nu-i}}\right) > 0, \quad \text{otherwise choose } U_n = 1 \quad (4.2\text{-}3)$$

For unequal energy signals, the branch metric of (4.2-2) would be modified to

$$\lambda_n\left(s_i\right) = \operatorname{Re}\left\{\int_{nT}^{(n+1)T} R^*\left(t\right) s_i\left(t - nT\right) dt\right\} - \frac{E_i}{2}$$

$$= \operatorname{Re}\left\{\int_0^T R^*\left(t + nT\right) s_i\left(t\right) dt\right\} - \frac{E_i}{2}, \quad i \in \left\{0, 1, \cdots, 2^{\nu+1} - 1\right\} \quad (4.2\text{-}4)$$

where $E_i = \int_0^T \left|s_i\left(t\right)\right|^2 dt$ is the energy of the ith signal in the set. Since the derivation of the conditions for finite decoding delay given in Ref. 3 relies on comparisons of sums of branch metrics, it is straightforward to substitute (4.2-4) for (4.2-2) in the steps of this derivation, which leads to an additional set of conditions on the energies of the signals. To illustrate the procedure, we first consider the simplest case corresponding to unit memory ($\nu = 1$).

Consider the two-state trellis (corresponding to the nth and $n+1$st intervals) in Fig. 4-6, where each branch is labeled with: (a) the input bit that causes the transition between states and (b) the baseband signal transmitted in accordance with the choice defined in Fig. 4-1(b). Assume first that we are in state "0" at time n (having gotten there as a result of decoding symbols in the previous intervals). Suppose now that the two paths (of length two branches) that survive at time $n+2$ are those that merge at (eminate from) the same node at time $n+1$ (thereby allowing unique decoding of the transmitted symbol, U_n). Since this node can correspond to either state "0" or state "1," there exist two possibilities, which are indicated by heavy lines in Figs. 4-5(a) and 4-5(b).

For Fig. 4-6(a), both surviving paths have a first branch corresponding to $U_n = 1$ and, thus, the decision $\hat{U}_n = 1$ is unique provided that

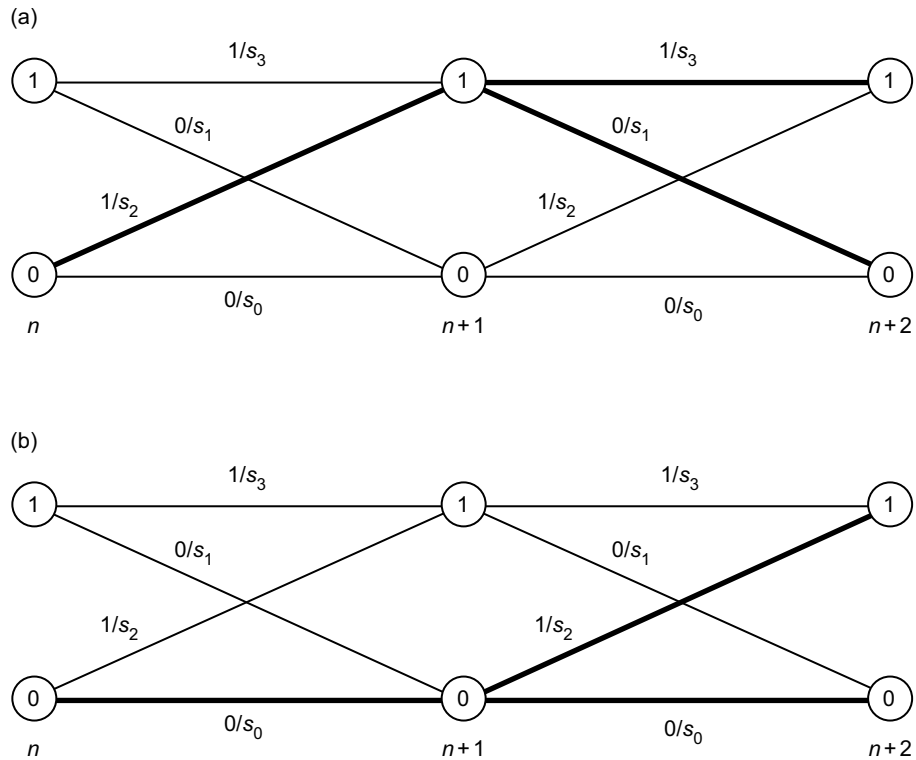


Fig. 4-6. A trellis diagram for memory one modulation, assuming state "0" at time *n*: (a) surviving paths merging at state "1" at time *n* + 1 and (b) surviving paths merging at state "0" at time *n* + 1.

$$\lambda_n(s_2) + \lambda_{n+1}(s_3) > \lambda_n(s_0) + \lambda_{n+1}(s_2) \tag{4.2-5a}$$

and

$$\lambda_n(s_2) + \lambda_{n+1}(s_1) > \lambda_n(s_0) + \lambda_{n+1}(s_0) \tag{4.2-5b}$$

or equivalently

$$\lambda_n(s_0) - \lambda_n(s_2) + \lambda_{n+1}(s_2) - \lambda_{n+1}(s_3) < 0 \tag{4.2-6a}$$

and

$$\lambda_n(s_0) - \lambda_n(s_2) + \lambda_{n+1}(s_0) - \lambda_{n+1}(s_1) < 0 \tag{4.2-6b}$$

