1 Introduction

1.1 Motivation

Claude E. Shannon was one of the great minds of the twentieth century. In the 1940s, he almost single-handedly created the field of information theory and gave the world a new way to look at information and communication. The channel-coding theorem, where he proved the *existence* of good error-correcting codes to transmit information at any rate below capacity with an arbitrarily small probability of error, was one of his fundamental contributions. Unfortunately, Shannon never described how to *construct* these codes. Ever since his 1948 landmark paper "A mathematical theory of communication" [1], the channel-coding theorem has tantalized researchers worldwide in their quest for the ultimate error-correcting code. After more than forty years, state-of-the-art error-correcting codes were still disappointingly far away from Shannon's theoretical capacity bound. No drastic improvement seemed to be forthcoming, and researchers were considering a more practical capacity benchmark, the cut-off rate [2], which could be achieved by practical codes.

In 1993, two until then little-known French researchers from the ENST in Bretagne, Claude Berrou and Alain Glavieux, claimed to have discovered a new type of code, which operated very close to Shannon capacity with reasonable decoding complexity. The decoding process consisted of two decoders passing information back and forth, giving rise to the name "turbo code." They first presented their results at the IEEE International Conference on Communications in Geneva, Switzerland [3]. Quite understandably, they were met with a certain amount of skepticism by the traditional coding community. Only when their findings were reproduced by other labs did the turbo idea really take off. It is fair to say that turbo codes have caused a paradigm shift in communications theory. The idea of passing information back and forth between different components in a receiver (so-called iterative processing or turbo processing) has become prevalent in state-of-the-art receiver design. Many books and international scientific conferences are currently devoted to turbo processing.

The original turbo decoder was developed in a somewhat ad-hoc way. An elegant mathematical framework was provided with the introduction of particular graphical models for statistical inference [4, 5] at the end of the twentieth century, describing how iterative receivers can be designed in an almost automatic fashion. The goal of this book is to show how various tasks in the receiver can be cast in this graphical framework. Important practical problems such as decoding, equalization, and multi-user detection

will be treated. Wherever possible, algorithms in pseudo-code will be provided to allow the reader to transfer knowledge from this book directly into a practical implementation. More importantly, it is my hope that the reader will discover the inherent beauty of these graphical models, and realize that they can be applied to a wide variety of problems far beyond Shannon's original coding problem.

1.2 The structure of this book

This book is organized as follows.

- In **Chapter 2**, we will give a brief overview of several important digital transmission schemes. We will cover single- and multi-carrier transmission, single- and multi-antenna transmission, and single- and multi-user transmission. For every one of these transmission schemes, a suitable receiver needs to be developed on the basis of some optimality criterion.
- **Chapter 3** deals with this optimality criterion. We will describe basic concepts from Bayesian estimation theory and Monte Carlo methods.
- In **Chapter 4** we will introduce the concept of factor graphs. Factor graphs are a way to represent graphically the factorization of a function. We will discuss factor graphs in detail in an abstract setting and show how marginals of functions can be computed by message-passing on the corresponding factor graph.
- **Chapter 5** ties together the knowledge from Chapter 3 and Chapter 4. We will show how to solve inference and estimation problems using factor graphs.
- Inference on state-space models appears in many engineering problems. **Chapter 6** is devoted to the application of factor graphs to such models. We will treat hidden Markov models, Kalman filters, and particle filters.
- After a detour into estimation theory and factor graphs, we return to receiver design in **Chapter 7.** We will show how, at least in principle, a digital receiver based on factor graphs can be built. Four critical functions will be revealed: decoding, demapping, equalization, and the conversion of the received waveform into a suitable observation.
- Decoding will be the topic of **Chapter 8**, where we will discuss four important types of error-correcting codes: repeat–accumulate codes, low-density parity-check codes, convolutional codes, and turbo codes.
- In **Chapter 9** we will cover demapping for bit-interleaved coded modulation and trellis-coded modulation in the factor-graph framework.
- Equalization techniques will be treated in **Chapter 10** in a general setting. A variety of general-purpose equalizers will be derived.
- **Chapter 11** deals with equalization for single-user, single-antenna transmission. For every transmission scheme, we show how to convert the received waveform into a suitable observation, and point out which equalizers from Chapter 10 can be applied.
- Equalization for multi-antenna transmission will be discussed in **Chapter 12**. Conversion to suitable observations and factor-graph equalization strategies will be covered.

