# Appendix B Information theory from first principles

This appendix discusses the information theory behind the capacity expressions used in the book. Section 8.3.4 is the only part of the book that supposes an understanding of the material in this appendix. More in-depth and broader expositions of information theory can be found in standard texts such as [26] and [43].

## B.1 Discrete memoryless channels

Although the transmitted and received signals are continuous-valued in most of the channels we considered in this book, the heart of the communication problem is *discrete* in nature: the transmitter sends one out of a finite number of codewords and the receiver would like to figure out which codeword is transmitted. Thus, to focus on the essence of the problem, we first consider channels with discrete input and output, so-called *discrete memoryless channels* (DMCs).

Both the input $x[m]$ and the output $y[m]$ of a DMC lie in finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. (These sets are called the input and output alphabets of the channel respectively.) The statistics of the channel are described by conditional probabilities $\{p(j|i)\}_{i\in\mathcal{X}, j\in\mathcal{Y}}$. These are also called *transition probabilities*. Given an input sequence $\mathbf{x} = (x[1], \ldots, x[N])$, the probability of observing an output sequence $\mathbf{y} = (y[1], \ldots, y[N])$ is given by[1]

$$p(\mathbf{y}|\mathbf{x}) = \prod_{m=1}^{N} p(y[m]|x[m]). \tag{B.1}$$

The interpretation is that the channel noise corrupts the input symbols independently (hence the term *memoryless*).

[1] This formula is only valid when there is no feedback from the receiver to the transmitter, i.e., the input is not a function of past outputs. This we assume throughout.

**Example B.1 Binary symmetric channel**
The binary symmetric channel has binary input and binary output ($\mathcal{X} = \mathcal{Y} = \{0, 1\}$). The transition probabilities are $p(0|1) = p(1|0) = \epsilon$, $p(0|0) = p(1|1) = 1 - \epsilon$. A 0 and a 1 are both flipped with probability $\epsilon$. See Figure B.1(a).

**Example B.2 Binary erasure channel**
The binary erasure channel has binary input and ternary output ($\mathcal{X} = \{0, 1\}, \mathcal{Y} = \{0, 1, e\}$). The transition probabilities are $p(0|0) = p(1|1) = 1 - \epsilon$, $p(e|0) = p(e|1) = \epsilon$. Here, symbols cannot be flipped but can be erased. See Figure B.1(b).

An abstraction of the communication system is shown in Figure B.2. The sender has one out of several equally likely messages it wants to transmit to the receiver. To convey the information, it uses a codebook $\mathcal{C}$ of block length $N$ and size $|\mathcal{C}|$, where $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{C}|}\}$ and $\mathbf{x}_i$ are the codewords. To transmit the $i$th message, the codeword $\mathbf{x}_i$ is sent across the noisy channel. Based on the received vector $\mathbf{y}$, the decoder generates an estimate $\hat{i}$ of the correct message. The error probability is $p_e = \mathbb{P}\{\hat{i} \neq i\}$. We will assume that the maximum likelihood (ML) decoder is used, since it minimizes the error probability for a given code. Since we are transmitting one of $|\mathcal{C}|$ messages, the number of bits conveyed is $\log|\mathcal{C}|$. Since the block length of the code is $N$, the rate of the code is $R = \frac{1}{N}\log|\mathcal{C}|$ bits per unit time. The data rate $R$ and the ML error probability $p_e$ are the two key performance measures of a code.

$$R = \tfrac{1}{N}\log|\mathcal{C}|. \tag{B.2}$$

$$p_e = \mathbb{P}\{\hat{i} \neq i\}. \tag{B.3}$$

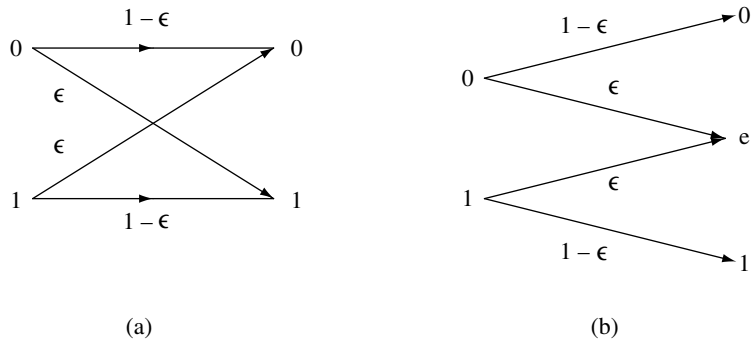


**Figure B.1** Examples of discrete memoryless channels: (a) binary symmetric channel; (b) binary erasure channel.

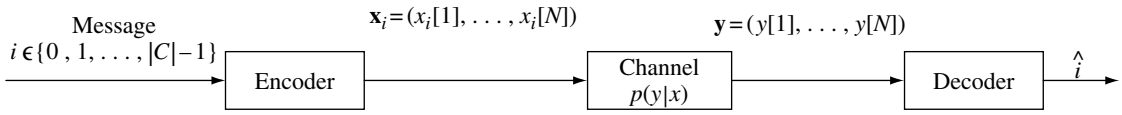


**Figure B.2** Abstraction of a communication system à la Shannon.

Information is said to be *communicated reliably* at rate $R$ if for every $\delta > 0$, one can find a code of rate $R$ and block length $N$ such that the error probability $p_e < \delta$. The capacity $C$ of the channel is the maximum rate for which reliable communication is possible.

Note that the key feature of this definition is that one is allowed to code over arbitrarily large block length $N$. Since there is noise in the channel, it is clear that the error probability cannot be made arbitrarily small if the block length is fixed a priori. (Recall the AWGN example in Section 5.1.) Only when the code is over long block length is there hope that one can rely on some kind of law of large numbers to average out the random effect of the noise. Still, it is not clear a priori whether a non-zero reliable information rate can be achieved in general.

Shannon showed not only that $C > 0$ for most channels of interest but also gave a simple way to compute $C$ as a function of $\{p(y|x)\}$. To explain this we have to first define a few statistical measures.

## B.2 Entropy, conditional entropy and mutual information

Let $x$ be a discrete random variable taking on values in $\mathcal{X}$ and with a probability mass function $p_x$. Define the *entropy* of $x$ to be[2]

$$H(x) := \sum_{i \in \mathcal{X}} p_x(i) \log(1/p_x(i)). \tag{B.4}$$

This can be interpreted as a measure of the amount of uncertainty associated with the random variable $x$. The entropy $H(x)$ is always non-negative and equal to zero if and only if $x$ is deterministic. If $x$ can take on $K$ values, then it can be shown that the entropy is maximized when $x$ is uniformly distributed on these $K$ values, in which case $H(x) = \log K$ (see Exercise B.1).

**Example B.3 Binary entropy**
The entropy of a binary-valued random variable $x$ which takes on the values with probabilities $p$ and $1-p$ is

$$H(p) := -p \log p - (1-p) \log(1-p). \tag{B.5}$$

[2] In this book, all logarithms are taken to the base 2 unless specified otherwise.

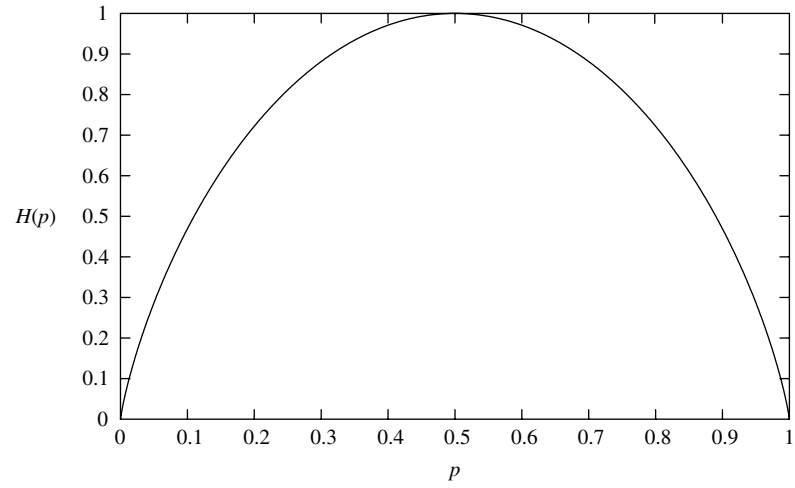


**Figure B.3** The binary entropy function.

The function $H(\cdot)$ is called the *binary entropy function*, and is plotted in Figure B.3. It attains its maximum value of 1 at $p = 1/2$, and is zero when $p = 0$ or $p = 1$. Note that we never mentioned the actual *values* $x$ takes on; the amount of uncertainty depends only on the probabilities.

