Chapter 4

The gap between uncoded performance and the Shannon limit

The channel capacity theorem gives a sharp upper limit $C_{[\text{b/2D}]} = \log_2(1 + \text{SNR})$ b/2D on the rate (nominal spectral efficiency) ρ b/2D of any reliable transmission scheme. However, it does not give constructive coding schemes that can approach this limit. Finding such schemes has been the main problem of coding theory and practice for the past half century, and will be our main theme in this book.

We will distinguish sharply between the power-limited regime, where the nominal spectral efficiency ρ is small, and the bandwidth-limited regime, where ρ is large. In the power-limited regime, we will take 2-PAM as our baseline uncoded scheme, whereas in the bandwidth-limited regime, we will take M-PAM (or equivalently $(M \times M)$-QAM) as our baseline.

By evaluating the performance of these simplest possible uncoded modulation schemes and comparing baseline performance to the Shannon limit, we will establish how much "coding gain" is possible.

4.1 Discrete-time AWGN channel model

We have seen that with orthonormal PAM or QAM, the channel model reduces to an analogous real or complex discrete-time AWGN channel model

$$\mathbf{Y} = \mathbf{X} + \mathbf{N},$$

where $\mathbf{X}$ is the random input signal point sequence, and $\mathbf{N}$ is an independent iid Gaussian noise sequence with mean zero and variance $\sigma^2 = N_0/2$ per real dimension. We have also seen that there is no essential difference between the real and complex versions of this model, so from now on we will consider only the real model.

We recapitulate the connections between the parameters of this model and the corresponding continuous-time parameters. If the symbol rate is $1/T$ real symbols/s (real dimensions per second), the bit rate per two dimensions is ρ b/2D, and the average signal energy per two dimensions is E_s, then:

- The nominal bandwidth is $W = 1/2T$ Hz;

- The data rate is $R = \rho W$ b/s, and the nominal spectral efficiency is ρ (b/s)/Hz;

- The signal power (average energy per second) is $P = E_s W$;

- The signal-to-noise ratio is $\mathrm{SNR} = E_s/N_0 = P/N_0 W$;

- The channel capacity in b/s is $C_{[\mathrm{b/s}]} = W C_{[\mathrm{b/2D}]} = W \log_2(1 + \mathrm{SNR})$ b/s.

4.2   Normalized SNR and E_b/N_0

In this section we introduce two normalized measures of SNR that are suggested by the capacity bound $\rho < C_{[\mathrm{b/2D}]} = \log_2(1 + \mathrm{SNR})$ b/2D, which we will now call the *Shannon limit*.

An equivalent statement of the Shannon limit is that for a coding scheme with rate ρ b/2D, if the error probability is to be small, then the SNR must satisfy

$$\mathrm{SNR} > 2^\rho - 1.$$

This motivates the definition of the *normalized signal-to-noise ratio* $\mathrm{SNR}_{\mathrm{norm}}$ as

$$\mathrm{SNR}_{\mathrm{norm}} = \frac{\mathrm{SNR}}{2^\rho - 1}. \tag{4.1}$$

$\mathrm{SNR}_{\mathrm{norm}}$ is commonly expressed in dB. Then the Shannon limit may be expressed as

$$\mathrm{SNR}_{\mathrm{norm}} > 1 \ (0 \text{ dB}).$$

Moreover, the value of $\mathrm{SNR}_{\mathrm{norm}}$ in dB measures how far a given coding scheme is operating from the Shannon limit, in dB (the "gap to capacity").

Another commonly used normalized measure of signal-to-noise ratio is E_b/N_0, where E_b is the average signal energy per information bit and N_0 is the noise variance per two dimensions. Note that since $E_b = E_s/\rho$, where E_s is the average signal energy per two dimensions, we have

$$E_b/N_0 = E_s/\rho N_0 = \mathrm{SNR}/\rho.$$

The quantity E_b/N_0 is sometimes called the "signal-to-noise ratio per information bit," but it is not really a signal-to-noise ratio, because its numerator and denominator do not have the same units. It is probably best just to call it "E_b/N_0" (pronounced "eebee over enzero" or "ebno"). E_b/N_0 is commonly expressed in dB.

Since $\mathrm{SNR} > 2^\rho - 1$, the Shannon limit on E_b/N_0 may be expressed as

$$E_b/N_0 > \frac{2^\rho - 1}{\rho}. \tag{4.2}$$

Notice that the Shannon limit on E_b/N_0 is a monotonic function of ρ. For $\rho = 2$, it is equal to $3/2$ (1.76 dB); for $\rho = 1$, it is equal to 1 (0 dB); and as $\rho \to 0$, it approaches $\ln 2 \approx 0.69$ (-1.59 dB), which is called the *ultimate Shannon limit* on E_b/N_0.

