Proc. Biennial Symp. Commun. (Kingston, ON), pp. 296-299, May 1992

Signaling in Multi-dimensional Signal Spaces†

A. K. Khandani[1] and P. Kabal[1,2]
[1]Dept. of Elec. Eng., McGill University, 3480 University, Montreal, Canada, H3A 2A7
[2]INRS-Telecommunications, 16 Place du Commerce, Verdun, Canada, H3E 1H6

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 C as the boundary instead of a hypercube is called the shape gain of 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. 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 discussed. Examples of the optimum tradeoff curves are given. 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. We discuss two addressing schemes with low complexity and good performance. In spectral shaping, the rate of the constellation is maximized subject to some constraints on its power spectrum. This results in a shaping region which has different values of power along different dimensions (unsymmetrical shaping). This spectral shaping also involves the selection of an appropriate basis (modulating waveform) for the space. Finally, we discuss the selection of a signal constellation for signaling over a partial-response channel using both continuous approximation and discrete analysis. We also present a closed form formula for the weight distribution of the scaled D_4 and E_8 lattices.

1 Introduction

In 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. To achieve the transmission, we select M points over the channel space. Each of the transmitter waveforms corresponds to one of these points. This is called a signal constellation.

In the design of a signal constellation, the overall objective is to minimize the probability 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). The figure of merit is the reduction in the required average energy with respect to a reference scheme. 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.

†This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

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.

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, [11].

Unfortunately, the situation is not as good as the conventional calculation methods based on the minimum distance to the nearest neighbor shows. Forney mentions in [7] 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 use of the region C as the boundary instead of using a hypercube is called the shape gain of C and is denoted as $\gamma_s(C)$, [5]. The price to be paid for shaping involves: (i) an increase in the factor CER_s[1], (Constellation-Expansion-Ratio) (ii) an increase in the factor PAR[2], (Peak-to-Average-power-Ratio) and (iii) an increase in the addressing complexity[3].

For a given dimensionality N, a spherical shaping region S_N, is the region with the highest possible γ_s but also with high values for CER_s and PAR. It is well known that as $N \to \infty$, the shape gain of S_N tends to 1.53 dB. This is an upper bound for the shape gain of all regions and is achievable at the price of $\text{CER}_s = \infty$ and also $\text{PAR} = \infty$. 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, 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.

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

[2]The PAR is ratio of the peak of the energy per two dimensions to the average energy per two dimensions.

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

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

The major problem associated with shaping in a high dimensional space is the addressing 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.

2 Previous work

In the work of Wei, [11], 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 [5]. Conway and Sloane in [4] 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 [8]. In [2], 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. The idea of the trellis shaping is introduced in [9]. This is based on using an infinite dimensional Voronoi region, determined by a convolutional code, to shape the constellation.

3 Optimum shaping

In the following we discuss the structure of the 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. Assuming that the CER_s and the PAR are measured on a 2-D basis, it can be shown that the optimum region is equal to,

$$\mathcal{A}_N(\psi) = \{\mathcal{S}_2(R_2)\}^n \cap \mathcal{S}_N(R_N), \tag{1}$$

where $\mathcal{S}_N(R)$ denotes an N-D sphere of radius R, $n=N/2$ and $\psi = R_N^2/nR_2^2$. For $\psi=1$, we have $\mathcal{A}_N = \{\mathcal{S}_2(R_2)\}^n$. This results in the starting point on the tradeoff curves. For $1/n < \psi < 1$, by decreasing ψ, we move along the optimum tradeoff curve. Finally, for $\psi = 1/n$, we obtain the spherical region $\mathcal{S}_N(R_N)$. By applying a change of variable denoted as the shell mapping, the region $\mathcal{A}_N$ is mapped to an $n=N/2$-D hypercube of edge length one truncated within a simplex of edge length $\beta = R_N^2/R_2^2$. This region is denoted as $\mathcal{TC}_n(1,\beta)$.

Shell mapping has the following properties:

- A uniform density within $\mathcal{A}_N$ results in a uniform density within $\mathcal{TC}_n$. This allows us to achieve the shaping, addressing on the equal volume partitions of $\mathcal{TC}_n$.
- 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 $\mathcal{TC}_n$ region is equal to the Voronoi region of the lattice D_n^* in the positive coordinates. This allows us to use a Voronoi constellation, [8], for the addressing.

