Chapter 5

Performance of small signal sets

In this chapter, we show how to estimate the performance of small-to-moderate-sized signal constellations on the discrete-time AWGN channel.

With equiprobable signal points in iid Gaussian noise, the optimum decision rule is a minimum-distance rule, so the optimum decision regions are minimum-distance (Voronoi) regions.

We develop useful performance estimates for the error probability based on the union bound. These are based on exact expressions for pairwise error probabilities, which involve the Gaussian probability of error $Q(\cdot)$ function. An appendix develops the main properties of this function.

Finally, we use the union bound estimate to find the "coding gain" of small-to-moderate-sized signal constellations in the power-limited and bandwidth-limited regimes, compared to the 2-PAM or $(M \times M)$-QAM baselines, respectively.

5.1 Signal constellations for the AWGN channel

In general, a coding scheme for the discrete-time AWGN channel model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ is a method of mapping an input bit sequence into a transmitted real symbol sequence $\mathbf{x}$, which is called encoding, and a method for mapping a received real symbol sequence $\mathbf{y}$ into an estimated transmitted signal sequence $\hat{\mathbf{x}}$, which is called decoding.

Initially we will consider coding schemes of the type considered by Shannon, namely block codes with a fixed block length N. With such codes, the transmitted sequence $\mathbf{x}$ consists of a sequence $(\ldots, \mathbf{x}_k, \mathbf{x}_{k+1}, \ldots)$ of N-tuples $\mathbf{x}_k \in \mathbb{R}^N$ that are chosen *independently* from some block code of length N with M codewords. Block codes are not the only possible kinds of coding schemes, as we will see when we study convolutional and trellis codes.

Usually the number M of codewords is chosen to be a power of 2, and codewords are chosen by some encoding map from blocks of $\log_2 M$ bits in the input bit sequence. If the input bit sequence is assumed to be an iid random sequence of equiprobable bits, then the transmitted sequence will be an iid random sequence $\mathbf{X} = (\ldots, \mathbf{X}_k, \mathbf{X}_{k+1}, \ldots)$ of equiprobable random codewords $\mathbf{X}_k$. We almost always assume equiprobability, because this is a worst-case (minimax) assumption. Also, the bit sequence produced by an efficient source coder must statistically resemble an iid equiprobable bit sequence.

In digital communications, we usually focus entirely on the code, and do not care what encoding map is used from bits to codewords. In other contexts the encoding map is also important; *e.g.,* in the "Morse code" of telegraphy.

If the block length N and the number of codewords M are relatively small, then a block code for the AWGN channel may alternatively be called a signal set, signal constellation, or signal alphabet. A scheme in which the block length N is 1 or 2, corresponding to a single signaling interval of PAM or QAM, may be regarded as an "uncoded" scheme.

Figure 1 illustrates some 1-dimensional and 2-dimensional signal constellations.

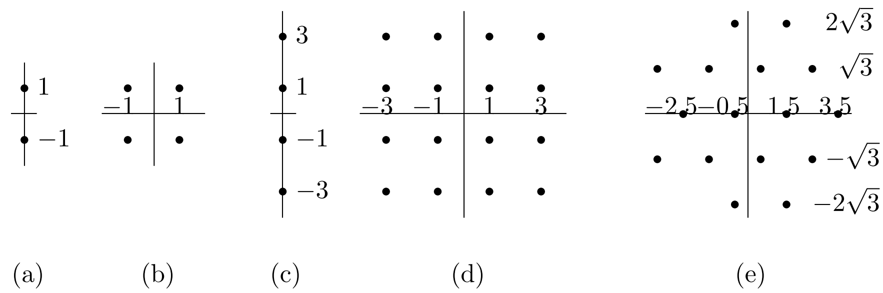


Figure 1. Uncoded signal constellations: (a) 2-PAM; (b) (2×2)-QAM; (c) 4-PAM; (d) (4×4)-QAM; (e) hexagonal 16-QAM.

An N-dimensional *signal constellation* (set, alphabet) will be denoted by

$$\mathcal{A} = \{\mathbf{a}_j, 1 \leq j \leq M\}.$$

Its M elements $\mathbf{a}_j \in \mathbb{R}^N$ will be called *signal points* (vectors, N-tuples).

The basic parameters of a signal constellation $\mathcal{A} = \{\mathbf{a}_j, 1 \leq j \leq M\}$ are its *dimension* N; its *size* M (number of signal points); its *average energy* $E(\mathcal{A}) = \frac{1}{M}\sum_j ||\mathbf{a}_j||^2$; and its *minimum squared distance* $d^2_{\min}(\mathcal{A})$, which is an elementary measure of its noise resistance. A secondary parameter is the average number $K_{\min}(\mathcal{A})$ of nearest neighbors (points at distance $d_{\min}(\mathcal{A})$).

From these basic parameters we can derive such parameters as:

- The bit rate (nominal spectral efficiency) $\rho = (2/N)\log_2 M$ b/2D;
- The average energy per two dimensions $E_s = (2/N)E(\mathcal{A})$,
  or the average energy per bit $E_b = E(\mathcal{A})/(\log_2 M) = E_s/\rho$;
- Energy-normalized figures of merit such as $d^2_{\min}(\mathcal{A})/E(\mathcal{A}), d^2_{\min}(\mathcal{A})/E_s$ or $d^2_{\min}(\mathcal{A})/E_b$, which are independent of scale.

For example, in Figure 1, the bit rate (nominal spectral efficiency) of the 2-PAM and (2×2)-QAM constellations is $\rho = 2$ b/2D, whereas for the other three constellations it is $\rho = 4$ b/2D. The average energy per two dimensions of the 2-PAM and (2×2)-QAM constellations is $E_s = 2$, whereas for the 4-PAM and (4×4)-QAM constellations it is $E_s = 10$, and for the hexagonal 16-QAM constellation it is $E_s = 8.75$. For all constellations, $d^2_{\min} = 4$. The average numbers of nearest neighbors are $K_{\min} = 1, 2, 1.5, 3$, and 4.125, respectively.

5.1.1 Cartesian-product constellations

Some of these relations are explained by the fact that an $(M \times M)$-QAM constellation is the Cartesian product of two M-PAM constellations. In general, a *Cartesian-product constellation* $\mathcal{A}^K$ is the set of all sequences of K points from an elementary constellation $\mathcal{A}$; *i.e.,*

$$\mathcal{A}^K = \{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K) \mid \mathbf{x}_k \in \mathcal{A}\}.$$

If the dimension and size of $\mathcal{A}$ are N and M, respectively, then the dimension of $\mathcal{A}' = \mathcal{A}^K$ is $N' = KN$ and its size is $M' = M^K$.

Exercise 1 (Cartesian-product constellations). (a) Show that if $\mathcal{A}' = \mathcal{A}^K$, then the parameters $N, \log_2 M, E(\mathcal{A}')$ and $K_{\min}(\mathcal{A}')$ of $\mathcal{A}'$ are K times as large as the corresponding parameters of $\mathcal{A}$, whereas the normalized parameters ρ, E_s, E_b and $d^2_{\min}(\mathcal{A})$ are the same as those of $\mathcal{A}$. Verify that these relations hold for the $(M \times M)$-QAM constellations of Figure 1. □

Notice that there is no difference between a random input sequence $\mathbf{X}$ with elements from $\mathcal{A}$ and a sequence $\mathbf{X}$ with elements from a Cartesian-product constellation $\mathcal{A}^K$. For example, there is no difference between a random M-PAM sequence and a random $(M \times M)$-QAM sequence. Thus Cartesian-product constellations capture in a non-statistical way the idea of independent transmissions. We thus may regard a Cartesian-product constellation $\mathcal{A}^K$ as equivalent to (or a "version" of) the elementary constellation $\mathcal{A}$. In particular, it has the same ρ, E_s, E_b and $d^2_{\min}$.

