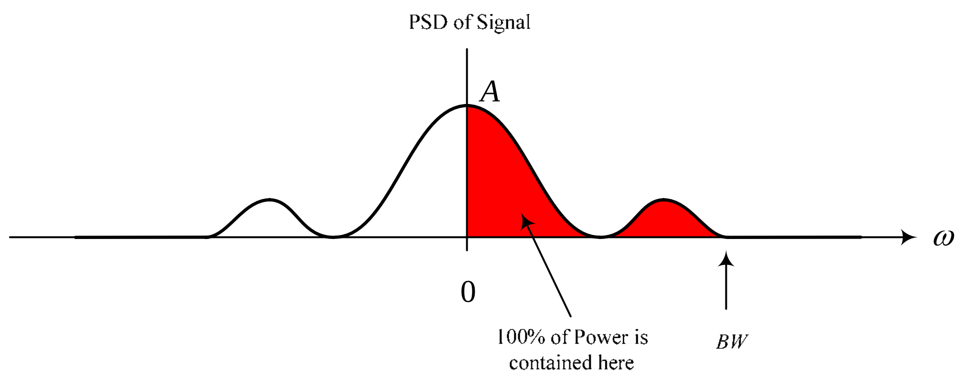
Digital Modulation Techniques

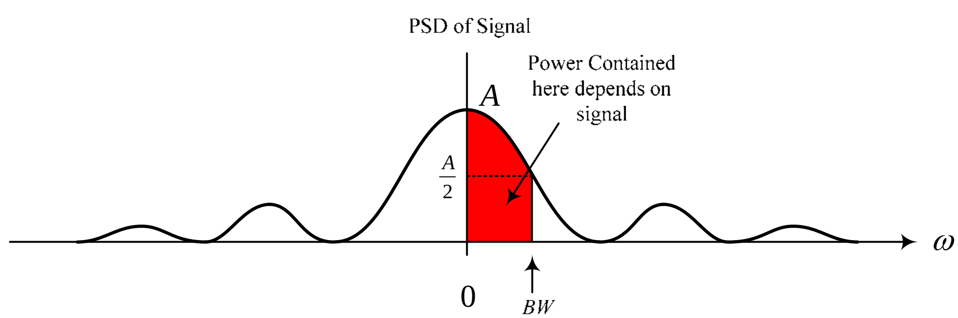
Bandwidth and Power Spectral Density of Digital Signals

Unlike analog signals, which are usually smooth signals, digital signals are composed of pulses with vertical transmissions. The fact that digital signals sometimes have vertical transmissions increases their bandwidth significantly. Compare for example the bandwidth of two baseband signals given by a sine wave with frequency f_0 and a square wave with frequency f_0. The sine wave has a single frequency component at f_0 Hz. However, the square wave has frequency components at f_0 and integer multiples of it. If we consider the bandwidth of a signal to be the minimum frequency above which the signal has no frequency components, then the sine wave will have a bandwidth of f_0 Hz because it has no frequency components above that frequency, while the square wave has an infinite bandwidth because it theoretically has frequency components that extend to infinity. There are many definitions for the bandwidth of the signal. The following definitions are those of Baseband signals. For passband signals, the same definitions exist but the bandwidth is usually double that for the corresponding baseband signal to cover both upper and lower sidebands:

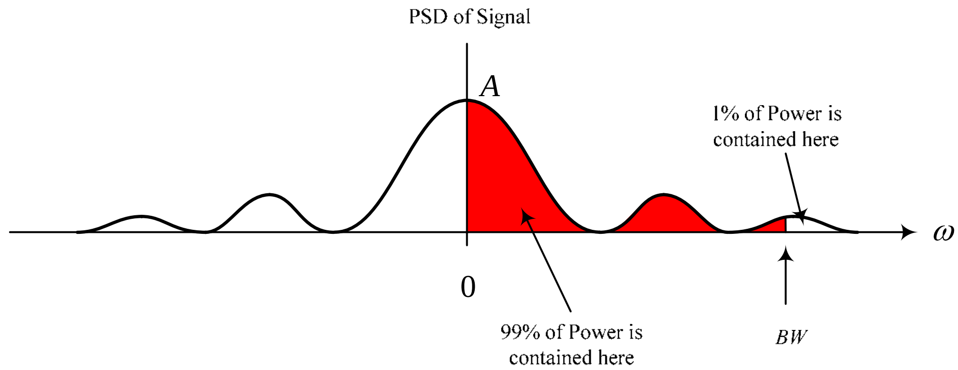
1. All Signal Components: in this definition, the signal bandwidth is the frequency of the highest component of the signal. With this definition, many signals will have infinite bandwidth such as a square wave.



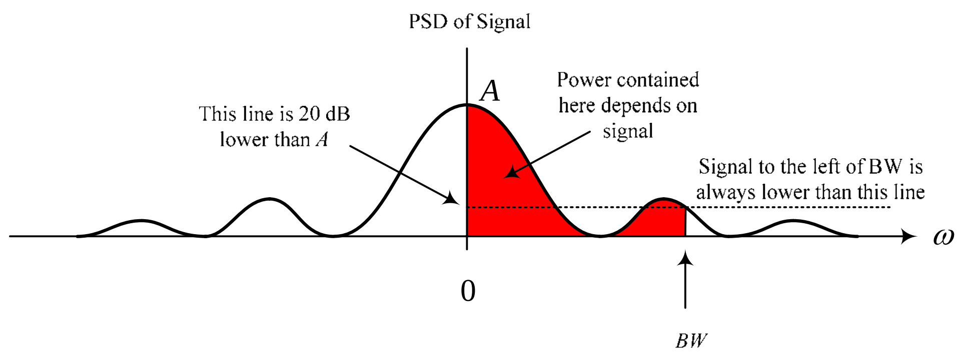
2. 3 dB: the bandwidth of the signal in this definition is the frequency at which the power of the signal drops to one half (or 3 dB) of its maximum value. This definition is generally not suitable for digital signals because a significant amount of the components of digital signals falls outside this bandwidth.



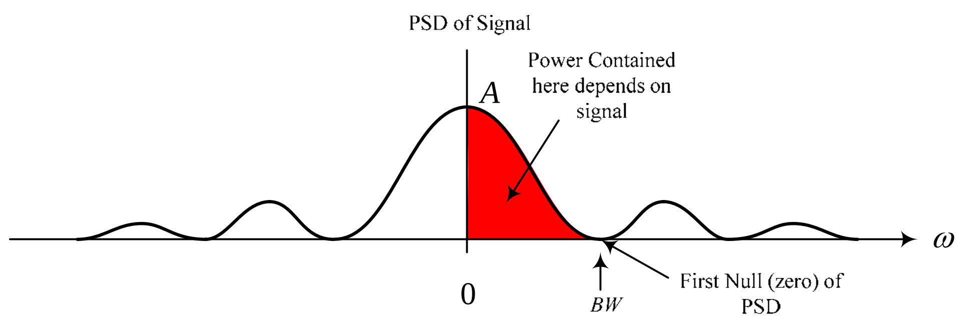
3. 99% of Signal Power: Here, the bandwidth is measured to be the frequency in which 99% of all the power of the signal is contained. So, only 1% of the power of the signal falls outside this bandwidth. Similar definitions with different power amounts also exist.



4. Signal drops below 20 dB of maximum: In this case, the bandwidth is defined to be the frequency after which all signal components have magnitude at least 20 dB below the maximum component. Similar definitions with different magnitude drops also exist.



