PDHonline Course L105 (12 PDH)

GPS Surveying

Instructor: Jan Van Sickle, P.L.S.

2012

PDH Online | PDH Center

5272 Meadow Estates Drive
Fairfax, VA 22030-6658
Phone & Fax: 703-988-0088
www.PDHonline.org
www.PDHcenter.com

An Approved Continuing Education Provider

Module 1

The General Idea

It is often said that GPS works by triangulation. Actually, it is a bit more like trilateration since distances, not angles determine positions, and to tell the truth the *tri* part could be misunderstood. Three distances aren't quite enough for good 3D positioning in GPS, it takes four. But I'm getting ahead of myself.

The distances in GPS are called ranges. And instead of being measured from control points on the earth they are measured to satellites orbiting nominally 20,183 km above the earth. In fact, the satellites really are the control points in GPS.

One often-used way of visualizing the system is to imagine a GPS satellite at the center of a huge sphere. The radius of that sphere is the distance between a GPS receiver on the earth to the satellite high above it. If that radius, that one distance, were the end of the story the receiver's position could be anywhere on that sphere. So, add another imaginary sphere to the thought experiment. Its radius is the distance from the same receiver up to a second GPS satellite. Now, the receiver might be anywhere the two spheres intersect, still a large area. But if a third sphere is added, the position of the receiver can only fall where all three come together. In other words, these three distances, three radii narrow down the receiver's position pretty tightly. Then why are four distances required?

The answer hinges on the way the distances, the ranges, are actually measured. The ranges in GPS are measured electronically, an idea well known to surveyors today. But even though the method is similar to the way distances are measured by an EDM, it is not exactly the same. In both cases, distance is a function of the speed of light, an electromagnetic signal of stable frequency, and elapsed time. While an EDM can derive all the information it requires because it's signal bounces off a reflector and returns to where it started, a GPS receiver can't. In GPS the satellites broadcast and the receivers only listen. That's why it's known as a passive system. It's like television. Millions of TV sets can tune in the signal without disrupting the broadcast, so can millions of GPS receivers listen to the satellites without affecting the GPS signal. And in both cases the signals travel one way. In GPS the signals travel from the satellite to the receiver, they don't come back, and there's the rub.

A clock in the satellite can mark the moment the signal departs, and a clock in the receiver can mark the moment it arrives. But exactly how long did it take the signal to make the trip? The accuracy of that measurement depends on how closely those two clocks are synched up. How close do they need to be? Well, it takes the GPS signal about 1/17th of a second to reach the receiver from the satellite and a clock error of 1/1000th of a second would bust the position of the receiver by 180 miles or so. That won't do, especially since GPS is supposed to be capable of millimeter level positioning. So, how closely do the clocks need to be synchronized to do that? They would need to be synchronized to near perfection because light travels 3 mm in just 0.01 nanoseconds. A nanosecond is a billionth of a second so that's 3 mm in one hundredth of a billionth of a second, and that's why you need the fourth satellite.

By tracking the fourth GPS satellite a receiver can synch up its clock with the satellites pretty well. The GPS satellites are the control points of the system, in more ways than one. Each satellite carries atomic clocks, which keep very accurate time. Running continuously they would be correct to within a second after more than 30,000 years, if anyone were around to care. And these clocks are made even more accurate by the periodic clock corrections uploaded to each satellite from the Department of Defense's facilities on the ground, more about that later. The GPS receivers, on the other hand, get along just fine with internal clocks that are reasonably accurate over short periods of time. Their clocks are not nearly as accurate, nor as expensive thank goodness, as those in the satellites. A GPS receiver has no need of an extremely stable clock because it can correct its relatively inaccurate clock by tracking the fourth GPS satellite.

Remember the intersecting spheres? If the fourth distance doesn't meet the other three precisely, something is wrong. In other words, if the receiver's clock were right, the fourth intersecting sphere would merge at exactly one point along with the other three. But if the fourth distance doesn't fit, the receiver assumes its internal clock is out of synch with the clocks in the satellites and adjusts it until all four spheres meet at the same point.

