Shaping Multi-dimensional Signal Spaces

by:

Amir K. Khandani

Department of Electrical Engineering

McGill University

Montreal, Canada

March 1992

A thesis submitted to the Faculty of the Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

To my mother and to the memory of my father

Acknowledgement

I am greatly indebted to my supervisor Professor Peter Kabal for his guidance throughout the course of my research. The period of my study at McGill was the most peaceful period of my life and this is mainly due to him. He provided me with the technical and financial support and peace of mind to concentrate.

I am also indebted to Dr. M. R. Soleymani who gave me the chance to start my Ph.D. .

Most of all, I am grateful to my mother who always provided the family with care and love, especially, when I was writting this thesis.

Finally, I am grateful to INRS-Télécommunications and to McGill University for the technical support.

Abstract

In selecting the boundary of a signal constellation used for data transmission, the objective is to minimize the average energy of the set for a given number of points from a given packing. Reduction in the average energy because of using the region $\mathcal{C}$ as the boundary instead of a hypercube is called the shape gain of $\mathcal{C}$. The price to be paid for shaping is: (i) an increase in the factor CER_s (Constellation-Expansion-Ratio), (ii) an increase in the factor PAR (Peak-to-Average-power-Ratio), and (iii) an increase in the addressing complexity. In this thesis, the structure of the region which optimizes the tradeoff between the shape gain and the CER_s and also between the shape gain and the PAR in a finite dimensional space is found. Analytical expressions are derived for the optimum tradeoff. The optimum shaping region can be mapped to a hypercube truncated within a simplex. This mapping has properties which facilitate the addressing of the signal points. We introduce several addressing schemes with low complexity and good performance. The concept of the unsymmetrical shaping is discussed. This is the selection of the boundary of a constellation which has different values of power along different dimensions. The rate of the constellation is maximized subject to some constraints on its power spectrum. This spectral shaping also involves the selection of an appropriate basis (modulating waveform) for the space. Finally, we discuss the selection a signal constellation for signaling over a partial-response channel. In the continuous approximation, we introduce a method to select the nonempty dimensions. This method is based on minimizing the degradation caused by the channel memory. In the discrete case, shaping and coding depend on each other. In this case, a combined shaping and coding method is used. This concerns the joint selection of the shaping and coding to minimize the probability of the symbol error.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminaries

This thesis is concerned with the problem of designing a data transmission system. The data is encoded such that in each signaling interval one of M equiprobable waveforms is transmitted. The overall transmission system can be modeled as a discrete-time system. In the discrete model, the channel provides us with a given number of dimensions per signaling interval. For instance, in a conventional quadrature modulation, a 2-D (2-dimensional) array of signal values is generated for each channel use. If a signaling interval consists of $N/2$ channel uses, we get an effective N-D signal space.

To achieve the transmission, we select M points over the channel space. Each of the transmitter waveforms (source symbols) corresponds to one of these points. This is called a signal constellation. We restrict our consideration to signal points which are selected from a regular array of points, specifically a lattice. Different arrangement of points give rise to lattices with designations such as Z^8 and E_8. An excellent reference work for the theory of lattices is [4]. For instance Z^8 is a rectangular grid of points in 8 dimensions and E_8 is a subset of points of Z^8 which are congruent modulo 2 to the codewords of the binary Reed-Muller code $\mathrm{RM}(1,3) = (8,4,4)$, [13].

In the design of a signal constellation, the overall objective is to minimize the prob-

ability of the symbol error at the receiver side. Our tools are: (i) the selection of the internal structure of the constellation (channel coding), (ii) the selection of the constellation boundary (shaping), and (iii) the selection of the constellation basis (modulating waveforms). We propose implementable schemes which have some benefits over the existing methods. The figure of merit is the reduction in the required average energy with respect to a reference scheme. This is denoted as the overall coding gain which is composed of the channel coding gain and the shaping gain.

In the process, if the channel is nonflat, the constellation shaping in conjunction with an appropriate modulator can produce a nonflat power spectrum to match the channel characteristics.

The problem of the channel coding is a well established subject in the theory of communications. For example, by selecting the constellation points from the lattice E_8, we obtain a channel coding gain of 3 dB over the uncoded case (lattice Z^8). This lattice has a minimum distance of 4 resulting in 6 dB gain, and a redundancy of 4 bits per 8 dimensions ($|Z^8/E_8| = 2^4$) resulting in 3 dB loss, $3 = 6 - 3$.

Ungerboeck proposed the idea of producing dense packings by the use of a trellis diagram. The use of trellis-based packings resulted in a breakthrough in coding theory. For example, by using the lattice E_8 with a 64-state trellis, the coding redundancy reduces from four bits to one bit. This scheme, in conjunction a simple shaping method, results in an overall gain of 5.4 dB, [42].

Unfortunately, the situation is not as good as the conventional calculation methods based on the minimum distance to the nearest neighbor shows. For example, Forney mentions in [12] that, in a general coset coding scheme, considering the effect of the error coefficient, after the initial 3–4 dB, it takes on the order of a doubling of complexity to achieve each 0.4 dB further increase in the effective coding gain. Consequently, to achieve higher gains, it is worthwhile to invest part of the complexity in shaping rather than in more complex channel codes.

In shaping, one tries to minimize the average energy of the constellation for a given number of points from a given packing. The reduction in the average energy due to the

CER_s	PAR	γ_s	M
1.02	2.18	0.48	131
1.07	2.41	0.72	137
1.19	2.86	1.00	153
1.41	3.53	1.18	181
12.04	33.0	1.31	1542

Table 1.1: Performance of the optimum shaping region in dimensionality $N=64$, the last row corresponds to a spherical region.

use of the region $\mathcal{C}$ as the boundary instead of using a hypercube is called the shape gain of $\mathcal{C}$ and is denoted as $\gamma_s(\mathcal{C})$. The price to be paid for shaping involves: (i) an increase in the factor CER_s (Constellation-Expansion-Ratio)[1], (ii) an increase in the factor PAR (Peak-to-Average-power-Ratio), and (iii) an increase in the addressing complexity[2].

For a given dimensionality N, a spherical shaping region $\mathcal{S}_N$, is the region with the highest possible γ_s but also with high values for CER_s and PAR. As $N \rightarrow \infty$, the shape gain of $\mathcal{S}_N$ tends to 1.53 dB, [11]. This is an upper bound for the shape gain of all regions and is achievable at the price of $CER_s = \infty$ and also $PAR = \infty$. However, as we will see later, an appreciable amount of this upperbound can be achieved over a reasonable dimensionality and with low values of CER_s and PAR. Table 1.1, contains some examples of the achievable shaping performance over dimensionality $N=64$. Column M denotes the required number of points of the 2-D (two dimensional) subconstellations in a scheme carrying 7 bits per two dimensions. For the same bit rate, an unshaped constellation needs $128 = 2^7$ points per 2-D subconstellations.

The major problem associated with shaping in a high dimensional space is the ad-

[1]This is the ratio of the number of points per two dimensions to the minimum necessary number of points per two dimensions.

[2]Addressing is the assignment of the data bits to the constellation points.

dressing complexity. For example, for 2-D subconstellations composed of 128 points, in an $N = 32$ dimensional space, a direct addressing scheme using a lookup table requires a block of memory with 112×2^{112} bits per N dimensions, where 112 arises from 7 bits per channel use times 16 channel uses per signaling interval. This is undoubtly impractical.

1.2 A simple example of shaping

In Figs. 1.1 and 1.2, we see two examples of a 64-points 2-D signal constellation from the half integer grid. The one shown in Fig. 1.1 has a cubic shaping region. The average energy of this constellation is equal to 5.25 . The constellation in Fig. 1.2 is obtained by replacing the four points of the highest energy in the cubic constellation with another four points of lower energy. These are the points marked by the circles. As a result of this replacement, the average energy has reduced to 5 . This corresponds to the shape gain, $\gamma_s = 10 \times \log_{10}(5.25/5) = 0.2$ dB.

The cubic constellation employs 8 points per dimension. This is the minimum necessary number of points per dimension to have 64 points in two dimensions, $8 \times 8 = 64$. However, in the shaped constellation, we have employed 10 points per dimension. Assuming that CER_s is measured on a one dimensional basis, this corresponds to, $\text{CER}_s = 10/8 = 1.25$.

The peak of energy per dimension of the cubic and the shaped constellations are equal to 12.25 and 20.25, respectively. The average energies per dimension are equal to 5.25 and 5, respectively. Assuming that PAR is measured on a one dimensional basis, the PAR's are equal to, $12.25/5.25 = 2.33$ and $20.25/5 = 4.05$, respectively.

In the cubic constellation, to map the six bits of data to a constellation point, we can use three bits to select a point along one dimension and another three bits to select a point along the other dimension. However, in the shaped constellation, this method is not applicable. In this simple example, we can use a lookup table with 64 memory locations to achieve the addressing. However, for the same number of bits per dimension, namely 3, in dimensionality 24, we need a lookup table with 2^{72} memory locations which is not

practical.

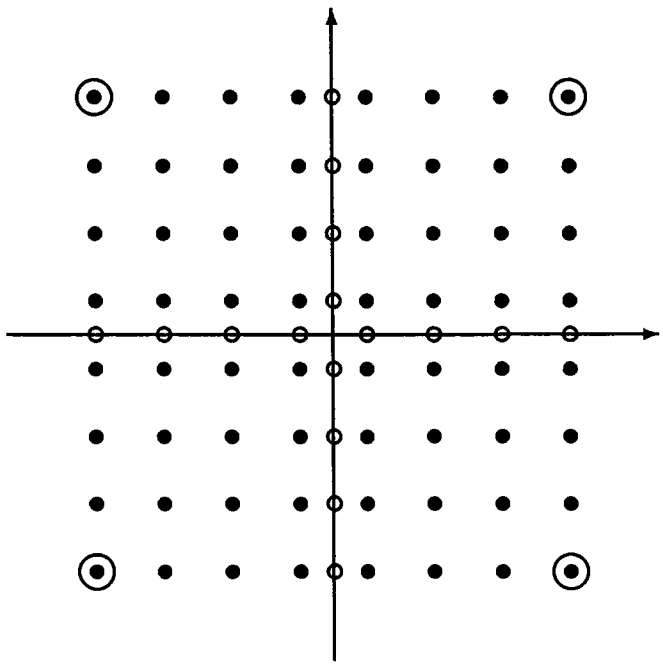


Fig. 1.1: A 64-points 2-D cubic constellation from half integer grid.

1.3 Previous relevant works

1.3.1 Constellation Shaping

In the work of Wei, [42], shaping is a side effect of the method employed to transmit a nonintegral number of bits per two dimensions. This method provides moderate shape gain for low values of CER_s. The addressing of this method is achieved by a table lookup. Forney and Wei elaborate and generalize this method under the topic of the generalized cross constellations in [9]. They also present an existence argument for the optimum shaping region in an infinite dimensional space and calculate the corresponding tradeoff. This is based on finding the optimum induced probability distribution on the 2-D

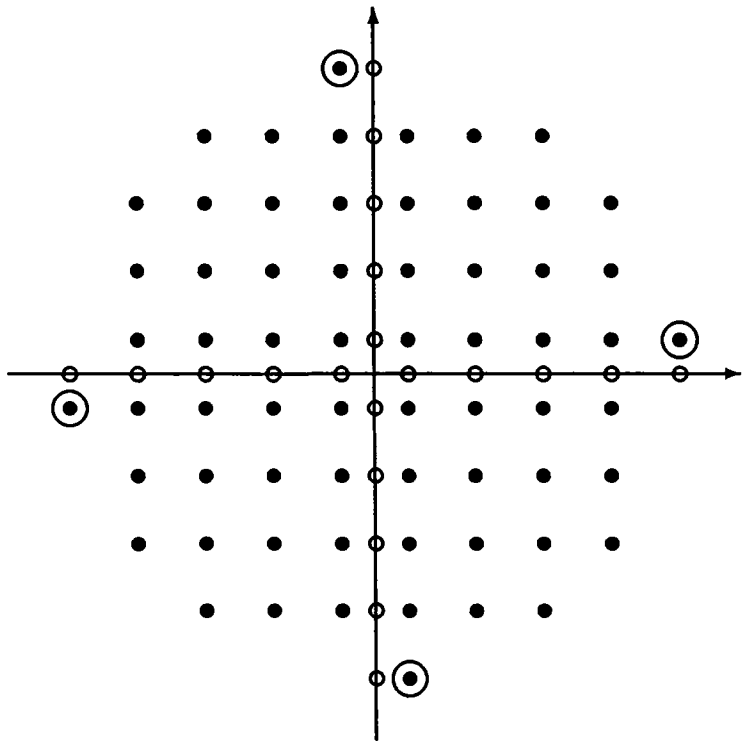


Fig. 1.2: A 64-points 2-D shaped constellation from half integer grid.

subspaces. Conway and Sloane in [5] introduced the idea of the Voronoi constellation based on using the Voronoi region of a lattice Λ_s as the shaping region. In these constellations the set of the points form a group under vector addition modulo Λ_s. This property is used to achieve the addressing. The complexity of the addressing is that of a linear mapping plus the decoding of the shaping lattice Λ_s. The Voronoi constellations are further considered by Forney in [10]. In [1], Calderbank and Ozarow introduced a shaping method which is directly achieved on the 2-D subconstellations. In this method, the 2-D subconstellations are partitioned into equal sized subregions of increasing average energy. A shaping code is then used to specify the sequence of the subregions. The shaping code is designed so that the lower energy subregions are used more frequently. They also calculate the optimum probability distribution on the 2-D subregions. Lang and Longstaff in [40] use an addressing scheme which first divides the final constellation into shells. Then, a point in a shell is found by successively decomposing the space into lower-dimensional subspaces via generating function techniques. The idea of the trellis shaping is introduced

in [14]. This is based on using an infinite dimensional Voronoi region, determined by a convolutional code, to shape the constellation.

1.3.2 Block-based signaling over partial response channels

The structure of the optimum modulator for a zero state block-based signaling scheme over a partial response channel together with a method to select the nonempty dimensions is introduced in [24]. The problem of designing codes for partial response channels is discussed in [24], [44] and [41]. The selection of a signal constellation by performing an optimization procedure over the discrete set of the constellation points was proposed for the first time in [16]. In a more general approach, [19] (also refer to [20]) discusses the optimization of a signal constellation for signaling over a partial response channel. In [19], two methods are studied. In the first method, the constellation at channel output is a finite portion of a dense lattice and the shaping region at the channel input is a sphere. In the second method, a gradient search algorithm is used to find a locally optimum constellation.

1.3.3 Spectral Shaping

The problem of the spectral shaping is a well established subject. The major difference between our approach and most of the works reported in the literature (two excellent references are [22] and [2]) is that in our case the memory of the code is confined to the elements within a block. This fact facilitates the calculation of the entropy (rate of the constellation). For a general code, the exact calculation of the entropy appears to be very difficult, [22].

1.4 Major contributions of the thesis

The major contributions of this thesis are:

- Finding the structure of the shaping region which optimizes the tradeoff between the γ_s and the CER_s and also the tradeoff between the γ_s and the PAR in a finite dimensional spaces.

- Finding analytical expressions for the optimum tradeoff curves. These curves provide an upperbound to the achievable shaping performance in a finite dimensional space. This upperbound, in the same way as the channel capacity in the channel coding or the rate distortion function in the source coding, serves as a benchmark against which practical schemes can be measured.

- Finding practical addressing schemes to achieve tradeoff points near to the knee of optimum curves. Specially, the shell mapped Voronoi constellation based on the lattice D_n^* and the address decomposition method are good candidates for practical implementation.

  As an example of the complexity of our addressing methods, for $N = 32$, we use a block of memory with $M_s = 44$ kilo-bytes per N dimensions to achieve a shape gain of $\gamma_s = 0.89$ dB with $\text{CER}_s = 1.19$ and $\text{PAR} = 2.8$ or, $M_s = 36$ kilo-bytes per N dimensions to achieve $\gamma_s = 1.02$, $\text{CER}_s = 1.41$ and $\text{PAR} = 3.42$. With VLSI technology, these memory sizes are available in a single memory chip.

- Discussing the concept of the unsymmetrical shaping. This is the selection of a signal constellation which has nonequal second moments along different dimension. For example, this nonequal energy allocation in conjunction with a nondiagonal modulating matrix can be used to shape the power spectrum of the transmitted signal.

- Introducing some methods to combine shaping and coding for signaling over a nonflat channel. This concerns using an optimization procedure, partly integer, to jointly select the shaping and coding.

- Finding analytical expressions for the eigensystem of the $1 \pm D$ and $1 - D^2$ partial-response channels.

- Finding analytical expressions for the weight distribution of the scaled D_4 and E_8 lattices.

- Finding an analytical method to calculate the absolute first moment of a lattice Voronoi region and presenting closed form formulas for the specific examples of the D_n and $\Re D_n$ lattices.

1.5 Organization of the thesis

A common thread for this work is the use of a lattice to define the signal points.

In Chapter 2, we have some definitions.

In Chapter 3, we discuss the structure of a general coset coding scheme with the emphasis on the shaping aspects, specifically, the interaction between the shaping and the channel coding.

In Chapter 4, two approaches for shaping based on a continuous approximation are introduced. In the first method, a 2-D sphere is the boundary of the 2-D subspaces and an N-D sphere is the boundary of the whole space. This method results in the optimum tradeoff between γ_s and CER_s and also between γ_s and PAR. Analytical expressions determining the optimum tradeoff are calculated for a finite number of dimensions. In the second method, a 2-D sphere is the boundary of the 2-D subspaces, an N'-D sphere, $N' \geq 2$, is the boundary of the N'-D subspaces and an N-D sphere is the boundary of the whole space. Analytical expressions determining the tradeoff are calculated.

In Chapter 5, practical methods of addressing are discussed. One class of the addressing schemes is based on using a lookup table. We introduce two methods to facilitate the hardware realization of the lookup table. The first method makes use of a specific property of the optimum shaping region and results in a logical table with many don't care entries. The second method, denoted as the address decomposition, is based on decomposing the addressing into a hierarchy of the addressing steps each of a low dimensionality. This avoids the exponential growth of the memory size. This method has a

negligible suboptimality and is easy to implement. Another class of addressing schemes is based on using a Voronoi constellation in a space of half the original dimensionality. Using this method, we achieve a single point with low addressing complexity on the optimum tradeoff curve. We also introduce a hybrid multi-level addressing scheme which combines the two classes. This scheme provide single points with moderate addressing complexity near to the knee of the optimum tradeoff curve.

In Chapter 6, the concept of the unsymmetrical shaping is discussed. This concerns the selection of the boundary of a constellation which has nonequal values of power along different dimensions. The objective is to maximize the rate of the constellation subject to some constraints on its power spectrum. We also consider the selection of a basis (modulating waveforms) for the space. An unsymmetrical region is obtained by scaling of a symmetrical baseline region. We show that the baseline region can be selected independently of the scale factors and the basis. The structure of the optimum baseline region with the corresponding addressing scheme is discussed. The scale factors and the constellation basis are computed by an optimization procedure. This reduces to maximizing the determinant of the correlation matrix subject to linear constraints on its elements. The optimum scheme and also a computationally efficient scheme based on the sine transform are studied.

In Chapter 7, we discuss the selection a signal constellation for the signaling over a partial-response channel. The design steps are: (i) selecting the internal structure of the constellation (channel coding), and (ii) selecting the constellation boundary (shaping). The objective is to minimize the degradation caused by the combined effect of the additive Gaussian noise and the channel memory. Assuming continuous approximation, shaping and coding are selected independently. The selection procedure is similar to the case of a flat channel with the difference that here some of the dimensions can be empty. We introduce a method to select the nonempty dimensions. In the discrete case, shaping and coding depend on each other. In this case, we introduce two joint optimization methods, partly integer, to select the shaping and coding (combined shaping and coding). In the first method, the minimum distance to noise ratio along all the nonempty dimensions is

the same. In the second method, this restriction is relaxed. As part of the calculations, we have found a closed form formula for the weight distribution of the scaled D_4 and E_8 lattices.

Finally, Chapter 8 is devoted to some concluding remarks.

Chapter 2

Shaping of a Constellation, Definitions

A 2-D (two-dimensional) signal constellation C_2 is a finite set of 2-D points bounded within a shaping region $\mathcal{C}_2$. In using such a constellation for signaling over a channel, the energy associated with different signal points is not the same. By using the points of the higher energy less frequently, one can obtain a higher entropy for a given average energy, [1]. Such a nonuniform probability distribution reduces the entropy of the set and, consequently, one needs more points to transmit the same rate. Increasing the number of signal points in the constellation is a price to be paid for the reduction in the average energy and is denoted by the factor CER_s (shaping-Constellation-Expansion-Ratio). To expand the constellation, some points of higher energy are added around the existing points. This increases the PAR (Peak-to-Average-Power-Ratio) of the constellation. The PAR affects the sensitivity of the constellation to the nonlinearities and other signal-dependent perturbations.

In using a nonequiprobable signaling scheme with a signal constellation C_2, we are potentially faced with the following problems: First, the rate associated with some points may not be an integer. Second, as the rate transmitted per channel use is not constant, we may have variable delay in the transmission. A practical way to avoid these problems is

given in [1]. Another method is to use an N-D signal constellation, C_N ($N > 2$), selected as an appropriate subset of the $N/2$-fold cartesian product of C_2 with itself. This subset is selected by the shaping region $\mathcal{C}_N$. In this case, using the points of the C_N with equal probability induces a nonuniform probability distribution on the points of the C_2's. This is the concept of the constellation shaping. The reduction in the average energy per two dimensions due to the use of the region $\mathcal{C}$ as the boundary instead of using a hypercube is called the shape gain of $\mathcal{C}$ and is denoted as $\gamma_s(\mathcal{C})$.

Addressing is the assignment of the data bits to the constellation points. If C_N is equal to, $\{C_2\}^{N/2}$ ($N/2$-fold cartesian product of C_2), the addressing in C_N can be achieved independently along each C_2. For a shaped constellation, which is a subset of $\{C_2\}^{N/2}$, independent addressing is not applicable and we need a means to specify that certain elements of $\{C_2\}^{N/2}$ are not allowed. This means that the use of shaping increases the addressing complexity. For a fixed number of bits per dimension, a multidimensional constellation can have a huge number of points. This makes the addressing of such a constellation a complicated problem.

In all our discussions, we assume that the dimensionality is even and the constellation points are used with equal probability.

By selecting $\mathcal{C}_N$ as the boundary of the constellation (instead of a hypercube), the average energy per two dimensions, P_2, reduces by the factor, [9],

$$\gamma_s(\mathcal{C}_N) = \frac{[V(\mathcal{C}_N)]^{2/N}}{6\,P_2(\mathcal{C}_N)}, \tag{2.1}$$

where $V(\mathcal{C}_N)$ is the volume of $\mathcal{C}_N$. This is called the shape gain of $\mathcal{C}_N$.

For a given integer l, assume that the space dimensions are indexed by $j = lp + m$, where $p = 0, \ldots, (N/l) - 1$ and $m = 0, \ldots, l - 1$. The subspaces spanned by the set of vectors with the same index p are called the (constituent) l-D subspaces, [9]. The region $\mathcal{C}_N$ is called 2-dimensionally symmetrical if its projection on the constituent 2-D subspaces is the same, [9]. All our discussions are based on 2-dimensionally symmetrical regions.

The Constellation-Expansion-Ratio, CER_s, is the ratio of the number of points per

two dimensions to the minimum necessary number of points per two dimensions, [9], i.e.,

$$\mathrm{CER}_s(\mathcal{C}_N) = \frac{V(\mathcal{C}_2)}{[V(\mathcal{C}_N)]^{2/N}} \,. \tag{2.2}$$

As an alternative to CER_s, we define the shaping redundancy in bits per N dimensions by,

$$r_s(\mathcal{C}_N) = (\frac{N}{2}) \log_2(\mathrm{CER}_s) \,. \tag{2.3}$$

Let's $E_P(\mathcal{C}_2)$ denotes the peak energy of $\mathcal{C}_2$. The Peak-to-Average-power-Ratio, PAR, is defined by, [9],

$$\mathrm{PAR}(\mathcal{C}_N) = \frac{E_P(\mathcal{C}_2)}{P_2(\mathcal{C}_N)} \,. \tag{2.4}$$

In general, there exists a tradeoff between the γ_s and the CER_s (or equivalently r_s) and also between the γ_s and the PAR. By an optimally shaped region, we mean a region which optimizes both of these tradeoffs. This region has the minimum second moment for a given volume, given CER_s and given PAR.

Chapter 3

Shaping and Coding on Lattices

Part of this chapter have been reported in [25].

A general coset coding scheme based on the lattice partition Z^N/Λ, Z^N is the N-D integer lattice and Λ is a sublattice of Z^N, is composed of two different parts. The first part selects a finite number of points from Z^N as the signal constellation. This selection is based on minimizing the average energy of the set for a given number of points and given CER_s. The second part selects a coset of Λ within the signal constellation. In the case that the selected coset has more than one point, a third part is used to address one point within that coset. The first and second parts have to do with shaping and coding, respectively.

In continuous approximation, the discrete set of the constellation points is approximated by a continuous uniform density within the shaping region. Assuming continuous approximation, all the parameters concerning shaping like γ_s, CER_s and PAR are determined by the first part and all the parameters concerning channel coding are determined by the second part. The third part scales the number of the constellation points.

Figure 3.1 shows the block diagram of the coding scheme under consideration. Signal space has $N = 2n$ dimensions and carries Q bits per two dimensions. There is also one bit of coding redundancy per N dimensions. The 2-D subconstellations are selected from the cross constellation, [11]. In the case that we need a nonintegral bit rate per two

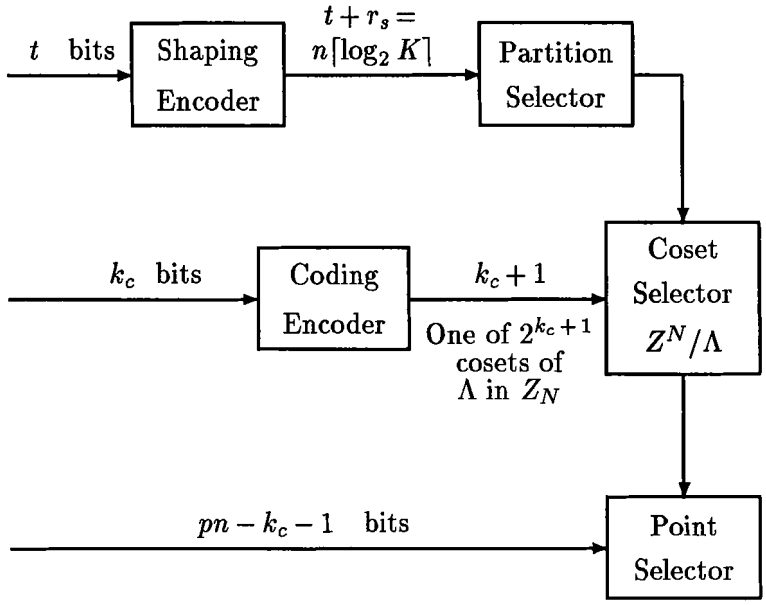


Fig. 3.1: Block diagram of the coding system.

dimensions, the necessary number of points of the least energy from the larger constellation are added around the existing points. Each 2-D point is labeled by a two part label. The first part of the label is determined by the shaping block. The second part of the label is determined by coding block.

For shaping, the two dimensional subconstellation containing M points are partitioned into K shaping shells of equal size and increasing average energy. The shells have four way symmetry. Each shell contains $P = 2^p$ points, $M = K \times P$, and is referenced by $\lceil \log_2 K \rceil$ bits. All the P points within a shaping shell use these bits as their shaping label. We refer to this partitioning/labeling as the shaping partitioning/labeling. Figure 3.2 shows an example of a 256 points constellation divided into 4 shells. In all cases, a finer partitioning of $2K$ shells can be obtained from a constellation already divided into K shells by subdividing each shell into two subshells.

The 2-D shaping shells partition the N-D space into K^n, $n = N/2$, shaping clusters. Shaping is achieved by selecting $T = 2^t$ clusters of the least average energy. The t shaping

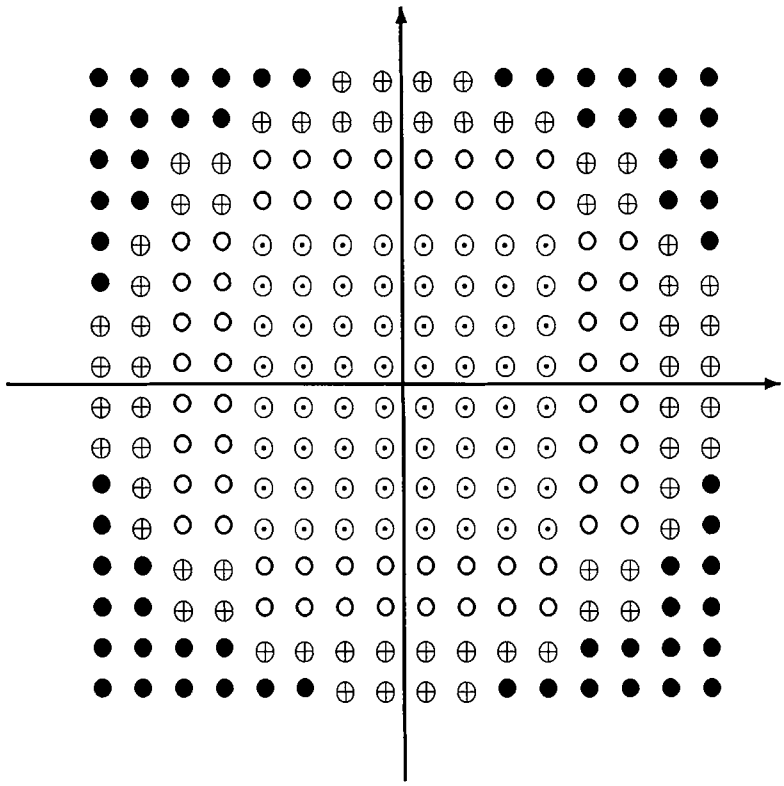


Fig. 3.2: Example of the 2-D shaping shells.

bits entering the shaping encoder are used to address one of these T clusters. The shaping encoder adds r_s redundant bits to the incoming bits and the $t + r_s = n\lceil \log_2 K \rceil$ bits at its output are used in parallel to address one shaping shell within each 2-D subconstellation. The total information rate is equal to $pn + t - 1$ bits.

For coding, the signal constellation is partitioned into the cosets of Λ in Z^N. Assuming that Λ is a binary lattice, each coset of Λ is labeled by $\log_2 |Z^N/\Lambda| = k_c + 1$ bits. The points within each coset have its label as their coding label. This is denoted as the coding partitioning/labeling. In the case that Λ is one of the Barnes-Wall lattices, [13], coding partitioning/labeling is achieved by applying the Ungerboeck partitioning/labeling rules to the two dimensional subconstellations, [42]. Most of the interesting lattices, like D_4, Schläffi lattice, and E_8, Gosset lattice, belong to this group, [13]. In the coding part, k_c

bits enter the encoder. After adding one bit redundancy, the k_c+1 bits at the encoder output are used to select one of the 2^{k_c+1} cosets of Λ in the shaping cluster already selected by the shaping part. For this selection to be possible, each N-D cluster should have an equal number of points from each coset of Λ. If Λ is one of the Barnes-Wall lattices, this condition is satisfied if in the shaping partitioning of the 2-D subconstellations, each shaping shell contains an equal number of points from each partition of the Ungerboeck partition chain, [42]. This is the point where the shaping and coding potentially interfere with each other.

To transmit Q bits per two dimensions with one bit redundancy, we should have, $pn+t=nQ+1$, or, $t=n(Q-p)+1$. For the coding partitioning to be possible, we should have $2^{k_c+1} \leq 2^{pn}$ or $p \geq (k_c+1)/n$. To transmit the total (coded) rate of $nQ+1$, we should also have, $n(p+\log_2 K) \geq nQ+1$, or, $\log_2 K \geq Q-p+(1/n)$.

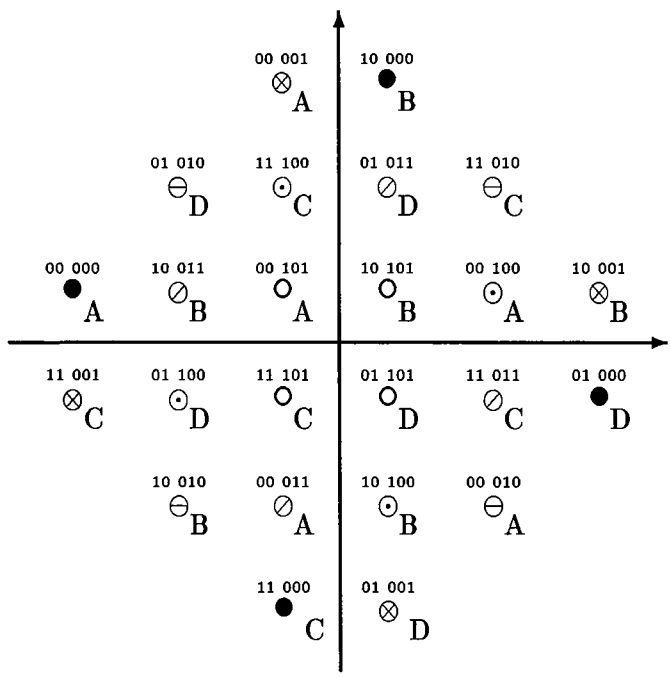


Fig. 3.3: An $M=24$ point constellation divided into $K=6$ shells.

An example of this partitioning with $M=24$ and $K=6$ ($P=4$) is shown in Fig. 3.3. The coding partitions are denoted by A$\leftrightarrow$ 00, B$\leftrightarrow$ 10, C$\leftrightarrow$ 11, D$\leftrightarrow$ 01. The shaping

shells are denoted by the $\bullet \leftrightarrow$ 000, $\otimes \leftrightarrow$ 001, $\ominus \leftrightarrow$ 010, $\oslash \leftrightarrow$ 011, $\odot \leftrightarrow$ 100, and $\circ \leftrightarrow$ 101. The first two bits of the label of each point is the coding label and the last three bits is the shaping label. Considering the condition $2^{k_c+1} \leq 2^{pn}$, this constellation can be used as long as $|Z^N/\Lambda| < 2^N$, for example when $\Lambda = E_8$, $|Z^8/E_8| = 2^4$, or when $\Lambda = D_4$, $|Z^4/D_4| = 2^3$.

The CER_s of this scheme is equal to,

$$\mathrm{CER}_s = M \times 2^{-Q-(1/n)}\,. \tag{3.1}$$

In the receiver, we first do the channel decoding and decide which point is transmitted along each two dimensional subspace. After that, shaping labels of these points are concatenated and are passed through a system which inverts the effect of the shaping encoder to recover the original shaping bits.

Chapter 4

Optimum Shaping, Shell Mapping

Part of this chapter have been reported in [26], [27].

4.1 Introduction

In this chapter, we study two methods for shaping. In the first method, a 2-D sphere is the boundary of the 2-D subspaces and an N-D sphere is the boundary of the whole space. The final region, which is denoted as $\mathcal{A}_N$, results in the optimum tradeoff between the γ_s and CER_s and also between the γ_s and PAR. In this case, the ratio of the radii of the two spheres determines the tradeoff. By applying a shell mapping, the optimum shaping region is mapped to a hypercube truncated within a simplex. This mapping has properties which facilitate the addressing of the signal points. Analytical expressions are presented for the tradeoff as a function of the dimensionality. We describe a method to achieve a point with low addressing complexity on the optimum tradeoff curves. When the dimensionality increases, the point achieved moves towards the initial parts of the curve (low γ_s). In this case, to achieve points with higher γ_s, a second shaping method with two degrees of freedom is used. In this method, first an $\mathcal{A}_{N'}$ region is employed along the N'-D subspaces and then the cartesian product $\{\mathcal{A}_{N'}\}^{n'}$, $n' = N/N'$ is further shaped by the use of an N-D sphere. Using this structure, we introduce a method to achieve a

curve or single points with low addressing complexity near to the optimum curves.

4.2 Shaping using one level of shell mapping

In an optimally shaped region, a 2-D sphere of radius R_2, $\mathcal{S}_2(R_2)$, is the boundary of the 2-D subspaces and an N-D sphere of radius R_N, $\mathcal{S}_N(R_N)$, is the boundary of the whole space. The final region is denoted by $\mathcal{A}_N$, i.e.,

$$\begin{aligned} \mathcal{A}_N &= \{X_j,\ j = 0, \ldots, N-1\} \\ &: \quad 0 \le X_{2p}^2 + X_{2p+1}^2 \le R_2^2, \qquad p = 0, \ldots, n-1 \\ &\quad\ 0 \le \sum_{j=0}^{N-1} X_j^2 \le R_N^2, \qquad 0 \le R_N^2 \le nR_2^2 \ . \end{aligned} \tag{4.1}$$

We say that the shaping of $\mathcal{A}_N$ is achieved in two steps; the first step uses the $\mathcal{S}_2$'s and the second step uses the $\mathcal{S}_N$. The projection of the region $\mathcal{A}_N$ on any constituent l-D subspace, l being an even number greater than two, is the region $\mathcal{A}_l$ with the same value of β.

The optimality of $\mathcal{A}_N$ as far as trading off γ_s versus CER_s is due to:

1. Considering (2.2), in so far as $V(\mathcal{C}_2)$ and $V(\mathcal{C}_N)$ are constant, changing the $\mathcal{C}_2$ does not change the CER_s. On the other hand, among all 2-D figures, a sphere has the least second moment for a given volume. This implies that a sphere should be the boundary of the 2-D subspaces.

2. The final region should be selected as a subset of $\{\mathcal{S}_2(R_2)\}^{N/2}$ which has the minimum second moment for a given volume. This implies that the boundary of the whole space is an N-D hypersphere.

Furthermore, it can be shown, [9],

$$\text{PAR}(\mathcal{C}_N) = \frac{\text{PAR}(\mathcal{C}_2)}{\gamma_s(\mathcal{C}_2)} \times \gamma_s(\mathcal{C}_N) \times \text{CER}_s(\mathcal{C}_N) \ . \tag{4.2}$$

But, a sphere is the 2-D figure which maximizes $\gamma_s(\mathcal{C}_2)$ and minimizes the $\text{PAR}(\mathcal{C}_2)$. Using these facts and also the optimality of the tradeoff between γ_s and CER_s in (4.2), proves the optimality of the tradeoff between γ_s and PAR.

