Chapter 1

Introduction

The advent of cheap high-speed global communications ranks as one of the most important developments of human civilization in the second half of the twentieth century.

In 1950, an international telephone call was a remarkable event, and black-and-white television was just beginning to become widely available. By 2000, in contrast, an intercontinental phone call could often cost less than a postcard, and downloading large files instantaneously from anywhere in the world had become routine. The effects of this revolution are felt in daily life from Boston to Berlin to Bangalore.

Underlying this development has been the replacement of analog by digital communications.

Before 1948, digital communications had hardly been imagined. Indeed, Shannon's 1948 paper [7] may have been the first to use the word "bit."[1]

Even as late as 1988, the authors of an important text on digital communications [5] could write in their first paragraph:

> Why would [voice and images] be transmitted digitally? Doesn't digital transmission squander bandwidth? Doesn't it require more expensive hardware? After all, a voice-band data modem (for digital transmission over a telephone channel) costs ten times as much as a telephone and (in today's technology) *is incapable of transmitting voice signals* with quality comparable to an ordinary telephone [authors' emphasis]. This sounds like a serious indictment of digital transmission for analog signals, but for most applications, the advantages outweigh the disadvantages ...

But by their second edition in 1994 [6], they were obliged to revise this passage as follows:

> Not so long ago, digital transmission of voice and video was considered wasteful of bandwidth, and the cost ... was of concern. [More recently, there has been] a complete turnabout in thinking ... In fact, today virtually all communication is either already digital, in the process of being converted to digital, or under consideration for conversion.

[1] Shannon explains that "bit" is a contraction of "binary digit," and credits the neologism to J. W. Tukey.

The most important factor in the digital communications revolution has undoubtedly been the staggering technological progress of microelectronics and optical fiber technology. For wireline and wireless radio transmission (but not optical), another essential factor has been progress in channel coding, data compression and signal processing algorithms. For instance, data compression algorithms that can encode telephone-quality speech at 8–16 kbps and voiceband modem algorithms that can transmit 40–56 kbps over ordinary telephone lines have become commodities that require a negligible fraction of the capacity of today's personal-computer microprocessors.

This book attempts to tell the channel coding part of this story. In particular, it focusses on coding for the point-to-point additive white Gaussian noise (AWGN) channel. This choice is made in part for pedagogical reasons, but also because in fact almost all of the advances in practical channel coding have taken place in this arena. Moreover, performance on the AWGN channel is the standard benchmark for comparison of different coding schemes.

1.1 Shannon's grand challenge

The field of information theory and coding has a unique history, in that many of its ultimate limits were determined at the very beginning, in Shannon's founding paper [7].

Shannon's most celebrated result is his channel capacity theorem, which we will review in Chapter 3. This theorem states that for many common classes of channels there exists a channel capacity C such that there exist codes at any rate $R < C$ that can achieve arbitrarily reliable transmission, whereas no such codes exist for rates $R > C$. For a band-limited AWGN channel, the capacity C in bits per second (b/s) depends on only two parameters, the channel bandwidth W in Hz and the signal-to-noise ratio SNR, as follows:

$$C = W \log_2(1 + \text{SNR}) \quad \text{b/s}.$$

Shannon's theorem has posed a magnificent challenge to succeeding generations of researchers. Its proof is based on randomly chosen codes and optimal (maximum likelihood) decoding. In practice, it has proved to be remarkably difficult to find classes of constructive codes that can be decoded by feasible decoding algorithms at rates which come at all close to the Shannon limit. Indeed, for a long time this problem was regarded as practically insoluble. Each significant advance toward this goal has been awarded the highest accolades the coding community has to offer, and most such advances have been immediately incorporated into practical systems.

In the next two sections we give a brief history of these advances for two different practical channels: the deep-space channel and the telephone channel. The deep-space channel is an unlimited-bandwidth, power-limited AWGN channel, whereas the telephone channel is very much bandwidth-limited. (We realize that many of the terms used here may be unfamiliar to the reader at this point, but we hope that these surveys will give at least an impressionistic picture. After reading later chapters, the reader may wish to return to reread these sections.)

Within the past decade there have been remarkable breakthroughs, principally the invention of turbo codes [1] and the rediscovery of low-density parity check (LDPC) codes [4], which have allowed the capacity of AWGN and similar channels to be approached in a practical sense. For example, Figure 1 (from [2]) shows that an optimized rate-1/2 LDPC code on an AWGN channel can approach the relevant Shannon limit within 0.0045 decibels (dB) in theory, and within 0.04 dB with an arguably practical code of block length 10^7 bits. Practical systems using block lengths of the order of 10^4–10^5 bits now approach the Shannon limit within tenths of a dB.

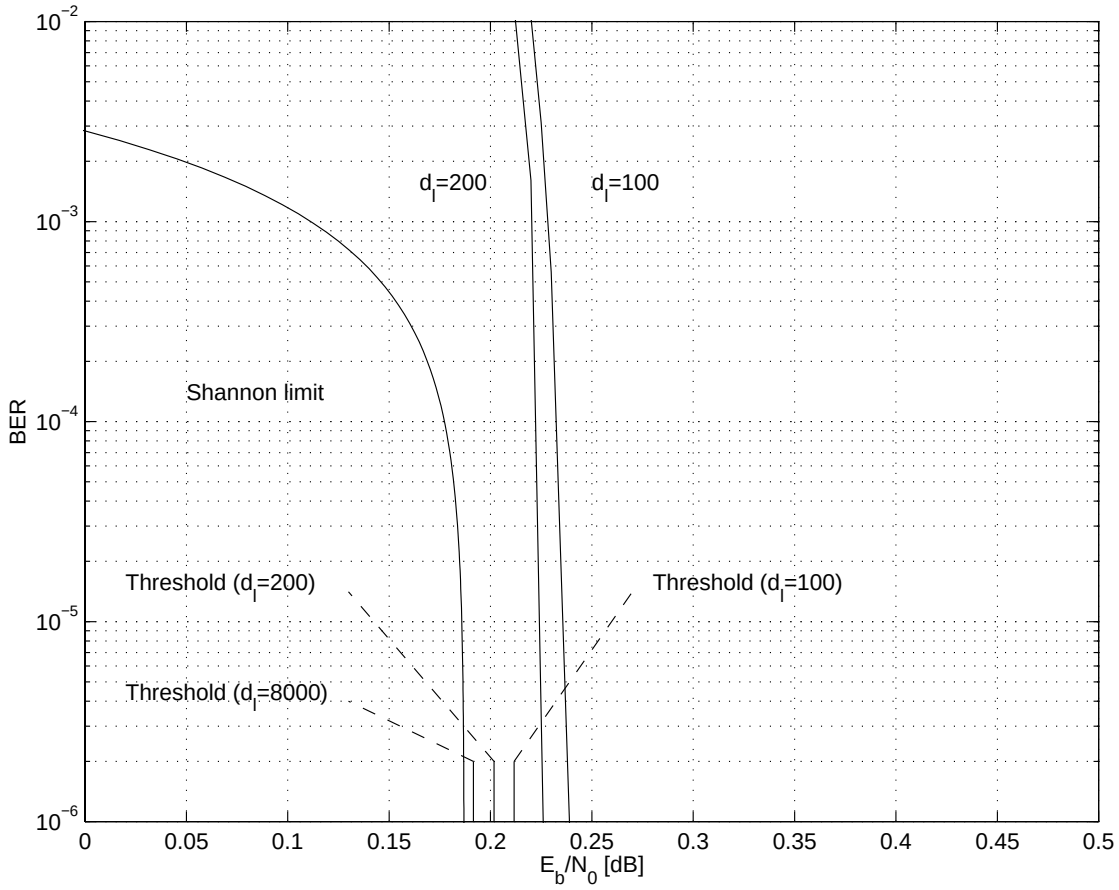

Figure 1. Bit error rate *vs.* E_b/N_0 in dB for optimized irregular rate-1/2 binary LDPC codes with maximum left degree d_l. Threshold: theoretical limit as block length $\to \infty$. Solid curves: simulation results for block length $= 10^7$. Shannon limit: binary codes, $R = 1/2$. (From [2].)

Here we will tell the story of how Shannon's challenge has been met for the AWGN channel, first for power-limited channels, where binary codes are appropriate, and then for bandwidth-limited channels, where multilevel modulation must be used. We start with the simplest schemes and work up to capacity-approaching codes, which for the most part follows the historical sequence.

1.2 Brief history of codes for deep-space missions

The deep-space communications application has been the arena in which most of the most powerful coding schemes for the power-limited AWGN channel have been first deployed, because:

- The only noise is AWGN in the receiver front end;
- Bandwidth is effectively unlimited;
- Fractions of a dB have huge scientific and economic value;
- Receiver (decoding) complexity is effectively unlimited.

For power-limited AWGN channels, we will see that there is no penalty to using binary codes with binary modulation rather than more general modulation schemes.

The first coded scheme to be designed was a simple $(32, 6, 16)$ biorthogonal code for the Mariner missions (1969), decoded by efficient maximum-likelihood decoding (the fast Hadamard transform, or "Green machine;" see Exercise 2, below). We will see that such a scheme can achieve a nominal coding gain of 3 (4.8 dB). At a target error probability per bit of $P_b(E) \approx 5 \cdot 10^{-3}$, the actual coding gain achieved was only about 2.2 dB.

The first coded scheme actually to be launched was a rate-1/2 convolutional code with constraint length $\nu = 20$ for the Pioneer 1968 mission. The receiver used 3-bit soft decisions and sequential decoding implemented on a general-purpose 16-bit minicomputer with a 1 MHz clock rate. At 512 b/s, the actual coding gain achieved at $P_b(E) \approx 5 \cdot 10^{-3}$ was about 3.3 dB.

During the 1970's, the NASA standard became a concatenated coding scheme based on a $\nu = 6$, rate-1/3 inner convolutional code and a $(255, 223, 33)$ Reed-Solomon outer code over $\mathbb{F}_{256}$. Such a system can achieve a real coding gain of about 8.3 dB at $P_b(E) \approx 10^{-6}$.

When the primary antenna failed to deploy on the Galileo mission (*circa* 1992), an elaborate concatenated coding scheme using a $\nu = 14$ rate-1/4 inner code with a Big Viterbi Decoder (BVD) and a set of variable-strength RS outer codes was reprogrammed into the spacecraft computers. This scheme was able to operate at $E_b/N_0 \approx 0.8$ dB at $P_b(E) \approx 2 \cdot 10^{-7}$, for a real coding gain of about 10.2 dB.

Turbo coding systems for deep-space communications have been developed by NASA's Jet Propulsion Laboratory (JPL) and others to get within 1 dB of the Shannon limit, and have now been standardized.

For a more comprehensive history of coding for deep-space channels, see [3].

1.3 Brief history of telephone-line modems

For several decades the telephone channel was the arena in which the most powerful coding and modulation schemes for the bandwidth-limited AWGN channel were first developed and deployed, because:

- The telephone channel is fairly well modeled as a band-limited AWGN channel;
- One dB has a significant commercial value;
- Data rates are low enough that a considerable amount of processing can be done per bit.

To approach the capacity of bandwidth-limited AWGN channels, multilevel modulation must be used. Moreover, it is important to use as much of the available bandwidth as possible.

The earliest modems developed in the 1950s and 1960s (Bell 103 and 202, and international standards V.21 and V.23) used simple binary frequency-shift keying (FSK) to achieve data rates of 300 and 1200 b/s, respectively. Implementation was entirely analog.

The first synchronous "high-speed" modem was the Bell 201 (later V.24), a 2400 b/s modem which was introduced about 1962. This modem used four-phase (4-PSK) modulation at 1200 symbols/s, so the nominal (Nyquist) bandwidth was 1200 Hz. However, because the modulation pulse had 100% rolloff, the actual bandwidth used was closer to 2400 Hz.

The first successful 4800 b/s modem was the Milgo 4400/48 (later V.27), which was introduced about 1967. This modem used eight-phase (8-PSK) modulation at 1600 symbols/s, so the nominal (Nyquist) bandwidth was 1600 Hz. "Narrow-band" filters with 50% rolloff kept the actual bandwidth used to 2400 Hz.

The first successful 9600 b/s modem was the Codex 9600C (later V.29), which was introduced in 1971. This modem used quadrature amplitude modulation (QAM) at 2400 symbols/s with an unconventional 16-point signal constellation (see Exercise 3, below) to combat combined "phase jitter" and AWGN. More importantly, it used digital adaptive linear equalization to keep the actual bandwidth needed to not much more than the Nyquist bandwidth of 2400 Hz.

All of these modems were designed for private point-to-point conditioned voice-grade lines, which use four-wire circuits (independent transmission in each direction) whose quality is higher and more consistent than that of the typical telephone connection in the two-wire (simultaneous transmission in both directions) public switched telephone network (PSTN).

The first international standard to use coding was the V.32 standard (1986) for 9600 b/s transmission over the PSTN (later raised to 14.4 kb/s in V.32*bis*). This modem used an 8-state, two-dimensional (2D) rotationally invariant Wei trellis code to achieve a coding gain of about 3.5 dB with a 32-QAM (later 128-QAM) constellation at 2400 symbols/s, again with an adaptive linear equalizer. Digital echo cancellation was also introduced to combat echoes on two-wire channels.

The "ultimate modem standard" was V.34 (1994) for transmission at up to 28.8 kb/s over the PSTN (later raised to 33.6 kb/s in V.34*bis*). This modem used a 16-state, 4D rotationally invariant Wei trellis code to achieve a coding gain of about 4.0 dB with a variable-sized QAM constellation with up to 1664 points. An optional 32-state, 4D trellis code with an additional coding gain of 0.3 dB and four times (4x) the decoding complexity and a 64-state, 4D code with a further 0.15 dB coding gain and a further 4x increase in complexity were also provided. A 16D "shell mapping" constellation shaping scheme provided an additional shaping gain of about 0.8 dB (see Exercise 4, below). A variable symbol rate of up to 3429 symbols/s was used, with symbol rate and data rate selection determined by "line probing" of individual channels. Nonlinear transmitter precoding combined with adaptive linear equalization in the receiver was used for equalization, again with echo cancellation. In short, this modem used almost every tool in the AWGN channel toolbox.