We may further define a "code over $\mathcal{A}$" as a subset $\mathcal{C} \subset \mathcal{A}^K$ of a Cartesian-product constellation $\mathcal{A}^K$. In general, a code $\mathcal{C}$ over $\mathcal{A}$ will have a lower bit rate (nominal spectral efficiency) ρ than $\mathcal{A}$, but a higher minimum squared distance $d^2_{\min}$. Via this tradeoff, we hope to achieve a "coding gain." Practically all of the codes that we will consider in later chapters will be of this type.

5.1.2 Minimum-distance decoding

Again, a decoding scheme is a method for mapping the received sequence into an estimate of the transmitted signal sequence. (Sometimes the decoder does more than this, but this definition will do for a start.)

If the encoding scheme is a block scheme, then it is plausible that the receiver should decode block-by-block as well. That there is no loss of optimality in block-by-block decoding can be shown from the theorem of irrelevance, or alternatively by an extension of the exercise involving Cartesian-product constellations at the end of this subsection.

We will now recapitulate how for block-by-block decoding, with equiprobable signals and iid Gaussian noise, the optimum decision rule is a minimum-distance (MD) rule.

For block-by-block decoding, the channel model is $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where all sequences are N-tuples. The transmitted sequence $\mathbf{X}$ is chosen equiprobably from the M N-tuples $\mathbf{a}_j$ in a signal constellation $\mathcal{A}$. The noise pdf is

$$p_N(\mathbf{n}) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-||\mathbf{n}||^2/2\sigma^2},$$

where the symbol variance is $\sigma^2 = N_0/2$.

In digital communications, we are usually interested in the *minimum-probability-of-error* (MPE) decision rule: given a received vector $\mathbf{y}$, choose the signal point $\hat{\mathbf{a}} \in \mathcal{A}$ to minimize the probability of decision error $\Pr(E)$.

Since the probability that a decision $\hat{\mathbf{a}}$ is correct is simply the *a posteriori* probability $p(\hat{\mathbf{a}} \mid \mathbf{y})$, the MPE rule is equivalent to the *maximum-a-posteriori-probability* (MAP) rule: choose the $\hat{\mathbf{a}} \in \mathcal{A}$ such that $p(\hat{\mathbf{a}} \mid \mathbf{y})$ is maximum among all $p(\mathbf{a}_j \mid \mathbf{y}), \mathbf{a}_j \in \mathcal{A}$.

By Bayes' law,

$$p(\mathbf{a}_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{a}_j)p(\mathbf{a}_j)}{p(\mathbf{y})}.$$

If the signals $\mathbf{a}_j$ are equiprobable, so $p(\mathbf{a}_j) = 1/M$ for all j, then the MAP rule is equivalent to the *maximum-likelihood* (ML) rule: choose the $\hat{\mathbf{a}} \in \mathcal{A}$ such that $p(\mathbf{y} \mid \hat{\mathbf{a}})$ is maximum among all $p(\mathbf{y} \mid \mathbf{a}_j), \mathbf{a}_j \in \mathcal{A}$.

Using the noise pdf, we can write

$$p(\mathbf{y} \mid \mathbf{a}_j) = p_N(\mathbf{y} - \mathbf{a}_j) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-||\mathbf{y}-\mathbf{a}_j||^2/2\sigma^2}.$$

Therefore the ML rule is equivalent to the *minimum-distance* (MD) rule: choose the $\hat{\mathbf{a}} \in \mathcal{A}$ such that $||\mathbf{y} - \hat{\mathbf{a}}||^2$ is minimum among all $||\mathbf{y} - \mathbf{a}_j||^2, \mathbf{a}_j \in \mathcal{A}$.

In summary, under the assumption of equiprobable inputs and iid Gaussian noise, the MPE rule is the minimum-distance rule. Therefore from this point forward we consider only MD detection, which is easy to understand from a geometrical point of view.

Exercise 1 (Cartesian-product constellations, cont.).

(b) Show that if the signal constellation is a Cartesian product $\mathcal{A}^K$, then MD detection can be performed by performing independent MD detection on each of the K components of the received KN-tuple $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$. Using this result, sketch the decision regions of the (4×4)-QAM signal set of Figure 1(d).

(c) Show that if $\Pr(E)$ is the probability of error for MD detection of $\mathcal{A}$, then the probability of error for MD detection of $\mathcal{A}'$ is

$$\Pr(E)' = 1 - (1 - \Pr(E))^K,$$

Show that $\Pr(E)' \approx K \Pr(E)$ if $\Pr(E)$ is small. □

Example 1. The K-fold Cartesian product $\mathcal{A}' = \mathcal{A}^K$ of a 2-PAM signal set $\mathcal{A} = \{\pm\alpha\}$ corresponds to independent transmission of K bits using 2-PAM. Geometrically, $\mathcal{A}'$ is the vertex set of a K-cube of side 2α. For example, for $K = 2$, $\mathcal{A}'$ is the (2×2)-QAM constellation of Figure 1(b).

From Exercise 1(a), the K-cube constellation $\mathcal{A}' = \mathcal{A}^K$ has dimension $N' = K$, size $M' = 2^K$, bit rate (nominal spectral efficiency) $\rho = 2$ b/2D, average energy $E(\mathcal{A}') = K\alpha^2$, average energy per bit $E_b = \alpha^2$, minimum squared distance $d^2_{\min}(\mathcal{A}') = 4\alpha^2$, and average number of nearest neighbors $K'_{\min}(\mathcal{A}') = K$. From Exercise 1(c), its probability of error is approximately K times the single-bit error probability:

$$\Pr(E)' \approx KQ\left(\sqrt{2E_b/N_0}\right).$$

Consequently, if we define the probability of error per bit as $P_b(E) = \Pr(E)'/K$, then we obtain the curve of (4.3) for all K-cube constellations:

$$P_b(E) \approx Q\left(\sqrt{2E_b/N_0}\right),$$

including the (2×2)-QAM constellation of Figure 1(b).

A code over the 2-PAM signal set $\mathcal{A}$ is thus simply a subset of the vertices of a K-cube. □

5.1.3 Decision regions

Under a minimum-distance (MD) decision rule, real N-space $\mathbb{R}^N$ is partitioned into M *decision regions* $\mathcal{R}_j, 1 \leq j \leq M$, where $\mathcal{R}_j$ consists of the received vectors $\mathbf{y} \in \mathbb{R}^N$ that are at least as close to $\mathbf{a}_j$ as to any other point in $\mathcal{A}$:

$$\mathcal{R}_j = \{\mathbf{y} \in \mathbb{R}^N : ||\mathbf{y} - \mathbf{a}_j||^2 \leq ||\mathbf{y} - \mathbf{a}_{j'}||^2 \text{ for all } j' \neq j\}. \quad (5.1)$$

The minimum-distance regions $\mathcal{R}_j$ are also called *Voronoi regions.* Under the MD rule, given a received sequence $\mathbf{y}$, the decision is $\mathbf{a}_j$ only if $\mathbf{y} \in \mathcal{R}_j$. The decision regions $\mathcal{R}_j$ cover all of N-space $\mathbb{R}^N$, and are disjoint except on their boundaries.

Since the noise vector $\mathbf{N}$ is a continuous random vector, the probability that $\mathbf{y}$ will actually fall precisely on the boundary of $\mathcal{R}_j$ is zero, so in that case it does not matter which decision is made.

