Chapter 14

Introduction to lattice and trellis codes

In this chapter we discuss coding techniques for bandwidth-limited (high-SNR) AWGN channels.

On bandwidth-limited channels, nonbinary signal alphabets such as M-PAM must be used to approach capacity. Furthermore, the signals should be used with a nonuniform, Gaussian-like probability distribution.

Using large-alphabet approximations, we show that the total coding gain of a coded modulation scheme for the bandwidth-limited AWGN channel is the sum of a coding gain due to a denser packing than the baseline M-PAM scheme, plus a shaping gain due to constellation shaping (or equivalently to use of a nonuniform distribution). At high SNRs, the coding and shaping problems are separable.

The maximum possible shaping gain is a factor of $\pi e/6$ (1.53 dB). Simple shaping methods such as shell mapping and trellis shaping can easily obtain of the order of 1 dB of shaping gain.

For moderate coding gains at moderate complexity, the two principal classes of packings are lattices and trellis codes, which are analogous to block and convolutional codes, respectively. By now the principles of construction of the best such codes are well understood, and it seems likely that the best codes have been found. We plot the effective coding gains of these known moderate-complexity lattices and trellis codes versus the branch complexity of their minimal trellises, assuming ML decoding. Trellis codes are somewhat superior, due mainly to their lower error coefficients.

We briefly mention higher-performance schemes, including multilevel schemes with multistage decoding and bit-interleaved coded modulation, which allow the use of high-performance binary codes such as those described in the previous chapter to approach capacity.

14.1 Lattices

It is clear from Shannon's capacity theorem that an optimal block code for a bandwidth-limited AWGN channel consists of a dense packing of code points within a sphere in a high-dimensional Euclidean space. Most of the densest known packings are lattices.

In this section we briefly describe lattice constellations, and analyze their performance using the union bound estimate and large-constellation approximations.

An n-dimensional (n-D) *lattice* Λ is a discrete subset of n-space $\mathbb{R}^n$ that has the group property. Without essential loss of generality, Λ may be assumed to span $\mathbb{R}^n$. The points of the lattice then form a uniform infinite packing of $\mathbb{R}^n$.

Example 1. The set of integers $\mathbb{Z}$ is a one-dimensional lattice, since $\mathbb{Z}$ is a discrete subgroup of $\mathbb{R}$. Any 1-dimensional lattice is of the form $\Lambda = \alpha\mathbb{Z}$ for some scalar $\alpha > 0$. □

Example 2. The *integer lattice* $\mathbb{Z}^n$ (the set of integer n-tuples) is an n-dimensional lattice for any $n \geq 1$. □

Example 3. The *hexagonal lattice* $A_2 = \{a(1,0) + b(\frac{1}{2}, \frac{\sqrt{3}}{2}) \mid (a,b) \in \mathbb{Z}^2\}$ is illustrated in Figure 1. This lattice is the densest packing of $\mathbb{R}^2$. □

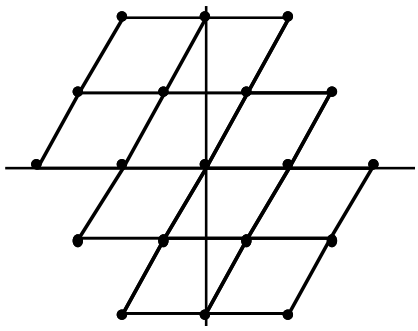

Figure 1. *The hexagonal lattice A_2.*

Exercise 1. Let $\mathcal{C}$ be an (n,k,d) binary linear block code. Show that

$$\Lambda_{\mathcal{C}} = \{\mathbf{x} \in \mathbb{Z}^n \mid \mathbf{x} \equiv \mathbf{c} \bmod 2 \text{ for some } \mathbf{c} \in \mathcal{C}\} \tag{14.1}$$

is an n-dimensional sublattice of $\mathbb{Z}^n$ (called a "Construction A" or "mod-2" lattice).

A general n-dimensional lattice Λ that spans $\mathbb{R}^n$ may be characterized by a set of linearly independent generators $G = \{\mathbf{g}_j,\ 1 \leq j \leq n\}$ such that Λ is the set of all integer linear combinations of the generators:

$$\Lambda = \{\mathbf{a}G = \sum_j a_j \mathbf{g}_j \mid \mathbf{a} \in \mathbb{Z}^n\}. \tag{14.2}$$

Thus Λ may be viewed as the image of the integer lattice $\mathbb{Z}^n$ under a linear transformation of n-space $\mathbb{R}^n$ by the linear operator G, as illustrated by Figure 1.

By the group property of Λ, any translate $\Lambda + \mathbf{x}$ by a lattice point $\mathbf{x} \in \Lambda$ is just Λ again. This implies that a lattice is "geometrically uniform;" every point of the lattice has the same number of neighbors at each distance, and all decision regions of a minimum-distance decoder ("Voronoi regions") are congruent and form a tessellation of $\mathbb{R}^n$. Indeed, any lattice translate $\Lambda + \mathbf{t}$ is geometrically uniform.

The key geometrical parameters of a lattice are:

- the *minimum squared distance* $d^2_{\min}(\Lambda)$ between lattice points;
- the *kissing number* $K_{\min}(\Lambda)$ (the number of nearest neighbors to any lattice point);
- the *volume* $V(\Lambda)$ of n-space per lattice point. As indicated in Figure 1, this volume is the volume of the *fundamental parallelotope*

$$[0,1)^n G = \{\mathbf{a}G \mid \mathbf{a} \in [0,1)^n\}.$$

Since the volume of the n-cube $[0,1)^n$ is 1 and the Jacobian of the linear transformation G is its determinant $|G|$, it follows that $V(\Lambda) = |G|$ for any generator matrix G of Λ.

The *Hermite parameter* of Λ is the normalized density parameter

$$\gamma_c(\Lambda) = \frac{d^2_{\min}(\Lambda)}{V(\Lambda)^{2/n}}, \tag{14.3}$$

which we will shortly identify as its nominal coding gain. The quantity $V(\Lambda)^{2/n}$ may be thought of as the normalized volume of Λ per two dimensions.

