Lecture 11-12: On-off Keying and shaping the spectrum

ENGR 76 lecture notes — May 9, 2024

Ayfer Özgür, Stanford University

In the second part of the course, we will focus on the communication problem. As formalized by Claude Shannon in his seminal paper, "A Mathematical Theory of Communication", 1948, *the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.* We will almost always assume that the message consists of a sequence of bits. You can think of these bits as produced by a source encoder, where the source itself can be discrete, continuous or continuous-time. The meaning of these bits is irrelevant to the communication problem; the goal is to simply reproduce these bits at the receiver. In this lecture, we would like to understand how we can map a sequence bits to a continuous-time transmit signal, since for many physical channels the transmitted signal will have to be a function of continuous time.

1 On-off keying

For most of our purposes in this course, we will adopt the simplest binary signaling scheme called on-off keying: a communication system transmits one of two voltages, mapping a "1" to the voltage V and mapping a "0" to voltage 0. The appropriate voltage is held steady over a fixed duration time-slot T seconds that is reserved for transmission of this bit, then moved to the appropriate voltage for the bit associated with the next time slot, and so on. This is called on-off keying and creates a transmit signal that looks like this:

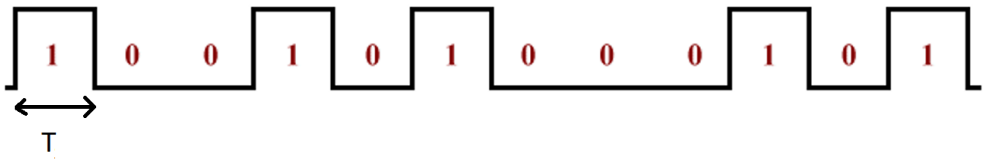
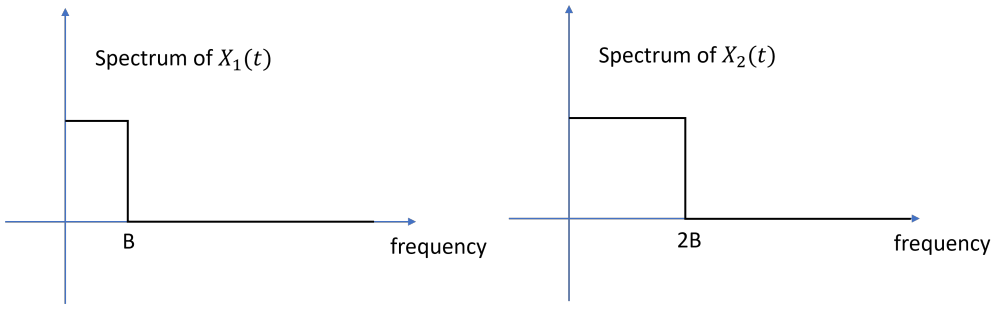


Figure 1: On-off Keying

There is one parameter we need to choose in this scheme: the slot or the symbol duration T. (We will often refer to the real voltage level corresponding to a bit as a symbol, and that is why T is called the symbol duration.) How do we choose T?

First, note that T has a direct impact on the rate of communication. The number of bits per second communicated by the above scheme is given by $1/T$ bits/second. Decreasing T to $T/2$ will double the bit rate. Higher bit rate is always desirable. Can we make T arbitrarily small?

Note that decreasing T will result in a faster-changing transmit signal and from our discussions on the Fourier Transform we know that the spectrum of a faster changing signal will contain higher frequencies. For example, assume that we have a signal $X_1(t)$ and we make it two times faster by scaling the time axis by a factor of 2. i.e. $X_2(t) = X_1(2t)$. When $X_1(t)$ is a pure sinusoid, i.e. $X_1(t) = a\sin(2\pi ft + \phi)$, $X_2(t) = a\sin(2\pi(2f)t + \phi)$, i.e. it is a sinusoid with twice the frequency. In general, since the time-scaling operation applies directly to each of the harmonic components of $X_1(t)$, we may easily conclude that it doubles the frequency of each spectral component without changing the coefficients themselves. As a result, if $X_1(t)$ has bandwidth B, $X_2(t)$ will have bandwidth $2B$ as illustrated below:



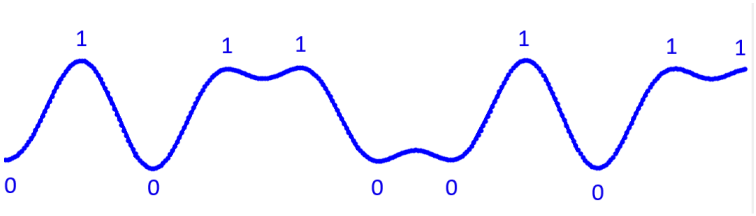
Hence, when we decrease T to increase the bit rate, we increase the bandwidth of our transmit signal. For example, decreasing T to $T/2$ doubles the bit rate but also doubles the bandwidth of the transmitted signal in Figure 1.

Why do we care about the bandwidth of our transmitted signals, and more generally their frequency spectrum? There are several reasons. First, most channels are frequency-selective meaning different frequency components of the signal experience different attenuation when the signal is transmitted over the channel. Hence, we may want to shape the spectrum of our transmitted signal so that it matches the frequency band where the channel's response is most favorable. The spectrum plays an even more critical role for wireless communication systems that have an inherently broadcast nature. If two wireless systems transmit signals with overlapping spectrum (using the same or around the same frequencies) then a wireless receiver would pick up a combination of both and be unable to sort the signals out. Allocating different bands of frequencies which do not overlap with each other to different wireless transmitters allows the signals to be separated at the receiver even if transmissions occur simultaneously. The electromagnetic spectrum is a valuable public commodity and its allocation is vested in the national government, specifically in the Federal Communication Commission in the Executive branch.