The decision region $\mathcal{R}_j$ is the intersection of the $M-1$ pairwise decision regions $\mathcal{R}_{jj'}$ defined by

$$\mathcal{R}_{jj'} = \{\mathbf{y} \in \mathbb{R}^N : ||\mathbf{y} - \mathbf{a}_j||^2 \leq ||\mathbf{y} - \mathbf{a}_{j'}||^2\}.$$

Geometrically, it is obvious that $\mathcal{R}_{jj'}$ is the half-space containing $\mathbf{a}_j$ that is bounded by the perpendicular bisector hyperplane $\mathcal{H}_{jj'}$ between $\mathbf{a}_j$ and $\mathbf{a}_{j'}$, as shown in Figure 2.

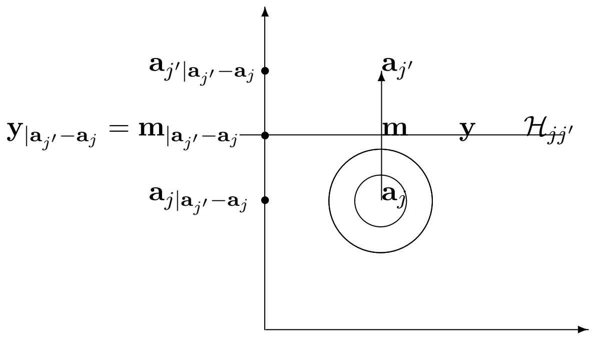


Figure 2. The boundary hyperplane $\mathcal{H}_{jj'}$ is the perpendicular bisector between $\mathbf{a}_j$ and $\mathbf{a}_{j'}$.

Algebraically, since $\mathcal{H}_{jj'}$ is the set of points in $\mathbb{R}^N$ that are equidistant from $\mathbf{a}_j$ and $\mathbf{a}_{j'}$, it is characterized by the following equivalent equations:

$$\begin{aligned} ||\mathbf{y} - \mathbf{a}_j||^2 &= ||\mathbf{y} - \mathbf{a}_{j'}||^2; \\ -2\langle \mathbf{y}, \mathbf{a}_j\rangle + ||\mathbf{a}_j||^2 &= -2\langle \mathbf{y}, \mathbf{a}_{j'}\rangle + ||\mathbf{a}_{j'}||^2; \\ \langle \mathbf{y}, \mathbf{a}_{j'} - \mathbf{a}_j\rangle &= \langle \frac{\mathbf{a}_j + \mathbf{a}_{j'}}{2}, \mathbf{a}_{j'} - \mathbf{a}_j\rangle = \langle \mathbf{m}, \mathbf{a}_{j'} - \mathbf{a}_j\rangle. \end{aligned} \quad (5.2)$$

where $\mathbf{m}$ denotes the midvector $\mathbf{m} = (\mathbf{a}_j + \mathbf{a}_{j'})/2$. If the difference vector between $\mathbf{a}_{j'}$ and $\mathbf{a}_j$ is $\mathbf{a}_{j'} - \mathbf{a}_j$ and

$$\boldsymbol{\phi}_{j\to j'} = \frac{\mathbf{a}_{j'} - \mathbf{a}_j}{||\mathbf{a}_{j'} - \mathbf{a}_j||}$$

is the normalized difference vector, so that $||\boldsymbol{\phi}_{j\to j'}||^2 = 1$, then the projection of any vector $\mathbf{x}$ onto the difference vector $\mathbf{a}_{j'} - \mathbf{a}_j$ is

$$\mathbf{x}_{|\mathbf{a}_{j'}-\mathbf{a}_j} = \langle \mathbf{x}, \boldsymbol{\phi}_{j\to j'}\rangle \boldsymbol{\phi}_{j\to j'} = \frac{\langle \mathbf{x}, \mathbf{a}_{j'} - \mathbf{a}_j\rangle}{||\mathbf{a}_{j'} - \mathbf{a}_j||^2}(\mathbf{a}_{j'} - \mathbf{a}_j).$$

The geometric meaning of Equation (5.2) is thus that $\mathbf{y} \in \mathcal{H}_{jj'}$ if and only if the projection $\mathbf{y}_{|\mathbf{a}_{j'}-\mathbf{a}_j}$ of $\mathbf{y}$ onto the difference vector $\mathbf{a}_{j'}-\mathbf{a}_j$ is equal to the projection $\mathbf{m}_{|\mathbf{a}_{j'}-\mathbf{a}_j}$ of the midvector $\mathbf{m} = (\mathbf{a}_j + \mathbf{a}_{j'})/2$ onto the difference vector $\mathbf{a}_j - \mathbf{a}_{j'}$, as illustrated in Figure 2.

The decision region $\mathcal{R}_j$ is the intersection of these $M-1$ half-spaces:

$$\mathcal{R}_j = \bigcap_{j' \neq j} \mathcal{R}_{jj'}.$$

(Equivalently, the complementary region $\overline{\mathcal{R}_j}$ is the union of the complementary half-spaces $\overline{\mathcal{R}_{jj'}}$.) A decision region $\mathcal{R}_j$ is therefore a convex polytope bounded by portions of a subset $\{\mathcal{H}_{jj'}, \mathbf{a}_{j'} \in \mathcal{N}(\mathbf{a}_j)\}$ of the boundary hyperplanes $\mathcal{H}_{jj'}$, where the subset $\mathcal{N}(\mathbf{a}_j) \subseteq \mathcal{A}$ of neighbors of $\mathbf{a}_j$ that contribute boundary faces to this polytope is called the *relevant subset*. It is easy to see that the relevant subset must always include the nearest neighbors to $\mathbf{a}_j$.

5.2 Probability of decision error

The probability of decision error given that $\mathbf{a}_j$ is transmitted is the probability that $\mathbf{Y} = \mathbf{a}_j + \mathbf{N}$ falls outside the decision region $\mathcal{R}_j$, whose "center" is $\mathbf{a}_j$. Equivalently, it is the probability that the noise variable $\mathbf{N}$ falls outside the translated region $\mathcal{R}_j - \mathbf{a}_j$, whose "center" is $\mathbf{0}$:

$$\Pr(E \mid \mathbf{a}_j) = 1 - \int_{\mathcal{R}_j} p_Y(\mathbf{y} \mid \mathbf{a}_j)\, d\mathbf{y} = 1 - \int_{\mathcal{R}_j} p_N(\mathbf{y} - \mathbf{a}_j)\, d\mathbf{y} = 1 - \int_{\mathcal{R}_j - \mathbf{a}_j} p_N(\mathbf{n})\, d\mathbf{n}.$$

Exercise 2 (error probability invariance). (a) Show that the probabilities of error $\Pr(E \mid \mathbf{a}_j)$ are unchanged if $\mathcal{A}$ is translated by any vector $\mathbf{v}$; *i.e.,* the constellation $\mathcal{A}' = \mathcal{A} + \mathbf{v}$ has the same error probability $\Pr(E)$ as $\mathcal{A}$.

(b) Show that $\Pr(E)$ is invariant under orthogonal transformations; *i.e.,* the constellation $\mathcal{A}' = U\mathcal{A}$ has the same $\Pr(E)$ as $\mathcal{A}$ when U is any orthogonal $N \times N$ matrix (*i.e.,* $U^{-1} = U^T$).

(c) Show that $\Pr(E)$ is unchanged if both the constellation $\mathcal{A}$ and the noise $\mathbf{N}$ are scaled by the same scale factor $\alpha > 0$. □

Exercise 3 (optimality of zero-mean constellations). Consider an arbitrary signal set $\mathcal{A} = \{\mathbf{a}_j, 1 \le j \le M\}$. Assume that all signals are equiprobable. Let $\mathbf{m}(\mathcal{A}) = \frac{1}{M}\sum_j \mathbf{a}_j$ be the average signal, and let $\mathcal{A}'$ be $\mathcal{A}$ translated by $\mathbf{m}(\mathcal{A})$ so that the mean of $\mathcal{A}'$ is zero:

$$\mathcal{A}' = \mathcal{A} - \mathbf{m}(\mathcal{A}) = \{\mathbf{a}_j - \mathbf{m}(\mathcal{A}), 1 \le j \le M\}.$$

Let $E(\mathcal{A})$ and $E(\mathcal{A}')$ denote the average energies of $\mathcal{A}$ and $\mathcal{A}'$, respectively.

