Coset Codes—Part I: Introduction and Geometrical Classification

G. DAVID FORNEY, JR., FELLOW, IEEE

Invited Paper

Abstract—**Practically all known good constructive coding techniques for band-limited channels, including lattice codes and various recently proposed trellis-coded modulation schemes, can be characterized as coset codes. A coset code is defined by a lattice partition Λ/Λ' and by a binary encoder C that selects a sequence of cosets of the lattice Λ'. The fundamental coding gain of a coset code, as well as other important parameters such as the error coefficient, the decoding complexity, and the constellation expansion factor, are purely geometric parameters determined by C and Λ/Λ'. The known types of coset codes, as well as a number of new classes that systematize and generalize known codes, are classified and compared in terms of these parameters.**

I. Introduction

A. History

THE FACT that the channel capacity of band-limited channels with white Gaussian noise is some 9 dB beyond what can be achieved with simple pulse amplitude modulation (PAM) was an immediate consequence of Shannon's original work (see [1]). It is therefore remarkable that approximately three decades passed before serious work began on developing constructive coding techniques that could achieve sizable fractions of this potential gain.

The field was not completely inactive during this period. Shannon's work had indicated that there must be sphere packings in spaces of high dimension with sufficiently high density to approach channel capacity. In the 1950's and 1960's mathematicians developed some constructive dense sphere packings based on lattices in spaces of moderate to high dimension, notably the 2^n-dimensional Barnes–Wall lattices [2], and the ultradense 24-dimensional Leech lattice [3]. The earliest advocate of the use of lattices for communications appears to have been Lang in Canada in the early 1960's (see, e.g., [4, preface]). In an interesting interplay between mathematics and communications, it seems that Lang's calculations of bounds on maximum lattice density for 8–32 dimensions helped to motivate Leech to discover his now-famous lattice [3, p. 265]. Another longtime proponent of lattices in communications, also in Canada, has been deBuda, who proved that the coding theorem applies to lattice codes [5]. Connections between lattice theory and coding theory were made by Leech and Sloane [6], and Sloane has since continued to develop bridges between these disciplines. This work is authoritatively and comprehensively summarized in [7].

It is probably fair to say, however, that it was the trellis-coded modulation schemes of Ungerboeck [8] that captured the attention of the modulation community and inspired widespread practical application as well as intensified research. Through a technique called "mapping by set partitioning" of one- and two-dimensional signal constellations, combined with binary convolutional codes, Ungerboeck showed how coding gains of the order of 3 dB could be obtained with simple four-state codes, while gains of up to 6 dB could be obtained with more complex (128-state) codes. A variant [9] of Ungerboeck's eight-state two-dimensional scheme has been adopted in international standards for both 14.4 kbit/s private-line modems and 9.6 kbit/s switched-network modems and is coming into wide commercial use.

More recently, trellis-coded modulation schemes using multidimensional signal constellations have been developed. A simple four-dimensional scheme of Gallager was presented by Forney *et al.* in [1], and a similar scheme was discovered independently by Calderbank and Sloane [10]. Wei [11] has developed a class of multidimensional schemes that are highly suited for implementation, one of which is used in a Codex 19.2-kbit/s modem. Calderbank and Sloane [12], [13] have also developed a variety of classes of new trellis codes.

In an earlier paper [1] Forney *et al.* pointed out that all schemes known at that time, including the most important lattice codes, could be generated by the same basic elements:

1) a conventional binary encoder, block or convolutional, operates on certain of the data bits to provide a larger number of coded bits;
2) these coded bits select one of the subsets of a partitioned signal constellation;
3) additional uncoded bits select an individual signal point from the selected subset.

Manuscript received September 2, 1986; revised September 18, 1987. This paper was presented at the 1987 Information Theory Workshop, Bellagio, Italy, June 24, 1987.

The author is with the Codex Corporation, 7 Blue Hill River Road, Canton, MA 02021.

IEEE Log Number 8824503.

0018-9448/88/0900-1123$01.00

This way of looking at coding schemes has several important consequences:

1) The code distance properties, and thus the fundamental coding gain, are determined by the binary encoder and the subset partitioning, which are largely decoupled from the choice of signal constellation in the third step. In this last step there is a trade-off between optimal shaping of the signal constellation and implementation simplicity, but this is almost independent of the fundamental coding scheme and has only a minor effect on the overall coding gain. Finally, in decoding, the first operation can always be to determine the best signal point in each subset and its metric; after that step, decoding depends again only on the fundamental code structure determined by the first two encoding operations.

2) Different classes of codes can be readily compared and contrasted in this common framework. In this paper, which is to a large extent a sequel to [1], we concern ourselves only with the code structure imposed by the first two encoding operations, regarding the constellation shaping as peripheral; in our view this clarifies the similarities and differences between various schemes.

B. Introduction to Coset Codes

Calderbank and Sloane [13] have made the important observations that the signal constellation should be regarded as a finite set of points taken from an infinite lattice, and that the partitioning of the constellation into subsets corresponds to the partitioning of that lattice into a sublattice and its cosets. This lattice/coset language is both illuminating and powerful, and we have found that all of the good coded modulation schemes mentioned above can be put into this framework. (These codes may be more specifically characterized as lattice-type coset codes, which are the only type considered in this paper, apart from a brief mention of phase-modulated coset codes in the next section.)

We call this general class of coded modulation schemes *coset codes*. They seem to provide a general approach to the construction of implementable codes for band-limited channels that approach channel capacity, just as conventional codes (both block and convolutional) do for the power-limited case.

We now give a quick preview of the elements of coset codes, and of key terms and concepts that will figure in the rest of the paper. Fig. 1 illustrates the general structure of an encoder for a coset code, embodying the three principal elements just described and using the language of lattices and cosets (compare [1, fig. 10] or [13, fig. 2]). The main ingredients are as follows:

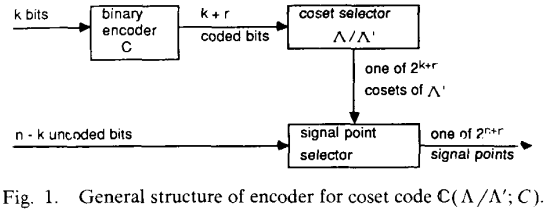


Fig. 1. General structure of encoder for coset code $\mathbb{C}(\Lambda/\Lambda'; C)$.

1) An N-dimensional *lattice* Λ, which we can think of as an infinite regular array of points in N-space. The signal points will all be taken from a finite subset of points lying within a translate (coset) of Λ, and the set of all possible signal points is called the *signal constellation*.

2) A *sublattice* Λ' of Λ, i.e., a subset of the points of Λ which is itself an N-dimensional lattice. The sublattice induces a *partition* Λ/Λ' of Λ into $|\Lambda/\Lambda'|$ *cosets* of Λ', where $|\Lambda/\Lambda'|$ is the *order* of the partition; when Λ and Λ' are *binary lattices*, the order of the partition is a power of 2, say 2^{k+r}, and correspondingly, the partition divides the signal constellation into 2^{k+r} subsets, each corresponding to a distinct coset of Λ'.

3) A *rate-$k/(k+r)$ binary encoder* C, which takes in k bits per N dimensions and puts out $k+r$ *coded bits*; the latter select one of the cosets of Λ' in the partition Λ/Λ'. The *redundancy* $r(C)$ of C is r bits per N dimensions; the *normalized redundancy* per two dimensions is $\rho(C) = 2r(C)/N$.

The *coset code* $\mathbb{C}(\Lambda/\Lambda'; C)$ is the set of all sequences of signal points that lie within a sequence of cosets of Λ' that could be specified by a sequence of coded bits from C. Some lattices, including the most useful ones, can be generated as *lattice codes* $\mathbb{C}(\Lambda/\Lambda'; C)$, where Λ and Λ' are lattices of lower dimension, and C is a binary block code. If C is a convolutional encoder, then $\mathbb{C}(\Lambda/\Lambda'; C)$ is a *trellis code*.

The *fundamental coding gain* of the coset code is denoted by $\gamma(\mathbb{C})$ and is defined by two elementary geometrical parameters: the *minimum squared distance* $d^2_{\min}(\mathbb{C})$ between signal point sequences in $\mathbb{C}$ and the *fundamental volume* $V(\mathbb{C})$ per N dimensions, which is equal to $2^{r(\mathbb{C})}$ where the *redundancy* $r(\mathbb{C})$ is equal to the sum of the redundancy $r(C)$ of the encoder C and the redundancy $r(\Lambda)$ of the lattice Λ. In fact,

$$\gamma(\mathbb{C}) = 2^{-\rho(\mathbb{C})} d^2_{\min}(\mathbb{C})$$

where the *normalized redundancy* $\rho(\mathbb{C})$ (per two dimensions) is equal to $2r(\mathbb{C})/N$.

To transmit n bits per N dimensions, the signal constellation must consist of 2^{n+r} points from a coset of Λ, partitioned into 2^{k+r} subsets, each consisting of 2^{n-k} points from a different coset of Λ'. Given a selected coset of Λ', $n-k$ *uncoded bits* select a particular signal point from that coset. The *constellation expansion factor* (compared to an uncoded constellation of 2^n points from a coset of Λ) is thus $2^{r(C)}$ per N dimensions, or $2^{\rho(C)}$ per two dimensions. This translates into an average power cost of a factor of $2^{\rho(C)}$ (or $\rho(C) \cdot 3.01$ dB), which is reflected in the formula for the fundamental coding gain $\gamma(\mathbb{C})$ just given.

The *total coding gain* $\gamma_{\text{tot}}(\mathbb{C})$ is the product of the fundamental coding gain $\gamma(\mathbb{C})$ with the *shape gain* γ_s of the finite constellation (γ_s is defined as $\gamma_{\text{tot}}(\mathbb{C})/\gamma(\mathbb{C})$ and is approximately equal to the ratio of the *normalized second moment* [7] of an N-cube to that of the region of N-space in which the constellation is contained). If the constella-

tion is the set of points in a coset of Λ that lie within an N-cube, $\gamma_s = 1$ and $\gamma_{tot}(\mathbb{C}) = \gamma(\mathbb{C})$. If the constellation is chosen more like an N-sphere to reduce average power, γ_s measures this reduction. Thus γ_s, unlike $\gamma(\mathbb{C})$, reflects finite constellation effects. Even for the same code $\mathbb{C}$, γ_s will in general vary with n, unlike $\gamma(\mathbb{C})$. γ_s is usually much smaller than $\gamma(\mathbb{C})$, being upper-bounded by the shape gain of an N-sphere (see next paragraph); however, calculating γ_s (or $\gamma_{tot}(\mathbb{C})$) is usually more cumbersome than calculating $\gamma(\mathbb{C})$. Finally, γ_s is determined not by $\mathbb{C}$ but by the choice of constellation boundary; in general, a similar gain can be achieved in uncoded systems by choosing a similar boundary for an uncoded constellation in N dimensions. For all these reasons we prefer to focus on the fundamental coding gain $\gamma(\mathbb{C})$ in this paper. (Multidimensional constellations will be considered in [14].) We feel that the various values for $\gamma_{tot}(\mathbb{C})$ that have appeared in the prior literature have confused rather than clarified the fundamental properties of these codes and the comparisons between them.

(Calderbank and Sloane [13] also calculate an asymptotic coding gain whose value is independent of finite constellation effects. This gain is the combination of our "fundamental coding gain" and the shape gain of an N-sphere over an N-cube, which for N even is $G_{\otimes} = \pi(n+1)/[6(n!)^{1/n}]$, where $n = N/2$; thus $G_{\otimes} = \pi/3$ (0.20 dB) for $N = 2$, $\pi/2^{3/2}$ (0.46 dB) for $N = 4$, $5\pi/6(24)^{1/4}$ (0.73 dB) for $N = 8$, with a limit of $\pi e/6$ (1.53 dB) as $N \to \infty$ [1].)

As a simple example of a coset code, let us consider the four-state two-dimensional Ungerboeck code illustrated in Fig. 2, transmitting 5 bits per two dimensions with the "square" 64-point constellation of Fig. 3(a). In this case the lattice Λ is the two-dimensional integer lattice $\boldsymbol{Z}^2$, i.e., the set of all integer 2-tuples (the signal constellation is actually chosen from its translate $\boldsymbol{Z}^2 + (1/2, 1/2)$ for symmetry). Thus the scale is such that the minimum (squared) distance between points in the constellation is one. The sublattice Λ' is the two-dimensional lattice $2\boldsymbol{Z}^2$, i.e., the set of all even integer 2-tuples. The partition $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$ has order 4, i.e., $\boldsymbol{Z}^2$ is the union of four cosets of $2\boldsymbol{Z}^2$, which correspond to the points labeled A, B, C, and D in Fig. 3. There are 16 points in the constellation from each of the four cosets. The minimum squared distance between points in any coset of $2\boldsymbol{Z}^2$ is four. The encoder C is a rate-1/2 four-state convolutional encoder. Thus the redundancy per two dimensions is $r(C) = \rho(C) = 1$. One of the five input bits per two dimensions goes into the encoder, and the two resulting coded bits select one of the

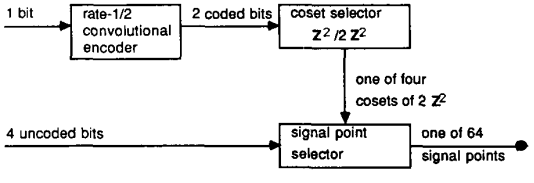
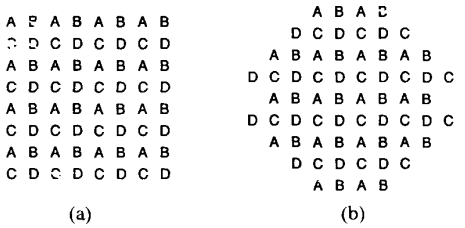


Fig. 2. Four-state two-dimensional Ungerboeck code ($n = 5$).



(a) (b)

Fig. 3. Two 64-point signal constellations based on partition $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$. (a) Square. (b) Cross.

four cosets A, B, C, and D; the remaining four uncoded bits select one of the 16 points from the selected coset.

As we show below, C can be chosen so that $d^2_{min}(\mathbb{C}) = d^2_{min}(\Lambda') = 4$; since $\rho(\mathbb{C}) = \rho(C) = 1$ (the integer lattice has zero redundancy), the fundamental coding gain $\gamma(\mathbb{C})$ is $2^{-1} \cdot 4 = 2$ (3.01 dB) (the minimum squared distance gain is a factor of four, but this is offset by a constellation expansion power cost of a factor of two, leaving a net gain of a factor of 2). Because the constellation is square, this is also the total coding gain. The total coding gain could be improved slightly by use of a more circular constellation, e.g., the 64-point "cross" constellation of Fig. 3(b), which is about 0.1 dB better [1].

C. Other Coset Codes

The coset codes described in this paper are based on partitions of binary lattices. More generally, a coset code $\mathbb{C}(S/T; C)$ can be defined whenever S is some set of discrete elements that forms an algebraic group, with some distance measure between elements of S, T is a subgroup of S such that the quotient group S/T has finite order $|S/T|$, and C is an appropriate code whose codewords select sequences of cosets of T in the partition S/T.

For instance, S can be a binary block code, with T a subcode, and Hamming distance as the distance measure. We shall see some examples of the construction of convolutional codes in this way in this paper (other such constructions appear in [15]). In [18] we shall see how such codes as the Reed–Muller and Golay codes can be built up from short codes in this way. In [16], we shall show how ternary codes, lattices, and trellis codes can be constructed as coset codes, where S and T are ternary block codes or lattices, and C is a block or convolutional code over the ternary field GF(3).

Finally, phase-modulated codes can be constructed as coset codes as follows. The signal constellation for m-ary PSK (mPSK) can be regarded as the m complex mth roots of unity, which forms a group S under complex multiplication. If n divides m, the nPSK constellation is a subgroup T. Thus, for example, 16PSK/8PSK/4PSK/2PSK/1PSK is a chain of two-way partitions. Although the minimum squared distances within these constellations are somewhat different from those in lattice partitions (e.g., 0.152/0.586/2/4/∞ for these constellations, compared to 1/2/4/8/16 for the comparable two-dimensional lattice partition $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2/2R\boldsymbol{Z}^2/4\boldsymbol{Z}^2$), similar constructions to those presented here often yield good phase-

modulated codes (e.g., the phase-modulated codes of Ungerboeck [8] or of LaFanchere *et al.* [17]).

In general, these code constructions rely very little on the linearity properties of the groups (e.g., lattices, sublattices) on which they are based, and the codes so constructed are often not linear, particularly the trellis codes. The essential properties of these sets seem to be their partition structure and related distance properties, which of course were the basis for Ungerboeck's constructions via 'mapping by set partitioning.' The primary benefit of starting with sets that are groups seems to be that their subgroups naturally induce useful partitions via coset decompositions.

D. Outline

The primary subject of this paper is the categorization of various lattice-type coset codes in terms of their key parameters and, ultimately, in terms of their performance as measured by their fundamental coding gains.

Since these codes are based on partitions of *binary lattices*, we begin with an introduction to such lattices in Section II, which is intended to be self-contained. A more comprehensive introduction to this family of lattices is given in the companion paper [18], hereafter referred to as part II. (By far the best general reference on lattices is the forthcoming encyclopedic book by Conway and Sloane [7].) Section III summarizes those results from part II most relevant to this paper, particularly the performance of lattices as lattice codes, their partition/distance structure, and the decoding complexity of lattices and lattice partitions as measured by the number of binary operations of the trellis-based decoding algorithms given in part II.

Section IV then extends lattice theory to trellis codes. Section V characterizes the principal trellis codes that have been introduced to date as coset codes. Section VI introduces some generic classes of trellis codes whose main parameters can be easily determined and which include codes similar to (and in many cases equivalent to) the principal known codes. Finally, Section VII compares and contrasts all of these codes in terms of performance vs. complexity.

It is intended that this paper and part II may be read independently; as a result, there is some overlap. The reader who desires to read both papers in the most logical order is advised to skim this paper quickly through Section II, omitting proofs; then to read part II, with primary focus on the material relating to Barnes–Wall lattices; and then to return to the rest of this paper. The mathematical level is kept as elementary as possible; for the more mathematically inclined reader, we recommend learning about lattices by reading Conway and Sloane [7].

II. A Lattice Primer

A. Definitions

A real *lattice* Λ is simply a discrete set of vectors (points, N-tuples) in real Euclidean N-space $\boldsymbol{R}^N$ that forms a group under ordinary vector addition, i.e., the sum or difference of any two vectors in Λ is in Λ. Thus Λ necessarily includes the all-zero N-tuple $\mathbf{0}$, and if λ is in Λ, then so is its additive inverse $-\lambda$. The vectors in a lattice may possibly span fewer than N dimensions; however, this will not be the case for any lattice considered here, so there will be no confusion if we call a lattice of real N-tuples an *N-dimensional* real lattice.

As an example, the set $\boldsymbol{Z}$ of all integers is essentially the only one-dimensional real lattice, up to scaling, and the prototype of all lattices. The set $\boldsymbol{Z}^N$ of all integer N-tuples is an N-dimensional real lattice for any N.

Lattices have only two principal structural characteristics. Algebraically, a lattice is a group; this property leads to the study of subgroups (sublattices) and partitions (coset decompositions) induced by such subgroups. Geometrically, a lattice is endowed with the properties of the space in which it is embedded, such as the Euclidean distance metric and the notion of volume in $\boldsymbol{R}^N$. The following two sections are concerned with these two aspects of lattice structure.

Lattices closely related to a given real N-dimensional lattice Λ are obtained by the following operations.

1) Scaling: If r is any real number, then $r\Lambda$ is the lattice consisting of all multiples $r\lambda$ of vectors λ in Λ by the scalar r.

2) Orthogonal Transformation: More generally, if T is any scaled orthogonal transformation of N-space, then $T\Lambda$ is the lattice consisting of all transformations $T\lambda$ of vectors λ in Λ by T. We say that $T\Lambda$ is a *version* of Λ.

3) Cartesian Product: The M-fold Cartesian product of Λ with itself—i.e., the set of all MN-tuples $(\lambda_1, \lambda_2, \cdots, \lambda_M)$ where each λ_j is in Λ—is an MN-dimensional lattice denoted by Λ^M.

For example, $\boldsymbol{Z}^N$ is the N-fold Cartesian product of $\boldsymbol{Z}$ with itself, and $r\boldsymbol{Z}^N$ is a scaled version of $\boldsymbol{Z}^N$ for any r and N. The two-dimensional lattice $\boldsymbol{Z}^2$ is illustrated in Fig. 4.

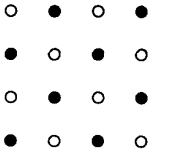


Fig. 4. Lattice $\boldsymbol{Z}^2$ and its sublattice $R\boldsymbol{Z}^2$ (black dots).

The most important scaled orthogonal transformation for our purposes is the *rotation operator* R, defined by the 2×2 matrix

$$R \triangleq \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

$R\boldsymbol{Z}^2$ is a version of $\boldsymbol{Z}^2$ obtained by rotating $\boldsymbol{Z}^2$ by 45° and scaling by $2^{1/2}$ and is also illustrated in Fig. 4. The points in $R\boldsymbol{Z}^2$ are a subset of the points in $\boldsymbol{Z}^2$, meaning that $R\boldsymbol{Z}^2$ is a sublattice of $\boldsymbol{Z}^2$. Note that $R^2 = 2I$, where I

is the identity operator (in two dimensions), so that $R^2\boldsymbol{Z}^2 = 2\boldsymbol{Z}^2$.

We can define a $2N$-dimensional rotation operator by letting R operate on each pair of coordinates in a $2N$-tuple; with a slight abuse of notation, we denote by R any such rotation operator. For instance, in four dimensions,

$$R \triangleq \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Note that $R^2 = 2I$ for any N, where I is the identity operator in $2N$ dimensions, so that $R^2\Lambda = 2\Lambda$ for any real $2N$-dimensional lattice Λ.

B. Group Properties

A *coset* of a lattice Λ, denoted by $\Lambda + \boldsymbol{c}$, is the set of all N-tuples of the form $\lambda + \boldsymbol{c}$, where λ is any point in Λ and $\boldsymbol{c}$ is some constant N-tuple that specifies the coset. Geometrically, the coset $\Lambda + \boldsymbol{c}$ is therefore a *translate* of Λ by $\boldsymbol{c}$ (if $\boldsymbol{c}$ is in Λ, then $\Lambda + \boldsymbol{c} = \Lambda$). Two N-tuples are *equivalent modulo* Λ if their difference is a point in Λ. Thus the coset $\Lambda + \boldsymbol{c}$ is the set of all points equivalent to $\boldsymbol{c}$ modulo Λ.

A *sublattice* Λ' of a lattice Λ is a subset of the elements of Λ that is itself a lattice, i.e., Λ' is a subgroup of the additive group Λ. Thus, by elementary group theory, a sublattice Λ' induces a *partition* (denoted by Λ/Λ') of Λ into equivalence classes modulo Λ' (the equivalence classes may be added modulo Λ' and form the *quotient group* Λ/Λ'). We shall say that the *order* of the partition (or quotient group) Λ/Λ' is the number $|\Lambda/\Lambda'|$ of such equivalence classes (in the mathematical literature, $|\Lambda/\Lambda'|$ is usually called the *index* of Λ' in Λ). Each equivalence class is a coset of Λ' (one being Λ' itself), or, geometrically, a translate of Λ'. For example, the partition $\boldsymbol{Z}^2/R\boldsymbol{Z}^2$ has order $|\boldsymbol{Z}^2/R\boldsymbol{Z}^2| = 2$, and Fig. 4 illustrates $\boldsymbol{Z}^2$ as the union of two cosets of $R\boldsymbol{Z}^2$. Of course, any N-dimensional integer lattice Λ is a sublattice of $\boldsymbol{Z}^N$.

If we take one element from each equivalence class, we obtain a system of *coset representatives* for the partition Λ/Λ', denoted by $[\Lambda/\Lambda']$. (In general, there are many ways of selecting such a system $[\Lambda/\Lambda']$, so the notation does not entirely specify the system.) Then every element of Λ can be written uniquely as a sum $\lambda = \lambda' + \boldsymbol{c}$, where $\boldsymbol{c} \in [\Lambda/\Lambda']$ is the coset representative of the equivalence class in which λ lies, and $\lambda' = \lambda - \boldsymbol{c}$ is an element of Λ' (because $\lambda \equiv \boldsymbol{c} \bmod \Lambda'$). This is called a *coset decomposition* of Λ and will be written here as

$$\Lambda = \Lambda' + [\Lambda/\Lambda'].$$

For example, the two 2-tuples $(0,0)$ and $(1,0)$ are a system of coset representatives for the partition $\boldsymbol{Z}^2/R\boldsymbol{Z}^2$, and every element of $\boldsymbol{Z}^2$ may be written as the sum of one of these two 2-tuples with an element of $R\boldsymbol{Z}^2$, i.e., $\boldsymbol{Z}^2$ is the union of $R\boldsymbol{Z}^2 + (0,0) = R\boldsymbol{Z}^2$ and $R\boldsymbol{Z}^2 + (1,0)$ (the black dots and white dots in Fig. 4, respectively).

As another example, if m is any integer, the lattice $m\boldsymbol{Z}$ of integer multiples of m is a sublattice of $\boldsymbol{Z}$. The partition $\boldsymbol{Z}/m\boldsymbol{Z}$ is the partition of the integers into m equivalence classes modulo $m\boldsymbol{Z}$ (modulo m), and the order of the partition is m. The integers $\{0,1,\cdots,m-1\}$ form a system of coset representatives for the partition $\boldsymbol{Z}/m\boldsymbol{Z}$, and every integer n can be written uniquely as $n = am + c$, where am is an element of $m\boldsymbol{Z}$ and $c \in \{0,1,\cdots,m-1\} = [\boldsymbol{Z}/m\boldsymbol{Z}]$ (thus $[\boldsymbol{Z}/m\boldsymbol{Z}]$ is essentially the ring $\boldsymbol{Z}_m$ of integers modulo m). In particular, the partition $\boldsymbol{Z}/2\boldsymbol{Z}$ has order 2 and divides the integers into two subsets, $2\boldsymbol{Z}$ (the even integers) and $2\boldsymbol{Z}+1$ (the odd integers).