However, this standard was shortly superseded by V.90 (1998). V.90 is based on a completely different, non-AWGN model for the telephone channel: namely, it recognizes that within today's PSTN, analog signals are bandlimited, sampled and quantized to one of 256 amplitude levels at 8 kHz, transmitted digitally at 64 kb/s, and then eventually reconstructed by pulse amplitude modulation (PAM). By gaining direct access to the 64 kb/s digital data stream at a central site, and by using a well-spaced subset of the pre-existing nonlinear 256-PAM constellation, data can easily be transmitted at 40–56 kb/s (see Exercise 5, below). In V.90, such a scheme is used for downstream transmission only, with V.34 modulation upstream. In V.92 (2000) this scheme has been extended to the more difficult upstream direction.

Neither V.90 nor V.92 uses coding, nor the other sophisticated techniques of V.34. In this sense, the end of the telephone-line modem story is a bit of a fizzle. However, techniques similar to those of V.34 are now used in higher-speed wireline modems, such as digital subscriber line (DSL) modems, as well as on wireless channels such as digital cellular. In other words, the story continues in other settings.

1.4 Exercises

In this section we offer a few warm-up exercises to give the reader some preliminary feeling for data communication on the AWGN channel.

In these exercises the underlying channel model is assumed to be a discrete-time AWGN channel whose output sequence is given by $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{X}$ is a real input data sequence and $\mathbf{N}$ is a sequence of real independent, identically distributed (iid) zero-mean Gaussian noise variables. This model will be derived from a continuous-time model in Chapter 2.

We will also give the reader some practice in the use of decibels (dB). In general, a dB representation is useful wherever logarithms are useful; *i.e.,* wherever a real number is a multiplicative factor of some other number, and particularly for computing products of many factors. The dB scale is simply the logarithmic mapping

$$\text{ratio or multiplicative factor of } \alpha \leftrightarrow 10 \log_{10} \alpha \text{ dB},$$

where the scaling is chosen so that the decade interval 1–10 maps to the interval 0–10. (In other words, the value of α in dB is $\log_\beta \alpha$, where $\beta = 10^{0.1} = 1.2589....$) This scale is convenient for human memory and calculation. It is often useful to have the little log table below committed to memory, even in everyday life (see Exercise 1, below).

α	dB (round numbers)	dB (two decimal places)
1	0	0.00
1.25	1	0.97
2	3	3.01
2.5	4	3.98
e	4.3	4.34
3	4.8	4.77
π	5	4.97
4	6	6.02
5	7	6.99
8	9	9.03
10	10	10.00

Exercise 1. (Compound interest and dB) How long does it take to double your money at an interest rate of P%? The bankers' "Rule of 72" estimates that it takes about $72/P$ years; *e.g.,* at a 5% interest rate compounded annually, it takes about 14.4 years to double your money.

(a) An engineer decides to interpolate the dB table above linearly for $1 \le 1 + p \le 1.25$; *i.e.,*

$$\text{ratio or multiplicative factor of } 1 + p \leftrightarrow 4p \text{ dB}.$$

Show that this corresponds to a "Rule of 75;" *e.g.,* at a 5% interest rate compounded annually, it takes 15 years to double your money.

(b) A mathematician linearly approximates the dB table for $p \approx 0$ by noting that as $p \to 0$, $\ln(1+p) \to p$, and translates this into a "Rule of N" for some real number N. What is N? Using this rule, how many years will it take to double your money at a 5% interest rate, compounded annually? What happens if interest is compounded continuously?

(c) How many years will it actually take to double your money at a 5% interest rate, compounded annually? [Hint: $10 \log_{10} 7 = 8.45$ dB.] Whose rule best predicts the correct result?

Exercise 2. (Biorthogonal codes) A $2^m \times 2^m$ $\{\pm 1\}$-valued Hadamard matrix H_{2^m} may be constructed recursively as the m-fold tensor product of the 2×2 matrix

$$H_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix},$$

as follows:

$$H_{2^m} = \begin{bmatrix} +H_{2^{m-1}} & +H_{2^{m-1}} \\ +H_{2^{m-1}} & -H_{2^{m-1}} \end{bmatrix}.$$

(a) Show by induction that:

(i) $(H_{2^m})^T = H_{2^m}$, where ^{T} denotes the transpose; *i.e.,* H_{2^m} is symmetric;

(ii) The rows or columns of H_{2^m} form a set of mutually orthogonal vectors of length 2^m;

(iii) The first row and the first column of H_{2^m} consist of all $+1$s;

(iv) There are an equal number of $+1$s and -1s in all other rows and columns of H_{2^m};

(v) $H_{2^m} H_{2^m} = 2^m I_{2^m}$; *i.e.,* $(H_{2^m})^{-1} = 2^{-m} H_{2^m}$, where $^{-1}$ denotes the inverse.

(b) A biorthogonal signal set is a set of real equal-energy orthogonal vectors and their negatives. Show how to construct a biorthogonal signal set of size 64 as a set of $\{\pm 1\}$-valued sequences of length 32.

(c) A simplex signal set S is a set of real equal-energy vectors that are equidistant and that have zero mean $\mathbf{m}(S)$ under an equiprobable distribution. Show how to construct a simplex signal set of size 32 as a set of 32 $\{\pm 1\}$-valued sequences of length 31. [Hint: The fluctuation $O - \mathbf{m}(O)$ of a set O of orthogonal real vectors is a simplex signal set.]

(d) Let $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ be the received sequence on a discrete-time AWGN channel, where the input sequence $\mathbf{X}$ is chosen equiprobably from a biorthogonal signal set B of size 2^{m+1} constructed as in part (b). Show that the following algorithm implements a minimum-distance decoder for B (*i.e.,* given a real 2^m-vector $\mathbf{y}$, it finds the closest $\mathbf{x} \in B$ to $\mathbf{y}$):

(i) Compute $\mathbf{z} = H_{2^m}\mathbf{y}$, where $\mathbf{y}$ is regarded as a column vector;

(ii) Find the component z_j of $\mathbf{z}$ with largest magnitude $|z_j|$;

(iii) Decode to $\mathrm{sgn}(z_j)\mathbf{x}_j$, where $\mathrm{sgn}(z_j)$ is the sign of the largest-magnitude component z_j and $\mathbf{x}_j$ is the corresponding column of H_{2^m}.

(e) Show that a circuit similar to that shown below for $m = 2$ can implement the $2^m \times 2^m$ matrix multiplication $\mathbf{z} = H_{2^m}\mathbf{y}$ with a total of only $m \times 2^m$ addition and subtraction operations. (This is called the "fast Hadamard transform," or "Walsh transform," or "Green machine.")

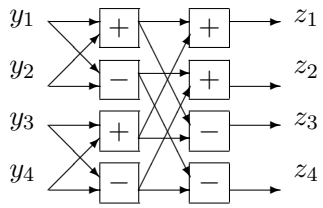

Figure 2. Fast $2^m \times 2^m$ Hadamard transform for $m = 2$.

Exercise 3. (16-QAM signal sets) Three 16-point 2-dimensional quadrature amplitude modulation (16-QAM) signal sets are shown in Figure 3, below. The first is a standard 4×4 signal set; the second is the V.29 signal set; the third is based on a hexagonal grid and is the most power-efficient 16-QAM signal set known. The first two have 90° symmetry; the last, only 180°. All have a minimum squared distance between signal points of $d^2_{\min} = 4$.

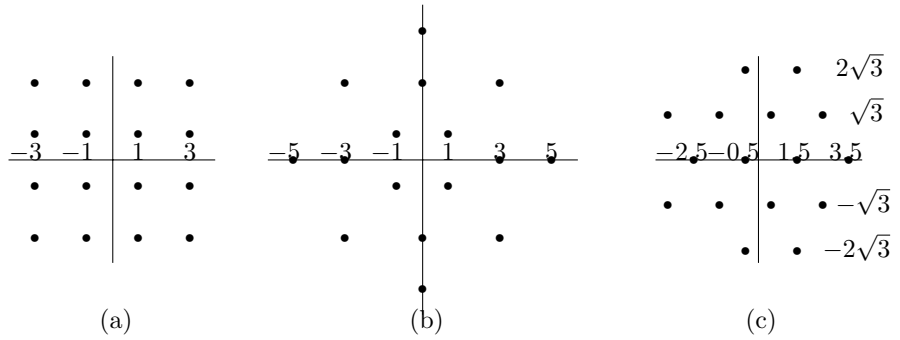

Figure 3. 16-QAM signal sets. (a) (4×4)-QAM; (b) V.29; (c) hexagonal.

(a) Compute the average energy (squared norm) of each signal set if all points are equiprobable. Compare the power efficiencies of the three signal sets in dB.

(b) Sketch the decision regions of a minimum-distance detector for each signal set.

(c) Show that with a phase rotation of $\pm 10°$ the minimum distance from any rotated signal point to any decision region boundary is substantially greatest for the V.29 signal set.

Exercise 4. (Shaping gain of spherical signal sets) In this exercise we compare the power efficiency of n-cube and n-sphere signal sets for large n.

An n-cube signal set is the set of all odd-integer sequences of length n within an n-cube of side $2M$ centered on the origin. For example, the signal set of Figure 3(a) is a 2-cube signal set with $M = 4$.

An n-sphere signal set is the set of all odd-integer sequences of length n within an n-sphere of squared radius r^2 centered on the origin. For example, the signal set of Figure 3(a) is also a 2-sphere signal set for any squared radius r^2 in the range $18 \leq r^2 < 25$. In particular, it is a 2-sphere signal set for $r^2 = 64/\pi = 20.37$, where the area πr^2 of the 2-sphere (circle) equals the area $(2M)^2 = 64$ of the 2-cube (square) of the previous paragraph.

Both n-cube and n-sphere signal sets therefore have minimum squared distance between signal points $d^2_{\min} = 4$ (if they are nontrivial), and n-cube decision regions of side 2 and thus volume 2^n associated with each signal point. The point of the following exercise is to compare their average energy using the following large-signal-set approximations:

- The number of signal points is approximately equal to the volume $V(\mathcal{R})$ of the bounding n-cube or n-sphere region $\mathcal{R}$ divided by 2^n, the volume of the decision region associated with each signal point (an n-cube of side 2).
- The average energy of the signal points under an equiprobable distribution is approximately equal to the average energy $E(\mathcal{R})$ of the bounding n-cube or n-sphere region $\mathcal{R}$ under a uniform continuous distribution.

(a) Show that if $\mathcal{R}$ is an n-cube of side $2M$ for some integer M, then under the two above approximations the approximate number of signal points is M^n and the approximate average energy is $nM^2/3$. Show that the first of these two approximations is exact.

(b) For n even, if $\mathcal{R}$ is an n-sphere of radius r, compute the approximate number of signal points and the approximate average energy of an n-sphere signal set, using the following known expressions for the volume $V_{\otimes}(n, r)$ and the average energy $E_{\otimes}(n, r)$ of an n-sphere of radius r:

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

(c) For $n = 2$, show that a large 2-sphere signal set has about 0.2 dB smaller average energy than a 2-cube signal set with the same number of signal points.

(d) For $n = 16$, show that a large 16-sphere signal set has about 1 dB smaller average energy than a 16-cube signal set with the same number of signal points. [Hint: $8! = 40320$ (46.06 dB).]

(e) Show that as $n \to \infty$ a large n-sphere signal set has a factor of $\pi e/6$ (1.53 dB) smaller average energy than an n-cube signal set with the same number of signal points. [Hint: Use Stirling's approximation, $m! \to (m/e)^m$ as $m \to \infty$.]

Exercise 5. (56 kb/s PCM modems)

This problem has to do with the design of "56 kb/s PCM modems" such as V.90 and V.92.

In the North American telephone network, voice is commonly digitized by low-pass filtering to about 3.8 KHz, sampling at 8000 samples per second, and quantizing each sample into an 8-bit byte according to the so-called "μ law." The μ law specifies 255 distinct signal levels, which are a quantized, piecewise-linear approximation to a logarithmic function, as follows:

- 1 level at 0;
- 15 positive levels evenly spaced with $d = 2$ between 2 and 30 (*i.e.,* 2, 4, 6, 8, ..., 30);
- 16 positive levels evenly spaced with $d = 4$ between 33 and 93;
- 16 positive levels evenly spaced with $d = 8$ between 99 and 219;
- 16 positive levels evenly spaced with $d = 16$ between 231 and 471;
- 16 positive levels evenly spaced with $d = 32$ between 495 and 975;
- 16 positive levels evenly spaced with $d = 64$ between 1023 and 1983;
- 16 positive levels evenly spaced with $d = 128$ between 2079 and 3999;
- 16 positive levels evenly spaced with $d = 256$ between 4191 and 8031;
- plus 127 symmetric negative levels.

The resulting 64 kb/s digitized voice sequence is transmitted through the network and ultimately reconstructed at a remote central office by pulse amplitude modulation (PAM) using a μ-law digital/analog converter and a 4 KHz low-pass filter.

For a V.90 modem, one end of the link is assumed to have a direct 64 kb/s digital connection and to be able to send any sequence of 8000 8-bit bytes per second. The corresponding levels are reconstructed at the remote central office. For the purposes of this exercise, assume that the reconstruction is exactly according to the μ-law table above, and that the reconstructed pulses are then sent through an ideal 4 KHz additive AWGN channel to the user.

(a) Determine the maximum number M of levels that can be chosen from the 255-point μ-law constellation above such that the minimum separation between levels is $d = 2$, 4, 8, 16, 64, 128, 256, 512, or 1024, respectively.

(b) These uncoded M-PAM subconstellations may be used to send up to $r = \log_2 M$ bits per symbol. What level separation can be obtained while sending 40 kb/s? 48 kb/s? 56 kb/s?

(c) How much more SNR in dB is required to transmit reliably at 48 kb/s compared to 40 kb/s? At 56 kb/s compared to 48 kb/s?

References

[1] C. Berrou, A. Glavieux and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," *Proc. 1993 Int. Conf. Commun.* (Geneva), pp. 1064–1070, May 1993.

