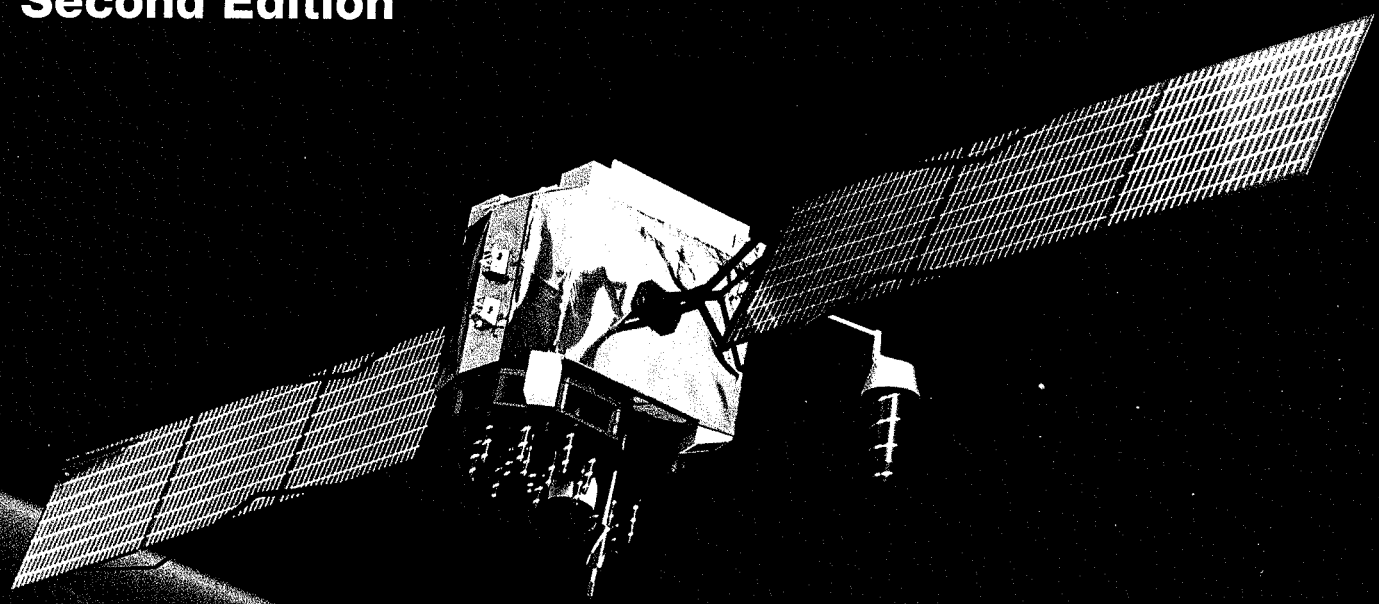
GLOBAL POSITIONING SYSTEM

Signals, Measurements, and Performance

Second Edition



Pratap Misra
Per Enge

GLOBAL POSITIONING SYSTEM

Signals, Measurements, and Performance

Second Edition

Pratap Misra
Lincoln Laboratory
Massachusetts Institute of Technology

Per Enge
Department of Aeronautics and Astronautics
Stanford University

❀
Ganga-Jamuna Press

To Prakash,
Hülya, Ray, and Tara

To Elaine and Nick

ISBN: 0-9709544-1-7

Designed by Randall Warniers

Printed in the United States of America

Cover illustration courtesy of The Boeing Company

Ganga-Jamuna Press
P.O. Box 692
Lincoln, Massachusetts 01773
GPStextbook@G-JPress.com

Preface

GPS in many ways is like the Internet. Both are gifts of the U.S. Department of Defense to the civil world. Both continue to transform the way we do ordinary, everyday things as individuals and society, delivering wide-ranging economic and social benefits far beyond anything their designers could have dreamed of. And, of course, the Internet now plays an important role in newer GPS applications.

The value of GPS to the military, its main sponsor, developer, and keeper, is illustrated by noting that GPS receiver will become the most widely used military radio (albeit one way), to be carried by every aircraft, ship, tank, truck, soldier, and bomb, even an artillery shell. In civil use, of greater interest in this book, GPS has already become indispensable to millions of car and truck drivers, sailors, pilots, sportsmen, and outdoorsmen. In a less visible role, GPS has become the source of precise time for telecommunications, banking, power generation, and the Internet. A one-chip GPS receiver will soon be built into every cell phone and PDA.

We are gratified by the response of teachers and students of GPS to our First Edition, which appeared in 2001. In this Second Edition, we have corrected, updated, revised, and expanded our treatment of the GPS signals, measurements, and performance, but the basic objective remains unchanged: to offer an introductory textbook for students of engineering and applied sciences and a self-study guide for practicing engineers.

Some background in linear algebra, probability theory, signal theory, and linear system theory is required for this book. An ideal preparation would be an upper-division course in each of these areas. Vector-matrix notation is used throughout. The concepts of mean, standard deviation, covariance, and correlation of random variables are used without explanation. We do, however, review the results from signal theory and system theory before beginning our discussion of the GPS signals and receivers.

The book is divided into four parts. Parts I and II have been revised and expanded but have retained their structure from the First Edition. Part I, now consisting of four chapters, introduces the basic framework of a global satellite navigation system, including coordinate frames, time references, and orbits, and provides an overview of GPS, GLONASS, and Galileo. The three chapters of Part II deal with the measurements, errors therein, and estimation of position, velocity, and time. The biggest change in this edition is in our discussion of the GPS signals and receivers. The treatment of both topics has been expanded significantly. Part III of the First Edition, with its three chapters, has grown into Parts III and IV, with three chapters each.

The treatment of GPS signals and receivers has been expanded in both breadth and depth.

The main results from signal theory and system theory useful to our discussion have been brought together in a review chapter (Chapter 8). An entire chapter is now devoted to the structure of the GPS signals, including the new binary offset carrier modulation (Chapter 9). A chapter each is devoted to signal acquisition and tracking (Chapters 11 and 12). A new chapter has been added for discussion of techniques to cope with signal obstruction and interference (Chapter 13).

We believe this book offers an instructor several options for a one-semester senior- or graduate-level course. The two simplest options are: Chapters 1–7 for a course with emphasis on positioning and navigation algorithms, and Chapters 1, 2, 8–13 for a course with emphasis on signals and receivers. In either case, we hope the instructor would apply and extend the basic ideas of positioning algorithms or receiver design to a specialized application, e.g., geodesy, aviation, land navigation, or assisted GPS.

The CD accompanying the Second Edition has been brought up to date. The GPS data sets used in homework problems and simple MATLAB tools (M-files) to manipulate these data sets are essentially unchanged from the First Edition. Even though the Internet now offers access to a wide variety of raw and processed GPS data collected around the world, we have retained the data sets on the CD for ease of use. The GPS-related documents have been updated.

The First Edition showed signs of smoldering resentment against Selective Availability (SA), the U.S Government policy throughout the 1990s of purposefully degrading the civil signals. Now with SA receding in memory and civil users of GPS feeling indulged, there seems no reason for side-by-side comparison of what the system was capable of and what it was delivering to the civil users in the late 1990s while the First Edition was taking shape. We have, however, retained for the historical record a couple of figures in Chapters 2 and 5 showing effects of SA so a newcomer would know what the fuss was about.

There are now many websites useful to GPS engineers and users, and we have included references to several in this edition. By definition, a website is a changeable entity, sometimes disappearing without notice. We have included sites which we thought would be maintained during the useful life of this book.

We are grateful to the instructors who used the First Edition for classroom instruction and offered ideas for improvement. We thank especially Professors Penina Axelrad and Dennis Akos (both of the University of Colorado at Boulder), Professor Kai Borre (Aalborg University), Professors Michael Braasch and Chris Bartone (both of Ohio University), Professors Elizabeth Cannon and Gérard Lachapelle (both of University of Calgary), Professor Demoz Gebre-Egziabher (University of Minnesota), Professor Richard Langley (University of New Brunswick), Professor Jade Morton (Miami University), Professor Boris Pervan (Illinois Institute of Technology), Dr. Marvin May (Penn State), Dr. Albert Paradis (MITRE Corporation), Professor John Racquet (AFIT), and Professor Christian Tiberius (TU Delft) for their comments and suggestions. We also owe a debt of gratitude to Professor Akio Yasuda and his research team at the Tokyo University of Marine Science and Technology for translating the First Edition into Japanese.

We are indebted to Grace Gao and Fiona Walter, both of Stanford University, for reviewing the manuscript. We thank the following colleagues for reviewing one or more chapters: Dr. Frank Bauregger (Novariant), Amy Englehart, Dr. Jacob Sauer, and Dr. Jay Sklar (all of MIT Lincoln Laboratory), Professor David Powell (Stanford University), and Dr. Frank van Diggelen (Global Locate). several current and former graduate students at Stanford and MIT have made significant contributions to this book. We thank Dr. Keith Alter, Alan Chen, Seebany Dutta-Barua, Michael Koenig, Dr. Guttorm Opshaug, Dr. Gang Xie, and Alan Zorn (all of Stanford), and Sean Bednarz and Tom Temple (both of MIT).

The authors will be grateful to the readers who point out errors or offer ideas for improvement <GPStextbook@G-JPress.com>.

15 March 2006

Pratap Misra
Per Enge

Contents

PART I

Fundamentals

Navigation is the art and science of charting a course from point A to point B and staying on that course. The subject has a long and fascinating history, what with some of our forebears crossing vast oceans guided only by the stars. But these were brave souls for the stars couldn't be counted upon to be visible. The technology of the twentieth century has now solved this problem *nearly* completely by placing artificial stars in the sky. These stars shine all the time, radiating extraordinarily faint radio signals. What the signals lack in raw power, however, they make up in cleverness of design, and provide far more information than the sailors of old ever got from the stars on the clearest of nights.

The NAVSTAR Global Positioning System (GPS) is the first of this new breed of global navigation satellite systems to become operational. Buy a $100 GPS receiver and a map, and you wouldn't be lost as long as you have a clear view of the sky. Or, buy a pair of more capable receivers for about $5000 each and, with a careful analysis of the measurements, you would be able to tell if the earth under your feet moved a few millimeters while you weren't looking.

In Part I we begin our study of the engineering principles of GPS with a discussion of the fundamentals. Chapter 1 is brief survey of the history of navigation from watching the heavens to measuring radio signals. Chapter 2 offers an overview of GPS, describing briefly the system, signals, receivers, performance, and applications. Chapter 3 gives a similar overview of the global navigation satellite systems now under development, including the GPS upgrade. Chapter 4 examines three topics basic to GPS: coordinate frames in which to specify a position, time scales, and satellite orbits.

Chapter 1
Introduction

Magellan set out on his voyage in 1519 to circumnavigate the globe equipped, according to Bowditch (1802), with "sea charts, a terrestrial globe, wooden and metal theodolites, wooden and wood-and-bronze quadrants, compasses, magnetic needles, hour glasses and 'timepieces,' and a log to be towed astern." With these instruments, and great personal skills, he could estimate the ship's speed, direction, and latitude, but not the longitude. It took another 250 years for the sailors to be able to determine their longitude at sea. And, two hundred years later yet, at the end of the 20th century, extraordinarily accurate estimates of position, velocity, and time became available to all instantaneously, continuously, inexpensively, and effortlessly, thanks to GPS.

GPS represents the fruition of several technologies, which matured and came together in the second half of the 20th century. In particular, stable space-borne platforms, ultra-stable atomic frequency standards, spread spectrum signaling, and microelectronics are the key developments in the realization and success of GPS. These technologies have been used to implement an ancient idea for positioning: *trilateration*, or position determination by measuring distances from known points.

In this chapter, we attempt to place GPS in the context of the navigation systems that preceded it. Section 1.1 offers a breezy tour of the developments spanning a thousand years and

culminating in inertial navigation and, of greater interest in this book, radionavigation. Section 1.2 deals with the principles of radionavigation. The terrestrial and satellite radionavigation systems developed during and after World War II are described in Section 1.3.

1.1 A Brief History of Navigation

We begin with a brief discussion of three areas, which, at first, may appear unrelated. In fact, these three fields are at the heart of positioning and navigation systems, ancient and modern.

- Geodesy, the study of the size and shape of the earth and mapping of its surface.
- Timekeeping (or, horology, the art and science of measuring time).
- Astronomy (and, in the 20th century, astronautics, the science and technology of space flight).

In order to navigate from point A to point B smartly, it's essential to know the position of each in some form, ideally represented on a map. It took about two thousand years and the imperatives of commerce and colonization in the sixteenth and seventeenth centuries, and hot and cold wars in the twentieth century, to accomplish this goal. We can now specify the position of a point on the earth with millimeter-level accuracy. It took many incremental steps and giant leaps in the development of mathematics, astronomy, and clock making to reach this point. It is a long and fascinating story filled with memorable characters and great human drama, and told masterfully by Williams (1992). We will simply touch upon a few main ideas in this chapter and get on with the story of GPS.

The early navigators and mapmakers relied on celestial observations to determine both time and positions on earth. This reliance provided the impetus for study of the basic laws that govern the motion of the stars and planets. It is also interesting to note that navigation provided the impetus for development of accurate clocks in the seventeenth and eighteenth centuries. The roles were eventually reversed with extraordinary developments in timekeeping technology in the twentieth century, fueled mainly by the requirements of the burgeoning telecommunication industry, which, in turn, led to development of a new class of radionavigation systems, culminating in GPS.

In order to specify the position of a point unambiguously, we need a reference system or coordinate frame. This idea is not new. Two thousand years ago, the Greeks knew the earth to be spherical. They apparently had a good idea of its size, and understood the concept of representing the position of a point on earth as so many degrees north or south of the equator, and so many degrees east or west of some chosen meridian (see Figure 1.1). This coordinate system, with some subtle modifications, discussed in Section 4.1, is still in use.

The latitude of a point turned out to be easy both to define and measure. The equator was a natural origin, and the two poles, equidistant from the equator, were at 90° north and south. The latitude could be measured from the elevation of the Pole star (see Figure 1.2), or the sun at its highest. The longitude, however, turned out to be much trickier to deal with. There was no natural, fixed line from which to define longitude. But this was not a hindrance as a group of users could adopt a meridian convenient to them as the reference meridian. (A meridian plane is defined as a plane containing the center of the earth, the pole of rotation, and the local vertical.) The measurement of longitude turned out to be a challenge which was to occupy

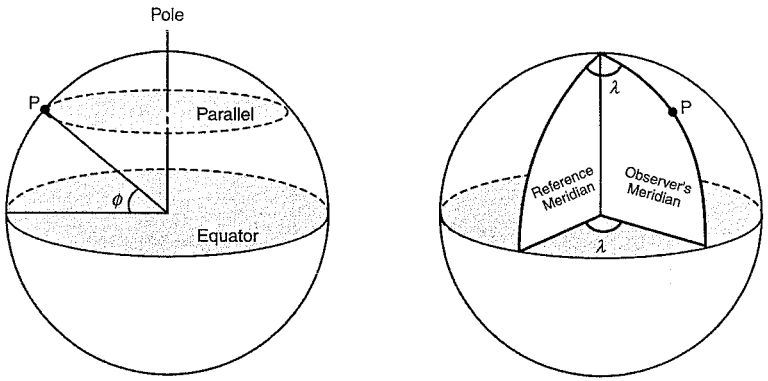


Figure 1.1 Geocentric latitude (ϕ) and longitude (λ).

some of the greatest scientists and craftsmen of the seventeenth and eighteenth centuries. Practical methods for determination of longitude on land had to wait until the 1650s and, at sea, until the 1770s.

Columbus, like many early sailors, certainly knew of latitude, though it's not clear if he knew how to measure it accurately. Measuring longitude, of course, would have been an impossible dream. How, then, did these early explorers manage to cross oceans and return to their home ports? The answer is that they did it mostly by dead reckoning (see below) and seamanship, which included pluck and luck in equal parts. The incentives to find sea routes to new and old lands were great. The spice trade in fifteenth-century Europe was hot, and there was money to be made by cutting out the middlemen involved in land routes. There was even more money to be made by trading in human beings kidnapped from the western coast of Africa, if the sailors only knew which way to sail beyond the Canaries.

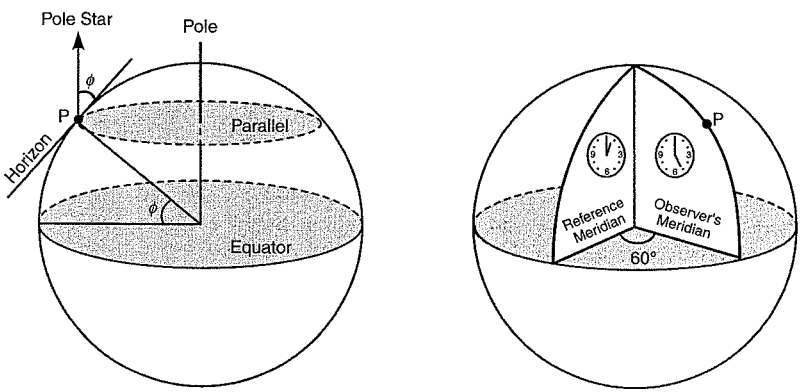


Figure 1.2 Measurement of latitude and longitude.

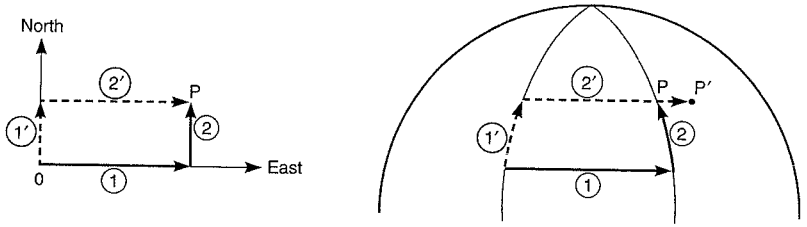


Figure 1.3 Dead reckoning on a plane and a spherical surface. The northerly and easterly legs executed in any order on a flat surface will get a navigator to P. On a spherical surface, the northerly lines converge and dead reckoning is trickier. [Adapted from Williams (1992)]

The early explorers navigated using a simple idea that a position could be estimated relative to the departure point by keeping track of the direction and distance traveled on each leg of a voyage. The technique is called *dead reckoning* (DR). This cheerless term may once have been ded reckoning, short for deduced reckoning. On the other hand, it may have been coined to warn the foolhardy sailors of the likely outcome. Dead reckoning requires measurement of direction, speed and time. The magnetic compass, introduced in the twelfth century in China, served to provide the direction. Distance was measured by estimating the ship's speed and time. Dead reckoning is a simple exercise in vector addition. The problem, however, arises in two-dimensional (2-D) reckoning of distances on a spherical surface because the east-west movement cannot be repeated at different latitudes to same effect as shown in Figure 1.3, and the errors accumulate.

There was also the problem of mapping the surface of the earth on a flat surface. There is no perfect solution as there is no way to reproduce the features of a curved surface on a flat surface. Something has got to give. In 1569, Gerardus Mercator (1512–1594) devised a projection of the spherical earth onto a cylindrical surface. In this projection, the spacing between the meridians of longitude and parallels of latitude is expanded in the same ratio as one moved from the equator to a pole. As a result, map features retain their shapes but the sizes are distorted. An important feature of the projection is that lines of constant bearing appear as straight lines. This was an important advance giving navigators a simple device to chart a course [Snyder (1984)].

1.1.1 Longitude and Time

The difference in longitudes between two places is directly related to the differences in their local times. The earth rotates about its axis 360° in 24 hours, or 15° per hour. The difference in longitudes of two places can, therefore, be determined if the difference in the local times between them is known. A difference of one hour in local times at two places would translate into a difference of 15° in their longitudes (see Figure 1.2). The local time could be measured by using the earth's rotation as a clock with a sundial during the day and star positions at night. The challenge, however, lay in knowing local times at an instant at two places *simultaneously*. And it was an extraordinary challenge. Two approaches were pursued in parallel for well over 200 years pitting the highbrow academics and astronomers against mechanics and builders of contrivances.

