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

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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

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

  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.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

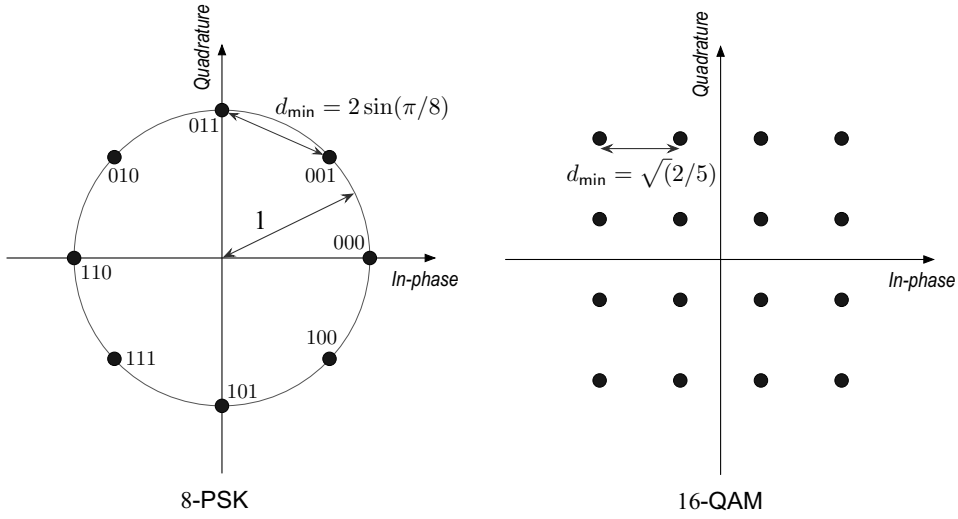

Fig. 1.1 Unit-variance 8-PSK and 16-QAM constellations.

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

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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 $p_{x_0x_1}(\cdot,\cdot)$ and marginals $p_{x_0}(\cdot)$ and $p_{x_1}(\cdot)$. The joint entropy of x_0 and x_1 is

$$\mathcal{H}(x_0, x_1) = -\sum_{\mathrm{x}_0} \sum_{\mathrm{x}_1} p_{x_0x_1}(\mathrm{x}_0, \mathrm{x}_1) \log_2 p_{x_0x_1}(\mathrm{x}_0, \mathrm{x}_1) \tag{1.10}$$

$$= -\mathbb{E}\big[\log_2 p_{x_0x_1}(x_0, x_1)\big]. \tag{1.11}$$

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

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

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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),

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

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

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

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

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),

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

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 dis-

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

tributed (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 $g(\cdot)$,

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

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

1.4.2 Mutual information

A quantity that lies at the heart of information theory is the *mutual information* between two or more random variables. Although present already in Shannon's original formulation [58], the mutual information did not acquire its current name until years later [67, 69]. Given two random variables s and y, the mutual information between them, denoted by $I(s;y)$, quantifies the reduction in uncertainty about the value of s that occurs when y is observed, and vice versa. The mutual information is symmetric and thus $I(s;y) = I(y;s)$. Put in the simplest terms, the mutual information measures the information that one random variable contains about another. As one would expect, $I(s;y)$ is nonnegative, equaling zero if and only if s and y are independent. At the other extreme, $I(s;y)$ cannot exceed the uncertainty contained in either s or y.

For discrete random variables, the mutual information can be computed on the basis of entropies as

$$I(s;y) = \mathcal{H}(s) - \mathcal{H}(s|y) \tag{1.37}$$

$$= \mathcal{H}(y) - \mathcal{H}(y|s), \tag{1.38}$$

or also as the information divergence between the joint PMF of s and y, on the one hand, and the product of their marginals on the other, i.e.,

$$I(s;y) = \mathcal{D}(p_{sy}||p_s p_y) \tag{1.39}$$

$$= \sum_{\mathrm{s}} \sum_{\mathrm{y}} p_{sy}(\mathrm{s},\mathrm{y}) \, \log_2 \frac{p_{sy}(\mathrm{s},\mathrm{y})}{p_s(\mathrm{s})\, p_y(\mathrm{y})} \tag{1.40}$$

$$= \sum_{\mathrm{s}} \sum_{\mathrm{y}} p_{sy}(\mathrm{s},\mathrm{y}) \, \log_2 \frac{p_{y|s}(\mathrm{y}|\mathrm{s})}{p_y(\mathrm{y})}. \tag{1.41}$$

Recalling that the information divergence measures the similarity between distributions, the intuition behind (1.39) is as follows: if the joint distribution is "similar" to the product of the marginals, it must be that s and y are essentially independent and thus one can hardly inform about the other. Conversely, if the joint and marginal distributions are "dissimilar," it must be that s and y are highly dependent and thus one can provide much information about the other.

For continuous random variables, relationships analogous to (1.37) and (1.39) apply, precisely

$$I(s;y) = \mathfrak{h}(s) - \mathfrak{h}(s|y) \tag{1.42}$$

$$= \mathfrak{h}(y) - \mathfrak{h}(y|s) \tag{1.43}$$

and

$$I(s;y) = \mathcal{D}(f_{sy}||f_s f_y) \tag{1.44}$$

$$= \iint f_{sy}(\mathrm{s},\mathrm{y}) \log_2 \frac{f_{sy}(\mathrm{s},\mathrm{y})}{f_s(\mathrm{s}) f_y(\mathrm{y})} \, \mathrm{ds}\, \mathrm{dy} \tag{1.45}$$

$$= \iint f_{sy}(\mathrm{s},\mathrm{y}) \log_2 \frac{f_{y|s}(\mathrm{y}|\mathrm{s})}{f_y(\mathrm{y})} \, \mathrm{ds}\, \mathrm{dy}. \tag{1.46}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

In contrast with the differential entropies, which cannot be obtained as the limit of the entropy of the discretized variables, $I(s;y)$ can be perfectly computed as the limit of the mutual information between discretized versions of s and y. Albeit the entropies and conditional entropies of the discretized variables diverge, their differences remain well behaved.

Since, because of their translation invariance, the entropies and differential entropies are not influenced by the mean of the corresponding random variables, neither is the mutual information. In the derivations that follow, therefore, we can restrict ourselves to zero-mean distributions.

As shorthand notation, we introduce the informal term *Gaussian mutual information* to refer to the function $\mathcal{I}(\rho) = I(s;\sqrt{\rho}s+z)$ when z is complex Gaussian and ρ is a fixed parameter. If we interpret s as a transmit symbol and z as noise, then ρ plays the role of the signal-to-noise ratio (SNR) and the mutual information between s and the received symbol $\sqrt{\rho}s+z$ is given by $\mathcal{I}(\rho)$. Because of this interpretation, attention is paid to how $\mathcal{I}(\rho)$ behaves for small and large ρ, in anticipation of low-SNR and high-SNR analyses later on. We examine these specific behaviors by expanding $\mathcal{I}(\rho)$ and making use of the Landau symbols $\mathcal{O}(\cdot)$ and $o(\cdot)$ described in Appendix F.

Example 1.7 (Gaussian mutual information for a complex Gaussian scalar)

Let us express, as a function of ρ, the mutual information between s and $y=\sqrt{\rho}s+z$ with s and z independent standard complex Gaussians, i.e., $s\sim\mathcal{N}_{\mathbb{C}}(0,1)$ and $z\sim\mathcal{N}_{\mathbb{C}}(0,1)$. Noting that $y\sim\mathcal{N}_{\mathbb{C}}(0,1+\rho)$ and $y|s\sim\mathcal{N}_{\mathbb{C}}(\sqrt{\rho}s,1)$, and invoking Example 1.5,

$$\begin{aligned}\mathcal{I}(\rho) &= I\big(s;\sqrt{\rho}s+z\big) && (1.47)\\ &= \mathfrak{h}\big(\sqrt{\rho}s+z\big)-\mathfrak{h}\big(\sqrt{\rho}s+z\,|s\big) && (1.48)\\ &= \mathfrak{h}\big(\sqrt{\rho}s+z\big)-\mathfrak{h}(z) && (1.49)\\ &= \log_2\big(\pi e(1+\rho)\big)-\log_2(\pi e) && (1.50)\\ &= \log_2(1+\rho). && (1.51)\end{aligned}$$

For small ρ,

$$\mathcal{I}(\rho)=\left(\rho-\frac{1}{2}\,\rho^2\right)\log_2 e+o(\rho^2), \tag{1.52}$$

which turns out to apply in rather wide generality: provided that s is proper complex as per the definition in Appendix C.1, its second-order expansion of $\mathcal{I}(\cdot)$ abides by (1.52) [64].

In turn, for complex Gaussian s and large ρ,

$$\mathcal{I}(\rho)=\log_2\rho+\mathcal{O}\left(\frac{1}{\rho}\right). \tag{1.53}$$

Example 1.8 (Gaussian mutual information for ∞-PSK)

Let us reconsider Example 1.7, only with s drawn from the ∞-PSK distribution defined in Section 1.2. The corresponding mutual information cannot be expressed in closed form, but meaningful expansions can be given. For low ρ, (1.52) holds verbatim because of the

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

properness of ∞-PSK. In turn, the high-ρ behavior is [70]

$$\mathcal{I}^{\infty\text{-PSK}}(\rho) = \frac{1}{2}\log_2\rho + \frac{1}{2}\log_2\left(\frac{4\pi}{e}\right) + \mathcal{O}\left(\frac{1}{\rho}\right). \tag{1.54}$$

Example 1.9 (Gaussian mutual information for ∞-QAM)

Let us again reconsider Example 1.7, this time with s drawn from the ∞-QAM distribution. As with ∞-PSK, the mutual information cannot be expressed in closed form, but meaningful expansions can be found. For low ρ, and since s is proper complex, (1.52) holds whereas for high ρ [71]

$$\mathcal{I}^{\infty\text{-QAM}}(\rho) = \log_2\rho - \log_2\left(\frac{\pi e}{6}\right) + \mathcal{O}\left(\frac{1}{\rho}\right). \tag{1.55}$$

With respect to the high-ρ mutual information in (1.53), ∞-QAM suffers a power penalty of $\frac{\pi e}{6}|_{\text{dB}} = 1.53$ dB, where we have introduced the notation $a|_{\text{dB}} = 10\log_{10} a$ that is to appear repeatedly in the sequel.

Example 1.10 (Gaussian mutual information for BPSK)

Let us reconsider Example 1.7 once more, now with s drawn from a BPSK distribution, i.e., $s = \pm 1$. The PDF of y equals

$$f_y(\mathrm{y}) = \frac{1}{2\pi}\left(e^{-|\mathrm{y}+\sqrt{\rho}|^2} + e^{-|\mathrm{y}-\sqrt{\rho}|^2}\right) \tag{1.56}$$

whereas $y|s \sim \mathcal{N}_{\mathbb{C}}(\sqrt{\rho}s, 1)$. Thus,

$$\begin{aligned}\mathcal{I}^{\text{BPSK}}(\rho) &= \mathfrak{h}(y) - \mathfrak{h}(y|s) && (1.57)\\ &= -\int f_y(\mathrm{y})\,\log_2 f_y(\mathrm{y})\,\mathrm{d}\mathrm{y} - \log_2(\pi e) && (1.58)\\ &= 2\rho\log_2 e - \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-\xi^2}\log_2\cosh\left(2\rho - 2\sqrt{\rho}\xi\right)\,\mathrm{d}\xi && (1.59)\end{aligned}$$

where, by virtue of the real nature of s, the integration over the complex plane in (1.58) reduces, after some algebra, to the integral on the real line in (1.59). In turn, this integral can be alternatively expressed as the series [72, example 4.39]

$$\begin{aligned}\mathcal{I}^{\text{BPSK}}(\rho) = 1 + \Bigg[&(4\rho-1)\,Q\left(\sqrt{2\rho}\right) - \sqrt{\frac{4\rho}{\pi}}\,e^{-\rho}\\ &+ \sum_{\ell=1}^{\infty}\frac{(-1)^\ell}{\ell(\ell+1)}\,e^{4\ell(\ell+1)\rho}\,Q\left((2\ell+1)\sqrt{2\rho}\right)\Bigg]\log_2 e\end{aligned} \tag{1.60}$$

where $Q(\cdot)$ is the Gaussian Q-function (see Appendix E.5).

For small ρ, using the identity

$$\log_e\cosh\left(2\rho - 2\sqrt{\rho}\xi\right) = 2\xi^2\rho - 4\xi\rho^{3/2} + \left(2 - \frac{4\xi^4}{3}\right)\rho^2 + o(\rho^2) \tag{1.61}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

we can reduce (1.59) to

$$\mathcal{I}^{\text{BPSK}}(\rho) = \left(\rho - \rho^2\right) \log_2 e + o(\rho^2) \tag{1.62}$$

whereas, for large ρ [71]

$$\mathcal{I}^{\text{BPSK}}(\rho) = 1 - \frac{e^{-\rho}}{\sqrt{\rho/\pi}} + \epsilon \tag{1.63}$$

with $\log \epsilon = o(\rho)$.

Example 1.11 (Gaussian mutual information for QPSK)

Since QPSK amounts to two BPSK constellations in quadrature with the power evenly divided between them,

$$\mathcal{I}^{\text{QPSK}}(\rho) = 2\,\mathcal{I}^{\text{BPSK}}\left(\frac{\rho}{2}\right). \tag{1.64}$$

Another way to see this equivalence is by considering that, given a BPSK symbol, we can add a second BPSK symbol of the same energy in quadrature without either BPSK symbol perturbing the other. The mutual information doubles while twice the energy is spent, i.e., $\mathcal{I}^{\text{QPSK}}(2\rho) = 2\,\mathcal{I}^{\text{BPSK}}(\rho)$.

Discrete constellations beyond QPSK, possibly nonequiprobable, are covered by the following example.

Example 1.12 (Gaussian mutual information for an arbitrary constellation)

Let s be a zero-mean unit-variance discrete random variable taking values in $\mathrm{s}_0, \ldots, \mathrm{s}_{M-1}$ with probabilities $p_0, \ldots, p_{M-1}$. This subsumes M-PSK, M-QAM, and any other discrete constellation. The PDF of $y = \sqrt{\rho}s + z$ equals

$$f_y(\mathrm{y}) = \frac{1}{\pi} \sum_{m=0}^{M-1} p_m\, e^{-|\mathrm{y} - \sqrt{\rho}\,\mathrm{s}_m|^2} \tag{1.65}$$

whereas $y|s \sim \mathcal{N}_{\mathbb{C}}(\sqrt{\rho}s, 1)$. Thus,

$$\mathcal{I}^{M\text{-ary}}(\rho) = I\left(s; \sqrt{\rho}s + z\right) \tag{1.66}$$

$$= -\int f_y(\mathrm{y}) \log_2 f_y(\mathrm{y})\, \mathrm{d}\mathrm{y} - \log_2(\pi e) \tag{1.67}$$

with integration over the complex plane.

For low ρ, an arduous expansion of $f_y(\cdot)$ and the subsequent integration leads, provided that s is proper complex, again to (1.52). For high ρ, it can be shown [71] that

$$\mathcal{I}^{M\text{-ary}}(\rho) = \log_2 M - \epsilon \tag{1.68}$$

with

$$\log \epsilon = -\frac{d_{\min}^2}{4}\rho + o(\rho) \tag{1.69}$$

where, recall,

$$d_{\min} = \min_{k \neq \ell} |\mathrm{s}_k - \mathrm{s}_\ell| \tag{1.70}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

is the minimum distance between constellation points. The mutual information is capped at $\log_2 M$, as one would expect, and the speed at which this limit is approached for $\rho \to \infty$ is regulated by d_{min}.

The definition of mutual information extends also to vectors. For continuous random vectors $\boldsymbol{s}$ and $\boldsymbol{y}$, specifically,

$$I(\boldsymbol{s};\boldsymbol{y}) = \mathfrak{h}(\boldsymbol{y}) - \mathfrak{h}(\boldsymbol{y}|\boldsymbol{s}) \tag{1.71}$$
$$= \mathfrak{h}(\boldsymbol{s}) - \mathfrak{h}(\boldsymbol{s}|\boldsymbol{y}) \tag{1.72}$$
$$= \mathcal{D}(f_{\boldsymbol{s}\boldsymbol{y}}||f_{\boldsymbol{s}}f_{\boldsymbol{y}}). \tag{1.73}$$

Example 1.13 (Gaussian mutual information for a complex Gaussian vector)

Let $\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}$ where $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{s}})$ and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{z}})$ while $\boldsymbol{A}$ is a deterministic matrix. With $\boldsymbol{s}$ and $\boldsymbol{z}$ mutually independent, let us express $I(\boldsymbol{s};\boldsymbol{y})$ as a function of ρ. Since $\boldsymbol{y} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^* + \boldsymbol{R}_{\boldsymbol{z}})$ and $\boldsymbol{y}|\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}, \boldsymbol{R}_{\boldsymbol{z}})$, leveraging Example 1.6,

$$\mathcal{I}(\rho) = I\big(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}\big) \tag{1.74}$$
$$= \mathfrak{h}\big(\sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}\big) - \mathfrak{h}\big(\sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}\,|\boldsymbol{s}\big) \tag{1.75}$$
$$= \mathfrak{h}\big(\sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}\big) - \mathfrak{h}(\boldsymbol{z}) \tag{1.76}$$
$$= \log_2 \det\big(\pi e\,(\rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^* + \boldsymbol{R}_{\boldsymbol{z}})\big) - \log_2 \det\big(\pi e \boldsymbol{R}_{\boldsymbol{z}}\big) \tag{1.77}$$
$$= \log_2 \det\big(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\big). \tag{1.78}$$

For low ρ, using

$$\frac{\partial}{\partial\rho} \log_{\mathrm{e}} \det(\boldsymbol{I} + \rho\boldsymbol{B})\bigg|_{\rho=0} = \mathrm{tr}(\boldsymbol{B}) \tag{1.79}$$
$$\frac{\partial^2}{\partial\rho^2} \log_{\mathrm{e}} \det(\boldsymbol{I} + \rho\boldsymbol{B})\bigg|_{\rho=0} = -\mathrm{tr}\big(\boldsymbol{B}^2\big) \tag{1.80}$$

it is found that

$$\mathcal{I}(\rho) = \left[\mathrm{tr}\big(\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\big)\,\rho - \frac{1}{2}\,\mathrm{tr}\Big(\big(\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\big)^2\Big)\,\rho^2\right] \log_2 e + o\left(\rho^2\right) \tag{1.81}$$

whose applicability extends beyond complex Gaussian vectors to any proper complex vector $\boldsymbol{s}$. For high ρ, in turn, provided $\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}$ is nonsingular,

$$\mathcal{I}(\rho) = \min(N_{\boldsymbol{s}}, N_{\boldsymbol{y}}) \log_2 \rho + \log_2 \det\big(\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\big) + \mathcal{O}\left(\frac{1}{\rho}\right), \tag{1.82}$$

where $N_{\boldsymbol{s}}$ and $N_{\boldsymbol{y}}$ are the dimensions of $\boldsymbol{s}$ and $\boldsymbol{y}$, respectively.

