Elements of Information Theory
Thomas M. Cover, Joy A. Thomas
Copyright © 1991 John Wiley & Sons, Inc.
Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 10

The Gaussian Channel

The most important continuous alphabet channel is the Gaussian channel depicted in Figure 10.1. This is a time discrete channel with output Y_i at time i, where Y_i is the sum of the input X_i and the noise Z_i. The noise Z_i is drawn i.i.d. from a Gaussian distribution with variance N. Thus

$$Y_i = X_i + Z_i, \qquad Z_i \sim \mathcal{N}(0, N). \tag{10.1}$$

The noise Z_i is assumed to be independent of the signal X_i. This channel is a good model for some common communication channels. Without further conditions, the capacity of this channel may be infinite. If the noise variance is zero, then the receiver receives the transmitted symbol perfectly. Since X can take on any real value, the channel can transmit an arbitrary real number with no error.

If the noise variance is non-zero and there is no constraint on the input, we can choose an infinite subset of inputs arbitrarily far apart, so that they are distinguishable at the output with arbitrarily small probability of error. Such a scheme has an infinite capacity as well. Thus if the noise variance is zero or the input is unconstrained, the capacity of the channel is infinite.

The most common limitation on the input is an energy or power constraint. We assume an average power constraint. For any codeword $(x_1, x_2, \ldots, x_n)$ transmitted over the channel, we require

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P. \tag{10.2}$$

This communication channel models many practical channels, including

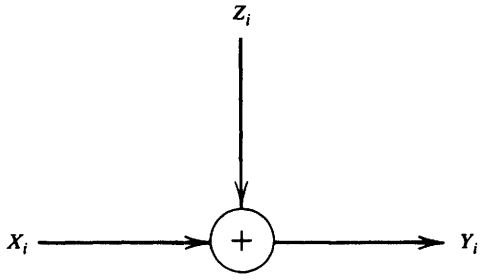


Figure 10.1. The Gaussian channel.

radio and satellite links. The additive noise in such channels may be due to a variety of causes. However, by the central limit theorem, the cumulative effect of a large number of small random effects will be approximately normal, so the Gaussian assumption is valid in a large number of situations.

We first analyze a simple suboptimal way to use this channel. Assume that we want to send 1 bit over the channel in 1 use of the channel. Given the power constraint, the best that we can do is to send one of two levels $+\sqrt{P}$ or $-\sqrt{P}$. The receiver looks at the corresponding received Y and tries to decide which of the two levels was sent. Assuming both levels are equally likely (this would be the case if we wish to send exactly 1 bit of information), the optimum decoding rule is to decide that $+\sqrt{P}$ was sent if $Y > 0$ and decide $-\sqrt{P}$ was sent if $Y < 0$. The probability of error with such a decoding scheme is

$$P_e = \frac{1}{2}\Pr(Y<0|X=+\sqrt{P}) + \frac{1}{2}\Pr(Y>0|X=-\sqrt{P}) \tag{10.3}$$

$$= \frac{1}{2}\Pr(Z<-\sqrt{P}|X=+\sqrt{P}) + \frac{1}{2}\Pr(Z>\sqrt{P}|X=-\sqrt{P}) \tag{10.4}$$

$$= \Pr(Z>\sqrt{P}) \tag{10.5}$$

$$= 1-\Phi\left(\sqrt{\frac{P}{N}}\right), \tag{10.6}$$

where $\Phi(x)$ is the cumulative normal function

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt. \tag{10.7}$$

Using such a scheme, we have converted the Gaussian channel into a discrete binary symmetric channel with crossover probability P_e. Similarly, by using a four level input signal, we can convert the Gaussian

channel into a discrete four input channel. In some practical modulation schemes, similar ideas are used to convert the continuous channel into a discrete channel. The main advantage of a discrete channel is ease of processing of the output signal for error correction, but some information is lost in the quantization.

10.1 THE GAUSSIAN CHANNEL: DEFINITIONS

We now define the (information) capacity of the channel as the maximum of the mutual information between the input and output over all distributions on the input that satisfy the power constraint.

Definition: The *information capacity* of the Gaussian channel with power constraint P is

$$C = \max_{p(x): EX^2 \le P} I(X; Y). \tag{10.8}$$

We can calculate the information capacity as follows: Expanding $I(X; Y)$, we have

$$I(X; Y) = h(Y) - h(Y|X) \tag{10.9}$$

$$= h(Y) - h(X + Z|X) \tag{10.10}$$

$$= h(Y) - h(Z|X) \tag{10.11}$$

$$= h(Y) - h(Z), \tag{10.12}$$

since Z is independent of X. Now, $h(Z) = \frac{1}{2} \log 2\pi eN$. Also,

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2 = P + N, \tag{10.13}$$

since X and Z are independent and $EZ = 0$. Given $EY^2 = P + N$, the entropy of Y is bounded by $\frac{1}{2} \log 2\pi e(P + N)$ by Theorem 9.6.5 (the normal maximizes the entropy for a given variance).

Applying this result to bound the mutual information, we obtain

$$I(X; Y) = h(Y) - h(Z) \tag{10.14}$$

$$\le \frac{1}{2} \log 2\pi e(P + N) - \frac{1}{2} \log 2\pi eN \tag{10.15}$$

$$= \frac{1}{2} \log\left(1 + \frac{P}{N}\right). \tag{10.16}$$

Hence the information capacity of the Gaussian channel is

$$C = \max_{EX^2 \le P} I(X; Y) = \frac{1}{2} \log\left(1 + \frac{P}{N}\right), \tag{10.17}$$

and the maximum is attained when $X \sim \mathcal{N}(0, P)$.

We will now show that this capacity is also the supremum of the achievable rates for the channel. The arguments are similar to the arguments for a discrete channel. We will begin with the corresponding definitions.

Definition: A (M, n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $x: \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$, satisfying the power constraint P, i.e., for every codeword

$$\sum_{i=1}^{n} x_i^2(w) \le nP, \qquad w = 1, 2, \ldots, M. \tag{10.18}$$

3. A decoding function

$$g: \mathcal{Y}^n \to \{1, 2, \ldots, M\}. \tag{10.19}$$

The rate and probability of error of the code are defined as in Chapter 8 for the discrete case.

Definition: A rate R is said to be *achievable* for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates.