It's a pretty neat trick, but even that isn't quite enough. Just having a receiver's clock synchronized exactly to the satellite's clock can't measure a range unless there is some way for the receiver to also know when the signal leaves the satellite. The PRN code is one way to do it.

The PRN code is a pseudo-random noise code. That is a code that is designed to look like random noise, but isn't. It is a very predictable stream of ones and zeroes. In fact, the civilian version of the code - yes, there is a more accurate military version too - repeats itself every millisecond. And every GPS satellite has its own unique PRN code, so when a receiver hears one it knows which satellite it came from. Not only that, but the receiver just happens to have a duplicate of each of all 32 civilian codes, or C/A, coarse/acquisition, codes. Each one is pre-programmed into its own code generator. No matter which codes it receives, it has a copy. When a PRN code comes in from a particular satellite it will be out of synch with the duplicate of itself by the time it gets to the receiver. So the receiver pulls out the copy of that particular satellite's code and just keeps shifting it in time little by little until the two match up. When they match, the receiver can look at how much it had to shift its code to synchronize with the one coming in from the satellite and, Presto! That shift is exactly the travel time of the GPS signal.

Well, almost exactly. Actually it is very close, but not quite right. It used to be that a C/A code pseudorange points position - that's what you call that technique - was good to about ±100 meters. That was before 2000, when the intentional dithering of the satellites' clocks, called Selective Availability, was switched off by the Department of Defense. Now it can be expected to yield an accuracy of ±20 to ±40 meters. A big improvement, but still not surveying grade accuracy. This is the sort of GPS position available from a handheld receiver you can buy at a sporting goods store. So, how do you get extraordinary accuracy from GPS? Answering that question takes a moment. I'll start with GPS codes and phase measurements.

A First Look at GPS Codes and Phase Measurements

GPS codes are binary, strings of zeroes and ones, the language of computers. There are three basic codes in GPS. Two of these codes are directly involved in the measurement of ranges from the satellites to the receiver. They are the military (precise code or P) code, and the civilian (coarse/acquisition code or C/A) code. The third code, the Navigation code, is also known as the Navigation message. It carries a bunch of critical information about the GPS satellites and their signals to the receivers.

Each code has a wavelength with a different frequency. For example, the Navigation message comes into the receivers at the lowest frequency, 50 hertz (Hz). And each code is modulated onto one or both of the carrier waves, L1 and L2. Now what does **that** mean?

Frequency and Hertz

Let's talk about hertz, a measure of frequency, first. 1 hertz is a full wavelength that takes 1 second to cycle through 360 degrees. For example, the lowest sound a human can hear has a frequency of about 25 Hz, 25 cycles in one second. So, the Navigation code's 50 cycles per second, or 50 Hz, is pretty darn slow. At that rate the full 1500 bit Navigation message takes 30 seconds to send. If your computer downloaded this course at that rate, it would take hours.

The highest frequency which we can hear is about 15,000 hertz, or 15 kilohertz (kHz). But that's nothing compared to the frequencies that have been assigned to the P code and the C/A code. Their generation rate is measured in millions of cycles per second, or megahertz (MHz). The P code is generated at a rate of 10.23 MHz and the C/A code is generated at a rate 10 times slower at 1.023 MHz. And the frequency of the carrier waves, L1 and L2 are faster still. The two frequencies used by the GPS carriers are L1 at 1575.42 MHz and L2 at 1227.60 MHz.

Wavelengths

Second, let's talk about wavelengths. In both EDMs and GPS they're used to measure distances. In other words, wavelengths act like the links of an old Gunter's chain, except these links are coming out of oscillators.

The time measurement devices used in both EDM and GPS measurements are more correctly called oscillators, or frequency standards, instead of clocks. They really keep time by chopping a continuous beam of electromagnetic energy at extremely regular intervals and the result is a steady series of wavelengths. And as long as the rate of an oscillator's operation is very stable both the length and elapsed time between the beginning and end of every wavelength it produces will be very stable too.