Example 1.14 (Gaussian mutual information for a discrete vector)

Reconsider Example 1.13, only with $\boldsymbol{s}$ an $N_{\boldsymbol{s}}$-dimensional discrete complex random vector and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$. The vector $\boldsymbol{y} = [y_0 \cdots y_{N_{\boldsymbol{y}}-1}]^{\mathrm{T}}$ is $N_{\boldsymbol{y}}$-dimensional and hence $\boldsymbol{A}$ is $N_{\boldsymbol{y}} \times N_{\boldsymbol{s}}$. Each entry of $\boldsymbol{s}$ can take one of M possible values and therefore $\boldsymbol{s}$ can take one

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

of $M^{N_{\boldsymbol{s}}}$ values, $\mathbf{s}_0, \ldots, \mathbf{s}_{M^{N_{\boldsymbol{s}}}-1}$, with probabilities $p_0, \ldots, p_{M^{N_{\boldsymbol{s}}}-1}$. With a smattering of algebra, the PDF of $\boldsymbol{y}$ can be found to be

$$f_{\boldsymbol{y}}(\mathbf{y}) = \frac{1}{\pi^{N_{\boldsymbol{y}}}} \sum_{m=0}^{M^{N_{\boldsymbol{s}}}-1} p_m \, e^{-\|\mathbf{y}-\sqrt{\rho}\boldsymbol{A}\mathbf{s}_m\|^2} \tag{1.83}$$

whereas $\boldsymbol{y}|\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}, \boldsymbol{I})$. Then,

$$\mathcal{I}^{M\text{-ary}}(\rho) = \mathfrak{h}(\boldsymbol{y}) - \log_2 \det(\pi e \boldsymbol{I}) \tag{1.84}$$

$$= -\int \ldots \int f_{\boldsymbol{y}}(\mathbf{y}) \log_2 f_{\boldsymbol{y}}(\mathbf{y}) \, \mathrm{d}y_0 \cdots \mathrm{d}y_{N_{\boldsymbol{y}}-1} - N_{\boldsymbol{y}} \log_2(\pi e). \tag{1.85}$$

The number of terms in the summation in (1.83) grows exponentially with $N_{\boldsymbol{s}}$, whereas the integration in (1.85) becomes unwieldy as $N_{\boldsymbol{y}}$ grows large. Except in very special cases, numerical integration techniques are called for [73]. Alternatively, it is possible to resort to approximations of the integral of a Gaussian function multiplied with an arbitrary real function [74].

For low ρ, and as long as $\boldsymbol{s}$ is proper complex, $\mathcal{I}^{M\text{-ary}}(\rho)$ expands as in (1.81) [75]. For $\rho \to \infty$, in turn, $\mathcal{I}(\rho) \to N_{\boldsymbol{s}} \log_2 M$.

Like the entropy and differential entropy, the mutual information satisfies a chain rule, specifically

$$I(x_0, \ldots, x_{N-1}; y) = \sum_{n=0}^{N-1} I(x_n; y | x_0, \ldots, x_{n-1}), \tag{1.86}$$

which applies verbatim to vectors.

1.5 Reliable communication

1.5.1 Information-theoretic abstraction

One of the enablers of Shannon's ground-breaking work was his ability to dissect a problem into simple pieces, which he could solve and subsequently put together to construct the full solution to the original problem [76]. This ability was manifest in the extremely simple abstraction of a communication link from which he derived quantities of fundamental interest, chiefly the capacity. This simple abstraction, echoed in Fig. 1.2, indeed contained all the essential ingredients.

- An encoder that parses the bits to be communicated into *messages* containing N_{bits}, meaning that there are $2^{N_{\mathsf{bits}}}$ possible such messages, and then maps each message onto a *codeword* consisting of N unit-power complex symbols, $s[0], \ldots, s[N-1]$. The codeword is subsequently amplified, subject to the applicable constraints, into the transmit signal $x[0], \ldots, x[N-1]$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

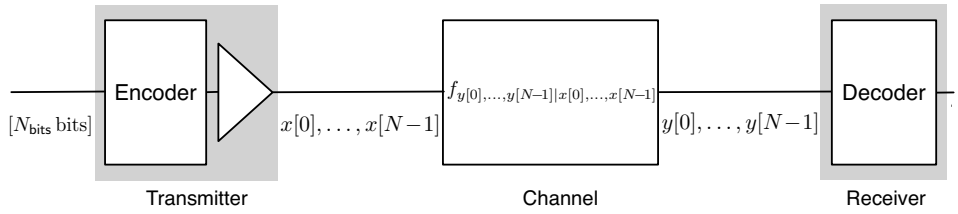

Fig. 1.2 Basic abstraction of a communication link.

- The channel, viewed as the random transformation experienced by the transmit signal and fully described, from such viewpoint, by the conditional probability of its output given every possible input, the *channel law* $f_{y[0],\ldots,y[N-1]|x[0],\ldots,x[N-1]}(\cdot)$. Accounting for the power amplification, and for any other transformation involved in converting the codeword into the transmit signal, $f_{y[0],\ldots,y[N-1]|s[0],\ldots,s[N-1]}(\cdot)$ readily derives from $f_{y[0],\ldots,y[N-1]|x[0],\ldots,x[N-1]}(\cdot)$.
- A decoder that, cognizant of the channel law, maps its observation of the channel output $y[0]\ldots, y[N-1]$ onto a guess of which codeword, and thereby which of the $2^{N_{\text{bits}}}$ possible messages, has been transmitted.

The functions used by the encoder and decoder to map messages (N_{bits} bits) onto codewords (N symbols) define the channel code. The set of all possible codewords is termed a *codebook* and the rate of information being transmitted (in bits/symbol) is N_{bits}/N.

Two observations are in order with respect to the foregoing abstraction.

(1) The abstraction is discrete in time, yet actual channels are continuous in time. As long as the channel is bandlimited, though, the sampling theorem ensures that a discrete-time equivalent can be obtained [77]. This discretization is tackled in Chapter 2 and its implications for time-varying channels are further examined in Section 3.4.5. To reconcile this discrete-time abstraction with the continuous-time nature of actual channels, the "channel" in Fig. 1.2 can be interpreted as further encompassing the transmit and receive filters, $g_{\text{tx}}(\cdot)$ and $g_{\text{rx}}(\cdot)$, plus a sampling device; this is reflected in Fig. 1.3.

(2) The abstraction is digital, i.e., the information to be transmitted is already in the form of bits. The digitization of information, regardless of its nature, underlies all modern forms of data storage and transmission, and is yet again a legacy of Shannon's work. We do not concern ourselves with the origin and meaning of the information, or with how it was digitized. Furthermore, we regard the bits to be transmitted as IID, sidestepping the source encoding process that removes data redundancies and dependencies before transmission as well as the converse process that reintroduces them after reception.

For the sake of notational compactness, we introduce vector notation for time-domain sequences (and in other chapters also for frequency-domain sequences). And, to distinguish these vectors from their space-domain counterparts, we complement the bold font types

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

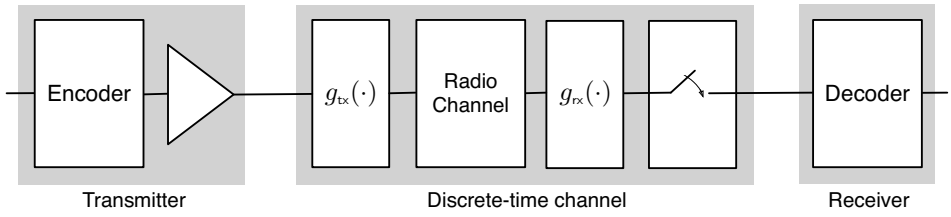

Fig. 1.3 Basic abstraction of a communication link, including the discrete-to-continuous and continuous-to-discrete interfaces.

with an overbar. The sequence $s[0], \ldots, s[N-1]$, for instance, is assembled into the vector

$$\bar{\boldsymbol{s}} = \begin{bmatrix} s[0] \\ \vdots \\ s[N-1] \end{bmatrix}. \tag{1.87}$$

The channel law $f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\cdot)$ is a key element in the computation of the capacity, and the mutual information between $s[0], \ldots, s[N-1]$ and $y[0], \ldots, y[N-1]$ can be expressed as a function thereof. Recalling (1.46), we can write

$$I\big(\bar{\boldsymbol{s}}; \bar{\boldsymbol{y}}\big) = \mathbb{E}\left[\log_2 \frac{f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}})}{f_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{y}})}\right] \tag{1.88}$$

$$= \mathbb{E}\left[\log_2 \frac{f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}})}{\frac{1}{2^{N_{\text{bits}}}} \sum_{m=0}^{2^{N_{\text{bits}}}-1} f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\boldsymbol{y}}|\bar{\mathbf{s}}_m)}\right], \tag{1.89}$$

where the expectations are over $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{y}}$. In (1.89), the $2^{N_{\text{bits}}}$ codewords have been assumed equiprobable, with $\bar{\mathbf{s}}_m$ the mth such codeword.

Example 1.15 (Channel law with Gaussian noise)

Let

$$y[n] = \sqrt{\rho}\,[\boldsymbol{A}]_{n,\mathsf{n}}\, s[\mathsf{n}] + z[n] \qquad n, \mathsf{n} = 0, \ldots, N-1 \tag{1.90}$$

or, more compactly, $\bar{\boldsymbol{y}} = \sqrt{\rho}\boldsymbol{A}\bar{\boldsymbol{s}} + \bar{\boldsymbol{z}}$ where $\boldsymbol{A}$ is a fixed matrix whose (n, n)th entry determines how the nth transmit symbol affects the nth received one, while $\bar{\boldsymbol{z}} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$. For this linear channel impaired by Gaussian noise,

$$f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}) = \frac{1}{\pi^N}\, e^{-\|\bar{\mathbf{y}} - \sqrt{\rho}\boldsymbol{A}\bar{\mathbf{s}}\|^2}. \tag{1.91}$$

If the channel law factors as $f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\cdot) = \prod_{n=0}^{N-1} f_{y[n]|s[n]}(\cdot)$, meaning that its output at symbol n depends only on the input at symbol n, the channel is said to be *memoryless*. Then, there is no loss of optimality in having codewords with statistically independent entries [14] and thus $f_{\bar{\boldsymbol{s}}}(\cdot)$, and subsequently $f_{\bar{\boldsymbol{y}}}(\cdot)$, can also be factored as a product of

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

per-symbol marginals to obtain

$$I\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big) = \sum_{n=0}^{N-1} I\big(s[n];y[n]\big) \tag{1.92}$$

with

$$I(s[n];y[n]) = \mathbb{E}\left[\log_2 \frac{f_{y[n]|s[n]}(y[n]|s[n])}{f_{y[n]}(y[n])}\right]. \tag{1.93}$$

For channels both memoryless and stationary, we can drop the index n and write

$$I(s;y) = \mathbb{E}\left[\log_2 \frac{f_{y|s}(y|s)}{f_y(y)}\right] \tag{1.94}$$

$$= \mathbb{E}\left[\log_2 \frac{f_{y|s}(y|s)}{\sum_{m=0}^{M-1} f_{y|s}(y|\mathrm{s}_m)\, p_m}\right], \tag{1.95}$$

where (1.95) applies if the signal conforms to an M-point constellation; with those constellation points further equiprobable, $p_m = 1/M$ for $m = 0,\ldots,M-1$. This convenient formulation involving a single symbol is said to be *single-letter*. Conversely, the codeword-wise formulation that is needed for channels with memory such as the one in Example 1.15 is termed *nonsingle-letter*. Although the direct discretization of a wireless channel generally does not yield a memoryless law, with equalizing countermeasures at the receiver the effects of the memory can be reduced to a minimum (see Chapter 2). Moreover, with OFDM, the signals are structured such that their joint discretization in time and frequency ends up being basically memoryless. Altogether, most—but not all—settings in this book are memoryless.

Example 1.16 (Memoryless channel law with Gaussian noise)

Let

$$y[n] = \sqrt{\rho}\, s[n] + z[n] \qquad n = 0,\ldots,N-1 \tag{1.96}$$

where $z[0],\ldots,z[N-1]$ are IID with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$. For this memoryless channel impaired by Gaussian noise,

$$f_{y|s}(\mathrm{y}|\mathrm{s}) = \frac{1}{\pi}\, e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2}. \tag{1.97}$$

1.5.2 Capacity

In an arbitrary channel, not necessarily memoryless, the average probability of making an error when decoding a codeword equals

$$p_{\mathrm{e}} = \sum_{\mathsf{m}=0}^{2^{N_{\mathrm{bits}}}-1} \mathbb{P}[\hat{\mathsf{w}} \neq \mathsf{m}\,|\mathsf{w}=\mathsf{m}]\,\mathbb{P}[\mathsf{w}=\mathsf{m}] \tag{1.98}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

where w is the index of the codeword actually transmitted while $\hat{\mathsf{w}}$ is the index guessed by the decoder. With the codewords equiprobable, the above reduces to

$$p_{\mathrm{e}} = \frac{1}{2^{N_{\mathsf{bits}}}} \sum_{\mathsf{m}=0}^{2^{N_{\mathsf{bits}}}-1} \mathbb{P}[\hat{\mathsf{w}} \neq \mathsf{m} \,|\mathsf{w}=\mathsf{m}]. \tag{1.99}$$

We term p_{e} the *error probability*, noting that it can be alternatively referred to as *word error probability*, *block error probability*, or *frame error probability*. A rate of information N_{bits}/N (in bits/symbol) can be communicated reliably if there exists a code of such rate for which $p_{\mathrm{e}} \to 0$ as $N \to \infty$. Note that we do not require the error probability to be zero for arbitrary N, but only that it vanishes as $N \to \infty$. In the channels of interest to this text, error-free communication at positive rates is possible only asymptotically in the codeword length.

The capacity C (in bits/symbol) is then the highest rate achievable reliably and, once exceeded, the error probability rises rapidly [60, section 10.4]. Most importantly, if the channel is *information stable* then the capacity is the maximum mutual information between the transmit and receive sequences. The concept of information stability can be explained by means of the so-called information density

$$i\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big) = \log_2 \frac{f_{\bar{\boldsymbol{s}},\bar{\boldsymbol{y}}}\big(\bar{\mathbf{s}},\bar{\mathbf{y}}\big)}{f_{\bar{\boldsymbol{s}}}\big(\bar{\mathbf{s}}\big) f_{\bar{\boldsymbol{y}}}\big(\bar{\mathbf{y}}\big)}, \tag{1.100}$$

which is the quantity whose expectation, recalling (1.46), equals the mutual information. The channel is information stable if [78]

$$\lim_{N\to\infty} \frac{1}{N}\, i\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big) = \lim_{N\to\infty} \frac{1}{N}\, \mathbb{E}\Big[i\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big)\Big] \tag{1.101}$$

$$= \lim_{N\to\infty} \frac{1}{N}\, I\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big), \tag{1.102}$$

which means that the information density does not deviate (asymptotically) from the mutual information. Intuitively, this indicates that the information that $y[0],\ldots,y[N-1]$ conveys about $s[0],\ldots,s[N-1]$ is invariant provided that N is large enough. This seemingly abstract concept is best understood by examining specific manifestations of stable and unstable channels, such as the ones encountered later in the context of fading. For our purposes, it is enough to point out that a sufficient condition for information stability is that the channel be stationary and ergodic, conditions that, as reasoned in Chapter 3, are satisfied within a certain time horizon by virtually all wireless channels of interest. For a more general capacity formulation that encompasses channels that are not information stable, the reader is referred to [79, 80].

If the channel is stationary and ergodic, then [81],

$$C = \max_{\text{signal constraints}} \lim_{N\to\infty} \frac{1}{N}\, I\big(\bar{\boldsymbol{s}};\bar{\boldsymbol{y}}\big), \tag{1.103}$$

where the maximization is over the joint distribution of the unit-power codeword symbols $s[0],\ldots,s[N-1]$, with subsequent amplification subject to whichever constraints apply to the signal's power and/or magnitude (see Section 2.3.5).

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

Shannon originally dealt with channels not only stationary and ergodic, but also memoryless, in which case [58]

$$C = \max_{\text{signal constraints}} I(s; y), \tag{1.104}$$

with the maximum taken over the distribution of the unit-power variable s, and with the subsequent amplification subject to the applicable constraints. The capacity then entails optimizing the marginal distribution of the symbols that make up the codewords. Because of the memorylessness and stationarity of the channel, such symbols may be not only independent but IID and thus the optimization is over any one of them. In this case, the capacity admits a *single-letter* formulation.

As argued earlier, the mean of the symbols $s[0], \ldots, s[N-1]$ does not contribute to the mutual information. However, a nonzero-mean would increase the power of the transmit signal. It follows that, irrespective of the specific type of power constraint, the maximizations of mutual information invariably yield signals that are zero-mean and hence only zero-mean signals are contemplated throughout the book.

From C (in bits/symbol) and from the symbol period T, the bit rate R (in bits/s) that can be communicated reliably satisfies $R \leq C/T$. And, since the sampling theorem dictates that $1/T \leq B$ with B the (passband) bandwidth, we have that

$$\frac{R}{B} \leq C, \tag{1.105}$$

evidencing the alternative measure of C in bits/s/Hz, often preferred to bits/symbol.[1] With a capacity-achieving codebook and $1/T = B$ symbols/s, the inequality in (1.105) becomes (asymptotically) an equality. If the pulse shape induced by the transmit and receive filters $g_{\mathsf{tx}}(\cdot)$ and $g_{\mathsf{rx}}(\cdot)$ incurs a bandwidth larger than $1/T$, the resulting shortfall from capacity must be separately accounted for. Indeed, as discussed in Chapter 2, pulse shapes with a modicum of excess bandwidth are common to diminish the sensitivity to synchronization inaccuracies.

Throughout this text, we resist utilizing the term "capacity" to describe the performance for specific distributions of $s[0], \ldots, s[N-1]$ that may be of interest but that are not optimum in the sense of maximizing (1.103) or (1.104). Rather, we then apply the term "spectral efficiency" and the description R/B, reserving "capacity" and C for the highest value over all possible signal distributions.

1.5.3 Coding and decoding

Before proceeding, let us establish some further terminology concerning the probabilistic relationship over the channel.

- We have introduced $f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\cdot)$ as the channel law, a function of both the transmit codeword and the observation at the receiver. For a fixed codeword, this defines the distribution of

[1] Our conversion of bits/symbol to bits/s/Hz, perfectly sufficient for complex baseband symbols representing real passband signals, can be generalized to real baseband signals and to spread-spectrum signals through the notion of *Shannon bandwidth* [82].