Theorem 10.1.1: *The capacity of a Gaussian channel with power constraint P and noise variance N is*

$$C = \frac{1}{2} \log\left(1 + \frac{P}{N}\right) \textit{ bits per transmission}. \tag{10.20}$$

Remark: We will first present a plausibility argument as to why we may be able to construct $(2^{nC}, n)$ codes with low probability of error. Consider any codeword of length n. The received vector is normally distributed with mean equal to the true codeword and variance equal to

the noise variance. With high probability, the received vector is contained in a sphere of radius $\sqrt{n(N+\epsilon)}$ around the true codeword. If we assign everything within this sphere to the given codeword, then when this codeword is sent, there will be an error only if the received vector falls outside the sphere, which has low probability.

Similarly we can choose other codewords and their corresponding decoding spheres. How many such codewords can we choose? The volume of an n-dimensional sphere is of the form $A_n r^n$ where r is the radius of the sphere. In this case, each of the decoding spheres has radius $\sqrt{nN}$. These spheres are scattered throughout the space of received vectors. The received vectors have energy no greater than $n(P+N)$ so they lie in a sphere of radius $\sqrt{n(P+N)}$. The maximum number of non-intersecting decoding spheres in this volume is no more than

$$\frac{A_n(n(P+N))^{\frac{n}{2}}}{A_n(nN)^{\frac{n}{2}}} = 2^{\frac{n}{2}\log\left(1+\frac{P}{N}\right)} \tag{10.21}$$

and the rate of the code is $\frac{1}{2}\log(1+\frac{P}{N})$. This idea is illustrated in Figure 10.2.

This sphere packing argument indicates that we cannot hope to send at rates greater than C with low probability of error. However, we can actually do almost as well as this, as is proved next.

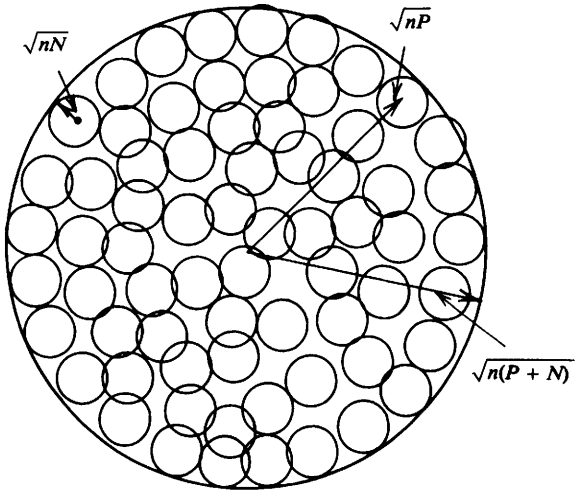


Figure 10.2. Sphere packing for the Gaussian channel.

Proof (*Achievability*): We will use the same ideas as in the proof of the channel coding theorem in the case of discrete channels, namely, random codes and joint typicality decoding. However, we must make some modifications to take into account the power constraint and the fact that the variables are continuous and not discrete.

1. *Generation of the codebook.* We wish to generate a codebook in which all the codewords satisfy the power constraint. To ensure this, we generate the codewords with each element i.i.d. according to a normal distribution with variance $P - \epsilon$. Since for large n, $\frac{1}{n} \Sigma X_i^2 \to P - \epsilon$, the probability that a codeword does not satisfy the power constraint will be small. However, we do not delete the bad codewords, as this will disturb the symmetry of later arguments.
   Let $X_i(w)$, $i = 1, 2, \ldots, n$, $w = 1, 2, \ldots, 2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P - \epsilon)$, forming codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR}) \in \mathcal{R}^n$.
2. *Encoding.* After the generation of the codebook, the codebook is revealed to both the sender and the receiver. To send the message index w, the transmitter sends the wth codeword $X^n(w)$ in the codebook.
3. *Decoding.* The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector. If there is one and only one such codeword, the receiver declares it to be the transmitted codeword. Otherwise the receiver declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.
4. *Probability of error.* Without loss of generality, assume that codeword 1 was sent. Thus $Y^n = X^n(1) + Z^n$.

Define the following events:

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2(1) > P \right\} \tag{10.22}$$

and

$$E_i = \{(X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}\} . \tag{10.23}$$

Then an error occurs if E_0 occurs (the power constraint is violated) or E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}}$ occurs (some wrong codeword is jointly typical with the received sequence). Let $\mathcal{E}$ denote the event $\hat{W} \neq W$ and let P denote the conditional probability given $W = 1$. Hence

$$\Pr(\mathcal{E} | W = 1) = P(\mathcal{E}) = P(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \tag{10.24}$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i), \tag{10.25}$$

by the union of events bound for probabilities. By the law of large numbers, $P(E_0) \to 0$ as $n \to \infty$. Now, by the joint AEP (which can be proved using the same argument used in the discrete case), $P(E_1^c) \to 0$, and hence

$$P(E_1^c) \leq \epsilon \quad \text{for } n \text{ sufficiently large}. \tag{10.26}$$

Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n will be jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP. Hence

$$P_e^{(n)} = \Pr(\mathcal{E}) = \Pr(\mathcal{E} \mid W = 1) = P(\mathcal{E}) \tag{10.27}$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \tag{10.28}$$

$$\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \tag{10.29}$$

$$= 2\epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \tag{10.30}$$

$$\leq 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \tag{10.31}$$

$$\leq 3\epsilon \tag{10.32}$$

for n sufficiently large and $R < I(X;Y) - 3\epsilon$.

This proves the existence of a good $(2^{nR}, n)$ code.

Now choosing a good codebook and deleting the worst half of the codewords, we obtain a code with low maximal probability of error. In particular, the power constraint is satisfied by each of the remaining codewords (since the codewords that do not satisfy the power constraint have probability of error 1 and must belong to the worst half of the codewords).

Hence we have constructed a code which achieves a rate arbitrarily close to capacity. The forward part of the theorem is proved. In the next section, we show that the rate cannot exceed the capacity. □

10.2 CONVERSE TO THE CODING THEOREM FOR GAUSSIAN CHANNELS

In this section, we complete the proof that the capacity of a Gaussian channel is $C = \frac{1}{2}\log(1 + \frac{P}{N})$ by proving that rates $R > C$ are not achiev-

able. The proof parallels the proof for the discrete channel. The main new ingredient is the power constraint.