For example, suppose you wanted to know the distance covered by a particular wavelength at a particular frequency. You can calculate it with this little formula:

$$l = \frac{c_a}{f}$$

Where: ë = the length of each complete wavelength in meters;

c_a = the speed of light corrected for atmospheric effects;

f = the frequency in hertz.

Suppose an oscillator produces a wavelength with a frequency of 30 MHz which is transmitted at the speed of light (approximately 300,000,000 meters per second, a more accurate value is 299,792,458 meters per second, but what the heck.), then:

$$l = \frac{c_a}{f}$$

$$l = \frac{300{,}000{,}000mps}{30{,}000{,}000Hz}$$

$$l = 10m$$

The wavelength is about 10 meters.

Modulation

Ok, what's modulation? I said L1 and L2 are modulated. The modulations are information that L1 and L2 carry from the satellites to the receivers, which is why they're called carrier waves. It's like the signal from your favorite radio station. When you tune into 92.5 AM, for example, you tune into a signal with a frequency of 92.5 kHz. That is the frequency of the radio station's carrier wave. But if that were all you got from them you would hear nothing but white noise. Fortunately they modulate the amplitude of their carrier, AM right? So you hear Patsy Cline singing, "Your Cheatin' Heart," instead of a steady hiss. Now, when it comes to carrier modulation you get three options: you can vary the amplitude (AM), the frequency (FM) or the phase. An AM, amplitude modulation, radio station has a carrier whose amplitude changes along with the music. It's that modulation that becomes the music you hear when the radio translates the information it receives back into sound.

As it happens, information, like the Navigation Code, the C/A code and the P code, gets on the GPS carriers, L1 and L2, by modulation too, but GPS uses phase modulation.

Phase

So, let's talk about phase. A wavelength is divided into phase angles.

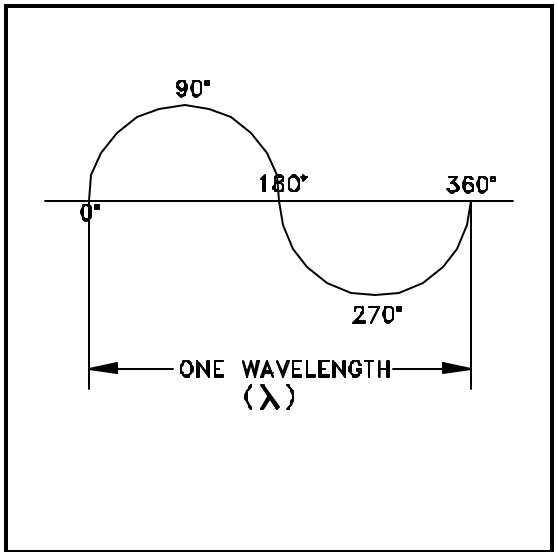


Figure 1.1 0° to 360° = 1 wavelength

The first minimum is called a 0° phase angle. The first maximum is called the 90 degree phase angle and it returns to minimum at the 180 degrees phase angle. But the wavelength isn't yet complete. It continues through 270 degrees and 360 degrees. The 360 degrees phase angle marks the end of one wavelength and the beginning of the next one. The time and distance between every other minimum, that is from the 0 degree to the 360 degree phase angle, is a wavelength and usually symbolized by the Greek letter lambda, ë.

Ok, we can use wavelengths to measure distances because we can know how long each wavelength is. So all we need to know is how many wavelengths there are from here to there

and we have the distance. But like the surveyors who used the old Gunter's link chain, one cannot depend that a particular measurement will end conveniently at the end of a complete link (or wavelength). A measurement is much more likely to end at some fractional part of a link (or wavelength). The question is, where?