By applying the change of variable,

$$Y_p = (X_{2p}^2 + X_{2p+1}^2)/R_2^2, \quad p = 0, \ldots, n-1, \quad n = N/2, \tag{4.3}$$

to (4.1), the region $\mathcal{A}_N$ is mapped to the following n-D solid,

$$\begin{aligned} \mathcal{TC}_n(1,\beta) &= \{Y_p,\ p = 0, \ldots, n-1\} \\ &: \ 0 \le Y_p \le 1 \\ &\quad 0 \le \sum_{p=0}^{n-1} Y_p \le \beta, \qquad 0 \le \beta = R_N^2/R_2^2 \le n\ . \end{aligned} \tag{4.4}$$

This is a hypercube of edge length one truncated within a simplex of edge length β. We refer to the N-D space as the N-domain and to the n-D space as the n-domain. This mapping, which is denoted as the shells mapping, is the key point to most of our discussions. Figure 4.1 shows an example for $N = 4$.

Shell mapping has the following properties:

- A uniform density of points within $\mathcal{A}_N$ results in a uniform density of points within $\mathcal{TC}_n$. This property allows us to achieve the shaping and the addressing on the equal volume partitions of $\mathcal{TC}_n$. This property can be developed from the basic fact that for a uniform density in a spherically symmetric region in 2-D (a circle in our case), the transformation from the rectangular coordinates (X_0, X_1) to the spherical coordinates $(U = X_0^2 + X_1^2, \Theta)$ gives a uniform U.

- Unlike the $\mathcal{A}_N$ region, the boundaries of $\mathcal{TC}_n$ are hyperplanes. This makes the partitioning and addressing of $\mathcal{TC}_n$ an easier task than that of $\mathcal{A}_N$.

- For $\beta = n/2$, the simplex in (4.4) divides the hypercube into two congruent partitions each of volume $1/2$. The $\mathcal{TC}_n$ region is equal to one of them. This is equal to the Voronoi region of the lattice D_n^* in the positive coordinates. This allows us to use a Voronoi constellation[1], [10], for the addressing.

[1]For a brief description of the Voronoi constellations refer to Appendix G.

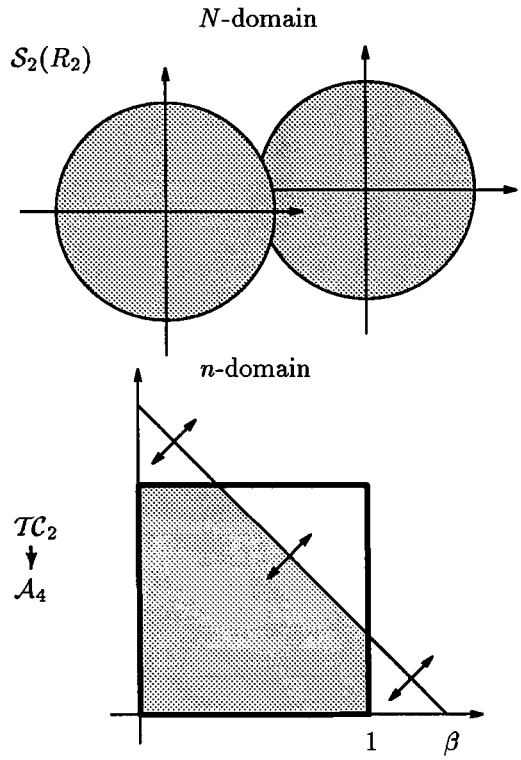


Fig. 4.1: Example of A_4 constellation, one-level shell mapping. Each 2-D subspace in the 4-D space is mapped to one of the axes of the $\mathcal{TC}_2$.

4.2.1 Shape gain tradeoff

In Appendix A, the integral of a function of the general form $F(X_0^2 + \ldots + X_{N-1}^2)$ over the $\mathcal{A}_N$ region is calculated as,

$$\int_{\mathcal{A}_N} F(X_0^2 + \ldots + X_{N-1}^2) dX_0 \ldots dX_{N-1} =$$

$$(\pi R_2^2)^n \sum_{k=0}^{\lfloor\beta\rfloor} (-1)^k C_n^k \frac{(\beta-k)^n}{(n-1)!} \int_0^1 F\left\{R_2^2 [(\beta-k)\tau + k]\right\} \tau^n d\tau \ . \tag{4.5}$$

This integral is used to calculate the volume and the second moment of the $\mathcal{A}_N$ region. The results, together with $V(\mathcal{C}_2) = \pi R_2^2$ and $E_P(\mathcal{C}_2) = R_2^2$, are used in (2.1), (2.2) and (2.4) to calculate the γ_s the CER_s and the PAR. Figure 4.2 shows the corresponding tradeoff curves for different values of N. The curves corresponding to $N = \infty$ are extracted from [9]. In Appendix B, it is proved that as $N \to \infty$, the induced probability distribution

along 2-D subspaces of the $\mathcal{A}_N$ region tends to a truncated Gaussian distribution. This justifies the use of the curves obtained in [9] for the $\mathcal{A}_\infty$ regions. Obviously, this is a consequence of the optimality of these regions.

Referring to Fig. 4.2, it is seen that in general the initial parts of the optimum tradeoff curves have a steep slope. This means that an appreciable portion of the maximum shape gain, corresponding to a spherical region, can be obtained with a small value for CER_s and PAR.

We use $\psi = \beta/n$ $(n = N/2)$ as the normalized parameter for the $\mathcal{A}_N$ region. The complete notation for the region is $\mathcal{A}_N(\psi)$. For $\psi = 1$, $(\beta = n,\ R_N = \sqrt{n}R_2)$, we have, $\mathcal{A}_N = \{\mathcal{S}_2(R_2)\}^n$. This results in the starting point on the tradeoff curves. For $1/n < \psi < 1$, $(1 < \beta < n,\ R_2 < R_N < \sqrt{n}R_2)$, by decreasing ψ, we move along the tradeoff curve. Finally, for $\psi = 1/n$ $(\beta = 1,\ R_N = R_2)$, we obtain the spherical region $\mathcal{S}_N(R_N)$. This case corresponds to the final point on the tradeoff curves. The two cases of $\psi > 1$, and $0 < \psi < 1/n$, result in the regions $\{\mathcal{S}_2\}^{N/2}$ and $\mathcal{S}_N(\sqrt{n\psi}R_2)$, respectively.

	A			*B*			*L*			*K*			*S*		
N	CER_s	PAR	γ_s dB	CER_s	PAR	γ_s dB	CER_s	PAR	γ_s dB	CER_s	PAR	γ_s dB	CER_s	PAR	γ_s dB
4	1.41	3.00	0.46	—	—	—	—	—	—	1.41	3.00	0.46	1.41	3.00	0.46
8	1.19	2.60	0.60	—	—	—	1.19	2.60	0.60	1.41	3.19	0.70	2.21	5.00	0.73
12	1.12	2.47	0.61	—	—	—	1.19	2.68	0.70	1.41	3.26	0.82	2.99	7.00	0.88
16	1.09	2.39	0.60	1.30	3.04	0.85	1.19	2.71	0.76	1.41	3.33	0.90	3.76	9.00	0.98
24	1.06	2.31	0.57	1.19	2.76	0.84	1.19	2.76	0.84	1.41	3.42	1.00	5.29	13.00	1.10
32	1.04	2.26	0.55	1.14	2.62	0.81	1.19	2.80	0.89	1.41	3.45	1.06	6.80	17.00	1.17
48	1.03	2.22	0.52	1.09	2.48	0.76	1.19	2.83	0.96	1.41	3.51	1.14	9.80	25.00	1.26
64	1.02	2.18	0.48	1.07	2.41	0.72	1.19	2.86	1.00	1.41	3.53	1.18	12.04	33.00	1.31
∞	1.00	2.00	0.20	1.00	2.00	0.20	1.19	3.00	1.20	1.41	3.75	1.40	∞	∞	1.54

Table 4.1: A set of the important points from the optimum tradeoff curves.

Table 4.1 contains a set of points from the optimum curves. These are the points marked on the curves in Fig. 4.2. The S-points correspond to a spherical region and achieve the maximum shape gain in a given dimensionality. The K-points correspond to $r_s = N/4$ bits per N dimensions $(\text{CER}_s = (2)^{1/2} = 1.41)$. They achieve almost all of

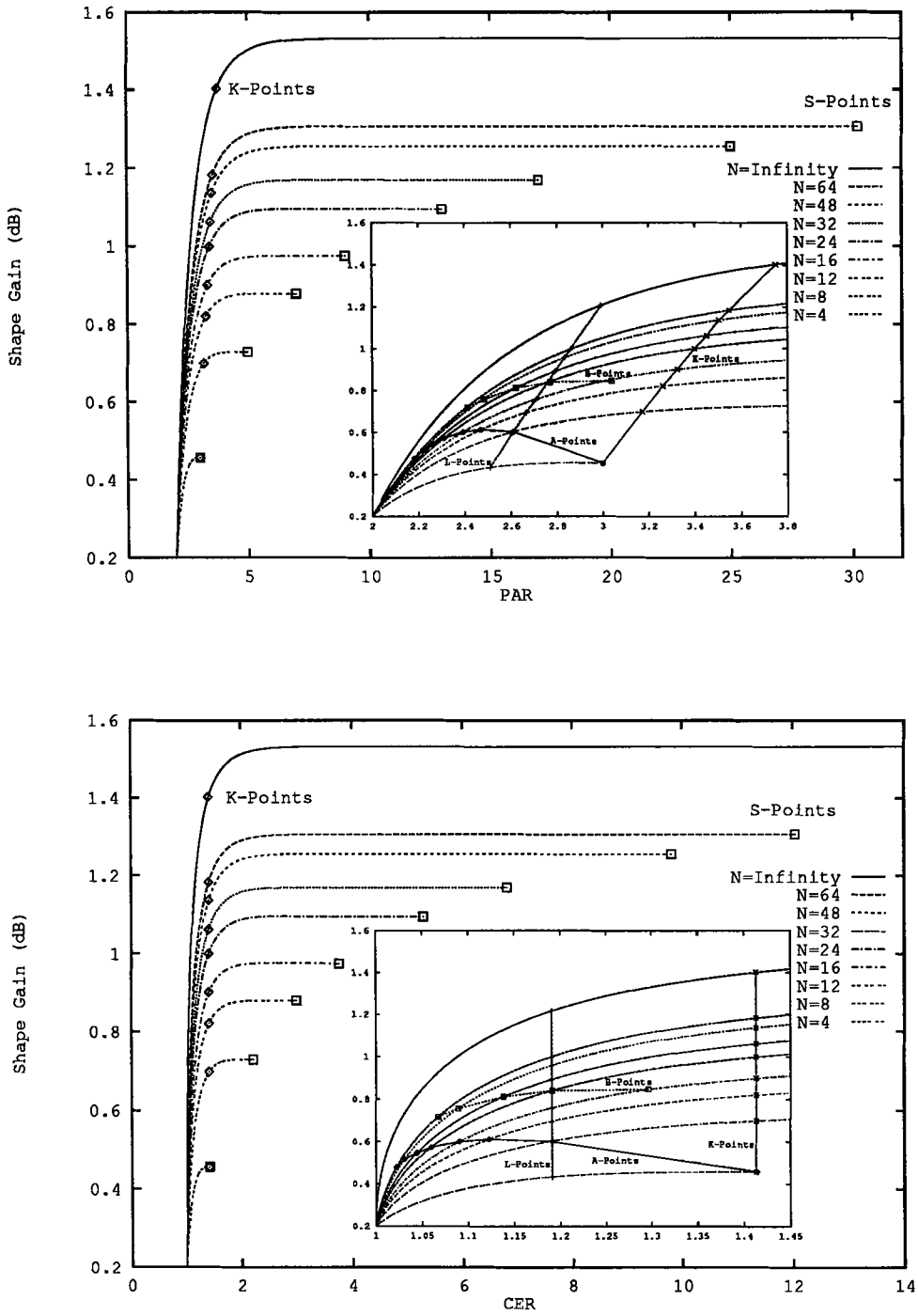


Fig. 4.2: The optimum curves as well as a set of important points on them.

shape gain of the S-points but with a much lower value of CER_s and PAR. The L-points correspond to $r_s = N/8$ bits per N dimensions ($\text{CER}_s = (2)^{1/4} = 1.19$). They achieve a significant γ_s with a very low CER_s and PAR. The A-points correspond to the addressing scheme based on the lattice D_n^*. They result in $r_s = 1$ bits per N dimensions, $\text{CER}_s = (2)^{2/N}$. For $N = 4$, this point corresponds to a spherical region. We will discuss the B-points later.

From Fig. 4.2, it is seen that for $N < 16$, the A-points with $r_s = 1$ bits per N dimensions are located near the knee of the optimum curves. For larger dimensionality, they are closers to the initial parts of the curves. This is due to the fact that one bit of redundancy per N dimensions is too small for $N > 16$. A solution in a space of dimensionality $N = n' \times N'$ is to use the lattice D_n^*, $n = N'/2$, to shape the N'-D subspaces and then achieve another level of shaping on the $n' = N/N'$–fold cartesian product of these subspaces. This is one example for the application of a mutli-level shaping/addressing scheme.

More generally, consider an $\mathcal{A}_N(\psi)$ region. This region has an $\mathcal{A}_{N'}(N\psi/N')$ region along each of its N'-D subspaces. The basic idea is that we can modify the $\mathcal{A}_{N'}(N\psi/N')$ subregions such that the complexity of the addressing is decreased while the overall suboptimality is small. Specifically, in some of our schemes, (i) the $\mathcal{A}_{N'}(N\psi/N')$ region is replaced by the region $\mathcal{A}_{N'}(1/2)$ and/or (ii) this region is partitioned into a finite number of the energy shells and then the cartesian product of the N'-D subspaces is shaped by a lookup table. These are the basis for the multi-level shell mapped constellations and the address decomposition method of the next chapter.

In the following, this idea is explained by the use of a more general approach.

4.3 Shaping using two level of shell mapping

In this method, shaping is achieved in three steps. In the first two steps an $\mathcal{A}_{N'}$ region is employed along the N'-D subspaces. In the third step, from the $n' = N/N'$–fold cartesian product of the $\mathcal{A}_{N'}$ with itself, a subset with a given volume and least second moment

is selected. As before, such a subset is selected by a hypersphere. This results in two degrees of freedom in selecting the final region. This region is denoted by $\mathcal{A}_N^{N'}$ and we have,

$$\mathcal{A}_N^{N'} = \{\mathcal{A}_{N'}\}^{n'} \cap \mathcal{S}_N(R_N) = \{\mathcal{S}_2(R_2)\}^{nn'} \cap \{\mathcal{S}_{N'}(R_{N'})\}^{n'} \cap \mathcal{S}_N(R_N) \ . \tag{4.6}$$

In the case that $\mathcal{A}_{N'}$ is selected as $\{\mathcal{S}_2\}^n$, this method is equivalent to the previous method.

The space dimensions are indexed by $j = N'q + 2p + m$, where $q = 0, \ldots, n' - 1$, $n' = N/N'$, $p = 0, \ldots, n-1$, $n = N'/2$ and $m = 0, 1$. Using the change of variable,

$$Z_q = \sum_{p=0}^{n-1} \left[X_{2(nq+p)}^2 + X_{2(nq+p)+1}^2\right] / R_2^2, \quad q = 0, \ldots, n'-1 \ , \quad p = 0, \ldots, n-1 \ , \tag{4.7}$$

and defining β and β' by,

$$R_{N'}^2 = \beta R_2^2, \qquad R_N^2 = \beta\beta' R_2^2 \ , \tag{4.8}$$

the region $\mathcal{A}_N^{N'}$ reduces to,

$$\begin{aligned} \mathcal{TC}_{n'}(\beta, \beta\beta') \ &= \{Z_q, \ q = 0, \ldots, n'-1\} \\ &: \ 0 \le Z_q \le \beta \\ &\quad 0 \le \sum_{q=0}^{n'-1} Z_q \le \beta\beta' . \end{aligned} \tag{4.9}$$

This is an n'-D simplex of edge length $\beta\beta'$ truncated within a hypercube of edge length β. The n'-D space is denoted as the n'-domain.

The normalized parameters are selected as $\psi = \beta/n$ and $\psi' = \beta'/n'$. The complete notation for the region is $\mathcal{A}_N^{N'}(\psi, \psi')$. For $\psi' = 1$, we have, $\mathcal{A}_N^{N'}(\psi, 1) = \{\mathcal{A}_{N'}(\psi)\}^{n'}$. In this case, γ_s, CER_s and PAR are equal to their corresponding values in $\mathcal{A}_{N'}(\psi)$.

Now, consider the region $\mathcal{A}_N^{N'}(1/2, \psi')$. This region has an $A_{N'}(1/2)$ region along the N'-D subspaces[2]. In Appendix C, the integral of a function of the general form $F(X_0^2 + \ldots + X_{N-1}^2)$ over the region $\mathcal{A}_N^{N'}$ is calculated. This integral is used to calculate

[2]The corresponding shaping region in the n-domain is the Voronoi region of the lattice D_n^* in the positive coordinates. This is a useful property and is used in the next chapter to partition the N'-D subspaces by the use of a lattice partition chain.

the tradeoff in the $\mathcal{A}_N^{N'}(1/2,\psi')$ region. The result of these calculations for $N=16,32$ is shown in Fig.4.3. The starting point of the curves ($\psi'=1$) corresponds to the region $\mathcal{A}_{N'}(1/2)$. It is seen that for relatively high γ_s, and for $n'=2$, the curves are very near to the optimum tradeoff curves.

In practice, we partition each $\mathcal{A}_{N'}(1/2)$ region into K energy shells of equal volume and select a subset in their n'–fold cartesian product. These partitions correspond to equal volume partitions in the n-domain produced by the radial hyperplanes and are denoted by a set of the points $\{U_i,\ i=0,\ldots,K\}$ along each dimension of the $\mathcal{TC}_{n'}$. An example for $N=8$, $N'=4$ and $K=4$ is shown in Fig.4.4.

A point U_i on a dimension of $\mathcal{TC}_{n'}$ corresponds to the region $\mathcal{A}_{N'}$ with $\beta=U_i$. Using (4.5), the volume of this region is equal to,

$$V\left(\mathcal{A}_{N'}(n/U_i)\right)=(\pi R_2^2)^n\sum_{k=0}^{\lfloor U_i\rfloor}(-1)^k\,C_n^k\,\frac{(U_i-k)^n}{n!}\,, \tag{4.10}$$

where $n=N'/2$. To obtain partitions of equal volume, the points U_i should satisfy,

$$\sum_{k=0}^{\lfloor U_i\rfloor}(-1)^k\,C_n^k\,\frac{(U_i-k)^n}{n!}=\frac{i}{K}\sum_{k=0}^{n/2}(-1)^k C_n^k\,\frac{[(n/2)-k]^n}{n!}\,. \tag{4.11}$$

The summation on the right hand is equal to the volume of the region $\mathcal{TC}_n(1,n/2)$ which is equal to $1/2$. Substituting in (4.11), we obtain the following equations for the points U_i,

$$\sum_{k=0}^{\lfloor U_i\rfloor}(-1)^k C_n^k\frac{(U_i-k)^n}{n!}=\frac{i}{2K},\ i=1,\ldots,K\,. \tag{4.12}$$

The partitioning of the N'-D subspaces results in $K^{n'}$ equal volume partitions in the N-domain. Each of these partitions corresponds to a parallelepiped in the n'-domain. A parallelepiped located at point $(U_{I_0},\ldots,U_{I_{n'-1}})$, $I_{j\in\{0,\ldots,n'-1\}}\in\{0,\ldots,K\}$ is shown by,

$$\begin{aligned}\mathcal{P}(U_{I_0},\ldots,U_{I_{n'-1}}) &= \{Z_q,\ q=0,\ldots,n'-1\}\\ &:\ U_{I_0}\le Z_0\le U_{I_0+1},\\ &\cdots\cdots\\ &U_{I_{n'-1}}\le Z_{n'-1}\le U_{I_{n'-1}+1}\,.\end{aligned} \tag{4.13}$$

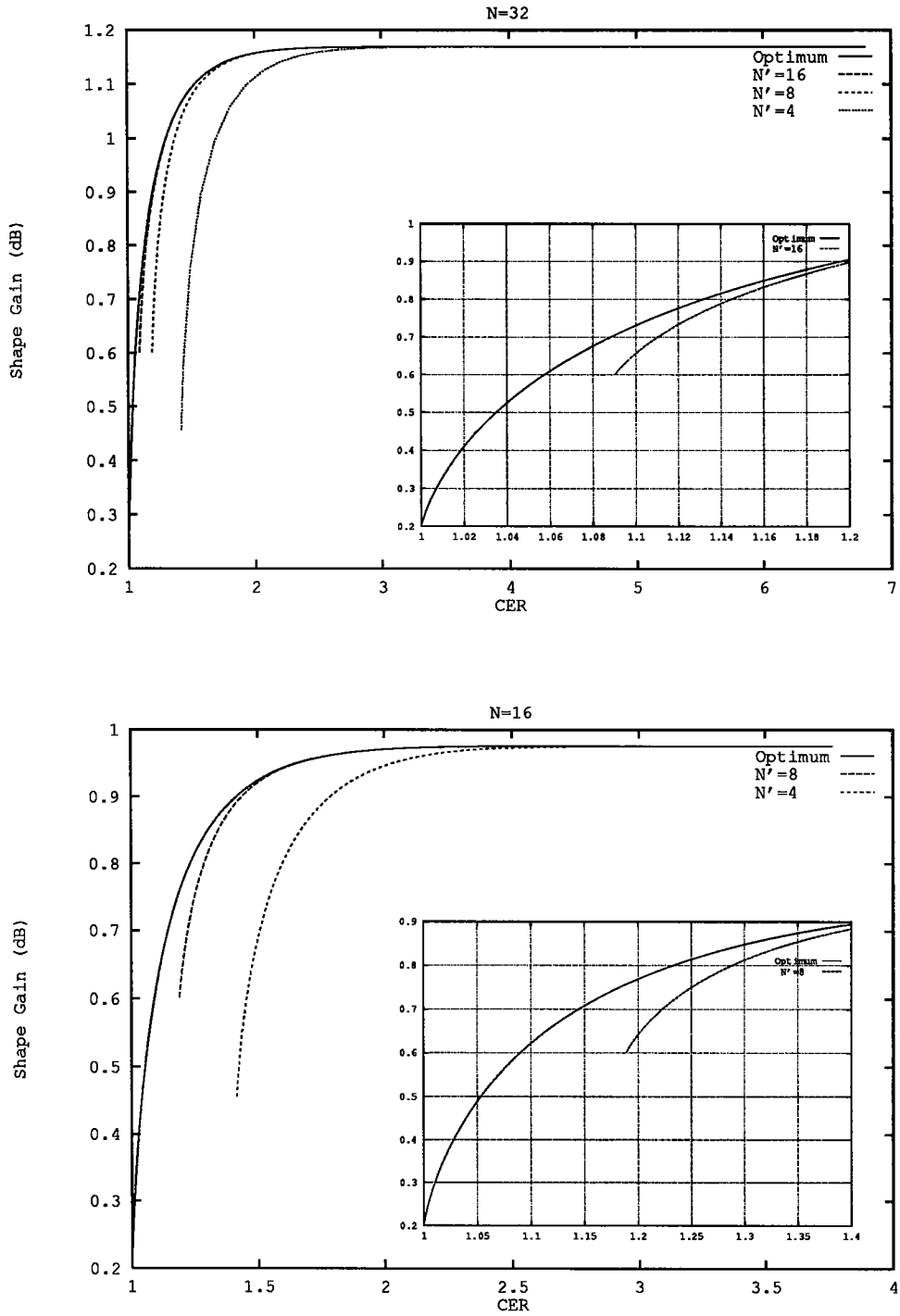


Fig. 4.3: Tradeoff between the CER_s and γ_s in the $\mathcal{A}_N^{N'}(1/2, \psi'')$ regions, $N = 16, 32$.

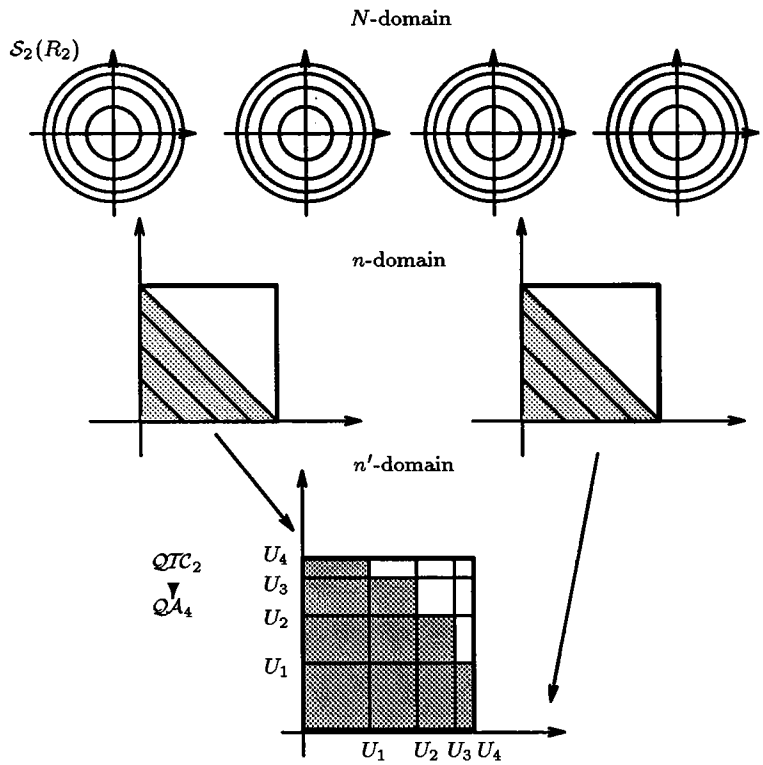


Fig. 4.4: Example of the two-level shell mapping.

Shaping is achieved by selecting T of the N-D partitions with the least second moment. In the example of Fig. 4.4, we have $T = 10$. Considering that the first moment of $\mathcal{TC}_{n'}$ is proportional to the second moment of $\mathcal{A}_N^{N'}$, the selected subset corresponds to the parallelepipeds with the least average first moment. This procedure in fact uses a quantized version of $\mathcal{TC}_{n'}$, denoted by $\mathcal{QTC}_{n'}$, as the shaping region in the n'-domain. The final region is denoted by $\mathcal{QA}_N^{N'}(K,T)$.

In a parallelepiped the average first moment is equal to the sum of the average first moments along different dimensions. Using this fact, we obtain,

$$F_m\left(\mathcal{P}(U_{I_0},\ldots,U_{I_{n'-1}})\right) = \left(\frac{1}{2K}\right)^{n'-1} \sum_{q=0}^{n'-1} \left\{ \sum_{k=0}^{\lfloor U_{I_q+1} \rfloor} (-1)^k C_n^k \right.$$

$$\left. \times \frac{n(\beta-k)^{n+1} + k(n+1)(\beta-k)^n}{(n+1)!} - \sum_{k=0}^{\lfloor U_{I_q} \rfloor} (-1)^k C_n^k \frac{(n)(\beta-k)^{n+1} + k(n+1)(\beta-k)^n}{(n+1)!} \right\}. \tag{4.14}$$

This is used to calculate the average first moment of the selected subset of the parallelepipeds, $F_m(\mathcal{QTC}_{n'})$. The average energy of the N-domain is equal to,

$$P_2\left(\mathcal{QA}_N^{N'}(K,T)\right) = \frac{2R_2^2}{N}\frac{(2K)^{n'}}{T}F_m(\mathcal{QTC}_{n'})\,. \tag{4.15}$$

It is easy to show that the volume of $\mathcal{QA}_N^{N'}$ is equal to,

$$V\left(\mathcal{QA}_N^{N'}(K,T)\right) = (\pi R_2^2)^n\frac{T}{(2K)^{n'}}\,. \tag{4.16}$$

The equations (4.15) and (4.16) can be used to calculate the tradeoff.

From Fig. 4.3, it is seen that for $N'=N/2$ $(n'=2)$, the tradeoff curve for the $\mathcal{A}_N^{N'}$ regions lies very near to the optimum curve. This suggests selecting $N'=N/2$ for the $\mathcal{A}_N^{N'}$ and also for the $\mathcal{QA}_N^{N'}$ regions. Figure 4.5 shows the tradeoff curves of the $\mathcal{QA}_N^{N/2}$ regions as a function of K. It is seen that in general, the suboptimality caused by applying a coarse quantization to $\mathcal{TC}_{n'}$ is negligible. In the following we explain an addressing scheme to achieve the marked points.

For $\psi'=1/2$, the region $\mathcal{TC}_{n'}$ is equal to the Voronoi region of the lattice $\mathcal{D}_{n'}^*$ in the positive coordinates. Unlike the case of the $\mathcal{A}_N$ regions, we cannot use this property to achieve a point on the corresponding tradeoff curve. This is due to the fact that in this case the density of points within $\mathcal{TC}_{n'}$ is no longer constant. However, we can still use this property to achieve points very near to these curves.

To do this, the equal volume partitions of $\mathcal{A}_{N'}$ are mapped to the *equally spaced points* along a dimension. The Voronoi region of $D_2^*=\Re Z^2$ is used to select half of the points in the cartesian product. This results in $r_s=3$ bits per N dimensions, $\mathrm{CER}_s=(8)^{2/N}$. The point marked in Fig. 4.5 corresponds to such a region for $K=128$. It is seen that the degradation is small. The B-points in Table 4.1 and Fig. 4.2 have the same CER_s as the point achieved here but with about 0.1 dB degradation in γ_s.

4.4 Summary and conclusions

In this chapter, we have found the structure of the regions which provide the optimum tradeoff between the CER_s and γ_s and also between the PAR and γ_s in a finite dimensional

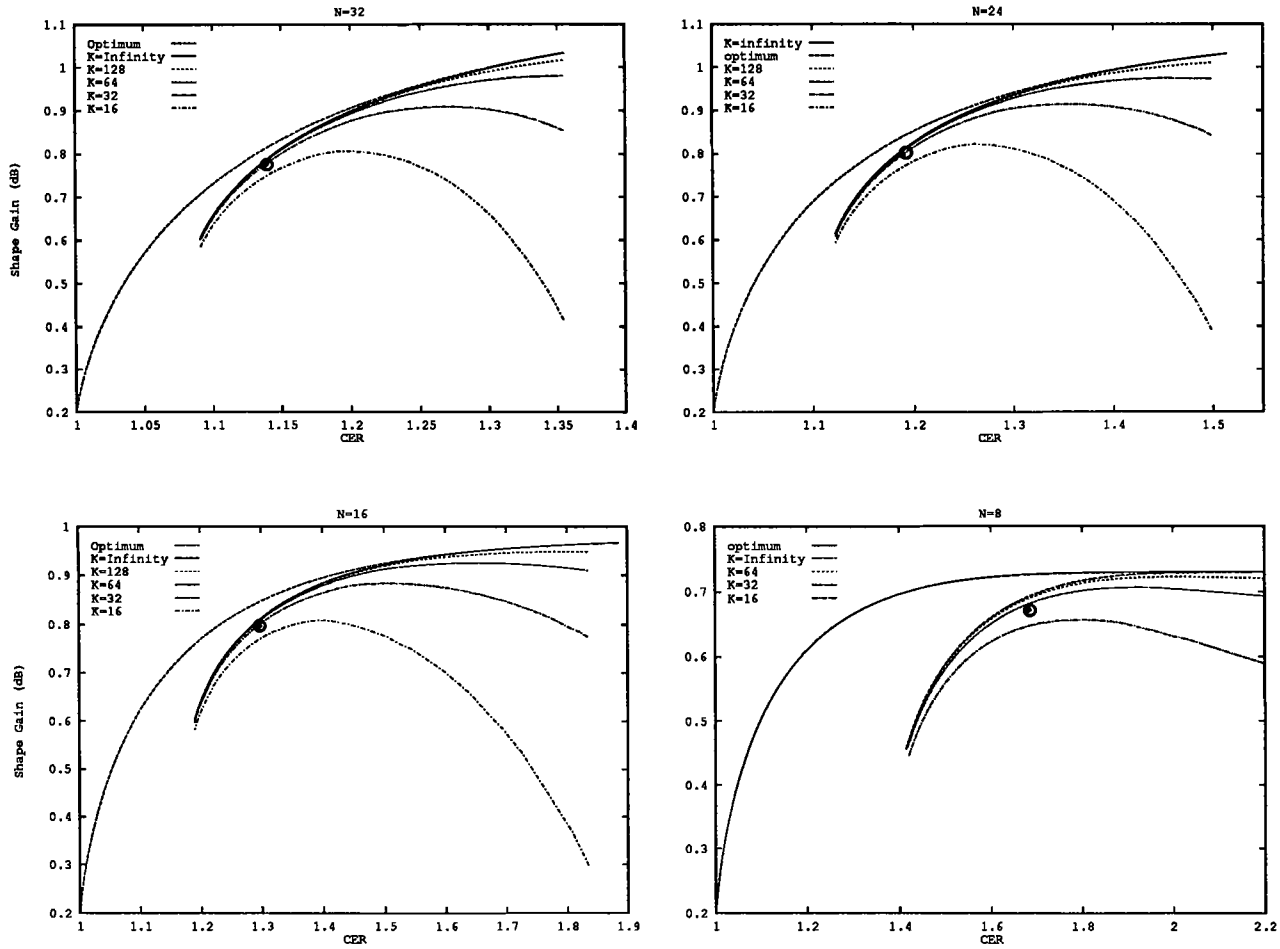


Fig. 4.5: Tradeoff between the CER_s and γ_s in the $\mathcal{QA}_N^{N/2}(K,T)$ regions.

space. Analytical expressions are derived for the corresponding tradeoff. In general, the initial part of the tradeoff curves has a steep slope. This means that an appreciable amount of the maximum shape gain, corresponding to a spherical region, can be achieved by a small value of CER_s and PAR. We have presented an addressing scheme to achieve a point on the optimum tradeoff curves. For dimensionality less than 16, the point achieved is near the knee of the curve. For higher dimensionalities, we use a more general shaping region to achieve a point near to the knee of the optimum curves.

Chapter 5

Shaping of Sets, Addressing Decomposition

Part of this chapter have been reported in [25], [28], [29], [30], [31].

5.1 Introduction

The structure of the optimum shaping regions is discussed in the previous chapter. In general, the initial portion of the optimum tradeoff curves has a steep slope. This means that an appreciable part of the maximum shape gain, γ_s, corresponding to a spherical region, can be achieved with a relatively small CER_s and PAR. In this chapter, we introduce some practical methods to achieve this goal.

We discuss using a lookup table to move along the optimum tradeoff curves and a method to facilitate the hardware realization of this lookup table. We also introduce a method to decompose the addressing by a lookup table into a hierarchy of the addressing steps each of a low dimensionality. As the memory size has an exponential growth with the dimensionality, this scheme results in a substantial decrease in the complexity. In this case, by using a memory of a practical size, we can move along a tradeoff curve which has a negligible suboptimality. This scheme is completely adaptable to the coding schemes of

[42]. This scheme makes it possible to achieve the shaping in spaces of high dimensionality and use the subspaces of lower dimensionality for the channel coding.

Another class of addressing schemes is based on using a Voronoi constellation in a space of half the original dimensionality. In this case, we introduce a method to achieve a single point on the optimum tradeoff curves. This point has significant shape gain with low addressing complexity and low CER_s and PAR. Finally, in a multi-level addressing scheme, we combine the Voronoi constellations of the previous method with a lookup table to move along a curve which is nearly optimum. To further reduce the complexity of this method, we replace the lookup table by a Voronoi constellation and thereby achieve a single point near to the knee of the optimum tradeoff curve.

We use both continuous approximation and discrete analysis to calculate the performance of our schemes. Usually, the continuous approximation underestimates the actual performance. To justify this effect, consider a shaping region which is the union of some unity volume Voronoi regions. The centroid of the Voronoi regions are the constellation points. Using the orthogonality principle, the average second moment of the region with respect to the origin is the sum of two terms, namely, the average second moment of the centroids and a second term which is the second moment of a single Voronoi region. In the discrete analysis, the second term is not present.

First, we give some definitions.

Discrete cube, $C_n(K)$:

This is equal to the cartesian product $\{C_1(K)\}^n$ where $C_1(K) = \{J + 0.5, J = 0, \ldots, K-1\}$.

The point $\vec{J} + (0.5)^n \in C_n(K)$ is indexed by $\vec{J} = (J_0, \ldots, J_{n-1})$.

We have $|C_n(K)| = |C_1(K)|^n = K^n$.

Symmetrical discrete cube, $SC_n(K)$:

This is equal to $\{SC_1(K)\}^n$ where, $SC_1(K) = \{\pm(J + 0.5), J = 0, \ldots, K-1\}$.

We have $|SC_n(K)| = |SC_1(K)|^n = (2K)^n$.

First moment shell, $F_n(K, L)$: This is the first moment shell of $C_n(K)$, i.e.,

$$F_n(K, L) \equiv \left\{ \vec{J} + (0.5)^n \in C_n(K) : \sum_{p=0}^{n-1} J_p = L \right\} . \tag{5.1}$$

Truncated discrete cube, $TC_n(K,L,T)$:

This is the set of $F_n(K,l)$'s, $0 \le l \le L-1$, and a selected subset of $F_n(K,L)$ such that $|TC_n(K,L,T)| = T$. We use the notation $TC_n(K,L)$ when the $F_n(K,L)$ is completely included. It is easy to show that,

$$\begin{aligned} |F_n(K,L)| &= \sum_{l=0}^{K-1} |F_{n-1}(K,L-l)|\,, \\ |TC_n(K,L)| &= \sum_{l=0}^{L} |F_n(K,l)|\,. \end{aligned} \tag{5.2}$$

5.2 Shell mapped constellations, A_N

The A_N constellations are based on a shaping region as close as possible to the optimum shaping region $\mathcal{A}_N(\psi)$ introduced in the previous chapter. Let's $\mathcal{S}_N(R)$ denotes an N-D (N-dimensional) sphere of radius R. In $\mathcal{A}_N(\psi)$ region, an $\mathcal{S}_2(R)$ is the boundary of the 2-D subspaces and an $\mathcal{S}_N(\sqrt{n\psi}R)$, $n = N/2$, is the boundary of the whole space. In Chapter 4, by a change of variable denoted as the shell mapping, the energy shells of $\mathcal{S}_2$'s are mapped to the points along a dimension. As a result, the N-D space (N-domain) is mapped to an $n = N/2$-D space (n-domain). The $\mathcal{A}_N$ region is mapped to an n-D hypercube of edge length one truncated to a simplex of edge length $n\psi$. This is denoted by $\mathcal{TC}_n(1, n\psi)$. This mapping has a useful property that a uniform density of points within $\mathcal{A}_N$ results in a uniform density of points within the $\mathcal{TC}_n$. Using this property, shaping is achieved by partitioning the n-domain into equal volume partitions and then selecting a subset of them. For $\psi = 1/2$, we have another interesting property that $\mathcal{TC}_n(1, n/2)$ is equal to the Voronoi region of the lattice D_n^* in the positive coordinates. In this chapter, this property is extensively used to facilitate the addressing and also the partitioning of the n-domain.