- Mechanical clocks: Set a timekeeper to local time at a reference point, and transport it to other places to compare with local times. To sailors, this meant keeping the time at the home port for the length of the voyage. The problem and the solution were well understood in theory, but the map makers and navigators had to wait for the technology to be developed. In the 1500s, the error in a good clock was about ten minutes a day. Considering that a difference of four minutes in local time is equivalent to 1° difference in longitude, within a few weeks of a voyage a sailor would have been placed over dry land. The timekeeper had to be more accurate and reliable than anybody could build in the sixteenth century.
- Astronomical observations: Observe a celestial event and compare the time of the same event as observed at a reference point. The mapmakers could get together for comparison of the local times weeks or months after the event, and determine longitude differences. In order for this approach to be useful to navigators, it had to be known in advance when the said celestial event would take place as seen from the point of reference. This meant predicting the times of occurrence of these celestial events on the basis of past observations and deciphering the pattern of movement of the heavenly bodies.

Both approaches were combined for land mapping. Accurate celestial measurements at sea, however, were difficult and, therefore, for navigation on the seas, the effort focused on development of accurate clocks, or chronometers, as they came to be called.

Key theoretical advances in the design of timekeepers (pendulum controllers and balance spring) came from Christiaan Huygens (1629–1696) in 1657 leading to clocks with an error of about ten seconds a day. But it was clear that no pendulum clock, no matter how isolated from the motion of a ship, could perform on rough seas. The solution lay in spring-driven, non-pendulous chronometers. John Harrison (1693–1776), a carpenter by trade, built a pendulum clock in 1726 that had an error of about one second a month. His crowning achievement, however, was his No. 4 marine chronometer, which saw sea trials in 1761, losing five seconds over an 81-day voyage [Brown (1956)]. The No. 4 met the requirements for the £20,000 prize, a very large sum in the eighteenth century, decreed in 1714 by the Longitude Act of the British Parliament for "discovering the longitude at sea." Chronometers continued to be used for longitude determination at observatories in Europe and the United States until the coming of the telegraph.

It is interesting to note that of all physical quantities, time is now measured with the greatest precision and accuracy [Allan, Ashby, and Hodge (1997)]. It will take clocks costing thousands of dollars to match the accuracy of time obtained with a $100 GPS receiver. GPS allows for this time to be distributed across the globe with ease, serving effectively as a precise grandfather clock in the sky.

1.1.2 Astronomical Methods

An idea that was pushed hard by the scientific establishment, which included Isaac Newton (1642–1727) and Edmond Halley (1656–1742) of comet fame, in competition with the clock technology, was tied to lunar orbits. The idea was to exploit the rapidly changing position of the moon relative to a fixed star—the "lunar-distance" method—to determine date and time. To make the method practicable, what was needed was a table giving the local times of the dis-

tance between the moon and certain stars at a location of known longitude. A navigator could then compare, for example, the time he observed the moon graze a star with the time such event was predicted at the reference point. The problem was that the star positions were unmapped, and lunar motion was not fully understood. Prediction of lunar orbits for their use in navigation was the impetus for the founding of the Paris Observatory in 1667 and the Greenwich Observatory in 1675.

A prerequisite for computing star tables was a theory of the motion of the moon that, with adequate observational data, could be used to predict position of the moon relative to the distant stars with sufficient accuracy. Newton's theory based on his solution of a two-body problem [Section 4.3.1] gave an error of 5 arc-minutes, an unacceptable level. The lunar problem was a three-body problem: motion of the moon which revolves around the earth, which in turn revolves around the sun. Development of such a theory by Leonhard Euler (1707–1783) in 1748 was one of the great mathematical achievements of the eighteenth century. The idea never became practical at sea even though the proponents kept pushing for it long after the battle had been lost to chronometers, and later radio. (Astronomer Royal Nevil Maskelyne (1732–1811), a champion of the lunar-distance method, was seen by some as a villain who schemed to deprive Harrison of the longitude award.)

The lunar-distance method was not the only candidate for positioning and navigation based on celestial observations. Development of the telescope early in the seventeenth century led to discovery of the moons of Jupiter. As seen from the earth, the satellites appear and disappear as they pass behind Jupiter's shadow. These eclipses occur at precisely the same moment for all observers on earth. If the eclipses could be predicted accurately, Jupiter's moons could serve as a celestial timepiece for determining longitude. Galileo Galilei (1564–1642), having established the periods of the satellites and having calculated tables of their motion, proposed to the Spanish Crown to teach sailors to observe Jupiter's moons to find their longitude. The problem occupied Galileo to the end of his days but never became practical at sea. Toward the end of the seventeenth century, however, the satellites of Jupiter had become the official method of the French Crown for longitude determination on land.

The seventeenth and eighteenth centuries also saw development of the technology for precise measurement of angles at sea, required by the astronomical methods of navigation. Robert Hookes' reflecting quadrant and Newton's reflecting octant, both developed in the second half of the seventeenth century, were revolutionary instruments which allowed a star and the horizon to be brought together, one seen directly and one through a reflecting mirror. Quadrants, octants, and eventually sextants, equipped with telescopes, mirrors, prisms, and vernier scales, were developed as precision instruments. At the end of the eighteenth century, the essential elements were in place for precise celestial navigation: a sextant to measure the elevation of celestial bodies above the horizon, an accurate clock to determine time of the observations, an almanac to find the predicted position of the body, and a magnetic compass to determine azimuth and to maintain course continuity between celestial observations.

In the nineteenth century, wooden ships had made way for iron hulls. The magnetization of the ship and its cargo interfered with the workings of the magnetic compass. What was needed was a non-magnetic compass. The problem was not solved completely until early in the twentieth century with the invention of the gyrocompass, discussed briefly below.

1.1.3 Twentieth Century Developments: Inertial Navigation and Radio

Two early events in the 20th century foreshadowed its technological direction. The first was the bridging of the Atlantic Ocean in 1901 by a radio signal transmitting the letter "S" in Morse code from England to Newfoundland. Guglielmo Marconi (1874–1937) was responsible for this feat. The 20th century would be the century of radio. In 1903, the Wright Brothers flew an airplane. The 20th century would also be the century of aviation, but not without the radio. In time, jetliners loaded with hundreds of passengers were flying halfway around the world navigating with radio navigation aids (navaids) and inertial navigation systems, and being guided by radio signals to narrow strips of land shrouded in fog. Nuclear-powered submarines were using inertial navigation not only to navigate but to initialize missile position, velocity, and azimuth. In this section, we offer brief introductions to inertial and radio navigation.

Guidance and Positioning Systems: A Taxonomy

It is instructive to consider the following classification of navigation systems, old and new.

- *Dead reckoning systems*, which we have met already, and will discuss further below, but only briefly.
- *Guidance systems*, which provide the user a course to steer by toward a destination without necessarily knowing his position. Examples are lighthouses and radio beacons. The Instrument Landing System (ILS) and Microwave Landing System (MLS), used by aircraft to land under poor visibility conditions, are radio guidance systems [Forssell (1991), Enge *et al.* (1995)]. Heat sensors which guide anti-aircraft missiles are also guidance systems. These systems are not discussed in this book.
- *Position finding systems*, which determine the position of a user in a well-defined coordinate frame. Examples are Loran, Omega, and Transit, which are described briefly in this chapter. GPS, the main subject of this book, is also a position finding system. Actually, GPS is more than that because it also gives velocity and time.

Inertial Navigation

A gyroscope is simply a spinning mass, usually mounted on gimbals so that the rotation axis is free to turn in any direction. The beauty of a gyroscope is that if its spin axis is pointed at a star, it will continue to point at the star as the earth rotates and the apparent position of the star changes. In other words, the spin axis maintains its direction in space (or inertial space). The idea of a gyroscope has been around since the mid-nineteenth century, but building one with precision for navigation had to wait until the twentieth century. A north-seeking gyroscope came to be called a *gyrocompass*.

The ability of a gyroscope to provide a stable or inertial platform which maintains its orientation in space led to inertial navigation systems (INS). An INS consists essentially of three accelerometers mounted along mutually orthogonal directions on such a stable platform. The instrument senses rotations and accelerations, and keeps track of them. The rotations (yaw, pitch, and roll) of a vehicle are sensed in comparison with the orientation of the stable platform. The system numerically integrates the acceleration components in real time to provide

velocity components. A second integration provides the current position coordinates. An INS is a small, self-contained system, an attractive feature for military use. Aircraft, submarines, spacecraft, and missiles have all depended upon an INS for navigation, and it would be difficult to overestimate the role played by this technology of the twentieth century.

The INS is a dead reckoning system and, as is characteristic of dead reckoning systems, the errors accumulate. A moderately accurate INS can accumulate a 2 km position error in an hour, and requires updating with accurate position estimates obtained by some other external means, for example, star sightings and, in recent years, GPS. GPS and INS are complementary technologies in the sense that the weakness of one is the strength of the other: GPS is vulnerable to jamming and interference to which INS is immune; errors accumulate in INS-based position estimates, but an INS can be reset with GPS-based estimates which are free of such drifts. Integration of INS and GPS leads to a particularly attractive and robust system which can flywheel through GPS outages due to signal interference or satellite blockage over short periods [Farrell and Barth (1999)].

An exciting recent development in inertial technology is the arrival of micromachined, silicon-based inertial sensors fabricated by integrated circuit fabrication methods. Microaccelerometers are being produced by the millions, though not yet for navigation. The main market is automotive safety systems, in particular, air bag deployment. Micromachined gyroscopes for measuring angle and angle rate of rotation are finding applications in the automotive market (e.g., ride stabilization and rollover detection) and consumer devices (e.g., video-camera stabilization, wheelchair balance, and robotics), but are not yet accurate enough for navigation. In a few years, look for these microelectromechanical systems (MEMS) to be integrated with GPS for a whole range of civil and military applications.

Radionavigation

Radionavigation systems exploit the basic principles of propagation of radio waves, which we review in this section. The radio waves correspond to a range of electromagnetic wave frequencies from 10 kHz to 300 GHz, where Hz (hertz) is the unit of frequency (cycles per second). Prefix 'k' stands for kilo (10^3), 'M' for mega (10^6), and 'G' for giga (10^9). The radio frequencies are classified into bands for convenience of nomenclature (see Table 1.1). The speed of electromagnetic waves in free space (c) is approximately 3×10^8 m/s, and the wavelength in meters is obtained as c/f, where f is the frequency in Hz. Radio waves with wavelengths in the range 0.1 cm to 10 cm are called *microwaves*.

The electromagnetic spectrum is a precious and finite resource with many competing claims for usage. It is also a fragile resource. The need for cooperation and regulation in the use of this resource was recognized early in this century with the founding in 1934 of the International Telecommunication Union (ITU), now a specialized agency of the United Nations (U.N.). The mission of ITU is to achieve international consensus on allocation of radio frequencies to different user groups to prevent mutual interference.

ITU divides the electromagnetic spectrum into frequency bands, with different radio services assigned to particular bands. The radio services competing for frequency allocation include navigation, telephone, telegraph, radar, TV, AM and FM radio, mobile communications, various aspects of satellite communications (satellite to satellite, satellite to ground, ground to satellite), and others. The trick is to maximize value without mutual interference. It is not easy. With the current explosion of mobile communications, the potential providers of these

Table 1.1 Classification of radio frequencies

Band	*Frequency*	*Wavelength*
Very Low Frequency (VLF)	< 30 kHz	> 10 km
Low Frequency (LF)	30–300 kHz	1–10 km
Medium Frequency (MF)	300 kHz–3 MHz	100 m–1 km
High Frequency (HF)	3–30 MHz	10–100 m
Very High Frequency (VHF)	30–300 MHz	1–10 m
Ultra High Frequency (UHF)	300 MHz–3 GHz	10 cm–1 m
Super High Frequency (SHF)	3–30 GHz	1–10 cm

services are competing fiercely for a larger portion of the spectrum. In the United States, the National Telecommunications and Information Administration (NTIA) promulgates policies, regulations, and technical standards for all Federal government users of the radio spectrum, including GPS. The Federal Communications Commission (FCC) manages the non-government use of the spectrum.

The radio signals travel in free space in straight lines at the speed of light. Free space is an idealization with no electric or magnetic fields, nor any obstructions. Space meets these requirements, but the earth's atmosphere does not. Signal propagation in the earth's atmosphere can be very complicated depending upon the signal frequency and the environment.

Radio signals propagating in the earth's environment are reflected by the ground, buildings, and surface of water, much as light is reflected. Like light, radio waves are refracted when propagating from one medium to another of a different density. The radio signals are also diffracted, and interfere with each other, like optical signals. The signals may also be attenuated by the earth's atmosphere.

The VHF and higher frequency signals travel in straight lines and do not propagate beyond the horizon. The frequencies below the HF range travel around the curvature of the earth hugging the terrain. These are called *ground waves*. Radio waves in the HF range and below are also reflected by the ionized layer of the atmosphere called the ionosphere [Section 5.3.2], and are called *sky waves*. These signals are beamed into the sky and, after reflection, return to the earth, sometimes at great distances. Before satellite communications, ground wave and sky wave radios provided the means for beyond-the-horizon propagation (see Figure 1.4).

As we will see, in order to be useful for navigation, the propagation times of the signals and, therefore, the propagation conditions, have to be predictable and repeatable. In the ground wave mode, the signal is affected by the terrain: mountains, desert, snow, salt water, lakes, and farm land, all have distinct characteristics. The propagation velocity is generally stable and the propagation losses are generally moderate. The sky wave can be received at greater distances but the ionospheric height from which it is reflected is not constant and, therefore, adds uncertainty to signal propagation times.

We discuss below methods of radionavigation and introduce radionavigation systems. The long-range terrestrial radionavigation systems to date (Loran, Omega, and Decca) have used

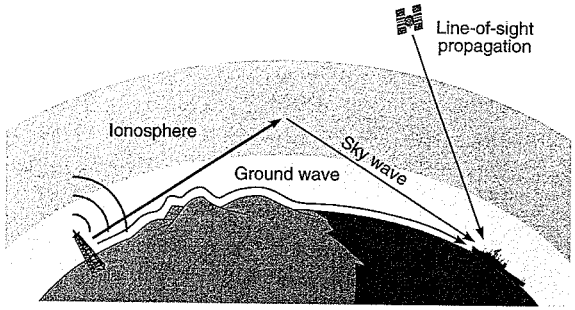


Figure 1.4 Propagation of radio waves.

VLF and LF bands. The signals of ground-based, short-range, line-of-sight radio guidance systems used in aviation (VOR, ILS, and MLS) range from VHF to SHF. The satellite navigation systems (Transit, GPS and GLONASS) have used VHF and UHF bands.

Radionavigation systems have historically been owned and operated by governments for reasons of safety and security, and to promote commerce. Loran, Omega, Transit, and GPS all started out as military systems which were subsequently opened for civil use. An exception was Decca, a commercial hyperbolic navigation system developed in the U.K. during and after World War II which guided maritime traffic in coastal areas in many parts of the world for almost fifty years.

1.2 Methods of Radionavigation

The use of radio waves for obtaining position estimates (also called *fixes*) is nearly as old as that for communication. The first radio aids determined position of a ship or aircraft by measuring bearings to two or more radio beacons with a directional antenna. The development of radio navaids for aviation accelerated during and after World War II resulting in VHF omnidirectional radio range (VOR), Tactical Air Navigation (Tacan), Instrument Landing System (ILS), and Microwave Landing System (MLS). All are ground-based, short-range, line-of-sight systems [Enge *et al.* (1995)], which are not discussed in this book. We'll narrow our focus to long-range, global radionavigation systems and, then, quickly to GPS.

Navigators on land or over oceans are generally satisfied with the horizontal or two-dimensional (2-D) position estimates. However, the geodesists and surveyors preparing topographic maps require heights. The aviators need altitude in real time. In both cases, the horizontal and vertical positions were determined in the past separately, using different techniques and technologies. It takes a satellite-based navigation system to offer three-dimensional (3-D) position.

1.2.1 Trilateration

The simplest of the principles of radio wave propagation is that these waves travel at a known speed. Therefore, if the transit time of a signal from a transmitting station can be measured, the distance between the transmitter and the observer can be determined. Given distances to three transmitters at known locations, the observer can compute his position unambiguously.

Estimation of a position based on measurement of distances is referred to as *trilateration*. A radionavigation system based on this idea is referred to as a *time-of-arrival* (TOA) system. GPS is a TOA system.

The basic idea of trilateration is illustrated in Figure 1.5 for 2-D positioning. An observer measures his distance (or range) from station S1, whose position coordinates are known. It follows that the observer must be located somewhere on a circle of radius r_1 centered at S1. In other words, we have identified a so-called line of position (LOP), which in this case is a circle. Range measurement from a second station S2 gives another circular LOP, and reduces the uncertainty in the observer's position to the two points (P and P') where the circles intersect. The observer may be able to reject one of the estimates on the basis of prior information. If not, a third measurement would determine the position uniquely.

Mathematical modeling is straightforward. The observer's position is determined by solving a set of quadratic equations

$$\sqrt{(x_k - x)^2 + (y_k - y)^2} = r_k \tag{1.1}$$

where (x_k, y_k) are the known station coordinates; $k = 1, 2, 3$; (x, y) are the observer's coordinates to be determined; and r_k is the measured range.

Extension to 3-D positioning is straightforward, in principle. In order to estimate the vertical position, it is required that the elevation angle of at least one of the stations relative to the observer be 'large' [Section 6.1.2]. The terrestrial radionavigation systems are limited in terms of transmitter heights and, therefore, are limited to 2-D navigation. With space-based signal transmitters, each range measurement would specify a surface of position, a sphere. Intersection of three such spheres would identify a point on or near the earth unambiguously.

In order to measure the signal transit time, it is required that the clocks at the transmitters and the receiver be maintained in synchronization. The radio signals travel at about 3×10^8 m/s, and a synchronization error of 1 μs would result in an error of 300 m in distance measurement. A transmitter can be synchronized to another unit at a known distance from which

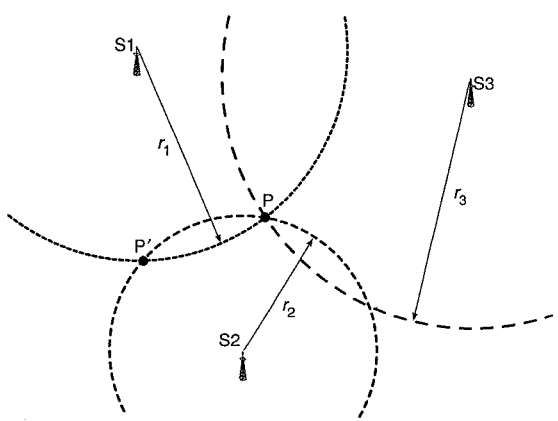


Figure 1.5 Trilateration.

it receives a signal. The demands on a receiver clock, however, must be kept to a minimum to keep the price of a receiver down if the system is planned for wide use. As discussed later in this chapter, GPS minimizes demands on the receiver clock at the expense of an additional radio transmitter and increased computational complexity in the receiver. The latter is easily accommodated in this age of microprocessors, but lack of raw compute power in the early radionavigation receivers put this approach at a disadvantage in comparison to an alternate approach discussed next.

1.2.2 Hyperbolic Positioning

These systems are based on measurement of the *difference* in the times of arrival of signals from two transmitting stations. The transmitters are synchronized; a receiver clock has to measure a time difference, typically a few milliseconds, and does not have to be synchronized to the transmitter clocks. A radionavigation system based on this idea is called a *time-difference-of-arrival* (TDOA) system. Loran is such a system. Omega, the first truly global radionavigation system, now discontinued, was also a TDOA system. Both are described briefly in the next section.

