RICE UNIVERSITY

Power Issues in Communication Systems

by

Nasir Ahmed

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Science

Approved, Thesis Committee:

Dr. Behnaam Aazhang, Chair
J.S. Abercrombie Professor of Electrical
and Computer Engineering

Dr. Joseph Cavallaro
Associate Professor of Electrical and
Computer Engineering

Dr. Edward Knightly
Assistant Professor of Electrical and
Computer Engineering

Houston, Texas

April, 2002

ABSTRACT

Power Issues in Communication Systems

by

Nasir Ahmed

In this work, we are concerned with power efficient communication techniques for wireless mobile devices. In this direction the effects of nonlinear amplifiers in communications systems are first studied, as they tend to consume a large fraction of available battery power. Next, a new information theoretic upper bound is proposed that takes into account power efficiency, and the bound achieving input distributions are analyzed and shown to be strikingly different than those that achieve channel capacity. Next, in order to examine methods to achieve the power efficient input distributions, results on shaping gains are restated, with particular emphasis on non-equiprobable signaling in two dimensions. With slight changes in classical techniques we are able achieve power efficient signaling schemes that reduce the wasted power from RF amplification. Finally, simulation results are shown to verify the savings in wasted energy obtained through the use of these power efficient signaling schemes.

Acknowledgments

I would like to thank my helpful advisor Dr. Behnaam Aazhang for his guidance throughout the course of this work. Also I would like to thank Dr. Ashutosh Sabharwal and the helpful physical layer team members Tarik Muharemovic and Jaber Borran. Finally, I wish to thank my family for their encouragement and support.

Contents

Illustrations

Chapter 1

Introduction

When looking at issues related to power efficient communications, an important area to focus on is the power amplifier. The power amplifier consumes a major fraction of the system power, due to the required transmit power, and the loss due to inefficiencies. These devices are used to provide the necessary power levels to overcome propagation loss and noise effects of the channel, but suffer from major drawbacks such as nonlinear transfer characteristics and inefficiencies. The inefficiency of the amplifier is seen as heat dissipation, something that needs to be minimized.

The objective of this work is to develop communication techniques with high efficiency in order to minimize heat dissipation and lengthen battery life. In chapter 2 of this thesis, issues related to power amplifiers such as their properties and effects on communication systems are explored. In chapter 3, channel capacity is discussed, and a new information theoretic metric similar to channel capacity that takes into account amplifier efficiency is proposed. Chapter 3 concludes with a description of the probability distribution on constellation points needed to obtain power efficient transmission. The implementation of such distributions on the input is the subject of non-equiprobable signaling, which is described in chapter 4. Chapter 5 shows simulation results of the proposed transmission scheme, and chapter 6 contains concluding remarks and a discussion on areas for future work.

Chapter 2

Amplifiers for Communication Systems

With the onset of more computationally expensive algorithms in mobile handsets, these devices will become increasingly in need of battery power. The life of a communications device is directly related to the capacity of the battery. It has been shown that the energy density of a battery doubles roughly every 35 years, whereas for A/D converters performance doubles every 8 years and for digital integrated circuits complexity doubles every two years [1]. This shows that the rate of improvement of battery technology is relatively slow, compared to advances in other areas relevant to communications devices. In addition to battery capacity, battery weight is also a constraint placed on small hand held communication devices, which additionally limits the energy available. For this reason in our study we assume battery technology is a constant, and look at techniques done at the physical layer that can potentially improve the talk time and reduce the heat dissipation of communication devices. The focus in this work will be on power amplifiers, as they are a major consumer of available battery power. In the remainder of this chapter, properties of amplifiers such as efficiency and distortion are discussed, followed by simulation results showing the effects of these devices on communication systems.

2.1 Amplifier Fundamentals

2.1.1 Types of Amplifiers

For communication systems, power amplifiers can be of two types: solid state power amplifiers (SSPA), and travelling wave tube (TWT) amplifiers. SSPA's are generally used for mobile handsets due to their small size. In this work, we consider mobile handset battery usage, and thus limit our focus to the SSPA.

Amplifiers can be further classified as being either narrowband or wideband. A narrowband device typically has a bandwidth up to a several hundred kilohertz, while the wideband type have a bandwidth on the order of several megahertz. To help clarify the distinction, consider a simple example, taken from [14]. In AM radio broadcasting, the receiver is tuned to cover a 15kHz portion of the band, corresponding to the spread of a single radio station. The required bandwidth of the RF amplifier is 15kHz in this case, since the amplifier is tuned to one station at a time. Similar results hold for FM radio, where each station has a 200kHz separation. Such applications use narrowband amplifiers. However, in television, stations are spaced 6MHz apart, and the amplifiers used typically have a bandwidth of 4.5MHz; this is known as a wideband RF amplifier.

2.1.2 Amplifier Properties

In this section important properties of amplifiers are defined, in order to better understand their effects on communication systems. The gain is found by computing the ratio of the output and input powers of the amplifier. These devices also have frequency responses, and the response is said to be flat over the range of frequencies for which the gain is a constant. It is of interest to the designer over which frequencies the response is flat, and the behavior of the response outside this range. Distortion is also an important consideration to the designer and is simply the degree to which the output signal varies from the input signal. Since in general, the amplitude transfer characteristic is never perfectly linear, distortion is always present. Efficiency is the final characteristic worth mentioning. Any additional output power not coming from the input, must be drawn from a DC supply. However, in all amplifiers, part of this DC power is wasted, for no device is perfectly efficient. The efficiency can be simply defined as the ratio of the output power to the power drawn from the DC supply.

2.1.3 Amplifier Distortion

We next look at two different types of distortion that arise from the use of amplifiers. Frequency distortion occurs when the gain at different frequencies vary. The resulting signal seen by the channel can be completely different than what was intended. As previously described, if the frequency content of the input signal is within the flat part of the amplifiers response, then no frequency distortion will occur. Amplitude distortion occurs when the system is operating in the nonlinear region of the amplitude characteristic curve. Once again, the signal seen by the channel will be a different shape than the intended signal as a result of the nonlinearities.

2.2 Typical Amplifier Behavior

Near the end of this chapter, some results regarding amplifier effects in a typical communication system are shown. Prior to analyzing performance in communication scenarios, it is instructive to consider a typical amplifier seen in mobile devices. An examination of a typical amplifier input/output characteristic curve (from [4]) shown in figure 2.1 reveals interesting properties that are worth noting. At small input power levels, observe that linearity can be assumed. As the input power approaches

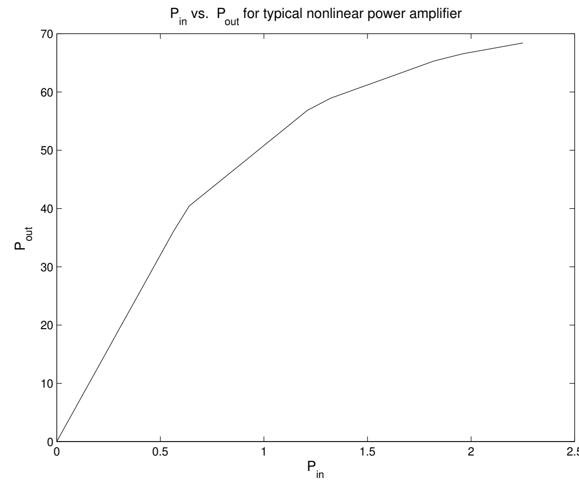

Figure 2.1 : Typical amplifier characteristic curve

the saturation level, the curve becomes extremely nonlinear. An input waveform would experience little distortion for small signal powers, and this distortion would increase for increasing input powers. The warped signal can cause disastrous effects, most notably the introduction of adjacent channel interference.

