Digital Communication Using GNU Radio

Prof Kumar Appaiah

Department of Electrical Engineering

Indian Institute of Technology Bombay

Week-02

Lecture-09

Welcome to this lecture on Digital Communication Using GNU radio. My name is Kumar Appiah and I belong to the department of Electrical Engineering at IIT Bombay. In this lecture, we are going to take a look at modulation, specifically digital modulation. Modulation can largely be considered as the means by which we map bits and symbols to waveforms. As we have been seeing, to perform practical communication, you need to be able to transmit and receive waveforms, but the task of converting your discrete bits or symbols into waveforms that can be deciphered at the receiver is achieved by modulation. Naturally, as in any communication system, we have constraints to deal with, the most common ones being bandwidth and power.

Bandwidth is something you have already heard of in the form of spectral constraints. Power as well because there are limits as to how much you can radiate into the air or transmit into a media and so on. One concept that we will be discussing in this lecture is the concept of degrees of freedom that concerns itself with the flexibility as to how much data you can send per unit time or with these constraints. Finally, we will be looking at the very basic aspects of signalling for digital communication.

When it comes to modulation degrees of freedom, an intuition is that this corresponds to the flexibility or the constraints on flexibility of modulation. At a very simple level, how much data can you send, let us say per unit time? This can be decided by the number of modulation levels that you can use. For example, within an interval of

$0 \text{to} T$,

you could use only two levels, 0 and 1 or you could use four levels, five levels or as many levels as you want. That corresponds with the vertical dimension and what values of amplitude you can use. But another aspect which the modulation degrees of freedom concerns itself with is the flexibility that you get in terms of given a time constraint, let us say a symbol duration of

T,

then

0 to T

you can send a symbol,

T to $2T$

you can send a symbol.

Let us say that you fix

T,

what is the maximum rate at which you can send symbols? That is how many symbols can you send per unit time? How small can you make this T so that you can send faster? These aspects are governed by the digital modulation degrees of freedom. So let us take a simple example. We have a band limited channel between

$f_c + W/2$ and $f_c - W/2$.

As you have seen in the context of the complex baseband equivalent, this can be looked upon as a baseband channel between

$-W/2$ to $W/2$.

Of course, you can keep

f_c

in the background, it is there.

But for analysing the system, you just need to look at a signal which is between

$-W/2$ to $W/2$.

Now the point is that if you look at all signals that are between

$-W/2$ to $W/2$,

you can vary this particular amplitude and get a different class of signals. But these are fundamentally band limited. The maximum frequency that you can have in your signal is bounded above by

$W/2$.

This means that you are limiting the range of waveforms and limiting the rate of data symbols that can be sent.

How? Intuitively, let us say that you have this particular region and you want to signal at the fastest rate possible or you want to send the maximum amount of data that you can send. If you look at this particular waveform which goes up like this, it has a slope which is reasonably fast rising. This particular waveform has a slope that is quite slower. Intuitively, the faster and faster you go up, the more kind of abrupt jumps that you have, the higher the frequency content that you have. In the limit, if you take something like an impulse or the unit step function, these have a large amount of frequency content in them and are not band limited.

So what do you want to conclude? Because of the fact that you are looking for band limited signals, the rate at which you can rise and fall is limited by the bandwidth of the system. In the context of electronics, you would have heard of something like slew rate for an op-amp or frequency limits of an op-amp. This is the similar concept. The moment you have a bandwidth limitation, you cannot change the amplitude very quickly. This is what is defined by the modulation degrees of freedom.

How do we make this more formal? To make this more formal, we will recall the Nyquist sampling theorem from the digital signal processing that you have already seen. Put using terminology that we have been using so far, if

$s(t)$

is a signal that is band limited to be between

$-W/2$to $W/2$,

then the Nyquist criterion says that this signal can be completely described by the samples

$s(n/W)$.

In other words, if you sample the signal

$s(t)$

at W samples per second, those samples can be used to reconstruct

$s(t)$

without any loss of information. This is the same concept that you have come across in the context of digital signal processing. To make this more formal, we will recall the

Nyquist sampling theorem that you have studied in the context of digital signal processing.

That is, if

$s(t)$

is a signal that is band limited to between

$-W/2$ to $W/2$,

then this signal can be completely described or captured by the discrete samples

$s[n]$

obtained as

$s(n/W)$.

