PART I

WIRELESS COMMUNICATION THEORY

Cambridge University Press & Assessment
978-0-521-76228-1 — Foundations of MIMO Communication
Robert W. Heath Jr. , Angel Lozano
Excerpt
More Information

1 A primer on information theory and MMSE estimation

> Theory is the first term in the Taylor series expansion of practice.
>
> Thomas Cover

1.1 Introduction

Information theory deals broadly with the science of information, including compressibility and storage of data, as well as reliable communication. It is an exceptional discipline in that it has a precise founder, Claude E. Shannon, and a precise birthdate, 1948. The publication of Shannon's seminal treatise, "A mathematical theory of communication" [58], represents one of the scientific highlights of the twentieth century and, in many respects, marks the onset of the information age. Shannon was an engineer, yet information theory is perhaps best described as an outpost of probability theory that has extensive applicability in electrical engineering as well as substantial overlap with computer science, physics, economics, and even biology. Since its inception, information theory has been distilling practical problems into mathematical formulations whose solutions cast light on those problems. A staple of information theory is its appreciation of elegance and harmony, and indeed many of its results possess a high degree of aesthetic beauty. And, despite their highly abstract nature, they often do reveal much about the practical problems that motivated them in the first place.

Although Shannon's teachings are by now well assimilated, they represented a radical departure from time-honored axioms [52]. In particular, it was believed before Shannon that error-free communication was only possible in the absence of noise or at vanishingly small transmission rates. Shannon's channel coding theorem was nothing short of revolutionary, as it proved that every channel had a characterizing quantity (the capacity) such that, for transmission rates not exceeding it, the error probability could be made arbitrarily small. Ridding the communication of errors did not require overwhelming the noise with signal power or slowing down the transmission rate, but could be achieved in the face of noise and at positive rates—as long as the capacity was not exceeded—by embracing the concept of coding: information units should not be transmitted in isolation but rather in coded blocks, with each unit thinly spread over as many symbols as possible; redundancy and interdependency as an antidote to the confusion engendered by noise. The notion of channel capacity is thus all-important in information theory, being something akin to the speed of light in terms of reliable communication. This analogy with the speed of light, which is common and enticing, must however be viewed with perspective. While, in the early years of information theory, the capacity might have been perceived as remote (wireline modems were transmitting on the order of 300 bits/s in telephone channels whose Shannon capacity was computed as being 2–3 orders of magnitude higher), nowadays it can be closely approached in important channels. Arguably, then, to the daily lives of people the capacity is a far more relevant limitation than the speed of light.

The emergence of information theory also had an important unifying effect, proving an umbrella under which all channels and forms of communication—each with its own toolbox of methodologies theretofore—could be studied on a common footing. Before Shannon, something as obvious today as the transmission of video over a telephone line would have been inconceivable.

As anecdotal testimony of the timeless value and transcendence of Shannon's work, we note that, in 2016, almost seven decades after its publication, "A mathematical theory of communication" ranked as a top-three download in IEEE *Xplore*, the digital repository that archives over four million electrical engineering documents—countlessly many of which elaborate on aspects of the theory spawned by that one paper.

This chapter begins by describing certain types of signals that are encountered throughout the text. Then, the chapter goes on to review those concepts in information theory that are needed throughout, with readers interested in more comprehensive treatments of the matter referred to dedicated textbooks [14, 59, 60]. In addition to the relatively young discipline of information theory, the chapter also touches on the much older subject of MMSE estimation. The packaging of both topics in a single chapter is not coincidental, but rather a choice that is motivated by the relationship between the two—a relationship made of bonds that have long been known, and of others that have more recently been unveiled [61]. Again, we cover only those MMSE estimation concepts that are needed in the book, with readers interested in broader treatments referred to estimation theory texts [62].

1.2 Signal distributions

The signals described next are in general complex-valued. The interpretation of complex signals, as well as complex channels and complex noise, as baseband representations of real-valued passband counterparts is provided in Chapter 2, and readers needing background on this interpretation are invited to peruse Section 2.2 before proceeding. We advance that the real and imaginary parts of a signal are respectively termed the *in-phase* and the *quadrature* components.

