BIBLIOGRAPHY

DEMODULATORS

This article introduces demodulators, physical devices situated in the receivers of communication systems. Demodulators undo (as best they can) the modulation performed at the transmitter as well as the impairments introduced by the channel. To help understand demodulators, this introduction presents the modulator, channel, and demodulator components of a communication system, as illustrated in Fig. 1.

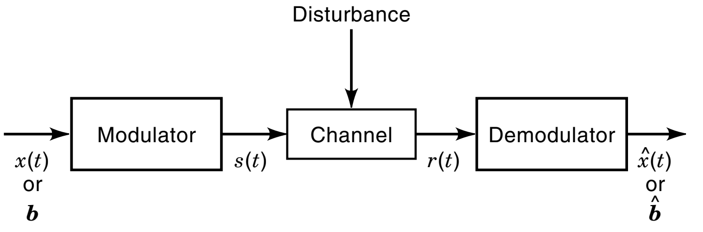

Figure 1. A communication system.

Modulators

Modulators are physical devices that map an information-bearing signal (either digital or analog) into a waveform ready for transmission over the channel, that is, into a waveform compatible with the channel characteristics. Some of the many uses of modulators include (1) radio transmission, where modulators ensure that the electromagnetic (EM) wave of the appropriate frequency is generated, (2) sharing a single channel, where modulators are used to separate signals so that they share the channel (e.g., frequency division multiplexing), and (3) placing a signal at frequencies where system design requirements are easily met.

A modulator maps the information-bearing signal to the channel-matching waveform as follows. The modulator codes the information in the phase, frequency, and/or amplitude of a sinusoid, called a carrier. Specifically, when the information is digital and hence representable by bits (binary digits), the modulator generally takes each block of $k = \log_2 M$ bits and maps this information into phase, frequency, and/or amplitude. For analog information, the modulator generally maps the analog information directly into one of these three carrier characteristics.

Channel

In this article, the channel refers to the propagation medium connecting the modulator to the demodulator. Typical communication channels include the atmosphere, free space, and physical media, such as twisted pair cables, coaxial cables, and fiber-optic cables. We also consider devices, such as antennas, lasers, and photodetectors, to be part of the channel.

Channels introduce a number of impairments that can significantly corrupt a modulated signal. Perhaps the best-known channel impairment is noise, an unwanted signal superimposed by the channel on the transmitted signal. Noise is modeled statistically and is most often modeled as additive white Gaussian noise (AWGN). Another channel impairment is signal attenuation, the reduction in energy of the transmitted signal caused, for example, by the absorption of signal energy by the atmosphere or the attenuation introduced by a twisted pair cable. Additionally, channels introduce intersymbol interference, the smearing of one transmitted symbol into its neighboring transmitted symbols, as a result of finite bandwidth and other filtering effects of the channel. Some channels also introduce fading, a peaking and nulling (in time) of the transmitted amplitude; this appears mainly in mobile communication systems, and results because the mobile receiver picks up many reflected versions of the originally transmitted signal.

Some other impairments that we associate with the channel include phase noise and timing noise. The local oscillator in the modulator and that of the demodulator cannot be matched exactly; this results in phase noise (also called phase jitter), which can cause detector degradation and hence signal loss. Additionally, the bit or symbol clock at the modulator can never exactly match that of the demodulator; this results in timing noise, which interferes with optimal detector performance and causes information loss.

Demodulator

The demodulator is the component of the communication system, located at the receiver side, that maps the received signal [$r(t)$ in Fig. 1: the modulated signal corrupted by the channel] into an information-bearing signal [$\hat{x}(t)$ or $\hat{\boldsymbol{b}}$ in Fig. 1]. It can be thought of as undoing the effects of both the modulator and the channel. The ultimate goal of a demodulator is to create an information-bearing signal [$\hat{x}(t)$ or $\hat{\boldsymbol{b}}$] that is as close as possible to the information-bearing signal input at the modulator [$x(t)$ or $\boldsymbol{b}$]. When the information-bearing signal is digital, closeness of information (i.e., closeness of $\boldsymbol{b}$ and $\hat{\boldsymbol{b}}$) is measured by the probability of error, that is, the probability that a binary digit output by the demodulator does not match the corresponding bit input at the modulator. When the information-bearing signal is analog, the proximity of information signals $\hat{x}(t)$ and $x(t)$ is measured by the signal-to-noise ratio, that is, the ratio of information-bearing signal power to noise power at the demodulator output.

Demodulators often get help in their task of information recovery from other devices in the receiver. Receivers typically use tracking loops to minimize the effects of phase noise and timing noise; such loops can reduce these noise effects to the point of negligibility. Furthermore, when channels introduce intersymbol interference, equalizers are commonly introduced to undo this channel corruption. Oftentimes, the task of the demodulator is reduced to extracting the information-bearing signal from the modulated signal in the presence of channel noise (typically AWGN).

Demodulators can be classified into a number of different categories. First, they can be classified according to the type of information signal output by the demodulator. If this information signal is digital (i.e., $\hat{\boldsymbol{b}}$), the demodulator is called a digital data demodulator (because it extracts digital information). Correspondingly, if the information is analog [i.e., $\hat{x}(t)$], the demodulator is called an analog data demodulator.

Second, a demodulator can be classified by modulation type. If the modulator maps the information into the frequency of the carrier, then the role of the demodulator is to extract information from the frequency (in the presence of channel impairments). Similarly, if the modulator maps information into carrier amplitude/phase, then the role of the demodulator is to extract information from the carrier amplitude/phase. Typically, demodulators are labeled to match the modulator, for example, a demodulator is called an amplitude modulation (AM) demodulator whenever the modulator is an AM modulator. In summary, demodulators are classified in accordance with the carrier parameter from which they extract information.

Third, demodulators are classified as coherent, noncoherent, or differentially coherent. Coherent demodulators are demodulators that require accurate knowledge of the carrier phase to achieve reliable information recovery. Noncoherent demodulators are demodulators that can detect information reliably without carrier phase information. Coherent demodulators typically require phase-tracking circuitry to estimate the carrier phase accurately, whereas noncoherent demodulators do not. Consequently, receivers with coherent demodulators are generally more complex. However, they also detect the information-bearing signal more reliably. A final category is differentially coherent detection, in which the carrier phase of the previous symbol(s) creates the carrier phase information for the current symbol.

Finally, digital data demodulators can be classified as either burst mode or continuous mode. Demodulators that operate in burst mode detect information in blocks of N symbols at a time, where typical values of N range between 100 and 500. Demodulators that operate in continuous mode detect symbols in an ongoing, continuous fashion.

DIGITAL DATA DEMODULATORS

This section introduces the most commonly implemented digital data demodulator. This demodulator detects digital data in the presence of channel noise, specifically AWGN. All other channel impairments are assumed to be accounted for by other receiver components before demodulation.

The demodulator presented is classified as coherent and operates in continuous mode. Understanding the workings and derivation of this demodulator will allow the reader to generate other optimal demodulator structures for detection in the presence of other channel impairments (as discussed in the section entitled "Advanced Issues").

Overview

The digital data demodulator for data detection in the presence of AWGN has a general implementation corresponding to Fig. 2. Here, the demodulator receives the channel output $r(t)$, which is assumed to correspond to the modulator output $s(t)$ in the presence of channel noise $n(t)$ (assumed AWGN). That is,

$$r(t) = s(t) + n(t) \tag{1}$$

The demodulator implementation is best explained by a division into two essential components. The component to the left of the dashed divider line, usually called the receiver front end, maps $r(t)$ into the vector $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)$. This vector provides sufficient information about $r(t)$ to allow the rest of the demodulator to carry out optimal data detection. The vector $\boldsymbol{r}$ is known as a sufficient statistic for detection. The remaining demodulator component (to the right of the divider), called the data detector, maps the vector $\boldsymbol{r}$ into a best guess at the transmitted information. This best guess ensures the smallest likelihood of error between transmitted and received bits.

The set $\{\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)\}$ in the demodulator represents an orthonormal basis for the set of possible modulator signals (as detailed in the next subsection). Typically, this set of signals corresponds to a cosine and a sine waveform, namely $\{\cos \omega_c t, \sin \omega_c t\}$, which are matched in frequency and phase to the carrier waveforms at the modulator. Additionally, the integrator component in the demodulator integrates over the exact duration of one symbol (also detailed next). Hence, to ensure optimal working of this demodulator, it must be provided with accurate frequency, phase, and timing information.

Derivation of the Demodulator Structure

In this section, the optimal receiver structure, introduced in the overview, is derived by mathematical and statistical arguments. This derivation is subdivided into a number of subsections to simplify its presentation. First, the modulation is described in some detail. This is followed by a derivation of the receiver front end, and then the data detector is derived. Later subsections present simplified versions of the digital data demodulator for a number of important practical modulations.

Modulator Details. Before explaining the digital data demodulator of Fig. 2, it is important first to understand digital data modulators. A digital data modulator maps each block of $k = \log_2 M$ binary digits (bits) into the phase, frequency, or amplitude of a carrier. More generally, the modulator can be described as mapping each block of k binary digits into one of $M = 2^k$ deterministic, finite-energy waveforms, $\{s_1(t), s_2(t),$

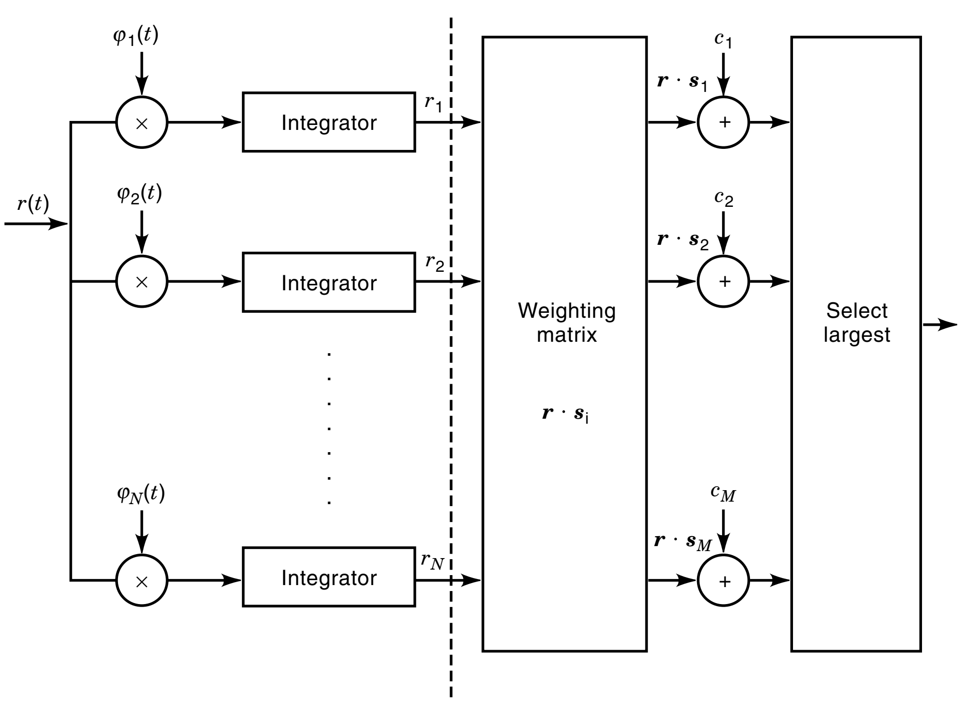

Figure 2. Digital data demodulator for detection in AWGN.