[2] S.-Y. Chung, G. D. Forney, Jr., T. J. Richardson and R. Urbanke, "On the design of low-density parity-check codes within 0.0045 dB from the Shannon limit," *IEEE Commun. Letters*, vol. 5, pp. 58–60, Feb. 2001.

[3] D. J. Costello, Jr., J. Hagenauer, H. Imai and S. B. Wicker, "Applications of error-control coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2531–2560, Oct. 1998.

[4] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1962.

[5] E. A. Lee and D. G. Messerschmitt, *Digital Communication* (first edition). Boston: Kluwer, 1988.

[6] E. A. Lee and D. G. Messerschmitt, *Digital Communication* (second edition). Boston: Kluwer, 1994.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, 1948.

Chapter 2

Discrete-time and continuous-time AWGN channels

In this chapter we begin our technical discussion of coding for the AWGN channel. Our purpose is to show how the continuous-time AWGN channel model $Y(t) = X(t) + N(t)$ may be reduced to an equivalent discrete-time AWGN channel model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, without loss of generality or optimality. This development relies on the sampling theorem and the theorem of irrelevance. More practical methods of obtaining such a discrete-time model are orthonormal PAM or QAM modulation, which use an arbitrarily small amount of excess bandwidth. Important parameters of the continuous-time channel such as SNR, spectral efficiency and capacity carry over to discrete time, provided that the bandwidth is taken to be the nominal (Nyquist) bandwidth. Readers who are prepared to take these assertions on faith may skip this chapter.

2.1 Continuous-time AWGN channel model

The continuous-time AWGN channel is a random channel whose output is a real random process

$$Y(t) = X(t) + N(t),$$

where $X(t)$ is the input waveform, regarded as a real random process, and $N(t)$ is a real white Gaussian noise process with single-sided noise power density N_0 which is independent of $X(t)$.

Moreover, the input $X(t)$ is assumed to be both power-limited and band-limited. The average input power of the input waveform $X(t)$ is limited to some constant P. The channel band B is a positive-frequency interval with *bandwidth* W Hz. The channel is said to be baseband if $B = [0, W]$, and passband otherwise. The (positive-frequency) support of the Fourier transform of any sample function $x(t)$ of the input process $X(t)$ is limited to B.

The *signal-to-noise ratio* SNR of this channel model is then

$$\text{SNR} = \frac{P}{N_0 W},$$

where $N_0 W$ is the total noise power in the band B. The parameter N_0 is defined by convention to make this relationship true; *i.e.,* N_0 is the noise power per positive-frequency Hz. Therefore the double-sided power spectral density of $N(t)$ must be $S_{nn}(f) = N_0/2$, at least over the bands $\pm B$.

The two parameters W and SNR turn out to characterize the channel completely for digital communications purposes; the absolute scale of P and N_0 and the location of the band B do not affect the model in any essential way. In particular, as we will show in Chapter 3, the capacity of any such channel in bits per second is

$$C_{[\text{b/s}]} = W \log_2(1 + \text{SNR}) \quad \text{b/s}.$$

If a particular digital communication scheme transmits a continuous bit stream over such a channel at rate R b/s, then the *spectral efficiency* of the scheme is said to be $\rho = R/W$ (b/s)/Hz (read as "bits per second per Hertz"). The Shannon limit on spectral efficiency is therefore

$$C_{[(\text{b/s})/\text{Hz}]} = \log_2(1 + \text{SNR}) \quad (\text{b/s})/\text{Hz};$$

i.e., reliable transmission is possible when $\rho < C_{[(\text{b/s})/\text{Hz}]}$, but not when $\rho > C_{[(\text{b/s})/\text{Hz}]}$.

2.2 Signal spaces

In the next few sections we will briefly review how this continuous-time model may be reduced to an equivalent discrete-time model via the sampling theorem and the theorem of irrelevance. We assume that the reader has seen such a derivation previously, so our review will be rather succinct.

The set of all real finite-energy signals $x(t)$, denoted by $\mathcal{L}_2$, is a real vector space; *i.e.,* it is closed under addition and under multiplication by real scalars. The inner product of two signals $x(t), y(t) \in \mathcal{L}_2$ is defined by

$$\langle x(t), y(t) \rangle = \int x(t) y(t) \, dt.$$

The squared Euclidean norm (energy) of $x(t) \in \mathcal{L}_2$ is defined as $||x(t)||^2 = \langle x(t), x(t) \rangle < \infty$, and the squared Euclidean distance between $x(t), y(t) \in \mathcal{L}_2$ is $d^2(x(t), y(t)) = ||x(t) - y(t)||^2$. Two signals in $\mathcal{L}_2$ are regarded as the same ($\mathcal{L}_2$-equivalent) if their distance is 0. This allows the following strict positivity property to hold, as it must for a proper distance metric:

$$||x(t)||^2 \geq 0, \quad \text{with strict inequality unless } x(t) \text{ is } \mathcal{L}_2\text{-equivalent to 0.}$$

Every signal $x(t) \in \mathcal{L}_2$ has an $\mathcal{L}_2$ Fourier transform

$$\hat{x}(f) = \int x(t) e^{-2\pi i f t} dt,$$

such that, up to $\mathcal{L}_2$-equivalence, $x(t)$ can be recovered by the inverse Fourier transform:

$$x(t) = \int \hat{x}(f) e^{2\pi i f t} df.$$

We write $\hat{x}(f) = \mathcal{F}(x(t)), x(t) = \mathcal{F}^{-1}(\hat{x}(f))$, and $x(t) \leftrightarrow \hat{x}(f)$.

It can be shown that an $\mathcal{L}_2$ signal $x(t)$ is $\mathcal{L}_2$-equivalent to a signal which is continuous except at a discrete set of points of discontinuity ("almost everywhere"); therefore so is $\hat{x}(f)$. The values of an $\mathcal{L}_2$ signal or its transform at points of discontinuity are immaterial.

By Parseval's theorem, the Fourier transform preserves inner products:

$$\langle x(t), y(t)\rangle = \langle \hat{x}(f), \hat{y}(f)\rangle = \int \hat{x}(f)\hat{y}^*(f)\, df.$$

In particular, $||x(t)||^2 = ||\hat{x}(f)||^2$.

A signal space is any subspace $\mathcal{S} \subseteq \mathcal{L}_2$. For example, the set of $\mathcal{L}_2$ signals that are time-limited to an interval $[0, T]$ ("have support $[0, T]$") is a signal space, as is the set of $\mathcal{L}_2$ signals whose Fourier transforms are nonzero only in $\pm B$ ("have frequency support $\pm B$").

Every signal space $\mathcal{S} \subseteq \mathcal{L}_2$ has an orthogonal basis $\{\phi_k(t), k \in \mathcal{I}\}$, where $\mathcal{I}$ is some discrete index set, such that every $x(t) \in \mathcal{S}$ may be expressed as

$$x(t) = \sum_{k\in\mathcal{I}} \frac{\langle x(t), \phi_k(t)\rangle}{||\phi_k(t)||^2}\phi_k(t),$$

up to $\mathcal{L}_2$ equivalence. This is called an orthogonal expansion of $x(t)$.

Of course this expression becomes particularly simple if $\{\phi_k(t)\}$ is an orthonormal basis with $||\phi_k(t)||^2 = 1$ for all $k \in \mathcal{I}$. Then we have the orthonormal expansion

$$x(t) = \sum_{k\in\mathcal{I}} x_k\phi_k(t),$$

where $\mathbf{x} = \{x_k = \langle x(t), \phi_k(t)\rangle, k \in \mathcal{I}\}$ is the corresponding set of orthonormal coefficients. From this expression, we see that inner products are preserved in an orthonormal expansion; *i.e.,*

$$\langle x(t), y(t)\rangle = \langle \mathbf{x}, \mathbf{y}\rangle = \sum_{k\in\mathcal{I}} x_k y_k.$$

In particular, $||x(t)||^2 = ||\mathbf{x}||^2$.

2.3 The sampling theorem

The sampling theorem allows us to convert a continuous signal $x(t)$ with frequency support $[-W, W]$ (*i.e.,* a baseband signal with bandwidth W) to a discrete-time sequence of samples $\{x(kT), k \in \mathbb{Z}\}$ at a rate of $2W$ samples per second, with no loss of information.

The sampling theorem is basically an orthogonal expansion for the space $\mathcal{L}_2[0, W]$ of signals that have frequency support $[-W, W]$. If $T = 1/2W$, then the complex exponentials $\{\exp(2\pi i f kT), k \in \mathbb{Z}\}$ form an orthogonal basis for the space of Fourier transforms with support $[-W, W]$. Therefore their scaled inverse Fourier transforms $\{\phi_k(t) = \mathrm{sinc}_T(t - kT), k \in \mathbb{Z}\}$ form an orthogonal basis for $\mathcal{L}_2[0, W]$, where $\mathrm{sinc}_T(t) = (\sin \pi t/T)/(\pi t/T)$. Since $||\mathrm{sinc}_T(t)||^2 = T$, every $x(t) \in \mathcal{L}_2[0, W]$ may therefore be expressed up to $\mathcal{L}_2$ equivalence as

$$x(t) = \frac{1}{T}\sum_{k\in\mathbb{Z}} \langle x(t), \mathrm{sinc}_T(t - kT)\rangle \mathrm{sinc}_T(t - kT).$$

Moreover, evaluating this equation at $t = jT$ gives $x(jT) = \frac{1}{T}\langle x(t), \mathrm{sinc}_T(t - jT)\rangle$ for all $j \in \mathbb{Z}$ (provided that $x(t)$ is continuous at $t = jT$), since $\mathrm{sinc}_T((j - k)T) = 1$ for $k = j$ and $\mathrm{sinc}_T((j - k)T) = 0$ for $k \neq j$. Thus if $x(t) \in \mathcal{L}_2[0, W]$ is continuous, then

$$x(t) = \sum_{k\in\mathbb{Z}} x(kT)\mathrm{sinc}_T(t - kT).$$

This is called the sampling theorem.

Since inner products are preserved in an orthonormal expansion, and here the orthonormal coefficients are $x_k = \frac{1}{\sqrt{T}}\langle x(t), \mathrm{sinc}_T(t-kT)\rangle = \sqrt{T}x(kT)$, we have

$$\langle x(t), y(t)\rangle = \langle \mathbf{x}, \mathbf{y}\rangle = T\sum_{k\in\mathbb{Z}} x(kT)y(kT).$$

The following exercise shows similarly how to convert a continuous passband signal $x(t)$ with bandwidth W (*i.e.,* with frequency support $\pm[f_c - W/2, f_c + W/2]$ for some center frequency $f_c > W/2$) to a discrete-time sequence of sample pairs $\{(x_{c,k}, x_{s,k}), k \in \mathbb{Z}\}$ at a rate of W pairs per second, with no loss of information.

Exercise 2.1 (Orthogonal bases for passband signal spaces)

(a) Show that if $\{\phi_k(t)\}$ is an orthogonal set of signals in $\mathcal{L}_2[0, W]$, then $\{\phi_k(t)\cos 2\pi f_c t, \phi_k(t)\sin 2\pi f_c t\}$ is an orthogonal set of signals in $\mathcal{L}_2[f_c - W, f_c + W]$, the set of signals in $\mathcal{L}_2$ that have frequency support $\pm[f_c - W, f_c + W]$, provided that $f_c \geq W$.

[Hint: use the facts that $\mathcal{F}(\phi_k(t)\cos 2\pi f_c t) = (\hat{\phi}_k(f - f_c) + \hat{\phi}_k(f + f_c))/2$ and $\mathcal{F}(\phi_k(t)\sin 2\pi f_c t) = (\hat{\phi}_k(f - f_c) - \hat{\phi}_k(f + f_c))/2i$, plus Parseval's theorem.]

(b) Show that if the set $\{\phi_k(t)\}$ is an orthogonal basis for $\mathcal{L}_2[0, W]$, then the set $\{\phi_k(t)\cos 2\pi f_c t, \phi_k(t)\sin 2\pi f_c t\}$ is an orthogonal basis for $\mathcal{L}_2[f_c - W, f_c + W]$, provided that $f_c \geq W$.

[Hint: show that every $x(t) \in \mathcal{L}_2[f_c - W, f_c + W]$ may be written as $x(t) = x_c(t)\cos 2\pi f_c t + x_s(t)\sin 2\pi f_c t$ for some $x_c(t), x_s(t) \in \mathcal{L}_2[0, W]$.]

(c) Conclude that every $x(t) \in \mathcal{L}_2[f_c - W, f_c + W]$ may be expressed up to $\mathcal{L}_2$ equivalence as

$$x(t) = \sum_{k\in\mathbb{Z}} (x_{c,k}\cos 2\pi f_c t + x_{s,k}\sin 2\pi f_c t)\,\mathrm{sinc}_T(t - kT), \quad T = \frac{1}{2W},$$

for some sequence of pairs $\{(x_{c,k}, x_{s,k}), k \in \mathbb{Z}\}$, and give expressions for $x_{c,k}$ and $x_{s,k}$. □

2.4 White Gaussian noise

The question of how to define a white Gaussian noise (WGN) process $N(t)$ in general terms is plagued with mathematical difficulties. However, when we are given a signal space $\mathcal{S} \subseteq \mathcal{L}_2$ with an orthonormal basis as here, then defining WGN with respect to $\mathcal{S}$ is not so problematic. The following definition captures the essential properties that hold in this case:

Definition 2.1 (White Gaussian noise with respect to a signal space $\mathcal{S}$) *Let $\mathcal{S} \subseteq \mathcal{L}_2$ be a signal space with an orthonormal basis $\{\phi_k(t), k \in \mathcal{I}\}$. A Gaussian process $N(t)$ is defined as white Gaussian noise with respect to $\mathcal{S}$ with single-sided power spectral density N_0 if*

(a) The sequence $\{N_k = \langle N(t), \phi_k(t)\rangle, k \in \mathcal{I}\}$ is a sequence of iid Gaussian noise variables with mean zero and variance $N_0/2$;

(b) Define the "in-band noise" as the projection $N_{|\mathcal{S}}(t) = \sum_{k\in\mathcal{I}} N_k\phi_k(t)$ of $N(t)$ onto the signal space $\mathcal{S}$, and the "out-of-band noise" as $N_{|\mathcal{S}^\perp}(t) = N(t) - N_{|\mathcal{S}}(t)$. Then $N_{|\mathcal{S}^\perp}(t)$ is a process which is jointly Gaussian with $N_{|\mathcal{S}}(t)$, has sample functions which are orthogonal to $\mathcal{S}$, is uncorrelated with $N_{|\mathcal{S}}(t)$, and thus is statistically independent of $N_{|\mathcal{S}}(t)$.