The modern EDM uses a measurement wave modulated onto a carrier and those wavelengths are the links in its electromagnetic chain. But since the wavelengths of an EDM's measurement wave are not tangible, the EDM must find the fractional part electronically. It compares the phase angle of the returning signal to that of a replica of the transmitted signal to determine the phase shift. That phase shift represents the fractional part of the measurement. Both EDM and GPS systems use this principle in distance measurement.

Measuring Fractional Distance by Comparing Phase

When two waves reach exactly the same phase angle at exactly the same time, they are said to be in phase, coherent or phase locked. However, when two waves reach the same phase angle at different times, they are out of phase or phase shifted.

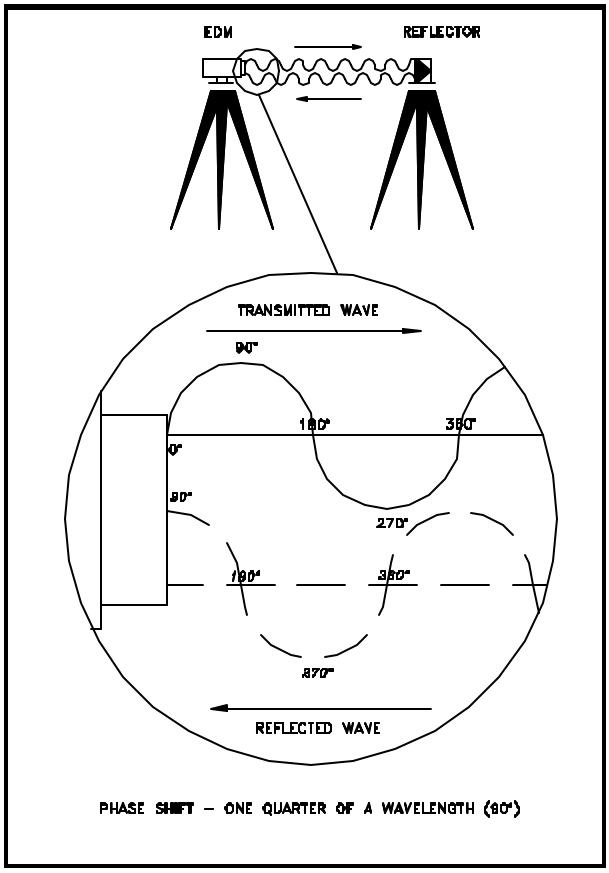


Figure 1.2 An EDM Measurement

For example Figure 1.2, shows a measurement wave from an EDM. The measurement wave shown in the dashed line has returned to the EDM from a reflector. Compared with the transmitted wave shown in the solid line, it is out of phase by one quarter of a wavelength. The distance between the EDM and the reflector, ñ, is then:

$$r = \frac{(n\lambda + d)}{2}$$

Where: n = the number of full wavelengths the modulated carrier has completed

d = the fractional part of a wavelength at the end that completes the doubled distance.

In this example, d is three-quarters of a wavelength, but how would the EDM know that?

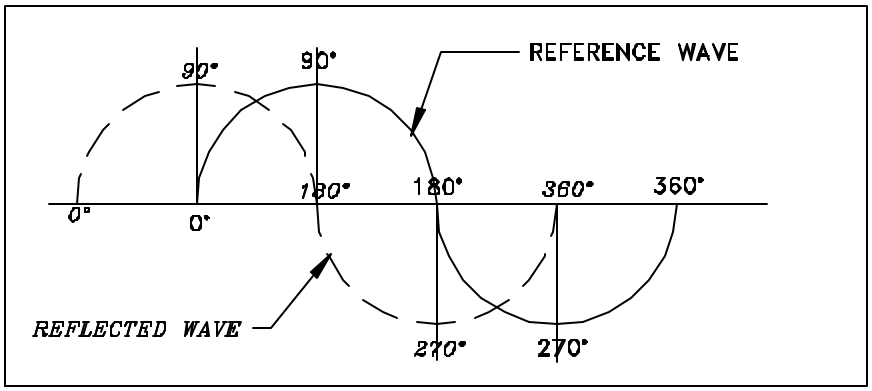


Figure 1.3 Reference and Reflected Waves

It knows because at the same time the external measurement wave, the dashed line, is sent to the reflector, the EDM keeps an identical internal reference wave, the solid line, at home in its receiver circuits. In Figure 1.3 the external measurement wave returned from the reflector is compared to the reference wave and the difference in phase between the two can be measured. It is that phase difference that reveals the fractional part of the whole distance.