More generally, for any $m \in \boldsymbol{Z}$, the lattice $m\boldsymbol{Z}^N$ of all N-tuples of integer multiples of m is a sublattice of $\boldsymbol{Z}^N$ of order m^N, and $[\boldsymbol{Z}/m\boldsymbol{Z}]^N$ is a system of coset representatives for $\boldsymbol{Z}^N/m\boldsymbol{Z}^N$; hence $\boldsymbol{Z}^N = m\boldsymbol{Z}^N + [\boldsymbol{Z}/m\boldsymbol{Z}]^N$.

A partition Λ/Λ' also induces a coset decomposition of any coset of Λ, say $\Lambda + \boldsymbol{c}$; for $\Lambda + \boldsymbol{c} = \Lambda' + [\Lambda/\Lambda'] + \boldsymbol{c}$.

A *partition chain* $\Lambda/\Lambda'/\Lambda''/\cdots$ is a sequence of lattices such that each is a sublattice of the previous one (in other words, $\Lambda \supseteq \Lambda' \supseteq \Lambda'' \supseteq \cdots$). For example, $\boldsymbol{Z}/2\boldsymbol{Z}/4\boldsymbol{Z}/\cdots$ is an infinite sequence of two-way partitions of the integers. A partition chain induces a multiterm coset decomposition chain, with a term corresponding to each partition; e.g., if $\Lambda/\Lambda'/\Lambda''$ is a partition chain, then

$$\Lambda = \Lambda'' + [\Lambda'/\Lambda''] + [\Lambda/\Lambda'],$$

meaning that every element of Λ can be expressed as an element of Λ'' plus a coset representative from $[\Lambda'/\Lambda'']$ plus a coset representative from $[\Lambda/\Lambda']$. For example, the chain $\boldsymbol{Z}/2\boldsymbol{Z}/4\boldsymbol{Z}/\cdots$ leads to the *standard binary representation* of an integer m:

$$m = a_0 + 2a_1 + 4a_2 + \cdots$$

where $a_0, a_1, a_2, \cdots \in \{0,1\}$, and a_0 specifies the coset in the partition $\boldsymbol{Z}/2\boldsymbol{Z}$, $2a_1$ specifies the coset in the partition $2\boldsymbol{Z}/4\boldsymbol{Z}$, and so forth. That is,

$$\boldsymbol{Z} = [\boldsymbol{Z}/2\boldsymbol{Z}] + [2\boldsymbol{Z}/4\boldsymbol{Z}] + [4\boldsymbol{Z}/8\boldsymbol{Z}] + \cdots.$$

For a related example with a finite chain, we can specify one of the eight cosets of $\boldsymbol{Z}/8\boldsymbol{Z}$ (one of the equivalence classes of integers modulo 8) by 3 bits (a_0, a_1, a_2), where $a_2a_1a_0$ is the standard binary representation of the coset representative $c \in \{0,1,\cdots,7\}$.

We may illustrate such a decomposition chain by a *partition tower*, as shown in Fig. 5(a). Each block in the tower represents one partition Λ/Λ' in the chain, and the input to that block is a variable which selects one of the $|\Lambda/\Lambda'|$ cosets of Λ' in that partition (or, equivalently, one of the coset representatives in $[\Lambda/\Lambda']$). The standard bi-

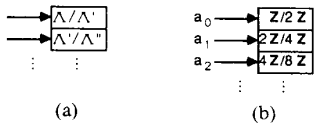


Fig. 5. Partition towers illustrating coset decomposition chains induced by lattice partition chains. (a) $\Lambda/\Lambda'/\Lambda''/\cdots$. (b) $\boldsymbol{Z}/2\boldsymbol{Z}/4\boldsymbol{Z}/8\boldsymbol{Z}/\cdots$.

nary representation is illustrated in this way in Fig. 5(b) (note that the "least significant bit" a_0 appears at the top).

C. Geometric Properties

The geometry of a real lattice Λ arises from the geometry of a real Euclidean N-space $\boldsymbol{R}^N$. The two principal geometrical parameters of Λ are the minimum squared distance $d^2_{\min}(\Lambda)$ between its points and its fundamental volume $V(\Lambda)$; these determine its fundamental coding gain $\gamma(\Lambda)$.

The *norm* $\|\boldsymbol{x}\|^2$ of a vector $\boldsymbol{x}$ in $\boldsymbol{R}^N$ is the sum of the squares of its coordinates. Norms are nonnegative and in fact nonzero unless $\boldsymbol{x}=\boldsymbol{0}$. The *squared distance* between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the norm of their difference $\|\boldsymbol{x}-\boldsymbol{y}\|^2$.

Because a lattice Λ consists of discrete points, the norms of all lattice points are an infinite set of discrete values that can be enumerated in ascending order We call this the *weight distribution* of the lattice (*theta series*, in the lattice literature). The weight distribution is also the squared distance distribution between any point in the lattice and all other points, since any point λ in Λ can be taken as the origin $\boldsymbol{0}$ by translation of Λ by λ (looking out from any point in Λ, the lattice looks the same).

The minimum nonzero norm is thus the *minimum squared distance* $d^2_{\min}(\Lambda)$ between any two points in Λ. The number of elements of Λ with this norm is the number of nearest neighbors of any lattice point (also called the *kissing number*, or *multiplicity*), and will be called here the *error coefficient* $N_0(\Lambda)$.

For example, for any N, the integer lattice $\boldsymbol{Z}^N$ has $d^2_{\min}(\boldsymbol{Z}^N)=1$. The set of all integer N-tuples of norm 1 is the set of all permutations and sign changes of the vector $(1,0,\cdots,0)$, so $N_0(\boldsymbol{Z}^N)=2N$.

Loosely, the *fundamental volume* $V(\Lambda)$ is the volume of N-space per lattice point, or the reciprocal of the number of lattice points per unit volume. More precisely, if we can partition N-space into regions of equal volume, one associated with each lattice point, then $V(\Lambda)$ is the volume of each such region. For example, it is easy to see that we may partition N-space into N-cubes of side 1, one associated with each point of $\boldsymbol{Z}^N$, so $V(\boldsymbol{Z}^N)=1$.

To treat the general case, note that $\boldsymbol{R}^N$ is itself a group under ordinary vector addition (but not a lattice), because its points are not discrete). Any real N-dimensional lattice Λ is a subgroup of $\boldsymbol{R}^N$. Thus there is a partition $\boldsymbol{R}^N/\Lambda$ of N-space into equivalence classes modulo Λ (cosets of Λ) (in our original definition of a coset of Λ, implicitly we meant a coset in the partition $\boldsymbol{R}^N/\Lambda$). Define a *fundamental region* $\mathbb{R}(\Lambda)$ as a region of N-space that contains one and only one point from each such equivalence class modulo Λ; thus $\mathbb{R}(\Lambda)$ is a system of coset representatives for the partition $\boldsymbol{R}^N/\Lambda$. Every point $\boldsymbol{x}$ in $\boldsymbol{R}^N$ is thus uniquely representable as $\boldsymbol{x}=\lambda+\boldsymbol{c}$, where $\lambda\in\Lambda$ and $\boldsymbol{c}\in\mathbb{R}(\Lambda)$, i.e., there is a coset decomposition $\boldsymbol{R}^N=\Lambda+\mathbb{R}(\Lambda)$. Geometrically, this is a tesselation of N-space by translates of fundamental regions of Λ. While there is no unique fundamental region, every fundamental region $\mathbb{R}(\Lambda)$ must have the same volume $V(\Lambda)$ (if it is measurable), since it is congruent to any other fundamental region modulo Λ; this uniquely defines the fundamental volume $V(\Lambda)$.

For example, one fundamental region of the one-dimensional integer lattice $\boldsymbol{Z}$ is the half-open interval $[0,1)=\{c: 0\le c<1\}$; another is the half-open interval $(-1/2,1/2]$. Whatever fundamental region we take, however, its volume (length) must be one. Similarly, for $\boldsymbol{Z}^N$, we may take $\mathbb{R}(\boldsymbol{Z}^N)$ as the half-open N-cube $[0,1)^N$ or as $(-1/2,1/2]^N$; again the volume $V(\boldsymbol{Z}^N)$ is one for any N.

Happily, the computation of fundamental volumes of an integer lattice Λ may be completely avoided by use of the following lemma, if we know the order $|\boldsymbol{Z}^N/\Lambda|$ of the partition $\boldsymbol{Z}^N/\Lambda$.

Lemma 1: If Λ' is a sublattice of Λ of order $|\Lambda/\Lambda'|$, then $V(\Lambda')=|\Lambda/\Lambda'|V(\Lambda)$.

Proof: Since only one of every $|\Lambda/\Lambda'|$ points in Λ is in Λ', the fundamental volume of Λ' must be $|\Lambda/\Lambda'|$ times larger than that of Λ for its fundamental regions to fill N-space. In fact, we may take a union of $|\Lambda/\Lambda'|$ fundamental regions of Λ as a fundamental region of Λ', one $(\mathbb{R}(\Lambda)+\boldsymbol{c})$ associated with each member $\boldsymbol{c}$ of a set of coset representatives for Λ/Λ', in view of the decomposition chain $\boldsymbol{R}^N=\Lambda'+[\Lambda/\Lambda']+\mathbb{R}(\Lambda)$.

Corollary: If Λ is an integer lattice, then $V(\Lambda)=|\boldsymbol{Z}^N/\Lambda|$.

Notice that Lemma 1 does not essentially depend on Λ being a lattice; as long as Λ is a union of some number $|\Lambda/\Lambda'|$ of cosets of Λ', the points of Λ are $|\Lambda/\Lambda'|$ times as dense in N-space as those of Λ'.

From the two geometrical parameters $d^2_{\min}(\Lambda)$ and $V(\Lambda)$, we define the *fundamental coding gain* $\gamma(\Lambda)$ of a lattice Λ as follows:

$$\gamma(\Lambda)\triangleq d^2_{\min}(\Lambda)/V(\Lambda)^{2/N}$$

(in the mathematical literature this is called *Hermite's parameter* and is also denoted by the symbol γ). The fundamental coding gain is a normalized measure of the density of a lattice in the following various senses.

a) It is dimensionless. Both $d^2_{\min}(\Lambda)$ and $V(\Lambda)^{2/N}$ have the dimensions of a two-dimensional volume (area). We shall often find that the most appropriate normalization of other parameters is to two dimensions.

b) The fundamental coding gain is invariant to scaling, $\gamma(r\Lambda)=\gamma(\Lambda)$, because $d^2_{\min}(r\Lambda)=r^2d^2_{\min}(\Lambda)$ and $V(r\Lambda)=r^NV(\Lambda)$.

c) More generally, the fundamental coding gain is invariant to any scaled orthogonal transformation T, $\gamma(T\Lambda)=\gamma(\Lambda)$, because $d^2_{\min}(T\Lambda)=|\det T|^{2/N}d^2_{\min}(\Lambda)$ and $V(T\Lambda)=|\det T|V(\Lambda)$, where $\det T$ is the determinant of T. Thus any version of Λ has the same fundamental coding gain.

d) The fundamental coding gain is invariant to the Cartesian product operation, $\gamma(\Lambda^M)=\gamma(\Lambda)$, because $d^2_{\min}(\Lambda^M)=d^2_{\min}(\Lambda)$ and $V(\Lambda^M)=V(\Lambda)^M$. Thus if Λ_N is an N-dimensional lattice and Λ_M is an M-dimensional lattice, then in MN-space (where they can be compared

directly) Λ_N^M has a greater or lesser density of points per unit volume than does Λ_N^M according to whether $\gamma(\Lambda_N)$ is greater or less than $\gamma(\Lambda_M)$, provided that they are scaled so that their minimum squared distances are the same. We therefore say that Λ is *denser* than Λ' if $\gamma(\Lambda) > \gamma(\Lambda')$, regardless of whether Λ and Λ' have the same dimension.

e) For any N, $\gamma(\boldsymbol{Z}^N) = 1$. An *uncoded system* may be defined as one that uses constellations based on $\boldsymbol{Z}^N$ (e.g., PAM uses constellations based on $\boldsymbol{Z}$, and narrow-sense quadrature amplitude modulation (QAM) uses constellations based on $\boldsymbol{Z}^2$). Thus the fundamental coding gain $\gamma(\Lambda)$ of an arbitrary lattice Λ may be considered to be the gain using constellations based on Λ over an uncoded system using constellations based on $\boldsymbol{Z}^N$.

f) More concretely, suppose we form a 2^n-point N-dimensional constellation by taking all points in an N-dimensional lattice Λ (or a coset of Λ) that lie within an N-sphere with radius chosen just large enough to enclose the desired number of points. The volume of the sphere must then be about $2^n V(\Lambda)$ for large n. For the integer lattice $\boldsymbol{Z}^N$, the volume of such a sphere must be about 2^n. The ratio of the radii of the two spheres is thus about $V(\Lambda)^{-1/N}$ (this dimensional argument in fact holds for constellation boundaries of any shape). If we scale Λ so that $d^2_{\min}(\Lambda) = 1$, then the minimum distance is the same as for $\boldsymbol{Z}^N$, but we achieve an average power reduction of about $V(\Lambda)^{-2/N}$, or $\gamma(\Lambda)$, by using the constellation based on Λ rather than that based on $\boldsymbol{Z}^N$. Thus $\gamma(\Lambda)$ is normalized properly to measure a power gain, and we will often give its value in decibels.

For example, $R\boldsymbol{Z}^2$ is a version of $\boldsymbol{Z}^2$ with $d^2_{\min} = 2$ (the rotation operator R always doubles norms, in any number of dimensions). The partition $\boldsymbol{Z}^2/R\boldsymbol{Z}^2$ has order 2, and thus $V(R\boldsymbol{Z}^2) = 2$. Thus we verify that $\gamma(R\boldsymbol{Z}^2) = 1$.

As an example of a denser lattice, the *Schläfli lattice* D_4 may be defined as the four-dimensional integer lattice consisting of all integer 4-tuples with an even number of odd coordinates or, equivalently, with even norm. The order of the partition $\boldsymbol{Z}^4/D_4$ is two, because $\boldsymbol{Z}^4$ is the union of D_4 and its coset $D_4 + (1,0,0,0)$ (the set of all integer 4-tuples with an odd number of odd coordinates or, equivalently, with odd norm). Thus $V(D_4) = 2$. Clearly, $d^2_{\min}(D_4) = 2$; therefore, the fundamental coding gain of D_4 is

$$\gamma(D_4) = 2/2^{2/4} = 2^{1/2}.$$

Thus D_4 is denser than $\boldsymbol{Z}$ or $\boldsymbol{Z}^4$ by a factor of $2^{1/2}$ (or 1.51 dB). The elements of norm 2 are the 24 points obtained by permutations and coordinate sign changes of the 4-tuple $(1,1,0,0)$, so the error coefficient $N_0(D_4)$ is 24.

Another way of comparing the density of D_4 to that of $\boldsymbol{Z}^4$ is the following. The lattice $R\boldsymbol{Z}^4$ is a version of $\boldsymbol{Z}^4$ with $d^2_{\min}(R\boldsymbol{Z}^4) = 2$ (since R doubles norms) and with $V(R\boldsymbol{Z}^4) = 4$ (since $\gamma(R\boldsymbol{Z}^4) = 1$). Moreover, $R\boldsymbol{Z}^4$ is a sublattice of D_4, which must be of order 2 (by Lemma 1); in fact, D_4 is the union of $R\boldsymbol{Z}^4$ (the lattice of integer 4-tuples with even norms in both pairs of coordinates) and its coset $R\boldsymbol{Z}^4 + (1,0,1,0)$ (the set of integer 4-tuples with odd norms in both pairs of coordinates). However, D_4 has the same minimum squared distance as $R\boldsymbol{Z}^4$. Thus it is possible to take a version of $\boldsymbol{Z}^4$ and insert a translate of that version into the interstices between points without reducing the minimum distance. So D_4 has twice as many points as $R\boldsymbol{Z}^4$ per unit volume, with no decrease in $d^2_{\min}$; hence D_4 is twice as dense as $R\boldsymbol{Z}^4$ in four dimensions, or $2^{1/2}$ times as dense per two dimensions, which is the normalization used in the definition of γ.

We see that $\boldsymbol{Z}^4/D_4/R\boldsymbol{Z}^4$ is a chain of two-way partitions with distances 1/2/2 (for short). However, since D_4 is a sublattice of $\boldsymbol{Z}^4$, it follows that RD_4 is a sublattice of $R\boldsymbol{Z}^4$, $R^2D_4 = 2D_4$ is a sublattice of $R^2\boldsymbol{Z}^4 = 2\boldsymbol{Z}^4$, and so forth. Hence $\boldsymbol{Z}^4/D_4/R\boldsymbol{Z}^4/RD_4/2\boldsymbol{Z}^4/2D_4/2R\boldsymbol{Z}^4/2RD_4/\cdots$ is an infinite chain of two-way partitions, with distances $1/2/2/4/4/8/8/16/\cdots$.

D. Complex Lattices and Gaussian Integers

A *complex lattice* Λ is a discrete set of points in complex Euclidean N-space $\boldsymbol{C}^N$ that forms a group under ordinary (complex) vector addition. Again, we stipulate that the only such lattices to be considered here will actually span N dimensions, so we shall feel free to call such a Λ an *N-dimensional* complex lattice.

An obvious isomorphism (written $\Lambda_r \simeq \Lambda_c$) exists between any $2N$-dimensional real lattice Λ_r and a corresponding N-dimensional complex lattice Λ_c, formed by taking each pair of coordinates of Λ_r to specify the real and imaginary parts of each coordinate of Λ_c, or vice versa. Addition of two points gives the same result in either case. Sublattices, cosets, and all such group properties carry over. Even the norm of two corresponding vectors is the same, so distances are not affected. Thus for most purposes it makes no difference whether we consider a lattice to be real or complex. For all parameters previously defined (e.g., $d^2_{\min}(\Lambda), V(\Lambda), \gamma(\Lambda)$), we may define the values for a complex lattice to be the same as those for the corresponding real lattice.

The only difference of any significance arises when we consider multiplicative operations, such as scaling, or the taking of inner products. A complex lattice Λ_c may be scaled by either a real number r or a complex number α, the latter operation involving an equal phase rotation of each coordinate of Λ_c by the phase of α (as well as a scaling of lengths by $|\alpha|$, or norms by $|\alpha|^2$). The inner product $(\boldsymbol{x}, \boldsymbol{y})$ of two real vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the sum of the products of their coordinates and must be real; the (Hermitian) inner product $(\boldsymbol{x}, \boldsymbol{y})$ of two complex vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the sum of the products of the coordinates of $\boldsymbol{x}$ with the complex conjugates of the coordinates of $\boldsymbol{y}$ and may be complex. Thus there may arise differences in definitions of orthogonality, duality, and so forth. In general, for the lattices considered in this paper, we shall prefer the complex definitions.

The simplest example of a complex lattice is the one-dimensional complex lattice $\boldsymbol{G}$ corresponding to the two-dimensional real lattice $\boldsymbol{Z}^2$. The point (a, b) in $\boldsymbol{Z}^2$ corre-

sponds to the point $a+bi$ in $\boldsymbol{G}$, where a and b may be any pair of integers. The set $\boldsymbol{G}$ is called the set of *Gaussian integers*.

The Gaussian integers $\boldsymbol{G}$ actually form a system of complex integers analogous to the ordinary real integers $\boldsymbol{Z}$. Multiplication of two elements of $\boldsymbol{G}$ (using complex arithmetic) yields another element of $\boldsymbol{G}$, which cannot be 0 unless one of the two elements is 0 (in fact, their norms multiply as real integers). Thus $\boldsymbol{G}$ is a ring and, in fact, an integral domain. Indeed, we have unique factorization in $\boldsymbol{G}$: every element of $\boldsymbol{G}$ can be expressed uniquely as a product of primes, up to units, where the units (invertible elements) are ± 1 and $\pm i$, and the primes are the elements that have no divisors other than themselves, up to units. The primes of $\boldsymbol{G}$, in order of increasing norm, are $1+i$, $2\pm i, 3, \cdots$, with norms $2, 5, 9, \cdots$. We denote the prime of least norm by $\phi \triangleq 1+i$. (Note that $|\phi|^2 = \phi\phi^* = 2$, and thus two is not a prime in $\boldsymbol{G}$.)

We may scale $\boldsymbol{G}$ by any element $g \in \boldsymbol{G}$ and obtain a sublattice $g\boldsymbol{G}$ of $\boldsymbol{G}$. By Lemma 1, the partition $\boldsymbol{G}/g\boldsymbol{G}$ must have order $|g|^2$ (the norm of g). There are thus $|g|^2$ equivalence classes of $\boldsymbol{G}$ modulo g.

For example, $\phi\boldsymbol{G}$ is a sublattice of $\boldsymbol{G}$ of order $|\phi|^2 = 2$ and, in fact, is the complex lattice corresponding to the real lattice $R\boldsymbol{Z}^2$. As with $R\boldsymbol{Z}^2$, $\phi\boldsymbol{G}$ consists of all the elements of $\boldsymbol{G}$ with even norm, its coset $\phi\boldsymbol{G}+1$ consists of all the elements of $\boldsymbol{G}$ with odd norm, and the union of $\phi\boldsymbol{G}$ and $\phi\boldsymbol{G}+1$ is $\boldsymbol{G}$ (Fig. 4 may equally well be taken to illustrate this partition of $\boldsymbol{G}$). The coset representatives $[\boldsymbol{G}/\phi\boldsymbol{G}]$ may thus be taken as $\{0,1\}$, and are isomorphic to $\boldsymbol{Z}_2 = \mathrm{GF}(2)$ using modulo ϕ arithmetic (since $2 \equiv 0 \bmod \phi$).

More generally, $\phi^\mu\boldsymbol{G}$ is a sublattice of $\boldsymbol{G}$ of order $|\phi|^{2\mu} = 2^\mu$ and, in fact, is the complex lattice corresponding to the real lattice $R^\mu\boldsymbol{Z}^2$, which is equal to $2^{\mu/2}\boldsymbol{Z}^2$ for μ even and $2^{(\mu-1)/2}R\boldsymbol{Z}^2$ for μ odd. As with $R^\mu\boldsymbol{Z}^2$, $\phi^\mu\boldsymbol{G}$ consists of all the elements of $\boldsymbol{G}$ whose norms are multiples of 2^μ, and thus $d^2_{\min}(\phi^\mu\boldsymbol{G}) = 2^\mu$. There is then an infinite chain $\boldsymbol{G}/\phi\boldsymbol{G}/\phi^2\boldsymbol{G}/\phi^3\boldsymbol{G}/\phi^4\boldsymbol{G}/\cdots$ of two-way partitions with distances $1/2/4/8/16/\cdots$, corresponding to the real chain $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2/2R\boldsymbol{Z}^2/4\boldsymbol{Z}^2/\cdots$. In analogy to the chain $\boldsymbol{Z}/2\boldsymbol{Z}/4\boldsymbol{Z}/\cdots$, this chain suggests a *complex binary representation* of a Gaussian integer g:

$$g = a_0 + \phi a_1 + \phi^2 a_2 + \cdots$$

where $a_0, a_1, a_2, \cdots \in \{0,1\}$, and a_0 specifies the coset of $\phi\boldsymbol{G}$ in the partition $\boldsymbol{G}/\phi\boldsymbol{G}$, ϕa_1 specifies the coset of $\phi^2\boldsymbol{G}$ in the partition $\phi\boldsymbol{G}/\phi^2\boldsymbol{G}$, and so forth. That is, the complex binary representation is based on the coset decomposition

$$\boldsymbol{G} = [\boldsymbol{G}/\phi\boldsymbol{G}] + [\phi\boldsymbol{G}/\phi^2\boldsymbol{G}] + [\phi^2\boldsymbol{G}/\phi^3\boldsymbol{G}] + \cdots.$$

For any lattice Λ, if λ is any lattice point and m is any integer, then $\pm m\lambda = \pm(\lambda + \lambda + \cdots + \lambda)$ is a lattice point, so $m\Lambda$ is a sublattice of Λ, and Λ (like any additive group) is a module over the ring $\boldsymbol{Z}$ of ordinary integers. However, a complex lattice Λ is not necessarily a module over the ring $\boldsymbol{G}$ of Gaussian integers (for example, the two-dimensional hexagonal lattice is not). It is so if and only if $\lambda \in \Lambda$ implies $i\lambda \in \Lambda$; for then if $g = a + bi$ is any Gaussian integer, $g\lambda = a\lambda + b(i\lambda)$ is a lattice point. Then $g\Lambda$ is a sublattice of Λ for any $g \in \boldsymbol{G}$. In particular, $i\Lambda$ is a sublattice of Λ; but since $i(i\Lambda) = -\Lambda = \Lambda$ is a sublattice of $i\Lambda$, in fact $i\Lambda = \Lambda$. When necessary, we shall call such a complex lattice a *$\boldsymbol{G}$-lattice*.

In general, multiplication of a $\boldsymbol{G}$-lattice Λ_c by the complex scalar ϕ has much the same effect as a transformation of the corresponding real lattice Λ_r by the rotation operator R. The correspondence is not exact because R includes a reflection as well as rotation and scaling, so that $R^2 = 2$, whereas $\phi^2 = 2i$. (We could have avoided this difficulty by exchanging columns in the definition of R.) However, if $\Lambda = \Lambda^*$—i.e., if $\lambda \in \Lambda$ implies $\lambda^* \in \Lambda$, where λ^* is the complex conjugate of λ—as will be true for all lattices to be considered here, then $\phi\Lambda_c \simeq R\Lambda_r$. The difference is slight, but we regard multiplication by the complex scalar ϕ as fundamentally a more natural operation than rotation by R.