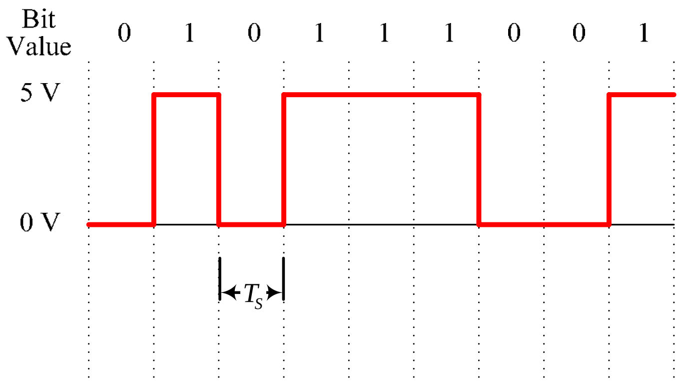
5. First Null: In this definition, the bandwidth is defined as the frequency at which the power spectral density hits zero (first null), which means the bandwidth of the signal is equal to the width of the main lube of the power spectral density. **This is the most widely used bandwidth definition for digital signals.**

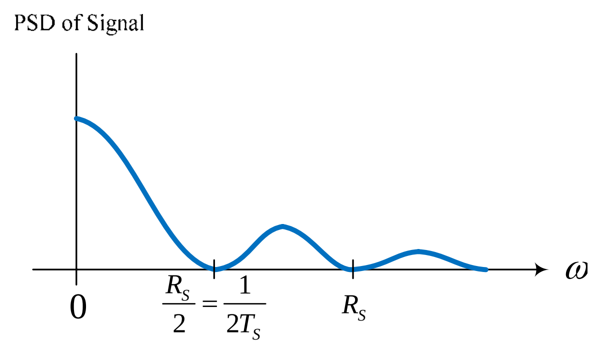


Line coding

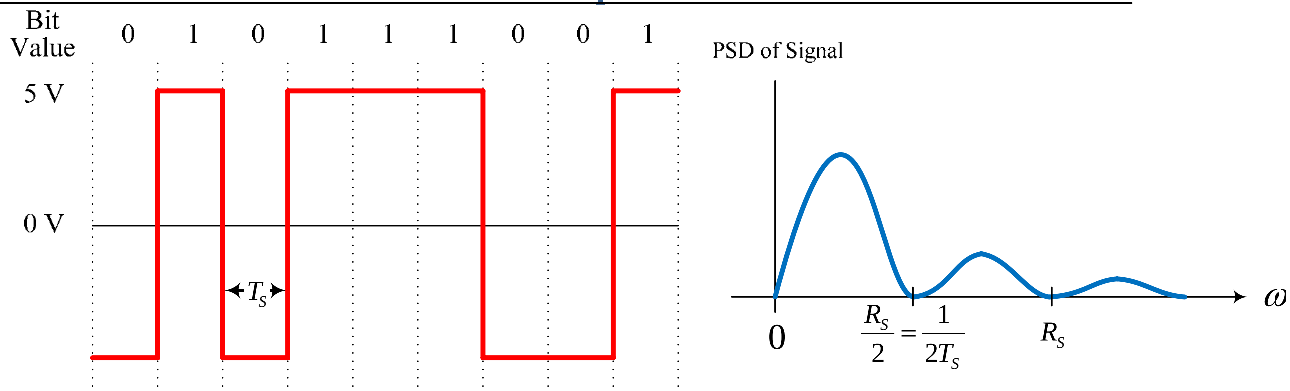
In fact, information sources that provide digital data actually provide numbers. An analog to digital converter (ADC) that converts an analog audio signal to a digital format provides sample values every 20 or 30 microseconds. The sample values must be formatted in a proper way to make them suitable for transmission through the communication channel. When each digital value is represented using a pulse for each bit (in binary communication) or a pulse for multiple bits simultaneously (for M-ary communication), the produces signal generated by stacking different pulses is called a line code. To put it in a simple way, the digital data obtained from an information source are used to generate a voltage signal that represents the information. There are different forms of line codes that can be used to represent the information. The terms Return to Zero (RZ) and Non–Return to Zero (NRZ) will be used in describing these signal. Several famous line codes are shown next along with their power spectral densities:

1. Unipolar (On–Off) Non-Return to Zero (NRZ): In this form of line codes, a bit of 1 is represented by some positive voltage (+5 volts for example) and a bit of 0 by 0 volts (justifying calling this signal On–Off). The pulses corresponding to binary 1 remain at the positive voltage for the whole duration of the bit period (it does not return to zero at any time during the bit period justifying calling this line code NRZ).