Consider a complex scalar s, zero-mean and normalized to be of unit variance, which is to serve as a signal. From a theoretical vantage, a distribution that is all-important because of its optimality in many respects is the complex Gaussian, $s \sim \mathcal{N}_{\mathbb{C}}(0, 1)$, details of which are offered in Appendix C.1.9. In practice though, a scalar signal s is drawn from a discrete distribution defined by M points, say $\mathrm{s}_0, \ldots, \mathrm{s}_{M-1}$, taken with probabilities $p_0, \ldots, p_{M-1}$. These points are arranged into constellations such as the following.

Table 1.1 Constellation minimum distances

Constellation	$d_{\min}$
M-PSK	$2\,\sin\left(\frac{\pi}{M}\right)$
Square M-QAM	$\sqrt{\frac{6}{M-1}}$

- M-ary phase shift keying (M-PSK), where

$$\mathrm{s}_m = e^{\mathrm{j}2\pi\frac{m}{M}+\phi_0} \qquad m = 0, \ldots, M-1 \tag{1.1}$$

  with ϕ_0 an arbitrary phase. Because of symmetry, the points are always equiprobable, $p_m = 1/M$ for $m = 0, \ldots, M-1$. Special mention must be made of binary phase-shift keying (BPSK), corresponding to $M = 2$, and quadrature phase-shift keying (QPSK), which corresponds to $M = 4$.

- Square M-ary quadrature amplitude modulation (M-QAM), where the in-phase and quadrature components of s independently take values in the set

$$\left\{\sqrt{\frac{3}{2\,(M-1)}}\left(2m-1-\sqrt{M}\right)\right\} \qquad m = 0, \ldots, \sqrt{M}-1 \tag{1.2}$$

  with $\sqrt{M}$ integer. (Nonsquare M-QAM constellations also exist, and they are employed regularly in wireline systems, but seldom in wireless.) Although making the points in a M-QAM constellation equiprobable is not in general optimum, it is commonplace. Note that, except for perhaps an innocuous rotation, 4-QAM coincides with QPSK.

For both M-PSK and square M-QAM, the minimum distance between constellation points is provided in Table 1.1.

Example 1.1

Depict the 8-PSK and 16-QAM constellations and indicate the distance between nearest neighbors within each.

Solution

See Fig. 1.1.

It is sometimes analytically convenient to approximate discrete constellations by means of continuous distributions over a suitable region on the complex plane. These continuous distributions can be interpreted as limits of dense M-ary constellations for $M \to \infty$. For equiprobable M-PSK and M-QAM, the appropriate unit-variance continuous distributions are:

- ∞-PSK, where $s = e^{\mathrm{j}\phi}$ with ϕ uniform on $[0, 2\pi)$.
- ∞-QAM, where s is uniform over the square $\left[-\sqrt{3/2}, \sqrt{3/2}\right] \times \left[-\sqrt{3/2}, \sqrt{3/2}\right]$ on the complex plane.