In [10], 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. This integral is used to calculate the γ_s, the CER_s and the PAR. Figure 1 shows the optimum tradeoff between the CER_s and the γ_s for different values of N. The curve corresponding to $N=\infty$ is extracted from [5]. The marked point in each case corresponds to the addressing scheme based on the lattice D_n^*.

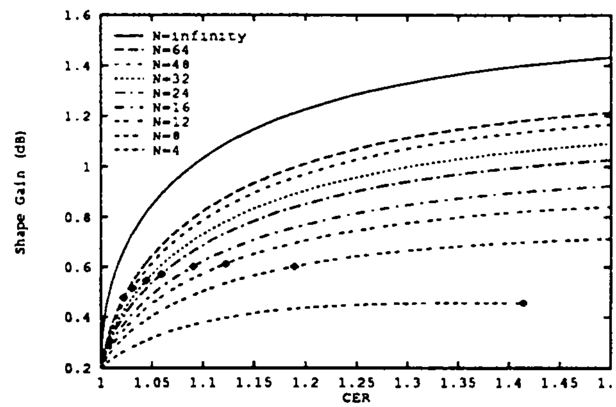

Fig. 1 Optimum tradeoff between CER_s and γ_s.

In general, the initial part of the optimum tradeoff curve has 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.

4 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., $\mathcal{A}_{N'} = \{\mathcal{S}_2\}^{N'/2}$. 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. We have found analytical expressions for the corresponding tradeoff, [10]. The calculations show that by selecting $\{K_i, i=1,\ldots\} = \{64, 64, 128, 256, \ldots\}$ we can essentially achieve the optimum tradeoff. 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. We have called 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. 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. Figure 2 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. It has no associated computation and is easy to implement. Also, it can be easily used in conjunction with the coding schemes of [11].

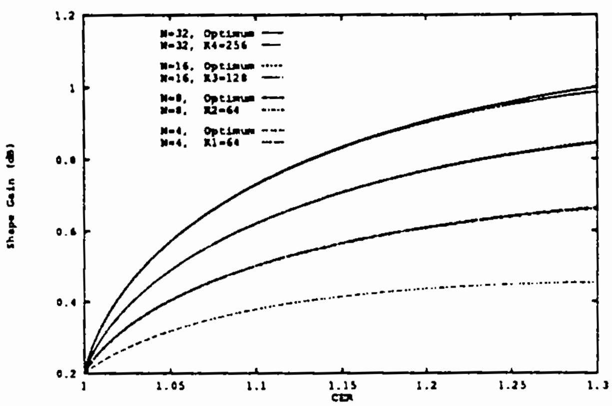

Fig. 2 Tradeoff between CER_s and γ_s using the address decomposition method.

As an example, for $N = 32$, and $M = 128$ points per 2-D subconstellations, we need $M_s = 44$ kilo-bytes per N dimensions to achieve $\gamma_s = 0.89$, $\text{CER}_s = 1.19$, $\text{PAR} = 2.8$. This point can not be distinguished from the optimum curve. As an alternate point, our method needs $M_s = 36$ kilo-bytes per N dimensions to achieve $\gamma_s = 1.02$, $\text{CER}_s = 1.41$ and $\text{PAR} = 3.42$ while the optimum point corresponding to $N = 32$ and $\text{CER}_s = 1.41$ satisfy, $\text{PAR} = 3.45$ and $\gamma_s = 1.06$.

5 Unsymmetrical Boundary Shaping, Spectral Shaping

Assuming continuous approximation, the selection of the constellation is composed of selecting a basis for the space and a shaping region for the points. 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 dimension. Without additional constraints, elliptical regions are optimum.

The unsymmetrical region can be obtained by scaling of a symmetrical baseline region. The baseline region can be selected independently of the scale factors and the basis. The scale factors and the constellation basis are computed by an optimization procedure. This procedure maximizes the rate of the constellation 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 the zero and/or at the Nyquist frequency. It is shown that this maximization is equivalent to maximizing the determinant of the correlation matrix subject to some linear constraints on its elements. In an optimized basis analysis, the optimum correlation matrix is found. In a fixed basis analysis, the eigenvectors of the correlation 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. We also given analytical expressions for the eigensystems of the $1 \pm D$ and $1 - D^2$ partial response channels. Appendix A contains the final results. The output eigenvectors provide an orthonormal basis with spectral null(s) at zero and/or Nyquist frequency. The important point is that the eigenvectors are closely related to the sine basis. This allows for the use of the fast transform algorithms for the modulation.