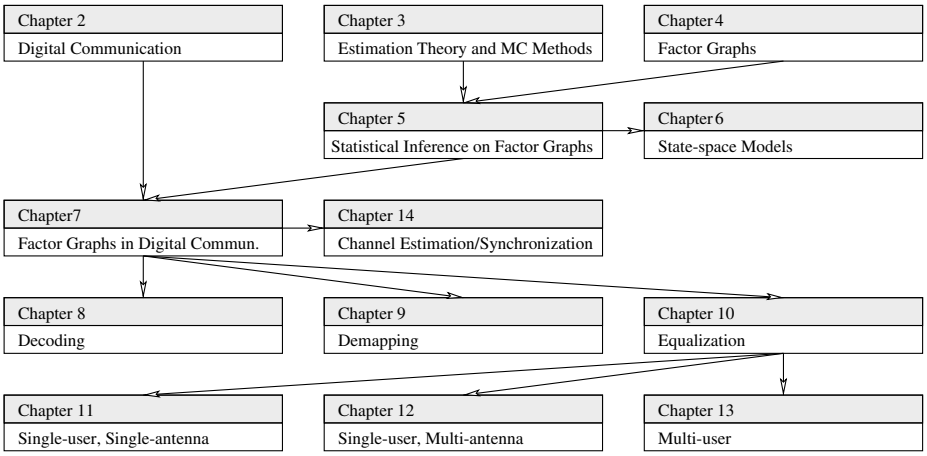

Figure 1.1. A graph representing the interdependencies between chapters.

- Multi-user transmission is the topic of **Chapter 13**. As in the previous two chapters, suitable observations will be determined and equalizers from Chapter 10 will be selected.
- **Chapter 14** will deal with channel estimation and synchronization. We will show how the unknown channel parameters can be incorporated into the factor-graph framework.
- The appendices make up **Chapter 15**. The reader is encouraged to browse through them at this point.

The logical relations between chapters are depicted in Fig. 1.1.

2 Digital communication

2.1 Introduction

As with any good story, it is best to start at the very beginning. Digital communication deals with the transmission of binary information (obtained as the output of a source encoder) from a *transmitter* to a *receiver*. The transmitter converts the binary information to an analog waveform and sends this waveform over a physical medium, such as a wire or open space, which we will call the *channel*. As we shall see, the channel modifies the waveform in several ways. At the receiver, this modified waveform is further corrupted due to thermal noise. Not only does the receiver have to recover the original binary information, but also it must deal with channel effects, thermal noise, and synchronization issues. All in all, the receiver has the bad end of the deal in digital communications. For this reason, this book deals mainly with receiver design, and only to a very small extent with the transmitter.

In this chapter we will describe several digital transmission schemes, detailing how binary information is converted into a waveform at the transmitter side, and how a corrupted version of this waveform arrives at the receiver side. Right now, our focus is not on how the corresponding receivers should be designed. Since there is a myriad of digital transmission schemes, we are obliged to limit ourselves to some of the most important ones. It is my hope that the receivers we will design for this limited set of transmission schemes will give the reader inspiration in developing novel receivers and understand existing receivers. We will assume that the reader has at least a passing familiarity with these transmission schemes. Textbooks and reference books on digital communications include [6–15].

This chapter is organized as follows.

- We will start with the basic principles of converting a sequence of bits into a baseband complex waveform in **Section 2.2**. This process consists of encoding the information, followed by mapping onto a signaling constellation and pulse-shaping.
- In **Section 2.3**, we then move on to the classical transmission scheme with a single transmitter using a single antenna. Both single-carrier and multi-carrier transmission will be described.
- These transmission schemes are then generalized to multi-antenna (**Section 2.4**) and multi-user (**Section 2.5**) scenarios.
- After this brief overview of transmission schemes, we will outline the main goals and working assumptions of this book in **Section 2.6**.