To simultaneously satisfy (4.2-6a) and (4.2-6b), we need to have

$$\lambda_{n+1}(s_2) - \lambda_{n+1}(s_3) = \lambda_{n+1}(s_0) - \lambda_{n+1}(s_1) \tag{4.2-7}$$

which is the identical requirement found by Li and Rimoldi [3] when treating the equal signal energy case. Using instead now the metric definition in (4.2-4) for unequal energy signals, then analogous to the results in Ref. 3, the condition of (4.2-7) can be satisfied by the first equality in (4.1-5a), namely,

$$s_0(t) - s_1(t) = s_2(t) - s_3(t) \tag{4.2-8}$$

and, furthermore,

$$E_0 - E_1 = E_2 - E_3 \tag{4.2-9}$$

Note that the relation in (4.2-9) is identical in form to that in (4.2-8) if each of the signals in the latter is replaced by its energy. This observation will carry over when we consider modulations with memory greater than one.

For Fig. 4-6(b), both surviving paths have a first branch corresponding to $U_n = 0$ and thus the decision $\hat{U}_n = 0$ is unique provided that

$$\lambda_n(s_0) + \lambda_{n+1}(s_2) > \lambda_n(s_2) + \lambda_{n+1}(s_3) \tag{4.2-10a}$$

and

$$\lambda_n(s_0) + \lambda_{n+1}(s_0) > \lambda_n(s_2) + \lambda_{n+1}(s_1) \tag{4.2-10b}$$

or equivalently

$$\lambda_n(s_0) - \lambda_n(s_2) + \lambda_{n+1}(s_2) - \lambda_{n+1}(s_3) > 0 \tag{4.2-11a}$$

and

$$\lambda_n(s_0) - \lambda_n(s_2) + \lambda_{n+1}(s_0) - \lambda_{n+1}(s_1) > 0 \tag{4.2-11b}$$

It is clear that the condition in (4.2-7) will also simultaneously satisfy (4.2-11a) and (4.2-11b).

Finally, if we assume that we were in state "1" at time n, then it is straightforward to show that the conditions on the signal set that produce a unique decision on U_n would be identical to those in (4.2-8) and (4.2-9). Thus, we conclude that *for a memory one modulation of the type described by Fig. 4-1(b) with unequal energy signals, the conditions on the signal set to guarantee unique decodability with one symbol delay are those given in (4.2-8) and (4.2-9).*

To extend the above to modulations with memory ν greater than one, we proceed as follows: As was observed in Ref. 3, what we now seek are the inequality conditions on the sums of branch metrics such that the 2^ν surviving paths at time $n+\nu$ merge at a single node at time $n+1$. Given a particular state at time n, this set of 2^ν conditions then allows for uniquely decoding U_n. Since these conditions are expressed entirely in terms of the branch metrics for the surviving paths, and, as such, do not depend on the form of the metric itself (i.e., whether it be (4.2-2) for equal energy signals or (4.2-4) for unequal energy signals), then it is straightforward to conclude that the finite decoding delay conditions on the signal set derived in Ref. 3 for the equal energy case also apply now to the signal energies in the nonequal energy case. Specifically, in addition to (4.1-6), the signal set must satisfy the energy conditions

$$E_0 - E_{2^m} = E_{2^{m+1}l} - E_{2^{m+1}l+2^m}, \; m = 0, 1, 2, \cdots, \nu - 1, \; l = 1, 2, \cdots, 2^{\nu-m} - 1 \tag{4.2-12}$$

For the equal energy case, (4.2-12) is trivially satisfied.

Having now specified the conditions for achieving finite decoding delay with unequal energy signals, we now investigate the impact of this relaxed restriction on the minimum-squared Euclidean distance (power efficiency) of the modulation. Again consider first the memory one case. For the trellis diagram of Fig. 4-5(a), the unnormalized squared Euclidean distance between the length 2 error event path and the all zeros path (corresponding to $U_n = 0, U_{n+1} = 0$) is

$$\begin{aligned} D^2 &= \int_0^T |s_0(t) - s_2(t)|^2 \, dt + \int_0^T |s_0(t) - s_1(t)|^2 \, dt \\ &= 2E_0 + E_1 + E_2 - 2 \operatorname{Re}\left\{ \int_0^T s_0^*(t) \left(s_1(t) + s_2(t)\right) dt \right\} \end{aligned} \tag{4.2-13}$$