Example 3 (cont.) For the hexagonal lattice A_2, the minimum squared distance is $d^2_{\min}(A_2) = 1$, the kissing number is $K_{\min}(A_2) = 6$, the volume is $V(A_2) = \sqrt{3}/2$, and the Hermite parameter is $\gamma_c(A_2) = 2/\sqrt{3} = 1.155$ (0.62 dB). Therefore A_2 is denser than the integer lattice $\mathbb{Z}^2$, for which $d^2_{\min}(\mathbb{Z}^2) = V(\mathbb{Z}^2) = \gamma_c(\mathbb{Z}^2) = 1$. □

Exercise 1 (cont.) Show that if $\mathcal{C}$ is an (n,k,d) binary linear block code with N_d weight-d words, then the mod-2 lattice $\Lambda_{\mathcal{C}}$ has the following geometrical parameters:

$$d^2_{\min}(\Lambda_{\mathcal{C}}) = \min\{d, 4\}; \tag{14.4}$$

$$K_{\min}(\Lambda_{\mathcal{C}}) = \begin{cases} 2^d N_d, & \text{if } d < 4; \\ 2n, & \text{if } d > 4; \\ 2^d N_d + 2n, & \text{if } d = 4; \end{cases} \tag{14.5}$$

$$V(\Lambda_{\mathcal{C}}) = 2^{n-k}; \tag{14.6}$$

$$\gamma_c(\Lambda_{\mathcal{C}}) = \frac{d^2_{\min}(\Lambda_{\mathcal{C}})}{2^{\eta(\mathcal{C})}}, \tag{14.7}$$

where $\eta(\mathcal{C}) = 2(n-k)/n$ is the redundancy of $\mathcal{C}$ in bits per two dimensions. □

Exercise 2. Show that $\gamma_c(\Lambda)$ is invariant to scaling, orthogonal transformations, and Cartesian products; *i.e.*, $\gamma_c(\alpha U \Lambda^m) = \gamma_c(\Lambda)$, where $\alpha > 0$ is any scale factor, U is any orthogonal matrix, and $m \geq 1$ is any positive integer. Show that $\gamma_c(\alpha U \mathbb{Z}^n) = 1$ for any version $\alpha U \mathbb{Z}^n$ of any integer lattice $\mathbb{Z}^n$. □

14.2 Lattice constellations

A *lattice constellation*

$$\mathcal{C}(\Lambda, \mathcal{R}) = (\Lambda + \mathbf{t}) \cap \mathcal{R} \tag{14.8}$$

is the finite set of points in a lattice translate $\Lambda + \mathbf{t}$ that lie within a compact bounding region $\mathcal{R}$ of n-space.

Example 4. An M-PAM constellation $\alpha\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$ is a one-dimensional lattice constellation $\mathcal{C}(2\alpha\mathbb{Z}, \mathcal{R})$ with $\Lambda + \mathbf{t} = 2\alpha(\mathbb{Z} + 1)$ and $\mathcal{R} = [-\alpha M, \alpha M]$. □

The key geometric properties of the region $\mathcal{R}$ are

- its *volume* $V(\mathcal{R}) = \int_{\mathcal{R}} d\mathbf{x}$;
- the *average energy* $P(\mathcal{R})$ per dimension of a uniform probability density function over $\mathcal{R}$:

$$P(\mathcal{R}) = \int_{\mathcal{R}} \frac{\|\mathbf{x}\|^2}{n} \frac{d\mathbf{x}}{V(\mathcal{R})}. \tag{14.9}$$

The *normalized second moment* of $\mathcal{R}$ is defined as the dimensionless parameter

$$G(\mathcal{R}) = \frac{P(\mathcal{R})}{V(\mathcal{R})^{2/n}}. \tag{14.10}$$

Example 4 (cont.). The key geometrical parameters of $\mathcal{R} = [-\alpha M, \alpha M]$ are $V(\mathcal{R}) = 2\alpha M$, $P(\mathcal{R}) = \alpha^2 M^2/3$, and $G(\mathcal{R}) = 1/12$. □

Exercise 3. Show that $G(\mathcal{R})$ is invariant to scaling, orthogonal transformations, and Cartesian products; *i.e.,* $G(\alpha U \mathcal{R}^m) = G(\mathcal{R})$, where $\alpha > 0$ is any scale factor, U is any orthogonal matrix, and $m \geq 1$ is any positive integer. Show that $G(\alpha U[-1,1)^n) = 1/12$ for any version $\alpha U[-1,1)^n$ of any n-cube $[-1,1)^n$ centered at the origin. □

For performance analysis of large lattice constellations, one may use the following approximations, the first two of which are together known as the *continuous approximation*:

- The *size* of the constellation is

$$|\mathcal{C}(\Lambda, \mathcal{R})| \approx \frac{V(\mathcal{R})}{V(\Lambda)}; \tag{14.11}$$

- The *average energy per dimension* of a uniform discrete distribution over $\mathcal{C}(\Lambda, \mathcal{R})$ is

$$P(\mathcal{C}(\Lambda, \mathcal{R})) \approx P(\mathcal{R}); \tag{14.12}$$

- The *average number of nearest neighbors* to any point in $\mathcal{C}(\Lambda, \mathcal{R})$ is $\approx K_{\min}(\Lambda)$.

Again, the union bound estimate (UBE) on probability of block decoding error is

$$\Pr(E) \approx K_{\min}(\Lambda) Q^{\surd}\left(\frac{d^2_{\min}(\Lambda)}{4\sigma^2}\right). \tag{14.13}$$

Since

$$\begin{aligned}
\rho &= \frac{2}{n}\log_2|\mathcal{C}(\Lambda,\mathcal{R})| \approx \frac{2}{n}\log_2\frac{V(\mathcal{R})}{V(\Lambda)};\\
\mathrm{SNR} &= \frac{P(\mathcal{C}(\Lambda,\mathcal{R}))}{\sigma^2} \approx \frac{P(\mathcal{R})}{\sigma^2};\\
\mathrm{SNR}_{\mathrm{norm}} &\approx \frac{\mathrm{SNR}}{2^\rho} = \frac{V(\Lambda)^{2/n}}{V(\mathcal{R})^{2/n}}\frac{P(\mathcal{R})}{\sigma^2},
\end{aligned}$$

we may write the UBE as

$$\Pr(E) \approx K_{\min}(\Lambda) Q^{\surd}\left(\gamma_{\mathrm{c}}(\Lambda)\gamma_s(\mathcal{R})(3\,\mathrm{SNR}_{\mathrm{norm}})\right), \tag{14.14}$$

where the *nominal coding gain* of Λ and the *shaping gain* of $\mathcal{R}$ are defined respectively as

$$\gamma_{\mathrm{c}}(\Lambda) = \frac{d^2_{\min}(\Lambda)}{V(\Lambda)^{2/n}}; \tag{14.15}$$

$$\gamma_s(\mathcal{R}) = \frac{V(\mathcal{R})^{2/n}}{12P(\mathcal{R})} = \frac{1/12}{G(\mathcal{R})}. \tag{14.16}$$

For a baseline M-PAM constellation with $\Lambda = 2\alpha\mathbb{Z}$ and $\mathcal{R} = [-\alpha M, \alpha M]$, we have $\gamma_{\mathrm{c}}(\Lambda) = \gamma_s(\mathcal{R}) = 1$ and $K_{\min}(\Lambda) \approx 2$, so the UBE reduces to the baseline expression

$$\Pr(E) \approx 2Q^{\surd}(3\,\mathrm{SNR}_{\mathrm{norm}}).$$

The nominal coding gain $\gamma_{\mathrm{c}}(\Lambda)$ measures the increase in density of Λ over the baseline integer lattice $\mathbb{Z}$ (or $\mathbb{Z}^n$). The shaping gain $\gamma_s(\mathcal{R})$ measures the decrease in average energy of $\mathcal{R}$ relative to an interval $[-\alpha,\alpha]$ (or an n-cube $[-\alpha,\alpha]^n$). Both contribute a multiplicative factor of gain to the argument of the $Q^{\surd}(\cdot)$ function.