Amplifier efficiency is also an important figure of merit in determining operating points for communication systems. No electronic device is completely efficient, and this loss is seen as heat dissipation. The efficiency can simply be defined as $\eta = P_{out}/P_{dc}$, where P_{dc} is the power drawn from the battery source. As the input power increases, so does the efficiency, up to the saturation point. For the class AB device shown in figure 2.1, the DC power also is an increasing function of input power, but this is not the case for all power amplifiers. The highly linear class A amplifier, for example, is characterized by a constant DC power drawn from the battery and low efficiency. The fact that efficiency increases as the saturation point is approached is a key point in this thesis and will be used for theoretical analysis in later sections. An inherent tradeoff can be seen, since to increase the average efficiency, more distortion

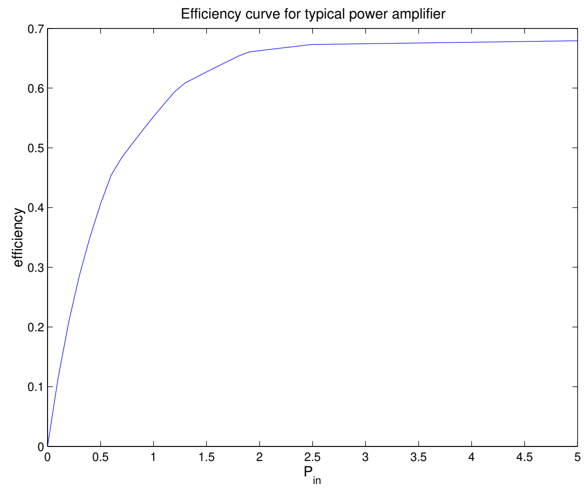

Figure 2.2 : Typical amplifier efficiency curve

must be incurred. For the characteristic curve shown in figure 2.1, figure 2.2 shows the associated efficiency curve. As can be seen, the device actually behaves more efficient in the nonlinear region of operation.

2.3 Amplifier Models

In this section, a typical mathematical model of a power amplifier will be defined that can be helpful in analyzing their effects in communication systems. First, it should be noted that the phase (AM/PM) effects of SSPA's are generally much smaller than those found in TWT amplifiers. It was shown in [6] through simulations that the amplitude characteristics have a more significant effect than the phase, especially when an SSPA is considered. Since the this work is concerned with mobile handsets which typically use SSPA's, the AM/PM effects are assumed to be negligible.

A bandpass modulated signal can be written as

$$s(t) = A(t)\cos(2\pi f_c t + \phi(t)) \tag{2.1}$$

Here $A(t)$ is the envelope of transmitted signal, $\phi(t)$ is the phase and f_c is the carrier frequency. The output of the amplifier can be written as [4]

$$s_{out}(t) = F(A(t))\cos(2\pi f_c t + \phi(t) + \Psi(F(t))) \tag{2.2}$$

This model assumes no memory, and quite often the effects of the phase $\Psi(A(t))$ are neglected. In [5], a polynomial bandpass memoryless model was given for the amplifier as

$$F(A(t)) = \alpha_1 A(t) + \alpha_3 A^3(t) + ... + \alpha_n A^n(t) \tag{2.3}$$

Here n is an odd integer. The coefficients α_i are found by fitting the model to a known (measured) characteristic curve.

2.4 Amplifier Effects

In this section, the effects of a nonlinear amplifier on the bit error performance of a typical communication system are explored. First, we consider the case of M-ary pulse amplitude modulation (M-PAM), which is a modulation scheme that can be represented by one orthonormal basis function. The simulation is of a typical single user communication system, with a power amplifier inserted just after modulation at the transmitter. This device upconverts the signal to a power level necessary for transmission through the channel. In M-PAM, the constellation points are equally spaced in terms of Euclidean distance. For the simulations, initially the constellation is given a small average power so that all the transmitted signals are within the linear region of the power amplifier. The noise power is assumed to be a constant σ^2. Next, the average power of the constellation is increased, and the effects of nonlinearities and noise on the error rate are observed. At the receiver, knowledge of the nonlinear characteristics are assumed, and a minimum distance receiver structure is used to

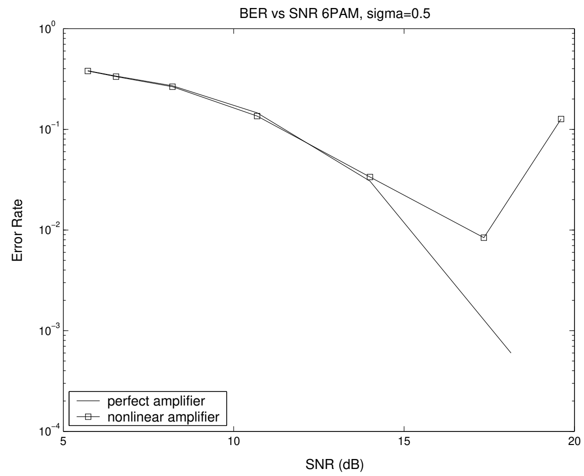

Figure 2.3 : SER vs SNR for 6-PAM signaling

estimate the transmitted symbol. In terms of the channel, we assume the only effects are that of additive white Gaussian noise (AWGN).

Figure 2.3 shows two scenarios, one assuming perfect linearity, and the other using the nonlinear amplifier shown in [4]. For small SNR's, these two curves follow each other closely. However, as the SNR approaches 15dB, the curves diverge. An explanation for this is that before this point, all signal points are operating in the approximately linear region of the amplitude characteristic curve. After this SNR, for the nonlinear case, higher power signal points experience smaller minimum distances with respect to the linear curve. The error rate for the nonlinear case decreases up to a minimum value, then begins to increase. The curve begins to rise once more than one symbol is brought into the saturation region. When this occurs, these symbols cannot be correctly decoded, and the probability of error rises sharply. It can be seen that in terms of error rate, there is a gain in having some symbols operating in the nonlinear portion of the active region. Additionally, with such a configuration, the average power efficiency of the transmission increases due to the nature of power

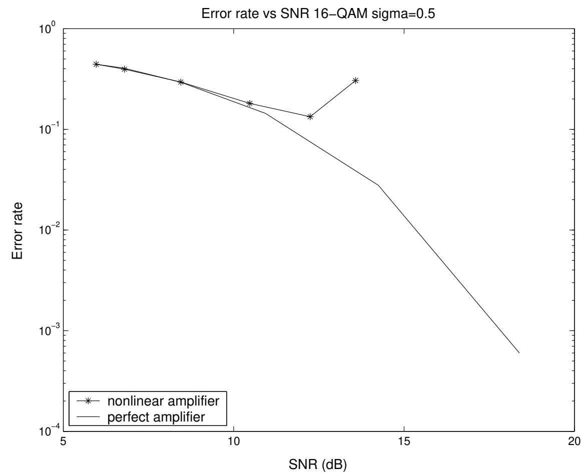

Figure 2.4 : SER vs SNR for 16-QAM signaling

amplifiers. Similar results are shown in figure 2.4 for a 16-QAM transmission system.

2.5 Challenges posed by Amplifier Limitations

Previously, it was shown how amplifier nonlinearities can greatly affect performance in communication systems. Generally, specifications require the amplifier to meet requirements on linearity, output power, gain and efficiency. Linearity is essential to limit adjacent channel interference, as was described previously. Given a modulation scheme, amplifiers are operated in a 'backed off' mode in terms of the peak power available. The degree of backoff is determined by the linearity requirements and the modulation format used. Constant envelope schemes immediately have the advantage that they can operate near the peak efficiency, but they suffer in terms of spectral efficiency. So, in general there is a tradeoff between power efficiency and spectral efficiency.

In this chapter, we first identified the RF amplifiers as a major power consuming device. Next, some basic properties of these devices were discussed. Finally, the effects of amplifiers in communication systems were investigated through simulation. Some interesting properties regarding operating points and nonlinearities were established. In the next chapters, knowledge gained through these simple experiments will be used to look at the problems of battery conservation and heat dissipation.

Chapter 3

Capacity

3.1 Introduction

In the late 1940's, Shannon developed a mathematical theory of communications that set limits on how much information can be reliably sent across a noisy communication channel. The capacity of a noisy communication channel is defined as the maximum bit rate that can be transmitted with arbitrarily small probability of error. In this chapter we will first restate the definition of channel capacity, then propose a modified capacity like definition that is suitable for our purpose of investigating battery consumption and heat dissipation in the power amplifier.

3.2 Channel Capacity

Channel capacity can be defined as the maximum data rate that can be supported by a channel with arbitrarily low probability of error. We will keep our interest to the additive Gaussian channel in this work. The capacity of this channel can be expressed as

$$C = \frac{1}{2} \log_2 (1 + \frac{2E_s}{No}) \tag{3.1}$$

bits per channel use. Here, E_s is the average energy per symbol, and $N_o/2$ is the variance of a zero mean Gaussian random noise variable that is added to the transmitted symbol. This theoretical limit has not been achieved, but in practice it can be approached through the use of error-correcting coding [13]. The capacity is calculated through the maximization of the mutual information between the transmitted and received random variable. This will be further described in the next section.