For example, any stationary Gaussian process whose single-sided power spectral density is equal to N_0 within a band B and arbitrary elsewhere is white with respect to the signal space $\mathcal{L}_2(B)$ of signals with frequency support $\pm B$.

Exercise 2.2 (Preservation of inner products) Show that a Gaussian process $N(t)$ is white with respect to a signal space $\mathcal{S} \subseteq \mathcal{L}_2$ with psd N_0 if and only if for any signals $x(t), y(t) \in \mathcal{S}$,

$$\mathsf{E}[\langle N(t), x(t)\rangle\langle N(t), y(t)\rangle] = \frac{N_0}{2}\langle x(t), y(t)\rangle. \qquad \square$$

Here we are concerned with the detection of signals that lie in some signal space $\mathcal{S}$ in the presence of additive white Gaussian noise. In this situation the following theorem is fundamental:

Theorem 2.1 (Theorem of irrelevance) *Let $X(t)$ be a random signal process whose sample functions $x(t)$ lie in some signal space $\mathcal{S} \subseteq \mathcal{L}_2$ with an orthonormal basis $\{\phi_k(t), k \in \mathcal{I}\}$, let $N(t)$ be a Gaussian noise process which is independent of $X(t)$ and white with respect to $\mathcal{S}$, and let $Y(t) = X(t) + N(t)$. Then the set of samples*

$$Y_k = \langle Y(t), \phi_k(t)\rangle, \quad k \in \mathcal{I},$$

is a set of sufficient statistics for detection of $X(t)$ from $Y(t)$.

Sketch of proof. We may write

$$Y(t) = Y_{|\mathcal{S}}(t) + Y_{|\mathcal{S}^\perp}(t),$$

where $Y_{|\mathcal{S}}(t) = \sum_k Y_k\phi_k(t)$ and $Y_{|\mathcal{S}^\perp}(t) = Y(t) - Y_{|\mathcal{S}}(t)$. Since $Y(t) = X(t) + N(t)$ and

$$X(t) = \sum_k \langle X(t), \phi_k(t)\rangle\phi_k(t),$$

since all sample functions of $X(t)$ lie in $\mathcal{S}$, we have

$$Y(t) = \sum_k Y_k\phi_k(t) + N_{|\mathcal{S}^\perp}(t),$$

where $N_{|\mathcal{S}^\perp}(t) = N(t) - \sum_k\langle N(t), \phi_k(t)\rangle\phi_k(t)$. By Definition 2.1, $N_{|\mathcal{S}^\perp}(t)$ is independent of $N_{|\mathcal{S}}(t) = \sum_k\langle N(t), \phi_k(t)\rangle\phi_k(t)$, and by hypothesis it is independent of $X(t)$. Thus the probability distribution of $X(t)$ given $Y_{|\mathcal{S}}(t) = \sum_k Y_k\phi_k(t)$ and $Y_{|\mathcal{S}^\perp}(t) = N_{|\mathcal{S}^\perp}(t)$ depends only on $Y_{|\mathcal{S}}(t)$, so without loss of optimality in detection of $X(t)$ from $Y(t)$ we can disregard $Y_{|\mathcal{S}^\perp}(t)$; *i.e.,* $Y_{|\mathcal{S}}(t)$ is a sufficient statistic. Moreover, since $Y_{|\mathcal{S}}(t)$ is specified by the samples $\{Y_k\}$, these samples equally form a set of sufficient statistics for detection of $X(t)$ from $Y(t)$. $\square$

The sufficient statistic $Y_{|\mathcal{S}}(t)$ may alternatively be generated by filtering out the out-of-band noise $N_{|\mathcal{S}^\perp}(t)$. For example, for the signal space $\mathcal{L}_2(B)$ of signals with frequency support $\pm B$, we may obtain $Y_{|\mathcal{S}}(t)$ by passing $Y(t)$ through a brick-wall filter which passes all frequency components in B and rejects all components not in B.[1]

[1]Theorem 2.1 may be extended to any model $Y(t) = X(t) + N(t)$ in which the out-of-band noise $N_{|\mathcal{S}^\perp}(t) = N(t) - N_{|\mathcal{S}}(t)$ is independent of both the signal $X(t)$ and the in-band noise $N_{|\mathcal{S}}(t) = \sum_k N_k\phi_k(t)$; *e.g.,* to models in which the out-of-band noise contains signals from other independent users. In the Gaussian case, independence of the out-of-band noise is automatic; in more general cases, independence is an additional assumption.

Combining Definition 2.1 and Theorem 2.1, we conclude that for any AWGN channel in which the signals are confined to a sample space $\mathcal{S}$ with orthonormal basis $\{\phi_k(t), k \in \mathcal{I}\}$, we may without loss of optimality reduce the output $Y(t)$ to the set of samples

$$Y_k = \langle Y(t), \phi_k(t)\rangle = \langle X(t), \phi_k(t)\rangle + \langle N(t), \phi_k(t)\rangle = X_k + N_k, \quad k \in \mathcal{I},$$

where $\{N_k, k \in \mathcal{I}\}$ is a set of iid Gaussian variables with mean zero and variance $N_0/2$. Moreover, if $x_1(t), x_2(t) \in \mathcal{S}$ are two sample functions of $X(t)$, then this orthonormal expansion preserves their inner product:

$$\langle x_1(t), x_2(t)\rangle = \langle \mathbf{x}_1, \mathbf{x}_2\rangle,$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the orthonormal coefficient sequences of $x_1(t)$ and $x_2(t)$, respectively.

2.5 Continuous time to discrete time

We now specialize these results to our original AWGN channel model $Y(t) = X(t) + N(t)$, where the average power of $X(t)$ is limited to P and the sample functions of $X(t)$ are required to have positive frequency support in a band B of width W. For the time being we consider the baseband case in which $B = [0, W]$.

The signal space is then the set $\mathcal{S} = \mathcal{L}_2[0, W]$ of all finite-energy signals $x(t)$ whose Fourier transform has support $\pm B$. The sampling theorem shows that $\{\phi_k(t) = \frac{1}{\sqrt{T}}\text{sinc}_T(t - kT), k \in \mathbb{Z}\}$ is an orthonormal basis for this signal space, where $T = 1/2W$, and that therefore without loss of generality we may write any $x(t) \in \mathcal{S}$ as

$$x(t) = \sum_{k\in\mathbb{Z}} x_k \phi_k(t),$$

where x_k is the orthonormal coefficient $x_k = \langle x(t), \phi_k(t)\rangle$, and equality is in the sense of $\mathcal{L}_2$ equivalence.

Consequently, if $X(t)$ is a random process whose sample functions $x(t)$ are all in $\mathcal{S}$, then we can write

$$X(t) = \sum_{k\in\mathbb{Z}} X_k \phi_k(t),$$

where $X_k = \langle X(t), \phi_k(t)\rangle = \int X(t)\phi_k(t)\ dt$, a random variable that is a linear functional of $X(t)$. In this way we can identify any random band-limited process $X(t)$ of bandwidth W with a discrete-time random sequence $\mathbf{X} = \{X_k\}$ at a rate of $2W$ real variables per second. Hereafter the input will be regarded as the sequence $\mathbf{X}$ rather than $X(t)$.

Thus $X(t)$ may be regarded as a sum of amplitude-modulated orthonormal pulses $X_k\phi_k(t)$. By the Pythagorean theorem,

$$||X(t)||^2 = \sum_{k\in\mathbb{Z}} ||X_k\phi_k(t)||^2 = \sum_{k\in\mathbb{Z}} X_k^2,$$

where we use the orthonormality of the $\phi_k(t)$. Therefore the requirement that the average power (energy per second) of $X(t)$ be less than P translates to a requirement that the average energy of the sequence $\mathbf{X}$ be less than P per $2W$ symbols, or equivalently less than $P/2W$ per symbol.[2]

[2]The requirement that the sample functions of $X(t)$ must be in $\mathcal{L}_2$ translates to the requirement that the sample sequences $\mathbf{x}$ of $\mathbf{X}$ must have finite energy. This requirement can be met by requiring that only finitely many elements of $\mathbf{x}$ be nonzero. However, we do not pursue such finiteness issues.

Similarly, the random Gaussian noise process $N(t)$ may be written as

$$N(t) = \sum_{k\in\mathbb{Z}} N_k\phi_k(t) + N_{|\mathcal{S}^\perp}(t)$$

where $\mathbf{N} = \{N_k = \langle N(t), \phi_k(t)\rangle\}$ is the sequence of orthonormal coefficients of $N(t)$ in $\mathcal{S}$, and $N_{|\mathcal{S}^\perp}(t) = N(t) - \sum_k N_k\phi_k(t)$ is out-of-band noise. The theorem of irrelevance shows that $N_{|\mathcal{S}^\perp}(t)$ may be disregarded without loss of optimality, and therefore that the sequence $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ is a set of sufficient statistics for detection of $X(t)$ from $Y(t)$.

In summary, we conclude that the characteristics of the discrete-time model $\mathbf{Y} = \mathbf{X}+\mathbf{N}$ mirror those of the continuous-time model $Y(t) = X(t) + N(t)$ from which it was derived:

- The symbol interval is $T = 1/2W$; equivalently, the symbol rate is $2W$ symbols/s;
- The average signal energy per symbol is limited to $P/2W$;
- The noise sequence $\mathbf{N}$ is iid zero-mean (white) Gaussian, with variance $N_0/2$ per symbol;
- The signal-to-noise ratio is thus SNR $= (P/2W)/(N_0/2) = P/N_0W$, the same as for the continuous-time model;
- A data rate of ρ bits per two dimensions (b/2D) translates to a data rate of $R = W\rho$ b/s, or equivalently to a spectral efficiency of ρ (b/s)/Hz.

This important conclusion is the fundamental result of this chapter.

2.5.1 Passband case

Suppose now that the channel is instead a passband channel with positive-frequency support band $B = [f_c - W/2, f_c + W/2]$ for some center frequency $f_c > W/2$.

The signal space is then the set $\mathcal{S} = \mathcal{L}_2[f_c - W/2, f_c + W/2]$ of all finite-energy signals $x(t)$ whose Fourier transform has support $\pm B$.

In this case Exercise 2.1 shows that an orthogonal basis for the signal space is a set of signals of the form $\phi_{k,c}(t) = \mathrm{sinc}_T(t - kT)\cos 2\pi f_c t$ and $\phi_{k,s}(t) = \mathrm{sinc}_T(t - kT)\sin 2\pi f_c t$, where the symbol interval is now $T = 1/W$. Since the support of the Fourier transform of $\mathrm{sinc}_T(t - kT)$ is $[-W/2, W/2]$, the support of the transform of each of these signals is $\pm B$.

The derivation of a discrete-time model then goes as in the baseband case. The result is that the sequence of real pairs

$$(Y_{k,c}, Y_{k,s}) = (X_{k,c}, X_{k,s}) + (N_{k,c}, N_{k,s})$$

is a set of sufficient statistics for detection of $X(t)$ from $Y(t)$. If we compute scale factors correctly, we find that the characteristics of this discrete-time model are as follows:

- The symbol interval is $T = 1/W$, or the symbol rate is W symbols/s. In each symbol interval a pair of two real symbols is sent and received. We may therefore say that the rate is $2W = 2/T$ real dimensions per second, the same as in the baseband model.
- The average signal energy per dimension is limited to $P/2W$;

- The noise sequences $\mathbf{N}_c$ and $\mathbf{N}_s$ are independent real iid zero-mean (white) Gaussian sequences, with variance $N_0/2$ per dimension;
- The signal-to-noise ratio is again SNR $= (P/2W)/(N_0/2) = P/N_0W$;
- A data rate of ρ b/2D again translates to a spectral efficiency of ρ (b/s)/Hz.

Thus the passband discrete-time model is effectively the same as the baseband model.

In the passband case, it is often convenient to identify real pairs with single complex variables via the standard correspondence between $\mathbb{R}^2$ and $\mathbb{C}$ given by $(x, y) \leftrightarrow x + iy$, where $i = \sqrt{-1}$. This is possible because a complex iid zero-mean Gaussian sequence $\mathbf{N}$ with variance N_0 per complex dimension may be defined as $\mathbf{N} = \mathbf{N}_c + i\mathbf{N}_s$, where $\mathbf{N}_c$ and $\mathbf{N}_s$ are independent real iid zero-mean Gaussian sequences with variance $N_0/2$ per real dimension. Then we obtain a complex discrete-time model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ with the following characteristics:

- The symbol interval is $T = 1/W$, or the rate is W complex dimensions/s.
- The average signal energy per complex dimension is limited to P/W;
- The noise sequence $\mathbf{N}$ is a complex iid zero-mean Gaussian sequence, with variance N_0 per complex dimension;
- The signal-to-noise ratio is again SNR $= (P/W)/N_0 = P/N_0W$;
- A data rate of ρ bits per complex dimension translates to a spectral efficiency of ρ (b/s)/Hz.

This is still the same as before, if we regard one complex dimension as two real dimensions.

Note that even the baseband real discrete-time model may be converted to a complex discrete-time model simply by taking real variables two at a time and using the same map $\mathbb{R}^2 \to \mathbb{C}$.

The reader is cautioned that the correspondence between $\mathbb{R}^2$ and $\mathbb{C}$ given by $(x, y) \leftrightarrow x + iy$ preserves some algebraic, geometric and probabilistic properties, but not all.

