3 Receiver design for the continuous-time AWGN channel: Second layer

3.1 Introduction

In Chapter 2 we focused on the receiver for the discrete-time AWGN (additive white Gaussian noise) channel. In this chapter, we address the same problem for a channel model closer to reality, namely the *continuous-time AWGN channel*. Apart from the channel model, the assumptions and the goal are the same: We assume that the source statistic, the transmitter, and the channel are given to us and we seek to understand what the receiver has to do to minimize the error probability. We are also interested in the resulting error probability, but this follows from Chapter 2 with no extra work. The setup is shown in Figure 3.1.

The channel of Figure 3.1 captures the most important aspect of all real-world channels, namely the presence of additive noise. Owing to the central limit theorem, the assumption that the noise is Gaussian is often a very good one. In Section 3.6 we discuss additional channel properties that also affect the design and performance of a communication system.

EXAMPLE 3.1 *A cable is a good example of a channel that can be modeled by the continuous-time AWGN channel. If the cable's frequency response cannot be considered as constant over the signal's bandwidth, then the cable's filtering effect also needs to be taken into consideration. We discuss this in Section 3.6. Another good example is the channel between the antenna of a geostationary satellite and the antenna of the corresponding Earth station. For the communication in either direction we can consider the model of Figure 3.1.* □

Although our primary focus is on the receiver, in this chapter we also gain valuable insight into the transmitter structure. First we need to introduce the notion of signal's energy and specify two mild technical restrictions that we impose on the signal set $\mathcal{W} = \{w_0(t), \ldots, w_{m-1}(t)\}$.

EXAMPLE 3.2 *Suppose that $w_i(t)$ is the voltage feeding the antenna of a transmitter when $H = i$. An antenna has an internal impedance Z. A typical value for Z is 50 ohms. Assuming that Z is purely resistive, the current at the feeding point is $w_i(t)/Z$, the instantaneous power is $w_i^2(t)/Z$, and the energy transferred to the antenna is $\frac{1}{Z}\int w_i^2(t)dt$. Alternatively, if the $w_i(t)$ is the current feeding the antenna when $H = i$, the voltage at the feeding point is $w_i(t)Z$, the instantaneous power is*

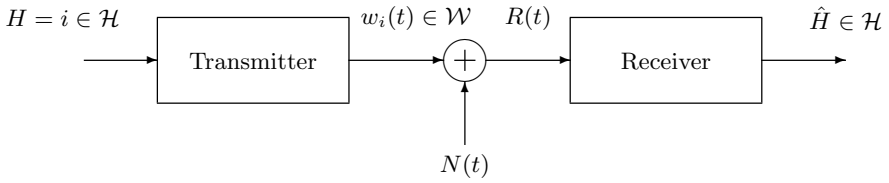


Figure 3.1. Communication across the continuous-time AWGN channel.

$w_i^2(t)Z$, and the energy is $Z \int w_i^2(t)dt$. In both cases the energy is proportional to $\|w_i\|^2 = \int |w_i(t)|^2 dt$. □

As in the above example, the squared norm of a signal $w_i(t)$ is generally associated with the signal's energy. It is quite natural to assume that we communicate via finite-energy signals. This is the first restriction on $\mathcal{W}$. A linear combination of a finite number of finite-energy signals is itself a finite-energy signal. Hence, every vector of the vector space $\mathcal{V}$ spanned by $\mathcal{W}$ is a square-integrable function. The second requirement is that if $v \in \mathcal{V}$ has a vanishing norm, then $v(t)$ vanishes for all t. Together, these requirements imply that $\mathcal{V}$ is an inner product space of square-integrable functions. (See Example 2.39.)

EXAMPLE 3.3 (Continuous functions) *Let $v : \mathbb{R} \to \mathbb{R}$ be a continuous function and suppose that $|v(t_0)| = a$ for some t_0 and some positive a. By continuity, there exists an $\epsilon > 0$ such that $|v(t)| > \frac{a}{2}$ for all $t \in \mathcal{I}$ where $\mathcal{I} = (t_0 - \frac{\epsilon}{2}, t_0 + \frac{\epsilon}{2})$. It follows that*

$$\|v(t)\|^2 \geq \|v(t)\mathbb{1}\{t \in \mathcal{I}\}\|^2 \geq \frac{a^2\epsilon}{4} > 0.$$

We conclude that if a continuous function has a vanishing norm, then the function vanishes everywhere. □

All signals that represent real-world communication signals are finite-energy and continuous. Hence the vector space they span is always an inner product space.

This is a good place to mention the various reasons we are interested in the signal's energy or, somewhat equivalently, in the signal's power, which is the energy per second. First, for safety and for spectrum reusability, there are regulations that limit the power of a transmitted signal. Second, for mobile devices, the energy of the transmitted signal comes from the battery: a battery charge lasts longer if we decrease the signal's power. Third, with no limitation to the signal's power, we can transmit across a continuous-time AWGN channel at any desired rate, regardless of the available bandwidth and of the target error probability. Hence, it would be unfair to compare signaling methods that do not use the same power.

For now, we assume that $\mathcal{W}$ is given to us. The problem of choosing a suitable set $\mathcal{W}$ of signals will be studied in subsequent chapters.

The highlight of the chapter is the power of abstraction. The receiver design for the *discrete-time* AWGN channel relied on geometrical ideas that can be

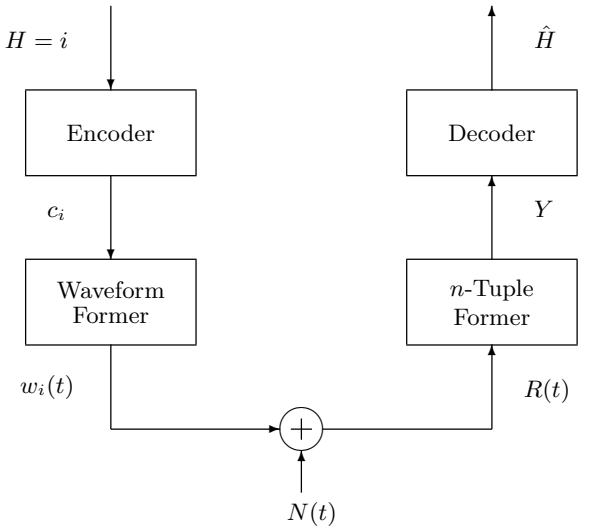


Figure 3.2. Waveform channel abstraction.

formulated whenever we are in an inner product space. We will use the same ideas for the *continuous-time* AWGN channel.

The main result is a decomposition of the sender and the receiver into the building blocks shown in Figure 3.2. We will see that, without loss of generality, we can (and should) think of the transmitter as consisting of an *encoder* that maps the message $i \in \mathcal{H}$ into an n-tuple c_i, as in the previous chapter, followed by a *waveform former* that maps c_i into a waveform $w_i(t)$. Similarly, we will see that the receiver can consist of an *n-tuple former* that takes the channel output and produces an n-tuple Y. The behavior from the waveform former input to the n-tuple former output is that of the discrete-time AWGN channel considered in the previous chapter. Hence we know already what the *decoder* of Figure 3.2 should do with the n-tuple former output.

In this chapter (like in the previous one) the vectors (functions) are real-valued. Hence, we could use the formalism that applies to real inner product spaces. Yet, in preparation of Chapter 7, we use the formalism for complex inner product spaces. This mainly concerns the standard inner product between functions, where we write $\langle a, b\rangle = \int a(t)b^*(t)dt$ instead of $\langle a, b\rangle = \int a(t)b(t)dt$. A similar comment applies to the definition of covariance, where for zero-mean random variables we use $\mathrm{cov}(Z_i, Z_j) = \mathbb{E}\left[Z_i Z_j^*\right]$ instead of $\mathrm{cov}(Z_i, Z_j) = \mathbb{E}\left[Z_i Z_j\right]$.

3.2 White Gaussian noise

The purpose of this section is to introduce the basics of *white Gaussian noise* $N(t)$. The standard approach is to give a mathematical description of $N(t)$, but this

requires measure theory if done rigorously. The good news is that a mathematical model of $N(t)$ is not needed because $N(t)$ is not observable through physical experiments. (The reason will become clear shortly.) Our approach is to model what we can actually measure. We assume a working knowledge of Gaussian random vectors (reviewed in Appendix 2.10).

A receiver is an electrical instrument that connects to the channel output via a cable. For instance, in wireless communication, we might consider the channel output to be the output of the receiving antenna; in which case, the cable is the one that connects the antenna to the receiver. A cable is a linear time-invariant filter. Hence, we can assume that all the observations made by the receiver are through some linear time-invariant filter.

So if $N(t)$ represents the noise introduced by the channel, the receiver sees, at best, a filtered version $Z(t)$ of $N(t)$. We model $Z(t)$ as a stochastic process and, as such, it is described by the statistic of $Z(t_1), Z(t_2), \ldots, Z(t_k)$ for any positive integer k and any finite collection of sampling times $t_1, t_2, \ldots, t_k$.

If the filter impulse response is $h(t)$, then linear system theory suggests that

$$Z(t) = \int N(\alpha)h(t-\alpha)d\alpha$$

and

$$Z(t_i) = \int N(\alpha)h(t_i-\alpha)d\alpha, \tag{3.1}$$

but the validity of these expressions needs to be justified, because $N(t)$ is not a deterministic signal. It is possible to define $N(t)$ as a stochastic process and prove that the (Lebesgue) integral in (3.1) is well defined; but we avoid this path which, as already mentioned, requires measure theory. In this text, equation (3.1) is shorthand for the statement "$Z(t_i)$ is the random variable that models the output at time t_i of a linear time-invariant filter of impulse response $h(t)$ fed with white Gaussian noise $N(t)$". Notice that $h(t_i-\alpha)$ is a function of α that we can rename as $g_i(\alpha)$. Now we are in the position to define white Gaussian noise.