2.2 Digital communication

2.2.1 From bits to waveform

In digital communication, the goal is to convey a sequence of binary information digits (bits, belonging to the set $\mathbb{B} = \{0, 1\}$) from the transmitter to the receiver. The binary information sequence (which may be infinitely long) is segmented into blocks of length N_b. A single block, say $\mathbf{b} \in \mathbb{B}^{N_\mathrm{b}}$, is referred to as an information word. Every information word is protected against channel effects by encoding it with a channel encoder, which converts the N_b information bits into N_c coded bits, with $N_\mathrm{c} > N_\mathrm{b}$, using an encoding function $f_\mathrm{c} : \mathbb{B}^{N_\mathrm{b}} \to \mathbb{B}^{N_\mathrm{c}}$. For instance, we can apply a convolutional code, a turbo code, or a low-density parity-check code. This results in a (longer) binary sequence $\mathbf{c} = f_\mathrm{c}(\mathbf{b})$.

Example 2.1 (Repetition code). *Probably the easiest way of encoding an information stream is by using a repetition code, whereby we simply repeat every bit K times. For instance, when $N_\mathrm{b} = 3$, $K = 2$, and $\mathbf{b} = [011]$, this yields*

$$\begin{aligned}\mathbf{c} &= f_\mathrm{c}([011]) \\ &= [001111]\end{aligned}$$

so that $N_\mathrm{c} = N_\mathrm{b} \times K = 6$. In practice, N_b can be very large.

After encoding, the coded bits are converted to a sequence $\mathbf{a} = f_\mathrm{a}(\mathbf{c})$ of N_s complex coded symbols using a mapping function f_a, where the kth symbol, a_k, belongs to a signaling constellation Ω_k. For instance, we can use bit-interleaved coded modulation (BICM) or trellis-coded modulation (TCM).

Example 2.2 (Mapping). *The most simple (and most common) way of mapping $\mathbf{c}$ to $\mathbf{a}$ is as follows. Suppose that we have a signaling constellation $\Omega = \{-1, +1, -\mathrm{j}, +\mathrm{j}\}$, where j is the imaginary unit ($\mathrm{j} = \sqrt{-1}$). We select the same constellation for every symbol a_k. With every element in Ω, we associate a unique bit-string. Since there are $|\Omega| = 4$ elements in Ω, we can associate $\log_2|\Omega| = 2$ bits with every element in Ω. This is known as quadrature phase-shift keying (QPSK). We can now define the following mapping $\phi : \mathbb{B} \times \mathbb{B} \to \Omega$*

$$\begin{aligned}\phi(0,0) &= +1, \\ \phi(0,1) &= -1, \\ \phi(1,0) &= -\mathrm{j}, \\ \phi(1,1) &= +\mathrm{j}.\end{aligned}$$

We break up $\mathbf{c}$ into blocks of length $\log_2|\Omega| = 2$ and map every block to a constellation point using the function $\phi(\cdot)$. In our case, using the sequence from the previous example

$\mathbf{c} = [001111]$, *we obtain*

$$\begin{aligned}\mathbf{a} &= [\phi(0,0)\phi(1,1)\phi(1,1)] \\ &= [+1 \; +\mathrm{j} \; +\mathrm{j}].\end{aligned}$$

Observe that $N_\mathrm{s} = N_\mathrm{c}/\log_2|\Omega|$.