The basic idea is illustrated in Figure 1.6 for 2-D positioning. An observer measures the difference between his ranges to two stations at known locations. The locus of the points having the same range difference from the pair of stations is a hyperbola. That's the first LOP. The measurement of a range difference from a second pair of stations would specify another hyperbola as a second LOP. The observer lies at the intersection of the two hyperbolas and, therefore, this approach to positioning is called *hyperbolic positioning*. At a minimum, we need three stations forming two station pairs. There are position ambiguities due to multiple intersections of the LOPs. These ambiguities can be resolved on the basis of prior knowledge (i.e., a rough idea of where the user is), or additional measurements.

For transmitter pairs (M, X) and (M, Y), Figure 1.6 shows two families of hyperbolic LOPs x_i and y_i corresponding to the different values of the range difference. For example, hyperbola

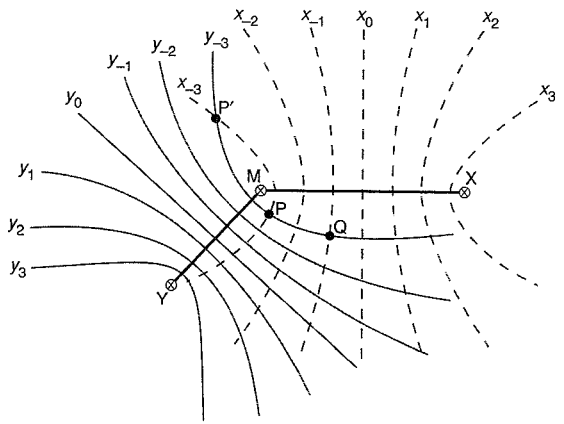


Figure 1.6 Hyperbolic positioning.

marked x_i is the locus of points that are i units closer to transmitter X than to transmitter M. A user at position P may find it difficult to rule out position estimate P′, and may require another measurement. A user at Q would have an easier time of dealing with such an ambiguity.

As before, we represent the user coordinates as (x, y) and the known coordinates of the three transmitters M, X, and Y in Figure 1.6 as (x_k, y_k), $k = 1, 2$, and 3, respectively. Let d_{21} be the measured difference between the observer's distances to stations 2 and 1; d_{31} is similarly defined. The user position is obtained by solving the following equations for (x, y).

$$\sqrt{(x_2 - x)^2 + (y_2 - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2} = d_{21}$$
$$\sqrt{(x_3 - x)^2 + (y_3 - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2} = d_{31} \qquad (1.2)$$

1.2.3 Doppler Positioning

The *Doppler effect* is another principle of propagation of radio waves (actually waves, in general) exploited for radionavigation. Recall that the Doppler effect is the change in the apparent frequency of a signal received by an observer due to relative motion between a transmitter and the observer. *Doppler shift*, defined as the difference between the frequency of the received signal and the frequency at the source, is a part of our everyday experience, but is less intuitive as a basis for a global navigation system. In fact, the first satellite navigation system, called Transit, was based on the measurement of Doppler shift. This approach to positioning is called *Doppler positioning*.

The basic principle of Doppler positioning is illustrated in Figure 1.7 for one-dimensional (1-D) positioning based on the apparent pitch of the whistle of a locomotive on a straight and level track passing an observer at a steady high speed. For an observer standing in the middle of the track (not recommended), the pitch starts out high and then drops sharply as the locomotive runs over him and recedes. The classical explanation of the apparent change in pitch

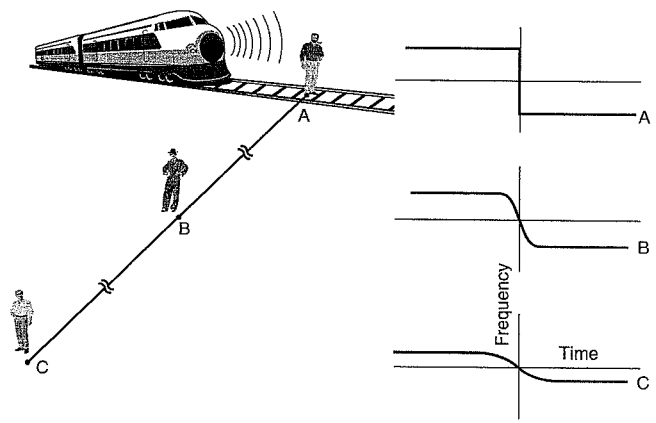


Figure 1.7 Doppler positioning: The Doppler shift profile can be related to the observer's distance from the track.

is as follows. As the locomotive approaches, more cycles are received in a time interval than transmitted due to the shrinking distance between the transmitter and the observer, and the pitch seems higher; the opposite is true as the locomotive recedes.

The transmitted and received frequencies, f_T and f_R, respectively, are related by

$$f_R = f_T\left(1 - \frac{\dot{r}}{v_s}\right) \tag{1.3}$$

where r is the (changing) line-of-sight distance between the transmitter and the receiver, $\dot{r}$ is the range rate, and v_s is the speed of propagation of the waves. The range rate can also be thought of as the line-of-sight component of the transmitter-receiver relative velocity vector. The Doppler shift is

$$(f_R - f_T) = -\frac{\dot{r}}{\lambda} \tag{1.4}$$

where λ is the wavelength of the transmitted signal.

The change in pitch for the observer standing in the middle of the track is a step function (see Figure 1.7). An observer standing away from the track also experiences Doppler shift, but the change between the high and low pitch is smooth. The farther away the observer is from the track, the lower the range rate, and the smaller and slower the change in pitch, as shown in Figure 1.7. The shape of the Doppler shift curve determines the observer's distance from the track. This is 1-D positioning, except for the fact that we have an ambiguity; we can't tell which side of the track the observer is on.

Armed with a watch and the train's schedule, we can turn the above scheme into a 2-D positioning system as follows. Record the time when the pitch of the whistle was midway between the high and the low. Check the train's schedule to determine its position along the track. The along-track position would be in error to the extent the train does not follow the time table, or the observer's watch is biased. Now we have the along-track and cross-track positions of the observer. That's essentially how Transit, the first satellite navigation system, discussed below, worked.

1.3 Radionavigation Systems

1.3.1 Terrestrial Radionavigation Systems: Loran and Omega

Loran

The radionavigation era began in earnest during World War II with the development of *Gee* in Great Britain to guide aircraft and *Loran* (Long-range navigation system) in the United States primarily for ships. Both systems were designed as high-power hyperbolic systems transmitting pulse signals. We discuss Loran, a later version of which remains in service.

Loran-A was developed at the Radiation Laboratory (the legendary RadLab) of Massachusetts Institute of Technology (MIT). It was developed, tested, and deployed with extraordinary speed. By the end of the war, the North Atlantic and parts of the Pacific were covered by Loran-A. Loran-C, the system in use today, was developed in the late 1950s and early 1960s to

expand the range and to improve the accuracy of Loran-A. Loran-C remained a military system until 1974 when it became a Federally provided radionavigation system for civil marine use in the U.S. coastal areas. The system was subsequently expanded to cover all of the coastal and inland navigable waters. In the mid-1980s, the coverage was increased further to fill the mid-continent gap, and the Federal Aviation Administration (FAA) approved the use of Loran-C for navigation in the U.S. airspace.

Loran-C is a hyperbolic system operating in the LF band at 90–110 kHz. The system comprises a set of chains of transmitters. A typical chain consists of a master and two to three secondary transmitting stations separated by about 1000 km, each transmitting at about 1 MW of peak power. The U.S. Coast Guard operates 29 transmitting stations composing 13 chains covering the U.S. coastal waters, conterminous 48 states, the Aleutians, and stretching into the Bering Sea [Loran-C (1992)]. Loran-C never became a global system, though it had the capability. Currently, the system covers much of the northern hemisphere. Loran-C provides 2-D rms positioning accuracy of about 250 m.

The synchronized transmitters of a chain radiate pulses of RF energy. A receiver measures the time difference (TD) between arrival of pulses from the master and secondary stations. Each measured TD defines a hyperbolic LOP for the user. Intersection of two LOPs defines the user position in 2-D. The Loran signal has both ground wave and sky wave components. The ground wave is generally stable and predictable. The sky wave component is not suited for precise positioning because of uncertainties associated with the height of the ionosphere, and the Loran pulse is designed to prevent the delayed sky wave from interfering with the TD measurements from the ground wave [Enge *et al.* (1995)].

In the mid-1990s, the U.S. Government decided to decommission Loran in 2000, but subsequently reversed the decision. The current view is that Loran can serve as a sort of insurance against catastrophic loss of GPS service. The GPS-Loran combination is seen as a robust system, both for navigation in safety-of-life applications such as civil aviation and as a source of precise time. The Loran infrastructure in the United States is now being upgraded, and the system is likely to be around until 2010, and perhaps later. A number of Loran stations continue to be operated in Europe, Asia, and the Middle East. A Russian system similar to Loran, called *Chayka*, continues to be operational.

Omega

Omega was the first worldwide, continuously available radionavigation system. It was developed in the early sixties by the U.S. Navy, the U.S. Coast Guard, and six international partners. The system became operational in the 1970s although the last permanent station (located in Australia) did not come on line until 1982. Omega was decommissioned in 1997.

Omega was realized with eight ground-based transmitters, two of which were located in the United States. Each station operated in continuous-wave mode, transmitting synchronized time-shared signals at five frequencies in the VLF band: 10.2 kHz, 11.05 kHz, 11.333 kHz, 13.6 kHz, and a unique frequency to identify the station. The radiated power of the transmitting stations was about 10 kW. The signals propagated around the globe between the ground and the ionosphere in sky wave mode, trapped in the earth-ionosphere 'duct.' The system was used for transoceanic shipping and saw use in civil aviation after the FAA approved its use as sole means for oceanic navigation and as supplemental for en route navigation [FRP (1994)].

The positioning error of a radionavigation system depends upon the uncertainties in the

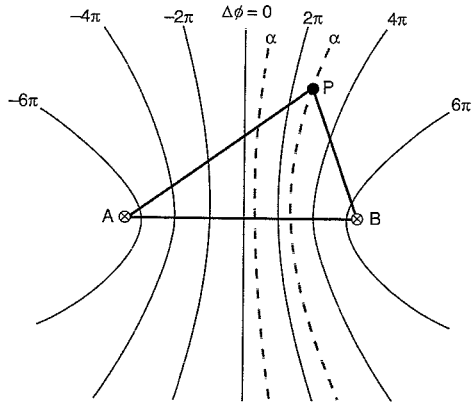


Figure 1.8 Omega lane boundaries and lane ambiguity. A user measuring phase difference α (modulo 360°) has to narrow the position uncertainty enough to identify the correct lane.

signal propagation path. For Omega, empirical propagation models were developed on the basis of worldwide measurements over extended periods. These modes were built into the receivers in the form of propagation tables and formulas, and were used to correct the measurements. The system provided 2-D rms positioning accuracy of 2–4 km.

Omega provides us with a simple context in which to introduce two ideas which will come up again when we get to GPS. The first deals with the ambiguity of phase measurements, and the second with differential corrections to improve the quality of the position estimates. We discuss both below [Omega (1983), Enge *et al.* (1995)].

Omega was a hyperbolic navigation system with one subtle difference: Rather than measuring time difference of arrival, an Omega receiver measured the phase difference $\Delta\phi$ between the sinusoidal signals. A line of constant phase difference is a hyperbola, but not a unique one. Lines of constant phase difference recur over the coverage area. Figure 1.8 shows hyperbolic LOPs corresponding to 0° phase difference. In general, given the phase difference $\Delta\phi$, the 'correct' LOP could be any of the hyperbolas corresponding to phase differences $\Delta\phi \pm$ integer multiples of 360°. The region between the lines of zero phase difference is called a *lane*. In order for a phase measurement to provide an unambiguous position estimate, the user has to determine the correct lane first.

For 10.2 kHz signals, the width of a lane along the line joining the two transmitters is about 15 km (one-half of a wavelength). If the *a priori* position uncertainty is less than 7.5 km, the correct lane is identified and phase difference measurements can be used to determine position. If not, we have *lane ambiguity*, and must define wider lanes by using phase differences for a different pair of frequencies. Consider phase measurements at 10.2 and 13.6 kHz. Lines corresponding to constant phase difference at both frequencies repeat after about 45 km (one-half wavelength of the difference frequency of 3.4 kHz). If the initial position uncertainty is greater yet, phase measurements at 10.2 and 11.333 kHz can be used to give a lane width of about 130 km. Given an initial user position, the receiver can keep track of the lanes crossed in the course of a voyage and produce subsequent position estimates without any ambiguity.

The Omega signals propagated over thousands of kilometers under conditions that could change significantly. The propagation models used to correct the measurements basically accounted for the average conditions. The actual propagation conditions could deviate considerably from the average, and this uncertainty was the largest source of error in the Omega measurements. The conditions, however, change slowly, and are similar for users within tens of kilometers of each other.

By monitoring Omega signals at a fixed, known position, the measured carrier phase difference could be compared with that predicted by the models used to build the propagation tables. The discrepancy between the two was the measurement error which is common to all Omega users in the area. The measured error, if broadcast to the users in the area as a *differential correction*, allowed them to mitigate their measurement errors and improve the quality of their position estimates. This system, called *Differential Omega*, was implemented in parts of the Atlantic Ocean and the Mediterranean and Caribbean Seas, where the corrections were broadcast over marine radiobeacons.

The idea of differential corrections is relevant to other radionavigation systems as well. The performance of Loran can be improved significantly in differential mode. The use of differential GPS is already widespread [Section 5.8].

1.3.2 Satellite Navigation Systems: Transit, GPS, and GNSS

The space age began in 1957 with the launch of *Sputnik I* by the Soviet Union, opening a new frontier for both the Cold War and commerce. Both the United States and the Soviet Union allocated vast resources over the next ten years to the space race. An important benefit to flow from this intense competition was that the potential for space-based communication, navigation, and surveillance systems was recognized and realized quickly.

Transit

The genesis of Transit was a brilliant discovery followed by an equally brilliant deduction. First the discovery: The pattern of Doppler shifts in the signals transmitted by *Sputnik I* measured from a single ground station at a known position was enough to determine the satellite's orbit. Now the deduction: If the satellite orbit were known, a radio receiver measuring Doppler shifts could determine its position on the earth. It was not immediately clear, however, if the idea could actually be realized.

In 1958, there was much to be learned about designing reliable, long-life spacecraft, launching them into space, keeping the antennas pointed toward the earth, and determining accurate orbits, given the non-uniform nature of the earth's gravity field. The U.S. Navy, however, had a vital need for a navigation system to guide a new class of submarines carrying Polaris missiles, and the first satellite navigation program was born. The system was named *Navy Navigation Satellite System*, but came to be better known as *Transit*.

Transit was a pioneering achievement of the Applied Physics Laboratory of The Johns Hopkins University, which was responsible for the whole thing from the initial concept developed in 1958 to the experimental satellites launched in 1961–1962, and the final system which became operational in 1964. Of course, the credit must be shared with the enterprising leadership of the U.S. Navy, which backed a revolutionary idea and moved swiftly to turn it into an operational system within six years. Our description is based on a lucid account by Stansell (1983).

Transit was realized with four to seven satellites in low-altitude (1100 km), nearly circular, polar orbits. Each satellite broadcast signals at 150 MHz and 400 MHz with a total transmitted power of one watt. Only one satellite was in view at a time and a user waited up to 100 minutes between successive satellite passes to determine position. After a satellite came in view, the receiver recorded continuously the Doppler shift of the received signal and the navigation message giving the satellite position during the satellite pass lasting about 10–20 minutes. Actually, the receivers determined a Doppler count, or integrated Doppler, over an interval instead.

$$\text{Doppler count} = \int_{t_{i-1}}^{t_i} (f_T - f_R)dt \tag{1.5}$$

where f_T and f_R are the transmitted and received frequencies, respectively. In measuring Doppler count, the receiver effectively formed a beat-frequency signal by differencing the received signal from a sinusoidal signal generated by the receiver clock with a nominal frequency f_T. The cycles of the beat signal, counted as so many whole cycles and a partial cycle, gave the Doppler count. From the definition of Doppler shift (1.4),

$$\begin{aligned}\text{Doppler count} &= \frac{1}{\lambda}\int_{t_{i-1}}^{t_i} \dot{r}(t)dt \\ &= \frac{1}{\lambda}[r(t_i) - r(t_{i-1})]\end{aligned} \tag{1.6}$$

The Doppler count is thus a direct measure of change in distance between the receiver and the satellite over a time interval. For each wavelength the satellite moves away, the Doppler count goes up by one, and vice versa.

The measurements from a satellite pass were subsequently processed to compute 2-D position of a stationary or slow-moving user. Dual-frequency measurements allowed for correction for the ionospheric propagation delay [Section 5.3.2]. The user's velocity, if known, could be included in the mathematical model for positioning. Any error in the velocity estimate resulted in penalty in positioning accuracy. (Rule of thumb: 1 km/h velocity error could translate into a 200 m error in the position estimate.) As in the previous example [Section 1.2.3], the shape of the Doppler shift curve provided a measure of the user's distance from the satellite's ground track, and the receiver clock and the satellite almanac determined the position along the ground track. The Doppler shift profiles for two points placed symmetrically about the satellite ground track are different due to the earth's rotation. This automatically resolved the position ambiguity we had observed in relation to the train track in our earlier example.

Transit was used by the U.S. submarine fleet to update a ship's position and reset the inertial navigation system. The 2-D positioning accuracy of Transit was about 25 m (rms) for a stationary user. The system saw limited civil use in the maritime industry and geodesy starting in 1967. A Transit receiver tracking and recording measurements from multiple satellite passes from a fixed location over several days could achieve a 3-D accuracy of 5 m. This mode of usage is called *absolute positioning* or *fixed-site survey*. An alternative is *relative positioning*: estimating the position of a point relative to another point. Accuracy of 1 m was achieved in 3-D relative positioning with Transit at a distance of hundreds of kilometers by tracking multiple satellite passes concurrently at both points.

There was a Cold War echo from the Soviet Union in the form of two systems virtually identical to Transit: *Parus* for the Soviet Navy and *Tsikada* for merchant ships. Transit was decommissioned in 1996. Parus appears to be active in 2005, but not Tsikada.

The Doppler-based system worked well for ships at sea which required infrequent position updates and could track a signal continuously during the 10–20 minute satellite pass. This technique is not suited for aircraft or mobile users requiring frequent or continuous positioning.

GPS

The success of Transit begat follow-on programs for space-based navigation systems. The U.S. Navy and the Air Force each had a development program in the late 1960s. The two were eventually combined into GPS. A dozen years had passed since the idea of Transit was conceived, and much had been learned about designing spacecraft, launching them in space, and tracking and maintaining them in orbit.

It is instructive to reconstruct the considerations of the GPS designers that led to the choice of system architecture. We do so below, and also introduce some new concepts.

- Active or passive system: A passive system broadcasts signals and a user determines his position by simply 'listening' to them. An active system interacts with each user (e.g., a user may be required to interrogate the system). An active system can handle a fixed, finite number of concurrent users; a passive system can service an unlimited number of users. Besides, a military user wouldn't want to give away his position by radiating signals.
- Positioning Method: Doppler, hyperbolic, or trilateration. In 1970, the clock technology had improved to a point that time-synchronized signals could be transmitted from satellites. Trilateration was the logical choice. GPS is, therefore, described as a *passive ranging system.*
- Pulsed versus continuous wave (CW) signals: Pulsed signals could be time-coordinated so that each satellite would have a separate transmission time. Spread spectrum signaling, however, allows simultaneous signal transmissions on one radio frequency. GPS is the first widespread application of this signaling concept called code division multiple access (CDMA) [Chapter 9].
- Carrier frequency: L-band (1 GHz–2 GHz) was optimum. GPS needed 20 MHz of spectrum, and L-band was relatively uncluttered in the early 1970s (no longer true). At higher frequencies, ranging error due to ionospheric refraction decreases but space loss (i.e., attenuation of signal power due to distance traveled) [Section 10.1] and atmospheric attenuation increase.
- Satellite constellation and orbits: Each user needs four or more satellites in view in order to determine her position (see below). Such worldwide coverage had to be provided economically. The choice of orbits determines satellite design, number of satellites, and costs of launch and maintenance.
  - Low earth orbit (LEO): With altitude < 2000 km, a LEO satellite is only visible for 10–20 minutes at a time, and the receiver would be acquiring new satellites constantly. Doppler rates would be high. Orbital perturbations due

to the atmospheric drag would be high, too. It would take a constellation of 100–200 satellites to provide global coverage for navigation. On the plus side, launch costs would be lower and the satellites could get by with low-power transmitters.