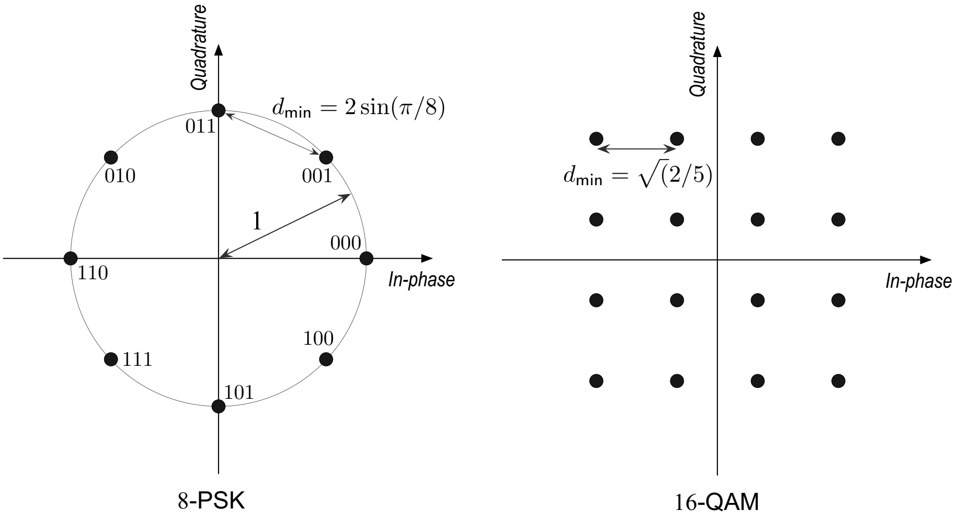

Fig. 1.1 Unit-variance 8-PSK and 16-QAM constellations.

Except for BPSK, all the foregoing distributions, both continuous and discrete, are *proper complex* in the sense of Section C.1.4.

Lastly, a distribution that is relevant for ultrawideband communication is "on–off" keying [63, 64]

$$s = \begin{cases} 0 & \text{with probability } 1-\epsilon \\ \sqrt{1/\epsilon} & \text{with probability } \epsilon \end{cases} \tag{1.3}$$

parameterized by ϵ. Practical embodiments of this distribution include pulse-position modulation [65] and impulse radio [66]. Generalizations of (1.3) to multiple "on" states are also possible.

1.3 Information content

Information equals uncertainty. If a given quantity is certain, then knowledge of it provides no information. It is therefore only natural, as Shannon recognized, to model information and data communication using probability theory. All the elements that play a role in communications (signals, channel, noise) are thereby abstracted using random variables and random processes. For the reader's convenience, reviews of the basic results on random variables and random processes that are necessary for the derivations in this chapter are respectively available in Appendices C.1 and C.3.

As the starting point of our exposition, let us see how to quantify the information content of random variables and processes. We adopt the *bit* as our information currency and, consequently, all applicable logarithms are to the base 2; other information units can be obtained by merely modifying that base, e.g., the *byte* (base 256), the *nat* (base e), and the *ban* (base 10).

All the summations and integrals that follow should be taken over the support of the corresponding random variables, i.e., the set of values on which their probabilities are nonzero.

1.3.1 Entropy

Let x be a discrete random variable with PMF $p_x(\cdot)$. Its *entropy*, denoted by $\mathcal{H}(x)$, is defined as

$$\mathcal{H}(x) = -\sum_{\mathrm{x}} p_x(\mathrm{x}) \log_2 p_x(\mathrm{x}) \tag{1.4}$$

$$= -\mathbb{E}\big[\log_2 p_x(x)\big]. \tag{1.5}$$

Although the entropy is a function of $p_x(\cdot)$ rather than of x, it is rather standard to slightly abuse notation and write it as $\mathcal{H}(x)$. The entropy is nonnegative and it quantifies the amount of uncertainty associated with x: the larger the entropy, the more unpredictable x. Not surprisingly then, the uniform PMF is the entropy-maximizing one. If the cardinality of x is M, then its entropy under a uniform PMF trivially equals $\mathcal{H}(x) = \log_2 M$ bits and thus we can affirm that, for any x with cardinality M, $\mathcal{H}(x) \leq \log_2 M$ bits. At the other extreme, variables with only one possible outcome (i.e., deterministic quantities) have an entropy of zero. The entropy $\mathcal{H}(x)$ gives the number of bits required to describe x on average. Note that the actual values taken by x are immaterial in terms of $\mathcal{H}(x)$; only the probabilities of those values matter.

Similar to Boltzmann's entropy in statistical mechanics, the entropy was introduced as a measure of information by Shannon with the rationale of being the only measure that is continuous in the probabilities, increasing in the support if $p_x(\cdot)$ is uniform, and additive when x is the result of multiple choices [67].

Example 1.2

Express the entropy of the Bernoulli random variable

$$x = \begin{cases} 0 & \text{with probability } \; p \\ 1 & \text{with probability } \; 1-p. \end{cases} \tag{1.6}$$

Solution

The entropy of x is the so-called binary entropy function,

$$\mathcal{H}(x) = -p \log_2 p - (1-p) \log_2 (1-p), \tag{1.7}$$

which satisfies $\mathcal{H}(x) \leq 1$ with equality for $p = 1/2$.