As before, the effective coding gain is reduced by the error coefficient $K_{\min}(\Lambda)$. The probability of block decoding error per two dimensions is

$$P_s(E) \approx K_s(\Lambda) Q^{\surd}(\gamma_{\mathrm{c}}(\Lambda)\gamma_s(\mathcal{R})(3\,\mathrm{SNR}_{\mathrm{norm}})), \tag{14.17}$$

in which the normalized error coefficient per two dimensions is $K_s(\Lambda) = 2K_{\min}(\Lambda)/n$.

Graphically, a curve of the form $P_s(E) \approx K_s(\Lambda) Q^{\surd}(\gamma_{\mathrm{c}}(\Lambda)\gamma_s(\mathcal{R})(3\,\mathrm{SNR}_{\mathrm{norm}}))$ may be obtained simply by moving the baseline curve $P_s(E) = 4Q^{\surd}(3\,\mathrm{SNR}_{\mathrm{norm}})$ to the left by $\gamma_{\mathrm{c}}(\Lambda)$ and $\gamma_s(\mathcal{R})$ (in dB), and upward by a factor of $K_s(\Lambda)/4$. Such simple manipulations of the baseline curve as a function of $\gamma_{\mathrm{c}}(\Lambda)$, $\gamma_s(\mathcal{R})$ and $K_s(\Lambda)$ again are an easy and useful design tool for lattice constellations of moderate complexity.

14.3 Shaping gain and shaping techniques

Although shaping is a newer and less important topic than coding, we discuss it first because its story is quite simple.

The n-dimensional shaping region $\mathcal{R}$ that minimizes $G(\mathcal{R})$ is obviously an n-sphere. The key geometrical parameters of an n-sphere of radius r (for n even) are:

$$\begin{aligned} V_{\otimes}(n,r) &= \frac{(\pi r^2)^{n/2}}{(n/2)!}; \\ P_{\otimes}(n,r) &= \frac{r^2}{n+2}; \\ G_{\otimes}(n,r) &= \frac{P_{\otimes}(n,r)}{V_{\otimes}(n,r)^{2/n}} = \frac{((n/2)!)^{2/n}}{\pi(n+2)}. \end{aligned}$$

By Stirling's approximation, $m! \rightarrow (m/e)^m$ as $m \rightarrow \infty$, which implies

$$\begin{aligned} G_{\otimes}(n,r) &\rightarrow \frac{1}{2\pi e}; \\ \gamma_{s\otimes}(n,r) &= \frac{1/12}{G_{\otimes}(n,r)} \rightarrow \frac{\pi e}{6} \ (1.53 \text{ dB}). \end{aligned}$$

Thus shaping gain is limited to a finite value as $n \rightarrow \infty$, namely $\pi e/6$ (1.53 dB), which is called the *ultimate shaping gain*.

The shaping gain of an n-sphere is plotted for dimensions $n \leq 24$ in Figure 2. Note that the shaping gain of a 16-sphere already exceeds 1 dB.

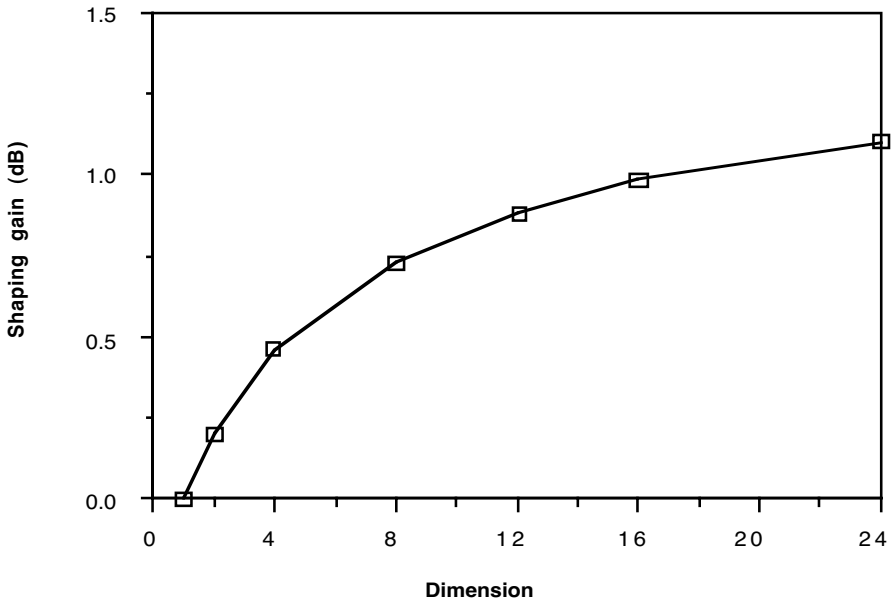

Figure 2. *Shaping gains of n-spheres for $n \leq 24$.*

The projection of a uniform probability distribution over an n-sphere onto one or two dimensions is a nonuniform probability distribution that approaches a Gaussian distribution

as $n \to \infty$. The ultimate shaping gain of $\pi e/6$ (1.53 dB) may alternatively be derived as the difference between the average power of a uniform distribution over an interval and that of a Gaussian distribution with the same differential entropy.

Shaping thus induces a Gaussian-like probability distribution on a one-dimensional PAM or two-dimensional QAM constellation, rather than an equiprobable distribution. In principle, with spherical shaping, the lower-dimensional constellation will become arbitrarily large, even with fixed average power. In practice, the lower-dimensional constellation is constrained by design to a certain region $\mathcal{R}$ to limit "shaping constellation expansion." The n-dimensional shape then only approximates spherical shaping subject to this constraint, and the lower-dimensional probability distribution approaches a truncated Gaussian distribution within the region $\mathcal{R}$.

With large constellations, shaping can be implemented almost independently of coding by operations on the "most significant bits" of M-PAM or $(M \times M)$-QAM constellation labels, which affect the gross shape of the n-dimensional constellation. In contrast, coding affects the "least significant bits" and determines fine structure.