3.3 Traditional Analysis

A useful channel model for communication systems occurs when their is a discrete input alphabet and a continuous output. Three general methods exist for the calculation of channel capacity. One is to not put any constraint on the input alphabet except possibly average power, and find the distribution that maximizes the mutual information. Another approach is to choose a particular input alphabet, and observe which distribution on this alphabet maximizes the mutual information. This can be taken one step further by assuming a particular distribution on the input alphabet, and finding the associated mutual information that is conveyed by the channel. This is sometimes practical because the distribution of source symbols is quite often uniformly distributed. In this work, we will consider the second scenario, where the input alphabet is well defined, but its distribution is to be determined.

In a Gaussian noise channel where the discrete transmitted random variable is X and the noise variable is N, the continuous received random variable Z can be defined as $Z = X + N$. The mutual information between X and Z is defined as follows

$$I(X;Z) = \sum_{k=0}^{M-1} Q(k) \int_{-\infty}^{+\infty} p(z/x^k) \log_2 \frac{p(z/x^k)}{\sum_{i=0}^{M-1} Q(i) p(z/x^i)} dz \tag{3.2}$$

Here $(x^0, ..., x^{M-1})$ are the M discrete input signals (realizations of X), and $Q(k)$ is the probability of occurrence for symbol x^k. The capacity of this channel is simply the maximization of the mutual information over all possible probability distributions of X

$$C = \max_{Q(0)...Q(M-1)} \sum_{k=0}^{M-1} Q(k) \int_{-\infty}^{+\infty} p(z/x^k) \log_2 \frac{p(z/x^k)}{\sum_{i=0}^{M-1} Q(i) p(z/x^i)} dz \tag{3.3}$$

Once again, $(x^0, ..., x^{M-1})$ are the M discrete input signals, and $Q(k)$ is the probability of occurrence for symbol x^k. In his classic paper on trellis coded modulation (TCM), using equation 3.3, Ungerboeck computed the capacity for various modulation schemes as a function of the signal to noise ratio, assuming a uniform distribution on the input alphabet [12]. The capacity calculation gives us an idea of how many bits per symbol can be transmitted over the channel theoretically with arbitrarily low probability of error. If a transmission rate R is chosen that is greater than capacity, the probability of error approaches one. However, if the transmission rate is lower than R, an arbitrarily small probability of transmission error is possible. Clearly in this equation, the higher the average power of the input alphabet, the higher the resultant capacity. Generalizing equation 3.3, the problem boils down to the maximization of the mutual information subject to an average power constraint on the input alphabet. This can be written as follows

$$\max_{E(X^2) \leq P} I(X;Z) \tag{3.4}$$

Figure 3.1 shows the distribution on signal points achieved for such a maximization. As can be seen, the resultant distribution is that of a truncated Gaussian. This is analogous to the Gaussian capacity achieving distribution for the situation where the input can take on any values (subject to the average power constraint).

3.4 Capacity Analysis with Efficiency Constraint

The traditional analysis of capacity only takes into account average power. Our goal in this work is to look at ways to maximize battery life and reduce heat dissipation in a mobile device. We thus need to modify the maximization shown in equation 3.4 to take into account efficiency, and to see if this leads to a different solution in terms

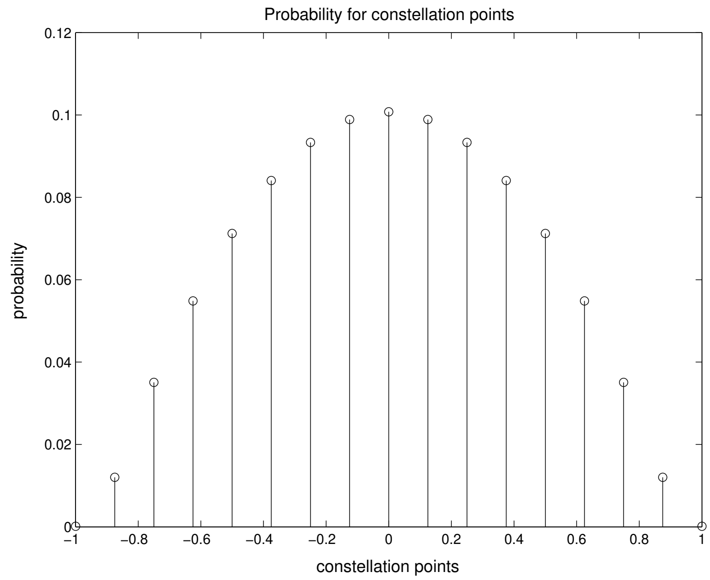

Figure 3.1 : Capacity achieving distribution

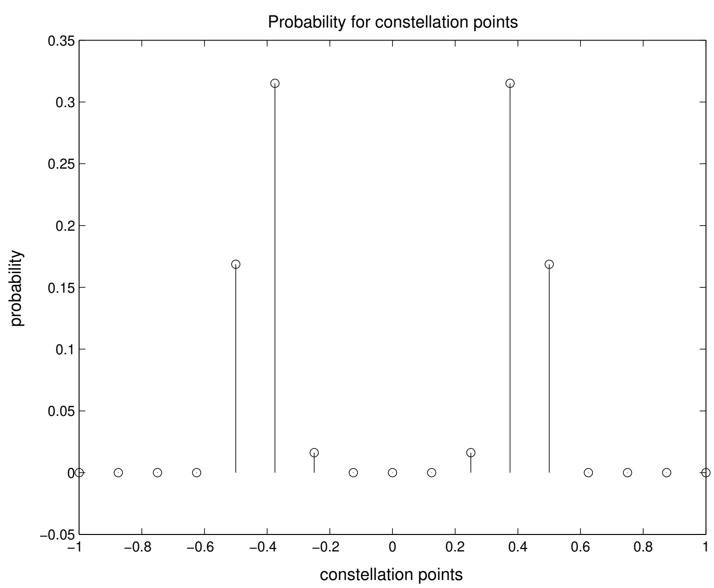

Figure 3.2 : Capacity achieving distribution with efficiency constraint

of the capacity achieving distribution. Our new maximization problem is as follows

$$\max_{\substack{E(X^2)\leq P \\ E(\eta(X))\geq\eta_0}} I(X;Z) \tag{3.5}$$

where P is the upper bound on the average power, and η_0 is the lower bound on the expected efficiency of the transmission scheme. This maximization was done numerically in MATLAB, and the results are shown in figure 3.2. It can be seen that the capacity achieving distribution is much different than that obtained from the traditional analysis. In the truncated Gaussian case, there is a high probability of transmitting the lower power signals. The new capacity achieving distribution has little mass at low powers, and appears to be bi-modal. For the same average power, the distribution in figure 3.2 has a 5% increase in efficiency, while the loss in mutual information is only 0.19% (for a very low SNR used in the simulation). However, as the SNR increases, the loss in mutual information should increase, with the maximum being the entropy difference between the two input distributions. This is a significant result since for the same average power the transmission scheme is more efficient, and as a result it dissipates less power as heat. The power constraint and the efficiency constraint are two opposing forces. The power constraint tends to force the distribution to be large at lower powers, while the efficiency constraint favors higher powered signal points. However, since in general the signals sent to the modulator are from an equiprobable source, the resultant distribution on the transmitted signal points will be uniform. This analysis suggests modulation in the phase shift keying (PSK) formats should be useful in terms of power efficiency at the expense of spectral efficiency. In order to design spectrally efficient modulation schemes, non-equiprobable signaling needs to be considered. Non-equiprobable signaling is required to obtain a power efficient transmission scheme with high spectral efficiency, and in the next chapter methods to achieve such non-uniform distributions will be explored. Another point to note about the resultant distribution is that since it is not a truncated Gaussian, it will not maximize capacity in the traditional sense. The plot in figure 3.1 maximizes the capacity, and is far different than the plot in figure 3.2. However, our goal in this work is to look at power efficiency and battery life, so potential loss in terms of achievable rate is allowable for a gain in efficiency of the transmission scheme.