Proof (*Converse to Theorem 10.1.1*): We must show that if $P_e^{(n)} \to 0$ for a sequence of $(2^{nR}, n)$ codes for a Gaussian channel with power constraint P, then

$$R \le C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right). \tag{10.33}$$

Consider any $(2^{nR}, n)$ code that satisfies the power constraint, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \le P, \tag{10.34}$$

for $w = 1, 2, \ldots, 2^{nR}$. Proceeding as in the converse for the discrete case, the uniform distribution over the index set $w \in \{1, 2, \ldots, 2^{nR}\}$ induces a distribution on the input codewords, which in turn induces a distribution over the input alphabet. Since we can decode the index W from the output vector Y^n with low probability of error, we can apply Fano's inequality to obtain

$$H(W|Y^n) \le 1 + nRP_e^{(n)} = n\epsilon_n, \tag{10.35}$$

where $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$. Hence

$$nR = H(W) = I(W; Y^n) + H(W|Y^n) \tag{10.36}$$

$$\le I(W; Y^n) + n\epsilon_n \tag{10.37}$$

$$\le I(X^n; Y^n) + n\epsilon_n \tag{10.38}$$

$$= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \tag{10.39}$$

$$= h(Y^n) - h(Z^n) + n\epsilon_n \tag{10.40}$$

$$\le \sum_{i=1}^{n} h(Y_i) - h(Z^n) + n\epsilon_n \tag{10.41}$$

$$= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{10.42}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n. \tag{10.43}$$

Here $X_i = X_i(W)$, where W is drawn according to the uniform distribution on $\{1, 2, \ldots, 2^{nR}\}$. Now let P_i be the average power of the ith column of the codebook, i.e.,

$$P_i = \frac{1}{2^{nR}} \sum_w x_i^2(w) . \tag{10.44}$$

Then, since $Y_i = X_i + Z_i$ and since X_i and Z_i are independent, the average power of Y_i is $P_i + N$. Hence, since entropy is maximized by the normal distribution,

$$h(Y_i) \le \frac{1}{2} \log 2\pi e(P_i + N) . \tag{10.45}$$

Continuing with the inequalities of the converse, we obtain

$$nR \le \sum (h(Y_i) - h(Z_i)) + n\epsilon_n \tag{10.46}$$

$$\le \sum \left(\frac{1}{2} \log(2\pi e(P_i + N)) - \frac{1}{2} \log 2\pi eN \right) + n\epsilon_n \tag{10.47}$$

$$= \sum \frac{1}{2} \log\left(1 + \frac{P_i}{N}\right) + n\epsilon_n . \tag{10.48}$$

Since each of the codewords satisfies the power constraint, so does their average, and hence

$$\frac{1}{n} \sum_i P_i \le P . \tag{10.49}$$

Since $f(x) = \frac{1}{2} \log(1 + x)$ is a concave function of x, we can apply Jensen's inequality to obtain

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log\left(1 + \frac{P_i}{N}\right) \le \frac{1}{2} \log\left(1 + \frac{1}{n} \sum_{i=1}^{n} \frac{P_i}{N}\right) \tag{10.50}$$

$$\le \frac{1}{2} \log\left(1 + \frac{P}{N}\right) . \tag{10.51}$$

Thus $R \le \frac{1}{2} \log(1 + \frac{P}{N}) + \epsilon_n$, $\epsilon_n \to 0$, and we have the required converse. □

Note that the power constraint enters the standard proof in (10.44).

10.3 BAND-LIMITED CHANNELS

A common model for communication over a radio network or a telephone line is a band-limited channel with white noise. This is a continuous time channel. The output of such a channel can be described as

$$Y(t) = (X(t) + Z(t)) * h(t) , \tag{10.52}$$

where $X(t)$ is the signal waveform, $Z(t)$ is the waveform of the white Gaussian noise, and $h(t)$ is the impulse response of an ideal bandpass filter, which cuts out all frequencies greater than W. In this section, we give simplified arguments to calculate the capacity of such a channel.

We begin with a representation theorem due to Nyquist [199] and Shannon [240], which shows that sampling a band-limited signal at a sampling rate $\frac{1}{2W}$ is sufficient to reconstruct the signal from the samples. Intuitively, this is due to the fact that if a signal is band-limited to W, then it cannot change by a substantial amount in a time less than half a cycle of the maximum frequency in the signal, that is, the signal cannot change very much in time intervals less than $\frac{1}{2W}$ seconds.

Theorem 10.3.1: *Suppose a function $f(t)$ is band-limited to W, namely, the spectrum of the function is 0 for all frequencies greater than W. Then the function is completely determined by samples of the function spaced $\frac{1}{2W}$ seconds apart.*

Proof: Let $F(\omega)$ be the frequency spectrum of $f(t)$. Then

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} \, d\omega \tag{10.53}$$

$$= \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega) e^{i\omega t} \, d\omega \,, \tag{10.54}$$

since $F(\omega)$ is 0 outside the band $-2\pi W \leq \omega \leq 2\pi W$. If we consider samples spaced $\frac{1}{2W}$ seconds apart, the value of the signal at the sample points can be written

$$f\left(\frac{n}{2W}\right) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega) e^{i\omega \frac{n}{2W}} \, d\omega \,. \tag{10.55}$$

The right hand side of this equation is also the definition of the coefficients of the Fourier series expansion of the periodic extension of the function $F(\omega)$, taking the interval $-2\pi W$ to $2\pi W$ as the fundamental period. Thus the sample values $f(\frac{n}{2W})$ determine the Fourier coefficients and, by extension, they determine the value of $F(\omega)$ in the interval $(-2\pi W, 2\pi W)$. Since a function is uniquely specified by its Fourier transform, and since $F(\omega)$ is 0 outside the band W, we can determine the function uniquely from the samples.

Consider the function

$$\text{sinc}(t) = \frac{\sin(2\pi W t)}{2\pi W t} \,. \tag{10.56}$$

This function is 1 at $t = 0$ and is 0 for $t = n/2W$, $n \neq 0$. The spectrum of this function is constant in the band $(-W, W)$ and is zero outside this

band. Now define

$$g(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right) \operatorname{sinc}\left(t - \frac{n}{2W}\right). \tag{10.57}$$

From the properties of the sinc function, it follows that $g(t)$ is band-limited to W and is equal to $f(n/2W)$ at $t = n/2W$. Since there is only one function satisfying these constraints, we must have $g(t) = f(t)$. This provides an explicit representation of $f(t)$ in terms of its samples. □

A general function has an infinite number of degrees of freedom—the value of the function at every point can be chosen independently. The Nyquist-Shannon sampling theorem shows that a band-limited function has only $2W$ degrees of freedom per second. The values of the function at the sample points can be chosen independently, and this specifies the entire function.