- Medium earth orbit (MEO): With altitude of 5000–20,000 km and 2–4 orbits per day, the satellites are typically visible for several hours in each pass. Launch costs are higher than those for a LEO, but a smaller satellite constellation (24–36 satellites) would be adequate.
- Geostationary orbit (GEO): A satellite in a 36,000-km orbit over the equator appears fixed to an observer on the earth. Global coverage can be provided with a small number of satellites. On the negative side, GEO satellites offer poor coverage at higher latitudes, and have higher launch costs.

These considerations led to a choice of a MEO constellation of 24 satellites for GPS. The Department of Defense (DoD) approved the basic architecture in 1973, and the first satellite was launched in 1978. The system was declared operational in 1995. The cost of development of GPS has been reported to be about $10 billion; the annual operation and maintenance cost is estimated to be about $500 million. As an aside, we should note that the full name given to GPS was *NAVSTAR, the Global Positioning System*. NAVSTAR is sometimes represented as an acronym for Navigation System with Time and Ranging. Though apt, the DoD did not intend NAVSTAR as an acronym [Parkinson (1996)].

Three engineers/scientists have been honored widely for their roles in the development of GPS: Roger L. Easton [Easton (1980)], Ivan A. Getting [Getting (1993)], and Bradford W. Parkinson [Parkinson (1996)]. Among the many honors Easton has received is the 2004 National Medal of Technology for "his invention of 'Navigation System using Satellites and Passive Ranging Techniques' and his subsequent development of Time Navigation and Navigation Technology Satellites that formed the technological basis for the modern GPS."

Getting (1912–2003), President of the Aerospace Corporation in the 1960s and 1970s, is credited with advocacy and leadership in the development of GPS. The 50th GPS satellite (GPS IIR-11), launched on 20 March 2004, was dedicated to him. Parkinson, the first Director of the GPS Joint Program Office (GPS JPO) in the 1970s, served as the chief architect during its development and implementation, and shepherded the program through the DoD acquisition process with great skill. Among the many honors Getting and Parkinson received is the shared $500,000 Charles Stark Draper Prize of the National Academy of Engineering in 2003 "for the concept and development of the Global Positioning System." (It's interesting that the White House and the NAE saw the responsibility for the development of GPS so differently.)

Four technologies which saw spectacular developments in the second half of the twentieth century and turned the ancient idea of trilateration into GPS are:

- stable space platforms in predictable orbits,
- ultra-stable clocks,
- spread spectrum signaling,
- integrated circuits.

Recall that trilateration required measurement of ranges to three position references (i.e., objects at known positions). In GPS, the position references are the satellites, which are actually moving in space at a speed of about 4 km/s. The position of a satellite at any instant, however, can be estimated with an error no larger than a few meters based on predictions made 24–48 hours earlier. The distance between the user and a satellite is measured in terms of transit time of the signal from the satellite to the user. The transmission times are imprinted upon the signals in accordance with nearly perfect and nearly perfectly synchronized atomic clocks carried aboard the satellites. The precise estimation of the arrival times is made possible by transmitting spread spectrum signals, which have wide bandwidths but each satellite can transmit its unique signal on the common frequency band. Finally, the spectacular developments in microelectronics have made the receiver light, compact, and an order-of-magnitude less inexpensive than thought possible in 1980.

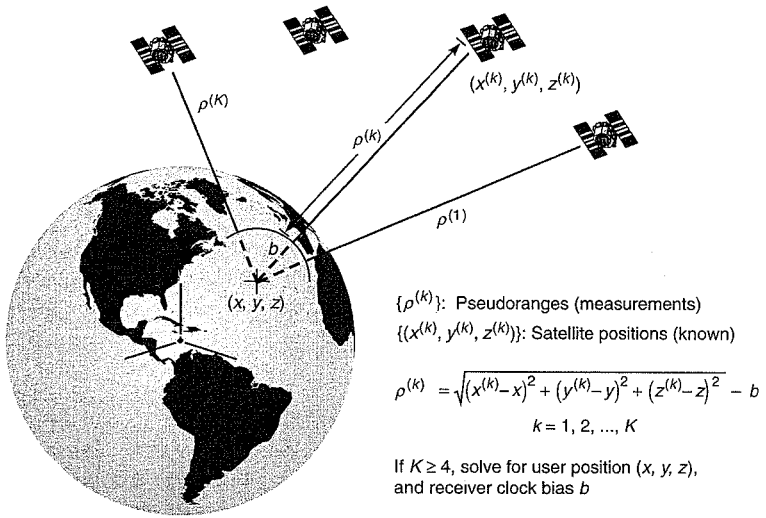


Figure 1.9 The principle of satellite navigation. The user-satellite range measurements based on the times of transmission and receipt of signals are biased by a common amount and are called pseudoranges. Pseudorange measurements are needed from at least four satellites to estimate the user position and receiver clock bias.

In order to measure the true transit time of a signal from a satellite to a receiver, clearly, the clocks in the satellite and the receiver must be maintained in synchronism. Fortunately, this onerous requirement is easily sidestepped, allowing use of inexpensive quartz oscillators in the receivers. The bias in the receiver clock at the instant of the measurements affects the observed transit times for all satellites equally. The corresponding measured ranges are thus all too short, or too long, by a common amount, and are called *pseudoranges*. The receiver clock bias thus becomes the fourth unknown to be estimated, in addition to the three coordinates of position. A user therefore needs a minimum of four satellites in view to estimate his four-dimensional position: three coordinates of spatial position, plus time. An idealized geometrical view of the pseudorange measurements and the resulting equations to be solved for the user position and receiver clock bias are given in Figure 1.9.

Table 1.2 A performance summary of radionavigation systems

System	Coverage		Fix dimensions	Positioning accuracy (rms)
	Worldwide	Continuous		
Loran-C	No	Yes	2-D	250 m
Omega	Yes	Yes	2-D	2–4 km
Transit	Yes	No	2-D	25 m
GPS	Yes	Yes	3-D plus time	Horizontal: 5 m Vertical: 7.5 m

We have now introduced GPS and the radionavigation systems which were its progenitors: Loran-C, Omega, and Transit. The performance characteristics of these systems are summarized in Table 1.2. The positioning accuracy numbers are based on FRP (1994).

GNSS

GPS is the first of a new generation of navigation satellite systems to become operational. Similar systems and satellite-based augmentations are being developed and deployed by governments, international consortia, and commercial interests. The generic name given to these systems is *Global Navigation Satellite Systems* (GNSS). A brief survey appears below. An expanded discussion is found in Chapter 3.

While GPS was under development, the Soviet Union undertook to develop a similar system called *GLONASS*. GLONASS, like GPS, was designed primarily for the military, and a subset of its signals was offered for civil use in the era of *glasnost* and *perestroika* in the late 1980s, apparently as an afterthought. Since the dissolution of the Soviet Union, the Russian Federation has assumed responsibility for GLONASS. The civil user community was excited at the prospect of having access to two autonomous systems as GLONASS built up to a full constellation of 24 prototype satellites in 1996. Unfortunately it didn't last [Langley (1997)]. This once high-prestige system has suffered from lack of resources in the changed political and economic climate, and appears to have been kept going in skeletal form by a dedicated GLONASS cadre. Between 2000 and 2005, the Russians managed only one launch per year, each placing three new satellites in orbit. While a couple of GPS+GLONASS receiver models have been on the market since the mid-1990s, the user community and receiver manufacturers have seemed wary of GLONASS. But there appears to be new optimism in 2005 among GLONASS watchers [Section 3.5].

A more promising GNSS now under development is the European system, Galileo, planned as "an open, global system, fully compatible with GPS, but independent from it." After several years of intense discussions and serious haggling among the member states, the European Union (EU) gave the final green light in 2002 to proceed with the development of the system. The EU has two pressing reasons for this initiative. First, concerns of sovereignty and security require that Europe must control its safety-critical navigation systems, and EU has no control over GPS. Secondly, this step is intended to help the EU countries become players in

the technology and commerce associated with satellite navigation. As was to be expected, the United States, with a vested interest in maintaining the preeminence of GPS, was not enthusiastic about the idea of a rival, but now appears to have made peace with it.

Galileo was conceived as a joint public-private enterprise under civilian control, to be financed and managed by the European Commission (EC), the European Space Agency (ESA), and industry. The estimated cost of developing and deploying the system is €3.2 billion. While the industry was slow to commit funds in the early phases of the program, EC and ESA committed over €1.1 billion in 2002 to proceed with the development phase of the system. The deployment phase will begin in 2006. The system could be operational in 2008, though 2010 appears more realistic [Section 3.6].

Galileo would be a second-generation GNSS, following the trail blazed by GPS and benefiting from the technological advances since the mid-1970s when GPS was designed. As we'll see in Chapter 3, GPS isn't sitting still either. All agree that Galileo and GPS must be compatible and interoperable. The United States, fearing that Galileo could complicate the situation with regard to selective denial of satellite navigation service to an adversary, pressured the Europeans with a demand that the Galileo signals must also be 'jammable' without harm to GPS military signals.

China has developed a regional satellite navigation system called *BeiDou*. They launched two GEO satellites late in 2000 and another in May 2003, apparently completing the system which provides positioning capability over China and neighboring areas. BeiDou (Big Dipper-1) is an active system, which, unlike GPS, GLONASS, and Galileo, interacts with the users. The Mission Control Center determines a position estimate and transmits it to each user. The system accuracy is reported to be "dozens of meters."

The Japanese government in conjunction with an industry consortium is undertaking development of *Quasi-Zenith Satellite System* (QZSS), a regional system to transmit ranging signals over Japan from satellites in highly elliptical orbits, and to transmit differential corrections to GPS and other GNSS satellites. The first satellite launch is planned for 2008, with two more to follow. The cost of development, deployment, and maintenance of the system is expected to be ¥170 billion ($1.57 billion) over twelve years. Other augmentation systems are described briefly in Chapter 2.

The idea of satellite navigation is not limited to the earth. After two recent high-profile failures in the Mars program, NASA has considered a system of low-orbiting satellites around Mars to provide GPS-like navigation signals and a better communications link to the control center.

GPS is likely to remain the only operational system of its kind at least until 2008, when Galileo is planned to be ready. GPS, however, is evolving into a more capable system. Satellites to be launched in 2005 and thereafter will broadcast additional signals. The completed system with considerably expanded capabilities for both military and civil users is expected to emerge between 2015 and 2020. The plans for GPS modernization are described in Chapter 3.

1.4 Summary

In this chapter, we surveyed the history of navigational tools and technology developed over centuries. Especially noteworthy are the efforts to find longitude at sea. There are heroic tales of craftsmen who built clocks with great mechanical ingenuity to keep time at sea, and the

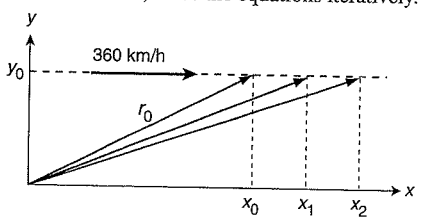
competing astronomers who preferred to measure lunar distance from a star and then do a prodigious amount of calculation by hand. It's all fascinating stuff, and we literally ran through it in a half-dozen pages. The twentieth century brought new technologies of radio and inertial navigation with enormous consequences for war and commerce. To all who would look impatiently at their wristwatches, cell phones, PDAs, or car navigation systems with moving maps for directions to their destinations, we recommend J.E.D. Williams' wonderful monograph for balance and perspective. As for GPS, we'll get plenty of it in the 500-plus pages that follow.

Homework Problems

1-1. Given that 1 minute of latitude is approximately equal to 1 nautical mile (1852 meters), how many significant digits after the decimal must be included for a latitude represented in *degrees* to describe a fix that is accurate to 1 cm? How many significant digits are required after the decimal in the *arc-seconds* field if the latitude is represented in degrees, arc-minutes, and arc-seconds to describe a fix that is accurate to 1 cm? Note: 1 degree = 60 arc-minutes, 1 arc-minute = 60 arc-seconds (you may find arc-minutes referred to as 'minutes' and arc-seconds referred to as 'seconds').

1-2. Suppose you start from location 45° N, 120° W (lat, long) and fly at an altitude of 10 km with ground speed of 885 km/h for eight hours at a constant heading of 45° from true north. Where would you end up? Assume a spherical earth with radius of 6371 km. Two helpful notes: (i) Ground speed in aviation means the speed of an aircraft relative to the surface of the earth (to be distinguished from air speed, which means the speed of an aircraft relative to its surrounding air mass). (ii) It's safe to say that you'll end up pretty far north. Don't worry too much about precision—you get full credit if your answer is within a kilometer of the exact answer.

1-3. An aircraft is carrying a transmitter that is broadcasting a single tone at 100 MHz. The aircraft flies away from you on a straight line at constant altitude with ground speed of 360 km/h. You measure the following Doppler shifts from 100 MHz spaced 0.1 s apart: –33.1679 Hz, –33.1711 Hz, and –33.1743 Hz. Determine the range rates in m/s that correspond to these Doppler measurements. In the figure below, the aircraft altitude is y_0, and its horizontal distance from observer at the three instants of Doppler measurements is shown as x_0, x_1, and x_2, respectively. Set up two linear equations that relate x_1 and x_2 to x_0. Can you set up two non-linear equations that relate x_0 and y_0 to the measurements? For extra credit, solve the equations iteratively.



1-4. A *pseudolite* (short for pseudo-satellite) consists of a generator of a GPS-like signal and a transmitter. Pseudolites are used to augment the GPS signals. Suppose an observer is constrained to be on the line joining two pseudolites PL1 and PL2, which are separated by 1 km. The pseudolite clocks are perfectly synchronized but the observer's clock may have an unknown bias with respect to the pseudolite clocks. Estimate the observer's position and clock bias given that the pseudoranges to PL1 and PL2 are (a) 550 m and 500 m, respectively, and (b) 400 m and 1400 m, respectively.

1-5. You set out from town A and head east to town B 120 km away. Your vehicle has an odometer that is not particularly accurate (it could be off by 1–2 km, or more, after driving 50 km). You carry a decent watch, which is great at keeping time over short intervals, but it has been months since you last reset it. In other words, you can measure time intervals accurately, but do not know exactly what time it is. A short time into your journey, the car breaks down.

(a) According to your odometer, you have traveled 56 km. Estimate your position.

(b) As you push your car to the shoulder of the road, a red bus zooms by heading from A to B. You glance at your watch and notice that it is exactly 21 minutes past the hour. You know that the red buses are prompt, and they leave town A every hour on the hour traveling at exactly 3 km/min. Can you estimate your position without using the odometer information?

(c) Estimate your position and clock bias based on all the information so far. (Hint: Write two equations that relate your position and clock bias to the available information. These equations are sometimes referred to as navigation equations.)

(d) At 25 minutes past the hour by your watch, you observe a blue bus zoom past at 2.5 km/min, going from B to A. Blue buses leave town B every hour on the hour promptly and drive to town A at a constant speed. Estimate your position and clock bias based on all the information so far. (Hint: Recall that your watch can measure time intervals accurately.)

(e) How would your solution be affected if your watch were exactly five minutes faster and all the clocks in town A and town B were running five minutes fast?

(f) Now suppose that your odometer never worked and the only vehicles you see are the identical yellow cabs of carrier L1. These cabs leave town A every minute on the minute and travel at exactly 1 km/min to town B. Can you estimate your position and clock error? Would it help if there were identical green cabs of carrier L2 leaving town B every minute on the minute and traveling at 1 km/min to town A? Explain briefly.

References

Allan, David W., Neil Ashby, and Clifford C. Hodge (1997). *The Science of Timekeeping*, Application Note 1289, Hewlett-Packard Corporation.

Bowditch, Nathaniel (1802). *The American Practical Navigator*, issued by the Defense Mapping Agency Hydrographic Office, Washington, DC, 1984.

Brown, Lloyd A. (1956). The Longitude, in *The World of Mathematics*, J.R. Newman (ed.), Simon and Schuster, pp. 780–819.

Easton, R.L. (1980). The Navigation Technology Program, in *Global Positioning System, Vol. I*, The Institute of Navigation, pp. 15–20.

Enge, Per, Eric Swanson, Richard Mullin, Ken Ganther, Anthony Bommarito, and Robert Kelly (1995). Terrestrial Radio Navigation Technologies, *Navigation*, vol. 42, no. 1, pp. 61–108.

Farrell, Jay A., and Matthew Barth (1999). *The Global Positioning System and Inertial Navigation*, McGraw Hill.

Federhen, H.M. (1993). *Worldwide Radionavigation Systems*, IDA Paper P-2852, Institute for Defense Analysis, Alexandria, VA.

Forssell, Börje (1992). *Radionavigation Systems*, Prentice Hall.

FRP (1994). *Federal Radionavigation Plan*, U.S. Departments of Defense and Transportation.

Getting, Ivan A. (1993). The Global Positioning System, *IEEE Spectrum*, vol. 30, no. 12, pp. 36-47.

Greenspan, Richard L. (1995). Inertial Navigation Technology from 1970 to 1995, *Navigation*, vol. 42, no. 1, pp. 165–185.

Howse, Derek (1980). *Greenwich Time and the Discovery of the Longitude*, Oxford University Press.

IEEE (1983). *Proceedings of the IEEE*, Special Issue on Global Navigation Systems, vol. 71, no. 10.

Kayton, Myron (ed.) (1989). *Navigation: Land, Sea, Air, and Space*, IEEE Press.

Langley, Richard B. (1997). GLONASS: Review and Update, *GPS World*, vol. 8, no. 7, pp. 46–51.

Loran-C (1992). *Loran-C User Handbook*, U.S. Coast Guard.

Omega (1983). *Omega Global Radio Navigation, A Guide for Users*, U.S. Coast Guard.

Parkinson, Bradford W. (1996). Introduction and Heritage of NAVSTAR, the Global Positioning System, in *Global Positioning System: Theory and Applications I*, B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 3–28.

Snyder, John P. (1987). *Map Projections—A Working Manual*, U.S. Geological Survey Professional Paper 1395, U.S. Government Printing Office.

Stansell, Thomas A. (1983). *The TRANSIT Navigation Satellite System*, Magnavox Technical Report, R-5933A.

Williams, J.E.D. (1992). *From Sails to Satellites: The Origin and Development of Navigational Science*, Oxford University Press.

Chapter 2
GPS in 2005: An Overview

Our objective in this chapter is quite ambitious. We aim to provide a compact, technical overview of GPS as it exists in the summer of 2005: the system, signals, measurements, performance, and applications. The subsequent chapters pick up many of the main ideas introduced herein and provide a comprehensive treatment. As such, this chapter is intended to serve both as an introduction to GPS and a road map to this book.

The satellites now scheduled for launch starting in September 2005 will broadcast additional signals for the military and civil users. But it'll be a half-dozen years before there are enough satellites broadcasting the new signals to make them widely useful. We describe these new signals in Chapter 3, where we also introduce the signals and services planned for GLONASS and Galileo.

2.1 Objectives, Policies, and Status

The principal objective of the U.S. Department of Defense (DoD) in developing GPS was to offer the U.S. military accurate estimates of position, velocity, and time. In quantitative terms, this statement was interpreted broadly as providing estimates with position error of 10 meters, velocity error of 0.1 meters/second, and time error of 100 nanoseconds, all in the root-mean-square (rms) sense. These estimates were to be available to an unlimited number of users all over the globe continuously and nearly instantaneously. The planned military use also required the system to be usable on high-dynamics platforms, and the signals to have a measure of resistance to jamming and interference. And, finally, the adversaries of the United States were to be denied the full benefits of the system.