Using (4.2-8) and (4.2-9) in (4.2-13) enables rewriting it in the form

$$\left.\begin{aligned} D^2 &= 2E_{av} - 2\ \mathrm{Re}\left\{\int_0^T s_0^*(t)\, s_3(t)\, dt\right\} \\ E_{av} &= \frac{E_0 + E_1 + E_2 + E_3}{4} = \frac{E_0 + E_3}{2} \end{aligned}\right\} \tag{4.2-14}$$

which when normalized by the average energy of the signal set, E_{av}, gives

$$d^2 \triangleq \frac{D^2}{2E_{av}} = 1 - \frac{\mathrm{Re}\left\{\int_0^T s_0^*(t)\, s_3(t)\, dt\right\}}{E_{av}} = 1 - \frac{\mathrm{Re}\left\{\int_0^T s_0^*(t)\, s_3(t)\, dt\right\}}{(E_0 + E_3)/2} \tag{4.2-15}$$

Following steps analogous to (4.2-13)–(4.2-15) and using the signal difference property in (4.2-8), it is straightforward to show that the unnormalized squared Euclidean distance between any pair of length 2 paths beginning and ending at the same node (i.e., other pairwise error events) is given by (4.2-15), i.e., the trellis has a uniform error probability (UEP) property. It can also be shown using a combination of (4.2-8) and (4.2-9) in (4.2-13) that (4.2-15) can be expressed as

$$d^2 \triangleq \frac{D^2}{2E_{av}} = 1 - \frac{\mathrm{Re}\left\{\int_0^T s_1^*(t)\, s_2(t)\, dt\right\}}{(E_1 + E_2)/2} \tag{4.2-16}$$

Finally noting that $-1 \leq \mathrm{Re}\left\{\int_0^T s_0^*(t)\, s_3(t)\, dt\right\} / [(E_0 + E_3)/2]$ with equality achieved when $s_0(t) = -s_3(t)$ and, likewise, $-1 \leq \mathrm{Re}\left\{\int_0^T s_1^*(t)\, s_2(t)\, dt\right\} / [(E_1 + E_2)/2]$ with equality achieved when $s_1(t) = -s_2(t)$, then, in order to achieve the maximum value, $d^2_{\min} = 2$, we would need to choose $s_0(t) = -s_3(t)$, which produces $E_0 = E_3$ and also $s_1(t) = -s_2(t)$, which produces $E_1 = E_2$. However, from (4.2-9), $E_0 + E_3 = E_1 + E_2$ and, thus, $E_0 = E_1 = E_2 = E_3 = E$, i.e., all signals have equal energy. Therefore, we conclude that *for memory one, an unequal energy signal set necessarily results in a value of* $d^2_{\min} < 2$.

For arbitrary memory, ν, by a straightforward extension of the procedure for memory one, it can be shown that the distance between any pair of length $\nu + 1$ paths beginning and ending at the same node (i.e., pairwise error events) is, analogous to (4.2-12), given by

$$d^2 \triangleq 1 - \frac{\mathrm{Re}\left\{\int_0^T s_0^*(t)\, s_{2^{\nu+1}-1}(t)\, dt\right\}}{(E_0 + E_{2^{\nu+1}-1})/2} \tag{4.2-17}$$

Thus, to achieve the maximum value, $d^2_{\min} = 2$, we would need to choose $s_0(t) = -s_{2^{\nu+1}-1}(t)$, which produces $E_0 = E_{2^{\nu+1}-1}$. However, in view of the other forms [analogous to (4.2-16)] that (4.2-17) can be expressed as, it can also be shown that achieving $d^2_{\min} = 2$ also requires choosing $s_i(t) = -s_{2^{\nu+1}-1-i}(t), i = 1, 2, \cdots, 2^\nu - 1$, which produces $E_i = E_{2^{\nu+1}-1-i}, i = 1, 2, \cdots, 2^\nu - 1$. Finally, using the energy conditions in (4.2-12), we arrive at the fact that $d^2_{\min} = 2$ can only be achieved when $E_0 = E_1 = E_2 = \cdots = E_{2^{\nu+1}-1} = E$, i.e., all signals have equal energy. Thus, we conclude that *for arbitrary memory, an unequal energy signal set necessarily results in a value of* $d^2_{\min} < 2$.

References

[1] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. IT-13, no. 2, pp. 260–269, April 1967.

[2] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, New York: McGraw-Hill, Inc., 1979.

[3] Q. Li and B. E. Rimoldi, "Bandwidth-efficient trellis-coded modulation scheme with prescribed decoding delay," *International Symposium on Information Theory*, Ulm, Germany, June 29–July 4, 1997.

[4] G. D. Forney, Jr., "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.

[5] B. E. Rimoldi, "A decomposition approach to CPM," *IEEE Transactions on Information Theory*, vol. IT-34, no. 3, pp. 260–270, May 1988.

[6] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Upper Saddle River, New Jersey: Prentice Hall, 1995.

[7] M. K. Simon, "A generalization of MSK-type signaling based upon input data symbol pulse shaping," *IEEE Transactions on Communications*, vol. COM-24, no. 8, pp. 845–856, August l976.

[8] F. Amoroso, "Pulse and spectrum manipulation in the minimum (frequency) shift keying (MSK) format," *EEE Transactions on Communications*, vol. COM-24, no. 3, pp. 381–384, March 1976.