If a function is band-limited, it cannot be limited in time. But we can consider functions that have most of their energy in bandwidth W and have most of their energy in a finite time interval, say $(0, T)$. We can describe these functions using a basis of *prolate spheroidal functions*. We do not go into the details of this theory here; it suffices to say that there are about $2TW$ orthonormal basis functions for the set of almost time-limited, almost band-limited functions, and we can describe any function within the set by its coordinates in this basis. The details can be found in a series of papers by Slepian, Landau and Pollak [169], [168], [253]. Moreover, the projection of white noise on these basis vectors forms an i.i.d. Gaussian process. The above arguments enable us to view the band-limited, time-limited functions as vectors in a vector space of $2TW$ dimensions.

Now we return to the problem of communication over a band-limited channel. Assuming that the channel has bandwidth W, we can represent both the input and the output by samples taken $1/2W$ seconds apart. Each of the input samples is corrupted by noise to produce the corresponding output sample. Since the noise is white and Gaussian, it can be shown that each of the noise samples is an independent, identically distributed Gaussian random variable. If the noise has power spectral density $N_0/2$ and bandwidth W, then the noise has power $\frac{N_0}{2} 2W = N_0 W$ and each of the $2WT$ noise samples in time T has variance $N_0 WT/2WT = N_0/2$. Looking at the input as a vector in the $2TW$ dimensional space, we see that the received signal is spherically normally distributed about this point with covariance $\frac{N_0}{2} I$.

Now we can use the theory derived earlier for discrete time Gaussian channels, where it was shown that the capacity of such a channel is

$$C = \frac{1}{2} \log\left(1 + \frac{P}{N}\right) \text{ bits per transmission}. \tag{10.58}$$

Let the channel be used over the time interval $[0, T]$. In this case, the power per sample is $PT/2WT = P/2W$, the noise variance per sample is $\frac{N_0}{2} 2W \frac{T}{2WT} = N_0/2$, and hence the capacity per sample is

$$C = \frac{1}{2} \log\left(1 + \frac{\frac{P}{2W}}{\frac{N_0}{2}}\right) = \frac{1}{2} \log\left(1 + \frac{P}{N_0 W}\right) \text{ bits per sample}. \tag{10.59}$$

Since there are $2W$ samples each second, the capacity of the channel can be rewritten as

$$C = W \log\left(1 + \frac{P}{N_0 W}\right) \text{ bits per second}. \tag{10.60}$$

This equation is one of the most famous formulae of information theory. It gives the capacity of a band-limited Gaussian channel with noise spectral density $N_0/2$ watts/Hz and power P watts.

If we let $W \to \infty$ in (10.60), we obtain

$$C = \frac{P}{N_0} \log_2 e \text{ bits per second}, \tag{10.61}$$

as the capacity of a channel with an infinite bandwidth, power P and noise spectral density $N_0/2$. Thus for infinite bandwidth channels, the capacity grows linearly with the power.

Example 10.3.1 (*Telephone line*): To allow multiplexing of many channels, telephone signals are band-limited to 3300 Hz. Using a bandwidth of 3300 Hz and a SNR (signal to noise ratio) of 20 dB (i.e., $P/N_0W = 100$), in (10.60), we find the capacity of the telephone channel to be about 22,000 bits per second. Practical modems achieve transmission rates up to 19,200 bits per second. In real telephone channels, there are other factors such as crosstalk, interference, echoes, non-flat channels, etc. which must be compensated for to achieve this capacity.

10.4 PARALLEL GAUSSIAN CHANNELS

In this section, we consider k independent Gaussian channels in parallel with a common power constraint. The objective is to distribute the total power among the channels so as to maximize the capacity. This channel models a non-white additive Gaussian noise channel where each parallel component represents a different frequency.

Assume that we have a set of Gaussian channels in parallel as illustrated in Figure 10.3. The output of each channel is the sum of the input and Gaussian noise. For channel j,

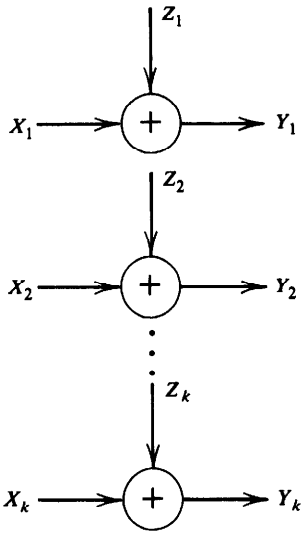


Figure 10.3. Parallel Gaussian channels.

$$Y_j = X_j + Z_j, \quad j = 1, 2, \ldots, k, \tag{10.62}$$

with

$$Z_j \sim \mathcal{N}(0, N_j), \tag{10.63}$$

and the noise is assumed to be independent from channel to channel. We assume that there is a common power constraint on the total power used, i.e.,

$$E \sum_{j=1}^{k} X_j^2 \leq P. \tag{10.64}$$

We wish to distribute the power among the various channels so as to maximize the total capacity.

The information capacity of the channel C is

$$C = \max_{f(x_1, x_2, \ldots, x_k): \sum EX_i^2 \leq P} I(X_1, X_2, \ldots, X_k; Y_1, Y_2, \ldots, Y_k). \tag{10.65}$$

We calculate the distribution that achieves the information capacity for this channel. The fact that the information capacity is the supremum of achievable rates can be proved by methods identical to those in the proof of the capacity theorem for single Gaussian channels and will be omitted.