The DoD also planned to provide the civil users of GPS a 'reasonable' accuracy consistent with the national security considerations. These considerations were formalized in a policy to offer two kinds of service:

- *Standard Positioning Service* (SPS) for peaceful civil use,
- *Precise Positioning Service* (PPS) for the DoD-authorized users.

Access to the full capabilities of the system (i.e., PPS) is restricted by cryptographic techniques. The system transmits encrypted signals intended for DoD-authorized users equipped with the appropriate decryption keys. This feature is called *Anti-Spoofing* (AS). Each satellite transmits one signal for civil use under SPS. It rankled the civil user community that the SPS signal was degraded throughout the 1990s by introducing controlled errors to reduce its precision and limit the civil users to accuracy of 100 meters. Such errors could be removed by the DoD-authorized users. This feature, called *Selective Availability* (SA), was deactivated by a Presidential Order on 2 May 2000. Both SA and AS are discussed below.

We should note as an aside that there was considerable tension between DoD and the potential civil users in the early years of GPS. The U.S. military at first appeared reluctant—even hostile—to the idea of sharing their marvelous new toy. During 1980–1981, when Charles C. Counselman III (MIT) and Peter MacDoran (Jet Propulsion Laboratory) built receivers and demonstrated that GPS could be used for precise positioning even if the signals were encrypted, they were required to register with the U.S. Government as arms manufacturers. Happily, the situation changed, albeit slowly, and the present relationship between the U.S. military and the civil users can be described as harmonious.

The policy for civil use of GPS was first announced by the DoD in the late 1970s with an assurance that the SPS signals will remain freely available. In 1983, Korean Airline Flight 007 went off course into Soviet airspace, apparently due to navigation problems, and was shot down. This disaster drew attention to the potential benefits of GPS for civil aviation, and the U.S. policy on civil use of GPS was reaffirmed by President Reagan. At the time, the system

was ten years away from being declared operational. Subsequently, the United States made a formal commitment in 1991 to the International Civil Aviation Organization (ICAO), a specialized agency of the United Nations, to make "GPS-SPS available for the foreseeable future on a continuous, worldwide basis, and free of direct user fees."

While free signals and growing markets for GPS products and services created general enthusiasm in the civil sector worldwide, the governments, particularly in Europe, were reluctant to dismantle the existing terrestrial navigation aids (navaids) and become dependent upon a system controlled by the military of a foreign government. Many held the view that technology as basic and vital as GPS required an international institutional framework for development of policies, regulations, and standards, but the U.S. Government wasn't about to share control of GPS. The European Union (EU) therefore decided to develop a satellite navigation system of their own, to be called *Galileo*. As was to be expected, the U.S. Government was not enthusiastic about this idea, but appears now to have made peace with it.

The U.S. policy on GPS is based on balancing the basic requirement of retaining the military advantage of this technology with considerations of commerce and international policy. An interesting account of the issues and considerations is found in the report of a study conducted in 1994–1995 at the direction of the U.S. Congress by the National Academy of Public Administration (NAPA) and the National Research Council (NRC) of the National Academy of Sciences and Engineering [NAPA-NRC (1995)]. This report has been influential in shaping the evolution of both GPS and the U.S. policy on its use. The report concluded that SA was not serving its intended purpose, and recommended that it be discontinued immediately. The report recommended that "immediate steps should be taken to obtain authorization to use an L-band frequency for an additional GPS signal" for civil use. The report also recommended several measures on governance of GPS to achieve the national goals and to promote its international acceptance.

The Presidential Decision Directive (PDD) of 1996 [US GPS Policy (1996)] adopted most of the recommendations of the NAPA-NRC report, and paved the way for discontinuance (in 2000) of SA. According to the PDD, one of the U.S. policy objectives is to promote integration of GPS into peaceful civil, commercial, and scientific applications worldwide, and to advocate acceptance of GPS as an international standard. Accordingly, changes were made in the governance of GPS. The management responsibility was assigned to the Interagency GPS Executive Board (IGEB), co-chaired by representatives of the DoD and the Department of Transportation (DOT). But that didn't mean DoD and DOT had the same clout in managing GPS. The funds to maintain and modernize GPS, and to operate IGEB, all came out of the DoD budget. We are reminded of an old proverb, something about paying the piper and calling the tune.

The Presidential Decision Directive of 1996 was superseded in December 2004 by the U.S. Space-Based Positioning, Navigation, and Timing Policy [US PNT Policy (2004)]. The new complexities are reflected in the fact that while the PDD barely filled three pages, the PNT Policy takes up 11 pages. There is no sharp departure in policy, only more things to address, especially on security. Navigation warfare [Section 3.4], not mentioned in the 1996 document, makes repeated appearances. The new policy includes a restatement of issues of free and open use of civil signals, denial of hostile use, and a clear statement of cooperation with Galileo. The new policy identifies GPS as a component of the "U.S. Critical Infrastructure" to give the system a special status for the purpose of funding under U.S. laws. A National Space-based Positioning, Navigation, and Timing Executive Committee, co-chaired by the DoD and

DOT, will take over the responsibilities of IGEB and may project a higher profile. Its members will include representatives from the various Departments and Agencies of the Federal Government [http://www.pnt.gov/]. FRP (2005) is the official source of the U.S. radionavigation policy and plans.

GPS was declared operational in 1995 and, since then, it has performed virtually flawlessly, offering a service better than that promised in its performance specifications [Section 2.4.2]. The system emerged unscathed from two "rollover" events. First, there was the end-of-week rollover (EOW or W1K) in August 1999 [Section 4.2.4]; the second was the Y2K event. It is to the credit of the U.S. Air Force Space Command, which is responsible for launching the satellites and operating the system, that glitches due to equipment malfunction and operational lapses have been rare. The United States appears committed to providing GPS SPS without interruption or degradation, but may deny the service in selected geographical areas if required by military needs.

GPS has now proven its worth as a military system in three conflicts. The first test came in the Gulf War (1990–1991) in which GPS helped ground troops navigate in the featureless desert while the system only had a partial satellite constellation and was years away from being declared ready for operational use. In Kosovo in 1999, GPS-guided munition ("smart bombs") appear to have changed the rules of war by proving that a war can be won with minimum risk of casualties for the winning side through an air campaign by targeting bombs with care and delivering them with precision so as to minimize collateral damage on the ground (a euphemism for civilian deaths). And, in Iraq War (2003), GPS was critical to the campaign of "shock and awe," guiding and tracking aircraft, ships, missiles, armored vehicles, and troops. Based on accounts in the press, the Iraq War also saw the first attempt, albeit clumsy and futile, at jamming GPS signals.

Literature on GPS has grown to be voluminous in recent years, and includes several influential collections of papers and books offering in-depth treatments of the system and its applications from different viewpoints [El-Rabbany (2002), Farrell and Barth (1999), Forssell (1992), Grewal *et al.* (2001), Hoffmann-Wellenhof *et al.* (1997), IEEE (1999), ION (1980), Kaplan (1996), Leick (2004), Parkinson *et al.* (1996), Seeber (2003), Strang and Borre (1997), Teunissen and Kleusberg (1998), Tsui (2004), Van Dierendonck (1997), Wells (1986), and Xu (2003)]. We owe a debt to the authors of these books and papers which have influenced the account that follows.

2.2 System Architecture

GPS consists of three segments: the Space Segment, the Control Segment, and the User Segment. The Space Segment comprises the satellites and the Control Segment deals with the management of the satellite operations. The DoD is responsible for both. The User Segment covers activities related to the development of military and civil GPS user equipment (i.e., receivers). The development of receivers and services in the civil sector is essentially left to market forces. The civil sector, however, clearly benefited from the investments by the DoD in the 1970s and 1980s in development of military receivers. This favor is now being returned as the design of a new generation of GPS military user equipment benefits from the extraordinary developments in the 1990s in civil receiver design and manufacturing.

	Block II/IIA	Block IIR	Block IIF
Number	28	21	12*
First launch	1989	1997	2007*
Satellite weight (kg)	1000	1100	1700*
Power/Solar panels (W)	700	1140	2900*
Design life (yrs)	7.5	10	12.5
Unit cost	$43M	$30M	$28M*

* Estimates

Figure 2.1 The GPS Space Segment consists of a baseline constellation of 24 satellites distributed in six orbital planes. Each new batch of satellites was designed to provide additional capabilities and, it should be noted, came with a lower price tag.

2.2.1 Space Segment

The baseline constellation comprises 24 satellites fielded in nearly circular orbits with a radius of 26,560 km, a period of approximately twelve hours, and stationary ground tracks. The constellation is shown in Figure 2.1. The satellites are arranged in six orbital planes inclined at 55° relative to the equatorial plane, with four primary satellite slots distributed unevenly in each orbit. Each satellite is identified with a two-character code: a letter identifies the orbital plane (A through F); a number identifies the satellite number in the plane (1 through 4 in the baseline constellation). There is a spare satellite slot in each orbital plane, designated by the two-character code as A5, B5, etc. The system can support a constellation of up to thirty satellites on orbit.

With the baseline constellation, almost all users with a clear view of the sky have a minimum of four satellites in view. It's more likely that a user would see six to eight satellites. The satellites broadcast ranging signals and navigation data allowing the users to measure their pseudoranges and to estimate their positions in passive, listen-only mode.

An initial batch of ten prototype or developmental satellites, called Block I satellites, was launched between 1978 and 1985, and used to demonstrate the feasibility of GPS. The prototypes were followed by production models named Block II and IIA, the first launch of which occurred in February 1989. Over the next six years, the system built up to 24 Block II/IIA satellites on orbit, and was declared operational in April 1995. The next generation of satellites,

called Block IIR (the 'R' stands for replenishment), entered service in 1997. The Block IIR satellites now on order will sustain the constellation at least until 2007.

A new generation of GPS satellites, called Block IIF (the 'F' stands for follow-on), is now in production, with delivery of the first six planned for 2007. The DoD had contracted in 1996 with Rockwell International (now a part of The Boeing Company) for six Block IIF satellites with an option to buy up to 27 more. Early in 2000, the DoD announced its decision to change course. It appears now that the DoD plans to acquire only twelve IIFs. The decision was apparently motivated by changes since 1996 in the requirements under the GPS modernization plan. As we'll discuss in Chapter 3, the modernization plan calls for two new civil signals and two new military signals. These changes would have required extensive modifications to the Block IIF design. The DoD will now start with a clean slate the process of acquiring a new generation of GPS satellites to be called GPS III.

The table in Figure 2.1 summarizes the basic features of Block II/IIA, IIR, and IIF satellites. The satellite weight refers to on-orbit weight. The power generated by the solar panels refers to levels early in mission life. Each successive batch of GPS satellites was designed with higher capabilities, longer service life and, it is worth noting, a lower price tag [Fisher and Ghassemi (1999)]. A satellite in orbit represents (in round numbers) a $100 million asset, counting the cost of building and launching it

GPS SPS doesn't promise or require a 24-satellite constellation and the United States retains some latitude on maintenance of the satellite constellation. "The intent of the U.S. Government is to manage the constellation such that it never drops below 22 healthy satellites in nominal slots" [SPS (2001)]. Actually, the number of active satellites on orbit since 1995 has exceeded 24, and there have been serious proposals to raise the constellation strength to 27, or higher.

Early in September 2005, the constellation comprised 29 satellites: 17 Block II/IIAs and 12 Block IIRs. The number of working satellites at any time can vary due to the requirements of maintenance. Satellites are taken off-line for short periods, announced in advance. Sudden catastrophic failures can reduce the constellation size as well, but the GPS satellites have been consistently reliable, and their life expectancy so far has exceeded design life significantly. Out of ten Block II/IIA satellites (design life: 7.5 years) launched in 1989–1990, all but one were in service at the beginning of 2000, and two made it to 2005.

GPS constellation status and scheduled outage updates are available from the website of the U.S. Coast Guard Navigation Center (NAVCEN) [http://www.navcen.uscg.gov].

2.2.2 Control Segment

At the heart of the Control Segment is the Master Control Station (MCS), located at the Schriever (formerly named Falcon) Air Force Base near Colorado Springs, Colorado. The Master Control Station operates the system and provides command and control functions. The specific functions of the Control Segment are:

- to monitor satellite orbits,
- to monitor and maintain satellite health,
- to maintain GPS Time,
- to predict satellite ephemerides and clock parameters,
- to update satellite navigation messages,

- to command small maneuvers of satellites to maintain orbit, and relocations to compensate for failures, as needed.

The satellite signals are tracked from U.S. Air Force monitor stations spread around the globe in longitude: Hawaii, Colorado Springs, Cape Canaveral, Ascension Island, Diego Garcia, and Kwajalein. In 2005, this network was expanded to include six monitor stations operated by the National Geospatial-Intelligence Agency (NGA) of the DoD: Washington, D.C., United Kingdom, Argentina, Ecuador, Bahrain, and Australia. The twelve-station network allows the system operators to watch each satellite from at least two monitor stations at all times. Five more NGA-operated monitor stations will be added to the network in the future.

The monitor stations are unmanned and are operated remotely from the Master Control Station. The Control Segment monitors the PPS signals and removes a satellite from service if anomalous behavior is detected. Equipment at a monitor station consists essentially of GPS receivers with cesium standards, meteorological instruments, and communications gear to transmit the measurements to the Master Control Station via ground and satellite links.

Dedicated ground antennas for communications with the satellites via S-band radio links are co-located with monitor stations at Ascension Island, Cape Canaveral, Diego Garcia, and Kwajalein. These ground antennas are operated remotely from the Master Control Station to receive telemetry from the satellites on the status of its subsystems and functions, to uplink commands, and to upload data to update the navigation messages broadcast by the satellites. [The Air Force Satellite Control Network (AFSCN) Automated Remote Tracking Station (ARTS) at Schriever Air Force Base can also be operated as a GPS ground antenna for command and control.]

The elements of the Control Segment (monitor stations, ground antennas, Master Control Station) and their functions are shown in Figure 2.2. The Control Segment is also referred to as the Operational Control Segment or Operational Control System, both abbreviated as OCS.

The estimation and prediction of satellite orbits and clock biases are based on data from the monitor stations. GPS Time [Section 4.2.4] is defined on the basis of a set of atomic clocks in the satellites and monitor stations. Synchronization of the satellite clocks is accomplished by estimating the time offset, drift, and drift rate of each satellite clock relative to GPS Time, and transmitting the parameters of a model of this bias in the satellite's navigation message. These parameters are a part of the navigation message transmitted by each satellite [Section 4.3.6]. The ephemeris and clock parameters broadcast by the satellites are computed by the Master Control Station and uploaded to the satellites via one of the ground antennas.

The navigation messages broadcast by the satellites are currently uploaded once daily at a minimum. The cross-link communication and ranging capability in Block IIR satellites was designed to allow the satellites to update their navigation messages autonomously and operate over extended periods without contact with the Control Segment. This Autonav function appears not to be exercised and may not be operative.

GPS Coordinate Frame and Time Reference

While on the subject of infrastructure required to support GPS operations, we mention two important ideas briefly and develop them further in Chapter 4. First, in order to represent a position or velocity, we need a coordinate frame. Before the advent of the satellite navigation systems, such coordinate frames were defined on a country or regional basis. A global positioning

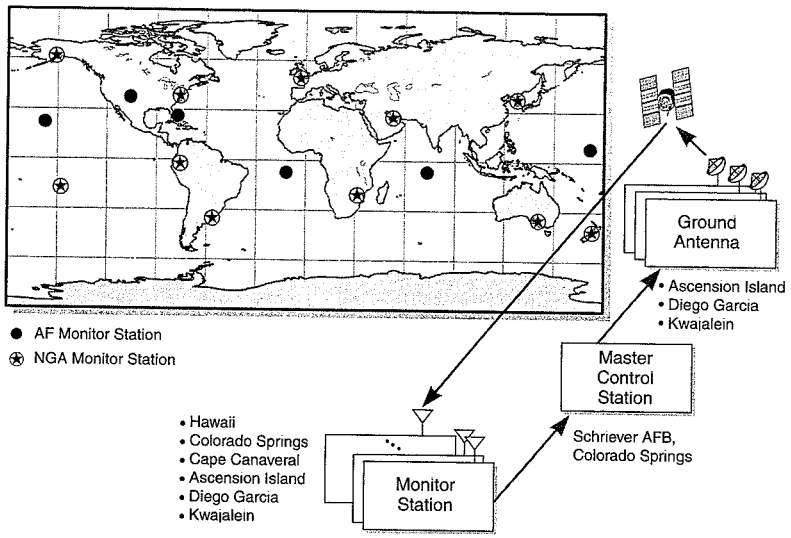


Figure 2.2 The elements of the GPS Control Segment and their functions.

system requires a global coordinate frame in which to express the position of its satellites and the users. GPS provided the impetus for development by the Defense Mapping Agency (DMA) of the DoD of the World Geodetic System 1984 (WGS 84), which has now become a *de facto* international standard, a great accomplishment in itself [Section 4.1]. DMA was subsequently reorganized and renamed the National Imagery and Mapping Agency (NIMA) and, in 2003, the National Geospatial-Intelligence Agency (NGA).

Clocks are at the heart of GPS. A receiver measures ranges to satellites by determining the transit times of signals generated aboard the satellites by synchronized clocks and carrying precise timing marks. Roughly speaking, meter-level positioning accuracy requires measurement of ranges accurate to meter level, which, in turn, requires nanosecond-level synchronization of the satellite clocks and measurement of transit time at nanosecond level. Our second important idea, therefore, deals with the definition and measurement of time. GPS defines a time reference of its own called GPS Time (GPST) and keeps track of the offset between GPST and the international civil standard called UTC [Section 4.2].

2.2.3 User Segment

The success of GPS in large-scale civil use is attributable almost entirely to the revolution in integrated circuits, which has made the receivers compact, light, and an order-of-magnitude less expensive than thought possible twenty years ago. The first receivers designed for precise positioning were introduced in the mid-1980s, and were priced at over $100,000. Receivers with much higher capabilities are now available for less than $5000. As late as 1980, it had only been hoped that the receiver manufacturers would be able to produce a basic GPS receiver for the mass market for about $2000. Price barriers, however, fell quickly. An important

industry milestone was reached in 1992 with the introduction of the first hand-held receiver priced below $1000. In 1997, the industry breached the $100 barrier with a pocket-size receiver running on two AA batteries.

Commerce in GPS-related products and services grew rapidly in the 1990s. There are now hundreds of GPS receiver models on the market. It is estimated that more than one million receivers have been produced each year since 1997. According to the U.S. Department of Commerce, the annual worldwide sales of GPS products and services exceeded $16 billion in 2003. GPS is on its way to becoming a part of our daily lives as an essential element of the commercial and public infrastructure.

While the GPS industry is yet to produce its Bill Gates and Michael Dell, there is money to be made in this business, as shown by Gary Burrell and Min Kao, the co-founders of Garmin International, manufacturer since 1991 of GPS receivers for the consumer and aviation markets. Both made the Forbes magazine's list of "400 Richest People" in 2003 as near-billionaires.

In the last twenty years, several generations of receivers based on newer technologies have been produced, but we are not ready to discuss the receivers yet. First, we have to understand the nature of the signals, which we take up next.

2.3 Signals

Currently, each GPS satellite transmits continuously using two radio frequencies in the L-band referred to as Link 1 (L1) and Link 2 (L2). The L-band covers frequencies between 1 GHz and 2 GHz, and is a subset of the ultra-high frequency (UHF) band. The center frequencies of L1 and L2 are as below:

$$\text{L1: } f_{L1} = 1575.42 \text{ MHz, and L2: } f_{L2} = 1227.60 \text{ MHz.}$$

Two signals are transmitted on L1, one for civil users, and the other for DoD-authorized users. The lone signal on L2 is intended for the DoD-authorized users only. The structure of these signals is discussed briefly below. [Actually, the satellites transmit additional RF signals at frequencies referred to as L3 and L4. These signals are associated with classified payloads aboard the satellites: L3 is associated with the Nuclear Detonation Detection System (NDS or Nudet), and L4 is associated with a payload called Reserve Auxiliary Package (RAP).]