In the following, the idea of shell mapping is extended to the discrete case. This is achieved in two steps. In the first step, the region $\mathcal{TC}_n$ is replaced by a discrete set TC_n. In the second step, the circular region $\mathcal{S}_2(R)$ is replaced by the circular constellation $S_2(M)$, M denotes the cardinality.

Step I:

Using K concentric circles, each $\mathcal{S}_2(R)$ is partitioned into K shells of equal volumes. These are indexed by $J = 0, \ldots, K-1$. The radius and the average energy of the J'th shell satisfy, $R(J) = \sqrt{J+1}\Delta R$ where $\Delta R = R/\sqrt{K}$ and $E(J) = (J+0.5)(\Delta R)^2$. The shells are mapped to the points $Y = J + 0.5$. This results in the set $C_n(K)$ in the n-domain. Each point of $C_n(K)$ corresponds to a shaping cluster of volume $\pi^n(\Delta R)^N$ in the N-domain. The average energy of the cluster indexed by $\vec{J}$ is, $E(\vec{J}) = (\Delta R)^2(0.5n + \sum_p J_p)$. This means that the points located on $F_n(K,l)$'s represent the clusters with the equal energy. Using this fact, the shaping is *optimally* achieved by selecting a subset of these points with the least $\sum_p J_p$. This results in the shaping set $TC_n(K,L)$ in the n-domain. The overall shaping is optimum to the extent that the resolution of the partitioning of the 2-D subspaces allows.

It is easy to show that assuming $R = 1$, the average energy per two dimensions of the selected subset is equal to,

$$P_2 = \frac{1}{nK|TC_n(k,L)|}\left[\sum_{l=0}^{L} l\ |F_n(K,l)| + \frac{n}{2}|TC_n(k,L)|\right]. \tag{5.3}$$

This formula together with $\mathrm{CER}_s = K/|TC_n(K,L)|^{1/n}$ are used to calculate the tradeoff curves given in Fig. 5.1. The two discrete set of points correspond to the discrete analysis with $M = 128$ points per two dimensions. The computational method will be explained later. The optimum curves are extracted from Fig. 4.2.

The lookup table is a block of $|TC_n(K,L)|$ memory locations each with $n\log_2 K$ bits. Fig. 5.2 shows the tradeoff between the γ_s and the size of the memory. Referring to Fig. 5.1, for small values of CER_s (which are of practical interest), $K = 4$ achieves almost all the shape gain. This effect is also observed in [1]. However, referring to Fig. 5.2, $K = 4$ results in a substantial decrease in the memory size comparing to $K = 8$.

Step II:

We assume that the projection of the constellation on the 2-D subspaces is a finite portion of $Z^2 + (1/2)^2$ where $Z^N + (1/2)^N$ denotes the N-D half integer grid. This assumption

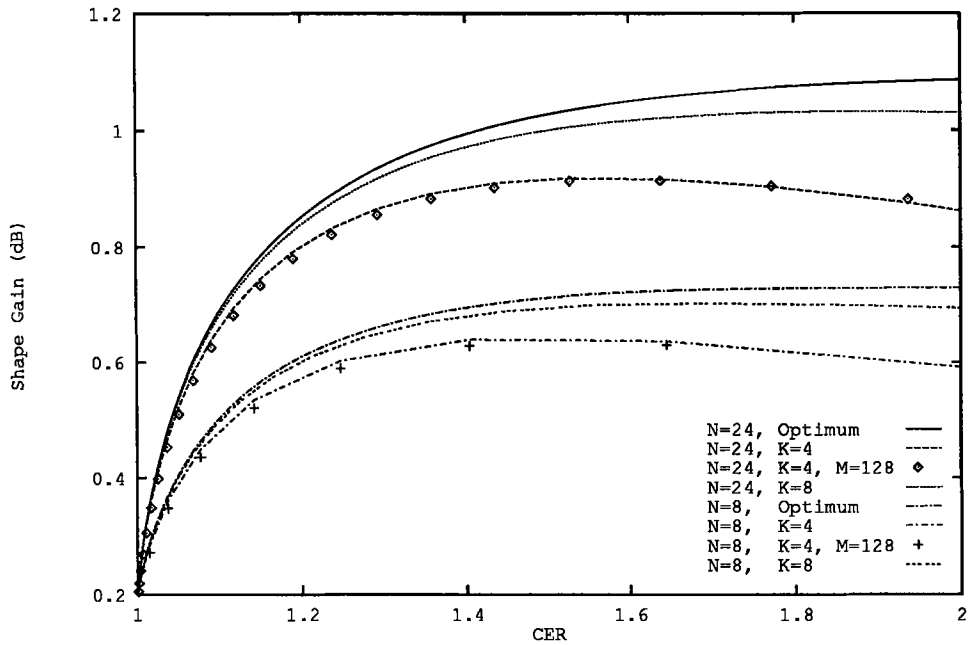


Fig. 5.1: Tradeoff between CER_s and γ_s in A_N constellations $N = 8, 24$, $K = 4, 8$.

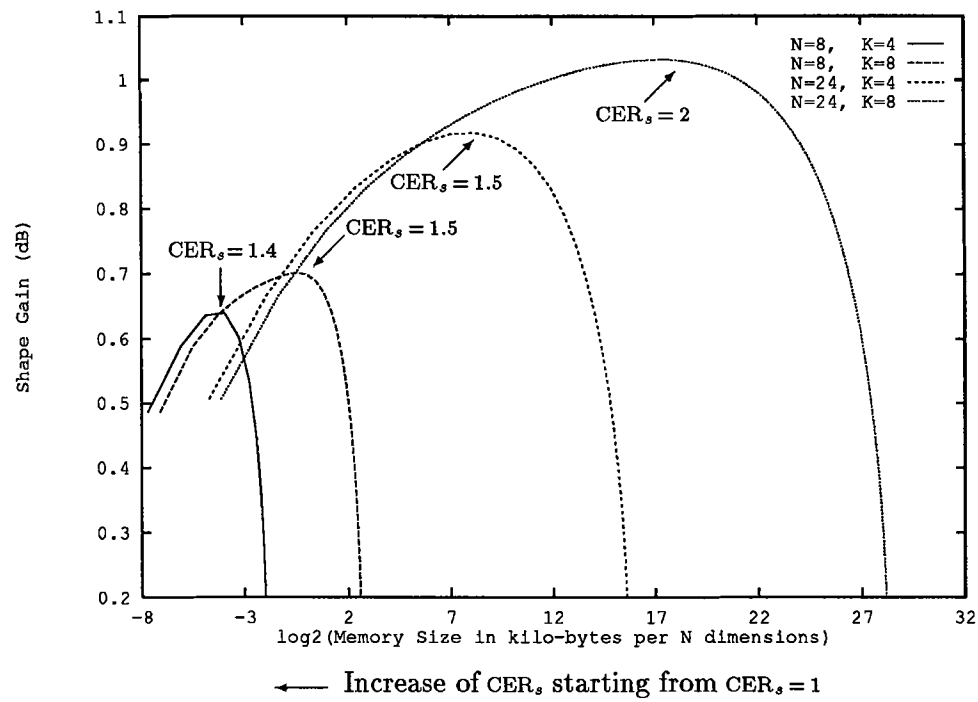


Fig. 5.2: Tradeoff between the memory size and γ_s in A_N constellations, $N = 8, 24$, $K = 4, 8$.

holds for most of the dense lattices, [42]. The points of $Z^2+(1/2)^2$ are grouped in the order of the increasing energy into K shells each with $2^p = M/K$ points. Each shell has four way symmetry and contains an equal number of points from each partition in an Ungerboeck partition chain. These are important issues for practical implementation of a multi-dimensional trellis coding scheme, [42]. The J'th shell is mapped to the point $Y = J + 0.5$ along a dimension. The shaping set in the n-domain is selected as,

$$TC_n(K, L, 2^t) = \{\vec{Y} : \vec{Y} \in F_n(K, l),\ l < L \ \text{ or } \ \vec{Y} \in F_n(K, L),\ E(\vec{Y}) \leq \Theta\}\ , \qquad (5.4)$$

where L selects the first fully filled shells and Θ selects the least energy points on the last shell. This method is not necessarily optimum because, unlike the case of Step I, the points located on the F_n's do not represent the N-D clusters with the equal energy. But, since the energy differences are small, the suboptimality is negligible.

For $p=2$ (shells of four points), the energy of the points within each shell are the same. This results in the best shaping ability in the n-domain. By changing M for a fixed K, we can change the total rate of the constellation for fixed lookup table complexity, fixed CER_s and *essentially* fixed γ_s. For continuous approximation in the 2-D subspaces (Step I), γ_s remains fixed.

The whole constellation is denoted as $A_N(M, K, 2^t)$, $|A_N(M, K, 2^t)| = 2^{t+pn}$, $2^p = M/K$, $n = N/2$, and $\text{CER}_s = K \times 2^{-t/n}$. The lookup table has t input and $n\lceil \log_2 K \rceil$ output lines.

Example:

Figure 5.3 shows the structure of the A_4 constellation, $M = 32$, $K = 4$. The average energy of the shells are, $\{E(J)\} = \{1.5, 3.5, 6.5, 8.5\}$. Assuming continuous approximation (Step I), we obtain, $(\Delta R)^2 = 8/\pi$ and $\{E(J)\} = \{1.27, 3.82, 6.37, 8.91\}$. The difference between the elements of these two sets is the main cause for any suboptimality. The available signal space in the n-domain is the set $C_2(4)$. Each point of the n-domain corresponds to $8 \times 8 = 64$ points in the N-domain. The shaping set in the n-domain is equal to $TC_2(4, L)$, $1 \leq L \leq 6$. The two dotted lines corresponds to shaping sets $TC_2(4, 1)$ and $TC_2(4, 4)$, respectively. For the solid line, we have the Voronoi region of $D_2^* = \Re Z^2$ where $\Re$ denotes the rotational operator, [12].

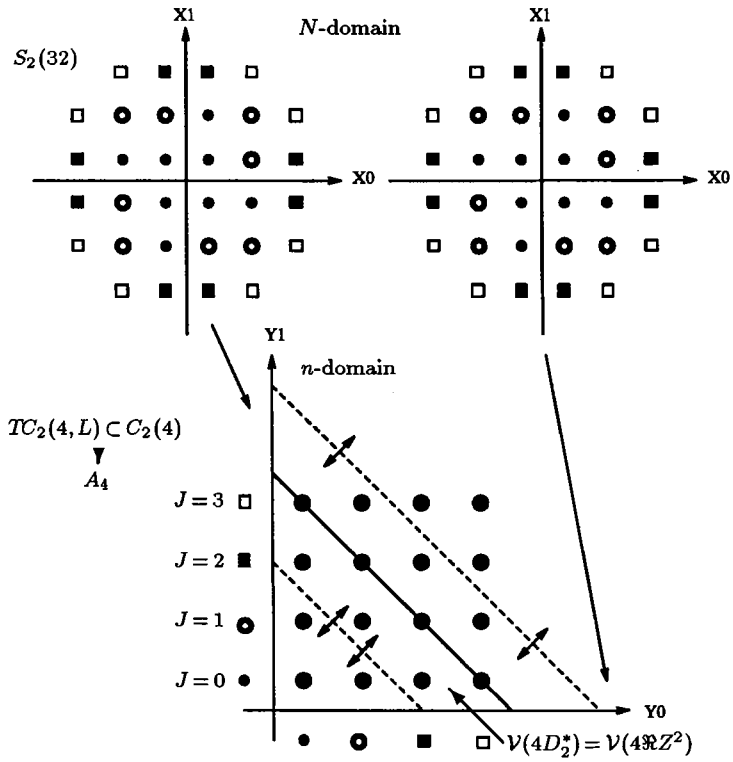


Fig. 5.3: Example of A_4 constellation, one-level shell mapping, $\mathcal{V}(\Lambda)$ denotes the Voronoi region around the origin of the lattice Λ.

In the following, we introduce a method to facilitate the hardware realization of the addressing lookup table in the A_N constellations.

5.2.1 Structure of the lookup table for the A_N constellations

We assume that the shaping set in the n-domain is equal to $TC_n(K,L) \subset C_n(K)$. We use a property of $TC_n(K,L)$ to simplify the hardware realization of the lookup table. In general, the projection of $TC_n(K,L)$ on any m-D subspace is equal to the set $TC_m(K,L)$. We partition the projections on the 2-D subspaces, which are TC_2's, into subsets such that the addressing can be achieved directly on them. This results in a logical table with many 'don't care' entries. We first show that for a uniform probability density on the points of the A_N, the induced probability density on the TC_2's depends only on the sum of the coordinates.

The dimensions of the n-domain are labeled by $Y_0, \ldots, Y_{n-1}$. To compute the induced probability density on the points of $TC_2(K,L)$ in the Y_0, Y_1 subspace, we draw from every point $(Y_2, \ldots, Y_{n-1}) \in TC_{n-2}(K,L)$, a 2-D plane parallel to the Y_0, Y_1 subspace and find the part of it which is located inside of $TC_n(K,L)$. The intersection of such a plane with the $TC_n(K,L)$ is the set $TC_2(K, L - \sum_{p=2}^{n-1} J_p)$ where $J_p = Y_p - 0.5$. The points of this set are mapped to the Y_0, Y_1 subspace. By counting the number of times that a given point is used, the induced probability density is calculated. The total number of times that the set $F_2(K,b)$ is used is equal to the total number of times that the sets $TC_2(K,a)$, $a \geq b$ are used. The number of times that the set $TC_2(K,a)$ is used is equal to $|F_{n-2}(K, L-a)|$. Using this fact, the frequency of $F_2(K,b)$ is found as,

$$N_2(b) = \sum_{b \leq a \leq L} |F_{n-2}(K, L-a)| = \sum_{l=0}^{L-b} |F_{n-2}(K,l)| = |TC_{n-2}(K, L-b)| \,. \tag{5.5}$$

From 5.5, it is seen that the frequency of the points of $F_2(K,b)$ are equal to each other. This means that if we partition the 2-D subspaces of $TC_n(K,L)$ into $F_2(K,l)$, l=0,...,min$(2K-2, L)$, the set $TC_n(K,L)$ can be expressed as a subset of the $n/2$-fold cartesian product of these partitions. In practice, we should further partition each $F_2(K,l)$ such that the number of points in each of the final partitions is an integral power of two.

Similarly, the frequency of the 2-D shells are found as,

$$N_1(J) = |TC_{n-1}(K, L-J)| \,. \tag{5.6}$$

The corresponding probabilities are,

$$P(J) = \frac{|TC_{n-1}(K, L-J)|}{\sum_{J=0}^{K-1} |TC_{n-1}(K, L-J)|}, \tag{5.7}$$

which results in the average power, $P_2 = \sum_{J=0}^{K-1} P(J)E(J)$. This is used to calculate the discrete set of points in Fig. 5.1.

We already claimed that the continuous approximation underestimates the actual performance. In Fig. 5.1, for low values of CER$_s$, this is not the case. This is mainly

due to the suboptimality of the shaping set in the n-domain. However, it is seen that the difference is negligible.

Example:

In the constellation $A_N(M,K,2^t)=A_8(128,4,64)$, the 2-D subconstellations of 128 point are partitioned into $K=4$ shells each of $2^p=32$ points. The average energy of the shells are equal to, $\{E(J)\}=\{5,15.5,25.5,36\}$. Assuming continuous approximation, we obtain, $\{E(J)\}=\{5.09,15.28,25.46,36.65\}$. There are $K^n=4^4=256$, N-D partitions and shaping is achieved by selecting $2^t=64$ of them. We have $|A_N|=26$ bits and $r_s=2$ bits, $\text{CER}_s=1.41$. The lookup table has 6 input and 8 output lines. The output lines are divided into four groups each with two lines. These are used to select a shell within each 2-D subspace. Another group of 20 bits, divided into 4 groups of 5 bits, is used to select a point within each 2-D shell.

To specify the $TC(4,L,64)$, we need to find L. Using (5.2), we obtain,

$$\{F_4(4,l)\}=\{1,4,10,20,31,40,44,40,31,20,10,4,1,0,0,\ldots\}\ . \tag{5.8}$$

It is seen that $\sum_{l=0}^{4}F_4(4,l)=66$. This means that $L=4$ and only 29 points of the 31 points in $F_4(4,4)$ are included in TC_4. In the following we discuss how to select these points.

If all the points of $F_4(4,4)$ were included, the frequencies of the 2-D shells would be, $\{N_1(J)\}=\{32,20,10,4\}$. The two points of $F_4(4,4)$ with the highest $E(\vec{J})$ are the points $\vec{J}=(1,1,1,1)$ with $E(\vec{J})=62$ and the point $\vec{J}=(0,1,1,2)$ with $E(\vec{J})=61.5$,[1]. If we discard these two points, the induced probability density on different dimensions of $TC_4(4,4,64)$ will be no longer the same. For the first dimension the frequencies are $\{32,18,10,4\}$ and for the second, third and the fourth dimensions $\{32,18,10,4\}$, $\{32,19,9,4\}$ and $\{31,19,10,4\}$, respectively. This results in the average energy of 13.094, 13.094, 12.938 and 13.254 along the first to the fourth 2-D subspaces. The overall average

[1]There are 24 points with $E(\vec{J})=61.5$. These are the points indexed by the permutations of $(0,1,1,2)$ and $(0,0,1,3)$, twelve points from each.

energy is $P_2 = 13.095$ resulting in $\gamma_s = 0.614$ dB. Using continuous approximation, the maximum shape gain for $N = 8$ and $\text{CER}_s = \sqrt{2}$ is $\gamma_s = 0.698$ dB.

The finest partitioning is obtained by $K = 32$ which results in $\gamma_s = 0.727$ dB. This requires a lookup table with 18 input and 20 output lines. The size of the memory with respect to $K = 4$ (6 input and 8 output lines) has increased by the multiplicative factor 10240. As a result of this large increase in the complexity, the shape gain has just increased by about 0.1 dB. This justifies our previous claim that $K = 4$ is a reasonable choice.

To build up the lookup table, we partition $F_2(4, l), l = 0, 1, 2, 3$, of $TC_2(4, 4)$ into subsets each with an integral bit rate, Fig. 5.4. The first moment shells are denoted by A, B, C, D and E. The addressing subsets have the same subscript and also the same sign. If $F_4(4, 4)$ were included completely in $TC_4(4, 4, 64)$, we could specify all the points of

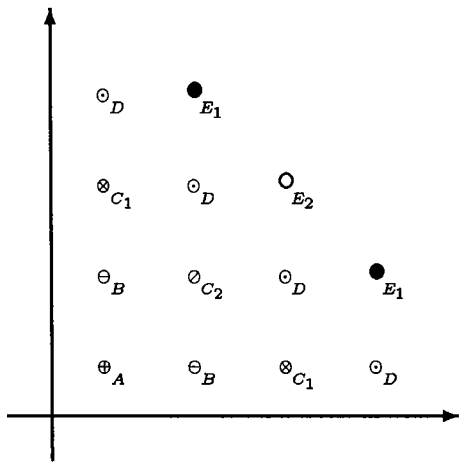


Fig. 5.4: The $TC_2(4, 4)$ partitioned into the addressing subsets.

$TC_4(4, 4, 64)$ with this partitioning. For this example, we need a finer partitioning. This can be avoided by discarding the point indexed by $(0, 0, 2, 2)$ with $E = 61$,[2] instead of the point indexed by $(1, 1, 2, 0)$ with $E = 61.5$. This results in the average power $P_2 = 13.097$. The loss in the shape gain comparing to the previous case ($P_2 = 13.096$) is negligible. Table 5.1 shows the index vector of the points of TC_4. The indices are obtained from the

[2]There are six points with $E = 61$. These are the points indexed by the permutations of $(0, 0, 2, 2)$.

Points	$\mathcal{N}$	E	Points	$\mathcal{N}$	E
0000	1	20.0	0012	12	51.0
0001	4	30.5	0111	4	51.5
0002	4	40.5	0022	5	61.0
0011	6	41.0	0013	12	61.5
0003	4	51.0	0112	12	61.5

Table 5.1: Points of the shaping set in the n-domain of the $A_8(128, 4, 2^6)$ constellation.

permutations of the given vectors. Also given is the number of points obtained from each vector, $\mathcal{N}$, and the average energy of the N-domain clusters, E.

We reach the conclusion that the points of $F_n(K, L)$ have almost the same E and selecting an arbitrary subset of them results in negligible degradation. This subset should be selected on the basis of the addressing complexity.

Table 5.2 contains a prefix code for the addressing in the n-domain. The don't care entries can be used to construct a logical table of reduced complexity. Of course, for a constellation like $A_{24}(128, 4, 2^{22})$ which needs a lookup table with 22 input and 24 output lines, the effect of the 'don't care' entries will be more pronounced.

5.3 Address decomposition

For a fixed rate per dimension, the complexity of an addressing scheme using a lookup table grows exponentially with the dimensionality. This can result in an impractically large memory. In this section, we describe a method to decompose the addressing into steps of low dimensionality and thereby avoid the exponential growth of the complexity. Consider an N'-D unshaped constellation, i.e., $A_{N'} = \{S_2\}^{N'/2}$. This constellation is partitioned into K energy shells of equal volume. The 2-fold cartesian product of the set of the partitions is shaped by using a lookup table. Assuming continuous approximation,

BD	$\times$	$\times$	$\times$	0	0	0	C_1A	$\times$	1	1	1	0	1
DB	$\times$	$\times$	$\times$	0	0	1	AE_1	$\times$	0	0	1	1	0
AD	$\times$	$\times$	0	0	1	0	E_1A	$\times$	0	1	1	1	0
DA	$\times$	$\times$	1	0	1	0	C_1C_2	$\times$	1	0	1	1	0
BC_1	$\times$	$\times$	0	0	1	1	C_2C_1	$\times$	1	1	1	1	0
C_1B	$\times$	$\times$	1	0	1	1	BC_2	$\times$	0	0	1	1	1
BB	$\times$	$\times$	0	1	0	0	C_2B	$\times$	0	1	1	1	1
C_1C_1	$\times$	$\times$	1	1	0	0	AC_2	0	0	1	1	1	1
AB	$\times$	0	0	1	0	1	C_2A	1	0	1	1	1	1
BA	$\times$	1	0	1	0	1	AA	0	1	1	1	1	1
AC_1	$\times$	0	1	1	0	1	AE_2	1	1	1	1	1	1

Table 5.2: A prefix code for the addressing in the n-domain of the $A_8(128, 4, 2^6)$ constellation. The '$\times$' denotes the 'don't care' entries.

the calculation of the tradeoff is quite similar to the one presented in section (4.3). The final result is shown in Fig. 5.5.

The main point is that for a moderate value of K, we can essentially achieve the optimum tradeoff. This phenomenon can be considered as a generalization of a similar effect observed over dimensionality two in [1]. This property allows us to decompose the addressing of a constellation into some intermediate steps achieved on the 2-fold cartesian product of a set with low cardinality[3]. We call this method as the *address decomposition*. For a dimensionality $N = 2^u$, this results in $u-1$ addressing steps. The i'th step, $i \in [0, u-1]$, is achieved on the 2^i-D subspaces and results in dimensionality 2^{i+1}. We assume that the subspaces involved in the i'th step are partitioned into $K_i = 2^{k_i}$ shells. Referring to Fig. 5.5, we select $\{K_i, i = 1, \ldots\} = \{64, 64, 128, 256, \ldots\}$. Actually,

[3]We already observed in Fig. 4.5 that the same property is valid for the $\mathcal{A}_N^{N/2}(1/2, \psi')$ regions. This means that the address decomposition procedure discussed here can be applied to that case too.

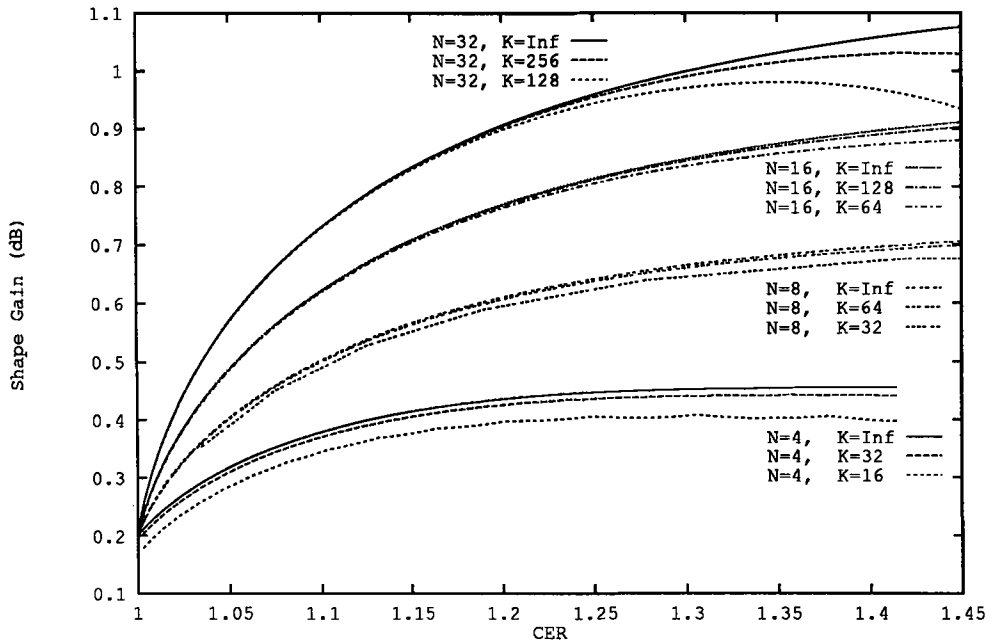


Fig. 5.5: Tradeoff between CER_s and γ_s using a finite number of the energy shells in the $N/2$-D subspaces.

for a 2-D subconstellation with M points, one can achieve the maximum shape gain with $K_1 = M/4$ shells. This is usually less than $K_1 = 64$.

The i'th addressing step requires a memory with $2k_i \times 2^{2k_i}$ bits. The last step requires $2k_i \times 2^{2k_i - r_s}$ bits. An upperbound to total memory size, M_s, is obtained by setting $r_s = 0$. As an example, for $N = 32$, the upperbound is equal to, 168 kilo-bytes per N dimensions.

Figure 5.6 shows the final tradeoff curves. It is seen that the suboptimality is negligible. This addressing scheme does not have the problem of ties or the constraint on the constellation total rate as encountered in the Voronoi constellations. Also, it can be easily used in conjunction with the coding schemes of [42].

An alternative to the addressing by a lookup table is the use of a Voronoi constellation, [5], [10]. In the following, the idea of the shell mapping is extended to this case.

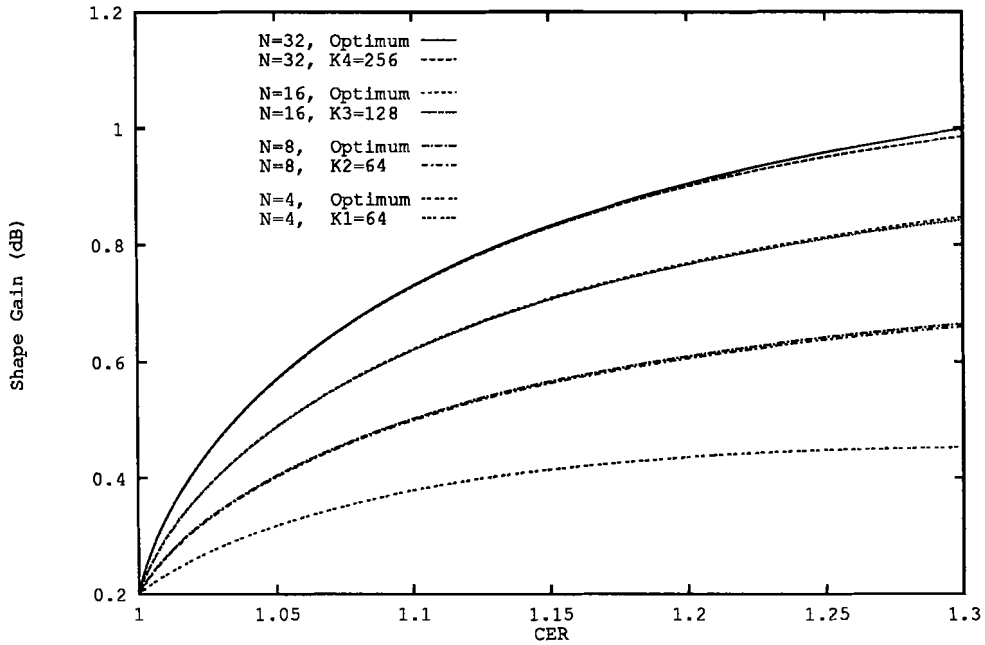


Fig. 5.6: Tradeoff between CER_s and γ_s using the address decomposition method.

5.4 Shell-mapped Voronoi constellations

Consider a lattice Λ_n^s, such that the projection of its Voronoi region on any dimension is the region $[-1,1]$. Let's $V_n(K\Lambda_n^s)$, denote a subset of $Z^n+(1/2)^n$ bounded within the Voronoi region around the origin of $K\Lambda_n^s$. This is the Voronoi constellation based on the half integer grid $Z^n+(1/2)^n$ and the shaping lattice $K\Lambda_n^s$,[4]. The shell-addressed Voronoi constellations are based on using the points of such a set to shape the n-domain. The complexity of the addressing is that of a linear mapping plus the complexity of decoding of Λ_n^s. Assuming binary lattices, to have an integral total rate, M and K should be integral powers of two, i.e., $M=2^m$ and $K=2^k$.

In the A_N constellations, the available signal space in the n-domain is restricted to positive coordinates. To obtain symmetry, the 2-D shells are further partitioned into two subshells each with 2^{p-1} points. The two subshells of the J'th shell are mapped to the point $Y=\pm(J+0.5)$. This results in the set $SC_n(2^k)$ in the n-domain. Shaping is

[4]For a brief description of the Voronoi constellations refer to Appendix G.

achieved by selecting the set $V_n(2^k\Lambda_n^s) \subset \mathcal{SC}_n(2^k)$ to shape the n-domain. It is easy to show that,

$$\mathrm{CER}_s = \frac{2}{(|Z^n/\Lambda_n^s|)^{1/n}}\,, \tag{5.9}$$

and,

$$\gamma_s = \frac{n\pi\left[V\left(|Z^n/\Lambda_n^s|\right)\right]^{1+1/n}}{12F_m\left[\mathcal{V}_n(\Lambda_n^s)\right]}\,, \tag{5.10}$$

where,

$$F_m(\mathcal{R}) = \int_{\mathcal{R}} |Y_0| + \ldots + |Y_{n-1}|\; dY_0 \ldots dY_{n-1}\,, \tag{5.11}$$

is the absolute first moment of the region $\mathcal{R}$.

It is seen that γ_s is determined by the absolute first moment of the n-domain. Consequently, in selecting Λ_n^s, one should try to minimize the absolute first moment of the lattice Voronoi region for a given volume. A lattice with a pyramid Voronoi region results in a spherical constellation. We know that such a lattice exists *only* in dimensionality two, [6]. An analytical method to calculate the absolute first moment of the lattices is given in Appendix D. Numerical results are presented for the lattices D_n and $\Re D_n$.

For the lattice D_n, using $|Z^n/D_n| = 2$ in 5.9, we obtain,

$$\mathrm{CER}_s = 2 \times 2^{-1/n}\,, \tag{5.12}$$

and using the result of Appendix D for F_m in (5.10), we obtain,

$$\gamma_s = \frac{\pi}{3} \times \frac{n(n+1)(2)^{1/n}}{n^2+n+2}\,. \tag{5.13}$$

For example, for $N = 8$ $(n = 4)$, we have, $\mathrm{CER}_s = 1.682$ and $\gamma_s = 0.54$ dB.

For the lattice $\Re D_n$, using $|Z^n/\Re D_n| = 2 \times (2)^{n/2}$ in (5.9), we obtain,

$$\mathrm{CER}_s = \sqrt{2} \times (2)^{-1/n}\,, \tag{5.14}$$

and using the result of Appendix D for F_m in (5.10), results in the shape gains of Table 5.3.

N	CER_s	γ_s dB
8	1.189	0.602
12	1.260	0.658
16	1.297	0.655

Table 5.3: Shape gain of the shell-addressed Voronoi regions based on the lattice $\Re D_n$.

For $\Lambda_n^s = D_n^*$, we obtain the constellation $A_N(2^m, 2^k, 2^{kn-1})$. This corresponds to the point $r_s = 1$ bit/N-D, $\text{CER}_s = (2)^{1/n}$, on the optimum tradeoff curves. These are the A-points in Table 4.1. For $N = 4$, $n = 2$, we have $D_2^* = \Re Z^2$ and the constellation is spherical. This is the case in Fig. 5.3.

The major complexity in a Voronoi constellation is that of decoding the shaping lattice. The decoding of D_n^* is efficiently achieved using the following definition, [4],

$$D_n^* = \{(2Z)^n\} \cup \{(2Z)^n + (1)^n\} \ . \tag{5.15}$$

To decode a vector $\mathbf{x}$, we first find the two nearest integers on the two sides of each component of $\mathbf{x}$. Let's x_i^e and x_i^o denotes these integers along the i'th dimension where superscripts e/o stands for even/odd. The point $\mathbf{x}^e = (x_i^e, i = 0, \ldots, n-1)$ is the nearest point of $2Z^n$ to $\mathbf{x}$. Similarly, the point $\mathbf{x}^o = (x_i^o, i = 0, \ldots, n-1)$ is the nearest point of $2Z^n + (1)^n$ to $\mathbf{x}$. The nearest of the two points $\mathbf{x}^e$, $\mathbf{x}^o$ is the nearest point of the lattice D_n^* to $\mathbf{x}$.

The decoding of D_n^* is much simpler than the decoding of the popular lattices used in the standard Voronoi constellations. For example, consider the lattices E_8 and Λ_{24} (Leech lattice). The decoding of these lattices is achieved by decoding their trellis diagram. The E_8 lattice has a 4-state trellis and the Λ_{24} lattice has a 256-state trellis, [13].

In the following, we show that the point corresponding to the lattice D_n^* is the only nontrivial point that a shell-mapped Voronoi constellation can achieve on the optimum tradeoff curve. Another point is the trivial case of a cubic constellation.

Referring to the definition of the region $\mathcal{TC}_n$ in (4.4), to achieve an optimum point,

the first condition is that the points $[0^{n-1}, \pm 2]$, [.] denotes the set of all the points obtained by the permutations of the components within [.], should be the nearest points to the origin along each dimension of Λ_n^s. Also, to realize a point with the parameter $1/n \leq \psi \leq 1$, we should have, $[(\pm 2\psi)^n] \in \Lambda_n^s$. Using the group property of the lattice, this requires that $[0^{n-1}, \pm(4-4\psi)] \in \Lambda_n^s$ and also $[0^{n-1}, \pm 4\psi] \in \Lambda_n^s$. This contradicts the first condition for all the range of $\psi \leq 1$ except for $\psi = 1/2$ and $\psi = 1$ where $\psi = 1$ results in a cubic constellation.

For $N < 16$, the point achieved by the shell-mapped Voronoi constellation based on the lattice D_n^* is located near the knee of the tradeoff curve. As N increases, this point moves toward the initial parts of the curve. It should be mentioned that in a practical scheme, this part of the curve may be the most interesting part.

5.5 Two-level shell-mapped constellations

In the following, we introduce a method to achieve a higher γ_s for $N \geq 16$. This is based on a multi-level addressing procedure which combines the shell-mapped Voronoi constellation method with a lookup table to move along a curve which is nearly optimum. In this case, the addressing by the lookup table is achieved in a space of dimensionality two and has a low complexity. In the rest of the chapter, we make frequent use of Fig. 5.7 to explain our schemes. The actual values corresponding to this figure will be written inside of double braces $\{\!\{ \ . \ \}\!\}$.