4.3 Power-limited and bandwidth-limited channels

Ideal band-limited AWGN channels may be classified as bandwidth-limited or power-limited according to whether they permit transmission at high spectral efficiencies or not. There is no sharp dividing line, but we will take $\rho = 2$ b/2D or (b/s)/Hz as the boundary, corresponding to the highest spectral efficiency that can be achieved with binary transmission.

We note that the behavior of the Shannon limit formulas is very different in the two regimes. If SNR is small (the *power-limited regime*), then we have

$$\rho < \log_2(1 + \text{SNR}) \approx \text{SNR} \log_2 e;$$
$$\text{SNR}_{\text{norm}} \approx \frac{\text{SNR}}{\rho \ln 2} = (E_b/N_0) \log_2 e.$$

In words, in the power-limited regime, the capacity (achievable spectral efficiency) increases linearly with SNR, and as $\rho \to 0$, SNR_{norm} becomes equivalent to E_b/N_0, up to a scale factor of $\log_2 e = 1/\ln 2$. Thus as $\rho \to 0$ the Shannon limit $\text{SNR}_{\text{norm}} > 1$ translates to the ultimate Shannon limit on E_b/N_0, namely $E_b/N_0 > \ln 2$.

On the other hand, if SNR is large (the *bandwidth-limited regime*), then we have

$$\rho < \log_2(1 + \text{SNR}) \approx \log_2 \text{SNR};$$
$$\text{SNR}_{\text{norm}} \approx \frac{\text{SNR}}{2^\rho}.$$

Thus in the bandwidth-limited regime, the capacity (achievable spectral efficiency) increases logarithmically with SNR, which is dramatically different from the linear behavior in the power-limited regime. In the power-limited regime, every doubling of SNR doubles the achievable rate, whereas in the bandwidth-limited regime, every additional 3 dB in SNR yields an increase in achievable spectral efficiency of only 1 b/2D or 1 (b/s)/Hz.

Example 1. A standard voice-grade telephone channel may be crudely modeled as an ideal band-limited AWGN channel with $W \approx 3500$ Hz and SNR ≈ 37 dB. The Shannon limit on spectral efficiency and bit rate of such a channel are roughly $\rho < 37/3 \approx 12.3$ (b/s)/Hz and $R < 43{,}000$ b/s. Increasing the SNR by 3 dB would increase the achievable spectral efficiency ρ by only 1 (b/s)/Hz, or the bit rate R by only 3500 b/s. $\square$

Example 2. In contrast, there are no bandwidth restrictions on a deep-space communication channel. Therefore it makes sense to use as much bandwidth as possible, and operate deep in the power-limited region. In this case the bit rate is limited by the ultimate Shannon limit on E_b/N_0, namely $E_b/N_0 > \ln 2$ (-1.59 dB). Since $E_b/N_0 = P/RN_0$, the Shannon limit becomes $R < (P/N_0)/(\ln 2)$. Increasing P/N_0 by 3 dB will now double the achievable rate R in b/s. $\square$

We will find that the power-limited and bandwidth-limited regimes differ in almost every way. In the power-limited regime, we will be able to use binary coding and modulation, whereas in the bandwidth-limited regime we must use nonbinary ("multilevel") modulation. In the power-limited regime, it is appropriate to normalize everything "per information bit," and E_b/N_0 is a reasonable normalized measure of signal-to-noise ratio. In the bandwidth-limited regime, on the other hand, we will see that it is much better to normalize everything "per two dimensions," and SNR_{norm} will become a much more appropriate measure than E_b/N_0. Thus the first thing to do in a communications design problem is to determine which regime you are in, and then proceed accordingly.

4.4   Performance of M-PAM and $(M \times M)$-QAM

We now evaluate the performance of the simplest possible uncoded systems, namely M-PAM and $(M \times M)$-QAM. This will give us a baseline. The difference between the performance achieved by baseline systems and the Shannon limit determines the maximum possible gain that can be achieved by the most sophisticated coding systems. In effect, it defines our playing field.

4.4.1   Uncoded 2-PAM

We first consider the important special case of a binary 2-PAM constellation

$$\mathcal{A} = \{-\alpha, +\alpha\},$$

where $\alpha > 0$ is a scale factor chosen such that the average signal energy per bit, $E_b = \alpha^2$, satisfies the average signal energy constraint.

For this constellation, the bit rate (nominal spectral efficiency) is $\rho = 2$ b/2D, and the average signal energy per bit is $E_b = \alpha^2$.