Since $Z_1, Z_2, \ldots, Z_k$ are independent,

$$\begin{aligned} I(X_1, X_2, \ldots, X_k; Y_1, Y_2, \ldots, Y_k) &= h(Y_1, Y_2, \ldots, Y_k) - h(Y_1, Y_2, \ldots, Y_k | X_1, X_2, \ldots, X_k) \\ &= h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k | X_1, X_2, \ldots, X_k) \\ &= h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k) \qquad (10.66) \\ &= h(Y_1, Y_2, \ldots, Y_k) - \sum_i h(Z_i) \qquad (10.67) \\ &\leq \sum_i h(Y_i) - h(Z_i) \qquad (10.68) \\ &\leq \sum_i \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right), \qquad (10.69) \end{aligned}$$

where $P_i = EX_i^2$, and $\Sigma P_i = P$. Equality is achieved by

$$(X_1, X_2, \ldots, X_k) \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & P_k \end{bmatrix}\right). \qquad (10.70)$$

So the problem is reduced to finding the power allotment that maximizes the capacity subject to the constraint that $\Sigma P_i = P$. This is a standard optimization problem and can be solved using Lagrange multipliers. Writing the functional as

$$J(P_1, \ldots, P_k) = \sum \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right) + \lambda\left(\sum P_i\right) \qquad (10.71)$$

and differentiating with respect to P_i, we have

$$\frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0, \qquad (10.72)$$

or

$$P_i = \nu - N_i. \qquad (10.73)$$

However since the P_i's must be non-negative, it may not always be possible to find a solution of this form. In this case, we use the Kuhn-Tucker conditions to verify that the solution

$$P_i = (\nu - N_i)^+ \qquad (10.74)$$

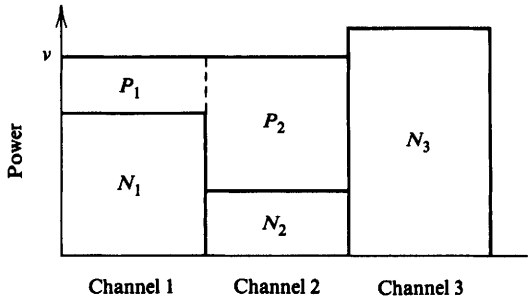


Figure 10.4. Water-filling for parallel channels.

is the assignment that maximizes capacity, where ν is chosen so that

$$\sum (\nu - N_i)^+ = P. \tag{10.75}$$

Here $(x)^+$ denotes the positive part of x, i.e.,

$$(x)^+ = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{10.76}$$

This solution is illustrated graphically in Figure 10.4. The vertical levels indicate the noise levels in the various channels. As signal power is increased from zero, we allot the power to the channels with the lowest noise. When the available power is increased still further, some of the power is put into noisier channels. The process by which the power is distributed among the various bins is identical to the way in which water distributes itself in a vessel. Hence this process is sometimes referred to as "water-filling."

10.5 CHANNELS WITH COLORED GAUSSIAN NOISE

In the previous section, we considered the case of a set of parallel independent Gaussian channels in which the noise samples from different channels were independent. Now we will consider the case when the noise is dependent. This represents not only the case of parallel channels, but also the case when the channel has Gaussian noise with memory. For channels with memory, we can consider a block of n consecutive uses of the channel as n channels in parallel with dependent noise. As in the previous section, we will only calculate the information capacity for this channel.

Let K_Z be the covariance matrix of the noise, and let K_X be the input covariance matrix. The power constraint on the input can then be written as

$$\frac{1}{n}\sum_i EX_i^2 \le P\,, \tag{10.77}$$

or equivalently,

$$\frac{1}{n}\operatorname{tr}(K_X) \le P\,. \tag{10.78}$$

Unlike the previous section, the power constraint here depends on n; the capacity will have to be calculated for each n.

Just as in the case of independent channels, we can write

$$I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) = h(Y_1, Y_2, \ldots, Y_n) - h(Z_1, Z_2, \ldots, Z_n)\,. \tag{10.79}$$

Here $h(Z_1, Z_2, \ldots, Z_n)$ is determined only by the distribution of the noise and is not dependent on the choice of input distribution. So finding the capacity amounts to maximizing $h(Y_1, Y_2, \ldots, Y_n)$. The entropy of the output is maximized when Y is normal, which is achieved when the input is normal. Since the input and the noise are independent, the covariance of the output Y is $K_Y = K_X + K_Z$ and the entropy is

$$h(Y_1, Y_2, \ldots, Y_n) = \frac{1}{2}\log((2\pi e)^n |K_X + K_Z|)\,. \tag{10.80}$$

Now the problem is reduced to choosing K_X so as to maximize $|K_X + K_Z|$, subject to a trace constraint on K_X. To do this, we decompose K_Z into its diagonal form,

$$K_Z = Q\Lambda Q^t, \quad \text{where } QQ^t = I\,. \tag{10.81}$$

Then

$$|K_X + K_Z| = |K_X + Q\Lambda Q^t| \tag{10.82}$$

$$= |Q||Q^t K_X Q + \Lambda||Q^t| \tag{10.83}$$

$$= |Q^t K_X Q + \Lambda| \tag{10.84}$$

$$= |A + \Lambda|\,, \tag{10.85}$$

where $A = Q^t K_X Q$. Since for any matrices B and C,

$$\operatorname{tr}(BC) = \operatorname{tr}(CB)\,, \tag{10.86}$$

we have

$$\text{tr}(A) = \text{tr}(Q^t K_X Q) \tag{10.87}$$

$$= \text{tr}(QQ^t K_X) \tag{10.88}$$

$$= \text{tr}(K_X) . \tag{10.89}$$

Now the problem is reduced to maximizing $|A + \Lambda|$ subject to a trace constraint $\text{tr}(A) \leq nP$.

Now we apply Hadamard's inequality, mentioned in Chapter 9. Hadamard's inequality states that the determinant of any positive definite matrix K is less than the product of its diagonal elements, i.e.,

$$|K| \leq \prod_i K_{ii} \tag{10.90}$$

with equality iff the matrix is diagonal. Thus

$$|A + \Lambda| \leq \prod_i (A_{ii} + \lambda_i) \tag{10.91}$$

with equality iff A is diagonal. Since A is subject to a trace constraint,

$$\frac{1}{n} \sum_i A_{ii} \leq P , \tag{10.92}$$

and $A_{ii} \geq 0$, the maximum value of $\Pi_i (A_{ii} + \lambda_i)$ is attained when

$$A_{ii} + \lambda_i = \nu . \tag{10.93}$$

However, given the constraints, it may not be always possible to satisfy this equation with positive A_{ii}. In such cases, we can show by standard Kuhn-Tucker conditions that the optimum solution corresponds to setting

$$A_{ii} = (\nu - \lambda_i)^+ , \tag{10.94}$$

where ν is chosen so that $\Sigma A_{ii} = nP$. This value of A maximizes the entropy of Y and hence the mutual information. We can use Figure 10.4 to see the connection between the methods described above and "water-filling".