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

$y[0], \ldots, y[N-1]$ given that such codeword is transmitted while, for a fixed observation, it defines the *likelihood function* of $s[0], \ldots, s[N-1]$.

- With the conditioning reversed, $f_{\bar{\boldsymbol{s}}|\bar{\boldsymbol{y}}}(\cdot)$ is the *posterior probability* of a codeword given the observation at the receiver.

Optimum decoding rules

To establish the decoding rule that minimizes p_{e}, let us rewrite (1.98) into [83]

$$p_{\mathrm{e}} = \sum_{\mathsf{m}=0}^{2^{N_{\mathrm{bits}}}-1} \mathbb{P}[\bar{\boldsymbol{y}} \notin \mathcal{R}_{\mathsf{m}} \,|\mathsf{w}=\mathsf{m}]\, \mathbb{P}[\mathsf{w}=\mathsf{m}] \tag{1.106}$$

$$= \sum_{\mathsf{m}=0}^{2^{N_{\mathrm{bits}}}-1} \Big(1 - \mathbb{P}[\bar{\boldsymbol{y}} \in \mathcal{R}_{\mathsf{m}} \,|\mathsf{w}=\mathsf{m}]\Big)\, \mathbb{P}[\mathsf{w}=\mathsf{m}] \tag{1.107}$$

$$= 1 - \sum_{\mathsf{m}=0}^{2^{N_{\mathrm{bits}}}-1} \int_{\mathcal{R}_{\mathsf{m}}} f_{\bar{\boldsymbol{s}},\bar{\boldsymbol{y}}}(\bar{\mathbf{s}}_{\mathsf{m}}, \bar{\mathbf{y}})\, \mathrm{d}\bar{\mathbf{y}} \tag{1.108}$$

$$= 1 - \sum_{\mathsf{m}=0}^{2^{N_{\mathrm{bits}}}-1} \int_{\mathcal{R}_{\mathsf{m}}} f_{\bar{\boldsymbol{s}}|\bar{\boldsymbol{y}}}(\bar{\mathbf{s}}_{\mathsf{m}}|\bar{\mathbf{y}})\, f_{\bar{\boldsymbol{y}}}(\bar{\mathbf{y}})\, \mathrm{d}\bar{\mathbf{y}} \tag{1.109}$$

where $\mathcal{R}_{\mathsf{m}}$ denotes the decision region associated with codeword m, that is, the set of observations $y[0], \ldots, y[N-1]$ being mapped by the receiver onto message m. The $2^{N_{\mathrm{bits}}}$ decision regions are disjoint. To minimize p_{e}, each term in (1.109) can be separately maximized. By inspection, the mth term is maximized by defining $\mathcal{R}_{\mathsf{m}}$ as the region that contains all observations $\bar{\mathbf{y}}$ for which the posterior probability $f_{\bar{\boldsymbol{s}}|\bar{\boldsymbol{y}}}(\bar{\mathbf{s}}_{\mathsf{m}}|\bar{\mathbf{y}})$ is maximum. The optimum decoding strategy is thus to select the most probable codeword given what has been observed, a rule that is naturally termed maximum a-posteriori (MAP).

Applying Bayes' theorem (see Appendix C.1.1),

$$f_{\bar{\boldsymbol{s}}|\bar{\boldsymbol{y}}}(\bar{\mathbf{s}}_{\mathsf{m}}|\bar{\mathbf{y}}) = \frac{f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}_{\mathsf{m}})\, f_{\bar{\boldsymbol{s}}}(\bar{\mathbf{s}}_{\mathsf{m}})}{f_{\bar{\boldsymbol{y}}}(\bar{\mathbf{y}})} \tag{1.110}$$

and, when the codewords are equiprobable,

$$f_{\bar{\boldsymbol{s}}|\bar{\boldsymbol{y}}}(\bar{\mathbf{s}}_{\mathsf{m}}|\bar{\mathbf{y}}) = \frac{f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}_{\mathsf{m}})}{2^{N_{\mathrm{bits}}}\, f_{\bar{\boldsymbol{y}}}(\bar{\mathbf{y}})}, \tag{1.111}$$

where the right-hand side denominator does not depend on m and is thus irrelevant to a maximization over m. It follows that, with equiprobable codewords, maximizing the posterior probability on the left-hand side is equivalent to maximizing the likelihood function on the right-hand side numerator. MAP decoding is then equivalent to maximum-likelihood (ML) decoding, which, faced with an observation $\bar{\mathbf{y}}$, guesses the message m that maximizes $f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}_{\mathsf{m}})$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

Example 1.17 (ML decoding rule with Gaussian noise)

Consider the channel with memory and Gaussian noise in Example 1.15. The likelihood function to maximize is

$$f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}_{\mathsf{m}}) = \frac{1}{\pi^N}\, e^{-\|\bar{\mathbf{y}}-\sqrt{\rho}\boldsymbol{A}\bar{\mathbf{s}}_{\mathsf{m}}\|^2} \tag{1.112}$$

and, because the logarithm is a monotonic function, the ensuing maximization yields the same result as the maximization of

$$\log_e f_{\bar{\boldsymbol{y}}|\bar{\boldsymbol{s}}}(\bar{\mathbf{y}}|\bar{\mathbf{s}}_{\mathsf{m}}) = -N\log_e \pi - \|\bar{\mathbf{y}} - \sqrt{\rho}\boldsymbol{A}\bar{\mathbf{s}}_{\mathsf{m}}\|^2 \tag{1.113}$$

whose first term is constant and so inconsequential to the maximization. The decision made by an ML decoder is thus the message m whose codeword $\bar{\mathbf{s}}_{\mathsf{m}}$ minimizes $\|\bar{\mathbf{y}} - \sqrt{\rho}\boldsymbol{A}\bar{\mathbf{s}}_{\mathsf{m}}\|^2$, i.e., the codeword $\bar{\mathbf{s}}_{\mathsf{m}}$ that induces the channel output $\sqrt{\rho}\boldsymbol{A}\bar{\mathbf{s}}_{\mathsf{m}}$ closest in Euclidean distance to the observation $\bar{\mathbf{y}}$. This rule is therefore termed *minimum-distance* (or *nearest-neighbor*) decoding.

Example 1.18 (ML decoding rule for a memoryless channel with Gaussian noise)

For

$$y[n] = \sqrt{\rho}\, s[n] + z[n] \qquad n = 0, \ldots, N-1 \tag{1.114}$$

with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$, the ML guess when the receiver observes $\mathrm{y}[0], \ldots, \mathrm{y}[N-1]$ is the codeword $\mathrm{s}[0], \ldots, \mathrm{s}[N-1]$ that minimizes $\sum_{n=0}^{N-1} \left|\mathrm{y}[n] - \sqrt{\rho}\,\mathrm{s}[n]\right|^2$.

From hard to soft decoding

In classic receivers of yore, the decoding rules were applied upfront on a symbol-by-symbol basis. From the observation of $y[n]$, a hard decision was made on the value of $s[n]$. This procedure, whereby the MAP or ML rules were applied to individual symbols, was regarded as the demodulation of the underlying constellation. Subsequently, the N hard decisions for $s[0], \ldots, s[N-1]$ were assembled and fed into a decoder, with two possible outcomes. If the block of hard decisions was a valid codeword, success was declared. Alternatively, some of the hard decisions were taken to be erroneous and an attempt was made, exploiting the algebraic structure of the code, to correct them by modifying the block into a valid codeword. In these receivers, then, the decoder was essentially a corrector for the mistakes made by the demodulator. Moreover, in making a hard decision on a given symbol, the demodulator was throwing away information that could have been valuable to the decoder when deciding on other symbols [83, 84].

The extreme instance of this approach is uncoded transmission, where the message bits are directly mapped onto a constellation at the transmitter and recovered via ML-based hard decision at the receiver. Each bit is then at the mercy of the channel experienced by the particular symbol in which it is transmitted, without the protection that being part of a

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

long codeword can afford. Only a strong SNR or a low spectral efficiency could guarantee certain reliability in this pre-Shannon framework.

Except when simplicity is the utmost priority or no latency can be tolerated, transmissions are nowadays heavily coded and decoders operate directly on $y[0], \ldots, y[N-1]$, avoiding any preliminary discarding of information.

Near-capacity coding

As far as the codebooks are concerned, the coding theorems that establish the capacity as the maximum mutual information rely on random coding arguments—championed by Shannon—whereby the codewords are constructed by drawing symbols at random from a to-be-optimized distribution. However, because such codebooks have no structure, their optimum decoding would require an exhaustive search through the $2^{N_{\mathsf{bits}}}$ codewords making up the codebook in order to find the one codeword that maximizes the MAP or ML criteria. This is an impossible task even for modest values of N_{bits}; with $N_{\mathsf{bits}} = 30$, a meager value by today's standards, the number of codewords is already over 1000 million. Thus, random coding arguments, while instrumental to establishing the capacity, do not provide viable ways to design practical codes for large N_{bits}. For decades after 1948, coding theorists concentrated on the design of codebooks with algebraic structures that could be decoded optimally with a complexity that was polynomial, rather than exponential, in the codeword length [85]. Then, in the 1990s, with the serendipitous discovery of *turbo codes* [86] and the rediscovery of low-density parity check (LDPC) codes—formulated by Robert Gallager in the 1960s but computationally unfeasible at that time—the emphasis shifted to codebook constructions that could be decoded *suboptimally* in an efficient fashion. Staggering progress has been made since, and today we have powerful codes spanning hundreds to thousands of symbols and operating very close to capacity. These codes offer the random-like behavior leveraged by coding theorems with a relatively simple inner structure; in particular, turbo codes are obtained by concatenating lower-complexity codes through a large pseudo-random interleaver.

A comprehensive coverage of codebook designs and decoding techniques is beyond the scope of this book, and the interested reader is referred to dedicated texts [83, 87]. Here, we mostly regard encoders and decoders as closed boxes and discuss how these boxes ought to be arranged and/or modified to fit the MIMO *transceivers* (our term to compactly subsume both transmitters and receivers) under consideration.

Signal-space coding versus binary coding

Thus far we have implicitly considered codes constructed directly over the signal alphabet, say the points of a discrete constellation. The art of constructing such codes is referred to as *signal-space coding* (or *coded modulation*). Practical embodiments of signal-space coding exist, chiefly the trellis-coded modulation (TCM) schemes invented by Ungerboeck in the 1970s [88]. Signal-space coding conforms literally to the diagram presented in Fig. 1.3.

As an alternative to signal-space coding, it is possible to first run the message bits through a binary encoder, which converts messages onto binary codewords; subsequently,

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

the coded bits are mapped onto symbols $s[0], \ldots, s[N-1]$ having the desired distribution, in what can be interpreted as a modulation of the constellation. This alternative is attractive because it keeps the signal distribution arbitrary while allowing the codebook to be designed over the simple and convenient binary alphabet. If the rate of the binary code is r message bits per coded bit and the spectral efficiency is R/B (in message b/s), the constellation must accommodate $\frac{1}{\mathsf{r}} R/B$ coded bits. When the M constellation points are equiprobable, the number of bits it can accommodate equals $\log_2 M$ and thus we can write

$$\frac{R}{B} = \mathsf{r} \log_2 M. \tag{1.115}$$

This expression suggests how the transmit rate can be controlled by adjusting r and M, a procedure that is explored in detail later in the book.

Fundamentally, there is no loss of optimality in implementing signal-space coding by mapping the output of a binary encoder onto constellation points as long as the receiver continues to decode as if those constellation points were the actual coding alphabet. Indeed, if we take a string of random bits, parse them onto groups, and map each such group to a constellation point, the resulting codeword is statistically equivalent to a codeword defined randomly on the constellation alphabet itself. At the transmitter end, therefore, coding and mapping can be separated with no penalty as long as the receiver performs joint demapping and decoding. Ironically, then, what defines signal-space coding is actually the signal-space *decoding*.

Rather than decoding on the signal alphabet, however, the preferred approach is to first demap the binary code from the constellation and then separately decode it by means of a binary decoder. However, to avoid the pitfalls of hard demodulation and prevent an early loss of information, what is fed to the decoder is not a hard decision on each bit but rather a soft value.

Soft-input binary decoding

Consider a bit b. We can characterize the probability that b is 0 or 1 directly via $\mathbb{P}[b=0]$ or $\mathbb{P}[b=1] = 1 - \mathbb{P}[b=0]$ but also, equivalently, through the ratio $\frac{\mathbb{P}[b=1]}{\mathbb{P}[b=0]}$ [89]. More conveniently (because products and divisions become simpler additions and subtractions), we may instead use a logarithmic version of this ratio, the so-called *L-value*

$$\mathrm{L}(b) = \log \frac{\mathbb{P}[b=1]}{\mathbb{P}[b=0]}, \tag{1.116}$$

where, as done with entropies and differential entropies, notation has been slightly abused by expressing $\mathrm{L}(\cdot)$ as a function of b when it is actually a function of its distribution. A positive L-value indicates that the bit in question is more likely to be a 1 than a 0, and vice versa for a negative value, with the magnitude indicating the confidence of the decision. An L-value close to zero indicates that the decision on the bit is unreliable.

Now denote by $b_\ell[n]$ the ℓth coded bit within $s[n]$. A soft demapper should feed to the binary decoder a value quantifying how close $b_\ell[n]$ is to being a 0 or a 1 in light of what the

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

receiver has observed, and that information can be conveyed through the posterior L-value

$$\mathrm{L_D}\big(b_\ell[n] \,|\, \bar{\boldsymbol{y}}\big) = \log \frac{\mathbb{P}\big[b_\ell[n] = 1 \,|\, \bar{\boldsymbol{y}}\big]}{\mathbb{P}\big[b_\ell[n] = 0 \,|\, \bar{\boldsymbol{y}}\big]}. \tag{1.117}$$

Applying Bayes' theorem, we have that

$$\mathbb{P}\big[b_\ell[n] = 0 \,|\, \bar{\boldsymbol{y}} = \bar{\mathrm{y}}\big] = \frac{f_{\bar{\boldsymbol{y}}|b_\ell[n]}(\bar{\mathrm{y}}|0)}{f_{\bar{\boldsymbol{y}}}(\bar{\mathrm{y}})}\, \mathbb{P}\big[b_\ell[n] = 0\big] \tag{1.118}$$

$$\mathbb{P}\big[b_\ell[n] = 1 \,|\, \bar{\boldsymbol{y}} = \bar{\mathrm{y}}\big] = \frac{f_{\bar{\boldsymbol{y}}|b_\ell[n]}(\bar{\mathrm{y}}|1)}{f_{\bar{\boldsymbol{y}}}(\bar{\mathrm{y}})}\, \mathbb{P}\big[b_\ell[n] = 1\big] \tag{1.119}$$

and thus

$$\mathrm{L_D}\big(b_\ell[n] \,|\, \bar{\boldsymbol{y}} = \bar{\mathrm{y}}\big) = \underbrace{\log \frac{\mathbb{P}[b_\ell[n] = 1]}{\mathbb{P}[b_\ell[n] = 0]}}_{\mathrm{L_A}(b_\ell[n])} + \underbrace{\log \frac{f_{\bar{\boldsymbol{y}}|b_\ell[n]}(\bar{\mathrm{y}}|1)}{f_{\bar{\boldsymbol{y}}|b_\ell[n]}(\bar{\mathrm{y}}|0)}}_{\mathrm{L_E}(b_\ell[n] \,|\, \bar{\boldsymbol{y}})}, \tag{1.120}$$

whose first term is whatever a-priori information the receiver may already have about $b_\ell[n]$; in the absence of any such information, $\mathrm{L_A}\big(b_\ell[n]\big) = 0$. The second term, in turn, captures whatever fresh information the demapper supplies about $b_\ell[n]$ in light of what the receiver observes. More precisely, $\mathrm{L_E}\big(b_\ell[n] \,|\, \bar{\boldsymbol{y}}\big)$ is the logarithm of the ratio of the likelihood function for $b_\ell[n]$ evaluated at its two possible values, hence it is a *log-likelihood ratio*. This convenient separation into a sum of two terms conveying old (or *intrinsic*) and new (or *extrinsic*) information is what makes L-values preferable to probabilities and sets the stage for iterative decoding schemes.