In other words, if you sample the signal

$s(t)$

at W samples per second, then those samples completely capture all the information about the signal

$s(t)$

even in the missing portions in between. This is just a restatement of the Nyquist sampling theorem that you have seen in DSP using the notation that we have been using so far in these lectures. The interesting part is that you can reconstruct the complete

$s(t)$

without any loss of information from

$s[n]$

by using this reconstruction formula which is basically the sinc interpolation formula. That is,

$$s(t) = \sum_{n=-\infty}^{\infty} s(n/W)\, sinc(t - n/W).$$

Now, why are we looking at this? The intuition is that over a time duration of let us say

T_0,

the number of complex samples that need to be specified is

WT_0

and thus the modulation degrees of freedom is

WT_0.

What does this mean? This means that suppose that you have an interval of time, let us say

0 to T_0,

the question which we are asking is how many symbols can be sent within this time? Now because we are forced to be band limited to between

$-W/2$ to $W/2$,

Nyquist sampling theorem says if you want to reconstruct the signal, the maximum number of symbols that you can send in this interval is

WT_0.

Why? If you signal something like this, let us say that this is some wave form where you are putting these symbols, this is a symbol, this is a symbol, this is a symbol, this is a symbol, this is a symbol, this is a symbol, this is a symbol, this is the best you can do. If you try to go any faster, then you are violating the Nyquist criterion that your bandwidth is between

$-W/2$ to $W/2$.

The moment you try to signal any faster, you are going to go out of that bandwidth.

So in other words, because you are constrained in bandwidth, this signal cannot rise and fall very quickly, so the maximum number of symbols or complex symbols that you can send in a duration of

T_0 is WT_0.

Intuitively, that means that if you have a channel whose bandwidth is

W,

you can send data only at

W

samples per second based on this criterion. The next thing we must recall is the signal space description that we have already seen in the previous lectures. Suppose that you want to send one of m possible messages, m possible, let us say it is a bit sequence or symbol. As we just discussed in the context of modulation, you are going to send this message by sending a signal, a wave form.

Let us say that these

M

wave forms are titled

$s_1(t)$, $s_2(t)$up to $s_M(t)$.

It turns out that you can often have a more compact description of these wave forms using vectors. That is, if you have any M wave forms, then you can express these

M

wave forms as the linear combination of some

n

basis signals, which of course

n

is at most

M

and each signal can be represented therefore as an

n

dimensional vector. To recall, let us say that if you have this example which we looked at previously,

$\psi_1(t)$

is basically 1 between 0 and half,

$\psi_2(t)$

is 1 between half and 1, then let us say that you have this particular wave form which is half from 0 to half, 1 from 0 to 1. Let us say this is half, 1, half and 1.

This particular wave form is actually

$1/2\psi_1(t)+1/2\psi_2(t)$.

In other words, we can express this wave form, oh this should not be half, it should be

$\psi_2(t)$.

We can express this wave form as the

$$\begin{bmatrix} 1/2 \\ 1 \end{bmatrix}.$$

Of course, there is an amplitude change that I am ignoring, but this is basically a way by which you can express these signals as a linear combination of some basis signals. Therefore, it is more convenient for us to look at the signal space description because if you have an orthonormal basis of

$\psi_1(t)$, $\psi_2(t)$ up to $\psi_n(t)$,

which are orthonormal implies that the

$\int \psi_k(t)\psi_l^*(t)\,dt=1$

only when

$k=l$

and 0 otherwise, which is what this

δ_{kl}

stands for.

Then all our signals can be replaced with this kind of vectors. These vectors capture the full signals. So, the pair of these vectors along with the basis signals give you the complete descriptions of the actual

$s_1$up to s_M.

This is very convenient and we will do this often as we shall see. The signal space description emphasizes the separability of the vectors as

ψ

and the basis signals

ψ_l.

That is, the vectors themselves are something, the basis functions are something. What does this mean? You have the flexibility to now take the same vectors corresponding to signals that you design and you can change the basis to some other basis. For example, instead of using the same

$\psi_1(t)$, $\psi_2(t)$ up to $\psi_n(t)$,

you can choose some

$\psi_l{}'$up to $\psi_n{}'$

which are different. Yet, you can get similar performance and all those things, all those benefits will prevail. In other words, the signal space description says let us design our vectors separately and then use the basis signals and combine them.

So, you have flexibility because if during the design you need to change the basis, still your vector design remains the same. How? Because the vector geometry is preserved when you go to signals, that is, overlap of

$s_k(t)$and $s_l(t)$

is the same as

$s_k^H s_l$.