DEFINITION 3.4 *$N(t)$ is white Gaussian noise of power spectral density $\frac{N_0}{2}$ if, for any finite collection of real-valued $\mathcal{L}_2$ functions $g_1(\alpha), \ldots, g_k(\alpha)$,*

$$Z_i = \int N(\alpha)g_i(\alpha)d\alpha, \quad i = 1, 2, \ldots, k \tag{3.2}$$

is a collection of zero-mean jointly Gaussian random variables of covariance

$$\operatorname{cov}\big(Z_i, Z_j\big) = \mathbb{E}\left[Z_i Z_j^*\right] = \frac{N_0}{2}\int g_i(t)g_j^*(t)dt = \frac{N_0}{2}\langle g_i, g_j\rangle. \tag{3.3}$$

□

If we are not evaluating the integral in (3.2), how do we know if $N(t)$ is white Gaussian noise? In this text, when applicable, we say that $N(t)$ is white Gaussian noise, in which case we can use (3.3) as we see fit. In the real world, often we know enough about the channel to know whether or not its noise can be modeled as

white and Gaussian. This knowledge could come from a mathematical model of the channel. Another possibility is that we perform measurements and verify that they behave according to Definition 3.4.

Owing to its importance and frequent use, we formulate the following special case as a lemma. It is the most important fact that should be remembered about white Gaussian noise.

LEMMA 3.5 *Let $\{g_1(t), \ldots, g_k(t)\}$ be an orthonormal set of real-valued functions. Then $Z = (Z_1, \ldots, Z_k)^{\mathsf{T}}$, with Z_i defined as in (3.2), is a zero-mean Gaussian random vector with iid components of variance $\sigma^2 = \frac{N_0}{2}$.* □

Proof The proof is a straightforward application of the definitions. □

EXAMPLE 3.6 *Consider two bandpass filters that have non-overlapping frequency responses but are otherwise identical, i.e. if we frequency-translate the frequency response of one filter by the proper amount we obtain the frequency response of the other filter. By Parseval's relationship, the corresponding impulse responses are orthogonal to one another. If we feed the two filters with white Gaussian noise and sample their output (even at different times), we obtain two iid Gaussian random variables. We could extend the experiment (in the obvious way) to n filters of non-overlapping frequency responses, and would obtain n random variables that are iid – hence of identical variance. This explains why the noise is called white: like for white light, white Gaussian noise has its power equally distributed among all frequencies.* □

Are there other types of noise? Yes, there are. For instance, there are natural and man-made electromagnetic noises. The noise produced by electric motors and that produced by power lines are examples of man-made noise. Man-made noise is typically neither white nor Gaussian. The good news is that a careful design should be able to ensure that the receiver picks up a negligible amount of man-made noise (if any). Natural noise is unavoidable. Every conductor (resistor) produces thermal (Johnson) noise. (See Appendix 3.10.) The assumption that thermal noise is white and Gaussian is an excellent one. Other examples of natural noise are solar noise and cosmic noise. A receiving antenna picks up these noises, the intensity of which depends on the antenna's gain and pointing direction. A current in a conductor gives rise to shot noise. Shot noise originates from the discrete nature of the electric charges. Wikipedia is a good reference to learn more about various noise sources.

3.3 Observables and sufficient statistics

Recall that the setup is that of Figure 3.1, where $N(t)$ is white Gaussian noise. As discussed in Section 3.2, owing to the noise, the channel output $R(t)$ is not observable. What we can observe via physical experiments (measurements) are k-tuples $V = (V_1, \ldots, V_k)^{\mathsf{T}}$ such that

$$V_i = \int_{-\infty}^{\infty} R(\alpha) g_i^*(\alpha) d\alpha, \qquad i = 1, 2, \ldots, k, \tag{3.4}$$

where k is an arbitrary positive integer and $g_1(t), \ldots, g_k(t)$ are arbitrary finite-energy waveforms. The complex conjugate operator "*" on $g_i^*(\alpha)$ is superfluous for real-valued signals but, as we will see in Chapter 7, the baseband representation of a passband impulse response is complex-valued.

Notice that we assume that we can perform an arbitrarily large but *finite* number k of measurements. By disallowing infinite measurements we avoid distracting mathematical subtleties without losing anything of engineering relevance.

It is important to point out that the kind of measurements we consider is quite general. For instance, we can pass $R(t)$ through an ideal lowpass filter of cutoff frequency B for some huge B (say 10^{10} Hz) and collect an arbitrary large number of samples taken every $\frac{1}{2B}$ seconds so as to fulfill the sampling theorem (Theorem 5.2). In fact, by choosing $g_i(t) = h(\frac{i}{2B} - t)$, where $h(t)$ is the impulse response of the lowpass filter, V_i becomes the filter output sampled at time $t = \frac{i}{2B}$. As stated by the sampling theorem, from these samples we can reconstruct the filter output. If $R(t)$ consists of a signal plus noise, and the signal is bandlimited to less than B Hz, then from the samples we can reconstruct the signal plus the portion of the noise that has frequency components in $[-B, B]$.

Let $\mathcal{V}$ be the inner product space spanned by the elements of the signal set $\mathcal{W}$ and let $\{\psi_1(t), \ldots, \psi_n(t)\}$ be an arbitrary orthonormal basis for $\mathcal{V}$. We claim that the n-tuple $Y = (Y_1, \ldots, Y_n)^\mathsf{T}$ with ith component

$$Y_i = \int R(\alpha)\psi_i^*(\alpha)d\alpha$$

is a sufficient statistic (for the hypothesis H) among any collection of measurements that contains Y. To prove this claim, let $V = (V_1, \ldots, V_k)^\mathsf{T}$ be the collection of additional measurements made according to (3.4). Let $\mathcal{U}$ be the inner product space spanned by $\mathcal{V} \cup \{g_1(t), \ldots, g_k(t)\}$ and let $\{\psi_1(t), \ldots, \psi_n(t), \phi_1(t), \ldots, \phi_{\tilde{n}}(t)\}$ be an orthonormal basis for $\mathcal{U}$ obtained by extending the orthonormal basis $\{\psi_1(t), \ldots, \psi_n(t)\}$ for $\mathcal{V}$. Define

$$U_i = \int R(\alpha)\phi_i^*(\alpha)d\alpha, \qquad i = 1, \ldots, \tilde{n}.$$

It should be clear that we can recover V from Y and U. This is so because, from the projections onto a basis, we can obtain the projection onto any waveform in the span of the basis. Mathematically,

$$\begin{aligned} V_i &= \int_{-\infty}^{\infty} R(\alpha)g_i^*(\alpha)d\alpha \\ &= \int_{-\infty}^{\infty} R(\alpha)\left[\sum_{j=1}^{n} \xi_{i,j}\psi_j(\alpha) + \sum_{j=1}^{\tilde{n}} \xi_{i,j+n}\phi_j(\alpha)\right]^* d\alpha \\ &= \sum_{j=1}^{n} \xi_{i,j}^* Y_j + \sum_{j=1}^{\tilde{n}} \xi_{i,j+n}^* U_j, \end{aligned}$$

where $\xi_{i,1}, \ldots, \xi_{i,n+\tilde{n}}$ is the unique set of coefficients in the orthonormal expansion of $g_i(t)$ with respect to the basis $\{\psi_1(t), \ldots, \psi_n(t), \phi_1(t), \phi_2(t), \ldots, \phi_{\tilde{n}}(t)\}$.

Hence we can consider (Y, U) as *the observable* and it suffices to show that Y is a sufficient statistic. Note that when $H = i$,

$$Y_j = \int R(\alpha)\psi_j^*(\alpha) = \int \big(w_i(\alpha) + N(\alpha)\big)\psi_j^*(\alpha)d\alpha = c_{i,j} + Z_{|\mathcal{V},j},$$

where $c_{i,j}$ is the jth component of the n-tuple of coefficients c_i that represents the waveform $w_i(t)$ with respect to the chosen orthonormal basis, and $Z_{|\mathcal{V},j}$ is a zero-mean Gaussian random variable of variance $\frac{N_0}{2}$. The notation $Z_{|\mathcal{V},j}$ is meant to remind us that this random variable is obtained by "projecting" the noise onto the jth element of the chosen orthonormal basis for $\mathcal{V}$. Using n-tuple notation, we obtain the following statistic

$$H = i, \qquad Y = c_i + Z_{|\mathcal{V}},$$

where $Z_{|\mathcal{V}} \sim \mathcal{N}(0, \frac{N_0}{2} I_n)$. Similarly,

$$U_j = \int R(\alpha)\phi_j^*(\alpha) = \int \big(w_i(\alpha) + N(\alpha)\big)\phi_j^*(\alpha)d\alpha = \int N(\alpha)\phi_j^*(\alpha)d\alpha = Z_{\perp\mathcal{V},j},$$

where we used the fact that $w_i(t)$ is in the subspace spanned by $\{\psi_1(t), \ldots, \psi_n(t)\}$ and therefore it is orthogonal to $\phi_j(t)$ for each $j = 1, 2, \ldots, \tilde{n}$. The notation $Z_{\perp\mathcal{V},j}$ reminds us that this random variable is obtained by "projecting" the noise onto the jth element of an orthonormal basis that is orthogonal to $\mathcal{V}$. Using n-tuple notation, we obtain

$$H = i, \qquad U = Z_{\perp\mathcal{V}},$$

where $Z_{\perp\mathcal{V}} \sim \mathcal{N}(0, \frac{N_0}{2} I_{\tilde{n}})$. Furthermore, $Z_{|\mathcal{V}}$ and $Z_{\perp\mathcal{V}}$ are independent of each other and of H. The conditional density of Y, U given H is

$$f_{Y,U|H}(y, u|i) = f_{Y|H}(y|i) f_U(u).$$

From the Fisher–Neyman factorization theorem (Theorem 2.13, Chapter 2, with $h(y, u) = f_U(u)$, $T(y, u) = y$, and $g_i(T(y, u)) = f_{Y|H}(y|i)$), we see that Y is a sufficient statistic and U is irrelevant as claimed.

Figure 3.3 depicts what is going on, which we summarize as follows:

$Y = c_i + Z_{\|\mathcal{V}}$	is a sufficient statistic: it is the projection of $R(t)$ onto the signal space $\mathcal{V}$;
$U = Z_{\perp\mathcal{V}}$	is irrelevant: it contains only independent noise.