Two practical schemes that can easily obtain shaping gains of 1 dB or more while limiting 2D shaping constellation expansion to a factor of 1.5 or less are "trellis shaping," a kind of dual to trellis coding, and "shell mapping," which uses generating-function techniques to enumerate the points in a Cartesian product constellation in approximate increasing order of energy.

14.4 Coding gains of dense lattices

Finding the densest lattice packings in a given number of dimensions is a mathematical problem of long standing. A summary of the densest known packings is given in [Conway and Sloane, *Sphere Packings, Lattices and Groups*]. The nominal coding gains of these lattices in up to 24 dimensions is plotted in Figure 3.

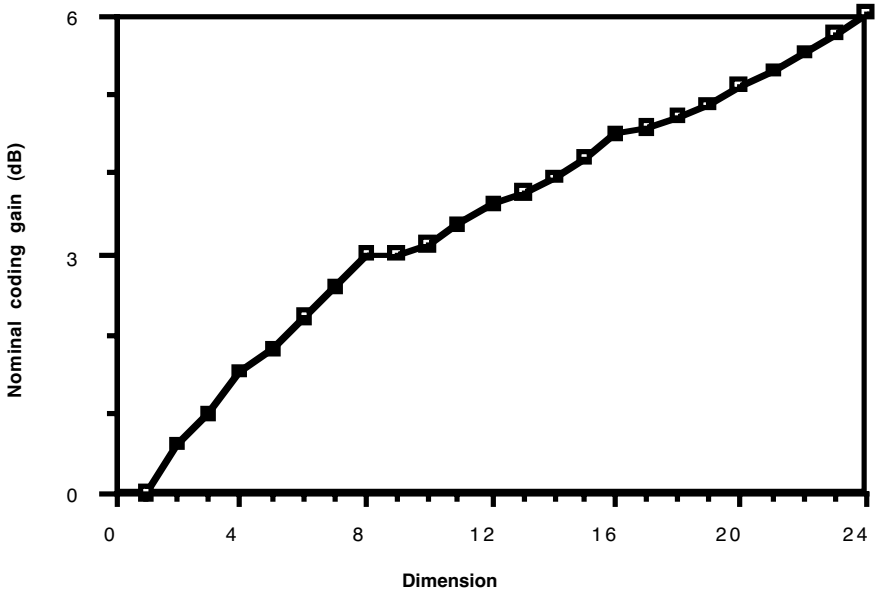

Figure 3. *Nominal coding gains of densest lattices in dimensions $n \leq 24$.*

In contrast to shaping gain, the nominal coding gains of dense n-dimensional lattices become infinite as $n \to \infty$.

Example 5 (Barnes-Wall lattices). For all integer $m \geq 0$, there exists a 2^{m+1}-dimensional *Barnes-Wall lattice* $BW_{2^{m+1}}$ whose nominal coding gain is $2^{m/2}$ (see next subsection). The two-dimensional BW lattice is $\mathbb{Z}^2$. In 4, 8, and 16 dimensions the BW lattices (denoted by D_4, E_8 and Λ_{16}, respectively) are the densest lattices known. For large m, considerably denser lattices are known. □

Exercise 1 (cont.) Show that the mod-2 lattices corresponding to the $(4, 3, 2)$ and $(4, 1, 4)$ binary linear block codes have coding gain $2^{1/2}$ (1.51 dB); these lattices are in fact versions of D_4. Show that the mod-2 lattice corresponding to the $(8, 4, 4)$ binary linear block code has coding gain 2 (3.01 dB); this lattice is in fact a version of E_8. Show that no mod-2 lattice has a nominal coding gain more than 4 (6.02 dB). □

However, effective coding gains cannot become infinite. Indeed, the Shannon limit shows that no lattice can have a combined effective coding gain and shaping gain greater than 9 dB at $P_s(E) \approx 10^{-6}$. This limits the maximum possible effective coding gain to 7.5 dB, since shaping gain can contribute up to 1.53 dB.

What limits effective coding gain is the number of near neighbors, which becomes very large for high-dimensional dense lattices.

Example 5 (cont.) The kissing number of the 2^{m+1}-dimensional Barnes-Wall lattice is

$$K_{\min}(BW_{2^{m+1}}) = \prod_{1 \leq i \leq m+1} (2^i + 2).$$

For $m = 0, 1, 2, 3, 4, \ldots$ these numbers are 4, 24, 240, 4320, 146880, ... Thus while BW_{32} has a nominal coding gain of 4 (6.02 dB), its kissing number is 146880, so its effective coding gain by our rule of thumb is only about 3.8 dB. BW_{128} has a nominal coding gain of 8 (9.03 dB), but a kissing number of 1 260 230 400, so its effective coding gain by our rule of thumb is only about 4.6 dB. These calculations indicate how the effective coding gain of higher-dimensional lattices eventually saturates. □

Example 6 (Leech Lattice). The *Leech lattice* L_{24}, a remarkably dense lattice in 24 dimensions, has a nominal coding gain of 4 (6.02 dB), but it has a kissing number of 196560, so its effective coding gain by our rule of thumb is only about 3.6 dB. □

14.4.1 Barnes-Wall lattices

The Barnes-Wall lattices (1959) are an infinite family of n-dimensional lattices that are analogous to the Reed-Muller binary block codes. For $n \leq 16$, they are the best lattices known. For greater n, they are not in general the best lattices known, but in terms of performance *vs.* decoding complexity they are still quite good, since they admit relatively simple decoding algorithms.

For any integer $m \geq 0$, there exists an $(n = 2^{m+1})$-dimensional BW lattice, denoted $BW_{2^{m+1}}$, that has minimum squared Euclidean distance $d^2_{\min}(BW_{2^{m+1}}) = 2^m$, normalized volume $V(BW_{2^{m+1}})^{2/n} = 2^{m/2}$, and therefore nominal coding gain $\gamma_c(BW_{2^{m+1}}) = 2^{m/2}$.

In 2 dimensions, the Barnes-Wall lattice BW_2 is the integer lattice $\mathbb{Z}^2$, which is the mod-2 lattice corresponding to the $(2, 2, 1)$ code.

The mod-2 lattice $R\mathbb{Z}^2$ corresponding to the $(2,1,2)$ code is a sublattice of $\mathbb{Z}^2$; it is the set of all integer 2-tuples in which both integers are even or both integers are odd. It can be obtained by rotating $\mathbb{Z}^2$ by 45° and scaling by $\sqrt{2}$; *i.e.,* by transforming $\mathbb{Z}^2$ by the 2×2 Hadamard matrix

$$R = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Consequently $d^2_{\min}(R\mathbb{Z}^2) = 2$ and $V(R\mathbb{Z}^2) = 2$.