E. Binary Lattices

Binary lattices have proved to be the most useful class of lattices in applications. On the one hand, this is because they are a natural extension of binary block codes and are well suited to the bit-oriented real world. On the other hand, in many cases they give the best performance, both as lattices and as the basis of trellis codes, a result which would have been harder to predict *a priori*. For instance, the densest known lattices in 1, 4, 8, 16, and 24 dimensions (among others) are binary lattices. We provide a brief introduction here; part II discusses binary lattices in more detail.

A real N-dimensional lattice Λ is a *binary lattice* if it is an integer lattice that has $2^m\boldsymbol{Z}^N$ as a sublattice for some m. The least such m is called the *2-depth* of the lattice. Thus $\boldsymbol{Z}^N/\Lambda/2^m\boldsymbol{Z}^N$ is a partition chain. It turns out that all of the binary lattices that have proved to be useful to date have 2-depth equal to one or two; we shall call such lattices mod-2 and mod-4 lattices, respectively.

A complex N-dimensional lattice Λ is a *binary lattice* if it is a Gaussian integer $\boldsymbol{G}$-lattice that has $\phi^\mu\boldsymbol{G}^N$ as a sublattice for some μ. The least such μ is called the *ϕ-depth* of the lattice. Thus $\boldsymbol{G}^N/\Lambda/\phi^\mu\boldsymbol{G}^N$ is a partition chain.

If Λ is a $2N$-dimensional real binary lattice, then the corresponding N-dimensional complex lattice is also a complex binary lattice (if it is a $\boldsymbol{G}$-lattice), and vice versa, since $2^m\boldsymbol{Z}^{2N} \simeq \phi^{2m}\boldsymbol{G}^N \subset \phi^{2m-1}\boldsymbol{G}^N$. So we may speak of the ϕ-depth of a real $2N$-dimensional binary lattice. A real $2N$-dimensional binary lattice with 2-depth m has ϕ-depth $2m$ or $2m-1$; thus the ϕ-depth is twice as fine-grained a parameter, and we shall henceforth call it simply the *depth* μ of a binary lattice. A mod-2 binary lattice thus has depth 1 or 2, and a mod-4 binary lattice has depth 3 or 4. For example, since $\boldsymbol{Z}^4/D_4/R\boldsymbol{Z}^4 \simeq \boldsymbol{G}^2/D_4/\phi\boldsymbol{G}^2$ is a partition chain, D_4 is a mod-2 binary lattice with depth $\mu = 1$.

Since the order of the partition $\mathbf{Z}^N/2^m\mathbf{Z}^N$ (resp. $\mathbf{G}^N/\phi^\mu\mathbf{G}^N$) is a power of two, the orders of $\mathbf{Z}^N/\Lambda$ and $\Lambda/2^m\mathbf{Z}^N$ (resp. $\mathbf{G}^N/\Lambda$ and $\Lambda/\phi^\mu\mathbf{G}^N$) must be powers of two, since their product is $|\mathbf{Z}^N/2^m\mathbf{Z}^N|$ (resp. $|\mathbf{G}^N/\phi^\mu\mathbf{G}^N|$). The *redundancy* $r(\Lambda)$ of a binary lattice Λ is defined as the binary logarithm of $|\mathbf{Z}^N/\Lambda|$, so that $|\mathbf{Z}^N/\Lambda| = 2^{r(\Lambda)}$. In view of the corollary to Lemma 1, the fundamental volume of a binary lattice is therefore $V(\Lambda) = 2^{r(\Lambda)}$, and the fundamental coding gain is

$$\gamma(\Lambda) = 2^{-\rho(\Lambda)} d^2_{\min}(\Lambda)$$

where $\rho(\Lambda)$ is the *normalized redundancy* (per two dimensions) of Λ, $\rho(\Lambda) = r(\Lambda)/N$, where $2N$ is the dimension of Λ as a real lattice, or N is the dimension of Λ as a complex lattice.

If we choose a constellation of (say) 2^n points from Λ, they will occupy a volume in $2N$-space approximately $2^{r(\Lambda)}$ times as large as the same number of points chosen from $\mathbf{Z}^{2N}$ would. We therefore say that the *constellation expansion factor* is $2^{r(\Lambda)}$ in $2N$ dimensions, or $2^{\rho(\Lambda)}$ in two dimensions. As previously discussed, this translates into a power cost due to constellation expansion of a factor of $2^{\rho(\Lambda)}$, or $\rho(\Lambda) \cdot 3.01$ dB. The formula for fundamental coding gain just given therefore has an interpretation as follows: the minimum squared distance gain of a factor of $d^2_{\min}(\Lambda)$ (relative to $d^2_{\min}(\mathbf{Z}^N) = 1$) is partially offset by a constellation expansion power cost of a factor of $2^{-\rho(\Lambda)}$, leaving a net coding gain of $\gamma(\Lambda) = 2^{-\rho(\Lambda)} d^2_{\min}(\Lambda)$.

The order of $\Lambda/2^m\mathbf{Z}^{2N}$ is $2^{2Nm - r(\Lambda)}$, and the order of $\Lambda/\phi^\mu\mathbf{G}^N$ is $2^{N\mu - r(\Lambda)}$. We may give $k(\Lambda) \triangleq N\mu(\Lambda) - r(\Lambda)$ the somewhat ugly but dual name of the '*informativity*' of Λ, where N is the dimension of Λ as a complex lattice; the *normalized informativity* (per two dimensions) of Λ is $\kappa(\Lambda) \triangleq k(\Lambda)/N$, where N is the dimension of Λ as a complex lattice, or $2N$ is the dimension of Λ as a real lattice. Thus the depth of a binary lattice Λ is the sum of its normalized redundancy and informativity:

$$\mu(\Lambda) = \rho(\Lambda) + \kappa(\Lambda).$$

For example, the depth of D_4 is 1, its redundancy and informativity are both equal to 1, and its normalized redundancy and informativity are both equal to 1/2.

Moreover, if a sublattice Λ' of a binary lattice Λ is also a binary lattice, then Λ/Λ' is a partition whose order is a power of two, since $\mathbf{Z}^N/\Lambda/\Lambda'/2^m\mathbf{Z}^N$ is a partition chain for some m. We define the *depth* of a partition Λ/Λ' of binary lattices as the depth of Λ', $\mu(\Lambda/\Lambda') \triangleq \mu(\Lambda')$; the *redundancy* as the redundancy of Λ, $\rho(\Lambda/\Lambda') \triangleq \rho(\Lambda)$; and the *informativity* as the informativity of Λ', $\kappa(\Lambda/\Lambda') \triangleq \kappa(\Lambda')$. It follows that the order of the partition is

$$|\Lambda/\Lambda'| = 2^{N(\mu(\Lambda') - \rho(\Lambda) - \kappa(\Lambda'))}$$

where $2N$ is the dimension of Λ as a real lattice.

F. Labelings of Partitions of Binary Lattices

If Λ and Λ' are binary lattices such that Λ' is a sublattice of Λ, then the partition Λ/Λ' has order 2^K for some integer K. Any map from binary K-tuples $\boldsymbol{a}$ to unique cosets of Λ' in this partition is called a *labeling*. The labeling may be defined by a function $\boldsymbol{c}(\boldsymbol{a})$, where $\boldsymbol{a}$ is any binary K-tuple, called a *label*, and $\boldsymbol{c}(\boldsymbol{a})$ is a coset representative of the coset of Λ' specified by $\boldsymbol{a}$. We always assume that $\boldsymbol{c}(\mathbf{0}) = \mathbf{0}$, i.e., that the zero label maps to the zero coset, namely, Λ' itself.

The following lemma shows, first, that any such partition can be broken up into a chain of K two-way partitions Λ_k/Λ_{k+1}, $0 \le k \le K-1$; second, that there is then a labeling

$$\boldsymbol{c}(\boldsymbol{a}) = \sum a_k \boldsymbol{g}_k$$

where a_k is the kth coordinate of the binary K-tuple $\boldsymbol{a}$ and $\boldsymbol{g}_k$ is an element of Λ_k but not of Λ_{k+1} such that the two vectors $\{a_k\boldsymbol{g}_k, a_k \in \{0,1\}\}$, are a system of coset representatives for the cosets of Λ_{k+1} in the two-way partition Λ_k/Λ_{k+1}. Thus the 2^K *binary linear combinations* $\{\sum a_k\boldsymbol{g}_k\}$ of the *generators* $\boldsymbol{g}_k$ are a system of coset representatives $[\Lambda_k/\Lambda_{k+1}]$ for the partition Λ_k/Λ_{k+1} (this is a special case of a general result for groups of order 2^K—binary groups—discussed in part II). In the coding context, we call such a labeling an *Ungerboeck labeling*.

Lemma 2: Let Λ and Λ' be binary lattices such that Λ' is a sublattice of Λ, and let $|\Lambda/\Lambda'| = 2^K$. Then there is a sequence of lattices $\Lambda_0 = \Lambda, \Lambda_1, \cdots, \Lambda_K = \Lambda'$ such that $\Lambda_0/\Lambda_1/\cdots/\Lambda_K$ is a lattice partition chain and each partition Λ_k/Λ_{k+1} is two-way, $0 \le k \le K-1$. Λ has the coset decomposition

$$\Lambda = \Lambda' + \left\{\sum a_k\boldsymbol{g}_k\right\}$$

where $a_k \in \{0,1\}$ and $\{a_k\boldsymbol{g}_k, a_k \in \{0,1\}\}$ is a system of coset representatives for $[\Lambda_k/\Lambda_{k+1}]$, $0 \le k \le K-1$.

Sketch of proof (by induction, from $k = K-1$ down to $k = 0$): Assume that Λ_{k+1} is a binary lattice such that $\Lambda/\Lambda_{k+1}/\Lambda'$ is a partition chain; $\Lambda_K = \Lambda'$ is certainly such a lattice. If $\Lambda_{k+1} \ne \Lambda$, then a vector $\boldsymbol{g}_k$ exists in Λ that is not in Λ_{k+1} but has order 2 mod Λ_{k+1}, i.e., such that $\boldsymbol{g}_k + \boldsymbol{g}_k \in \Lambda_{k+1}$ (see part II). Let Λ_k then be the union of Λ_{k+1} and its coset $\Lambda_{k+1} + \boldsymbol{g}_k$; Λ_k is clearly a lattice, which by construction is a sublattice of Λ and has Λ_{k+1} (and thus Λ') as a sublattice. The two vectors $\{a_k\boldsymbol{g}_k, a_k \in \{0,1\}\}$ are a system of coset representatives for the cosets of Λ_{k+1} in the two-way partition Λ_k/Λ_{k+1}. By induction, the 2^{K-k} binary linear combinations $\{\sum_{k \le j \le K-1} a_j\boldsymbol{g}_j\}$ are a system of coset representatives for Λ_k/Λ'. The induction terminates when $k = 0$.

Fig. 6 portrays an Ungerboeck labeling in two ways: as a partition tower, as in Fig. 5, and as a binary partition tree. In the tower, the first bit a_0 in the label $\boldsymbol{a}$ selects one

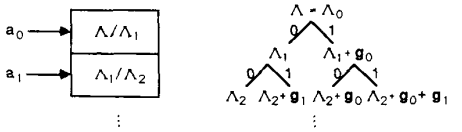


Fig. 6. Illustration of Ungerboeck labeling by partition tower and by partition tree.

of the two cosets of Λ_1 in the partition Λ/Λ_1, the second bit a_1 selects one of the two cosets of Λ_2 in the partition Λ_1/Λ_2, and so forth. In the tree, these are shown as two-way branches. Thus an Ungerboeck labeling is *nested*, in the sense that the first k bits of the labeling of the partition Λ/Λ' are a labeling of the partition Λ/Λ_k. If λ is a vector in Λ in the coset of Λ' whose label is $\boldsymbol{a}$, then λ is in a coset of Λ_k determined by the first k bits of $\boldsymbol{a}$.

Consequently, we have the lattice/coset version of the *Ungerboeck distance bound*: if λ and λ' are two different points of Λ that lie in cosets of Λ' whose labels agree in their first k bits, then $\|\lambda - \lambda'\|^2 \geq d^2_{\min}(\Lambda_k)$, because both λ and λ' are in the same coset of Λ_k (this is a special case of the partition distance lemma of part II).

G. Decomposition of Binary Lattices

This section shows that the structure of binary lattices (whether real or complex) is a generalization of that of binary block codes. Mod-2 binary lattices are essentially isomorphic to linear binary block codes, as we can see in the following lemma and proof (this is "Construction A" of Leech and Sloane [6]).

Recall that an (N, K) *linear binary block code* is any K-dimensional subspace of the N-dimensional linear vector space over $\boldsymbol{Z}_2 = \mathrm{GF}(2)$ of all binary N-tuples, where the coordinates of each of the 2^K *codewords* are regarded as elements of the binary field $\boldsymbol{Z}_2 = \mathrm{GF}(2)$. The codewords may be expressed as linear combinations $\boldsymbol{c}(\boldsymbol{a}) = \sum a_k \boldsymbol{g}_k$ of a set of K binary N-tuples $\boldsymbol{g}_k$, $1 \leq k \leq K$, called *generators*. The set is called the *generator matrix* G, so $\boldsymbol{c}(\boldsymbol{a}) = \boldsymbol{a}G$. The *minimum Hamming distance* of the code is the minimum number of nonzero coordinates in any nonzero codeword; we shall sometimes call an (N, K) code with minimum Hamming distance d_H an (N, K, d_H) code. (Unless otherwise specified, an (N, K) code will always mean a linear binary (N, K) code.)

Lemma 3: An N-dimensional real lattice Λ is a mod-2 binary lattice if and only if it is the set of all integer N-tuples that are congruent modulo 2 to one of the codewords $\boldsymbol{c}$ in a linear binary (N, K) block code C. The redundancy of Λ is $r(\Lambda) = N - K$, and its minimum squared distance is $d^2_{\min}(\Lambda) = \min[4, d_H(C)]$, where $d_H(C)$ is the minimum Hamming distance of the code C.

Proof: If Λ is a mod-2 binary lattice, then $\boldsymbol{Z}^N/\Lambda/2\boldsymbol{Z}^N$ is a partition chain, $\Lambda/2\boldsymbol{Z}^N$ has order 2^K for some integer K, and $\boldsymbol{Z}^N/\Lambda$ then has order 2^{N-K}; so the redundancy of Λ is $r(\Lambda) = N - K$. Λ is thus the union of 2^K cosets of $2\boldsymbol{Z}^N$ (in the partition $\boldsymbol{Z}^N/2\boldsymbol{Z}^N$); any such coset of $2\boldsymbol{Z}^N$ has a binary coset representative $\boldsymbol{c}$, i.e., an N-tuple of ones and zeros, and the coset $2\boldsymbol{Z}^N + \boldsymbol{c}$ consists of all integer N-tuples congruent to $\boldsymbol{c}$ modulo 2. By Lemma 2, $\Lambda = 2\boldsymbol{Z}^N + \{\sum a_k \boldsymbol{g}_k\}$, where the $\boldsymbol{g}_k$ can be taken as binary N-tuples, $1 \leq k \leq K$, and addition may be taken modulo $2\boldsymbol{Z}^N$; i.e., modulo 2. The set C of 2^K coset representatives $\{\boldsymbol{c}(\boldsymbol{a}) = \sum a_k \boldsymbol{g}_k\}$ is thus a linear binary (N, K) block code. Conversely, if C is such a code, then the set of all integer N-tuples that are congruent modulo 2 to codewords $\boldsymbol{c}$ in C is easily seen to be a lattice, with $2\boldsymbol{Z}^N$ as a sublattice (the set of all integer N-tuples congruent to $\boldsymbol{0}$ modulo 2).

If $\boldsymbol{c} \neq \boldsymbol{0}$, the minimum norm in the coset $2\boldsymbol{Z}^N + \boldsymbol{c}$ is the norm $\|\boldsymbol{c}\|^2$ of $\boldsymbol{c}$ itself, which is also the Hamming weight of $\boldsymbol{c}$. The minimum such weight is the minimum Hamming distance of the code, $d_H(C)$. If $\boldsymbol{c} = \boldsymbol{0}$, the minimum nonzero norm in $2\boldsymbol{Z}^N$ itself is four. Thus $d^2_{\min}(\Lambda) = \min[4, d_H(C)]$.

Lemma 3 says that any mod-2 lattice Λ may be constructed as a coset code $\mathbb{C}(\boldsymbol{Z}^N/2\boldsymbol{Z}^N; C)$, where C is a linear binary (N, K) block code. We can express Λ by the *code formula* (coset decomposition) $\Lambda = 2\boldsymbol{Z}^N + C$. This explicitly exhibits Λ as the union of 2^K cosets of $2\boldsymbol{Z}^N$. The labeling $\boldsymbol{c}(\boldsymbol{a}) = \sum a_k \boldsymbol{g}_k = \boldsymbol{a}G \pmod 2$ is *linear* in the sense that $\boldsymbol{c}(\boldsymbol{a} \oplus \boldsymbol{a}') = \boldsymbol{c}(\boldsymbol{a}) \oplus \boldsymbol{c}(\boldsymbol{a}')$. This picture of a mod-2 lattice as a coset code is illustrated in Fig. 7.

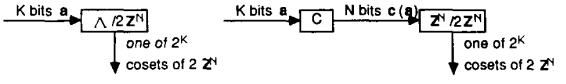


Fig. 7. Illustration of mod-2 binary lattice Λ as union of 2^K cosets of $2\boldsymbol{Z}^N$, each coset corresponding to codeword $\boldsymbol{c}(\boldsymbol{a})$ in linear binary (N, K) code C.

For example, the mod-2 lattices $\boldsymbol{Z}^2$, $R\boldsymbol{Z}^2$, and $2\boldsymbol{Z}^2$ correspond to the $(2,2,1)$, $(2,1,2)$, and $(2,0,\infty)$ binary codes, respectively, and consequently have minimum squared distances 1, 2, and 4. The mod-2 lattices $\boldsymbol{Z}^4$, D_4, $R\boldsymbol{Z}^4$, RD_4 and $2\boldsymbol{Z}^4$ correspond to $(4,4,1)$, $(4,3,2)$, $(4,2,2) = (2,1,2)^2$, $(4,1,4)$, and $(4,0,\infty)$ binary codes, with distances 1, 2, 2, 4, and 4.

In general, the lattice corresponding to any single-parity-check $(N, N-1, 2)$ code is a mod-2 lattice, the so-called "*checkerboard lattice*" D_N, with minimum squared distance 2 and redundancy 1, a sublattice of $\boldsymbol{Z}^N$ of order 2. The lattice D_2 is the lattice $R\boldsymbol{Z}^2 \simeq \phi\boldsymbol{G}$.

Mod-2 lattices are limited to minimum squared distances of four or less because N-tuples in $2\boldsymbol{Z}^N$ such as $(2,0,\cdots,0)$ are lattice points. Binary block codes with Hamming distance 4 are therefore of special interest. The *Gosset lattice* E_8 is the mod-2 lattice corresponding to the $(8,4,4)$ Reed–Muller code; E_8 thus has $d^2_{\min}(E_8) = 4$, $r(E_8) = 4$, $\rho(E_8) = 1$, and thus $\gamma(E_8) = 2^{-1} \cdot 4 = 2$ (3.01 dB). (E_8 has many remarkable properties; it is perhaps the second most important lattice in lattice theory.)

If Λ' is a sublattice of Λ, where both are mod-2 lattices, it is easy to see that the code C' associated with Λ' must be a subcode of the code C associated with Λ.

In general, if Λ is a mod-4 lattice, then $4\boldsymbol{Z}^N$ is a sublattice, and Lemma 2 allows us to write

$$\Lambda = 4\boldsymbol{Z}^N + \left\{\sum a_k \boldsymbol{g}_k\right\}$$

where each $\boldsymbol{g}_k$ may be taken as a coset representative of $4\boldsymbol{Z}^N$, i.e., as an N-tuple of integers modulo 4, but where the label $\boldsymbol{a}$ is still a $\{0,1\}$-valued integer K-tuple. Addition may be taken modulo 4. This exhibits Λ as a union of 2^K

cosets of $4\mathbf{Z}^N$, where the coset representatives are $\mathbf{c}(\mathbf{a}) = \sum a_k \mathbf{g}_k \pmod 4$. If we take the coordinates of $\mathbf{c}(\mathbf{a})$ to be from the set $\{0, \pm 1, 2\}$, then $\mathbf{c}(\mathbf{a})$ is also a *coset leader* of its coset, i.e., an element of minimum norm.

For a further refinement, let Λ_e be the set of all points in Λ whose coordinates are all even. Then Λ_e is a lattice, a sublattice of Λ, with $4\mathbf{Z}^N$ as a sublattice, so $\Lambda/\Lambda_e/4\mathbf{Z}^N$ is a partition chain, and there is a coset decomposition of the form $\Lambda = 4\mathbf{Z}^N + [\Lambda_e/4\mathbf{Z}^N] + [\Lambda/\Lambda_e]$. The lattice Λ_e is clearly a mod-2 lattice scaled by a factor of 2; consequently, $\Lambda_e = 4\mathbf{Z}^N + 2C$, where C is a binary (N, K') code for some K', by Lemma 3; in other words, the coset representatives $[\Lambda_e/4\mathbf{Z}^N]$ may be taken as $2C$, or $\{2\sum a_k \mathbf{g}_k\}$, where the $\mathbf{g}_k$ constitute a set of K' binary generators for the code C. Thus we may take K' of the generators to be $2\mathbf{g}_k$, $1 \le k \le K'$, and we may write

$$\Lambda = 4\mathbf{Z}^N + 2C + \left\{\sum a_k \mathbf{g}_k\right\}$$

where the $\mathbf{g}_k$, $K'+1 \le k \le K$, are N-tuples of integers modulo 4 that are not all even, such that $\{\sum a_k \mathbf{g}_k\}$ is a system of coset representatives $[\Lambda/\Lambda_e]$. The generators $\{2\mathbf{g}_k,\ K'+1 \le k \le K\}$ generate a lattice Λ' that is a sublattice of Λ_e, whose elements are congruent to $2\mathbf{c}'$ modulo 4, where $\mathbf{c}'$ is a codeword in a binary $(N, K-K')$ block code C' that is a subcode of the code C, so $K - K' \le K'$.

It is sometimes possible to find a set of generators $\{\mathbf{g}_k,\ K'+1 \le k \le K\}$ for Λ/Λ_e such that each $\mathbf{g}_k$ is an N-tuple of ones and zeros and such that the "carries" (the twos-coefficients in the vector sum) of any sum $\mathbf{g}_k + \mathbf{g}_{k'}$ are a codeword in C; then we say that Λ is *decomposable*. The coset decomposition then becomes the code formula $\Lambda = 4\mathbf{Z}^N + 2C_1 + C_0$, where C_0 is a subcode of C_1. This means that Λ consists of all integer N-tuples whose coordinate ones-coefficients in the standard binary representation form a binary N-tuple $\mathbf{a}_0$ that is a codeword in C_0, and whose coordinate twos-coefficients form a binary N-tuple $\mathbf{a}_1$ that is a codeword in C_1 (this is the "coordinate array" idea of Leech and Sloane [6]).

The minimum squared distance of a mod-4 decomposable real lattice Λ is

$$d^2_{\min}(\Lambda) = \min\left[16, 4d_H(C_1), d_H(C_0)\right]$$

because, on the one hand, there are N-tuples in $4\mathbf{Z}^N$ of norm 16, in $2C_1$ of norm $4d_H(C_1)$ and in C_0 of norm $d_H(C_0)$; on the other hand, if $\lambda \equiv 2\mathbf{a}_1 + \mathbf{a}_0 \bmod 4$ and $\mathbf{a}_0 \ne \mathbf{0}$, then $\|\lambda\|^2 \ge d_H(C_0)$; if $\mathbf{a}_0 = \mathbf{0}$ but $\mathbf{a}_1 \ne \mathbf{0}$, then $\|\lambda\|^2 \ge 4d_H(C_1)$; finally, if $\mathbf{a}_0 = \mathbf{a}_1 = \mathbf{0}$, then $\|\lambda\|^2 \ge 16$ (if $\lambda \ne \mathbf{0}$). This suggests that we shall want to choose codes C_1 and C_0 for which the Hamming distances are in the ratio 1:4.

Most of the mod-4 binary lattices useful in practice are decomposable. For example, the lattice RE_8 is a version of the Gosset lattice which is mod-4 and decomposable, with the code formula $RE_8 = 4\mathbf{Z}^8 + 2(8,7,2) + (8,1,8)$. The standard binary representation of an integer modulo 4 is a more elementary example of a decomposition of this type, with the code formula $\mathbf{Z} = 4\mathbf{Z} + 2(1,1) + (1,1)$; this simply means that every integer modulo 4 can be expressed as a binary linear combination of the two generators 2 and 1, where the combination can be specified by a 2-bit label.

The simplest example of an indecomposable mod-4 lattice is the two-dimensional lattice $\Lambda = \{(x_1, x_2)\colon x_1 + x_2 \equiv 0 \bmod 4\}$, for which C is the (2,1) code with generator (1,1), so $2C$ has generator $\mathbf{g}_1 = (2,2)$, but $\mathbf{g}_2$ must be taken as $(1,-1)$ or $(-1,1)$; however, this lattice can be made decomposable by inverting the sign of one coordinate. The Leech lattice Λ_{24} is an example of a fundamentally indecomposable mod-4 lattice.