Finally, once encoding and mapping are completed, the symbols in $\mathbf{a}$ are embedded in a (possibly infinitely long) data stream and pulse-shaped, giving rise to a complex baseband signal

$$s(t) = \sqrt{E_\mathrm{s}} \sum_{k=-\infty}^{+\infty} a_k p_k(t), \tag{2.1}$$

where E_s is the energy per transmitted coded symbol and $p_k(t)$ is a unit-energy transmit pulse corresponding to the kth symbol. This implies that

$$\int_{-\infty}^{+\infty} |p_k(t)|^2 \mathrm{d}t = 1. \tag{2.2}$$

The signal $s(t)$ is modulated onto a carrier waveform with carrier frequency f_C, yielding a real (as opposed to complex) radio-frequency (RF) signal

$$s_\mathrm{RF}(t) = \Re\left\{\sqrt{2}s(t)\mathrm{e}^{\mathrm{j}2\pi f_\mathrm{C} t}\right\}. \tag{2.3}$$

The RF signal propagates through a physical medium (e.g., a wireless channel, an optical fiber), and is corrupted by thermal noise in the receiver, resulting in a received signal $r_\mathrm{RF}(t)$. The received RF signal is down-converted to complex baseband, giving rise to an equivalent complex baseband received signal $r(t)$. We will follow the common practice of working only with equivalent baseband signals. This leads to the following observed signal at the receiver:

$$r(t) = \sqrt{E_\mathrm{s}} \sum_{k=-\infty}^{+\infty} a_k h_k(t) + n(t), \tag{2.4}$$

where $h_k(t)$ is the equivalent channel for the kth symbol, encompassing the transmit pulse and the equivalent baseband physical channel $h_\mathrm{ch}(t)$, and $n(t)$ is a zero-mean complex white Gaussian process with power-spectral density $N_0/2$ both for the real and for the imaginary component:

$$\mathbb{E}\{N(t)N^*(u)\} = N_0\delta(t-u). \tag{2.5}$$

The equivalent channel $h_k(t)$ can be written as the convolution of the transmit pulse and the physical channel:

$$h_k(t) = \int_{-\infty}^{+\infty} p_k(u) h_{\text{ch}}(t-u) \mathrm{d}u. \tag{2.6}$$

Example 2.3 (Pulse-shaping). *A simple type of complex baseband signal is formed by setting $p_k(t) = p(t - kT)$, where $p(t)$ is a unit-energy square pulse, defined as*

$$p(t) = \begin{cases} 1/\sqrt{T} & 0 \le t < T \\ 0 & \textit{else} \end{cases}$$

so that we transmit one data symbol every T seconds. Going back to our previous example, where $N_s = 3$ and $\mathbf{a} = [+1 \; +\mathrm{j} \; +\mathrm{j}]$, this results in a signal shown in Fig. 2.1 (left-hand side), where we depict the real and imaginary parts of the signal $s(t)$. Let us assume that the equivalent baseband channel is given by $h_{ch}(t) = \exp(\mathrm{j}\pi/4)\delta(t-\tau)$, for some propagation delay $\tau \in \mathbb{R}$, then

$$r(t) = \sqrt{E_s} \sum_{k=0}^{2} a_k \mathrm{e}^{\mathrm{j}\pi/4} p(t - kT - \tau) + n(t).$$

A noiseless version of $r(t)$ is also shown in Fig. 2.1 (right-hand side).

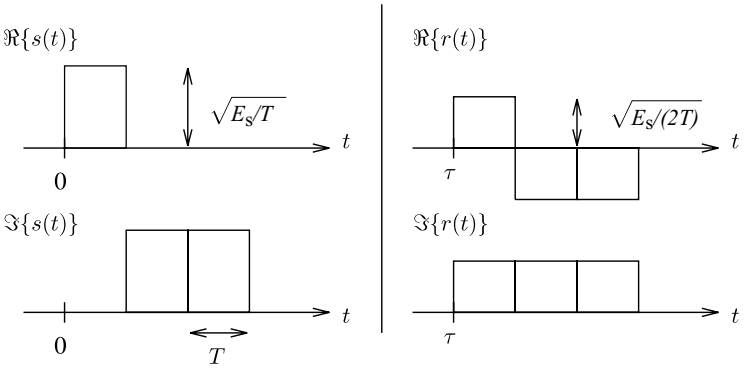

Figure 2.1. Pulse-shaping using square pulses. On the left is the transmitted signal with transmitted sequence $[+1 \; +\mathrm{j} \; +\mathrm{j}]$. On the right is the (noiseless) received signal, after a phase rotation of $\pi/4$ and a delay of τ.