In other words, the inner products are preserved when you look at, go from vectors to signals. Therefore, as we just mentioned, different modulation formats can be paired with different basis signals depending on application which is why often you shall see that the most common modulation formats are repeated depending on which application you use. For example, the same format may be very useful in wireless systems, in wired systems, in underwater systems, in acoustic systems because all you need to do is to change the appropriate basis signals.

So, you have reusable design and as we shall see, predictable noise performance and you can use the same design across a large number of scenarios. The next concept that we are dealing with is the concept of linear modulation. We will initially look at linear modulation because linear modulation is very convenient and easy to understand and design and analyze especially because when you perform signaling and you have a linear time invariant system, your design and analysis becomes very convenient and in fact, in many practical systems, this assumption actually holds true. So, the way we will go about this is that at every symbol interval, we will be considering a symbol

$b[n]$,

a complex signal

$b[n]$

and we will have a transmit pulse

$g_{TX}(t)$.

For a simplified analysis, we will say the same

$g_{TX}(t)$

is being used and we are transmitting at one symbol per T seconds.

We will assume that this

$T > 1/W$,

that is you are satisfying the bandwidth criterion. So, the transmit signal is you basically take your

$b[n]$

and multiply them by

$g_{TX}(t - nT)$

and you get the resulting waveform. What is the way to visualize this? Let us say that you have this

$-2T, -T, 0, T, 2T, 3T$

and so on. How have you chosen T? You have chosen T so that you satisfy the constraint given by the modulation degrees of freedom, that is in general your

$T > 1/W$,

greater than or equal to greater than 1 upon w, that is you cannot signal any faster than W symbols per second. So, your T must be more than $1/W$.

In the interval

$[0, T]$,

you are sending

$b[0]$.

In the interval

$[T, 2T]$,

you are sending

$b[1]$.

So, how are you sending them? You are basically saying, let us say I choose

$g_{TX}(t)$

to be a rectangular signal, let us say from 0 to T and let us say that my

$b[0]=1, b[1]=3$.

That means you are sending 1 here from 0 to T and 3 from T to 2T.

So, I chose a rectangular signal. I need not choose rectangular

$g_{TX}(t)$.

I can choose

$g_{TX}(t)$

to be something else which is like, which goes even beyond 0 to T and so on, but in a sense I have to choose

$g_{TX}(t)$

such that my bandwidth constraints and modulation degrees of freedom constraints are honored. In general,

$b[n]$

is complex and said to belong to a constellation. In other words, if you take B to be a complex constellation, suppose if you take

$b[n]$

to be a complex number, typically we lay out

$b[n]$

on the complex plane and mark where the real part and imaginary part are and call it a constellation. There are many examples of constellations that we will see subsequently.

To have a visual picture of the, similar to what we saw in the previous step, if you use

$g_{TX}(t)$

to be a rectangular pulse and in fact precisely you are using

$g_{TX}(t)$

to be this particular pulse

$-T/2$ to $T/2$

and let us say that you are sending values let us say 0, 1 and 2. What you are sending over here, though I have not marked it, this is 1, this is 2. So, this is essentially 1, 2, 2, 1, 1 and you have 2, 2, 2, 1, 2, 2, 1, 1, 2, 0. This is the data that you are sending. See our discrete data was 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 0.

So, if you look at this

$g_{TX}(t)$,

1 multiplies this

$g_{TX}(t)$

and is placed over here. It is actually

$g_{TX}(t)$

shifted 5T to the left. Then 2 multiplied by

$g_{TX}(t)$

shifted 4T to the left and the next will be 2 and so on. Now, as you remember we have been discussing band limited and you know you need to be, you need to honor the constraint that you have to be within the bandwidth minus w to w and so on. Unfortunately, if you choose

$g_{TX}(t)$

to be a rectangular signal, you are definitely violating the bandwidth constraint because

$g_{TX}(t)$

is band unlimited.

So, as we have been discussing in the past, let's actually use a band limited signal. The band limited signal that we will use is

$sinc(t/T)$.

That is we choose

$g_{TX}(t)=sinc(t/T)$.

Let's actually not do the complete waveform but just place a sinc instead of the rectangle at those locations. So, if you now just place the syncs and just in the individual pulses, let's see what is happening.