Fig. 8(a) illustrates the coset decomposition $\Lambda/\Lambda_e/4\mathbf{Z}^N$ of a mod-4 binary lattice that is used above. The K'-bit label $\mathbf{a}_1$ specifies a coset of $4\mathbf{Z}^N$ in the partition $\Lambda_e/4\mathbf{Z}^N$, and the $(K-K')$-bit label $\mathbf{a}_0$ specifies a coset of Λ_e in the partition Λ/Λ_e. Altogether, therefore, the K-bit label $(\mathbf{a}_0, \mathbf{a}_1)$ specifies a coset of $4\mathbf{Z}^N$ in the partition $\Lambda/4\mathbf{Z}^N$. Fig. 8(b) illustrates the same decomposition when Λ is a decomposable mod-4 binary lattice with code formula $\Lambda = 4\mathbf{Z}^N + 2C_1 + C_0$; the form is then that of a coset code $\mathbb{C}(\mathbf{Z}^N/2\mathbf{Z}^N/4\mathbf{Z}^N; C_0, C_1)$. (If C_1 is the (N,N) code, then $\Lambda_e = 2\mathbf{Z}^N$, Λ is actually a mod-2 lattice, and Fig. 8 reduces to Fig. 7.)

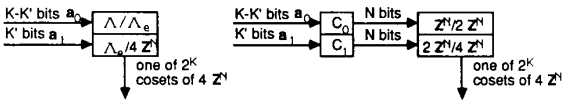


Fig. 8. Illustration of mod-4 binary lattice Λ as union of 2^K cosets of $4\mathbf{Z}^N$, each coset corresponding to binary label $(\mathbf{a}_0, \mathbf{a}_1)$. (a) General case. (b) Case where Λ is decomposable.

It should be clear how to extend Fig. 8 to real binary lattices of any 2-depth m, using a partition chain $\Lambda/\Lambda_e/\cdots/2^m\mathbf{Z}^N$ of sublattices of Λ consisting of all elements of Λ whose coordinates are in $2\mathbf{Z}^N, 4\mathbf{Z}^N$, and so forth. The decomposition of complex binary lattices is similar. Mod-2 complex binary $\mathbf{G}$-lattices are always decomposable, with a code formula of the form $\Lambda = \phi^2\mathbf{G}^N + \phi C_1 + C_0$, as we can see from the following lemma. Complex binary lattices with depths of three or more are not necessarily decomposable, but the ones that we are concerned with generally are.

Lemma 4: An N-dimensional complex $\mathbf{G}$-lattice Λ is a mod-2 binary lattice if and only if it is the set of all Gaussian integer N-tuples that are congruent modulo ϕ^2 to an N-tuple of the form $\phi\mathbf{c}_1 + \mathbf{c}_0$, where $\mathbf{c}_1$ is a codeword in a binary (N, K) code C_1, and $\mathbf{c}_0$ is a codeword in a binary $(N, J-K)$ code C_0 which is a subcode of C_1. The redundancy of Λ is $r(\Lambda) = 2N - J$, and its minimum squared distance is

$$d^2_{\min}(\Lambda) = \min\left[4, 2d_H(C_1), d_H(C_0)\right].$$

Proof: If Λ is a mod-2 binary lattice, then $\mathbf{G}^N/\Lambda/2\mathbf{G}^N = \phi^2\mathbf{G}^N$ is a partition chain, and $\Lambda/2\mathbf{G}^N$ has order 2^J for some integer J. Since $\mathbf{G}^N/2\mathbf{G}^N$ has order 2^{2N}, $\mathbf{G}^N/\Lambda$ has order 2^{2N-J}, so the redundancy of Λ is $r(\Lambda) = 2N - J$. Λ is the union of 2^J cosets of $2\mathbf{G}^N = \phi^2\mathbf{G}^N$, and we may take the coordinates of the coset representative of

any such coset to be of the form $\phi \boldsymbol{c}_1 + \boldsymbol{c}_0$, where $\boldsymbol{c}_1$ and $\boldsymbol{c}_0$ are binary N-tuples. Sums of coset representatives may be taken mod $2\boldsymbol{G}^N$ (or modulo 2).

Consider the set Λ_ϕ of all points in Λ whose coordinates are all multiples of ϕ. Then $\Lambda/\Lambda_\phi/\phi^2\boldsymbol{G}^N$ is a partition chain, and there is a coset decomposition of the form $\Lambda = \phi^2\boldsymbol{G}^N + [\Lambda_\phi/\phi^2\boldsymbol{G}^N] + [\Lambda/\Lambda_\phi]$. The generators of the lattice Λ_ϕ modulo $\phi^2\boldsymbol{G}^N$ may be taken as $\phi\boldsymbol{g}_k$, $1 \le k \le K$, for some K, where each $\boldsymbol{g}_k$ is a binary N-tuple; consequently, $\Lambda_\phi = \phi^2\boldsymbol{G}^N + \phi C_1$, where C_1 is the binary (N, K) code generated by these $\boldsymbol{g}_k$. Similarly, the generators of the lattice Λ modulo Λ_ϕ may be taken as $\boldsymbol{g}_k$, $K+1 \le k \le J$, where each $\boldsymbol{g}_k$ is a binary N-tuple; consequently, $\Lambda = \Lambda_\phi + C_0$, where C_0 is the binary $(N, J-K)$ code generated by the $\boldsymbol{g}_k$, $K+1 \le k \le J$. Since Λ is a $\boldsymbol{G}$-lattice, $\lambda \in \Lambda$ implies $\phi\lambda \in \Lambda$. Therefore, C_0 must be a subcode of C_1.

The minimum distance expression arises from the fact that there are N-tuples of norm 4 in $\phi^2\boldsymbol{G}^N$, of norm $2d_H(C_1)$ in ϕC_1, and of norm $d_H(C_0)$ in C_0; conversely, if $\lambda \equiv \phi\boldsymbol{c}_1 + \boldsymbol{c}_0 \bmod \phi^2$ and $\boldsymbol{c}_0 \neq \boldsymbol{0}$, then $\|\lambda\|^2 \ge d_H(C_0)$; if $\boldsymbol{c}_0 = \boldsymbol{0}$ but $\boldsymbol{c}_1 \neq \boldsymbol{0}$, then $\|\lambda\|^2 \ge 2d_H(C_1)$; finally, if $\boldsymbol{c}_0 = \boldsymbol{c}_1 = \boldsymbol{0}$, then $\|\lambda\|^2 \ge 4$ (if $\lambda \neq \boldsymbol{0}$).

For the converse, let Λ be the set of all Gaussian integer N-tuples $\lambda \equiv \phi\boldsymbol{c}_1 + \boldsymbol{c}_0 \bmod \phi^2$. To show that Λ is a $\boldsymbol{G}$-lattice, we must show that if $\lambda_1, \lambda_2 \in \Lambda$, then $\lambda_1 + \lambda_2 \in \Lambda$, and also if $\lambda \in \Lambda$, then $-\lambda \in \Lambda$ and $i\lambda \in \Lambda$. The first two propositions follow immediately from $2\lambda \equiv \boldsymbol{0} \bmod \phi^2$. The third follows from $i\lambda = \phi\lambda - \lambda \in \Lambda$, where the fact that $\phi\lambda \in \Lambda$ depends on C_0 being a subcode of C_1. Λ is a Gaussian integer lattice with $2\boldsymbol{G}^N$ as a sublattice (by construction).

A mod-2 complex $\boldsymbol{G}$-lattice Λ has depth 1 if and only if the code C_1 of Lemma 4 is the (N, N) code; then its code formula can be simplified to $\Lambda = \phi\boldsymbol{G}^N + C_0$. Otherwise, it has depth 2.

For example, as a complex lattice, the mod-2 depth-1 Schläfli lattice D_4 is decomposable with the code formula

$$D_4 = \phi\boldsymbol{G}^2 + (2,1,2).$$

Thus $\mu(D_4) = 1$, $r(D_4) = 1$, $\rho(D_4) = 1/2$, and $d^2_{\min}(D_4) = 2$, in agreement with the values obtained earlier for D_4 as a real lattice. As a complex lattice, the mod-2 depth-2 Gosset lattice E_8 is decomposable with the code formula

$$E_8 = \phi^2\boldsymbol{G}^4 + \phi(4,3,2) + (4,1,4)$$

(see part II). Thus $\mu(E_8) = 2$, $r(E_8) = 4$, $\rho(E_8) = 1$, and $d^2_{\min}(E_8) = 4$, in agreement with the values obtained earlier for E_8 as a real lattice. The complex binary representation of a Gaussian integer mod ϕ^2 is a more elementary example of a decomposition of this type.

As in the real case, complex binary lattices that are not mod-2 are not necessarily decomposable, i.e., expressible purely in terms of binary codes. However, the lattices that are useful in applications generally are as follows. Let $C_{\mu-1}/C_{\mu-2}/\cdots/C_0$ be a partition chain of binary (N, K_j) codes, $0 \le j \le \mu - 1$, i.e., C_j is a subcode of C_{j+1}. Then let Λ be a complex binary lattice whose elements are the set of Gaussian integer N-tuples λ that are congruent to $\phi^{\mu-1}\boldsymbol{c}_{\mu-1} + \cdots + \boldsymbol{c}_0$ modulo ϕ^μ, where $\boldsymbol{c}_j$ is a codeword in the code C_j, i.e., the coefficients of ϕ^j in the complex binary representation of λ are codewords in C_j, $0 \le j \le \mu - 1$. We represent this by the (complex) code formula

$$\Lambda = \phi^\mu\boldsymbol{G}^N + \phi^{\mu-1}C_{\mu-1} + \cdots + C_0.$$

This is the coordinate array idea again, but using the complex binary representation rather than the standard binary representation.

Some of the principal properties of such a complex decomposable lattice are as follows.

a) Λ is a $\boldsymbol{G}$-lattice, because if $\lambda \in \Lambda$, then $\phi\lambda \in \Lambda$, in view of the subcode structure of the code formula; thus $i\lambda = (\phi - 1)\lambda \in \Lambda$.

b) Λ has depth μ (assuming that $K_{\mu-1} < N$).

c) The order of the partition $\Lambda/\phi^\mu\boldsymbol{G}^N$ is the product of the 2^{K_j}; the informativity of Λ is $k(\Lambda) = \Sigma K_j$; the redundancy of Λ is $r(\Lambda) = N\mu - \Sigma K_j = \Sigma(N - K_j)$; and the normalized redundancy of Λ is $\rho(\Lambda) = \mu - \Sigma K_j/N = \Sigma(1 - K_j/N)$.

d) The minimum squared distance of Λ is

$$d^2_{\min}(\Lambda) = \min\left[2^\mu, 2^{\mu-1}d_H(C_{\mu-1}), \cdots, d_H(C_0)\right]$$

because, on the one hand, we can exhibit N-tuples with all of these norms by appropriate choice of codewords; on the other hand, if $j(\lambda)$ is the smallest index such that $\boldsymbol{c}_j \neq \boldsymbol{0}$ in the above congruence, then λ has at least $d_H(C_j)$ coordinates with norm at least 2^j. This suggests that we shall want to choose a code partition chain $C_{\mu-1}/C_{\mu-2}/\cdots/C_0$ for which the Hamming distances are in the ratio $2/4/\cdots$.

A decomposable complex lattice of depth μ may be depicted as a coset code $\mathbb{C}(\boldsymbol{G}^N/\phi\boldsymbol{G}^N/\cdots/\phi^\mu\boldsymbol{G}^N; C_0, \cdots, C_{\mu-1})$, analogously to Fig. 8(b), where the coset of $\phi^j\boldsymbol{G}^N$ in the partition $\phi^j\boldsymbol{G}^N/\phi^{j+1}\boldsymbol{G}^N$ is determined by a codeword from C_j, $0 \le j \le \mu - 1$.

H. Dual Lattices

If Λ is a binary lattice of depth μ, either $2N$-dimensional real or N-dimensional complex, we define its *dual lattice* $\Lambda^\perp$ as the set of all Gaussian integer N-tuples $\boldsymbol{y}$ that are orthogonal to all vectors $\boldsymbol{x} \in \Lambda$ modulo ϕ^μ, where we use the complex inner product $(\boldsymbol{x}, \boldsymbol{y})$. That is, $(\boldsymbol{x}, \boldsymbol{y})$ is a Gaussian integer in the lattice $\phi^\mu\boldsymbol{G}$, or a multiple of ϕ^μ. For example, it is easy to verify that D_4 is self-dual, using the fact that D_4 is the set of all pairs of Gaussian integers that are both even (in $\phi\boldsymbol{G}$) or both odd (in $\phi\boldsymbol{G} + 1$).

If Λ is a binary lattice of depth μ, we may regard $\boldsymbol{G}^N/\Lambda/\phi^\mu\boldsymbol{G}^N$ and $\boldsymbol{G}^N/\Lambda^\perp/\phi^\mu\boldsymbol{G}^N$ as *dual partition chains*, since $\phi^\mu\boldsymbol{G}^N$ is the lattice of all vectors orthogonal to all vectors in $\boldsymbol{G}^N$ modulo ϕ^μ, and vice versa. It is straightforward to verify that the order of $\boldsymbol{G}^N/\Lambda$ is the same as the order of $\Lambda^\perp/\phi^\mu\boldsymbol{G}^N$, which implies that the order of $\boldsymbol{G}^N/\Lambda^\perp$ is the same as the order of $\Lambda/\phi^\mu\boldsymbol{G}^N$. Therefore, the redundancy of $\Lambda^\perp$ is the informativity of Λ, and vice versa. Moreover, if Λ/Λ' is a partition of binary lattices of

depth μ, then $\boldsymbol{G}^N/\Lambda/\Lambda'/\phi^\mu\boldsymbol{G}^N$ and $\boldsymbol{G}^N/\Lambda'^\perp/\Lambda^\perp/\phi^\mu\boldsymbol{G}^N$ are dual partition chains, and $\Lambda'^\perp/\Lambda^\perp$ is a partition of binary lattices with the same order and depth as Λ/Λ'.

If Λ is a decomposable N-dimensional complex binary lattice of depth μ, with code formula

$$\Lambda = \phi^\mu\boldsymbol{G}^N + \phi^{\mu-1}C_{\mu-1} + \cdots + C_0,$$

then its dual lattice $\Lambda^\perp$ is a decomposable complex binary lattice of depth μ, with code formula

$$\Lambda^\perp = \phi^\mu\boldsymbol{G}^N + \phi^{\mu-1}C_0^\perp + \cdots + C_{\mu-1}^\perp$$

where $C_j^\perp$ is the dual code to C_j. (Recall that the dual code to a linear binary (N, K) block code is an $(N, N-K)$ code and that, if C' is a subcode of C, then $C^\perp$ is a subcode of $(C')^\perp$.) This proposition may be verified by noting that $C_0^\perp/\cdots/C_{\mu-1}^\perp$ is a code partition chain, that every generator $\phi^{\mu-j-1}\boldsymbol{g}_j^\perp$ of $\Lambda^\perp/\phi^\mu\boldsymbol{G}^N$ is orthogonal mod ϕ^μ to every generator $\phi^{j'}\boldsymbol{g}_{j'}$ of $\Lambda/\phi^\mu\boldsymbol{G}^N$, and that the dimensions are such that the informativity of $\Lambda^\perp$ is equal to the redundancy of Λ and vice versa.

For example, the Schläfli lattice D_4 is self-dual, because its complex code formula is $D_4 = \phi\boldsymbol{G}^2 + (2,1,2)$, and the $(2,1,2)$ code is self-dual. The Gosset lattice E_8 is self-dual, because its complex code formula is $E_8 = \phi^2\boldsymbol{G}^4 + \phi(4,3,2) + (4,1,4)$, and the $(4,3,2)$ and $(4,1,4)$ codes are duals.

An alternative definition of the dual of an N-dimensional real lattice Λ with 2-depth m is the lattice $\Lambda^\perp$ consisting of all integer N-tuples orthogonal to all N-tuples in Λ modulo 2^m, using the real inner product. For the lattices that we will be considering, this definition coincides with the definition given above when the depth (ϕ-depth) μ of Λ is even, so that $\mu = 2m$; when μ is odd, $\mu = 2m-1$, the dual lattice under this definition will be a version of the dual lattice as defined earlier, rotated by the rotation operator R. When a real lattice is decomposable, the code formula for its dual has the same relation to its code formula as is given in the complex case above. For example, the Gosset lattice E_8 has depth 2 and has real code formula $E_8 = 2\boldsymbol{Z}^8 + (8,4,4)$; its dual under this alternative definition is still itself, because the $(8,4,4)$ code is self-dual. The Schläfli lattice D_4 has depth 1 and has real code formula $D_4 = 2\boldsymbol{Z}^4 + (4,3,2)$; its dual under this alternative definition is RD_4, which has code formula $RD_4 = 2\boldsymbol{Z}^4 + (4,1,4)$, since the $(4,3,2)$ and $(4,1,4)$ codes are duals.

III. Useful Lattices and Their Partitions

In this section we list the lattices that have proved useful in applications. These are primarily the sequence of *Barnes-Wall lattices* and their *principal sublattices*, all of which are closely interrelated. Another close relative is the *Leech lattice*, probably the most important lattice in lattice theory. We give the principal properties of these lattices, including their geometrical parameters and their partition properties. For further details, see part II.

The Barnes–Wall lattices are a family of 2^n-dimensional binary lattices. The nth member of the family, which we denote by $\Lambda(0,n)$, may be regarded as either a 2^n-dimensional complex $\boldsymbol{G}$-lattice or a 2^{n+1}-dimensional real lattice. The first few members of the family are $\Lambda(0,0) = \boldsymbol{G} \simeq \boldsymbol{Z}^2$ (the Gaussian integer lattice or the two-dimensional real integer lattice), $\Lambda(0,1) = D_4$ (the Schläfli lattice), $\Lambda(0,2) = E_8$ (the Gosset lattice), $\Lambda(0,3) = \Lambda_{16}$ (the 16-dimensional Barnes–Wall lattice), and $\Lambda(0,4) = \Lambda_{32}$ (the 32-dimensional Barnes–Wall lattice). These are the densest lattices known in 4, 8, and 16 dimensions (and, until recently, 32) [7].

The Barnes–Wall lattices are decomposable, with code formulas that involve the family of Reed–Muller codes. Recall that the *Reed-Muller code* $\mathrm{RM}(r,n)$, $0 \le n$, $0 \le r \le n$, is a code of length $N = 2^n$, minimum distance $d_H = 2^{n-r}$, and with $K(r,n) = \sum_{0 \le j \le r} C_{nj}$ information bits, where C_{nj} is the combinatorial coefficient $(n!)/[(j!)((n-j)!)]$; further, the Reed–Muller codes of a given length are nested, in the sense that $\mathrm{RM}(n,n)/\mathrm{RM}(n-1,n)/\cdots/\mathrm{RM}(0,n)$ is a code partition chain.

The Barnes–Wall lattice $\Lambda(0,n)$ has depth μ equal to n and may be defined as the complex $\boldsymbol{G}$-lattice that has the code formula

$$\Lambda(0,n) = \phi^n\boldsymbol{G}^N + \phi^{n-1}\mathrm{RM}(n-1,n) + \cdots + \mathrm{RM}(0,n),$$

where $N = 2^n$. For example, as we have already seen, $\Lambda(0,0) = \boldsymbol{G}$, $\Lambda(0,1) = D_4 = \phi\boldsymbol{G}^2 + (2,1,2)$, $\Lambda(0,2) = E_8 = \phi^2\boldsymbol{G}^4 + \phi(4,3,2) + (4,1,4)$, and so forth. Thus the complex code formula involves all Reed–Muller codes of length $N = 2^n$ (this construction is due to Cusack [19]). There are similar real code formulas involving alternate Reed–Muller codes of length $2N$, as we shall tabulate below.

From the properties of decomposable binary lattices and Reed–Muller codes, the redundancy and informativity of $\Lambda(0,n)$ are both equal to $n2^{n-1} = nN/2$, the normalized redundancy and informativity are both equal to $n/2$, and the minimum squared distance $d_{\min}^2$ is equal to 2^n. Consequently, the fundamental coding gain is $\gamma = 2^{n/2}$. Because $\mathrm{RM}(n-r-1,n)$ and $\mathrm{RM}(r,n)$ are dual codes, the dual of $\Lambda(0,n)$ has the same code formula as $\Lambda(0,n)$ itself, and $\Lambda(0,n)$ is self-dual for any n.

The principal sublattices of the Barnes–Wall lattices are a family of lattices $\Lambda(r,n)$, $0 \le n$, $0 \le r \le n$, which may be defined as decomposable 2^n-dimensional complex $\boldsymbol{G}$-lattices of depth $\mu = n-r$ with the code formulas

$$\Lambda(r,n) = \phi^{n-r}\boldsymbol{G}^N + \phi^{n-r-1}\mathrm{RM}(n-1,n) + \cdots + \mathrm{RM}(r,n)$$

where $N = 2^n$. Thus $\Lambda(n,n)$ is $\boldsymbol{G}^N$, or the integer lattice $\boldsymbol{Z}^{2N}$, and $\Lambda(0,n)$ is the Barnes–Wall lattice as previously defined. The lattice $\Lambda(n-1,n)$, $n \ge 1$, is the "checkerboard lattice" D_N, with code formula $D_N = \phi\boldsymbol{G}^N + \mathrm{RM}(n-1,n) = \phi\boldsymbol{G}^N + (N, N-1, 2)$, where $N = 2^n$ and $(N, N-1, 2)$ is the single-parity-check code of length N.

In view of the general formula for the dual of a decomposable complex $\boldsymbol{G}$-lattice and the duality properties of Reed–Muller codes, the duals of the principal sublattices are the decomposable 2^n-dimensional complex $\boldsymbol{G}$-lattices

TABLE I
USEFUL BINARY LATTICES

(r,n)	Λ	$2N$	μ	Real Code Formula	Complex Code Formula
(0,0)	$\boldsymbol{Z}^2$	2	0	$\boldsymbol{Z}^2$	$\boldsymbol{G}$
(1,1)	$\boldsymbol{Z}^4$	4	0	$\boldsymbol{Z}^4$	$\boldsymbol{G}^2$
(0,1)	D_4	4	1	$2\boldsymbol{Z}^4+(4,3,2)$	$\phi\boldsymbol{G}^2+(2,1,2)$
(2,2)	$\boldsymbol{Z}^8$	8	0	$\boldsymbol{Z}^8$	$\boldsymbol{G}^4$
(1,2)	D_8	8	1	$2\boldsymbol{Z}^8+(8,7,2)$	$\phi\boldsymbol{G}^4+(4,3,2)$
(0,2)	E_8	8	2	$2\boldsymbol{Z}^8+(8,4,4)$	$\phi^2\boldsymbol{G}^4+\phi(4,3,2)+(4,1,4)$
(3,3)	$\boldsymbol{Z}^{16}$	16	0	$\boldsymbol{Z}^{16}$	$\boldsymbol{G}^8$
(2,3)	D_{16}	16	1	$2\boldsymbol{Z}^{16}+(16,15,2)$	$\phi\boldsymbol{G}^8+(8,7,2)$
(1,3)	H_{16}	16	2	$2\boldsymbol{Z}^{16}+(16,11,4)$	$\phi^2\boldsymbol{G}^8+\phi(8,7,2)+(8,4,4)$
(0,3)	Λ_{16}	16	3	$4\boldsymbol{Z}^{16}+2(16,15,2)+(16,5,8)$	$\phi^3\boldsymbol{G}^8+\phi^2(8,7,2)+\phi(8,4,4)+(8,1,8)$
(4,4)	$\boldsymbol{Z}^{32}$	32	0	$\boldsymbol{Z}^{32}$	$\boldsymbol{G}^{16}$
(3,4)	D_{32}	32	1	$2\boldsymbol{Z}^{32}+(32,31,2)$	$\phi\boldsymbol{G}^{16}+(16,15,2)$
(2,4)	X_{32}	32	2	$2\boldsymbol{Z}^{32}+(32,26,4)$	$\phi^2\boldsymbol{G}^{16}+\phi(16,15,2)+(16,11,4)$
(1,4)	H_{32}	32	3	$4\boldsymbol{Z}^{32}+2(32,31,2)+(32,16,8)$	$\phi^3\boldsymbol{G}^{16}+\phi^2(16,15,2)+\phi(16,11,4)+(16,5,8)$
(0,4)	Λ_{32}	32	4	$4\boldsymbol{Z}^{32}+2(32,26,4)+(32,6,16)$	$\phi^4\boldsymbol{G}^{16}+\phi^3(16,15,2)+\phi^2(16,11,4)+\phi(16,5,8)+(16,1,16)$
—	$\boldsymbol{Z}^{24}$	24	0	$\boldsymbol{Z}^{24}$	$\boldsymbol{G}^{12}$
—	D_{24}	24	1	$2\boldsymbol{Z}^{24}+(24,23,2)$	$\phi\boldsymbol{G}^{12}+(12,11,2)$
—	X_{24}	24	2	$2\boldsymbol{Z}^{24}+(24,18,4)$	$\phi^2\boldsymbol{G}^{12}+\phi(12,11,2)+(12,7,4)$
—	H_{24}	24	3	$4\boldsymbol{Z}^{24}+2(24,23,2)+(24,12,8)$	$\phi^3\boldsymbol{G}^{12}+\phi^2(12,11,2)+\phi(12,7,4)+(12,5,8)'$
—	Λ_{24}	24	4	$4\boldsymbol{Z}^{24}+2(24,18,4)+(24,6,16)'$	$\phi^4\boldsymbol{G}^{12}+\phi^3(12,11,2)+\phi^2(12,7,4)+\phi(12,5,8)'+(12,1,16)'$

of depth $\mu = n - r$ with the code formulas

$$\Lambda(r,n)^{\perp} = \phi^{n-r}\boldsymbol{G}^N + \phi^{n-r-1}\mathrm{RM}(n-r-1,n) + \cdots + \mathrm{RM}(0,n)$$

where $N = 2^n$. Thus $\Lambda(n,n)^{\perp} = \boldsymbol{G}^N \simeq \boldsymbol{Z}^{2N}$, $\Lambda(0,n)^{\perp} = \Lambda(n)$, and $\Lambda(n-1,n)^{\perp}, n \geq 1$, is the dual $D_N^{\perp}$ of the checkerboard lattice D_N, with code formula $D_N^{\perp} = \phi\boldsymbol{G}^N + \mathrm{RM}(0,n) = \phi\boldsymbol{G}^N + (N,1,N)$, where $N = 2^n$ and $(N,1,N)$ is the repetition code of length N.