2.2.2 Channel model

A common channel model for wireless communication is the multi-path model, whereby the channel impulse response consists of a number of distinct paths [16]:

$$h_{\text{ch}}(t) = \sum_{l=0}^{L-1} \alpha_l \delta(t - \tau_l), \tag{2.7}$$

where α_l and τ_l are the complex gain and the propagation delay of the lth path. We usually consider these paths to be resolvable (meaning that $\tau_{l+1} - \tau_l \gg 1/B$, where B represents the bandwidth of $p(t)$). This is without loss of generality, since unresolvable paths can be combined into a single path, and the complex gains added.

A channel is said to be *frequency-selective* when it has at least two resolvable paths. Otherwise the channel is frequency-non-selective (also known as a *frequency-flat channel*). The delay of the first path, τ_0, corresponds to the *propagation delay* of the signal through the channel.

2.2.3 Communication schemes

We can distinguish between single-antenna and multi-antenna transmission, between single-carrier and multi-carrier transmission, and between single-user and multi-user transmission. In the next few sections, we will describe these schemes in more detail. Our aim is not completeness, but rather to touch on a few selected schemes for which we will later design appropriate receivers.

2.3 Single-user, single-antenna communication

We first consider a system where there is only a single user transmitting. Both the transmitter and the receiver are equipped with a single antenna. The transmitter wishes to transmit a long data stream (consisting of many codewords).

2.3.1 Single-carrier modulation

In single-carrier modulation we transmit a symbol every T seconds over a single carrier. This gives rise to the following transmitted signal:

$$s(t) = \sqrt{E_{\text{s}}} \sum_{k=-\infty}^{+\infty} a_k p(t - kT). \tag{2.8}$$

The transmit pulse $p(t)$ is usually selected according to specific criteria. For instance, we like our pulses to have finite bandwidth and to be unit-energy square-root Nyquist pulses for a rate $1/T$. We remind the reader that a unit-energy square-root Nyquist pulse

$p(t)$ for a rate $1/T$ satisfies, for $k \in \mathbb{Z}$,

$$g(kT) = \begin{cases} 1 & k = 0, \\ 0 & \text{else}, \end{cases} \tag{2.9}$$

where

$$g(t) = \int_{-\infty}^{+\infty} p(u)p^*(t+u)\mathrm{d}u. \tag{2.10}$$

The pulse $g(t)$ is then a Nyquist pulse for a rate $1/T$. The received signal can be expressed as

$$r(t) = \sqrt{E_s} \sum_{k=-\infty}^{+\infty} a_k h(t - kT) + n(t), \tag{2.11}$$

where $h(t)$ is the equivalent channel (given by the convolution of the transmit pulse and the physical channel $h_{\text{ch}}(t)$, see (2.6)) and $n(t)$ is a complex white Gaussian noise process.

2.3.2 Multi-carrier modulation – OFDM

Orthogonal frequency-division multiplexing (OFDM) is a popular multi-carrier transmission technique that avoids inter-symbol interference over frequency-selective channels [17,18]. Intuitively, the idea of OFDM is to break up the transmission bandwidth into narrow subbands (subcarriers) such that the channel is frequency-flat on every subcarrier. Inter-symbol interference is avoided by pre-appending a cyclic prefix to the transmitted symbols.

More formally, we break up the data stream into segments of length N_{FFT}, where N_{FFT} is usually a power of 2 (for instance, 256 or 1024). Let us call one such segment $\mathbf{a}$, an $N_{\text{FFT}} \times 1$ vector. We multiply $\mathbf{a}$ by an N_{FFT}-point inverse discrete Fourier transform (IDFT) matrix $\mathbf{F}$, yielding a vector

$$\check{\mathbf{a}} = \mathbf{F}\mathbf{a}, \tag{2.12}$$

where

$$F_{m,n} = \frac{1}{\sqrt{N_{\text{FFT}}}} \exp\left(\mathrm{j}2\pi \frac{m \times n}{N_{\text{FFT}}}\right), \tag{2.13}$$

for $m, n \in \{0, \ldots, N_{\text{FFT}} - 1\}$. Usually the operation (2.12) is performed by means of a computationally efficient inverse fast Fourier transform (FFT). We then pre-append the last N_{CP} of $\check{\mathbf{a}}$ symbols to $\check{\mathbf{a}}$, and obtain a vector of length $N_{\text{FFT}} + N_{\text{CP}}$, known as an *OFDM*