Let us now consider two random variables $x$ and $y$. The joint entropy of $x$ and $y$ is defined to be

$$H(x, y) := \sum_{i \in \mathcal{X}, j \in \mathcal{Y}} p_{x,y}(i, j) \log(1/p_{x,y}(i, j)). \tag{B.6}$$

The entropy of $x$ conditional on $y = j$ is naturally defined to be

$$H(x|y = j) := \sum_{i \in \mathcal{X}} p_{x|y}(i|j) \log(1/p_{x|y}(i|j)). \tag{B.7}$$

This can be interpreted as the amount of uncertainty left in $x$ after observing that $y = j$. The conditional entropy of $x$ given $y$ is the expectation of this quantity, averaged over all possible values of $y$:

$$H(x|y) := \sum_{j \in \mathcal{Y}} p_y(j) H(x|y = j) = \sum_{i \in \mathcal{X}, j \in \mathcal{Y}} p_{x,y}(i, j) \log(1/p_{x|y}(i|j)). \tag{B.8}$$

The quantity $H(x|y)$ can be interpreted as the average amount of uncertainty left in $x$ after observing $y$. Note that

$$H(x, y) = H(x) + H(y|x) = H(y) + H(x|y). \tag{B.9}$$

This has a natural interpretation: the total uncertainty in $x$ and $y$ is the sum of the uncertainty in $x$ plus the uncertainty in $y$ conditional on $x$. This is called the *chain rule for entropies.* In particular, if $x$ and $y$ are independent, $H(x|y) = H(x)$ and hence $H(x, y) = H(x) + H(y)$. One would expect that conditioning reduces uncertainty, and in fact it can be shown that

$$H(x|y) \leq H(x), \tag{B.10}$$

with equality if and only if $x$ and $y$ are independent. (See Exercise B.2.) Hence,

$$H(x, y) = H(x) + H(y|x) \leq H(x) + H(y), \tag{B.11}$$

with equality if and only if $x$ and $y$ are independent.

The quantity $H(x) - H(x|y)$ is of special significance to the communication problem at hand. Since $H(x)$ is the amount of uncertainty in $x$ before observing $y$, this quantity can be interpreted as the *reduction* in uncertainty of $x$ from the observation of $y$, i.e., the amount of information in $y$ about $x$. Similarly, $H(y) - H(y|x)$ can be interpreted as the reduction in uncertainty of $y$ from the observation of $x$. Note that

$$H(y) - H(y|x) = H(y) + H(x) - H(x, y) = H(x) - H(x|y). \tag{B.12}$$

So if one defines

$$I(x; y) := H(y) - H(y|x) = H(x) - H(x|y), \tag{B.13}$$

then this quantity is symmetric in the random variables $x$ and $y$. $I(x; y)$ is called the *mutual information* between $x$ and $y$. A consequence of (B.10) is that the mutual information $I(x; y)$ is a non-negative quantity, and equal to zero if and only if $x$ and $y$ are independent.

We have defined the mutual information between scalar random variables, but the definition extends naturally to random vectors. For example, $I(x_1, x_2; y)$ should be interpreted as the mutual information between the random vector $(x_1, x_2)$ and $y$, i.e., $I(x_1, x_2; y) = H(x_1, x_2) - H(x_1, x_2|y)$. One can also define a notion of conditional mutual information:

$$I(x; y|z) := H(x|z) - H(x|y, z). \tag{B.14}$$

Note that since

$$H(x|z) = \sum_k p_z(k) H(x|z = k), \tag{B.15}$$

and

$$H(x|y,z) = \sum_k p_z(k) H(x|y, z=k), \tag{B.16}$$

it follows that

$$I(x;y|z) = \sum_k p_z(k) I(x;y|z=k). \tag{B.17}$$

Given three random variables $x_1$, $x_2$ and $y$, observe that

$$\begin{aligned} I(x_1,x_2;y) &= H(x_1,x_2) - H(x_1,x_2|y) \\ &= H(x_1) + H(x_2|x_1) - [H(x_1|y) + H(x_2|x_1,y)] \\ &= I(x_1;y) + I(x_2;y|x_1). \end{aligned}$$

This is the *chain rule for mutual information*:

$$\boxed{I(x_1,x_2;y) = I(x_1;y) + I(x_2;y|x_1).} \tag{B.18}$$

In words: the information that $x_1$ and $x_2$ jointly provide about $y$ is equal to the sum of the information $x_1$ provides about $y$ plus the additional information $x_2$ provides about $y$ after observing $x_1$. This fact is very useful in Chapters 7 to 10.

## B.3 Noisy channel coding theorem

Let us now go back to the communication problem shown in Figure B.2. We convey one of $|\mathcal{C}|$ equally likely messages by mapping it to its $N$-length codeword in the code $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{C}|}\}$. The input to the channel is then an $N$-dimensional random vector $\mathbf{x}$, uniformly distributed on the codewords of $\mathcal{C}$. The output of the channel is another $N$-dimensional vector $\mathbf{y}$.

### B.3.1 Reliable communication and conditional entropy

To decode the transmitted message correctly with high probability, it is clear that the conditional entropy $H(\mathbf{x}|\mathbf{y})$ has to be close to zero[3]. Otherwise, there is too much uncertainty in the input, given the output, to figure out what the right message is. Now,

$$H(\mathbf{x}|\mathbf{y}) = H(\mathbf{x}) - I(\mathbf{x};\mathbf{y}), \tag{B.19}$$

[3] This statement can be made precise in the regime of large block lengths using Faro's inequality.

i.e., the uncertainty in $\mathbf{x}$ subtracting the reduction in uncertainty in $\mathbf{x}$ by observing $\mathbf{y}$. The entropy $H(\mathbf{x})$ is equal to $\log|\mathcal{C}| = NR$, where $R$ is the data rate. For reliable communication, $H(\mathbf{x}|\mathbf{y}) \approx 0$, which implies

$$R \approx \frac{1}{N} I(\mathbf{x}; \mathbf{y}). \tag{B.20}$$

Intuitively: for reliable communication, the *rate of flow* of mutual information across the channel should match the rate at which information is generated. Now, the mutual information depends on the distribution of the random input $\mathbf{x}$, and this distribution is in turn a function of the code $\mathcal{C}$. By optimizing over all codes, we get an upper bound on the reliable rate of communication:

$$\max_{\mathcal{C}} \frac{1}{N} I(\mathbf{x}; \mathbf{y}). \tag{B.21}$$

### B.3.2 A simple upper bound

The optimization problem (B.21) is a high-dimensional combinatorial one and is difficult to solve. Observe that since the input vector $\mathbf{x}$ is uniformly distributed on the codewords of $\mathcal{C}$, the optimization in (B.21) is over only a subset of possible input distributions. We can derive a further upper bound by relaxing the feasible set and allowing the optimization to be over *all* input distributions:

$$\bar{C} := \max_{p_{\mathbf{x}}} \frac{1}{N} I(\mathbf{x}; \mathbf{y}), \tag{B.22}$$

Now,

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}) &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) && \text{(B.23)} \\ &\leq \sum_{m=1}^{N} H(y[m]) - H(\mathbf{y}|\mathbf{x}) && \text{(B.24)} \\ &= \sum_{m=1}^{N} H(y[m]) - \sum_{m=1}^{N} H(y[m]|x[m]) && \text{(B.25)} \\ &= \sum_{m=1}^{N} I(x[m]; y[m]). && \text{(B.26)} \end{aligned}$$

The inequality in (B.24) follows from (B.11) and the equality in (B.25) comes from the memoryless property of the channel. Equality in (B.24) is attained if the output symbols are independent over time, and one way to achieve this is to make the inputs independent over time. Hence,

$$\bar{C} = \frac{1}{N} \sum_{m=1}^{N} \max_{p_{x[m]}} I(x[m]; y[m]) = \max_{p_{x[1]}} I(x[1]; y[1]). \tag{B.27}$$

Thus, the optimizing problem over input distributions on the $N$-length block reduces to an optimization problem over input distributions on single symbols.

### B.3.3 Achieving the upper bound

To achieve this upper bound $\bar{C}$, one has to find a code whose mutual information $I(\mathbf{x};\mathbf{y})/N$ per symbol is close to $\bar{C}$ *and* such that (B.20) is satisfied. A priori it is unclear if such a code exists at all. The cornerstone result of information theory, due to Shannon, is that indeed such codes exist *if the block length N is chosen sufficiently large*.

**Theorem B.1** *(Noisy channel coding theorem [109]) Consider a discrete memoryless channel with input symbol x and output symbol y. The capacity of the channel is*

$$C = \max_{p_x} I(x;y). \tag{B.28}$$

Shannon's proof of the existence of optimal codes is through a randomization argument. Given any symbol input distribution $p_x$, we can randomly generate a code $\mathcal{C}$ with rate $R$ by choosing each symbol in each codeword independently according to $p_x$. The main result is that with the rate as in (B.20), the code with large block length $N$ satisfies, with high probability,

$$\frac{1}{N} I(\mathbf{x};\mathbf{y}) \approx I(x;y). \tag{B.29}$$

In other words, reliable communication is possible at the rate of $I(x;y)$. In particular, by choosing codewords according to the distribution $p_x^*$ that maximizes $I(x;y)$, the maximum reliable rate is achieved. The smaller the desired error probability, the larger the block length $N$ has to be for the law of large numbers to average out the effect of the random noise in the channel as well as the effect of the random choice of the code. We will not go into the details of the derivation of the noisy channel coding theorem in this book, although the sphere-packing argument for the AWGN channel in Section B.5 suggests that this result is plausible. More details can be found in standard information theory texts such as [26].