Could we prove that a subset of the components of Y is *not* a sufficient statistic? Yes, we could. Here is the outline of a proof. Without loss of essential generality, let us think of Y as consisting of two parts, Y_a and Y_b. Similarly, we decompose every c_i into the corresponding parts c_{ia} and c_{ib}. The claim is that H followed by Y_a followed by (Y_a, Y_b) does *not* form a Markov chain in that order. In fact when $H = i$, Y_b consists of c_{ib} plus noise. Since c_{ib} cannot be deduced from Y_a in

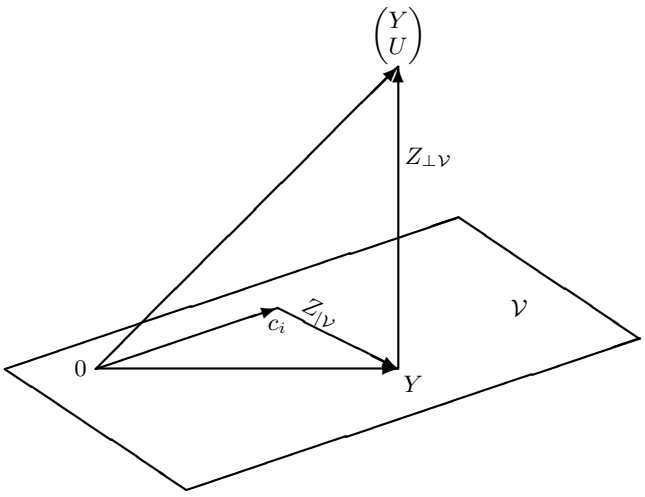


Figure 3.3. The vector of measurements $(Y^\mathsf{T}, U^\mathsf{T})^\mathsf{T}$ describes the projection of the received signal $R(t)$ onto $\mathcal{U}$. The vector Y describes the projection of $R(t)$ onto $\mathcal{V}$.

general (or else we would not bother sending c_{ib}), it follows that the statistic of Y_b depends on i even if we know the realization of Y_a.

3.4 Transmitter and receiver architecture

The results of the previous section tell us that a MAP receiver for the waveform AWGN channel can be structured as shown in Figure 3.4. We see that the receiver front end computes $Y \in \mathbb{R}^n$ from $R(t)$ in a block that we call *n-tuple former.* (The name is not standard.) Thus the n-tuple former performs a huge data reduction from the channel output $R(t)$ to the sufficient statistic Y. The hypothesis testing problem based on the observable Y is

$$H = i: \qquad Y = c_i + Z,$$

where $Z \sim \mathcal{N}(0, \frac{N_0}{2} I_n)$ is independent of H. This is precisely the hypothesis testing problem studied in Chapter 2 in conjunction with a transmitter that sends $c_i \in \mathbb{R}^n$ to signal message i across the *discrete-time* AWGN channel. As shown in the figure, we can also decompose the transmitter into a module that produces c_i, called *encoder*, and a module that produces $w_i(t)$, called *waveform former.* (Once again, the terminology is not standard.) Henceforth the n-tuple of coefficients c_i will be referred to as the *codeword* associated to $w_i(t)$. Figure 3.4 is the main result of the chapter. It implies that the decomposition of the transmitter and the receiver as depicted in Figure 3.2 is indeed completely general and it gives details about the waveform former and the n-tuple former.

Everything that we learned about a decoder for the discrete-time AWGN channel is applicable to the decoder of the continuous-time AWGN channel. Incidentally,

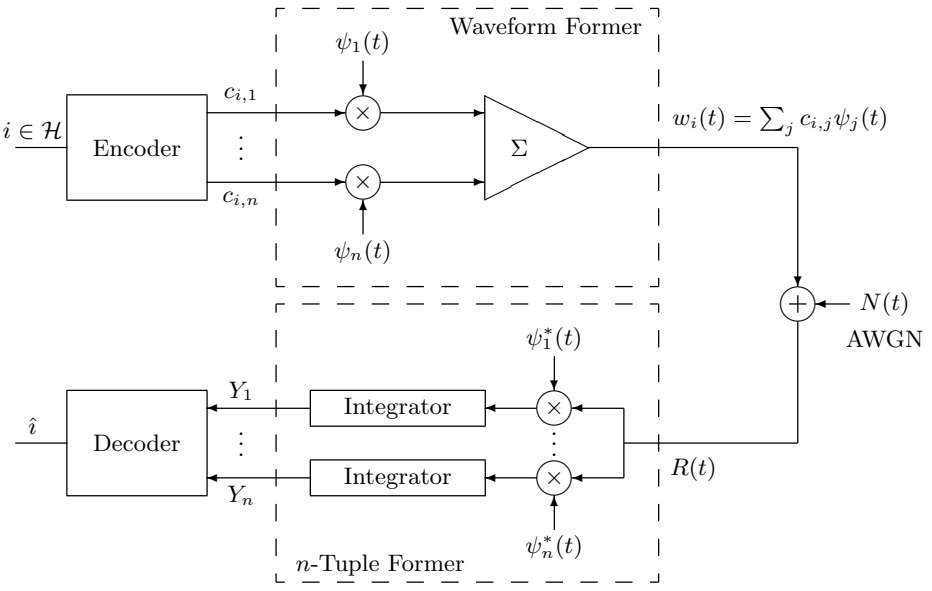


Figure 3.4. Decomposed transmitter and receiver.

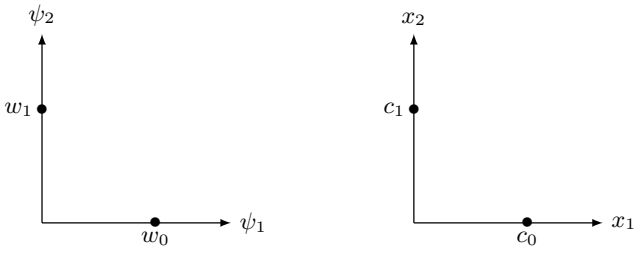
the decomposition of Figure 3.4 is consistent with the layering philosophy of the OSI model (Section 1.1), in the sense that the encoder and decoder are designed as if they were talking to each other directly via a discrete-time AWGN channel. In reality, the channel seen by the encoder/decoder pair is the result of the "service" provided by the waveform former and the n-tuple former.

The above decomposition is useful for the system conception, for the performance analysis, as well as for the system implementation; but of course, we always have the option of implementing the transmitter as a straight map from the message set $\mathcal{H}$ to the waveform set $\mathcal{W}$ without passing through the codebook $\mathcal{C}$. Although such a straight map is a possibility and makes sense for relatively unsophisticated systems, the decomposition into an encoder and a waveform former is standard for modern designs. In fact, information theory, as well as coding theory, devote much attention to the study of encoder/decoder pairs.

The following example is meant to make two important points that apply when we communicate across the continuous-time AWGN channel and make an ML decision. First, sets of continuous-time signals may "look" very different yet they may share the same codebook, which is sufficient to guarantee that the error probability be the same; second, for binary constellations, what matters for the error probability is the distance between the two signals and nothing else.

EXAMPLE 3.7 (Orthogonal signals) *The following four choices of* $\mathcal{W} = \{w_0(t), w_1(t)\}$ *look very different yet, upon an appropriate choice of orthonormal basis, they share the same codebook* $\mathcal{C} = \{c_0, c_1\}$ *with* $c_0 = (\sqrt{\mathcal{E}}, 0)^\mathsf{T}$ *and* $c_1 = (0, \sqrt{\mathcal{E}})^\mathsf{T}$.

To see this, it suffices to verify that $\langle w_i, w_j \rangle$ *equals* $\mathcal{E}$ *if* $i = j$ *and equals* 0 *otherwise. Hence the two signals are orthogonal to each other and they have squared norm* $\mathcal{E}$*. Figure 3.5 shows the signals and the associated codewords.*



(a) $\mathcal{W}$ in the signal space. (b) $\mathcal{C}$ in $\mathbb{R}^2$.

Figure 3.5. $\mathcal{W}$ and $\mathcal{C}$.

Choice 1 (Rectangular pulse position modulation):

$$w_0(t) = \sqrt{\frac{\mathcal{E}}{T}}\, \mathbb{1}\{t \in [0, T]\}$$
$$w_1(t) = \sqrt{\frac{\mathcal{E}}{T}}\, \mathbb{1}\{t \in [T, 2T]\},$$

where we have used the indicator function $\mathbb{1}\{t \in [a, b]\}$ *to denote a rectangular pulse which is* 1 *in the interval* $[a, b]$ *and* 0 *elsewhere. Rectangular pulses can easily be generated, e.g. by a switch. They are used, for instance, to communicate a binary symbol within an electrical circuit. As we will see, in the frequency domain these pulses have side lobes that decay relatively slow, which is not desirable for high data rate over a channel for which bandwidth is at a premium.*

Choice 2 (Frequency-shift keying):

$$w_0(t) = \sqrt{\frac{2\mathcal{E}}{T}} \sin\left(\pi k \frac{t}{T}\right) \mathbb{1}\{t \in [0, T]\}$$
$$w_1(t) = \sqrt{\frac{2\mathcal{E}}{T}} \sin\left(\pi l \frac{t}{T}\right) \mathbb{1}\{t \in [0, T]\},$$

where k *and* l *are positive integers,* $k \neq l$*. With a large value of* k *and* l*, these signals could be used for wireless communication. To see that the two signals are orthogonal to each other we can use the trigonometric identity* $\sin(\alpha)\sin(\beta) = 0.5[\cos(\alpha - \beta) - \cos(\alpha + \beta)]$.