The lattice $2\mathbb{Z}^2$ (the mod-2 lattice corresponding to the $(2,0,\infty)$ code) is a sublattice of $R\mathbb{Z}^2$ with $d^2_{\min}(2\mathbb{Z}^2) = 4$ and $V(2\mathbb{Z}^2) = 4$. Note that $2\mathbb{Z}^2 = R(R\mathbb{Z}^2)$, since $R^2 = 2I$.

In fact, we see that there is a lattice chain $\mathbb{Z}^2/R\mathbb{Z}^2/2\mathbb{Z}^2/2R\mathbb{Z}^2/4\mathbb{Z}^2/\ldots$ with minimum squared distances $1/2/4/8/16/\ldots$.

The remaining BW lattices may be constructed recursively from this chain by the $|u|u+v|$ construction. $BW_{2^{m+1}}$ is constructed from BW_{2^m} and RBW_{2^m} as

$$BW_{2^{m+1}} = \{(\mathbf{u}, \mathbf{u}+\mathbf{v}) \mid \mathbf{u} \in BW_{2^m}, \mathbf{v} \in RBW_{2^m}\}.$$

More generally, for any $j \geq 0$, $R^j BW_{2^{m+1}} = \{(\mathbf{u}, \mathbf{u}+\mathbf{v}) \mid \mathbf{u} \in R^j BW_{2^m}, \mathbf{v} \in R^{j+1} BW_{2^m}\}$.

It is then easy to prove the following facts by recursion:

(a) The dimension of $BW_{2^{m+1}}$ is $n = 2^{m+1}$.

(b) The volume of $BW_{2^{m+1}}$ is

$$V(BW_{2^{m+1}}) = V(BW_{2^m})V(R(BW_{2^m})) = 2^{2^{m-1}} V(BW_{2^m})^2.$$

This recursion yields $V(BW_{2^{m+1}}) = 2^{m2^{m-1}}$, or $V(BW_{2^{m+1}})^{2/n} = 2^{m/2}$.

(c) The minimum squared distance of $BW_{2^{m+1}}$ is $d^2_{\min}(BW_{2^{m+1}}) = 2^m$.

(d) $\{R^j BW_{2^{m+1}}, j \geq 1\}$ is a chain of sublattices with minimum squared distances and normalized volumes increasing by a factor of 2 for each increment of j.

We verify that these assertions hold for $BW_2 = \mathbb{Z}^2$. For $m \geq 1$, the dimension and volume follow from the construction. We verify the distance as follows:

(a) if $\mathbf{u} = \mathbf{0}$, then $||(\mathbf{0}, \mathbf{v})||^2 = ||\mathbf{v}||^2 \geq 2^m$ if $\mathbf{v} \neq \mathbf{0}$, since $\mathbf{v} \in RBW_{2^m}$.

(b) if $\mathbf{u} + \mathbf{v} = \mathbf{0}$, then $\mathbf{u} = -\mathbf{v} \in RBW_{2^m}$ and $||(-\mathbf{v}, \mathbf{0})||^2 \geq 2^m$ if $\mathbf{v} \neq \mathbf{0}$.

(c) if $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{u} + \mathbf{v} \neq \mathbf{0}$, then both $\mathbf{u}$ and $\mathbf{u} + \mathbf{v}$ are in BW_{2^m} (since RBW_{2^m} is a sublattice of BW_{2^m}), so

$$||(\mathbf{u}, \mathbf{u}+\mathbf{v})||^2 = ||\mathbf{u}||^2 + ||\mathbf{u}+\mathbf{v}||^2 \geq 2 \cdot 2^{m-1} = 2^m.$$

Equality clearly holds for $(\mathbf{0}, \mathbf{v})$, $(\mathbf{v}, \mathbf{0})$ or $(\mathbf{u}, \mathbf{u})$ if we choose $\mathbf{v}$ or $\mathbf{u}$ as a minimum-weight vector from their respective lattices.

Finally, the sublattice chain for m follows from the sublattice chain for $m-1$ by construction.

The $|u|u+v|$ construction suggests the following tableau of BW lattices. Here $D_4 = BW_4$, $E_8 = BW_8$, and $\Lambda_n = BW_n$ for $n = 2^{m+1} \geq 16$. Also, we use $R^2 = 2I_{2^m}$.

$$
\begin{array}{ccccccccc}
\mathbb{Z}^2 & & & & & & & & \\
& D_4 & & & & & & & \\
R\mathbb{Z}^2 & & E_8 & & & & & & \\
& RD_4 & & \Lambda_{16} & & & & & \\
2\mathbb{Z}^2 & & RE_8 & & \Lambda_{32} & & & & \\
& 2D_4 & & R\Lambda_{16} & & \Lambda_{64} & & & \\
2R\mathbb{Z}^2 & & 2E_8 & & R\Lambda_{32} & & \Lambda_{128} & & \\
& 2RD_4 & & 2\Lambda_{16} & & R\Lambda_{64} & & \Lambda_{256} & \\
4\mathbb{Z}^2 & & 2RE_8 & & 2\Lambda_{32} & & R\Lambda_{128} & & \Lambda_{512}
\end{array}
$$

Figure 4. *Tableau of Barnes-Wall lattices.*

In this tableau each BW lattice lies halfway between the two lattices of half the dimension that are used to construct it in the $|u|u+v|$ construction, from which we can immediately deduce its normalized volume.

For example, E_8 has the same normalized volume as $R\mathbb{Z}^2$, namely $V(E_8)^{2/8} = 2$. However, $d^2_{\min}(E_8) = 4$, whereas $d^2_{\min}(R\mathbb{Z}^2) = 2$. Therefore the nominal coding gain of E_8 is twice that of $R\mathbb{Z}^2$, namely $\gamma_c(E_8) = 2$ (3.01 dB).

14.5 Trellis codes

Trellis codes are dense packings of Euclidean-space sequences in a sequence space which is in principle infinite-dimensional. Trellis codes are to lattices as convolutional codes are to block codes. We will see that, just as binary convolutional codes provide a better performance/complexity tradeoff than binary block codes in the power-limited regime, trellis codes provide a better performance/complexity tradeoff than lattices in the bandwidth-limited regime, although the difference is not as dramatic.

The key ideas in the invention of trellis codes were:

- use of minimum squared Euclidean distance as the design criterion;
- coding on subsets of signal sets using convolutional coding principles (*e.g.,* trellises and the Viterbi algorithm).

A typical large-constellation trellis code is designed as follows. One starts with a large low-dimensional constellation, which in practice is almost always a lattice constellation $\mathcal{C}(\mathbb{Z}^n, \mathcal{R})$ based on a version of an n-dimensional integer lattice $\mathbb{Z}^n$, such as M-PAM or $(M \times M)$-QAM. (M-PSK constellations are sometimes used in the intermediate ($\rho \approx 2$ b/2D) regime because of their constant-energy property, but we will not discuss M-PSK trellis codes here.)