The usual symbol-by-symbol detector (which is easily seen to be optimum) makes an independent decision on each received symbol y_k according to whether the sign of y_k is positive or negative. The probability of error per bit is evidently the same regardless of which of the two signal values is transmitted, and is equal to the probability that a Gaussian noise variable of variance $\sigma^2 = N_0/2$ exceeds α, namely

$$P_b(E) = Q\left(\alpha/\sigma\right),$$

where the Gaussian probability of error $Q(\cdot)$ function is defined by

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

Substituting the energy per bit $E_b = \alpha^2$ and the noise variance $\sigma^2 = N_0/2$, the probability of error per bit is

$$P_b(E) = Q\left(\sqrt{2E_b/N_0}\right). \tag{4.3}$$

This gives the performance curve of $P_b(E)$ *vs.* E_b/N_0 for uncoded 2-PAM that is shown in Figure 1, below.

4.4.2   Power-limited baseline *vs.* the Shannon limit

In the power-limited regime, we will take binary pulse amplitude modulation (2-PAM) as our baseline uncoded system, since it has $\rho = 2$. By comparing the performance of the uncoded baseline system to the Shannon limit, we will be able to determine the maximum possible gains that can be achieved by the most sophisticated coding systems.

In the power-limited regime, we will primarily use E_b/N_0 as our normalized signal-to-noise ratio, although we could equally well use SNR_{norm}. Note that when $\rho = 2$, since $E_b/N_0 = \text{SNR}/\rho$ and $\text{SNR}_{\text{norm}} = \text{SNR}/(2^\rho - 1)$, we have $2E_b/N_0 = 3\text{SNR}_{\text{norm}}$. The baseline performance curve can therefore be written in two equivalent ways:

$$P_b(E) = Q\left(\sqrt{2E_b/N_0}\right) = Q\left(\sqrt{3\text{SNR}_{\text{norm}}}\right).$$
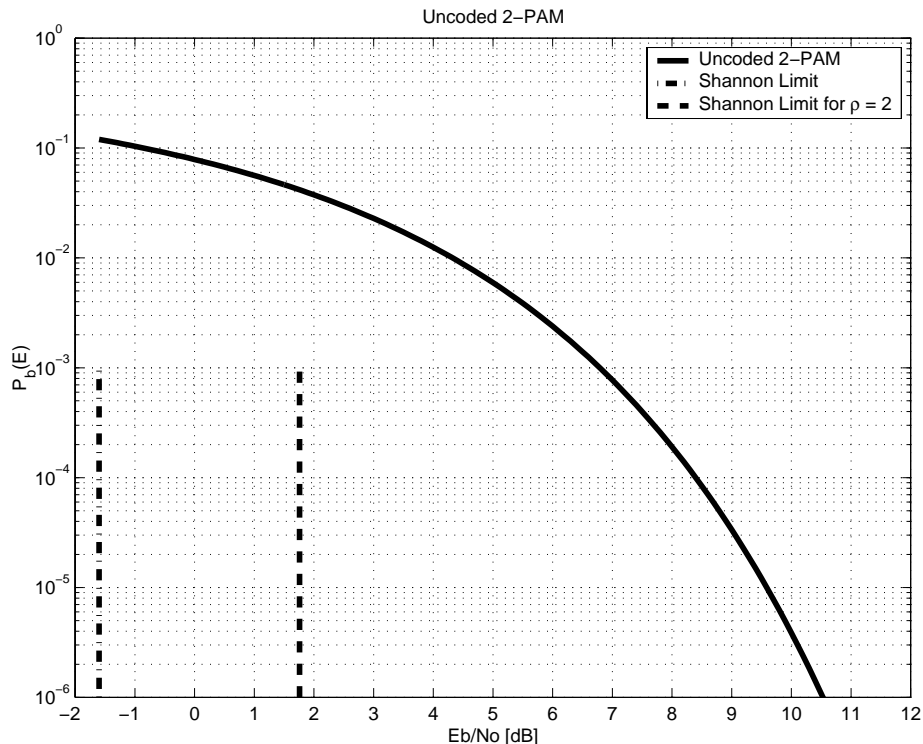
Figure 1. $P_b(E)$ *vs.* E_b/N_0 for uncoded binary PAM.

Figure 1 gives us a universal design tool. For example, if we want to achieve $P_b(E) \approx 10^{-5}$ with uncoded 2-PAM, then we know that we will need to achieve $E_b/N_0 \approx 9.6$ dB.