. . ., $s_M(t)$}. This is best explained by example, and hence we briefly introduce three examples of great practical interest, namely, amplitude-shift keying (ASK), phase-shift keying (PSK), and frequency-shift keying (FSK).

ASK refers to the mapping of binary information into the amplitude of the carrier waveform. Specifically, ASK refers to the mapping of each block of k bits into one of the $M = 2^k$ signals $\{s_1(t), s_2(t), \ldots, s_M(t)\}$, where

$$s_i(t) = A_i \cos(\omega_c t + \theta), \qquad jT \le t < (j+1)T \tag{2}$$

or, equivalently,

$$s_i(t) = A_i \cos(\omega_c t + \theta) \cdot \Pi(t - jT) \tag{3}$$

where $\Pi(t) = 1$, $0 \le t < T$, and is 0 elsewhere. For example, for $k = 1$, each single bit is mapped to one of $M = 2^1 = 2$ symbols $\{s_1(t), s_2(t)\} = \{A_1 \cos(\omega_c t + \theta) \cdot \Pi(t - jT), A_2 \cos(\omega_c t + \theta) \cdot \Pi(t - jT)\}$. Specifically, the bit 0 is mapped to $s_1(t) = -A \cos(\omega_c t + \theta) \cdot \Pi(t - jT)$ (letting $A_1 = -A$), and the bit 1 is mapped to $s_2(t) = +A \cos(\omega_c t + \theta) \cdot \Pi(t - jT)$ (letting $A_2 = A$). This is shown in Fig. 3(a).

PSK refers to the mapping of the binary information into the phase of the carrier waveform. In PSK, each block of k bits is mapped to one of the $M = 2^k$ signals $\{s_1(t), s_2(t), \ldots, s_M(t)\}$, where

$$s_i(t) = A \cos(\omega_c t + \theta_i) \cdot \Pi(t - jT) \tag{4}$$

Here, $\theta_i = (2\pi/M)i$. For example, with $k = 1$, PSK corresponds to mapping the bit 0 to the waveform $s_1(t) = A \cos(\omega_c t + \pi) \cdot \Pi(t - jT)$ and mapping the bit 1 to $s_2(t) = A \cos(\omega_c t) \cdot \Pi(t - jT)$. This is shown in Fig. 3(b).

FSK corresponds to the mapping of binary information to carrier frequency. Specifically, FSK corresponds to mapping each block of k bits into one of the $M = 2^k$ signals $\{s_1(t), s_2(t), \ldots, s_M(t)\}$, where

$$s_i(t) = A \cos[(\omega_c + \Delta\omega_i)t + \theta] \cdot \Pi(t - jT) \tag{5}$$

An example of this, with $k = 1$, is provided in Fig. 3(c).

Derivation of the Receiver Front End. The first component of the demodulator (to the left of the divider line in Fig. 2) is known as the receiver front end. This sub-subsection derives the receiver front end, showing how it produces a sufficient statistic for detection.

Orthonormal Basis. Deriving the receiver front end requires some preliminary information, borrowed from linear algebra and provided here. Any set of M finite-energy signals, say, $\{s_1(t), s_2(t), \ldots, s_M(t)\}$, can be fully represented on an orthonormal set of N ($N \le M$) signals $\{\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)\}$. The orthonormal set of signals is so named because they satisfy the property

$$\int_{-\infty}^{\infty} \varphi_i(t)\varphi_j(t)\,dt = \begin{cases} 0, & i \ne j \\ 1, & i = j \end{cases} \tag{6}$$

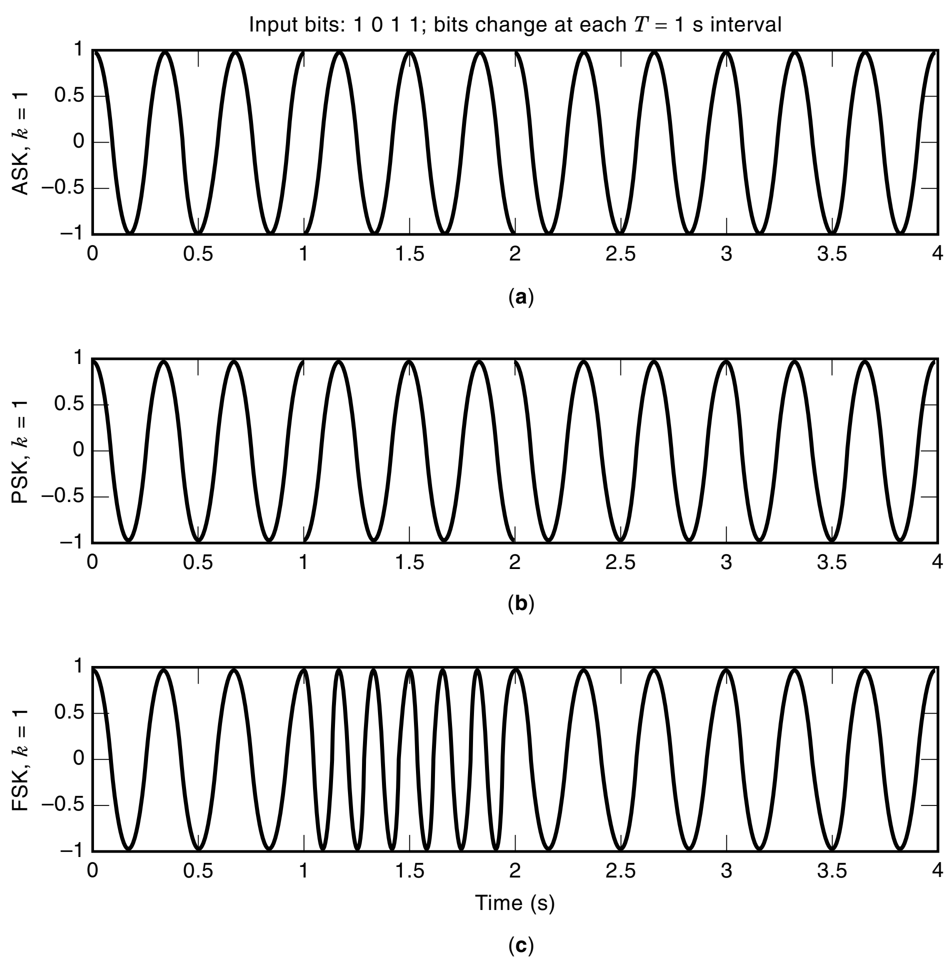

Figure 3. (a) ASK modulation with $k = 1$; (b) PSK modulation with $k = 1$; (c) FSK modulation with $k = 1$.

For any set of M signals, the corresponding orthonormal set can be established by the Gram–Schmidt orthogonalization procedure (1, Appendix 4A).

The signal set $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ can be expressed using the orthonormal basis according to

$$\begin{aligned} s_1(t) &= s_{11}\varphi_1(t) + s_{12}\varphi_2(t) + \cdots + s_{1N}\varphi_N(t) \\ s_2(t) &= s_{21}\varphi_1(t) + s_{22}\varphi_2(t) + \cdots + s_{2N}\varphi_N(t) \\ &\vdots \\ s_M(t) &= s_{M1}\varphi_1(t) + s_{M2}\varphi_2(t) + \cdots + s_{MN}\varphi_N(t) \end{aligned} \tag{7}$$

where $s_{ij} = \int s_i(t)\varphi_j(t)\,dt$; that is, $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ can be represented by vectors $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_M\}$ where $\boldsymbol{s}_i = (s_{i1}, s_{i2}, \ldots, s_{iN})$.

Orthonormal Basis for PSK. When the set of M signals $\{s_1(t), \ldots, s_M(t)\}$ corresponds to those generated by PSK modulation [i.e., $s_i(t) = A\cos(\omega_c t + \theta_i)\cdot\Pi(t - jT)$], the orthonormal basis functions $\{\varphi_1(t), \ldots, \varphi_N(t)\}$ can be established as follows. First, apply the trigonometric rule $\cos(A + B) = \cos A\cos B - \sin A\sin B$ in Eq. (4). This leads to

$$s_i(t) = A\cos\theta_i\cos\omega_c t\cdot\Pi(t - jT) - A\sin\theta_i\sin\omega_c t\cdot\Pi(t - jT) \tag{8}$$

Next, we can easily show that $\{\varphi_1(t), \varphi_2(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos\omega_c t\cdot\Pi(t - jT)$ and $\varphi_2(t) = -\sqrt{2/T}\sin\omega_c t\cdot\Pi(t - jT)$, is an orthonormal basis (assuming $\omega_c \gg 1/T$. It follows that $s_i(t)$ can be fully represented using $\{\varphi_1(t), \varphi_2(t)\}$ as

$$s_i(t) = \left(\sqrt{\frac{T}{2}}A\cos\theta_i\right)\varphi_1(t) + \left(\sqrt{\frac{T}{2}}A\sin\theta_i\right)\varphi_2(t) \tag{9}$$

Equivalently, we can represent $s_i(t)$ by the vector $\boldsymbol{s}_i = (\sqrt{T/2}\,A\cos\theta_i, \sqrt{T/2}\,A\sin\theta_i)$, where it is understood that the first component in the vector is along $\varphi_1(t)$ and the second component is along $\varphi_2(t)$.

It follows that the set of vectors $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ corresponding to PSK modulation can be fully represented on the basis $\{\varphi_1(t), \varphi_2(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos\omega_c t\cdot\Pi(t - jT)$ and $\varphi_2(t) = -\sqrt{2/T}\sin\omega_c t\cdot\Pi(t - jT)$. Along this basis, each $s_i(t)$ is fully represented by $s_i = (s_{i1}, s_{i2}) = (\sqrt{T/2}\,A\cos\theta_i, \sqrt{T/2}\,A\sin\theta_i)$. An example of this is shown in Fig. 4(a).

Orthonormal Basis for ASK. When the set of M signals $\{s_1(t), \ldots, s_M(t)\}$ corresponds to ASK modulation [i.e., $s_i(t) = A_i\cos(\omega_c t + \theta)\cdot\Pi(t - jT)$], the orthonormal basis functions are easily established. Specifically, we simply note that $\{\varphi_1(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos(\omega_c t + \theta)\cdot\Pi(t - jT)$, forms a single-element orthonormal basis. It follows that the $s_i(t)$ of ASK, described by Eq. (3), can be expressed simply as

$$s_i(t) = \sqrt{\frac{T}{2}}A_i\cdot\varphi_1(t) \tag{10}$$

Hence, the set $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ corresponding to ASK modulation can be fully represented in the orthonormal basis $\{\varphi_1(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos(\omega_c t + \theta)\cdot\Pi(t - jT)$. The signal $s_i(t)$ is expressed in this basis as $\boldsymbol{s}_i = s_i = \sqrt{T/2}\,A_i$, where it is understood that this represents the projection of $s_i(t)$ along $\varphi_1(t)$. An example of this is shown in Fig. 4(b).