(a) Show that the error probability of an optimum detector is the same for $\mathcal{A}'$ as it is for $\mathcal{A}$.

(b) Show that $E(\mathcal{A}') = E(\mathcal{A}) - ||\mathbf{m}(\mathcal{A})||^2$. Conclude that removing the mean $\mathbf{m}(\mathcal{A})$ is always a good idea.

(c) Show that a binary antipodal signal set $\mathcal{A} = \{\pm\mathbf{a}\}$ is always optimal for $M = 2$. □

In general, there is no closed-form expression for the Gaussian integral $\Pr(E \mid \mathbf{a}_j)$. However, we can obtain an upper bound in terms of pairwise error probabilities, called the union bound, which is usually quite sharp. The first term of the union bound, called the union bound estimate, is usually an excellent approximation, and will be the basis for our analysis of coding gains of small-to-moderate-sized constellations. A lower bound with the same exponential behavior may be obtained by considering only the worst-case pairwise error probability.

5.2.1 Pairwise error probabilities

We now show that each pairwise probability has a simple closed-form expression that depends only on the squared distance $d^2(\mathbf{a}_j, \mathbf{a}_{j'}) = ||\mathbf{a}_j - \mathbf{a}_{j'}||^2$ and the noise variance $\sigma^2 = N_0/2$.

From Figure 2, it is clear that whether $\mathbf{y} = \mathbf{a}_j + \mathbf{n}$ is closer to $\mathbf{a}_{j'}$ than to $\mathbf{a}_j$ depends only on the projection $\mathbf{y}_{|\mathbf{a}_{j'}-\mathbf{a}_j}$ of $\mathbf{y}$ onto the difference vector $\mathbf{a}_{j'} - \mathbf{a}_j$. In fact, from (5.2), an error can occur if and only if

$$|\mathbf{n}_{|\mathbf{a}_{j'}-\mathbf{a}_j}| = |\langle \mathbf{n}, \boldsymbol{\phi}_{j\to j'}\rangle| = \frac{|\langle \mathbf{n}, \mathbf{a}_{j'} - \mathbf{a}_j\rangle|}{||\mathbf{a}_{j'} - \mathbf{a}_j||} \quad \geq \quad \frac{\langle \mathbf{a}_{j'} - \mathbf{a}_j, \mathbf{a}_{j'} - \mathbf{a}_j\rangle}{2||\mathbf{a}_{j'} - \mathbf{a}_j||} = \frac{||\mathbf{a}_{j'} - \mathbf{a}_j||}{2}.$$

In other words, an error can occur if and only if the magnitude of the one-dimensional noise component $n_1 = \mathbf{n}_{|\mathbf{a}_{j'}-\mathbf{a}_j}$, the projection of $\mathbf{n}$ onto the difference vector $\mathbf{a}_{j'} - \mathbf{a}_j$, exceeds half the distance $d(\mathbf{a}_{j'}, \mathbf{a}_j) = ||\mathbf{a}_{j'} - \mathbf{a}_j||$ between $\mathbf{a}_{j'}$ and $\mathbf{a}_j$.

We now use the fact that the distribution $p_N(\mathbf{n})$ of the iid Gaussian noise vector $\mathbf{N}$ is spherically symmetric, so the pdf of any one-dimensional projection such as n_1 is

$$p_N(n_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-n_1^2/2\sigma^2}.$$

In other words, $\mathbf{N}$ is an iid zero-mean Gaussian vector with variance σ^2 in any coordinate system, including a coordinate system in which the first coordinate axis is aligned with the vector $\mathbf{a}_{j'} - \mathbf{a}_j$.

Consequently, the pairwise error probability $\Pr\{\mathbf{a}_j \to \mathbf{a}_{j'}\}$ that if $\mathbf{a}_j$ is transmitted, the received vector $\mathbf{y} = \mathbf{a}_j + \mathbf{n}$ will be at least as close to $\mathbf{a}_{j'}$ as to $\mathbf{a}_j$ is given simply by

$$\Pr\{\mathbf{a}_j \to \mathbf{a}_{j'}\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{d(\mathbf{a}_{j'},\mathbf{a}_j)/2}^{\infty} e^{-x^2/2\sigma^2} dx = Q\left(\frac{d(\mathbf{a}_{j'}, \mathbf{a}_j)}{2\sigma}\right), \tag{5.3}$$

where $Q(\cdot)$ is again the Gaussian probability of error function.

As we have seen, the probability of error for a 2-PAM signal set $\{\pm\alpha\}$ is $Q(\alpha/\sigma)$. Since the distance between the two signals is $d = 2\alpha$, this is just a special case of this general formula.

In summary, the spherical symmetry of iid Gaussian noise leads to the remarkable result that the pairwise error probability from $\mathbf{a}_j$ to $\mathbf{a}_{j'}$ depends only on the squared distance $d^2(\mathbf{a}_{j'}, \mathbf{a}_j) = ||\mathbf{a}_{j'} - \mathbf{a}_j||^2$ between $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ and the noise variance σ^2.

Exercise 4 (non-equiprobable signals).

Let $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ be two signals that are not equiprobable. Find the optimum (MPE) pairwise decision rule and pairwise error probability $\Pr\{\mathbf{a}_j \to \mathbf{a}_{j'}\}$. □

5.2.2 The union bound and the UBE

The *union bound* on error probability is based on the elementary union bound of probability theory: if A and B are any two events, then $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$. Thus the probability of detection error $\Pr(E \mid \mathbf{a}_j)$ with minimum-distance detection if $\mathbf{a}_j$ is sent— *i.e.,* the probability that $\mathbf{y}$ will be closer to some other $\mathbf{a}_{j'} \in \mathcal{A}$ than to $\mathbf{a}_j$– is upperbounded by the sum of the pairwise error probabilities to all other signals $\mathbf{a}_{j'} \neq \mathbf{a}_j \in \mathcal{A}$:

$$\Pr(E \mid \mathbf{a}_j) \leq \sum_{\mathbf{a}_j \neq \mathbf{a}_{j'} \in \mathcal{A}} \Pr\{\mathbf{a}_j \to \mathbf{a}_{j'}\} = \sum_{\mathbf{a}_j \neq \mathbf{a}_{j'} \in \mathcal{A}} Q\left(\frac{d(\mathbf{a}_j, \mathbf{a}_{j'})}{2\sigma}\right).$$

Let D denote the set of distances between signal points in $\mathcal{A}$; then we can write the union bound as

$$\Pr(E \mid \mathbf{a}_j) \leq \sum_{d \in D} K_d(\mathbf{a}_j) Q\left(\frac{d}{2\sigma}\right), \tag{5.4}$$

where $K_d(\mathbf{a}_j)$ is the number of signals $\mathbf{a}_{j'} \neq \mathbf{a}_j \in \mathcal{A}$ at distance d from $\mathbf{a}_j$. Because $Q(x)$ decreases exponentially as $e^{-x^2/2}$ (see Appendix), the factor $Q(d/2\sigma)$ will be largest for the minimum Euclidean distance

$$d_{\min}(\mathcal{A}) = \min_{\mathbf{a}_j \neq \mathbf{a}_{j'} \in \mathcal{A}} ||\mathbf{a}_{j'} - \mathbf{a}_j||,$$

and will decrease rapidly for larger distances.