Under the GPS modernization program, additional navigation signals will be broadcast by Block IIR-M (the "M" is for modernized) satellites to be launched starting in 2005 and by Block IIF satellites to be launched starting in 2007. These signals are described briefly in Chapter 3.

2.3.1 Signal Structure

The structure of the GPS signals is described exhaustively in Interface Specifications. The Interface Specification for the SPS signals is public [IS-GPS-200D (2004)]. Each GPS signal consists of three components (see Figure 2.3):

- *Carrier*: RF sinusoidal signal with frequency f_{L1} or f_{L2}
- *Ranging code*: Associated with each Service (i.e., SPS and PPS) is a family of

binary codes called *pseudo-random noise* (PRN) sequences or PRN codes. The PRN codes have special mathematical properties which allow all satellites to transmit at the same frequency without interfering with each other [Chapter 9]. These codes also allow precise range measurements, and mitigate the deleterious effects of reflected and interfering signals received by a GPS antenna [Chapter 10]. The SPS codes are called coarse/acquisition codes (C/A-codes) and PPS codes are referred to as precision (encrypted) codes, or P(Y)-codes. Each satellite transmits a unique C/A-code on L1 and unique P(Y)-codes on both L1 and L2.

Each C/A-code is a unique sequence of 1023 bits, called *chips* in this context, which is repeated each millisecond. The duration of each C/A-code chip is about 1 μs. Equivalently, the *chip width* or wavelength is about 300 m. The rate of the C/A-code chips, called *chipping rate*, is 1.023 MHz [or megachips/second (Mcps)]. A coarse/acquisition code is so called because (i) it's coarse as compared to a precision code, and (ii) it was originally intended to serve as a necessary stepping stone for acquisition of a P(Y)-code. Newer technology now allows direct acquisition of the P(Y)-code.

A P-code is a unique segment of an extremely long ($\approx 10^{14}$ chips) PRN sequence. The chipping rate is 10.23 Mcps, ten times that for a C/A-code, and the chip width is about 30 m. The smaller wavelength results in greater precision in the range measurements than that for the C/A-codes. The P-codes repeat after one week. Actually, since 1994, the satellites have transmitted encrypted versions of the P-codes, called Y-codes, about which we know nothing.

We note here in passing that the length of a PRN code and its chipping rate determine the acquisition and tracking characteristics of the signal in the presence of noise and interference [Sections 3.3, 9.3, 9.4].

- *Navigation data*: a binary-coded message consisting of data on the satellite health status, ephemeris (satellite position and velocity), clock bias parameters, and an almanac giving reduced-precision ephemeris data on all satellites in the constellation. The navigation message is transmitted at a leisurely 50 bits per second (bps), with a bit duration of 20 ms. It takes 12.5 minutes for the entire message to be received. The essential satellite ephemeris and clock parameters are repeated each thirty seconds [Section 4.3.6]. A faster data rate would get the message out quicker, but would require a stronger signal to ensure demodulation with low error rate.

These three components of a signal (namely, carrier, code, and navigation data) are derived coherently from one of the atomic standards aboard the satellite. The frequency of the atomic standards aboard a satellite is 10.23 MHz. The relationship with chipping rates is obvious. Note also the relationship with the L1 and L2 carrier frequencies:

$$f_{L1} = 1575.42 \text{ MHz} = 2 \times 77 \times 10.23 \text{ MHz}$$
$$f_{L2} = 1227.60 \text{ MHz} = 2 \times 60 \times 10.23 \text{ MHz}$$

The code is combined with the binary navigation data using modulo-2 addition: If the code chip and the data bit are the same (both are 0s or both are 1s), the result is 0; and if both are

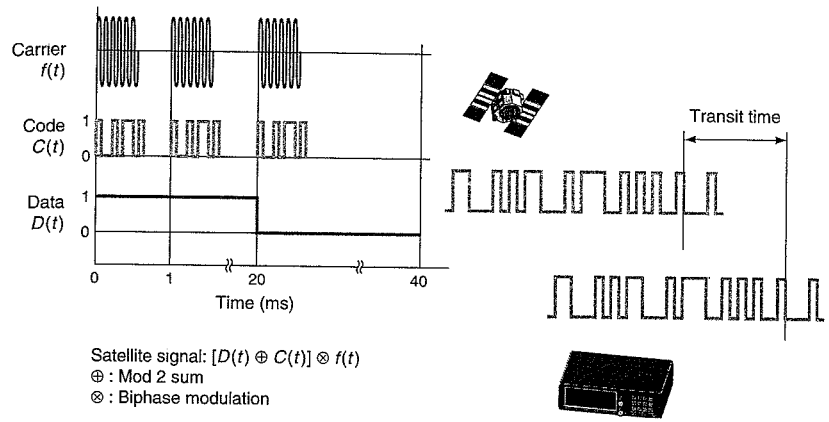


Figure 2.3 The structure of the signal available for civil use and estimation of its transit time from the satellite to a user. Each GPS signal comprises three components: an RF carrier, a unique binary pseudo-random noise (PRN) code, and a binary navigation message. The transit time is estimated by correlating the received signal with its replica generated by the receiver.

different, the result is 1. The composite binary signal is then impressed upon the carrier in a process called modulation. The specific form of modulation used is called *binary phase shift keying* (BPSK): a 0 bit leaves the carrier signal unchanged; and a 1 bit multiplies the carrier by –1, which is equivalent to shifting the phase of the sinusoidal signal by 180°. At bit transitions from 0 to 1, or from 1 to 0, the phase of the carrier signal is shifted by 180°.

The carrier frequency in Figure 2.3 is 1575.42 MHz (L1). During a 20-ms data bit duration of the navigation message, the code repeats twenty times. In addition to a C/A-code, each satellite also transmits a P(Y)-code on L1. This is accomplished by generating two carrier signals on L1: one as generated by the clock (in-phase component), and the other is obtained by shifting it in phase by 90° (quadrature component). The in-phase component is modulated by a C/A-code, and the quadrature component is modulated by a P(Y)-code. The phase shift makes the two carriers 'orthogonal' in a sense, allowing a receiver to separate their modulating signals. The signal on L2 carries a P(Y)-code only. The three BPSK-modulated signals transmitted by each satellite are shown in Figure 2.4 showing 180° shifts in carrier phase associated with bit transitions.

The L1 and L2 signals leaving the antenna of the kth satellite can be modeled as

$$\begin{aligned} s_{L1}^{(k)}(t) &= \sqrt{2P_C}\,x^{(k)}(t)D^{(k)}(t)\cos(2\pi f_{L1}t + \theta_{L1}) \\ &\quad + \sqrt{2P_{Y1}}\,y^{(k)}(t)D^{(k)}(t)\sin(2\pi f_{L1}t + \theta_{L1}) \\ s_{L2}^{(k)}(t) &= \sqrt{2P_{Y2}}\,y^{(k)}(t)D^{(k)}(t)\sin(2\pi f_{L2}t + \theta_{L2}) \end{aligned} \tag{2.1}$$

where P_C, P_{Y1}, and P_{Y2} are the signal powers for signals carrying C/A-code on L1, and P(Y)-codes on L1 and L2, and, respectively; $x^{(k)}$ and $y^{(k)}$ are the C/A- and P(Y)-code sequences assigned to satellite number k; $D^{(k)}$ denotes the navigation data bit stream; f_{L1} and f_{L2} are the

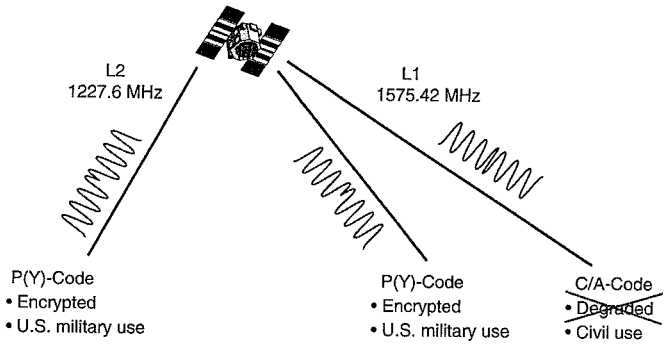


Figure 2.4 In mid-2005, each GPS satellite transmitted three signals, two on L1 and one on L2 frequency. The BPSK-modulated signals are shown. The signal carrying C/A-code on L1 was degraded purposely throughout the 1990s, but this practice has now ended. Access to P(Y)-code is limited to the DoD-authorized users via encryption.

carrier frequencies corresponding to L1 and L2, respectively; and θ_{L1} and θ_{L2} are phase offsets. The signal powers are discussed below. Note, however, that in order to express the BPSK signals as (2.1), we have switched the binary values of the codes and navigation data to ±1. From our old notation, a bit 0 maps into +1; and a bit 1 maps into –1. It is left as an exercise to show that (2.1) in fact represents the BPSK signals described earlier.

The modulation of a carrier by a binary code spreads the signal energy, initially concentrated at a single frequency, over a wide frequency band: over 2 MHz for the C/A-code and about 20 MHz for the P(Y)-code, centered at the carrier frequency, as shown in Figure 2.5. While the signal power is unchanged, this step reduces the power spectral density below that for the background RF radiation. Such signals, referred to as spread spectrum signals, have many properties which make them attractive for use in communication and navigation. The signal energy can be 'de-spread' in a receiver, if the code is known. In principle, keeping a code secret limits access to the signal. A comprehensive discussion of the spread spectrum signals is given in Chapter 9.

Let us take a quick look at two special properties of the PRN sequences which take values ±1. First, PRN sequences are nearly *orthogonal* to each other: The sum of the term-by-term products of the two sequences, shifted arbitrarily relative to each other, is nearly zero. For satellites k and l, which are assigned unique PRN sequences called C/A-codes $x^{(k)}$ and $x^{(l)}$,

$$\sum_{i=0}^{1022} x^{(k)}(i) \cdot x^{(l)}(i+n) \approx 0\,, \quad \text{for all } n, k \neq l \tag{2.2}$$

where $x^{(\bullet)}(1023+m) = x^{(\bullet)}(m)$. The left-hand side of (2.2) defines the *cross-correlation function* of the two sequences for shift n. The PRN sequences are nearly uncorrelated for all shifts. The PRN sequences associated with the P(Y)-codes are also similarly orthogonal to each other. The orthogonality of these codes allows all satellites to broadcast simultaneously at the same frequency without interfering with each other.

The second important property is that a PRN sequence is nearly uncorrelated with itself, except for zero shift. For a C/A-code

$$\sum_{i=0}^{1022} x^{(k)}(i) \cdot x^{(k)}(i+n) \approx 0\,, \quad \text{for all } |n| \geq 1 \tag{2.3}$$

The left-hand side of (2.3) defines the *auto-correlation function* of a sequence for shift n. The auto-correlation function of a PRN is nearly zero except for zero shift where it has a sharp peak. These cross- and auto-correlation properties of a PRN are exploited in acquisition and tracking of signals, discussed briefly in Section 2.4.1 and in greater detail in Chapters 9 and 10.

2.3.2 Anti-Spoofing (AS) and (the Late) Selective Availability (SA)

The main mechanism for limiting access to the full capabilities of GPS has been encryption of the P-code broadcast on both L1 and L2. This feature is referred to as Anti-spoofing (AS). An encrypted P-code is referred to as a Y-code. The main purpose of AS is to protect a user from spurious GPS signals with misleading data that an adversary may be tempted to transmit. AS has been active nearly continuously since 1994. Access to Y-coded signals requires the cryptographic key, which the DoD makes available to authorized users.

SPS limits civil users to the C/A-coded signal on L1 but, as we will see in Chapters 5–7, dual-frequency measurements are essential for precise positioning. Receiver manufacturers have, therefore, devised proprietary techniques to gain access to measurements on both L1 and L2. These techniques exploit in different ways the fact that the same P(Y)-code is being transmitted by a satellite on both frequencies. But there is a price: The L2 measurements are more fragile and noisier than they would be if the code were known, and the dual-frequency receivers cost a lot more. These techniques do not compromise AS because the structure of the Y-codes remains unrevealed and no spoofing signals can be generated.

Throughout the 1990s, the signals available for unrestricted use were purposefully degraded under the policy of Selective Availability (SA) by adding controlled errors in the measurements. These errors were significantly larger than the errors inherent to the system. *The net result of SA was about a five-fold increase in positioning error.* The signal degradation was achieved by 'dithering' the satellite clock and, therefore, the timing marks on the ranging signals, affecting the C/A-code, P(Y)-code, and carrier phase measurements equally. Another mechanism designed to degrade performance, though apparently not used often, was to broad-

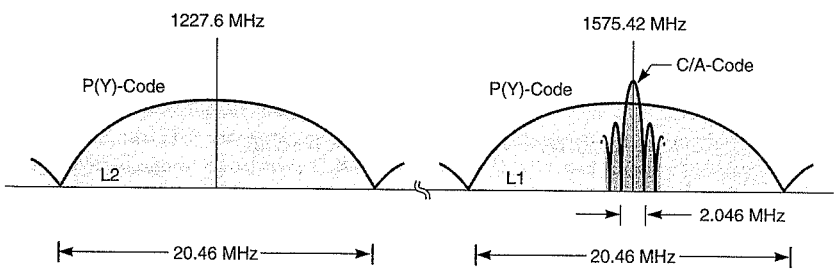


Figure 2.5 Power spectra of signals transmitted by a GPS satellite in mid-2005. The energy of the signal for civil users carrying a C/A-code on L1 is spread mainly over a 2-MHz-wide frequency band. The bandwidths of the signals for military users on L1 and L2 carrying a P(Y)-code are ten times wider.

cast erroneous or imprecise values of the ephemeris parameters. The DoD-authorized users were able to undo SA. Actually, civil users could undo SA as well, but it required additional expense. A user with several hundred dollars to spare for a radio receiver was able to obtain differential corrections in real time, as discussed below. As might have been expected, SA made few friends among civil users.

SA was deactivated on 2 May 2000 in accordance with a Presidential decree. Perhaps the European plans to develop Galileo accelerated the U.S. move to drop SA. Regardless of the political calculation, the departure of SA marked an important milestone for the civil users of GPS. Accuracy is addictive and the return of SA is unthinkable. To give an idea to a newcomer of the bad old days, we present several examples of SA-corrupted data and the corresponding position estimates later in this chapter.

With SA gone, the main advantage of PPS over SPS is robustness: The jamming resistance of PPS is higher, and spoofing it is a challenge. PPS also offers improved positioning performance due to dual-frequency measurements which allow compensation for ionospheric propagation effects [Section 5.3.2]; faster codes which lead to higher precision of range measurements [Section 10.6]; and lower error due to multipath [Section 10.7].

2.3.3 Signal Power

The GPS signals received on the earth are extremely weak. The RF power at the antenna input port of a satellite is about 50 watts, of which about half is allocated to the C/A-code. The satellite antenna is designed to spread the RF signal roughly evenly over the surface of the earth below. The satellites are said to bathe the earth with a gentle, uniform radiation. That, however, may be too sensuous an imagery for engineers.

Before we discuss the signal power levels, let us digress to introduce some terminology. In order to deal simply with a wide range of power levels, electrical engineers express power ratios on a logarithmic scale in units of decibel (dB), defined as

$$\left(\frac{P_1}{P_0}\right)_{dB} = 10 \log_{10}\left(\frac{P_1}{P_0}\right) \tag{2.4}$$

where P_1 and P_0 are the power levels to be compared. Absolute values of power can be expressed similarly in relation to 1 watt or 1 milliwatt in units of dBW or dBm, respectively. Consider a signal with power P_1 of 0.1 watts. This power level can also be represented as –10 dBW or 20 dBm. A second signal, with a power P_2 of 100 watts is 30 dB more powerful than the first signal. A third signal, with 200-watt power (P_3), is 3 dB stronger than the second signal. We can capture these relationships as follows.

$$P_1 = -10 \text{ dBW}, \quad \frac{P_2}{P_1} = +30 \text{ dB}, \quad \frac{P_2}{P_3} = -3 \text{ dB}$$

The GPS specifications on the minimum received power level for the users on the earth are given in Table 2.1. The signal powers are only around 10^{-16} watts! Interestingly, 10^{-16} watts is enough power to navigate with if we were among friends and people of good will. Adversaries and malicious hackers, however, may be tempted to exploit this vulnerability of the system.

The GPS signals are well below the background RF noise level sensed by an antenna

Table 2.1 GPS minimum received signal power specifications

Frequency band	*C/A-code (dBW)*	*P(Y)-code (dBW)*
L1	–158.5	–161.5
L2	—	–164.5

[Chapter 10]. It is the knowledge of the signal structure (i.e., PRN code) that allows a receiver to extract the signal buried in noise and make precise measurements. The signal boost so realized is called *processing gain*. If the noise level is raised by RF interference or jamming, the available processing gain may not be enough to extract the signal.

The low signal power is the Achilles' heel of GPS, especially in military use. Even in civil use, there is concern about the vulnerability of the signals as the national commercial and public infrastructure comes to rely more and more on GPS. Frequency diversity and increased signal power are important considerations in the modernization plans to make the GPS service more robust [Chapter 3].

2.4 Receivers, Measurements, and Performance

2.4.1 Evolution of Receiver Technology

Several generations of GPS receivers came to market between 1980 and 2005. The receivers available today bear the same resemblance to the early receivers as the laptop and palm computers do to the minicomputers of the early 1980s. The advent of very large scale integration (VLSI) has led to powerful microprocessors and memory chips, which have changed the look and feel of all electronic equipment, including GPS receivers.

The first set of GPS receivers was built in the 1970s for the DoD to prove that GPS would work (see Figure 2.6). In the early 1980s, the GPS constellation consisted of a few Block I satellites. There was no rush to produce receivers for the potentially huge automobile market because the system was at least ten years away from being ready for operational use and the U.S. policy on civil use was still unclear. Besides, the requisite digital street maps and databases did not yet exist. The geodesy and surveying community, however, saw the potential of GPS to offer extraordinary precision in positioning and was eager to use the system even though four satellites were in view typically for only a few hours per day. The market was limited and the focus was on precise measurements.

In 1982, two GPS receivers designed for precise geodesy came on the market: Macrometer V-1000, designed by Charles C. Counselman III of MIT and built by the Steinbrecher Corporation, and the Texas Instruments TI 4100. These receivers, bulky by today's standards, were revolutionary instruments which demonstrated millimeter-level positioning accuracy with GPS using carrier phase measurements [Chapter 7]. A receiver took up quite a bit of rack space, weighed tens of kilograms, and consumed over 100 watts of power. Each was priced at over $100 000.

The Macrometer did not require knowledge of the PRN codes and offered carrier phase

Figure 2.6 Early GPS receivers were bulky as shown in this U.S. Air Force photograph (circa 1980).

measurements on L1 from up to six satellites by implementing the so-called squaring channel: by squaring a signal, the binary modulation due to the code and navigation message disappears, and what we are left with is a carrier at twice the frequency. Satellite ephemeris and clock synchronization, available with code measurements, had to be provided externally by other means. The TI 4100 provided C/A- and P-code pseudoranges and carrier phase measurements from four satellites at a time. The receiver actually had a single hardware channel which was switched rapidly among the satellites.

In 1989, the DoD began to launch Block II satellites to populate the constellation fully and announced plans to declare the system operational. The U.S. policy on civil use of GPS had become clear by then. Digital maps and databases were beginning to appear, and receiver manufacturers switched into high gear in anticipation of large-volume production. The focus shifted to digital processing of signals using application-specific integrated circuits (ASIC) which would lead eventually to credit card-size GPS receivers and single-chip designs to be incorporated in pagers and cellular telephones. These receivers typically provide code and carrier phase measurements for the SPS signals from all satellites in view.