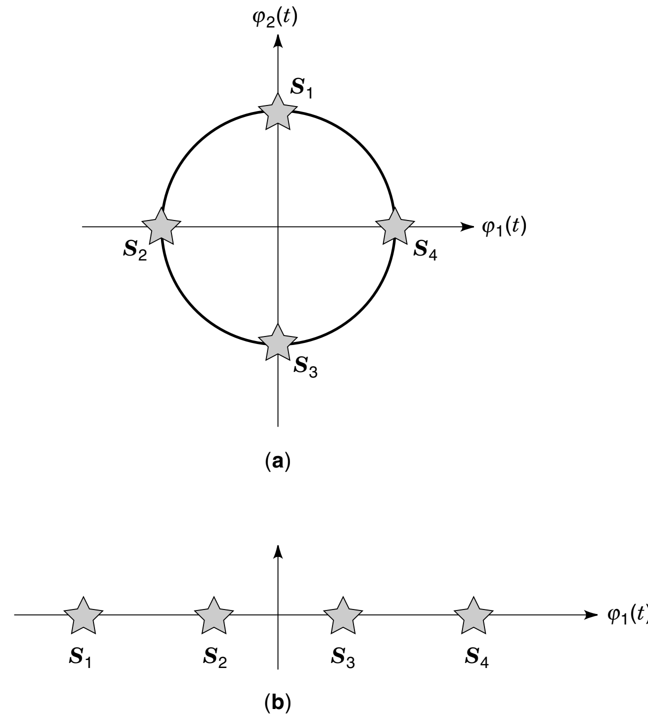

Figure 4. (a) PSK symbols with $M = 4$; (b) ASK symbols with $M = 4$. Both are represented along their orthonormal basis.

Orthonormal Basis for FSK. Typically, in FSK, $\Delta\omega_i$ is chosen so that

$$\int s_i(t)s_j(t)\,dt = 0, \qquad i \neq j \tag{11}$$

In this case, an orthonormal basis for FSK is simply $\{\varphi_1(t), \varphi_2(t), \ldots, \varphi_M(t)\}$, with $\varphi_i(t) = s_i(t)/K$ and $K = A\sqrt{T/2}$.

Representing r(t) along an Orthonormal Basis Achieves the Receiver Front End. In this subsection, the received signal $r(t) = s(t) + n(t)$ is represented along an orthonormal basis. We show how this leads to the receiver front end (the left side of the demodulator implemented in Fig. 2).

For simplicity in presentation, we assume that $s(t)$ corresponds to a PSK signal, that is, an element in the set of PSK modulated signals $\{s_1(t), \ldots, s_M(t)\}$. Hence, $s(t)$ can be represented fully on $\{\varphi_1(t), \varphi_2(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos\omega_c t\cdot\Pi(t - jT)$ and $\varphi_2(t) = -\sqrt{2/T}\sin\omega_c t\cdot\Pi(t - jT)$.

Consider representing $r(t) = s(t) + n(t)$ along the orthonormal basis $\{\varphi_1(t), \varphi_2(t), \varphi_3(t), \ldots\}$, where elements $\{\varphi_1(t), \varphi_2(t)\}$ correspond to the orthonormal basis of $s(t)$, and $\{\varphi_3(t), \ldots\}$ corresponds to whatever other signals are required in the orthonormal basis to represent $r(t)$. In this case, $r(t)$ can be expressed as $r(t) = r_1\varphi_i(t) + r_2\varphi_2(t) + r_3\varphi_3(t) + \cdots$, that is, $r(t)$ can be represented by $(r_1, r_2, r_3, \ldots)$, where r_i represents the component of the signal $r(t)$ along the signal $\varphi_i(t)$. Specifically,

$$\begin{aligned} r_1 &= \int_{-\infty}^{\infty} r(t)\varphi_1(t)\,dt \\ &= \int_{-\infty}^{\infty} s(t)\varphi_1(t)\,dt + \int_{-\infty}^{\infty} n(t)\varphi_1(t)\,dt = s_1 + n_1 \end{aligned} \tag{12}$$

Here, s_1 represents the projection of $s(t)$ on $\varphi_1(t)$, known from our earlier result to be $\sqrt{T/2}\,A\cos\theta_i$; also, n_1 is shorthand for the integral $\int_{-\infty}^{\infty} n(t)\varphi_1(t)\,dt$, and because $n(t)$ is AWGN, n_1 represents a Gaussian random variable. Similarly,

$$\begin{aligned} r_2 &= \int_{-\infty}^{\infty} r(t)\varphi_2(t)\,dt \\ &= \int_{-\infty}^{\infty} s(t)\varphi_2(t)\,dt + \int_{-\infty}^{\infty} n(t)\varphi_2(t)\,dt = s_2 + n_2 \end{aligned} \tag{13}$$

Here, s_2 is the projection of $s(t)$ on $\varphi_2(t)$, known (see our earlier result) to be $\sqrt{T/2}\,A\sin\theta_i$. Also, $n_2 = \int_{-\infty}^{\infty} n(t)\varphi_2(t)\,dt$, and with $n(t)$ corresponding to AWGN, n_2 is a Gaussian random variable independent of n_1. Additionally,

$$\begin{aligned} r_3 &= \int_{-\infty}^{\infty} r(t)\varphi_3(t)\,dt \\ &= \int_{-\infty}^{\infty} s(t)\varphi_3(t)\,dt + \int_{-\infty}^{\infty} n(t)\varphi_3(t)\,dt = 0 + n_3 = n_3 \end{aligned} \tag{14}$$

Here, $\int_{-\infty}^{\infty} s(t)\varphi_3(t)\,dt = \int_{-\infty}^{\infty} [s_1\varphi_1(t) + s_2\varphi_2(t)]\varphi_3(t)\,dt = s_1\int_{-\infty}^{\infty} \varphi_1(t)\varphi_3(t)\,dt + s_2\int_{-\infty}^{\infty} \varphi_2(t)\varphi_3(t)\,dt = s_1\cdot 0 + s_2\cdot 0 = 0$; in words, because $s(t)$ is fully represented on $\varphi_1(t)$ and $\varphi_2(t)$, there is no remaining component to project on $\varphi_3(t)$. Also, n_3 is a Gaussian random variable independent of n_1 and n_2. Similarly,

$$r_4 = n_4 \tag{15}$$

$$r_5 = n_5 \tag{16}$$

and so on. Now, since $r_3, r_4, r_5, \ldots$ represent only noise terms and these noise terms are independent of r_1 and r_2, they are simply not useful in deciding which signal in the set $\{s_1(t), \ldots, s_M(t)\}$ was sent by the modulator. Hence, the only terms required for detection are

$$r_1 = \int_{-\infty}^{\infty} r(t)\varphi_1(t)\,dt = s_1 + n_1 \tag{17}$$

$$r_2 = \int_{-\infty}^{\infty} r(t)\varphi_2(t)\,dt = s_2 + n_2 \tag{18}$$

or, in shorthand notation,

$$\boldsymbol{r} = \boldsymbol{s} + \boldsymbol{n} \tag{19}$$

where $\boldsymbol{r} = (r_1, r_2)$, $\boldsymbol{s} = (s_1, s_2)$, and $\boldsymbol{n} = (n_1, n_2)$. That is, the only terms required for detection are the projections of $r(t)$ along the orthonormal basis of $s(t)$, which in the case of PSK signaling is $\{\varphi_1(t), \varphi_2(t)\}$, with $\varphi_1(t) = \sqrt{2/T}\cos\omega_c t\cdot\Pi(t - jT)$ and $\varphi_2(t) = -\sqrt{2/T}\sin\omega_c t\cdot\Pi(t - jT)$.

Denoting the orthonormal basis for a general $\{s_1(t), \ldots, s_M(t)\}$ as $\{\varphi_1(t), \ldots, \varphi_N(t)\}$, then the creation of the vector $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)$ is all that is required for detection. The creation of this vector is implemented as shown in the receiver front end in Fig. 2.

Derivation of the Data Detector. The remainder of the demodulator is derived by starting from a simple premise and introducing statistical and mathematical arguments.

The demodulator wants to determine which element in the set $\{s_1(t), \ldots, s_M(t)\}$, or equivalently, which vector in the set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$, was sent across the channel by the modulator. If it can determine this, then it knows which bits were input to the modulator. Specifically, the demodulator can be described as wanting to output the element in the set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$ that is most likely to be correct, that is, least likely to be incorrect. Mathematically, the demodulator wants to output the element $\hat{\boldsymbol{s}}_i$ according to

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} P(\epsilon) \tag{20}$$

where $P(\epsilon)$ denotes the probability of error. This is accomplished when the demodulator chooses $\hat{\boldsymbol{s}}_i$ using

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} p(\boldsymbol{s}_i|\boldsymbol{r}) \tag{21}$$

That is, the demodulator wants to choose the element in the set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$ that is most likely, given that $\boldsymbol{r}$ was received. Applying Bayes's rule, namely, $P(A|B) = P(B|A)P(A)/P(B)$, leads to

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \frac{p(\boldsymbol{r}|\boldsymbol{s}_i)\,p(\boldsymbol{s}_i)}{p(\boldsymbol{r})} \tag{22}$$

Next, observing that $p(\boldsymbol{r})$ is independent of $\boldsymbol{s}_i$ and hence plays no role in the optimization, we have

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} p(\boldsymbol{r}|\boldsymbol{s}_i)\,p(\boldsymbol{s}_i) \tag{23}$$

Now, $p(\boldsymbol{r}|\boldsymbol{s}_i)$ denotes the probability that $\boldsymbol{r}$ is received, given that $\boldsymbol{s}_i$ is sent. However, since $\boldsymbol{r} = \boldsymbol{s}_i + \boldsymbol{n}$ (i.e., $\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s}_i$, then $\boldsymbol{r}$ is received, given that $\boldsymbol{s}_i$ is sent, if and only if the noise $\boldsymbol{n}$ equals $\boldsymbol{r} - \boldsymbol{s}_i$, that is, $p(\boldsymbol{r}|\boldsymbol{s}_i) = p(\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s}_i)$. It follows that

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} p(\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s}_i)\,p(\boldsymbol{s}_i) \tag{24}$$

Next, because the noise $n(t)$ is AWGN, the vector $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$ represents a set of independent, identically distributed (i.i.d.) Gaussian random variables, that is, $\boldsymbol{n}$ has a probability density function (pdf) given by

$$p(\boldsymbol{n}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|\boldsymbol{n}|^2}{2\sigma_n^2}\right) \tag{25}$$

Applying this distribution to Eq. (24) results in

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|\boldsymbol{r} - \boldsymbol{s}_i|^2}{2\sigma_n^2}\right)\cdot p(\boldsymbol{s}_i) \tag{26}$$

Now, $\ln f(x)$ increases whenever $f(x)$ increases, and hence the optimization of $f(x)$ is equivalent to that of $\ln f(x)$. It follows that $\hat{\boldsymbol{s}}_i$ can be expressed by

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \ln\left[\frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|\boldsymbol{r}-\boldsymbol{s}_i|^2}{2\sigma_n^2}\right) p(\boldsymbol{s}_i)\right] \tag{27}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \left\{\ln\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right) + \ln\left[\exp\left(-\frac{|\boldsymbol{r}-\boldsymbol{s}_i|^2}{2\sigma_n^2}\right)\right] + \ln p(\boldsymbol{s}_i)\right\} \tag{28}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \left(-\frac{|\boldsymbol{r}-\boldsymbol{s}_i|^2}{2\sigma_n^2} + \ln p(\boldsymbol{s}_i)\right) \tag{29}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} \left[|\boldsymbol{r}-\boldsymbol{s}_i|^2 - 2\sigma_n^2 \ln p(\boldsymbol{s}_i)\right] \tag{30}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} \left[|\boldsymbol{r}|^2 + |\boldsymbol{s}_i|^2 - 2\boldsymbol{r}\cdot\boldsymbol{s}_i - 2\sigma_n^2 \ln p(\boldsymbol{s}_i)\right] \tag{31}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} \left\{\boldsymbol{r}\cdot\boldsymbol{s}_i + \tfrac{1}{2}[2\sigma_n^2 \ln p(\boldsymbol{s}_i) - |\boldsymbol{s}_i|^2]\right\} \tag{32}$$

where $\boldsymbol{r}\cdot\boldsymbol{s}_i = \sum_{j=1}^{N} r_j s_{ij}$. Letting $c_i = \frac{1}{2}[2\sigma_n^2 \ln p(\boldsymbol{s}_i) - |\boldsymbol{s}_i|^2]$, we have

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} (\boldsymbol{r}\cdot\boldsymbol{s}_i + c_i) \tag{33}$$

Hence, the second part of the demodulator chooses $\hat{\boldsymbol{s}}_i$ according to the criteria of Eq. (33). This is implemented in the right half (right of the divider line) of the receiver of Fig. 2.