An Ambiguity

While this technique discloses the fractional part of a wavelength, a problem remains: determining the number of full wavelengths of the EDM's measurement wave between the transmitter and the receiver. In other words, like a surveyor looking at the last link in a Gunter's link chain measurement, the EDM knows the very last part of the distance, but what about the number of full links or wavelengths between here and there? This ambiguity is solved in the EDM by using additional measurement waves with longer wavelengths.

For example, the meter and part of meter aspects of a measured distance are resolved by measuring the phase difference of a 10-meter wavelength. This procedure may be followed by the resolution of the tens of meters using a wavelength of 100 meters. The hundreds of meters can then be resolved with a wavelength of 1000 meters, and so on. Actually three wavelengths, 10 meters, 1,000 meters and 10,000 meters, are used in most EDMs.

For example, suppose an EDM measures a 5m distance with a 10m wavelength. It sends out a wavelength of 10m, the wave returns after completing exactly one full wavelength. Remember it had to travel 5m to the reflector and 5m back. So when it is compared with the reference wavelength it matches, it is "in phase". Ok, great, it works. Well, there's a problem. What if the distance weren't really 5m? What if it was 10m, or 20m, or 25m? In each case and many more the 10m wavelength would still come back in phase and match the reference wave.

There's ambiguity. The ambiguity stems from the EDM's inability to figure out how many full wavelengths the signal went through on its trip by just looking at the fractional part at the end.

In an EDM sending out a 100m wavelength could solve this ambiguity. Suppose the EDM measures a 25m distance starting with both a 10m wavelength and a 100m wavelength. It sends out a wavelength of 10m, the wave returns after completing exactly 5 full wavelengths. It had to travel 50m, 25m to the reflector and 25m back. So it comes back in phase, matching the reference wave. But now the EDM sends out a wavelength of 100m, and this wave returns after completing exactly ½of a full wavelength. It also had to travel 50m, 25m to the reflector and 25m back. So it comes back ½of a wavelength out of phase with the reference wave. This same principle is extended using 1000m and 10,000m measurement waves to finally confirm the distance.

Such a method is convenient for the EDM's two-way ranging system, but impossible in the one-way ranging used in GPS measurements. GPS has exactly the same problem. But GPS ranging must use an entirely different strategy for solving the ambiguity problem because the satellites broadcast only two carriers of constant wavelengths, in one direction: from the satellites to the receivers. Unlike an EDM measurement the wavelengths of these carriers in GPS cannot be changed to resolve the number of cycles between transmission and reception. This problem is solved another way in GPS, more about that process later.

With frequency, wavelength, modulation, and phase defined, we can look at how the GPS codes get on the L1 and L2 carriers.