Table I gives the real and complex code formulas of the Barnes–Wall lattices and their principal sublattices for up to 32 real dimensions (16 complex dimensions). In addition to the designations already introduced, we designate $\Lambda(1,3)$ as H_{16}, $\Lambda(1,4)$ as H_{32}, and $\Lambda(2,4)$ as X_{32}. For reference, we also give "code formulas" for the Leech lattice Λ_{24} and its principal sublattices H_{24}, X_{24}, D_{24}, and $\boldsymbol{Z}^{24}$, which are closely related to their 32-dimensional relatives. (Λ_{24} and its principal sublattice H_{24} are indecomposable complex binary lattices of depths 4 and 3, respectively; in these two cases a notation such as $(24,6,16)'$ means the set of all binary linear combinations modulo 4 of a set of six generators whose coordinates are integers modulo 4, such that the minimum nonzero norm in any coset with such a representative is 16.)

Table II gives additional information on these lattices. The informativity $k(\Lambda)$ and the normalized informativity $\kappa(\Lambda)$ follow from the dimensionality of the Reed–Muller codes in the complex code formula, and from $k(\Lambda)$ (or from the code formulas) we can compute the redundancy $r(\Lambda)$ and the normalized redundancy $\rho(\Lambda)$. For all of these lattices, the minimum squared distance $d^2_{\min}(\Lambda)$ is equal to 2^{μ}; it follows that the fundamental coding gain is given by $\gamma(\Lambda) = 2^{\mu-\rho} = 2^{\kappa}$. (This expression for the fundamental coding gain has the following interpretation: both Λ and $\phi^{\mu}\boldsymbol{G}^N$ have the same minimum squared distance 2^{μ}, but Λ is the union of $2^{k(\Lambda)}$ cosets of $\phi^{\mu}\boldsymbol{G}^N$ and is therefore $2^{k(\Lambda)}$ times as dense as $\phi^{\mu}\boldsymbol{G}^N$ in $2N$-space. A constellation based on Λ will therefore be a factor of $2^{k(\Lambda)}$ smaller in $2N$ dimensions, or $2^{\kappa(\Lambda)}$ in two dimensions, which translates into power savings of a factor of $2^{\kappa(\Lambda)}$.) We also give the normalized error coefficient $\tilde{N}_0 = N_0(\Lambda)/N$ (normalized to two dimensions), which may be obtained from the lattice literature or from the weight distributions of Reed–Muller codes. Finally, we give the number 2^{ν} of states in the trellis diagrams for Λ that are derived in Part II, as well as the corresponding normalized decoding complexity $\tilde{N}_D = N_D(\Lambda)/N$, per two dimensions, where $N_D(\Lambda)$ is the number of binary operations (additions or comparisons of two numbers) required to decode Λ (find the closest element of Λ to an arbitrary point $\boldsymbol{r}$ in $2N$-space), using the trellis-based algorithms of Part II.

(Note that the constellation expansion factor $2^{\rho(\Lambda)}$ is almost a factor of two smaller for H_{2N} than for Λ_{2N}, while the fundamental coding gain is only slightly inferior (by a factor of $2^{-1/N}$). The error coefficient and decoding complexity are also about a factor of two smaller. Therefore, the lattices H_{2N} may be attractive alternatives to the densest lattices Λ_{2N} in practical applications. In addition,

TABLE II
USEFUL BINARY LATTICES

Λ	$k(\Lambda)$	$\kappa(\Lambda)$	$r(\Lambda)$	$\rho(\Lambda)$	$d^2_{\min}(\Lambda)$	$\gamma(\Lambda)$	dB	$\tilde{N}_0$	2^ν	$\tilde{N}_D$
$\boldsymbol{Z}^2$	0	0	0	0	1	1	0.00	4	1	1
$\boldsymbol{Z}^4$	0	0	0	0	1	1	0.00	4	1	1
D_4	1	1/2	1	1/2	2	$2^{1/2}$	1.51	12	2	3.5
$\boldsymbol{Z}^8$	0	0	0	0	1	1	0.00	4	1	1
D_8	3	3/4	1	1/4	2	$2^{3/4}$	2.26	28	2	5.75
E_8	4	1	4	1	4	2	3.01	60	4	11.75
$\boldsymbol{Z}^{16}$	0	0	0	0	1	1	0.00	4	1	1
D_{16}	7	7/8	1	1/8	2	$2^{7/8}$	2.63	60	2	~ 7
H_{16}	11	11/8	5	5/8	4	$2^{11/8}$	4.14	284	8	~ 32
Λ_{16}	12	3/2	12	3/2	8	$2^{3/2}$	4.52	540	16	~ 64
$\boldsymbol{Z}^{32}$	0	0	0	0	1	1	0.00	4	1	1
D_{32}	15	15/16	1	1/16	2	$2^{15/16}$	2.82	124	2	7.5
X_{32}	26	13/8	6	3/8	4	$2^{13/8}$	4.89	1244	16	~ 76
H_{32}	31	31/16	17	17/16	8	$2^{31/16}$	5.83	5084	128	~ 792
Λ_{32}	32	2	32	2	16	4	6.02	9180	256	~ 1584
$\boldsymbol{Z}^{24}$	0	0	0	0	1	1	0.00	4	1	1
D_{24}	11	11/12	1	1/12	2	$2^{11/12}$	2.76	92	2	7.25
X_{24}	18	3/2	6	1/2	4	$2^{3/2}$	4.52	508	8	~ 42
H_{24}	23	23/12	13	13/12	8	$2^{23/12}$	5.77	8188	128	~ 632
Λ_{24}	24	2	24	2	16	4	6.02	16380	256	~ 1264

the Leech half-lattice H_{24} is decomposable as a real lattice, with the (24,12,8) Golay code appearing in the code formula, whereas the Leech lattice Λ_{24} itself is not.)

Because of the nested character of Reed–Muller codes, the code formulas show that $\boldsymbol{Z}^{2N} \simeq \Lambda(n,n)/\Lambda(n-1,n)/\cdots/\Lambda(0,n)$ is a partition chain of 2^n-dimensional complex lattices of depths $0/1/\cdots/n$ and with distances $1/2/\cdots/2^n$ (for short). Also, we may verify that $\Lambda(0,n)/\phi\Lambda(1,n)/\cdots/\phi^n\Lambda(n,n) = \phi^n\boldsymbol{G}^N \simeq R^n\boldsymbol{Z}^{2N}$ is a partition chain of 2^n-dimensional complex lattices of depths $n/n/\cdots/n$ and with distances $2^n/2^n/\cdots/2^n$. Similarly, $\boldsymbol{Z}^{2N} \simeq \Lambda(n,n)^\perp/\Lambda(n-1,n)^\perp/\cdots/\Lambda(0,n)^\perp = \Lambda(0,n)$ is a partition chain of 2^n-dimensional complex lattices of depths $0/1/\cdots/n$ and with distances $1/2/\cdots/2^n$, and $\Lambda(0,n) = \Lambda(0,n)^\perp/\phi\Lambda(1,n)^\perp/\cdots/\phi^n\Lambda(n,n)^\perp = \phi^n\boldsymbol{G}^N \simeq R^n\boldsymbol{Z}^{2N}$ is a partition chain of 2^n-dimensional complex lattices of depths $n/n/\cdots/n$ and with distances $2^n/2^n/\cdots/2^n$.

Fig. 9 is an illustration of these partition chains in dimensions 2, 4, 8, and 16, extended indefinitely using the lattices $\phi^j\Lambda(r,n) \simeq R^j\Lambda(r,n)$ for all $j \geq 0$. The lattices are arranged in columns according to depth, where for the purposes of this diagram we regard $\phi^j\Lambda(r,n) \simeq R^j\Lambda(r,n)$ as having depth $n-r$ for any $j \geq 0$. One unit of vertical distance corresponds to a two-way partition. The lines indicate sublattice relationships. From this diagram, we can easily find the least μ for which $\phi^\mu\boldsymbol{G}^N \simeq R^\mu\boldsymbol{Z}^{2N}$ is a sublattice of any given lattice, and thus verify the depths of lattices and partition chains.

In the rest of this paper, we will be considering coset codes $\mathbb{C}(\Lambda/\Lambda'; C)$ based on partitions Λ/Λ' of lattices that appear in Fig. 9. The partitions that we will use are generally those with Λ' at least as dense as Λ, depths no greater than four, and orders no greater than 2^{12}. Table III summarizes some of the principal properties of such partitions, including: the (real) dimension $2N$; the order $|\Lambda/\Lambda'|$

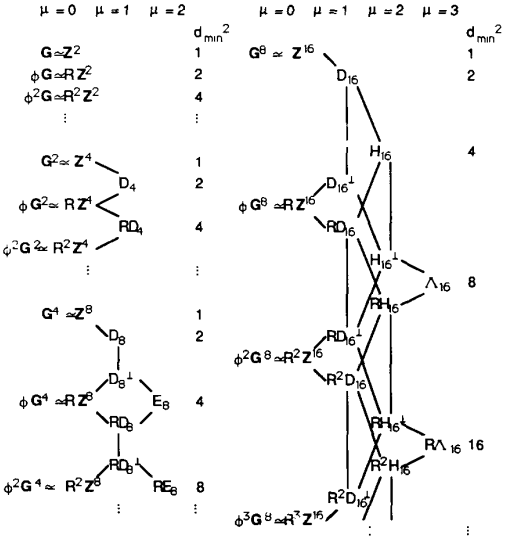


Fig. 9. Partition chains involving Barnes–Wall lattices, principal sublattices, and duals of principal sublattices in 2, 4, 8, and 16 dimensions.

of the partition; its depth $\mu(\Lambda/\Lambda') = \mu(\Lambda')$, normalized informativity $\kappa(\Lambda/\Lambda') = \kappa(\Lambda')$, and normalized redundancy $\rho(\Lambda/\Lambda') = \rho(\Lambda)$; the minimum squared distances of Λ and Λ'; and the normalized (per two dimensions) complexity $\tilde{N}_D = N_D/N$ of decoding the partition, where N_D is the number of binary operations required by the trellis-based decoding algorithms of part II to determine the closest element of each of the $|\Lambda/\Lambda'|$ cosets of Λ' in the partition Λ/Λ' to an arbitrary point $\boldsymbol{r}$ in $2N$-space. (*Note*: if Λ/Λ' is a partition, then so is $R\Lambda/R\Lambda' \simeq \phi\Lambda/\phi\Lambda'$; if it is simpler to decode $R\Lambda/R\Lambda'$ than Λ/Λ', the lesser $\tilde{N}_D$ is given.) The final column gives $\tilde{N}_D/|\Lambda/\Lambda'|$, to show that $\tilde{N}_D$ is approximated by $\alpha|\Lambda/\Lambda'|$, where α is a small number in the range from one to six.

TABLE III
USEFUL LATTICE PARTITIONS

Λ	Λ'	$2N$	$\lvert\Lambda/\Lambda'\rvert$	μ	κ	ρ	$d^2_{\min}(\Lambda)$	$d^2_{\min}(\Lambda')$	$\tilde{N}_D$	$\tilde{N}_D/\lvert\Lambda/\Lambda'\rvert$
$\boldsymbol{Z}^2$	$R\boldsymbol{Z}^2$	2	2	1	0	0	1	2	2	1
$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	2	4	2	0	0	1	4	4	1
$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	2	8	3	0	0	1	8	8	1
$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	2	16	4	0	0	1	16	16	1
$\boldsymbol{Z}^4$	D_4	4	2	1	1/2	0	1	2	5	2.5
D_4	RD_4	4	4	2	1/2	1/2	2	4	10	2.5
$\boldsymbol{Z}^4$	RD_4	4	8	2	1/2	0	1	4	16	2
D_4	$2D_4$	4	16	3	1/2	1/2	2	8	32	2
$\boldsymbol{Z}^4$	$2D_4$	4	32	3	1/2	0	1	8	56	1.75
D_4	$2RD_4$	4	64	4	1/2	1/2	2	16	112	1.75
$\boldsymbol{Z}^4$	$2RD_4$	4	128	4	1/2	0	1	16	208	1.6
$\boldsymbol{Z}^8$	D_8	8	2	1	3/4	0	1	2	6.5	3.25
D_8	E_8	8	8	2	1	1/4	2	4	30	3.75
$\boldsymbol{Z}^8$	E_8	8	16	2	1	0	1	4	44	2.75
E_8	RE_8	8	16	3	1	1	4	8	60	3.75
$D_8^{\perp}$	RE_8	8	32	3	1	3/4	2	8	88	2.75
D_8	RE_8	8	128	3	1	1/4	2	8	280	2.2
$\boldsymbol{Z}^8$	RE_8	8	256	3	1	0	1	8	504	~2
E_8	$2E_8$	8	256	4	1	1	4	16	560	2.2
D_8	$2E_8$	8	2^{11}	4	1	1/4	2	16	3792	~2
$\boldsymbol{Z}^8$	$2E_8$	8	2^{12}	4	1	0	1	16	7376	~2
$\boldsymbol{Z}^{16}$	D_{16}	16	2	1	7/8	0	1	2	7.25	3.6
D_{16}	H_{16}	16	16	2	11/8	1/8	2	4	74	4.6
$\boldsymbol{Z}^{16}$	H_{16}	16	32	2	11/8	0	1	4	104	3.25
H_{16}	Λ_{16}	16	128	3	3/2	5/8	4	8	776	~6
D_{16}	Λ_{16}	16	2^{11}	3	3/2	1/8	2	8	8440	~4
$\boldsymbol{Z}^{16}$	Λ_{16}	16	2^{12}	3	3/2	0	1	8	16376	~4
Λ_{16}	$R\Lambda_{16}$	16	256	4	3/2	3/2	8	16	1552	~6

IV. TRELLIS CODES

A. Introduction to Trellis Codes

A *trellis code* is a coset code $\mathbb{C}(\Lambda/\Lambda'; C)$ as shown in Fig. 1, where C is a rate-$k/(k+r)$ convolutional code. In this paper C will always be a binary convolutional code, and Λ and Λ' binary lattices, generally mod-2 or mod-4.

The *codewords* in a rate-$k/(k+r)$ convolutional code may be expressed as sequences $(\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \cdots)$ of binary $(k+r)$-tuples $\boldsymbol{a}_j$, which serve as labels that select cosets $\Lambda' + \boldsymbol{c}(\boldsymbol{a}_j)$ of Λ' in the partition Λ/Λ'. The *code sequences* in a trellis code $\mathbb{C}(\Lambda/\Lambda'; C)$ therefore consist of the sequences of elements of Λ that are congruent to some coset representative sequence $(\boldsymbol{c}(\boldsymbol{a}_t), \boldsymbol{c}(\boldsymbol{a}_{t+1}), \cdots)$ modulo Λ', where $(\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \cdots)$ is a codeword in C.

For technical reasons, all sequences $(s_t, s_{t+1}, \cdots)$ are assumed to have a definite starting time t, although they may continue indefinitely. We may associate with any such sequence a formal power series in the delay operator D,

$$s(D) = s_t D^t + s_{t+1} D^{t+1} + \cdots$$

where t may be any integer, positive or negative. Thus a coset code maps a label sequence $\boldsymbol{a}(D)$ to a coset representative sequence $\boldsymbol{c}(D)$.

The important properties of a convolutional code are linearity and time-invariance. Linearity means that the mod-2 sum of any two codewords is a codeword. Time-invariance means that the time shift of any codeword is a codeword, i.e., if $\boldsymbol{a}(D)$ is a codeword, then so is $D\boldsymbol{a}(D)$. (It follows that a convolutional code is a vector space over the field of binary formal power series $f(D)$, $f_t \in \{0,1\}$; the dimension of this vector space is k, and any codeword $\boldsymbol{a}(D)$ can be written as $\boldsymbol{a}(D) = \Sigma f_j(D)\boldsymbol{g}_j(D)$, where the $\boldsymbol{g}_j(D)$, $1 \le j \le k$, are a set of k generator sequences that form a generator matrix G; see [20].) We assume that the labeling function $\boldsymbol{c}(\boldsymbol{a}_t)$ is time-invariant; then a trellis code is also time-invariant, although, as we shall see in more detail below, not necessarily linear.

A convolutional code C has a well-defined state space, which is a vector space over the binary field of some dimension ν. The parameter ν is called the overall constraint length, or just *constraint length*, of C. The code C can be generated by a linear (binary) finite-state machine with k inputs, $k+r$ outputs, and ν binary memory elements; such an encoder has 2^ν *states*.

A *trellis diagram* for a 2^ν-state, rate-$k/(k+r)$ convolutional code is an extended state transition diagram for the encoder that generates C. For each time t, it has 2^ν *nodes*, or states, representing the possible states at time t. For each possible state transition, it has a *branch* connecting the two corresponding nodes. There are 2^k branches leaving and entering each node, and each is labeled with the $(k+r)$-tuple $\boldsymbol{a}$ that represents the encoder output associated with that state transition. Thus we may obtain a trellis diagram for a trellis code $\mathbb{C}(\Lambda/\Lambda'; C)$ by taking a trellis diagram for C and replacing each label $\boldsymbol{a}$ by the corresponding coset representative $\boldsymbol{c}(\boldsymbol{a})$, representing the coset $\Lambda' + \boldsymbol{c}(\boldsymbol{a})$.

The *minimum Hamming distance* $d_H(C)$ of a convolutional code C is the minimum Hamming distance between

any two codewords in C, i.e., the minimum number of coordinate differences between the outputs $\boldsymbol{a}$ on any two paths in the trellis that start and end in a common state. Because C is linear, this is also the minimum Hamming weight of any codeword, which is the minimum weight of any path that starts and ends on a zero state.

The *minimum squared distance* $d^2_{\min}(\mathbb{C})$ of a trellis code $\mathbb{C}(\Lambda/\Lambda'; C)$ is the minimum squared distance between any two code sequences in $\mathbb{C}$, which is the lesser of a) the minimum distance $\|\lambda_1(D)-\lambda_2(D)\|^2$ between sequences $\lambda_1(D)$ and $\lambda_2(D)$ that correspond to two distinct paths that begin and end in a common state; and b) the minimum distance $d^2_{\min}(\Lambda')$ between elements of Λ', corresponding to "parallel transitions" associated with any given branch. (If $\lambda(D)$ is a code sequence, then so is $\lambda(D)+\lambda' D^j$ for any $\lambda' \in \Lambda'$.)

For example, the four-state Ungerboeck code shown in Figs. 2 and 3 uses the four-state rate-1/2 convolutional code whose encoder and trellis diagram are illustrated in Fig. 10. Contrary to convention, the encoder is shown in coset code form, using the partition $(2,2,1)/(2,1,2)/(2,0,\infty)$ of binary codes of length 2. Let $\boldsymbol{g}_0$ be a coset representative for the nonzero coset in the partition $(2,2)/(2,1)$, e.g., $\boldsymbol{g}_0=[10]$, and let $\boldsymbol{g}_1$ be the coset representative for the nonzero coset in the partition $(2,1)/(2,0)$, i.e., $\boldsymbol{g}_1=[11]$. The two bits $\boldsymbol{a}=(a_0,a_1)$ select a 2-tuple $\boldsymbol{c}(\boldsymbol{a})=a_0\boldsymbol{g}_0+a_1\boldsymbol{g}_1$, representing one of the four cosets of $(2,0)$ (the single codeword [00]) in the four-way partition $(2,2)/(2,0)$. In the trellis diagram, branches are labeled by both $\boldsymbol{a}$ and $\boldsymbol{c}(\boldsymbol{a})$.

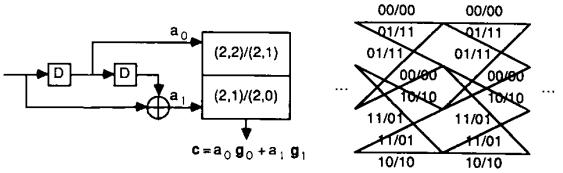


Fig. 10. Convolutional encoder C for four-state Ungerboeck code of Figs. 2 and 3, and trellis diagram labeled with both $\boldsymbol{a}$ and $\boldsymbol{c}(\boldsymbol{a})$.

The minimum Hamming distance of this code (taking the outputs as $\boldsymbol{c}(\boldsymbol{a})$) is five, because the distance between distinct paths is at least two where they diverge, two where they merge, and one somewhere in between. (This is because the difference between paths is a codeword of the $(2,1,2)$ code where they merge and diverge, and a codeword in the $(2,2,1)$ code somewhere in between; that is, we are exploiting the Ungerboeck distance bound for this code partition chain.)

The trellis code of Figs. 2 and 3 is obtained by replacing the code partition chain $(2,2)/(2,1)/(2,0)$ by the corresponding partition of mod-2 lattices, $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2 \simeq \boldsymbol{G}/\phi\boldsymbol{G}/\phi^2\boldsymbol{G}$. Note that $\boldsymbol{g}_0$ and $\boldsymbol{g}_1$ are still coset representatives for the nonzero cosets of $R\boldsymbol{Z}^2$ in the partition $\boldsymbol{Z}^2/R\boldsymbol{Z}^2 \simeq \boldsymbol{G}/\phi\boldsymbol{G}$ and of $2\boldsymbol{Z}^2$ in the partition $R\boldsymbol{Z}^2/2\boldsymbol{Z}^2 \simeq \phi\boldsymbol{G}/\phi^2\boldsymbol{G}$, respectively, if we regard them as integers modulo 2. The trellis diagram of Fig. 10 then continues to represent this trellis code, where we now regard the 2-tuples $\boldsymbol{c}(\boldsymbol{a})$ as coset representatives of cosets of $2\boldsymbol{Z}^2$.

It is easy to see that the minimum squared distance between code sequences corresponding to distinct paths in the trellis is the minimum Hamming distance between sequences $\boldsymbol{c}(D)$ of coset representatives and thus is equal to $d_H(C)=5$. However, since $d^2_{\min}(2\boldsymbol{Z}^2)=4$, the minimum squared distance of the trellis code is $d^2_{\min}(\mathbb{C})=4$. If $\lambda(D)$ is any code sequence, the only code sequences at distance 4 from $\lambda(D)$ are the sequences $\lambda(D)+\lambda' D^j$, where λ' is one of the four elements of $2\boldsymbol{Z}^2$ of norm 4, namely $\pm(2,0)$ and $\pm(0,2)$. These are special cases of general results for trellis codes based on partitions of mod-2 lattices that will be given below.

B. Geometrical Parameters

As with lattices, the two principal geometrical parameters of a trellis code are the minimum squared distance $d^2_{\min}(\mathbb{C})$ between its code sequences and its fundamental volume $V(\mathbb{C})$ (per N dimensions); these determine its fundamental coding gain $\gamma(\mathbb{C})$.

We have already introduced $d^2_{\min}(\mathbb{C})$ and noted that $d^2_{\min}(\mathbb{C})=\min[d_0^2, d^2_{\min}(\Lambda')]$, where d_0^2 is the minimum squared distance $\|\lambda_1(D)-\lambda_2(D)\|^2$ between code sequences $\lambda_1(D)$ and $\lambda_2(D)$ that correspond to two distinct paths that begin and end in a common state in the code trellis. If, as with convolutional codes, the distribution of distances from a given code sequence to all other code sequences does not depend on the given code sequence, then $d^2_{\min}(\mathbb{C})$ is the minimum squared distance from the all-zero code sequence to any other code sequence, i.e., the minimum norm of any code sequence. We call such codes *distance-invariant*. All codes in this paper are distance-invariant.

The *error coefficient* $N_0(\mathbb{C})$ is the average number of code sequences that differ from a given sequence by $d^2_{\min}(\mathbb{C})$ and that first differ from the given sequence at a given time t. By time-invariance, $N_0(\mathbb{C})$ does not depend on the time t, and we may take $t=0$, say. If the code is distance-invariant, then the number of code sequences that differ from a given sequence by $d^2_{\min}(\mathbb{C})$ does not depend on the given sequence, and we may take the given sequence as the all-zero sequence. Thus in a distance-invariant code $N_0(\mathbb{C})$ is the number of sequences of norm $d^2_{\min}(\mathbb{C})$ that start at time zero.

The fundamental volume is a trickier concept. Intuitively, since the code C conveys k bits of information per unit time (per N dimensions) and has r bits of redundancy, it is clear that in some sense the trellis code $\mathbb{C}$ is 2^k times as dense as Λ' and a factor of 2^r less dense than Λ, per N dimensions. Therefore, the fundamental volume $V(\mathbb{C})$ should be equal to $2^{-k}V(\Lambda')$, or to $2^rV(\Lambda)$.

To substantiate this proposition, we argue as follows. Define $\mathbb{C}_t$ as the set of all code sequences that start at time t or later. By time-invariance, all such sets are isomorphic to each other and to a particular such set, say $\mathbb{C}_0$. How-

ever, $\mathbb{C}_t$ is also a proper subset of $\mathbb{C}_{t'}$ if $t > t'$; e.g., $\mathbb{C}_1$ is a proper subset of $\mathbb{C}_0$.

Define two code sequences as equivalent modulo $\mathbb{C}_t$ if their first difference is at time t or later. Two sequences in $\mathbb{C}_0$ are then equivalent modulo $\mathbb{C}_1$ if and only if their first element $\boldsymbol{c}_0$ is the same. Let Λ_0 then be the set of all possible first elements $\boldsymbol{c}_0$; that is, $\Lambda_0 = \{\lambda\colon\ \lambda \in \Lambda' + \boldsymbol{c}(\boldsymbol{a}_0)\}$, where $\boldsymbol{a}_0$ is a possible time-zero output from the encoder C given that all previous outputs were zero or, equivalently, given that the encoder starts in the zero state. There are 2^k such $\boldsymbol{a}_0$, and thus Λ_0 is the union of 2^k cosets of Λ'. Ordinarily, Λ_0 is a lattice, which we call the *time-zero lattice*. By Lemma 1, $V(\Lambda_0) = 2^{-k}V(\Lambda')$; since $V(\Lambda') = 2^{k+r}V(\Lambda)$, we also have $V(\Lambda_0) = 2^rV(\Lambda)$.