The *union bound estimate* (UBE) of $\Pr(E \mid \mathbf{a}_j)$ is based on the idea that the nearest neighbors to $\mathbf{a}_j$ at distance $d_{\min}(\mathcal{A})$ (if there are any) will dominate this sum. If there are $K_{\min}(\mathbf{a}_j)$ neighbors at distance $d_{\min}(\mathcal{A})$ from $\mathbf{a}_j$, then

$$\Pr(E \mid \mathbf{a}_j) \approx K_{\min}(\mathbf{a}_j)\, Q\left(\frac{d_{\min}(\mathcal{A})}{2\sigma}\right). \tag{5.5}$$

Of course this estimate is valid only if the next nearest neighbors are at a significantly greater distance and there are not too many of them; if these assumptions are violated, then further terms should be used in the estimate.

The union bound may be somewhat sharpened by considering only signals in the relevant subset $\mathcal{N}(\mathbf{a}_j)$ that determine faces of the decision region $\mathcal{R}_j$. However, since $\mathcal{N}(\mathbf{a}_j)$ includes all nearest neighbors at distance $d_{\min}(\mathcal{A})$, this will not affect the UBE.

Finally, if there is at least one neighbor $\mathbf{a}_{j'}$ at distance $d_{\min}(\mathcal{A})$ from $\mathbf{a}_j$, then we have the *pairwise lower bound*

$$\Pr(E \mid \mathbf{a}_j) \geq \Pr\{\mathbf{a}_j \to \mathbf{a}_{j'}\} = Q\left(\frac{d_{\min}(\mathcal{A})}{2\sigma}\right), \tag{5.6}$$

since there must be a detection error if $\mathbf{y}$ is closer to $\mathbf{a}_{j'}$ than to $\mathbf{a}_j$. Thus we are usually able to obtain upper and lower bounds on $\Pr(E \mid \mathbf{a}_j)$ that have the same "exponent" (argument of the $Q(\cdot)$ function) and that differ only by a small factor of the order of $K_{\min}(\mathbf{a}_j)$.

We can obtain similar upper and lower bounds and estimates for the total error probability

$$\Pr(E) = \overline{\Pr(E \mid \mathbf{a}_j)},$$

where the overbar denotes the expectation over the equiprobable ensemble of signals in $\mathcal{A}$. For example, if $K_{\min}(\mathcal{A}) = \overline{K_{\min}(\mathbf{a}_j)}$ is the average number of nearest neighbors at distance $d_{\min}(\mathcal{A})$, then the union bound estimate of $\Pr(E)$ is

$$\Pr(E) \approx K_{\min}(\mathcal{A}) Q\left(\frac{d_{\min}(\mathcal{A})}{2\sigma}\right). \tag{5.7}$$

Exercise 5 (UBE for M-PAM constellations). For an M-PAM constellation $\mathcal{A}$, show that $K_{\min}(\mathcal{A}) = 2(M-1)/M$. Conclude that the union bound estimate of $\Pr(E)$ is

$$\Pr(E) \approx 2\left(\frac{M-1}{M}\right) Q\left(\frac{d}{2\sigma}\right).$$

Show that in this case the union bound estimate is exact. Explain why. □

5.3 Performance analysis in the power-limited regime

Recall that the power-limited regime is defined as the domain in which the nominal spectral efficiency ρ is not greater than 2 b/2D. In this regime we normalize all quantities "per bit," and generally use E_b/N_0 as our normalized measure of signal-to-noise ratio.

The baseline uncoded signal set in this regime is the one-dimensional 2-PAM signal set $\mathcal{A} = \{\pm\alpha\}$, or equivalently a K-cube constellation $\mathcal{A}^K$. Such a constellation has bit rate (nominal spectral efficiency) $\rho = 2$ b/2D, average energy ber bit $E_b = \alpha^2$, minimum squared distance $d^2_{\min}(\mathcal{A}) = 4\alpha^2$, and average number of nearest neighbors per bit $K_b(\mathcal{A}) = 1$. By the UBE (5.7), its error probability per bit is given by

$$P_b(E) \approx Q^{\surd}(2E_b/N_0), \tag{5.8}$$

where we now use the "Q-of-the-square-root-of" function $Q^{\surd}$, defined by $Q^{\surd}(x) = Q(\sqrt{x})$ (see Appendix). This baseline curve of $P_b(E)$ *vs.* E_b/N_0 is plotted in Chapter 4, Figure 1.

The *effective coding gain* $\gamma_{\text{eff}}(\mathcal{A})$ of a signal set $\mathcal{A}$ at a given target error probability per bit $P_b(E)$ will be defined as the difference in dB between the E_b/N_0 required to achieve the target $P_b(E)$ with $\mathcal{A}$ and the E_b/N_0 required to achieve the target $P_b(E)$ with 2-PAM (*i.e.,* no coding).

For example, we have seen that the maximum possible effective coding gain at $P_b(E) \approx 10^{-5}$ is approximately 11.2 dB. For lower $P_b(E)$, the maximum possible gain is higher, and for higher $P_b(E)$, the maximum possible gain is lower.

In this definition, the effective coding gain includes any gains that result from using a lower nominal spectral efficiency $\rho < 2$ b/2D, which as we have seen can range up to 3.35 dB. If ρ is held constant at $\rho = 2$ b/2D, then the maximum possible effective coding gain is lower; *e.g.,* at $P_b(E) \approx 10^{-5}$ it is approximately 8 dB. If there is a constraint on ρ (bandwidth), then it is better to plot $P_b(E)$ *vs.* $\mathrm{SNR}_{\text{norm}}$, especially to measure how far $\mathcal{A}$ is from achieving capacity.

The UBE allows us to estimate the effective coding gain as follows. The probability of error per bit (not in general the same as the bit error probability!) is

$$P_b(E) = \frac{\Pr(E)}{\log_2 |\mathcal{A}|} \approx \frac{K_{\min}(\mathcal{A})}{\log_2 |\mathcal{A}|} Q^{\surd}\left(\frac{d^2_{\min}(\mathcal{A})}{2N_0}\right),$$

since $Q^{\surd}\left(d^2_{\min}(\mathcal{A})/2N_0\right) = Q(d_{\min}(\mathcal{A})/2\sigma)$. In the power-limited regime, we define the *nominal coding gain* $\gamma_c(\mathcal{A})$ as

$$\gamma_c(\mathcal{A}) = \frac{d^2_{\min}(\mathcal{A})}{4E_b}. \tag{5.9}$$

This definition is normalized so that for 2-PAM, $\gamma_c(\mathcal{A}) = 1$. Because nominal coding gain is a multiplicative factor in the argument of the $Q^{\surd}(\cdot)$ function, it is often measured in dB. The UBE then becomes

$$P_b(E) \approx K_b(\mathcal{A}) Q^{\surd}(2\gamma_c(\mathcal{A}) E_b/N_0), \tag{5.10}$$

where $K_b(\mathcal{A}) = K_{\min}(\mathcal{A})/\log_2 |\mathcal{A}|$ is the average number of nearest neighbors per transmitted bit. Note that for 2-PAM, this expression is exact.

Given $\gamma_c(\mathcal{A})$ and $K_b(\mathcal{A})$, we may obtain a plot of the UBE (5.10) simply by moving the baseline curve (Figure 1 of Chapter 4) to the left by $\gamma_c(\mathcal{A})$ (in dB), and then up by a factor of $K_b(\mathcal{A})$, since $P_b(E)$ is plotted on a log scale. (This is an excellent reason why error probability curves are always plotted on a log-log scale, with SNR measured in dB.)

