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:

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

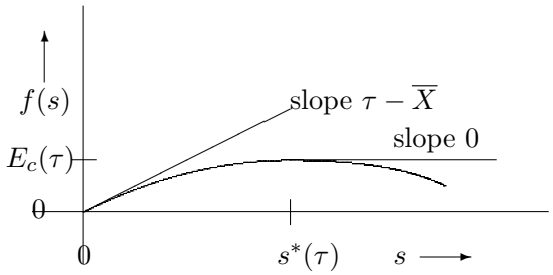

Figure 1. General form of function $f(s) = s\tau - \mu(s)$ when $\tau > \overline{X}$.

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.

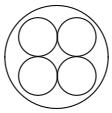

Figure 2. Packing $\approx (S_y/S_n)^{N/2}$ typical regions $T_\varepsilon(\mathbf{c}_i)$ of squared radius $\approx NS_n$ into a large typical region T_Y of squared radius $\approx NS_y$.

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.