In an appropriate sense, therefore, the equivalence classes of $\mathbb{C}_0$ modulo $\mathbb{C}_1$, which we may write as $\mathbb{C}_0/\mathbb{C}_1$, are isomorphic to the time-zero lattice Λ_0. The set $\mathbb{C}_0$ has the decomposition

$$\mathbb{C}_0 = \mathbb{C}_0/\mathbb{C}_1 + \mathbb{C}_1/\mathbb{C}_2 + \cdots$$

which by time-invariance is isomorphic to the Cartesian product $\Lambda_0 \times \Lambda_0 \times \cdots$. In other words, $\mathbb{C}_0$ fills space as densely as does $\Lambda_0 \times \Lambda_0 \times \cdots$. Thus it is reasonable to define the *fundamental volume* of $\mathbb{C}$ per N dimensions as $V(\mathbb{C}) = V(\Lambda_0)$.

Now we may define the *fundamental coding gain* of a trellis code in the same way as we did for a lattice:

$$\gamma(\mathbb{C}) = d^2_{\min}(\mathbb{C})/V(\mathbb{C})^{2/N}.$$

Let us now write $k(C)$ and $r(C)$ for the parameters k and r of a convolutional code C, and $\kappa(C)$ and $\rho(C)$ for the normalized informativity and redundancy $\kappa(C) = 2k(C)/N$ and $\rho(C) = 2r(C)/N$, respectively. Since $V(\mathbb{C}) = V(\Lambda_0) = 2^{-k(C)}V(\Lambda') = 2^{r(C)}V(\Lambda)$, we also have the expressions

$$\begin{aligned}\gamma(\mathbb{C}) &= 2^{\kappa(C)}d^2_{\min}(\mathbb{C})/V(\Lambda')^{2/N} \\ &= 2^{\kappa(C)}\left[d^2_{\min}(\mathbb{C})/d^2_{\min}(\Lambda')\right]\gamma(\Lambda') \\ &= 2^{\kappa(\mathbb{C})}\left[d^2_{\min}(\mathbb{C})/2^{\mu(\mathbb{C})}\right]\end{aligned}$$

where $\kappa(\mathbb{C}) = \kappa(C) + \kappa(\Lambda')$ and $\mu(\mathbb{C}) = \mu(\Lambda')$. Also,

$$\begin{aligned}\gamma(\mathbb{C}) &= 2^{-\rho(C)}d^2_{\min}(\mathbb{C})/V(\Lambda)^{2/N} \\ &= 2^{-\rho(C)}\left[d^2_{\min}(\mathbb{C})/d^2_{\min}(\Lambda)\right]\gamma(\Lambda) \\ &= 2^{-\rho(\mathbb{C})}d^2_{\min}(\mathbb{C})\end{aligned}$$

where $\rho(\mathbb{C}) = \rho(C) + \rho(\Lambda)$. Thus if we define the normalized redundancy, informativity, and depth of the trellis code $\mathbb{C}$ as the sums of the corresponding quantities for the code C and partition Λ/Λ', where we regard the depth of C as 0, then we get expressions analogous to those that we obtained for lattices.

The following lemma is both useful in itself and also gives an intuitive explanation of these formulas.

Lemma 5: If $\mathbb{C}(\Lambda/\Lambda'; C)$ is a trellis code based on a partition Λ/Λ' of binary lattices, where the depth of Λ' is μ, and a 2^ν-state, rate-$k/(k+r)$ convolutional code C, then there is an equivalent trellis code $\mathbb{C}(\boldsymbol{G}^N/\phi^\mu\boldsymbol{G}^N; C')$ based on the partition $\boldsymbol{G}^N/\phi^\mu\boldsymbol{G}^N$, where C' is a 2^ν-state, rate-$[k + k(\Lambda')]/N\mu$ convolutional code, and N is the dimension of Λ or Λ' as complex lattices.

Proof: If Λ' is a binary lattice of depth μ, then $\boldsymbol{G}^N/\Lambda/\Lambda'/\phi^\mu\boldsymbol{G}^N$ is a partition chain, with $|\boldsymbol{G}^N/\Lambda| = 2^{r(\Lambda)}$, $|\Lambda/\Lambda'| = 2^{k+r}$, $|\Lambda'/\phi^\mu\boldsymbol{G}^N| = 2^{k(\Lambda')}$, and $|\boldsymbol{G}^N/\phi^\mu\boldsymbol{G}^N| = 2^{N\mu}$. In view of the coset decomposition $\boldsymbol{G}^N = \phi^\mu\boldsymbol{G}^N + [\Lambda'/\phi^\mu\boldsymbol{G}^N] + [\Lambda/\Lambda'] + [\boldsymbol{G}^N/\Lambda]$, we may select a coset of $\phi^\mu\boldsymbol{G}^N$ in the partition $\boldsymbol{G}^N/\phi^\mu\boldsymbol{G}^N$ by the following set of $N\mu$ bits: an all-zero $r(\Lambda)$-tuple $\boldsymbol{0}$, which selects the zero coset of Λ in the partition $\boldsymbol{G}^N/\Lambda$, namely, Λ itself; a $(k+r)$-tuple $\boldsymbol{a}$, which selects the coset $\Lambda' + \boldsymbol{c}(\boldsymbol{a})$ of Λ' in the partition Λ/Λ' as in the original trellis code $\mathbb{C}(\Lambda/\Lambda'; C)$; and finally, a $k(\Lambda')$-tuple $\boldsymbol{a}'$ of "uncoded bits" which selects one of the $2^{k(\Lambda')}$ cosets of $\phi^\mu\boldsymbol{G}^N$ whose union is Λ'. These $N\mu$ bits can be regarded as the outputs of an augmented convolutional encoder C', which has $k + k(\Lambda')$ information bits, $r + r(\Lambda)$ redundant bits, and the same number of states as the original encoder for C, as illustrated in Fig. 11. The set of code sequences that may be generated by this augmented encoder are the same as those in the original code $\mathbb{C}(\Lambda/\Lambda'; C)$.

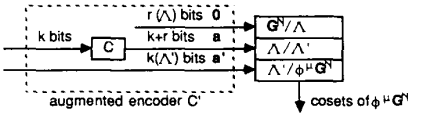


Fig. 11. Augmented encoder C' of Lemma 5.

An alternative form of Lemma 5 is as follows. If the 2-depth of Λ' is m, then $\mathbb{C}(\Lambda/\Lambda'; C)$ is equivalent to a code $\mathbb{C}(\boldsymbol{Z}^N/2^m\boldsymbol{Z}^N; C')$ based on the partition $\boldsymbol{Z}^N/2^m\boldsymbol{Z}^N$, where C' is a 2^ν-state, rate-$[Nm - r - r(\Lambda)]/Nm$ convolutional code, and N is the dimension of Λ or Λ' as real lattices. The proof is essentially the same. Indeed, if μ is even, $\mu = 2m$, then the partition $\boldsymbol{Z}^{2N}/\Lambda/\Lambda'/2^m\boldsymbol{Z}^{2N}$ is the same as $\boldsymbol{G}^N/\Lambda/\Lambda'/\phi^\mu\boldsymbol{G}^N$, and the augmented encoder C' is the same; if μ is odd, $\mu = 2m - 1$, then the partition $\boldsymbol{Z}^{2N}/\Lambda/\Lambda'/2^m\boldsymbol{Z}^{2N}$ is an extension of $\boldsymbol{G}^N/\Lambda/\Lambda'/\phi^\mu\boldsymbol{G}^N$ by the partition $\phi^\mu\boldsymbol{G}^N/\phi^{\mu+1}\boldsymbol{G}^N$, and the augmented encoder C' just uses N more uncoded bits.

The coding gain $\gamma(\mathbb{C})$ may now be related to that of the lattices $\boldsymbol{G}^N$, Λ, Λ', and $\phi^\mu\boldsymbol{G}^N$ as follows. Relative to $\gamma(\Lambda') = 2^{\kappa(\Lambda')}[d^2_{\min}(\Lambda')/2^{\mu(\Lambda')}]$, the gain $\gamma(\mathbb{C})$ is greater by a factor of $2^{\kappa(C)}$ due to the fact that $\mathbb{C}$ conveys $\kappa(C)$ more bits of information per two dimensions, offset by a distance loss factor of $d^2_{\min}(\mathbb{C})/d^2_{\min}(\Lambda')$ (if any). Relative to $\gamma(\Lambda) = 2^{-\rho(\Lambda)}d^2_{\min}(\Lambda')$, the gain $\gamma(\mathbb{C})$ is greater by the distance gain factor of $d^2_{\min}(\mathbb{C})/d^2_{\min}(\Lambda)$, offset by a power loss of $2^{-\rho(C)}$ due to constellation expansion. If $\Lambda' = \phi^\mu\boldsymbol{G}^N$, then $\gamma(\Lambda') = 1$ and $d^2_{\min}(\Lambda') = 2^{\mu(\Lambda')}$, so $\gamma(\mathbb{C})$ is simply $2^{\kappa(\mathbb{C})}$, offset by a distance loss factor of $d^2_{\min}(\mathbb{C})/2^{\mu(\Lambda')}$ (if any). If $\Lambda = \boldsymbol{G}^N$, then $\gamma(\Lambda) = 1$ and $d^2_{\min}(\Lambda) = 1$, so $\gamma(\mathbb{C})$

is simply $d^2_{\min}(\mathbb{C})$, offset by a power loss of $2^{-\rho(\mathbb{C})}$ due to constellation expansion.

This last expression is the simplest and shows that we need to know only the minimum squared distance $d^2_{\min}(\mathbb{C})$ and the normalized redundancy $\rho(\mathbb{C})$ to compute the fundamental coding gain $\gamma(\mathbb{C}) = 2^{-\rho(\mathbb{C})} d^2_{\min}(\mathbb{C})$, where the normalized redundancy is simply the sum of the normalized redundancies of the code C and the lattice Λ. If $\Lambda = \boldsymbol{G}^N \simeq \boldsymbol{Z}^{2N}$, then it suffices to know the normalized redundancy r/N of the code C. A small normalized redundancy $\rho(\mathbb{C})$ is thus desirable to both minimize constellation expansion and maximize coding gain for a given $d^2_{\min}(\mathbb{C})$, as was recognized by Wei [11].

Lemma 5 shows that all trellis codes based on partitions of binary lattices are equivalent to trellis codes based on partitions $\boldsymbol{G}^N/\phi^{\mu}\boldsymbol{G}^N$, or, by extension, partitions $\boldsymbol{Z}^N/2^m\boldsymbol{Z}^N$ of the integer lattice $\boldsymbol{Z}^N$, so that in principle only these kinds of partitions need to be considered to discover all binary trellis codes. In practice, consideration of more general partitions Λ/Λ' both facilitates the search for good codes and simplifies their encoding and decoding.

C. Linear Trellis Codes

In general, if $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are two code sequences in a trellis code $\mathbb{C}$, it is not necessarily true that $\boldsymbol{c}_1 + \boldsymbol{c}_2$ and $\boldsymbol{c}_1 - \boldsymbol{c}_2$ are also code sequences. When this property does hold, we say that $\mathbb{C}$ is a *linear trellis code*. In this section we give some examples of linear trellis codes, including the important case where Λ/Λ' is a partition of mod-2 binary lattices.

If a trellis code is linear, then it is a lattice, albeit an infinite-dimensional lattice. Because it is a time-invariant infinite-dimensional lattice, it is usually possible to define its parameters on a per-unit-time or per-two-dimensions basis, so that they are perfectly finite and analogous to the parameters of a finite-dimensional lattice, as we have already seen in the previous section. In this sense, linear trellis codes are to finite-dimensional lattices as convolutional codes are to block codes. (To extend this analogy, nonlinear trellis codes are a generalization of nonlattice finite-dimensional sphere packings.)

A trivial example of a linear trellis code is the repeated use of a lattice code. If Λ is any lattice, let Λ^{∞} be defined as the set of all sequences $(\lambda_t, \lambda_{t+1}, \cdots)$, where $\lambda_t \in \Lambda$, $j \geq t$. This may be regarded as a trellis code based on the "dummy partition" Λ/Λ where the convolutional encoder C disappears. The trellis associated with Λ^{∞} has one state for each time t, and the branch in each time interval is labeled by Λ.

If $\mathbb{C}(\Lambda/\Lambda'; C)$ is a linear trellis code, it is a sublattice of Λ^{∞} and has $(\Lambda')^{\infty}$ as a sublattice, so $\Lambda^{\infty}/\mathbb{C}/(\Lambda')^{\infty}$ is a partition chain. If the partition Λ/Λ' has depth μ, then $(\boldsymbol{G}^N)^{\infty}/\mathbb{C}/(\phi^{\mu}\boldsymbol{G}^N)^{\infty}$ is a partition chain; if it has 2-depth m, then $(\boldsymbol{Z}^N)^{\infty}/\mathbb{C}/(2^m\boldsymbol{Z}^N)^{\infty}$ is a partition chain.

A trellis code $\mathbb{C}$ has little chance of being linear unless the mapping $\boldsymbol{c}(\boldsymbol{a})$ from encoder output $(k+r)$-tuples (labels) $\boldsymbol{a}$ to coset representatives $\boldsymbol{c}$ is linear modulo Λ', as in Lemma 2; i.e., $\boldsymbol{c}(\boldsymbol{a}) = \boldsymbol{a}G$, where $G = \{\boldsymbol{g}_j, 1 \leq j \leq k+r\}$ is a generator matrix of $k+r$ vectors of Λ that span Λ, modulo Λ'. The following lemma shows that when Λ/Λ' is a partition of mod-2 lattices and the labeling map is linear, $\mathbb{C}$ is linear, and indeed isomorphic to a binary convolutional code, in the same sense as a mod-2 binary lattice Λ is isomorphic to a binary block code, described in Lemma 3 (recall that if $\boldsymbol{c}(\boldsymbol{a})$ is an Ungerboeck labeling, it is linear modulo Λ').

Lemma 6: If Λ' is a mod-2 lattice, C is a 2^r-state, rate-$k/(k+r)$ convolutional code and the labeling map $\boldsymbol{c}(\boldsymbol{a})$ is linear modulo Λ', then a trellis code $\mathbb{C}(\Lambda/\Lambda'; C)$ is the set of all sequences of integer N-tuples that are congruent modulo 2 to one of the words in a 2^r-state rate-$[N - r(\mathbb{C})]/N$ convolutional code C'. The redundancy of $\mathbb{C}$ is $r(\mathbb{C}) = r + r(\Lambda)$, and its minimum squared distance is $d^2_{\min}(\mathbb{C}) = \min[4, d_H(C')]$.

Sketch of proof: By the extension of Lemma 5, $\mathbb{C}$ is equivalent to a code based on the partition $\boldsymbol{Z}^N/2\boldsymbol{Z}^N$, where the augmented encoder C' has N output bits and redundancy $r(C') = r + r(\Lambda)$. In the augmented encoder, the $(k+r)$-tuple $\boldsymbol{a}$ and the uncoded bits $\boldsymbol{a}'$ specify a coset of $2\boldsymbol{Z}^N$ in the partition $\boldsymbol{Z}^N/2\boldsymbol{Z}^N$, which may be specified by a binary N-tuple $\boldsymbol{c}'(\boldsymbol{a}, \boldsymbol{a}')$. The mapping $\boldsymbol{c}'(\boldsymbol{a}, \boldsymbol{a}')$ may be taken to be linear mod 2 if $\boldsymbol{c}(\boldsymbol{a})$ is linear modulo Λ'. Thus $\mathbb{C}$ is set of all sequences of integer N-tuples that are congruent modulo 2 to one of the words in the convolutional code C'. The minimum squared distance between code sequences corresponding to distinct codewords of C' is $d_H(C')$, and $d^2_{\min}(2\boldsymbol{Z}^N) = 4$.

The four-state Ungerboeck code shown in Figs. 2, 3, and 10 is an example of a code of this type. The encoder of Fig. 10 is of the form of the augmented encoder of Fig. 11. Many of the important known codes to be listed in the next section, including Gallager–Calderbank–Sloane (GCS)-type codes and most of the Wei codes, are of this type.

Even when a code is not linear, it may still be regular in the sense of Calderbank and Sloane [13]. A labeling $\boldsymbol{c}(\boldsymbol{a})$ is defined as *regular* if the minimum squared distance between points in two cosets $\Lambda' + \boldsymbol{c}(\boldsymbol{a})$ and $\Lambda' + \boldsymbol{c}(\boldsymbol{a}')$ is a function only of the mod-2 sum $\boldsymbol{a} \oplus \boldsymbol{a}'$ of their labels or, equivalently, if the minimum norm in the coset $\Lambda' + \boldsymbol{c}(\boldsymbol{a}) - \boldsymbol{c}(\boldsymbol{a}')$ is equal to the minimum norm in the coset $\Lambda' + \boldsymbol{c}(\boldsymbol{a} \oplus \boldsymbol{a}')$. For example, the 2-bit standard binary representation is a regular labeling of the four cosets in the partition $\boldsymbol{Z}/4\boldsymbol{Z}$. If $\mathbb{C}(\Lambda/\Lambda'; C)$ is a code based on a partition Λ/Λ' with a regular labeling, the minimum squared distance between code sequences in the coset sequences $\boldsymbol{c}(\boldsymbol{a}(D))$ and $\boldsymbol{c}(\boldsymbol{a}'(D))$ corresponding to codewords $\boldsymbol{a}(D)$ and $\boldsymbol{a}'(D)$ is then equal to the minimum norm of any code sequence in the coset sequence $\boldsymbol{c}(\boldsymbol{a}(D) \oplus \boldsymbol{a}'(D))$ corresponding to the codeword $\boldsymbol{a}(D) \oplus \boldsymbol{a}'(D)$. Therefore, the distribution of distances from any given code sequence to all other code sequences is the same as the norm distribution of code sequences, and the code is thus distance-invariant.

A labeling is regular under any of the following conditions:

a) if it is linear, in the sense that $c(a)-c(a')=c(a\oplus a')$ modulo Λ', e.g., whenever Λ and Λ' are mod-2 lattices;
b) if it is an Ungerboeck labeling and the Ungerboeck distance bound always holds with equality, e.g., any Ungerboeck labeling for $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2/2R\boldsymbol{Z}^2 \simeq \boldsymbol{G}/\phi\boldsymbol{G}/\phi^2\boldsymbol{G}/\phi^3\boldsymbol{G}$;
c) if Λ and Λ' are N-fold Cartesian products Λ^N and $(\Lambda')^N$, and the labeling for $\Lambda^N/(\Lambda')^N$ is the N-fold Cartesian product of a regular labeling for Λ/Λ', e.g., when the partition is $\boldsymbol{G}^N/\phi^3\boldsymbol{G}^N$ or $\boldsymbol{Z}^N/4\boldsymbol{Z}^N$.

In fact, regular labelings (although not necessarily regular Ungerboeck labelings) exist for all partitions used in all the codes covered in this paper.

V. Known Classes of Trellis Codes

We shall now categorize the principal classes of trellis codes that have so far appeared in the literature according to the parameters of the previous sections. In the next section, we give further generic classes. We shall then compare and contrast all of those schemes, including lattice codes.

Ungerboeck [8] developed classes of one- and two-dimensional trellis codes using rate-$k/(k+1)$ binary convolutional codes. From the viewpoint of this paper, his one-dimensional schemes are based on the four-way partition $\boldsymbol{Z}/4\boldsymbol{Z}$ of the integers into the four residue classes modulo 4, in combination with a binary rate-1/2 convolutional coder to select cosets of $4\boldsymbol{Z}$. His two-dimensional schemes for rectangular constellations, which have the greatest practical importance, are based on either the four-way partition $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$ in combination with a rate-1/2 convolutional encoder, or the eight-way partition $\boldsymbol{Z}^2/2R\boldsymbol{Z}^2$ with a rate-2/3 convolutional encoder. He also gives codes using phase-modulated constellations that are based on similar principles and may be regarded as coset codes (see Section I-C), but that will not be covered here.

Table IV gives the characteristics of the Ungerboeck one- and two-dimensional schemes. The codes achieve increasing $d^2_{\min}$ as the number 2^ν of states increases from 4 to 512, up to the maximum possible value of $d^2_{\min}(\Lambda')=2^\mu$. (*Note:* We use the codes listed in [21], where minor corrections have been made in the earlier code tables.) The depth is $\mu=4$ for the one-dimensional schemes, while in two dimensions $\mu=2$ or 3. The redundancy r is one for both classes, but the normalized redundancy ρ (per two dimensions) is thus two in the one-dimensional case, versus one in the two-dimensional case. The fundamental coding gain γ is given by the formula $2^{-\rho}d^2_{\min}$ and is also given in decibels. N_0 is the number of nearest neighbors, and $\tilde{N}_0=2N_0/N$ is the error coefficient normalized to two dimensions. N_D is the number of decoding operations using the trellis-based decoding algorithms of the partition Λ/Λ' whose complexity is given in Table III, followed by a conventional Viterbi algorithm for the convolutional code, and $\tilde{N}_D=2N_D/N$ is the decoding complexity per two dimensions. (For each unit of time, for each of the 2^ν states, the Viterbi algorithm requires 2^k additions and a comparison of 2^k numbers, or 2^k-1 binary comparisons, so that its complexity is $\beta 2^{k+\nu}$, where $\beta=2-2^{-k}$, and $2^{k+\nu}$ is the number of branches per stage of the trellis, which is the measure of complexity used by Ungerboeck [21], following Wei [11].)

The error coefficient reduces the effective coding gain by an amount that depends on the steepness of the error probability curve. In this paper, we will use the rule of thumb that every factor of two increase in the error coefficient reduces the coding gain by about 0.2 dB (at error rates of the order of 10^{-6}); this will enable us to compute an effective coding gain γ_{eff} (in dB), normalized for the error coefficient $\tilde{N}_0$. In principle, the error coefficients at every distance ought to be considered, and the effective coding gain evaluated in the same way for each; if

TABLE IV
Ungerboeck Codes

N	Λ	Λ'	2^ν	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$
				One-dimensional						
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	4	1/2	2	9	2.25	3.52	8	24
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	8	1/2	2	10	2.5	3.98	8	48
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	16	1/2	2	11	2.75	4.39	16	96
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	32	1/2	2	13	3.25	5.12	24	192
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	64	1/2	2	14	3.5	5.44	72	384
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	128	1/2	2	16	4	6.02	132	768
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	256	1/2	2	16	4	6.02	4	1536
1	$\boldsymbol{Z}$	$4\boldsymbol{Z}$	512	1/2	2	16	4	6.02	4	3072
				Two-dimensional						
2	$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	4	1/2	1	4	2	3.01	4	16
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	8	2/3	1	5	2.5	3.98	16	64
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	16	2/3	1	6	3	4.77	56	120
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	32	2/3	1	6	3	4.77	16	232
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	64	2/3	1	7	3.5	5.44	56	456
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	128	2/3	1	8	4	6.02	344	902
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	256	2/3	1	8	4	6.02	44	1800
2	$\boldsymbol{Z}^2$	$2R\boldsymbol{Z}^2$	512	2/3	1	8	4	6.02	4	3592

they grow too large too rapidly, they can dominate performance. We will not go beyond the error coefficient $\tilde{N}_0$ in this paper, except for Ungerboeck-type codes, where we can present results of Eyuboglu and Li (unpublished) that take into account the next two normalized coefficients, $\tilde{N}_1$ and $\tilde{N}_2$.

Honig [22] has performed a search for one-dimensional Ungerboeck-type codes based on the four-way partition $\boldsymbol{Z}/4\boldsymbol{Z}$, using a criterion that includes the effect of the error coefficient, and has obtained an improvement at 64 states (the apparently improved 16-state code is actually catastrophic). Similarly, Pottie and Taylor [23] have searched for two-dimensional Ungerboeck-type codes based on the eight-way partition $\boldsymbol{Z}^2/2R\boldsymbol{Z}^2$ and have obtained improvements at 64 and 128 states; the 128-state code has a lower fundamental coding gain γ but a greater effective coding gain γ_{eff} due to its much lower error coefficient. Finally, Eyuboglu and Li have made a reasonably exhaustive search for the best codes of both classes in terms of the effective coding gain criterion, for up to 256 states, with modest improvements at as few as 16 states.

Table V gives the performance parameters $d^2_{\min}$, $\tilde{N}_0$, $\tilde{N}_1$, $\tilde{N}_2$ and the consequent effective coding gains γ_{eff} for a number of the codes of Ungerboeck, Honig, Pottie and Taylor, and Eyuboglu and Li. When the dominant error coefficient is other than $\tilde{N}_0$, it is starred. The parity-check polynomials $\boldsymbol{h}^j$ for these codes are also given, in the octal notation of Ungerboeck [8], [21]. All parameters not given (including decoding complexity) are the same as for the Ungerboeck code with the same number of states.

Fig. 12(a) plots the effective coding gain γ_{eff} versus the normalized complexity $\tilde{N}_D$ for the best of these Ungerboeck-type one- and two-dimensional codes. We see that the graphs are fairly linear on this log–log plot over most of their range. An increase of a factor of two in complexity yields an increase of about 0.4 dB in coding gain, until the effective coding gain passes 5 dB. The one-dimensional codes are of the order of 0.2 dB better over this linear range (however, the two-dimensional codes have been generally preferred in practice because their constellation expansion factor 2^ρ is only two, not four).