Chapter 4

Non-equiprobable Signaling

4.1 Introduction

Through the theory of Shannon, it can be shown that there is a 9dB gap between the capacity of the ideal bandlimited Gaussian channel and QAM signaling formats. To make up this 9dB gab, two methods are generally employed. Coding gain is the gain that can be obtained through the use of error correcting coding. The second type of gain, shaping gain, occurs by changing the shape of the constellation. More specifically, shaping gain refers to the reduction in average power through the use of a constellation shape that is not an N-cube. As the shape of the constellation becomes more spherical, a non-equiprobable distribution is forced upon the resultant two-dimensional projection of this N-dimensional constellation. This is a relevant result for situations such as QAM, where information is sent in sequences of two dimensional coordinates. In the limit, as the number of dimensions becomes infinite, the N-sphere has a 1.53dB gain over the N cube. The resultant two dimensional distribution will be that of a truncated Gaussian, similar to the distributions seen in the previous chapter. Traditionally, the way to achieve shaping gains was through the use of multi-dimensional constellations. However, recent work in [8] proposed a new method to force this truncated Gaussian distribution on the constituent two dimensional constellation. The idea was to use what is known as a shaping code to force a non-uniform distribution on the constellation points. The approach was more direct in the sense it used a lower dimensional constellation and in fact was able to achieve the theoretical 1.53dB shaping gain in any fixed dimension.

Before restating the results in [8], some terms need to be defined. The authors in [8] made use of what is known as the continuous approximation. The continuous approximation states that the average power of a constellation whose points lie in a region $\Re$ can be approximated by the average power $P(\Re)$ of a continuous distribution uniform in $\Re$ and zero everywhere else. In two dimensions, this result implies that the power of region $\Re$, $P(\Re)$, is proportional to the regions volume, $V(\Re)$. Another useful performance parameter is the peak to average power ratio (PAPR). For peak power limited systems, this ratio should be minimized to maximize the power efficiency of the transmission scheme. It can be shown that for the N-sphere, the PAPR is $(N+2)/2$.

4.2 Non-equiprobable Signaling

Shaping gains refer to the reduction in average power obtained by the choice of signaling scheme and the design of the constellation. Biasing gains, introduced in [8], refers to the reduction in average power from transmitting the constellation symbols with non-uniform probabilities. In [8], the following equality was shown

$$\gamma_s(S^N)\gamma_N = \frac{\pi e}{6} \tag{4.1}$$

which, in decibels, is 1.53dB. Here, $\gamma_S(S^N)$ is the shaping gain of the N-sphere, and γ_N is the maximum biasing gain in N dimensions. Thus, even in two dimensions, there exists a spherical constellation with a biasing scheme that can achieve the 1.53dB shaping gain. This result shows that shaping gains could be achieved without using higher dimensional constellations. The following is a restatement of the problem setup

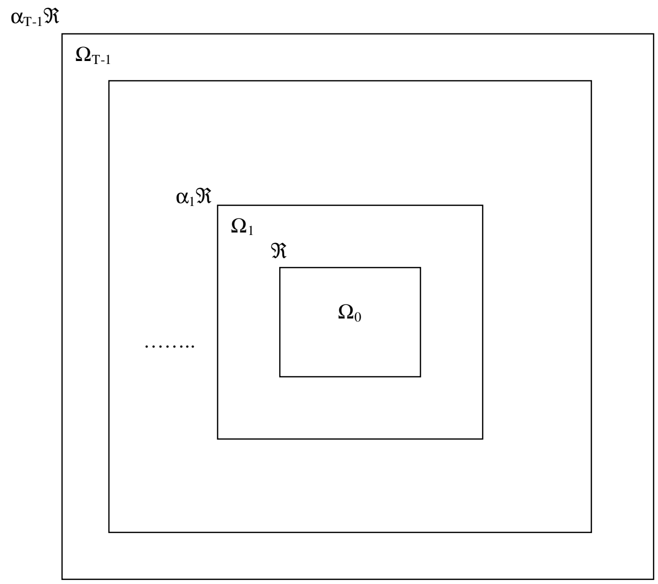

Figure 4.1 : Decomposition into Subconstellations

in [8] and [10], and is included for reasons apparent in later sections.

Non-equiprobable signaling is achieved by firstly dividing the constellation into T subconstellations, as shown in figure 4.1. The subdivisions are based on the inner most region, denoted the fundamental region $\Re$. As can be seen in the figure, the outer subconstellations are scaled versions of the fundamental region. Denoting the scaled copies of the fundamental regions as $\Re, \alpha_1\Re, ..., \alpha_{T-1}\Re$, the constellation Ω may be redefined as the union of T subconstellations, $\Omega_i, i = 0...T-1$, where $\Omega_i = \Omega \cap (\alpha_{i+1}\Re/\alpha_i\Re)$. Each subconstellation Ω_i is selected with probability f_i, and within each subconstellation, the symbols are used with equal probability. In [8], results were obtained when each subconstellation was of equal size. This constraint was relaxed in [10] by Livingston, where biasing gains were calculated for variable size regions. Although the work of Livingston allowed for variable size regions, results were not derived for scenarios in which subconstellations of decreasing size were required.

4.3 Generalization of Results

Next, the results of Livingston are generalized to allow for regions of decreasing size. We know that subconstellation i is chosen with frequency f_i, so given that the average power in subconstellation Ω_i is P_i, the average power of the non-equiprobable signaling technique can be defined as

$$P_{avg} = \sum_{i=0}^{T-1} f_i P_i \tag{4.2}$$

Let us next define

$$k_i = \frac{V(\alpha_i \Re) - V(\alpha_{i-1} \Re)}{V(\Re)} = \frac{|\Omega_i|}{|\Omega_0|} \tag{4.3}$$

In equation 4.3, $V(\alpha_i \Re) - V(\alpha_{i-1} \Re)$ is the volume of subregion i. Lets now define t_i so that $P_i = t_i P_0$, which leads to

$$P_{avg} = \sum_{i=0}^{T-1} t_i f_i P_0 \tag{4.4}$$

Here, P_0 is the average power of region Ω_0. To the accuracy of the continuous approximation, it can be shown that

$$\frac{\sum_{i=0}^{T-1} t_i P_0}{\sum_{i=0}^{T-1} k_i} = \sum_{i=0}^{T-1} k_i P_0 \tag{4.5}$$

since $V(\alpha_{T-1} \Re) = \sum_{i=0}^{T-1} k_i V(\Re)$ and $P(\alpha_i \Re) \propto V(\alpha_i \Re)$. Solving for t_j gives

$$t_j = \frac{(K_j^2 - K_{j-1}^2)}{k_j} \tag{4.6}$$

where $K_j = \sum_{i=0}^{j} k_i$. The total average power can then be expressed as

$$P_{avg} = [\sum_{i=0}^{T-1} (K_i^2 - K_{i-1}^2) f_i / k_i] P_0 \tag{4.7}$$

Using this construction, we can transmit $H(f_0, ..., f_{T-1}) + \sum_{i=0}^{T-1} f_i \log_2 |\Omega_i|$ bits of information. If this same rate of transfer is used with equiprobable signaling, and the signal constellation is taken from the same underlying lattice, the constellation would need to be of size [8]

$$2^{\log|\Omega_0|+H(f_0,\ldots,f_{T-1})+\sum_{i=0}^{T-1} f_i \log k_i} \tag{4.8}$$

To the accuracy of the continuous approximation, this leads to an average signal power of

$$P' = 2^{H(f_0,\ldots,f_{T-1})+\sum_{i=0}^{T-1} f_i \log k_i} P_0 \tag{4.9}$$

Hence, the biasing gain $\gamma_2 = P'/P_{avg}$ can be expressed as

$$\gamma_2(f_0, \ldots, f_{T-1}) = \frac{2^{H(f_0,\ldots,f_{T-1})+\sum_{i=0}^{T-1} f_i \log k_i}}{\sum_{i=0}^{T-1}(K_i^2 - K_{i-1}^2) f_i/k_i} \tag{4.10}$$

We next wish to find the peak to average power ratio (PAPR) of the constellation. With the use of the continuous approximation, it can be shown that $P(\alpha_{T-1}\Re) = \sum_{i=0}^{T-1} k_i P(\Re)$. The PAPR as a function of the basic region $\Re$ can be expressed as

$$PAPR = \frac{P_{max}(\Re)\sum_{i=0}^{T-1} k_i}{P_{avg}(\Re)\sum_{i=0}^{T-1}(K_i^2 - K_{i-1}^2) f_i/k_i} \tag{4.11}$$

where $P_{max}(\Re)$ is the maximum power in the fundamental region, and $P_{avg}(\Re)$ is the associated average power.

Transmitting at equal rates, the size of the constellation for non-equiprobable signaling is larger than that for uniform signaling. To quantify this difference, the constellation expansion ratio (CER) is derived next. Using the fact that $V(\alpha_{T-1}\Re) = \sum_{i=0}^{T-1} k_i V(\Re)$ and equation 4.8, the CER can be defined as

$$CER = \frac{\sum_{i=0}^{T-1} k_i}{2^{H(f_0,\ldots,f_{T-1})+\sum_{i=0}^{T-1} f_i \log k_i}} \tag{4.12}$$

Clearly, this figure of merit plays a role in the distortion due to nonlinearities and the efficiency of the amplifier.