Exercise 2.3 (Properties of the correspondence $\mathbb{R}^2 \leftrightarrow \mathbb{C}$) Verify the following assertions:

(a) Under the correspondence $\mathbb{R}^2 \leftrightarrow \mathbb{C}$, addition is preserved.

(b) However, multiplication is not preserved. (Indeed, the product of two elements of $\mathbb{R}^2$ is not even defined.)

(c) Inner products are not preserved. Indeed, two orthogonal elements of $\mathbb{R}^2$ can map to two collinear elements of $\mathbb{C}$.

(d) However, (squared) Euclidean norms and Euclidean distances are preserved.

(e) In general, if $\mathbf{N}_c$ and $\mathbf{N}_s$ are real jointly Gaussian sequences, then $\mathbf{N}_c + i\mathbf{N}_s$ is not a proper complex Gaussian sequence, even if $\mathbf{N}_c$ and $\mathbf{N}_s$ are independent iid sequences.

(f) However, if $\mathbf{N}_c$ and $\mathbf{N}_s$ are independent real iid zero-mean Gaussian sequences with variance $N_0/2$ per real dimension, then $\mathbf{N}_c + i\mathbf{N}_s$ is a complex zero-mean Gaussian sequence with variance N_0 per complex dimension. □

2.6 Orthonormal PAM and QAM modulation

More generally, suppose that $X(t) = \sum_k X_k \phi_k(t)$, where $\mathbf{X} = \{X_k\}$ is a random sequence and $\{\phi_k(t) = p(t - kT)\}$ is an orthonormal sequence of time shifts $p(t - kT)$ of a basic modulation pulse $p(t) \in \mathcal{L}_2$ by integer multiples of a symbol interval T. This is called *orthonormal pulse-amplitude modulation (PAM).*

The signal space $\mathcal{S}$ is then the subspace of $\mathcal{L}_2$ spanned by the orthonormal sequence $\{p(t-kT)\}$; *i.e.,* $\mathcal{S}$ consists of all signals in $\mathcal{L}_2$ that can be written as linear combinations $\sum_k x_k p(t - kT)$.

Again, the average power of $X(t) = \sum_k X_k p(t-kT)$ will be limited to P if the average energy of the sequence $\mathbf{X}$ is limited to PT per symbol, since the symbol rate is $1/T$ symbol/s.

The theorem of irrelevance again shows that the set of inner products

$$Y_k = \langle Y(t), \phi_k(t) \rangle = \langle X(t), \phi_k(t) \rangle + \langle N(t), \phi_k(t) \rangle = X_k + N_k$$

is a set of sufficient statistics for detection of $X(t)$ from $Y(t)$. These inner products may be obtained by filtering $Y(t)$ with a *matched filter* with impulse response $p(-t)$ and sampling at integer multiples of T as shown in Figure 1 to obtain

$$Z(kT) = \int Y(\tau) p(\tau - kT)\, d\tau = Y_k,$$

Thus again we obtain a discrete-time model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where by the orthonormality of the $p(t - kT)$ the noise sequence $\mathbf{N}$ is iid zero-mean Gaussian with variance $N_0/2$ per symbol.

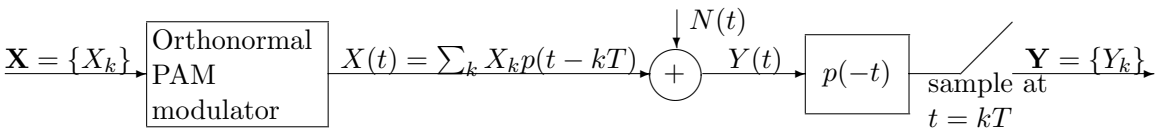

Figure 1. Orthonormal PAM system.

The conditions that ensure that the time shifts $\{p(t-kT)\}$ are orthonormal are determined by Nyquist theory as follows. Define the composite response in Figure 1 as $g(t) = p(t) * p(-t)$, with Fourier transform $\hat{g}(f) = |\hat{p}(f)|^2$. (The composite response $g(t)$ is also called the autocorrelation function of $p(t)$, and $\hat{g}(f)$ is also called its power spectrum.) Then:

Theorem 2.2 (Orthonormality conditions) *For a signal $p(t) \in \mathcal{L}_2$ and a time interval T, the following are equivalent:*

(a) The time shifts $\{p(t - kT), k \in \mathbb{Z}\}$ are orthonormal;

(b) The composite response $g(t) = p(t) * p(-t)$ satisfies $g(0) = 1$ and $g(kT) = 0$ for $k \neq 0$;

(c) The Fourier transform $\hat{g}(f) = |\hat{p}(f)|^2$ satisfies the Nyquist criterion for zero intersymbol interference, namely

$$\frac{1}{T} \sum_{m \in \mathbb{Z}} \hat{g}(f - m/T) = 1 \quad \text{for all } f.$$

Sketch of proof. The fact that (a) $\Leftrightarrow$ (b) follows from $\langle p(t - kT), p(t - k'T) \rangle = g((k - k')T)$. The fact that (b) $\Leftrightarrow$ (c) follows from the aliasing theorem, which says that the discrete-time Fourier transform of the sample sequence $\{g(kT)\}$ is the aliased response $\frac{1}{T} \sum_m \hat{g}(f - m/T)$. □

It is clear from the Nyquist criterion (c) that if $p(t)$ is a baseband signal of bandwidth W, then

(i) The bandwidth W cannot be less than $1/2T$;

(ii) If $W = 1/2T$, then $\hat{g}(f) = T, -W \leq f \leq W$, else $\hat{g}(f) = 0$; *i.e.,* $g(t) = \mathrm{sinc}_T(t)$;

(iii) If $1/2T < W \leq 1/T$, then any real non-negative power spectrum $\hat{g}(f)$ that satisfies $\hat{g}(1/2T + f) + \hat{g}(1/2T - f) = T$ for $0 \leq f \leq 1/2T$ will satisfy (c).

For this reason $W = 1/2T$ is called the *nominal* or *Nyquist bandwidth* of a PAM system with symbol interval T. No orthonormal PAM system can have bandwidth less than the Nyquist bandwidth, and only a system in which the modulation pulse has autocorrelation function $g(t) = p(t)*p(-t) = \mathrm{sinc}_T(t)$ can have exactly the Nyquist bandwidth. However, by (iii), which is called the *Nyquist band-edge symmetry condition*, the Fourier transform $|\hat{p}(f)|^2$ may be designed to roll off arbitrarily rapidly for $f > W$, while being continuous and having a continuous derivative.

Figure 2 illustrates a raised-cosine frequency response that satisfies the Nyquist band-edge symmetry condition while being continuous and having a continuous derivative. Nowadays it is no great feat to implement such responses with excess bandwidths of 5–10% or less.

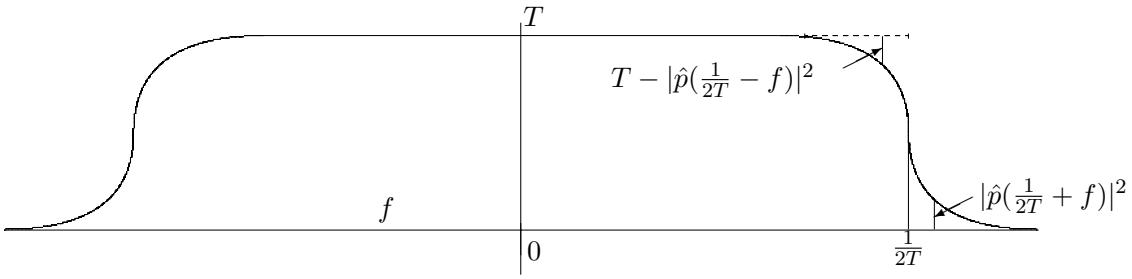

Figure 2. Raised-cosine spectrum $\hat{g}(f) = |\hat{p}(f)|^2$ with Nyquist band-edge symmetry.

We conclude that an orthonormal PAM system may use arbitrarily small excess bandwidth beyond the Nyquist bandwidth $W = 1/2T$, or alternatively that the power in the out-of-band frequency components may be made to be arbitrarily small, without violating the practical constraint that the Fourier transform $\hat{p}(f)$ of the modulation pulse $p(t)$ should be continuous and have a continuous derivative.

In summary, if we let W denote the Nyquist bandwidth $1/2T$ rather than the actual bandwidth, then we again obtain a discrete-time channel model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ for any orthonormal PAM system, not just a system with the modulation pulse $p(t) = \frac{1}{\sqrt{T}}\mathrm{sinc}_T(t)$, in which:

- The symbol interval is $T = 1/2W$; equivalently, the symbol rate is $2W$ symbols/s;
- The average signal energy per symbol is limited to $P/2W$;
- The noise sequence $\mathbf{N}$ is iid zero-mean (white) Gaussian, with variance $N_0/2$ per symbol;
- The signal-to-noise ratio is $\mathrm{SNR} = (P/2W)/(N_0/2) = P/N_0W$;
- A data rate of ρ bits per two dimensions (b/2D) translates to a data rate of $R = \rho/W$ b/s, or equivalently to a spectral efficiency of ρ (b/s)/Hz.

Exercise 2.4 (Orthonormal QAM modulation)

Figure 3 illustrates an orthonormal quadrature amplitude modulation (QAM) system with symbol interval T in which the input and output variables X_k and Y_k are complex, $p(t)$ is a complex finite-energy modulation pulse whose time shifts $\{p(t-kT)\}$ are orthonormal (the inner product of two complex signals is $\langle x(t), y(t)\rangle = \int x(t)y^*(t)\ dt$), the matched filter response is $p^*(-t)$, and $f_c > 1/2T$ is a carrier frequency. The box marked $2\Re\{\cdot\}$ takes twice the real part of its input— *i.e.,* it maps a complex signal $f(t)$ to $f(t) + f^*(t)$— and the Hilbert filter is a complex filter whose frequency response is 1 for $f > 0$ and 0 for $f < 0$.

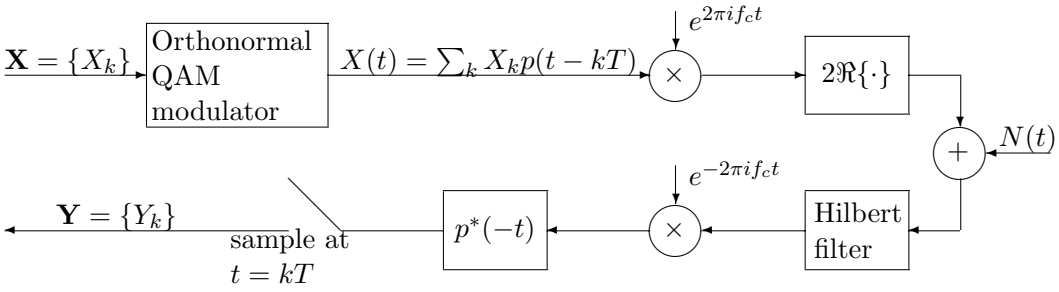

Figure 3. Orthonormal QAM system.

(a) Assume that $\hat{p}(f) = 0$ for $|f| \geq f_c$. Show that the Hilbert filter is superfluous.

(b) Show that Theorem 2.2 holds for a complex response $p(t)$ if we define the composite response (autocorrelation function) as $g(t) = p(t) * p^*(-t)$. Conclude that the bandwidth of an orthonormal QAM system is lowerbounded by its Nyquist bandwidth $W = 1/T$.

(c) Show that $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{N}$ is an iid complex Gaussian noise sequence. Show that the signal-to-noise ratio in this complex discrete-time model is equal to the channel signal-to-noise ratio $\mathrm{SNR} = P/N_0W$, if we define $W = 1/T$. [Hint: use Exercise 2.1.]

(d) Show that a mismatch in the receive filter— *i.e.,* an impulse response $h(t)$ other than $p^*(-t)$— results in linear intersymbol interference— *i.e.,* in the absence of noise $Y_k = \sum_j X_j h_{k-j}$ for some discrete-time response $\{h_k\}$ other than the ideal response δ_{k0} (Kronecker delta).

(e) Show that a phase error of θ in the receive carrier— *i.e.,* demodulation by $e^{-2\pi i f_c t + i\theta}$ rather than by $e^{-2\pi i f_c t}$— results (in the absence of noise) in a phase rotation by θ of all outputs Y_k.

(f) Show that a sample timing error of δ— *i.e.,* sampling at times $t = kT + \delta$— results in linear intersymbol interference. □

2.7 Summary

To summarize, the key parameters of a band-limited continuous-time AWGN channel are its bandwidth W in Hz and its signal-to-noise ratio SNR, regardless of other details like where the bandwidth is located (in particular whether it is at baseband or passband), the scaling of the signal, etc. The key parameters of a discrete-time AWGN channel are its symbol rate W in two-dimensional real or one-dimensional complex symbols per second and its SNR, regardless of other details like whether it is real or complex, the scaling of the symbols, etc. With orthonormal PAM or QAM, these key parameters are preserved, regardless of whether PAM or QAM is used, the precise modulation pulse, etc. The (nominal) spectral efficiency ρ (in (b/s)/Hz or in b/2D) is also preserved, and (as we will see in the next chapter) so is the channel capacity (in b/s).

Chapter 3

Capacity of AWGN channels

In this chapter we prove that the capacity of an AWGN channel with bandwidth W and signal-to-noise ratio SNR is $W \log_2(1+\text{SNR})$ bits per second (b/s). The proof that reliable transmission is possible at any rate less than capacity is based on Shannon's random code ensemble, typical-set decoding, the Chernoff-bound law of large numbers, and a fundamental result of large-deviation theory. We also sketch a geometric proof of the converse. Readers who are prepared to accept the channel capacity formula without proof may skip this chapter.

3.1 Outline of proof of the capacity theorem

The first step in proving the channel capacity theorem or its converse is to use the results of Chapter 2 to replace a continuous-time AWGN channel model $Y(t) = X(t) + N(t)$ with bandwidth W and signal-to-noise ratio SNR by an equivalent discrete-time channel model $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ with a symbol rate of $2W$ real symbol/s and the same SNR, without loss of generality or optimality.