Choice 3 (Sinc pulse position modulation):

$$w_0(t) = \sqrt{\frac{\mathcal{E}}{T}}\, \operatorname{sinc}\left(\frac{t}{T}\right)$$
$$w_1(t) = \sqrt{\frac{\mathcal{E}}{T}}\, \operatorname{sinc}\left(\frac{t - T}{T}\right).$$

An advantage of sinc *pulses is that they have a finite support in the frequency domain. By taking their Fourier transform, we quickly see that they are orthogonal to each other. See Appendix 5.10 for details.*

Choice 4 (Spread spectrum):

$$w_0(t) = \sqrt{\mathcal{E}}\psi_1(t), \quad \text{with } \psi_1(t) = \sqrt{\frac{1}{T}} \sum_{j=1}^{n} s_{0,j} \mathbb{1}\left\{t - j\frac{T}{n} \in \left[0, \frac{T}{n}\right]\right\}$$

$$w_1(t) = \sqrt{\mathcal{E}}\psi_2(t), \quad \text{with } \psi_2(t) = \sqrt{\frac{1}{T}} \sum_{j=1}^{n} s_{1,j} \mathbb{1}\left\{t - j\frac{T}{n} \in \left[0, \frac{T}{n}\right]\right\},$$

where $(s_{0,1}, \ldots, s_{0,n}) \in \{\pm 1\}^n$ *and* $(s_{1,1}, \ldots, s_{1,n}) \in \{\pm 1\}^n$ *are orthogonal. This signaling method is called spread spectrum. It is not hard to show that it uses much bandwidth but it has an inherent robustness with respect to interfering (non-white and possibly non-Gaussian) signals.*

Now assume that we use one of the above choices to communicate across a continuous-time AWGN channel and that the receiver implements an ML decision rule. Since the codebook $\mathcal{C}$ *is the same in all cases, the decoder and the error probability will be identical no matter which choice we make.*

Computing the error probability is particularly easy when there are only two codewords. From the previous chapter we know that $P_e = Q\big(\frac{\|c_1 - c_0\|}{2\sigma}\big)$*, where* $\sigma^2 = \frac{N_0}{2}$*. The distance*

$$\|c_1 - c_0\| := \sqrt{\sum_{i=1}^{2} (c_{1,i} - c_{0,i})^2} = \sqrt{\mathcal{E} + \mathcal{E}} = \sqrt{2\mathcal{E}}$$

can also be computed as

$$\|w_1 - w_0\| := \sqrt{\int [w_1(t) - w_0(t)]^2 dt},$$

which requires neither an orthonormal basis nor the codebook. Yet another alternative is to use Pythagoras' theorem. As we know already that our signals have squared norm $\mathcal{E}$ *and are orthogonal to each other, their distance is* $\sqrt{\|w_0\|^2 + \|w_1\|^2} = \sqrt{2\mathcal{E}}$*. Inserting, we obtain*

$$P_e = Q\left(\sqrt{\frac{\mathcal{E}}{N_0}}\right).$$

□

EXAMPLE 3.8 (Single-shot PAM) *Let* $\psi(t)$ *be a unit-energy pulse. We speak of single-shot pulse amplitude modulation when the transmitted signal is of the form*

$$w_i(t) = c_i \psi(t),$$

where c_i *takes a value in some discrete subset of* $\mathbb{R}$ *of the form* $\{\pm a, \pm 3a, \pm 5a, \ldots, \pm(m-1)a\}$ *for some positive number* a*. An example for* $m = 6$ *is shown in Figure 2.9, where* $d = 2a$*.* □

EXAMPLE 3.9 (Single-shot PSK) *Let T and f_c be positive numbers and let m be a positive integer. We speak of single-shot phase-shift keying when the signal set consists of signals of the form*

$$w_i(t) = \sqrt{\frac{2\mathcal{E}}{T}} \cos\left(2\pi f_c t + \frac{2\pi}{m} i\right) \mathbb{1}\{t \in [0,T]\}, \quad i = 0, 1, \ldots, m-1. \tag{3.5}$$

For mathematical convenience, we assume that $2f_cT$ is an integer, so that $\|w_i\|^2 = \mathcal{E}$ for all i. (When $2f_cT$ is an integer, $w_i^2(t)$ has an integer number of periods in a length-T interval. This ensures that all $w_i(t)$ have the same norm, regardless of the initial phase. In practice, f_cT is very large, which implies that there are many periods in an interval of length T, in which case the energy difference due to an incomplete period is negligible.) The signal space representation can be obtained by using the trigonometric identity $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$ to rewrite (3.5) as

$$w_i(t) = c_{i,1}\psi_1(t) + c_{i,2}\psi_2(t),$$

where

$$c_{i,1} = \sqrt{\mathcal{E}} \cos\left(\frac{2\pi i}{m}\right), \quad \psi_1(t) = \sqrt{\frac{2}{T}} \cos(2\pi f_c t) \mathbb{1}\{t \in [0,T]\},$$
$$c_{i,2} = \sqrt{\mathcal{E}} \sin\left(\frac{2\pi i}{m}\right), \quad \psi_2(t) = -\sqrt{\frac{2}{T}} \sin(2\pi f_c t) \mathbb{1}\{t \in [0,T]\}.$$

The reader should verify that $\psi_1(t)$ and $\psi_2(t)$ are normalized functions and, because $2f_cT$ is an integer, they are orthogonal to each other. This can easily be verified using the trigonometric identity $\sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha+\beta) + \sin(\alpha-\beta)]$. Hence the codeword associated to $w_i(t)$ is

$$c_i = \sqrt{\mathcal{E}} \begin{pmatrix} \cos 2\pi i/m \\ \sin 2\pi i/m \end{pmatrix}.$$

In Example 2.15, we have already studied this constellation for the discrete-time AWGN channel. □

EXAMPLE 3.10 (Single-shot QAM) *Let T and f_c be positive numbers such that $2f_cT$ is an integer, let m be an even positive integer, and define*

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos(2\pi f_c t) \mathbb{1}\{t \in [0,T]\}$$
$$\psi_2(t) = \sqrt{\frac{2}{T}} \sin(2\pi f_c t) \mathbb{1}\{t \in [0,T]\}.$$

(We have already established in Example 3.9 that $\psi_1(t)$ and $\psi_2(t)$ are orthogonal to each other and have unit norm.) If the components of $c_i = (c_{i,1}, c_{i,2})^\mathsf{T}$, $i = 0, \ldots, m^2-1$, take values in some discrete subset of the form $\{\pm a, \pm 3a, \pm 5a, \ldots, \pm(m-1)a\}$ for some positive a, then

$$w_i(t) = c_{i,1}\psi_1(t) + c_{i,2}\psi_2(t),$$

is a single-shot quadrature amplitude modulation (QAM). The values of c_i for $m = 2$ and $m = 4$ are shown in Figures 2.10 and 2.22, respectively. □

The signaling methods discussed in this section are the building blocks of many communication systems.

3.5 Generalization and alternative receiver structures

It is interesting to explore a refinement and a variation of the receiver structure shown in Figure 3.4. We also generalize to an arbitrary message distribution. We take the opportunity to review what we have so far.

The source produces $H = i$ with probability $P_H(i)$, $i \in \mathcal{H}$. When $H = i$, the channel output is $R(t) = w_i(t) + N(t)$, where $w_i(t) \in \mathcal{W} = \{w_0(t), w_1(t), \ldots, w_{m-1}(t)\}$ is the signal constellation composed of finite-energy signals (known to the receiver) and $N(t)$ is white Gaussian noise. Throughout this text, we make the natural assumption that the vector space $\mathcal{V}$ spanned by $\mathcal{W}$ forms an inner product space (with the standard inner product). This is guaranteed if the zero signal is the only signal that has vanishing norm, which is always the case in real-world situations. Let $\{\psi_1(t), \ldots, \psi_n(t)\}$ be an orthonormal basis for $\mathcal{V}$. We can use the Gram–Schmidt procedure to find an orthonormal basis, but sometimes we can pick a more convenient one "by hand". At the receiver, we obtain a sufficient statistic by taking the inner product of the received signal $R(t)$ with each element of the orthonormal basis. The result is

$$\begin{aligned} Y &= (Y_1, Y_2, \ldots, Y_n)^{\mathsf{T}}, \text{ where} \\ Y_i &= \langle R, \psi_i \rangle, \quad i = 1, \ldots, n. \end{aligned}$$

We now face a hypothesis testing problem with prior $P_H(i)$, $i \in \mathcal{H}$, and observable Y distributed according to

$$f_{Y|H}(y|i) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|y - c_i\|^2}{2\sigma^2}\right),$$

where $\sigma^2 = \frac{N_0}{2}$. A MAP receiver that observes $Y = y$ decides $\hat{H} = i$ for one of the $i \in \mathcal{H}$ that maximize $P_H(i) f_{Y|H}(y|i)$ or any monotonic function thereof. Since $f_{Y|H}(y|i)$ is an exponential function of y, we simplify the test by taking the natural logarithm. We also remove terms that do not depend on i and rescale, keeping in mind that if we scale with a negative number, we have to change the maximization into minimization.

If we choose negative N_0 as the scaling factor we obtain the first of the following equivalent MAP tests.

(i) Choose $\hat{H}$ as one of the j that minimizes $\|y - c_j\|^2 - N_0 \ln P_H(j)$.

(ii) Choose $\hat{H}$ as one of the j that maximizes $\langle y, c_j \rangle - \frac{\|c_j\|^2}{2} + \frac{N_0}{2} \ln P_H(j)$.

(iii) Choose $\hat{H}$ as one of the j that maximizes $\int r(t) w_j^*(t) dt - \frac{\|w_j\|^2}{2} + \frac{N_0}{2} \ln P_H(j)$.

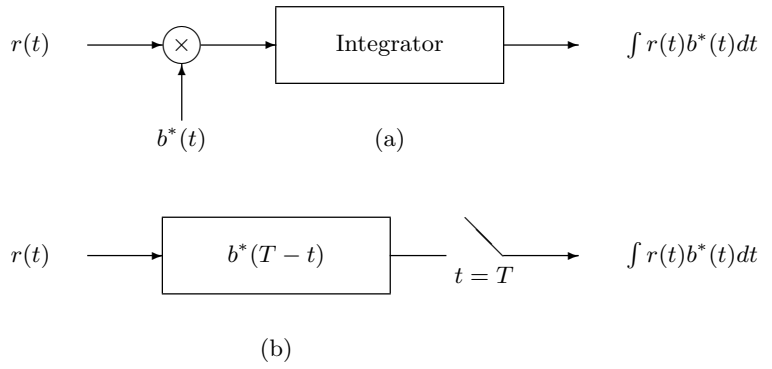


Figure 3.6. Two ways to implement $\int r(t)b^*(t)dt$, namely via a correlator (a) and via a matched filter (b) with the output sampled at time T.

The second is obtained from the first by using $\|y-c_i\|^2 = \|y\|^2 - 2\Re\{\langle y, c_i\rangle\} + \|c_i\|^2$. Once we drop the $\Re\{\cdot\}$ operator (the vectors are real-valued), remove the constant $\|y\|^2$, scale by $-1/2$, we obtain (ii).

Rules (ii) and (iii) are equivalent since $\int r(t)w_i^*(t)dt = \int r(t)\big(\sum_j c_{i,j}^*\psi_j^*(t)\big)dt = \sum_j y_j c_{i,j}^* = \langle y, c_i\rangle$.