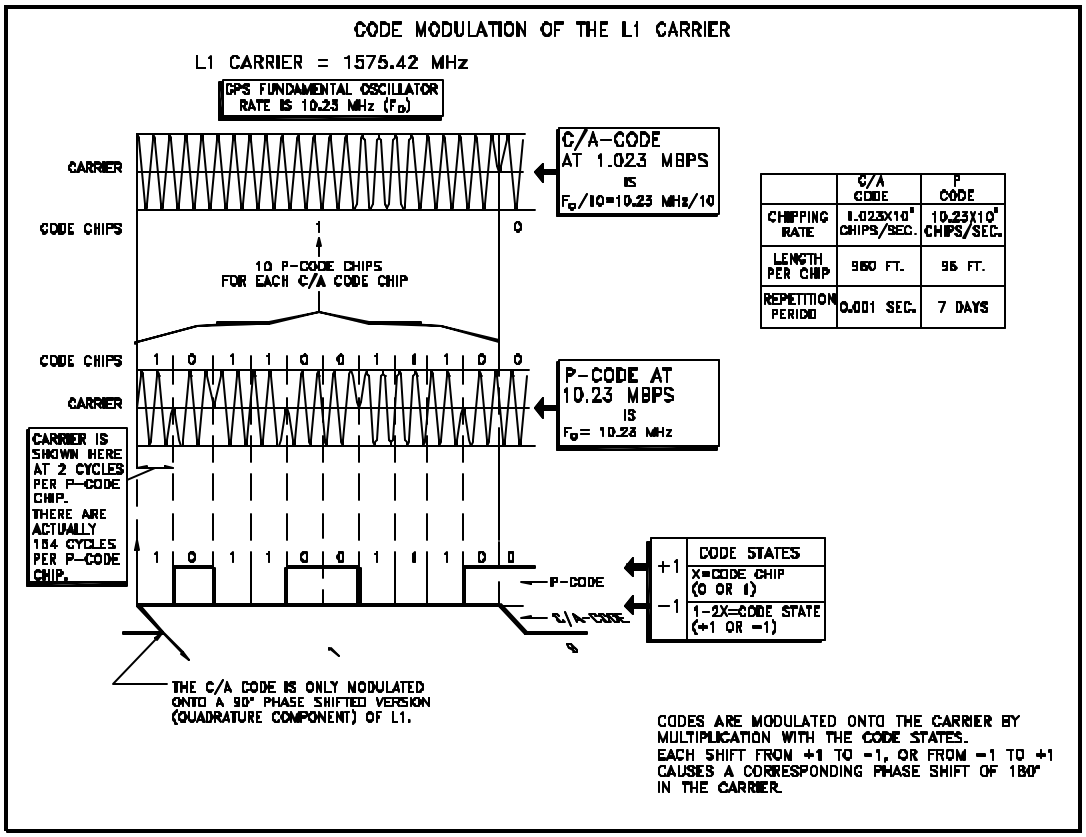
Phase Modulation

The GPS measurement codes could have been modulated onto the carriers L1 and L2 in a variety of ways. They get on by phase modulation. So, while the frequency and amplitude of the L1 and L2 carrier waves don't ever change, there are instantaneous 180° changes in their phase. It is these changes in phase, modulations from zero to one and from one to zero, that make codes. You can see them in Figure 1.4. Each shift from zero to one and from one to zero in the code is accompanied by a corresponding change in the phase of the carrier.

The rates of all of the components of GPS signals are multiples of the fundamental clock rates of the oscillators, 10.23 MHz. This rate is symbolized F_o. For example, the GPS carriers are 154 times F_o, or 1575.42 MHz, and 120 times F_o, or 1227.60 MHz, L1 and L2 respectively.

The codes are also based on F_o. And 10.23 code chips of the P code, zeros or ones, occur every microsecond. In other words, the chipping rate of the P code is 10.23 million bits per second, 10.23 MBPS, exactly the same as F_o, 10.23 MHz.

The chipping rate of the C/A code is 10 times slower than the P code, a tenth of F_o, and 1.023 MBPS. Ten P code chips occur in the time it takes to generate one C/A code chip, allowing P code derived pseudoranges to be much more precise; this is one reason the C/A code is known as the coarse/acquisition code.



	C/A CODE	P CODE
CHIPPING RATE	1.023X10^6 CHIPS/SEC.	10.23X10^6 CHIPS/SEC.
LENGTH PER CHIP	960 FT.	95 FT.
REPETITION PERIOD	0.001 SEC.	7 DAYS

Figure 1.4 Code Modulation of the L1 Carrier

Even though both codes are broadcast on L1, they are distinguishable from one another by their transmission in quadrature. That means that the C/A code modulation on the L1 carrier is phase shifted 90° from the P code modulation on the same carrier.