Thus if $K_b(\mathcal{A}) = 1$, then the effective coding gain $\gamma_{\rm eff}(\mathcal{A})$ is equal to the nominal coding gain $\gamma_{\rm c}(\mathcal{A})$ for all $P_b(E)$, to the accuracy of the UBE. However, if $K_b(\mathcal{A}) > 1$, then the effective coding gain is less than the nominal coding gain by an amount which depends on the steepness of the $P_b(E)$ *vs.* E_b/N_0 curve at the target $P_b(E)$. At $P_b(E) \approx 10^{-5}$, a rule of thumb which is fairly accurate if $K_b(\mathcal{A})$ is not too large is that an increase of a factor of two in $K_b(\mathcal{A})$ costs about 0.2 dB in effective coding gain; *i.e.*,

$$\gamma_{\rm eff}(\mathcal{A}) \approx \gamma_{\rm c}(\mathcal{A}) - (0.2)(\log_2 K_b(\mathcal{A})) \quad \text{(in dB)}. \tag{5.11}$$

A more accurate estimate may be obtained by a plot of the union bound estimate (5.10).

Exercise 6 (invariance of coding gain). Show that the nominal coding gain $\gamma_{\rm c}(\mathcal{A})$ of (5.9), the UBE (5.10) of $P_b(E)$, and the effective coding gain $\gamma_{\rm eff}(\mathcal{A})$ are invariant to scaling, orthogonal transformations and Cartesian products. □

5.4 Orthogonal and related signal sets

Orthogonal, simplex and biorthogonal signal sets are concrete examples of large signal sets that are suitable for the power-limited regime when bandwidth is truly unconstrained. Orthogonal signal sets are the easiest to describe and analyze. Simplex signal sets are believed to be optimal for a given constellation size M when there is no constraint on dimension. Biorthogonal signal sets are slightly more bandwidth-efficient. For large M, all become essentially equivalent.

The following exercises develop the parameters of these signal sets, and show that they can achieve reliable transmission for E_b/N_0 within 3 dB from the ultimate Shannon limit.[1] The drawback of these signal sets is that the number of dimensions (bandwidth) becomes very large and the spectral efficiency ρ very small as $M \to \infty$. Also, even with the "fast" Walsh-Hadamard transform (see Chapter 1, Problem 2), decoding complexity is of the order of $M \log_2 M$, which increases exponentially with the number of bits transmitted, $\log_2 M$, and thus is actually "slow."

Exercise 7 (Orthogonal signal sets). An *orthogonal signal set* is a set $\mathcal{A} = \{\mathbf{a}_j, 1 \le j \le M\}$ of M orthogonal vectors in $\mathbb{R}^M$ with equal energy $E(\mathcal{A})$; *i.e.*, $\langle \mathbf{a}_j, \mathbf{a}_{j'} \rangle = E(\mathcal{A})\delta_{jj'}$ (Kronecker delta).

(a) Compute the nominal spectral efficiency ρ of $\mathcal{A}$ in bits per two dimensions. Compute the average energy E_b per information bit.

(b) Compute the minimum squared distance $d^2_{\min}(\mathcal{A})$. Show that every signal has $K_{\min}(\mathcal{A}) = M - 1$ nearest neighbors.

(c) Let the noise variance be $\sigma^2 = N_0/2$ per dimension. Show that the probability of error of an optimum detector is bounded by the UBE

$$\Pr(E) \le (M-1) Q^{\surd}(E(\mathcal{A})/N_0).$$

(d) Let $M \to \infty$ with E_b held constant. Using an asymptotically accurate upper bound for the $Q^{\surd}(\cdot)$ function (see Appendix), show that $\Pr(E) \to 0$ provided that $E_b/N_0 > 2 \ln 2$ (1.42 dB). How close is this to the ultimate Shannon limit on E_b/N_0? What is the nominal spectral efficiency ρ in the limit? □

[1]Actually, it can be shown that with optimum detection orthogonal signal sets can approach the ultimate Shannon limit on E_b/N_0 as $M \to \infty$; however, the union bound is too weak to prove this.

Exercise 8 (Simplex signal sets). Let $\mathcal{A}$ be an orthogonal signal set as above.

(a) Denote the mean of $\mathcal{A}$ by $\mathbf{m}(\mathcal{A})$. Show that $\mathbf{m}(\mathcal{A}) \neq \mathbf{0}$, and compute $||\mathbf{m}(\mathcal{A})||^2$.

The zero-mean set $\mathcal{A}' = \mathcal{A} - \mathbf{m}(\mathcal{A})$ (as in Exercise 2) is called a *simplex signal set*. It is universally believed to be the optimum set of M signals in AWGN in the absence of bandwidth constraints, except at ridiculously low SNRs.

(b) For $M = 2, 3, 4$, sketch $\mathcal{A}$ and $\mathcal{A}'$.

(c) Show that all signals in $\mathcal{A}'$ have the same energy $E(\mathcal{A}')$. Compute $E(\mathcal{A}')$. Compute the inner products $\langle \mathbf{a}_j, \mathbf{a}_{j'} \rangle$ for all $\mathbf{a}_j, \mathbf{a}_{j'} \in \mathcal{A}'$.

(d) [Optional]. Show that for ridiculously low SNRs, a signal set consisting of $M - 2$ zero signals and two antipodal signals $\{\pm\mathbf{a}\}$ has a lower $\Pr(E)$ than a simplex signal set. [Hint: see M. Steiner, "The strong simplex conjecture is false," IEEE Transactions on Information Theory, pp. 721-731, May 1994.] □

Exercise 9 (Biorthogonal signal sets). The set $\mathcal{A}'' = \pm\mathcal{A}$ of size $2M$ consisting of the M signals in an orthogonal signal set $\mathcal{A}$ with symbol energy $E(\mathcal{A})$ and their negatives is called a *biorthogonal signal set*.

(a) Show that the mean of $\mathcal{A}''$ is $\mathbf{m}(\mathcal{A}'') = \mathbf{0}$, and that the average energy per symbol is $E(\mathcal{A})$.

(b) How much greater is the nominal spectral efficiency ρ of $\mathcal{A}''$ than that of $\mathcal{A}$, in bits per two dimensions?

(c) Show that the probability of error of $\mathcal{A}''$ is approximately the same as that of an orthogonal signal set with the same size and average energy, for M large.

(d) Let the number of signals be a power of 2: $2M = 2^k$. Show that the nominal spectral efficiency is $\rho(\mathcal{A}'') = 4k2^{-k}$ b/2D, and that the nominal coding gain is $\gamma_c(\mathcal{A}'') = k/2$. Show that the number of nearest neighbors is $K_{\min}(\mathcal{A}'') = 2^k - 2$. □

Example 2 (Biorthogonal signal sets). Using Exercise 9, we can estimate the effective coding gain of a biorthogonal signal set using our rule of thumb (5.11), and check its accuracy against a plot of the UBE (5.10).

The $2^k = 16$ biorthogonal signal set $\mathcal{A}$ has dimension $N = 2^{k-1} = 8$, rate $k = 4$ b/sym, and nominal spectral efficiency $\rho(\mathcal{A}) = 1$ b/2D. With energy $E(\mathcal{A})$ per symbol, it has $E_b = E(\mathcal{A})/4$ and $d^2_{\min}(\mathcal{A}) = 2E(\mathcal{A})$, so its nominal coding gain is

$$\gamma_c(\mathcal{A}) = d^2_{\min}(\mathcal{A})/4E_b = 2 \text{ (3.01 dB)},$$

The number of nearest neighbors is $K_{\min}(\mathcal{A}) = 2^k - 2 = 14$, so $K_b(\mathcal{A}) = 14/4 = 3.5$, and the estimate of its effective coding gain at $P_b(E) \approx 10^{-5}$ by our rule of thumb (5.11) is thus

$$\gamma_{\text{eff}}(\mathcal{A}) \approx 3 - 2(0.2) = 2.6 \text{ dB}.$$

A more accurate plot of the UBE (5.10) may be obtained by shifting the baseline curve (Figure 1 of Chapter 4) left by 3 dB and up by half a vertical unit (since $3.5 \approx \sqrt{10}$), as shown in Figure 3. This plot shows that the rough estimate $\gamma_{\text{eff}}(\mathcal{A}) \approx 2.6$ dB is quite accurate at $P_b(E) \approx 10^{-5}$.