The maximization in (B.28) is over all distributions of the input random variable $x$. Note that the input distribution together with the channel transition probabilities specifies a joint distribution on $x$ and $y$. This determines the

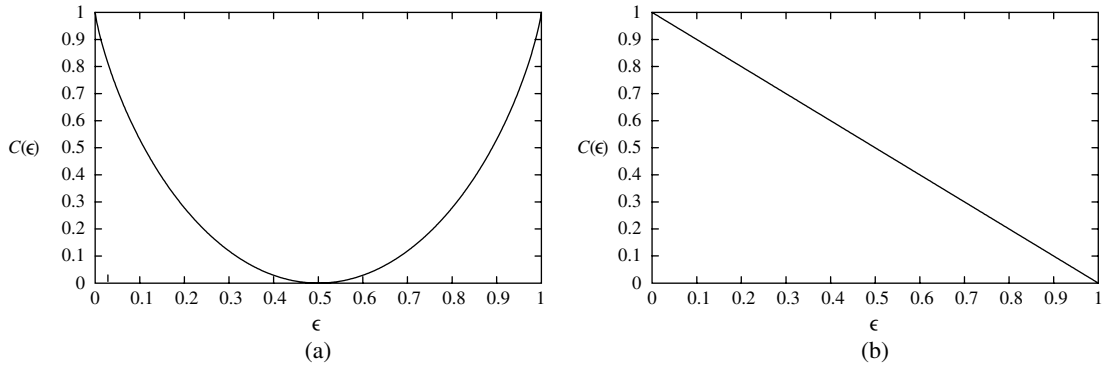


**Figure B.4** The capacity of (a) the binary symmetric channel and (b) the binary erasure channel.

value of $I(x; y)$. The maximization is over all possible input distribution It can be shown that the mutual information $I(x; y)$ is a concave function of the input probabilities and hence the input maximization is a convex optimization problem, which can be solved very efficiently. Sometimes one can even appeal to symmetry to obtain the optimal distribution in closed form.

**Example B.4 Binary symmetric channel**

The capacity of the binary symmetric channel with crossover probability $\epsilon$ is

$$\begin{aligned} C &= \max_{p_x} H(y) - H(y|x) \\ &= \max_{p_x} H(y) - H(\epsilon) \\ &= 1 - H(\epsilon) \text{ bits per channel use} \end{aligned} \tag{B.30}$$

where $H(\epsilon)$ is the binary entropy function (B.5). The maximum is achieved by choosing $x$ to be uniform so that the output $y$ is also uniform. The capacity is plotted in Figure B.4. It is 1 when $\epsilon = 0$ or 1, and 0 when $\epsilon = 1/2$.

Note that since a fraction $\epsilon$ of the symbols are flipped in the long run, one may think that the capacity of the channel is $1 - \epsilon$ bits per channel use, the fraction of symbols that get through unflipped. However, this is too naive since the receiver does not know which symbols are flipped and which are correct. Indeed, when $\epsilon = 1/2$, the input and output are independent and there is no way we can get any information across the channel. The expression (B.30) gives the correct answer.

**Example B.5 Binary erasure channel**
The optimal input distribution for the binary symmetric channel is uniform because of the symmetry in the channel. Similar symmetry exists in the binary erasure channel and the optimal input distribution is uniform too. The capacity of the channel with erasure probability $\epsilon$ can be calculated to be

$$C = 1 - \epsilon \text{ bits per channel use.} \tag{B.31}$$

In the binary symmetric channel, the receiver does not know which symbols are flipped. In the erasure channel, on the other hand, the receiver knows exactly which symbols are erased. If the *transmitter* also knows that information, then it can send bits only when the channel is not erased and a long-term throughput of $1 - \epsilon$ bits per channel use is achieved. What the capacity result says is that no such feedback information is necessary; (forward) coding is sufficient to get this rate reliably.

## B.3.4 Operational interpretation

There is a common misconception that needs to be pointed out. In solving the input distribution optimization problem (B.22) for the capacity $C$, it was remarked that, at the optimal solution, the outputs $y[m]$ should be independent, and one way to achieve this is for the inputs $x[m]$ to be independent. Does that imply no coding is needed to achieve capacity? For example, in the binary symmetric channel, the optimal input yields i.i.d. equally likely symbols; does it mean then that we can send equally likely information bits raw across the channel and still achieve capacity?

Of course not: to get very small error probability one needs to code over many symbols. The fallacy of the above argument is that reliable communication *cannot* be achieved at *exactly* the rate $C$ and when the outputs are *exactly* independent. Indeed, when the outputs and inputs are i.i.d.,

$$H(\mathbf{x}|\mathbf{y}) = \sum_{m=1}^{N} H(x[m]|y[m]) = NH(x[m]|y[m]), \tag{B.32}$$

and there is a lot of uncertainty in the input given the output: the communication is hardly reliable. But once one shoots for a rate *strictly* less than $C$, no matter how close, the coding theorem guarantees that reliable communication is possible. The mutual information $I(\mathbf{x}; \mathbf{y})/N$ per symbol is close to $C$, the outputs $y[m]$ are *almost* independent, but now the conditional entropy $H(\mathbf{x}|\mathbf{y})$ is reduced abruptly to (close to) zero since reliable decoding is possible. But to achieve this performance, *coding* is crucial; indeed the entropy per input symbol is close to $I(\mathbf{x}; \mathbf{y})/N$, less than $H(x[m])$ under uncoded transmission.

For the binary symmetric channel, the entropy per coded symbol is $1 - H(\epsilon)$, rather than 1 for uncoded symbols.

The bottom line is that while the *value* of the input optimization problem (B.22) has operational meaning as the maximum rate of reliable communication, it is incorrect to interpret the i.i.d. input distribution which attains that value as the statistics of the input symbols which achieve reliable communication. Coding is *always* needed to achieve capacity. What *is* true, however, is that if we randomly pick the codewords according to the i.i.d. input distribution, the resulting code is very likely to be good. But this is totally different from sending uncoded symbols.

## B.4 Formal derivation of AWGN capacity

We can now apply the methodology developed in the previous sections to formally derive the capacity of the AWGN channel.

### B.4.1 Analog memoryless channels

So far we have focused on channels with discrete-valued input and output symbols. To derive the capacity of the AWGN channel, we need to extend the framework to analog channels with continuous-valued input and output. There is no conceptual difficulty in this extension. In particular, Theorem B.1 can be generalized to such analog channels.[4] The definitions of entropy and conditional entropy, however, have to be modified appropriately.

For a continuous random variable $x$ with pdf $f_x$, define the *differential entropy* of $x$ as

$$h(x) := \int_{-\infty}^{\infty} f_x(u) \log(1/f_x(u)) \mathrm{d}u. \tag{B.33}$$

Similarly, the conditional differential entropy of $x$ given $y$ is defined as

$$h(x|y) := \int_{-\infty}^{\infty} f_{x,y}(u, v) \log(1/f_{x|y}(u|v)) \mathrm{d}u \mathrm{d}v. \tag{B.34}$$

The mutual information is again defined as

$$I(x; y) := h(x) - h(x|y). \tag{B.35}$$

[4] Although the underlying channel is analog, the communication process is still digital. This means that discrete symbols will still be used in the encoding. By formulating the communication problem directly in terms of the underlying analog channel, this means we are not constraining ourselves to using a particular symbol constellation (for example, 2-PAM or QPSK) a priori.

Observe that the chain rules for entropy and for mutual information extend readily to the continuous-valued case. The capacity of the continuous-valued channel can be shown to be

$$C = \max_{f_x} I(x; y). \tag{B.36}$$

This result can be proved by discretizing the continuous-valued input and output of the channel, approximating it by discrete memoryless channels with increasing alphabet sizes, and taking limits appropriately.

For many channels, it is common to have a *cost constraint* on the transmitted codewords. Given a cost function $c : \mathcal{X} \to \Re$ defined on the input symbols, a cost constraint on the codewords can be defined: we require that every codeword $\mathbf{x}_n$ in the codebook must satisfy

$$\frac{1}{N} \sum_{m=1}^{N} c(x_n[m]) \leq A. \tag{B.37}$$

One can then ask: what is the maximum rate of reliable communication subject to this constraint on the codewords? The answer turns out to be

$$C = \max_{f_x : E[c(x)] \leq A} I(x; y). \tag{B.38}$$

## B.4.2 Derivation of AWGN capacity

We can now apply this result to derive the capacity of the power-constrained (real) AWGN channel:

$$y = x + w, \tag{B.39}$$

The cost function is $c(x) = x^2$. The differential entropy of a $\mathcal{N}(\mu, \sigma^2)$ random variable $w$ can be calculated to be

$$h(w) = \frac{1}{2} \log(2\pi e \sigma^2). \tag{B.40}$$

Not surprisingly, $h(w)$ does not depend on the mean $\mu$ of $W$: differential entropies are invariant to translations of the pdf. Thus, conditional on the input $x$ of the Gaussian channel, the differential entropy $h(y|x)$ of the output $y$ is just $(1/2)\log(2\pi \mathrm{e}\sigma^2)$. The mutual information for the Gaussian channel is, therefore,

$$I(x; y) = h(y) - h(y|x) = h(y) - \frac{1}{2} \log(2\pi \mathrm{e}\sigma^2). \tag{B.41}$$

The computation of the capacity

$$C = \max_{f_x : E[x^2] \leq P} I(x; y) \tag{B.42}$$

is now reduced to finding the input distribution on $x$ to maximize $h(y)$ subject to a second moment constraint on $x$. To solve this problem, we use a key fact about Gaussian random variables: they are differential entropy maximizers. More precisely, given a constraint $E[u^2] \leq A$ on a random variable $u$, the distribution $u$ is $\mathcal{N}(0, A)$ maximizes the differential entropy $h(u)$. (See Exercise B.6 for a proof of this fact.) Applying this to our problem, we see that the second moment constraint of $P$ on $x$ translates into a second moment constraint of $P + \sigma^2$ on $y$. Thus, $h(y)$ is maximized when $y$ is $\mathcal{N}(0, P + \sigma^2)$, which is achieved by choosing $x$ to be $\mathcal{N}(0, P)$. Thus, the capacity of the Gaussian channel is

$$C = \frac{1}{2}\log(2\pi e(P+\sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right), \tag{B.43}$$

agreeing with the result obtained via the heuristic sphere-packing derivation in Section 5.1. A capacity-achieving code can be obtained by choosing each component of each codeword i.i.d. $\mathcal{N}(0, P)$. Each codeword is therefore isotropically distributed, and, by the law of large numbers, with high probability lies near the surface of the sphere of radius $\sqrt{NP}$. Since in high dimensions most of the volume of a sphere is near its surface, this is effectively the same as picking each codeword uniformly from the sphere.