The first multidimensional code seems to have been developed by Gallager [1]. In this code, an eight-state rate-3/4 convolutional encoder selects two successive cosets from the four-way two-dimensional partition $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$ or, equivalently, one coset from the 16-way four-dimensional lattice partition $\boldsymbol{Z}^4/2\boldsymbol{Z}^4$ (equivalently, four successive cosets from the two-way one-dimensional partition $\boldsymbol{Z}/2\boldsymbol{Z}$). The basic idea, as in Lemma 6, is that with such partitions the minimum squared distance $d^2_{\min}$ between code sequences is simply the minimum Hamming distance d_H of the binary code, as long as $d_H \leq 4$. Quite independently, Calderbank and Sloane [10] discovered a very similar code, although with improved error coefficient $\tilde{N}_0$ due to the choice of an eight-state rate-3/4 binary code with a lower error coefficient. We shall call codes based

TABLE V
EFFECTIVE CODING GAINS

2^ν	h^2	h^1	h^0	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_1$	$\tilde{N}_2$	γ_{eff} (dB)	Type
					For $\boldsymbol{Z}/4\boldsymbol{Z}$ codes						
4		2	5	9	2.25	3.52	8	16	32	3.32	U
8		04	13	10	2.5	3.98	8	16	32	3.78	U
16		04	23	11	2.75	4.39	16	16	32	3.99	U
16		10	23	11	2.75	4.39	8	16	48	4.19	EL
32		10	45	13	3.25	5.12	24	56	112	4.60	U
64		024	103	14	3.5	5.44	72	0	180	4.61	U
64		054	161	14	3.5	5.44	16	*64	132	4.94	H
128		126	235	16	4	6.02	132	0	512	5.01	U
128		160	267	15	3.75	5.74	16	68	*200	5.16	EL
128		124	207	14	3.5	5.44	8	16	28	5.24	EL
256		362	515	16	4	6.02	4	64	*160	5.47	U
256		370	515	15	3.75	5.74	8	12	*80	5.42	EL
512		0342	1017	16	4	6.02	4	0	*112	5.57	U
					For $\boldsymbol{Z}^2/2R\boldsymbol{Z}^2$ codes						
4	—	2	5	4	2	3.01	4	32	128	3.01	U
8	04	02	11	5	2.5	3.98	16	72	320	3.58	U
16	16	04	23	6	3	4.77	56	160	820	4.01	U
32	10	06	41	6	3	4.77	16	104	404	4.37	U
32	34	16	45	6	3	4.77	8	*128	404	4.44	EL
64	064	016	101	7	3.5	5.44	56	260	1008	4.68	U
64	060	004	143	7	3.5	5.44	48	292	1184	4.72	PT
64	036	052	115	7	3.5	5.44	40	252	992	4.78	EL
128	042	014	203	8	4	6.02	344	0	5900	4.74	U
128	056	150	223	8	4	6.02	172	624	2568	4.94	EL
128	024	100	245	7	3.5	5.44	8	*188	968	4.91	PT
128	164	142	263	7	3.5	5.44	8	*132	752	5.01	EL
256	304	056	401	8	4	6.02	44	*304	1316	5.28	U
256	370	272	417	8	4	6.02	36	*308	1224	5.28	EL
256	274	162	401	7	3.5	5.44	4	*64	248	5.22	EL
512	0510	0346	1001	8	4	6.02	4	128	*700	5.50	U

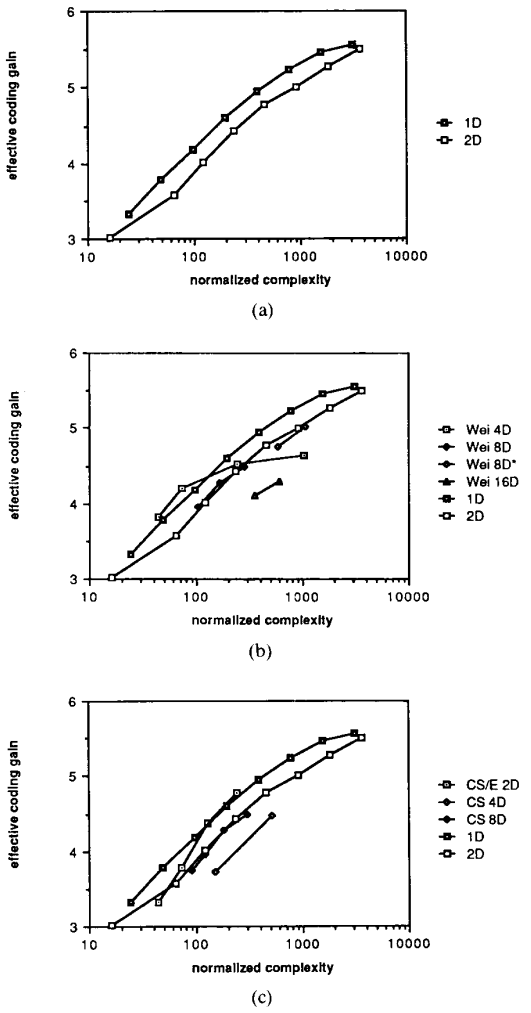


Fig. 12. Performance versus complexity. (a) For Ungerboeck-type one-dimensional and two-dimensional codes (as improved by Eyuboglu and Li). (b) For Wei codes. (c) For Calderbank–Sloane-type codes.

on partitions $\boldsymbol{Z}^N/2\boldsymbol{Z}^N$ (with $N > 2$) GCS-type codes. Table VI gives the parameters of the GCS-type code just described, with the error coefficient for the Calderbank–Sloane (CS) version (as given in [13]).

Wei [11] has developed a variety of multidimensional codes. We shall say that a code is "Wei-type" if Λ/Λ' is a lattice partition where Λ' is a denser lattice than Λ. Wei stresses that Λ' should be dense to maximize coding gain and to simplify code construction but that Λ should have low redundancy $\rho(\Lambda)$ to maximize coding gain and to minimize the constellation expansion factor. Wei is also willing to increase the number of states 2^r (and decoding complexity) to minimize the error coefficient $\tilde{N}_0$, even when the fundamental coding gain γ is not improved thereby. Table VII gives the parameters for the codes discussed in [11]; Wei also mentions that any of his codes that are based on the 16-way partition $\boldsymbol{Z}^8/E_8$ can be translated into a code based on the 16-way partition E_8/RE_8, with the same error coefficient, but with twice the constellation size. (Wei's lattice "DE_8" is here called $D_8^{\perp}$, in keeping with the notation of Section III.) Note that the 32-state code based on $\boldsymbol{Z}^4/2\boldsymbol{Z}^4$ is a GCS-type code. Indeed, all of the codes in which Λ' is a lattice of depth 2 are equivalent to GCS-type codes in view of Lemma 6; in particular, the eight-state code based on $\boldsymbol{Z}^4/RD_4$ is equivalent to the CS version of the GCS-type code just described, although its decoding complexity is less because of the symmetries of RD_4 (the blank error coefficients are unknown but large).

Ungerboeck [21] has found additional 128-state Wei-type codes that extend the four- and eight-dimensional families; these codes are listed in Table VIII. Fig. 12(b) plots the effective coding gain γ_{eff} versus the normalized complexity $\tilde{N}_D$ for the Wei codes for which the error coefficient is known, versus the codes of Fig. 12(a) as benchmarks.

Calderbank and Sloane [12], [13] have also developed a number of multidimensional codes. We shall say that a

TABLE VI
GCS-TYPE CODE

N	Λ	Λ'	2^r	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff}(dB)
4	$\boldsymbol{Z}^4$	$2\boldsymbol{Z}^4$	8	3/4	1/2	4	$2^{3/2}$	4.52	44	64	3.82

TABLE VII
WEI CODES

N	Λ	Λ'	2^r	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff}(dB)
4	$\boldsymbol{Z}^4$	RD_4	8	2/3	1/2	4	$2^{3/2}$	4.52	44	44	3.82
4	$\boldsymbol{Z}^4$	RD_4	16	2/3	1/2	4	$2^{3/2}$	4.52	12	72	4.20
4	$\boldsymbol{Z}^4$	$2\boldsymbol{Z}^4$	32	3/4	1/2	4	$2^{3/2}$	4.52	4	244	4.52
4	$\boldsymbol{Z}^4$	$2D_4$	64	4/5	1/2	5	$5/2^{1/2}$	5.48	72	1048	4.65
8	$\boldsymbol{Z}^8$	E_8	16	3/4	1/4	4	$2^{7/4}$	5.27	316	104	4.01
8	$\boldsymbol{Z}^8$	E_8	32	3/4	1/4	4	$2^{7/4}$	5.27	124	164	4.28
8	$\boldsymbol{Z}^8$	E_8	64	3/4	1/4	4	$2^{7/4}$	5.27	60	284	4.49
16	$\boldsymbol{Z}^{16}$	H_{16}	32	4/5	1/8	4	$2^{15/8}$	5.64		228	
16	$\boldsymbol{Z}^{16}$	H_{16}	64	4/5	1/8	4	$2^{15/8}$	5.64	796	352	4.12
16	$\boldsymbol{Z}^{16}$	H_{16}	128	4/5	1/8	4	$2^{15/8}$	5.64	412	600	4.31
8	$D_8^{\perp}$	RE_8	32	4/5	1	8	4	6.02		336	
8	$D_8^{\perp}$	RE_8	64	4/5	1	8	4	6.02	316	584	4.76
8	$D_8^{\perp}$	RE_8	128	4/5	1	8	4	6.02	124	1080	5.03

TABLE VIII
FURTHER WEI-TYPE CODES (UNGERBOECK)

N	Λ	Λ'	2^ν	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff}(dB)
4	$\boldsymbol{Z}^4$	$2D_4$	128	4/5	1/2	6	$6/2^{1/2}$	6.28	728	2040	4.77
8	$\boldsymbol{Z}^8$	RD_8	128	4/5	1/4	4	$2^{7/4}$	5.27	28	1032	4.71

TABLE IX
CALDERBANK–SLOANE CODES

N	Λ	Λ'	2^ν	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff}(dB)
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	4	2/4	2	8	2	3.01	4	44	3.01
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	8	2/4	2	11	2.75	4.39	32	72	3.79
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	16	2/4	2	12	3	4.77	48	128	4.05
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	64	2/4	2	14	3.5	5.44	48	464	4.72
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	128	2/4	2	16	4	6.02	228	912	4.85
4	D_4	$2D_4$	16	3/4	1	6	3	4.77	152	152	3.72
4	D_4	$2D_4$	64	3/4	1	8	4	6.02	828	512	4.48
8	E_8	RE_8	8	3/4	5/4	8	$2^{7/4}$	5.27	764	90	3.75
8	E_8	RE_8	16	3/4	5/4	8	$2^{7/4}$	5.27	316	120	4.01
8	E_8	RE_8	32	3/4	5/4	8	$2^{7/4}$	5.27	124	180	4.28
8	E_8	RE_8	64	3/4	5/4	8	$2^{7/4}$	5.27	60	300	4.49

Note added in proof: J. Chow (private communication) has obtained values of $\tilde{N}_0 = 88$ and $\gamma_{\text{eff}} = 3.88$ dB for the 16-state D_4/RD_4 code, and of $d^2_{\min} = 6$, $\tilde{N}_0 = 16$, and $\gamma_{\text{eff}} = 4.37$ dB for the 64-state D_4/RD_4 code.

TABLE X
FURTHER $\boldsymbol{Z}^2/4\boldsymbol{Z}^2$ CODES (EYUBOGLU)

N	Λ	Λ'	2^ν	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff}(dB)
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	4	2/4	2	9	2.25	3.52	8	44	3.32
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	16	2/4	2	12	3	4.77	16	128	4.37
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	32	2/4	2	12	3	4.77	4	240	4.77
2	$\boldsymbol{Z}^2$	$4\boldsymbol{Z}^2$	32	2/4	2	13	3.25	5.12	16	240	4.72

code is "CS-type" if Λ/Λ' is a lattice partition where Λ and Λ' are versions of the same lattice. Some of their codes are shown in Table IX. (They also consider the following: Ungerboeck-type codes based on partitions $\boldsymbol{Z}/4\boldsymbol{Z}$, $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$ and $\boldsymbol{Z}^2/2R\boldsymbol{Z}^2$, but without improvement over Ungerboeck either in fundamental coding gain γ or in error coefficient $\tilde{N}_0$, except for the $\nu = 6$ case also found by Pottie and Taylor; the GCS-type code based on the partition $\boldsymbol{Z}^4/2\boldsymbol{Z}^4$, as previously mentioned; and codes using the nonbinary two-dimensional hexagonal lattice A_2, for which the results are not particularly encouraging. The last three codes appear to be equivalent to the aforementioned translation of Wei's $\boldsymbol{Z}^8/E_8$ codes.)

Finally, Eyuboglu has also searched for codes based on the 16-way two-dimensional partition $\boldsymbol{Z}^2/4\boldsymbol{Z}^2$. The additional codes found that improve on codes already listed are summarized in Table X.

Fig. 12(c) plots the effective coding gain γ_{eff} versus the normalized complexity $\tilde{N}_D$ for the Calderbank–Sloane codes, as improved by the codes in Table X (up to 32 states), again with the codes of Fig. 12(a) for comparison. Note that the $\boldsymbol{Z}^2/4\boldsymbol{Z}^2$ codes ought to be compared to the one-dimensional $\boldsymbol{Z}/4\boldsymbol{Z}$ codes, since they have the same redundancy and depth, and in fact include the latter as a subset; their performance improvement is in fact small, and, taking complexity into account, they are no better.

VI. FURTHER CLASSES OF TRELLIS CODES

In this section we present a number of additional generic classes of codes that can be described relatively simply. Our objective is more to round out the picture than to improve on earlier results; in general, these codes have parameters comparable to those of the known codes of the previous section (or indeed of lattice codes). Our main point, in fact, is that "there are many ways to modulate," and that the complexity of the encoder and decoder required to achieve a given coding gain and error coefficient remains remarkably constant across a wide variety of codes.

We describe eight different classes of codes $\mathcal{C}(\Lambda/\Lambda'; C)$, based on all possible choices of the three following binary characteristics. The codes are based either on a lattice partition Λ/Λ' with minimum squared distances $d^2_{\min}(\Lambda)/d^2_{\min}(\Lambda')$ in the ratio 1:2, or on a partition chain $\Lambda/\Lambda'/\Lambda''$ with distances $d^2_{\min}(\Lambda)/d^2_{\min}(\Lambda')/d^2_{\min}(\Lambda'')$ in the ratio 1:2:4 (in the latter case, we use an Ungerboeck labeling and exploit the Ungerboeck distance bound). The convolutional code C is either a rate-$k/2k$ code, with $|\Lambda/\Lambda'| = 2^{2k}$ or else $|\Lambda/\Lambda'| = |\Lambda'/\Lambda''| = 2^k$, or a rate-$k/(k+1)$ code, with $|\Lambda/\Lambda'| = 2^{k+1}$ or else $|\Lambda/\Lambda'| = 2$ and $|\Lambda'/\Lambda''| = 2^k$. The constraint length ν of C is either k or $2k$, with each of the k input bits being held in memory for

either one or two time units. All codes are *noncatastrophic*, i.e., there is no infinite sequence of nonzero inputs that leads to a finite sequence of nonzero outputs.

The resulting codes have many characteristics similar to those of the Wei and Calderbank–Sloane codes, as well as the four-state two-dimensional Ungerboeck code. Except for error coefficient, the rate-$k/2k$ codes also have parameters resembling those of the Barnes–Wall lattices.

Class I Codes

Let Λ/Λ' be a 2^{2k}-way lattice partition with $d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$. Let C be a "unit-memory" rate-$k/2k$ binary convolutional encoder, as shown in Fig. 13(a). In each time unit k information bits enter the encoder and are stored for one time unit; the encoder output is the combination of the k current and k previous bits, or b bits altogether, which select one of the 2^{2k} cosets of Λ'. The encoder has 2^k states, and the code has a trellis diagram in which every current state is connected to every next state, so there are a total of 2^{2k} branches in each time unit, one corresponding to each coset of Λ'. The code is thus not catastrophic, because only one branch corresponds to the zero coset of Λ' (Λ' itself). Any two paths through the trellis must differ in at least two time units, so the minimum squared distance between paths is $2d^2_{\min}(\Lambda)$, which is the same as the minimum squared distance $d^2_{\min}(\Lambda')$ within any coset of Λ'; thus $d^2_{\min}(\mathbb{C}) = d^2_{\min}(\Lambda')$.

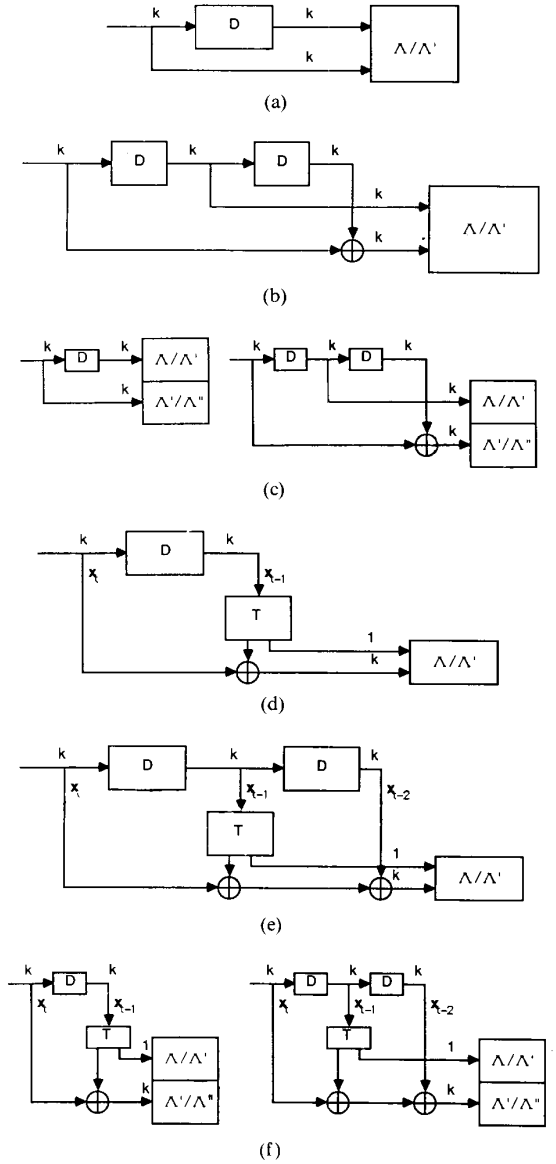


Fig. 13. Encoders. (a) For Class I code. (b) For Class II codes. (c) For Class III and IV codes. (d) For Class V codes. (e) For Class VI codes. (f) For Class VII and VIII codes.

The multiplicity N_0 of sequences at distance $d^2_{\min}(\mathbb{C})$ from a given sequence that start at a given time is

$$N_0 = N_{\Lambda'} + (2^k - 1)N_\Lambda^2$$

where $N_{\Lambda'}$ is the number of points of weight $d^2_{\min}(\Lambda')$ in Λ', and N_Λ is the number of points of weight $d^2_{\min}(\Lambda)$ in any nonzero coset of Λ' (if it is the same for all such cosets).

Table XI gives the parameters for Class I codes based on the partitions $\boldsymbol{Z}^4/R\boldsymbol{Z}^4$, D_4/RD_4, E_8/RE_8, and $\Lambda_{16}/R\Lambda_{16}$, which have orders 4, 4, 16, and 256, and depths 1, 2, 3, and 4, respectively (the D_4/RD_4 code was developed as a phase-modulated code by Divsalar and Simon [24]). These codes are closely related to the lattices D_4, E_8, Λ_{16}, and Λ_{32}, and have the same principal parameters μ, ρ, and γ. In the case of the two-state code based on $\boldsymbol{Z}^4/R\boldsymbol{Z}^4$, it is possible to make the minimum squared distance between distinct paths equal to $3d^2_{\min}(\Lambda)$ (use the partition chain $\boldsymbol{Z}^4/D_4/R\boldsymbol{Z}^4$, and let one of the two bits control each of these two-way partitions; see Class III), so that the error coefficient achieves its minimum possible value, $\tilde{N}_0 = 4$.

Class II Codes

Let Λ/Λ' again be a 2^{2k}-way lattice partition with $d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$. Let C be a rate-1/2, 2^{2k}-state convolutional encoder as shown in Fig. 13(b), with k information bits entering in each time unit, two units of memory, and $2k$ output bits, k representing the inputs one time unit earlier, and $2k$ representing the mod-2 sum of the current inputs with those two time units earlier, which together select one of the 2^{2k} cosets of Λ'. The trellis diagram has 2^{2k} states, with 2^k branches leaving and entering every state. Because the distance between paths is at least $d^2_{\min}(\Lambda)$ when they diverge, $d^2_{\min}(\Lambda)$ when they merge, and $d^2_{\min}(\Lambda)$ in some other time unit, the distance between distinct paths is at least $3d^2_{\min}(\Lambda)$. Thus $d^2_{\min}(\mathbb{C}) = d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$, the distance within cosets of Λ', and the error coefficient $\tilde{N}_0$ is the same as that of Λ' (the minimum possible for any coset code based on the partition Λ/Λ'). The code is noncatastrophic because every nonzero input creates a nonzero output one time unit later.

TABLE XI

N	Λ	Λ'	2^r	$k/(k+r)$	ρ	$d^2_{\min}$	γ	dB	$\tilde{N}_0$	$\tilde{N}_D$	γ_{eff} (dB)
					Class I codes						
2	$\boldsymbol{Z}^4$	$R\boldsymbol{Z}^4$	2	1/2	1/2	2	$2^{1/2}$	1.51	4	5	1.51
2	D_4	RD_4	2	1/2	1	4	2	3.01	44	13	2.32
4	E_8	RE_8	4	2/4	3/2	8	$2^{3/2}$	4.52	252	67	3.24
8	Λ_{16}	$R\Lambda_{16}$	16	4/8	2	16	4	6.02		1614	
					Class II codes						
4	D_4	RD_4	4	1/2	1	4	2	3.01	12	16	2.69
8	E_8	RE_8	16	2/4	3/2	8	$2^{3/2}$	4.52	60	88	3.73
16	Λ_{16}	$R\Lambda_{16}$	256	4/8	2	16	4	6.02	540	2544	4.61
					Class III codes						
2	$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	2	1/2	1	3	3/2	1.76	8	8	1.56
4	D_4	$2D_4$	4	2/4	3/2	6	$3/2^{1/2}$	3.27		46	
8	E_8	$2E_8$	16	4/8	2	16	4	6.02	1020	684	4.42
					Class IV codes						
2	$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	4	1/2	1	4	2	3.01	4	16	3.01
4	D_4	$2D_4$	16	2/4	3/2	8	$2^{3/2}$	4.52	12	88	4.20
8	E_8	$2E_8$	256	4/8	2	16	4	6.02	60	2544	5.24
					Class V codes						
4	$\boldsymbol{Z}^4$	$R\boldsymbol{Z}^4$	2	1/2	1/2	2	$2^{1/2}$	1.51	4	5	1.51
4	D_4	RD_4	2	1/2	1	4	2	3.01	44	13	2.32
8	E_8	RE_8	8	3/4	5/4	8	$2^{7/4}$	5.27	764	90	3.75
16	Λ_{16}	$R\Lambda_{16}$	128	7/8	13/8	16	$2^{19/8}$	7.15		5632	
8	D_8	E_8	4	2/3	1/2	4	$2^{3/2}$	4.52	316	37	3.25
16	D_{16}	H_{16}	8	3/4	1/4	4	$2^{7/4}$	5.27	1692	89	3.52
16	H_{16}	Λ_{16}	64	6/7	3/4	8	$2^{9/4}$	6.77		1792	
					Class VI codes						
4	D_4	RD_4	4	1/2	1	4	2	3.01	12	16	2.69
8	E_8	RE_8	64	3/4	5/4	8	$2^{7/4}$	5.27	60	300	4.49
16	Λ_{16}	$R\Lambda_{16}$	2^{14}	7/8	13/8	16	$2^{19/8}$	7.15	540	$\sim 2^{19}$	5.73
8	D_8	E_8	16	2/3	1/2	4	$2^{3/2}$	4.52	60	58	3.73
8	D_{16}	H_{16}	64	3/4	1/4	4	$2^{7/4}$	5.27	284	194	4.04
16	H_{16}	Λ_{16}	2^{12}	6/7	3/4	8	$2^{9/4}$	6.77	540	$\sim 2^{16}$	5.36
					Class VII codes						
2	$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	2	1/2	1	3	3/2	1.76	8	8	1.56
4	$\boldsymbol{Z}^4$	RD_4	4	2/3	1/2	3	$3/2^{1/2}$	3.27	24	30	2.75
8	$\boldsymbol{Z}^8$	E_8	8	3/4	1/4	3	$3/2^{1/4}$	4.02		74	
16	$\boldsymbol{Z}^{16}$	H_{16}	16	4/5	1/8	3	$3/2^{1/8}$	4.39		166	
8	$D_8^\perp$	RE_8	16	4/5	1	6	3	4.77		212	
					Class VIII codes						
2	$\boldsymbol{Z}^2$	$2\boldsymbol{Z}^2$	4	1/2	1	4	2	3.01	4	16	3.01
4	$\boldsymbol{Z}^4$	RD_4	16	2/3	1/2	4	$2^{3/2}$	4.52	12	72	4.20
8	$\boldsymbol{Z}^8$	E_8	64	3/4	1/4	4	$2^{7/4}$	5.27	60	284	4.49
16	$\boldsymbol{Z}^{16}$	H_{16}	256	4/5	1/8	4	$2^{15/8}$	5.64	284	1096	4.41
8	$D_8^\perp$	RE_8	256	4/5	1	8	4	6.02	60	2072	5.24

Table XI gives the parameters for Class II codes based on the partitions D_4/RD_4, E_8/RE_8, and $\Lambda_{16}/R\Lambda_{16}$. Their parameters, including coding gain, are the same as those of Class I codes, except that the increase of the number of states to equal the order 2^{2k} of the partition Λ/Λ' results in a reduction of the error coefficient $\tilde{N}_0$ to its minimum value. The decoding complexity increases only modestly because it is dominated by the complexity of decoding Λ/Λ'.