The two level shell-mapped constellations $A_N^{N'}$ are based on a shaping region as close as possible to the region $\mathcal{A}_N^{N'}$ introduced in the previous chapter. These constellations provide a means to move along a curve which is nearly optimum. Examples of such curves are given in Figs. 4.3 and 4.5. The structure of the $\mathcal{A}_N^{N'}$ $\{\!\{A_8^4\}\!\}$constellations is as follows: A constellation $A_{N'}(2^m, 2^k, 2^{kn-1})$, $n = N'/2$, is employed along each N'-D subspace. The shaping set in the n-domain is equal to $V_n(2^k D_n^*)$ $\{\!\{V_2(4\Re Z^2)\}\!\}$. By using a partitioning lattice Λ_n^p $\{\!\{2Z^2\}\!\}$which has $2^k D_n^*$ $\{\!\{4\Re Z^2\}\!\}$as a sublattice, $V_n(2^k D_n^*)$ is partitioned into $2^{k'} = |\Lambda_n^p / 2^k D_n^*|$ $\{\!\{2^3 = |2Z^2/4\Re Z^2|\}\!\}$shaping clusters each with $2^{k''} = |Z^n/\Lambda_n^p|$

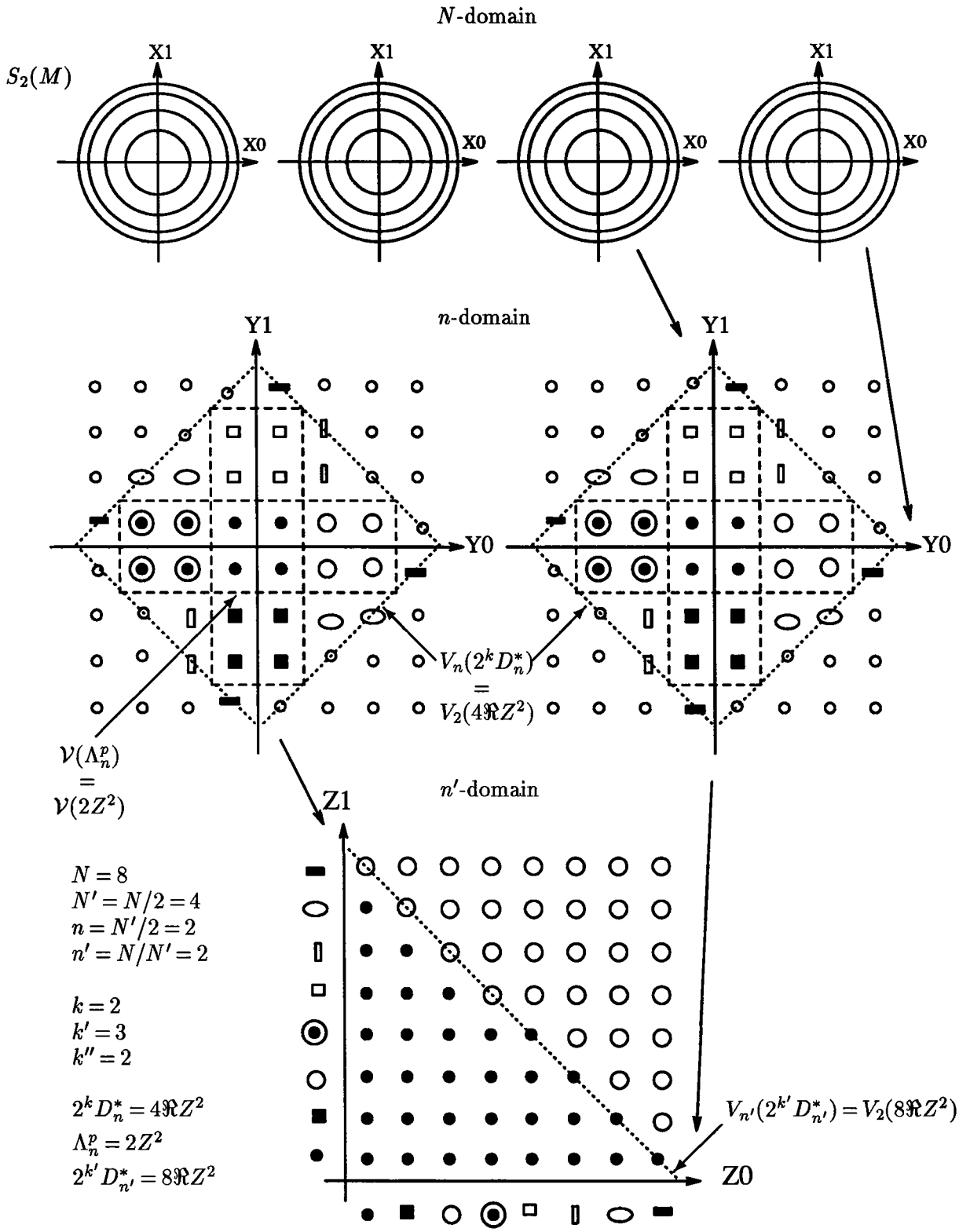


Fig. 5.7: Example of a multi-level constellation, $N = 8$, $N' = 4$, $n = 2$, $n' = 2$, $k = 2$, $k' = 3$, $k'' = 2$, $2^k D_n^* = 4\Re Z^2$, $\Lambda_n^p = 2Z^2$ and $2^{k'} D_{n'}^* = 8\Re Z^2$.

$\{\!\{2^2 = |Z^2/2Z^2|\}\!\}$points. This is based on the decomposition,

$$Z^n = 2^k D_n^* + [Z^n/\Lambda_n^p] + [\Lambda_n^p/2^k D_n^*] \ , \tag{5.16}$$

obtained from the partition chain $Z^n/\Lambda_n^p/2^k D_n^*$ $\{\!\{Z^2/2Z^2/4\Re Z^2\}\!\}$. In the matrix notation, we have,

$$Z^n = 2^k D_n^* + \mathbf{aG} + \mathbf{bH} \ , \tag{5.17}$$

where $\mathbf{G}$ and $\mathbf{H}$ are generators for $[Z^n/\Lambda_n^p]$ $\{\!\{[Z^2/2Z^2]\}\!\}$and $[\Lambda_n^p/2^k D_n^*]$ $\{\!\{[2Z^2/4\Re Z^2]\}\!\}$, respectively, $\mathbf{a}$ is a binary k''-tuple and $\mathbf{b}$ is a binary k'-tuple. Using (5.17), each coset of $[Z^n/2^k D_n^*]$ $\{\!\{[Z^2/4\Re Z^2]\}\!\}$is labeled by $(\mathbf{a}, \mathbf{b})$. Each shaping cluster in the n-domain is the set of the points with the same $\mathbf{a}$. In other words, $\mathbf{b}$ determines a cluster within the n-domain and $\mathbf{a}$ determines a point within that cluster.

The partitioning of the N'-D subspaces results in $2^{k'n'}$, $n' = (N/N')$ $\{\!\{2^6\}\!\}$shaping partitions in the N-domain. The third step of shaping is achieved by using a lookup table to select 2^t of these N-D partitions of the least average energy. The whole constellation is denoted by, $A_N^{N'}(2^m, 2^k, \Lambda_n^p, 2^t)$. The lookup table has t input lines and $k'n'$ output lines. The output lines are divided into n' groups. Each group is assigned to one of the N'-D subspaces and is used as the $\mathbf{b}$ part of the label in (5.17). Another $k''n'$ data bits, divided into n' groups, are used as the $\mathbf{a}$ part of the label in (5.17). Finally, another $N(p-1)/2$ data bits select one point within each 2-D subspace.

Referring to Figs. 4.3 and 4.5, it is seen that for relatively high CER_s, the sub-optimality of this addressing method is negligible. In this case, assuming the simplest partitioning lattice namely D_n with $|Z^n/D_n| = 2$, the size of the memory with respect to a direct addressing scheme decreases by the factor $2^{2n'}$. By using other partitioning lattices, one can further decrease the size of the memory at the cost of a small loss in γ_s.

The $B_N^{N'}$ $\{\!\{B_8^4\}\!\}$constellation is devised to achieve a single point near to the tradeoff curve of $A_N^{N'}$ *without using the lookup table.* These are the marked points in Fig. 4.5. The shaping/partitioning of the N'-D subspaces is the same as in the $A_N^{N'}$ constellations. However, in this case, the shaping clusters of the N'-D subspace are mapped in the order of the increasing energy to the points of $Z + (1/2)$ bounded within $\left[-2^{k'}, 2^{k'}\right]$ $\{\!\{[-2^3, 2^3]\}\!\}$.—In

Fig. 5.7 just the positive part of the n'-domain is shown. The positive and the negative points are mapped to the same cluster. Each point is labeled by the label of the corresponding cluster, namely the k'-tuple $\mathbf{b}$ in (5.17), and an extra bit which is selected according to the sign of the point. This sign bit is used as one bit of the k''-tuple $\mathbf{a}$ in (5.17). This results in the set $SC_{n'}(2^{k'})$ $\{\!\{SC_2(8)\}\!\}$, in the n'-domain. The shaping set in the n'-domain is selected as, $V_{n'}(2^{k'}D^*_{n'}) \subset SC_{n'}(2^{k'})$ $\{\!\{V_2(8\Re Z^2) \subset SC_2(8)\}\!\}$. In the $A_N^{N'}$ constellation, this part of the shaping was achieved by the lookup table.

In each signaling interval $n'(k'+1)-1$ data bits are used to select one point in the n'-domain. The label of each component of the selected point (which is the k'-tuple $\mathbf{b}$ plus a sign bit) with another $k''-1$ data bits are used in (5.17). To store the labels, we require a block of memory with $M_s = k' \times 2^{k'}$ bits (comparing to the $A_N^{N'}$ constellations which requires $n'k' \times 2^{n'k'}$ bits.).

The whole constellation is denoted by, $B_N^{N'}(2^m, 2^k, \Lambda_n^p)$. The total rate is $mN/2 - n' - 1$ and the shaping redundancy is $r_s = 1 + n'$. As in the case of the $A_N^{N'}$ constellations, we have the appropriate choice of $n' = 2$. In the sequel, we assume that $n' = 2$.

5.5.1 Performance measure

To calculate the shape gain, first, by using (5.17), the average energy of the N'-D points mapped to each shaping cluster in the n-domain are calculated. Then, by adding the average energies along different dimensions of the n'-domain, the average energy of the final subset of the N-domain is found.

As an example, Table 5.4 shows the shape gain of the $B_{16}^8(32, 4, \Lambda_4^p)$ constellation for different partitioning lattices. For this constellation, we have $r_s = 3$ ($\mathrm{CER}_s = 1.3$). By changing m, we can change the total rate of the constellation for fixed lookup table complexity, fixed CER_s and *essentially* fixed γ_s. As an example $B_{16}^8(64, 4, Z^4)$ results in $\gamma_s = 0.72$ dB and $B_{16}^8(128, 4, Z^4)$ results in $\gamma_s = 0.71$ dB.

Assuming continuous approximation, the maximum shape gain for $\mathrm{CER}_s = 1.3$, $N = 16$ is equal to, $\gamma_s = 0.85$ dB. Assuming $M = 128$ points per 2-D subconstellations, the ad-

Λ_4^p	k'	M_s	γ_s dB
D_4^*	8	0.25	0.62
$\Re Z^4$	9	0.57	0.66
D_4	10	1.25	0.70
Z^4	11	2.75	0.73

Table 5.4: Shape gain of the $B_{16}^8(32,4,\Lambda_4^p)$ constellation for $\mathrm{CER}_s = 1.3$, M_s denotes the required memory size in kilo-bytes per N dimensions.

dress decomposition method needs $M_s = 11$ kilo-bytes per N dimensions to achieve the optimum shape gain (within a small fraction of dB).

5.6 Multi-level shell-mapped constellations

As long as the shaping region in a domain is selected as the Voronoi region of a lattice, it can be easily partitioned into shaping clusters of equal volume. This provides us with a way to achieve another level of shaping/addressing on the cartesian product of the clusters. This can be done several times to produce a multi-level (nested) form of shaping. Similarly to the $B_N^{N'}$ constellations, this can be used to achieve *single points* with high shaping redundancy near to the optimum tradeoff curves.

The notation $B_N^{N_1,\ldots,N_q}(2^m, 2^k, \Lambda_{n_1}^p, \Lambda_{n_1}^p, \ldots, \Lambda_{n_q}^p)$ is used as the complete notation for this constellation. This constellation has a constellation $B_{N_q}^{N_1,\ldots,N_{q-1}}(2^m, 2^k, \Lambda_{n_1}^p, \ldots, \Lambda_{n_{q-1}}^p)$ along each of its N_q-D subspaces and the lattices $\Lambda_{n_q}^p$, and $D_{n_q}^*$, $n_q = N/N_q$, are used to partition/shape the cartesian product of the $B_{N_q}^{N_1,\ldots,N_{q-1}}$'s. Addressing in $B_N^{N_1,\ldots,N_q}$ requires a set of q memory blocks with $k_i', i = 1,\ldots,q$ input and output lines where $2^{k_{i+1}'} = |\Lambda_{n_i}^p / 2^{k_i'} D_{n_i}^*|$, $k_1' = k$, $n_1 = N_1/2$.

The total rate is equal to $mN/2 - r_s$ and $r_s = 1 + N\sum_{i=1}^{q}(1/N_i)$.

5.7 Comparison with other techniques

In the following, we compare our addressing schemes with the pioneering works of [10], [1] and [14].

In the Voronoi constellations, the Voronoi region of a lattice is used as the shaping region, [5], [10]. The complexity of the addressing is that of a linear mapping plus that of the decoding of the shaping lattice. In [1], the 2-D subspaces are partitioned into the circular shells of equal volume. Then, a multi-level shaping code is used to specify the sequence of the 2-D subregions. In [14], the Voronoi region of an infinite dimensional lattice obtained from a convolutional code is used as the shaping region. The addressing complexity is that of a linear mapping plus the decoding of the code trellis diagram.

The major problem in the Voronoi constellations based on the binary lattices is that they have a cubic 2-D subconstellation (instead of spherical). For a given CER_s, this decreases the achievable γ_s and also increases the PAR. The Voronoi constellation also suffer from the the problem of ties which occurs when some points are located on the boundary of the shaping region. The ties complicate the addressing procedure and potentially may result in a constellation which is not symmetrical.

The shell-addressed Voronoi constellations introduced here have a spherical 2-D subconstellation. Their addressing is achieved by a Voronoi constellation of *half* the original dimensionality. This reduces the addressing complexity. Also, an important class of our schemes achieving a point on the optimum tradeoff curve are based on the lattice D_n^* which has a simple decoding algorithm. In a shell-addressed Voronoi constellation, the ties, although still existing, do not result in addressing problem or unsymmetry.

The schemes of [1] also use a spherical 2-D subconstellation and do not have the problem of ties. But, to have a fair comparison of [1] with this work or with [10] and [14], it remains: (i) to find an appropriate shaping code which has an integral bit rate per signaling interval (to avoid the problem of the nonintegral bit rate) and (ii) to find an addressing scheme to map the data bits to the code words. As mentioned in [1], the addressing problem is not a major issue. However, the problem of the nonintegral bit

rate, needs to be further discussed.

As we are essentially able to achieve any point up to the knee of the optimum tradeoff curves, in table (5.5), we have compared some of the values obtained in [10] and [1] with the optimum values calculated in the fourth chapter.

	VC, [10]			*C/O*, [1]		
N	γ_s	CER_s	PAR	γ_s	CER_s	PAR
4	0.37	1.41 (1.09)	4.62 (2.27)	——	————	————
8	0.65	2.00 (1.26)	6.98 (2.81)	0.60	1.27 (1.19)	2.80 (2.61)
12	0.75	3.00 (1.26)	8.24 (2.86)	0.63	1.55 (1.13)	3.42 (2.51)
16	0.81	1.54 (1.24)	5.58 (2.88)	0.69	1.45 (1.14)	3.02 (2.55)
24	1.03	5.20 (1.50)	15.2 (3.67)	0.80	1.50 (1.16)	3.46 (2.67)
32	0.85	1.35 (1.16)	4.96 (2.70)	0.86	1.46 (1.17)	3.40 (2.72)

Table 5.5: Comparison between the the Voronoi constellations (*VC*) and the Calderbank, Ozarow method (*C/O*) with the optimum constellations, the values in the parenthesis are the optimum values of CER_s, PAR for the given γ_s.

It should be mentioned that by extending the *peak constraint* technique introduced in [14] to the case of the finite dimensional lattices, it is possible to modify the Voronoi constellation in such a way that the 2-D points outside a circle of selected radius are not allowed. This constraint can be applied to the minimum distance decoder, [13], of the lattice. Such a modification, to some extent, remedies the deficiencies caused by a cubic 2-D subconstellation. For example, our simulation results show that for the lattice E_8, one can achieve almost all the shape gain given in table (5.5) but with $\text{CER}_s = 1.7$ and $\text{PAR} = 4$ instead of $\text{CER}_s = 2$ and $\text{PAR} = 6.98$. It should be mentioned that most probably for the higher dimensional lattices (like Λ_{24}), the improvement due to this technique will be more pronounced.

As a more detailed comparison, a four state trellis diagram of [14] (in conjunction with the peak constraint technique) achieves $\gamma_s = 0.97$ dB, $\text{CER}_s = 1.5$, $\text{PAR} = 3.75$. For

$N=32$, a two-level shaping code of [1] achieves, $\gamma_s=0.86$ dB, $\text{CER}_s=1.46$, $\text{PAR}=3.40$. For $N=32$ and $M=128$ points per 2-D subconstellations, our address decomposition method needs $M_s=44$ kilo-bytes per N dimensions to achieve $\gamma_s=0.89$, $\text{CER}_s=1.19$ ($r_s=4$ bits per N dimensions) and $\text{PAR}=2.8$. This tradeoff point can not be distinguished from the L-Point in table 4.1. As an alternative, our method needs $M_s=36$ kilo-bytes per N dimensions to achieve $\gamma_s=1.02$, $\text{CER}_s=1.41$ ($r_s=8$ bits per N dimensions) and $\text{PAR}=3.42$. This is very near to the K-Point in Table 4.1. On the other hand, to realize the L-Points/K-Points, the appropriate number of shells per 2-D subspaces is equal to 4/8. As an example refer to Fig. (5.1). For these numbers of partitions, a direct addressing scheme requires a lookup table with $M_s=1.05\times10^6/6.5\times10^9$ kilo-bytes per N dimensions.

In addition to the better (nearly optimum) performance, the address decomposition method has two other advantages over the examples given in [1] and [14]:

- The *examples* given in [1] and [14] achieve tradeoff points with relatively high CER_s (about 1.5). The achieved points are relatively far from the knee of the optimum tradeoff curves. Also, in a coding scheme carrying a high bit rate per dimension, $\text{CER}_s\simeq1.5$ may be hard to implement. However, our method is not confined to a specific tradeoff point. Specifically, for $\text{CER}_s=1.19$ (L-Points in Table 4.1), we can achieve a higher γ_s than the examples of [1] or almost all the γ_s of the examples of [14] with a substantial decrease in CER_s and PAR. It should be mentioned that it is possible to find other examples for the application of the ideas introduced in [1] and [14] achieving different, possibly better, tradeoff points.

- It seems that our address decomposition method which is just a block of memory (no associated computation) is easier to implement.

However, the schemes of [14] and [10] have an advantage over all other known shaping methods that in their case the constellation points form a group under vector addition modulo the shaping lattice. This property can be used to combine the shaping and the precoding for signaling over the partial response channels, [15], [8].

5.8 Summary and conclusions

We have introduced several practical addressing methods. These are based on partitioning the constellation into clusters of equal volume and selecting a subset of them with the low average energy. In one class, the addressing is achieved by a lookup table. By using the address decomposition method, we have substantially decreased the complexity while the suboptimality is negligible. In another class, the addressing is based on the use of the Voronoi constellations. One can also implement hybrid multi-level schemes which combine both classes.

Chapter 6

Unsymmetrical Boundary Shaping, Spectral Shaping

Part of this chapter have been reported in [32], [33], [34], [35].

6.1 Introduction

In this chapter the concept of the unsymmetrical shaping is discussed. This is the selection of the boundary of a constellation which has nonequal values of power along different dimensions. The objective is to maximize the rate of the constellation subject to some constraints on its power spectrum.

Assuming continuous approximation, the selection of the constellation is composed of selecting a basis for the space (modulating matrix) and a boundary (shaping region) for the points. This is formulated in terms of an optimization problem. The objective function (to be maximized) is the rate of the constellation. There is always a constraint on the total energy. We impose an additional constraint on a factor denoted as the Constellation-Expansion-Ratio, CER_s, and also some constraints on the resulting power spectrum. Due to the continuous approximation, the structure of the shaping region appears as an independent factor in the objective function. This reduces the complexity

of the optimization procedure. Selection of the basis depends on the specific set of the spectral constraints. However, shaping can be studied in a more general context. Because of this reason, we first consider the shaping.

In the continuous approximation, the entropy and the energy per signaling interval are determined by the volume and the second moment of the shaping region. In a conventional shaping problem, one tries to maximize the volume of the shaping region subject to a fixed total second moment. This leads to equal energy being allocated in each dimension. Without additional constraints, spherical regions are optimum.

In some applications, we need a constellation which has nonequal second moments along different dimensions. For example, this nonequal energy allocation in conjunction with a nondiagonal modulating matrix can be used to shape the power spectrum of the transmitted signal. This results in an unsymmetrical shaping problem. In this case, one tries to maximize the volume of the shaping region subject to having the second moment λ_i along the i'th dimensions. Without additional constraint, elliptical regions are optimum. The case of the elliptical shaping region is already discussed in [24] and [19]. In this work, we impose an additional constraint on the factor CER_s of the shaping region. The factor CER_s for the case of the unsymmetrical shaping will be defined later.

The body of the chapter is as follows: In Section 6.2, the block diagram of the system is introduced. In Section 6.3, we discuss the idea of the unsymmetrical shaping in more detail. In Section 6.4, we discuss how to maximize the rate of a constellation (volume of the shaping region) subject to some constraints on the power spectrum. The following constraints are considered in detail: (i) A fraction of the total power equal to F_p is located in the frequency band $[0, \omega_c]$, and/or (ii) The spectrum has spectral nulls at zero and/or at the Nyquist frequency. It is shown that this maximization is equivalent to maximizing the determinant of the autocorrelation matrix subject to some linear constraints on its elements. In an optimized basis analysis, the optimum autocorrelation matrix is found. In a fixed basis analysis, the eigenvectors of the autocorrelation matrix are fixed and the eigenvalues are optimized. The eigenvectors are selected to reduce the computational complexity of the modulation by using fast transform algorithms.

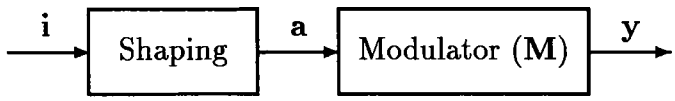


Fig. 6.1: System block diagram.

6.2 System block diagram

Figure 6.1 shows the block diagram of the system under consideration. We use a discrete time model and block based processing. Each signaling interval is composed of M time multiplexed impulses. The available energy per time impulse is normalized to unity. In each signaling interval, a binary data vector $\mathbf{i}$ is encoded. The shaping block maps the vector $\mathbf{i}$ to the point $\mathbf{a}$ in the baseband constellation. This is a finite set of the N-D points bounded within the shaping region $\mathcal{R}_a$. We assume that the points $\mathbf{a}$ are uniformly distributed within $\mathcal{R}_a$ and are used with equal probability. The second moment along the i'th dimension of $\mathcal{R}_a$ is equal to λ_i. The diagonal matrix $\mathbf{\Lambda}_a$ is defined as, $\mathbf{\Lambda}_a = \text{diag}\,[\lambda_0, \ldots, \lambda_{N-1}]$.

Normalizing the volume of the Voronoi region around each constellation point to unity, the entropy of $\mathbf{a}$ is found as,

$$H(\mathbf{a}) = \log\,[V(\mathcal{R}_a)], \tag{6.1}$$

where $V(\mathcal{R}_a)$ is the volume of $\mathcal{R}_a$. This is due to the assumption that the points $\mathbf{a}$ are uniformly distributed within $\mathcal{R}_a$ and are used with equal probability.

The columns of the $M \times N$, $M \geq N$ (modulating) matrix $\mathbf{M}$ are the dimensions of the constellation (line codes). We have $\mathbf{M}^t\mathbf{M} = \mathbf{I}$ where $\mathbf{I}$ is the $N \times N$ identity matrix. This results in $H(\mathbf{a}) = H(\mathbf{y})$. The objective is to maximize the $H(\mathbf{y})$, or equivalently $V(\mathcal{R}_a)$, subject to a constraint on the CER_s of $\mathcal{R}_a$ and also some constraints on the power spectrum of $\mathbf{y}$. This is denoted as the *spectral shaping*. Our tools are the selection of the region $\mathcal{R}_a$ and the matrix $\mathbf{M}$. In the following, we first consider the selection of the shaping region $\mathcal{R}_a$.

6.3 Unsymmetrical shaping

For a given set of second moments λ_i's, the shape gain (γ_s) of the region $\mathcal{R}_a$ is defined as the reduction in the average energy comparing to a using cubic shaping region with the same volume and with the second moments proportional to λ_i's. Using continuous approximation over the hypercube, we obtain,

$$\gamma_s(\mathcal{R}_a) = \frac{1}{12}\left[\frac{V^2(\mathcal{R}_a)}{\prod_{i=0}^{N-1}\lambda_i}\right]^{\frac{1}{N}}. \tag{6.2}$$

Assume that the projection of $\mathcal{R}_a$ on the i'th dimension is the region $[-L_i, L_i]$. Define the binit rate along the i'th dimension of $\mathcal{R}_a$ as $\log(2L_i)$. The shaping redundancy, r_s, is defined as the difference between the average binit rate and the entropy per dimension of $\mathcal{R}_a$. The CER_s is defined as, $\text{CER}_s = e^{r_s}$. This results in,

$$\text{CER}_s(\mathcal{R}_a) = 2\left[\frac{\prod_{i=0}^{N-1} L_i}{V(\mathcal{R}_a)}\right]^{\frac{1}{N}}. \tag{6.3}$$

A higher γ_s is achieved at the price of a higher CER_s and there exists a tradeoff between these two factors. An *optimally shaped* region is the one which optimizes this tradeoff.

From (6.2) and (6.3), it is seen that CER_s and γ_s are invariant to the scaling of the coordinates. This means that the regions obtained by scaling have the same shaping performance. This allows us to relax the constraints on the individual λ_i's to a single constraint on $\sum_i \lambda_i$. This reduces the problem of the unsymmetrical shaping to a conventional shaping problem.

Define a symmetrical region as a region which is closed under the permutations and sign changes of the coordinates. The unsymmetrical versions of a symmetrical region are obtained by scaling that region along different dimensions. For a given total second moment, a symmetrical region has a larger volume than its unsymmetrical versions. The reduction in the volume is the price associated with the unsymmetrical shaping.

We assume that region $\mathcal{R}_a$ is obtained by scaling a symmetrical baseline region $\mathcal{B}$ by the scale factor S_i along the i'th dimension. Projection of the region $\mathcal{B}$ on the space

dimensions is the region $[-1,1]$. The scale factors are equal to $S_i = \sqrt{\lambda_i / E}$ where E denotes the average energy per dimension of $\mathcal{B}$. In this case, as we will see later, the power spectrum of $\mathbf{y}$ depends only on the matrices $\mathbf{\Lambda}_a$ and $\mathbf{M}$. This allows us to select the baseline region and the scale factors independently. The baseline region is selected to maximize the $\gamma_s(\mathcal{B})$ for a given $\mathrm{CER}_s(\mathcal{B})$. Substituting $\lambda_i = E$ and $L_i = 1$ in (6.2) and (6.3), it is easy to verify that this objective is equivalent to maximizing the volume for a given second moment. The scale factors (or equivalently $\mathbf{\Lambda}_a$) and $\mathbf{M}$ are selected to shape the power spectrum.

6.3.1 Optimum baseline region

An optimally shaped $\mathcal{B}$ should be selected as a subset of the hypercube $[-1,1]^N$, N–fold product of $[-1,1]$, which has the maximum volume for a given second moment. This subset is selected by a hypersphere. This is based on the same general idea as in the fourth chapter. By changing the radius of the hypersphere, β, we can tradeoff γ_s and CER_s. This region is denoted by $\mathcal{A}_N^{(1)}$. If we label the dimensions by X_i, $i \in [0, N-1]$, we have,

$$\mathcal{A}_N^{(1)} = \begin{cases} -1 \le X_i \le 1, & i \in [0, N-1], \\ \sum_{i=0}^{N-1} X_i^2 \le \beta, & 1 \le \beta \le N. \end{cases} \tag{6.4}$$

Figure 6.2 shows the region $\mathcal{A}_3^{(1)}$ for three different values of β.

For $\beta \le 1$, region $\mathcal{A}_N^{(1)}$ is a sphere. It has the maximum γ_s (for a given N) but also large value for CER_s. For $\beta \ge N$, region $\mathcal{A}_N^{(1)}$ is a hypercube. This corresponds to no shaping, i.e., $\gamma_s = 1$ and $\mathrm{CER}_s = 1$. By changing β in the range $1 \le \beta \le N$, we move along the optimum tradeoff curve between these two extreme points.

6.3.2 Addressing

The main difference between the addressing of an unsymmetrical shaping region and the addressing in a conventional shaping problem is that in the unsymmetrical case *there are*

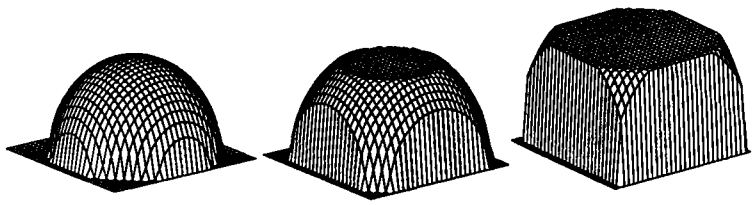


Fig. 6.2: The $\mathcal{A}_3^{(1)}$ region for three different values of β.

different number of points along different dimensions.

We assume that the addressing is achieved in two steps. The first step is on the one-D subconstellations and the second step is on their cartesian product. We also assume that the points of the one-D subconstellations belong to the half integer grid $Z + (1/2)$. This is the case for the coding schemes of [42]. The points of the $Z + (1/2)$'s are divided into K energy shells of equal cardinality. There are P_i points in the shells of the i'th dimension. Each shell contains an equal number of points from each partition in an Ungerboeck partition chain. This is an important issue in using the constellation in a multi-dimensional trellis coding scheme, [42].

The one-D energy shells divide the available signal space into K^N, N-D shaping clusters of equal cardinality. The second level of the addressing is achieved by selecting T of the N-D clusters with the least average energy. The CER_s of this scheme is equal to,

$$\mathrm{CER}_s(\mathcal{A}_N^{(1)}) = \frac{K}{T^{\frac{1}{N}}}. \tag{6.5}$$

To transmit an integral number of bits per signaling interval, the T and P_i's should be integral powers of two, i.e., $T = 2^t$ and $P_i = 2^{p_i}$. In this case, each N-D shaping cluster corresponds to the *constant integral* rate $\sum_i p_i$. The total rate is equal to, $t + \sum_i p_i$.

Addressing is achieved by a lookup table with t input lines and $N\lceil \log K \rceil$ output lines. In each signaling interval, t bits enter the lookup table and the $N\lceil \log K \rceil$ bits at the output selects an energy shell within each one-D subconstellation. Another $\sum_i p_i$ data bits select the final point within each of the one-D shells. The complexity of the addressing lookup table can be substantially decreased by using the address decomposition method of the fifth chapter.

In the shaping regions of the fourth chapter, the boundary of the 2-D subconstellations is a circle. The 2-D subconstellations are partitioned into the energy shells of equal volume. The main property is that in this case, the average energy of the 2-D shells depends linearly on their index. This is the key point to a set of the addressing schemes in the fifth chapter. In the case of the $\mathcal{A}_N^{(1)}$ region, by partitioning the 2-D subconstellations (which are 2-D cubes) into the energy shells of equal volume, the same addressing schemes are applicable. However, in this case, the linear relationship between the average energy and the index is slightly violated. This results in a small degradation in the shaping performance.

6.4 Spectral shaping

6.4.1 Preliminaries

The autocorrelation matrix of a sequence of M-D, blockwise uncorrelated, real vector $\mathbf{y}$ is equal to, $\mathbf{R}_y = \mathbf{E}[\mathbf{y}\mathbf{y}^t]$, where $\mathbf{E}[\ .\ ]$ denotes the expectation. Obviously, $\mathbf{R}_y$ is a symmetrical matrix. Define $d_y(k)$ as,

$$d_y(k) = \sum_{|i-j|=k} R_y(i,j), \tag{6.6}$$

where $R_y(i,j)$'s are the elements of $\mathbf{R}_y$. Using the results of [3], the power spectrum of $\mathbf{y}$ is equal to,

$$S_y(\omega) = \frac{1}{M} \sum_{k=0}^{M-1} d_y(k) \cos(\omega k). \tag{6.7}$$

For the model of Fig. 6.1, we have $\mathbf{R}_y = \mathbf{M}\mathbf{R}_a\mathbf{M}^t$ where $\mathbf{R}_a$ is the autocorrelation matrix of $\mathbf{a}$. It can be shown that if the shaping region $\mathcal{R}_a$ is obtained by the scaling of a symmetrical baseline region $\mathcal{B}$, the autocorrelation matrix $\mathbf{R}_a$ will be diagonal, i.e., $\mathbf{R}_a = \mathbf{\Lambda}_a$. In this case, the spectrum of $\mathbf{y}$ is equal to,

$$S_y(\omega) = \sum_{i=0}^{N-1} \lambda_i S_i(\omega), \tag{6.8}$$

where $S_i(\omega)$ is the power spectrum of the i'th dimension of $\mathbf{M}$. This means that $S_y(\omega)$ *depends only* on matrices $\mathbf{M}$ and $\mathbf{\Lambda}_a$. More specifically, it *does not depend* on the structure of the baseline region $\mathcal{B}$.

Assuming $\mathbf{R}_a = \mathbf{\Lambda}_a$, to realize a given $\mathbf{R}_y$, it is enough to select the matrices $\mathbf{M}$ and $\mathbf{\Lambda}_a$ as the matrices of the eigenvectors and the eigenvalues of $\mathbf{R}_y$, respectively. All our following discussions are based on this structure.

Using (6.2), and considering that $H(\mathbf{y}) = H(\mathbf{a}) = \log[V(\mathcal{R}_a)]$, we obtain,

$$H(\mathbf{y}) = \frac{N}{2}\log[12\gamma_s(\mathcal{B})] + \frac{1}{2}\sum_{\lambda_i(\mathbf{R}_y)\neq 0}\log[\lambda_i(\mathbf{R}_y)]. \tag{6.9}$$

If all the eigenvalues are nonzero, we have $\sum_i \log[\lambda_i(\mathbf{R}_y)] = \log(|\mathbf{R}_y|)$ where $|\ .\ |$ denotes the determinant. Considering (6.9), it is seen that the baseline region $\mathcal{B}$ and the matrix $\mathbf{\Lambda}_a$ have independent effects on $H(\mathbf{y})$. *This property allows us to select them independently.* The baseline region $\mathcal{B}$ is selected by the boundary shaping considerations. The objective is to maximize the $\gamma_s(\mathcal{B})$ for a given $\mathrm{CER}_s(\mathcal{B})$. The matrices $\mathbf{M}$ and $\mathbf{\Lambda}_a$ are selected by the spectral shaping considerations. The objective is to maximize the second term in (6.9) subject to some constraints on the power spectrum of $\mathbf{y}$.

6.4.2 Linear filtering

Consider a linear system with the $M \times N$, $N \leq M$, transfer matrix $\mathbf{A}$. The input eigenvectors of $\mathbf{A}$, $\mathbf{m}_i$, $i \in [0, N-1]$, are the eigenvectors of $\mathbf{A}^t\mathbf{A}$ with the eigenvalues ϕ_i. Similarly, the output eigenvectors $\hat{\mathbf{m}}_i$, $i \in [0, M-1]$ are the eigenvectors of $\mathbf{A}\mathbf{A}^t$. Assuming $M > N$, $\mathbf{A}\mathbf{A}^t$ has N nonzero eigenvalues equal to the same ϕ_i's and $M_0 = M - N$ eigenvalues equal to zero. The eigensystem of $\mathbf{A}$ satisfies,

$$\begin{aligned} \mathbf{A}\mathbf{m}_i &= \sqrt{\phi_i}\,\hat{\mathbf{m}}_i, \\ \mathbf{A}^t\hat{\mathbf{m}}_i &= \sqrt{\phi_i}\,\mathbf{m}_i. \end{aligned} \tag{6.10}$$

Since $\mathbf{A}^t\mathbf{A}$ and $\mathbf{A}\mathbf{A}^t$ are both symmetrical, the input and the output eigenvectors form an orthonormal basis. Obviously, the input eigenvalues are nonnegative. Furthermore,

all our subsequent discussions are based on systems which have strictly positive input eigenvalues.

For the system $\mathbf{A}$, the autocorrelation matrix of the N-D input vector $\mathbf{x}$ and the M-D output vector $\mathbf{y}$ are related by, $\mathbf{R}_y = \mathbf{A}\mathbf{R}_x\mathbf{A}^t$. For a positive-definite $\mathbf{R}_x$ ($|\mathbf{R}_x| \neq 0$), $M - N$ eigenvalues of $\mathbf{R}_y$ are zero and the product of the N nonzero eigenvalues is equal to,

$$\prod_{\lambda_i(\mathbf{R}_y)\neq 0} \lambda_i(\mathbf{R}_y) = |\mathbf{A}^t\mathbf{A}| \times |\mathbf{R}_x|. \tag{6.11}$$

By the appropriate selection of $\mathbf{A}$, we can impose some constraints on the spectrum of $\mathbf{y}$.

6.4.3 Performance loss of a nonflat spectrum

The price to be paid for a nonflat spectrum (unsymmetrical shaping) is a reduction in the signal space volume. This is measured in terms of the power loss with respect to a reference scheme with the same dimensionality and with a white spectrum. In this case, equating the entropies, the power loss, P_l, is defined as the ratio of the total energies. As already mentioned, the energy per time interval is normalized to unity, i.e., $\sum_i \lambda_i = M$. For a cubic baseline region, assuming a cubic shaping region for the reference scheme, we obtain,

$$P_l = (12)^{1-\frac{N}{M}} \left(\prod_{i=0}^{N-1} \lambda_i \right)^{-\frac{1}{M}}. \tag{6.12}$$

For $N = M$, or $N \simeq M$ when N and M are large, (6.12) reduces to,

$$P_l = \left(\prod_{i=0}^{N-1} \lambda_i \right)^{-\frac{1}{N}}. \tag{6.13}$$

6.4.4 Asymptotic behavior assuming a spherical baseline region

Assume that the shaping region is elliptical (spherical $\mathcal{B}$). The volume is equal to,

$$V(\mathcal{R}_a) = \frac{[\pi(N+2)]^{\frac{N}{2}}}{\Gamma(0.5N+1)} \times \prod_{i=0}^{N-1} \sqrt{\lambda_i}, \tag{6.14}$$

where $\Gamma(\ .\ )$ is the gamma function. This is the maximum volume obtainable for a given set of λ_i's.

For all the shaping regions, as N tends to infinity, the distributions along different dimensions become independent of each other. For an elliptical region, the distributions tend to independent Gaussian ones. In this case, using (6.14), it can be shown the entropy along the i'th dimension is equal to $0.5 \log (2\pi e \lambda_i)$. This is an upper bound to the entropy of a process with the energy λ_i. Also, for $N \rightarrow \infty$, the eigenvectors tend to complex exponentials, $\exp(-j\omega)$, and the eigenvalues tend to the power spectrum, $S_y(\omega)$. In this case, using (6.14), it is easy to show that the average entropy per time dimension of $\mathbf{y}$ tends to,

$$H_0(\mathbf{y}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log\ [2\pi e S_y(\omega)]\ d\omega = \frac{1}{2} \log (2\pi e \lambda), \tag{6.15}$$

where,

$$\lambda = \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\ [S_y(\omega)]\ d\omega \right\}. \tag{6.16}$$

To have the same entropy with an infinite dimensional spherical region, the required energy per dimension is equal to λ given in (6.16). This is closely related to the innovation power of $\mathbf{y}$.