Now consider a *complex* baseband AWGN channel:

$$y = x + w \tag{B.44}$$

where $w$ is $\mathcal{CN}(0, N_0)$. There is an average power constraint of $P$ per (complex) symbol. One way to derive the capacity of this channel is to think of each use of the complex channel as two uses of a real AWGN channel, with $\mathsf{SNR} = (P/2)/(N_0/2) = P/N_0$. Hence, the capacity of the channel is

$$\frac{1}{2}\log\left(1+\frac{P}{N_0}\right) \text{ bits per real dimension,} \tag{B.45}$$

or

$$\log\left(1+\frac{P}{N_0}\right) \text{ bits per complex dimension.} \tag{B.46}$$

Alternatively we may just as well work directly with the complex channel and the associated complex random variables. This will be useful when we deal with other more complicated wireless channel models later on. To this end, one can think of the differential entropy of a complex random variable $x$ as that of a real random vector $(\Re(x), \Im(x))$. Hence, if $w$ is $\mathcal{CN}(0, N_0)$, $h(w) = h(\Re(w)) + h(\Im(w)) = \log(\pi e N_0)$. The mutual information $I(x; y)$ of the complex AWGN channel $y = x + w$ is then

$$I(x; y) = h(y) - \log(\pi e N_0). \tag{B.47}$$

With a power constraint $E[|x|^2] \leq P$ on the complex input $x$, $y$ is constrained to satisfy $E[|y|^2] \leq P + N_0$. Here, we use an important fact: among all complex random variables, the *circular symmetric* Gaussian random variable maximizes the differential entropy for a given second moment constraint. (See Exercise B.7.) Hence, the capacity of the complex Gaussian channel is

$$C = \log(\pi e(P + N_0)) - \log(\pi e N_0) = \log\left(1 + \frac{P}{N_0}\right), \tag{B.48}$$

which is the same as Eq. (5.11).

## B.5 Sphere-packing interpretation

In this section we consider a more precise version of the heuristic sphere-packing argument in Section 5.1 for the capacity of the real AWGN channel. Furthermore, we outline how the capacity as predicted by the sphere-packing argument can be achieved. The material here is particularly useful when we discuss precoding in Chapter 10.

### B.5.1 Upper bound

Consider transmissions over a block of $N$ symbols, where $N$ is large. Suppose we use a code $\mathcal{C}$ consisting of $|\mathcal{C}|$ equally likely codewords $\{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{C}|}\}$. By the law of large numbers, the $N$-dimensional received vector $\mathbf{y} = \mathbf{x} + \mathbf{w}$ will with high probability lie approximately[5] within a $y$-sphere of radius $\sqrt{N(P + \sigma^2)}$, so without loss of generality we need only to focus on what happens inside this $y$-sphere. Let $\mathcal{D}_i$ be the part of the maximum-likelihood decision region for $\mathbf{x}_i$ within the $y$-sphere. The sum of the volumes of the $\mathcal{D}_i$ is equal to $V_y$, the volume of the $y$-sphere. Given this total volume, it can be shown, using the spherical symmetry of the Gaussian noise distribution, that the error probability is lower bounded by the (hypothetical) case when the $\mathcal{D}_i$ are all perfect spheres of equal volume $V_y/|\mathcal{C}|$. But by the law of large numbers, the received vector $\mathbf{y}$ lies near the surface of a *noise sphere* of radius $\sqrt{N\sigma^2}$ around the transmitted codeword. Thus, for reliable communication, $V_y/|\mathcal{C}|$ should be no smaller than the volume $V_w$ of this noise sphere, otherwise even in the ideal case when the decision regions are all spheres of equal volume, the error probability will still be very large. Hence, the number of

[5] To make this and other statements in this section completely rigorous, appropriate $\delta$ and $\varepsilon$ have to be added.

codewords is at most equal to the ratio of the volume of the $y$-sphere to that of a noise sphere:

$$\frac{V_y}{V_w} = \frac{\left[\sqrt{N(P+\sigma^2)}\right]^N}{\left[\sqrt{N\sigma^2}\right]^N}.$$

(See Exercise B.10(3) for an explicit expression of the volume of an $N$-dimensional sphere of a given radius.) Hence, the number of bits per symbol time that can be reliably communicated is at most

$$\frac{1}{N}\log\left(\frac{\left[\sqrt{N(P+\sigma^2)}\right]^N}{\left[\sqrt{N\sigma^2}\right]^N}\right) = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right). \tag{B.49}$$

The geometric picture is in Figure B.5.

## B.5.2 Achievability

The above argument only gives an upper bound on the rate of reliable communication. The question is: can we design codes that can perform this well?

Let us use a codebook $\mathcal{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{C}|}\}$ such that the $N$-dimensional codewords lie in the sphere of radius $\sqrt{NP}$ (the "$x$-sphere") and thus satisfy the power constraint. The optimal detector is the maximum likelihood nearest neighbor rule. For reasons that will be apparent shortly, we instead consider the following suboptimal detector: given the received vector $\mathbf{y}$, decode to the codeword $\mathbf{x}_i$ nearest to $\alpha\mathbf{y}$, where $\alpha := P/(P+\sigma^2)$.

It is not easy to design a specific code that yields good performance, but suppose we just randomly and independently choose each codeword to be

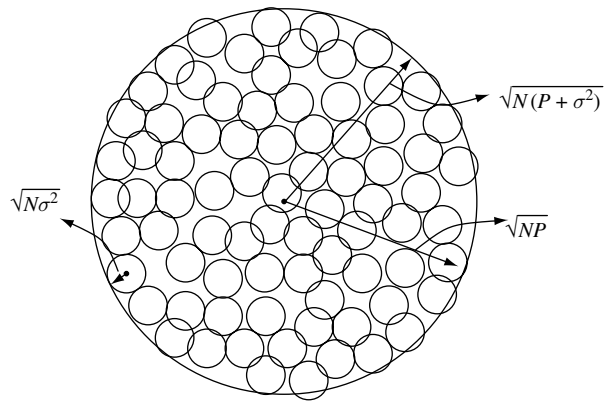


**Figure B.5** The number of noise spheres that can be packed into the $y$-sphere yields the maximum number of codewords that can be reliably distinguished.

uniformly distributed in the sphere[6]. In high dimensions, most of the volume of the sphere lies near its surface, so in fact the codewords will with high probability lie near the surface of the $x$-sphere.

What is the performance of this random code? Suppose the transmitted codeword is $\mathbf{x}_1$. By the law of large numbers again,

$$\begin{aligned} \|\alpha\mathbf{y} - \mathbf{x}_1\|^2 &= \|\alpha\mathbf{w} + (\alpha - 1)\mathbf{x}_1\|^2, \\ &\approx \alpha^2 N\sigma^2 + (\alpha - 1)^2 NP, \\ &= N\frac{P\sigma^2}{P + \sigma^2}, \end{aligned}$$

i.e., the transmitted codeword lies inside an *uncertainty* sphere of radius $\sqrt{NP\sigma^2/(P+\sigma^2)}$ around the vector $\alpha\mathbf{y}$. Thus, as long as all the other codewords lie outside this uncertainty sphere, then the receiver will be able to decode correctly (Figure B.6). The probability that the random codeword $\mathbf{x}_i$ $(i \neq 1)$ lies inside the uncertainty sphere is equal to the ratio of the volume of the uncertainty sphere to that of the $x$-sphere:

$$p = \frac{\left(\sqrt{NP\sigma^2/(P+\sigma^2)}\right)^N}{(\sqrt{NP})^N} = \left(\frac{\sigma^2}{P+\sigma^2}\right)^{\frac{N}{2}}. \tag{B.50}$$

By the union bound, the probability that *any* of the codewords $(\mathbf{x}_2, \ldots, \mathbf{x}_{|\mathcal{C}|})$ lie inside the uncertainty sphere is bounded by $(|\mathcal{C}| - 1)p$. Thus, as long as the number of codewords is much smaller than $1/p$, then the probability of error is small (in particular, we can take the number of codewords $|\mathcal{C}|$ to be

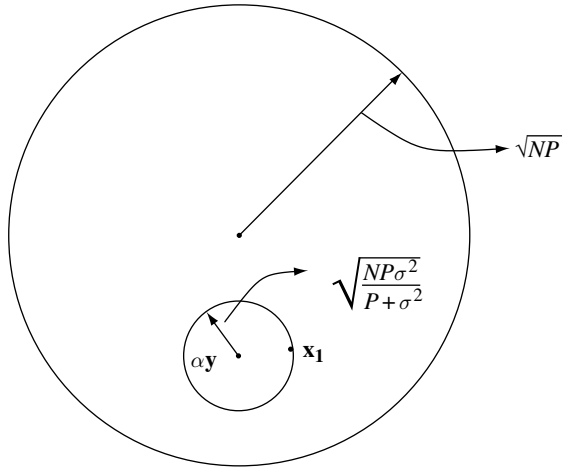


**Figure B.6** The ratio of the volume of the uncertainty sphere to that of the $x$-sphere yields the probability that a given random codeword lies inside the uncertainty sphere. The inverse of this probability yields a lower bound on the number of codewords that can be reliably distinguished.