2.4.2 Signal Acquisition and Tracking

The basic functions of a GPS receiver are:

- to capture the RF signals transmitted by the satellites spread out in the sky,
- to separate the signals from satellites in view,
- to perform measurements of signal transit time and Doppler shift,
- to decode the navigation message to determine the satellite position, velocity, and clock parameters,
- to estimate the user position, velocity, and time.

The signals are gathered by a hemispherical antenna. Given a recent almanac and a rough idea of the user location, the receiver determines which satellites are in view. Given the satellite ID, the receiver knows the structure of the C/A-code being transmitted by it, and attempts to 'tune' to it to *acquire* the signal, and from then on *track* changes in it continuously.

To acquire a signal, the receiver generates a replica of the known C/A-code, and attempts to align it with the incoming code by sliding the replica in time and computing the correlation. From the auto-correlation property of the signal (2.2), the correlation function exhibits a sharp peak when the code replica is aligned with the code received from the satellite. The uncertainty in matching the replica with the incoming code is limited to only 1023 code chips, and the process of aligning them is generally quick. Direct acquisition of a P(Y)-code is difficult by design due to the length of the code. The signal acquisition is accomplished in two steps and is based on the known timing relationship between the C/A- and P(Y)-codes. First, the receiver acquires the C/A-code and, then, with the aid of the timing information in the navigation message, acquires the P(Y)-code. Direct acquisition of the P(Y)-code would be accomplished in the newer receivers with better clocks and thousands of parallel correlators

Code tracking is implemented as a feedback control loop, called a *delay lock loop*, which continuously adjusts the replica code to keep it aligned with the code in the incoming signal [Section 12.2]. After the alignment is accomplished, the PRN code is removed from the signal, leaving the carrier modulated by the navigation message. This signal is now tracked with another feedback control loop called a *phase lock loop* [Section 12.3]. Essentially, the receiver generates a sinusoidal signal to match the frequency and phase of the incoming signal, and in the process extracts the navigation message. The Doppler shift is measured in the phase-lock loop. Receivers to be used for precise positioning also keep a continuous track of the Doppler count [Section 1.3.2] or delta pseudorange. Figure 2.7 represents a conceptual view of these tracking loops.

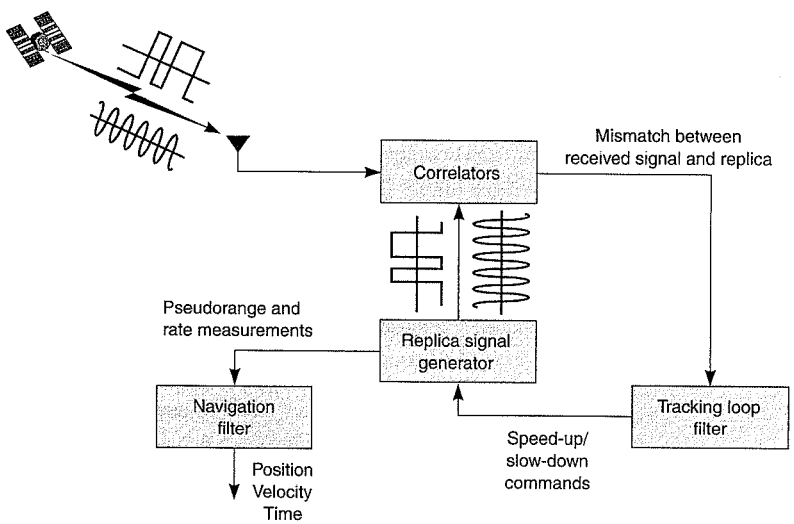


Figure 2.7 A conceptual view of a GPS receiver.

The measurement of transit time for a signal modulated by a C/A-code is conceptually quite simple. The time shift required to align the receiver-generated code replica and the signal received from the satellite is the apparent transit time of the signal modulo 1 ms (see Figure 2.3). The PRN code chips are generated at the satellite at precisely known instants in accordance with the satellite clock and, therefore, the receiver can 'read' the satellite clock time to determine when a chip was generated. The time of reception is determined from the receiver clock. Multiplying the apparent transit time by the speed of light gives pseudorange. Pseudoranges measured from four, or preferably more, satellites are used to compute position [Section 6.1]. The Doppler shift, caused by the relative motion of a satellite and the user, is the projection of the relative velocity on the line of sight, and can be converted into pseudorange rate. Given the pseudorange rates corresponding to four satellites, and the satellite velocity vectors (derived from the navigation message), a user can compute his velocity [Section 6.2]. A GPS receiver does this automatically, continuously, and virtually instantaneously.

2.4.3 Estimation of Position, Velocity, and Time (PVT)

The quality of the PVT estimates obtained by a user from GPS depends basically upon two factors: (i) number of satellites in view and their spatial distribution in the sky, and (ii) quality of the range and range rate measurements. The spatial distribution of the satellites relative to the user is referred to as *satellite geometry*. The satellite geometry changes with time as the satellites rise, move across the sky, and set. Roughly speaking, the geometry is good if the satellites are on all sides of the user, some high in the sky and several low. If a significant part of the sky is somehow blocked, the user may still be able to compute PVT estimates if four or more satellites are in view, but there would generally be an accuracy penalty for poor geometry [Section 6.1.2].

The second factor determining the quality of the PVT estimates is the quality of the pseudorange and Doppler measurements. There are several sources of biases and random errors, which affect the measurements:

- Errors in the navigation message parameters which specify satellite position and signal transmission time introduce errors in the pseudorange measurements. These errors are often referred to as signal-in-space (SIS) errors.
- Propagation delays in the ionosphere and the troposphere, signal distortion due to multipath, and receiver noise also introduce measurement errors. Ionospheric propagation delay, in particular, can be large, ranging from several meters to several tens of meters, depending upon the state of the ionosphere and the elevation of the satellite. This error, however, can be removed substantially by a user equipped with a dual-frequency receiver.

We focus on these errors in Chapter 5, and examine ways to mitigate them. For now, as an academic exercise, we illustrate the nature and size of the error due to Selective Availability (SA), the now-discontinued policy of purposeful degradation of GPS signals referred to earlier. The pseudorange measurement errors are plotted in Figure 2.8 for three satellites, or space vehicles (SV), two with SA active and one with SA inactive (or SA off). These measurements were obtained in 1997 from an antenna at a surveyed location. Use of an external cesium atomic standard with the GPS receiver essentially removed the receiver clock bias.

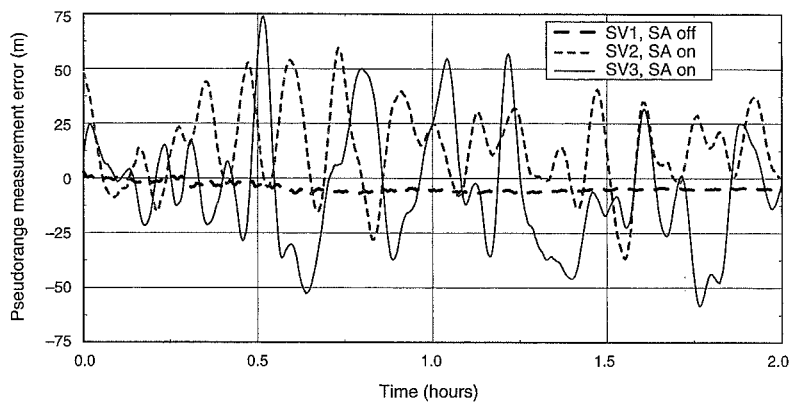


Figure 2.8 Pseudorange errors measured for three satellites, two with SA active and one with SA inactive. The SA-induced pseudorange error remained about 25 m rms throughout the 1990s. SA was terminated on 2 May 2000.

Figure 2.8 shows the reason for the civil community's resentment over SA, which was the dominant source of error. Range measurement error for the satellite with SA off appears essentially to be a 5-m bias holding steady for about ninety minutes. The explanation is that, unlike the SA error, the signal propagation errors change slowly and, if not compensated for, appear as biases. As expected, the positioning accuracy of SPS improved dramatically when SA was 'set to zero' on 2 May 2000, as illustrated in Figure 2.9.

We now revise the simple model for pseudorange measurements used earlier in Figure 1.9 to include the measurement errors discussed above. The pseudorange measurement from the kth satellite is modeled as

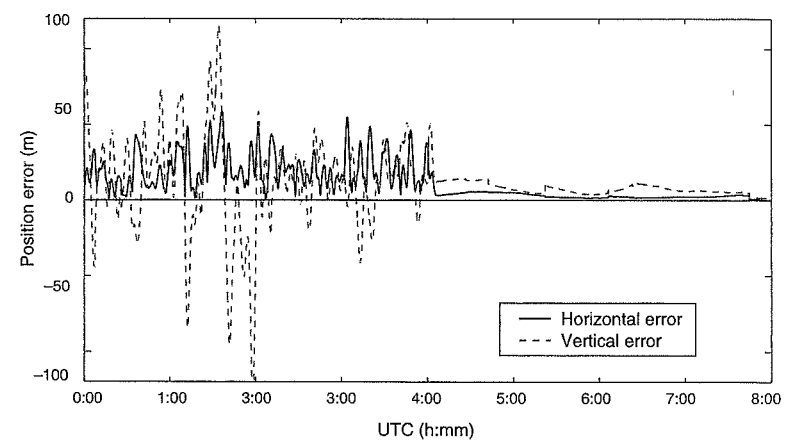


Figure 2.9 Change observed in positioning accuracy of SPS when SA was terminated on 2 May 2000 at about 04:00 hours UTC. (Courtesy of Dr. Sherman Lo, Stanford University)

$$\rho^{(k)} = \sqrt{\left(x^{(k)} - x\right)^2 + \left(y^{(k)} - y\right)^2 + \left(z^{(k)} - z\right)^2} + b + \varepsilon^{(k)} \qquad (2.5)$$

$k = 1, 2, \ldots, K$, where $(x^{(k)}, y^{(k)}, z^{(k)})$ are the satellite coordinates broadcast in the navigation message; (x, y, z) are the user coordinates to be determined; b is the unknown receiver clock bias; and $\varepsilon^{(k)}$ represents the unknown error in the measurement. We need a minimum of four such equations to solve for the four unknowns. Indeed, if $K = 4$, the best we can do is to pretend that there is no error term present in (2.5). For $K > 4$, we solve the system of equations recognizing that the measurements are 'noisy' and look for a solution (x, y, z, b) which fits the measurements best in some sense. There are several approaches. The most commonly used approach goes all the way back to Gauss and is referred to as *least squares*.

The *residual* is defined as the discrepancy in (2.5) when we substitute the estimates for the unknown parameters. It's the difference between the pseudorange computed from the estimated values of the user position and clock bias, and its measured value. The parameter values which minimize the sum of the squares of the residuals are called least-squares estimates. The least-squares approach turns the estimation problem into a simple minimization problem: Determine the vector (x, y, z, b) which minimizes

$$\sum_{k=1}^{K}\left(\sqrt{\left(x^{(k)} - x\right)^2 + \left(y^{(k)} - y\right)^2 + \left(z^{(k)} - z\right)^2} + b - \rho^{(k)}\right)^2$$

The problem is solved easily by using the *Newton-Raphson method*, as described in Chapter 6. The components of the minimizing vector, which we represent as $(\hat{x}, \hat{y}, \hat{z}, \hat{b})$, are our "best" estimates of position and clock bias in least-squares sense. We also analyze in Chapter 6 how the measurement errors and satellite geometry determine the quality of these position estimates.

2.4.4 Positioning Accuracy

The performance obtained from GPS is dynamic, changing both with time and place as the satellite geometry and measurement errors change. A global characterization of the performance requires knowledge of the satellite constellation and receiver capabilities, and is given in statistical terms, e.g., as root-mean-square (rms) error or 95th percentile of the error distribution. The latter is often referred to as 95% error. In the 1990s, SA was the dominant source of error, and the performance specifications of 95th percentile for the horizontal and vertical error were 100 m and 156 m, respectively, defined for measurements taken over 24 hours at any point in the world with an unobstructed view of the sky.

In the post-SA world, the performance specifications account only for the satellite orbit and clock errors for which the Control Segment assumes responsibility. These are referred to as signal-in-space (SIS) errors. A user would typically also suffer the consequences of errors in ionospheric and tropospheric modeling, multipath, and receiver noise, all discussed in Chapter 5.

Table 2.2 gives the current SPS performance specifications [SPS (2001)] and the accuracy commonly achieved, the latter in round numbers. The specifications only consider the signal-in-space errors and appear overly conservative. Happily, most users achieve better performance. No performance specifications are given for velocity estimation. For moderate dynamics, velocity accuracy of 0.01 m/s has been reported [Section 6.2.1].

The SPS performance will improve in a few years when the satellites start transmitting

Table 2.2 GPS/SPS performance for global positioning and time dissemination

Error (95%)	*Specifications**	*Empirical data*
Position		
Horizontal	13 m	10 m
Vertical	22 m	15 m
Time	40 ns	30 ns

* Signal-in-space errors only

additional civil signals. And, of course, much better performance is available today to the users of differential GPS, discussed next.

2.5 Differential GPS (DGPS)

The accuracy requirements of the different navigation and positioning applications vary widely. Horizontal positioning accuracy of tens of meters is generally more than adequate for navigation in wide-open spaces: maritime navigation on the open seas; aircraft navigation in en route, terminal, and non-precision approach phases of flight; and recreational use by hikers and backpackers. Many important applications, however, require greater accuracy. Under poor visibility conditions, harbor entry by ships and taxiway guidance on airport surface require meter-level horizontal accuracy, and precision approaches by aircraft typically require meter-level vertical accuracy. Automobile navigation over roads and highways has a similar accuracy requirement. [FRS (2001) defines the various phases of flight and states the navigation accuracy requirements of each.]

In view of the discussion in Section 2.4.2, the quality of the GPS position estimates can be improved by reducing measurement errors and/or improving satellite geometry. We discuss below how this may be done.

- Mitigation of the measurement errors turns out to be simpler: The errors associated with the worst error sources are similar for users located 'not far' from each other, and change 'slowly' in time. In other words, the errors are correlated both spatially and temporally. Clearly, we can estimate the error in a measurement if the receiver location is known. Such error estimates can be used as *differential corrections* if made available to the GPS users in the area, allowing them to mitigate errors in their measurements. That's differential GPS, generally abbreviated as dGPS or DGPS.
- Satellite geometry can be improved by adding satellites to the constellation to provide additional ranging signals. A user can also improve the geometry by deploying pseudo-satellites, called *pseudolites*, which transmit GPS-like signals. The pseudolites can be deployed on the ground, in the air, or on a ship. A GPS receiver has to be modified to receive and process these signals.

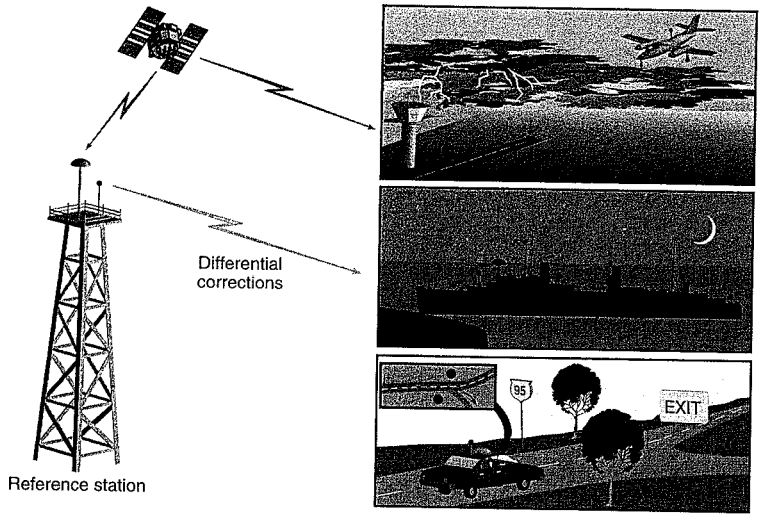


Figure 2.10 Local-area differential GPS (LADGPS). A reference receiver at a known location measures GPS signals, determines errors therein, and transmits corrections on a radio link to the users in the area.

Local-Area Differential GPS

An implementation of LADGPS is shown in Figure 2.10. A reference station is set up and the antenna is placed at a surveyed location. In this stylized picture, the reference station and each user is equipped with a saucer-shaped antenna, used to receive the GPS signals, and a bulb-shaped antenna, used to transmit or receive the differential corrections on a radio link. (DGPS radio links have used frequencies from 100 kHz to 1.5 GHz.) The DGPS users in the picture are: an aircraft in clouds preparing to land; a ship entering a harbor at night; and a motorist on I-95 in an unfamiliar city. DGPS can provide meter-level and, even sub-meter-level, position estimates depending upon the closeness of the user to a reference station and the delay in the corrections transmitted over a radio link. Such performance can meet most requirements of land transportation and maritime traffic, and DGPS services, both commercial and government-provided, are now widely available.

The U.S. Coast Guard (USCG) provides differential corrections for free on marine radiobeacon frequencies (285–325 kHz) from about fifty broadcast sites in the coastal areas and around the Great Lakes and inland waterways in the conterminous U.S. (CONUS), and in parts of Hawaii, Alaska, and Puerto Rico. The differential corrections to GPS signals piggyback on the radiobeacon broadcasts. This service, called *Maritime DGPS*, was declared operational in March 1999. Maritime DGPS has been demonstrated to provide horizontal positioning accuracy of 1–3 meters at a distance of 100 km, or more, from a reference station. A number of countries have implemented systems compliant with the Maritime DGPS standards to enhance safety in their waterways.

The success of Maritime DGPS service has led to development of *Nationwide Differential GPS* (NDGPS) in the United States by converting the Air Force's Ground Wave Emergency Network (GWEN) sites to GPS reference stations and adding new installations, modeled after the Maritime DGPS stations. These sites mount transmitting antennas on 100-m towers and have signal ranges of about 400 km. This service will add about a hundred stations distributed over CONUS to the network of maritime radiobeacons, benefiting operations of railroads, agriculture, environment, forestry, and emergency response services. At the end of 2004, the operational NDGPS sites covered nearly 90% of the lower 48 states. [http://www.navcen.uscg.gov/dgps]

In the Maritime DGPS and NDGPS model, each reference station serves an area around it, with coverage ranging from a few kilometers to hundreds of kilometers, depending upon the accuracy requirements of the application. Such DGPS service is often referred to as *Local-Area DGPS*. The most ambitious of such services is being developed by the FAA. The Local Area Augmentation Service (LAAS) is planned to guide aircraft during approach and landing when visibility is poor [http://gps.faa.gov].

Wide-Area Differential GPS

An alternative to Local-Area DGPS is a centralized system which provides differential corrections usable over continent-wide areas or oceans. Such systems are referred to as *Wide-Area DGPS* (WADGPS). A WADGPS system collects measurements from a network of reference stations covering the area of interest and processes them to decompose the measurement error into its constituent parts. The differential corrections are then transmitted in a form so that each user can determine her corrections on the basis of her position estimate obtained from (unaugmented) GPS. A number of commercial services now provide such differential corrections via communication satellites and FM subcarrier, and use of such services is common in offshore oil exploration and fleet management.

The FAA's Wide Area Augmentation System (WAAS), declared operational in 2003, broadcasts corrections applicable to North America from geostationary satellites in the form of navigation messages modulated on GPS-like signals on L1 [http://gps.faa.gov]. A receiver requires only small software changes in order to exploit the WAAS corrections, and the newer receivers are generally designed to be "WAAS-capable."

The Canada-Wide Differential GPS (CDGPS), which became operational in 2004, also uses geostationary satellites to broadcast corrections which are available to all who can invest in a radio receiver [http://www.cdgps.com]. NASA's Global Differential GPS (GDGPS) transmits corrections on the Internet (to authorized users) and offers even greater accuracy, mostly for scientific activities [http://www.gdgps.net].