The question then is the following. Granted the use of a specific portion of the electromagnetic spectrum, how can we ensure that the frequency of our transmitted signal is limited to this designated band? More precisely, given a sequence of bits, how can we map it to a transmit signal $X(t)$ while ensuring that the spectrum of $X(t)$ is limited to a desired interval, say $[0, B]$ Hz? The sampling theorem suggests a solution. Check the discussion at the end of the Lecture 9-10 handout on the stroboscopic effect. The reconstruction formula

$$X(t) = \sum_{m \in \mathbb{Z}} X[m] \text{sinc}\left(\frac{t - mT}{T}\right) \tag{1}$$

allows us to synthesize signals with samples $X(mT) = X[m]$ and spectrum limited to the interval $[0, 1/2T]$. By taking $X[m] = V\, b_m$, where b_m denotes the m'th bit in the sequence, either "0" or "1", and $T = 1/2B$, we can generate a continuous-time signal $X(t)$ with spectrum limited to the interval $[0, B]$. Moreover, by sampling $X(t)$ at $t = mT$ for integer m at the receiver side we can read out the transmitted bit sequence from the samples $X(mT)$ (assuming there is no noise in the channel – we will come back to noise in the following lectures.) The signal $X(t)$ in eq:synthesis looks like this:



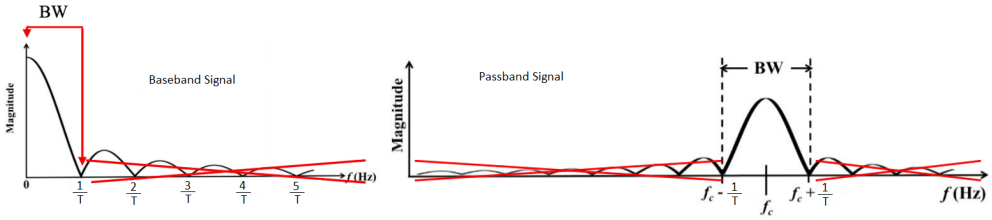
However, the ideal interpolation function sinc(t) in (1) is difficult to use in practice because it is not time-limited. Instead we will stick with the on-off keying approach depicted in Figure1. Note that on-off keying

corresponds to replacing the sinc function in (1) with the rectangular interpolation function:

$$F(t) = \begin{cases} 1, & \text{if } -1/2 \leq t \leq 1/2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

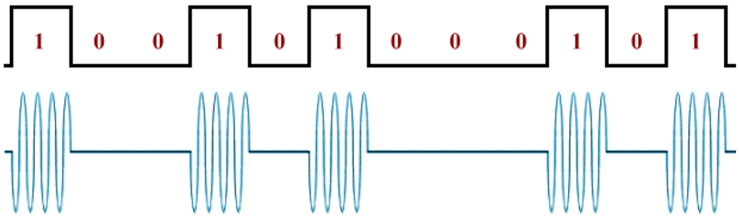
(We talked about this interpolation function when we discussed the sampling theorem.)

It turns out that the spectrum of the on-off keying signal in Figure 1 has the following spectrum:



This corresponds to the magnitude of a sinc function with the first zero crossing at $\frac{1}{T}$, beyond which the spectrum is considered small and can be ignored. So $\frac{1}{T}$ can be considered as the effective bandwidth of on-off keying. Note that for the same T, this is twice the bandwidth of the signal in (1) obtained with sinc interpolation. This is the penalty we pay for not using the ideal interpolation function.

In radio lingo, signals with spectrum limited (or approximately limited) to an interval $[0, B]$ are called baseband signals. What if the spectrum of our signal needs to be concentrated in a band $[f_c - B, f_c + B]$, where f_c is typically much larger than the bandwidth B? More radio lingo, such signals are called passband signals and f_c is called the carrier frequency. One way to generate a passband signal is to first generate the baseband signal and then multiply it with the carrier wave $\cos(2\pi f_c t)$. This leads to a transmit signal that looks like this:



The multiplication with the carrier wave is called upconversion. The upconversion process shifts the spectrum to f_c along with a mirror image copy, with half the amplitude, but double the effective bandwidth, which is now $2/T$. (See the second picture at the top of the page.) To understand why the spectrum of the passband signal looks like this, it's useful to first consider what happens to a pure sinusoid $\sin(2\pi ft)$ on upconversion - it gets converted to two sinusoids with half the amplitude and frequencies $f_c - f$ and $f_c + f$. For more general signals, the effect is simply a combination of effects on all pure sinusoids the signal comprises of. This way of generating the passband signal allows the system to operate with the simpler rectangular pulse for the processing and only perform the *upconversion* just before the transmission. In some systems, the carrier frequency is allocated dynamically to the transmitting device and so it is important to operate in the baseband signal for the most part and only do the upconversion when it's required.

1.1 Decoding at the receiver

In the on-off keying scheme above, `1` is transmitted as a sinusoid with some carrier frequency f_c, and for some symbol duration T. A `0` is simply transmitted as a zero-amplitude signal for the same symbol duration T. At the decoder, the received signal might include some noise, and we use energy detection to determine whether a given symbol is a `0` or a `1`. The energy for a symbol can be computed simply as $E_m = \int_{(m-1)T}^{mT} X^2(t)dt$

where $X(t)$ is the received signal, and the integration limits are changed to compute the energy of the received signal over the corresponding time slot. A simple energy detector that is theoretically optimal when the noise has a bell shape distribution is to determine a threshold E_T and decode to a `0` if $E_m < E_T$ and decode to a `1` otherwise. Note that this follows from the fact that the transmitted energy is zero for symbol `0` and it is a fixed non-zero value for the symbol `1`.