The MAP rules (i)–(iii) require performing operations of the kind

$$\int r(t)b^*(t)dt, \tag{3.6}$$

where $b(t)$ is some function ($\psi_j(t)$ or $w_j(t)$). There are two ways to implement (3.6). The obvious way, shown in Figure 3.6a is by means of a so-called *correlator*. A correlator is a device that multiplies and integrates two input signals. The other way to implement (3.6) is via a so-called *matched filter*. This is a filter that takes $r(t)$ as the input and has $h(t) = b^*(T-t)$ as impulse response (Figure 3.6b), where T is an arbitrary design parameter selected in such a way as to make $h(t)$ a causal impulse response. The matched filter output $y(t)$ is then

$$\begin{aligned} y(t) &= \int r(\alpha)\, h(t-\alpha)\, d\alpha \\ &= \int r(\alpha)\, b^*(T+\alpha-t)\, d\alpha, \end{aligned}$$

and at $t = T$ it is

$$y(T) = \int r(\alpha)\, b^*(\alpha)\, d\alpha.$$

We see that the latter is indeed (3.6).

EXAMPLE 3.11 (Matched filter) *If $b(t)$ is as in Figure 3.7, then $y = \langle r(t), b(t)\rangle$ is the output at $t = 0$ of a linear time-invariant filter that has input $r(t)$ and has $h_0(t) = b(-t)$ as the impulse response (see the figure). The impulse response $h_0(t)$ is non-causal. We obtain the same result with a causal filter by delaying the*

impulse response by $3T$ and by sampling the output at $t = 3T$. The delayed impulse response is $h_{3T}(t)$, also shown in the figure. □

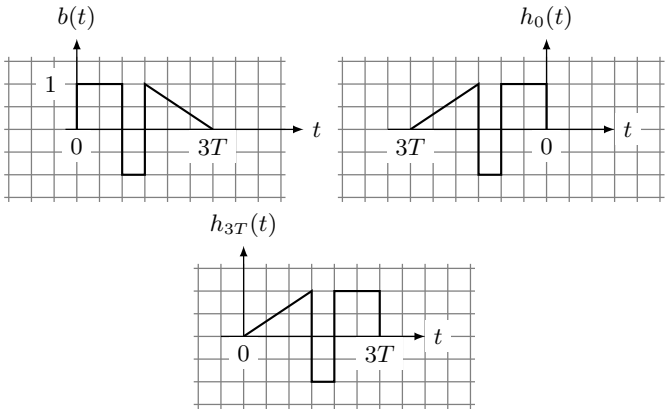


Figure 3.7.

It is instructive to plot the matched filter output as we do in the next example.

EXAMPLE 3.12 *Suppose that the signals are $w_0(t) = a\psi(t)$ and $w_1(t) = -a\psi(t)$, where a is some positive number and*

$$\psi(t) = \sqrt{\frac{1}{T}}\mathbb{1}\{0 \le t \le T\}.$$

The signals are plotted on the left of Figure 3.8. The n-tuple former consists of the matched filter of impulse response $h(t) = \psi^*(T - t) = \psi(t)$, with the output sampled at $t = T$. In the absence of noise, the matched filter output at the sampling time should be a when $w_0(t)$ is transmitted and $-a$ when $w_1(t)$ is transmitted. The

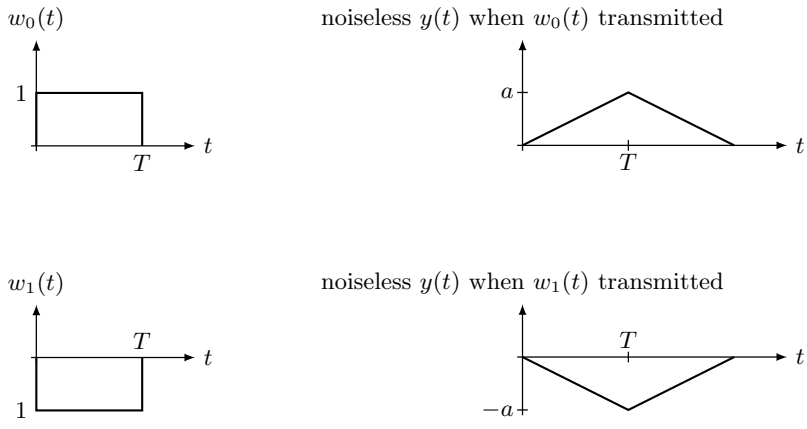
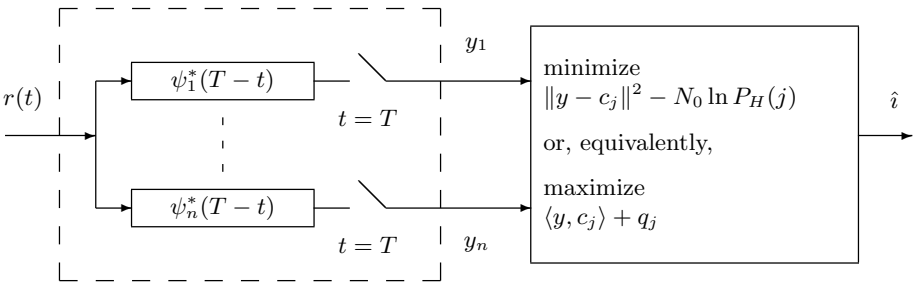


Figure 3.8. Matched filter response (right) to the input on the left.

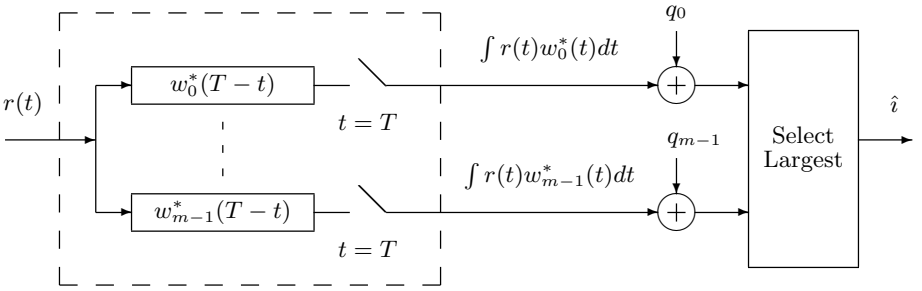
plots on the right of the figure show the matched filter response $y(t)$ to the input on the left. Indeed, at $t = T$ we have a or $-a$. At any other time we have b or $-b$, for some b such that $0 \leq b \leq a$. This, and the fact that the noise variance does not depend on the sampling time, implies that $t = T$ is the sampling time at which the error probability is minimized. □

Figure 3.9 shows the block diagrams for the implementation of the three MAP rules (i)–(iii). In each case the front end has been implemented by using matched filters, but correlators could also be used, as in Figure 3.4.

Whether we use matched filters or correlators depends on the technology and on the waveforms. Implementing a correlator in analog technology is costly. But, if the processing is done by a microprocessor that has enough computational power, then a correlation can be done at no additional hardware cost. We would be inclined to use matched filters if there were easy-to-implement filters of the desired impulse response. In Exercise 3.10 of this chapter, we give an example where the matched filters can be implemented with passive components.



n-Tuple Former.



Alternative to the *n*-Tuple Former.

Figure 3.9. Block diagrams of a MAP receiver for the waveform AWGN channel, with $y = (y_1, \ldots, y_n)^\mathsf{T}$ and $q_j = -\|w_j\|^2/2 + (N_0/2) \ln P_H(j)$. The dashed boxes can alternatively be implemented via correlators.

Notice that the bottom implementation of Figure 3.9 requires neither an orthonormal basis nor knowledge of the codebook, but it does require m as opposed to n matched filters (or correlators). We know that $m \geq n$, and often m is much larger than n. Notice also that this implementation does *not* quite fit into the decomposition of Figure 3.2. In fact the receiver bypasses the need for the n-tuple Y. As a byproduct, this proves that the receiver performance is not affected by the choice of an orthonormal basis.

In a typical communication system, n and m are very large. So large that it is not realistic to have n or m matched filters (or correlators). Even if we disregard the cost of the matched filters, the number of operations required by a decoder that performs a "brute force" search to find the distance-minimizing index (or inner-product-maximizing index) is typically exorbitant. We will see that clever design choices can dramatically reduce the complexity of a MAP receiver.

The equivalence of the two operations of Figure 3.6 is very important. It should be known by heart.

3.6 Continuous-time channels revisited

Every channel adds noise and this is what makes the communication problem both challenging and interesting. In fact, noise is the only reason there is a fundamental limitation to the maximal rate at which we can communicate reliably through a cable, an optical fiber, and most other channels of practical interest. Without noise we could transmit reliably as many bits per second as we want, using as little energy as desired, even in the presence of the other channel imperfections that we describe next.

Attenuation and amplification Whether wireline or wireless, a passive channel always attenuates the signal. For a wireless channel, the attenuation can be of several orders of magnitude. Much of the attenuation is compensated for by a cascade of amplifiers in the first stage of the receiver, but an amplifier scales both the information-carrying signal and the noise, and adds some noise of its own.

The fact that the receiver front end incorporates a cascade of amplifiers needs some explanation. Why should the signal be amplified if the noise is amplified by the same factor? A first answer to this question is that electronic devices, such as an n-tuple former, are designed to process electrical signals that are in a certain range of amplitudes. For instance, the signal's amplitude should be large compared to the noise added by the circuit. This explains why the first amplification stage is done by a so-called *low-noise amplifier*. If the receiving antenna is connected to the receiver via a relatively long cable, as it would be the case for an outdoor antenna, then the low-noise amplifier is typically placed between the antenna and the cable.

The low-noise amplifier (or the stage that follows it) contains a noise-reduction filter that removes the out-of-band noise. With perfect electronic circuits, such a filter is superfluous because the out-of-band noise is removed by the n-tuple former.

But the out-of-band noise increases the chance that the electronic circuits – up to and including the n-tuple former – saturate, i.e. that the amplitude of the noise exceeds the range that can be tolerated by the circuits.

The typical next stage is the so-called *automatic gain control* (AGC) amplifier, designed to bring the signal's amplitude into the desired range. Hence the AGC amplifier introduces a scaling factor that depends on the strength of the input signal.