T	k_0	k_1	k_2	k_3	γ_2	$PAPR$
2	1	3	–	–	0.6247	2.6666
4	1	1.4	2.3	7	0.9675	3.3548
2	1	0.75	–	–	-0.3444	1.8667
4	1	0.75	0.75	1.5	0.2431	2.2069
2	1	5	–	–	0.4846	3
4	1	0.6	1	3	0.6382	2.8

Table 4.1 : Biasing and Shaping Gains for Equiprobable Selection of Subconstellations (T=2, T=4)

4.4 Results

As was done in [10], with regions of increasing area (and equal probabilities of selecting these areas), one can obtain biasing gain with a modest increase in PAPR. As was seen in chapter 3, when the average efficiency constraint of the amplifier is increased, the capacity achieving distributions also change. The result is a distribution with little mass at low power. For this purpose, we propose having some regions smaller than the fundamental region, in order to increase the probability of using the higher powered constellation points. If $f_0 = f_1 =, ..., = f_{T-1}$ then variable size regions can achieve non-equiprobable signaling. This scheme should have drawbacks in that in general, there is an expected gain in total average power compared to the situation with regions of increasing area, but the PAPR should be much closer to unity. The main gain is in power efficiency, as inefficient constellation points are now transmitted with lower probability.

Referring to table 4.1, some interesting results can be observed. The first row of the table shows the best ratios of region sizes for two subconstellations ($T = 2$) as was demonstrated in [10]. These region sizes are optimal in terms of maximizing the biasing gain. However, there is a price to pay in the increased PAPR. The third row has a scheme in which the second region is smaller than the first, giving a high probability of transmitting the high power points. It can be seen in this case the PAPR is actually less than 2, the value for a spherical constellation in two dimensions. Now, suppose that the average power of the shaped constellation is to be kP_0, where P_0 is the average power in the basic region $\Re$. The fifth row shows the results when a two subregion scheme is used, with $k = 4$. If the number of subregions is doubled (still with $k = 4$), the resultant PAPR decreases, as can be seen in row six. From these results it is apparent that the PAPR can be reduced using regions which are smaller than the fundamental region. Additionally, given a value of k, the appropriate amount of subregions (value of T) can be determined to minimize the PAPR, and thus increase the power efficiency. Through experiments, it was determined that as k increased, the optimal amount of subregions converged to $T = 3$.

4.5 Realization of Non-equiprobable Signaling

In this section, practical method of achieving shaping gains are explored. The scheme described next was proposed in [9] and involves partitioning the data into constant length blocks. The result of this technique is that the system transmits an equal number of symbols for each block. Calderbank and Klimesh proposed this transmission method, and make use of balanced codes, which are codes that contain an equal number of ones and zeros. In [9], some examples of balanced codes are given, with simple encoding and decoding mechanisms. Recalling the notations of the previous section, two methods can be used to obtain a non-uniform distribution on the input signals. The first method is to have subregions of equal size, and select these regions with different probabilities. Another method is to have subregions of varying size, and choose these subregions with equal probabilities. This is the method we chose for our experiments. A block diagram of the setup can be seen in figure 4.2. Uncoded bits are sent into the shaping block, and are divided into two streams. The first stream is sent to a shaping coder, which produces the balanced codeword. Based on bits of the codeword, a subregion i, $0 \leq i \leq T-1$ is selected. The second branch in the shaping block selects a point in subregion i, and a serial to parallel converter chooses the appropriate amount of bits to select a point in the subregion. When regions are chosen with equal probability, balanced codewords can be used. Since the subregions are of varying size, the probability distribution on the resultant symbols is non-uniform. Since the balanced codeword is of a fixed size, and the distribution of zeros and ones in each codeword is known, the length of a frame is fixed for this transmission scheme. The decoding of this system can be performed in stages. The first stage is a minimum distance receiver that determines the actual symbol that was transmitted, and hence the bits corresponding to the transmitted symbol. The bits can be broken up into those corresponding to the shaping code, and those that selected points within subregions. The shaping code can then be decoded using the simple mechanism described in [9]. Finally, the decoded shaping bits are combined with the bits corresponding to points in the subregions to form the original frame. A block diagram of the receiver is shown in figure 4.3.

This method could be used in applications where packets are sent. In [10], Livingston describes how such a scheme could be used in ARQ systems. A cyclic redundancy check is applied at the receiver. If a frame or packet error occurs, a retransmission of the entire packet is requested. A single bit error can cause catastrophic

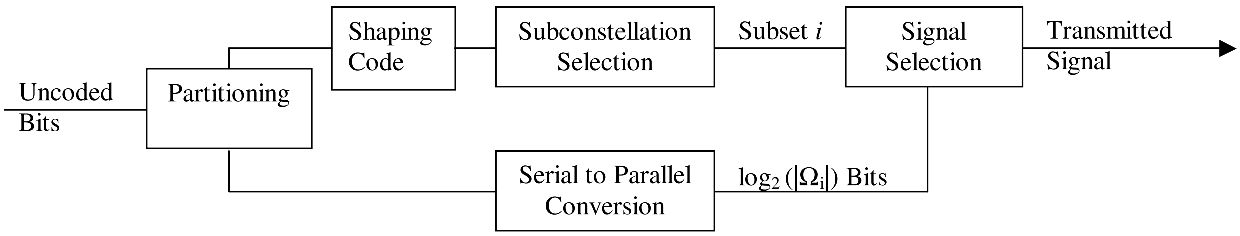

Figure 4.2 : Shaping Code Implementation

decoding for a given frame. However, with the use of a CRC, retransmission can be requested if necessary.

4.6 Other Shaping Code Implementations

It was shown in previous sections how through the use of a shaping code, non-equiprobable signaling could be achieved. Another method exists for achieving this non-uniform distribution, and will be described for completeness. Once again, it is assumed here that we wish to shape the output of a memoryless, binary equiprobable source. An obvious method to induce a non-uniform distribution is to represent constellation points with variable length codewords. Denoting the length of codeword i as l_i, the probability of transmitting codeword i is 2^{-l_i}. If $\sum_i 2^{-l_i} \leq 1$ (Kraft inequality), then a binary prefix code exists with l_i as the length of the codewords. This method could be applied to the subconstellations described previously. Within a given subconstellation, all symbols could have labels of the same length, for they are used with equal probability. However, this method suffers in that it has a variable rate transmission, which can cause error propagation at the decoder. The method using balanced codewords described previously is more desirable in that fixed length

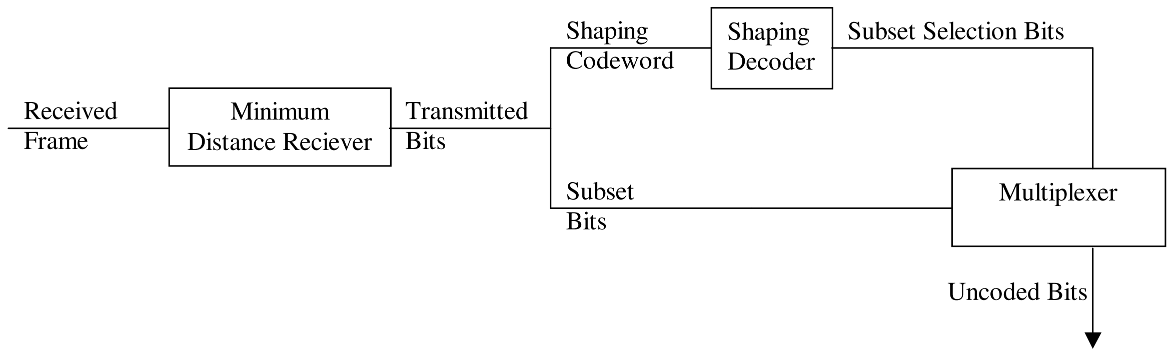

Figure 4.3 : Shaping code receiver structure

frames are used, eliminating the problem of variable rate transmission. For this reason, we use balanced codewords and the partitioning schemes described previously for our constellation shaping.

4.7 Shaping and Coded Modulation

The idea of jointly optimizing channel coding and modulation for performance has been around for many years, and two methods to accomplish this are trellis coded modulation (TCM) proposed by Ungerboeck, and multilevel codes (MLC) suggested by Imai. In both schemes, rather than traditionally optimizing the channel code in a hamming distance sense, it is dealt with in Euclidean space [11].