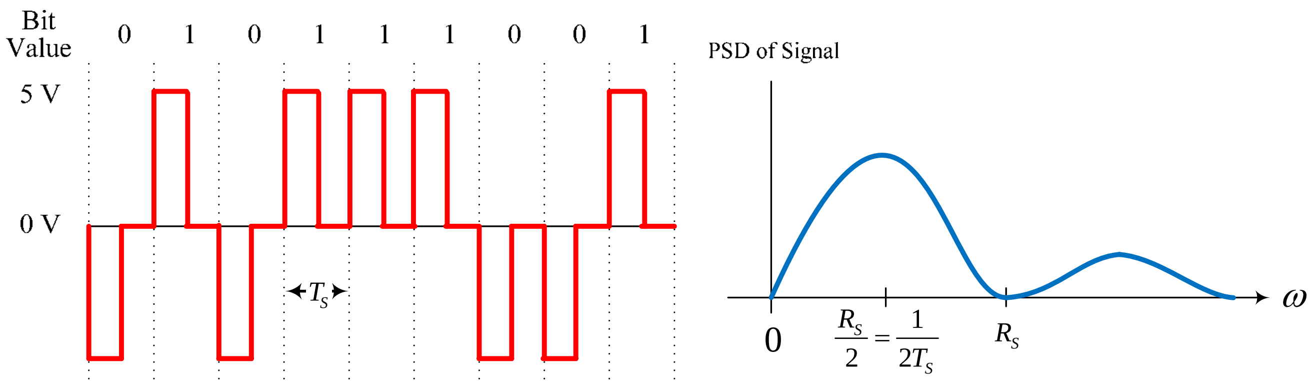




2. Bipolar Non-Return to Zero (NRZ): In this line codes, a bit of 1 is represented by some positive voltage (+5 volts for example) and a bit of 0 is represented by negative of that voltage (so it would be –5 volts). The pulses corresponding to binary 1 and binary 0 remain at the positive and negative voltages, respectively, for the whole duration of the bit period (they do not return to zero). The advantage of this line code over the On–Off (NRZ) is that it has zero–DC value when the number of binary 1's is equal to the number of binary 0's. A line code with zero–DC is desired in some applications that require that the transmitted signal to have no DC.