We may also compare the performance shown in Figure 1 to the Shannon limit. The rate of 2-PAM is $\rho = 2$ b/2D. The Shannon limit on E_b/N_0 at $\rho = 2$ b/2D is $E_b/N_0 > (2^\rho - 1)/\rho = 3/2$ (1.76 dB). Thus if our target error rate is $P_b(E) \approx 10^{-5}$, then we can achieve a coding gain of up to about 8 dB with powerful codes, at the same rate of $\rho = 2$ b/2D.

However, if there is no limit on bandwidth and therefore no lower limit on spectral efficiency, then it makes sense to let $\rho \to 0$. In this case the ultimate Shannon limit on E_b/N_0 is $E_b/N_0 > \ln 2$ (-1.59 dB). Thus if our target error rate is $P_b(E) \approx 10^{-5}$, then Shannon says that we can achieve a coding gain of over 11 dB with powerful codes, by letting the spectral efficiency approach zero.

4.4.3   Uncoded M-PAM and $(M \times M)$-QAM

We next consider the more general case of an M-PAM constellation

$$\mathcal{A} = \alpha\{\pm 1, \pm 3, \ldots, \pm(M - 1)\},$$

where $\alpha > 0$ is again a scale factor chosen to satisfy the average signal energy constraint. The bit rate (nominal spectral efficiency) is then $\rho = 2\log_2 M$ b/2D.

The average energy per M-PAM symbol is

$$E(\mathcal{A}) = \frac{\alpha^2(M^2 - 1)}{3}.$$

An elegant way of making this calculation is to consider a random variable $Z = X + U$, where X is equiprobable over $\mathcal{A}$ and U is an independent continuous uniform random variable over the interval $(-\alpha, \alpha]$. Then Z is a continuous random variable over the interval $(-M\alpha, M\alpha]$, and[1]

$$\overline{X^2} = \overline{Z^2} - \overline{U^2} = \frac{(\alpha M)^2}{3} - \frac{\alpha^2}{3}.$$

The average energy per two dimensions is then $E_s = 2E(\mathcal{A}) = 2\alpha^2(M^2 - 1)/3$.

Again, an optimal symbol-by-symbol detector makes an independent decision on each received symbol y_k. In this case the decision region associated with an input value αz_k (where z_k is an odd integer) is the interval $\alpha[z_k - 1, z_k + 1]$ (up to tie-breaking at the boundaries, which is immaterial), except for the two outermost signal values $\pm\alpha(M - 1)$, which have decision regions $\pm\alpha[M - 2, \infty)$. The probability of error for any of the $M - 2$ inner signals is thus equal to twice the probability that a Gaussian noise variable N_k of variance $\sigma^2 = N_0/2$ exceeds α, namely $2Q(\alpha/\sigma)$, whereas for the two outer signals it is just $Q(\alpha/\sigma)$. The average probability of error with equiprobable signals per M-PAM symbol is thus

$$\Pr(E) = \frac{M - 2}{M} 2Q(\alpha/\sigma) + \frac{2}{M} Q(\alpha/\sigma) = \frac{2(M - 1)}{M} Q(\alpha/\sigma).$$

For $M = 2$, this reduces to the usual expression for 2-PAM. For $M \geq 4$, the "error coefficient" $2(M - 1)/M$ quickly approaches 2, so $\Pr(E) \approx 2Q(\alpha/\sigma)$.

Since an $(M \times M)$-QAM signal set $\mathcal{A}' = \mathcal{A}^2$ is equivalent to two independent M-PAM transmissions, we can easily extend this calculation to $(M \times M)$-QAM. The bit rate (nominal spectral efficiency) is again $\rho = 2\log_2 M$ b/2D, and the average signal energy per two dimensions (per QAM symbol) is again $E_s = 2\alpha^2(M^2 - 1)/3$. The same dimension-by-dimension decision method and calculation of probability of error per dimension $\Pr(E)$ hold. The probability of error per $(M \times M)$-QAM symbol, or per two dimensions, is given by

$$P_s(E) = 1 - (1 - \Pr(E))^2 = 2\Pr(E) - (\Pr(E))^2 \approx 2\Pr(E).$$

Therefore for $(M \times M)$-QAM we obtain a probability of error per two dimensions of

$$P_s(E) \approx 2\Pr(E) \approx 4Q(\alpha/\sigma). \tag{4.4}$$

4.4.4 Bandwidth-limited baseline *vs.* the Shannon limit

In the bandwidth-limited regime, we will take $(M \times M)$-QAM with $M \geq 4$ as our baseline uncoded system, we will normalize everything per two dimensions, and we will use $\mathrm{SNR}_{\mathrm{norm}}$ as our normalized signal-to-noise ratio.