Applications of the Digital Data Demodulator to ASK, PSK, and FSK

The digital data demodulator, derived in the previous section and shown in Fig. 2, is generally applicable for demodulation, whenever the transmitted signal $s(t)$ generated by the modulator is a selection from the finite-energy signal set $\{s_1(t), \ldots, s_M(t)\}$ and the received signal $r(t)$ corresponds to $r(t) = s(t) + n(t)$ with $n(t)$ corresponding to AWGN. When the modulator, and hence the set $\{s_1(t), \ldots, s_M(t)\}$, corresponds specifically to ASK, PSK, or FSK, the digital data demodulator can be simplified. This subsection presents those simplifications.

Digital Data Demodulator for ASK Modulation. As highlighted by Eq. (3), ASK modulation corresponds to mapping k bits to one of the $M = 2^k$ symbols $\{s_1(t), \ldots, s_M(t)\}$, where $s_i(t) = A_i \cos(\omega_c t + \theta)\cdot\Pi(t - jT)$. Thus, using the orthonormal basis $\{\varphi_1(t)\}$ where $\varphi_1(t) = \sqrt{2/T}\cos(\omega_c t + \theta)\cdot\Pi(t - jT)$, $s_i(t) = (\sqrt{T/2}\,A_i)\cdot\varphi_1(t)$, that is, $s_i(t)$ can be represented by $s_i = \sqrt{T/2}\,A_i$.

The presentation in the previous section shows that the front end of the demodulator maps $r(t)$ onto the orthonormal basis of the transmitted signal $s(t) \in \{s_1(t), \ldots, s_M(t)\}$. Hence,

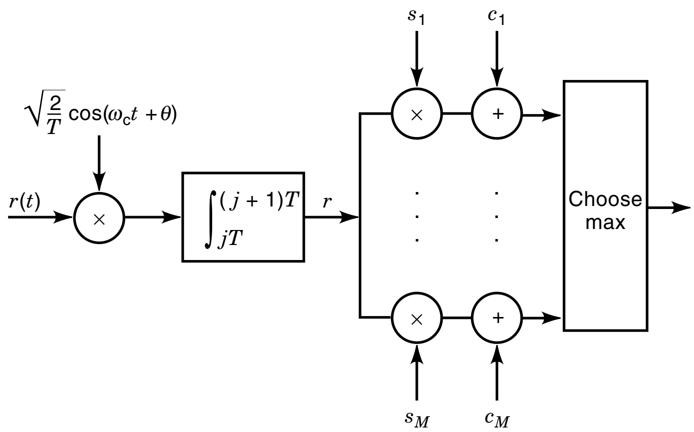

Figure 5. Digital data demodulator for ASK.

for the ASK case at hand, the receiver front end simply evaluates

$$\begin{aligned} r = r_1 &= \int_{-\infty}^{\infty} r(t)\varphi_1(t)\,dt \\ &= \int_{-\infty}^{\infty} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t + \theta)\cdot\Pi(t - jT)\,dt \\ &= \int_{jT}^{(j+1)T} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t + \theta)\,dt \end{aligned} \tag{34}$$

This is implemented as shown in the left side of Fig. 5.

Referring now to the derivation of the remainder of the demodulator and specifically Eq. (33), the remainder of the demodulator for the case at hand evaluates

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmax}} r s_i + c_i \tag{35}$$

This is implemented as shown on the left-hand side of Fig. 5.

Digital Data Demodulator for PSK Modulation. A simplification of the digital data demodulator of Fig. 2 also exists for PSK demodulators. First, recall that in PSK modulation k bits are mapped to one of $M = 2^k$ symbols $\{s_1(t), \ldots, s_M(t)\}$, where $s_i(t) = A\cos(\omega_c t + \theta_i)\cdot\Pi(t - jT)$ and $\theta_i = (2\pi/M)i$. Along the orthonormal basis $\{\varphi_1(t), \varphi_2(t)\}$ with $\varphi_1(t) = \sqrt{2/T}\cos\omega_c t\cdot\Pi(t - jT)$ and $\varphi_2(t) = -\sqrt{2/T}\sin\omega_c t\cdot\Pi(t - jT)$, $s_i(t) = (\sqrt{T/2}\,A\cos\theta_i)\varphi 1(t) + (\sqrt{T/2}\,A\sin\theta_i)\varphi 2(t)$, or, in vector notation, $s_i(t)$ is represented by $\boldsymbol{s}_i = (s_{i1}, s_{i2}) = (\sqrt{T/2}\,A\cos\theta_i, \sqrt{T/2}\,A\sin\theta_i)$.

Following the sub-subsection entitled "Derivation of the Receiver Front End" above, the receiver front end in the case at hand computes

$$\begin{aligned} r_1 &= \int_{-\infty}^{\infty} r(t)\varphi_1(t)\,dt = \int_{-\infty}^{\infty} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t)\cdot\Pi(t - jT)\,dt \\ &= \int_{jT}^{(j+1)T} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t)\,dt \end{aligned} \tag{36}$$

$$\begin{aligned} r_2 &= \int_{-\infty}^{\infty} r(t)\varphi_2(t)\,dt = \int_{-\infty}^{\infty} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t)\cdot\Pi(t - jT)\,dt \\ &= \int_{jT}^{(j+1)T} r(t)\sqrt{\frac{2}{T}}\cos(\omega_c t)\,dt \end{aligned} \tag{37}$$

This is implemented as shown in the left-hand side of Fig. 6.

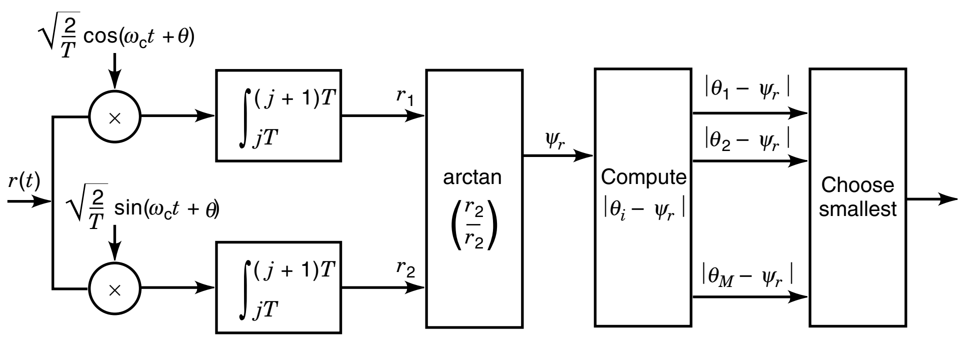

Figure 6. Digital data demodulator for PSK.

The remainder of the demodulator is usually implemented as follows. According to Eq. (30), the remainder of the receiver implements

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} |\boldsymbol{r}-\boldsymbol{s}_i|^2 - 2\sigma_n^2 \ln p(\boldsymbol{s}_i) \tag{38}$$

where $\boldsymbol{s}_i = (s_{i1}, s_{i2}) = (\sqrt{T/2}\,A\cos\theta_i, \sqrt{T/2}\,A\sin\theta_i) = (\sqrt{T/2}\,A\cos(2\pi/M)i, \sqrt{T/2}\,A\sin(2\pi/M)i)$ and $\boldsymbol{r} = (r_1, r_2)$. An example of this is shown graphically in Fig. 7 (using $M = 4$).

It is usually assumed, and we will make this assumption here, that all the transmitted symbols are equally likely, that is, $p(\boldsymbol{s}_i) = 1/M$. Applying this to Eq. (38) leads to

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} |\boldsymbol{r}-\boldsymbol{s}_i|^2 \tag{39}$$

This equation states simply that the selected $\hat{\boldsymbol{s}}_i$ should correspond to the $\boldsymbol{s}_i$ closest to $\boldsymbol{r}$. This is highlighted graphically by Fig. 8.

From Fig. 8, the criterion for choosing $\hat{\boldsymbol{s}}_i$ is based exclusively on the angle of $\boldsymbol{r}$, that is, exclusively on $\mu_r = \arctan(r_2/r_1)$. If the angle is closest to θ_1 (the angle of $\boldsymbol{s}_1$), then $\hat{\boldsymbol{s}}_i$ should be chosen as $\boldsymbol{s}_1$; if closest to θ_2 (the angle of $\boldsymbol{s}_2$), then $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_2$; and so on. The demodulator implementing this is built as shown in the right-hand side of Fig. 6.

Digital Data Demodulator for Orthogonal FSK. In this section we briefly present the digital data demodulator when the modulated signal is orthogonal FSK, that is, when the modulator maps k bits into one of the $M = 2^k$ signals $\{s_1(t), \ldots, s_M(t)\}$, where $s_i(t) = A\cos[(\omega_c + \Delta\omega_i)t + \theta]\cdot\Pi(t - jT)$ and $\Delta\omega_i$ is chosen such that $\int_{-\infty}^{\infty} s_i(t)s_k(t)\,dt = 0$, $i \neq k$. In this case, the orthonormal basis for the transmitted signal set is $\{\varphi_1(t), \ldots, \varphi_M(t)\}$ where $\varphi_i = s_i(t)/K$. Hence, $s_i(t)$ can be expressed as $s_i(t) = K\varphi_i(t)$, or, in vector notation, $\boldsymbol{s}_i = (0, \ldots, 0, K, 0, \ldots, 0)$, where K is the ith element in the M-element vector.

Following the general derivation of the receiver front end, the receiver front end for the case at hand computes

$$\begin{aligned} r_1 &= \int_{-\infty}^{\infty} r(t)\varphi_1(t)\,dt = \int_{jT}^{(j+1)T} r(t)\frac{A}{K}\cos[(\omega_c+\Delta\omega_1)t+\theta]\,dt \\ &\;\;\vdots \\ r_M &= \int_{-\infty}^{\infty} r(t)\varphi_M(t)\,dt = \int_{jT}^{(j+1)T} r(t)\frac{A}{K}\cos[(\omega_c+\Delta\omega_M)t+\theta]\,dt \end{aligned} \tag{40}$$

This is implemented as shown in the left side of Fig. 9.

The remainder of the demodulator evaluates

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i\in\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} \boldsymbol{r}\cdot\boldsymbol{s}_i + c_i \tag{41}$$

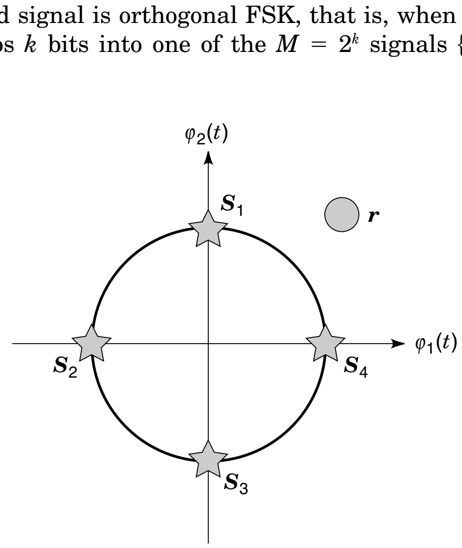

Figure 7. PSK symbols ($M = 4$) and received signal.