Example 1.3

Express the entropy of an equiprobable M-ary constellation.

Solution

For s conforming to a discrete constellation with M equiprobable points,

$$\mathcal{H}(s) = -\sum_{m=0}^{M-1} \frac{1}{M} \log \frac{1}{M} \tag{1.8}$$

$$= \log_2 M. \tag{1.9}$$

These $\log_2 M$ bits can be mapped onto the M constellation points in various ways. Particularly relevant is the so-called *Gray mapping*, characterized by nearest-neighbor constellation points differing by a single bit. This ensures that, in the most likely error event, when a constellation point is confused with its closest neighbor, a single bit is flipped. Gray mapping is illustrated for a PSK constellation in Fig. 1.1.

Having seen how to quantify the amount of information in an individual variable, we now extend the concept to multiple ones. Indeed, because of the multiple inputs and outputs, the most convenient MIMO representation uses vectors for the signals and matrices for the channels.

Let x_0 and x_1 be discrete random variables with joint PMF $p_{x_0 x_1}(\cdot,\cdot)$ and marginals $p_{x_0}(\cdot)$ and $p_{x_1}(\cdot)$. The joint entropy of x_0 and x_1 is

$$\mathcal{H}(x_0, x_1) = -\sum_{\mathrm{x}_0}\sum_{\mathrm{x}_1} p_{x_0 x_1}(\mathrm{x}_0, \mathrm{x}_1) \log_2 p_{x_0 x_1}(\mathrm{x}_0, \mathrm{x}_1) \tag{1.10}$$

$$= -\mathbb{E}\big[\log_2 p_{x_0 x_1}(x_0, x_1)\big]. \tag{1.11}$$

If x_0 and x_1 are independent, then $\mathcal{H}(x_0, x_1) = \mathcal{H}(x_0) + \mathcal{H}(x_1)$. Furthermore, by regarding x_0 and x_1 as entries of a vector, we can claim (1.10) as the entropy of such a vector. More generally, for any discrete random vector $\boldsymbol{x}$,

$$\mathcal{H}(\boldsymbol{x}) = -\mathbb{E}\big[\log_2 p_{\boldsymbol{x}}(\boldsymbol{x})\big]. \tag{1.12}$$

Often, it is necessary to appraise the uncertainty that remains in a random variable x once a related random variable y has been observed. This is quantified by the conditional entropy of x given y,

$$\mathcal{H}(x|y) = -\sum_{\mathrm{x}}\sum_{\mathrm{y}} p_{xy}(\mathrm{x}, \mathrm{y}) \log_2 p_{x|y}(\mathrm{x}|\mathrm{y}). \tag{1.13}$$

If x and y are independent, then naturally $\mathcal{H}(x|y) = \mathcal{H}(x)$ whereas, if x is a deterministic function of y, then $\mathcal{H}(x|y) = 0$.

The joint and conditional entropies are related by the chain rule

$$\mathcal{H}(x, y) = \mathcal{H}(x) + \mathcal{H}(y|x), \tag{1.14}$$

which extends immediately to vectors. When more than two variables are involved, the chain rule generalizes as

$$\mathcal{H}(x_0, \ldots, x_{N-1}) = \sum_{n=0}^{N-1} \mathcal{H}(x_n | x_0, \ldots, x_{n-1}). \tag{1.15}$$

1.3.2 Differential entropy

A quantity seemingly analogous to the entropy, the *differential entropy*, can be defined for continuous random variables. If $f_x(\cdot)$ is the probability density function (PDF) of x, its differential entropy is

$$\mathfrak{h}(x) = -\int f_x(\mathrm{x}) \log_2 f_x(\mathrm{x})\, \mathrm{dx} \tag{1.16}$$