One can then form an m-fold Cartesian product constellation

$$\mathcal{C}(\mathbb{Z}^n, \mathcal{R})^m = \mathcal{C}(\mathbb{Z}^{mn}, \mathcal{R}^m),$$

which is still based on an mn-dimensional integer lattice $\mathbb{Z}^{mn}$.

The constellation $\mathcal{C}(\mathbb{Z}^{mn}, \mathcal{R}^m)$ is partitioned into subsets of equal size, where the number of subsets is typically a power of two, say 2^b. Initially this was done by a sequence of two-way partitions in which the minimum squared distance within subsets was maximized at each level. Subsequently it was recognized that the resulting constellations were almost always lattice constellations $\mathcal{C}(\Lambda', \mathcal{R}^m)$ based on a sublattice Λ' of index $|\mathbb{Z}^{mn}/\Lambda'| = 2^b$ in $\mathbb{Z}^{mn}$. In other words, $\mathbb{Z}^{mn}$ is the union of 2^b cosets of Λ', and the 2^b subsets are the points of $\mathcal{C}(\mathbb{Z}^{mn}, \mathcal{R}^m)$ that lie in each such coset. The sublattice Λ' is usually chosen to be as dense as possible.

Example 7 (1D partitions). In one dimension, there is a chain of sublattices of $\mathbb{Z}$ as follows:

$$\mathbb{Z} \supseteq 2\mathbb{Z} \supseteq 4\mathbb{Z} \supseteq 8\mathbb{Z} \supseteq \cdots,$$

which may alternatively be written as $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}/8\mathbb{Z}/\cdots$. Each partition is two-way; that is, each lattice is the union of two cosets of the next sublattice. The corresponding minimum squared distances are $1/4/16/64/\cdots$. Thus an M-PAM constellation $\mathcal{C}(\mathbb{Z}, [-M/2, M/2])$ with minimum squared distance 1 may be partitioned into 2 subsets of the form $\mathcal{C}(2\mathbb{Z}, [-M/2, M/2])$ with minimum squared distance 4 within subsets, or 4 subsets of the form $\mathcal{C}(4\mathbb{Z}, [-M/2, M/2])$ with minimum squared distance 16 within subsets, and so forth. □

Example 8 (2D partitions). In two dimensions, there is a chain of sublattices of $\mathbb{Z}^2$ as follows:

$$\mathbb{Z}^2 \supseteq R\mathbb{Z}^2 \supseteq 2\mathbb{Z}^2 \supseteq 2R\mathbb{Z}^2 \supseteq \cdots,$$

where R is the 2×2 Hadamard matrix as above. This chain may alternatively be written as $\mathbb{Z}^2/R\mathbb{Z}^2/2\mathbb{Z}^2/2R\mathbb{Z}^2/\cdots$. Each partition is two-way. The corresponding minimum squared distances are $1/2/4/8/\cdots$. Thus a QAM constellation $\mathcal{C}(\mathbb{Z}^2, \mathcal{R})$ with minimum squared distance 1 may be partitioned into 2 subsets of the form $\mathcal{C}(R\mathbb{Z}^2, \mathcal{R})$ with minimum squared distance 2 within subsets, or 4 subsets of the form $\mathcal{C}(2\mathbb{Z}^2, \mathcal{R})$ with minimum squared distance 4 within subsets, and so forth. The bounding region $\mathcal{R}$ should contain an equal number of points in each subset. □

Example 9 (4D partitions). In four dimensions, there is a chain of sublattices of $\mathbb{Z}^4$ as follows:

$$\mathbb{Z}^4 \supseteq D_4 \supseteq R\mathbb{Z}^4 \supseteq RD_4 \supseteq \cdots,$$

where D_4 is the 4-dimensional Barnes-Wall lattice and R is the 4×4 matrix

$$R = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

(Alternatively, this is the chain of mod-2 lattices corresponding to the $(4,4,1)$, $(4,3,2)$, $(4,2,2)$ and $(4,1,4)$ binary linear block codes.) This chain may alternatively be written as $\mathbb{Z}^4/D_4/R\mathbb{Z}^4/RD_4/\cdots$. Each partition is two-way. The corresponding minimum squared distances are $1/2/2/4/\cdots$. Thus a 4D constellation $\mathcal{C}(\mathbb{Z}^4, \mathcal{R})$ with minimum squared distance 1 may be partitioned into 2 subsets of the form $\mathcal{C}(D_4, \mathcal{R})$ with minimum squared distance 2 within subsets, 8 subsets of the form $\mathcal{C}(2D_4, \mathcal{R})$ with minimum squared distance 4 within subsets, etc. Again, the bounding region $\mathcal{R}$ should contain an equal number of points in each subset. □

A trellis code encoder then operates as shown in Figure 5. Some of the input data bits are encoded in a rate-k/b 2^ν-state binary convolutional encoder. Almost always k is chosen to equal $b-1$, so the code redundancy is 1 bit per mn dimensions. The encoder output sequence of b-tuples selects a corresponding sequence of subsets of $\mathcal{C}(\mathbb{Z}^{mn}, \mathcal{R}^m)$ (cosets of Λ'). The convolutional code and the labeling of the subsets are chosen primarily to maximize the minimum squared distance $d^2_{\min}(\mathcal{C})$ between signal point sequences in any possible encoded subset sequence, and secondarily to minimize the maximum possible number $K_{\min}(\mathcal{C})$ of nearest-neighbor sequences. Finally, other input data bits select the actual signal points to be transmitted from the selected subsets. If there is any shaping, it is done at this level.