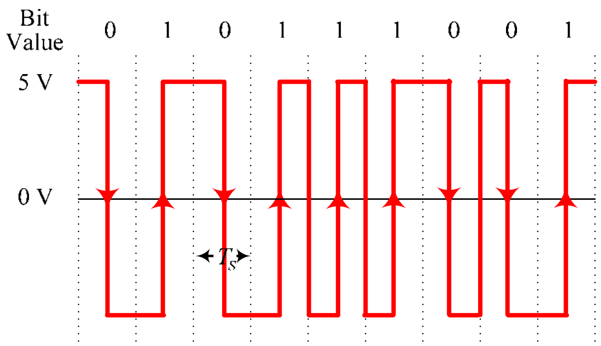


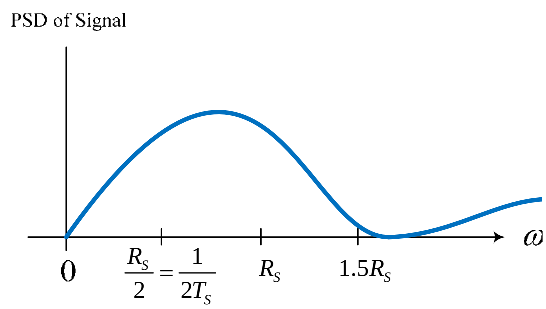
3. Bipolar Return to Zero (RZ): In this line code, a bit of 1 is represented by some positive voltage (+5 volts for example) for half of the bit period and zero in the other half of the bit period and a bit of 0 is represented by the negative of that voltage for half of the period and zero for the other half. The advantage of this line code over the previous ones is that long sequences of ones or zeros have transitions at the center of each bit and therefore bit synchronization becomes easy for long sequences of ones or zeros. Also, this line code has zero DC when the number of ones and zeros is the same.



4. Manchester: In this line code, a bit of 0 is represented by some positive voltage for the first half of the bit period and some negative voltage for the second half of the bit period. A bit of 1 is simply the negative of the zero bit so it is represented by the negative voltage for the first half of the bit period and the positive voltage for the second half of the bit. Since each of the two bits has part of it with negative voltage an the part with positive voltage, the information is not carrier in the levels but in the transition from high to low voltage or vice versa in the middle of the bit. A transition from high to low may represent a zero while a transition from low to high may represent a one. This line code is very good for insuring synchronization between the transmitter and receiver. For consecutive bits that are equal, a

transition may occur at the border of bits. These transitions are simply ignored and do not carry information.





Pulse Shaping

The use of rectangular-shape pulses to transmit digital information makes sense because they have flat tops which fit the shapes of digital signals perfectly. In addition, a rectangular pulse that extends over a bit (or symbol) period avoids interference between consecutive pulses. However, the power spectral density of rectangular-shape pulses is very wide (remember that the spectrum of a rectangular pulse is a "sinc" function). The wide spectrum of rectangular pulses means that such pulses must be transmitted over very wideband channels even for relatively low bit (or symbol) rates or else part of the transmitted signal will be filtered out by the channel and the received signal will be a distorted version of the transmitted signal. Filtering out part of the transmitted signal results in the rectangular pulses getting mixed up with preceding and succeeding pulses in what is known as Inter-Symbol Interference (ISI).

To combat ISI, the pulses that we use to transmit data must have limited bandwidth so that when transmitted over limited bandwidth channels, the complete spectrum of these signals is retained and no part of it is filtered out. This will guarantee that the signal does not change as it is transmitted through the channel. However, limiting the bandwidth of the pulses we use transmit data causes their duration in time to be infinite. A pulse with an infinite time duration (or at least very long time duration) means that each pulse extends over a very large number of bit periods. This is not necessarily bad if the pulse is designed properly. What we mean by designed properly is that each pulse needs to be equal to a constant (1 V) at the time instant of the start of the bit that that pulse represents and be zero (0 V) at all time instants of future and past bits so not to interfere with these bits at the moments that they are detected. A class of pulses called "Nyquest Pulses" satisfies all these requirements. A famous class of Nyquest pulses is called "Raised Cosine" pulses

Raised Cosine Pulses

The class of Raised Cosine pulses include the famous "sinc" function. Although the "sinc" The "sinc" function has the narrowest bandwidth of all Nyquest pulses, it decays a very slow rate that is proportional to 1/t. This means that the generation of the "sinc" pulse corresponding to a specific bit must start many bit periods before the arrival of that bit and must continue many bit periods after the arrival of that bit. This exerts a relatively large computational requirements on the system in additional to a delay before and after the transmission of data. Other Raised Cosine pulses provide a compromise between the bandwidth (they require more bandwidth than the "sinc" pulse) with the length of tails of the pulse (they have much shorter tails than the "sinc" pulse).