We then wish to prove that arbitrarily reliable transmission can be achieved on the discrete-time channel at any rate (nominal spectral efficiency)

$$\rho < C_{[\text{b/2D}]} = \log_2(1 + \text{SNR}) \quad \text{b/2D}.$$

This will prove that reliable transmission can be achieved on the continuous-time channel at any data rate

$$R < C_{[\text{b/s}]} = WC_{[\text{b/2D}]} = W \log_2(1 + \text{SNR}) \quad \text{b/s}.$$

We will prove this result by use of Shannon's random code ensemble and a suboptimal decoding technique called typical-set decoding.

Shannon's random code ensemble may be defined as follows. Let $S_x = P/2W$ be the allowable average signal energy per symbol (dimension), let ρ be the data rate in b/2D, and let N be the code block length in symbols. A block code $\mathcal{C}$ of length N, rate ρ, and average energy S_x per dimension is then a set of $M = 2^{\rho N/2}$ real sequences (codewords) $\mathbf{c}$ of length N such that the expected value of $||\mathbf{c}||^2$ under an equiprobable distribution over $\mathcal{C}$ is NS_x.

For example, the three 16-QAM signal sets shown in Figure 3 of Chapter 1 may be regarded as three block codes of length 2 and rate 4 b/2D with average energies per dimension of $S_x = 5, 6.75$ and 4.375, respectively.

In Shannon's random code ensemble, every symbol c_k of every codeword $\mathbf{c} \in \mathcal{C}$ is chosen independently at random from a Gaussian ensemble with mean 0 and variance S_x. Thus the average energy per dimension over the ensemble of codes is S_x, and by the law of large numbers the average energy per dimension of any particular code in the ensemble is highly likely to be close to S_x.

We consider the probability of error under the following scenario. A code $\mathcal{C}$ is selected randomly from the ensemble as above, and then a particular codeword $\mathbf{c}_0$ is selected for transmission. The channel adds a noise sequence $\mathbf{n}$ from a Gaussian ensemble with mean 0 and variance $S_n = N_0/2$ per symbol. At the receiver, given $\mathbf{y} = \mathbf{c}_0 + \mathbf{n}$ and the code $\mathcal{C}$, a typical-set decoder implements the following decision rule (where ε is some small positive number):

- If there is one and only one codeword $\mathbf{c} \in \mathcal{C}$ within squared distance $N(S_n \pm \varepsilon)$ of the received sequence $\mathbf{y}$, then decide on $\mathbf{c}$;
- Otherwise, give up.

A decision error can occur only if one of the following two events occurs:

- The squared distance $||\mathbf{y} - \mathbf{c}_0||^2$ between $\mathbf{y}$ and the transmitted codeword $\mathbf{c}_0$ is not in the range $N(S_n \pm \varepsilon)$;
- The squared distance $||\mathbf{y} - \mathbf{c}_i||^2$ between $\mathbf{y}$ and some other codeword $\mathbf{c}_i \neq \mathbf{c}_0$ is in the range $N(S_n \pm \varepsilon)$.

Since $\mathbf{y} - \mathbf{c}_0 = \mathbf{n}$, the probability of the first of these events is the probability that $||\mathbf{n}||^2$ is not in the range $N(S_n - \varepsilon) \leq ||\mathbf{n}||^2 \leq N(S_n + \varepsilon)$. Since $\mathbf{N} = \{N_k\}$ is an iid zero-mean Gaussian sequence with variance S_n per symbol and $||\mathbf{N}||^2 = \sum_k N_k^2$, this probability goes to zero as $N \to \infty$ for any $\varepsilon > 0$ by the weak law of large numbers. In fact, by the Chernoff bound of the next section, this probability goes to zero exponentially with N.

For any particular other codeword $\mathbf{c}_i \in \mathcal{C}$, the probability of the second event is the probability that a code sequence drawn according to an iid Gaussian pdf $p_X(\mathbf{x})$ with symbol variance S_x and a received sequence drawn *independently* according to an iid Gaussian pdf $p_Y(\mathbf{y})$ with symbol variance $S_y = S_x + S_n$ are "typical" of the joint pdf $p_{XY}(\mathbf{x}, \mathbf{y}) = p_X(\mathbf{x})p_N(\mathbf{y} - \mathbf{x})$, where here we define "typical" by the distance $||\mathbf{x} - \mathbf{y}||^2$ being in the range $N(S_n \pm \varepsilon)$. According to a fundamental result of large-deviation theory, this probability goes to zero as e^{-NE}, where, up to terms of the order of ε, the exponent E is given by the relative entropy (Kullback-Leibler divergence)

$$D(p_{XY}||p_X p_Y) = \int dx\ dy\ p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}.$$

If the logarithm is binary, then this is the mutual information $I(X;Y)$ between the random variables X and Y in bits per dimension (b/D).

In the Gaussian case considered here, the mutual information is easily evaluated as

$$I(X;Y) = \mathsf{E}_{XY}\left[-\frac{1}{2}\log_2 2\pi S_n - \frac{(y-x)^2 \log_2 e}{2S_n} + \frac{1}{2}\log_2 2\pi S_y + \frac{y^2 \log_2 e}{2S_y}\right] = \frac{1}{2}\log_2 \frac{S_y}{S_n} \quad \text{b/D}.$$

Since $S_y = S_x + S_n$ and $\text{SNR} = S_x/S_n$, this expression is equal to the claimed capacity in b/D.

Thus we can say that the probability that any incorrect codeword $\mathbf{c}_i \in \mathcal{C}$ is "typical" with respect to $\mathbf{y}$ goes to zero as $2^{-N(I(X;Y)-\delta(\varepsilon))}$, where $\delta(\varepsilon)$ goes to zero as $\varepsilon \to 0$. By the union bound, the probability that any of the $M-1 < 2^{\rho N/2}$ incorrect codewords is "typical" with respect to $\mathbf{y}$ is upperbounded by

$$\Pr\{\text{any incorrect codeword "typical"}\} < 2^{\rho N/2} 2^{-N(I(X;Y)-\delta(\varepsilon))},$$

which goes to zero exponentially with N provided that $\rho < 2I(X;Y)$ b/2D and ε is small enough.

In summary, the probabilities of both types of error go to zero exponentially with N provided that

$$\rho < 2I(X;Y) = \log_2(1+\text{SNR}) = C_{[\text{b/2D}]} \quad \text{b/2D}$$

and ε is small enough. This proves that an arbitrarily small probability of error can be achieved using Shannon's random code ensemble and typical-set decoding.

To show that there is a particular code of rate $\rho < C_{[\text{b/2D}]}$ that achieves an arbitrarily small error probability, we need merely observe that the probability of error over the random code ensemble is the average probability of error over all codes in the ensemble, so there must be at least one code in the ensemble that achieves this performance. More pointedly, if the average error probability is $\Pr(E)$, then no more than a fraction of $1/K$ of the codes can achieve error probability worse than $K\Pr(E)$ for any constant $K > 0$; *e.g.,* at least 99% of the codes achieve performance no worse than $100\Pr(E)$. So we can conclude that almost all codes in the random code ensemble achieve very small error probabilities. Briefly, "almost all codes are good" (when decoded by typical-set or maximum-likelihood decoding).

3.2 Laws of large numbers

The channel capacity theorem is essentially an application of various laws of large numbers.

3.2.1 The Chernoff bound

The weak law of large numbers states that the probability that the sample average of a sequence of N iid random variables differs from the mean by more than $\varepsilon > 0$ goes to zero as $N \to \infty$, no matter how small ε is. The Chernoff bound shows that this probability goes to zero exponentially with N, for arbitrarily small ε.

Theorem 3.1 (Chernoff bound) *Let S_N be the sum of N iid real random variables X_k, each with the same probability distribution $p_X(x)$ and mean $\overline{X} = \mathsf{E}_X[X]$. For $\tau > \overline{X}$, the probability that $S_N \geq N\tau$ is upperbounded by*

$$\Pr\{S_N \geq N\tau\} \leq e^{-NE_c(\tau)},$$

where the Chernoff exponent $E_c(\tau)$ is given by

$$E_c(\tau) = \max_{s \geq 0} s\tau - \mu(s),$$

where $\mu(s)$ denotes the semi-invariant moment-generating function, $\mu(s) = \log \mathsf{E}_X[e^{sX}]$.

Proof. The indicator function $\Phi(S_N \geq N\tau)$ of the event $\{S_N \geq N\tau\}$ is bounded by

$$\Phi(S_N \geq N\tau) \leq e^{s(S_N - N\tau)}$$

for any $s \geq 0$. Therefore

$$\Pr\{S_N \geq N\tau\} = \overline{\Phi(S_N \geq N\tau)} \leq \overline{e^{s(S_N - N\tau)}}, \quad s \geq 0,$$

where the overbar denotes expectation. Using the facts that $S_N = \sum_k X_k$ and that the X_k are independent, we have

$$\overline{e^{s(S_N - N\tau)}} = \prod_k \overline{e^{s(X_k - \tau)}} = e^{-N(s\tau - \mu(s))},$$

where $\mu(s) = \log \overline{e^{sX}}$. Optimizing the exponent over $s \geq 0$, we obtain the Chernoff exponent

$$E_c(\tau) = \max_{s \geq 0} s\tau - \mu(s). \qquad \square$$

We next show that the Chernoff exponent is positive:

Theorem 3.2 (Positivity of Chernoff exponent) *The Chernoff exponent $E_c(\tau)$ is positive when $\tau > \overline{X}$, provided that the random variable X is nondeterministic.*

Proof. Define $X(s)$ as a random variable with the same alphabet as X, but with the tilted probability density function $q(x, s) = p(x)e^{sx - \mu(s)}$. This is a valid pdf because $q(x, s) \geq 0$ and

$$\int q(x, s)\ dx = e^{-\mu(s)} \int e^{sx} p(x)\ dx = e^{-\mu(s)} e^{\mu(s)} = 1.$$

Evidently $\mu(0) = \log \mathsf{E}_X[1] = 0$, so $q(x, 0) = p(x)$ and $X(0) = X$.

Define the moment-generating (partition) function

$$Z(s) = e^{\mu(s)} = \mathsf{E}_X[e^{sX}] = \int e^{sx} p(x)\ dx.$$

Now it is easy to see that

$$Z'(s) = \int x e^{sx} p(x)\ dx = e^{\mu(s)} \int x e^{sx} q(x, s)\ dx = Z(s)\overline{X(s)}.$$

Similarly,

$$Z''(s) = \int x^2 e^{sx} p(x)\ dx = Z(s)\overline{X^2(s)}.$$

Consequently, from $\mu(s) = \log Z(s)$, we have

$$\begin{aligned} \mu'(s) &= \frac{Z'(s)}{Z(s)} = \overline{X(s)}; \\ \mu''(s) &= \frac{Z''(s)}{Z(s)} - \left(\frac{Z'(s)}{Z(s)}\right)^2 = \overline{X^2(s)} - \overline{X(s)}^2. \end{aligned}$$

Thus the second derivative $\mu''(s)$ is the variance of $X(s)$, which must be strictly positive unless $X(s)$ and thus X is deterministic.

We conclude that if X is a nondeterministic random variable with mean $\overline{X}$, then $\mu(s)$ is a strictly convex function of s that equals 0 at $s = 0$ and whose derivative at $s = 0$ is $\overline{X}$. It follows that the function $s\tau - \mu(s)$ is a strictly concave function of s that equals 0 at $s = 0$ and whose derivative at $s = 0$ is $\tau - \overline{X}$. Thus if $\tau > \overline{X}$, then the function $s\tau - \mu(s)$ has a unique maximum which is strictly positive. □

Exercise 1. Show that if X is a deterministic random variable— *i.e.,* the probability that X equals its mean $\overline{X}$ is 1— and $\tau > \overline{X}$, then $\Pr\{S_N \geq N\tau\} = 0$. □

The proof of this theorem shows that the general form of the function $f(s) = s\tau - \mu(s)$ when X is nondeterministic is as shown in Figure 1. The second derivative $f''(s)$ is negative everywhere, so the function $f(s)$ is strictly concave and has a unique maximum $E_c(\tau)$. The slope $f'(s) = \tau - \overline{X(s)}$ therefore decreases continually from its value $f'(0) = \tau - \overline{X} > 0$ at $s = 0$. The slope becomes equal to 0 at the value of s for which $\tau = \overline{X(s)}$; in other words, to find the maximum of $f(s)$, keep increasing the "tilt" until the tilted mean $\overline{X(s)}$ is equal to τ. If we denote this value of s by $s^*(\tau)$, then we obtain the following parametric equations for the Chernoff exponent:

$$E_c(\tau) = s^*(\tau)\tau - \mu(s^*(\tau)); \quad \tau = \overline{X(s^*(\tau))}.$$

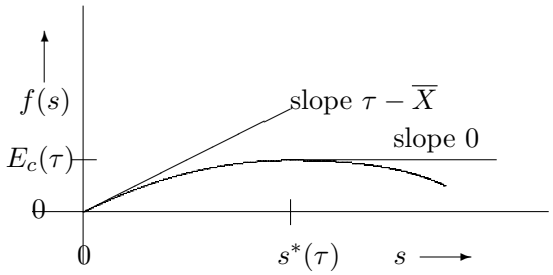

Figure 1. General form of function $f(s) = s\tau - \mu(s)$ when $\tau > \overline{X}$.

We will show below that the Chernoff exponent $E_c(\tau)$ is the correct exponent, in the sense that

$$\lim_{N\to\infty} \frac{\log \Pr\{S_N \geq N\tau\}}{N} = E_c(\tau).$$

The proof will be based on a fundamental theorem of large-deviation theory

We see that finding the Chernoff exponent is an exercise in convex optimization. In convex optimization theory, $E_c(\tau)$ and $\mu(s)$ are called conjugate functions. It is easy to show from the properties of $\mu(s)$ that $E_c(\tau)$ is a continuous, strictly convex function of τ that equals 0 at $\tau = \overline{X}$ and whose derivative at $\tau = \overline{X}$ is 0.