Chapter 5
Strictly Bandlimited Modulations with Large Envelope Fluctuation (Nyquist Signaling)

Nyquist signaling schemes, which by the very nature of their construction are strictly bandlimited, clearly result in the most bandwidth-efficient modulations of all the ones considered previously in this monograph; however, they also result in modulations with the largest envelope fluctuation. Since the theory of Nyquist signaling is well documented in many textbooks on digital communications, e.g., [1–3], we shall present here only a brief summary of the basic principles simply as a matter of completeness. Although most of the discussion will be focussed on single-channel binary signaling, the extension to multilevel and quadrature signaling schemes such as QAM will be immediately obvious and will receive a brief treatment.

5.1 Binary Nyquist Signaling

The Nyquist criterion is a condition imposed on a waveform that results in zero ISI when a sequence of such waveforms amplitude-modulated by the data is sequentially transmitted at a fixed data rate. Specifically, a binary Nyquist signal is one whose underlying pulse shape, $p(t)$, has uniform samples taken at the bit rate, $1/T_b$ (i.e., herein referred to as the Nyquist rate), that satisfy

$$p_n = p(nT_b) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)\, e^{j\omega n T_b} d\omega = \delta_n = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \tag{5.1-1}$$

Since the Nyquist criterion is derived based on the sampling theorem, the signals to which it is applied are inherently strictly bandlimited. To see this, we proceed as follows:

The integral in (5.1-1) can be written in terms of a partition of adjacent radian frequency intervals of width $2\pi\left(1/T_b\right) = 2\pi\left(2W\right)$, viz.,

$$p_n = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\int_{(\pi/T_b)(2k-1)}^{(\pi/T_b)(2k+1)} P\left(\omega\right)e^{j\omega nT_b}d\omega = \delta_n \tag{5.1-2}$$

Using the change of variables $v = \omega - 2k\pi/T_b$, (5.1-2) becomes

$$\begin{aligned} p_n &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\int_{-(\pi/T_b)}^{(\pi/T_b)} P\left(v + \frac{2k\pi}{T_b}\right)e^{jnT_b(v+[2k\pi/T_b])}dv \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\int_{-(\pi/T_b)}^{(\pi/T_b)} P\left(v + \frac{2k\pi}{T_b}\right)e^{jvnT_b}dv \\ &= \frac{1}{2\pi}\int_{-(\pi/T_b)}^{(\pi/T_b)}\sum_{k=-\infty}^{\infty} P\left(v + \frac{2k\pi}{T_b}\right)e^{jvnT_b}dv \end{aligned} \tag{5.1-3}$$

Next, define the equivalent Nyquist channel characteristic

$$P_{eq}\left(\omega\right) = \begin{cases} \sum\limits_{k=-\infty}^{\infty} P\left(\omega + \dfrac{2k\pi}{T_b}\right), & |\omega| \le \dfrac{\pi}{T_b} \\ 0, & \text{otherwise} \end{cases} \tag{5.1-4}$$

i.e., all of the translates of $P\left(\omega\right)$ folded into the interval $\left(-\pi/T_b, \pi/T_b\right)$ and superimposed on each other. Substituting (5.1-4) into (5.1-3) gives

$$p_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_{eq}\left(\omega\right)e^{j\omega nT_b}d\omega \tag{5.1-5}$$

But the inverse Fourier transform of $P_{eq}\left(\omega\right)$ is, by definition,

$$p_{eq}\left(t\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_{eq}\left(\omega\right)e^{j\omega t}d\omega = \frac{1}{2\pi}\int_{-(\pi/T_b)}^{(\pi/T_b)} P_{eq}\left(\omega\right)e^{j\omega t}d\omega \tag{5.1-6}$$

Thus, from (5.1-5) and (5.1-6), we see that the Nyquist rate samples of $p(t)$, namely, p_n, are also the Nyquist rate samples of $p_{eq}(t)$. Since, by the definition of (5.1-4), $P_{eq}(\omega)$ is a strictly bandlimited function on the interval $(-\pi/T_b, \pi/T_b) = (-2\pi W, 2\pi W)$, then, from the sampling theorem,

$$P_{eq}(\omega) = \begin{cases} \dfrac{1}{2W} \sum\limits_{n=-\infty}^{\infty} p\left(\dfrac{n}{2W}\right) \exp\left(-jn\dfrac{\omega}{2W}\right), & |\omega| \leq 2\pi W \\ 0, & \text{otherwise} \end{cases} \tag{5.1-7a}$$

or equivalently

$$P_{eq}(\omega) = \begin{cases} T_b \sum\limits_{n=-\infty}^{\infty} p_n \exp(-jn\omega T_b), & |\omega| \leq \dfrac{\pi}{T_b} \\ 0, & \text{otherwise} \end{cases} \tag{5.1-7b}$$