Pseudorange Point Positioning

Remember the example of the four imaginary spheres? Well, this idea of having a replica to match to an incoming signal is popular. Not only do EDMs use it; the GPS receiver uses it too. A GPS receiver pulls out the copy a particular satellite's code and just keeps shifting it in time little by little until the two match up in order to find the transmission time of the signal.

Actually lining up the code from the satellite with the replica in the GPS receiver is called autocorrelation, and depends on the transformation of code chips into code states. The formula used to derive code states (+1 and -1) from code chips (0 and 1) is:

$$code\ state = 1 - 2x$$

where x is the code chip value. For example, a normal code state is +1, and corresponds to a code chip value of 0. A mirror code state is -1, and corresponds to a code chip value of 1.

The function of these code states can be illustrated by asking two questions:

First, if a tracking loop of 10 code states generated in a receiver does not match 10 code states received from the satellite, how does the receiver know? In that case, the sum of the products of each of the receiver's 10 code states, with each of the 10 from the satellite, when divided by 10, does not equal 1.

Secondly, how does the receiver know when a tracking loop of 10 replica code states does match 10 code states from the satellite? In that case the sum of the products of each code state of the receiver's replica 10, with each of the 10 from the satellite, divided by 10, is exactly 1.

The autocorrelation function is:

$$\frac{1}{N}\int_0^T X(t)^* X(t-\tau)dt = \frac{1}{N}\sum_{i=1}^{N} X_i^* X_{i-j}$$

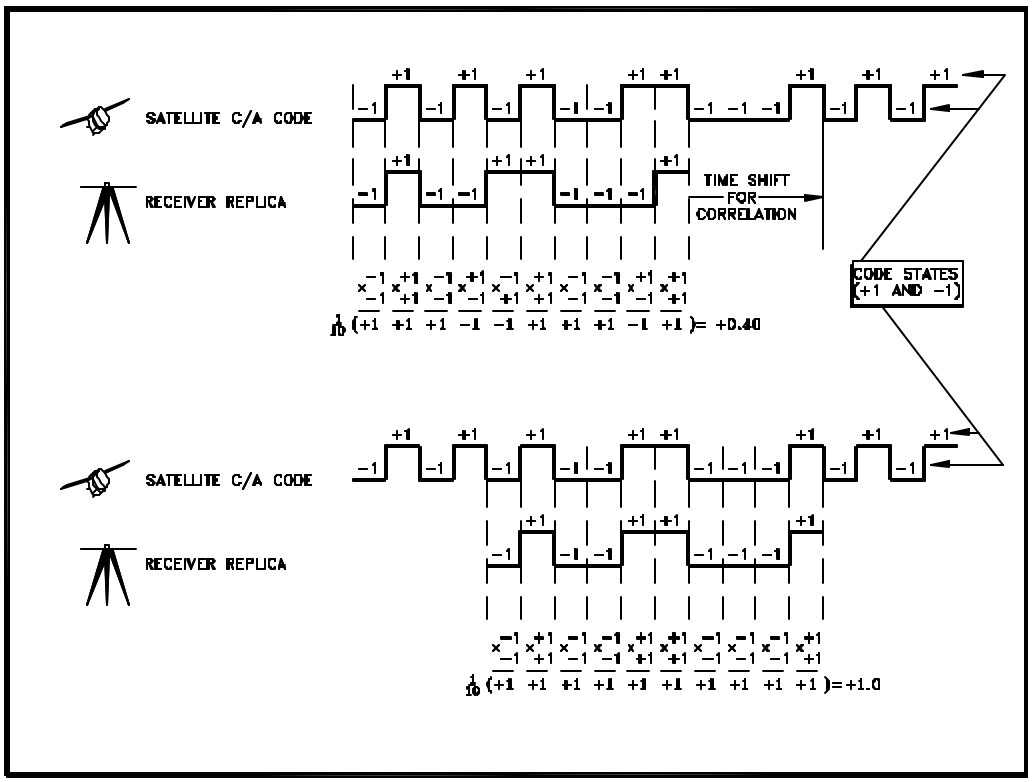


Figure 1.5 Code Correlation

$$\frac{1}{10}\sum_{i=1}^{10} X_i * X_i = \frac{1}{10}(+1+1+1-1-1+1+1+1-1+1) = +0.40$$