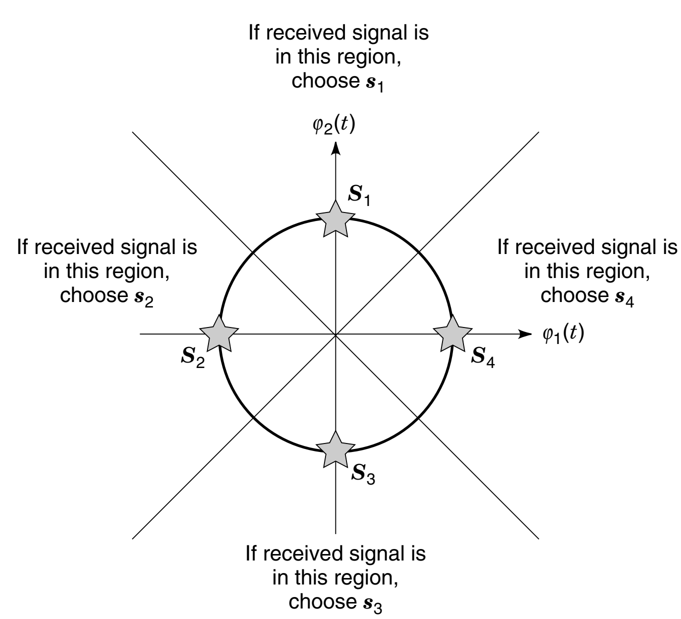

Figure 8. Explaining how a PSK demodulator can decide which symbol to output.

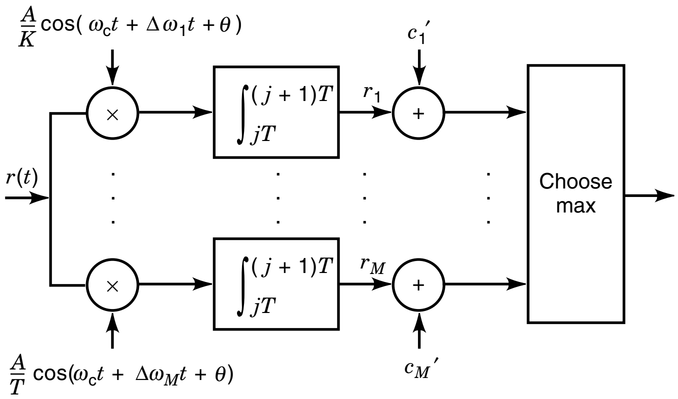

Figure 9. Digital data demodulator for FSK.

Using the $\boldsymbol{s}_i$ for FSK, namely $\boldsymbol{s}_i = (0, \ldots, 0, K, 0, \ldots, 0)$, leads to

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} \; r_i K + c_i \tag{42}$$

$$\hat{\boldsymbol{s}}_i = \underset{\boldsymbol{s}_i \in \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}}{\operatorname{argmin}} \; r_i + c_i' \tag{43}$$

This is implemented as shown in the right side of Fig. 9. Finally, note that if the signals are equally likely, then $c_1' = c_2' = \cdots = c_M'$, and hence the addition of c_i' can be removed from Eq. (43) and the implementation of Fig. 9.

Alternative Implementation of the Digital Data Demodulator

The digital data demodulator, shown in Fig. 2, is commonly referred to as the correlation receiver. Other implementations of this demodulator, known as matched-filter receivers, are also commonly used. These are detailed in this section.

For simplicity in presentation, we reduce the terms of the form $s_i(t) = g(t)\Pi(t - jT)$ to $s_i(t) = g(t)\Pi(t)$ (i.e., we assume $j = 0$).

Matched-Filter Implementation. An alternative implementation of the demodulator of Fig. 2 is shown in Fig. 10. This implementation is known as the matched-filter receiver. Quick comparison of Fig. 2 and 10 indicates that these receivers are equivalent if it can be demonstrated that $u_i(T) = r_i$ for $i = 1, \ldots, N$. This is shown as follows. First, following Fig. 10, $u_i(t)$ corresponds to the output of the filter with impulse response $\varphi_i(T - t)$ and input $r(t)$, that is,

$$u_i(t) = r(t) * \varphi_i(T - t) \tag{44}$$

$$u_i(t) = \int_{-\infty}^{\infty} r(t-\tau)\varphi_i(T-\tau)\,d\tau \tag{45}$$

Now, $u_i(T)$ is the sampling of $u_i(t)$ at time $t = T$, and hence

$$u_i(T) = \int_{-\infty}^{\infty} r(T-\tau)\varphi_i(T-\tau)\,d\tau \tag{46}$$

$$u_i(T) = \int_{-\infty}^{\infty} r(u)\varphi_i(u)\,du \tag{47}$$

$$u_i(T) = r_i \tag{48}$$

Since $u_i(T) = r_i$, we conclude that the matched-filter receiver of Fig. 10 is simply an alternative implementation of Fig. 2.

Matched-Filter Implementation in Terms of $s_i(t)$. Another common implementation of the matched-filter receiver is presented in Fig. 11.

Comparing Figs. 2 and 11, it is apparent that, if we can show the equivalence $p_i(T) = \boldsymbol{r} \cdot \boldsymbol{s}_i$, then these two implementations are equivalent.

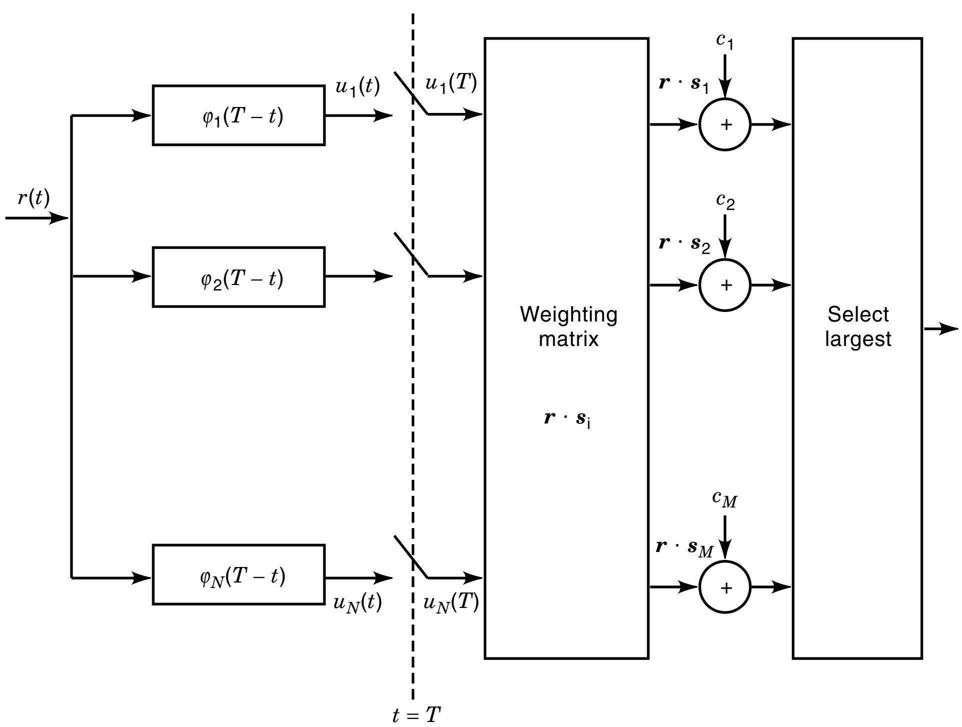

Figure 10. Matched-filter receiver.

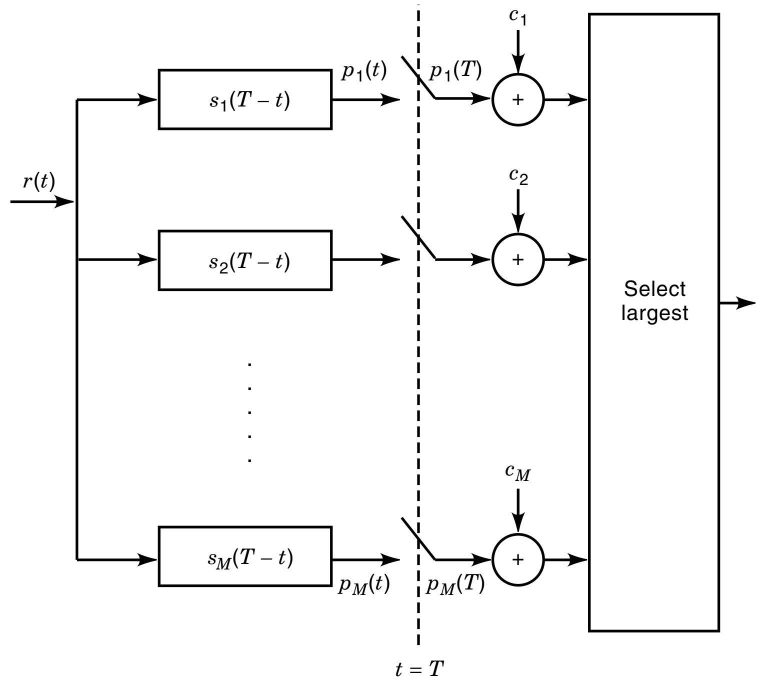

Figure 11. Matched-filter receiver in terms of the transmitted symbols.

The equality $p_i(T) = \boldsymbol{r} \cdot \boldsymbol{s}_i$ is shown as follows. First, referring to Fig. 11, it is apparent that $p_i(t)$ is the output of the filter with impulse response $s_i(T - t)$ and input $r(t)$, that is,

$$p_i(t) = r(t) * s_i(T - t) \tag{49}$$

$$p_i(t) = \int_{-\infty}^{\infty} r(t - \tau) s_i(T - \tau)\, d\tau \tag{50}$$

Next, $p_i(T)$ is simply $p_i(t)$ evaluated at $t = T$, that is,

$$p_i(T) = \int_{-\infty}^{\infty} r(T - \tau) s_i(T - \tau)\, d\tau \tag{51}$$

$$p_i(T) = \int_{-\infty}^{\infty} r(u) s_i(u)\, du \tag{52}$$

Applying Parseval's rule, namely, $\int_{-\infty}^{\infty} r(u)s_i(u)\, du = \boldsymbol{r} \cdot \boldsymbol{s}_i$, leads to the desired result

$$p_i(T) = \boldsymbol{r} \cdot \boldsymbol{s}_i \tag{53}$$

This equality confirms the equivalence of the two structures.

The demodulator structure of Fig. 11 eliminates the need for the explicit evaluation done in the weighting matrix of Figs. 2 and 10. However, M parallel filters are required, rather than the N computations/filterings required in Figs. 2 and 10. If $M \gg N$, then the implementation of Figs. 2 or 10 is usually preferred. Otherwise, the implementation of Fig. 11 is recommended.

Digital Implementation of Matched Filters. The availability and low expense of digital technology has led to digital implementation of matched-filter receivers. Specifically, matched filtering followed by sampling (see Fig. 12) is replaced by the processing shown in Fig. 13. Here, a band-pass filter removes the noise outside the signal bandwidth. This is followed by a sampler and then by a digital version of the matched filter.