Similarly, the 64-biorthogonal signal set $\mathcal{A}'$ has nominal coding gain $\gamma_c(\mathcal{A}') = 3$ (4.77 dB), $K_b(\mathcal{A}') = 62/6 \approx 10$, and effective coding gain $\gamma_{\text{eff}}(\mathcal{A}') \approx 4.8$ - $3.5(0.2) = 4.1$ dB by our rule of thumb. The 256-biorthogonal signal set $\mathcal{A}''$ has nominal coding gain $\gamma_c(\mathcal{A}'') = 4$ (6.02 dB), $K_b(\mathcal{A}'') = 254/8 \approx 32$, and effective coding gain $\gamma_{\text{eff}}(\mathcal{A}'') \approx 6$ - $5(0.2) = 5.0$ dB by our rule of thumb. Figure 3 also shows plots of the UBE (5.10) for these two signal constellations, which show that our rule of thumb continues to be fairly accurate.

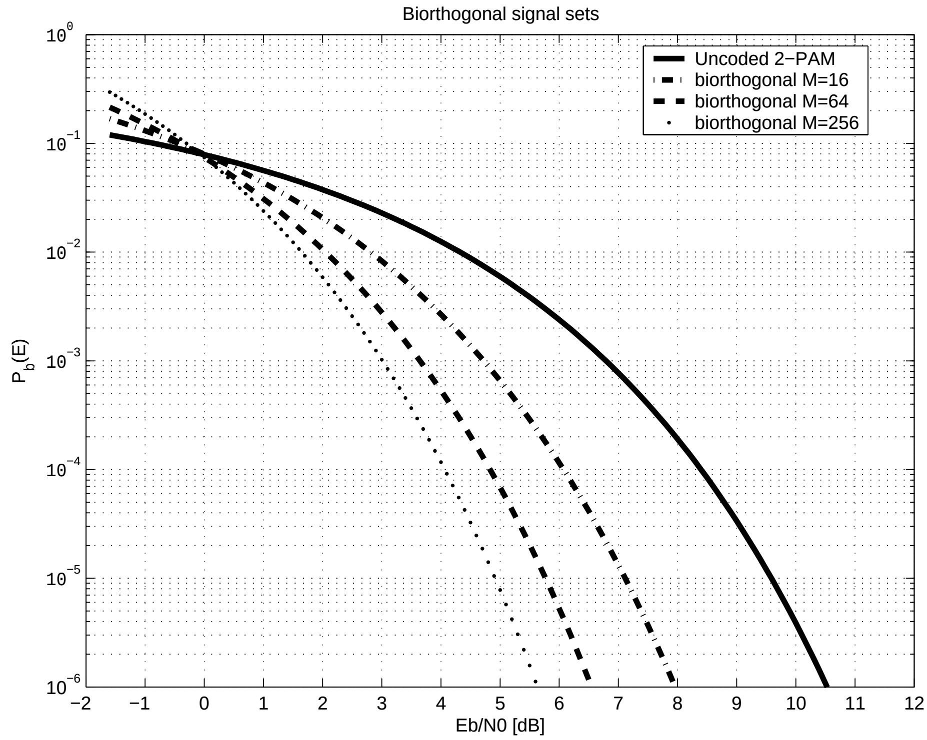


Figure 3. $P_b(E)$ *vs.* E_b/N_0 for biorthogonal signal sets with $2^k = 16, 64$ and 256.

5.5 Performance in the bandwidth-limited regime

Recall that the bandwidth-limited regime is defined as the domain in which the nominal spectral efficiency ρ is greater than 2 b/2D; *i.e.,* the domain of nonbinary signaling. In this regime we normalize all quantities "per two dimensions," and use $\mathrm{SNR}_{\mathrm{norm}}$ as our normalized measure of signal-to-noise ratio.

The baseline uncoded signal set in this regime is the M-PAM signal set $\mathcal{A} = \alpha\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$, or equivalently the $(M \times M)$-QAM constellation $\mathcal{A}^2$. Typically M is a power of 2. Such a constellation has bit rate (nominal spectral efficiency) $\rho = 2\log_2 M$ b/2D and minimum squared distance $d^2_{\min}(\mathcal{A}^2) = 4\alpha^2$. As shown in Chapter 4, its average energy per two dimensions is

$$E_s = \frac{2\alpha^2(M^2-1)}{3} = \frac{d^2_{\min}(\mathcal{A})(2^\rho - 1)}{6}. \quad (5.12)$$

The average number of nearest neighbors per two dimensions is twice that of M-PAM, namely $K_s(\mathcal{A}) = 4(M-1)/M$, which rapidly approaches $K_s(\mathcal{A}) \approx 4$ as M becomes large. By the UBE (5.7), the error probability per two dimensions is given by

$$P_s(E) \approx 4Q^{\surd}(3\mathrm{SNR}_{\mathrm{norm}}). \quad (5.13)$$

This baseline curve of $P_s(E)$ *vs.* $\mathrm{SNR}_{\mathrm{norm}}$ was plotted in Figure 2 of Chapter 4.

In the bandwidth-limited regime, the *effective coding gain* $\gamma_{\text{eff}}(\mathcal{A})$ of a signal set $\mathcal{A}$ at a given target error rate $P_s(E)$ will be defined as the difference in dB between the SNR_{norm} required to achieve the target $P_s(E)$ with $\mathcal{A}$ and the SNR_{norm} required to achieve the target $P_s(E)$ with M-PAM or $(M \times M)$-QAM (no coding).

For example, we saw from Figure 2 of Chapter 4 that the maximum possible effective coding gain at $P_s(E) \approx 10^{-5}$ is approximately 8.4 dB, which is about 3 dB less than in the power-limited regime (due solely to the fact that the bandwidth is fixed).

The effective coding gain is again estimated by the UBE, as follows. The probability of error per two dimensions is

$$P_s(E) = \frac{2\Pr(E)}{N} \approx \frac{2K_{\min}(\mathcal{A})}{N} Q^{\surd}\left(\frac{d^2_{\min}(\mathcal{A})}{2N_0}\right).$$

In the bandwidth-limited regime, we define the *nominal coding gain* $\gamma_{\text{c}}(\mathcal{A})$ as

$$\gamma_{\text{c}}(\mathcal{A}) = \frac{(2^\rho - 1)d^2_{\min}(\mathcal{A})}{6E_s}. \tag{5.14}$$

This definition is normalized so that for M-PAM or $(M \times M)$-QAM, $\gamma_{\text{c}}(\mathcal{A}) = 1$. Again, $\gamma_{\text{c}}(\mathcal{A})$ is often measured in dB. The UBE (5.10) then becomes

$$P_s(E) \approx K_s(\mathcal{A}) Q^{\surd}(3\gamma_{\text{c}}(\mathcal{A})\text{SNR}_{\text{norm}}), \tag{5.15}$$

where $K_s(\mathcal{A}) = 2K_{\min}(\mathcal{A})/N$ is the average number of nearest neighbors per two dimensions. Note that for M-PAM or $(M \times M)$-QAM, this expression reduces to (5.13).

Given $\gamma_{\text{c}}(\mathcal{A})$ and $K_s(\mathcal{A})$, we may obtain a plot of (5.15) by moving the baseline curve (Figure 2 of Chapter 4) to the left by $\gamma_{\text{c}}(\mathcal{A})$ (in dB), and up by a factor of $K_s(\mathcal{A})/4$. The rule of thumb that an increase of a factor of two in $K_s(\mathcal{A})$ over the baseline $K_s(\mathcal{A}) = 4$ costs about 0.2 dB in effective coding gain at $P_s(E) \approx 10^{-5}$ may still be used if $K_s(\mathcal{A})$ is not too large.