$$= -\mathbb{E}\big[\log_2 f_x(x)\big] \tag{1.17}$$

where the integration in (1.16) is over the complex plane. Care must be exercised when dealing with differential entropies, because they may be negative. Indeed, despite the similarity in their forms, the entropy and differential entropy do not admit the same interpretation: the former measures the information contained in a random variable whereas the latter does not. Tempting as it may be, $\mathfrak{h}(x)$ cannot be approached by discretizing $f_x(\cdot)$ into progressively smaller bins and computing the entropy of the ensuing discrete random variable. The entropy of a b-bit quantization of x is approximately $\mathfrak{h}(x)+b$, which diverges as $b \to \infty$. This merely confirms what one may have intuitively guessed, namely that the amount of information in a continuous variable, i.e., the number of bits required to describe it, is generally infinite. The physical meaning of $\mathfrak{h}(x)$ is thus not the amount of information in x. In fact, the differential entropy is devoid—from an engineering viewpoint—of operational meaning and ends up serving mostly as a stepping stone to the mutual information, which does have plenty of engineering significance.

Example 1.4

Calculate the differential entropy of a real random variable x uniformly distributed in $[0, b]$.

Solution

$$\mathfrak{h}(x) = -\int_0^b \frac{1}{b} \log_2\left(\frac{1}{b}\right) \mathrm{d}x \tag{1.18}$$

$$= \log_2 b. \tag{1.19}$$

Note that $\mathfrak{h}(x) < 0$ for $b < 1$.

Example 1.5 (Differential entropy of a complex Gaussian scalar)

Let $x \sim \mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$. Invoking the PDF in (C.14),

$$\mathfrak{h}(x) = \mathbb{E}\left[\frac{|x-\mu|^2}{\sigma^2} \log_2 e + \log_2\left(\pi\sigma^2\right)\right] \tag{1.20}$$

$$= \log_2\left(\pi e \sigma^2\right). \tag{1.21}$$

Observe how, in Example 1.5, the mean μ is immaterial to $\mathfrak{h}(x)$. This reflects the property of differential entropy being translation-invariant, meaning that $\mathfrak{h}(x+a) = \mathfrak{h}(x)$ for any constant a; it follows from this property that we can always translate a random variable and set its mean to zero without affecting its differential entropy.

In the context of information content, the importance of the complex Gaussian distribution stems, not only from its prevalence, but further from the fact that it is the distribution that maximizes the differential entropy for a given variance [14]. Thus, for any random variable x with variance σ^2, $\mathfrak{h}(x) \leq \log_2(\pi e \sigma^2)$.

As in the discrete case, the notion of differential entropy readily extends to the multivariate realm. If $\boldsymbol{x}$ is a continuous random vector with PDF $f_{\boldsymbol{x}}(\cdot)$, then

$$\mathfrak{h}(\boldsymbol{x}) = -\mathbb{E}\big[\log_2 f_{\boldsymbol{x}}(\boldsymbol{x})\big]. \tag{1.22}$$

Example 1.6 (Differential entropy of a complex Gaussian vector)

Let $\boldsymbol{x} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{R})$. From (C.15) and (1.22),

$$\begin{aligned}
\mathfrak{h}(\boldsymbol{x}) &= -\mathbb{E}\big[\log_2 f_{\boldsymbol{x}}(\boldsymbol{x})\big] \tag{1.23}\\
&= \log_2 \det(\pi \boldsymbol{R}) + \mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})^* \boldsymbol{R}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\big] \log_2 e \tag{1.24}\\
&= \log_2 \det(\pi \boldsymbol{R}) + \mathrm{tr}\Big(\mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})^* \boldsymbol{R}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\big]\Big) \log_2 e \tag{1.25}\\
&= \log_2 \det(\pi \boldsymbol{R}) + \mathrm{tr}\Big(\mathbb{E}\big[\boldsymbol{R}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^*\big]\Big) \log_2 e \tag{1.26}\\
&= \log_2 \det(\pi \boldsymbol{R}) + \mathrm{tr}\Big(\boldsymbol{R}^{-1} \mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^*\big]\Big) \log_2 e \tag{1.27}\\
&= \log_2 \det(\pi \boldsymbol{R}) + \mathrm{tr}(\boldsymbol{I}) \log_2 e \tag{1.28}\\
&= \log_2 \det(\pi e \boldsymbol{R}), \tag{1.29}
\end{aligned}$$

where in (1.25) we used the fact that a scalar equals its trace, while in (1.26) we invoked the commutative property in (B.26).