In Figure 1.5, before the code from the satellite and the replica from the receiver are matched:

the sum of the products of the code states is not 1.

Following the correlation of the two codes:

$$\frac{1}{10}\sum_{i=1}^{10} X_i * X_i = \frac{1}{10}(+1+1+1+1+1+1+1+1+1+1) = +1.0$$

the sum of the code states is exactly 1, and the receiver's replica code fits the code from the satellite like a key fits a lock.

That receiver is finding a pseudorange to that particular satellite when it does that, or as some folks say it is observing code-phase. So a pseudorange observable is based on a time shift.

The Limitation of Pseudorange Point Positioning

Pseudorange point positioning is the technique employed by inexpensive handheld GPS receivers. As I said at the top of this module it can yield positional accuracy of about ±20 to ±40 meters. The reasons for the relative inaccuracy are instructive because they affect all GPS. And it can be illustrated by the process of setting a watch from a time signal heard over a telephone. Imagine that a recorded voice said, "The time at the tone is 3 hours and 59 minutes, beep." If you set your watch the instant you heard that beep your watch would be wrong. Because the beep was broadcast at exactly 3 hours and 59 minutes, the moment you

heard it was later by precisely the amount of time it took the beep to travel to you through the telephone lines. In fact, you could measure the length of that telephone line, if you knew how long the beep was delayed, you could multiply by the speed of the light, and there you are. That's what the C/A code allows the receiver to do. It can calculate the range by calculating the approximately $1/17^{th}$ second delay from the moment the signal left the satellite to the moment it arrived by sliding the code from the satellite until it fits the replica code. But why is the result called a pseudorange, a false range?

Ok, let's carry this telephone analogy farther. Instead of just setting your watch, suppose you decided to try to measure the length of the telephone line back to the master clock. So you get a clock just like the one they have. You take it down there and you synch them up. When you get home you call and get the voice, "The time at the tone is 9 hours and 33 minutes, beep." All right, but this time you can tell right away that the beep is late. That time has come and gone when you hear the beep. But now you can measure the difference between what your clock shows and the beeps from the master clock. All you have to do is set your replica clock back little by little until it matches the beeps and note how big the delay has to be. You multiply the delay by the speed of light; there's the distance. But what if your clock weren't perfect? What if your replica clock strayed a little after you synched it up and drove home? What if the master clock weren't perfect and its rate strayed a little too? What if the telephone lines weren't perfect and the beep got delayed a little bit? You get the idea. If the clocks can't be perfectly synchronized and if the propagation of the signal isn't perfect, then the calculated distance must be false. That's why it's called a pseudorange.

A Pseudorange Equation

Ok, here's a formula presented by Langley in 1993 that neatly summarizes the errors that prevent a pseudorange from really being the true distance to a satellite.

$$p = \rho + c(dt - dT) + d_{ion} + d_{trop} + \varepsilon_p$$

Where: p = the pseudorange measurement

ñ = the true range.

c = the speed of light

dt = the satellite clock error

dT = the receiver clock error

d_{ion} = ionospheric delay

d_{trop} = tropospheric delay

ε_p = multipath, receiver noise and etc.

Please note that the pseudorange, p, and the true range, ñ, cannot be made equivalent, without consideration of clock offsets, atmospheric effects and other biases that are inevitably present.

In the next module, we'll talk about exactly where these errors come from and how they are mitigated to improve the accuracy of GPS positioning.