However, since for zero ISI we require $p_n = \delta_n$, then (5.1-7b) simplifies to

$$P_{eq}(\omega) = \begin{cases} T_b, & |\omega| \leq \dfrac{\pi}{T_b} \\ 0, & \text{otherwise} \end{cases} \tag{5.1-8}$$

i.e., the equivalent Nyquist channel characteristic is an ideal brick wall filter. Finally, combining (5.1-4) and (5.1-8), we see that the Nyquist channel (Fourier transform of the Nyquist pulse) $P(\omega)$ must satisfy

$$\sum_{k=-\infty}^{\infty} P\left(\omega + \frac{2k\pi}{T_b}\right) = T_b, \quad |\omega| \leq \frac{\pi}{T_b} \tag{5.1-9}$$

i.e., the superposition of all the translates of $P(\omega)$ must yield a flat spectrum in the Nyquist bandwidth $(-\pi/T_b, \pi/T_b)$. It can also be shown that the superposition of all the translates of $P(\omega)$ must yield a flat spectrum in the interval $((2k-1)\pi/T_b, (2k+1)\pi/T_b)$ for any k. Thus, combining the equation that would result from this fact with (5.1-9) gives

$$\sum_{k=-\infty}^{\infty} P\left(\omega + \frac{2k\pi}{T_b}\right) = T_b \tag{5.1-10}$$

for all ω. Note that the zero ISI criterion does not uniquely specify the pulse shape spectrum $P(\omega)$ unless its bandwidth happens to be limited to $(-\pi/T_b, \pi/T_b)$, in which case, it must itself be flat, since the sum in (5.1-10) reduces to a single term, namely, $k = 0$. The implication of this statement is (as we shall soon see) that there are many $P(\omega)$'s that satisfy the zero ISI condition.

Consider now a system transmitting a baseband signal of the form $s(t) = \sqrt{P}\sum_{n=-\infty}^{\infty} a_n p(t - nT_b)$ where $p(t)$ satisfies the Nyquist condition and $\{a_n\}$ are binary (± 1) symbols. Then, based on the above, the minimum lowpass, single-sided bandwidth needed to transmit this signal at rate $R = 1/T_b$ without ISI is $R/2 = 1/2T_b$. Such transmissions occur when the equivalent channel $P_{eq}(\omega)$ has a rectangular transfer function or equivalently

$$p_{eq}(t) = p(t) = \frac{\sin \dfrac{\pi t}{T_b}}{\dfrac{\pi t}{T_b}} \tag{5.1-11}$$

When the binary symbols are independent and the noise samples (spaced T_b s apart) are uncorrelated, each symbol can be recovered without resorting to past history of the waveform, i.e., with a zero memory receiver.

Since, in the above case, R b/s are transmitted without ISI over a baseband bandwidth $R/2$ hertz, then the throughput efficiency is R (b/s)/($R/2$) hertz = 2 (b/s)/hertz. To achieve this efficiency, one must generate the sin x/x pulse shape of (5.1-11), which, in theory, is a noncausal function and extends from $-\infty$ to ∞. This pulse shape is additionally impractical because of its very slowly decreasing tail, which will cause excessive ISI if any perturbations from the ideal sampling instants should occur. Stated another way, the price paid for the extreme bandwidth efficiency achieved with this Nyquist pulse is a large variation in the instantaneous amplitude of the pulse, resulting in a high sensitivity to timing (sampling instant) offset.

To reduce this sensitivity, one employs more practical shapes for $p(t)$, whose Fourier transforms, $P(\omega)$, have smoother transitions at the edges of the band, yet still satisfy the Nyquist condition, thereby resulting in zero ISI. As a consequence, these waveforms will not achieve the minimum Nyquist bandwidth, as we shall see momentarily. The raised cosine transfer function

$$P(\omega) = \begin{cases} T_b, & 0 \le |\omega| \le \dfrac{\pi}{T_b}(1-\alpha) \\ T\cos^2\left\{\dfrac{\pi}{4\alpha}\left[\dfrac{|\omega| T_b}{\pi} - 1 + \alpha\right]\right\}, & \dfrac{\pi}{T_b}(1-\alpha) \le |\omega| \le \dfrac{\pi}{T_b}(1+\alpha) \\ 0, & \dfrac{\pi}{T_b}(1+\alpha) \le |\omega| \le \infty \end{cases} \tag{5.1-12}$$

with excess bandwidth $\alpha R/2\,(0 \leq \alpha \leq 1)$ (see Fig. 5-1a) satisfies the Nyquist criterion and has a pulse shape whose tails decrease faster than the $\sin x/x$ function, i.e., they are the product of $\sin x/x$ and $\cos(\pi\alpha t/T_b)/[1-(2\alpha t/T_b)^2]$ [see Fig. 5-1(b)]. Note that these pulses are still noncausal and extend from $-\infty$ to ∞—properties that are a direct consequence of the strict bandlimitation of the Nyquist formulation. Since the bandwidth of this class of Nyquist pulses is $R/2\,(1+\alpha)$, the price paid for improved sensitivity to timing jitter is a reduction of the throughput efficiency to $R/[R/2\ (1+\alpha)] = 2/(1+\alpha)$. Ideally (perfect sampling), the error probability of all binary Nyquist signaling schemes is equivalent to that of ideal binary PSK, as given by (2.6-2).