Consider a channel in which the additive Gaussian noise forms a stochastic process with finite dimensional covariance matrix $K_Z^{(n)}$. If the process is stationary, then the covariance matrix is Toeplitz and the eigenvalues tend to a limit as $n \to \infty$. The density of eigenvalues on the real line tends to the power spectrum of the stochastic process [126]. In this case, the above "water-filling" argument translates to water-filling in the spectral domain.

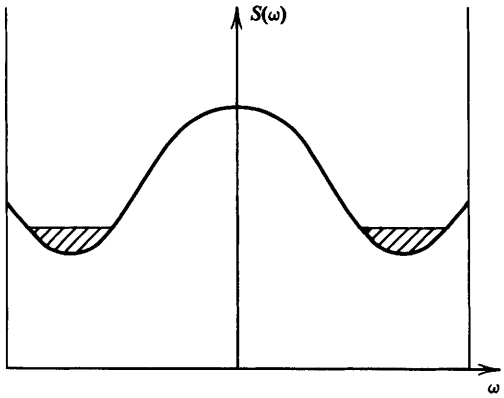


Figure 10.5. Water-filling in the spectral domain.

Hence for channels in which the noise forms a stationary stochastic process, the input signal should be chosen to be a Gaussian process with a spectrum which is large at frequencies where the noise spectrum is small. This is illustrated in Figure 10.5. The capacity of an additive Gaussian noise channel with noise power spectrum $N(f)$ can be shown to be [120]

$$C = \int \frac{1}{2} \log\left(1 + \frac{(\nu - N(f))^+}{N(f)}\right) df, \tag{10.95}$$

where ν is chosen so that $\int (\nu - N(f))^+ \, df = P$.

10.6 GAUSSIAN CHANNELS WITH FEEDBACK

In Chapter 8, we proved that feedback does not increase the capacity for discrete memoryless channels. It can greatly help in reducing the complexity of encoding or decoding. The same is true of an additive noise channel with white noise. As in the discrete case, feedback does not increase capacity for memoryless Gaussian channels. However, for channels with memory, where the noise is correlated from time instant to time instant, feedback does increase capacity. The capacity without feedback can be calculated using water-filling, but we do not have a simple explicit characterization of the capacity with feedback. In this section, we describe an expression for the capacity in terms of the covariance matrix of the noise Z. We prove a converse for this expression for capacity. We then derive a simple bound on the increase in capacity due to feedback.

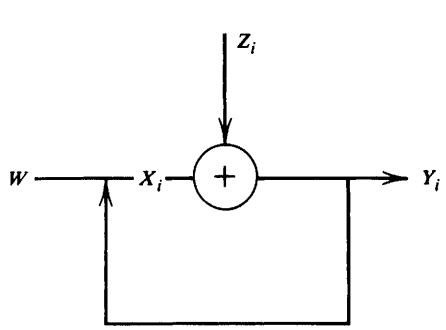


Figure 10.6. Gaussian channel with feedback.

The Gaussian channel with feedback is illustrated in Figure 10.6. The output of the channel Y_i is

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, K_Z^{(n)}). \tag{10.96}$$

The feedback allows the input of the channel to depend on the past values of the output.

A $(2^{nR}, n)$ code for the Gaussian channel with feedback consists of a sequence of mappings $x_i(W, Y^{i-1})$, where $W \in \{1, 2, \ldots, 2^{nR}\}$ is the input message and Y^{i-1} is the sequence of past values of the output. Thus $x(W, \cdot)$ is a code function rather than a codeword. In addition, we require that the code satisfy a power constraint,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2(w, Y^{i-1})\right] \le P, \quad w \in \{1, 2, \ldots, 2^{nR}\}, \tag{10.97}$$

where the expectation is over all possible noise sequences.

We will characterize the capacity of the Gaussian channel is terms of the covariance matrices of the input X and the noise Z. Because of the feedback, X^n and Z^n are not independent; X_i depends causally on the past values of Z. In the next section, we prove a converse for the Gaussian channel with feedback and show that we achieve capacity if we take X to be Gaussian.

We now state an informal characterization of the capacity of the channel with and without feedback.

1. *With feedback.* The capacity $C_{n,FB}$ in bits per transmission of the time-varying Gaussian channel with feedback is

$$C_{n,FB} = \max_{\frac{1}{n}\operatorname{tr}(K_X^{(n)}) \le P} \frac{1}{2n} \log \frac{|K_{X+Z}^{(n)}|}{|K_Z^{(n)}|} \tag{10.98}$$

where the maximization is taken over all X^n of the form

$$X_i = \sum_{j=1}^{i-1} b_{ij} Z_j + V_i, \quad i = 1, 2, \ldots, n, \tag{10.99}$$

and V^n is independent of Z^n.

To verify that the maximization over (10.99) involves no loss of generality, note that the distribution on $X^n + Z^n$ achieving the maximum entropy is Gaussian. Since Z^n is also Gaussian, it can be verified that a jointly Gaussian distribution on $(X^n, Z^n, X^n + Z^n)$ achieves the maximization in (10.98). But since $Z^n = Y^n - X^n$, the most general jointly normal causal dependence of X^n on Y^n is of the form (10.99), where V^n plays the role of the innovations process. Recasting (10.98) and (10.99) using $X = BZ + V$ and $Y = X + Z$, we can write

$$C_{n,FB} = \max \frac{1}{2n} \log \frac{|(B+I)K_Z^{(n)}(B+I)^t + K_V|}{|K_Z^{(n)}|} \tag{10.100}$$

where the maximum is taken over all nonnegative definite K_V and strictly lower triangular B such that

$$\operatorname{tr}(BK_Z^{(n)}B^t + K_V) \le nP. \tag{10.101}$$

(Without feedback, B is necessarily 0.)