When the channel is memoryless, $s[n]$ does not influence received symbols other than $y[n]$ and thus $\mathrm{L_D}\big(b_\ell[n] \,|\, \bar{\boldsymbol{y}}\big) = \mathrm{L_D}\big(b_\ell[n] \,|\, y[n]\big)$. Dropping the symbol index, the single-letter L-value can then be written as $\mathrm{L_D}\big(b_\ell \,|\, y\big)$ and further insight can be gained. Assuming that the coded bits mapped to each codeword symbol are independent—a condition discussed in the next section—such that their probabilities can be multiplied, and that the signal conforms to the discrete constellation defined by $\mathrm{s}_0, \ldots, \mathrm{s}_{M-1}$,

$$\begin{aligned}
\mathrm{L_D}\big(b_\ell | y = \mathrm{y}\big) &= \log \frac{\mathbb{P}\big[b_\ell = 1 \,|\, y = \mathrm{y}\big]}{\mathbb{P}\big[b_\ell = 0 \,|\, y = \mathrm{y}\big]} && (1.121)\\
&= \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} p_{s|y}(\mathrm{s}_m|\mathrm{y})}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} p_{s|y}(\mathrm{s}_m|\mathrm{y})} && (1.122)\\
&= \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, p_m}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, p_m} && (1.123)\\
&= \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \mathbb{P}[b_\ell = 1] \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \mathbb{P}[b_\ell = 0] \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]} && (1.124)\\
&= \log \frac{\mathbb{P}[b_\ell = 1] \sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]}{\mathbb{P}[b_\ell = 0] \sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]} && (1.125)\\
&= \mathrm{L_A}(b_\ell) + \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = \ell'\text{th bit of } \mathrm{s}_m]} && (1.126)
\end{aligned}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

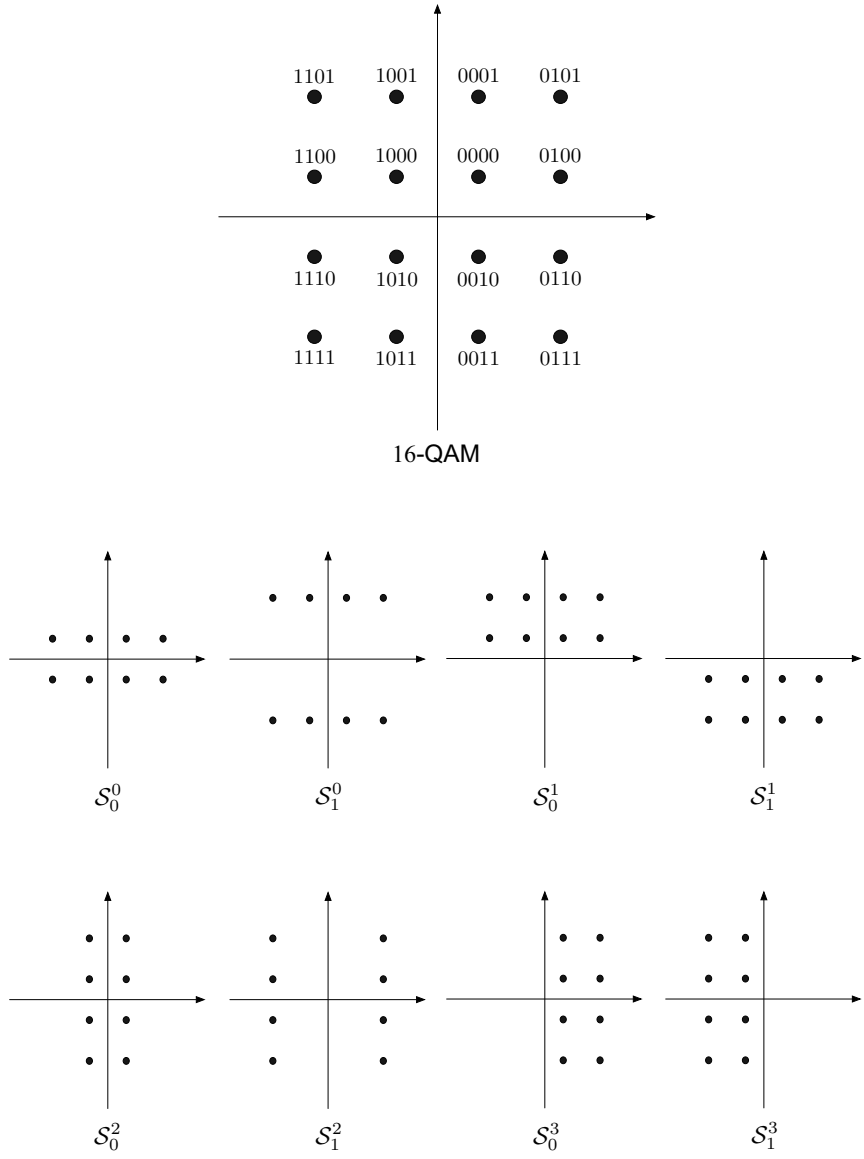

Fig. 1.4 Above, 16-QAM constellation with Gray mapping. Below, subsets $\mathcal{S}_0^\ell$ and $\mathcal{S}_1^\ell$ for $\ell = 0, 1, 2, 3$ with the bits ordered from right to left.

where $\mathcal{S}_0^\ell$ and $\mathcal{S}_1^\ell$ are the subsets of constellation points whose ℓth bit is 0 or 1, respectively (see Fig. 1.4). Dividing the second term's numerator and denominator by $\prod_{\ell' \neq \ell} \mathbb{P}[b_{\ell'} = 0]$, we further obtain

$$\mathrm{L_D}\big(b_\ell | y{=}\mathrm{y}\big) = \mathrm{L_A}(b_\ell) + \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' = \mathcal{B}_1^m} \frac{\mathbb{P}[b_{\ell'}=1]}{\mathbb{P}[b_{\ell'}=0]}}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m) \prod_{\ell' = \mathcal{B}_1^m} \frac{\mathbb{P}[b_{\ell'}=1]}{\mathbb{P}[b_{\ell'}=0]}} \tag{1.127}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

$$= \mathrm{L_A}(b_\ell) + \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \exp\Big(\sum_{\ell' \neq \ell, \ell' \in \mathcal{B}_1^m} \log \frac{\mathbb{P}[b_{\ell'}=1]}{\mathbb{P}[b_{\ell'}=0]}\Big)}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \exp\Big(\sum_{\ell' \neq \ell, \ell' \in \mathcal{B}_1^m} \log \frac{\mathbb{P}[b_{\ell'}=1]}{\mathbb{P}[b_{\ell'}=0]}\Big)}$$

$$= \mathrm{L_A}(b_\ell) + \underbrace{\log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \exp\Big(\sum_{\ell' \neq \ell, \ell' \in \mathcal{B}_1^m} \mathrm{L_A}(b_{\ell'})\Big)}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)\, \exp\Big(\sum_{\ell' \neq \ell, \ell' \in \mathcal{B}_1^m} \mathrm{L_A}(b_{\ell'})\Big)}}_{\mathrm{L_E}(b_\ell \,|\, y)} \tag{1.128}$$

where $\mathcal{B}_1^m$ is the subset of coded bits mapped to constellation point s_m that equal 1. The crucial insight here is that the extrinsic term $\mathrm{L_E}(b_\ell|y)$ depends on $\mathrm{L_A}(b_{\ell'})$ for $\ell' \neq \ell$ but not on $\mathrm{L_A}(b_\ell)$. Hence, $\mathrm{L_E}(b_\ell|y)$ contains the information that the demapper can gather about the ℓth bit in light of what the receiver observes and of whatever a-priori information may be available about the other bits within the same symbol. (Although assumed unconditionally independent, the coded bits may exhibit dependences when conditioned on y.) This opens the door to implementing iterative receivers, as detailed in the next section. For one-shot receivers, where the soft demapping takes place only once, there is no a-priori information and thus

$$\mathrm{L_D}\big(b_\ell \,|y{=}\mathrm{y}\big) = \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(\mathrm{y}|\mathrm{s}_m)}. \tag{1.129}$$

This log-likelihood ratio is what is fed into the decoder of a one-shot receiver.

Example 1.19 (Log-likelihood ratio for a memoryless channel with Gaussian noise)

For

$$y[n] = \sqrt{\rho}\, s[n] + z[n] \qquad n = 0, \ldots, N-1 \tag{1.130}$$

with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$ and equiprobable constellation points, the log-likelihood ratios fed into the decoder for each symbol are

$$\mathrm{L_D}\big(b_\ell \,|y{=}\mathrm{y}\big) = \log \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} e^{-|\mathrm{y}-\sqrt{\rho}\,\mathrm{s}_m|^2}}{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} e^{-|\mathrm{y}-\sqrt{\rho}\,\mathrm{s}_m|^2}} \qquad \ell = 0, \ldots, \log_2 M - 1. \tag{1.131}$$

From the log-likelihood ratios based on the receiver observations and from its own knowledge of the code structure, a decoder can then compute posterior L-values for each of the message bits, namely

$$\mathrm{L_D}\big(\mathsf{b}[\mathsf{n}] \,|\bar{\boldsymbol{y}}\big) = \log \frac{\mathbb{P}\big[\mathsf{b}[\mathsf{n}] = 1 \,|\bar{\boldsymbol{y}}\big]}{\mathbb{P}\big[\mathsf{b}[\mathsf{n}] = 0 \,|\bar{\boldsymbol{y}}\big]}, \tag{1.132}$$

where $\mathsf{b}[\mathsf{n}]$ for $\mathsf{n} = 0, \ldots, N_{\mathsf{bits}}-1$ are the bits making up the message. A processor producing these posterior L-values, a decidedly challenging task when the codewords are long, is referred to as an a-posteriori probability (APP) decoder, or also as a soft-input soft-output decoder. The APP decoder is one of the key engines of modern receivers, with different

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

flavors depending on the class of code, e.g., the Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm for convolutional codes [90]. In the case of turbo codes, where two constituent codes are concatenated, a breakdown of $\mathrm{L_D}\big(\mathsf{b}[\mathsf{n}]\,|\bar{\boldsymbol{y}}\big)$ into a-priori and extrinsic information about each message bit is the key to iterative decoding procedures whereby two APP decoders operate on the constituent codes exchanging information. Specifically, the extrinsic information generated by a first decoder is fed as a-priori information to the second decoder, allowing it to produce a better guess on the message bits as well as new extrinsic information for the first decoder, and so on. As the constituent codes are concatenated through an interleaver, the extrinsic information exchanged by the decoders must be interleaved and deinterleaved on each pass. This reduces the probability that the decoding process gets stuck in loops, and thus every iteration reduces the error probability with a handful of iterations sufficing to reach satisfactory levels. LDPC codes, although made up of a single block code, are also decoded iteratively.

Whichever the type of code, the sign of the L-values generated by an APP decoder for the message bits directly gives the MAP decisions,

$$\hat{\mathsf{b}}[\mathsf{n}] = \operatorname{sign}\Big(\mathrm{L_D}\big(\mathsf{b}[\mathsf{n}]\,|\bar{\boldsymbol{y}}\big)\Big) \qquad \mathsf{n} = 0, \ldots, N_{\mathsf{bits}} - 1. \tag{1.133}$$

Although it takes the entire codeword into account, an APP decoder maximizes the posterior probability on a bit-by-bit basis, thereby minimizing the average bit error probability rather than p_{e}. If the probabilities $\mathbb{P}\big[\hat{\mathsf{b}}[\mathsf{n}] = \mathsf{b}[\mathsf{n}]\,|\bar{\boldsymbol{y}}\big]$ for $\mathsf{n} = 0, \ldots, N_{\mathsf{bits}} - 1$ are conditionally independent given the observations, then the minimization of the bit error probability also minimizes p_{e}. Otherwise it need not, yet in practice it hardly matters: although there is no guarantee that capacity can then be achieved for $N \to \infty$, APP decoders perform superbly. In simple settings with turbo or LDPC codes, operation at the brink of capacity with satisfactorily small error probabilities has been demonstrated [91, 92].

1.5.4 Bit-interleaved coded modulation

As mentioned, there is no fundamental loss of optimality in the conjunction of binary encoding and constellation mapping: a signal-space decoder can recover from such signals as much information as could have been transmitted with a nonbinary code defined directly on the constellation alphabet. Is the same true when the receiver features a combination of soft demapping and binary decoding?

To shed light on this issue at a fundamental level, let us posit a stationary and memoryless channel as well as an M-point equiprobable constellation. In this setting, codewords with IID entries are optimum and thus bits mapped to distinct symbols can be taken to be independent. However, the channel does introduce dependencies among same-symbol bits. Even with the coded bits produced by the binary encoder being statistically independent, after demapping at the receiver dependencies do exist among the soft values for bits that traveled on the same symbol. Unaware, a binary decoder designed for IID bits ignores these dependencies and regards the channel as being memoryless, not only at a symbol level but further at a bit level [93]. Let us see how much information can be recovered under this premise.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

The binary-decoding counterpart to the memoryless mutual information in (1.94) and (1.95) is $\sum_{\ell=0}^{\log_2 M-1} I(b_\ell; y)$, and this binary-decoding counterpart can be put as a function of the channel law $f_{y|s}(\cdot)$ via [94]

$$\sum_{\ell=0}^{\log_2 M-1} I(b_\ell; y) = \sum_{\ell=0}^{\log_2 M-1} \mathbb{E}\left[\log_2 \frac{f_{y|b_\ell}(y|b_\ell)}{f_y(y)}\right] \tag{1.134}$$

$$= \sum_{\ell=0}^{\log_2 M-1} \frac{1}{2}\left(\mathbb{E}\left[\log_2 \frac{f_{y|b_\ell}(y|0)}{f_y(y)}\right] + \mathbb{E}\left[\log_2 \frac{f_{y|b_\ell}(y|1)}{f_y(y)}\right]\right) \tag{1.135}$$

$$= \sum_{\ell=0}^{\log_2 M-1} \frac{1}{2}\left(\mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(y|\mathrm{s}_m)\, \frac{1}{M/2}}{\sum_{m=0}^{M-1} f_{y|s}(y|\mathrm{s}_m)\, \frac{1}{M}}\right]\right.$$
$$\left. + \mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(y|\mathrm{s}_m)\, \frac{1}{M/2}}{\sum_{m=0}^{M-1} f_{y|s}(y|\mathrm{s}_m)\, \frac{1}{M}}\right]\right) \tag{1.136}$$

$$= \sum_{\ell=0}^{\log_2 M-1} \frac{1}{2}\left(\mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} f_{y|s}(y|\mathrm{s}_m)}{\frac{1}{2}\sum_{m=0}^{M-1} f_{y|s}(y|\mathrm{s}_m)}\right]\right.$$
$$\left. + \mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} f_{y|s}(y|\mathrm{s}_m)}{\frac{1}{2}\sum_{m=0}^{M-1} f_{y|s}(y|\mathrm{s}_m)}\right]\right), \tag{1.137}$$

where $\mathcal{S}_0^\ell$ and $\mathcal{S}_1^\ell$ are as defined in the previous section. In (1.136), the factors $1/(M/2)$ and $1/M$ correspond, respectively, to the probability of a constellation point s_m within the sets $\mathcal{S}_0^\ell$ and $\mathcal{S}_1^\ell$ (whose cardinality is $M/2$) and within the entire constellation (whose cardinality is M).

Whenever no dependencies among same-symbol coded bits are introduced by the channel, the binary decoder is not disregarding information and thus $\sum_\ell I(b_\ell; y) = I(s; y)$. This is the case with BPSK and QPSK, where a single coded bit is mapped to the in-phase and quadrature dimensions of the constellation. However, if each coded bit does acquire information about other ones within the same symbol, as is the case when multiple coded bits are mapped to the same dimension, then, with binary decoding not taking this information into account, $\sum_\ell I(b_\ell; y) < I(s; y)$.

Example 1.20

Consider a binary codeword mapped onto a QPSK constellation. The coded bits are parsed into pairs and the first and second bit within each pair are mapped, respectively, to the in-phase and quadrature components of the constellation, e.g., for a particular string 010010 within the binary codeword,

$$\cdots \underbrace{\underbrace{0}_{\mathrm{I}}\ \underbrace{1}_{\mathrm{Q}}}_{s[n-1]}\ \underbrace{\underbrace{0}_{\mathrm{I}}\ \underbrace{0}_{\mathrm{Q}}}_{s[n]}\ \underbrace{\underbrace{1}_{\mathrm{I}}\ \underbrace{0}_{\mathrm{Q}}}_{s[n+1]} \cdots \tag{1.138}$$

The resulting QPSK codeword $s[0], \ldots, s[N-1]$ is transmitted, contaminated by noise, and demapped back into a binary codeword at the receiver. The noise affects the bits as

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

follows:

$$\cdots \underbrace{0}_{\Re\{z[n-1]\}} \underbrace{1}_{\Im\{z[n-1]\}} \underbrace{0}_{\Re\{z[n]\}} \underbrace{0}_{\Im\{z[n]\}} \underbrace{1}_{\Re\{z[n+1]\}} \underbrace{0}_{\Im\{z[n+1]\}} \cdots \tag{1.139}$$

Provided the noise samples are independent and the real and imaginary parts of each noise sample are also mutually independent, no dependences are introduced among the bits. Even with binary coding and decoding, it is as if the code were defined on the QPSK alphabet itself and the performance limits are dictated by $I(s;y)$.

Example 1.21

Consider a binary codeword mapped onto a 16-QAM constellation. The bits are parsed into groups of four, from which the in-phase and quadrature components must be determined. Among the various possible mappings, suppose we choose to map the first two bits of each group to the in-phase component and the final two bits to the quadrature component, e.g., for a particular string 011010110100 within the binary codeword,

$$\cdots \underbrace{\underbrace{01}_{\text{I}} \underbrace{10}_{\text{Q}}}_{s[n-1]} \underbrace{\underbrace{10}_{\text{I}} \underbrace{11}_{\text{Q}}}_{s[n]} \underbrace{\underbrace{01}_{\text{I}} \underbrace{00}_{\text{Q}}}_{s[n+1]} \cdots \tag{1.140}$$

The resulting 16-QAM codeword $s[0], \ldots, s[N-1]$ is transmitted, contaminated by noise, and soft-demapped back into a binary codeword at the receiver. The noise affects the bits as

$$\cdots \underbrace{1}_{\Im\{z[n-1]\}} \underbrace{0}_{\Im\{z[n-1]\}} \underbrace{1}_{\Re\{z[n]\}} \underbrace{0}_{\Re\{z[n]\}} \underbrace{1}_{\Im\{z[n]\}} \underbrace{1}_{\Im\{z[n]\}} \underbrace{0}_{\Re\{z[n+1]\}} \underbrace{1}_{\Re\{z[n+1]\}} \cdots \tag{1.141}$$

and thus pairs of coded bits are subject to the same noise. While a signal-space decoder would take these additional dependences into account and be limited by $I(s;y)$, a binary decoder ignores them and is instead limited by $I(b_0;y) + I(b_1;y) + I(b_2;y) + I(b_3;y)$ where b_ℓ is the ℓth bit within each group of four.

Remarkably, the difference between $\sum_\ell I(b_\ell;y)$ and $I(s;y)$ is tiny provided the mapping of coded bits to constellation points is chosen wisely. Gray mapping, where nearest-neighbor constellation points differ by only one bit, has been identified as a robust and attractive choice [94, 95].

Once all the ingredients that lead to (1.137) are in place, only one final functionality is needed to have a complete information-theoretic abstraction of a modern wireless transmission chain: interleaving. Although symbol-level interleaving would suffice to break off bursts of poor channel conditions, bit-level interleaving has the added advantage of shuffling also the bits contained in a given symbol; if the interleaving were deep enough to push any bit dependencies beyond the confines of each codeword, then the gap between $\sum_\ell I(b_\ell;y)$ and $I(s;y)$ would be closed. Although ineffective for $N \to \infty$, and thus not captured by mutual information calculations, bit-level interleaving does improve the performance of actual codes with finite N.

The coalition of binary coding and decoding, bit-level interleaving, and constellation

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

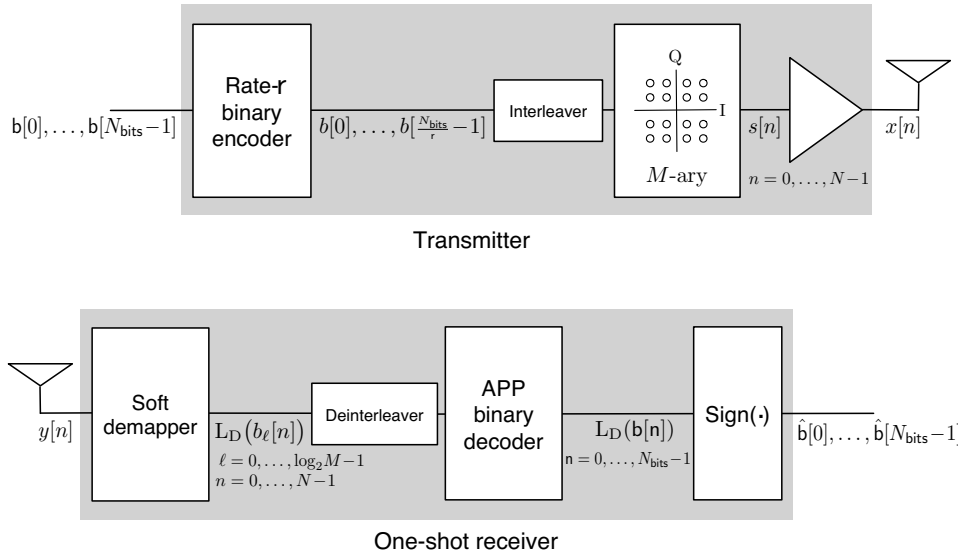

Fig. 1.5 BICM architecture with a one-shot receiver.

mappers and soft demappers constitutes the so-called *bit-interleaved coded modulation* (BICM) architecture, depicted in Fig. 1.5 and standard in wireless transceivers nowadays [96]. As mentioned, BICM is information-theoretically equivalent to signal-space coding for the cases of BPSK (single bit per symbol) and Gray-mapped QPSK (two quadrature bits per symbol). For higher-order constellations, even if the dependencies among same-symbol bits are not fully eradicated by interleaving and the receiver ignores them, it is only slightly inferior.