5.2 Multilevel and Quadrature Nyquist Signaling

To achieve higher throughput efficiencies, one can extend the above notions to multilevel and quadrature signaling schemes. First, since the Nyquist criterion does not impact the choice of levels for the data symbols, one may simply employ an M-ary alphabet for $\{a_n\}$, e.g., $a_n = \pm 1, \pm 3, \cdots, \pm(M-1)$, resulting in a form of M-ary pulse amplitude modulation. Using the raised cosine Nyquist pulse of (5.1-12), the throughput efficiency is increased to $2\log_2 M/(1+\alpha)$. If now one modulates independent Nyquist signals on I and Q carriers, resulting in a form of pulse-shaped M^2-QAM results, the throughput is further increased to $4\log_2 M/(1+\alpha)$. Of course, if one specifically chooses $M = 4$, what results is Nyquist-pulse-shaped QPSK.

References

[1] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Upper Saddle River, New Jersey: Prentice Hall, 1995.

[2] J. Proakis, *Digital Communications*, 3rd edition, New York: McGraw-Hill, 1995.

[3] E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd edition, Boston, Massachusetts: Kluwer Academic Publishers, 1994.

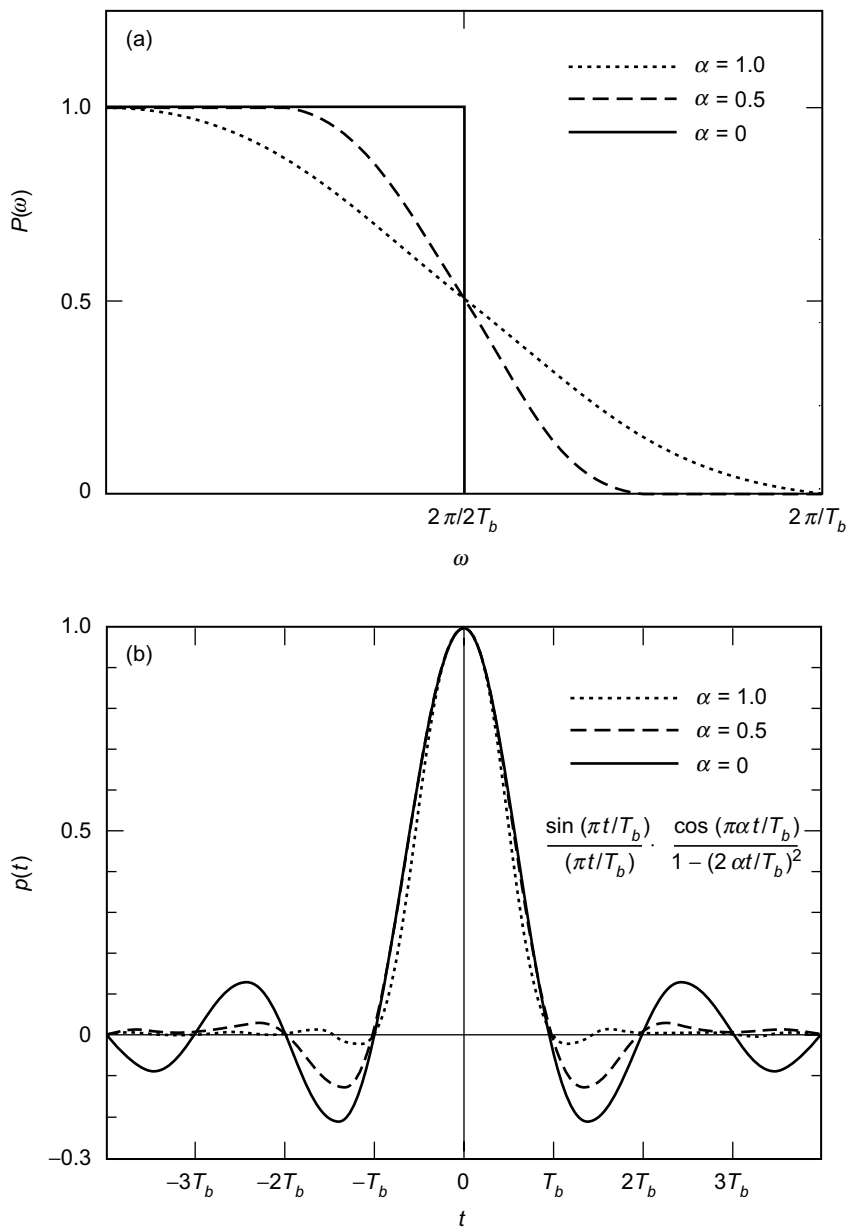


Fig. 5-1. The raised cosine pulse: (a) frequency function and (b) time function.