Ungerboeck's approach involves mapping by set partitioning. The set of constellation points is usually separated to maximize the intra-subset Euclidean distance. To encode the data, a binary convolutional code selects the subset, and the remaining uncoded input bits select a point within the subset. To optimize this scheme, exhaustive computer search is performed to find code parameters that maximize the minimum distance of coded sequences in a Euclidean distance sense [11].

Imai's multilevel coding provides flexible transmission rates because the code rate and signal set size are independent, as compared to TCM where they are intimately connected. Shaping can be incorporated into this coded modulation scheme quite easily, as will be shown next. For more details, the interested reader should refer to [8] and [11].

An L level partition of a signal constellation is a sequence of partitions $\Gamma_0, ..., \Gamma_L$ where the partition at level i is a refinement of the partition at level $i-1$. This partition can be thought of in a tree structure, where the root corresponds to the signal constellation (Γ_0), and the leaf nodes the individual constellation points. At each level, we divide the tree into two halves, and thus each branch in the tree can be given a binary label. A label from the root to the associated leaf node has as distinct binary address, given by $A = (a_1, a_2, ..., a_L)$. Since we are considering an AWGN channel, a useful figure of merit is the intrasubset squared Euclidean distance δ_i, $0 \leq i \leq L$, which can be defined as [8]

$$\delta_i = \max_{\substack{S \varepsilon \Gamma_i \\ a,b \varepsilon S \\ a \neq b}} d(a, b) \tag{4.13}$$

where $d(a, b)$ is the squared Euclidean distance between points a and b. Consider an L level code $C = [C_1, C_2, ..., C_L]$, where C_i are the component codes. In MLC, the codes are applied to the input bits corresponding to the appropriate branch level. For example, C_1 encodes the input bits used to define address a_1. If the minimum Hamming distance of code C_i is d_i, then the minimum Euclidean distance of the code C is [8]

$$d(C) = \min\left(d_1\delta_0, d_2\delta_2, ..., d_L\delta_{L-1}, \delta_L\right) \tag{4.14}$$

This multilevel approach can be integrated with shaping as follows. Consider a scheme with 3 layers, ie $L = 3$. The first two codes C_1, and C_2 can help produce

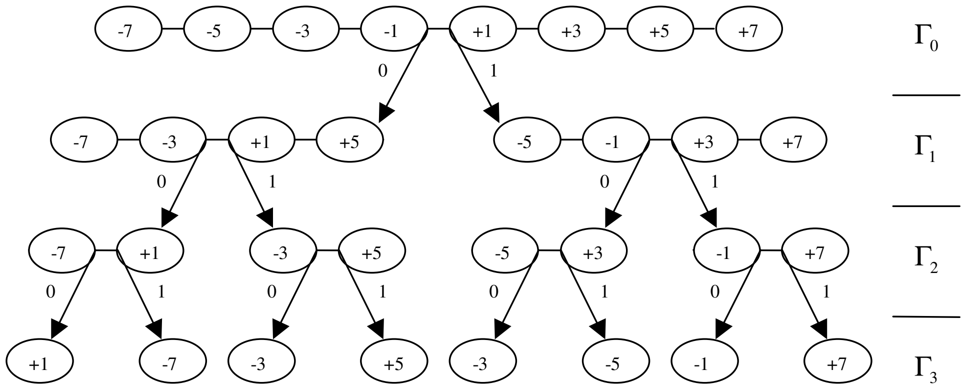

Figure 4.4 : Set Partitioning for Shaping

coding gain over uncoded transmission, while the third level, C_3 can be a shaping code, altering the distribution of constellation points to achieve a desired goal. Either the goal can be the reduction of average power, where a Gaussian type distribution would try to be induced on the signal points, or it would be the bimodal type distribution for power efficiency that was described previously. Ungerboeck's set partitioning method is used at each partitioning level where the goal is to maximize intra-subset Euclidean distance, and thus $\delta_2 \geq \delta_1 \geq \delta_0$. The hamming distance of the shaping code in most cases would not affect the coding gain, due to the high minimum distance at its partitioning level. This fits in naturally with the intuition how coding gain and shaping gain can be performed independently. For the multilevel approach, in general each subset is divided into equal sized subregions. For this reason, it is probably best to do non-equiprobable signaling with equal sized subregions. To induce a non-uniform distribution, select the subregions with different probabilities. A constant weight codeword could be used to accomplish this. In such a scheme, shaping would be done at the lowest level in the partition tree; the shaping code could differentiate between high and low power points. A simple 8-PAM example is shown in figure 4.4, which is similar to that shown in [8]. Ungerboeck's set partitioning is used at all layers, and it can be seen that each of the sets in Γ_2 contain one high power and one low power point. A shaping code can be used to differentiate between these points.

Chapter 5

Simulation Results

5.1 Introduction

Previously, amplifiers and their effects on communication systems were discussed. It was shown that an increase in power at the input of the amplifier produces a more distorted, yet more power efficient situation. Additionally, it was shown how a change in the definition of capacity to incorporate power efficiency brought about a bimodal distribution at the input. Next, shaping was discussed and it was described how it could be used to achieve different types of probability distributions on the constellation points. It was shown that having multiple subregions produced a reduction in PAPR, which is desirable for highly efficient systems. We now wish to study the effect of a constellation that is shaped for high efficiency, and its effect on a system with a nonlinear power amplifier.

5.2 Simulation Setup

The simulation setup is as follows. For the RF power amplifier, the model given by Stark in [4] will be used. The power transfer characteristic and efficiency curve for this amplifier is given in chapter 2. It can be seen that as the input power is increased, the amplifier becomes increasingly nonlinear. Additionally, the power efficiency increases nonlinearly with input power. Previously, it was shown that to transmit at high efficiency, the high powered signals must be sent with increased probability. This will be accomplished using the balanced codewords and non-equiprobable signaling described in chapter 4. In the simulations, a high efficiency scheme will be compared to one with lower efficiency. For these two schemes, results will be shown for power consumption and heat dissipation, the two quantities of interest in this work.

5.3 Results

5.3.1 Class AB Amplifiers

The simulations were performed for a scenario with a degree of back off, so that higher powered constellation points were not in the highly nonlinear region. In terms of performance, since one technique has higher average power than the other, a plot of symbol error rate (SER) versus the energy per transmitted bit to noise ratio (E_b/N_o) would have the higher efficiency curve shifted to the right, indicating inferior performance. However, in our study, maximizing battery life and minimizing heat dissipation are of great importance, so in this regard we wish to look at wasted energy and the power drawn from the battery, not necessarily the power seen by the channel. The metrics for wasted energy and power consumption will be described in the next paragraph. As a side note, the higher average power scheme is more efficient (and is called the high efficiency scheme) because the high power points are transmitted with increased probability, leading to high average transmit power and higher expected efficiency. The efficiency is increased in this case because as was seen in chapter 3, when high power constellation points are transmitted with increased probability, the average efficiency increases.

First, lets consider 16-QAM, a typical modulation scheme in communication systems. As stated earlier, when the low efficiency signaling scheme is compared with

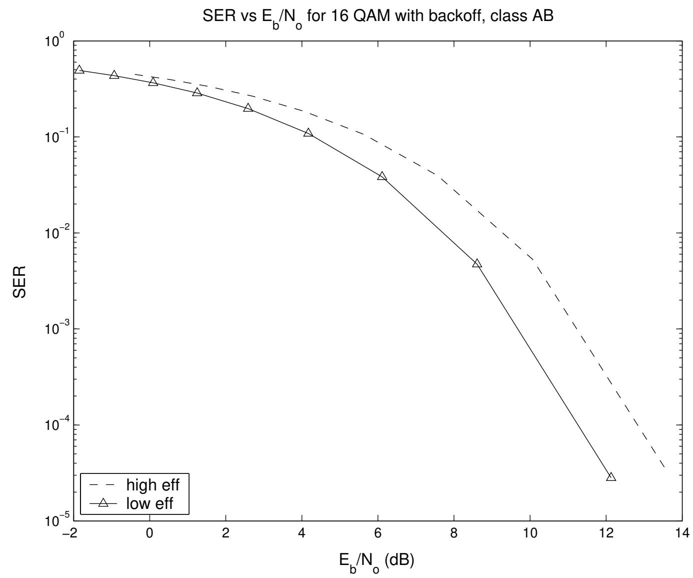

Figure 5.1 : SER vs E_b/N_o shaping code simulation for 16-QAM using class AB amplifier. The higher efficiency scheme has a higher average transmit power.