This amplitude is 2. So, this sinc is essentially let's say appearing first over here it's 1, over here it is 2. The observation which you must make is that since we have chosen our sinc carefully, the amplitude of the sinc is 1 at 5T and this sinc contributes 0 at every integer multiple of T. Similarly, the amplitude of this sinc is 2 at -4T and the contribution of this sinc at any other integer multiple of T is 0. You can similarly see that 1, 1, 2, 2, 1, 1, 2. So, this essentially has the same form as the rectangular pulse, but it uses the sinc.

So, as you can see it's essentially the same thing, same values except that the waveform is different and these waveforms are individually band limited to be between

$\frac{-1}{2T}$ and $\frac{1}{2T}$.

So, if you are minus

$-W$ to W

is basically just slightly above that you are safe. Now, but this is not the actual waveform because what we are essentially plotting is something like

$b[4]\, g_{TX}(t-4T)$.

We are plotting this

$b[-3]g_{TX}(t+3T)$

and etcetera, but we should actually be plotting the sum of these.

That is what we should be plotting. So, let us actually plot the correct pulse which you should get upon doing this correctly. So, the addition of those syncs essentially gives you this. Now, let us first check whether we are safe. How do we check whether we are correct? We have to check at integer multiples of T whether the correct data is obtained. What is the value of this signal at -5T? It is 1.

What is the value at -4T? It is 2. What is the value at -3T? It is 3. At other locations it is slightly different. Unlike in the rectangular pulse which you see faintly in the background, the information is actually held the same, but in the sinc case in between the amplitude is quite different, but at the sampling points say -3T at -2T, sorry this should be actually 2, sorry it should be 2.

At -2T it is 1. At -T it is 1. At 0 it is 2 again. At T it is 2 again and so on. If you now compare with the previous, the amplitudes remain the same, but what you see in this waveform that is very striking. The first thing that you see is that you exactly hit the correct values at integer values of T and the second thing that you see is that in between it moves gradually. You don't have, let's say for example between 1 and 2 you have some number between 1 and 2, but it is risky to take this particular value let's say, but exactly when you reach -4T you get 2 again.

The waveform moves gradually. There are no abrupt jumps like in the rectangular case. This is a side effect of the fact that we are using a band limited waveform and a band limited waveform cannot contain abrupt jumps. It has to have only gradual transitions. So, in this manner you can actually signal using various different pulses and this also emphasizes the fact that

$g_{TX}(t)$

need not be limited to be between 0 and T or

$-T/2$to $T/2$.

It can be time unlimited at least in theory. Of course, in practice we have to make some compromises and we will see that eventually. So, the summary of this discussion is that whenever you have a combined waveform in this fashion that satisfies the modulation degrees of freedom by choosing a waveform that you have that is band limited to be within the bandwidth constraint, you will hit these values and trying to signal any faster will violate the bandwidth constraint. That is if you try to make your t smaller and smaller to increase the data rate, you can make it only just so small that it doesn't exceed minus w to w. That is a key thing that you should remember. Therefore, often times sinc and other such waveforms are very good candidates for signaling in a band limited manner and discussing the family of such band limited waveforms is something we will do in the next few lectures.

Let us now summarize what we have seen. In the context of digital modulation, you must remember that digital modulation essentially means that you have a discrete set of values that you want to communicate but like in any realistic system, you can only send continuous waveforms. So, the mapping between these discrete set of values to waveforms is called digital modulation and thus the task of modulation is you have some set of symbols. For each symbol what waveform do you send? That is essentially modulation. There are more complicated versions where you can take collections of symbols to represent waveforms and so on but at a very simple level, taking the symbols, converting them to waveforms is modulation. So, one task for us is to establish the correspondence between symbols and waveforms.

So, the mapping between symbols and waveforms is one thing. Another thing that we have seen is that when we have waveforms, these waveforms can be many in number but expressing them in the form of suitable basis signals is something that makes your design and analysis much simpler and reusable across multiple scenarios. Going ahead, what we will see is to see the most common digital modulation approaches that is what kind of digital modulation you have. For example, we can change the amplitude, we can change the frequency, we can change a combination, we can change the phase and we can look at all those kinds of modulation techniques and study their advantages and disadvantages and we will be looking at what are the practical baseband and passband waveforms for digital modulation. That is at a very abstract level, there are symbols which become waveforms.

How do these waveforms look and do these waveforms implicitly have the characteristic of the symbol? By looking at the waveform, can you tell what symbol it was and so on? These are some questions that we will answer in subsequent lectures. Thank you.