[6] Randomly and independently choosing each codeword to have i.i.d. $\mathcal{N}(0, P)$ components would work too but the argument is more complex.

$1/pN$). In terms of the data rate $R$ bits per symbol time, this means that as long as

$$R = \frac{\log |\mathcal{C}|}{N} = \frac{\log 1/p}{N} - \frac{\log N}{N} < \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right),$$

then reliable communication is possible.

Both the upper bound and the achievability arguments are based on calculating the ratio of volumes of spheres. The ratio is the same in both cases, but the spheres involved are different. The sphere-packing picture in Figure B.5 corresponds to the following decomposition of the capacity expression:

$$\frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) = I(x; y) = h(y) - h(y|x), \tag{B.51}$$

with the volume of the $y$-sphere proportional to $2^{Nh(y)}$ and the volume of the noise sphere proportional to $2^{Nh(y|x)}$. The picture in Figure B.6, on the other hand, corresponds to the decomposition:

$$\frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) = I(x; y) = h(x) - h(x|y), \tag{B.52}$$

with the volume of the $x$-sphere proportional to $2^{Nh(x)}$. Conditional on $y$, $x$ is $N(\alpha y, \sigma^2_{\text{mmse}})$, where $\alpha = P/(P+\sigma^2)$ is the coefficient of the MMSE estimator of $x$ given $y$, and

$$\sigma^2_{\text{mmse}} = \frac{P\sigma^2}{P + \sigma^2},$$

is the MMSE estimation error. The radius of the uncertainty sphere considered above is $\sqrt{N\sigma^2_{\text{mmse}}}$ and its volume is proportional to $2^{Nh(x|y)}$. In fact the proposed receiver, which finds the nearest codeword to $\alpha\mathbf{y}$, is motivated precisely by this decomposition. In this picture, then, the AWGN capacity formula is being interpreted in terms of the number of MMSE error spheres that can be packed inside the $x$-sphere.

## B.6 Time-invariant parallel channel

Consider the parallel channel (cf. (5.33):

$$\tilde{y}_n[i] = \tilde{h}_n\tilde{d}_n[i] + \tilde{w}_n[i] \qquad n = 0, 1, \ldots, N_c - 1, \tag{B.53}$$

subject to an average power per sub-carrier constraint of $P$ (cf. (5.37)):

$$E[|\tilde{\mathbf{d}}[i]|^2] \le N_c P. \tag{B.54}$$

The capacity in bits per symbol is

$$C_{N_c} = \max_{\mathbb{E}[\|\tilde{\mathbf{d}}\|^2] \le N_c P} I(\tilde{\mathbf{d}}; \tilde{\mathbf{y}}). \tag{B.55}$$

Now

$$\begin{aligned} I(\tilde{\mathbf{d}}; \tilde{\mathbf{y}}) &= h(\tilde{\mathbf{y}}) - h(\tilde{\mathbf{y}}|\tilde{\mathbf{d}}) && \text{(B.56)} \\ &\le \sum_{n=0}^{N_c-1} \left( h(\tilde{y}_n) - h(\tilde{y}_n|\tilde{d}_n) \right) && \text{(B.57)} \\ &\le \sum_{n=0}^{N_c-1} \log\left(1 + \frac{P_n|\tilde{h}_n|^2}{N_0}\right). && \text{(B.58)} \end{aligned}$$

The inequality in (B.57) is from (B.11) and $P_n$ denotes the variance of $\tilde{d}_n$ in (B.58). Equality in (B.57) is achieved when $\tilde{d}_n, n = 0, \ldots, N_c - 1$, are independent. Equality is achieved in (B.58) when $\tilde{d}_n$ is $\mathcal{CN}(0, P_n), n = 0, \ldots, N_c - 1$. Thus, computing the capacity in (B.55) is reduced to a power allocation problem (by identifying the variance of $\tilde{d}_n$ with the power allocated to the $n$th sub-carrier):

$$C_{N_c} = \max_{P_0, \ldots, P_{N_c-1}} \sum_{n=0}^{N_c-1} \log\left(1 + \frac{P_n|\tilde{h}_n|^2}{N_0}\right), \tag{B.59}$$

subject to

$$\frac{1}{N_c} \sum_{n=0}^{N_c-1} P_n = P, \qquad P_n \ge 0, \quad n = 0, \ldots, N_c - 1. \tag{B.60}$$

The solution to this optimization problem is waterfilling and is described in Section 5.3.3.

## B.7 Capacity of the fast fading channel

### B.7.1 Scalar fast fading channnel

#### Ideal interleaving

The fast fading channel with ideal interleaving is modeled as follows:

$$y[m] = h[m]x[m] + w[m], \tag{B.61}$$

where the channel coefficients $h[m]$ are i.i.d. in time and independent of the i.i.d. $\mathcal{CN}(0, N_0)$ additive noise $w[m]$. We are interested in the situation when the receiver tracks the fading channel, but the transmitter only has access to the statistical characterization; the receiver CSI scenario. The capacity of the

power-constrained fast fading channel with receiver CSI can be written as, by viewing the receiver CSI as part of the output of the channel,

$$C = \max_{p_x:\mathbb{E}[x^2]\le P} I(x; y, h). \tag{B.62}$$

Since the fading channel $h$ is independent of the input, $I(x; h) = 0$. Thus, by the chain rule of mutual information (see (B.18)),

$$I(x; y, h) = I(x; h) + I(x; y|h) = I(x; y|h). \tag{B.63}$$

Conditioned on the fading coefficient $h$, the channel is simply an AWGN one, with SNR equal to $P|h|^2/N_0$, where we have denoted the transmit power constraint by $P$. The optimal input distribution for a power constrained AWGN channel is $\mathcal{CN}$, *regardless* of the operating SNR. Thus, the maximizing input distribution in (B.62) is $\mathcal{CN}(0, P)$. With this input distribution,

$$I(x; y|h = \mathtt{h}) = \log\left(1 + \frac{P|\mathtt{h}|^2}{N_0}\right),$$

and thus the capacity of the fast fading channel with receiver CSI is

$$C = \mathbb{E}_h\left[\log\left(1 + \frac{P|h|^2}{N_0}\right)\right], \tag{B.64}$$

where the average is over the stationary distribution of the fading channel.

### Stationary ergodic fading

The above derivation hinges on the i.i.d. assumption on the fading process $\{h[m]\}$. Yet in fact (B.64) holds as long as $\{h[m]\}$ is stationary and ergodic. The alternative derivation below is more insightful and valid for this more general setting.

We first fix a realization of the fading process $\{h[m]\}$. Recall from (B.20) that the rate of reliable communication is given by the average rate of flow of mutual information:

$$\frac{1}{N} I(\mathbf{x}; \mathbf{y}) = \frac{1}{N} \sum_{m=1}^{N} \log(1 + |h[m]|^2 \mathsf{SNR}). \tag{B.65}$$

For large $N$, due to the ergodicity of the fading process,

$$\frac{1}{N} \sum_{m=1}^{N} \log(1 + |h[m]|^2 \mathsf{SNR}) \to \mathbb{E}[\log(1 + |h|^2 \mathsf{SNR})], \tag{B.66}$$

for almost all realizations of the fading process $\{h[m]\}$. This yields the same expression of capacity as in (B.64).

### B.7.2 Fast fading MIMO channel

We have only considered the scalar fast fading channel so far; the extension of the ideas to the MIMO case is very natural. The fast fading MIMO channel with ideal interleaving is (cf. (8.7))

$$\mathbf{y}[m] = \mathbf{H}[m]\mathbf{x}[m] + \mathbf{w}[m], \qquad m = 1, 2, \ldots, \tag{B.67}$$

where the channel $\mathbf{H}$ is i.i.d. in time and independent of the i.i.d. additive noise, which is $\mathcal{CN}(0, N_0\mathbf{I}_{n_r})$. There is an average total power constraint of $P$ on the transmit signal. The capacity of the fast fading channel with receiver CSI is, as in (B.62),

$$C = \max_{p_{\mathbf{x}}:\mathbb{E}[|\mathbf{x}|^2]\le P} I(\mathbf{x}; \mathbf{y}, \mathbf{H}). \tag{B.68}$$

The observation in (B.63) holds here as well, so the capacity calculation is based on the conditional mutual information $I(\mathbf{x}; \mathbf{y}|\mathbf{H})$. If we fix the MIMO channel at a specific realization, we have

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}|\mathbf{H} = \mathtt{H}) &= h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) \\ &= h(\mathbf{y}) - h(\mathbf{w}) \qquad \text{(B.69)} \\ &= h(\mathbf{y}) - n_r \log(\pi e N_0). \qquad \text{(B.70)} \end{aligned}$$