6 Block-based signaling over a partial response channel, combined shaping and coding

We discuss the selection a signal constellation for use over a partial-response channel. The constellation design is composed of three parts: (i) selection of the internal structure of the constellation (channel coding), and (ii) selection of the constellation boundary (shaping), and (iii) selection of the constellation dimensions (modulation). The objective is to minimize the degradation caused by the channel memory and the additive white Gaussian noise.

Assuming continuous approximation, shaping, coding and modulation can be selected independently. The procedure is similar to the case of a flat channel. The main difference is that here some of the dimensions may be empty. In the scheme proposed in [10], the determination of the nonempty dimensions 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. The optimum modulator is the set of the channel input eigenvectors, [1]. In the conventional schemes, to reduce the computational complexity, the optimum basis is usually replaced by the Fourier basis. Referring to appendix A, it is seen that the optimum modulation over the $1 \pm D$ and $1 - D^2$ channels can be achieved by using a fast sine transform algorithm. In [10], numerical results for the optimum basis and also for the Fourier basis over $1 \pm D$ channels are presented. The optimum basis shows about 0.5 dB improvement with respect to the Fourier basis while the computational complexities are the same.

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 error. We introduce two methods to do this. 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. 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. The calculation of the error probability is based on using the weight distribution of the lattice. As part of the calculations, we have found a new closed form formula for the weight distribution of the scaled D_4 and E_8 lattices. The calculation is based on using the trellis diagram of the lattice, [6]. 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,

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

and,

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

where θ_2 and θ_3 are the Jacobi theta functions, [3], and D_j is the minimum square distance along the j'th 2-D subspace.

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

An N-dimensional $1 - D$ system have an $(N+1) \times N$ transfer matrix with the i'th column equal to, $[(0)^i, \sqrt{2}/2, \pm\sqrt{2}/2, (0)^{N-1-i}]$. An N-dimensional $1 - D$ system has an $(N+2) \times N$ transfer matrix with the i'th column equal to, $[(0)^i, \sqrt{2}/2, 0, -\sqrt{2}/2, (0)^{N-1-i}]$.

For an N-dimensional $1 - D$ system, the input eigenvectors are equal to the sine basis, [10], i.e.,

$$m_i(n) = \sqrt{\frac{2}{N+1}}\ \sin\frac{\pi(i+1)(n+1)}{N+1}, \quad (4)$$

where $i, n = 0, \ldots, N-1$. The corresponding eigenvalues are equal to,

$$\phi_i = 1 - \cos\frac{\pi(i+1)}{(N+1)}. \quad (5)$$

Using (4) in $\mathbf{A}\mathbf{m}_i = \sqrt{\phi_i}\hat{\mathbf{m}}_i$, the output eigenvectors of the $1 - D$ system are found as,

$$\hat{m}_i(n) = \sqrt{\frac{2}{N+1}}\ \cos\frac{\pi(i+1)(n+0.5)}{N+1}, \quad (6)$$

where $n = 0, \ldots, N$ and $i = 0, \ldots, N-1$.

The input and output eigenvectors of $1 + D$ system are obtained by multiplying (4) and (6) with $(-1)^n$. The eigenvalues are the same as the $1 - D$ system given in Eq. (5).

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

$$\phi_i = 1 - \cos\frac{\pi(i+1)}{0.5N+1},\ i = 0, \ldots, 0.5N-1. \quad (7)$$

The two eigenvectors corresponding to a pair of eigenvalues are of the general form $\alpha_1 m_i(2n) + \alpha_2 m_i(2n+1)$ where $\alpha_1^2 + \alpha_2^2 = 1$ and $m_i(n)$ is the eigenvector of an $N/2$-dimensional $1 - D$ system.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] G. D. Forney, "Coset codes–Part II: Binary lattices and related codes," *IEEE Trans. Inform. Theory*, vol. IT–34, pp. 1152–1187, September 1988.

[7] G. D. Forney, "Coset codes—Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1123–1151, September 1988.

[8] G. D. Forney, Jr., "Multidimensional constellations—Part II: Voronoi constellations," *IEEE J. Select. Areas Commun.*, vol. SAC-7, pp. 941–958, August 1989.

[9] G. D. Forney, "Trellis shaping," *IEEE Trans. Inform. Theory*, vol. IT–38, pp. 281–300, March 1992.

[10] A. K. Khandani *Shaping multi-dimensional signal spaces*, Ph.D. Dissertation, McGill Univ., March 1992.

[11] L. F. Wei, "Trellis coded modulation with multidimensional constellations," *IEEE Trans. Inform. Theory*, vol. IT–33, pp. 483–501, July 1987.