2. *Without feedback.* The capacity C_n of the time-varying Gaussian channel without feedback is given by

$$C_n = \max_{\frac{1}{n}\operatorname{tr}(K_X^{(n)}) \le P} \frac{1}{2n} \log \frac{|K_X^{(n)} + K_Z^{(n)}|}{|K_Z^{(n)}|}. \tag{10.102}$$

This reduces to water-filling on the eigenvalues $\{\lambda_i^{(n)}\}$ of $K_Z^{(n)}$. Thus

$$C_n = \frac{1}{2n} \sum_{i=1}^{n} \log\left(1 + \frac{(\lambda - \lambda_i^{(n)})^+}{\lambda_i^{(n)}}\right), \tag{10.103}$$

where $(y)^+ = \max\{y, 0\}$ and where λ is chosen so that

$$\sum_{i=1}^{n} (\lambda - \lambda_i^{(n)})^+ = nP. \tag{10.104}$$

We now prove an upper bound for the capacity of the Gaussian channel with feedback. This bound is actually achievable, and is therefore the capacity, but we do not prove this here.

Theorem 10.6.1: *The rate R_n for any $(2^{nR_n}, n)$ code with $P_e^{(n)} \to 0$ for the Gaussian channel with feedback satisfies*

$$R_n \le \frac{1}{n}\frac{1}{2}\log\frac{|K_Y^{(n)}|}{|K_Z^{(n)}|} + \epsilon_n , \tag{10.105}$$

with $\epsilon_n \to 0$ as $n \to \infty$.

Proof: By Fano's inequality,

$$H(W|Y^n) \le 1 + nR_nP_e^{(n)} = n\epsilon_n , \tag{10.106}$$

where $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$. We can then bound the rate as follows:

$$nR_n = H(W) \tag{10.107}$$

$$= I(W; Y^n) + H(W|Y^n) \tag{10.108}$$

$$\le I(W; Y^n) + n\epsilon_n \tag{10.109}$$

$$= \sum I(W; Y_i|Y^{i-1}) + n\epsilon_n \tag{10.110}$$

$$\overset{(a)}{=} \sum [h(Y_i|Y^{i-1}) - h(Y_i|W, Y^{i-1}, X_i, X^{i-1}, Z^{i-1})] + n\epsilon_n \tag{10.111}$$

$$\overset{(b)}{=} \sum [h(Y_i|Y^{i-1}) - h(Z_i|W, Y^{i-1}, X_i, X^{i-1}, Z^{i-1})] + n\epsilon_n \tag{10.112}$$

$$\overset{(c)}{=} \sum [h(Y_i|Y^{i-1}) - h(Z_i|Z^{i-1})] + n\epsilon_n \tag{10.113}$$

$$= h(Y^n) - h(Z^n) + n\epsilon_n , \tag{10.114}$$

where (a) follows from the fact that X_i is a function of W and the past Y_i's, and Z^{i-1} is $Y^{i-1} - X^{i-1}$, (b) follows from $Y_i = X_i + Z_i$ and the fact that $h(X + Z|X) = h(Z|X)$, and (c) follows from the fact Z_i and (W, Y^{i-1}, X^i) are conditionally independent given Z^{i-1}. Continuing the chain of inequalities after dividing by n, we have

$$R_n \le \frac{1}{n}[h(Y^n) - h(Z^n)] + \epsilon_n \le \frac{1}{2n}\log\frac{|K_Y^{(n)}|}{|K_Z^{(n)}|} + \epsilon_n , \tag{10.115}$$

by the entropy maximizing property of the normal. □

We have proved an upper bound on the capacity of the Gaussian channel with feedback in terms of the covariance matrix $K_{X+Z}^{(n)}$. We now derive bounds on the capacity with feedback in terms of $K_X^{(n)}$ and $K_Z^{(n)}$,

which will then be used to derive bounds in terms of the capacity without feedback. For simplicity of notation, we will drop the superscript n in the symbols for covariance matrices.

We first prove a series of lemmas about matrices and determinants.

Lemma 10.6.1: *Let X and Z be n-dimensional random vectors. Then*

$$K_{X+Z} + K_{X-Z} = 2K_X + 2K_Z \tag{10.116}$$

Proof:

$$K_{X+Z} = E(X+Z)(X+Z)^t \tag{10.117}$$

$$= EXX^t + EXZ^t + EZX^t + EZZ^t \tag{10.118}$$

$$= K_X + K_{XZ} + K_{ZX} + K_Z . \tag{10.119}$$

Similarly,

$$K_{X-Z} = K_X - K_{XZ} - K_{ZX} + K_Z . \tag{10.120}$$

Adding these two equations completes the proof. □

Lemma 10.6.2: *For two $n \times n$ positive definite matrices A and B, if $A - B$ is positive definite, then $|A| \geq |B|$.*

Proof: Let $C = A - B$. Since B and C are positive definite, we can consider them as covariance matrices. Consider two independent normal random vectors $\mathbf{X}_1 \sim \mathcal{N}(0, B)$ and $\mathbf{X}_2 \sim \mathcal{N}(0, C)$. Let $\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2$. Then

$$h(\mathbf{Y}) \geq h(\mathbf{Y}|\mathbf{X}_2) \tag{10.121}$$

$$= h(\mathbf{X}_1|\mathbf{X}_2) \tag{10.122}$$

$$= h(\mathbf{X}_1), \tag{10.123}$$

where the inequality follows from the fact that conditioning reduces differential entropy, and the final equality from the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Substituting the expressions for the differential entropies of a normal random variable, we obtain

$$\frac{1}{2}\log(2\pi e)^n|A| > \frac{1}{2}\log(2\pi e)^n|B|\,, \tag{10.124}$$

which is equivalent to the desired lemma. □

Lemma 10.6.3: *For two n-dimensional random vectors X and Z,*

$$|K_{X+Z}| \leq 2^n |K_X + K_Z|\,. \tag{10.125}$$

Proof: From Lemma 10.6.1,

$$2(K_X + K_Z) - K_{X+Z} = K_{X-Z} \geq 0\,, \tag{10.126}$$

where $A \geq 0$ means that A is non-negative definite. Hence, applying Lemma 10.6.2, we have

$$|K_{X+Z}| \leq |2(K_X + K_Z)| = 2^n |K_X + K_Z|\,, \tag{10.127}$$

which is the desired result. □

We are now in a position to prove that feedback increases the capacity of a non-white Gaussian additive noise channel by at most half a bit.