As in the scalar case, the complex Gaussian distribution maximizes the differential entropy for a given covariance matrix. For any complex random vector $\boldsymbol{x}$ with covariance $\boldsymbol{R}$, therefore, $\mathfrak{h}(\boldsymbol{x}) \leq \log_2 \det(\pi e \boldsymbol{R})$.

The conditional differential entropy of x given y equals

$$\mathfrak{h}(x|y) = -\mathbb{E}\big[\log_2 f_{x|y}(x|y)\big] \tag{1.30}$$

with expectation over the joint distribution of x and y. The chain rule that relates joint and conditional entropies is

$$\mathfrak{h}(x_0, \ldots, x_{N-1}) = \sum_{n=0}^{N-1} \mathfrak{h}(x_n | x_0, \ldots, x_{n-1}), \tag{1.31}$$

which extends verbatim to vectors.

1.3.3 Entropy rate

To close the discussion on information content, let us turn our attention from random variables to random processes. A discrete random process $x_0, \ldots, x_{N-1}$ is a sequence of discrete random variables indexed by time. If $x_0, \ldots, x_{N-1}$ are independent identically distributed (IID), then the entropy of the process grows linearly with N at a rate $\mathcal{H}(x_0)$. More generally, the entropy grows linearly with N at the so-called *entropy rate*

$$\mathcal{H} = \lim_{N\to\infty} \frac{1}{N}\mathcal{H}(x_0, \ldots, x_{N-1}). \tag{1.32}$$

If the process is stationary, then the entropy rate can be shown to equal

$$\mathcal{H} = \lim_{N\to\infty} \mathcal{H}(x_N|x_0, \ldots, x_{N-1}). \tag{1.33}$$

When the distribution of the process is continuous rather than discrete, the same definitions apply to the differential entropy and a classification that proves useful in the context of fading channels can be introduced: a process is said to be *nonregular* if its present value is perfectly predictable from noiseless observations of the entire past, while the process is *regular* if its present value cannot be perfectly predicted from noiseless observations of the entire past [68]. In terms of the differential entropy rate $\mathfrak{h}$, the process is regular if $\mathfrak{h} > -\infty$ and nonregular otherwise.

1.4 Information dependence

Although it could be—and has been—argued that Shannon imported the concept of entropy from statistical mechanics, where it was utilized to measure the uncertainty surrounding the state of a physical system, this was but a step toward something radically original: the idea of measuring with information (e.g., with bits) the interdependence among different quantities. In the context of a communication channel, this idea opens the door to relating transmit and receive signals, a relationship from which the capacity ultimately emerges.

1.4.1 Relative entropy

Consider two PMFs, $p(\cdot)$ and $q(\cdot)$. If the latter is nonzero over the support of the former, then their *relative entropy* is defined as

$$\mathcal{D}(p||q) = \sum_{\mathrm{x}} p(\mathrm{x}) \log_2 \frac{p(\mathrm{x})}{q(\mathrm{x})} \tag{1.34}$$

$$= \mathbb{E}\left[\log_2 \frac{p(x)}{q(x)}\right] \tag{1.35}$$

where the expectation is over $p(\cdot)$. The relative entropy, also referred to as the *Kullback–Leibler divergence* or the *information divergence*, can be interpreted as a measure of the similarity of $p(\cdot)$ and $q(\cdot)$. Note, however, that it is not symmetric, i.e., $\mathcal{D}(p||q) \neq \mathcal{D}(q||p)$ in general. It is a nonnegative quantity, and it is zero if and only if $p(\mathrm{x}) = q(\mathrm{x})$ for every x.

Similarly, for two PDFs $f(\cdot)$ and $g(\cdot)$,

$$\mathcal{D}(f||g) = \int f(\mathrm{x}) \log_2 \frac{f(\mathrm{x})}{g(\mathrm{x})}\, \mathrm{dx}. \tag{1.36}$$