Example 1.22 (BICM mutual information in Gaussian noise)

Let $y = \sqrt{\rho}s + z$ with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$. Recalling from Example (1.16) the corresponding channel law, (1.137) specializes to

$$\sum_{\ell=0}^{\log_2 M-1} I(b_\ell; y) = \sum_{\ell=0}^{\log_2 M-1} \frac{1}{2}\left(\mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_0^\ell} e^{-|y-\sqrt{\rho}\,\mathrm{s}_m|^2}}{\frac{1}{2}\sum_{m=0}^{M-1} e^{-|y-\sqrt{\rho}\,\mathrm{s}_m|^2}}\right] + \mathbb{E}\left[\log_2 \frac{\sum_{\mathrm{s}_m \in \mathcal{S}_1^\ell} e^{-|y-\sqrt{\rho}\,\mathrm{s}_m|^2}}{\frac{1}{2}\sum_{m=0}^{M-1} e^{-|y-\sqrt{\rho}\,\mathrm{s}_m|^2}}\right]\right). \tag{1.142}$$

Example 1.23

For 16-QAM and 64-QAM constellations, compare the mutual information $I(s;y)$ in (1.95) against (1.142) with Gray mapping of bits to constellation points.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

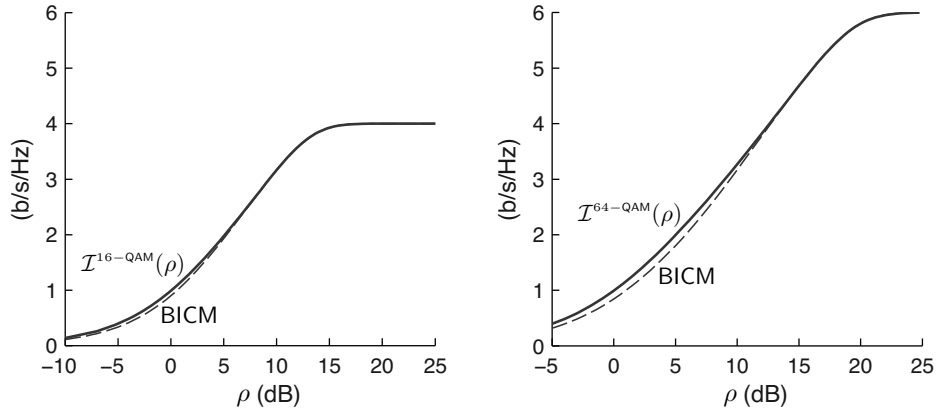

Fig. 1.6 Left-hand side, comparison between the mutual information in (1.95), shown in solid, and its BICM counterpart in (1.142), shown in dashed, for 16-QAM in Gaussian noise. Right-hand side, same comparison for 64-QAM.

Solution

The comparisons are shown in Fig. 1.6.

By incorporating iterative procedures at the receiver, most of the tiny loss incurred by a one-shot BICM receiver could be recovered at the price of decoding latency [97, 98]. In essence, an iterative BICM receiver can progressively learn the dependencies among bits given the observation of $y[0], \ldots, y[N-1]$, thereby closing the gap with signal-space coding. Although arguably not worthwhile given the tininess of this gap, the idea of iterative reception becomes more appealing in MIMO, where the gap broadens, and thus it is worthwhile that we introduce its basic structure here.

Figure 1.7 depicts an iterative BICM receiver, where the soft demapping is aided by a-priori information $\mathrm{L}_{\mathrm{A}}^{\mathsf{map}}(\cdot)$ about the coded bits. This improves the L-values $\mathrm{L}_{\mathrm{D}}^{\mathsf{map}}(\cdot)$ produced by the demapper, and the ensuing extrinsic information $\mathrm{L}_{\mathrm{E}}^{\mathsf{map}}(\cdot) = \mathrm{L}_{\mathrm{D}}^{\mathsf{map}}(\cdot) - \mathrm{L}_{\mathrm{A}}^{\mathsf{map}}(\cdot)$ is deinterleaved and fed to the APP decoder as $\mathrm{L}_{\mathrm{A}}^{\mathsf{cod}}(\cdot)$. The APP decoder then generates extrinsic information on the coded bits, $\mathrm{L}_{\mathrm{E}}^{\mathsf{cod}}(\cdot) = \mathrm{L}_{\mathrm{D}}^{\mathsf{cod}}(\cdot) - \mathrm{L}_{\mathrm{A}}^{\mathsf{cod}}(\cdot)$, which, properly interleaved, becomes the new a-priori information for the soft demapper, thereby completing an iteration. The APP decoder also generates L-values for the message bits and, once sufficient iterations have been run, the sign of these directly gives the final MAP decisions. Notice that only extrinsic L-values, representing newly distilled information, are passed around in the iterations. That prevents the demapper from receiving as a-priori information knowledge generated by itself in the previous iteration, and likewise for the decoder. Interestingly, with iterative reception, departing from Gray mapping is preferable so as to enhance the bit dependencies chased by the iterative process. Pushing things further in that direction, once could even consider multidimensional mappers operating, rather than on individual symbols, on groups of symbols [99, 100].

Altogether, the main take-away point from this section is the following: because of the

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

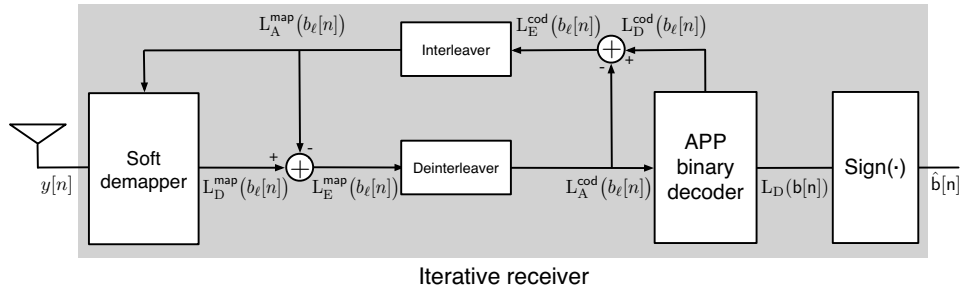

Fig. 1.7 BICM architecture with an iterative receiver.

coincidence of (1.137) and the actual mutual information for BPSK and QPSK, and their minute—and recoverable—difference for other constellations of interest, no distinction is customarily made between these quantities. This is also the principle followed in this text, where the performance limits of systems featuring BICM are investigated by means of the mutual information directly.

1.5.5 Finite-length codewords

For the most part we concern ourselves with the performance for $N \to \infty$, a stratagem that relies on this limit being representative of the performance of finite—but long—codewords. To substantiate this representativity, it is useful to briefly touch on an information-theoretic result that sheds light on the spectral efficiencies that can be fundamentally achieved when the length of the codewords is finite [101, 102]. Since error-free communication is generally not possible nonasymptotically, an acceptable error probability must be specified. If the acceptable codeword error probability is p_e, then, in many channels admitting a single-letter characterization it is possible to transmit at

$$\frac{R}{B} = C - \sqrt{\frac{V}{N}}\, Q^{-1}(p_\mathrm{e}) + \mathcal{O}\left(\frac{\log N}{N}\right), \tag{1.143}$$

where $Q(\cdot)$ is the Gaussian Q-function while V is the variance of the information density, i.e.,

$$V = \mathrm{var}\big[i(s;y)\big], \tag{1.144}$$

with s conforming to the capacity-achieving distribution. This pleasing result says that the backoff from capacity is approximately $\sqrt{V/N}\, Q^{-1}(p_\mathrm{e})$, which for codeword lengths and error probabilities of interest is generally small; quantitative examples for specific channels are given in Chapter 4. Hence, the capacity indeed retains its significance for finite—but long—codewords.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

Example 1.24

The turbo codes featured by 3G systems and by LTE, and the LDPC codes featured by the NR standard, have codeword lengths corresponding to values of N_{bits} that typically range from the few hundreds to the few thousands [103, chapter 12][104]. In certain simple channels, such codes can operate within a fraction of dB—in terms of SNR—of capacity.

Example 1.25

Over a bandwidth of $B = 20$ MHz, every 1000 codeword symbols incur a latency of $\frac{1000}{20 \cdot 10^6} = 0.05$ ms. If such bandwidth, typical of LTE, is shared by U users, then the latency is multiplied correspondingly. Given that LTE end-to-end latencies stand at about 10 ms, the contribution of coding to those latencies is minor.

For NR, the latency target is on the order of 1 ms [105, 106], but this reduction is to be accompanied by major increases in bandwidth and thus codeword lengths need not suffer major contractions.

The robustness of the capacity to finite codeword lengths, in conjunction with its computability for many relevant channels, renders it a quantity of capital importance. At the same time, for finite N and $p_{\mathrm{e}} > 0$, in addition to the transmit bit rate R and the ensuing spectral efficiency R/B, a companion quantity of interest is the *throughput* that measures the rate (in b/s) within the successfully decoded codewords; this throughput is given by $(1 - p_{\mathrm{e}})R$.

For small N, the expansion in (1.143) ceases to be precise and, in the limit of $N = 1$, the communication would become uncoded and every individual symbol would then be left at the mercy of the particular noise realization it experienced, without the protection that coding affords. The error probability would be comparatively very high and the throughput would suffer. Uncoded communication, the rule in times past, seems unnatural in the post-Shannon world and it is nowadays found only in systems priming simplicity.

1.5.6 Hybrid-ARQ

In relation to the codeword length, hybrid-ARQ has become established as an indispensable ingredient from 3G onwards. Blending channel coding with the traditional automatic repeat request (ARQ) procedure whereby erroneously received data are retransmitted, hybrid-ARQ turns the length and rate of the codewords into flexible—rather than fixed—quantities [107–109].

In a nutshell, hybrid-ARQ works as follows: when a received codeword is decoded incorrectly, rather than discarding it and receiving its retransmission anew as is the case in standard ARQ, the received codeword is stored and subsequently combined with the retransmission once it arrives at the receiver. This combination has a higher chance of successful decoding than either of the (re)transmissions individually. Moreover, the procedure may be repeated multiple times, until either decoding is indeed successful or the number

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

of retransmissions reaches a certain value and an error is declared. Two variants of hybrid-ARQ stand out:

- *Chase combining*, where every (re)transmission contains an identical copy of the codeword. The receiver simply adds its observations thereof, increasing the SNR with each new retransmission.
- *Incremental redundancy*, where every retransmission contains additional coded symbols that extend the codeword and lower its rate. Indeed, the result of appending each retransmission to the previous one(s) is a longer codeword that represents the original N_{bits} message bits with a larger number of symbols N.

Incremental redundancy is the most powerful incarnation of hybrid-ARQ and the one we implicitly refer to unless otherwise stated.

Example 1.26

How could incremental redundancy be implemented if every (re)transmission had length N and the maximum number of retransmissions were four?

Solution

The transmitter could generate a codeword of length $4N$ and then transmit N of those symbols, say every fourth one. The receiver, privy to the codebook and hybrid-ARQ scheme, would attempt decoding. If that failed, another set of N symbols could be sent and the receiver could again attempt decoding, this time the ensuing codeword of $2N$ symbols, and so on. If the final decoding with all $4N$ symbols failed, an error would be declared.

1.5.7 Extension to MIMO

How does the formulation of the channel capacity change with MIMO? In essence, the abstraction gets vectorized. Referring back to Fig. 1.2:

- The encoder parses the source bits into messages of N_{bits} and maps those onto codewords made up of N vector symbols, $\boldsymbol{s}[0], \ldots, \boldsymbol{s}[N-1]$. Each codeword is possibly transformed (e.g., via OFDM or MIMO precoding) and amplified into $\boldsymbol{x}[0], \ldots, \boldsymbol{x}[N-1]$ as per the applicable constraints.
- The channel, which connects every input (transmit antenna) with every output (receive antenna), is described by the conditional distribution $f_{\boldsymbol{y}[0],\ldots,\boldsymbol{y}[N-1]|\boldsymbol{x}[0],\ldots,\boldsymbol{x}[N-1]}(\cdot)$. With transformations and amplification accounted for, $f_{\boldsymbol{y}[0],\ldots,\boldsymbol{y}[N-1]|\boldsymbol{s}[0],\ldots,\boldsymbol{s}[N-1]}(\cdot)$ follows from $f_{\boldsymbol{y}[0],\ldots,\boldsymbol{y}[N-1]|\boldsymbol{x}[0],\ldots,\boldsymbol{x}[N-1]}(\cdot)$.
- The decoder maps every possible channel output, $\boldsymbol{y}[0], \ldots, \boldsymbol{y}[N-1]$, onto a guess of the original block of N_{bits} bits.

The encoder can be implemented as a true vector encoder, as a bank of parallel scalar encoders, or as a combination thereof, and the tradeoffs involved as well as the structure of the corresponding receivers are examined later in the text. At this point, we do not peek inside the encoder, but only posit that it produces codewords $\boldsymbol{s}[0], \ldots, \boldsymbol{s}[N-1]$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

Under information stability, the capacity is then

$$C = \max_{\text{power constraints}} \lim_{N\to\infty} \frac{1}{N} I\big(\boldsymbol{s}[0],\ldots,\boldsymbol{s}[N-1];\boldsymbol{y}[0],\ldots,\boldsymbol{y}[N-1]\big) \tag{1.145}$$

with maximization over the distribution of $\boldsymbol{s}[0],\ldots,\boldsymbol{s}[N-1]$, with the subsequent transformation and amplification having to respect the applicable constraints. If the channel is stationary and memoryless, then the transmit vector symbols are IID and the optimization in (1.145) becomes single-letter over a single vector symbol.

MIMO BICM

Recall that the norm in modern communication systems is to rely on powerful binary codes mapped to discrete constellations at the transmitter and with soft demapping and binary decoding at the receiver. With the complement of bit-level interleaving, this comprises the BICM architecture. The dependencies that may exist among same-symbol bits are disregarded in one-shot BICM receivers and progressively learned in their iterative counterparts.

BICM extends to the MIMO realm. With parallel scalar encoders, one per transmit antenna, the remarks made for single-input single-output (SISO) BICM apply verbatim. With a vector encoder, a one-shot BICM receiver regards as mutually independent all the bits transmitted from the various antennas at each symbol. The role of $f_{y|b_\ell}$ is then played by $f_{\boldsymbol{y}|b_{\ell,j}}(\cdot)$, defined as the PDF of $\boldsymbol{y}$ conditioned on the ℓth bit from the jth transmit antenna equaling 0 or 1. With M-ary equiprobable constellations,

$$f_{\boldsymbol{y}|b_{\ell,j}}(\mathbf{y}|0) = \frac{1}{\frac{1}{2}M^{N_{\boldsymbol{s}}}} \sum_{\mathbf{s}_m\in\mathcal{S}_0^{\ell,j}} f_{\boldsymbol{y}|\boldsymbol{s}}(\mathbf{y}|\mathbf{s}_m) \tag{1.146}$$

$$f_{\boldsymbol{y}|b_{\ell,j}}(\mathbf{y}|1) = \frac{1}{\frac{1}{2}M^{N_{\boldsymbol{s}}}} \sum_{\mathbf{s}_m\in\mathcal{S}_1^{\ell,j}} f_{\boldsymbol{y}|\boldsymbol{s}}(\mathbf{y}|\mathbf{s}_m), \tag{1.147}$$

where $N_{\boldsymbol{s}}$ is the dimensionality of $\boldsymbol{s}$ while $\mathcal{S}_0^{\ell,j}$ and $\mathcal{S}_1^{\ell,j}$ are the subsets of transmit vectors whose ℓth coded bit at the jth transmit antenna is 0 and 1, respectively. From the unconditioned equiprobability of the coded bits, the cardinality of each subset is $\frac{1}{2}M^{N_{\boldsymbol{s}}}$, hence the scaling factors in (1.146) and (1.147). Extending the SISO expression in (1.137), the information-theoretic performance of a one-shot MIMO BICM receiver is characterized by [110, 111]

$$\begin{aligned}\sum_{j=0}^{N_{\boldsymbol{s}}-1}\sum_{\ell=0}^{\log_2 M-1} I(b_{\ell,j};\boldsymbol{y}) = \sum_{j=0}^{N_{\boldsymbol{s}}-1}\sum_{\ell=0}^{\log_2 M-1} \frac{1}{2}\Bigg(&\mathbb{E}\left[\log_2 \frac{\sum_{\mathbf{s}_m\in\mathcal{S}_0^{\ell,j}} f_{\boldsymbol{y}|\boldsymbol{s}}(\boldsymbol{y}|\mathbf{s}_m)}{\frac{1}{2}\sum_{m=0}^{M^{N_{\boldsymbol{s}}}-1} f_{\boldsymbol{y}|\boldsymbol{s}}(\boldsymbol{y}|\mathbf{s}_m)}\right] \\ &+\mathbb{E}\left[\log_2 \frac{\sum_{\mathbf{s}_m\in\mathcal{S}_1^{\ell,j}} f_{\boldsymbol{y}|\boldsymbol{s}}(\boldsymbol{y}|\mathbf{s}_m)}{\frac{1}{2}\sum_{m=0}^{M^{N_{\boldsymbol{s}}}-1} f_{\boldsymbol{y}|\boldsymbol{s}}(\boldsymbol{y}|\mathbf{s}_m)}\right]\Bigg).\end{aligned} \tag{1.148}$$

In contrast with SISO, where BICM one-shot reception experiences no loss relative to signal-space coding for BPSK and QPSK, in MIMO there may be a nonzero loss even in those cases because of possible dependences introduced by the channel among the bits emitted from different transmit antennas. The loss is more significant than in SISO, yet still

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

relatively modest (about 1 dB at most) if the bit mappings are wisely chosen [96]. And, as in SISO, this loss can be largely recovered through the use of iterative decoding [112, 113].

Altogether, the mutual information $I(\boldsymbol{s};\boldsymbol{y})$ continues to be satisfyingly representative of the fundamental performance achievable with binary encoding and decoding.