For the rest of this text, we ignore hardware imperfections. Therefore, we can also ignore the presence of the low-noise amplifier, of the noise-reduction filter, and of the automatic gain control amplifier. If the channel scales the signal by a factor α, the receiver front end can compensate by scaling the received signal by α^{-1}, but the noise is also scaled by the same factor. This explains why, in evaluating the error probability associated to a signaling scheme, we often consider channel models that only add noise. In such cases, the scaling factor α^{-1} is implicitly accounted for in the noise parameter $N_0/2$. An example of how to determine $N_0/2$ is given in Appendix 3.11, where we work out a case study based on satellite communication.

Propagation delay and clock misalignment Propagation delay refers to the time it takes a signal to reach a receiver. If the signal set is $\mathcal{W} = \{w_0(t), w_1(t), \ldots, w_{m-1}(t)\}$ and the propagation delay is τ, then for the receiver it is as if the signal set were $\tilde{\mathcal{W}} = \{w_0(t-\tau), w_1(t-\tau), \ldots, w_{m-1}(t-\tau)\}$. The common assumption is that the receiver does not know τ when the communication starts. For instance, in wireless communication, a receiver has no way to know that the propagation delay has changed because the transmitter has moved while it was turned off. It is the responsibility of the receiver to adapt to the propagation delay. We come to the same conclusion when we consider the fact that the clocks of different devices are often not synchronized. If the clock of the receiver reads $t - \tau$ when that of the transmitter reads t then, once again, for the receiver, the signal set is $\tilde{\mathcal{W}}$ for some unknown τ. Accounting for the unknown τ at the receiver goes under the general name of *clock synchronization*. For reasons that will become clear, the clock synchronization problem decomposes into the symbol synchronization and into the phase synchronization problems, discussed in Sections 5.7 and 7.5. Until then and unless otherwise specified, we assume that there is no propagation delay and that all clocks are synchronized.

Filtering In wireless communication, owing to reflections and diffractions on obstacles, the electromagnetic signal emitted by the transmitter reaches the receiver via multiple paths. Each path has its own delay and attenuation. If $w_i(t)$ is transmitted, the receiver antenna output has the form $R(t) = \sum_{l=1}^{L} w_i(t-\tau_l)h_l$ plus noise, where τ_l and h_l are the delay and the attenuation along the lth path. Unlike a mirror, the rough surface of certain objects creates a large number of small reflections that are best accounted for by the integral form $R(t) = \int w_i(t-\tau)h(\tau)d\tau$ plus noise. This is the same as saying that the channel contains a filter of impulse response $h(t)$. For a different reason, the same channel model applies to wireline communication. In fact, due to dispersion, the channel output to a unit-energy pulse applied to the input at time $t = 0$ is some impulse

response $h(t)$. Owing to the channel linearity, the output due to $w_i(t)$ at the input is, once again, $R(t) = \int w_i(t-\tau)h(\tau)d\tau$ plus noise.

The possibilities we have to cope with the channel filtering depend on whether the channel impulse response is known to the receiver alone, to both the transmitter and the receiver, or to neither. It is often realistic to assume that the receiver can measure the channel impulse response. The receiver can then communicate it to the transmitter via the reversed communication link (if it exists). Hence it is hardly the case that only the transmitter knows the channel impulse response.

If the transmitter uses the signal set $\mathcal{W} = \{w_0(t), w_1(t), \ldots, w_{m-1}(t)\}$ and the receiver knows $h(t)$, from the receiver's point of view, the signal set is $\tilde{\mathcal{W}}$ with the ith signal being $\tilde{w}_i(t) = (w_i \star h)(t)$ and the channel just adds white Gaussian noise. This is the familiar case. Realistically, the receiver knows at best an estimate $\tilde{h}(t)$ of $h(t)$ and uses it as the actual channel impulse response.

The most challenging situation occurs when the receiver does not know and cannot estimate $h(t)$. This is a realistic assumption in bursty communication, when a burst is too short for the receiver to estimate $h(t)$ and the impulse response changes from one burst to the next.

The most favorable situation occurs when both the receiver and the transmitter know $h(t)$ or an estimate thereof. Typically it is the receiver that estimates the channel impulse response and communicates it to the transmitter. This requires two-way communication, which is typically available. In this case, the transmitter can adapt the signal constellation to the channel characteristic. Arguably, the best strategy is the so-called water-filling (see e.g. [19]) that can be implemented via orthogonal frequency division multiplexing (OFDM).

We have assumed that the channel impulse response characterizes the channel filtering for the duration of the transmission. If the transmitter and/or the receiver move, which is often the case in mobile communication, then the channel is still linear but *time-varying*. Excellent graduate-level textbooks that discuss this kind of channel are [2] and [17].

Colored Gaussian noise We can think of colored noise as filtered white noise. It is safe to assume that, over the frequency range of interest, i.e. the frequency range occupied by the information-carrying signals, there is no positive-length interval over which there is no noise. (A frequency interval with no noise is physically unjustifiable and, if we insist on such a channel model, we no longer have an interesting communication problem because we can transmit infinitely many bits error-free by signaling where there is no noise.) For this reason, we assume that the frequency response of the noise-shaping filter cannot vanish over a positive-length interval in the frequency range of interest. In this case, we can modify the aforementioned noise-reduction filter in such a way that, in the frequency range of interest, it has the inverse frequency response of the noise-shaping filter. The noise at the output of the modified noise-reduction filter, called *whitening filter*, is zero-mean, Gaussian, and white (in the frequency range of interest). The minimum error probability with the whitening filter cannot be higher than without, because the filter is invertible in the frequency range of interest. What we gain with the noise-whitening filter is that we are back to the familiar situation

where the noise is white and the signal set is $\tilde{\mathcal{W}} = \{\tilde{w}_0(t), \tilde{w}_1(t), \ldots, \tilde{w}_{m-1}(t)\}$, where $\tilde{w}_i(t) = (w_i \star h)(t)$ and $h(t)$ is the impulse response of the whitening filter.

3.7 Summary

In this chapter we have addressed the problem of communicating a message across a waveform AWGN channel. The importance of the continuous-time AWGN channel model comes from the fact that every conductor is a linear time-invariant system that smooths out and adds up the voltages created by the electron's motion. Owing to the central limit theorem, the result of adding up many contributions can be modeled as white Gaussian noise. No conductor can escape this phenomena, unless it is cooled to zero degrees kelvin. Hence every channel adds Gaussian noise. This does not imply that the continuous-time AWGN channel is the only channel model of interest. Depending on the situation, there can be other impairments such as fading, nonlinearities, and interference, that should be considered in the channel model, but they are outside the scope of this text.

As in the previous chapter, we have focused primarily on the receiver that minimizes the error probability assuming that the signal set is given to us. We were able to move forwards swiftly by identifying a sufficient statistic that reduces the receiver design problem to the one studied in Chapter 2. The receiver consists of an n-tuple former and a decoder. We have seen that the sender can also be decomposed into an encoder and a waveform former. This decomposition naturally fits the layering philosophy discussed in the introductory chapter: The waveform former at the sender and the n-tuple former at the receiver can be seen as providing a "service" to the encoder–decoder pair. The service consists in making the continuous-time AWGN channel look like a discrete-time AWGN channel.

Having established the link between the continuous-time and the discrete-time AWGN channel, we are in the position to evaluate the error probability of a communication system for the AWGN channel by means of simulation. An example is given in Appendix 3.8.

How do we proceed from here? First, we need to introduce the performance parameters we care mostly about, discuss how they relate to one another, and understand what options we have to control them. We start this discussion in the next chapter where we also develop some intuition about the kind of signals we want to use to transmit many bits.

Second, we need to start paying attention to cost and complexity because they can quickly get out of hand. For a brute-force implementation, the n-tuple former requires n correlators or matched filters and the decoder needs to compute and compare $\langle y, c_j \rangle + q_j$ for m codewords. With $k = 100$ (a very modest number of transmitted bits) and $n = 2k$ (a realistic relationship), the brute-force approach requires 200 matched filters or correlators and the decoder needs to evaluate roughly 10^{30} inner products. These are staggering numbers. In Chapter 5 we will learn how to choose the waveform former in such a way that the n-tuple former can be implemented with a single matched filter. In Chapter 6 we will see that

there are encoders for which the decoder needs to explore a number of possibilities that grows linearly rather than exponentially in k.

3.8 Appendix: A simple simulation

Here we give an example of a basic simulation. Instead of sending a continuous-time waveform $w(t)$, we send the corresponding codeword c; instead of adding a sample path of white Gaussian noise of power spectral density $N_0/2$, we add a realization z of a Gaussian random vector that consists of iid components that are zero mean and of variance $\sigma^2 = N_0/2$. The decoder observes $y = c + z$. MATLAB is a programming language that makes it possible to implement a simulation in a few lines of code. Here is how we can determine (by simulation) the error probability of m-PAM for $m = 6$, $d = 2$, and $\sigma^2 = 1$.

```
% define the parameters
m = 6 % alphabet size (positive even number)
d = 2 % distance between points
noiseVariance = 1
k = 1000 % number of transmitted symbols

% define the encoding function
encodingFunction = -(m-1)*d/2:d:(m-1)*d/2;

% generate the message
message = randi(m,k,1);

% encode
c = encodingFunction(message);

% generate the noise
z = normrnd(0,sqrt(noiseVariance),1,k);

% add the noise
y = c+z;

% decode
[distances,message_estimate] = min(abs(repmat(y',1,m)...
    -repmat(encodingFunction,k,1)),[],2);

% determine the symbol error probability and print
errorRate = symerr(message,message_estimate)/k
```

The above MATLAB code produces the following output (reformatted)

```
m = 6
d = 2
```

```
noiseVariance = 1
k = 1000
errorRate = 0.2660
```

3.9 Appendix: Dirac-delta-based definition of white Gaussian noise

It is common to define white Gaussian noise as a zero-mean WSS Gaussian random process $N(t)$ of autocovariance $K_N(\tau) = \frac{N_0}{2}\delta(\tau)$.

From the outset, the difference between this and the approach we chose (Section 3.2) lies on where we start with a mathematical model of the physical world. We chose to start with the measurements that the receiver can make about $N(t)$, whereas the standard approach starts with $N(t)$ itself.