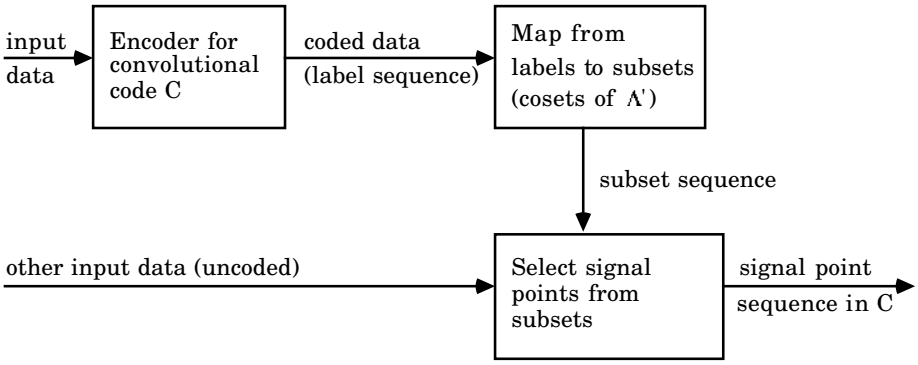

Figure 5. *Trellis code encoder.*

The nominal coding gain of such a trellis code is

$$\gamma_{\rm c}(\mathcal{C}) = d^2_{\min}(\mathcal{C}) 2^{-\eta(\mathcal{C})}, \tag{14.18}$$

where $\eta(\mathcal{C}) = 2/mn$ is the redundancy of the convolutional code in bits per two dimensions. The factor $2^{\eta(\mathcal{C})}$ may be thought of as the normalized volume of the trellis code per two dimensions, if the signal constellation is a lattice constellation based on an integer lattice $\mathbb{Z}^{mn}$. The effective coding gain is reduced by the amount that the error coefficient $2K_{\min}(\mathcal{C})/mn$ per two dimensions exceeds the baseline M-PAM error coefficient of 4 per two dimensions, again according to the rule of thumb that a factor of 2 increase costs 0.2 dB.

Exercise 1 (cont.) Let $\mathcal{C}$ be a rate-k/n binary linear convolutional code with free distance d and N_d minimum-weight code sequences per n dimensions. Define the corresponding mod-2 trellis code $\Lambda_{\mathcal{C}}$ to be the set of all integer sequences $\mathbf{x}$ with D-transform $x(D)$ such that $x(D) \equiv c(D) \bmod 2$ for some code sequence $c(D)$ in $\mathcal{C}$.

(a) Show that an encoder as in Figure 5 based on the convolutional code $\mathcal{C}$ and the lattice partition $\mathbb{Z}^n/2\mathbb{Z}^n$ is an encoder for this mod-2 trellis code.

(b) Show that $\Lambda_{\mathcal{C}}$ has the group property.

(c) Show that $\Lambda_{\mathcal{C}}$ has the following parameters:

$$d^2_{\min}(\Lambda_{\mathcal{C}}) = \min\{d, 4\}; \tag{14.19}$$

$$K_{\min}(\Lambda_{\mathcal{C}}) = \begin{cases} 2^d N_d, & \text{if } d < 4; \\ 2n, & \text{if } d > 4; \\ 2^d N_d + 2n, & \text{if } d = 4; \end{cases} \tag{14.20}$$

$$\gamma_{\rm c}(\Lambda_{\mathcal{C}}) = d^2_{\min}(\Lambda_{\mathcal{C}}) 2^{-\eta(\mathcal{C})}, \tag{14.21}$$

where $\eta(\mathcal{C}) = 2(n-k)/n$ is the redundancy of $\mathcal{C}$ in bits per two dimensions. □

The encoder redundancy $\eta(\mathcal{C})$ also leads to a "coding constellation expansion ratio" which is a factor of $2^{\eta(\mathcal{C})}$ per two dimensions—*i.e.,* a factor of $4, 2, \sqrt{2}, \ldots$ for 1D, 2D, 4D,... codes, respectively. Minimization of coding constellation expansion has motivated the increasing use of higher-dimensional trellis codes.

A trellis code may be decoded by a Viterbi algorithm (VA) decoder, as follows. Given a received point $\mathbf{r}$ in $\mathbb{R}^{mn}$, the received first finds the closest signal point to $\mathbf{r}$ in each subset. A VA decoder then finds the closest code sequence to the entire received sequence. The decoding complexity is usually dominated by the complexity of the VA decoder, which to first order is dominated by the branch complexity $2^{\nu+k}$ of the convolutional code, normalized by the dimension mn.

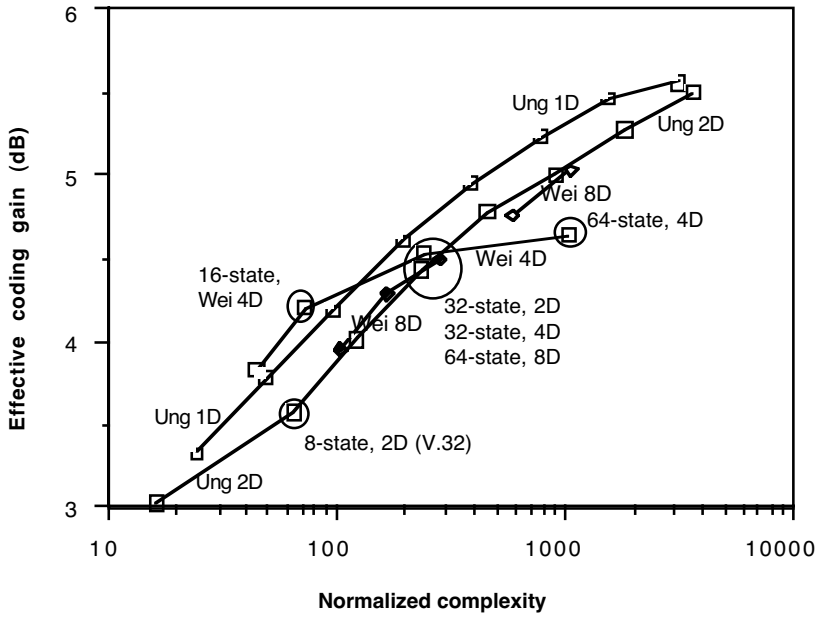

Figure 6. *Effective coding gain vs. complexity for Ungerboeck and Wei codes.*

Figure 6 shows the effective coding gains of certain important families of trellis codes versus their decoding complexity, measured by a detailed operation count. The codes considered are:

(a) The original 1D (PAM) trellis codes of Ungerboeck (1982), which are based on rate-1/2 convolutional codes ($\eta(\mathcal{C}) = 2$) with $2 \le \nu \le 9$ and the 4-way partition $\mathbb{Z}/4\mathbb{Z}$.

(b) The 2D (QAM) trellis codes of Ungerboeck, which (apart from the simplest 4-state code) are based on rate-2/3 convolutional codes ($\eta(\mathcal{C}) = 1$) with $3 \le \nu \le 9$ and the 8-way partition $\mathbb{Z}^2/2R\mathbb{Z}^2$.

(c) The 4D trellis codes of Wei (1987), all with $\eta(\mathcal{C}) = 1/2$, based on

  (a) rate-2/3 8- and 16-state convolutional codes and the 8-way partition $\mathbb{Z}^4/RD_4$;
  (b) a rate-3/4 32-state convolutional code and the 16-way partition $\mathbb{Z}^4/2\mathbb{Z}^4$;
  (c) a rate-4/5 64-state convolutional code and the 32-way partition $\mathbb{Z}^4/2D_4$.

(d) Two families of 8D trellis codes of Wei ($\eta(\mathcal{C}) = 1/4$).