1.6 MMSE estimation

Estimation theory deals with the questions of how and with what accuracy one can infer the value taken by a certain quantity on the basis of related observations. Normally built upon an underlying statistical model that connects those observations with the unknown quantity, estimation theory involves devising estimators according to different fidelity criteria and analyzing their performances [62, 114, 115].

To begin with, let us again consider the basic transformation that lies at the heart of any noisy linear channel, namely

$$y = \sqrt{\rho}s + z, \tag{1.149}$$

where ρ is fixed and the noise z can be taken to be zero-mean and of unit variance, but otherwise arbitrarily distributed for now. The problem at hand is to produce the "best possible" estimate $\hat{s}(y)$ for the variable s based on the following.

- The observation of y.
- A fidelity criterion specifying the sense in which "best possible" is to be understood.
- Some knowledge of the probabilistic relationship between s and y, and in particular knowledge of the posterior probability $f_{s|y}(\cdot)$ and the likelihood function $f_{y|s}(\cdot)$. The marginal distribution $f_s(\cdot)$, termed the *prior probability*, is further available to the estimator whenever s is part of the system design (say, if s is a signal) but may or may not be available otherwise (say, if s is a channel coefficient produced by nature).

Among the fidelity criteria that could be considered, a few are, for good reasons, prevalent in information theory and communications.

- The MAP criterion gives $\hat{s}(\mathrm{y}) = \arg\max_{\mathrm{s}} f_{s|y}(\mathrm{s}|\mathrm{y})$ with maximization over all values taken by s. Just like a MAP decoder identifies the most probable codeword, a MAP estimator returns the most probable value of s given what has been observed.
- The *maximum-likelihood* (ML) criterion gives $\hat{s}(\mathrm{y}) = \arg\max_{\mathrm{s}} f_{y|s}(\mathrm{y}|\mathrm{s})$, again with maximization over all values taken by s. Just like an ML decoder selects the most likely codeword, an ML estimator returns the value of s whose likelihood is highest. As argued via Bayes' theorem in the context of optimum decoding, if the prior is uniform, i.e., if s takes equiprobable values, then the ML and the MAP criteria coincide.
- The MMSE criterion, which is the one herein entertained.

The mean-square error measures the power of the estimation error, that is, the power of $|s - \hat{s}(y)|$. A rather natural choice in Gaussian-noise contexts, given how the defining

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

feature of such noise is its power, the mean-square error was introduced early in the nineteenth century (by Gauss himself, as well as by Legendre [116, 117]) and it is by now a ubiquitous metric. The mean-square error for a given estimate $\hat{s}(y)$ thus equals

$$\mathbb{E}\Big[\big|s-\hat{s}(y)\big|^2\Big], \tag{1.150}$$

with expectation over both s and y or, equivalently, over s and z. The minimization of this quantity gives the MMSE and the corresponding $\hat{s}(\cdot)$ is the MMSE estimator.

1.6.1 The conditional-mean estimator

As it turns out, and regardless of the noise distribution (refer to Problem 1.34), the MMSE estimator is

$$\hat{s}(y)=\mathbb{E}\big[s|y\big], \tag{1.151}$$

whose rather intuitive form indicates that, in the MMSE sense, the best guess for s is its expected value given whatever observations are available; if no observations are available, then the MMSE estimate is directly the mean. This *conditional-mean estimator* is unbiased in the sense that

$$\begin{aligned}\mathbb{E}\big[\hat{s}(y)\big] &= \mathbb{E}\big[\mathbb{E}[s|y]\big] &(1.152)\\ &= \mathbb{E}[s], &(1.153)\end{aligned}$$

but it is biased in the sense that, for a realization s, it may be that $\mathbb{E}\big[\hat{s}(\sqrt{\rho}\,\mathrm{s}+z)\big]\neq \mathrm{s}$. Put differently, the estimation error over all possible values of s is always zero-mean, but achieving the MMSE may require that the estimation error for given values of s be nonzero-mean. This dichotomy may cause confusion as the estimator can be declared to be both biased and unbiased, and it is important to make the distinction precise.

Crucially, the conditional-mean estimator $\hat{s}(y)=\mathbb{E}[s|y]$ complies with the *orthogonality principle* whereby $\mathbb{E}\big[g(y^*)\big(s-\hat{s}(y)\big)\big]=0$ for every function $g(\cdot)$. In particular,

$$\begin{aligned}\mathbb{E}\big[y^*(s-\hat{s})\big]&=0 &(1.154)\\ \mathbb{E}\big[\hat{s}^*(s-\hat{s})\big]&=0. &(1.155)\end{aligned}$$

Plugged into (1.150), the conditional-mean estimator yields

$$\mathsf{MMSE}(\rho)=\mathbb{E}\Big[s-\big|\mathbb{E}[s|y]\big|^2\Big] \tag{1.156}$$

with outer expectation over both s and y or, equivalently, over s and z. Alternatively, we can write $\mathsf{MMSE}(\rho)=\mathbb{E}\big[\mathrm{var}[s|y]\big]$ with expectation over y and with

$$\mathrm{var}[s|y]=\mathbb{E}\Big[\big|s-\mathbb{E}[s|y]\big|^2\,|y\Big] \tag{1.157}$$

the conditional variance of s given y. For given $f_s(\cdot)$ and $f_z(\cdot)$, i.e., for a certain signal format and some noise distribution, $\mathsf{MMSE}(\rho)$ is a decreasing function of ρ. Also, the mean of the signal being estimated does not influence the MMSE, and hence we can restrict ourselves to zero-mean signal distributions.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

1.6.2 MMSE estimation in Gaussian noise

Homing in on Gaussian-noise settings, with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$, we have that $y|s \sim \mathcal{N}_{\mathbb{C}}(\sqrt{\rho}\,\mathrm{s}, 1)$ and thus

$$f_{y|s}(\mathrm{y}|\mathrm{s}) = \frac{1}{\pi}\, e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2}. \tag{1.158}$$

Then, the posterior probability equals, via Bayes' theorem,

$$f_{s|y}(\mathrm{s}|\mathrm{y}) = \frac{f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})}{f_y(\mathrm{y})} \tag{1.159}$$

$$= \frac{f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})}{\int f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})\, \mathrm{ds}} \tag{1.160}$$

from which the conditional-mean estimator can be expressed as

$$\hat{s}(\mathrm{y}) = \mathbb{E}\big[s\,|y{=}\mathrm{y}\big] \tag{1.161}$$

$$= \int \mathrm{s}\, f_{s|y}(\mathrm{s}|\mathrm{y})\, \mathrm{ds} \tag{1.162}$$

$$= \int \frac{\mathrm{s}\, f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})}{\int f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})\, \mathrm{ds}}\, \mathrm{ds} \tag{1.163}$$

$$= \frac{\int \mathrm{s}\, f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})\, \mathrm{ds}}{\int f_{y|s}(\mathrm{y}|\mathrm{s})\, f_s(\mathrm{s})\, \mathrm{ds}} \tag{1.164}$$

$$= \frac{\int \mathrm{s}\, e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2}\, f_s(\mathrm{s})\, \mathrm{ds}}{\int e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2}\, f_s(\mathrm{s})\, \mathrm{ds}} \tag{1.165}$$

with integrations over the complex plane.

Example 1.27 (MMSE estimation of a complex Gaussian scalar)

Consider $y = \sqrt{\rho}s + z$ with $s \sim \mathcal{N}_{\mathbb{C}}(0,1)$. Then,

$$f_s(\mathrm{s}) = \frac{1}{\pi}\, e^{-|\mathrm{s}|^2} \tag{1.166}$$

and, applying (1.165),

$$\hat{s}(\mathrm{y}) = \frac{\int \mathrm{s}\, e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2} e^{-|\mathrm{s}|^2}\, \mathrm{ds}}{\int e^{-|\mathrm{y}-\sqrt{\rho}\mathrm{s}|^2} e^{-|\mathrm{s}|^2}\, \mathrm{ds}} \tag{1.167}$$

$$= \frac{\int \mathrm{s}\, e^{-\frac{|\mathrm{y}|^2}{1+\rho}}\, e^{-\left|\sqrt{1+\rho}\,\mathrm{s}-\sqrt{\frac{\rho}{1+\rho}}\,\mathrm{y}\right|^2}\, \mathrm{ds}}{\int e^{-\frac{|\mathrm{y}|^2}{1+\rho}}\, e^{-\left|\sqrt{1+\rho}\,\mathrm{s}-\sqrt{\frac{\rho}{1+\rho}}\,\mathrm{y}\right|^2}\, \mathrm{ds}} \tag{1.168}$$

$$= \frac{\int \mathrm{s}\, e^{-\left|\mathrm{s}-\frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}\right|^2 / \frac{1}{1+\rho}}\, \mathrm{ds}}{\int e^{-\left|\mathrm{s}-\frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}\right|^2 / \frac{1}{1+\rho}}\, \mathrm{ds}} \tag{1.169}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

$$= \frac{\int \mathrm{s} \, \frac{1}{\pi \left(\frac{1}{1+\rho}\right)} \, e^{-\left|\mathrm{s}-\frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}\right|^2 / \frac{1}{1+\rho}} \, \mathrm{ds}}{\int \frac{1}{\pi \left(\frac{1}{1+\rho}\right)} \, e^{-\left|\mathrm{s}-\frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}\right|^2 / \frac{1}{1+\rho}} \, \mathrm{ds}}. \tag{1.170}$$

Recognizing that

$$\frac{1}{\pi \left(\frac{1}{1+\rho}\right)} \, \exp\left(- \frac{\left|\mathrm{s} - \frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}\right|^2}{\frac{1}{1+\rho}} \right) \tag{1.171}$$

is the PDF of a complex Gaussian variable with mean $\frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}$, the expectation in the numerator of (1.170) equals that mean, whereas the denominator equals unity and thus

$$\hat{s}(\mathrm{y}) = \frac{\sqrt{\rho}}{1+\rho}\,\mathrm{y}, \tag{1.172}$$

which is a linear function of the observed value of y, hence the result that the MMSE estimator of a Gaussian quantity is *linear*. This estimator then yields

$$\mathsf{MMSE}(\rho) = \mathbb{E}\left[\left| s - \frac{\sqrt{\rho}}{1+\rho}\, y \right|^2 \right] \tag{1.173}$$

$$= \mathbb{E}\big[|s|^2\big] - 2\frac{\sqrt{\rho}}{1+\rho} \Re\big(\mathbb{E}[y s^*]\big) + \frac{\rho}{(1+\rho)^2} \mathbb{E}\big[|y|^2\big] \tag{1.174}$$

and, using $\mathbb{E}\big[|s|^2\big] = 1$ as well as $\mathbb{E}\big[y s^*\big] = \sqrt{\rho}$ and $\mathbb{E}\big[|y|^2\big] = 1 + \rho$, finally

$$\mathsf{MMSE}(\rho) = \frac{1}{1+\rho}. \tag{1.175}$$

Interestingly, the MMSE estimate of a Gaussian variable coincides with its MAP estimate (but not with the ML one). And, unsurprisingly given that the Gaussian distribution maximizes the differential entropy for a given variance, Gaussian variables are the hardest to estimate, meaning that any non-Gaussian variable of the same variance is bound to incur a lower estimation MMSE [118].

Example 1.28

Verify that the MMSE estimator in the previous example may be biased for a specific value of s but is unbiased over the distribution thereof.

Solution

For a given s,

$$\mathbb{E}\big[\hat{s}\big(\sqrt{\rho}s + z\big) \,|s{=}\mathrm{s}\big] = \mathbb{E}\left[\frac{\sqrt{\rho}}{1+\rho} (\sqrt{\rho}s + z) \,|s{=}\mathrm{s} \right] \tag{1.176}$$

$$= \mathbb{E}\left[\frac{\rho}{1+\rho}\mathrm{s} + \frac{\sqrt{\rho}}{1+\rho} z \right] \tag{1.177}$$

$$= \frac{\rho}{1+\rho}\,\mathrm{s} \tag{1.178}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

$$= \mathrm{s} - \frac{1}{1+\rho}\,\mathrm{s} \tag{1.179}$$

$$\neq \mathrm{s} \tag{1.180}$$

with a bias $-\frac{1}{1+\rho}\mathrm{s}$. The expectation of this bias over the distribution of s is zero.

Example 1.29 (MMSE estimation of a BPSK scalar)

Consider $y = \sqrt{\rho}s + z$ with s drawn from a BPSK constellation. The conditional-mean estimate (refer to Problem 1.36) is

$$\hat{s}(y) = \tanh\big(2\sqrt{\rho}\,\Re\{y\}\big), \tag{1.181}$$

while the corresponding MMSE reduces to the real integral

$$\mathsf{MMSE}^{\text{BPSK}}(\rho) = 1 - \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \tanh\big(2\rho - 2\sqrt{\rho}\,\xi\big)\, e^{-\xi^2}\,\mathrm{d}\xi. \tag{1.182}$$

Example 1.30 (MMSE estimation of a QPSK scalar)

Since QPSK amounts to two BPSK constellations in quadrature, each with half the power, the conditional-mean estimators for the in-phase and quadrature components are given by (1.181) applied to the real and imaginary parts of the observation, respectively, with $\rho/2$ in place of ρ. Then, the MMSE function equals

$$\mathsf{MMSE}^{\text{QPSK}}(\rho) = \mathsf{MMSE}^{\text{BPSK}}\left(\frac{\rho}{2}\right). \tag{1.183}$$

The low-ρ expansion of (1.175) reveals that, for a complex Gaussian variable,

$$\mathsf{MMSE}(\rho) = 1 - \rho + o(\rho), \tag{1.184}$$

which extends to the estimation of any variable that is proper complex, i.e., that occupies both noise dimensions in a balanced manner [119]. The prime example is QPSK.

In contrast,

$$\mathsf{MMSE}^{\text{BPSK}}(\rho) = 1 - 2\rho + o(\rho) \tag{1.185}$$

and this expansion applies, beyond BPSK, whenever a one-dimensional variable is being estimated in complex Gaussian noise.

In the high-ρ regime, in turn, the MMSE decays as $1/\rho$ when the variable being estimated is Gaussian and possibly faster otherwise [120]. In particular, for discrete constellations the decay is exponential and details on the exponents for certain types of constellations are given in [121, 122].

Generalization to vectors

The generalization of the preceding derivations to vector transformations is straightforward. Given

$$\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}, \tag{1.186}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

where $\boldsymbol{A}$ is fixed while $\boldsymbol{s}$ and $\boldsymbol{z}$ are independent with $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R_z})$, the conditional-mean estimator

$$\hat{\boldsymbol{s}}(\boldsymbol{y}) = \mathbb{E}\big[\boldsymbol{s}|\boldsymbol{y}\big] \tag{1.187}$$

attains the MMSE simultaneously for every entry of $\boldsymbol{s}$ and

$$\boldsymbol{E} = \mathbb{E}\Big[\big(\boldsymbol{s} - \hat{\boldsymbol{s}}(\boldsymbol{y})\big)\big(\boldsymbol{s} - \hat{\boldsymbol{s}}(\boldsymbol{y})\big)^*\Big] \tag{1.188}$$

is the MMSE matrix, which equals the covariance of the estimation error vector and fully describes the accuracy of the conditional-mean vector estimator. The jth diagonal entry of $\boldsymbol{E}$ indicates the MMSE incurred in the estimation of the jth entry of $\boldsymbol{s}$. From $\boldsymbol{E}$, scalar quantities with various significances may be derived as needed, say weighted arithmetic or geometric averages of the diagonal entries, or directly the largest diagonal entry [123].

Example 1.31 (MMSE estimation of a complex Gaussian vector)

Consider $\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}$ with $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R_s})$ and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$. Extending to the vector realm the derivations of Example 1.27, the MMSE estimator is found to be

$$\hat{\boldsymbol{s}}(\boldsymbol{y}) = \sqrt{\rho}\,\boldsymbol{R_s}\boldsymbol{A}^*(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)^{-1}\boldsymbol{y}, \tag{1.189}$$

which, as in the case of a Gaussian scalar, is linear in the observation. Then, from the above and (1.188),

$$\begin{aligned}
\boldsymbol{E} &= \mathbb{E}[\boldsymbol{s}\boldsymbol{s}^*] - \mathbb{E}[\boldsymbol{s}\hat{\boldsymbol{s}}^*] - \mathbb{E}[\hat{\boldsymbol{s}}\boldsymbol{s}^*] + \mathbb{E}[\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}^*] && (1.190)\\
&= \boldsymbol{R_s} - 2\rho\boldsymbol{R_s}\boldsymbol{A}^*(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)^{-1}\boldsymbol{A}\boldsymbol{R_s} \\
&\quad + \rho\boldsymbol{R_s}\boldsymbol{A}^*(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)^{-1}(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)^{-1}\boldsymbol{A}\boldsymbol{R_s} && (1.191)\\
&= \boldsymbol{R_s} - \rho\boldsymbol{R_s}\boldsymbol{A}^*(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R_s}\boldsymbol{A}^*)^{-1}\boldsymbol{A}\boldsymbol{R_s}. && (1.192)
\end{aligned}$$

Applying the matrix inversion lemma (see Appendix B.7) in a reverse fashion to (1.192), we can also rewrite $\boldsymbol{E}$ into the alternative form

$$\boldsymbol{E} = \big(\boldsymbol{R_s}^{-1} + \rho\boldsymbol{A}^*\boldsymbol{A}\big)^{-1}. \tag{1.193}$$

Expanding (1.192), the generalization of the low-ρ expansion in (1.184) to proper complex vectors comes out as

$$\boldsymbol{E} = \boldsymbol{R_s} - \rho\,\boldsymbol{R_s}\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R_s} + \mathcal{O}(\rho^2), \tag{1.194}$$

which holds whenever $\boldsymbol{s}$ is a proper complex vector, irrespective of its distribution.

1.6.3 The I-MMSE relationship in Gaussian noise

The random transformation invoked extensively throughout this chapter, namely

$$y = \sqrt{\rho}s + z, \tag{1.195}$$

where s and z are independent and $z \sim \mathcal{N}_{\mathbb{C}}(0, 1)$, is the cornerstone of any linear scalar channel impaired by Gaussian noise and, as we have seen in the formulation of the capacity,

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

the mutual information functions $\mathcal{I}(\rho) = I\big(s; \sqrt{\rho}s + z\big)$ for relevant distributions of s are exceedingly significant. The derivative of $\mathcal{I}(\rho)$ turns out to have significance as well. Regardless of the distribution of s, it holds that [124]

$$\frac{1}{\log_2 e} \cdot \frac{\mathrm{d}}{\mathrm{d}\rho}\mathcal{I}(\rho) = \mathsf{MMSE}(\rho), \tag{1.196}$$

where the right-hand side equals the MMSE when estimating s from its noisy observation, y. The identity in (1.196) is termed the *I-MMSE relationship*, and its integration yields an alternative form for the mutual information function, precisely

$$\frac{1}{\log_2 e}\,\mathcal{I}(\rho) = \int_0^{\rho} \mathsf{MMSE}(\xi)\,\mathrm{d}\xi. \tag{1.197}$$

Example 1.32 (I-MMSE relationship for a complex Gaussian scalar)

Consider $y = \sqrt{\rho}s + z$ with $s \sim \mathcal{N}_{\mathbb{C}}(0, 1)$. As derived in Examples 1.7 and 1.27,

$$\mathcal{I}(\rho) = \log_2(1 + \rho) \tag{1.198}$$

and

$$\mathsf{MMSE}(\rho) = \frac{1}{1+\rho}, \tag{1.199}$$

which satisfy the I-MMSE relationship in (1.196).