Theorem: 10.6.2:

$$C_{n,FB} \leq C_n + \frac{1}{2} \textit{ bits per transmission} \tag{10.128}$$

Proof: Combining all the lemmas, we obtain

$$C_{n,FB} \leq \max_{\mathrm{tr}(K_X) \leq nP} \frac{1}{2n} \log \frac{|K_Y|}{|K_Z|} \tag{10.129}$$

$$\leq \max_{\mathrm{tr}(K_X) \leq nP} \frac{1}{2n} \log \frac{2^n |K_X + K_Z|}{|K_Z|} \tag{10.130}$$

$$= \max_{\mathrm{tr}(K_X) \leq nP} \frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|} + \frac{1}{2} \tag{10.131}$$

$$\leq C_n + \frac{1}{2} \quad \text{bits per transmission}\,, \tag{10.132}$$

where the inequalities follow from Theorem 10.6.1, Lemma 10.6.3 and the definition of capacity without feedback, respectively. □

SUMMARY OF CHAPTER 10

Maximum entropy: $\max_{EX^2=\alpha} h(X) = \frac{1}{2}\log 2\pi e\alpha$.

The Gaussian channel: $Y_i = X_i + Z_i$, $Z_i \sim \mathcal{N}(0, N)$, power constraint $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$,

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \text{ bits per transmission.} \tag{10.133}$$

Band-limited additive white Gaussian noise channel: Bandwidth W, two-sided power spectral density $N_0/2$, signal power P,

$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \text{ bits per second.} \tag{10.134}$$

Water-filling (k parallel Gaussian channels): $Y_j = X_j + Z_j, j = 1, 2, \ldots, k$, $Z_j \sim \mathcal{N}(0, N_j)$, $\sum_{j=1}^{k} X_j^2 \le P$,

$$C = \sum_{i=1}^{k} \frac{1}{2}\log\left(1 + \frac{(\nu - N_i)^+}{N_i}\right) \tag{10.135}$$

where ν is chosen so that $\sum(\nu - N_i)^+ = P$.

Additive non-white Gaussian noise channel: $Y_i = X_i + Z_i$, $Z^n \sim \mathcal{N}(0, K_Z)$

$$C = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log\left(1 + \frac{(\nu - \lambda_i)^+}{\lambda_i}\right) \tag{10.136}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of K_Z and ν is chosen so that $\sum_i (\nu - \lambda_i)^+ = nP$.

Capacity without feedback:

$$C_n = \max_{\mathrm{tr}(K_X) \le nP} \frac{1}{2n}\log\frac{|K_X + K_Z|}{|K_Z|}. \tag{10.137}$$

Capacity with feedback:

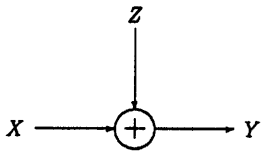
$$C_{n,FB} = \max_{\mathrm{tr}(K_X) \le nP} \frac{1}{2n}\log\frac{|K_{X+Z}|}{|K_Z|}. \tag{10.138}$$

Feedback bound:

$$C_{n,FB} \le C_n + \frac{1}{2}. \tag{10.139}$$

PROBLEMS FOR CHAPTER 10

1. *A mutual information game.* Consider the following channel:



Throughout this problem we shall constrain the signal power

$$EX = 0, \qquad EX^2 = P, \tag{10.140}$$

and the noise power

$$EZ = 0, \qquad EZ^2 = N, \tag{10.141}$$

and assume that X and Z are independent. The channel capacity is given by $I(X; X + Z)$.

Now for the game. The noise player chooses a distribution on Z to minimize $I(X; X + Z)$, while the signal player chooses a distribution on X to maximize $I(X; X + Z)$.

Letting $X^* \sim \mathcal{N}(0, P), Z^* \sim \mathcal{N}(0, N)$, show that X^* and Z^* satisfy the saddlepoint conditions

$$I(X; X + Z^*) \leq I(X^*; X^* + Z^*) \leq I(X^*; X^* + Z). \tag{10.142}$$

Thus

$$\min_Z \max_X I(X; X + Z) = \max_X \min_Z I(X; X + Z) \tag{10.143}$$

$$= \frac{1}{2} \log\left(1 + \frac{P}{N}\right), \tag{10.144}$$

and the game has a value. In particular, a deviation from normal for either player worsens the mutual information from that player's standpoint. Can you discuss the implications of this?

Note: Part of the proof hinges on the entropy power inequality from Section 16.7, which states that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$e^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq e^{\frac{2}{n}h(\mathbf{X})} + e^{\frac{2}{n}h(\mathbf{Y})}. \tag{10.145}$$

2. *A channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

(a) Show $I(X; Y_1, Y_2) = 2I(X; Y_1) - I(Y_1; Y_2)$.

(b) Conclude that the capacity of the channel

$$X \rightarrow \boxed{\quad} \rightarrow (Y_1, Y_2)$$

is less than twice the capacity of the channel

$$X \rightarrow \boxed{\quad} \rightarrow Y_1$$

3. *The two-look Gaussian channel.*

$$X \rightarrow \boxed{\quad} \rightarrow (Y_1, Y_2)$$

Consider the ordinary Shannon Gaussian channel with two correlated looks at X, i.e., $Y = (Y_1, Y_2)$, where

$$Y_1 = X + Z_1 \tag{10.146}$$

$$Y_2 = X + Z_2 \tag{10.147}$$

with a power constraint P on X, and $(Z_1, Z_2) \sim \mathcal{N}_2(\mathbf{0}, K)$, where

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}. \tag{10.148}$$

Find the capacity C for

(a) $\rho = 1$

(b) $\rho = 0$

(c) $\rho = -1$

4. *Parallel channels and waterfilling.* Consider a pair of parallel Gaussian channels, i.e.,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{10.149}$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \tag{10.150}$$

and there is a power constraint $E(X_1^2 + X_2^2) \leq 2P$. Assume that $\sigma_1^2 > \sigma_2^2$. At what power does the channel stop behaving like a single channel with noise variance σ_2^2, and begin behaving like a pair of channels?

HISTORICAL NOTES

The Gaussian channel was first analyzed by Shannon in his original paper [238]. The water-filling solution to the capacity of the colored noise Gaussian channel

was developed by Holsinger [135]. Pinsker [210] and Ebert [94] showed that feedback at most doubles the capacity of a non-white Gaussian channel; a simple proof can be found in Cover and Pombra [76]. Cover and Pombra also show that feedback increases the capacity of the non-white Gaussian channel by at most half a bit.