It can be shown that the asymptotic value of P_l for an elliptical shaping region with respect to a spherical reference scheme is the same as given in (6.13). This reduces to,

$$P_l = \exp \left\{ -\frac{1}{2\pi} \int_{-\pi}^{\pi} \log\ [S_y(\omega)]\ d\omega \right\}. \tag{6.17}$$

Equation (6.17) can be also deduced from (6.16).

6.4.5 Spectral shaping using an optimized basis

We are going to select $\mathbf{M}$ and $\mathbf{\Lambda}_a$ such that the power spectrum of $\mathbf{y}$ satisfies certain constraints and the entropy of the $\mathbf{y}$ is maximized. Considering (6.9), for a given $\gamma_s(\mathcal{B})$, $H(\mathbf{y})$ is maximized by maximizing the $\prod_i \lambda_i(\mathbf{R}_y)$, $\lambda_i(\mathbf{R}_y) \neq 0$.

Due to the linear relationship between the spectrum and the d_y's in (6.7), most of the spectral constraints can be formulated as linear constraints on the d_y's. The total

number of such constraints is denoted by L. There is always a constraint on the total energy. We study two other constraints, namely the F_p-constraint and/or the spectral null.

Define the power-ratio of a spectrum as the fraction of the total power in the frequency band $[0, \omega_c]$. The F_p-constraint is the constraint of having a power-ratio less than or equal to F_p. Integrating (6.7), the F_p-constraint is expressed as,

$$\sum_{i=0}^{M-1} d_y(i) \sin(\omega_c i)/i \leq \pi M F_p. \tag{6.18}$$

A spectral null at the zero frequency or at the Nyquist frequency results in at least one zero eigenvalue for $\mathbf{R}_y$. In this case, we consider $\mathbf{y}$ as the output of a system $\mathbf{A}$ with the same spectral null and reformulate the problem at the system input, $\mathbf{x}$. As $\mathbf{x}$ has no spectral null, $\mathbf{R}_x$ is positive-definite. Considering (6.11), for a given $\mathbf{A}$, to maximize $H(\mathbf{y})$, one should maximize $|\mathbf{R}_x|$. Using $\mathbf{R}_y = \mathbf{A}\mathbf{R}_x\mathbf{A}^t$, the linear constraints on the elements of $\mathbf{R}_y$ are transferred to the linear constraints on the elements of $\mathbf{R}_x$. The energy constraint reduces to,

$$\sum_{i=0}^{M-1} R_y(i,i) = \sum_{p=0}^{N-1}\sum_{q=0}^{N-1} U(p,q) R_x(p,q) = M, \tag{6.19}$$

where,

$$U(p,q) = \sum_{i=0}^{M-1} A(i,p) A(i,q). \tag{6.20}$$

These are the elements of the matrix $\mathbf{U} = \mathbf{A}^t\mathbf{A}$. Similarly, the F_p-constraint reduces to,

$$\sum_{p=0}^{N-1}\sum_{q=0}^{N-1} V(p,q) R_x(p,q) \leq \pi M F_p, \tag{6.21}$$

where,

$$V(p,q) = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} A(i,p) A(j,q) \sin[\omega_c(i-j)]/(i-j). \tag{6.22}$$

The final optimization problem is as follows:

$$\begin{cases} \text{Maximize} & \log(|\mathbf{R}_x|), \\ \text{Subject to:} & \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} B_l(i,j) R_x(i,j) \leq e_l, \quad l \in [0, L-1], \\ & \mathbf{R}_x \text{ is positive-definite.} \end{cases} \tag{6.23}$$

It can be shown that the logarithm of the determinant of a positive-definite matrix is a convex $\cap$ function, [21]. In Appendix E, it is shown that the set of the constraints in (6.23) determines a convex region. This results in a convex optimization problem. As a result, the maximum point is unique and can be found by using the Lagrange method.

Define an active constraint as a constraint for which the equality holds. The set of the active constraints are denoted by A_c. The Lagrange multipliers are denoted by ψ_l, $l \in A_c$. Calculating the derivatives and considering that the derivative of the determinant with respect to the (i,j)'th element is equal to the determinant of the corresponding adjoint matrix, we obtain,

$$\text{adj}\,[\mathbf{R}_x] = \sum_{l \in A_c} \psi_l \mathbf{B}_l, \tag{6.24}$$

where $\text{adj}\,[\mathbf{R}_x]$ is the adjoint matrix of $\mathbf{R}_x$ and $\mathbf{B}_l$ is the matrix of the elements $B_l(i,j)$ in (6.23). For the spectral null constraint, we have $\mathbf{B}_1 = \mathbf{U} = \mathbf{A}^t\mathbf{A}$. For the F_p constraint, we have $\mathbf{B}_2 = \mathbf{V}$ where the elements of $\mathbf{V}$ are given in (6.22).

To calculate the Lagrange multipliers, we first calculate $\mathbf{R}_x$ using,

$$\mathbf{R}_x = |\text{adj}\,\mathbf{R}_x|^{\frac{1}{N-1}} \times (\text{adj}\,[\mathbf{R}_x])^{-1}, \tag{6.25}$$

and then apply the active constraints to the result. By iteratively satisfying the constraints, the multipliers are found.

It is easy to show that for the spectral nulls and/or the F_p-constraint, the energy constraint is always active. For $F_p \in [F_{\min}, F_{\max}]$, the F_p-constraint is active. For $F_p < F_{\min}$, the optimization problem has no answer. For $F_p > F_{\max}$, the F_p-constraint is not active and the power-ratio is equal to $F_{\max}$. The $F_{\max}$ can be calculated by relaxing the F_p-constraint and finding the power-ratio of the answer. Without spectral null constraint, this results in a white spectrum and $F_{\max} = \omega_c/\pi$. Later, we show that for a spectral null at zero frequency, the optimum code is obtained by allocating equal energy to the output eigenvectors of the $1-D$ system. This will be used to calculate the $F_{\max}$ with a spectral null.

If the spectral null constraint is relaxed, we have $\mathbf{A} = \mathbf{I}$ $(M = N)$ resulting in,

$$\begin{cases} \mathbf{B}_1 = \mathbf{U} = \mathbf{I}, \\ \mathbf{B}_2 = \mathbf{V} = \left[V(p,q) = \sin \dfrac{\omega_c(p-q)}{(p-q)} \right]. \end{cases} \tag{6.26}$$

6.4.6 Spectral shaping using fixed basis

This concerns selecting a fixed $\mathbf{M}$ and using only $\mathbf{\Lambda}_a$ to maximize the entropy. This method has a lower degree of freedom and is suboptimum. However, by the appropriate selection of $\mathbf{M}$, one can decrease the computational complexity. For a spectrum with spectral nulls, $\mathbf{M}$ is selected as an orthonormal basis with the same set of nulls. For the case of no spectral null, sine basis is used. First, we discuss the basis with spectral nulls.

If the system $\mathbf{A}$ has spectral null at certain frequencies, its output eigenvectors form an orthonormal basis with the same set of nulls. For a null at zero/Nyquist frequency, $\mathbf{A}$ is taken as $1 - D/1 + D$ system. The $1 \pm D$ systems have an $(N+1) \times N$ transfer matrix with the i'th column equal to, $[(0)^i, \sqrt{2}/2, \pm\sqrt{2}/2, (0)^{N-1-i}]$. For a null at both zero and Nyquist frequency, $\mathbf{A}$ is taken as $1 - D^2$ system. This has an $(N+2) \times N$ transfer matrix with the i'th column equal to, $[(0)^i, \sqrt{2}/2, 0, -\sqrt{2}/2, (0)^{N-1-i}]$. These are three important examples of the partial response channels, [23]. The eigenvectors/eigenvalues of these systems are calculated in Appendix F. The eigenvectors are closely related to the sine basis. This reduces the computational complexity of the modulation by using a fast sine transform algorithm.

Using (6.8), the F_p-constraint is formulated as,

$$\sum_{i=0}^{N-1} \lambda_i B_i(\omega_c) \leq F_p, \tag{6.27}$$

where,

$$B_i(\omega_c) = 2 \int_0^{\omega_c} S_i(\omega) d\omega. \tag{6.28}$$

Similar to the case of the optimized basis, the energy constraint is always active. This

results in the following *convex* optimization problem,

$$\begin{cases} \text{Maximize} & \sum_{i=0}^{N-1} \log(\lambda_i), \\ \text{Subject to:} & \sum_{i=0}^{N-1} \lambda_i B_i(\omega_c) \leq F_p, \\ & \sum_{i=0}^{N-1} \lambda_i = M, \quad \lambda_i \geq 0. \end{cases} \tag{6.29}$$

Assuming that the F_p-constraint is active and using the Lagrange method, we obtain,

$$\lambda_i = \frac{1}{\psi_1 B_i(\omega_c) + \psi_2}, \tag{6.30}$$

where ψ_1 and ψ_2 are determined by solving,

$$\sum_{i=0}^{N-1} \frac{B_i(\omega_c)}{\psi_1 B_i(\omega_c) + \psi_2} = F_p, \qquad \text{and} \qquad \sum_{i=0}^{N-1} \frac{1}{\psi_1 B_i(\omega_c) + \psi_2} = M. \tag{6.31}$$

In the case that the F_p-constraint is not active, the answer is obtained by allocating equal energy to all the dimensions.

For a spectral null at zero frequency, $\mathbf{A}$ is selected as the $1-D$ system. In this case, by changing the λ_i's while keeping $\sum_i \lambda_i = N+1$, one can tradeoff the width of the null and the entropy of the code. One interesting case in this tradeoff corresponds to,

$$\lambda_i = \frac{N+1}{N} \times \phi_i, \tag{6.32}$$

where ϕ_i's are the eigenvalues of the $1-D$ system. This results in,

$$\mathbf{R}_y = \frac{N+1}{N} \times \mathbf{A}\mathbf{A}^t, \tag{6.33}$$

and,

$$S_y(\omega) = 1 - \cos(\omega). \tag{6.34}$$

In this case, using the results of Appendix F, we obtain,

$$\prod_{i=0}^{N-1} \lambda_i = \left(\frac{N+1}{N}\right)^N \times |\mathbf{A}^t\mathbf{A}| = \frac{(N+1)^{N+1}}{N^N} \times 2^{-N}. \tag{6.35}$$

Using Eqs. (6.17) and (6.34), or equivalently, Eqs. (6.13) and (6.35), the asymptotic value of P_l is found to be equal to 3 dB.

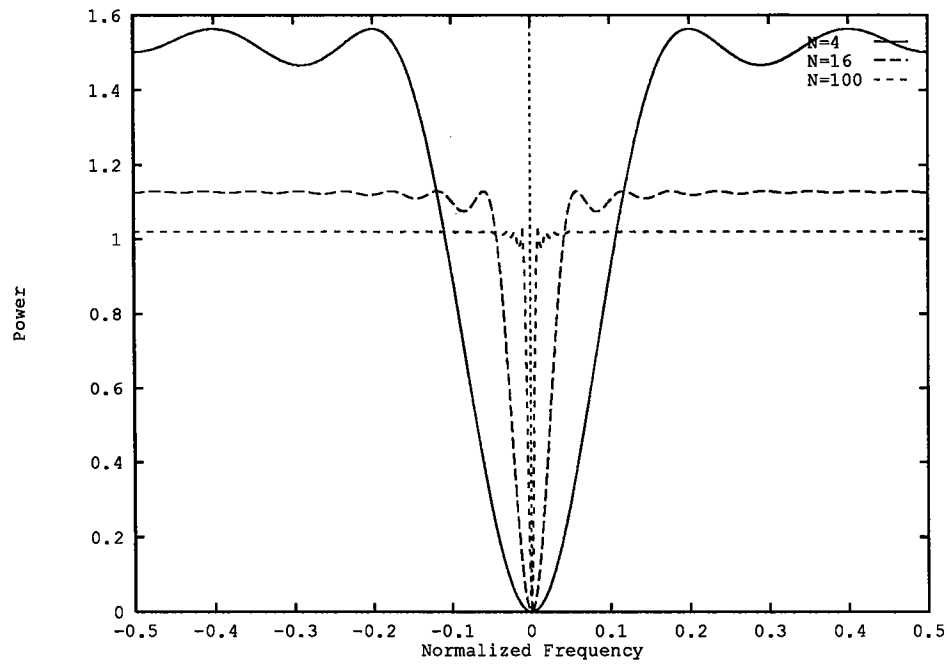


Fig. 6.3: Spectrum of the highest entropy (narrowest null width) with spectral null.

Another interesting case corresponds to the spectrum with the maximum entropy. This is obtained by using a symmetrical shaping region, i.e.,

$$\lambda_i = \frac{N+1}{N}, \quad i \in [0, N-1]. \tag{6.36}$$

Using Eq. (6.13), the asymptotic value for P_l is zero dB. The power spectrum is equal to,

$$S_y(\omega) = \frac{N+1}{N} \sum_{i=0}^{N-1} S_i(\omega), \tag{6.37}$$

where $S_i(\omega)$'s are the spectrum of the output eigenvectors of $1-D$ system. Figure 6.3 shows the resulting spectrum for different values of N.

This is the optimum spectrum with a spectral null at the zero frequency over dimensionality $M = N+1$. This means that the optimized basis analysis results in the same answer. To verify this claim, using (6.24), we obtain,

$$\text{adj}\,[\mathbf{R}_x] = \psi_1 \mathbf{U}, \tag{6.38}$$

where $\mathbf{U} = \mathbf{A}^t\mathbf{A}$ is a matrix with all the diagonal element equal to one, all the elements on the two subdiagonal adjacent to the main diagonal equal to $-1/2$ and all the other elements equal to zero. In this case, it can be shown that (6.38) results in the (symmetrical)

matrix,

$$\mathbf{R}_x = \left[R_x(i,j) = \frac{2}{N}(N-i)(j+1), \quad \text{for} \quad i \geq j \right]. \tag{6.39}$$

It can be also shown that,

$$\lambda_i(\mathbf{R}_x) = \frac{N+1}{2N} \times \sin^{-2} \frac{\pi\,(i+1)}{2\,(N+1)}, \tag{6.40}$$

and,

$$|\mathbf{R}_x| = \left(\frac{2}{N}\right)^N \times (N+1)^{N-1}. \tag{6.41}$$

This results in a matrix $\mathbf{R}_y = \mathbf{A}\mathbf{R}_x\mathbf{A}^t$, with all the diagonal elements equal to one and all the nondiagonal elements equal to $-1/N$. This matrix has one zero eigenvalue and N eigenvalues equal to $(N+1)/N$. Using the resulting $\mathbf{R}_y$, it can be shown that the $F_{\max}$ with the spectral null is equal to,

$$F_{\max} = \frac{\omega_c}{\pi} - \frac{2}{\pi} \sum_{i=1}^{N} \frac{N+1-i}{N\,(N+1)} \sin\,(\omega_c i)/i. \tag{6.42}$$

6.4.7 Numerical results

In this subsection, by a spectral null, we mean a first order null ($M = N+1$) at zero frequency.

Table 6.1 shows the performance loss of a nonflat spectrum, with or without spectral null, using optimized or fixed basis schemes, different normalized cutoff frequencies ($f_c = \omega_c/\pi$), cubic shaping region, $M = 5$ and $F_p = 0.1$.

Figure 6.4 shows the P_l as a function of the f_c, with and without spectral null, using optimized basis, cubic shaping region, $F_p = 0.1$ and $M = 4$.

Figure 6.5 shows the P_l as a function of the f_c, with and without spectral null, using fixed basis, cubic shaping region, $F_p = 0.1$, $M = 4, 8, 16$.

From the given results, it is seen that for high value of f_c, having a spectral null at zero frequency results in a better performance. In general, the curve corresponding to a spectral null of a given order, in a similar way as in Fig. 6.4 or Fig. 6.5, crosses the curves corresponding to a spectral null of a lower order.

f_c	$P_l(ON)$	$P_l(O)$	$P_l(FN)$	$P_l(F)$
0.15	1.7	1.1	1.7	1.2
0.20	2.4	2.1	2.5	2.3
0.25	3.4	3.4	3.6	3.5
0.30	4.7	4.9	5.0	5.0

Table 6.1: Performance loss (in dB) for a cubic shaping region, $M=5$, $F_p=0.1$, (O) means optimized basis, (F) means fixed basis, (N) means with spectral null.

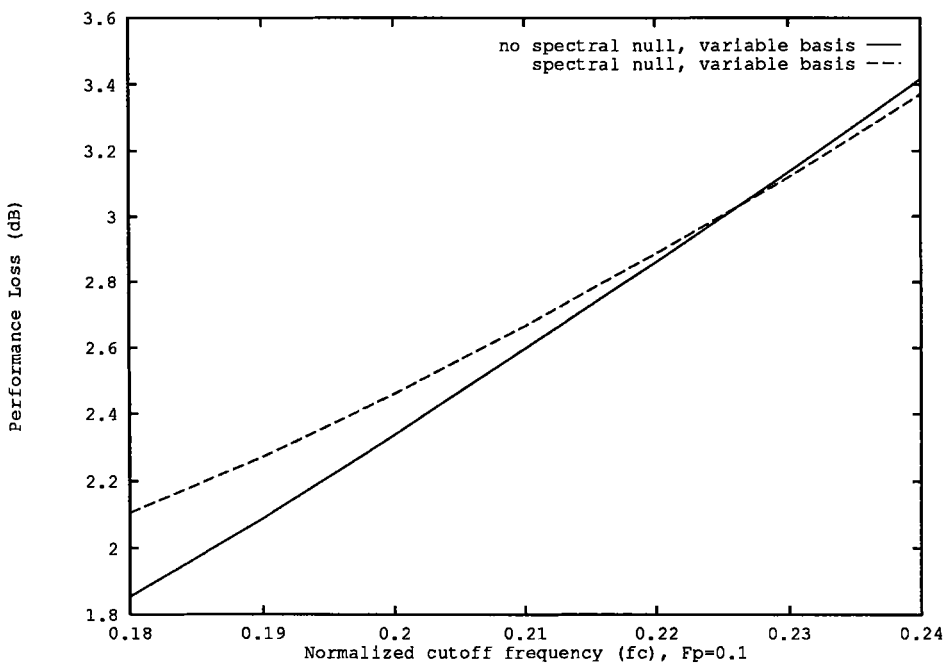


Fig. 6.4: Performance loss (in dB) as a function of the f_c, with and without spectral null, using optimized basis, cubic shaping region, $F_p=0.1$, $M=4$.

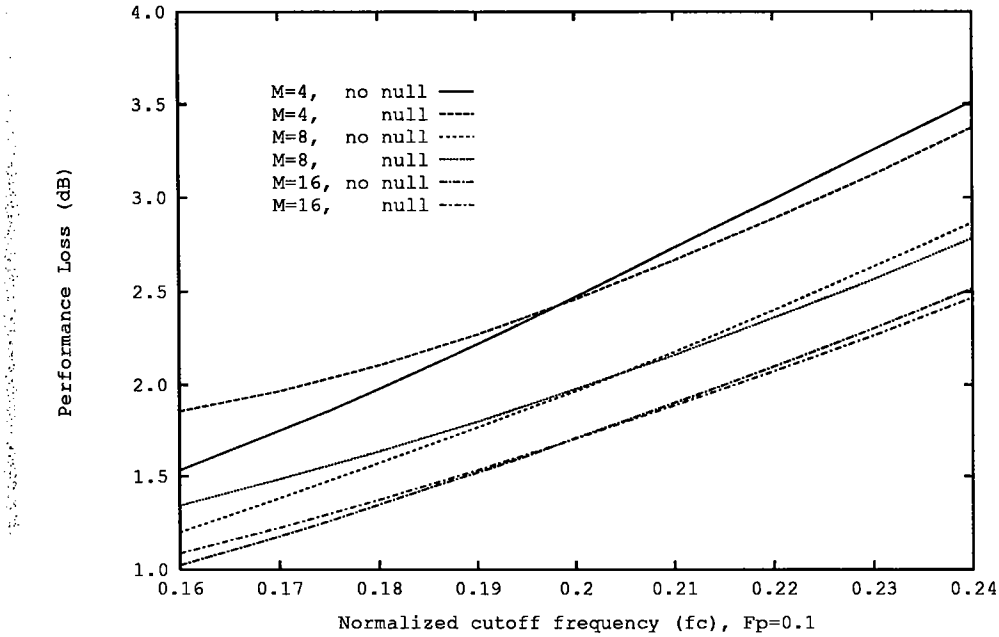


Fig. 6.5: Performance loss (in dB) as a function of the f_c, with and without spectral null, using fixed basis, cubic shaping region, $F_p = 0.1$, $M = 4, 8, 16$.

Figure 6.6 shows the P_l as a function of the f_c, without spectral null, using fixed (sine) and also optimized basis, cubic shaping region, $F_p = 0.1$, $M = N = 4, 8, 16$.

The other conclusion is that increasing the space dimensionality can be very useful, specifically, for higher values of f_c (wider null width). For example, referring to Fig. 6.5 and Fig. 6.6, for moderate values of f_c, increasing the dimensionality from 4 to 16 results in about 1 dB saving in the energy.

6.4.8 Example

We assume fixed basis with a spectral null at zero frequency. For $N = 2$, this basis (output eigenvectors of $1 - D$ system) is equal to,

$$\mathbf{M} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{6}/6 \\ 0 & -\sqrt{6}/3 \\ -\sqrt{2}/2 & \sqrt{6}/6 \end{bmatrix}. \tag{6.43}$$

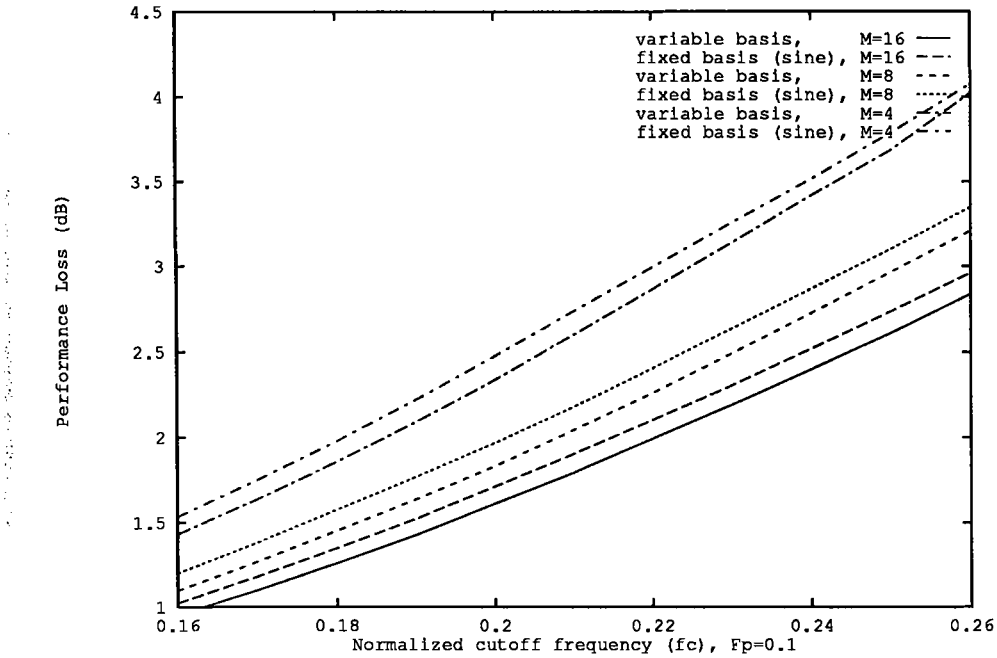


Fig. 6.6: Performance loss (in dB) as a function of the f_c, without spectral null, using fixed and optimized basis, cubic shaping region, $F_p = 0.1$, $M = N = 4, 8, 16$.

For this matrix, the shaping region of $\mathbf{y}$, denoted as $\mathcal{R}_y$, is located at the subspace $Y_0 + Y_1 + Y_2 = 0$. We assume that the baseline region is of $\mathcal{A}_2^{(1)}$ type. After some manipulation, projection of $\mathcal{R}_y$ into the (Y_0, Y_1) subspace is found as,

$$\begin{cases} (Y_0 + 0.5Y_1)^2/S_0^2 + 0.75(Y_1/S_1)^2 \leq 0.5\beta, \\ -\sqrt{2} \leq (Y_0 + 0.5Y_1)/S_0 \leq \sqrt{2}, \\ -\sqrt{6}/3 \leq Y_1/S_1 \leq \sqrt{6}/3. \end{cases} \tag{6.44}$$

where S_0, S_1 and S_2 are the scale factors. Similarly, projection on (Y_0, Y_2) subspace is equal to,

$$\begin{cases} (Y_0 - Y_2)^2/S_0^2 + 3(Y_0 + Y_2)^2/S_2^2 \leq 2\beta, \\ -\sqrt{2}/2 \leq (Y_0 - Y_2)/S_0 \leq \sqrt{2}/2, \\ -\sqrt{6}/3 \leq (Y_0 + Y_2)/S_2 \leq \sqrt{6}/3. \end{cases} \tag{6.45}$$

Projection on (Y_1, Y_2) subspace can be obtained by replacing Y_0 by Y_2 in (6.44). By changing β in the rage $1 \leq \beta \leq 2$, one obtains different regions corresponding to different points on the optimum tradeoff curve. Figures 6.7 and 6.8 show the regions of (6.44) and (6.45) for $\beta = 1.1$.

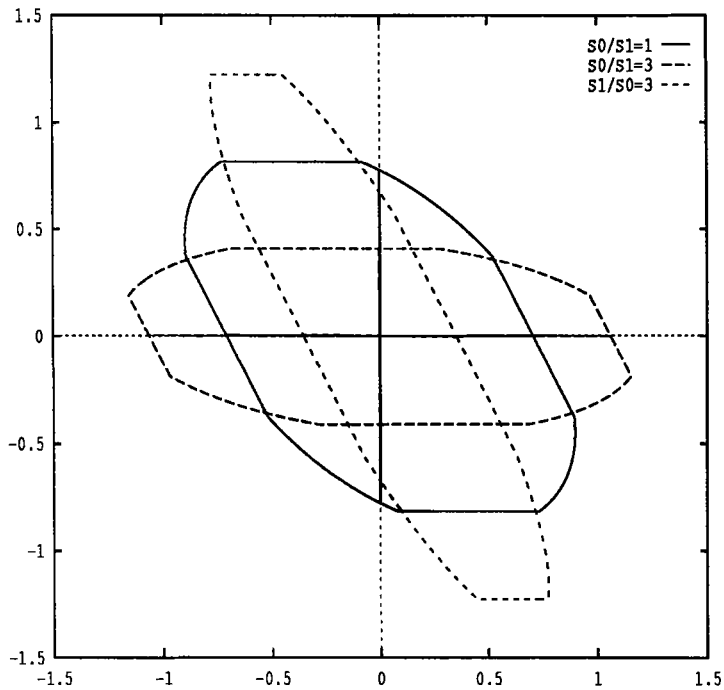


Fig. 6.7: Projection of $\mathcal{R}_y$ of the example on (Y_0, Y_1) subspace.

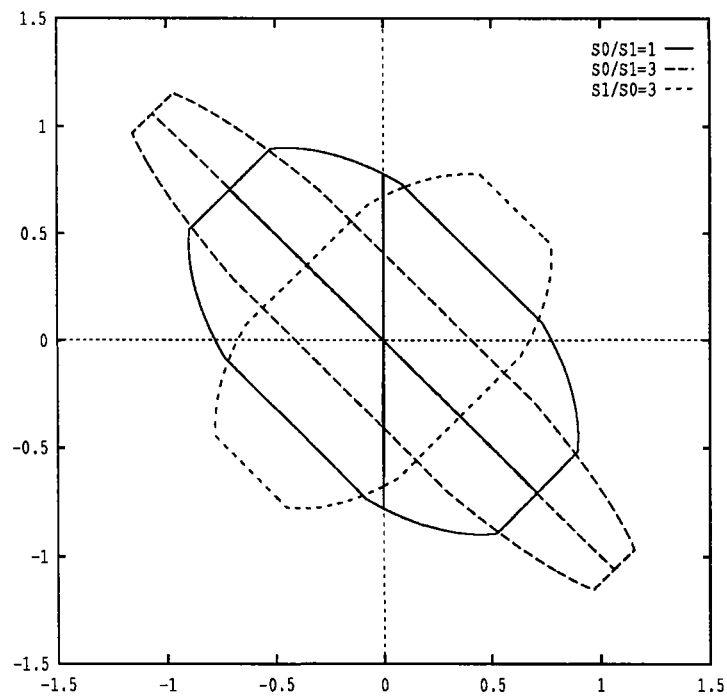


Fig. 6.8: Projection of $\mathcal{R}_y$ of the example on (Y_0, Y_2) subspace.

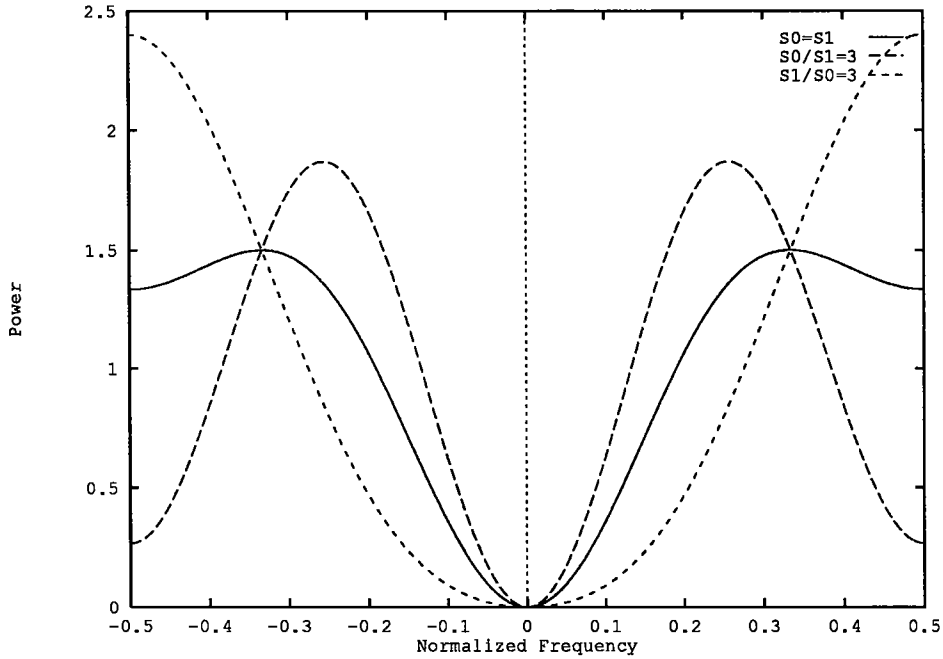


Fig. 6.9: Power spectrum of the example.

Within a scale factor, the shape of the power spectrum is determined by the ratio of S_0 and S_1. By changing this ratio, we can tradeoff the width of the null and the entropy of the code. Figure 6.9 shows the corresponding spectrums.

6.5 Summary and conclusions

We have studied the selection of a constellation for spectral shaping. This is composed of selecting a basis for the space and a boundary (shaping region) for the constellation. The constellation has the power λ_i along the i'th dimension. Shaping region is selected as a region with the second moments λ_i's and with a volume as large as possible. This is denoted as an unsymmetrical shaping problem. The unsymmetrical region is obtained by scaling of a symmetrical baseline region. The selection of the constellation is decomposed into two independent parts, namely, (i) selection of a baseline region, (ii) selection of a basis for the space together with a set of the scale factors for the dimensions. Part (i) is expressed in terms of a conventional shaping problem. The structure of the optimum baseline region with the corresponding addressing scheme is discussed. Part (ii) is com-

puted by an an optimization procedure. This optimization procedure maximizes the rate of the constellation subject to some constraints on the resulting power spectrum. In the optimized basis analysis, we considered the selection of the basis and also the corresponding scale factors. In the fixed basis analysis, the basis is fixed and is selected to reduce the computational complexity of the modulation.

Chapter 7

Block-based Signaling over Partial-Response Channels

Part this chapter have been reported in [36], [37], [38].

7.1 Introduction

In this chapter, we discuss the selection of a signal constellation for signaling over a partial-response channel. Using a baseband channel for L subsequent time intervals results in an L-D space. Selection of an N-D signal constellation, $N \leq L$, over this space is composed of three different parts, namely, channel coding, shaping and modulation. The objective is to minimize the probability of error between the constellation points for a given total rate and total power (energy per channel use). The error is caused by the combined effect of the channel memory and the additive noise. Shaping concerns the selection of the constellation boundary at the channel input. Channel coding concerns the selection of the internal structure of the constellation at the channel output. Modulation concerns the selection of a basis, including the dimensionality, N, for the constellation.

Consider the transmission system shown in Fig. 7.1. We assume a discrete time model and block-based processing. Each block is composed of L subsequent channel uses.

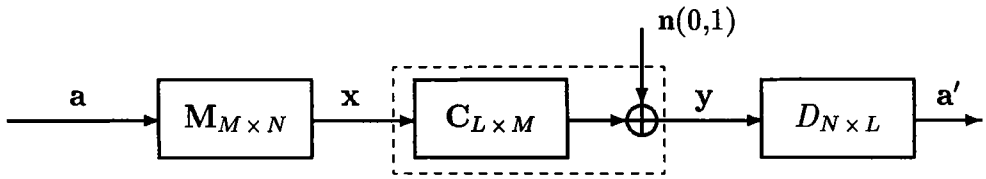


Fig. 7.1: System block diagram.

The additive noise is zero mean white Gaussian. Without loss of generality, the power gain of the channel and also the variance of the additive noise are normalized to unity.

The channel has a memory length of M_0 symbols. The channel memory results in intersymbol interference (ISI) between successive transmissions. The last M_0 transmissions of each block are zero. As a results, at the beginning of each block, the channel is in the zero state. In this case, the ISI is confined to the elements within a block and different blocks are uncorrelated. This property is obtained at the price of loosing M_0 dimensions per each block. We refer to this scheme as the zero state block-based signaling. This is the same scheme as used in [24]. Because of the zero transmissions, the transfer matrix of the channel, $\mathbf{C}$, is written in $L \times M$ form where $M = L - M_0$. The i'th column of $\mathbf{C}$ is the zero state impulse response of the channel to an impulse at time i. For a memoryless channel $M_0 = 0$ and $\mathbf{C}$ is the $L \times L$ identity matrix.

From the available $M = L - M_0$ dimensions, $N \leq M$ are used as a basis for the constellation. When N is strictly less than M, there are $M - N$ empty dimensions. In each block interval, the data bits select a point $\mathbf{a}$ within the input constellation. This is a set of N-D points bounded within a shaping region $\mathcal{R}_N$. We assume that the constellation points are used with equal probability. The columns of the $M \times N$, $M \geq N$ (modulating) matrix $\mathbf{M}$ are the basis for the constellation. We have $\mathbf{M}^t\mathbf{M} = \mathbf{I}$ where $\mathbf{I}$ is the $N \times N$ identity matrix.

At the channel output, we receive the L-D vector $\mathbf{y}$. The $N \times L$ demodulator matrix $\mathbf{D}$ is selected such that $\mathbf{DCM} = \mathbf{I}$ where $\mathbf{I}$ is the $N \times N$ identity matrix. The whole system is equivalent to an N-D identity channel with an additive Gaussian noise of autocorrelation matrix $\mathbf{DD}^t$. A nondiagonal $\mathbf{DD}^t$ corresponds to a colored noise. For a nondiagonal

$\mathbf{DD}^t$, the decisions along different dimensions are not independent of each other. We assume a suboptimum decoding method in which the decision for each dimension is made independently. In this case, the noise power along the i'th dimension, σ_i^2, is equal to the i'th diagonal element of $\mathbf{DD}^t$. For a channel with memory, $\mathbf{DD}^t$ depends on $\mathbf{M}$. Consequently, unlike the case of a flat channel, the selection of the modulating waveforms plays a role in the overall performance.

Our numerical examples are based on the $1 \pm D$ and $1 - D^2$ channels. These channels have a special importance in partial-response signaling, [23]. The $1 - D$ channel has a spectral null at zero frequency, $1 + D$ has a spectral null at the Nyquist frequency and $1 - D^2$ has spectral nulls at both zero frequency and the Nyquist frequency. The $1 \pm D$ channels have an $(M+1) \times M$ transfer matrix with the i'th column equal to $[(0)^i, \sqrt{2}/2, \pm\sqrt{2}/2, (0)^{M-1-i}]^t$. The $1 - D^2$ channel has an $(M+2) \times M$ transfer matrix with the i'th column equal to, $[(0)^i, \sqrt{2}/2, 0, -\sqrt{2}/2, (0)^{M-1-i}]^t$.

The organization of the chapter is as follows: In Section 7.2, we discuss the application of the continuous approximation. Assuming continuous approximation, shaping, channel coding and modulation can be selected independently. In this case, the selection of the basis is optimally achieved using the method of [24] and the selection of the shaping and the channel coding is similar to the case of a flat channel. The main difference is that here some of the dimensions may be empty. We propose a method to select the nonempty dimensions. This is based on minimizing the degradation caused by the channel memory. This degradation is measured in terms of the power loss with respect to a reference scheme over a unity gain flat channel with the same additive noise. Numerical results for the optimum basis and also for the Fourier basis over $1 \pm D$ channels are presented. It is shown that using the optimum basis over the $1 \pm D$ channels results in about 0.5 dB saving in energy with respect to the conventional case of the Fourier basis, while the modulation can be achieved by the use of the fast sine transform algorithms and has almost the same complexity. In the discrete case (practical restrictions on rate), shaping and coding depend on each other. In this case, a combined shaping and coding method is used. This concerns the joint selection of the shaping and coding to minimize

the probability of error. In Section 7.3, we propose two methods for this joint selection. In the first method, the minimum distance to noise ratio along all the dimensions is the same. In the second method, this restriction is relaxed. This freedom is used to reduce the effective number of the nearest neighbors of the coding lattice. Neither of these methods increases the complexity over the conventional schemes. The second method outperforms the first one.

7.2 Continuous approximation

In continuous approximation, as far as coding is concerned, the constellation is assumed to be an infinite array of points without boundary and as far as shaping is concerned, it is assumed that there are infinite points uniformly distributed within the shaping region. Assuming continuous approximation and zero state block-based signaling, it is easy to show that shaping, channel coding and modulation can be selected independently.