Chapter 6
Summary

This monograph has introduced and discussed a number of different bandwidth-efficient modulation schemes, in each case emphasizing the trade-off between their amount of envelope (or instantaneous amplitude) fluctuation and their bandwidth efficiency. While not specifically focused upon, the trade-off between power and bandwidth efficiency is also of importance. One means of illustrating this trade-off is via a plot of throughput efficiency (or its reciprocal) versus E_b/N_0 required to achieve a given error probability. In the next section, we offer such plots, obtained from a combination of simulation and analysis for many of the modulations (with and without error correction coding) discussed earlier. The measure of spectral containment used to arrive at the throughput is the 99 percent in-band power, which is equivalent to the -20-dB crossing on a fractional out-of-band power chart. Both unfiltered and filtered cases will be considered, the latter being of interest when the need arises to further restrict the transmitted RF bandwidth beyond that inherently achieved by the generic modulation technique.

6.1 Throughput Performance Comparisons

A 3-phase study [1–3] conducted by the CCSDS in response to an action item from the SFCG identified 10 modulations commonly used or planned by space agencies for bandwidth-efficient applications. The 10 modulations so identified were: PCM/PM/NRZ, PCM/PM/Biphase, QPSK, MSK, 8-PSK, BPSK/NRZ, BPSK/Biphase, OQPSK, GMSK, and FQPSK-B. The objective of the study was to compare these modulation methods, using a combination of simulation and analysis in terms of the E_b/N_0 required to maintain the data BEP at a given constant level. For the cases where very low BEPs were required, a concatenated coding scheme (a combination of a rate 1/2, constraint length 7 inner

convolutional code with a Reed-Solomon 223,255 outer block code) was used. Some results for turbo-coded and trellis-coded modulations were also obtained. Nonideal data and system parameters (e.g., data imbalance) were included in the simulation models to make the results appear as realistic as possible. Where filtering was employed, a three-pole Butterworth baseband filter was used. Finally, to simulate the hard-limiting (nonlinear) effect of an SSPA, the simulation model also used the characteristics of the European Space Agency's SSPA operating in full saturation.

Figures 6-1 and 6-2 are illustrations of the reciprocal of the throughput (the ratio of two-sided 99 percent bandwidth for RF transmission to the data rate) versus the E_b/N_0 required to maintain data BEPs of 10^{-3} and 10^{-4}. The following conclusions can be drawn from these numerical results: FQPSK-B delivers the narrowest bandwidth (highest throughput) with reasonable end-to-end loss compared with BPSK/NRZ while GMSK comes in a close second in terms of bandwidth efficiency.[1] At the other extreme, turbo-coded rate 1/3 BPSK/NRZ is the clear choice for achieving power efficiency at the expense of bandwidth that meets the requirements for deep-space applications. Trellis-coded 8-PSK with or without filtering is also an excellent choice for bandwidth efficient operations. Finally, combining the CCSDS-recommended error-correction coding with PCM/PM/NRZ and with BPSK/NRZ are reasonable choices when both power and bandwidth are considerations.

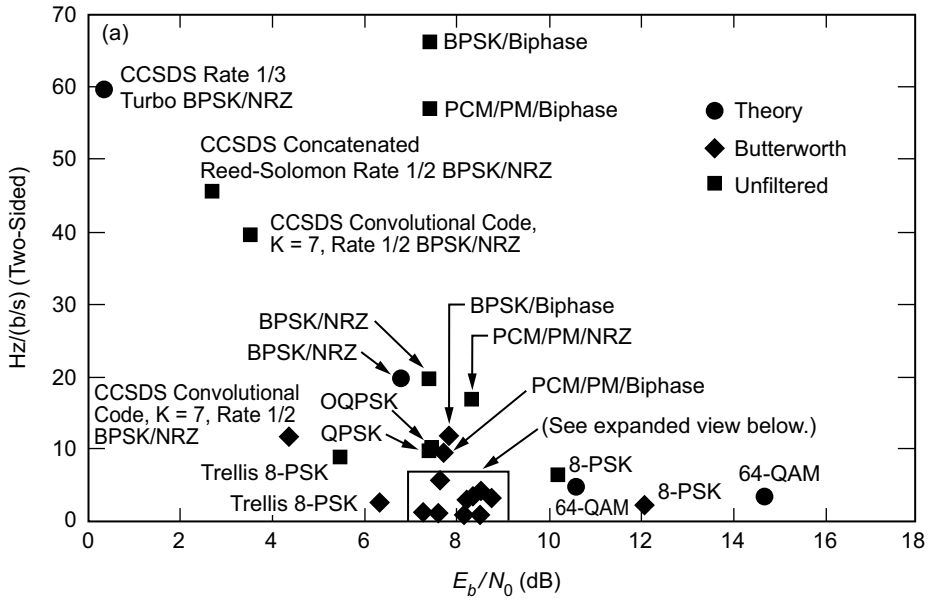
References

[1] W. L. Martin, T-Y. Yan, and L. V. Lam, "CCSDS-SFCG: Efficient modulation methods study at NASA/JPL, Phase 3: End-to end performance," *Proceedings of the SFGC Meeting*, Galveston, Texas, September 16–25, 1997.

[2] W. L. Martin and T-Y. Nguyen, "CCSDS-SFCG: Efficient modulation methods study at NASA/JPL, Phase 1: Bandwidth Utilization," *Proceedings of the SFGC Meeting*, Ottawa, Canada, October 13–21, 1993.