the higher efficiency technique, the first scheme is expected to have a a smaller E_b/N_o for a given SER. This effect is shown in figure 5.1. It can be seen that the higher efficiency schemes curve is shifted to the right, a result of the higher average energy for the same minimum distance between constellation points. Next, the SER versus the ratio of power per bit drawn from the batteries to the noise (P_{dc}/N_o) is considered. This figure of merit includes not only the power of the signal sent onto the channel, but also the wasted power due to amplifier inefficiencies. One would expect the curves for the low and high efficiency schemes to be closer in value than what appeared on the previous figure (SER vs. E_b/N_o). The reason for this is that as input drive level increases, the amplifiers DC consumption also rises, but not at the same rate. This explains why the wasted power actually decreases as output power increases. For a given minimum distance and noise power, the two schemes should have different average powers as described previously, but their average DC powers

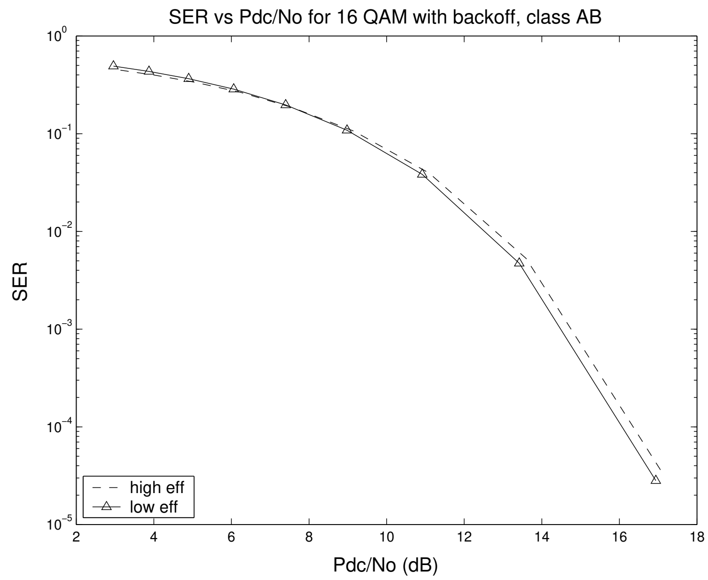

Figure 5.2 : SER vs P_{dc}/N_o shaping code simulation for 16-QAM using class AB amplifier. The two schemes have almost identical battery consumption for a given SER.

should be closer in value. This effect is shown in figure 5.2. Another reason for the decreased distance between the two curves has to do with the nature of the QAM constellation. Signaling with the high powered points has one main advantage, that being fewer nearest neighbors. For the low powered points, each is surrounded by many points separated by the minimum distance of the constellation. However, the corner points of the constellation have fewer neighbors at the minimum distance and as a result, the probability of error should be lower when higher power signals are used with more frequency. This effect will be seen more clearly when the performance of class A amplifiers are looked at later in this chapter.

Finally, for the same constellation, one more figure of merit is considered: probability of symbol error versus the ratio of the wasted power per bit to noise (P_{wasted}/N_o), shown in figure 5.3. It can be seen that for a given probability of error in the high efficiency scheme, less power is wasted (less heat is dissipated) in the mobile hand-

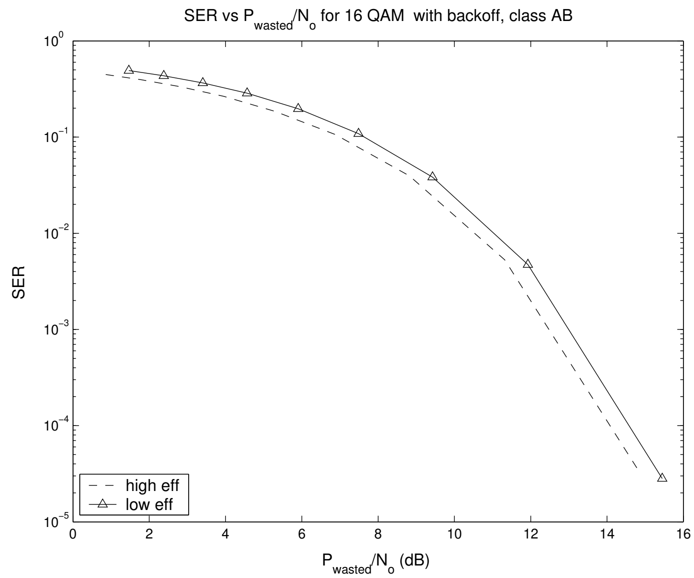

Figure 5.3 : SER vs P_{wasted}/N_o shaping code simulation for 16-QAM using class AB amplifier. The higher efficiency scheme dissipates less power as heat for a given SER.

set, which is a significant result consider both schemes draw approximately the same power from the batteries for a given SER.

5.3.2 Class A Amplifiers

Class A power amplifiers are characterized by constant DC powers over all input drive levels [15]. Additionally, due to the nature of their biasing, they tend to be linear amplifiers, at the expense of low efficiencies. This should immediately bring large savings in terms of wasted power by operating at higher efficiency. We performed similar experiments as described previously but now with a class A power amplifier. Once again, the SER vs E_b/N_o curve demonstrates poor performance for the high efficiency scheme (seen in figure 5.4) because of the difference in average powers for the same minimum distance. Looking at the SER vs P_{dc}/N_o curves in figure 5.5, for the high efficiency scheme it is evident that the battery power used is actually less

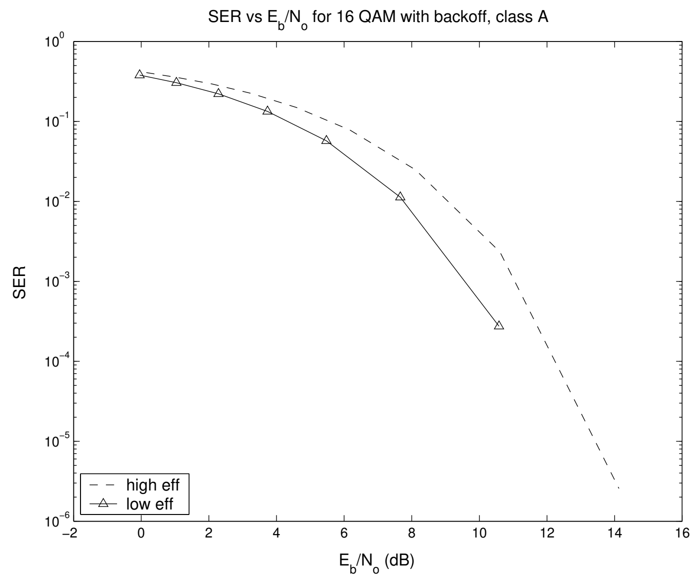

Figure 5.4 : SER vs E_b/N_o shaping code simulation for 16-QAM using class A amplifier. The higher efficiency scheme has a higher average transmit power.

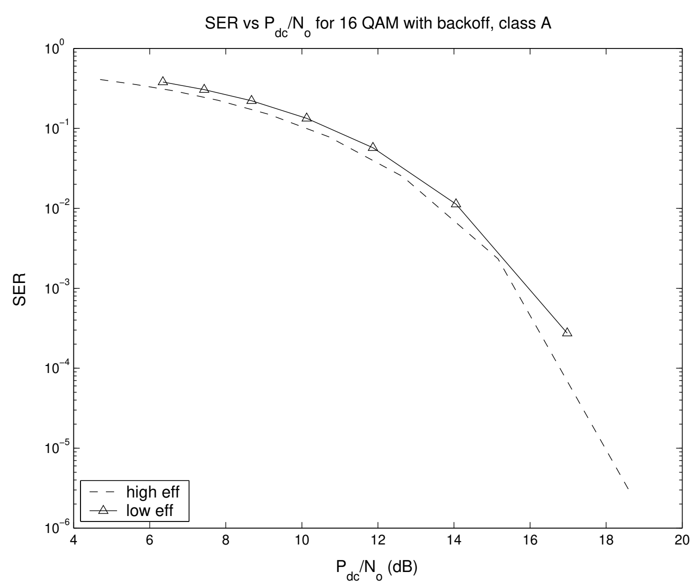

Figure 5.5 : SER vs P_{dc}/N_o shaping code simulation for 16-QAM using class A amplifier. The higher efficiency scheme consumes less DC power from the battery for a given SER.