Shaping is similar to that for a flat channel with the difference that here some of the dimensions may remain empty. The structure of the optimum shaping region is discussed in Chapter 4. This region has equal second moments (energy) along different dimensions. In using such a region over a partial response channel, the dimensions are used in an on-off manner. In other words, a dimension has either zero energy or an amount of energy equal to other nonempty dimensions. This is compatible with the results obtained for the case of the spherical shaping region in [24].

Channel coding in this case is similar to that for a flat channel with the difference that here the coding lattice is scaled along the i'th dimension with a factor proportional to σ_i. This results in equal minimum distance to noise ratio along all the nonempty dimensions at the demodulator output. The proportionality factor is selected to adjust the total rate.

The optimum modulating basis is the basis which minimizes the product of the noise powers, [24]. This basis is found in [24] to be composed of the input eigenvectors of $\mathbf{C}$, i.e., the eigenvectors of $\mathbf{C}^t\mathbf{C}$. In a generalization of [24], it can be shown that when some

of the dimensions are empty, the optimum basis is composed of the eigenvectors of $\mathbf{C}^t\mathbf{C}$ corresponding to the largest eigenvalues. Using the optimum basis results in $\sigma_i^2 = 1/\phi_i^2$ where ϕ_i^2's are the largest eigenvalues of $\mathbf{C}^t\mathbf{C}$. For the optimum basis, $\mathbf{D}\mathbf{D}^t$ is diagonal and the noise along different dimensions is uncorrelated (independent), [24]. Later, we will introduce a method to find the value of N where $N \leq M$.

In the conventional methods, to reduce the computational complexity, the optimum basis is usually replaced by the Fourier basis. Appendix F contains the eigensystem of the channels under consideration. The eigenvectors are closely related to the sine basis. This reduces the computational complexity of the modulation by using fast sine transform algorithms.

7.2.1 Performance loss, capacity

As the decision along different dimensions is made independently, the capacity is equal to the sum of the capacities along different dimensions. The capacity is calculated by using the water filling analogy, [17]. Also, it can be shown that the basis which maximizes the capacity is composed of the eigenvectors of $\mathbf{C}^t\mathbf{C}$ corresponding to the largest eigenvalues.

Mathematically, the capacity per channel use is calculated from,

$$\begin{cases} C = \dfrac{1}{2L} \displaystyle\sum_{i:\sigma_i^2 \leq B} \log_2 \left(\frac{B}{\sigma_i^2} \right) , \\ B = E_i + \sigma_i^2 , \\ \displaystyle\sum_{i:\sigma_i^2 \leq B} E_i = LE , \end{cases} \tag{7.1}$$

where E_i is the energy allocated to i'th dimension and E is the available energy per channel use.

Let's consider the $1 \pm D$ channels with the optimum modulator. Using the results of Appendix F for the eigenvalues, we obtain,

$$\sigma_i^2 = \frac{1}{2} \sin^{-2} \left[\frac{\pi(i+1)}{2L} \right] . \tag{7.2}$$

Using (7.2) in (7.1) and considering that the noise powers are in decreasing order, we obtain,

$$B = \frac{1}{N}\left\{LE + \frac{1}{2}\sum_{i=L-N-1}^{L-2} \sin^{-2}\left[\frac{\pi(i+1)}{2L}\right]\right\} < \frac{1}{2}\cos^{-2}\left[\frac{\pi(N+1)}{2L}\right], \qquad (7.3)$$

where N is the largest integer in the range $[1, L-1]$ satisfying the right hand side inequality.

As the reference scheme, we use a flat channel with the same additive noise. The loss with respect to the reference scheme is defined as the ratio of the total energy of the two systems when the capacities per dimension are the same. Figures 7.2 and 7.3 show the loss of $1 \pm D$ channels considering the optimum basis and also the Fourier basis for $L = 9, 28$. The two curves corresponding to block-based signaling will be explained later.

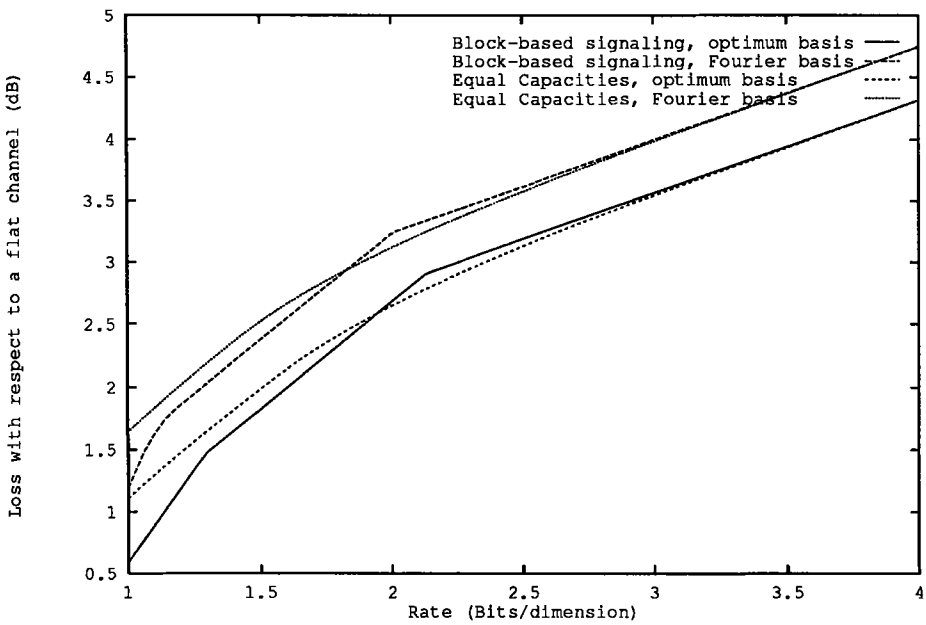


Fig. 7.2: Performance loss in $1 \pm D$ channels, $L = 9$.

An L-D, L even, $1 - D^2$ channel is obtained by time multiplexing two $L/2$-D, $1 - D$ channels. Consequently, the capacity of this channel is equal to two times the capacity of a $1 - D$ channel with a half block dimensionality.

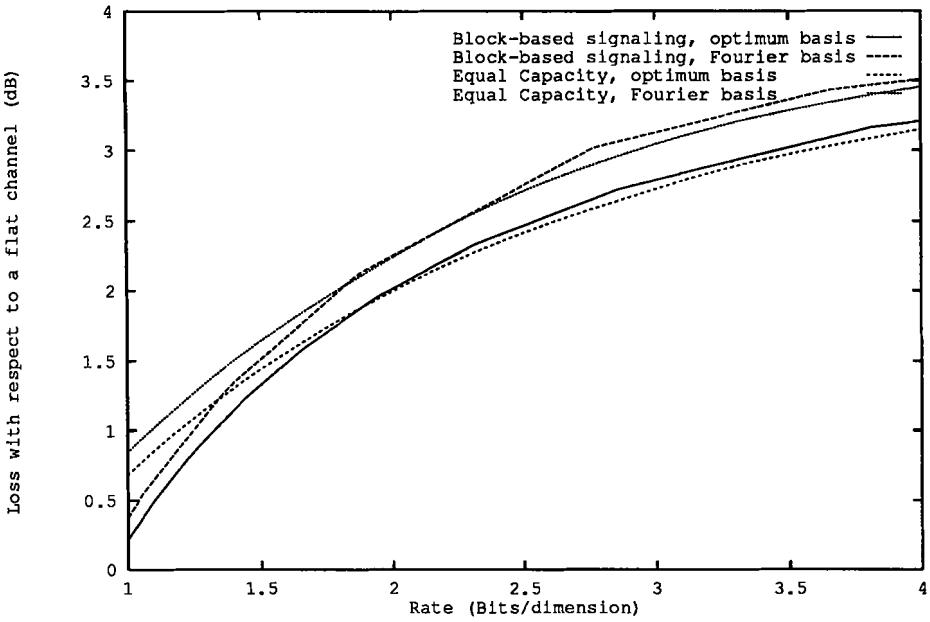


Fig. 7.3: Performance loss in $1 \pm D$ channels, $L = 28$.

Using the results of Appendix F, for $1 \pm D$ channels, we have,

$$\prod_{i=0}^{M-1} \sigma_i^2 = \frac{2^M}{M+1}, \tag{7.4}$$

and for $1 - D^2$ channel, we have,

$$\prod_{i=0}^{M-1} \sigma_i^2 = \frac{2^M}{[(M/2)+1]^2}. \tag{7.5}$$

As the block length tends to infinity, using (7.1), (7.4) and (7.5) and assuming that the E_i's are large enough such that $E_i \simeq B$, it is easy to verify that the capacity of the $1 \pm D$ and $1 - D^2$ channels are 0.5 bits per dimension less than the capacity of the reference scheme. This corresponds to 3 dB loss of energy.

We already mentioned that in using an optimum shaping region over a partial response channel, a dimension has either zero energy or an amount of energy equal to other nonempty dimensions. This strategy may seem to be in contradiction with the water filling method. This can be explained by considering that:

1. In achieving the capacity, each dimension can be extended infinitely in time. This allows for the use of the infinite dimensional shaping regions. As a result, the

calculation of capacity reduces to finding an appropriate probability distribution at the channel input[1]. However, in a finite dimensional block coding scheme, (i) the space dimensionality is fixed and, (ii) the constellation points are used with equal probability.

2. For a Gaussian channel subject to an average power constraint, capacity is achieved by a Gaussian distribution and for such a distribution there exists a one to one relationship between the energy and the entropy. This means that in maximizing the capacity, we have just one degree of freedom. On the other hand, in a zero state block based signaling scheme, the energy and the rate can be distributed independently.

3. Capacity is determined by the volume of the signal space at the channel output. This volume is maximized by having equal signal plus noise energy along all the nonempty dimensions. However, in a zero state block-based signaling scheme, the rate is determined by the volume of the signal space at the channel input. This is maximized by having equal energy along all the nonempty dimensions.

If the rate (or equivalently the energy) per dimension is sufficiently high, we obtain, (i) $M = N$, (ii) $B = E_i + \sigma_i^2 \simeq E_i$. In this case, the two strategies are equivalent.

For a fixed energy per dimension, the rate of a block-based signaling scheme is upper bounded by the rate associated with an infinite dimensional spherical shaping region. This induces an independent Gaussian distribution along each dimension. This is the same distribution as the one achieving capacity, but, here, the energy allocated to the nonempty dimensions is the same (instead of water filling).

[1]It is known that for any probability distribution on the space dimensions, there exists an infinite dimensional shaping region such that a uniform density of points within that region induces that distribution along each dimension while the distributions along different dimensions are independent.

7.2.2 Performance loss, zero state block-based signaling

Assuming continuous approximation, the rate of a block-based signaling scheme is determined by the ratio of the volume of the shaping region, $V(\mathcal{R}_N)$, and the volume of the Voronoi region around each constellation point, $V(\mathcal{V}_N)$. We assume that the constellation is composed of the points of the scaled half integer grid, $Z^N + (1/2)^N$, bounded within a hypercube. This can be easily generalized to other coding lattices and shaping regions. Without loss of generality, we assume that the scale factor along the i'th dimension is equal to σ_i. This results in a minimum distance to noise ratio of one along all the nonempty dimensions.

As the reference scheme, we use an L-D cubic constellation (no shaping) over a flat channel. The total rate and the minimum distance to noise ratio of the two schemes are equal. The rate to be transmitted per channel use is equal to $r_f = \log_2 T_f$ where T_f is the edge length of the shaping hypercube of the reference scheme. For the flat channel, we have, $V_f(\mathcal{R}_L) = (T_f)^L$ and $V_f(\mathcal{V}_L) = 1$. The edge length of the shaping hypercube over the partial-response channel is denoted by T_p. For the partial-response channel, we have $V(\mathcal{R}_N) = (T_p)^N$ and $V(\mathcal{V}_N) = \prod_{i=0}^{N-1} \sigma_i$. Equating the total rates, we obtain,

$$(T_f)^L = \frac{(T_p)^N}{\prod_{i=0}^{N-1} \sigma_i}. \tag{7.6}$$

Using continuous approximation over a hypercube, the average energy of the coded scheme and the reference scheme are equal to, $E_p = N(T_p)^2/12$ and $E_f = L(T_f)^2/12$, respectively. Using (7.6), the loss in energy with respect to the reference scheme is found as,

$$10\log_{10}\left(\frac{E_p}{E_f}\right) = 6r_f\left(\frac{L}{N} - 1\right) + \frac{10}{N}\sum_{i=0}^{N-1}\log_{10}\left(\sigma_i^2\right) - 10\log_{10}\left(\frac{L}{N}\right). \tag{7.7}$$

The number of nonempty dimensions is determined by minimizing (7.7) over $N \leq M$ where $M = L - M_0$. The result is affected by two conflicting phenomena: (i) A larger number of empty dimensions decreases $V(\mathcal{V}_N) = \prod_{i=0}^{N-1} \sigma_i$. (ii) For a fixed total second moment, decreasing the dimensionality results in a lower $V(\mathcal{R}_N)$. The results of these calculations for $L = 9, 28$ and $1 \pm D$ channels using the optimum basis and also the Fourier basis are

shown in Figs. 7.2 and 7.3. Different slopes correspond to different number of empty dimensions.

In general, for higher rate and/or for lower dimensionality, the difference between the optimum basis and the Fourier basis is more pronounced. For a moderate rate per dimension and a moderate dimensionality, the saving in energy due to the use of the optimum basis instead of the Fourier basis is in the order of 0.5 dB. As the computational complexity of the optimum basis (sine basis) and the Fourier basis are almost equivalent, this gain is obtained with no additional cost.

As we implicitly mentioned in the previous section, for large values of rate, the loss given in (7.7) is equal to the performance loss obtained by equating the capacities. Figures 7.2 and 7.3 show this fact in the specific case of $1 \pm D$ channels.

7.3 Discrete case

In the discrete case, shaping and coding are interrelated and one gains by using a joint optimization procedure. The major coupling is due to the addressing scheme. To obtain a tractable addressing complexity, one should impose restrictions on the rate distribution. For example, in our case, the number of the points in each 2-D subconstellation is of the form $2^{R_j}(1+2/N)$ where N (constellation dimensionality) is an integral power of two and the R_j's are integer numbers greater than or equal to $\log_2(N/2)$.

7.3.1 Shaping method

We generalize the shaping method introduced in [42] to the case that there are different number of points in different 2-D subconstellations. In general, to transmit R bits per two dimensions in an $N=2n$-D TCM (Trellis-Coded-Modulation) scheme with one bit redundancy, a signal constellation with 2^{nR+1} points is needed. Assume that the rate allocated to the j'th subspace is equal to R_j. To construct the constellation, the 2-D subconstellations are divided into an inner group and an outer group. The number of

R	$A(R)$	$A_c(R)$
4	3.0	3.2
5	5.9	6.3
6	11.9	12.7
7	23.6	25.4

Table 7.1: Average energy per two dimensions as a function of the rate for a minimum distance of one, $N = 8$.

points in the inner group of the j'th subconstellation is equal to 2^{R_j}. The integer numbers R_j satisfy $\sum_j R_j = nR$. The number of points in the outer group is $1/n$ of that in the inner group. This is possible if n is a power of two and $R_j \geq \log_2(n)$. The inner and the outer group of the 2-D subconstellations have the same structure as in [42]. The N-D constellation is constructed by concatenating n such 2-D subconstellations and excluding the N-D points corresponding to more than one 2-D outer point. Addressing is achieved by a lookup table with $(1 + n \log_2 n)$ input lines and $n\lceil \log_2(1+n) \rceil$ output lines.

Similar to the case of [42], for each 2-D subconstellation, the inner group is used $N-1$ times as often as the outer group. This means that the average energy per two dimensions is equal to $(N-1)/N$ times the average energy of the inner group plus $1/N$ times the average energy of the outer group. Table 7.1 shows the average energy per two dimensions, A, for $N = 8$ and for a minimum distance of one, as a function of the rate per two dimensions, R. The A values will be used later. Column $A_c(R)$ is the energy obtained by applying continuous approximation to a cubic shaping region, $A_c(R) = 2^{(R+0.25)}/6$. The $A_c(R)$ values are provided to indicate the accuracy of a continuous approximation.

7.3.2 Channel coding

The points of the 2-D subconstellations belong to the half integer grid. Different 2-D subconstellations are scaled with different scale factors. Consequently, the N-D points

belong to a scaled version of the N-D half integer grid. This is partitioned into the cosets of a (scaled) sublattice shifted to the point $(1/2)^N$. The unscaled version of this sublattice is denoted as the baseline lattice. It should be mentioned that scaling does not change the group property of the lattices. Consequently, the coset decomposition has the same properties as the unscaled case.

In each signaling interval some of the data bits are encoded and used to select one of the cosets. The rest of the data bits select a point within the selected coset. We assume that the dominant error event is the error within a coset. This is the case in most of the TCM schemes. The numerical examples are based on dimensionality eight and the baseline lattice E_8.

7.3.3 Weight distribution of the scaled lattices

The weight distribution of a set of points Λ with respect to a given center is defined as, [4, p.45],

$$\Theta_\Lambda(q) = \sum_{u\in\Lambda} q^{\|u\|^2} = \sum_x N(x) q^x, \tag{7.8}$$

where $\|u\|$ is the norm of vector associated with point u and $N(x)$ is the number of the points at square distance x from the center. For a set of points with the distance invariance property, the weight distribution function is independent of the center. This is the case for a scaled lattice (a consequence of the group property).

Assume that the square minimum distance along the j'th 2-D subconstellation ($j \in [0, N/2-1]$) is equal to D_j. We use the trellis diagram of the lattice, [13], to calculate the weight distribution of the scaled lattice. Each branch in the diagram is labeled by the weight distribution of the corresponding 2-D coset. The weight distribution of a path is obtained by multiplying the weight distribution of its branches. The weight distribution of the scaled lattice is obtained by adding the weight distribution of the parallel paths in the diagram. Using this approach, we have derived new results for the weight distribution of the scaled D_4 and E_8 lattices. The final results are,

$$\Theta_{D_4}(q) = \theta_3^2(q_0)\theta_3^2(q_1) + \theta_2^2(q_0)\theta_2^2(q_1),\ q_j = q^{2D_j},\ j = 0, 1\,, \tag{7.9}$$

and,

$$\begin{aligned}
&\Theta_{E_8}(q) = \\
&\theta_3^2(q_0)\theta_3^2(q_1)\theta_3^2(q_2)\theta_3^2(q_3) + \theta_3^2(q_0)\theta_3^2(q_1)\theta_2^2(q_2)\theta_2^2(q_3) + \\
&\theta_3^2(q_0)\theta_2^2(q_1)\theta_3^2(q_2)\theta_2^2(q_3) + \theta_3^2(q_0)\theta_2^2(q_1)\theta_2^2(q_2)\theta_3^2(q_3) + \\
&\theta_2^2(q_0)\theta_3^2(q_1)\theta_3^2(q_2)\theta_2^2(q_3) + \theta_2^2(q_0)\theta_3^2(q_1)\theta_2^2(q_2)\theta_3^2(q_3) + \\
&\theta_2^2(q_0)\theta_2^2(q_1)\theta_3^2(q_2)\theta_3^2(q_3) + \theta_2^2(q_0)\theta_2^2(q_1)\theta_2^2(q_2)\theta_2^2(q_3) + \\
&8\theta_2(q_0)\theta_2(q_1)\theta_2(q_2)\theta_2(q_3)\theta_3(q_0)\theta_3(q_1)\theta_3(q_2)\theta_3(q_3), \\
&q_j = q^{4D_j},\ j = 0,1,2,3\,,
\end{aligned} \tag{7.10}$$

where θ_2 and θ_3 are the Jacobi theta functions, [4, p. 101]. To the extent of our knowledge, this is the first time that the above weight distributions have appeared in the literature. Later, the weight distribution will be used to calculate the error probability.

7.3.4 Probability of error

For an additive Gaussian noise of power σ^2, the probability of error between two points with distance d is upper bounded by, [4, pp. 69–73],

$$p_0 \leq \frac{1}{2}\exp\left(-d^2/8\sigma^2\right), \tag{7.11}$$

Using (7.11) in the union bound results in an upper bound for error probability. This bound is equal to, [4, p. 73],

$$P_e \leq \frac{1}{2}\,\Theta_\Lambda\left[\exp\left(-1/8\sigma^2\right)\right] - \frac{1}{2}, \tag{7.12}$$

where Θ_Λ is given in (7.8).

In practice, we truncate the weight distribution to the set of the nearest neighbors. For the lattice E_8, truncating (7.10) to the set of the 240 nearest neighbors, results in,

$$\begin{aligned}
P_e \simeq \frac{1}{2}\,(&4Z_0^4 + 4Z_1^4 + 4Z_2^4 + 4Z_3^4 + 16Z_0^2Z_1^2 + 16Z_0^2Z_2^2 + 16Z_0^2Z_3^2 + \\
&16Z_1^2Z_2^2 + 16Z_1^2Z_3^2 + 16Z_2^2Z_3^2 + 128Z_0Z_1Z_2Z_3\Big)\,,
\end{aligned} \tag{7.13}$$

where $Z_j = \exp(-D_j/2N_j)$ and N_j,D_j are the noise power and the square minimum distance in the j'th 2-D subspace. We use the notation $P_e \simeq F(Z_j, j=0,1,2,3)$ to indicate the function in (7.13).

7.3.5 Problem statement

We have a zero state block-based signaling scheme with M dimensions. The number of nonempty dimensions is equal to N. The total energy is equal to E_t. We use the Fourier basis for modulation. The two dimensions with the equal noise power constitute a 2-D subchannel. A 2-D subconstellation is employed over each subchannel.

The nonempty subchannels are indexed by $i \in [0, \ldots, n-1]$, $n = N/2$. These are divided into K groups each of N_c dimensions. Each group uses an independent TCM scheme. The total rate is equal to $R_t = nR + K$ corresponding to R bits per each nonempty subchannel and one bit redundancy for each coding group.

The 2-D subconstellations are indexed by (k,j) where $k \in [0, K-1]$ is the index of the group and $j \in [0, n_c - 1], n_c = N_c/2$ is the index within the group. The noise power, the square minimum distance and the minimum distance to noise ratio (protection) of the (k,j)'th 2-D subconstellation are shown by, N_j^k, D_j^k and $P_j^k = D_j^k/N_j^k$, respectively. The corresponding rate and energy are related by,

$$E_j^k = A(R_j^k) \times D_j^k, \tag{7.14}$$

where $A(R)$ is given in Table 7.1.

The total gain of the system, γ_t, is defined as the savings in energy with respect to a reference system with the same probability of error. The reference system uses a one-D flat channel with unity gain and is composed of the points of the one-D half integer grid (no coding) bounded within $[-2^{(Q/2)-1}, 2^{(Q/2)-1}]$ (no shaping), Q denotes the rate per two dimensions of the reference scheme and is given by, $Q = NR/L$.

For the reference system, we assume continuous approximation and use (7.11) for the error probability. Assuming continuous approximation, the average energy per dimension of the reference scheme for a minimum distance of one is equal to, $2^Q/12$. The number

of the nearest neighbors in the reference system is equal to 2. For an additive Gaussian noise of unity power and total energy E_r, the probability of error of the reference system within a block of length L is approximately equal to,

$$P_e \simeq L \exp\left(-\frac{3E_r}{L\,2^{Q+1}}\right). \tag{7.15}$$

Equating the error probabilities, γ_t is equal to the ratio of the energies, i.e.,

$$\gamma_t = \frac{E_r}{E_t} = \frac{L\,2^{Q+1}}{3\,E_t} \log\left(\frac{L}{P_e}\right). \tag{7.16}$$

The γ_t reflects: (i) the shaping gain, (ii) the coding gain, and (iii) the degradation caused by the channel memory. This degradation is due to a loss in dimensionality and/or having $\prod_i \sigma_i$ greater than one. Unlike to the case of the continuous approximation, it is not possible to separately identify the effects of these three factors in γ_t

We are looking for R_j^k's, E_j^k's and a rule for grouping the subchannels. The grouping rule is expressed in terms of the one-to-one assignment $(j,k) \Longleftrightarrow i$, where (j,k) is the index of the 2-D subconstellation and i is the index of the 2-D subchannel. The objective is to maximize γ_t or equivalently to minimize P_e. As R is an integer, it is usually impossible to change N while keeping the total data rate, $NR/2$, constant. As a result, different systems obtained by changing N cannot be easily compared. This means that N should be considered as a fixed parameter and then the scheme can be compared with the other possibilities.

7.3.6 First method: Equal protection along the subchannels

For equal protection, we set $P_j^k = P_0$, $\forall j,k$. In this case, the optimization problem is formulated as:

$$\begin{cases} \text{Maximize} & P_0 \\ \text{Subject to:} & \displaystyle\sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} R_j^k = nR, \quad R_j^k \in \mathbf{N},\ R_j^k \geq \log_2(n_c) \\ & \displaystyle\sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} E_j^k = E_t, \quad E_j^k \geq 0, \end{cases} \tag{7.17}$$

where $\mathbf{N}$ is the set of integers. Combining (7.14) and (7.17) results in,

$$P_0 = \frac{E_t}{\sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} A(R_j^k)N_j^k}. \tag{7.18}$$

Using (7.18), the optimization problem in (7.17) reduces to,

$$\begin{cases} \text{Minimize} & \sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} A(R_j^k)N_j^k \\ \text{Subject to:} & \sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} R_j^k = nR, \quad R_j^k \in \mathbf{N}, \; R_j^k \geq \log_2(n_c). \end{cases} \tag{7.19}$$

In this case, the assignment $(j,k) \Longleftrightarrow i$ is arbitrary. This problem is solved by the following algorithm:

1. Set $R_j^k = \log_2(n_c)$, $\forall j,k$. Another $n[R - \log(n_c)]$ bits remain to be distributed.

2. Allocate one bit to the 2-D subconstellation with the smallest $\left[A(R_j^k+1) - A(R_j^k)\right] N_j^k$. Update the rates. If there are still bits to be distributed go to step 2, otherwise quit.

7.3.7 Second method: Nonequal protection along the subchannels

In this case, we minimize the average error probability of the whole system. This is formulated as,

$$\begin{cases} \text{Minimize} & \sum_{k=0}^{K-1} F(Z_j^k, j=0,1,2,3), & Z_j^k = \exp(-D_j^k/2N_j^k) \\ \text{Subject to:} & \sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} R_j^k = nR, & R_j^k \in \mathbf{N}, \; R_j^k \geq \log_2(n_c) \\ & \sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} E_j^k = E_t, & E_j^k \geq 0, \end{cases} \tag{7.20}$$

where D_j^k is related to E_j^k and R_j^k by (7.14). The rate/energy distribution and the assignment rule $(j,k) \Longleftrightarrow i$ are determined by a two part iterative procedure. The first part

itself is another iterative procedure and finds the optimum rate distribution for a given energy distribution and vice versa. As the starting point, we use the answer obtained by applying the first method. In the second part, we find the optimum assignment rule $(j,k) \Longleftrightarrow i$ for the final answer of the first step. Then the two parts repeat to improve on the solution.

Part 1: Optimum rate distribution for a given energy distribution

Following algorithm is used to find the rate distribution:

1. Set $R_j^k = \log_2(n_c)$, $\forall j,k$. Another $n[R - \log(n_c)]$ bits remain to be distributed.

2. Arrange the 2-D subconstellations according to the value of $E_j^k / A(R_j^k) N_j^k$ (protection) in the decreasing order and index them with $i_1 \in [0, n-1]$.

3. Arrange the 2-D subconstellations according to the value of $E_j^k / A(R_j^k + 1) N_j^k$ in decreasing order and index them with $i_2 \in [0, n-1]$.

4. Find the smallest integer $m \leq n$ such that for $i_1, i_2 \in [0, m-1]$ the elements in the set indexed by i_1 are obtained by the permutation of the elements in the set indexed by i_2. These 2-D subconstellations are the candidates for receiving the next bit.

5. Allocate one bit to the candidate which by receiving it will result in the least increase in the objective function. Update the rates. If there are still bits to be distributed go to step 2, otherwise quit.

It can be shown that this method gives the same answer as if the same search is performed over the set of all the subconstellations.

Part 2: Optimum energy distribution for a given rate distribution

The objective function in (7.20) is a convex $\cup$ function of E_j^k's. As a result, the global optimum point over the convex region determined by the energy constraint is determined

by the Lagrange method. This results in the following set of equations for E_j^k's,

$$\begin{cases} \dfrac{\partial}{\partial Z_j^k}\left[F(Z_j^k, j=0,1,2,3)\right] = \lambda N_j^k A(R_j^k) \\ \displaystyle\sum_{k=0}^{K-1}\sum_{j=0}^{n_c-1} E_j^k = E_t. \end{cases} \tag{7.21}$$

This set of equations is solved by an iterative method.

Assignment $(j,k) \Longleftrightarrow i$

This problem is solved by the following algorithm:

1. Arrange the nonempty subchannels according to the value of the noise power in increasing order and index them by $i_1 \in [0, n-1]$.

2. Arrange the 2-D subconstellations according to the value of P_j^k (protection) in the increasing order and index them by $i_2 \in [0, n-1]$.

3. Assign the members of the two sets with the same index to each other.

It can be shown that for a given rate and energy distribution, this assignment rule minimizes the probability of error.

Special cases

We can show that if there exists a rate distribution such that $A(R_j^k)N_j^k = \text{constant}$, then this is optimum for the energy distribution $E_j^k = E_t/n$ and vice versa. This results in equal protection along the subconstellations. Another special case arises when in a given energy updating step we obtain $E_j^k = E_t/n$. In this case, if the total rate is a multiple of n, the optimum rate distribution will be of the form $R_j^k = R$. The converse is true if the noise powers along different dimensions are equal.

Note: The proof of the optimality of the methods and the claims in subsubsection 7.3.7 are based on certain properties of the function $F(Z_j)$ given in (7.13). The same properties are valid in the case of the lattice D_4.

Example:

In this example, a zero state block-based signaling scheme over $(1 \pm D)$ channels is studied. The total number of dimensions is equal to $L = 28, 30$ and $N = 24$ dimensions are nonempty. We consider $R = 2, 3$, corresponding to a total data rate of $NR/2 = 24, 36$ bits. There are $K = 3$ coding groups each of dimensionality $N_c = 8$. Lattice E_8 is used as the baseline lattice. Over a flat channel, the corresponding TCM scheme results in a total gain of 5.41 dB (channel coding gain plus the shaping gain), [42]. This gain does not include the effect of the error multiplicity. We apply both of our design methods to this problem. The performance is measured in terms of the total gain, γ_t, and the probability of error. Figures 7.4 through 7.7 show these parameters as a function of the energy per dimension, E_t/L.

In general, the improvement of the second method is almost equivalent to multiplying the probability of error by a constant factor. This can be considered as reducing the number of the nearest neighbors to some smaller effective value. A justification of this phenomenon is obtained by referring to (7.13). This equation is composed of the sum of 240 terms. Each term corresponds to one of the nearest neighbors of the lattice E_8. It is seen that the protection (square minimum distance to noise ratio) from the center to 1/15 of the neighbors is determined by the protection along only one 2-D subconstellation. The protection to 6/15 of them is determined by the sum of the protections along two of the 2-D subconstellations. Finally, the protection to 8/15 of the neighbors is determined by the sum of the protections along all the 2-D subconstellations. When the protections are added it makes no difference which subconstellation has a larger effect on the sum. This flexibility is used by the optimization algorithm to reduce the effective number of the nearest neighbors. We note that for lattices like the Leech lattice with 196560 nearest neighbors, [4, p. 133], the improvement due to the second method will be more pronounced.

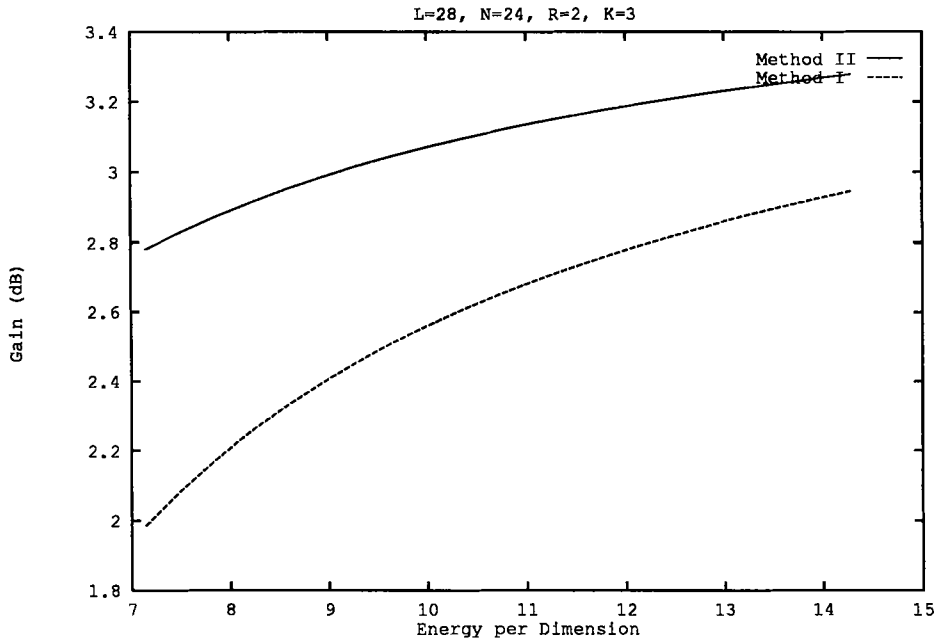


Fig. 7.4: Total gain as a function of the energy per dimension (E_t/L) for $L=28$, $N=24$, $R=2$.

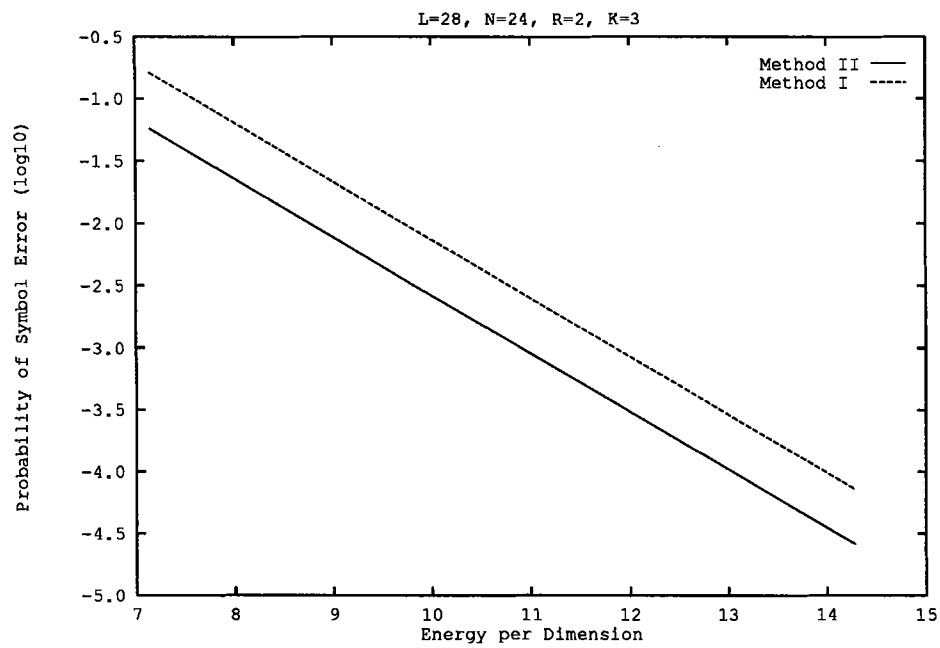


Fig. 7.5: Probability of symbol error as a function of the energy per dimension (E_t/L) for $L=28$, $N=24$, $R=2$.

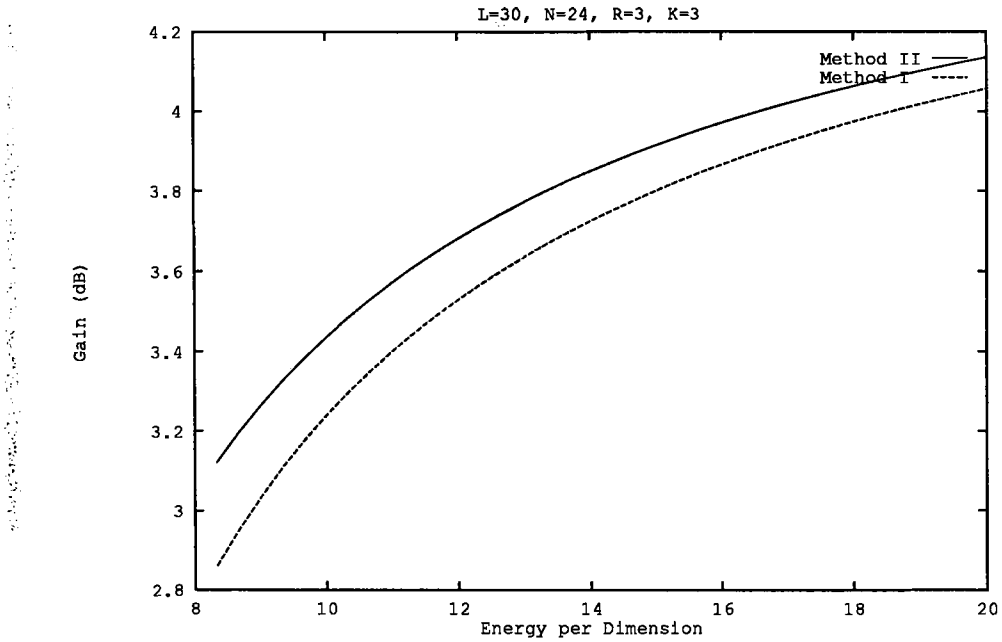


Fig. 7.6: Total gain as a function of the energy per dimension (E_t/L) for $L=30$, $N=24$, $R=3$.