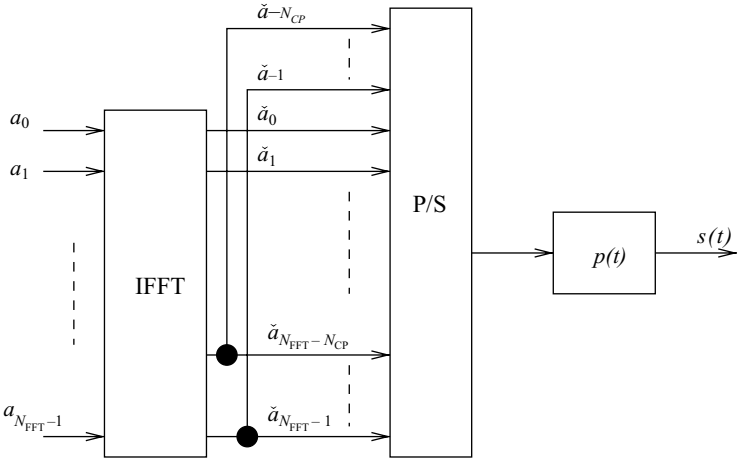

Figure 2.2. An OFDM transmitter. The data symbols are passed through an IFFT, a cyclic prefix is added and, after parallel-to-serial conversion, pulse-shaping is employed.

symbol:

$$\left[\underbrace{\check{a}_{-N_{\text{CP}}}\ldots\check{a}_{-2}\,\check{a}_{-1}}_{\text{cyclic prefix}}\;\underbrace{\check{a}_0\,\check{a}_1\,\check{a}_2\ldots\check{a}_{N_{\text{FFT}}-1}}_{\check{\mathbf{a}}^{\text{T}}}\right]^{\text{T}} \tag{2.14}$$

where $\check{a}_{-l}=\check{a}_{N_{\text{FFT}}-l}$, for $l=1,\ldots,N_{\text{CP}}$. The sequence $\left[\check{a}_{-N_{\text{CP}}}\ldots\check{a}_{-2}\check{a}_{-1}\right]^{\text{T}}$ of length N_{CP} is known as the *cyclic prefix*. Finally, we transmit one OFDM symbol using the following complex baseband signal (see Fig. 2.2):

$$s(t)=\sqrt{\frac{E_s N_{\text{FFT}}}{N_{\text{FFT}}+N_{\text{CP}}}}\sum_{l=-N_{\text{CP}}}^{N_{\text{FFT}}-1}\check{a}_l p(t-lT), \tag{2.15}$$

using a unit-energy transmit pulse $p(t)$. We can also transmit a sequence of OFDM symbols consecutively as

$$s(t)=\sqrt{\frac{E_s N_{\text{FFT}}}{N_{\text{FFT}}+N_{\text{CP}}}}\sum_{k=-\infty}^{+\infty}\sum_{l=-N_{\text{CP}}}^{N_{\text{FFT}}-1}\check{a}_{l,k}p(t-lT-kT_{\text{OFDM}}), \tag{2.16}$$

where $\check{a}_{l,k}$ is the lth component of the kth OFDM symbol and $T_{\text{OFDM}}=T(N_{\text{FFT}}+N_{\text{CP}})$, the symbol duration corresponding to a single OFDM symbol. The received signal is often written as

$$r(t)=\sum_{k=-\infty}^{+\infty}\sum_{l=-N_{\text{CP}}}^{N_{\text{FFT}}-1}\check{a}_l h(t-lT-kT_{\text{OFDM}})+n(t), \tag{2.17}$$