Exercise 6 (invariance of coding gain, cont.) Show that in the bandwidth-limited regime the nominal coding gain $\gamma_{\text{c}}(\mathcal{A})$ of (5.14), the UBE (5.15) of $P_s(E)$, and the effective coding gain $\gamma_{\text{eff}}(\mathcal{A})$ are invariant to scaling, orthogonal transformations and Cartesian products. □

5.6 Design of small signal constellations

The reader may now like to try to find the best constellations of small size M in N dimensions, using coding gain $\gamma_{\text{c}}(\mathcal{A})$ as the primary figure of merit, and $K_{\min}(\mathcal{A})$ as a secondary criterion.

Exercise 10 (small nonbinary constellations).

(a) For $M = 4$, the (2×2)-QAM signal set is known to be optimal in $N = 2$ dimensions. Show however that there exists at least one other inequivalent two-dimensional signal set $\mathcal{A}'$ with the same coding gain. Which signal set has the lower "error coefficient" $K_{\min}(\mathcal{A})$?

(b) Show that the coding gain of (a) can be improved in $N = 3$ dimensions. [Hint: consider the signal set $\mathcal{A}'' = \{(1,1,1), (1,-1,-1), (-1,1,-1), (-1,-1,1)\}$.] Sketch $\mathcal{A}''$. What is the geometric name of the polytope whose vertex set is $\mathcal{A}''$?

(c) For $M = 8$ and $N = 2$, propose at least two good signal sets, and determine which one is better. [Open research problem: Find the optimal such signal set, and prove that it is optimal.]

(d) [Open research problem.] For $M = 16$ and $N = 2$, the hexagonal signal set of Figure 1(e), Chapter 4, is thought to be near-optimal. Prove that it is optimal, or find a better one. □

5.7 Summary: Performance analysis and coding gain

The results of this chapter may be summarized very simply.

In the power-limited regime, the nominal coding gain is $\gamma_c(\mathcal{A}) = d^2_{\min}(\mathcal{A})/4E_b$. To the accuracy of the UBE, $P_b(E) \approx K_b(\mathcal{A})Q^{\surd}(2\gamma_c(\mathcal{A})E_b/N_0)$. This curve may be plotted by moving the power-limited baseline curve $P_b(E) \approx Q^{\surd}(2E_b/N_0)$ to the left by $\gamma_c(\mathcal{A})$ in dB and up by a factor of $K_b(\mathcal{A})$. An estimate of the effective coding gain at $P_b(E) \approx 10^{-5}$ is $\gamma_{\text{eff}}(\mathcal{A}) \approx \gamma_c(\mathcal{A}) - (0.2)(\log_2 K_b(\mathcal{A}))$ dB.

In the bandwidth-limited regime, the nominal coding gain is $\gamma_c(\mathcal{A}) = (2^\rho - 1)d^2_{\min}(\mathcal{A})/6E_s$. To the accuracy of the UBE, $P_s(E) \approx K_s(\mathcal{A})Q^{\surd}(3\gamma_c(\mathcal{A})\text{SNR}_{\text{norm}})$. This curve may be plotted by moving the bandwidth-limited baseline curve $P_s(E) \approx 4Q^{\surd}(3\text{SNR}_{\text{norm}})$ to the left by $\gamma_c(\mathcal{A})$ in dB and up by a factor of $K_s(\mathcal{A})/4$. An estimate of the effective coding gain at $P_s(E) \approx 10^{-5}$ is $\gamma_{\text{eff}}(\mathcal{A}) \approx \gamma_c(\mathcal{A}) - (0.2)(\log_2 K_s(\mathcal{A})/4)$ dB.

Appendix: The Q function

The Gaussian probability of error (or Q) function, defined by

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy,$$

arises frequently in error probability calculations on Gaussian channels. In this appendix we discuss some of its properties.

As we have seen, there is very often a square root in the argument of the Q function. This suggests that it might have been more useful to define a "Q-of-the-square-root-of" function $Q^{\surd}(x)$ such that $Q^{\surd}(x^2) = Q(x)$; *i.e.,*

$$Q^{\surd}(x) = Q(\sqrt{x}) = \int_{\sqrt{x}}^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

From now on we will use this $Q^{\surd}$ function instead of the Q function. For example, our baseline curves for 2-PAM and $(M \times M)$-QAM will be

$$\begin{aligned} P_b(E) &= Q^{\surd}(2E_b/N_0); \\ P_s(E) &\approx 4Q^{\surd}(3\text{SNR}_{\text{norm}}). \end{aligned}$$

The Q or $Q^{\surd}$ functions do not have a closed-form expression, but must be looked up in tables. Non-communications texts usually tabulate the complementary error function, namely

$$\text{erfc}(x) = \int_x^\infty \frac{1}{\sqrt{\pi}} e^{-y^2}\, dy.$$

Evidently $Q(x) = \text{erfc}(x/\sqrt{2})$, and $Q^{\surd}(x) = \text{erfc}(\sqrt{x/2})$.

The main property of the Q or $Q^{\surd}$ function is that it decays exponentially with x^2 according to

$$Q^{\surd}(x^2) = Q(x) \approx e^{-x^2/2}.$$

The following exercise gives several ways to prove this, including upper bounds, a lower bound, and an estimate.

Exercise A (Bounds on the $Q^{\surd}$ function).

(a) As discussed in Chapter 3, the Chernoff bound on the probability that a real random variable Z exceeds b is given by

$$\Pr\{Z \geq b\} \leq \overline{e^{s(Z-b)}}, \quad s \geq 0$$

(since $e^{s(z-b)} \geq 1$ when $z \geq b$, and $e^{s(z-b)} \geq 0$ otherwise). When optimized over $s \geq 0$, the Chernoff exponent is asymptotically correct.

Use the Chernoff bound to show that

$$Q^{\surd}(x^2) \leq e^{-x^2/2}. \tag{5.16}$$

(b) Integrate by parts to derive the upper and lower bounds

$$Q^{\surd}(x^2) < \frac{1}{\sqrt{2\pi x^2}} e^{-x^2/2}; \tag{5.17}$$

$$Q^{\surd}(x^2) > \left(1 - \frac{1}{x^2}\right) \frac{1}{\sqrt{2\pi x^2}} e^{-x^2/2}. \tag{5.18}$$

(c) Here is another way to establish these tight upper and lower bounds. By using a simple change of variables, show that

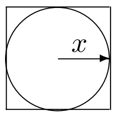
$$Q^{\surd}(x^2) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \int_0^{\infty} \exp\left(\frac{-y^2}{2} - xy\right) dy.$$

Then show that

$$1 - \frac{y^2}{2} \leq \exp\left(\frac{-y^2}{2}\right) \leq 1.$$

Putting these together, derive the bounds of part (b).

For (d)-(f), consider a circle of radius x inscribed in a square of side $2x$ as shown below.



(d) Show that the probability that a two-dimensional iid real Gaussian random variable $\mathbf{X}$ with variance $\sigma^2 = 1$ per dimension falls inside the square is equal to $(1 - 2Q^{\surd}(x^2))^2$.

(e) Show that the probability that $\mathbf{X}$ falls inside the circle is $1 - e^{-x^2/2}$. [Hint: write $p_X(\mathbf{x})$ in polar coordinates: *i.e.*, $p_{R\Theta}(r, \theta) = \frac{1}{2\pi} r e^{-r^2/2}$. You can then compute the integral $\int_0^{2\pi} d\theta \int_0^x dr \, p_{R\Theta}(r, \theta)$ in closed form.]

(f) Show that (d) and (e) imply that when x is large,

$$Q^{\surd}(x^2) \leq \frac{1}{4} e^{-x^2/2}. \qquad \square$$