The V.32 modem (1984) uses an 8-state 2D trellis code, also due to Wei (1984), whose performance/complexity tradeoff is the same as that as that of the original 8-state 2D Ungerboeck code, but which uses a nonlinear convolutional encoder to achieve 90° rotational invariance. This code has an effective coding gain of about 3.6 dB, a branch complexity of 2^5 (per two dimensions), and a coding constellation expansion ratio of 2.

The V.34 modem (1994) specifies three 4D trellis codes, with performance and complexity equivalent to the 4D Wei codes circled on Figure 6. All have a coding constellation expansion ratio of $\sqrt{2}$. The 16-state code is the original 16-state 4D Wei code, which has an effective coding gain of about 4.2 dB and a branch complexity of 2^6 (per four dimensions). The 32-state code is due to Williams and is based on the 16-way partition $\mathbb{Z}^4/H\mathbb{Z}^4$, where H is a 4×4 Hadamard matrix, to ensure that there are no minimum-distance error events whose length is only two dimensions; it has an effective coding gain of about 4.5 dB and a branch complexity of 2^8 (per four dimensions) The 64-state code is a modification of the original 4D Wei code, modified to prevent quasicatastrophic error propagation; it has an effective coding gain of about 4.7 dB and a branch complexity of 2^{10} (per four dimensions).

It is noteworthy that no one has improved on the performance vs. complexity tradeoff of the original 1D and 2D trellis codes of Ungerboeck or the subsequent multidimensional codes of Wei, and by this time it seems safe to predict that no one will ever do so. There have however been new trellis codes that enjoy other properties with about the same performance and complexity, such as those described in the previous two paragraphs, and there may still be room for further improvements of this kind.

Finally, we see that trellis codes have a performance/complexity advantage over lattice codes, when used with maximum-likelihood decoding. Effective coding gains of 4.2–4.7 dB, better than that of the Leech lattice L_{24} or of BW_{32}, are attainable with less complexity (and much less constellation expansion). 512-state 1D or 2D trellis codes can achieve effective coding gains of the order of 5.5 dB, which is superior to that of lattice codes of far greater complexity.

On the other hand, it seems very difficult to obtain effective coding gains of greater than 6 dB. This is not surprising, because at $P_s(E) \approx 10^{-6}$ the effective coding gain at the Shannon limit would be about 7.5 dB, and at the cutoff rate limit it would be about 5.8 dB. To approach the Shannon limit, much more complicated codes and decoding methods are necessary.

14.6 Sequential decoding in the high-SNR regime

In the bandwidth-limited regime, the cutoff rate limit is a factor of $4/e$ (1.68 dB) less than capacity. Therefore sequential decoders should be able to operate within about 1.7 dB of the Shannon limit; *i.e.,* sequential decoders should be able to achieve an effective coding gain of about 6 dB at $P_s(E) \approx 10^{-6}$. Several theses (Wang, Ljungberg, Maurer) have confirmed that sequential decoders are indeed capable of such performance.

14.7 Multilevel codes and multistage decoding

To approach the Shannon limit even more closely, it is clear that much more powerful codes must be used, with non-ML but near-ML decoding. Multilevel codes and multistage decoding may be used for this purpose. Multilevel coding may be based on a chain of sublattices of $\mathbb{Z}^n$,

$$\Lambda_0 = \mathbb{Z}^n \supseteq \Lambda_1 \supseteq \cdots \supseteq \Lambda_{r-1} \supseteq \Lambda_r,$$

which induce a chain of lattice partitions Λ_{j-1}/Λ_j, $1 \le j \le r$. A different encoder as in Figure 5 may be used independently on each such lattice partition. Moreover, with multistage decoding, each level is decoded independently.

Remarkably, such a multilevel scheme incurs no loss in channel capacity, compared to a single-level code based on the partition $\mathbb{Z}^n/\Lambda_r$; the capacity $C(\mathbb{Z}^n/\Lambda_r)$ of the partition $\mathbb{Z}^n/\Lambda_r$ is equal to the sum of the capacities $C(\Lambda_{j-1}/\Lambda_j)$ at each level. If the partition $\mathbb{Z}^n/\Lambda_r$ is "large enough" and appropriately scaled, then $C(\mathbb{Z}^n/\Lambda_r)$ approaches the capacity of the Gaussian channel.

All of the partitions Λ_{j-1}/Λ_j may even be binary; *e.g.,* one may use the standard one-dimensional or two-dimensional chains

$$\mathbb{Z} \supseteq 2\mathbb{Z} \supseteq 4\mathbb{Z} \supseteq 8\mathbb{Z} \supseteq \cdots;$$
$$\mathbb{Z}^2 \supseteq R\mathbb{Z}^2 \supseteq 2\mathbb{Z}^2 \supseteq 2R\mathbb{Z}^2 \supseteq 4\mathbb{Z}^2 \supseteq \cdots.$$

Then one can use a binary code of rate close to $C(\Lambda_{j-1}/\Lambda_j)$ at each level to approach the Shannon limit.

In particular, by using binary turbo codes of appropriate rate at each level, it has been shown that one can get within 1 dB of the Shannon limit (Wachsmann and Huber).

Powerful probabilistic coding methods such as turbo codes are really needed only at the higher levels. At the lower levels, the channels become quite clean and the capacity $C(\Lambda_{j-1}/\Lambda_j)$ approaches $\log_2 |\Lambda_{j-1}/\Lambda_j|$, so that the desired redundancy approaches zero. For these levels, algebraic codes and decoding methods may be more appropriate.

In summary, multilevel codes and multistage decoding allow the Shannon limit to be approached as closely in the bandwidth-limited regime as it can be approached in the power-limited regime with binary codes.

14.8 Multilevel turbo codes

A number of varieties of multilevel turbo codes based on multiple component trellis codes have been developed for the bandwidth-limited regime by several authors (*e.g.,* Berrou *et al.*, Benedetto *et al.*, Robertson and Wörz, Divsalar *et al.*). The performance of these codes seems to be comparable to that of binary turbo codes in the power-limited regime: *i.e.,* within about 1 dB of the Shannon limit. However, such capacity-approaching codes do not seem to have been implemented yet in practice, to the best of our knowledge.

14.9 Bit-interleaved coded modulation

In bit-interleaved coded modulation (BICM), the signals in a nonbinary constellation of size 2^b are selected by b randomly interleaved encoded bits from a binary encoder. The effective binary channel is then an equiprobable mixture of b parallel channels. The receiver knows which channel is used for each bit, and therefore can compute the correct APP vector for each symbol. Capacity-approaching codes may be designed for this mixture channel. While capacity is typically slightly reduced on an AWGN channel, the "pragmatic" BICM approach has become quite popular.