Example 1.33 (I-MMSE relationship for a BPSK scalar)

Let $y = \sqrt{\rho}s + z$ with s drawn from a BPSK constellation. From Examples 1.10 and 1.29,

$$\mathcal{I}(\rho) = 2\rho \log_2 e - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\xi^2} \log_2 \cosh\big(2\rho - 2\sqrt{\rho}\xi\big)\,\mathrm{d}\xi \tag{1.200}$$

and

$$\mathsf{MMSE}(\rho) = 1 - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \tanh\big(2\rho - 2\sqrt{\rho}\,\xi\big)\, e^{-\xi^2} \mathrm{d}\xi, \tag{1.201}$$

which satisfy the I-MMSE relationship.

Example 1.34 (I-MMSE relationship for a QPSK scalar)

As shown in Examples 1.11 and 1.30,

$$\mathcal{I}^{\mathsf{QPSK}}(\rho) = 2\,\mathcal{I}^{\mathsf{BPSK}}\left(\frac{\rho}{2}\right). \tag{1.202}$$

and

$$\mathsf{MMSE}^{\mathsf{QPSK}}(\rho) = \mathsf{MMSE}^{\mathsf{BPSK}}\left(\frac{\rho}{2}\right), \tag{1.203}$$

which are consistent with the I-MMSE relationship.

In the low-ρ regime, the I-MMSE relationship bridges the distinctness of proper complex

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

signals in terms of mutual information and MMSE. Indeed, recalling (1.52) and (1.185), for such signals

$$\mathcal{I}(\rho) = \left(\rho - \frac{1}{2}\,\rho^2\right) \log_2 e + o(\rho^2) \tag{1.204}$$

and

$$\mathsf{MMSE}(\rho) = 1 - 2\rho + o(\rho). \tag{1.205}$$

Generalization to vectors

The I-MMSE relationship also extends to the vector realm. Consider again the random transformation

$$\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}, \tag{1.206}$$

where $\boldsymbol{A}$ is fixed while $\boldsymbol{s}$ and $\boldsymbol{z}$ are independent with $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$. Then, regardless of the distribution of $\boldsymbol{s}$ [125]

$$\frac{1}{\log_2 e}\,\nabla_{\boldsymbol{A}}\, I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}) = \rho\boldsymbol{A}\boldsymbol{E}, \tag{1.207}$$

where $\nabla_{\boldsymbol{A}}$ denotes the gradient with respect to $\boldsymbol{A}$ (see Appendix D) while $\boldsymbol{E}$ is the MMSE matrix defined in (1.188) for the estimation of $\boldsymbol{s}$, i.e., the generalization to multiple dimensions of the scalar MMSE.

Example 1.35 (I-MMSE relationship for a complex Gaussian vector)

As established in Example 1.13, when the noise is $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$ while $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{s}})$,

$$?I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}) = \log_2 \det\left(\boldsymbol{I} + \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\right) \tag{1.208}$$

$$= \log_2 \det\left(\boldsymbol{I} + \rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\right) \tag{1.209}$$

and, applying the expression for the gradient of a log-determinant function given in Appendix D, we obtain

$$\frac{1}{\log_2 e}\,\nabla_{\boldsymbol{A}}\, I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}) = \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\left(\boldsymbol{I} + \rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\right)^{-1} \tag{1.210}$$

$$= \rho\boldsymbol{A}\left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho\boldsymbol{A}^*\boldsymbol{A}\right)^{-1}, \tag{1.211}$$

which indeed equals $\rho\boldsymbol{A}\boldsymbol{E}$ with $\boldsymbol{E} = \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho\boldsymbol{A}^*\boldsymbol{A}\right)^{-1}$ as determined in Example 1.31 for a complex Gaussian vector.

Example 1.36

Use the I-MMSE relationship to express $\frac{\partial}{\partial\rho} I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z})$ as a function of $\boldsymbol{A}$ and the MMSE matrix $\boldsymbol{E}$, for an arbitrarily distributed $\boldsymbol{s}$.

Solution

Let us first narrow the problem to $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{s}})$. Denoting by $\lambda_j(\cdot)$ the jth eigenvalue

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

of a matrix, we can rewrite (1.209) as

$$I(\boldsymbol{s};\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}+\boldsymbol{z}) = \log_2 \det(\boldsymbol{I}+\rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}) \tag{1.212}$$

$$= \log_2 \prod_j \lambda_j(\boldsymbol{I}+\rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}) \tag{1.213}$$

$$= \sum_j \log_2 \lambda_j(\boldsymbol{I}+\rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}) \tag{1.214}$$

$$= \sum_j \log_2\big(1+\rho\,\lambda_j(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}})\big). \tag{1.215}$$

Then, differentiating with respect to ρ, we obtain

$$\frac{\partial}{\partial\rho} I(\boldsymbol{s};\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}+\boldsymbol{z}) = \sum_j \frac{\lambda_j(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}})}{1+\rho\,\lambda_j(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}})}\log_2 e \tag{1.216}$$

$$= \sum_j \lambda_j\big(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}(\boldsymbol{I}+\rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}})^{-1}\big)\log_2 e \tag{1.217}$$

$$= \mathrm{tr}\big(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}(\boldsymbol{I}+\rho\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}})^{-1}\big)\log_2 e \tag{1.218}$$

$$= \mathrm{tr}\big(\boldsymbol{A}^*\boldsymbol{A}(\boldsymbol{R}_{\boldsymbol{s}}^{-1}+\rho\boldsymbol{A}^*\boldsymbol{A})^{-1}\big)\log_2 e \tag{1.219}$$

$$= \mathrm{tr}\big(\boldsymbol{A}\,(\boldsymbol{R}_{\boldsymbol{s}}^{-1}+\rho\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\big)\log_2 e \tag{1.220}$$

and thus we can write

$$\frac{1}{\log_2 e}\frac{\partial}{\partial\rho} I(\boldsymbol{s};\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}+\boldsymbol{z}) = \mathrm{tr}\big(\boldsymbol{A}\boldsymbol{E}\boldsymbol{A}^*\big). \tag{1.221}$$

Although derived for a complex Gaussian vector $\boldsymbol{s}$, as a corollary of the I-MMSE relationship this identity can be claimed regardless of the distribution of $\boldsymbol{s}$. Indeed, the evaluation of $\frac{\partial}{\partial\rho} I(\boldsymbol{s};\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}+\boldsymbol{z})$ for an arbitrary $\boldsymbol{s}$ can be effected through the gradient with respect to $\sqrt{\rho}\boldsymbol{A}$, and the application of (1.207) then leads to (1.221) all the same.

Evaluated at $\rho = 0$, the identity in (1.221) gives the formula

$$\frac{1}{\log_2 e}\frac{\partial}{\partial\rho} I(\boldsymbol{s};\sqrt{\rho}\boldsymbol{A}\boldsymbol{s}+\boldsymbol{z})\bigg|_{\rho=0} = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\right), \tag{1.222}$$

which is a generalization of (1.79).

1.7 LMMSE estimation

While the precise distribution of certain quantities (say, the signals being transmitted) is entirely within the control of the system designer, there are other quantities of interest (say, the channel gain) that are outside that control. When quantities of the latter type are to be estimated, it is often the case that we are either unable or unwilling to first obtain their distributions beyond the more accessible mean and variance. With the MMSE retained as the estimation criterion, a sensible approach is to regard the distribution as that whose MMSE

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

estimation is the hardest, namely the Gaussian distribution. This leads to the application of the linear MMSE (LMMSE) estimators derived in Examples 1.27 and 1.31 to quantities that need not be Gaussian.

Alternatively, LMMSE estimators may be featured as a design choice, even if the relevant distribution is known, simply because of the appeal and simplicity of a linear filter.

And then, of course, LMMSE estimators may be in place simply because the quantities to be estimated are known to be Gaussian (say, capacity-achieving signals).

For all the foregoing reasons, LMMSE estimators are prevalent in wireless communications and throughout this text. Except when estimating a truly Gaussian quantity, an LMMSE estimator is bound to be inferior to a conditional-mean estimator, but also more versatile and robust.

1.7.1 Random variables

Under the constraint of a linear structure, the LMMSE estimator for a vector $\boldsymbol{s}$ based on the observation of a related vector $\boldsymbol{y}$ is to be

$$\hat{\boldsymbol{s}} = \boldsymbol{W}^{\mathsf{MMSE}*}\boldsymbol{y} + \boldsymbol{b}^{\mathsf{MMSE}}. \tag{1.223}$$

The mean $\boldsymbol{\mu_s}$ can be regarded as known and the role of the constant term $\boldsymbol{b}^{\mathsf{MMSE}}$ is to ensure that $E[\hat{\boldsymbol{s}}(\boldsymbol{y})] = \boldsymbol{\mu_s}$ (refer to Problem 1.45). With the unbiasedness in that sense taken care of, the LMMSE estimator is embodied by the matrix $\boldsymbol{W}^{\mathsf{MMSE}}$, which can be inferred from (1.189) to equal

$$\boldsymbol{W}^{\mathsf{MMSE}} = \boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{ys}} \tag{1.224}$$

and that is indeed its form in broad generality. To see that, we can write the mean-square error on the estimation of the jth entry of $\boldsymbol{s}$ via a generic linear filter $\boldsymbol{W}$ as

$$\mathbb{E}\Big[\big|[\boldsymbol{s}-\hat{\boldsymbol{s}}]_j\big|^2\Big] = \mathbb{E}\Big[\big|[\boldsymbol{s}-\boldsymbol{W}^*\boldsymbol{y}]_j\big|^2\Big] \tag{1.225}$$

$$= \mathbb{E}\Big[\big|s_j - \boldsymbol{w}_j^*\boldsymbol{y}\big|^2\Big] \tag{1.226}$$

$$= \mathbb{E}\Big[\,|s_j|^2\Big] - \mathbb{E}\Big[\boldsymbol{w}_j^*\boldsymbol{y}\,s_j^*\Big] - \mathbb{E}\Big[s_j\,\boldsymbol{y}^*\boldsymbol{w}_j\Big] + \mathbb{E}\Big[\boldsymbol{w}_j^*\boldsymbol{y}\boldsymbol{y}^*\boldsymbol{w}_j\Big], \tag{1.227}$$

where $\boldsymbol{w}_j = [\boldsymbol{W}]_{:,j}$ is shorthand for the part of $\boldsymbol{W}$—its jth column—that is responsible for estimating that particular entry, $s_j = [\boldsymbol{s}]_j$. The gradient of (1.227) with respect to $\boldsymbol{w}_j$, obtained by applying (D.5)–(D.7), equals

$$\nabla_{\boldsymbol{w}_j}\mathbb{E}\Big[\big|[\boldsymbol{s}-\hat{\boldsymbol{s}}]_j\big|^2\Big] = -\mathbb{E}\Big[\boldsymbol{y}\,s_j^*\Big] + \mathbb{E}\Big[\boldsymbol{y}\boldsymbol{y}^*\boldsymbol{w}_j\Big] \tag{1.228}$$

$$= -\mathbb{E}\Big[\boldsymbol{y}\,\big(s_j - \boldsymbol{w}_j^*\boldsymbol{y}\big)^*\Big] \tag{1.229}$$

$$= -\mathbb{E}\Big[\boldsymbol{y}\,[\boldsymbol{s}-\hat{\boldsymbol{s}}]_j^*\Big], \tag{1.230}$$

which, equated to zero, is nothing but a manifestation of the orthogonality principle exposed earlier in this chapter. Assembling the expressions corresponding to (1.229) for every column of $\boldsymbol{W}$ and equating the result to zero, we find that the LMMSE filter must

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

satisfy

$$-\mathbb{E}\Big[\boldsymbol{y}\left(\boldsymbol{s}-\boldsymbol{W}^{\text{MMSE}*}\boldsymbol{y}\right)^*\Big] = \mathbb{E}\big[\boldsymbol{y}\boldsymbol{y}^*\big]\boldsymbol{W}^{\text{MMSE}} - \mathbb{E}\big[\boldsymbol{y}\boldsymbol{s}^*\big] \tag{1.231}$$

$$= \boldsymbol{0} \tag{1.232}$$

and, since the mean-square error is a quadratic—and thus convex—function of the linear filter, this condition is not only necessary but sufficient (see Appendix G). Rewritten as

$$\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{W}^{\text{MMSE}} - \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}} = \boldsymbol{0}, \tag{1.233}$$

its solution does give $\boldsymbol{W}^{\text{MMSE}} = \boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}$ as anticipated in (1.224).

Moving on, the covariance of the estimate $\hat{\boldsymbol{s}}$ emerges as

$$\boldsymbol{R}_{\hat{\boldsymbol{s}}} = \mathbb{E}[\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}^*] \tag{1.234}$$

$$= \boldsymbol{W}^{\text{MMSE}*}\,\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^*]\boldsymbol{W}^{\text{MMSE}} \tag{1.235}$$

$$= \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}^*\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}} \tag{1.236}$$

$$= \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}^*\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}, \tag{1.237}$$

while

$$\boldsymbol{R}_{\hat{\boldsymbol{s}}\boldsymbol{s}} = \mathbb{E}\big[\boldsymbol{W}^{\text{MMSE}*}\boldsymbol{y}\boldsymbol{s}^*\big] \tag{1.238}$$

$$= \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}^*\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}} \tag{1.239}$$

$$= \boldsymbol{R}_{\hat{\boldsymbol{s}}}. \tag{1.240}$$

It follows that the MMSE matrix is given by

$$\boldsymbol{E} = \mathbb{E}\Big[\big(\boldsymbol{s}-\hat{\boldsymbol{s}}(\boldsymbol{y})\big)\big(\boldsymbol{s}-\hat{\boldsymbol{s}}(\boldsymbol{y})\big)^*\Big] \tag{1.241}$$

$$= \boldsymbol{R}_{\boldsymbol{s}} - \boldsymbol{R}_{\hat{\boldsymbol{s}}\boldsymbol{s}} - \boldsymbol{R}_{\hat{\boldsymbol{s}}\boldsymbol{s}}^* + \boldsymbol{R}_{\hat{\boldsymbol{s}}} \tag{1.242}$$

$$= \boldsymbol{R}_{\boldsymbol{s}} - \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}^*\boldsymbol{R}_{\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{s}}. \tag{1.243}$$

Specialized to the linear random transformation $\boldsymbol{y} = \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}$, the foregoing expressions for $\boldsymbol{W}^{\text{MMSE}}$ and $\boldsymbol{E}$ become

$$\boldsymbol{W}^{\text{MMSE}} = \sqrt{\rho}\left(\boldsymbol{R}_{\boldsymbol{z}} + \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\right)^{-1}\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}} \tag{1.244}$$

and

$$\boldsymbol{E} = \boldsymbol{R}_{\boldsymbol{s}} - \rho\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\left(\boldsymbol{R}_{\boldsymbol{z}} + \rho\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}\boldsymbol{A}^*\right)^{-1}\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}}, \tag{1.245}$$

consistent with (1.189) and (1.192) if $\boldsymbol{R}_{\boldsymbol{z}} = \boldsymbol{I}$. Derived in the context of conditional-mean MMSE estimation for white Gaussian noise and Gaussian signals, within the broader confines of the LMMSE these expressions apply regardless of the distributions thereof. Only the second-order statistics of noise and signals enter the relationships, as a result of which the formulation is characterized by the presence of quadratic forms.

Applying the matrix inversion lemma (see Appendix B.7) to (1.244), we can rewrite $\boldsymbol{W}^{\text{MMSE}}$ into the alternative form

$$\boldsymbol{W}^{\text{MMSE}} = \sqrt{\rho}\left[\boldsymbol{R}_{\boldsymbol{z}}^{-1} - \rho\boldsymbol{R}_{\boldsymbol{z}}^{-1}\boldsymbol{A}\left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^*\boldsymbol{R}_{\boldsymbol{z}}^{-1}\right]\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{s}} \tag{1.246}$$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

$$= \sqrt{\rho}\, \boldsymbol{R}_{\boldsymbol{z}}^{-1} \left[\boldsymbol{I} - \rho \boldsymbol{A} \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \right] \boldsymbol{A} \boldsymbol{R}_{\boldsymbol{s}} \tag{1.247}$$

$$= \sqrt{\rho}\, \boldsymbol{R}_{\boldsymbol{z}}^{-1} \left[\boldsymbol{A} - \rho \boldsymbol{A} \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right] \boldsymbol{R}_{\boldsymbol{s}} \tag{1.248}$$

$$= \sqrt{\rho}\, \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \left[\boldsymbol{I} - \rho \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1} \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right] \boldsymbol{R}_{\boldsymbol{s}} \tag{1.249}$$

$$= \sqrt{\rho}\, \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1} \left[\left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right) - \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right] \boldsymbol{R}_{\boldsymbol{s}}$$

$$= \sqrt{\rho}\, \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1}, \tag{1.250}$$

while applying the matrix inversion lemma in a reverse fashion to (1.245), $\boldsymbol{E}$ can be rewritten as

$$\boldsymbol{E} = \left(\boldsymbol{R}_{\boldsymbol{s}}^{-1} + \rho \boldsymbol{A}^* \boldsymbol{R}_{\boldsymbol{z}}^{-1} \boldsymbol{A} \right)^{-1}. \tag{1.251}$$

If both noise and signal are scalars, rather than vectors, then the two expressions for $\boldsymbol{W}^{\mathsf{MMSE}}$ coincide, yielding

$$W^{\mathsf{MMSE}} = \frac{\sqrt{\rho}}{1+\rho}, \tag{1.252}$$

while the two expressions for $\boldsymbol{E}$ reduce to

$$\mathsf{MMSE} = \frac{1}{1+\rho}, \tag{1.253}$$

as derived earlier, in the context of conditional-mean MMSE estimation, for Gaussian noise and Gaussian signals. In LMMSE, these equations acquire broader generality.