3.2.2 Chernoff bounds for functions of rvs

If $g : \mathcal{X} \to \mathbb{R}$ is any real-valued function defined on the alphabet $\mathcal{X}$ of a random variable X, then $g(X)$ is a real random variable. If $\{X_k\}$ is a sequence of iid random variables X_k with the same distribution as X, then $\{g(X_k)\}$ is a sequence of iid random variables $g(X_k)$ with the same distribution as $g(X)$. The Chernoff bound thus applies to the sequence $\{g(X_k)\}$, and shows that the probability that the sample mean $\frac{1}{N}\sum_k g(X_k)$ exceeds τ goes to zero exponentially with N as $N \to \infty$ whenever $\tau > \overline{g(X)}$.

Let us consider any finite set $\{g_j\}$ of such functions $g_j : \mathcal{X} \to \mathbb{R}$. Because the Chernoff bound decreases exponentially with N, we can conclude that the probability that *any* of the sample means $\frac{1}{N}\sum_k g_j(X_k)$ exceeds its corresponding expectation $\overline{g_j(X)}$ by a given fixed $\varepsilon > 0$ goes to zero exponentially with N as $N \to \infty$.

We may define a sequence $\{X_k\}$ to be ε-typical with respect to a function $g_j : \mathcal{X} \to \mathbb{R}$ if $\frac{1}{N}\sum_k g_j(X_k) < \overline{g_j(X)} + \varepsilon$. We can thus conclude that the probability that $\{X_k\}$ is not ε-typical with respect to any finite set $\{g_j\}$ of functions g_j goes to zero exponentially with N as $N \to \infty$.

A simple application of this result is that the probability that the sample mean $\frac{1}{N}\sum_k g_j(X_k)$ is not in the range $\overline{g_j(X)} \pm \varepsilon$ goes to zero exponentially with N as $N \to \infty$ for any $\varepsilon > 0$, because this probability is the sum of the two probabilities $\Pr\{\sum_k g_j(X_k) \geq N(\overline{g_j(X)} + \varepsilon)\}$ and $\Pr\{\sum_k -g_j(X_k) \geq N(-\overline{g_j(X)} + \varepsilon)\}$.

More generally, if the alphabet $\mathcal{X}$ is finite, then by considering the indicator functions of each possible value of X we can conclude that the probability that all observed relative frequencies in a sequence are not within ε of the corresponding probabilities goes to zero exponentially with N as $N \to \infty$. Similarly, for any alphabet $\mathcal{X}$, we can conclude that the probability of any finite number of sample moments $\frac{1}{N}\sum_k X_k^m$ are not within ε of the corresponding expected moments $\overline{X^m}$ goes to zero exponentially with N as $N \to \infty$.

In summary, the Chernoff bound law of large numbers allows us to say that as $N \to \infty$ we will almost surely observe a sample sequence $\mathbf{x}$ which is typical in every (finite) way that we might specify.

3.2.3 Asymptotic equipartition principle

One consequence of any law of large numbers is the asymptotic equipartition principle (AEP): as $N \to \infty$, the observed sample sequence $\mathbf{x}$ of an iid sequence whose elements are chosen according to a random variable X will almost surely be such that $p_X(\mathbf{x}) \approx 2^{-N\mathcal{H}(X)}$, where $\mathcal{H}(X) = \mathsf{E}_X[-\log_2 p(x)]$. If X is discrete, then $p_X(x)$ is its probability mass function (pmf) and $\mathcal{H}(X)$ is its entropy; if X is continuous, then $p_X(x)$ is its probability density function (pdf) and $\mathcal{H}(X)$ is its differential entropy.

The AEP is proved by observing that $-\log_2 p_X(\mathbf{x})$ is a sum of iid random variables $-\log_2 p_X(x_k)$, so the probability that $-\log_2 p_X(\mathbf{x})$ differs from its mean $N\mathcal{H}(X)$ by more than $\varepsilon > 0$ goes to zero as $N \to \infty$. The Chernoff bound shows that this probability in fact goes to zero exponentially with N.

A consequence of the AEP is that the set T_ε of all sequences $\mathbf{x}$ that are ε-typical with respect to the function $-\log_2 p_X(x)$ has a total probability that approaches 1 as $N \to \infty$. Since for all sequences $\mathbf{x} \in T_\varepsilon$ we have $p_X(\mathbf{x}) \approx 2^{-N\mathcal{H}(X)}$— *i.e.,* the probability distribution $p_X(\mathbf{x})$ is approximately uniform over T_ε— this implies that the "size" $|T_\varepsilon|$ of T_ε is approximately $2^{N\mathcal{H}(X)}$. In the discrete case, the "size" $|T_\varepsilon|$ is the number of sequences in T_ε, whereas in the continuous case $|T_\varepsilon|$ is the volume of T_ε.

In summary, the AEP implies that as $N \to \infty$ the observed sample sequence $\mathbf{x}$ will almost surely lie in an ε-typical set T_ε of size $\approx 2^{N\mathcal{H}(X)}$, and within that set the probability distribution $p_X(\mathbf{x})$ will be approximately uniform.

3.2.4 Fundamental theorem of large-deviation theory

As another application of the law of large numbers, we prove a fundamental theorem of large-deviation theory. A rough statement of this result is as follows: if an iid sequence $\mathbf{X}$ is chosen according to a probability distribution $q(x)$, then the probability that the sequence will be typical of a second probability distribution $p(x)$ is approximately

$$\Pr\{\mathbf{x} \text{ typical for } p \mid q\} \approx e^{-ND(p||q)},$$

where the exponent $D(p||q)$ denotes the relative entropy (Kullback-Leibler divergence)

$$D(p||q) = \mathsf{E}_p\left[\log \frac{p(x)}{q(x)}\right] = \int_{\mathcal{X}} dx\ p(x) \log \frac{p(x)}{q(x)}.$$

Again, $p(x)$ and $q(x)$ denote pmfs in the discrete case and pdfs in the continuous case; we use notation that is appropriate for the continuous case.

Exercise 2 (Gibbs' inequality).

(a) Prove that for $x > 0$, $\log x \leq x - 1$, with equality if and only if $x = 1$.

(b) Prove that for any pdfs $p(x)$ and $q(x)$ over $\mathcal{X}$, $D(p||q) \geq 0$, with equality if and only if $p(x) = q(x)$. □

Given $p(x)$ and $q(x)$, we will now define a sequence $\mathbf{x}$ to be ε-typical with regard to $\log p(x)/q(x)$ if the log likelihood ratio $\lambda(\mathbf{x}) = \log p(\mathbf{x})/q(\mathbf{x})$ is in the range $N(D(p||q) \pm \varepsilon)$, where $D(p||q) = \mathsf{E}_p[\lambda(x)]$ is the mean of $\lambda(x) = \log p(x)/q(x)$ under $p(x)$. Thus an iid sequence $\mathbf{X}$ chosen according to $p(x)$ will almost surely be ε-typical by this definition.

The desired result can then be stated as follows:

Theorem 3.3 (Fundamental theorem of large-deviation theory) *Given two probability distributions $p(x)$ and $q(x)$ on a common alphabet $\mathcal{X}$, for any $\varepsilon > 0$, the probability that an iid random sequence $\mathbf{X}$ drawn according to $q(x)$ is ε-typical for $p(x)$, in the sense that $\log p(\mathbf{x})/q(\mathbf{x})$ is in the range $N(D(p||q) \pm \varepsilon)$, is bounded by*

$$(1 - \delta(N))e^{-N(D(p||q)+\varepsilon)} \leq \Pr\{\mathbf{x}\ \varepsilon\text{−typical for } p \mid q\} \leq e^{-N(D(p||q)-\varepsilon)},$$

where $\delta(N) \to 0$ as $N \to \infty$.

Proof. Define the ε-typical region

$$T_\varepsilon = \{\mathbf{x} \mid N(D(p||q) - \varepsilon) \leq \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \leq N(D(p||q) + \varepsilon)\}.$$

By any law of large numbers, the probability that $\mathbf{X}$ will fall in T_ε goes to 1 as $N \to \infty$; *i.e.,*

$$1 - \delta(N) \leq \int_{T_\varepsilon} d\mathbf{x}\ p(\mathbf{x}) \leq 1,$$

where $\delta(N) \to 0$ as $N \to \infty$. It follows that

$$\begin{aligned}
\int_{T_\varepsilon} d\mathbf{x}\ q(\mathbf{x}) &\leq \int_{T_\varepsilon} d\mathbf{x}\ p(\mathbf{x}) e^{-N(D(p||q)-\varepsilon)} \leq e^{-N(D(p||q)-\varepsilon)}; \\
\int_{T_\varepsilon} d\mathbf{x}\ q(\mathbf{x}) &\geq \int_{T_\varepsilon} d\mathbf{x}\ p(\mathbf{x}) e^{-N(D(p||q)+\varepsilon)} \geq (1 - \delta(N)) e^{-N(D(p||q)+\varepsilon)}.
\end{aligned}$$

□

Since we can choose an arbitrarily small $\varepsilon > 0$ and $\delta(N) > 0$, it follows the exponent $D(p||q)$ is the correct exponent for this probability, in the sense that

$$\lim_{N\to\infty} \frac{\log \Pr\{\mathbf{x} \ \varepsilon\text{−typical for } p \mid q\}}{N} = D(p||q).$$

Exercise 3 (Generalization of Theorem 3.3).

(a) Generalize Theorem 3.3 to the case in which $q(x)$ is a general function over $\mathcal{X}$. State any necessary restrictions on $q(x)$.

(b) Using $q(x) = 1$ in (a), state and prove a form of the Asymptotic Equipartition Principle. □

As an application of Theorem 3.3, we can now prove:

Theorem 3.4 (Correctness of Chernoff exponent) *The Chernoff exponent $E_c(\tau)$ is the correct exponent for $\Pr\{S_N \geq N\tau\}$, in the sense that*

$$\lim_{N\to\infty} \frac{\log \Pr\{S_N \geq N\tau\}}{N} = E_c(\tau),$$

where $S_N = \sum_k x_k$ is the sum of N iid nondeterministic random variables drawn according to some distribution $p(x)$ with mean $\overline{X} < \tau$, and $E_c(\tau) = \max_{s\geq 0} s\tau - \mu(s)$ where $\mu(s) = \log \overline{e^{sX}}$.

Proof. Let s^* be the s that maximizes $s\tau - \mu(s)$ over $s \geq 0$. As we have seen above, for $s = s^*$ the tilted random variable $X(s^*)$ with tilted distribution $q(x, s^*) = p(x)e^{s^*x-\mu(s^*)}$ has mean $\overline{X(s^*)} = \tau$, whereas for $s = 0$ the untilted random variable $X(0)$ with untilted distribution $q(x, 0) = p(x)$ has mean $\overline{X(0)} = \overline{X}$.

Let $q(0)$ denote the untilted distribution $q(x, 0) = p(x)$ with mean $\overline{X(0)} = \overline{X}$, and let $q(s^*)$ denote the optimally tilted distribution $q(x, s^*) = p(x)e^{s^*x-\mu(s^*)}$ with mean $\overline{X(s^*)} = \tau$. Then $\log q(x, s^*)/q(x, 0) = s^*x - \mu(s^*)$, so

$$D(q(s^*)||q(0)) = s^*\tau - \mu(s^*) = E_c(\tau).$$

Moreover, the event that $\mathbf{X}$ is ε-typical with respect to the variable $\log q(x, s^*)/q(x, 0) = s^*x - \mu(s^*)$ under $q(x, 0) = p(x)$ is the event that $s^*S_N - N\mu(s^*)$ is in the range $N(s^*\tau - \mu(s^*) \pm \varepsilon)$, since τ is the mean of X under $q(x, s^*)$. This event is equivalent to S_N being in the range $N(\tau \pm \varepsilon/s^*)$. Since ε may be arbitrarily small, it is clear that the correct exponent of the event $\Pr\{S_N \approx N\tau\}$ is $E_c(\tau)$. This event evidently dominates the probability $\Pr\{S_N \geq N\tau\}$, which we have already shown to be upperbounded by $e^{-NE_c(\tau)}$. □

Exercise 4 (Chernoff bound ⇒ divergence upper bound.)

Using the Chernoff bound, prove that for any two distributions $p(x)$ and $q(x)$ over $\mathcal{X}$,

$$\Pr\{\log \frac{p(\mathbf{x})}{q(\mathbf{x})} \geq ND(p||q) \mid q\} \leq e^{-N(D(p||q))}.$$

[Hint: show that the s that maximizes $s\tau - \mu(s)$ is $s = 1$.] □

3.2.5 Proof of the forward part of the capacity theorem

We now prove that with Shannon's random Gaussian code ensemble and with a slightly different definition of typical-set decoding, we can achieve reliable communication at any rate $\rho < C_{[\mathrm{b/2D}]} = \log_2(1 + \mathrm{SNR})$ b/2D.

We recall that under this scenario the joint pdf of the channel input X and output Y is

$$p_{XY}(x, y) = p_X(x)p_N(y - x) = \frac{1}{\sqrt{2\pi S_x}} e^{-x^2/2S_x} \frac{1}{\sqrt{2\pi S_n}} e^{-(y-x)^2/2S_n}.$$

Since $Y = X + N$, the marginal probability of Y is

$$p_Y(y) = \frac{1}{\sqrt{2\pi S_y}} e^{-y^2/2S_y},$$

where $S_y = S_x + S_n$. On the other hand, since incorrect codewords are independent of the correct codeword and of the output, the joint pdf of an incorrect codeword symbol X' and of Y is

$$q_{XY}(x', y) = p_X(x')p_Y(y) = \frac{1}{\sqrt{2\pi S_x}} e^{-(x')^2/2S_x} \frac{1}{\sqrt{2\pi S_y}} e^{-y^2/2S_y}.$$

We now redefine typical-set decoding as follows. An output sequence $\mathbf{y}$ will be said to be ε-typical for a code sequence $\mathbf{x}$ if

$$\lambda(\mathbf{x}, \mathbf{y}) = \log \frac{p_{XY}(\mathbf{x}, \mathbf{y})}{p_X(\mathbf{x})p_Y(\mathbf{y})} \geq N(D(p_{XY}||p_X p_Y) - \varepsilon).$$

Substituting for the pdfs and recalling that $D(p_{XY}||p_X p_Y) = \frac{1}{2} \log S_y/S_n$, we find that this is equivalent to

$$\frac{||\mathbf{y} - \mathbf{x}||^2}{S_n} \leq \frac{||\mathbf{y}||^2}{S_y} + 2N\varepsilon.$$

Since $||\mathbf{y}||^2/N$ is almost surely very close to its mean S_y, this amounts to asking that $||\mathbf{y}-\mathbf{x}||^2/N$ be very close to its mean S_n under the hypothesis that $\mathbf{x}$ and $\mathbf{y}$ are drawn according to the joint pdf $p_{XY}(x, y)$. The correct codeword will therefore almost surely meet this test.