Demodulation with Explicit Clock, Phase, and Frequency Recovery

The digital data demodulation shown for ASK in Fig. 5, for PSK in Fig. 6, and for FSK in Fig. 9 requires (1) exact knowledge of the phase and frequency information for use in the cosine and sine products, and (2) exact knowledge of the start time and end time of the transmitted symbol $s(t)$ for use in the integrators. These requirements are explicitly acknowledged in the more complete implementation shown in Fig. 14.

Phase is most commonly recovered with a *phase-locked loop* (PLL), frequency is typically recovered with an automatic frequency control (AFC) loop, and timing is recovered by one of a number of available feedforward and feedback schemes, such as early–late gate timing recovery, sample-derivative timing recovery, and in-phase/midphase timing recovery.

ANALOG DATA DEMODULATORS

In some communication systems, the information communicated between transmitter and receiver is not a sequence of bits, but rather an analog signal $x(t)$. In this case, an analog

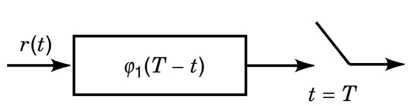

Figure 12. Matched filtering followed by sampling.

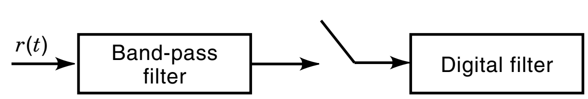

Figure 13. Digital implementation of matched filter and sampling.

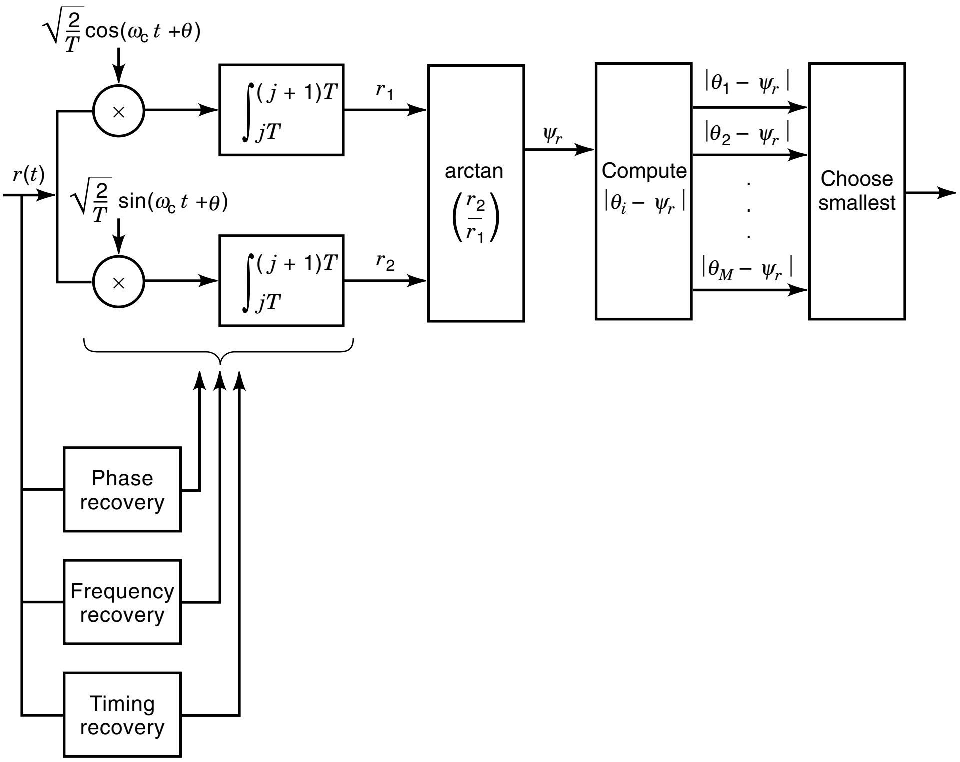

Figure 14. Digital data demodulator for PSK showing explicit use of phase, frequency, and timing recovery.

modulator maps the analog information signal $x(t)$ to the amplitude, phase, or frequency of a sinusoidal signal $s(t)$, called the carrier. The signal $s(t)$ is then sent across the channel. This section presents analog data demodulators: physical devices that recover the information signal $x(t)$ from the received signal. This is highlighted in Fig. 15.

Analog Modulator

Before we present the design of the analog data demodulator, a brief description of analog modulation is in order. For brevity, only the two most popular analog modulation schemes are introduced: amplitude modulation (AM) and frequency modulation (FM).

AM, as the name suggests, is the mapping of the information signal $x(t)$ into the amplitude $s(t)$ of the carrier. Specifically, in AM the carrier signal is described by

$$s(t) = A \cdot [1 + mx(t)]\cos(\omega_c t) \tag{54}$$

Here, the information signal $x(t)$ is assumed to be normalized so that $x(t) \in [-1, 1]$, and m refers to the modulation index, a number between 0 and 1. An example of this modulation is shown in Fig. 16(a) and (b).

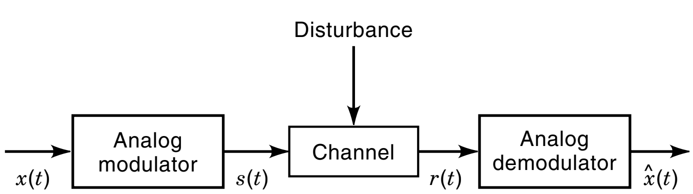

Figure 15. An analog communication system.

Another widely used analog modulation method is FM, where, as the name suggests, the analog signal $x(t)$ is mapped to the carrier frequency. In this case, the signal output by the FM modulator is described by

$$s(t) = A\cos[\omega_c t + \theta(t)] \tag{55}$$

where

$$\theta(t) = K_f \int_{-\infty}^{t} x(u)\,du \tag{56}$$

It may not be apparent from these two equations that the information signal $x(t)$ has been mapped into the frequency of $s(t)$. However, it is hoped that the following brief analysis will clarify this. The instantaneous frequency $f(t)$ of the signal $s(t)$ in Eq. (55) is evaluated as follows.

$$f(t) = \frac{1}{2\pi}\frac{d}{dt}[\omega_c t + \theta(t)] = \frac{1}{2\pi}\frac{d}{dt}\left(\omega_c t + K_f \int_{-\infty}^{t} x(u)\,du\right) \tag{57}$$

$$f(t) = \frac{\omega_c}{2\pi} + \frac{K_f}{2\pi}x(t) \tag{58}$$

$$f(t) = f_c + \frac{K_f}{2\pi}x(t) \tag{59}$$

It follows that the analog signal $x(t)$ effectively determines the instantaneous frequency of $s(t)$. This is highlighted in Fig. 16(c).

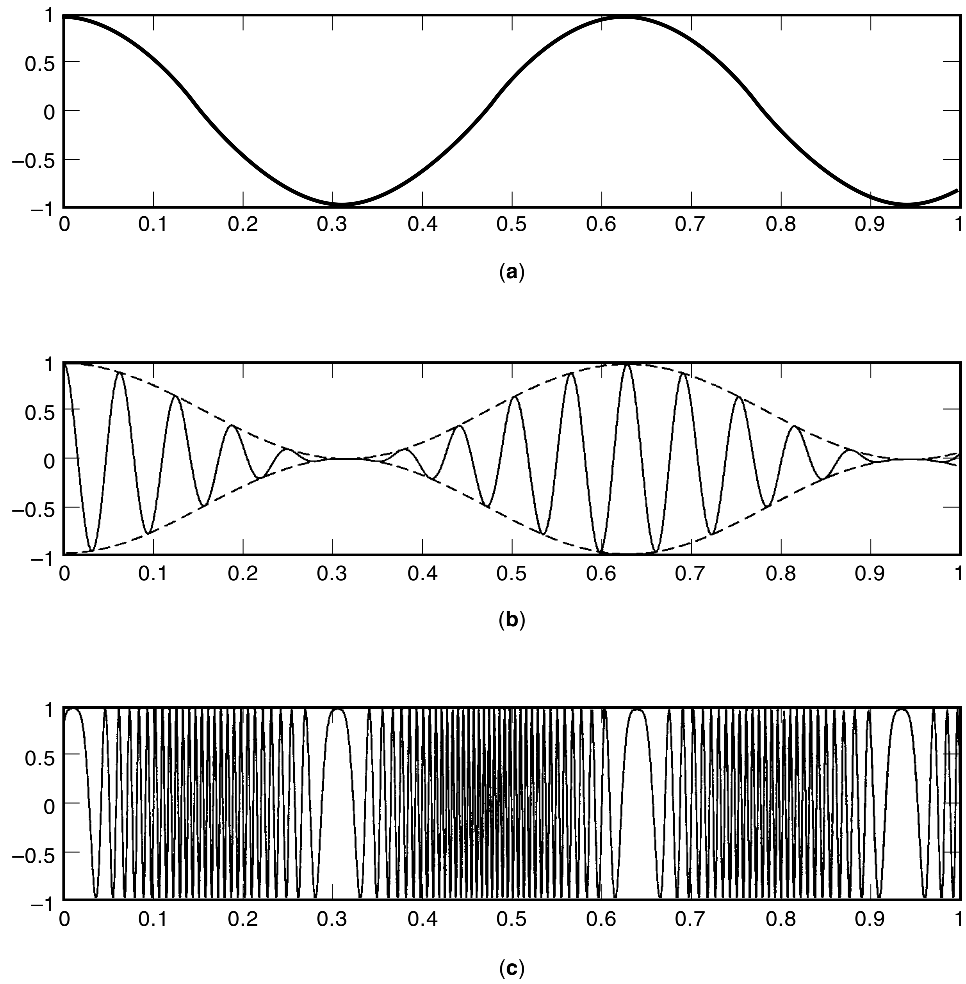

Figure 16. (a) Information signal; (b) AM signal; (c) FM signal.

Analog Data Demodulators

AM Demodulators (Envelope Detectors). In this section we introduce the AM demodulator, a physical device that extracts the information signal $x(t)$ from the AM signal $s(t)$.

The AM demodulator is best explained with the help of Fig. 16(a) and (b). A brief look at these two graphs reveals that the dashed line enveloping the AM signal $s(t)$ [Fig. 16(b)] is a replica of the information signal $x(t)$ [Fig. 16(a)]. It follows that an AM demodulator, achieving the recovery of $x(t)$ from $s(t)$, can be constructed using a device that, given $s(t)$, generates the envelope of $s(t)$. Devices that do this, called envelope detectors, are by far the most common type of AM demodulator.

A widely used envelope detector is shown in Fig. 17. The workings of this detector are best explained with the help of Fig. 18. Figure 18(a) shows the input to the envelope detector. This input traverses the diode (Fig. 17), which effectively maps all negative inputs to zero, while leaving positive inputs untouched. This is highlighted in Fig. 18(b). Next, the output of the diode passes through an RC circuit. This RC circuit acts as a low-pass filter, removing the rapid variations of the incoming signal, while leaving its slow variations (i.e., the envelope) essentially untouched. This is shown in Fig. 18(c).

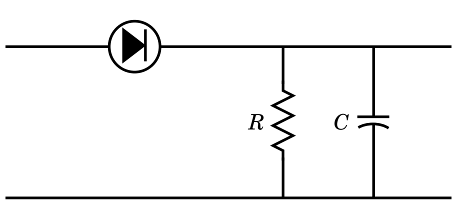

Figure 17. An envelope detector.