1.7.2 Random processes

The LMMSE estimation problem becomes richer when formulated for random processes, as it then splits into several variants:

- *Noncausal*. The value of some signal at time n is estimated on the basis of the entire observation of another signal, a procedure also termed *smoothing*. If the observed signal is decimated relative to its estimated brethren, then the smoothing can also be regarded as *interpolation* in the MMSE sense.
- *Causal*. The value of some signal at time n is estimated on the basis of observations of another signal at times $n-1, \ldots, n-N$. This variant, for which the term *filtering* is sometimes formally reserved in the estimation literature, and which can also be regarded as *prediction* in the MMSE sense, can be further subdivided depending on whether N is finite or unbounded.

For stationary processes, the problem was first tackled by Norbert Wiener in the 1940s [126], hence the common designation of the corresponding estimator as a *Wiener filter*. (For nonstationary processes, the more general Kalman filter was developed years later.)

Without delving extensively into the matter, on which excellent textbooks exist already [62, 114, 127], we introduce herein a couple of results that are invoked throughout the text. These results pertain to the discrete-time scalar channel $y[n] = \sqrt{\rho}\, s[n] + z[n]$ where $s[n]$

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

is a zero-mean unit-variance stationary signal with power spectrum $S(\cdot)$ while $z[n]$ is an IID noise sequence. For such a setting, the noncausal LMMSE filter yields [68]

$$\mathsf{MMSE} = 1 - \int_{-1/2}^{1/2} \frac{\rho\, S^2(\nu)}{1 + \rho\, S(\nu)}\, \mathrm{d}\nu, \tag{1.254}$$

while its causal counterpart gives

$$\mathsf{MMSE} = \frac{1}{\rho}\left[\exp\left(\int_{-1/2}^{1/2} \log_e\big(1 + \rho\, S(\nu)\big)\, \mathrm{d}\nu\right) - 1\right]. \tag{1.255}$$

Letting $\rho \to \infty$ in (1.255) returns the causal MMSE when predicting $s[n]$ based on past noiseless observations of the same process,

$$\mathsf{MMSE} = \exp\left(\int_{-1/2}^{1/2} \log_e\big[S(\nu)\big]\, \mathrm{d}\nu\right), \tag{1.256}$$

which is zero if $s[n]$ is nonregular while strictly positive if it is regular. By inspecting (1.256) it can be deduced that, in the context of stationary processes, nonregularity is tantamount to a bandlimited power spectrum—whereby the integrand diverges over part of the spectrum—while regularity amounts to a power spectrum that is not bandlimited and strictly positive.

1.8 Summary

From the coverage in this chapter, we can distill the points listed in the accompanying summary box.

Problems

1.1 Show that, for s to be proper complex, its in-phase and quadrature components must be uncorrelated and have the same variance.

1.2 Let s conform to a 3-PSK constellation defined by $\mathrm{s}_0 = \frac{1}{\sqrt{2}}(1-\mathrm{j})$, $\mathrm{s}_1 = \frac{1}{\sqrt{2}}(-1-\mathrm{j})$, and $\mathrm{s}_2 = \mathrm{j}$. Is this signal proper complex? Is it circularly symmetric?

1.3 Let s conform to a ternary constellation defined by $\mathrm{s}_0 = -1$, $\mathrm{s}_1 = 0$, and $\mathrm{s}_2 = 1$. Is this signal proper complex? Is it circularly symmetric?

1.4 Give an expression for the minimum distance between neighboring points in a one-dimensional constellation featuring M points equidistant along the real axis.

1.5 Let x be a discrete random variable and let $y = g(x)$ with $g(\cdot)$ an arbitrary function. Is $\mathcal{H}(y)$ larger or smaller than $\mathcal{H}(x)$?

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

Take-away points

1. The mutual information between two random variables measures the information that one of them can supply about the other.
2. The channel capacity is the highest spectral efficiency at which reliable communication is possible in the sense that the probability of erroneous codeword decoding vanishes as the codeword length N grows.
3. If the channel is information stable, meaning that the information that the received sequence $y[0], \ldots, y[N-1]$ conveys about the transmit sequence $x[0], \ldots, x[N-1]$ is invariable for large N, then the capacity equals the maximum mutual information between $x[0], \ldots, x[N-1]$ and $y[0], \ldots, y[N-1]$ for $N \to \infty$. This maximization entails finding the optimum distribution for $x[0], \ldots, x[N-1]$ subject to the applicable constraints on the transmit signal (e.g., the power).
4. The capacity is robust in that the spectral efficiencies with finite-length (but long) codewords and reasonably small error probabilities hardly depart from it. Moreover, such long codewords can be featured without incurring excessive latencies. And, through hybrid-ARQ, the codeword length can be made adaptive.
5. The use of binary codes with binary decoding incurs only a minute information-theoretic penalty with respect to coding on the constellation's alphabet. The penalty is actually nil for BPSK and QPSK, and can be largely recovered for other constellations through iterative reception. It is thus routine, in terms of performance limits, to treat binary codes mapped to arbitrary constellations as if the coding took place on that constellation's alphabet.
6. BICM is the default architecture for coding and modulation. At the transmitter, this entails binary coding, bit-level interleaving, and constellation mapping. At the receiver, it entails soft demapping, deinterleaving, and APP binary decoding.
7. In the MMSE sense, the best estimate of a quantity is the one delivered by the conditional-mean estimator. When both the quantity being estimated and the noise contaminating the observations are Gaussian, such conditional-mean estimator is a linear function of the observations, the LMMSE estimator.
8. The I-MMSE relationship establishes that the derivative of the Gaussian-noise mutual information between two quantities equals the MMSE when observing one from the other.
9. While inferior to the conditional-mean for non-Gaussian quantities, the LMMSE estimator remains attractive because of the simplicity of linear filtering and the robustness, as only second-order statistics are required.

1.6 Express the entropy of a discrete random variable x as a function of the information divergence between x and a uniformly distributed counterpart.

1.7 Express the differential entropy of a real Gaussian variable $x \sim \mathcal{N}(\mu, \sigma^2)$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

1.8 Compute the differential entropy of a random variable that takes the value 0 with probability $1/3$ and is otherwise uniformly distributed in the interval $[-1, 1]$.

1.9 Calculate the differential entropy of a random variable x that abides by the exponential distribution

$$f_x(\mathrm{x}) = \frac{1}{\mu} e^{-x/\mu}. \tag{1.257}$$

1.10 Consider a random variable s such that $\Re\{s\} \sim \mathcal{N}(0, 1/2)$ and $\Im\{s\} = q\,\Re(s)$ where $q = \pm 1$ equiprobably. Compute the differential entropy of s, which is complex and Gaussian but not proper, and compare it with that of a standard complex Gaussian.

1.11 Prove that $\mathfrak{h}(x + a) = \mathfrak{h}(x)$ for any constant a.

1.12 Prove that $\mathfrak{h}(ax) = \mathfrak{h}(x) + \log_2 |a|$ for any constant a.

1.13 Express the differential entropy of the real Gaussian vector $\boldsymbol{x} \sim \mathcal{N}(\mu, \boldsymbol{R})$.

1.14 Consider the first-order Gauss–Markov process

$$h[n] = \sqrt{1 - \varepsilon}\, h[n-1] + \sqrt{\varepsilon}\, w[n] \tag{1.258}$$

where $\{w[n]\}$ is a sequence of IID random variables with $w \sim \mathcal{N}_{\mathbb{C}}(0, 1)$.

(a) Express the entropy rate as a function of ε.

(b) Quantify the entropy rate for $\varepsilon = 10^{-3}$.

Note: The Gauss–Markov process underlies a fading model presented in Chapter 3.

1.15 Verify (1.79) and (1.80).

Hint: Express $\det(\cdot)$ *as the product of the eigenvalues of its argument.*

1.16 Show that $I(x_0; x_1; y) \geq I(x_0; y)$ for any random variables x_0, x_1, and y.

1.17 Let $y = \sqrt{\rho}\,(s_0 + s_1) + z$ where s_0, s_1, and z are independent standard complex Gaussian variables.

(a) Show that $I(s_0, s_1; y) = I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + z)$ for $\boldsymbol{s} = [s_0\ s_1]^{\mathrm{T}}$ and a suitable $\boldsymbol{A}$.

(b) Characterize $I(s_0, s_1; y) - I(s_0; y)$ and approximate its limiting behaviors for $\rho \ll 1$ and $\rho \gg 1$.

(c) Repeat part (b) for the case that s_0 and s_1 are partially correlated. What do you observe?

(d) Repeat part (b) for the modified relationship $y = \sqrt{\rho/2}\,(s_0 + s_1) + z$.

Can you draw any conclusion related to MIMO from this problem?

1.18 Let s be of unit variance and uniformly distributed on a disk while $z \sim \mathcal{N}_{\mathbb{C}}(0, 1)$.

(a) What is the first-order expansion of $\mathcal{I}(\rho) = I(s; \sqrt{\rho}s + z)$ for small ρ?

(b) What is the leading term in the expansion of $\mathcal{I}(\rho)$ for large ρ?

Note: The signal distribution in this problem can be interpreted as a dense set of concentric ∞*-PSK rings, conveying information in both phase and magnitude.*

1.19 Repeat Problem 1.18 with s conforming to a one-dimensional discrete constellation featuring M points equidistant along a line forming an angle ϕ with the real axis.

1.20 Let s and z conform to BPSK distributions. Express $\mathcal{I}(\rho) = I(s; \sqrt{\rho}s + z)$ and obtain expansions thereof for small and large ρ. How much is $\mathcal{I}(\rho)$ for $\rho = 5$?

1.21 Compute $I(s; \sqrt{\rho}s + z)$ with $s \sim \mathcal{N}_{\mathbb{C}}(0, 1)$ and with z having a BPSK distribution.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

1.22 Compute $I(s; \sqrt{\rho}s + z)$ with both s and z having BPSK distributions.

1.23 Verify that, as argued in Example 1.11,

$$\mathcal{I}^{\text{QPSK}}(\rho) = 2\,\mathcal{I}^{\text{BPSK}}\left(\frac{\rho}{2}\right). \tag{1.259}$$

1.24 Express the Gaussian mutual information of a square QAM signal as a function of the Gaussian mutual information of another signal whose points are equiprobable and uniformly spaced over the real line.
Note: This relationship substantially simplifies the computation of the Gaussian mutual information of square QAM signals, and it is exploited to perform such computations in this book.

1.25 Let $y = \sqrt{\rho}s + z$. If z were not independent of s, would that increase or decrease $I(s; y)$ relative to the usual situation where they are independent? Can you draw any communication-theoretic lesson from this?

1.26 Let $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$ while

$$\boldsymbol{A} = \begin{bmatrix} 0.7 & 1 + 0.5\,\mathrm{j} & 1.2\,\mathrm{j} \\ 0.2 + \mathrm{j} & -2.1 & 0 \end{bmatrix}. \tag{1.260}$$

(a) Plot the exact $I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z})$ against its low-ρ expansion for $\rho \in [0, 1]$. Up to which value of ρ is the difference below 10%?
(b) Plot the exact $I(\boldsymbol{s}; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z})$ against its high-ρ expansion for $\rho \in [10, 100]$. Beyond which value of ρ is the difference below 10%?

1.27 Let $\boldsymbol{s}$ have two independent unit-variance entries and let $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$ while $\boldsymbol{A} = [0.7 \quad 1 + 0.5\,\mathrm{j}]$. On a common chart, plot $\mathcal{I}(\rho) = I(s; \sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{z})$ for $\rho \in [0, 10]$ under the following distributions for the entries of $\boldsymbol{s}$:
(a) Real Gaussian.
(b) Complex Gaussian.
(c) BPSK.
(d) QPSK.

1.28 Compute and plot, as function of $\rho \in [-5, 25]$ dB, the Gaussian mutual information function for the following constellations:
(a) 8-PSK.
(b) 16-QAM.

1.29 Establish the law of the channel

$$\bar{\boldsymbol{y}} = \sqrt{\rho}\boldsymbol{A}\bar{\boldsymbol{s}} + \bar{\boldsymbol{z}}, \tag{1.261}$$

where $\boldsymbol{A}$ is a fixed matrix whose (n, n)th entry determines how the nth transmit symbol affects the nth received one, while $\bar{\boldsymbol{z}} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\bar{\boldsymbol{z}}})$ with the (n, n)th entry of $\boldsymbol{R}_{\bar{\boldsymbol{z}}}$ determining the correlation between the noise afflicting symbols n and n.

1.30 Consider the channel

$$y[n] = \sqrt{\rho}\,h[n]s[n] + z[n] \qquad n = 0, \ldots, N-1 \tag{1.262}$$

where $z[0], \ldots, z[N-1]$ are IID with $z \sim \mathcal{N}_{\mathbb{C}}(0, 1)$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

(a) If $h[0], \ldots, h[N-1]$ are also IID with $h \sim \mathcal{N}_{\mathbb{C}}(0,1)$, what is the channel law? Is the channel memoryless?

(b) Now suppose that $h[n+1] = h[n]$ for $n = 0, 2, 4, \ldots, N-2$ while $h[n+1]$ and $h[n]$ are independent for $n = 1, 3, 5, \ldots, N-1$, meaning that every pair of symbols shares the same coefficient but then the coefficients change across symbol pairs in an IID fashion. For $h \sim \mathcal{N}_{\mathbb{C}}(0,1)$, what is the channel law? Is the channel memoryless?

1.31 Express, to first order, the number of codeword symbols N required to achieve a certain share of the capacity C as a function of V and p_{e}. Then, for $V/C^2 = 4$, use the found expression to gauge the following.

(a) The value of N required to achieve 90% of capacity at $p_{\mathrm{e}} = 10^{-2}$.

(b) The value of N required to achieve 95% of capacity at $p_{\mathrm{e}} = 10^{-3}$.

1.32 Consider a system with $B = 100$ MHz, equally divided among $U = 10$, and with a coding latency target of 1 ms. If the operating point is $p_{\mathrm{e}} = 10^{-2}$ and $V/C^2 = 2$, what fraction of the capacity can each user attain?

1.33 Reproduce the BICM curve on the left-hand side of Fig. 1.6.

1.34 Consider the transformation $y = \sqrt{\rho}s + z$.

(a) Prove that, for any arbitrary function $g(\cdot)$, $\mathbb{E}\big[g(y)(s - \mathbb{E}[s|y])\big] = 0$. This is the so-called *orthogonality principle*.

(b) Taking advantage of the orthogonality principle, prove that the MMSE estimate is given by $\hat{s}(y) = \mathbb{E}[s|y]$.

1.35 Consider the transformation $y = \sqrt{\rho}s + z$ with z a standard complex Gaussian and with $s \sim \mathcal{N}_{\mathbb{C}}(\mu_s, \sigma_s^2)$.

(a) Obtain the conditional-mean estimator.

(b) Express the corresponding $\mathsf{MMSE}(\rho)$.

(c) Verify that, when $\mu_s = 0$ and $\sigma_s^2 = 1$, such estimator reduces to (1.172) while $\mathsf{MMSE}(\rho)$ reduces to (1.175).

(d) Verify that $\mathsf{MMSE}(\cdot)$ does not depend on μ_s.

1.36 Prove that, for the transformation $y = \sqrt{\rho}s + z$ with z a standard complex Gaussian and with s being BPSK-distributed, the following are true.

(a) The conditional-mean estimate equals (1.181).

(b) The MMSE as a function of ρ equals (1.182).

1.37 Given the transformation $y = \sqrt{\rho}s + z$ with z a standard complex Gaussian, derive the function $\mathsf{MMSE}(\rho)$ for s conforming to a 16-QAM distribution.

1.38 Consider the vector transformation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}$ where $\boldsymbol{A}$ is fixed while $\boldsymbol{s}$ and $\boldsymbol{z}$ are independent with $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{s}})$ and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{z}})$.

(a) Obtain the conditional-mean estimator.

(b) Express the corresponding MMSE matrix.

(c) Verify that, for $\boldsymbol{R}_{\boldsymbol{z}} = \boldsymbol{I}$, the MMSE matrix equals (1.192).

1.39 Let s be BPSK-distributed while $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$. Compute the dB-difference between the MMSEs achieved by conditional-mean and LMMSE estimates of s based on observations of $\sqrt{\rho}s + z$ for two cases:

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press

(a) $\rho = 1$.
(b) $\rho = 10$.

1.40 Verify that the application of (1.196) to (1.200) yields (1.201).

1.41 Let s be of unit variance while $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$. Provide first-order low-ρ expansions of $\mathsf{MMSE}(\rho)$ as achieved by the conditional-mean estimate of s based on observations of $\sqrt{\rho}s + z$ under the following distributions for s:
(a) Real Gaussian.
(b) Complex Gaussian.
(c) BPSK.
(d) QPSK.
(e) ∞-PSK.
(f) ∞-QAM.
What can be observed?

1.42 On a common chart, plot $\mathsf{MMSE}(\rho)$ for the estimation of s based on observing $\sqrt{\rho}s + z$ with $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$ and under the following distributions for s:
(a) Real Gaussian.
(b) Complex Gaussian.
(c) BPSK.
(d) QPSK.
Further plot, on the same chart, the corresponding low-ρ expansions of $\mathsf{MMSE}(\rho)$.

1.43 Let $y = \sqrt{\rho}s + z$ with s zero-mean unit-variance and with z a standard complex Gaussian. For $\rho \in [0, 10]$, plot the dB-difference between the mean-square error achieved by a regular LMMSE estimator and by a modified version thereof in which the estimation bias for each realization of s has been removed.

1.44 Consider the vector transformation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{z}$ where $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R_s})$ and $\boldsymbol{z} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{R_z})$.
(a) Express the MMSE matrix $\boldsymbol{E}$ when estimating $\boldsymbol{s}$ based on the observation of $\boldsymbol{y}$.
(b) Based on the expression obtained for $\boldsymbol{E}$, generalize to colored Gaussian noise the I-MMSE relationship for white Gaussian noise given in (1.207).
Note: Although derived for a Gaussian signal in this problem, the generalized version of the I-MMSE relationship does hold for arbitrarily distributed $\boldsymbol{s}$.

1.45 For the LMMSE estimator $\hat{\boldsymbol{s}}(\boldsymbol{y}) = \boldsymbol{W}^{\mathsf{MMSE}*}\boldsymbol{y} + \boldsymbol{b}^{\mathsf{MMSE}}$, determine the value of $\boldsymbol{b}^{\mathsf{MMSE}}$ as a function of the known means $\boldsymbol{\mu_s}$ and $\boldsymbol{\mu_y}$.

1.46 Let $\boldsymbol{s}$ be a vector containing two unit-variance entries exhibiting 50% correlation and let $z \sim \mathcal{N}_{\mathbb{C}}(0,1)$ while $\boldsymbol{A} = [0.7 \quad 1 + 0.5\,\mathrm{j}]$. Plot the MMSE as a function of $\rho \in [0, 10]$ when LMMSE-estimating $\boldsymbol{s}$ from $\sqrt{\rho}\boldsymbol{A}\boldsymbol{s} + z$.

https://doi.org/10.1017/9781139049276.002 Published online by Cambridge University Press