According to Exercise 4, the probability that any particular incorrect codeword meets the test

$$\lambda(\mathbf{x}, \mathbf{y}) = \log \frac{p_{XY}(\mathbf{x}, \mathbf{y})}{p_X(\mathbf{x})p_Y(\mathbf{y})} \geq ND(p_{XY}||p_X p_Y)$$

is upperbounded by $e^{-ND(p_{XY}||p_X p_Y)} = 2^{-NI(X;Y)}$. If we relax this test by an arbitrarily small number $\varepsilon > 0$, then by the continuity of the Chernoff exponent, the exponent will decrease by an amount $\delta(\varepsilon)$ which can be made arbitrarily small. Therefore we can assert that the probability that a random output sequence $\mathbf{Y}$ will be ε-typical for a random incorrect sequence $\mathbf{X}$ is upperbounded by

$$\Pr\{\mathbf{Y} \ \varepsilon\text{-typical for } \mathbf{X}\} \leq 2^{-N(I(X;Y)-\delta(\varepsilon))},$$

where $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Now if the random codes have rate $\rho < 2I(X;Y)$ b/2D, then there are $M = 2^{\rho N/2}$ codewords, so by the union bound the total probability of any incorrect codeword being ε-typical is upperbounded by

$$\Pr\{\mathbf{Y} \text{ } \varepsilon\text{-typical for any incorrect } \mathbf{X}\} \leq (M-1)2^{-N(I(X;Y)-\delta(\varepsilon))} < 2^{-N(I(X;Y)-\rho/2-\delta(\varepsilon))}.$$

If $\rho < 2I(X;Y)$ and ε is small enough, then the exponent will be positive and this probability will go to zero as $N \to \infty$.

Thus we have proved the forward part of the capacity theorem: the probability of any kind of error with Shannon's random code ensemble and this variant of typical-set decoding goes to zero as $N \to \infty$, in fact exponentially with N.

3.3 Geometric interpretation and converse

For AWGN channels, the channel capacity theorem has a nice geometric interpretation in terms of the geometry of spheres in real Euclidean N-space $\mathbb{R}^N$.

By any law of large numbers, the probability that the squared Euclidean norm $||\mathbf{X}||^2$ of a random sequence $\mathbf{X}$ of iid Gaussian variables of mean zero and variance S_x per symbol falls in the range $N(S_x \pm \varepsilon)$ goes to 1 as $N \to \infty$, for any $\varepsilon > 0$. Geometrically, the typical region

$$T_\varepsilon = \{\mathbf{x} \in \mathbb{R}^N \mid N(S_x - \varepsilon) \leq ||\mathbf{x}||^2 \leq N(S_x + \varepsilon)\}$$

is a spherical shell with outer squared radius $N(S_x + \varepsilon)$ and inner squared radius $N(S_x - \varepsilon)$. Thus the random N-vector $\mathbf{X}$ will almost surely lie in the spherical shell T_ε as $N \to \infty$. This phenomenon is known as "sphere hardening."

Moreover, the pdf $p_X(\mathbf{x})$ within the spherical shell T_ε is approximately uniform, as we expect from the asymptotic equipartition principle (AEP). Since $p_X(\mathbf{x}) = (2\pi S_x)^{-N/2} \exp -||\mathbf{x}||^2/2S_x$, within T_ε we have

$$(2\pi e S_x)^{-N/2} e^{-(N/2)(\varepsilon/S_x)} \leq p_X(\mathbf{x}) \leq (2\pi e S_x)^{-N/2} e^{(N/2)(\varepsilon/S_x)}.$$

Moreover, the fact that $p_X(\mathbf{x}) \approx (2\pi e S_x)^{-N/2}$ implies that the volume of T_ε is approximately $|T_\varepsilon| \approx (2\pi e S_x)^{N/2}$. More precisely, we have

$$1 - \delta(N) \leq \int_{T_\varepsilon} p_X(\mathbf{x}) \, d\mathbf{x} \leq 1,$$

where $\delta(N) \to 0$ as $N \to \infty$. Since $|T_\varepsilon| = \int_{T_\varepsilon} d\mathbf{x}$, we have

$$\begin{aligned} 1 \geq (2\pi e S_x)^{-N/2} e^{-(N/2)(\varepsilon/S_x)} |T_\varepsilon| &\Rightarrow |T_\varepsilon| \leq (2\pi e S_x)^{N/2} e^{(N/2)(\varepsilon/S_x)}; \\ 1 - \delta(N) \leq (2\pi e S_x)^{-N/2} e^{(N/2)(\varepsilon/S_x)} |T_\varepsilon| &\Rightarrow |T_\varepsilon| \geq (1-\delta(N))(2\pi e S_x)^{N/2} e^{-(N/2)(\varepsilon/S_x)}. \end{aligned}$$

Since these bounds hold for any $\varepsilon > 0$, this implies that

$$\lim_{N\to\infty} \frac{\log |T_\varepsilon|}{N} = \frac{1}{2} \log 2\pi e S_x = \mathcal{H}(X),$$

where $\mathcal{H}(X) = \frac{1}{2} \log 2\pi e S_x$ denotes the differential entropy of a Gaussian random variable with mean zero and variance S_x.

We should note at this point that practically all of the volume of an N-sphere of squared radius $N(S_x+\varepsilon)$ lies within the spherical shell $|T_\varepsilon|$ as $N\to\infty$, for any $\varepsilon>0$. By dimensional analysis, the volume of an N-sphere of radius r must be given by $A_N r^N$ for some constant A_N that does not depend on r. Thus the ratio of the volume of an N-sphere of squared radius $N(S_x-\varepsilon)$ to that of an N-sphere of squared radius $N(S_x+\varepsilon)$ must satisfy

$$\frac{A_N(N(S_x-\varepsilon))^{N/2}}{A_N(N(S_x+\varepsilon))^{N/2}}=\left(\frac{S_x-\varepsilon}{S_x+\varepsilon}\right)^{N/2}\to 0 \text{ as } N\to\infty, \text{ for any } \varepsilon>0.$$

It follows that the volume of an N-sphere of squared radius NS_x is also approximated by $e^{N\mathcal{H}(X)}=(2\pi eS_x)^{N/2}$ as $N\to\infty$.

Exercise 5. In Exercise 4 of Chapter 1, the volume of an N-sphere of radius r was given as

$$V_\otimes(N,r)=\frac{(\pi r^2)^{N/2}}{(N/2)!},$$

for N even. In other words, $A_N=\pi^{N/2}/((N/2)!)$. Using Stirling's approximation, $m!\to(m/e)^m$ as $m\to\infty$, show that this exact expression leads to the same asymptotic approximation for $V_\otimes(N,r)$ as was obtained above by use of the asymptotic equipartition principle. □

The sphere-hardening phenomenon may seem somewhat bizarre, but even more unexpected phenomena occur when we code for the AWGN channel using Shannon's random code ensemble.

In this case, each randomly chosen transmitted N-vector $\mathbf{X}$ will almost surely lie in a spherical shell T_X of squared radius $\approx NS_x$, and the random received N-vector $\mathbf{Y}$ will almost surely lie in a spherical shell T_Y of squared radius $\approx NS_y$, where $S_y=S_x+S_n$.

Moreover, given the correct transmitted codeword $\mathbf{c}_0$, the random received vector $\mathbf{Y}$ will almost surely lie in a spherical shell $T_\varepsilon(\mathbf{c}_0)$ of squared radius $\approx NS_n$ centered on $\mathbf{c}_0$. A further consequence of the AEP is that almost all of the volume of this nonzero-mean shell, whose center $\mathbf{c}_0$ has squared Euclidean norm $||\mathbf{c}_0||^2\approx NS_x$, lies in the zero-mean shell T_Y whose squared radius is $\approx NS_y$, since the expected squared Euclidean norm of $\mathbf{Y}=\mathbf{c}_0+\mathbf{N}$ is

$$\mathsf{E}_N[||\mathbf{Y}||^2]=||\mathbf{c}_0||^2+NS_n\approx NS_y.$$

"Curiouser and curiouser," said Alice.

We thus obtain the following geometrical picture. We choose $M=2^{\rho N/2}$ code vectors at random according to a zero-mean Gaussian distribution with variance S_x, which almost surely puts them within the shell T_X of squared radius $\approx NS_x$. Considering the probable effects of a random noise sequence $\mathbf{N}$ distributed according to a zero-mean Gaussian distribution with variance S_n, we can define for each code vector $\mathbf{c}_i$ a typical region $T_\varepsilon(\mathbf{c}_i)$ of volume $|T_\varepsilon(\mathbf{c}_i)|\approx(2\pi eS_n)^{N/2}$, which falls almost entirely within the shell T_Y of volume $|T_Y|\approx(2\pi eS_y)^{N/2}$.

Now if a particular code vector $\mathbf{c}_0$ is sent, then the probability that the received vector $\mathbf{y}$ will fall in the typical region $T_\varepsilon(\mathbf{c}_0)$ is nearly 1. On the other hand, the probability that $\mathbf{y}$ will fall in the typical region $T_\varepsilon(\mathbf{c}_i)$ of some other independently-chosen code vector $\mathbf{c}_i$ is approximately equal to the ratio $|T_\varepsilon(\mathbf{c}_i)|/|T_Y|$ of the volume of $T_\varepsilon(\mathbf{c}_i)$ to that of the entire shell, since if $\mathbf{y}$ is generated according to $p_y(\mathbf{y})$ independently of $\mathbf{c}_i$, then it will be approximately uniformly distributed over T_Y. Thus this probability is approximately

$$\Pr\{\mathbf{Y}\text{ typical for }\mathbf{c}_i\}\approx\frac{|T_\varepsilon(\mathbf{c}_i)|}{|T_Y|}\approx\frac{(2\pi eS_n)^{N/2}}{(2\pi eS_y)^{N/2}}=\left(\frac{S_n}{S_y}\right)^{N/2}.$$

As we have seen in earlier sections, this argument may be made precise.

It follows then that if $\rho < \log_2(1 + S_x/S_n)$ b/2D, or equivalently $M = 2^{\rho N/2} < (S_y/S_n)^{N/2}$, then the probability that $\mathbf{Y}$ is typical with respect to any of the $M - 1$ incorrect codewords is very small, which proves the forward part of the channel capacity theorem.

On the other hand, it is clear from this geometric argument that if $\rho > \log_2(1 + S_x/S_n)$ b/2D, or equivalently $M = 2^{\rho N/2} > (S_y/S_n)^{N/2}$, then the probability of decoding error must be large. For the error probability to be small, the decision region for each code vector $\mathbf{c}_i$ must include almost all of its typical region $T_\varepsilon(\mathbf{c}_i)$. If the volume of the $M = 2^{\rho N/2}$ typical regions exceeds the volume of T_Y, then this is impossible. Thus in order to have small error probability we must have

$$2^{\rho N/2}(2\pi e S_n)^{N/2} \le (2\pi e S_y)^{N/2} \quad \Rightarrow \quad \rho \le \log_2 \frac{S_y}{S_n} = \log_2(1 + \frac{S_x}{S_n}) \quad \text{b/2D}.$$

This argument may also be made precise, and is the converse to the channel capacity theorem.

In conclusion, we obtain the following picture of a capacity-achieving code. Let T_Y be the N-shell of squared radius $\approx NS_y$, which is almost the same thing as the N-sphere of squared radius NS_y. A capacity-achieving code consists of the centers $\mathbf{c}_i$ of M typical regions $T_\varepsilon(\mathbf{c}_i)$, where $||\mathbf{c}_i||^2 \approx NS_x$ and each region $T_\varepsilon(\mathbf{c}_i)$ consists of an N-shell of squared radius $\approx NS_n$ centered on $\mathbf{c}_i$, which is almost the same thing as an N-sphere of squared radius NS_x. As $\rho \to C_{[\text{b/2D}]} = \log_2(1 + \frac{S_x}{S_n})$ b/2D, these regions $T_\varepsilon(\mathbf{c}_i)$ form an almost disjoint partition of T_Y. This picture is illustrated in Figure 2.

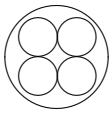

Figure 2. Packing $\approx (S_y/S_n)^{N/2}$ typical regions $T_\varepsilon(\mathbf{c}_i)$ of squared radius $\approx NS_n$ into a large typical region T_Y of squared radius $\approx NS_y$.

3.3.1 Discussion

It is natural in view of the above picture to frame the problem of coding for the AWGN channel as a sphere-packing problem. In other words, we might expect that a capacity-achieving code basically induces a disjoint partition of an N-sphere of squared radius NS_y into about $(S_y/S_n)^{N/2}$ disjoint decision regions, such that each decision region includes the sphere of squared radius NS_n about its center.

However, it can be shown by geometric arguments that such a disjoint partition is impossible as the code rate approaches capacity. What then is wrong with the sphere-packing approach? The subtle distinction that makes all the difference is that Shannon's probabilistic approach does not require decision regions to be disjoint, but merely probabilistically almost disjoint. So the solution to Shannon's coding problem involves what might be called "soft sphere-packing."

We will see that hard sphere-packing— *i.e.,* maximizing the minimum distance between code vectors subject to a constraint on average energy— is a reasonable approach for moderate-size codes at rates not too near to capacity. However, to obtain reliable transmission at rates near capacity, we will need to consider probabilistic codes and decoding algorithms that follow more closely the spirit of Shannon's original work.