The values of R and C are carefully chosen to ensure that

$$B \ll \frac{1}{2\pi RC} \ll f_c \tag{60}$$

where B refers to the bandwidth of $x(t)$, $f_c = \omega_c/2\pi$ is the carrier frequency, and $1/2\pi RC$ is the cutoff frequency of the low-pass filter (RC circuit). That is, the values of R and C are chosen to ensure that the cutoff frequency of the low-pass filter (RC circuit), namely, $1/2\pi RC$, is much smaller than the carrier frequency f_c (and hence eliminates the rapid carrier frequency) and yet much larger than the bandwidth of the desired waveform $x(t)$ (and hence transmits the envelope).

FM Demodulators. Typically FM demodulators consist of two main components, a limiter and a discriminator, as shown in Fig. 19.

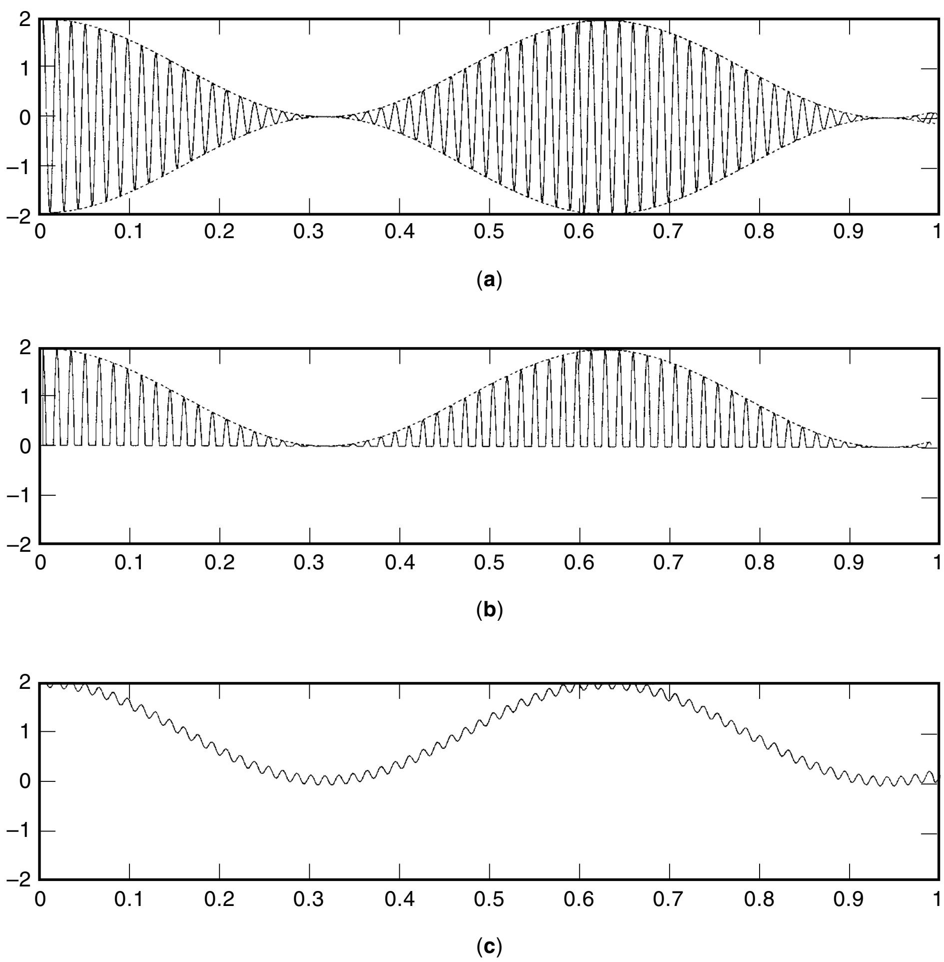

Figure 18. (a) AM signal input to envelope detector; (b) signal after the diode; (c) signal after RC circuit (LPF).

A limiter is a device that maps all positive values to $+A$ and all negative values to $-A$. Hence, if the FM signal shown in Fig. 16(c) is the input to the limiter, its output corresponds to Fig. 20. The limiter effectively removes any amplitude variations in the signal. More importantly, it effectively removes any amplitude distortions that have occurred in the channel. This operation does not cause any loss of the information $x(t)$, because $x(t)$ is stored in the frequency of the received signal, not in its amplitude.

After amplitude distortions are eliminated in the limiter, the FM signal enters the discriminator. A discriminator is any device that generates an output proportional to the instantaneous frequency of the input. In FM, the discriminator outputs are given by

$$y(t) = Kf(t) = K \cdot \left(f_c + \frac{K_f}{2\pi} x(t) \right) \tag{61}$$

Clearly, the output of the discriminator is easily mapped to the information signal $x(t)$.

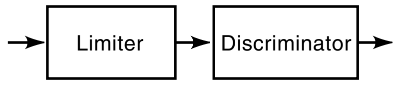

Figure 19. FM demodulator consisting of limiter followed by discriminator.

A number of different circuits can be used as discriminators. Some commonly used discriminators are (1) a differentiator followed by an envelope detector, and (2) a zero-crossing detector (counts the zero crossings of the input waveform). Implementations of these devices are found in Ref. 2, Chap. 4.

The Superheterodyne Receiver. The superheterodyne receiver is the most common AM radio receiver. This sub-subsection describes this receiver, which, in essence, is some filtering followed by an AM demodulator.

The superheterodyne receiver consists of two essential components: (1) a front end, consisting of filters that pass only the AM radio signal $s(t)$ that the radio dial is tuned to, and (2) an AM demodulator, which extracts the information signal $x(t)$ from the radio signal $s(t)$ (passing through the front end filtering). This receiver is shown in Fig. 21.

A detailed description of the superheterodyne receiver of Fig. 21 follows. For simplicity in presentation, we assume that the radio receiver is tuned to receive an AM signal at 900 kHz. The input signal to the receiver is the entire AM radio-frequency band, which includes, for example, radio stations at 600 kHz, 830 kHz, 900 kHz (desired), and 1210 kHz. A tunable RF filter $H_1(f)$ is the first component to greet the incoming AM band signal. With the receiver tuned to 900 kHz, $H_1(f)$ amplifies the signal at frequencies around 900 kHz and diminishes signals at all other frequencies. However, for reasons of cost, this tunable filter is not sharp. The signal

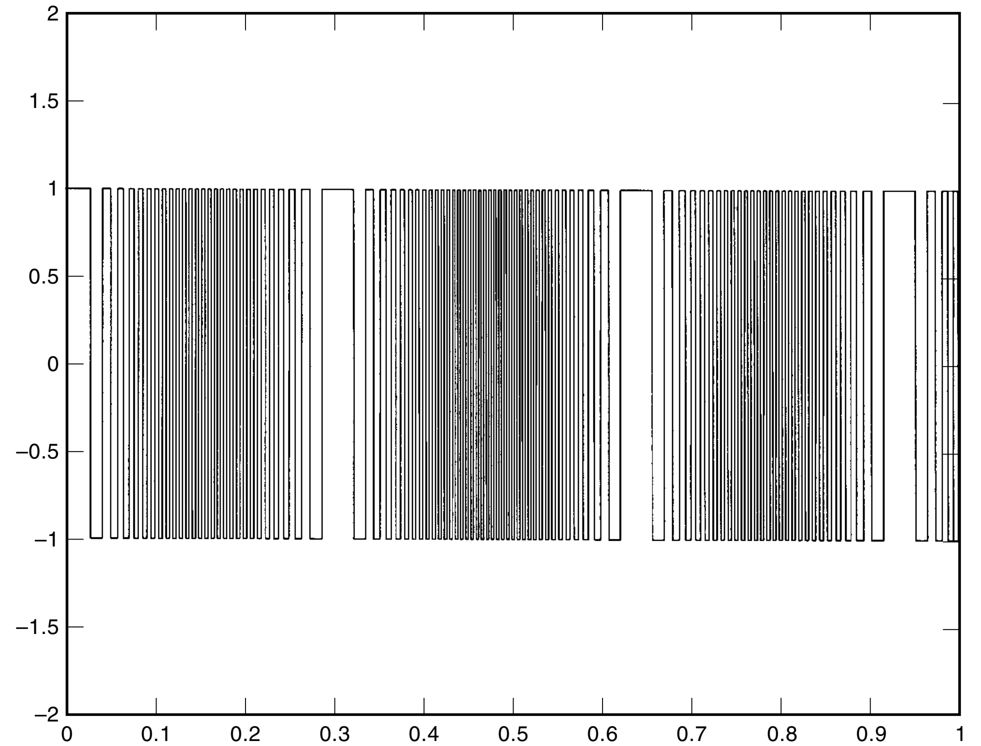

Figure 20. Limiter output for input of Fig. 16(c).

filtered by $H_1(f)$ is then passed through a tunable mixer, a device that shifts the input frequency. With the radio tuned to 900 kHz, the mixer components ensure that the signal at 900 kHz is shifted in frequency to the intermediate frequency (IF) of 455 kHz. Next, a very sharp filter, $H_2(f)$, is applied. It blocks all but the desired signal, now at 455 kHz. Finally, with only the desired radio signal present at 455 kHz, an AM demodulator extracts the information signal $x(t)$.

ADVANCED ISSUES

This section introduces advanced issues in digital data demodulation. Specifically, it highlights demodulators for trellis-coded modulation and demodulators capable of detecting data in rapidly changing phase-offset environments.

Demodulators for Trellis-Coded Modulation

Trellis-coded modulation (TCM) is a joint channel-coding and modulation scheme. A demodulator built to decode trellis-coded modulation acts as both a demodulator and a channel decoder.

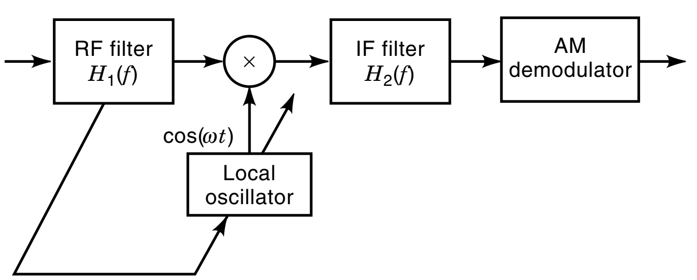

Figure 21. Superheterodyne receiver.

Trellis-Coded Modulation. Trellis-coded modulation, first proposed in 1982 by Ungerboeck (3), has since gained widespread popularity and usage, primarily because trellis-coded modulation achieves substantial improvements in the probability of error at a very reasonable cost in increased receiver complexity.

Trellis-coded modulation is most easily understood as a combination of convolutional channel coding and modulation, as explained next.

Consider a block of k bits that arrive at a convolutional channel coder. These k bits are mapped to a block of n bits, where $n > k$. The additional bits added are called redundant bits (because they are not necessary to communicate the original k bits). Specifically, in convolutional coding, the n output bits are generated using the k input bits as well as the previous $K - 1$ sets of k bits each.

An example of convolutional channel coding is shown in Fig. 22. Here, $k = 1$ bits arrive at the input and are mapped to an $n = 2$ bit output. The $n = 2$ bit output depends on both the current input bit and the two previous bits stored in the shift register.

The operation of the convolutional channel coder of Fig. 22 is completely characterized by the trellis diagram of Fig. 23. Here, the dots at times 0, 1, and 2 represent possible values held in the memory of the shift register. These sets are referred to as the *state.* A solid line connecting two states indicates the change in the shift-register memory if a zero is input. A dashed line indicates the change in the state if a 1 is input. The two numbers over each line indicate the output bits for the current input and state.