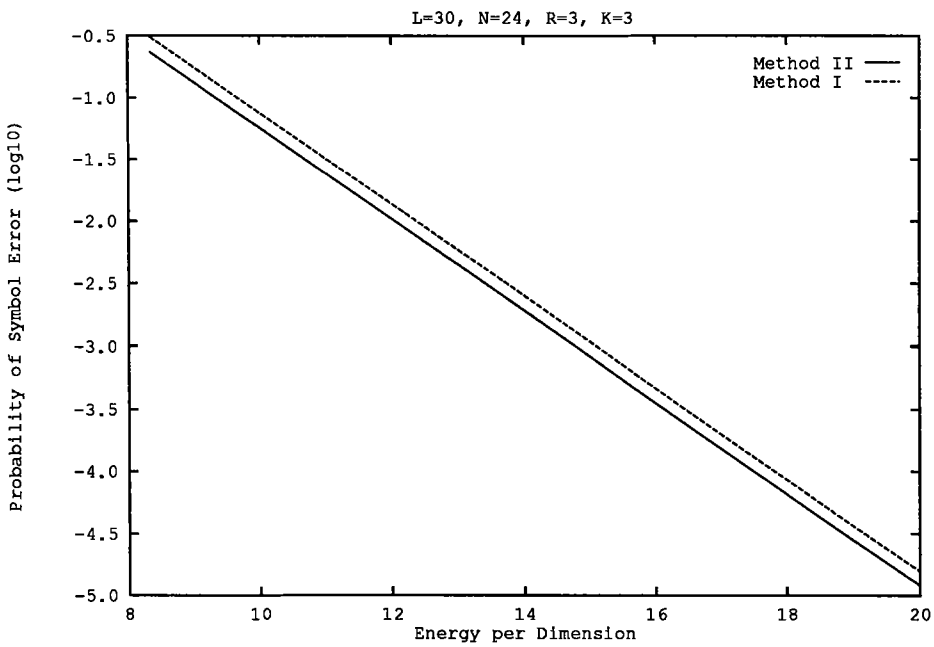


Fig. 7.7: Probability of symbol error as a function of the energy per dimension (E_t/L) for $L=30$, $N=24$, $R=3$.

7.4 Summary and conclusions

We have discussed the selection of a signal constellation for signaling over a partial response channel. Assuming continuous approximation, shaping, channel coding and modulation are selected independently. The only unknown in this case is the selection of the nonempty dimensions. In the proposed scheme, this selection is based on minimizing the loss with respect to a reference scheme. It is shown that using the optimum basis over the $1 \pm D$ channels results in about 0.5 dB saving in energy with respect to the Fourier basis, while the computational complexities are almost equivalent. In the discrete analysis, shaping and coding are jointly selected to minimize the error probability. This is denoted as the combined shaping and coding. In the combined case, instead of dealing with the rate as a continuous variable and then rounding the result, we have used an integer optimization procedure for the rate allocation. Two different schemes have been proposed. The first scheme has equal minimum distance to noise ratio along all the nonempty dimensions. In the second scheme, this constraint is removed. On the basis of the average error probability, the second method outperforms the first one. Neither of the two schemes increases the complexity over the conventional schemes. As part of the calculations, we have found a closed form formula for the weight distribution of the scaled D_4 and E_8 lattices.

Appendix A

Integral of $F(X_0^2 + \ldots + X_{N-1}^2)$ over the $\mathcal{A}_N$ region

We are going to calculate the integral of the function $F(X_0^2 + \ldots + X_{N-1}^2)$ over the region $\mathcal{A}_N$ defined by (4.1). The calculation is based on decomposing the region $\mathcal{TC}_n$, defined by (4.4), into the union and intersection of the simplexes and applying the Dirichlet's integral, [43], to each of them. An example of this decomposition for $N=4$, $n=2$ is shown in Fig. A.1.

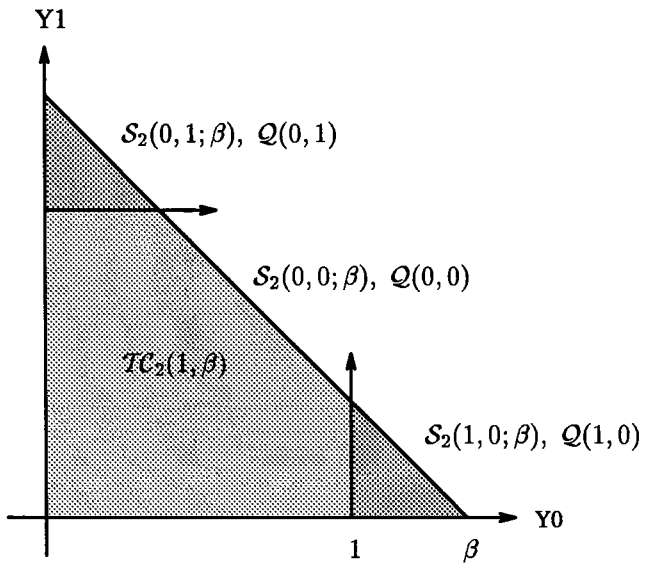
Applying the change of variable in (4.3), we obtain,

$$I = \int_{\mathcal{A}_N} F(X_0^2 + \ldots + X_{N-1}^2) = (\pi R_2^2)^n \int_{\mathcal{TC}_n} F\left[R_2^2(Y_0 + \ldots + Y_{n-1})\right] . \tag{A.1}$$

where $\mathcal{TC}_n$ is defined in (4.4).

Define the n-D regions,

$$\begin{aligned} \mathcal{S}_n(\alpha_0, \alpha_1, \ldots, \alpha_{n-1}; \beta) \; &= \{Y_p,\ p = 0, \ldots, n-1\} \\ &: \; \alpha_0 \le Y_0 \le B + \alpha_0, \\ &\quad \alpha_1 \le Y_1 \le B + \alpha_1 - Y_0, \\ &\quad \ldots\ldots \\ &\quad \alpha_{n-1} \le Y_{n-1} \le (B + \alpha_{n-1} - Y_0 - \ldots - Y_{n-2}); \\ &\quad \text{where } \; B = \beta - \textstyle\sum_i \alpha_i \; , \end{aligned} \tag{A.2}$$



$$\mathcal{TC}_2(1,\beta) = \mathcal{S}_2(0,0;\beta) - \mathcal{S}_2(0,1;\beta) - \mathcal{S}_2(1,0;\beta)$$

Fig. A.1: Example of decomposing $\mathcal{TC}_2$ into simplexes.

$$\begin{aligned} \mathcal{C}_n(\alpha_0, \alpha_1, \ldots, \alpha_{n-1}) &= \{Y_p,\ p = 0, 1, \ldots, n-1\} \\ &: \ \alpha_p \leq Y_p \leq 1 + \alpha_p \ , \end{aligned} \tag{A.3}$$

and,

$$\begin{aligned} \mathcal{Q}_n(\alpha_0, \alpha_1, \ldots, \alpha_{n-1}) &= \{Y_p,\ p = 0, 1, \ldots, n-1\} \\ &: \ Y_p \geq \alpha_p \ . \end{aligned} \tag{A.4}$$

Using these notations, the region $\mathcal{TC}_n$ can be written as,

$$\mathcal{TC}_n = \mathcal{S}_n\left((0)^n; \beta\right) \cap \mathcal{C}_n\left((0)^n\right) \ . \tag{A.5}$$

We also define,

$$\mathcal{Q}_n\left[(0)^{n-k}, (1)^k\right] = \sum \mathcal{Q}_n(i_0, \ldots, i_{n-1}) \ , \tag{A.6}$$

where $i_0, \ldots, i_{n-1}$ is an n-tuple with k ones and $n-k$ zeros and the summation is calculated over all the C_n^k possible combinations of this type. Using these definitions, we can write,

$$\mathcal{C}_n\left((0)^n\right) = \sum_{k=0}^{n} (-1)^k \mathcal{Q}_n\left[(0)^{n-k}, (1)^k\right] \ , \tag{A.7}$$

Using (A.7) in (A.5) we obtain,

$$\mathcal{TC}_n = \sum_{k=0}^{n} (-1)^k \left\{ \mathcal{S}_n \left((0)^n ; \beta \right) \cap \mathcal{Q}_n \left[(0)^{n-k}, (1)^k \right] \right\} . \tag{A.8}$$

It is easy to verify that,

$$\mathcal{S}_n \left((0)^n ; \beta \right) \cap \mathcal{Q}_n \left[(0)^{n-k}, (1)^k \right] = \emptyset, \quad \text{for} \quad k \geq \lfloor \beta \rfloor + 1 ,$$

where $\lfloor \beta \rfloor$ denotes the largest integer smaller than β. Combining (A.6) and (A.8) and using,

$$\mathcal{S}_n \left((0)^n ; \beta \right) \cap \mathcal{Q}_n (i_0, \ldots, i_{n-1}) = \mathcal{S}_n (i_0, \ldots, i_{n-1} ; \beta) . \tag{A.9}$$

we obtain,

$$\mathcal{TC}_n = \sum_{k=0}^{\lfloor \beta \rfloor} (-1)^k \sum \mathcal{S}_n (i_0, \ldots, i_{n-1} ; \beta) , \tag{A.10}$$

where the second summation is calculated over all the combinations of $(i_0, \ldots, i_{n-1})$ with k ones and $n-k$ zeros. An example of this summation is shown in Fig. (A.1).

The integrand in (A.1) is symmetric with respect to the variables and any permutation of the variables does not change its value. Consequently, the integral over the region $\mathcal{S}_n (i_0, \ldots, i_{n-1} ; \beta)$ is independent of the permutation applied to $i_0, \ldots, i_{n-1}$. We calculate this integral over one of these regions, say over $\mathcal{P}_n (k, \beta) = \mathcal{S}_n \left((1)^k, (0)^{n-k} ; \beta \right)$. and multiply the result by C_n^k. This results in,

$$I = (\pi R_2^2)^n \sum_{k=0}^{\lfloor \beta \rfloor} (-1)^k C_n^k \int_{\mathcal{P}_n (k,\beta)} F \left[R_2^2 (Y_0 + \ldots + Y_{n-1}) \right] . \tag{A.11}$$

The integral over the region $\mathcal{P}_n (k, \beta)$ in (A.11) can be written as,

$$\int_{\mathcal{P}_n (k,\beta)} F \left[R_2^2 (Y_0 + \ldots + Y_{n-1}) \right] = \int_0^{\beta-k} \int_0^{\beta-k-Y_0} \cdots \int_0^{\beta-k-Y_0-\ldots-Y_{n-2}} F \left[R_2^2 (Y_0 + \ldots + Y_{n-1} + k) \right] . \tag{A.12}$$

The region of integration in (A.12) is a simplex of edge length $\beta - k$. Applying the Dirichlet's integral, [43], to this simplex results in (4.5).

Appendix B

Limiting behavior for the $\mathcal{A}_N$ region

To study the asymptotic behavior of the optimum shaping region as the number of dimensions tend to infinity, we find the induced probability distribution along the 2-D subspaces and show that as the number of dimensions tends to infinity these distributions become independent truncated Gaussian distributions. It is easy to show that a uniform distribution within the N-D region $\mathcal{A}_N$ defined by (4.1) results in a uniform distribution within the n dimensional region $\mathcal{TC}_n$ defined by (4.4). Using (4.5) to find the volume, we calculate the induced probability distribution along a given dimension, say Y, of $\mathcal{TC}_n$. For $0 \leq Y < \beta - \lfloor \beta \rfloor$, we obtain,

$$P(Y) = \frac{\sum_{k=0}^{\lfloor \beta \rfloor} (-1)^k C_{n-1}^k (\beta - k - Y)^{n-1}}{\sum_{k=0}^{\lfloor \beta \rfloor} (-1)^k C_n^k (\beta - k)^n}, \tag{B.1}$$

and for $\beta - \lfloor \beta \rfloor \leq Y \leq 1$, we obtain,

$$P(Y) = \frac{\sum_{k=0}^{\lfloor \beta \rfloor - 1} (-1)^k C_{n-1}^k (\beta - k - Y)^{n-1}}{\sum_{k=0}^{\lfloor \beta \rfloor} (-1)^k C_n^k (\beta - k)^n}. \tag{B.2}$$

In (B.1) and (B.2), the denominator is equal to the volume of the region $\mathcal{TC}_n(1, \beta)$ defined by (4.4) and the numerator is equal to the volume of the region $\mathcal{TC}_{n-1}(1, \beta - Y)$. Equation

(B.2) can be written as,

$$P(Y) = \frac{\sum_{k=0}^{\lfloor\beta\rfloor}(-1)^k C_{n-1}^k(\beta-k-Y)^{n-1} - (-1)^{\lfloor\beta\rfloor} C_{n-1}^{\lfloor\beta\rfloor}(\beta-\lfloor\beta\rfloor-Y)^{n-1}}{\sum_{k=0}^{\lfloor\beta\rfloor}(-1)^k C_n^k(\beta-k)^n}. \tag{B.3}$$

In (B.3), we have, $0 \leq Y-\beta+\lfloor\beta\rfloor < 1$, as a result, as n tends to infinity, the second term in the numerator tends to zero. Assuming an integer value for β, we obtain the following general expression,

$$P(Y) = \frac{\sum_{k=0}^{\beta}(-1)^k C_{n-1}^k(\beta-k-Y)^{n-1}}{\sum_{k=0}^{\beta}(-1)^k C_n^k(\beta-k)^n}, \quad \beta \in \mathbf{N}. \tag{B.4}$$

Defining,

$$D(n,\beta,Y) = \frac{1}{n!}\sum_{k=0}^{\beta}(-1)^k C_n^k(\beta-k-Y)^n, \tag{B.5}$$

Equation (B.4) is expressed as,

$$P(Y) = \frac{D(n-1,\beta,Y)}{D(n,\beta,0)}. \tag{B.6}$$

Using the identity, $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k$, it is easy to show that the function $D(n,\beta,Y)$ and its derivative satisfy the following recursions,

$$D(n,\beta,Y) = \frac{\beta-Y}{n}\{D(n-1,\beta,Y) - D(n-1,\beta-1,Y)\} + D(n-1,\beta-1,Y), \tag{B.7}$$

$$\frac{\partial^m D(n,\beta,Y)}{\partial^m Y} = \sum_{i=0}^{m}(-1)^{m+i} C_m^i D(n-m,\beta-i,Y). \tag{B.8}$$

Recursion (B.7) can be used to express $D(n,\beta,Y)$ as a function of $D(\alpha,\alpha,Y)$ where α is an integer smaller than or equal to β. Using the properties of the Stirling numbers of the second kind, it can be shown that $D(\alpha,\alpha,Y) = 1, \ \forall\alpha$.

To show that the probability distribution on each 2-D subspace is a Gaussian distribution truncated within a circle, it is enough to show that $P(Y)$ is of exponential form, or equivalently,

$$\frac{1}{P(Y)}\frac{\partial^2 P(Y)}{\partial^2 Y} = \left[\frac{1}{P(Y)}\frac{\partial P(Y)}{\partial Y}\right]^2. \tag{B.9}$$

Using (B.6), we can write,

$$\frac{1}{P(Y)}\frac{\partial^2 P(Y)}{\partial^2 Y} = \frac{1}{D(n-1,\beta,Y)}\frac{\partial^2 D(n-1,\beta,Y)}{\partial^2 Y}\,, \tag{B.10}$$

$$\frac{1}{P(Y)}\frac{\partial P(Y)}{\partial Y} = \frac{1}{D(n-1,\beta,Y)}\frac{\partial D(n-1,\beta,Y)}{\partial Y}\,. \tag{B.11}$$

Using (B.7) for large values of n and defining $\psi = n/\beta$, we obtain,

$$D(n-1,\beta,Y) = \frac{1}{\psi}\left\{D(n-2,\beta,Y) - D(n-2,\beta-1,Y)\right\} + D(n-2,\beta-1,Y)\,. \tag{B.12}$$

Using (B.8), we also obtain,

$$\frac{\partial^2 D(n-1,\beta,Y)}{\partial^2 Y} = D(n-3,\beta,Y) - 2D(n-3,\beta-1,Y) + D(n-3,\beta-2,Y)\,, \tag{B.13}$$

$$\frac{\partial D(n-1,\beta,Y)}{\partial Y} = D(n-2,\beta-1,Y) - D(n-2,\beta,Y)\,. \tag{B.14}$$

Combining (B.12), (B.13) and (B.14), results in,

$$\frac{1}{\psi^2}\frac{\partial^2 D(n-1,\beta,Y)}{\partial^2 Y} = D(n-1,\beta,Y) - 2D(n-2,\beta-1,Y) + D(n-3,\beta-2,Y)\,, \tag{B.15}$$

$$-\frac{1}{\psi}\frac{\partial D(n-1,\beta,Y)}{\partial Y} = D(n-1,\beta,Y) - D(n-2,\beta-1,Y)\,. \tag{B.16}$$

Assuming,

$$\frac{D(n-2,\beta-1,Y)}{D(n-1,\beta,Y)} = A\,, \tag{B.17}$$

for large values of β and n, we can write,

$$\frac{D(n-3,\beta-2,Y)}{D(n-2,\beta-1,Y)} = A\,. \tag{B.18}$$

Combining (B.17) and (B.18), results in,

$$\frac{D(n-3,\beta-2,Y)}{D(n-1,\beta,Y)} = A^2\,. \tag{B.19}$$

Using (B.17) and (B.19) in (B.15), we obtain,

$$\frac{1}{D(n-1,\beta,Y)}\frac{\partial^2 D(n-1,\beta,Y)}{\partial^2 Y} = \psi^2(1-A)^2\,. \tag{B.20}$$

Using (B.17) in (B.16), we obtain,

$$\frac{1}{D(n-1,\beta,Y)}\frac{\partial D(n-1,\beta,Y)}{\partial Y} = -\psi(1-A)\,. \tag{B.21}$$

Combining (B.10), (B.11) (B.20) and (B.21) results in (B.9). As a result $P(Y)$ is of exponential form and considering that its integral over the range $0 \leq Y \leq 1$ should be equal to one, we obtain,

$$P(Y) = \frac{\lambda}{1-e^{-\lambda}}e^{-\lambda Y}\;;\;\; 0 \leq Y \leq 1\,. \tag{B.22}$$

From this argument, we can also say that the value of A in (B.17) does not depend on Y. To calculate the value of A, considering the definition of $D(n,\beta,Y)$ in (B.5), we can write,

$$D(n-2,\beta-1,0) = D(n-2,\beta,1)\,. \tag{B.23}$$

As A does not depend on Y, we can set $Y=0$ in (B.17) and obtain,

$$A = \frac{D(n-2,\beta-1,0)}{D(n-1,\beta,0)}\,. \tag{B.24}$$

Combining (B.23) and (B.24), we obtain,

$$A = \frac{D(n-2,\beta,1)}{D(n-1,\beta,0)}\,. \tag{B.25}$$

Finally, combining (B.6) for large values of n with (B.25), results in,

$$A = P(1) = \frac{\lambda}{1-e^{-\lambda}}e^{-\lambda}\,. \tag{B.26}$$

To calculate the constant λ, substituting,

$$\frac{1}{D(n-1,\beta,Y)}\frac{\partial D(n-1,\beta,Y)}{\partial Y} = \frac{1}{P(Y)}\frac{\partial P(Y)}{\partial Y} = -\lambda\,, \tag{B.27}$$

in (B.21) and using (B.26) for the value of A, we obtain,

$$\frac{\lambda}{\psi} + \frac{\lambda}{1-e^{-\lambda}}e^{-\lambda} = 1\,. \tag{B.28}$$

Equation (B.28) can be solved to obtain the constant λ as a function of $\psi = n/\beta$.

Finally, applying the change of variable (4.3) to (B.22), probability distribution along each 2-D subspace is calculated as,

$$P_2(X_0, X_1) = \begin{cases} C(\lambda) \exp\{-\lambda(X_0^2 + X_1^2)/R_2^2\} & (X_0^2 + X_1^2) \leq R_2^2 \\ 0 & (X_0^2 + X_1^2) > R_2^2 \end{cases},$$

where,

$$C(\lambda) = \frac{\lambda}{\pi R_2^2 (1 - e^{-\lambda})}. \tag{B.29}$$

In the extreme case of $\psi \to \infty$, which corresponds to a hypersphere as the boundary of the constellation, from (B.28), we obtain $\lambda \simeq \psi$. This results in a Gaussian distribution with variance $R_2^2/2\lambda$. In the other extreme case of $\psi \to 1$, which corresponds to no shaping, from (B.28), we obtain $\lambda \simeq 0$. This results in a uniform distribution on the 2-D subspaces as expected.

To show that for $N \to \infty$, the probability distributions along the subspaces become independent of each other, we proceed as follows: With an approach similar to that used in deriving (B.6), it is easy to verify that for large values of n, the joint probability distribution along j dimensions of the n-D region $\mathcal{TC}_n$ is equal to,

$$P_j(Y_0, \ldots, Y_{j-1}) = \frac{D(n-j, \beta, Y_0 + \ldots + Y_{j-1})}{D(n, \beta, 0)}. \tag{B.30}$$

Substituting the value of A from (B.26) in (B.17) and using the result together with (B.27) in (B.14), we obtain,

$$\frac{D(n-2, \beta, Y_0 + \ldots + Y_{j-1})}{D(n-1, \beta, Y_0 + \ldots + Y_{j-1})} = \frac{\lambda}{1 - e^{-\lambda}}. \tag{B.31}$$

For large values of n, we have,

$$\frac{D(n-2-i, \beta, Y_0 + \ldots + Y_{j-1})}{D(n-1-i, \beta, Y_0 + \ldots + Y_{j-1})} = \frac{D(n-2, \beta, Y_0 + \ldots + Y_{j-1})}{D(n-1, \beta, Y_0 + \ldots + Y_{j-1})}, \quad \forall i \in \mathbf{N}. \tag{B.32}$$

Using (B.32) recursively with the initial value of (B.31), we obtain,

$$\frac{D(n-j, \beta, Y_0 + \ldots + Y_{j-1})}{D(n-1, \beta, Y_0 + \ldots + Y_{j-1})} = \left(\frac{\lambda}{1 - e^{-\lambda}}\right)^{j-1}. \tag{B.33}$$

Using (B.6), we obtain,

$$P(Y_0 + \ldots + Y_{j-1}) = \frac{D(n-1, \beta, Y_0 + \ldots + Y_{j-1})}{D(n, \beta, 0)} = \frac{\lambda}{1 - e^{-\lambda}} e^{-\lambda(Y_0 + \ldots + Y_{j-1})}. \quad \text{(B.34)}$$

Multiplying (B.33) and (B.34) and using the result in (B.30), we obtain,

$$P_j(Y_0, \ldots, Y_{j-1}) = \frac{D(n-j, \beta, Y_0 + \ldots + Y_{j-1})}{D(n, \beta, 0)} = \frac{\lambda^j}{(1 - e^{-\lambda})^j} e^{-\lambda(Y_0 + \ldots + Y_{j-1})}, \quad \text{(B.35)}$$

which means that the random variables $Y_0, \ldots, Y_{j-1}$ are independent.

Appendix C

A generalization of the shaping regions $\mathcal{A}_N$ and $\mathcal{A}_N^{N'}$

Assume that the CER$_s$ and the PAR are measured on an l-D basis, l being an even integer greater than two. In the case, the optimum shaping is equal to,

$$\mathcal{A}_N^{(l)} = \{\mathcal{S}_l(R_l)\}^n \cap \mathcal{S}_N(R_N)\,. \tag{C.1}$$

The main difference is that here the first level of shaping is achieved by employing a sphere as the boundary of the l-D subspaces. With a similar approach as in Appendix A, the integral of the function $F(X_0^2 + \ldots + X_{N-1}^2)$ over the region $\mathcal{A}_N^{(l)}$ is calculated as,

$$I = (\pi R_l^2)^{\frac{N}{2}} \sum_{k=0}^{\lfloor\beta\rfloor} \sum_{i_0=0}^{\frac{l}{2}-1} \cdots \sum_{i_{k-1}=0}^{\frac{l}{2}-1} (-1)^{i_0+\ldots+i_{k-1}} C_n^k \frac{(\beta-k)^{\frac{N}{2}-i_0-\ldots-i_{k-1}}}{i_0!\ \ldots i_{k-1}!\Gamma(\frac{N}{2}-i_0-\ldots-i_{k-1})} \tag{C.2}$$

$$\times \int_0^1 F\left\{R_l^2\left[(\beta-k)\tau+k\right]\right\} \tau^{\frac{N}{2}-i_0-\ldots-i_{k-1}-1}\ d\tau\,.$$

Similarly, the two level shell mapped region can be generalized to the region,

$$\mathcal{A}_N^{(l,n)} = \{\mathcal{S}_l(R_l)\}^{n'n''} \cap \{\mathcal{S}_n(R_n)\}^{n''} \cap \mathcal{S}_N(R_N)\,, \tag{C.3}$$

where $n' = n/l$ and $n'' = N/n$. In this case, using the change of variable,

$$Y_{n'q+p} = \frac{\sum_{m=0}^{l-1} X_{l(n'q+p)+m}^2}{R_l^2},\quad p = 0,1,\ldots,n'-1\,,\quad q = 0,1,\ldots,n''-1\,, \tag{C.4}$$

and defining β' and β'' by,

$$R_n^2 = \beta' R_l^2, \quad R_N^2 = \beta'\beta'' R_l^2 , \tag{C.5}$$

the region $\mathcal{A}_N^{(l,n)}$ can be written in the following $n'n'' = N/l$ form,

$$\begin{aligned} \mathcal{A}_{n'\times n''} &= \{Y_{n'q+p},\ p = 0,1,\ldots,n'-1,\ q = 0,1,\ldots,n''-1\} \\ &: \ 0 \le Y_{n'q+p} \le 1 \\ &\quad 0 \le \sum_{p=0}^{n'-1} Y_{n'q+p} \le \beta' \\ &\quad 0 \le \sum_{q=0}^{n''-1}\sum_{p=0}^{n'-1} Y_{n'q+p} \le \beta'\beta'' . \end{aligned} \tag{C.6}$$

Using the change of variable,

$$\sum_{p=0}^{n'-1} Y_{n'q+p} = Z_q , \tag{C.7}$$

the region in (C.6) is mapped to,

$$\begin{aligned} \mathcal{TC}_{n''}(\beta', \beta'\beta'') &= \{Z_q,\ q = 0,1,\ldots,n''-1\} \\ &: \ 0 \le Z_q \le \beta' \\ &\quad 0 \le \sum_{q=0}^{n''-1} Z_q \le \beta'\beta'' , \end{aligned} \tag{C.8}$$

In the following, we calculate the integral of the function $F(X_0^2 + \ldots + X_{N-1}^2)$ over the region $\mathcal{A}_N^{(l,n)}$. To express the integration over the region $\mathcal{TC}_{n''}$, we need a relation between the incremental volume of the region $\mathcal{A}_N^{(l,n)}$ and the incremental volume of the region $\mathcal{TC}_{n''}$. To obtain such a relations, we use the fact that a point Z_q on the dimension q of $\mathcal{TC}_{n''}$ corresponds to the region $\mathcal{A}_n^{(l)}$ with $\beta = Z_q$. The volume of this region can be calculated as,

$$V = (\pi R_l^2)^{\frac{n}{2}} \sum_{K_q=0}^{\lfloor Z_q \rfloor} \sum_{I_0^q=0}^{\frac{l}{2}-1} \cdots \sum_{I_{K_q-1}^q=0}^{\frac{l}{2}-1} (-1)^{K_q} C_{n'}^{K_q} \frac{(Z_q - K_q)^{\frac{n}{2} - I_0^q - \ldots - I_{K_q-1}^q}}{I_0^q! \ \ldots I_{K_q-1}^q! \Gamma(\frac{n}{2} - I_0^q - \ \ldots - I_{K_q-1}^q + 1)} . \tag{C.9}$$

The dimension q of $\mathcal{TC}_{n''}$ corresponds to dimensions qn, $\ldots$,$qn+n-1$ of $\mathcal{A}_N^{(l,n)}$. Using this fact and differentiating (C.9) with respect to Z_q, we obtain,

$$\mathrm{d}X_{qn} \ \ldots \ \mathrm{d}X_{qn+n-1} = (\pi R_l^2)^{\frac{n}{2}} \sum_{K_q=0}^{\lfloor Z_q \rfloor} \sum_{I_0^q=0}^{\frac{l}{2}-1} \cdots \sum_{I_{K_q-1}^q=0}^{\frac{l}{2}-1}$$

$$(-1)^{K_q} C_{n'}^{K_q} \frac{(Z_q - K_q)^{\frac{n}{2} - I_0^q - \ \ldots - I_{K_q-1}^q - 1}}{I_0^q! \ \ldots I_{K_q-1}^q! \Gamma(\frac{n}{2} - I_0^q - \ \ldots - I_{K_q-1}^q)} \, \mathrm{d}Z_q \,. \tag{C.10}$$

Multiplying (C.10) for different values of q we obtain,

$$\begin{aligned} \mathrm{d}X_0 \ \ldots \ \mathrm{d}X_{N-1} = (\pi R_l^2)^{\frac{N}{2}} \sum_{K_0=0}^{\lfloor Z_0 \rfloor} \sum_{I_0^0=0}^{\frac{l}{2}-1} \ \ldots \sum_{I_{K_0-1}^0=0}^{\frac{l}{2}-1} \ \ldots \sum_{K_{n''-1}=0}^{\lfloor Z_{n''-1} \rfloor} \sum_{I_0^{n''-1}=0}^{\frac{l}{2}-1} \ \ldots \sum_{I_{K_{n''-1}-1}^{n''-1}=0}^{\frac{l}{2}-1} \\ (-1)^{K_0 + \ \ldots + K_{n''-1}} C_{n'}^{K_0} \ \ldots C_{n'}^{K_{n''-1}} \frac{(Z_0 - K_0)^{\frac{n}{2} - I_0^0 - \ \ldots - I_{K_0-1}^0 - 1}}{I_0^0! \ \ldots I_{K_0-1}^0! \Gamma(\frac{n}{2} - I_0^0 - \ \ldots - I_{K_0-1}^0)} \\ \times \frac{(Z_{n''-1} - K_{n''-1})^{\frac{n}{2} - I_0^{n''-1} - \ \ldots - I_{K_{n''-1}-1}^{n''-1} - 1}}{I_0^{n''-1}! \ \ldots I_{K_{n''-1}-1}^{n''-1}! \Gamma(\frac{n}{2} - I_0^{n''-1} - \ \ldots - I_{K_{n''-1}-1}^{n''-1})} \, \mathrm{d}Z_0 \ \ldots \mathrm{d}Z_{n''-1} \,. \end{aligned} \tag{C.11}$$

It is seen that Z_q, $q = 0, \ldots, n''-1$ appears in the limits of summations in (C.11). This makes the integration difficult. To avoid this problem, we subdivide each dimension of $\mathcal{TC}_{n''}$ with the points of integer value. These points subdivide each dimension into $\lceil \beta' \rceil$ regions. The first $\lfloor \beta' \rfloor$ regions are of length one and the last one is of length $a = \beta' - \lfloor \beta' \rfloor$. This subdivides the region $\mathcal{TC}_{n''}$ into $\lceil \beta' \rceil^{n''}$ disjoint subregions. A total of $\lfloor \beta' \rfloor^{n''}$ of these regions are hypercubes of edge length one located at points $(\lfloor Z_0 \rfloor, \ \ldots, \lfloor Z_{n''-1} \rfloor) \in \{0, 1, \ \ldots, \lfloor \beta' \rfloor - 1\}^{n''}$, where $\{0, 1, \ \ldots, \lfloor \beta' \rfloor - 1\}^{n''}$ is equal to the n''-fold cartesian product of the set $\{0, 1, \ \ldots, \lfloor \beta' \rfloor - 1\}$ with itself. The remaining $\lceil \beta' \rceil^{n''} - \lfloor \beta' \rfloor^{n''}$ regions are parallelepipeds. At least one edge length of each parallelepiped is equal to, $a = \beta' - \lfloor \beta' \rfloor$. In each of these regions, the value of $\lfloor Z_q \rfloor$, appearing in the limit of summation in (C.11), is constant.

To calculate the desired integral over a subregion which is a hypercube located at point $(\lfloor Z_0 \rfloor, \ldots, \lfloor Z_{n''-1} \rfloor)$, we first shift the origin to that point. The region of integration is an n''-D simplex with edge length,

$$L = \begin{cases} \beta'\beta'' - \lfloor Z_0 \rfloor - \ \ldots - \lfloor Z_{n''-1} \rfloor & \text{if} \quad 0 \le \beta'\beta'' - \lfloor Z_0 \rfloor - \ \ldots - \lfloor Z_{n''-1} \rfloor \le n'' \\ n'' & \text{if} \quad \beta'\beta'' - \lfloor Z_0 \rfloor - \ \ldots - \lfloor Z_{n''-1} \rfloor > n'' \\ 0 & \text{if} \quad \beta'\beta'' - \lfloor Z_0 \rfloor - \ \ldots - \lfloor Z_{n''-1} \rfloor < 0 \,, \end{cases} \tag{C.12}$$

truncated within a hypercube of edge length one ($\mathcal{TC}_{n''}(1, L)$). With a procedure similar

to the Appendix A, this integral is calculated as,

$$I_1(\lfloor Z_0\rfloor,\ldots,\lfloor Z_{n''-1}\rfloor) = (\pi R_l^2)^{\frac{N}{2}} \sum_{K_0=0}^{\lfloor Z_0\rfloor} \sum_{I_0^0=0}^{\frac{l}{2}-1} \cdots \sum_{I_{K_0-1}^0=0}^{\frac{l}{2}-1} \cdots \sum_{K_{n''-1}=0}^{\lfloor Z_{n''-1}\rfloor} \sum_{I_0^{n''-1}=0}^{\frac{l}{2}-1} \cdots \sum_{I_{K_{n''-1}-1}^{n''-1}=0}^{\frac{l}{2}-1}$$

$$\sum_{k=0}^{\lfloor L\rfloor} \sum_{i_0=0}^{J_0} \cdots \sum_{i_{n''-1}=0}^{J_{n''-1}} (-1)^{k+K_0+\ \ldots+K_{n''-1}} C_{n''}^{k} C_{n'}^{K_0} \ \ldots C_{n'}^{K_{n''-1}}$$

$$\frac{(\lfloor Z_0\rfloor - K_0+1)^{i_0} \ \ldots(\lfloor Z_{k-1}\rfloor - K_{k-1}+1)^{i_{k-1}}(\lfloor Z_k\rfloor - K_k)^{i_k} \ \ldots(\lfloor Z_{n''-1}\rfloor - K_{n''-1})^{i_{n''-1}}}{i_0!\ \ldots i_{n''-1}!}$$

$$\frac{(\beta'\beta''-A)^{M+1}}{\Gamma(M+1)} \int_0^1 F\left\{R_l^2\left[(\beta'\beta''-A)\tau+A\right]\right\} \tau^M d\tau\,, \tag{C.13}$$

where,

$$\begin{aligned}
&J_q = n/2-I_0^q- \ \ldots-I_{K_q-1}^q-1,\\
&M = N/2-i_0- \ \ldots-i_{n''-1}-I_0^0- \ \ldots-I_{K_0-1}^0- \ \ldots.-I_0^{n''-1}- \ \ldots-I_{K_{n''-1}-1}^{n''-1}-1,\\
&A = \lfloor Z_0\rfloor + \ \ldots+\lfloor Z_{n''-1}\rfloor + k\,.
\end{aligned}$$

It is seen that the result is independent of the permutation applied to $(\lfloor Z_0\rfloor,\ldots,\lfloor Z_{n''-1}\rfloor)$. This fact can be used to further simplify (C.13).

For a region which is a parallelepiped located at point $(\lfloor Z_0\rfloor,\ldots,\lfloor Z_{r-1}\rfloor,\lfloor\beta'\rfloor^{n''-r})$, again, the origin is shifted to that point. The region of integration in the shifted coordinates is written as,

$$\sum_{k_0=0}^{1} \cdots \sum_{k_{n''-1}=0}^{1} (-1)^{k_0+\ \ldots+k_{n''-1}} \mathcal{Q}_n(k_0,\ \ldots,k_{r-1},a\times k_r,\ \ldots,a\times k_{n''-1}) \cap \mathcal{S}_n((0)^{n''},L') =$$

$$\sum \mathcal{S}_n(k_0,\ldots,k_{r-1},a\times k_r,\ldots,a\times k_{n''-1};L')\,, \tag{C.14}$$

where,

$$L' = \begin{cases}
\beta'\beta''-\lfloor Z_0\rfloor- \ \ldots-\lfloor Z_{r-1}\rfloor-(n''-r)\times\lfloor\beta'\rfloor,\\
\text{if} \quad 0\le\beta'\beta''-\lfloor Z_0\rfloor- \ \ldots-\lfloor Z_{r-1}\rfloor-(n''-r)\times\lfloor\beta'\rfloor\le n'',\\
n'' \qquad \text{if} \quad \beta'\beta''-\lfloor Z_0\rfloor- \ \ldots-\lfloor Z_{r-1}\rfloor-(n''-r)\times\lfloor\beta'\rfloor\ge n'',\\
0 \qquad \text{if} \quad \beta'\beta''-\lfloor Z_0\rfloor- \ \ldots-\lfloor Z_{r-1}\rfloor-(n''-r)\times\lfloor\beta'\rfloor\le 0\,,
\end{cases} \tag{C.15}$$

and the summation on the right hand side of (C.14) is calculated over $(k_0, \ldots, k_{n''-1})$ such that,

$$\beta'\beta''-\lfloor Z_0\rfloor-\ldots-\lfloor Z_{r-1}\rfloor-(n''-r)\times\lfloor\beta'\rfloor-k_0-\ldots-k_{r-1}-a\times(k_r+\ldots+k_{n''-1}),$$

is positive, the regions $\mathcal{S}_n$ and $\mathcal{Q}_n$ are defined in Appendix A.