The general format for a raised cosine pulse is

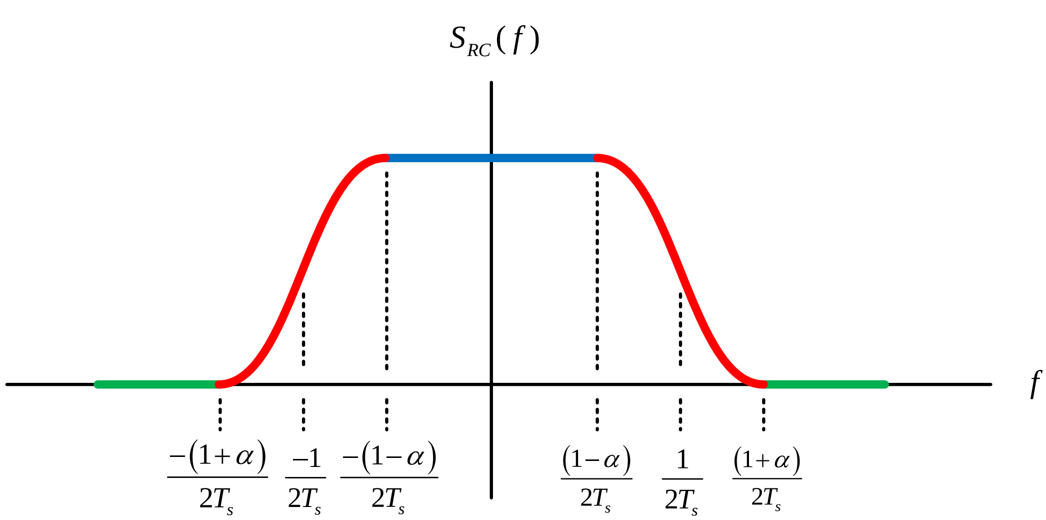
$$s_{RC}(t) = \frac{\sin\left(\frac{\pi t}{T_s}\right)}{\pi t} \cdot \frac{\cos\left(\frac{\pi \alpha t}{T_s}\right)}{1 - \left(\frac{4\alpha t}{2T_s}\right)^2}$$

where α is a parameter that provides the tradeoff between the bandwidth and tail length of the raised cosine function, and T_s is the symbol period. The first component in the raised cosine pulse shown above is a "sinc" pulse. The tails of the "sinc" pulse are attenuated further by the second component at the rate of t^2. So the raised cosine tails drop at the rate of t^3 which means that for a properly designed raised cosine, the tails die out after few (3 to 5) bit or symbol periods only. The raised cosine pulse becomes the "sinc" when the parameter $\alpha = 0$.