To model and use $N(t)$ in a rigorous way requires familiarity with the notion of stochastic processes (typically not a problem), the ability to manipulate the Dirac delta (not a problem until something goes wrong), and measure theory to guarantee that integrals such as $\int N(\alpha)g(\alpha)d\alpha$ are well-defined. Most engineers are not familiar with measure theory. This results in situations that are undesirable for the instructor and for the student. Nevertheless it is important that the reader be aware of the standard procedure, which is the reason for this appendix.

As the following example shows, it is a simple exercise to derive (3.3) from the above definition of $N(t)$. (We take it for granted that the integrals exist.)

EXAMPLE 3.13 *Let $g_1(t)$ and $g_2(t)$ be two finite-energy pulses and for $i = 1, 2$, define*

$$Z_i = \int N(\alpha)g_i(\alpha)d\alpha, \tag{3.7}$$

where $N(t)$ is white Gaussian noise as we just defined. We compute the covariance $\mathrm{cov}\big(Z_i, Z_j\big)$ as follows:

$$\begin{aligned}
\mathrm{cov}(Z_i, Z_j) &= \mathbb{E}\left[Z_i Z_j^*\right] \\
&= \mathbb{E}\left[\int N(\alpha)g_i(\alpha)d\alpha \int N^*(\beta)g_j^*(\beta)d\beta\right] \\
&= \int\int \mathbb{E}\left[N(\alpha)N^*(\beta)\right] g_i(\alpha)g_j^*(\beta)d\alpha d\beta \\
&= \int\int \frac{N_0}{2}\delta(\alpha-\beta)g_i(\alpha)g_j^*(\beta)d\alpha d\beta \\
&= \frac{N_0}{2}\int g_i(\beta)g_j^*(\beta)d\beta.
\end{aligned}$$

□

EXAMPLE 3.14 *Let $N(t)$ be white Gaussian noise at the input of a linear time-invariant circuit of impulse response $h(t)$ and let $Z(t)$ be the filter's output. Compute the autocovariance of the output $Z(t) = \int N(\alpha)h(t-\alpha)d\alpha$.*

Solution: The definition of autocovariance is $K_Z(\tau) := \mathbb{E}[Z(t+\tau)Z^*(t)]$. We proceed two ways. The computation using the definition of $N(t)$ given in this appendix mimics the derivation in Example 3.13. The result is $K_Z(\tau) = \frac{N_0}{2}\int h(t+\tau)h^*(t)dt$. If we use the definition of white Gaussian noise given in Section 3.2, we do not need to calculate (but we do need to know (3.3), which is part of the definition). In fact, the Z_i and Z_j defined in (3.2) and used in (3.3) become $Z(t+\tau)$ and $Z(t)$ if we set $g_i(\alpha) = h(t+\tau-\alpha)$ and $g_j(\alpha) = h(t-\alpha)$, respectively. Hence we can read the result directly out of 3.3, namely

$$K_Z(\tau) = \frac{N_0}{2}\int h(t+\tau-\alpha)h^*(t-\alpha)dt = \frac{N_0}{2}\int h(\beta+\tau)h^*(\beta)d\beta.$$

By defining the self-similarity function[1] *of $h(t)$*

$$R_h(\tau) = \int h(t+\tau)h^*(t)dt$$

we can summarize as follows

$$K_Z(\tau) = \frac{N_0}{2}R_h(\tau). \qquad \square$$

The definition of $N(t)$ given in this appendix is somewhat unsatisfactory also based on physical grounds. Recall that the Fourier transform of the autocovariance $K_N(\tau)$ is the *power spectral density* $S_N(f)$ (also called *power spectrum*). If $K_N(\tau) = \frac{N_0}{2}\delta(\tau)$ then $S_N(f) = \frac{N_0}{2}$, i.e. a constant. Integrating over the power spectral density yields the power, which in this case is infinite. The noise of a physical system cannot have infinite power. A related problem shows up from a different angle when we try to determine the variance of a sample $N(t_0)$ for an arbitrary time t_0. This is the autocovariance $K_N(\tau)$ evaluated at $\tau = 0$, but a Dirac delta is not defined as a stand-alone function.[2] Since we think of a Dirac delta as a very narrow and very tall function of unit area, we could argue that $\delta(0) = \infty$. This is also unsatisfactory because we would rather avoid having to define Gaussian random variables of infinite variance. More precisely, a stochastic process is characterized by specifying the joint distribution of each finite collection of samples, which implies that we would have to define the density of any collection of Gaussian random variables of *infinite* variance. Furthermore, we know that a random variable of infinite variance is not a good model for what we obtain when we sample noise. The reason the physically-unsustainable Dirac-delta-based model leads to physically meaningful results is that we use it only to describe filtered

[1] Also called the autocorrelation function. We reserve the term autocorrelation function for stochastic processes and use self-similarity function for deterministic pulses.

[2] Recall that a Dirac delta function is defined through what happens when we integrate it against a function, i.e. through the relationship $\int \delta(t)g(t) = g(0)$.

white Gaussian noise. (But then, why not bypass the mathematical description of $N(t)$ as we do in Section 3.2?)

As a final remark, note that defining an object indirectly through its behavior, as we have done in Section 3.2, is not new to us. We do something similar when we introduce the Dirac delta function by saying that it fulfills the relationship $\int f(t)\delta(t) = f(0)$. In both cases, we introduce the object of interest by saying how it behaves when integrated against a generic function.

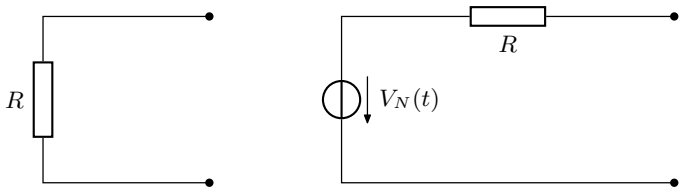
3.10 Appendix: Thermal noise

Any conductor at non-zero temperature produces thermal (Johnson) noise. The motion of charges (electrons) that move inside a conductor yields many tiny electrical fields, the sum of which can be measured as a voltage at the conductor's terminals. Owing to the central limit theorem, the aggregate voltage can be modeled as white Gaussian noise. (It looks white, up to very high frequencies.)

Thermal noise was first measured by Johnson (Bell Labs, 1926) who made the following experiment. He took a number of different conducting substances, such as solutions of salt in water, copper sulfate, etc., and measured the intrinsic voltage fluctuations across these substances. He found that the thermal noise expresses itself as a voltage source $V_N(t)$ in series with the noise-free conductor (Figure 3.10). The mean square voltage of $V_N(t)$ per hertz (Hz) of bandwidth (accounting only for positive frequencies) equals $4Rk_BT$, where $k_B = 1.381 \times 10^{-23}$ is Bolzmann's constant in joules/kelvin, T is the absolute temperature of the substance in kelvin (290 K at room temperature), and R its resistance in ohms.

Johnson described his findings to Nyquist (also at Bell Labs) who was able to explain the results by using thermodynamics and statistical mechanics. (Nyquist's paper [25] is only four pages and very accessible. A recommended reading.) The expression for the mean of $V_N^2(t)$ per Hz of bandwidth derived by Nyquist is

$$\frac{4Rhf}{e^{\frac{hf}{k_BT}} - 1}, \tag{3.8}$$



(a) Noisy conductor. (b) Equivalent electrical circuit.

Figure 3.10. (a) Conductor of resistance R; (b) equivalent electrical circuit, where $V_N(t)$ is a voltage source modeled as white Gaussian noise of (single-sided) power spectral density $N_0 = 4k_BTR$ and R is an ideal (noise-free) resistor.

where $h = 6.626 \times 10^{-34}$ joules×seconds is Planck's constant. This expression also holds for the mean square voltage at the terminals of an impedance Z with $\Re\{Z\} = R$.

For small values of x, $e^x - 1$ is approximately x. Hence, as long as hf is much smaller than k_BT, the denominator of Nyquist's expression is approximately $\frac{hf}{k_BT}$, in which case (3.8) simplifies to

$$4Rk_BT,$$

in exact agreement with Johnson's measurements.

EXAMPLE 3.15 *At room temperature ($T = 290$ kelvin), k_BT is about $4 \cdot 10^{-21}$. At 600 GHz, hf is about $4 \cdot 10^{-22}$. Hence, for applications in the frequency range from 0 to 600 GHz, we can pretend that $V_N(t)$ has a constant power spectral density.* □

EXAMPLE 3.16 *Consider a resistor of 50 ohms at $T = 290$ kelvin. The mean square voltage per Hz of bandwidth due to thermal noise is $4k_BTR = 4 \times 1.381 \times 10^{-23} \times 290 \times 50 = 8 \times 10^{-19}$ volts^2/Hz.* □

It is a straightforward exercise to check that the power per Hz of (single-sided) bandwidth that the voltage source of Figure 3.10b dissipates into a load of matching impedance is k_BT watts. Because this is a very small number, it is convenient to describe it by means of its temperature T.

Even other noises, such as the noise produced by an amplifier or the one picked up by an antenna, are often characterized by a "noise temperature", defined as the number T that makes k_BT equal to the spectral density of the noise injected into the receiver. (See Appendix 3.11.)

3.11 Appendix: Channel modeling, a case study

Once the signal set $\mathcal{W}$ is fixed (up to a scaling factor), the error probability of a MAP decoder for the AWGN channel depends only on the signal energy divided by the noise-power density (signal-to-noise ratio in short) $\mathcal{E}/N_0$ at the input of the n-tuple former. How do we determine $\mathcal{E}$ in terms of the design parameters that we can measure, such as the power P_T of the transmitter, the transmitting antenna gain G_T, the distance d between the transmitter and the receiver, the receiving antenna gain G_R? And how do we find the value for N_0? In this appendix, we work out a case study based on satellite communication.

Consider a transmitting antenna that radiates isotropically in free space at a power level of P_T watts. Imagine a sphere of radius d meters centered at the transmitting antenna. The surface of this sphere has an area of $4\pi d^2$, thus the power density at distance d is $\frac{P_T}{4\pi d^2}$ watts/m^2.

Satellites and the corresponding Earth stations use antennas that have directivity (typically a parabolic or a horn antenna for a satellite, and a parabolic antenna for an Earth station). Their directivity is specified by their gain G in the pointing direction. If the transmitting antenna has gain G_T, the power density in the pointing direction at distance d is $\frac{P_T G_T}{4\pi d^2}$ watts/m^2.