To proceed, we use the following fact about Gaussian random vectors: they are entropy maximizers. Specifically, among all $n$-dimensional complex random vectors with a given covariance matrix $\mathbf{K}$, the one that maximizes the differential entropy is complex circular-symmetric jointly Gaussian $\mathcal{CN}(0, \mathbf{K})$ (Exercise B.8). This is the *vector* extension of the result that Gaussian random variables are entropy maximizers for a fixed variance constraint. The corresponding maximum value is given by

$$\log(\det(\pi \mathrm{e}\mathbf{K})). \tag{B.71}$$

If the covariance of $\mathbf{x}$ is $\mathbf{K}_x$ and the channel is $\mathbf{H} = \mathtt{H}$, then the covariance of $\mathbf{y}$ is

$$N_0\mathbf{I}_{n_r} + \mathtt{H}\mathbf{K}_x\mathtt{H}^*. \tag{B.72}$$

Calculating the corresponding maximal entropy of $\mathbf{y}$ (cf. (B.71)) and substituting in (B.70), we see that

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}|\mathbf{H} = \mathtt{H}) &\le \log((\pi \mathrm{e})^{n_r} \det(N_0\mathbf{I}_{n_r} + \mathtt{H}\mathbf{K}_x\mathtt{H}^*)) - n_r \log(\pi \mathrm{e} N_0) \\ &= \log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathtt{H}\mathbf{K}_x\mathtt{H}^*\right), \qquad \text{(B.73)} \end{aligned}$$

with equality if $\mathbf{x}$ is $\mathcal{CN}(0, \mathbf{K}_x)$. This means that even if the transmitter does not know the channel, there is no loss of optimality in choosing the input to be $\mathcal{CN}$.

Finally, the capacity of the fast fading MIMO channel is found by averaging (B.73) with respect to the stationary distribution of $\mathbf{H}$ and choosing the appropriate covariance matrix subject to the power constraint:

$$C = \max_{\mathbf{K}_x : \mathrm{Tr}[\mathbf{K}_x] \le P} \mathbb{E}_{\mathbf{H}} \left[ \log \det \left( \mathbf{I}_{n_r} + \frac{1}{N_0} \mathbf{H} \mathbf{K}_x \mathbf{H}^* \right) \right]. \tag{B.74}$$

Just as in the scalar case, this result can be generalized to any stationary and ergodic fading process $\{\mathbf{H}[m]\}$.

## B.8 Outage formulation

Consider the slow fading MIMO channel (cf. (8.79))

$$\mathbf{y}[m] = \mathbf{H}\mathbf{x}[m] + \mathbf{w}[m]. \tag{B.75}$$

Here the MIMO channel, represented by $\mathbf{H}$ (an $n_r \times n_t$ matrix with complex entries), is random but not varying with time. The additive noise is i.i.d. $\mathcal{CN}(0, N_0)$ and independent of $\mathbf{H}$.

If there is a positive probability, however small, that the entries of $\mathbf{H}$ are small, then the capacity of the channel is zero. In particular, the capacity of the i.i.d. Rayleigh slow fading MIMO channel is zero. So we focus on characterizing the $\epsilon$-outage capacity: the largest rate of reliable communication such that the error probability is no more than $\epsilon$. We are aided in this study by viewing the slow fading channel in (B.75) as a *compound channel*.

The basic compound channel consists of a collection of DMCs $p_\theta(y|x)$, $\theta \in \Theta$ with the same input alphabet $\mathcal{X}$ and the same output alphabet $\mathcal{Y}$ and parameterized by $\theta$. Operationally, the communication between the transmitter and the receiver is carried out over one specific channel based on the (arbitrary) choice of the parameter $\theta$ from the set $\Theta$. The transmitter does not know the value of $\theta$ but the receiver does. The capacity is the largest rate at which a single coding strategy can achieve reliable communication regardless of which $\theta$ is chosen. The corresponding capacity achieving strategy is said to be *universal* over the class of channels parameterized by $\theta \in \Theta$. An important result in information theory is the characterization of the capacity of the compound channel:

$$\boxed{C = \max_{p_x} \inf_{\theta \in \Theta} I_\theta(x; y).} \tag{B.76}$$

Here, the mutual information $I_\theta(x; y)$ signifies that the conditional distribution of the output symbol $y$ given the input symbol $x$ is given by the

channel $p_\theta(y|x)$. The characterization of the capacity in (B.76) offers a natural interpretation: there exists a coding strategy, parameterized by the input distribution $p_x$, that achieves reliable communication at a rate that is the minimum mutual information among all the allowed channels. We have considered only discrete input and output alphabets, but the generalization to continuous input and output alphabets and, further, to cost constraints on the input follows much the same line as our discussion in Section B.4.1. The tutorial article [69] provides a more comprehensive introduction to compound channels.

We can view the slow fading channel in (B.75) as a compound channel parameterized by $\mathbf{H}$. In this case, we can simplify the parameterization of coding strategies by the input distribution $p_{\mathbf{x}}$: for any fixed $\mathbf{H}$ and channel input distribution $p_{\mathbf{x}}$ with covariance matrix $\mathbf{K}_x$, the corresponding mutual information

$$I(\mathbf{x};\mathbf{y}) \leq \log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right). \tag{B.77}$$

Equality holds when $p_{\mathbf{x}}$ is $\mathcal{CN}(0,\mathbf{K}_x)$ (see Exercise B.8). Thus we can reparameterize a coding strategy by its corresponding covariance matrix (the input distribution is chosen to be $\mathcal{CN}$ with zero mean and the corresponding covariance). For every fixed covariance matrix $\mathbf{K}_x$ that satisfies the power constraint on the input, we can reword the compound channel result in (B.76) as follows. Over the slow fading MIMO channel in (B.75), there exists a universal coding strategy at a rate $R$ bits/s/Hz that achieves reliable communication over all channels $\mathbf{H}$ which satisfy the property

$$\log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right) > R. \tag{B.78}$$

Furthermore, no reliable communication using the coding strategy parameterized by $\mathbf{K}_x$ is possible over channels that are in *outage*: that is, they do not satisfy the condition in (B.78). We can now choose the covariance matrix, subject to the input power constraints, such that we minimize the probability of outage. With a total power constraint of $P$ on the transmit signal, the outage probability when communicating at rate $R$ bits/s/Hz is

$$p_{\text{out}}^{\text{mimo}} := \min_{\mathbf{K}_x:\mathrm{Tr}[\mathbf{K}_x]\leq P} \mathbb{P}\left\{\log\det\left(\mathbf{I}_{n_r} + \frac{1}{N_0}\mathbf{H}\mathbf{K}_x\mathbf{H}^*\right) < R\right\}. \tag{B.79}$$

The $\epsilon$-outage capacity is now the largest rate $R$ such that $p_{\text{out}}^{\text{mimo}} \leq \epsilon$.

By restricting the number of receive antennas $n_r$ to be 1, this discussion also characterizes the outage probability of the MISO fading channel. Further, restricting the MIMO channel $\mathbf{H}$ to be diagonal we have also characterized the outage probability of the parallel fading channel.

## B.9 Multiple access channel

### B.9.1 Capacity region

The uplink channel (with potentially multiple antenna elements) is a special case of the multiple access channel. Information theory gives a formula for computing the capacity region of the multiple access channel in terms of mutual information, from which the corresponding region for the uplink channel can be derived as a special case.

The capacity of a memoryless point-to-point channel with input $x$ and output $y$ is given by

$$C = \max_{p_x} I(x; y),$$

where the maximization is over the input distributions subject to the average cost constraint. There is an analogous theorem for multiple access channels. Consider a two-user channel, with inputs $x_k$ from user $k$, $k = 1, 2$ and output $y$. For given input distributions $p_{x_1}$ and $p_{x_2}$ and *independent* across the two users, define the pentagon $\mathcal{C}(p_{x_1}, p_{x_2})$ as the set of all rate pairs satisfying:

$$R_1 < I(x_1; y|x_2), \tag{B.80}$$

$$R_2 < I(x_2; y|x_1), \tag{B.81}$$

$$R_1 + R_2 < I(x_1, x_2; y). \tag{B.82}$$

The capacity region of the multiple access channel is the convex hull of the union of these pentagons over all possible independent input distributions subject to the appropriate individual average cost constraints, i.e.,

$$\mathcal{C} = \text{convex hull of}(\cup_{p_{x_1}, p_{x_2}} \mathcal{C}(p_{x_1}, p_{x_2})). \tag{B.83}$$

The convex hull operation means that we not only include points in $\cup\mathcal{C}(p_{x_1}, p_{x_2})$ in $\mathcal{C}$, but also all their convex combinations. This is natural since the convex combinations can be achieved by time-sharing.

The capacity region of the uplink channel with single antenna elements can be arrived at by specializing this result to the scalar Gaussian multiple access channel. With average power constraints on the two users, we observe that Gaussian inputs for user 1 and 2 *simultaneously* maximize $I(x_1; y|x_2)$, $I(x_2; y|x_1)$ and $I(x_1; x_2; y)$. Hence, the pentagon from this input distribution is a superset of all other pentagons, and the capacity region itself is this pentagon. The same observation holds for the time-invariant uplink channel with single transmit antennas at each user and multiple receive antennas at the base-station. The expressions for the capacity regions of the uplink with a single receive antenna are provided in (6.4), (6.5) and (6.6). The capacity region of the uplink with multiple receive antennas is expressed in (10.6).