[3] W. L. Martin and T-Y. Nguyen, "CCSDS-SFCG: Efficient modulation methods study at NASA/JPL, Phase 2: Spectrum Shaping," *Proceedings of the SFGC Meeting*, Rothenburg, Germany, September 14–23, 1994.

[1] The demodulator used for GMSK was that based on the AMP representation as discussed in Sec. 2.8.2.6, i.e., a matched filter followed by a Wiener filter.



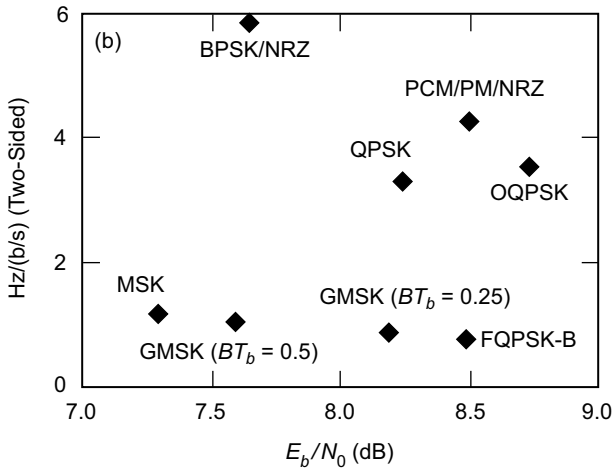


Fig. 6-1. The power-bandwidth trade-off at bit-error probability = 10^{-3}. Power = 99 percent and □□□ = 2, unless otherwise specified: (a) full view and (b) expanded view of the box in (a).

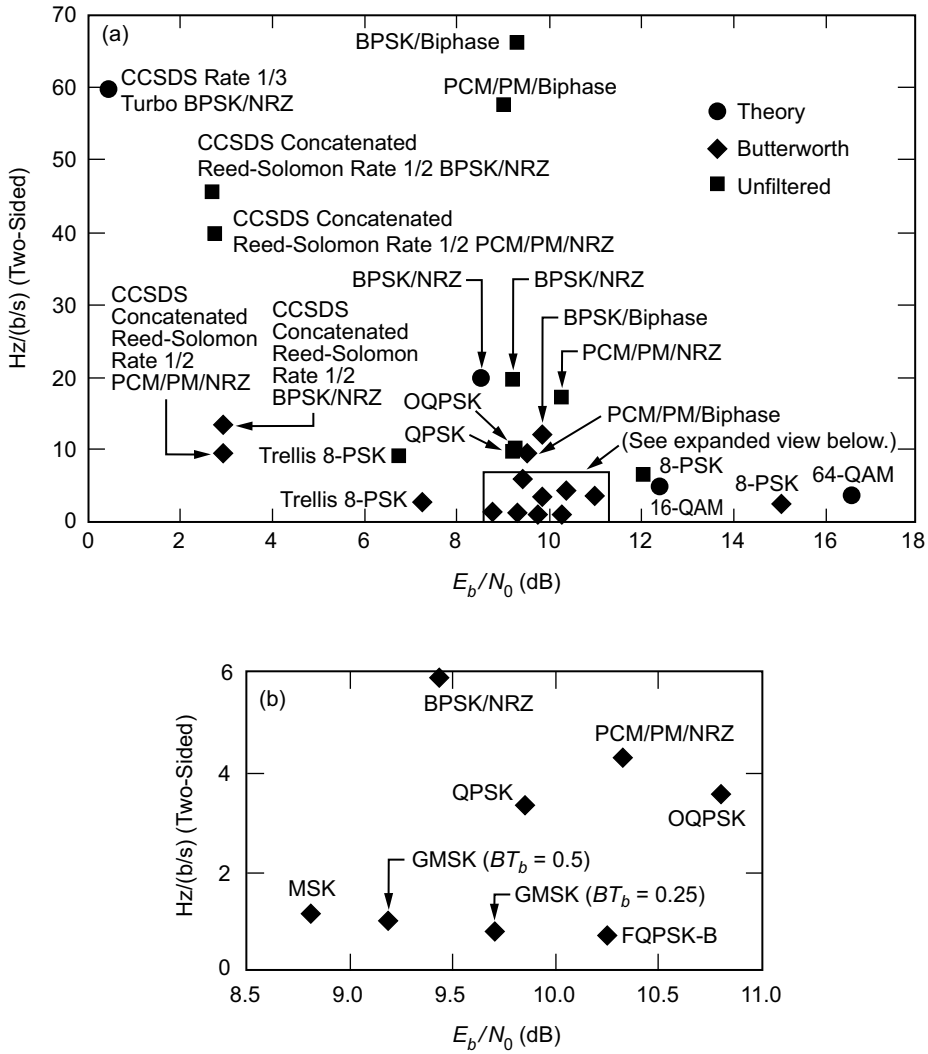


Fig. 6-2. The power-bandwidth trade-off at bit-error probability = 10^{-4}. Power = 99 percent and □□□ = 2, unless otherwise specified: (a) full view and (b) expanded view of the box in (a).