For $M \geq 4$, the probability of error per two dimensions is given by (4.4):

$$P_s(E) \approx 4Q(\alpha/\sigma).$$

---

[1]This calculation is actually somewhat fundamental, since it is based on a perfect one-dimensional sphere-packing and on the fact that the difference between the average energy of a continuous random variable and the average energy of an optimally quantized discrete version thereof is the average energy of the quantization error. As the same principle is used in the calculation of channel capacity, in the relation $S_y = S_x + S_n$, we can even say that the "1" that appears in the capacity formula is the same "1" as appears in the formula for $E(\mathcal{A})$. Therefore the cancellation of this term below is not quite as miraculous as it may at first seem.

Substituting the average energy $E_s = 2\alpha^2(M^2 - 1)/3$ per two dimensions, the noise variance $\sigma^2 = N_0/2$, and the normalized signal-to-noise ratio

$$\text{SNR}_{\text{norm}} = \frac{\text{SNR}}{2^\rho - 1} = \frac{E_s/N_0}{M^2 - 1},$$

we find that the factors of $M^2 - 1$ cancel (*cf.* footnote 1) and we obtain the performance curve

$$P_s(E) \approx 4Q\left(\sqrt{3\text{SNR}_{\text{norm}}}\right). \tag{4.5}$$

Note that this curve does not depend on M, which shows that SNR_{norm} is correctly normalized for the bandwidth-limited regime.

The bandwidth-limited baseline performance curve (4.5) of $P_s(E)$ *vs.* SNR_{norm} for uncoded $(M \times M)$-QAM is plotted in Figure 2.
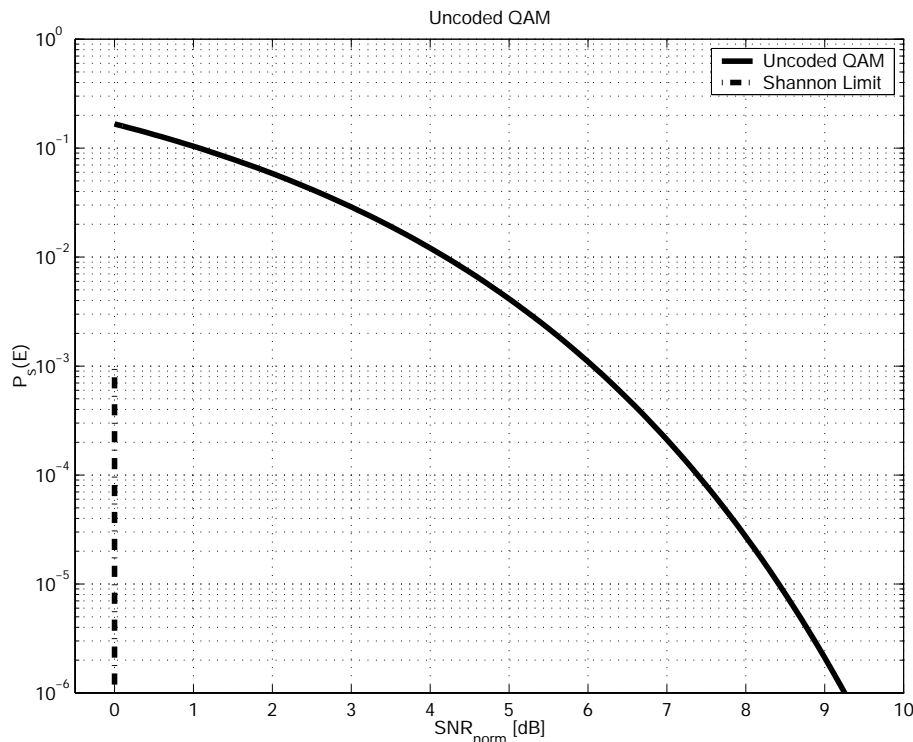

Figure 2. $P_s(E)$ *vs.* SNR_{norm} for uncoded $(M \times M)$-QAM.

Figure 2 gives us another universal design tool. For example, if we want to achieve $P_s(E) \approx 10^{-5}$ with uncoded $(M \times M)$-QAM (or M-PAM), then we know that we will need to achieve $\text{SNR}_{\text{norm}} \approx 8.4$ dB. (Notice that SNR_{norm}, unlike E_b/N_0, is already normalized for spectral efficiency.)

The Shannon limit on SNR_{norm} for any spectral efficiency is $\text{SNR}_{\text{norm}} > 1$ (0 dB). Thus if our target error rate is $P_s(E) \approx 10^{-5}$, then Shannon says that we can achieve a coding gain of up to about 8.4 dB with powerful codes, at any spectral efficiency. (This result holds approximately even for $M = 2$, as we have already seen in the previous subsection.)