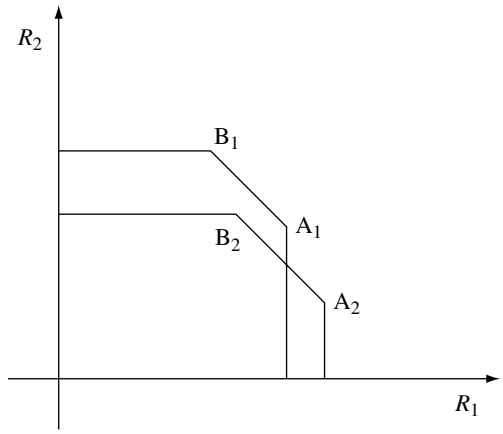


**Figure B.7** The achievable rate regions (pentagons) corresponding to two different input distributions may not fully overlap with respect to one another.

In the uplink with single transmit antennas, there was a unique set of input distributions that simultaneously maximized the different constraints ((B.80), (B.81) and (B.82)). In general, no single pentagon may dominate over the other pentagons, and in that case the overall capacity region may not be a pentagon (see Figure B.7). An example of this situation is provided by the uplink with multiple transmit antennas at the users. In this situation, zero mean circularly symmetric complex Gaussian random vectors still simultaneously maximize all the constraints, but with different covariance matrices. Thus we can restrict the user input distributions to be zero mean $\mathcal{CN}$, but leave the covariance matrices of the users as parameters to be chosen. Consider the two-user uplink with multiple transmit and receive antennas. Fixing the $k$th user input distribution to be $\mathcal{CN}(0, \mathbf{K}_k)$ for $k = 1, 2$, the corresponding pentagon is expressed in (10.23) and (10.24). In general, there is no single choice of covariance matrices that simultaneously maximize the constraints: the capacity region is the convex hull of the union of the pentagons created by all the possible covariance matrices (subject to the power constraints on the users).

## B.9.2 Corner points of the capacity region

Consider the pentagon $\mathcal{C}(p_{x_1}, p_{x_2})$ parameterized by fixed independent input distributions on the two users and illustrated in Figure B.8. The two corner points A and B have an important significance: if we have coding schemes that achieve reliable communication to the users at the rates advertised by these two points, then the rates at every other point in the pentagon can be achieved by appropriate time-sharing between the two strategies that achieved the points A and B. Below, we try to get some insight into the nature of the two corner points and properties of the receiver design that achieves them.

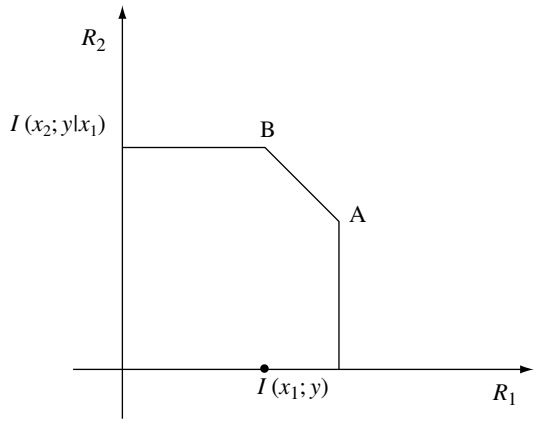


**Figure B.8** The set of rates at which two users can jointly reliably communicate is a pentagon, parameterized by the independent users' input distributions.

Consider the corner point B. At this point, user 1 gets the rate $I(x_1; y)$. Using the chain rule for mutual information we can write

$$I(x_1, x_2; y) = I(x_1; y) + I(x_2; y|x_1).$$

Since the sum rate constraint is tight at the corner point B, user 2 achieves its highest rate $I(x_2; y|x_1)$. This rate pair can be achieved by a successive interference cancellation (SIC) receiver: decode user 1 first, treating the signal from user 2 as part of the noise. Next, decode user 2 conditioned on the already decoded information from user 1. In the uplink with a single antenna, the second stage of the successive cancellation receiver is very explicit: given the decoded information from user 1, the receiver simply subtracts the decoded transmit signal of user 1 from the received signal. With multiple receive antennas, the successive cancellation is done in conjunction with the MMSE receiver. The MMSE receiver is information lossless (this aspect is explored in Section 8.3.4) and we can conclude the following intuitive statement: the MMSE–SIC receiver is optimal because it "implements" the chain rule for mutual information.

## B.9.3 Fast fading uplink

Consider the canonical two-user fast fading MIMO uplink channel:

$$\mathbf{y}[m] = \mathbf{H}_1[m]\mathbf{x}_1[m] + \mathbf{H}_2[m]\mathbf{x}_2[m] + \mathbf{w}[m], \tag{B.84}$$

where the MIMO channels $\mathbf{H}_1$ and $\mathbf{H}_2$ are independent and i.i.d. over time. As argued in Section B.7.1, interleaving allows us to convert stationary channels with memory to this canonical form. We are interested in the receiver CSI situation: the receiver tracks both the users' channels perfectly. For fixed

independent input distributions $p_{x_1}$ and $p_{x_2}$, the achievable rate region consists of tuples $(R_1, R_2)$ constrained by

$$R_1 < I(\mathbf{x}_1; \mathbf{y}, \mathbf{H}_1, \mathbf{H}_2 | \mathbf{x}_2), \tag{B.85}$$

$$R_2 < I(\mathbf{x}_2; \mathbf{y}, \mathbf{H}_1, \mathbf{H}_2 | \mathbf{x}_1), \tag{B.86}$$

$$R_1 + R_2 < I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}, \mathbf{H}_1, \mathbf{H}_2). \tag{B.87}$$

Here we have modeled receiver CSI as the MIMO channels being part of the output of the multiple access channel. Since the channels are independent of the user inputs, we can use the chain rule of mutual information, as in (B.63), to rewrite the constraints on the rate tuples as

$$R_1 < I(\mathbf{x}_1; \mathbf{y} | \mathbf{H}_1, \mathbf{H}_2, \mathbf{x}_2), \tag{B.88}$$

$$R_2 < I(\mathbf{x}_2; \mathbf{y} | \mathbf{H}_1, \mathbf{H}_2, \mathbf{x}_1), \tag{B.89}$$

$$R_1 + R_2 < I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y} | \mathbf{H}_1, \mathbf{H}_2). \tag{B.90}$$

Fixing the realization of the MIMO channels of the users, we see again (as in the time-invariant MIMO uplink) that the input distributions can be restricted to be zero mean $\mathcal{CN}$ but leave their covariance matrices as parameters to be chosen later. The corresponding rate region is a pentagon expressed by (10.23) and (10.24). The conditional mutual information is now the average over the stationary distributions of the MIMO channels: an expression for this pentagon is provided in (10.28) and (10.29).

## B.10 Exercises

**Exercise B.1** Suppose $x$ is a discrete random variable taking on $K$ values, each with probability $p_1, \ldots, p_K$. Show that

$$\max_{p_1, \ldots, p_K} H(x) = \log K,$$

and further that this is achieved only when $p_i = 1/K$, $i = 1, \ldots, K$, i.e., $x$ is uniformly distributed.

**Exercise B.2** In this exercise, we will study when conditioning does not reduce entropy.

1. A concave function $f$ is defined in the text by the condition $f''(x) \leq 0$ for $x$ in the domain. Give an alternative geometric definition that does not use calculus.
2. Jensen's inequality for a random variable $x$ states that for any concave function $f$

$$\mathbb{E}[f(x)] \leq f(\mathbb{E}[x]). \tag{B.91}$$

Prove this statement. *Hint*: You might find it useful to draw a picture and visualize the proof geometrically. The geometric definition of a concave function might come in handy here.

3. Show that $H(x|y) \leq H(x)$ with equality if and only if $x$ and $y$ are independent. Give an example in which $H(x|y = k) > H(x)$. Why is there no contradiction between these two statements?

**Exercise B.3** Under what condition on $x_1, x_2, y$ does it hold that

$$I(x_1, x_2; y) = I(x_1; y) + I(x_2; y)? \tag{B.92}$$

**Exercise B.4** Consider a continuous real random variable $x$ with density $f_x(\cdot)$ non-zero on the entire real line. Suppose the second moment of $x$ is fixed to be $P$. Show that among all random variables with the constraints as those on $x$, the Gaussian random variable has the maximum differential entropy. *Hint*: The differential entropy is a concave function of the density function and fixing the second moment corresponds to a linear constraint on the density function. So, you can use the classical Lagrangian techniques to solve this problem.

**Exercise B.5** Suppose $x$ is now a non-negative random variable with density non-zero for all non-negative real numbers. Further suppose that the mean of $x$ is fixed. Show that among all random variables of this form, the exponential random variable has the maximum differential entropy.

**Exercise B.6** In this exercise, we generalize the results in Exercises B.4 and B.5. Consider a continuous real random variable $x$ with density $f_x(\cdot)$ on a support set $S$ (i.e., $f_x(u) = 0, u \notin S$). In this problem we will study the structure of the random variable $x$ with maximal differential entropy that satisfies the following *moment* conditions:

$$\int_S r_i(u) f_x(u)\, \mathrm{d}u = A_i, \qquad i = 1, \ldots, m. \tag{B.93}$$

Show that $x$ with density

$$f_x(u) = \exp\left(\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(u)\right), u \in S, \tag{B.94}$$

has the maximal differential entropy subject to the moment conditions (B.93). Here $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen such that the moment conditions (B.93) are met and that $f_x(\cdot)$ is a density function (i.e., it integrates to unity).

**Exercise B.7** In this problem, we will consider the differential entropy of a vector of continuous random variables with moment conditions.