The integral over $\mathcal{S}_n(k_0, \ldots, k_{r-1}, a\times k_r, \ldots, a\times k_{n''-1}; L')$ can be calculated with a procedure quite similar to the Appendix A. The result is,

$$I_2(r;\lfloor Z_0\rfloor,\ldots,\lfloor Z_{r-1}\rfloor)=(\pi R_l^2)^{\frac{N}{2}}\sum_{K_0=0}^{\lfloor Z_0\rfloor}\sum_{I_0^0=0}^{\frac{l}{2}-1}\cdots\sum_{I_{K_0-1}^0=0}^{\frac{l}{2}-1}\cdots\sum_{K_{r-1}=0}^{\lfloor Z_{r-1}\rfloor}\sum_{I_0^{r-1}=0}^{\frac{l}{2}-1}\cdots\sum_{I_{K_{r-1}-1}^{r-1}=0}^{\frac{l}{2}-1}$$

$$\sum_{K_r=0}^{\lfloor\beta'\rfloor}\sum_{I_0^r=0}^{\frac{l}{2}-1}\cdots\sum_{I_{K_r-1}^r=0}^{\frac{l}{2}-1}\cdots\sum_{K_{n''-1}=0}^{\lfloor\beta'\rfloor}\sum_{I_0^{n''-1}=0}^{\frac{l}{2}-1}\cdots\sum_{I_{K_{n''-1}-1}^{n''-1}=0}^{\frac{l}{2}-1}\sum_{k_0=0}^{1}\cdots\sum_{k_{n''-1}=0}^{1}\sum_{i_0=0}^{J_0}\cdots\sum_{i_{n''-1}=0}^{J_{n''-1}}$$

$$(-1)^{k_0+\ldots+k_{n''-1}+K_0+\ldots+K_{n''-1}}C_{n''}^{r}C_{n'}^{K_0}\ldots C_{n'}^{K_{n''-1}}$$

$$\frac{(\lfloor Z_0\rfloor-K_0+k_0)^{i_0}\ldots(\lfloor Z_{r-1}\rfloor-K_{r-1}+k_{r-1})^{i_{r-1}}}{I_0^0!\ldots I_{K_0-1}^0!\ldots I_0^{r-1}!\ldots I_{K_{r-1}-1}^{r-1}!i_0!\ldots i_{r-1}!}$$

$$\frac{(\lfloor Z_r\rfloor-K_r+a\times k_r)^{i_r}\ldots(\lfloor Z_{n''-1}\rfloor-K_{n''-1}+a\times k_{n''-1})^{i_{n''-1}}}{I_0^r!\ldots I_{K_r-1}^r!\ldots I_0^{n''-1}!\ldots I_{K_{n''-1}-1}^{n''-1}!i_r!\ldots i_{n''-1}!}$$

$$\frac{\{\max[0,(\beta'\beta''-B)]\}^{M+1}}{\Gamma(M+1)}\int_0^1 F\left\{R_l^2\left[(\beta'\beta''-B)\tau+B\right]\right\}\tau^M d\tau\,, \qquad \text{(C.16)}$$

where,

$$B=\lfloor Z_0\rfloor+\ldots+\lfloor Z_{r-1}\rfloor+(n''-r)\times\lfloor\beta'\rfloor+k_0+\ldots+k_{r-1}+a\times(k_r+\ldots+k_{n''-1}).$$

Adding up (C.13) and (C.16) over different regions, the final result is found as,

$$\int_{\mathcal{A}_N^{(l,n)}}F(X_0^2+\ldots+X_{N-1}^2)dX_0\ldots dX_{N-1}=(\pi R_l^2)^{\frac{N}{2}}\times$$

$$\left\{\sum_{\lfloor Z_0\rfloor=0}^{\lfloor\beta'\rfloor-1}\cdots\sum_{\lfloor Z_{n''-1}\rfloor=0}^{\lfloor\beta'\rfloor-1}I_1(\lfloor Z_0\rfloor,\ldots,\lfloor Z_{n''-1}\rfloor)+\sum_{r=0}^{n''-1}\sum_{\lfloor Z_0\rfloor=0}^{\lfloor\beta'\rfloor-1}\cdots\sum_{\lfloor Z_{r-1}\rfloor=0}^{\lfloor\beta'\rfloor-1}C_{n''}^{r}I_2(r;\lfloor Z_0\rfloor,\ldots,\lfloor Z_{r-1}\rfloor)\right\}. \qquad \text{(C.17)}$$

In deriving (C.17), we have used the fact that there are $C_{n''}^r$ subregions with the integral $I_2(r;\lfloor Z_0\rfloor,\ldots,\lfloor Z_{r-1}\rfloor)$ given in (C.16). We also notice that the argument of the

first summation in (C.17) is independent of the permutation applied to $(\lfloor Z_0 \rfloor, \ldots, \lfloor Z_{n''-1} \rfloor)$ and the argument of the second summation is independent of the permutation applied to $(\lfloor Z_0 \rfloor, \ldots, \lfloor Z_{r-1} \rfloor)$. This facts can be used to simplify the calculations.

Appendix D

Calculation of the absolute first moment of a lattice Voronoi region

We are going to calculate the integral of,

$$F_m[\mathcal{V}(\Lambda_n^s)] = \int_{\mathcal{V}(\Lambda_n^s)} |Y_0| + \ldots + |Y_{n-1}| \; dY_0 \ldots dY_{n-1} \,, \tag{D.1}$$

where $\mathcal{V}(\Lambda_n^s)$ denotes the Voronoi region of the lattice Λ_n^s. The calculation is based on decomposing the Voronoi region into the congruent simplexes and applying the generalized midpoint method of integration, [18], to each simplex. According to this method the integral of any linear function over a simplex is equal to the volume of the simplex times the value of that function at its centroid. For the methods of calculating the centroid and the volume of the simplexes refer to [4]. The decomposition into simplexes is based on expressing the Voronoi region of the lattice by the Coxeter diagram. In the following, we first give a brief description of the Coxeter diagram, [4].

The Voronoi region of a root lattice Λ_n is the union of the reflections of a spherical simplex into its walls. This simplex is known as the fundamental simplex of the lattice. The resulting reflection group is shown by $W(\Lambda_n)$. The Coxeter diagram of a lattice represents its fundamental simplex in two different ways. In the first way, the nodes of the diagram represent the reflecting hyperplanes which are the walls of the fundamental simplex. The equation of each wall is written besides its corresponding node. The angle

between two walls is indicated by the branch of the diagram. If the hyperplanes are at an angle of $\pi/3$, the nodes are joined by a single branch. If the angle is $\pi/4$, they are joined by a double branch. If the angle is π/P, $P > 4$, they are joined by a branch labeled P. Finally, if the hyperplanes are perpendicular the nodes are not joined. In the second way, the nodes in the diagram are taken to represent the vertices of the fundamental simplex. Each node represent the vertex opposite to the corresponding hyperplane. The components of each vertex (vertex vector) are written besides its corresponding node. The node corresponding to origin is shown as a black node. Normally, there is no reflection hyperplane at this node. Adding this hyperplane to the spherical simplex results in an ordinary simplex. The corresponding reflection group is an infinite group shown by $W_a(\Lambda_n)$. The images of the simplex under $W_a(\Lambda_n)$ are distinct and tile the space. The image of a point located at origin under $W_a(\Lambda_n)$ is the set of all points of Λ_n. For further information about the Coxeter diagram refer to [7].

Case I: $\Lambda_n^s = D_n$

The two interpretations of the Coxeter diagram for lattice D_n is shown Fig. D.1. By

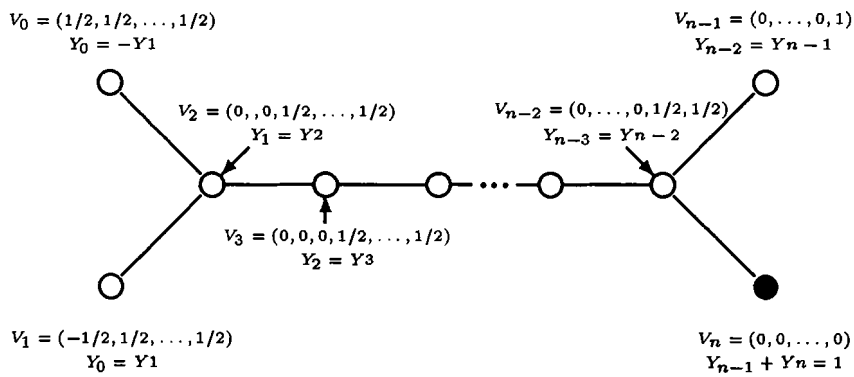


Fig. D.1: The two interpretations of the Coxeter diagram for the lattice D_n.

referring to the equations of the reflecting hyperplanes, it is seen that the reflection group is composed of n permutations and $n-1$ sign changes. This means that if we consider the vertex vector of the simplexes as the rows of an n-D matrix (denotes as the vertex

matrix), the effect of the reflection group can be considered as permutating the columns of this matrix (a subgroup of order $n!$) or changing the sign of $n-1$ columns (a subgroup of order 2^{n-1}). The reflection group is the product of these two subgroups. The order of such group is equal to, $|W(D_n)| = 2^{n-1} \times n!$. Neither permutation nor sign changings of the coordinates changes the first moment. As a result, all the images under the reflection group has the same value for the first moment.

This argument indicates that the first moment of $V_n(D)$ is equal to $2^{n-1} \times n!$ times the first moment of its fundamental simplex. The problem in applying the generalized midpoint method is that one vertex of the fundamental simplex has negative value for Y_0. This makes the function $|Y_0| + \ldots + |Y_{n-1}|$ nonlinear. To avoid this problem the fundamental simplex is divided into two regions. The first region corresponds to $Y_0 > 0$ and the second one to $Y_0 < 0$. In both of these regions, the function $|Y_0| + \ldots + |Y_{n-1}|$ is linear. The first region itself is a simplex. By writing the second region as the difference between the main simplex and the first region, the desired integrals is calculated. Using this procedure, we obtain,

$$F_m\left[V_n(D_n)\right] = \frac{1}{2} \times \frac{n^2+n+2}{(n+1)}. \tag{D.2}$$

Case II: $\Lambda_n^s = \Re D_n$

Figure D.2 shows the two interpretations of the Coxeter diagram for this lattice. The

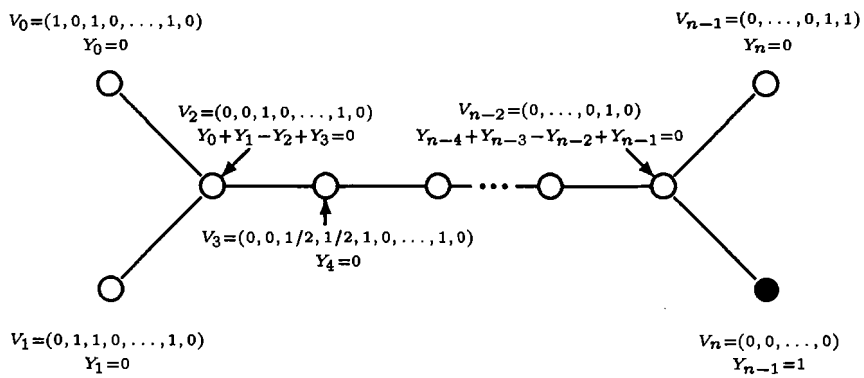


Fig. D.2: The two interpretations of the Coxeter diagram for the lattice $\Re D_n$.

diagram in Fig. D.2 is obtained by applying the rotation operator to the diagram of Fig. D.1. It is seen that the components of all the edges of the fundamental simplex are positive. To obtain the images of the fundamental simplex, we apply the operator $\Re$ to the corresponding images of D_n. As we already saw these images were obtained by permutating or $n-1$ sign changings of the columns of the vertex matrix. It is easy to verify that if we interchange the two columns $2q$ and $2q+1$, $q=0,\ldots,(n/2)-1$ with each other and apply $\Re$ to the result, it is equivalent to first applying $\Re$ and then changing the sign of the second column. This means that the second moment of the fundamental simplex and its image under this operation are equal. In a similar way, if we change the sign of some of the columns and then apply $\Re$, it is equivalent to first applying $\Re$ and then changing the sign of some columns or permutating some of the columns with each other. None of these operations changes the first moment. Consequently, the only operations which (possibly) change the first moment are those which change one of the columns $2q$, $2q+1$ with one of the columns $2q'$, $2q'+1$, $q,q'=0,\ldots,(n/2)-1$. The total number of these operations is equal to, $(2)^{-n/2}n!/(n/2)!$. We refer to this operations as the first moment changing operations.

It can be shown that if all the components of the vertex matrix, $\mathbf{V}$, of a simplex are positive, its first absolute moment using the midpoint method is equal to,

$$F_m = \frac{S(\mathbf{V}) \times d(\mathbf{V})}{(n+1) \times n!}, \tag{D.3}$$

where $S(\mathbf{V})$ denotes the sum of the elements of $\mathbf{V}$ and $d(\mathbf{V})$ denotes its determinant. To calculate $S(\mathbf{V})$ for different images of the fundamental simplex, we proceed as follows:

Using Fig. D.1, the columns of the vertex matrix of D_n are found as,

$$\begin{cases} C_0 = \left[0.5, -0.5, (0)^{n-2}\right] \\ C_p = \left[(0.5)^{p+1}, (0)^{n-p-1}\right] \quad 1 \le p \le n-2 \\ C_{n-1} = \left[(0.5)^{n-1}, (1)\right]. \end{cases} \tag{D.4}$$

In this case, if in the process of permutating the columns of the vertex matrix with each other, vectors C_p and $C_{p'}$, $p,p'=0,\ldots,n-1$ are assigned to the columns $2q$ and $2q+1$,

$q = 0, \ldots, (n/2) - 1$, the sum of the elements of these two columns after applying $\Re$ will be equal to,

$$\mathcal{S} = \begin{cases} \max(p, p') + 1 & \text{if} \quad \max(p, p') \neq n - 1 \\ \max(p, p') + 2 & \text{if} \quad \max(p, p') = n - 1 \,. \end{cases} \tag{D.5}$$

This facilitates the calculation of $S(\mathbf{V})$ necessary in (D.3). It can be shown that for the lattice $\Re D_n$, the determinant of the vertex matrix is equal to, $d(\mathbf{V}) = 4 \times (2)^{-n/2}$. If the average value of $S(\mathbf{V})$ over the set of the moment changing operations is equal to $E[S(\mathbf{V})]$, we obtain,

$$F_m[\mathcal{V}_n(\Re D_n)] = \frac{(2)^{1+\frac{n}{2}} \times E[S(\mathbf{V})]}{n+1} \,. \tag{D.6}$$

Appendix E

Proof of the convexity of the optimization region

In this appendix we prove that the region determined by the set of the constraints in (6.23) is convex. Obviously, the linear constraints result in a convex region. We consider the more general case of the set of the positive-semi-definite matrices, i.e., $\lambda_i(\mathbf{R}_x) \geq 0$, $\forall i$. To prove the convexity, we use the following theorem:

Let $\mathbf{A}$ and $\mathbf{B}$ be N-D symmetrical matrices and let the eigenvalues $\lambda_i(\mathbf{A})$, $\lambda_i(\mathbf{B})$ and $\lambda_i(\mathbf{A}+\mathbf{B})$, $i \in [0, N-1]$, be arranged in the increasing order. For each $k \in [0, N-1]$, we have,

$$\lambda_k(\mathbf{A}) + \lambda_0(\mathbf{B}) \leq \lambda_k(\mathbf{A}+\mathbf{B}) \tag{E.1}$$

Now, assume that $\mathbf{R}_x^1$ and $\mathbf{R}_x^2$ are two symmetrical positive-semi-definite matrices. Substituting $\mathbf{A} = \alpha\mathbf{R}_x^1$ and $\mathbf{B} = (1-\alpha)\mathbf{R}_x^2$ in (E.1) and considering that $\lambda_i(\alpha\mathbf{A}) = \alpha\lambda_i(\mathbf{A})$, results in,

$$\alpha\lambda_k(\mathbf{R}_x^1) + (1-\alpha)\lambda_0(\mathbf{R}_x^2) \leq \lambda_k[\alpha\mathbf{R}_x^1 + (1-\alpha)\mathbf{R}_x^2] \tag{E.2}$$

For $\alpha \in [0,1]$, the left hand side of (E.2) is nonnegative and consequently the right hand side is also nonnegative. This proves the desired result.

The final region is located at the intersection of two convex regions and is convex.

Appendix F

Block-based eigensystem of the $1 \pm D$ and $1 - D^2$ systems

This chapter have been reported in [35].

For the $1 - D$ system, the input eigenvectors are equal to,

$$m_k(n) = \sqrt{\frac{2}{N+1}} \sin \frac{\pi(k+1)(n+1)}{N+1}, \quad k, n = 0, \ldots, N-1, \tag{F.1}$$

and the corresponding eigenvalues are equal to,

$$\phi_k = 1 - \cos \frac{\pi(k+1)}{(N+1)}. \tag{F.2}$$

This can be verified by considering Eq. (F.1) as a periodic function with period $N+1$. This function is zero at $n = i(N+1) - 1$, $\forall i$. This means that the signal itself provides zero initial conditions for the N-D blocks. Consequently, the response of the system in each block is equal to its steady state response. Note that, in steady state, a sinusoid is the input eigenfunction of any linear system.

To give a formal proof, we consider $\mathbf{A}^t\mathbf{A}$ as the transform matrix of a linear time invariant system with the transfer function $H(D) = 0.5(1-D)(1-D^{-1})$. This is the transform of $c(n) * c(-n)$ where $c(n) = \{-1/\sqrt{2}, 1/\sqrt{2}\}$ is the impulse response of the $1-D$ system. To be consistent with the block-based processing, we apply a causal input and

truncate the output to positive time. In this case, if $m(n)$ is an eigenvector with the eigenvalue ϕ, we should have,

$$H(D)\,[M(D) - m(0)] + m(0)(1 - 0.5D) = \phi\, M(D)\,, \tag{F.3}$$

where $M(D)$ is the transform of $m(n)$. Calculating (F.3) at time zero, we obtain,

$$\phi = 1 - 0.5\frac{m(1)}{m(0)}\,. \tag{F.4}$$

Combining Eqs. (F.3) and (F.4), we obtain,

$$M(D) = \frac{m(0)}{1 - \dfrac{m(1)}{m(0)}D + D^2}\,. \tag{F.5}$$

Equations (F.5) and (F.4) are satisfied by the eigenvectors and eigenvalues given in Eqs. (F.1) and (F.2).

Let $\hat{\mathbf{m}}_i$ denote the output eigenvector corresponding to the i'th eigenvalue. Using (F.1) in $\mathbf{A}\mathbf{m}_i = \sqrt{\phi_i}\hat{\mathbf{m}}_i$, we obtain,

$$\hat{m}(n) = \sqrt{\frac{2}{N+1}}\,\cos\frac{\pi(k+1)(n+0.5)}{N+1}, \quad n = 0,\ldots,N, \quad k = 0,\ldots,N-1. \tag{F.6}$$

The input and output eigenvectors of the $1 + D$ system are obtained by multiplying (F.1) and (F.6) with $(-1)^n$. The eigenvalues are the same as the $1 - D$ system given in Eq. (F.2).

In general, the product of the nonzero eigenvalues is equal to,

$$\prod_{k=0}^{N-1} \phi_k = |\mathbf{A}^t\mathbf{A}|\,, \tag{F.7}$$

where $|\mathbf{A}^t\mathbf{A}|$ is the determinant of $\mathbf{A}^t\mathbf{A}$. This product is an important parameter of the systems based on the system $\mathbf{A}$. For example, in a transmission system using the optimum modulating basis, the volume of the Voronoi region at the system input is proportional to $(\prod \phi_k)^{-1/2}$ and the required energy is proportional to $(\prod \phi_k)^{-1/N}$. For the $1 \pm D$ systems, assuming $|\mathbf{A}^t\mathbf{A}| = 2^{-N}A_N$ and expanding the determinant, we obtain $A_N = A_{N-1} + 1$. Solving this recursive equation with the initial value $A_1 = 2$ results in $A_N = N + 1$. Consequently, for the $1 \pm D$ systems, we have,

$$\prod_{k=0}^{N-1} \phi_k = 2^{-N}(N+1)\,. \tag{F.8}$$

An N-D, N even, $1-D^2$ system can be considered as two time multiplexed $N/2$-D, $1-D$ systems. Consequently, the eigenvalues are in pair equal to,

$$\phi_k = 1 - \cos\frac{\pi(k+1)}{0.5N+1},\ k = 0,\ldots,0.5N-1. \tag{F.9}$$

The two eigenvectors corresponding to a pair of eigenvalues are of the general form $\alpha_1 m_k(2n) + \alpha_2 m_k(2n+1)$ where $\alpha_1^2 + \alpha_2^2 = 1$ and $m_k(n)$ is the eigenvector of the $1-D$ system given in (F.1). For the $1-D^2$ system, we have,

$$\prod_{k=0}^{N-1} \phi_k = 2^{-N}(0.5N+1)^2\,. \tag{F.10}$$

Appendix G

Voronoi constellations

A real n-D lattice Λ_n is a discrete set of n-D vectors in $\mathbf{R}^n$ which form a group under ordinary vector addition. A sublattice Λ_n^s of a lattice Λ_n is a subset of elements of Λ_n that is itself a lattice. A binary lattice of second depth m is an integer lattice (sublattice of Z^n) which has $2^m Z^n$ as a sublattice. Around each lattice point is its Voronoi region consisting of all points of the space which are closer to that point than to any other.

A sublattice Λ_n^s induces a partition of Λ_n into equivalence classes modulo Λ_n^s. The order of this partition is shown by $|\Lambda_n/\Lambda_n^s|$. The lattice Λ_n is the union of $|\Lambda_n/\Lambda_n^s|$ cosets of Λ_n^s. The set of the cosets form a group under addition modulo Λ_n^s. This is called the quotient group and is denoted by $[\Lambda_n/\Lambda_n^s]$. In general, any element of Λ_n can be written as the sum of an element of Λ_n^s plus an element of the quotient group, i.e.,

$$\Lambda_n = \Lambda_n^s + [\Lambda_n/\Lambda_n^s] \; . \tag{G.1}$$

A Voronoi constellation is the set of the coset leaders (minimum energy points) of these $|\Lambda_n/\Lambda_n^s|$ cosets. This means that a Voronoi constellation based on the lattice partition Λ_n/Λ_n^s is the set of points of Λ_n (or some translate $\Lambda_n + \mathbf{a}$ of Λ_n) that fall within the Voronoi region around the origin of the lattice Λ_n^s. The lattice Λ_n^s is denoted as the shaping lattice. If Λ_n and Λ_n^s are binary lattices, the order of the constellation will be an integral power of two. For binary lattices, the resulting 2-D subconstellation is a square constellation, [10].

G.1 Address decomposition for the Voronoi constellations

In the Voronoi constellation the set of the constellation points form a group under the vector addition modulo the shaping lattice. This is an important property of these constellations and is used to facilitate the addressing. The complexity of this addressing method is that of a linear mapping plus the decoding of Λ_n^s.

A partition chain $\Lambda_n^1/\Lambda_n^2/\ldots/\Lambda_n^q$ induces a multiterm coset decomposition chain with a term corresponding to each partition, i.e.,

$$\Lambda_n^1 = \Lambda_n^q + \left[\Lambda_n^1/\Lambda_n^q\right] = \Lambda_n^q + \left[\Lambda_n^1/\Lambda_n^2\right] + \left[\Lambda_n^2/\Lambda_n^3\right] + \cdots + \left[\Lambda_n^{q-1}/\Lambda_n^q\right] . \tag{G.2}$$

In the case that $\Lambda_n^1, \Lambda_n^2, \ldots, \Lambda_n^q$ are binary lattices with the second depth m, the order of the set $[\Lambda_n^i/\Lambda_n^j]$ will be a power of two, say 2^L, and each element of this set can be expressed as $\mathbf{aG} = \sum_l a_l g_l$ where $\mathbf{a} = (a_0, a_1, \ldots, a_{L-1})$ is a binary L-tuple and the generators g_l are taken from the coset representatives of $2^m Z^n$ in Z^n (n-tuples of integers modulo-2^m). This provides a way to label the cosets by a label which has $q-1$ parts obtained by concatenating different a's. The decoding is achieved by extracting the different parts of the label in steps. The label extraction is achieved by using the following property: If a point of the Voronoi constellation based on the partition Λ_n^1/Λ_n^q in Eq. G.2 is calculated modulo one of the intermediate lattices, say Λ_n^i, the result will be equal to a point of the Voronoi constellation based on Z^n/Λ_n^i and its label is the first $i-1$ parts of the original label. In the following, we describe two methods for labeling.

The first labeling method is based on the partition chain $Z^n/2^k\Lambda_n^s/2^{k+1}Z^n$. If $Z^n/\Lambda_n^s/2Z^n$ is a partition chain with $|Z^n/\Lambda_n^s| = 2^J$, the Voronoi constellation based on the partition $Z^n/2^k\Lambda_n^s$ will consist of 2^{kn+J} points and we have the partition chain $Z^n/\Lambda_n^s/2\Lambda_n^s/\ldots/2^k\Lambda_n^s$ which induces a chain decomposition of the form,

$$Z^n = 2^k\Lambda_n^s + [Z^n/\Lambda_n^s] + [\Lambda_n^s/2\Lambda_n^s] + \ldots + \left[2^{k-1}\Lambda_n^s/2^k\Lambda_n^s\right] . \tag{G.3}$$

Assume that the $(n \times n)$ matrix $\mathbf{H}$ is a generator for the set $[\Lambda_n^s/2\Lambda_n^s]$ and the $(J \times n)$ matrix $\mathbf{G}$ is a generator for the set $[Z^n/\Lambda_n^s]$. By considering the fact that the generators

of the set $[2^{i-1}\Lambda_n^s/2^i\Lambda_n^s]$ may be taken as 2^{i-1} times the generators of the set $[\Lambda_n^s/2\Lambda_n^s]$, we can write,

$$Z^n = 2^k\Lambda_n^s + \mathbf{aG} + \mathbf{bH}\,, \tag{G.4}$$

where $\mathbf{a}$ is a binary J-tuple and $\mathbf{b}$ is an n-tuple of integers modulo 2^k.

The second labeling method is based on the partition chain $Z^n/2^kZ^n/2^k\Lambda_n^s$ which induces the following decomposition,

$$Z^n = 2^k\Lambda_n^s + \left[Z^n/2^kZ^n\right] + \left[2^kZ^n/2^k\Lambda_n^s\right]\,. \tag{G.5}$$

By considering that $\left[Z^n/2^kZ^n\right] = \left[Z/2^kZ\right]^n$, where $\left[Z/2^kZ\right]$ is the set of integers modulo 2^k, and also $[2^kZ^n/2^k\Lambda_n^s] = 2^k[Z^n/\Lambda_n^s]$, we obtain the following matrix form for the decomposition,

$$Z^n = 2^k\Lambda_n^s + \mathbf{b}' + 2^k\mathbf{a}'\mathbf{G}\,. \tag{G.6}$$

Again, $\mathbf{a}'$ is a J-tuple modulo-2, $|Z^n/\Lambda_n^s| = 2^J$, $\mathbf{b}'$ is an n-tuple modulo 2^k and the $(J \times n)$ matrix $\mathbf{G}$ is a generator for $[Z^n/\Lambda_n^s]$. This form explicitly reflects the separability of the labeling as discussed by Forney in [10].

In both labeling methods, the obtained point is calculated modulo $2^k\Lambda_n^s$ to obtain the minimum energy point of the corresponding coset. Relations (G.4) or (G.6) provide a method for labeling the cosets of $2^k\Lambda_n^s$ in Z^n by a two part label $(\mathbf{a}, \mathbf{b})$ or $(\mathbf{a}', \mathbf{b}')$, respectively.

Next, we consider the complexity of using a lookup table for the addressing of a Voronoi constellation. We know that in the partition chain $Z^n/2^k\Lambda_n^s/2^{k+1}Z^n$, the Voronoi region of $2^k\Lambda_n^s$ is a subset of the Voronoi region of $2^{k+1}Z^n$ and as a result the Voronoi constellation based on the partition $Z^n/2^k\Lambda_n^s$ is a subset of the Voronoi constellation based on the partition $Z^n/2^{k+1}Z^n$. But, the Voronoi constellation based on the partition $Z^n/2^{k+1}Z^n$ is a cubic constellation with 2^{k+1} points along each dimension. This is the n-fold cartesian product of the Voronoi constellation based on the partition $Z/2^{k+1}Z$. In this case, labeling can be achieved independently along each dimension and have a trivial complexity. Now, all we need is a means to specify the desired 2^{kn+J} points of the

constellation based on the partition $Z^n/2^k\Lambda_n^s$ among the 2^{kn+n} points of the constellation based on the partition $Z^n/2^{k+1}Z^n$. This can be achieved by employing a lookup table with $kn+J$ input lines where J is given by $|Z^n/\Lambda_n^s| = 2^J$, and with words of length $kn+n$ bits.

G.2 Voronoi constellations based on the lattices D_n, $\Re D_n$ and D_n^*

The lattice D_n is defined as, [4],

$$D_n = \{(X_0, \ldots, X_{n-1}) \in Z^n;\ X_0 + \ldots + X_{n-1} \text{ even}\}\,. \tag{G.7}$$

We have $|Z^n/D_n| = 2$ and $|D_n/2Z^n| = 2^{n-1}$. The set of $2n(n-1)$ nearest neighbors in this lattice are located at points $[(\pm 1)^2, (0)^{n-2}]$ where a vector within [ ] sign denotes the set obtained by all the possible permutations of the components of that vector. The Voronoi cell in D_n is determined by the set of the nearest neighbors. The constituent 2-D sublattice of D_n is equal to $\Re Z^2$. The Voronoi constellation obtained from the partition $Z^n/2^kD_n$ consists of 2^{kn+1} points. Its constituent 2-D subconstellation is a Voronoi constellation based on the partition $Z^2/2^k\Re Z^2$.

The Voronoi constellations based on the partition $Z^n/2^k\Re D_n$ carry one more bit per two dimensions than the constellations based on the partition $Z^n/2^kD_n$ and its constituent 2-D subconstellation is the Voronoi constellation based on the partition $Z^2/2^{k+1}Z^2$.

The lattice D_n^* is defined as, [4],

$$D_n^* = \{(2Z)^n\} \cup \{(2Z)^n + (1)^n\}\,. \tag{G.8}$$

This lattice can be obtained by scaling the dual lattice of D_n by a factor of two. We have $|Z^n/D_n^*| = 2^{n-1}$ and $|D_n^*/2Z^n| = 2$. In this lattice the closest points to the origin in the first set consist of $2n$ points of the form $[(\pm 2), (0)^{n-1}]$ and the closest points in the second set consist of 2^n points of the form $[(\pm 1)^n]$. The Voronoi cell is the intersection of the

Voronoi cells determined by these two sets. The first of these is a hypercube centered at zero with the vertices $[(\pm 1)^n]$ and the second is a generalized octahedron with vertices $[\pm(n/2), (0)^{n-1}]$. The lattice D_n^* is a sublattice of D_n with $|D_n/D_n^*| = 2^{n-2}$.

The Voronoi constellation obtained from the partition $Z^n/2^k D_n^*$ consists of $2^{(k+1)n-1}$ points. Its constituent 2-D subconstellation is a Voronoi constellation based on the partition $Z^2/2^{k+1}Z^2$.

References

[1] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable signaling on the Gaussian channel," *IEEE Trans. Inform. Theory,* vol. IT-36, pp. 726–740, July 1990.

[2] A. R. Calderbank and J. E. Mazo, "Baseband line codes via spectral factorization", *IEEE J. Select. Areas Commun.,* vol. SAC–7, pp. 914–928, August 1989.

[3] G. L. Cariolaro and G. P. Tronca, "Spectra of block coded digital signals,". *IEEE Trans. on Commun.,* vol. COM–22, pp. 1555–1563, Oct 1974

[4] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups,* Springer-Verlag, 1988.

[5] J. H. Conway and N. J. A. Sloane, "A fast encoding method for lattice codes and quantizers," *IEEE Trans. Inform. Theory,* vol. IT-31, pp. 106–109, January 1985.

[6] J. H. Conway and N. J. A. Sloane, "A lower bound on the average error of vector quantizers," *IEEE Trans. Inform. Theory,* vol. IT-29, pp. 820–824, November 1983.

[7] H. S. M. Coxeter, *Regular Polytopes,* Macmillan Company, New York, 1962.

[8] M. V. Eyuboglu and G. D. Forney, Jr. "Trellis precoding: combined coding, precoding and shaping for intersymbol interference channels," *IEEE Trans. Inform. Theory,* vol. IT-38, pp. 301–314, March 1992.

[9] G. D. Forney, Jr. and L. F. Wei, "Multidimensional constellations–Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Select. Areas Commun.,* vol. SAC–7, pp. 877–892, August 1989.

[10] G. D. Forney, Jr., "Multidimensional constellations—Part II: Voronoi constellations," *IEEE J. Select. Areas Commun.,* vol. SAC-7, pp. 941–958, August 1989.

[11] G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff , and S. U. Quereshi, "Efficient modulation for bandlimited channels," *IEEE J. Select. Areas Commun.,* vol. SAC–2, pp.632–647, 1984.

[12] G. D. Forney, "Coset codes—Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory,* vol. IT-34, pp. 1123–1151, September 1988.

[13] G. D. Forney, "Coset codes–Part II: Binary lattices and related codes," *IEEE Trans. Inform. Theory,* vol. IT–34, pp. 1152–1187, September 1988.

[14] G. D. Forney, "Trellis shaping," *IEEE Trans. Inform. Theory,* vol. IT–38, pp. 281–300, March 1992.

[15] G. D. Forney, Jr., and A. R. Calderbank "Coset codes for partial response channels; or, coset codes with spectral nulls," *IEEE Trans. Inform. Theory,* vol. IT-35, pp. 925–943, September 1989.

[16] G. J. Foschini, R. D. Gitlin and S. B. Weinstein, "Optimization of two-dimensional signal constellations in the presence of Gaussian noise," *IEEE Trans. Commun.,* vol. COM-22, pp. 28–38, 1974.

[17] R. G. Gallager, *Information theory and reliable communication,* John Wiley & Sons, NY, 1968.

[18] I. J. Good and R. A. Gaskins, *The centroid method of numerical integration,* Num. Math. vol. 16, pp 343–359, 1971.

[19] M. L. Honig, K. Steiglitz and S. Norman, "Optimization of signal sets for partial-response channels—Part I: Numerical techniques," *IEEE Trans. Inform. Theory,* vol. IT-37, pp. 1327–1341, September 1991.

[20] M. L. Honig, K. Steiglitz, V. Balakrishnan and E. Rantapaa, l_∞/l_∞ signal design," *IEEE Int. Symp. Inform. Theory,* p. 71, June 1991, Budapest, Hungary.

[21] R. A. Horn and C. R. Johnson, *Matrix analysis,* Cambridge University Press, 1985.

[22] J. Justesen, "Information rate and power spectra of digital codes," *IEEE Trans. Inform. Theory,* vol. IT-28, pp. 457–472, May 1982.

[23] P. Kabal and S. Pasupathy, "Partial-Response signaling," *IEEE Trans. Commun.,* vol. 23, pp. 921–934, September 1975.

[24] S. Kasturia, J. T. Aslanis, and J. M. Cioffi, "Vector Coding for partial response channels," *IEEE Trans. Inform. Theory,* vol. IT-36, pp. 741–762, July 1989.

[25] A. K. Khandani and P. Kabal, "Shaping of multi–dimensional signal constellations using a lookup table," to be presented at the *IEEE International Conference on Communications 1992 (ICC'92)*, Chicago, IL, 14–18 June 1992.

[26] A. K. Khandani and P. Kabal, "Shaping multidimensional signal constellations," *IEEE Int. Symp. Inform. Theory,* p. 4, June 1991, Budapest, Hungary.

[27] A. K. Khandani and P. Kabal, "Shaping multi-dimensional signal spaces—Part I: optimum shaping, shell mapping," submitted to *IEEE Trans. Inform. Theory.*

[28] A. K. Khandani and P. Kabal, "Using a prefix code for the 2-dimensional addressing of the Voronoi constellations based on the lattices D_N and D_N^*," to be presented at the *IEEE International Conference on Communications 1992 (ICC'92)*, Chicago, IL, 14–18 June 1992.

[29] A. K. Khandani and P. Kabal, "Shaping multi-dimensional signal spaces—Part II: shell-addressed constellations," submitted to *IEEE Trans. Inform. Theory.*

[30] A. K. Khandani and P. Kabal, "Shaping of multi-dimensional signal constellations using a lookup table," submitted to *IEEE Trans. Inform. Theory.*

[31] A. K. Khandani and P. Kabal, "Two-dimensional statistics of the Voronoi constellations based on the shaping lattices D_N and D_N^*," to appear in the *Journal of Communications, China Institute of Communications.*

[32] A. K. Khandani and P. Kabal, "Spectral shaping with unequal power distribution," to be presented at *Twenty-sixth annual conference on information sciences and systems,* Princeton, NJ, 19–20 March 1992.

[33] A. K. Khandani and P. Kabal, "Unsymmetrical boundary shaping in multi-dimensional spaces," *Proc. Twenty-ninth Annual Allerton Conference on Communications, Control, and Computing,* Allerton, IL, pp. 798–808, Oct 1991.

[34] A. K. Khandani and P. Kabal, "Unsymmetrical boundary shaping with applications to spectral shaping," submitted to *IEEE Trans. Inform. Theory.*

[35] A. K. Khandani and P. Kabal, "Block-based eigensystem of the $1 \pm D$ and $1 - D^2$ partial-response channels," submitted to *IEEE Trans. Inform. Theory.*

[36] A. K. Khandani, P. Kabal and H. Leib, "Combined coding and shaping over a multitone channel," *Proc. IEEE Global Telecommunications Conference,* Phoenix, Arizona, pp. 1182–1186, Dec 1991,

[37] A. K. Khandani and P. Kabal, "Optimum block-based signaling over $1 \pm D$ and $1 - D^2$ partial-response channels," *Proc. Twenty-ninth Annual Allerton Conference on Communications, Control, and Computing,* Allerton, IL, pp. 779–789, Oct 1991.

[38] A. K. Khandani, P. Kabal H. Leib, "Block-based signaling over partial-response channels," submitted to *IEEE Trans. Inform. Theory.*

[39] A. K. Khandani and P. Kabal, "Nonuniform lattice-based vector quantization," to be presented at *Twenty-sixth annual conference on information sciences and systems,* Princeton, NJ, 19–20 March 1992.

[40] G. R. Lang and F. M. Longstaff, "A leech lattice modem," *IEEE J. Select. Areas Commun.,* vol. SAC-7, pp. 968–973, August 1989.

[41] J. Leichleider, "The optimum combination of block codes and receivers for arbitrary channels," *IEEE Trans. Commun.,* vol. COM-38, pp. 615–621, May 1990.

[42] L. F. Wei, "Trellis coded modulation with multidimensional constellations," *IEEE Trans. Inform. Theory,* vol. IT–33, pp. 483–501, July 1987.

[43] E. T. Whittaker and G. N. Watson, *A course of modern analysis,* Camb. Univ. Press, 4th ed., 1963.

[44] J. K. Wolf and G. Ungerboeck, "Trellis coding for partial-response channels," *IEEE Trans. Commun.,* vol. COM-34, pp. 765–773, August 1986.