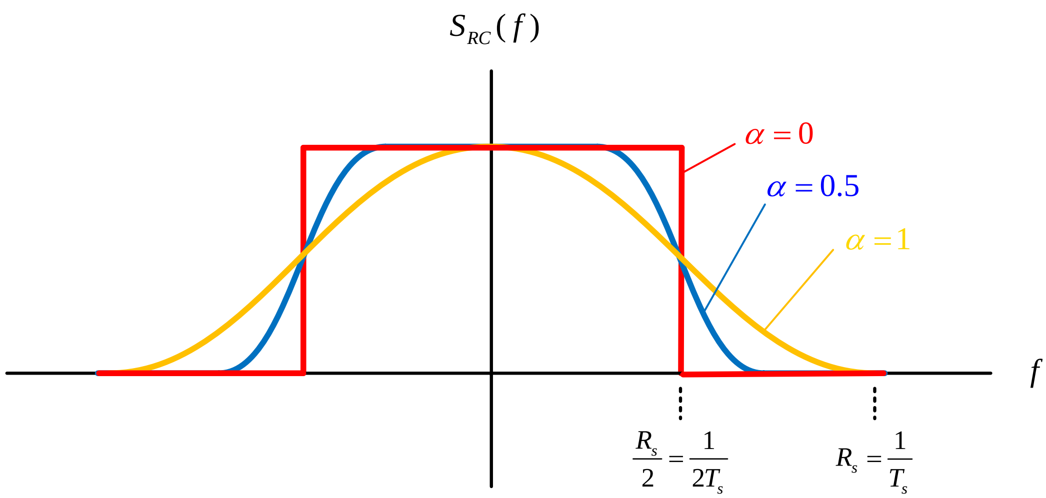
The spectrum of raised cosine pulses is

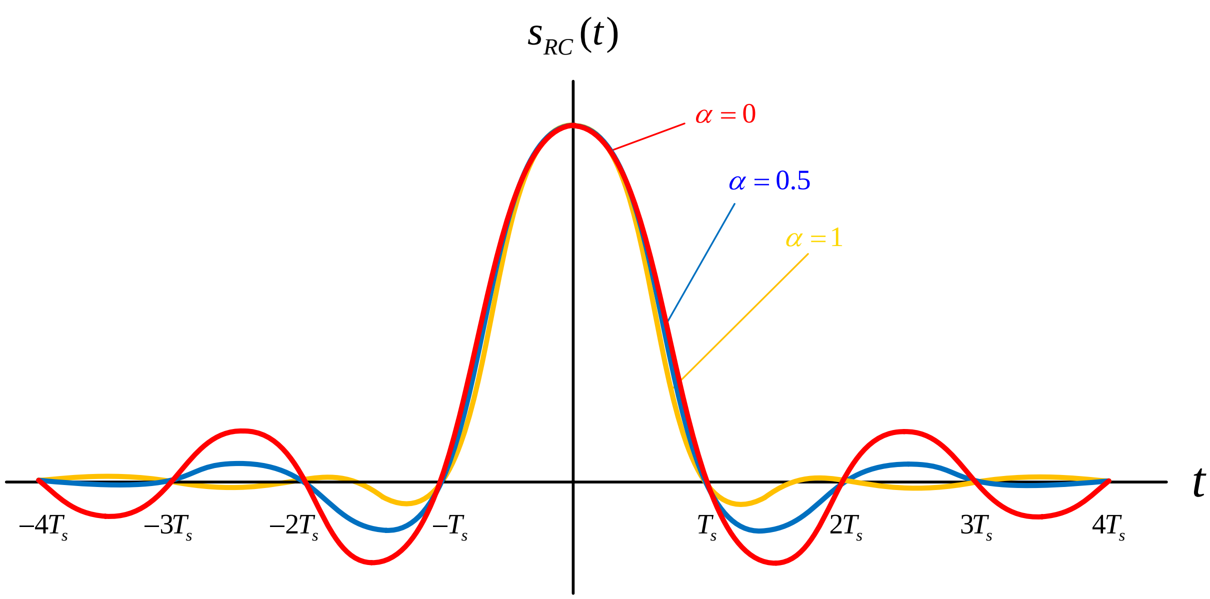
$$S_{RC}(f) = \begin{cases} 1 & 0 \le |f| \le \frac{(1-\alpha)}{2T_s} \\ \frac{1}{2}\left[1 + \cos\left(\frac{\pi\left(2T_s|f| - 1 + \alpha\right)}{2\alpha}\right)\right] & \frac{(1-\alpha)}{2T_s} \le |f| \le \frac{(1+\alpha)}{2T_s} \\ 0 & \frac{(1+\alpha)}{2T_s} < |f| \end{cases}$$

The spectrum is divided into three regions that are shown in the figure below



The spectrums and time-domain pulse shapes of several raised cosine pulses are shown below for different values of α.





For baseband transmission , the symbol rate of the transmitted data that can be transmitted using a Raised Cosine pulse is related to α and the bandwidth of the signal B by the relation

$$R_s = \frac{1}{T_s} = \frac{2B}{1+\alpha} \qquad \text{(baseband transmission)}$$

and for passband transmission, the rate is half of the above value, or

$$R_s = \frac{1}{T_s} = \frac{B}{1+\alpha} \qquad \text{(baseband transmission)}$$