Class III and IV Codes

Let $\Lambda/\Lambda'/\Lambda''$ now be a chain of two 2^k-way partitions with distances $d^2_{\min}(\Lambda'') = 2d^2_{\min}(\Lambda') = 4d^2_{\min}(\Lambda)$. For a Class III code (resp. Class IV), let C be the same rate-$k/2k$ encoder as in Class I (resp. Class II), but now let the first set of k output bits select the coset of Λ' in the partition Λ/Λ', and the second set of k output bits select the coset of Λ'' in the partition Λ'/Λ'', as shown in Fig. 13(c). These encoders are still noncatastrophic.

Because the label selecting the coset of Λ' in the partition Λ/Λ' is zero at the time of the first nonzero input, the set of possible initial code symbols (the time-zero lattice Λ_0) is Λ', so the minimum squared distance between distinct paths is $d^2_{\min}(\Lambda')$ where they first diverge. In the case of Class III codes, the minimum squared distance between distinct paths is $d^2_{\min}(\Lambda)$ where they merge, so $d^2_{\min}(\mathbb{C}) = 3d^2_{\min}(\Lambda)$. In the case of Class IV codes, the set of possible final code symbols is also Λ', so that the minimum squared distance between distinct paths is

$d^2_{\min}(\Lambda')$ where they merge, and also at least $d^2_{\min}(\Lambda)$ at some other time, so that the $d^2_{\min}$ between distinct paths is $5d^2_{\min}(\Lambda)$. This means that $d^2_{\min}(\mathbb{C}) = d^2_{\min}(\Lambda'') = 4d^2_{\min}(\Lambda)$, and furthermore the normalized error coefficient $\tilde{N}_0$ is the same as that of Λ''.

Table XI gives the parameters for Class III codes based on the partitions $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2$, $D_4/RD_4/2D_4$, and $E_8/RE_8/2E_8$, which have orders 4, 16, and 256, and depths 2, 3, and 4, respectively. The first is a noncatastrophic Ungerboeck-type two-state code with $\gamma = 1.5$ (1.76 dB) (this code appears as a phase-modulated code in Divsalar *et al.* [25]). The others are CS-type codes. For the $E_8/RE_8/2E_8$ code, we use the fact that there exists an alternative partition $E_8/R^*E_8/2E_8$, where R^*E_8, like RE_8, is a version of E_8 with $d^2_{\min} = 8$, such that the system of coset representatives $[R^*E_8/2E_8]$ is also a system of coset representatives for E_8/RE_8 (see part II); this ensures that $d^2_{\min} = 8$ when paths merge as well as when they diverge, so that $d^2_{\min}(\mathbb{C}) = d^2_{\min}(2E_8) = 16$.

Table XI gives the parameters for Class IV codes based on the same partitions $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2$, $D_4/RD_4/2D_4$, and $E_8/RE_8/2E_8$. The first is Ungerboeck's four-state code, which is the prototype of this class. These are CS-type codes that are closely related to Class III codes, except that they have twice the constraint length and 4/3 the minimum squared distance and coding gain (except in the last case, where only the error coefficient improves). Again, these codes are closely related to the lattices E_8, Λ_{16}, and Λ_{32} and to the corresponding Class I and II codes and have the same principal parameters. In fact, even the decoding complexity is the same as that of Class II codes, but the error coefficient is still further reduced.

Class V Codes

Let Λ/Λ' now be a 2^{k+1}-way partition with distances $d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$, and let C be a rate-$k/(k+1)$ convolutional encoder as shown in Fig. 13(d), with k information bits entering in each time unit, and $k+1$ output bits generated as follows. Let T be a linear (modulo 2) circuit with k input bits, namely the k-tuple $\boldsymbol{x}_{t-1}$ of input bits delayed by one time unit, and $k+1$ output bits. One output bit goes directly to the coset selector; the remaining k bits are added (modulo 2) to the k-tuple $\boldsymbol{x}_t$, and the k-bit sum forms the remaining inputs to the coset selector.

The circuit T need have only the following two properties: a) its outputs are all-zero only when its inputs are all-zero, and b) there is no infinite input sequence $(\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots)$ into C that generates a finite output sequence from C (so that the code is noncatastrophic). A simple circuit T with these properties is the one whose output is simply the $(k+1)$-tuple $(\boldsymbol{x}_{t-1}, 0)$, where the leftmost bit is the one that goes directly into the coset selector. Property a) is obvious. Property b) follows from the fact that if $\boldsymbol{x}_0 \neq \boldsymbol{0}$, then there is no sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$ such that $(\boldsymbol{x}_0, 0) \oplus (0, \boldsymbol{x}_1) = \boldsymbol{0}$, $(\boldsymbol{x}_1, 0) \oplus (0, \boldsymbol{x}_2) = \boldsymbol{0}$, etc., since $\boldsymbol{x}_1$ can only match the $k-1$ low-order bits in $\boldsymbol{x}_0$, $\boldsymbol{x}_2$ can then only match the $k-2$ low-order bits in $\boldsymbol{x}_0$, etc., and so eventually the highest order nonzero bit in $\boldsymbol{x}_0$ "shifts" to the highest order position, where it cannot be matched.

Property a) ensures that the minimum squared distance is at least $d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$, because two distinct paths differ by at least $d^2_{\min}(\Lambda)$ where they diverge and $d^2_{\min}(\Lambda)$ where they merge. The multiplicity N_0 of sequences at distance $d^2_{\min}(\Lambda')$ from any given sequence starting at any given time is

$$N_0 = N_{\Lambda'} + (2^{k+1} - k - 2) N_\Lambda^2$$

where $N_{\Lambda'}$ is the number of points of weight $d^2_{\min}(\Lambda')$ in Λ', and N_Λ is the number of points of weight $d^2_{\min}(\Lambda)$ in any nonzero coset of Λ' (if it is the same for all such cosets). The coefficient of N_Λ^2 follows from the observation that in the code trellis, starting from a given zero state and ending at some later zero state, there are $2^k - 1$ nonzero paths of length 2, $2^{k-1} - 1$ nonzero paths of length 3, and so forth, up to $2 - 1 = 1$ nonzero path of length $k+1$, so that the total number of nonzero paths is $2^{k+1} - k - 2$ (this is generally true for any noncatastrophic rate-$k/(k+1)$ encoder; see Forney [26]).

Table XI gives the parameters for Class V codes based on the partitions $\boldsymbol{Z}^4/R\boldsymbol{Z}^4$, D_4/RD_4, E_8/RE_8, and $\Lambda_{16}/R\Lambda_{16}$, as in Class I, as well as D_8/E_8, D_{16}/H_{16}, and H_{16}/Λ_{16}. The first two are just the two-state Class I codes again, since $k+1 = 2k = 2$. The third code is a code equivalent to the eight-state CS code based on E_8/RE_8, which may be considered as the prototype of this class. The fourth is a 128-state code with a coding gain in excess of 7 dB, but with a huge error coefficient and decoding complexity. The last three further illustrate that Class V codes attain large coding gains for relatively few states (and thus small decoding complexity) but with outsize error coefficients: the fifth code attains $\gamma = 2^{3/2}$ (4.52 dB) with only four states, while the last gets considerably beyond 6 dB with only 64 states. (In unpublished work, Wei had earlier constructed codes based on such 16-dimensional partitions as H_{16}/Λ_{16}, with comparable coding gains.) Even if the effective number of states is taken as the order of the partition Λ/Λ' rather than 2^ν (since the decoding complexity is dominated by decoding Λ/Λ'), so that the effective number of states doubles, these are still very good (ignoring $\tilde{N}_0$).

Class VI Codes

Once again, let Λ/Λ' be a 2^{k+1}-way partition with distances $d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$, but now let C be a rate-$k/(k+1)$ convolutional encoder as shown in Fig. 13(e). This is the same as the Class V encoder, including the circuit T, except that there is a second memory element, and its output $\boldsymbol{x}_{t-2}$ is further added to the k-tuple output of the encoder of Fig. 13(d).

Property b) of circuit T again ensures that the code is noncatastrophic. Furthermore, it ensures that if the input sequence has a finite number of nonzero $\boldsymbol{x}_t$, then the encoder outputs are nonzero at at least three different times: once when the sequence begins, once at some inter-

mediate time when T has a nonzero output, and once when the last nonzero input finally leaves the encoder. Consequently, the minimum distance between two distinct paths is at least $3d^2_{\min}(\Lambda)$, so that the minimum distance of the code is $d^2_{\min}(\mathbb{C}) = d^2_{\min}(\Lambda') = 2d^2_{\min}(\Lambda)$, and the error coefficient $\tilde{N}_0$ is the same as that of Λ'.

Table XI gives the parameters for Class VI codes based on the same partitions as for Class V (except for $\boldsymbol{Z}^4/R\boldsymbol{Z}^4$, where no improvement is achieved). The four-state D_4/RD_4 code is again a Class II code, and the 64-state E_8/RE_8 code is equivalent to the corresponding Wei/Calderbank–Sloane code. The remaining codes illustrate that Class VI codes have the same coding gains as Class V codes, but with reasonable error coefficients, at the cost of increased decoding complexity (vastly increased, for those codes with gains more than 6 dB).

Class VII and VIII Codes

Now let $\Lambda/\Lambda'/\Lambda''$ be a two-level partition chain with distances $d^2_{\min}(\Lambda'') = 2d^2_{\min}(\Lambda') = 4d^2_{\min}(\Lambda)$ and orders 2 and 2^k. Let C be a rate-$k/(k+1)$ encoder as in Fig. 13(d) and (e), but with one output bit selecting one of the two cosets of Λ' in the partition Λ/Λ', and the remaining k bits selecting a coset of Λ'' in the partition Λ'/Λ'', as shown in Fig. 13(f).

Again, the codes are noncatastrophic. With Class VII codes, as with Class III, two distinct paths differ by at least $d^2_{\min}(\Lambda')$ where they diverge and $d^2_{\min}(\Lambda)$ where they merge, so that $d^2_{\min}(\mathbb{C}) = 3d^2_{\min}(\Lambda)$. With Class VIII codes, as with Class IV, two distinct paths differ by at least $d^2_{\min}(\Lambda')$ where they diverge, by at least $d^2_{\min}(\Lambda)$ at some intermediate time, and by at least $d^2_{\min}(\Lambda')$ where they merge, so that the minimum squared distance between distinct paths is at least $5d^2_{\min}(\Lambda)$. Hence $d^2_{\min}(\mathbb{C}) = d^2_{\min}(\Lambda'') = 4d^2_{\min}(\Lambda)$, and the error coefficient $\tilde{N}_0$ is the same as that for Λ''.

Table XI gives the parameters for Class VII and Class VIII codes based on the partitions $\boldsymbol{Z}^2/R\boldsymbol{Z}^2/2\boldsymbol{Z}^2$, $\boldsymbol{Z}^4/D_4/RD_4$, $\boldsymbol{Z}^8/D_8/E_8$, $\boldsymbol{Z}^{16}/D_{16}/H_{16}$, and $D_8^{\perp}/E_8/RE_8$, which have orders 4, 8, 16, 32, and 32, and depths 2, 2, 2, 2, and 3, respectively. The first Class VII code is the two-state Class III code again, and the first Class VIII code is again Ungerboeck's four-state code. The remaining codes are Wei-type codes. In particular, the second and third Class VIII codes correspond to Wei's 16-state four-dimensional and 64-state eight-dimensional codes, which are the prototypes of this class, and the last two Class VIII codes are 256-state elaborations of codes that Wei investigated for $2^{\nu} = 32$, 64, and 128. The Class VII codes are closely related to Class VIII codes, except that they have half the state space dimension and 3/4 the minimum squared distance and coding gain.

VII. Discussion

A large number of codes have been discussed in a common framework in this paper. In this section we draw what conclusions seem warranted by the evidence.

1) *Trellis codes and lattice codes are comparable*, with respect to fundamental parameters such as fundamental coding gain γ versus number of states 2^{ν}. Considering the sequence of Barnes–Wall lattices, we see that it takes two states to get $\gamma = 2^{1/2}$ (1.51 dB), four states to get $\gamma = 2$ (3.02 dB), 16 states to get $\gamma = 2^{3/2}$ (4.52 dB), and 256 states to get $\gamma = 4$ (6.02 dB). The depths μ of these lattices are 1, 2, 3, and 4; their redundancies ρ are 1/2, 1, 3/2, and 2; and their minimum squared distances are 2, 4, 8, and 16.

These properties are shared by the generic trellis codes that we have called Class I, II, and IV, which include a two-state $\gamma = 2^{1/2}$ code based on the partition $\boldsymbol{Z}^4/R\boldsymbol{Z}^4$ and the four-state Ungerboeck $\gamma = 2$ code based on the partition $\boldsymbol{Z}^2/2\boldsymbol{Z}^2$, both with minimal error coefficient $\tilde{N}_0 = 4$.

All of the trellis codes that achieve $\gamma = 2^{3/2}$ (4.52 dB) require 16 states, except for the GCS/Wei eight-state four-dimensional code, which has an error coefficient of $\tilde{N}_0 = 44$, and four-state Class I and V codes, whose error coefficients are very large and whose decoding complexity is not that much less than that of the Wei/Class VIII 16-state four-dimensional code, for example. Note also that the lattices X_{24} and X_{32} achieve $\gamma = 2^{3/2}$ (4.52 dB) and $\gamma = 2^{13/8}$ (4.89 dB) with 8 and 16 states, respectively, but with $\mu = 2$, $\rho = 1/2$ and $d^2_{\min} = 4$, like the Wei codes.

There is a nearby cluster of codes that achieve $\gamma = 3$ (4.77 dB), with either $\mu = 3$, $\rho = 1$ and $d^2_{\min} = 6$, or $\mu = 4$, $\rho = 2$ and $d^2_{\min} = 12$; e.g., the 16- and 32-state two-dimensional Ungerboeck codes, the 16-state CS/Eyuboglu codes, or the 16-state Class III and Class VII codes.

There is another cluster of codes at $\gamma = 2^{7/4}$ (5.27 dB). While there are codes that achieve this fundamental coding gain with as few as eight states (e.g., the eight-dimensional CS code, or two of the Class V codes), it seems to take 32 or 64 states to get reasonable error coefficients (e.g., the Wei or CS eight-dimensional codes).

To achieve $\gamma = 4$ (6.02 dB), all of the trellis codes with reasonable error coefficient ($\tilde{N}_0 < 100$) require 256 states. There are such codes with as few as 16 states (e.g., Class I $\Lambda_{16}/R\Lambda_{16}$) but with very large error coefficients and without as much saving in decoding complexity as the low number of states would suggest (because most of the complexity occurs in decoding the lattice partition). There are a number of good 128-state codes, but there are also lattices (H_{24} and H_{32}) that achieve nearly 6 dB with 128 states. The 256-state one-dimensional Ungerboeck code is remarkable: it obtains $\gamma = 4$ with minimal error coefficient $\tilde{N}_0 = 4$, and with quite low decoding complexity. Note that it has $\mu = 4$, $\rho = 2$ and $d^2_{\min} = 16$, like Λ_{24} and Λ_{32}.

In summary, we propose a folk theorem: it takes two states to get 1.5 dB, four states to get 3 dB, 16 states to get 4.5 dB, perhaps 64 states to get 5.25 dB, and 256 states to get 6 dB, as long as we require a reasonably small error coefficient (for trellis codes).

2) *Trellis codes are better than lattice codes*, if we consider effective coding gain versus decoding complexity. Granted, our measure of effective coding gain is based on

a rule of thumb that is only approximately valid for moderately low error rates and that generally does not take into account neighbors other than the nearest; granted also, our measure of decoding complexity is based specifically on the algorithms of part II and is highly implementation-dependent. We have not even given an effective coding gain γ_{eff} for lattice codes because our rule of thumb is questionable when the number of nearest neighbors is so high. Nonetheless, it seems clear that the very large error coefficients of lattice codes will mean that their effective performance will be significantly inferior to that of comparable trellis codes. For codes with the same parameters, due to the many symmetries of the lattice codes, the decoding complexity does seem to increase slightly as we go from a lattice code to Class I to Class II to Class IV—i.e., as the code becomes "more convolutional"—but this slight effect is very much outweighed by the large reduction in error coefficient.

3) *It is best to keep the redundancy ρ as small as possible*, within reasonable limits. The densest lattices are all self-dual, so their redundancy is equal to half their depth (the comparable trellis codes use rate-$k/2k$ encoders). Ungerboeck [8] made the point, using channel capacity arguments, that there is little to be gained by going beyond 1 bit of redundancy per symbol, i.e., by using rate-$k/(k+r)$ encoders with $r>1$. Wei [11] recognized that, by going beyond two dimensions, the normalized redundancy ρ could be reduced below one and thus that good codes could be obtained with small constellation expansion. The evidence of the codes presented here is that, while the very best codes (e.g., Ungerboeck's four-state two-dimensional code, or all of his one-dimensional codes) may have informativity equal to redundancy, like the best lattices, there is very little loss if redundancy is reduced as long as we do not go to extremes (e.g., Wei's 16-dimensional codes, with $\rho=1/8$). Compare, for example, the Ungerboeck-type two-dimensional codes with the one-dimensional; or the codes (or lattices) with $\rho=1/2$ that achieve $\gamma=2^{3/2}$ (4.52 dB), versus those with $\rho=3/2$.

4) *The Ungerboeck codes are still the benchmark*. Comparing all codes shown in Fig. 12(a)–(c), we see that little improvement has been achieved over Ungerboeck's original results. The one-dimensional codes are generally slightly better than the two-dimensional codes, but this is offset by their normalized redundancy of $\rho=2$, which gives a constellation expansion factor of four, versus the two-dimensional redundancy of $\rho=1$, which gives a constellation expansion factor of two. Some of the $\boldsymbol{Z}^2/4\boldsymbol{Z}^2$ codes found by Eyuboglu are slightly better, but this is not surprising because any 2^ν-state rate-1/2 $\boldsymbol{Z}/4\boldsymbol{Z}$ code can also be regarded as a 2^ν-state rate-2/4 $\boldsymbol{Z}^2/4\boldsymbol{Z}^2$ code. Some of Wei's four-dimensional codes are also slightly better; this is more surprising and significant because these codes also have normalized redundancy $\rho=1/2$. (The Wei eight-dimensional codes are also in the vicinity of the two-dimensional Ungerboeck codes but with $\rho=1/4$; the comparable Wei/Calderbank–Sloane codes have slightly higher decoding complexity and, more importantly, $\rho=5/4$.)

Of all the codes we have considered, a few stand out as "special." The four-state two-dimensional Ungerboeck code is certainly in this category because it is the unique code with $\gamma=2$ and $\tilde{N}_0=4$ and because of its symmetries and close relationship to the special lattice E_8. As mentioned before, the 256-state one-dimensional Ungerboeck code is also special because it is a code with $\gamma=4$ and $\tilde{N}_0=4$, which makes it the trellis cousin of the very special lattice Λ_{24}. The 16-state four-dimensional Wei code is the single code that most clearly improves on the Ungerboeck-type codes; note that it has the same parameters as the lattice X_{24}. (However, could there be a 16-state code with $\mu=3$, $\rho=3/2$, $d^2_{\min}=8$—i.e., with the same parameters as Λ_{16}—that achieves $\gamma=2^{3/2}$, $\tilde{N}_0=4$ or 8? or a 16-state, $\mu=2$, $\rho=1$, $d^2_{\min}=6$ code that achieves $\gamma=3$, $\tilde{N}_0=4$ or 8?)

These results suggest that there is little likelihood of finding significantly better codes in terms of the parameters that we have considered. Above 5 dB, where the curves for the Ungerboeck-type codes begin to tail off, there could be codes with 64 or more states that are superior to those known, although it is also possible that this is close enough to channel capacity that the performance/complexity curve will tend to saturate for all codes. We do not expect to need depths more than three to four in this range, so in view of Lemma 5 a systematic search of $\boldsymbol{Z}^N/4\boldsymbol{Z}^N$ codes should settle the question. (Ternary codes may also be attractive in this region; see [16].)

VIII. Conclusion

We have defined coset codes in such a way as to embrace all of the good known codes and to suggest a large variety of extensions. Their characterization in terms of geometrical parameters like the fundamental coding gain turns out to be quite simple and allows us to sort out from the variety of schemes that have been proposed those that seem to have the best combinations of coding gain, decoding complexity, and constellation expansion.

With respect to those parameters, Ungerboeck's original codes continue to stand up very well *vis-à-vis* the rest of the codes considered. Wei's codes probably represent the most significant improvement, particularly because they reduce the constellation expansion factor below two, while achieving some gains in coding gain and decoding complexity. While the codes of Calderbank and Sloane do not rise to the top in any of our comparisons, their introduction of the lattice/coset viewpoint has clearly been the most significant conceptual contribution since Ungerboeck.

In the opinion of the author, while many of the best codes may have already been discovered, the fields of coset codes and trellis codes are no further developed than that of ordinary coding theory in the early 1960's. There may well be better codes still to be discovered in the 3–6-dB range, as indicated in the previous section. Suboptimal decoders should be investigated, as well as codes specifically tailored for such decoders. The design of good sphere

packings in large dimensions is a topic of active mathematical interest, and the development of still more powerful trellis codes is wide open. Codes which combine good coding gain with other properties, such as rotational invariance and decoding delay, will be important for applications. The theory of phase-modulated coset codes should be brought along in parallel with that of the lattice-type codes. The combination of these codes with spectral shaping requirements, e.g., signalling for partial-response or other band-limited channels, is an important topic. The vector quantization problem is dual to the sphere packing problem and in the block case has been attacked successfully with lattices; there should also be good trellis quantizers. The question of how to design good multidimensional constellations is not closed. Finally, it seems that mathematicians should be interested in trellis codes, particularly as infinite-dimensional generalizations of finite-dimensional sphere-packings. As in other parts of information theory, the interplay between theoretically and practically motivated research is likely to prove fruitful for some time.

Acknowledgment

This work was directly stimulated by the work of L.-F. Wei on multidimensional trellis codes. Remarks by A. R. Calderbank and preprints of the Calderbank–Sloane papers were most helpful in pointing the way to the lattice/coset viewpoint. G. R. Lang kindly provided references to the early history of lattices in communications. I am indebted to M. V. Eyuboglu for providing some of the code parameters and for permission to publish the improved codes cited in the text. I am grateful for comments on earlier versions of this paper by J. B. Anderson, A. R. Calderbank, R. G. Gallager, M. I. Klun, G. Ungerboeck, and L.-F. Wei.

References

[1] G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff, and S. U. Qureshi, "Efficient modulation for band-limited channels," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 632–647, 1984.

[2] E. S. Barnes and G. E. Wall, "Some extreme forms defined in terms of Abelian groups," *J. Australian Math. Soc.*, vol. 1, pp. 47–63, 1959.

[3] J. Leech, "Notes on sphere packings," *Can. J. Math.*, vol. 19, pp. 251–267, 1967.

[4] H. S. M. Coxeter, *Twelve Geometric Essays*. Carbondale, IL: Southern Illinois Univ. Press, 1968.

[5] R. deBuda, "The upper bound of a new near optimal code," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 441–445, 1975.

[6] J. Leech and N. J. A. Sloane, "Sphere packings and error-correcting codes," *Can. J. Math.*, vol. 23, pp. 718–745, 1971.

[7] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag, 1988.

[8] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 55–67, 1982.

[9] L. F. Wei, "Rotationally invariant convolutional channel coding with expanded signal space. Part II: Nonlinear coding," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 672–686, 1984.

[10] A. R. Calderbank and N. J. A. Sloane, "Four-dimensional modulation with an eight-state trellis code," *AT&T Tech. J.*, vol. 64, pp. 1005–1018, 1985.

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

[12] A. R. Calderbank and N. J. A. Sloane, "An eight-dimensional trellis code," *Proc. IEEE*, vol. 74, pp. 757–759, 1986.

[13] ——, "New trellis codes based on lattices and cosets," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 177–195, 1987.

[14] G. D. Forney, Jr., and L. F. Wei, "Multidimensional signal constellations," in preparation, 1989.

[15] F. Pollara, R. J. McEliece, and K. Abdel-Ghaffar, "On finite-state codes," submitted to *IEEE Trans. Inform. Theory*, 1987.

[16] G. D. Forney, Jr., "Coset codes—Part III: Ternary codes, lattices, and trellis codes," in preparation, 1989.

[17] A. LaFanchere, R. H. Deng, and D. J. Costello, Jr., "Multidimensional trellis coded phase modulation using unit-memory and partial-unit-memory convolutional codes," submitted to *IEEE Trans. Inform. Theory*, 1987.

[18] G. D. Forney, Jr., "Coset codes—Part II: Binary lattices and related codes," *IEEE Trans. Inform. Theory*, this issue, pp. 1152–1187.

[19] E. L. Cusack, "Error control codes for QAM signalling," *Electron. Lett.*, vol. 20, pp. 62–63, 1984.

[20] G. D. Forney, Jr., "Convolutional codes I: Algebraic structure," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 720–738, 1970.

[21] G. Ungerboeck, "Trellis-coded modulation with redundant signal sets. Part II: State of the art," *IEEE Commun. Mag.*, vol. 25, no. 2, pp. 12–21, 1987.

[22] M. L. Honig, "Optimization of trellis codes with multilevel amplitude modulation with respect to an error probability criterion," *IEEE Trans. Commun.*, vol. COM-34, pp. 821–825, 1986.

[23] G. J. Pottie and D. P. Taylor, "An approach to Ungerboeck coding for rectangular signal sets," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 285–290, 1987.

[24] D. Divsalar and M. K. Simon, "Multiple trellis-coded modulation (MTCM)," preprint, 1986.

[25] D. Divsalar, M. K. Simon, and J. H. Yuen, "Trellis coding with asymmetrical modulations," *IEEE Trans. Commun.*, vol. COM-35, pp. 130–141, 1987.

[26] G. D. Forney, Jr., "Structural analysis of convolutional codes via dual codes," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 512–518, 1973.