In trellis-coded modulation, each set of n bits output by the convolutional coder is mapped to one of $M = 2^n$ symbols by the modulator. The modulator selects the mapping very carefully, using a strategy known as *mapping by set partitioning.*

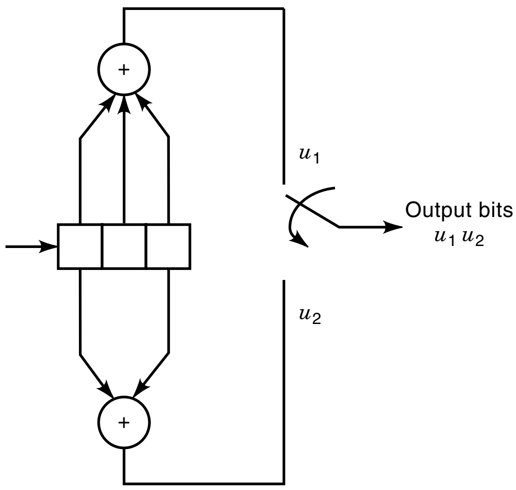

Figure 22. An example of a convolutional channel coder.

This strategy ensures that the distance between any two paths in the trellis with the same start and end node is as large as possible.

Demodulators for Trellis-Coded Modulation. This section presents the demodulator for trellis-coded modulation. First, we introduce some key notation. The modulator maps each n bits into one of the $M = 2^n$ symbols $\{s_1(t), s_2(t), \ldots, s_M(t)\}$, or, equivalently, to one of the $M = 2^n$ symbols $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$ (where vector notation indicates the representation in the orthonormal basis). Furthermore, we denote the mapping of the first n bits arriving at the modulator to a symbol in $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M\}$ as a mapping of the first n bits to $\boldsymbol{s}^1$; that of the second set of n bits, as a mapping to $\boldsymbol{s}^2$; and so on. Hence the entire mapping carried out by the modulator is summarized as a mapping of L sets of n bits to the vector $\boldsymbol{s} = (\boldsymbol{s}^1, \boldsymbol{s}^2, \ldots, \boldsymbol{s}^L)$.

The signal arriving at the data demodulator is most often modeled by $\boldsymbol{r} = \boldsymbol{s} + \boldsymbol{n}$, where $\boldsymbol{r} = (\boldsymbol{r}^1, \ldots, \boldsymbol{r}^L)$. Here, $\boldsymbol{r}^k = \boldsymbol{s}^k + \boldsymbol{n}^k$, where $\boldsymbol{n}^k$ is a vector of i.i.d. Gaussian random variables.

The data demodulator is constructed to minimize the probability of an error, that is, it is designed to output the sequence $\hat{\boldsymbol{s}}$ that achieves

$$\hat{\boldsymbol{s}} = \operatorname*{argmin}_{\boldsymbol{s}} P(\epsilon) \tag{62}$$

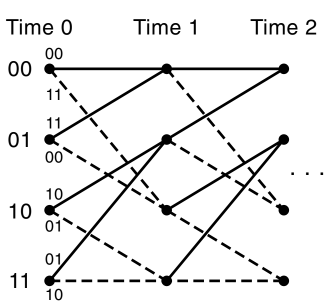

Figure 23. Trellis diagram characterizing the convolutional channel coder.

In what follows, we apply statistical and mathematical arguments to derive an equation suggesting the implementation of the data demodulator. First, an equivalent expression for $\hat{\boldsymbol{s}}$ is

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{r}) \tag{63}$$

That is, $\hat{\boldsymbol{s}}$ is the most likely sequence, given that $\boldsymbol{r}$ is received. Applying Bayes's rule results in

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{r}|\boldsymbol{s})\,p(\boldsymbol{s}) \tag{64}$$

Next, because logarithms are monotonic functions and consequently do not affect the outcome of a maximization, we express $\hat{\boldsymbol{s}}$ according to

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} \ln\, p(\boldsymbol{r}|\boldsymbol{s}) + \ln\, p(\boldsymbol{s}) \tag{65}$$

It is commonly assumed that transmitted sequences $\boldsymbol{s}$ are equally likely, that is, $p(\boldsymbol{s})$ is a constant. Applying this to the maximization generates

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} \ln\, p(\boldsymbol{r}|\boldsymbol{s}) \tag{66}$$

Using $\boldsymbol{r} = \boldsymbol{s} + \boldsymbol{n}$, the probability $p(\boldsymbol{r}|\boldsymbol{s})$ corresponds to the likelihood $p(\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s})$. Hence,

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} \ln\, p(\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s}) \tag{67}$$

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} \ln \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|\boldsymbol{r}-\boldsymbol{s}|^2}{2\sigma_n^2}\right) \tag{68}$$

where Eq. (68) results because the noise vector consists of i.i.d. Gaussian random variables. Simple mathematical manipulation, followed by a removal of terms not useful in the optimization, results in

$$\hat{\boldsymbol{s}} = \operatorname*{argmin}_{\boldsymbol{s}} |\boldsymbol{r}-\boldsymbol{s}|^2 \tag{69}$$

In words, this equation indicates that the demodulator output sequence that minimizes probability of error corresponds to the sequence of transmitted symbols closest to the received sequence $\boldsymbol{r}$. In other words, because the transmitter sequence corresponds to a path through the trellis, the demodulator output sequence corresponds to the sequence of symbols that make up the path (through the trellis) closest to the received $\boldsymbol{r}$.

We now turn our attention to implementing a demodulator that determines the path of symbols through the trellis closest to $\boldsymbol{r}$. The simplest implementation (conceptually) is to construct a device that compares every possible sequence of symbols through the trellis with the received $\boldsymbol{r}$, and selects the closest. However, this is unreasonably complex.

A common method for establishing the best path of symbols through the trellis (path closest to $\boldsymbol{r}$) is the *Viterbi algorithm* (VA). The VA is based on the rather simple idea high-

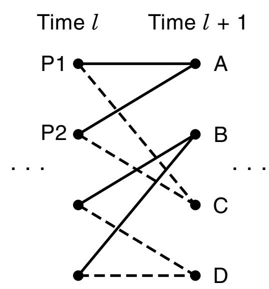

Figure 24. Explaining the key idea underlying the Viterbi algorithm.

lighted in Fig. 24. Here, the state A of the trellis at time $l + 1$ has two possible *parent* states (at time l), state P1 and state P2. The VA is based on the realization that a decision can be made at time $l + 1$ as to which parent state is the better parent for state A. Similar decisions can be made for states B, C, and D. Consequently, finding the best path through the trellis simplifies to finding the best path to each state at time 1, then the best path to each state at time 2, and so on. In the example of Fig. 24, only four paths through the trellis are maintained at any one time, one for each end state. A detailed description of the VA is available in Ref. 4, Chaps. 6, 7.

Demodulators and Synchronization: Demodulators for Data Detection of PSK in the Presence of Rapidly Changing Phase

The demodulators highlighted in the preceding section assume complete knowledge of the phase offset introduced by the channel. This was shown explicitly in Fig. 14. However, in environments such as those of many mobile communication systems, the channel phase changes are so rapid that a traditional phase-tracking device (e.g., a PLL) cannot provide an accurate estimate of the phase. In these cases, demodulators of the types shown in Section 2 are inadequate, and new demodulators are in order. This subsection briefly presents four such demodulators.

One of the first demodulators proposed for rapidly changing phase offset was conventional differential detection, also called differential PSK, or DPSK for short (4, Chap. 3). Here, the phase reference for a symbol is simply the previously received symbol. Consequently, DPSK can be used whenever the phase is constant over two or more symbols, a very mild constraint. Unfortunately, the performance of DPSK degrades substantially. Up to 3 dB may be lost in comparison with coherent detection. This degradation comes about because the phase reference (namely, the previous received symbol) is noisy.

The performance degradation of DPSK led researchers to search for alternatives for the rapidly changing phase offset. Among the first researchers to develop an alternative were Viterbi and Viterbi (5). They proposed a demodulator using a novel feedforward carrier phase estimator to track the channel phase. Their proposal demonstrates performance comparable to coherent detection, but only if the phase remains constant over a long enough period, say 20 symbols.

More recently, four groups of researchers independently generated a demodulator for phase-offset communication by extending the ideas of DPSK (6–9). Their demodulator is commonly called *multiple-symbol differential detection* (MSDD). The performance of MSDD is far superior to that of DPSK, even with the phase constant over as few as three symbols. As the number of symbols with constant phase increases from three, the performance of their demodulator tends rapidly toward the coherent. In fact, the researchers show that the performance achieved by their demodulator is optimal, in the sense of minimizing the symbol error rate, given an unknown channel phase over a block of N received symbols. However, a drawback of this scheme is its complexity. The complexity of their demodulator increases exponentially as N increases. Specifically, the complexity of MSDD, in terms of computation per decoded symbol, is in the order of $M^N N$. This limits the applicability of MSDD. Recently, however, low-complexity implementations of MSDD have been proposed (10).

Finally, a novel demodulator structure for data detection in the presence of rapidly changing phase offset was proposed in (11,12). Here, the demodulator assumes that the channel phase offset is discretized to one of eight values in the range $[0, 2\pi/M)$. With this assumption in hand, the demodulator effectively performs eight PSK demodulations, one for each possible phase value, and then uses simple processing (based on phase history) to determine which demodulation output is best. This demodulator outperforms DPSK by 1.5 dB, requires a low complexity comparable to that of the low-complexity implementation of MSDD, and, in cases of rapidly changing phase, easily outperforms MSDD.

BIBLIOGRAPHY

1. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering,* New York: Wiley, 1965.
2. L. W. Couch II, *Digital and Analog Communication Systems,* 2nd ed., New York: Macmillan, 1987.
3. G. Ungerboeck, Channel coding with multilevel/phase signals, *IEEE Trans. Inf. Theory,* **28**: 55–66, 1982.
4. B. Sklar, *Digital Communications: Fundamentals and Applications,* Englewood Cliffs, NJ: Prentice-Hall, 1988.
5. A. J. Viterbi and A. M. Viterbi, Nonlinear estimation of PSK-modulated carrier phase with application to burst digital transmission, *IEEE Trans. Inf. Theory,* **IT-29**: 543–551, 1983.
6. D. Divsalar and M. K. Simon, Multiple-symbol differential detection of MPSK, *IEEE Trans. Commun.,* **38**: 300–308, 1990.
7. S .G. Wilson, J. Freebersyser, and C. Marshall, Multi-symbol detection of M-DPSK, *GLOBECOM'89,* Dallas, TX, 1989, p. 1692–1697.
8. D. Makrakis and K. Feher, Optimal noncoherent detection of PSK signals, *Electron. Lett.,* **26**: 398–400, 1990.
9. H. Leib and S. Pasupathy, Optimal noncoherent block demodulation of differential phase shift keying (DPSK), *Arch. Electronik u. Übertragungstechnik,* **45**: 299–305, 1991.
10. K. M. Mackenthun, A fast algorithm for multiple-symbol differential detection of MPSK, *IEEE Trans. Commun.,* **42**: 1471–1474, 1994.
11. C. R. Nassar and M. R. Soleymani, Data detection of MPSK in the presence of rapidly changing carrier phase, *IEEE Trans. Veh. Technol.,* **45**: 484–490, 1996.
12. C. R. Nassar and M. R. Soleymani, Data detection of MPSK in the presence of unknown carrier phase at low complexity, *Electron. Lett.,* **31**: 945–947, 1995.

Carl R. Nassar
Colorado State University