1. Consider the class of continuous real random vectors $\mathbf{x}$ with the covariance condition: $\mathbb{E}[\mathbf{x}\mathbf{x}^t] = \mathbf{K}$. Show that the jointly Gaussian random vector with covariance $\mathbf{K}$ has the maximal differential entropy among this set of covariance constrained random variables.
2. Now consider a complex random variable $x$. Show that among the class of continuous complex random variables $x$ with the second moment condition $\mathbb{E}[|x|^2] \leq P$,

the *circularly symmetric* Gaussian complex random variable has the maximal differential entropy. *Hint*: View $\mathbf{x}$ as a length 2 vector of real random variables and use the previous part of this question.

**Exercise B.8** Consider a zero mean complex random vector $\mathbf{x}$ with fixed covariance $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \mathbf{K}$. Show the following upper bound on the differential entropy:

$$h(\mathbf{x}) \leq \log\det(\pi e\mathbf{K}), \tag{B.95}$$

with equality when $\mathbf{x}$ is $\mathcal{CN}(0, \mathbf{K})$. *Hint*: This is a generalization of Exercise B.7(2).

**Exercise B.9** Show that the structure of the input distribution in (5.28) optimizes the mutual information in the MISO channel. *Hint*: Write the second moment of $\mathbf{y}$ as a function of the covariance of $\mathbf{x}$ and see which covariance of $\mathbf{x}$ maximizes the second moment of $y$. Now use Exercise B.8 to reach the desired conclusion.

**Exercise B.10** Consider the real random vector $\mathbf{x}$ with i.i.d. $\mathcal{N}(0, P)$ components. In this exercise, we consider properties of the scaled vector $\tilde{\mathbf{x}} := (1/\sqrt{N})\mathbf{x}$. (The material here is drawn from the discussion in Chapter 5.5 in [148].)

1. Show that $(\mathbb{E}[\|\mathbf{x}\|^2])/N = P$, so the scaling ensured that the mean length of $\|\tilde{\mathbf{x}}\|^2$ is $P$, independent of $N$.
2. Calculate the variance of $\|\tilde{\mathbf{x}}\|^2$ and show that $\|\tilde{\mathbf{x}}\|^2$ converges to $P$ in probability. Thus, the scaled vector is *concentrated* around its mean.
3. Consider the event that $\tilde{\mathbf{x}}$ lies in the *shell* between two concentric spheres of radius $\rho - \delta$ and $\rho$. (See Figure B.9.) Calculate the volume of this shell to be

$$B_N\left(\rho^N - (\rho-\delta)^N\right), \quad \text{where} B_N = \begin{cases} \pi^{N/2}/(\frac{N}{2})! & N \text{ even} \\ (2^N\pi^{(N-1)/2})[(N-1)/2]!/N! & N \text{ odd.} \end{cases} \tag{B.96}$$

4. Show that we can approximate the volume of the shell by

$$NB_N\rho^{N-1}\delta, \qquad \text{for} \delta/\rho \ll 1. \tag{B.97}$$

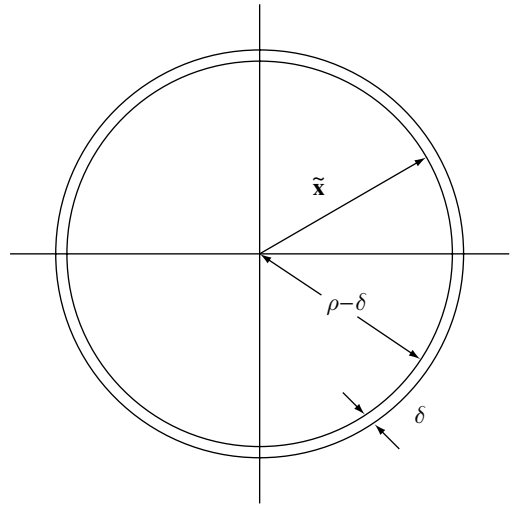


**Figure B.9** The shell between two concentric spheres of radius $\rho - \delta$ and $\rho$.

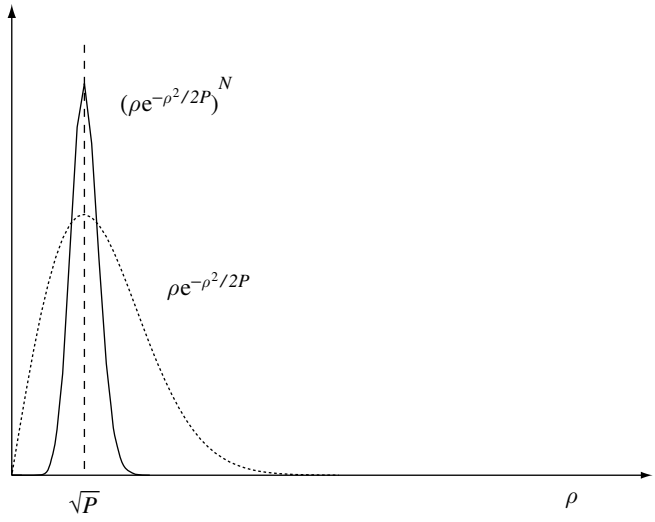


**Figure B.10** Behavior of $\mathbb{P}(\rho - \delta \leq \|\tilde{\mathbf{x}}\| < \rho)$ as a function of $\rho$.

5. Let us approximate the density of $\tilde{\mathbf{x}}$ inside this shell to be

$$f_{\tilde{\mathbf{x}}}(\mathbf{a}) \approx \left(\frac{N}{2\pi P}\right)^{N/2} \exp\left(-\frac{N\rho^2}{2P}\right), \qquad r - \delta < \|\mathbf{a}\| \leq \rho. \tag{B.98}$$

   Combining (B.98) and (B.97), show that for $\delta/\rho =$ a constant $\ll 1$,

$$\mathbb{P}(\rho - \delta \leq \|\tilde{\mathbf{x}}\| < \rho) \approx \left[\rho \exp\left(-\frac{\rho^2}{2P}\right)\right]^N. \tag{B.99}$$

6. Show that the right hand side of (B.99) has a single maximum at $\rho^2 = P$ (see Figure B.10).
7. Conclude that as $N$ becomes large, the consequence is that only values of $\|\tilde{\mathbf{x}}\|^2$ in the vicinity of $P$ have significant probability. This phenomenon is called *sphere hardening*.

**Exercise B.11** Calculate the mutual information achieved by the isotropic input distribution $\mathbf{x}$ is $\mathcal{CN}(0, P/L \cdot \mathbf{I}_L)$ in the MISO channel (cf. (5.27)) with given channel gains $h_1, \ldots, h_L$.

**Exercise B.12** In this exercise, we will study the capacity of the $L$-tap frequency-selective channel directly (without recourse to the cyclic prefix idea). Consider a length $N_c$ vector input $\mathbf{x}$ on to the channel in (5.32) and denote the vector output (of length $N_c + L - 1$) by $\mathbf{y}$. The input and output are linearly related as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \tag{B.100}$$

where $\mathbf{G}$ is a matrix whose entries depend on the channel coefficients $h_0, \ldots, h_{L-1}$ as follows: $G[i, j] = h_{i-j}$ for $i \geq j$ and zero everywhere else. The channel in (B.100) is a *vector* version of the basic AWGN channel and we consider the rate of reliable communication $I(\mathbf{x}; \mathbf{y})/N_c$.

1. Show that the optimal input distribution is $\mathbf{x}$ is $\mathcal{CN}(0, \mathbf{K}_x)$, for some covariance matrix $\mathbf{K}_x$ meeting the power constraint. (*Hint*: You will find Exercise B.8 useful.)
2. Show that it suffices to consider only those covariances $\mathbf{K}_x$ that have the same set of eigenvectors as $\mathbf{G}^*\mathbf{G}$. (*Hint*: Use Exercise B.8 to explicitly write the reliable rate of communiation in the vector AWGN channel of (B.100).)
3. Show that

$$(\mathbf{G}^*\mathbf{G})_{ij} = r_{i-j}, \tag{B.101}$$

where

$$r_n := \sum_{\ell=0}^{L-l-1} (h_\ell)^* h[\ell + n], \qquad n \geq 0, \tag{B.102}$$

$$r_n := r_{-n}^*, \qquad n \leq 0. \tag{B.103}$$

Such a matrix $\mathbf{G}^*\mathbf{G}$ is said to be *Toeplitz*.

4. An important result about the Hermitian Toeplitz matrix $\mathbf{G}\mathbf{G}^*$ is that the empirical distribution of its eigenvalues converges (weakly) to the discrete-time Fourier transform of the sequence $\{r_l\}$. How is the discrete-time Fourier transform of the sequence $\{r_l\}$ related to the discrete-time Fourier transform $H(f)$ of the sequence $h_0, \ldots, h_{L-1}$?
5. Use the result of the previous part and the nature of the optimal $\mathbf{K}_x^*$ (discussed in part (2)) to show that the rate of reliable communication is equal to

$$\int_0^W \log\left(1 + \frac{P^*(f)|H(f)|^2}{N_0}\right) \mathrm{d}f. \tag{B.104}$$

Here the waterfilling power allocation $P^*(f)$ is as defined in (5.47). This answer is, of course, the same as that derived in the text (cf. (5.49)). The cyclic prefix converted the frequency-selective channel into a parallel channel, reliable communication over which is easier to understand. With a direct approach we had to use analytical results about Toeplitz forms; more can be learnt about these techniques from [53].