A receiving antenna at distance d gathers a portion of the transmitted power that is proportional to the antenna's effective area A_R. If we assume that the transmitting antenna is pointed in the direction of the receiving antenna, the received power is $P_R = \frac{P_T G_T A_R}{4\pi d^2}$.

Like the transmitting antenna, the receiving antenna can be described by its gain G_R. For a given effective area A_R, the gain is inversely proportional to λ^2, where λ is the wavelength. (As the bandwidth of the transmitted signal is small compared to the carrier frequency, we can use the carrier frequency wavelength.) Notice that this relationship between area, gain, and wavelength is rather intuitive. A familiar case is that of a flashlight. Owing to the small wavelength of light, a flashlight can create a focused beam even with a relatively small parabolic reflector. As we know from experience, the bigger the flashlight reflector, the narrower the beam. The precise relationship is $G_R = \frac{4\pi A_R}{\lambda^2}$. (Thanks to the ratio $\frac{A_R}{\lambda^2}$, the gain G_R is dimension-free.) Solving for A_R and plugging into P_R yields

$$P_R = \frac{P_T G_T G_R}{(4\pi d/\lambda)^2}. \tag{3.9}$$

The factor $L_S = (4\pi d/\lambda)^2$ is commonly called the *free-space path loss*, but this is a misnomer. In fact the free-space attenuation is independent of the wavelength. It is the relationship between the antenna's effective area and its gain that brings in the factor λ^2. Nevertheless, being able to write

$$P_R = P_T \frac{G_T G_R}{L_S} \tag{3.10}$$

has the advantage of underlining the "gains" and the "losses". Notice also that L_S is a factor on which the system designer has little control (for a geostationary satellite the distance is fixed and the carrier frequency is often dictated by regulations), whereas P_T, G_T, and G_R are parameters that a designer might be able to choose (within limits).

Now suppose that the receiving antenna is connected to the receiver via a lossless coaxial cable. The antenna and the receiver input have an impedance and the connecting cable has a characteristic impedance. For best power transfer, the three impedances should be resistive and have the same value, typically 50 ohms (see, e.g., Wikipedia, impedance matching). We assume that it is indeed the case and let R ohms be its value. Then, the impedance seen by the antenna looking into the cable is also R as if the receiver were connected directly to the antenna (see, e.g., Wikipedia, transmission line, or [14]). Figure 3.11 shows the electrical model for the receiving antenna and its load.[3] It shows the voltage source $W(t)$ that represents the intended signal, the voltage source $V_N(t)$ that represents all noise sources, the antenna impedance R and the antenna's load R.

[3] The circuit of Figure 3.11 is a suitable model for determining the voltage (and the current) at the receiver input (the load in the figure). There is a more complete model [26] that enables us to associate the power dissipated by the antenna's internal impedance with the power that the antenna radiates back to space.

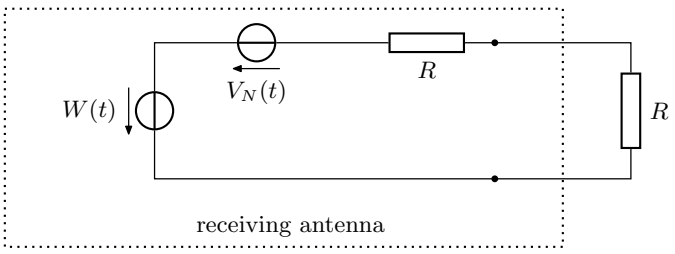


Figure 3.11. Electrical model for the receiving antenna and the load it sees looking into the first amplifier.

The advantage of having all the noise sources be represented by a single source which is co-located with the signal source $W(t)$ is that the signal-to-noise ratio at that point is the same as the signal-to-noise-ratio at the input of the n-tuple former. (Once all noise sources are accounted for at the input, the electronic circuits are considered as noise free). So, the $\mathcal{E}/N_0$ of interest to us is the signal energy absorbed by the load divided by the noise-power density absorbed by the same load.

The power harvested by the antenna is passed onto the load. This power is P_R, hence the energy is $P_R\tau$, where τ is the duration of the signals (assumed to be the same for all signals).

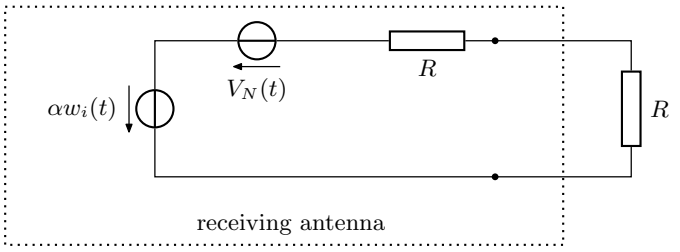
As mentioned in Appendix 3.10, it is customary to describe the noise-power density by the temperature T_N of a fictitious resistor that transfers the same noise-power density to the same load. This density is k_BT_N. If we know (for instance from measurements) the power density of each noise source, we can determine the equivalent density at the receiver input, sum all the densities, and divide by k_B to obtain the noise temperature T_N. Here we assume that this number is provided to us by the manufacturer of the receiver (see Example 3.17 for a numerical value). Putting things together, we obtain

$$\mathcal{E}/N_0 = \frac{P_R\tau}{k_BT_N} = \frac{P_T\tau G_TG_R}{L_Sk_BT_N}. \tag{3.11}$$

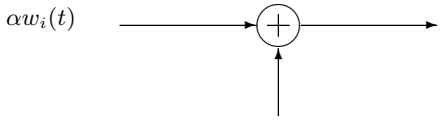
To go one step further, we characterize the two voltage sources of Figure 3.11. This is a calculation that the hardware designer might want to do to determine the range of voltages and currents at the antenna output.

Recall that a voltage of v volts applied to a resistor of R ohms dissipates the power $P = v^2/R$ watts. When $H = i$, $W(t) = \alpha w_i(t)$ for some scaling factor α. We determine α by computing the resulting average power dissipated by the load and by equating to P_R. Thus $P_R = \frac{\alpha^2\mathcal{E}}{4R\tau}$. Inserting the value of P_R and solving for α yields

$$\alpha = \sqrt{\frac{4RP_TG_TG_R}{L_S\mathcal{E}/\tau}}.$$



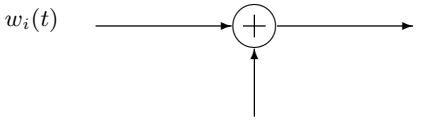
(a) Electrical circuit.



$V_N(t)$

$N_0/2 = 2Rk_BT_N$

(b) System-engineering viewpoint.



$N(t)$

$N_0/2 = \frac{k_BT_NL_S\mathcal{E}}{2P_T\tau G_TG_R}$

(c) Preferred channel model.

Figure 3.12. Various viewpoints under hypothesis $H = i$.

Hence, when $H = i$, the received signal (before noise) is

$$W(t) = \alpha w_i(t) = \sqrt{\frac{4RP_TG_TG_R}{L_S\mathcal{E}/\tau}}w_i(t).$$

Figure 3.12a summarizes the equivalent electrical circuit under the hypothesis $H = i$. As determined in Appendix 3.10, the mean square voltage of the noise source $V_N(t)$ per Hz of (single-sided) bandwidth is $N_0 = 4Rk_BT_N$. Figure 3.12b is the equivalent representation from the point of view of a system designer. The usefulness of these models is that they give us actual voltages. As long as we are not concerned with hardware limitations, for the purpose of the channel model, we are allowed to scale the signal and the noise by the same factor. Specifically, if we divide the signal by α and divide the noise-power density by α^2, we obtain the channel model of Figure 3.12c. Observe that the impedance R has fallen out of the picture.

As a "sanity check", if we compute $\mathcal{E}/N_0$ using Figure 3.12c we obtain $\frac{\tau P_T G_T G_R}{L_S k_B T_N}$, which corresponds to (3.11). The following example gives numerical values.

EXAMPLE 3.17 *The following parameters pertain to Mariner-10, an American robotic space probe launched by NASA in* 1973 *to fly to the planets Mercury and Venus.*

$P_T = 16.8$ *watts (12.25 dBW).*
$\lambda = 0.13$ *m (carrier frequency at* 2.3 *GHz).*
$G_T = 575.44$ *(*27.6 *dB).*
$G_R = 1.38 \times 10^6$ *(*61.4 *dB).*
$d = 1.6 \times 10^{11}$ *meters.*
$T_N = 13.5$ *kelvin.*
$R_b = 117.6$ *kbps.*

The bit rate $R_b = 117.6$ kbps (kilobits per second) is the maximum data rate at which the space probe could transmit information. This can be achieved via antipodal signals of duration $\tau = 1/R_b = 8.5 \times 10^{-6}$ seconds. Under this assumption, plugging into (3.11) yields $\mathcal{E}/N_0 = 2.54$. The error rate for antipodal signaling is

$$P_e = Q\left(\sqrt{\frac{2\mathcal{E}_s}{N_0}}\right) = 0.0120.$$

We see that the error rate is fairly high, but by means of coding techniques (Chapter 6), it is possible to achieve reliable communication at the expense of some reduction in the bit rate. □

3.12 Exercises

Exercises for Section 3.1

EXERCISE 3.1 (Gram–Schmidt procedure on tuples) *By means of the Gram–Schmidt orthonormalization procedure, find an orthonormal basis for the subspace spanned by the four vectors $\beta_1 = (1,0,1,1)^\mathsf{T}$, $\beta_2 = (2,1,0,1)^\mathsf{T}$, $\beta_3 = (1,0,1,-2)^\mathsf{T}$, and $\beta_4 = (2,0,2,-1)^\mathsf{T}$.*

EXERCISE 3.2 (Gram–Schmidt procedure on two waveforms) *Use the Gram–Schmidt procedure to find an orthonormal basis for the vector space spanned by the functions shown in Figure 3.13.*

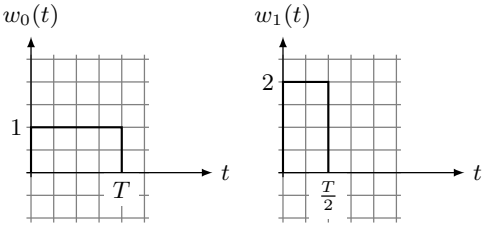


Figure 3.13.