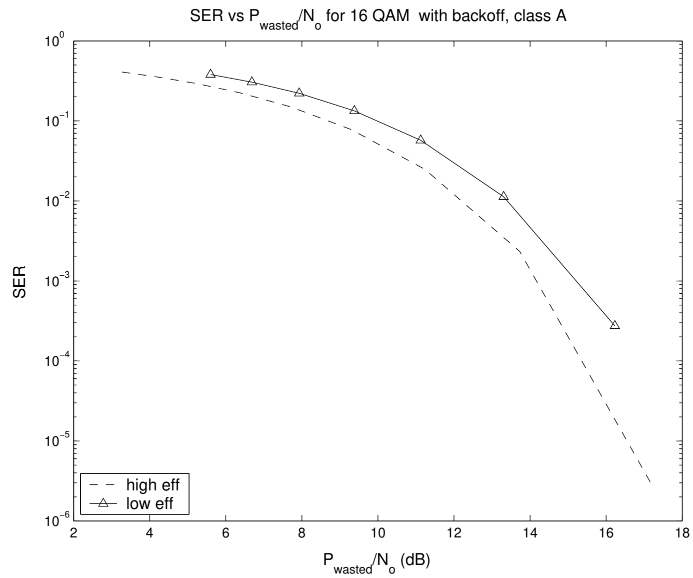

Figure 5.6 : SER vs P_{wasted}/N_o shaping code simulation for 16-QAM using class A amplifier. The higher efficiency scheme dissipates less power as heat for a given SER.

for a given SER. The reason for this is the drain from the batteries is a constant, and using the higher powered constellation points results in less nearest neighbors on the average and a smaller probability of error. Finally, the SER versus P_{wasted}/N_o is shown in figure 5.6 and demonstrates how the higher efficiency signaling scheme offers savings in terms of wasted power per bit for a given SER. It can be seen that their is a gain in terms of energy drained from the batteries in using high efficiency schemes, and this gain is most evident in the class A amplifier. The class AB model used has an increasing power requirement from the battery as the input power to the amplifier increases, but this increase is not as rapid as the change in output power. And because of the nearest neighbor effect, the low and high efficiency schemes can have similar performance in terms of drained power (this is only true when the two schemes have average powers that are relatively close). However, for the high efficiency scheme, wasted power is less, resulting in less dissipated heat.

In this chapter, an end-to-end communications system with non-equiprobable signaling and a nonlinear amplifier was simulated. It was observed that higher efficiency schemes suffer in terms of transmitted energy, but this loss is not as severe when power consumption from the batteries is considered. In fact for class AB, the power drained for a given SER was approximately equal for the high and low efficiency schemes. Finally, it was shown that for class A amplifiers, high efficiency schemes actually perform better in terms of reduced power consumption and heat dissipation.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

We have studied the effects of nonlinear amplifiers in communications systems. Firstly, it was shown how these nonlinearities can reduce system performance in terms of bit error probability as a result of a decrease in the minimum distance between points in the constellation. Next, a new mutual information maximization with the addition of a power efficiency constraint on the system was proposed. This maximization produced strikingly different capacity achieving distributions then the traditional analysis, and led us to investigate means by which to achieve this distribution on the transmitted signals. This was achieved through the use of non-equiprobable signaling and shaping codes, which allowed transmission with high power efficiency and less wasted power. These results were confirmed by simulations of the end-to-end communication system in chapter 5.

6.2 Future Work

In QAM signaling schemes where nonlinear logarithmic quantization is performed, higher powered constellation points are more susceptible to noise than lower powered ones. For this reason, occasionally systems utilize constellation warping, where the distance between adjacent constellation points increases with increasing distance from the origin. The result is an increase in minimum distance for points of high power, which combats the effects of logarithmic quantization. Additionally, there is another gain from performing such a technique. It was shown in previous chapters how a Gaussian like distribution at the input is needed to approach channel capacity. By using warping, the point density near zero increases, and decreases with increasing distance from the origin. This results in a reduction in average power, in a flavor similar to that for non-equiprobable signaling. In fact, it was shown in [17] that by using non-uniform spacing between constellation points but equiprobable signaling, it is possible to approach channel capacity without resorting to shaping.

A nonlinear amplifier can also be thought of as a warping function applied to a signal set, changing the set so that a varying distance between points results. The high powered points experience a decrease in the average distance between neighbors compared to the average neighbor distance for the low power points. This should produce an increase in average power compared to the case where no nonlinearities exist. Work has been done in the area of non-uniform constellations that analyzes the gain of performing warping techniques [16]; in this work, non-uniform signaling was also discussed in the context of coded modulation. Future work could involve extend this analysis to nonlinear amplifiers, looking at the loss associated with these devices, and possible coding techniques appropriate for the system.

Coding principles were briefly described for warped constellations in [16]. The basic idea is that by warping (or nonlinear amplification), only a fraction ψ of the points now have the minimum distance d_{min}. If the length of the minimum distance trellis event is L, then the path multiplicity is multiplied by a factor ψ^L. Multiplying an approximation to the error probability by this factor should serve as a good lower bound on the probability of error for the warped constellation. Hence, for channel coding with nonlinearities, it is best to have the dominant error a path in the trellis (as opposed $L = 1$), and to have the path as long as possible. Note however that this argument is valid for the range of SNR's in which the performance is determined by both the minimum distance and the average number of nearest neighbors. This simple example demonstrates an open research issue with respect to nonlinearities and coded modulation.

Bibliography

[1] L. Larson, *Radio Frequency Integrated Circuit Technology for Low-Power Wireless Communications,* IEEE Personal Comm., June 1998.

[2] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* John Wiley and Sons, Inc., 1991.

[3] J. G. Proakis, *Digital Communications,* McGraw-Hill, Inc., 1995.

[4] C. Kiang, J. Jong, W. Stark, J. East, "Nonlinear Amplifier Effects in Communications Systems,", *IEEE Trans. Micro. Theory Techn.,* vol. 47, pp.1461-1466, Aug. 1999.

[5] J. Jong and W. Stark, "Performance Analysis of Coded Multicarrier Spread-Spectrum Systems in the Presence of Multipath Fading and Nonlinearities,", *IEEE Trans. Comm.,* vol. 49, pp.168-179, Jan. 2001.

[6] M. Le and L. Thibault, "Performance evaluation of COFDM for digital audio broadcasting–Part II. Effects of HPA nonlinearities,", *IEEE Trans. Broadcast.,* vol. 44, pp.165-171, June 1998.

[7] S. Verdu, "On Channel Capacity per Unit Cost,", *IEEE Trans. Inform. Theory,* vol. 36, pp.1019-1030, Jan. 1990.

[8] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable Signaling on the Gaussian Channel,", *IEEE Trans. Inform. Theory,* vol. 36, pp.726-740, July. 1990.

[9] A. R. Calderbank and M. Klimesh, "Balanced Codes and Nonequiprobable Signaling,", *IEEE Trans. Inform. Theory,* vol. 38, pp.1119-1122, May. 1992.

[10] J. N. Livingston, "Shaping Using Variable-Size Regions,", *IEEE Trans. Inform. Theory,* vol. 38, pp.1347-1353, July. 1992.

[11] U. Wachsmann, R. Fischer, J. Huber, "Multilevel Codes: Theoretical Concepts and Practical Design Rules,", *IEEE Trans. Inform. Theory,* vol. 45, pp.1361-1391, July. 1999.

[12] G. Ungerboeck, "Channel Coding with Multilevel/Phase Signals,", *IEEE Trans. Inform. Theory,* vol. 28, pp.55-67, Jan. 1982.

[13] M. Valenti, "Turbo Codes and Iterative Processing,", *Proc. IEEE New Zealand Wireless Commun. Symp.,* Nov. 1998.

[14] J. Lenk, *Practical Guide to Electronic Amplifiers,* JohPrentice-Hall, Inc., 1991.

[15] T. Schubert and E. Kim, *Active and Non-linear Electronics,* John Wiley and Sons, Inc., 1996.

[16] W. Betts, A. Calderbank, R. Laroia, "Performance of Nonuniform Constellations on the Gaussian Channel,", *IEEE Trans. Inform. Theory,* vol. 40, pp. 1633-1638, Sept. 1994.

[17] F. Sun and C. Tilborg, "Approaching Capacity by Equiprobable Signaling on the Gaussian Channel,", *IEEE Trans. Inform. Theory,* vol. 39, pp.1714-1716, Sept. 1993.