DGPS Positioning Accuracy

An overview of typical positioning accuracy achievable with GPS in different modes is given in Figure 2.11. With code measurements (i.e., pseudoranges), real-time position estimates within several tens of meters of the true location were available from SPS while SA was active. With SA gone, the error has been reduced generally to less than ten meters. The PPS users with access to dual-frequency measurements can correct their measurements for ionospheric delay [Section 5.3.2] and do slightly better yet. To obtain a better performance, an additional investment is required in the form of augmentation of the GPS signals. Wide-area DGPS can

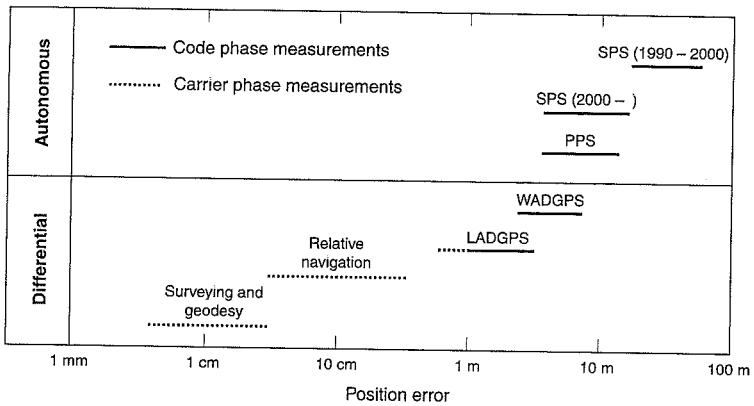


Figure 2.11 A summary of GPS performance in different modes.

offer meter-level positioning accuracy. Local-area DGPS can provide decimeter-to-meter level position estimates in real time.

Centimeter-level accuracy in positioning relative to a reference station within tens of kilometers is achievable in real time from GPS carrier phase measurements, as discussed in Chapter 7. Systems offering such capability are now finding application in agriculture, mining, and

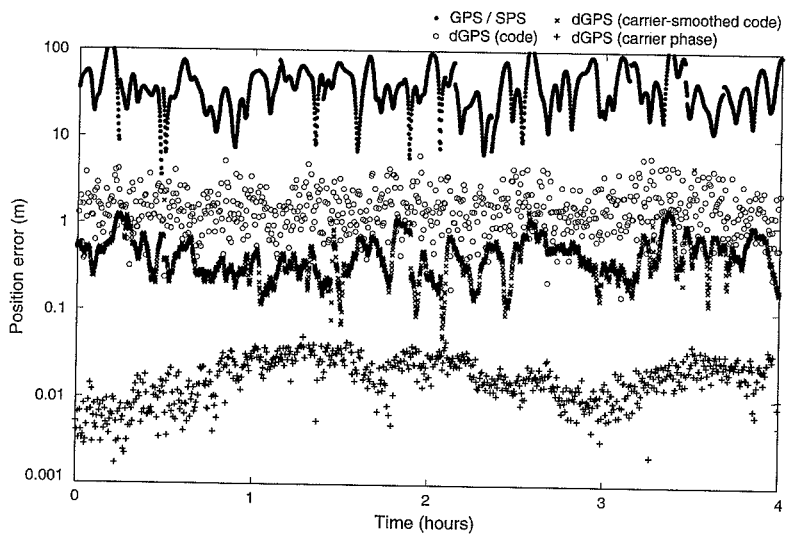


Figure 2.12 Accuracy of the GPS-based position estimates obtained in real time can range from tens of meters (SA active) to centimeters depending upon the user's resources. The four plots show 3-D error in position estimates obtained by processing a set of measurements from 1997 in different modes discussed in Chapters 6–7.

construction industries. Even higher accuracy is achievable, but not in real time. The geodesists studying, for example, the motion of the earth's tectonic plates, have now become accustomed to millimeter-level positioning accuracy. In these applications, however, the difficulty of the problem is somewhat ameliorated by the patience of the users. The answers can wait for minutes, hours, or even longer.

Four levels of positioning accuracy obtained from GPS in real time are shown in Figure 2.12 with position errors ranging from tens of meters to centimeters. The three-dimensional position error is plotted for measurements taken at five-second intervals over four hours early in 1997 while SA was active. The results are typical. The basic SPS provided position estimates with error of tens of meters using a single receiver. Access to concurrent measurements from a GPS receiver at a known location in the area reduced the errors in the position estimates to meter level and decimeter level, as shown, depending upon the data transferred over a radio link from the reference station to the user. Such accuracy is now routine. Finally, the full potential of the measurements is exploited to obtain centimeter-level position estimates. While the plot at the top is only of historical interest now that SA is history, the lower three plots are typical of performance achievable today in differential and relative positioning with code and carrier phase measurements, which we discuss in detail in Chapters 6–7.

2.6 Civil Applications

The growing civil applications of GPS may be divided into the following categories:

- precise time dissemination,
- high-precision (millimeter-to-centimeter level) positioning,
- specialized applications such as aviation and space navigation,
- land transportation and maritime uses,
- consumer products.

Each is reviewed briefly below.

Timing

While its positioning capability receives most attention, GPS is also a global source of precise time and its use as a timing service has grown rapidly. Indeed, GPS may have a larger impact on our commercial and industrial infrastructure as a source of precise time than for navigation or tracking. In recognition of this, some have seriously suggested renaming the system as Global Positioning and Timing Service (GPtS).

Consider simultaneous observation of events (perhaps celestial) at distant locations. Not too long ago, synchronizing the clocks at multiple sites would have been a significant challenge, but no longer. Inexpensive GPS receivers can provide estimates of time with an accuracy heretofore offered only by atomic standards. GPS-derived time has become a logical choice for recording times of events for scientific purposes, and correlating events recorded at different geographic locations. In fact, the national and international laboratories, which serve as standards for time and frequency, use GPS to compare their time standards.

Timing is everything in today's telecommunication systems, and GPS is being used increasingly to synchronize the elements of networks. Clocks at the major Internet nodes are

being set by GPS, and the time passed down to the local-area networks. Bank transactions are marked by GPS-derived time. The electric utilities are using GPS to analyze the state of a power grid via precisely synchronized measurements of the voltage phasor (amplitude and phase) at different substations. GPS-synchronized measurements are expected to become an essential element of power system control.

Precise Positioning

The community of geodesists and geophysicists appears to have been the most surprising and unintended beneficiary of GPS. The geodesists have adopted techniques developed previously for radio astronomy to achieve millimeter-level accuracy in relative positioning with GPS carrier phase measurements. These techniques are now being used widely to study plate motion and crustal deformation, earthquakes, volcanic processes, ice sheet processes and post-glacial rebound, and variations in the earth's rotation. In these studies, the answers are generally not needed in real time, and data collected over hours and days at different locations are post-processed to obtain estimates of relative position vectors good to millimeters.

The value of GPS in studies of the dynamics of the earth has been enhanced greatly by the establishment of the International GNSS Service (IGS), previously known as the International GPS Service, which coordinates collection and analysis of GPS and GLONASS measurements with geodetic-quality receivers at over 350 sites around the world. These data and products based upon them (e.g., satellite orbit and clock parameters, ionospheric total electron content, tropospheric zenith path delays) are available from http://igscb.jpl.nasa.gov.

The high-precision positioning capability of GPS is also being used to monitor deformations of large engineering structures in real time under actual loads, e.g., bridges and towers under actual traffic and wind-loading conditions. Such real-time estimates are typically good to centimeter level. GPS-based precise estimates of position and attitude of an aircraft are proving invaluable in airborne surveys and photogrammetry. Atmospheric scientists are using GPS measurements for precise, real-time characterization of electron densities in the ionosphere and water vapor content of the troposphere.

Actually, the market for high-precision GPS receivers for scientific studies is rather small. The techniques developed in these studies, however, have become indispensable tools of surveyors and mapmakers, creating a large market for high-end GPS receivers and services. Indeed, the field of surveying has been revolutionized by GPS with vast improvements in accuracy, speed, and economy.

The techniques of precise positioning have found applications in real-time control of machinery used in agriculture, construction, and mining industries. In agriculture, GPS is being used for accurate yield mapping and soil sampling in order to apply fertilizer or herbicide more effectively and economically. Consider a project for grading a large construction site. There are systems on the market that integrate GPS and computer-drawn grading plans with the grader's hydraulic system to allow automated control of blade elevation and cutting angle while the machine operator simply steers the vehicle.

The convergence of GPS and personal computer technologies has made it possible to collect vast amounts of positional data and to organize them into geographic information systems (GIS). A number of positioning and attribute collection systems are on the market. The precision required of the estimates depends upon the application. At its simplest, such a system typically consists of a backpack containing a battery-powered personal computer with a GPS card (and perhaps a radio modem to receive differential corrections for real-time DGPS) and a handheld keyboard and display unit. A user so-equipped can walk around gathering information and entering it through the keyboard to create or update a spatial database showing, for example, the location and status of every utility pole or man-hole in a town, or a map showing concentrations of toxic or radioactive wastes. Such geographic databases are becoming invaluable tools for management of forests and coastlines.

Aviation and Space Navigation

Until a few years ago, civil aviation relied entirely on on-board inertial navigation systems and ground-based radionavigation aids to assist pilots in navigating their assigned routes and guide them during approach and landing in bad weather. The radio navaids are expensive to operate and maintain, and large parts of the world lack even the basic radionavigation infrastructure. GPS is widely seen as the most important advance in civil aviation since the advent of radio, with a potential to enhance greatly both the economy and safety of air operations.

Civil aviation is a demanding enterprise with stringent requirements that go beyond accuracy. It requires that a pilot shall be warned in a timely manner if a navigation system anomaly creates a hazard (*integrity* requirement), and that the navigation system shall continue to provide service at the required level for the duration of a demanding operation once underway (*continuity of service* requirement). In order to ensure that these requirements would be met by a navigation system, the regulatory authorities adopt and promulgate avionics standards and certification criteria. GPS has been certified for use in the U.S. airspace as a supplemental system for en route, terminal, and nonprecision approach phases of flight, and as a primary system for oceanic and remote area operations. [See FRS (2001) for definition of any unfamiliar terms.] In the near future, GPS will be used as a primary system for navigation in all phases of flight, including autolandings, and for collision avoidance and surface traffic surveillance at airports.

In order to appreciate the economic and safety benefits of GPS augmentations for civil aviation, we take a little detour to describe approach and landing in bad weather. Category I approaches allow a pilot to descend in fog or clouds down to 200 feet above the ground, half a mile from the touchdown zone. If the runway can be seen at that point, the pilot can continue to descend and land using visual references. If the runway is not visible, the pilot must execute a missed-approach procedure and climb to a safer altitude, possibly diverting to another airport. Category II (III) precision approaches allow the pilot to descend to 100 feet (50 feet) and are, therefore, even more demanding of the navigation system. Such approaches today require Instrument Landing System (ILS) or Microwave Landing System (MLS), one for each runway end. Each system requires careful site preparation, expensive maintenance and calibration, and can cost over a million dollars. Boston's Logan Airport, for example, has five ILS installations, one of which is certified to Category II-III standards.

The United States, Europe, and Japan are developing "interoperable" space-based augmentation systems (SBAS) which would offer a seamless global coverage. The U.S. contribution, called Wide Area Augmentation System (WAAS), developed by the FAA, became operational in 2003. It augments GPS with GPS-like signals transmitted on L1 from geostationary satellites. These signals are modulated with data messages so that a user receives data on the integrity of the satellite signals and differential corrections to the measurements. These systems are planned to meet the accuracy, integrity, and continuity requirements of en route and

terminal phases of flight, and non-precision and near-Category I precision approaches. Such approaches wouldn't require any special navigation equipment at the airport.

An even more ambitious initiative undertaken by the FAA is to develop a ground-based augmentation system (GBAS) to be deployed at airports for use in Category II/III precision approaches, including autolandings. This system is known as the Local Area Augmentation System (LAAS). A single LAAS installation at an airport will allow Category III landings at all runways. GPS-LAAS will meet the stringent requirements of Category III precision approaches: sub-meter accuracy in the vertical position estimates along with a guaranteed integrity and continuity of service. The DoD has a similar development program called the Joint Precision Approach and Landing System (JPALS). The demanding requirements of WAAS, LAAS, and JPALS are described in FRS (2001).

The United States is considering use of GPS as the primary system for civil aviation navigation, backed up by a skeleton network of VHF omnidirectional range/distance measurement equipment (VOR/DME) stations and the newly revived Loran-C for en route and terminal airspace, and ILS at selected airports for precision approaches.

To appreciate the benefits to be derived from GPS in space operations, consider, for example, satellite orbit determination and orbit maintenance. Satellite orbits used to be determined from radar tracking data from ground antennas distributed around the world. The measurements were collected centrally and processed to estimate the orbits. Subsequently, any commands to adjust the orbits were generated and transmitted to the satellite. Onboard GPS measurements, however, have been shown to perform much better than any ground-based tracking system. A GPS receiver combined with onboard orbit determination software and a closed-loop propulsion system can maintain orbits autonomously.

GPS receivers aboard satellites in low-earth orbits would become the primary source for position, velocity, attitude, attitude rate, and time, replacing an array of sensors and reducing the cost and complexity of spacecraft. A number of successful experiments and demonstrations have been carried out in various space missions in the last ten years, and a few spacecraft have actually integrated GPS measurements into their operational control systems. The International Space Station is being designed to use GPS for navigation, attitude determination, tracking of vehicles approaching the Station, and as a source of time for scheduling vehicle operations. The navigation system of the Space Shuttle is being upgraded to use GPS as the primary system during re-entry and landing.

The concept of flying several satellites 'in perfect formation' to form a single sensor much larger than anything that could have been flown on a single satellite is now receiving much attention. GPS signals would play a vital role in precise relative positioning of the satellites and attitude control to create telescopes and imaging radars of unprecedented resolution to explore space or to watch over the earth.

Land and Maritime Navigation

The application of GPS in land transportation, especially vehicle navigation and tracking, has grown to be very large (perhaps second-largest, after cell phones). GPS-based systems and services for motorists, commercial fleets, public transit, and emergency response agencies are in great demand. Rental car companies now routinely offer GPS-based navigation and route-guidance systems. In the post-9/11 world, GPS-aided vehicle tracking would address the security concerns about monitoring in real time the tens of thousands of loads of hazardous materials on the move every day. Railroad companies would use GPS for positive train control. GPS-based toll collection systems which account for time and distance traveled by commercial trucks on toll roads are already in use in Europe.

As noted previously, at least four satellites are required to be in view for three-dimensional navigation with GPS. Terrain, foliage, and buildings, however, can obstruct parts of the sky, and maintaining four or more satellites in track continuously is often impractical for land vehicles. A land-vehicle navigation system must, therefore, supplement GPS with dead-reckoning sensors, e.g., gyroscopes, compasses, odometer, inclinometer, and accelerometers. These sensors cannot provide absolute position, but can measure change in position accurately over a short period. DGPS can provide absolute position, but only intermittently. Land navigation must, therefore, exploit the complementary nature of GPS and the dead-reckoning sensors to meet its requirements.

There are millions of pleasure boats, fishing boats, ferries, cruise lines, cargo lines, and oil tankers in the world. All would benefit from GPS. Since the oil spill caused in the Alaskan waters by the grounding of Exxon *Valdez* in 1989, the harbors are implementing active monitoring of oil tankers for safety and efficiency. The United States now requires each tanker to transmit its DGPS-derived position via a radio link to a control station monitored by the harbor authorities. Such GPS-based surveillance would also help countries enforce fishing boundaries in national waters, an increasingly problematic area as fish stocks come under greater pressure.

Consumer Market

Knowledge of one's precise three-dimensional position expressed in WGS 84 by itself is only of academic interest for most users. But the position information can be invaluable if given in relation to the intended path, showing points of interest and potential hazards: a hiker's position in relation to a trail, a car on a moving street map, and a boat in relation to islands and obstacles. Combined with communication technology, say, a wireless phone, the knowledge of position can be life saving, reducing search and rescue to simply rescue. Athletes and outdoor adventurers have found much to love in GPS and millions of GPS receivers are sold each year to hikers, bikers, runners, and kayakers, replacing the compass, speedometer, charts, and maps.

The full power of GPS is being realized in civil applications in combination with other technologies, especially wireless communication systems, the Internet, and geographic databases. The consumer market for GPS products is seeing an explosive growth fueled by inexpensive, single-chip GPS receivers integrated into an array of consumer products: cellular phones, personal digital assistants (PDA), and security devices for personal possessions ranging from cars to computers.

The technology of telematics, offering information-on-the-move, has taken off now that the position coordinates of the moving party are so easy to obtain. Two-way messaging devices incorporating GPS technology are on the market now for a user to navigate with, and to communicate his position or course, or any other information, to anyone on earth who has an e-mail address. High-end cars now incorporate GPS-based features such as roadside assistance, automatic crash notification, and automatic vehicle location as standard. With access to data bases for services, this trend leads naturally to development and growth of location-based information services (LBS) directing a user in an unfamiliar place to a gas station, an Indian

restaurant, or a tourist attraction. GPS-based services now offer motorists the best routes to their destinations, navigating around traffic accidents and road closures.

A potentially huge application of GPS is for Enhanced-911 (E911). The Federal Communications Commission (FCC) has mandated that by end of 2005 all cellular telephones in the United States be equipped so that the location of each can be determined accurately in an emergency. A GPS receiver built into each handset would easily achieve the FCC-mandated accuracy (50 meters 67% of the time and 150 meters 95% of the time) while it is outdoors with an unobstructed view of the sky. The challenge is to acquire and track the severely attenuated signals under foliage, in urban areas, and indoors. A promising approach is to relieve the GPS receiver in the handset of the onerous function of decoding the navigation message. This function would be performed instead on a server at a base station and the required information would be transmitted to the receiver via wireless signals. This approach is referred to as handset-based assisted GPS (AGPS) [Section 13.4]. Nextel, Sprint, and Verizon have announced plans to adopt AGPS solution to meet the FCC mandate. (AT&T and Cingular are reported to be exploring alternative approaches based on measurements of distances or angles by a cellular network.)

GPS is an important element of what's been called "whereware" technology to track people and things indoor and out, and to know where everyone and everything is at all times. Big Brother issues aside, few would care to be targeted for ads and services on their "sell phones" broadcasting their whereabouts. While millions of GPS-enabled cell phones are now in use in Japan and South Korea, the acceptance and long-term growth of location-based services would depend upon legal protections to address the privacy issues, giving users the option not to be found. One group not entitled to that option is the people on parole or probation, who have to keep out of certain "exclusion zones" and whose movements have to be watched, in general. Sadly, there is market for a million such surveillance units in the United States alone.

In summary, the industrial, commercial, scientific, and personal everyday applications of GPS appear simply limitless. Applications are being developed at an astonishing rate. Examples of applications not falling cleanly in any of the categories listed above include: automatic tracking of oil spills and flooding with especially fitted buoys; sounding of the upper atmosphere with radiosondes (instrumented weather balloons) equipped with GPS receivers; tracking of wild animals with special GPS collars for research on habits and habitats; blind navigation, as in helping a visually impaired person get around outdoors with help from a talking map ("you are on the Main Street heading west, 25 m from the Park Avenue intersection"); making crop circles to mystify the gullible; and keeping the geocaching crowd occupied harmlessly with their treasure hunts [http://www.geocaching.com].

The growing dependence on GPS raises fears of deliberate disruption of GPS service by terrorists and hackers. Such threats and their impact on the civil infrastructure are assessed in the Volpe Center Report [Volpe Report (2001)]. Attempts to make the GPS service more robust are discussed in the next chapter.

2.7 GPS at a Glance

- **Basic Description**
  Space-based radionavigation system broadcasting, synchronized timing signals to provide estimates of position, velocity, and time based on passive, one-way ranging to satellites.
- **Milestones**
  1973: Architecture approved
  1978: First satellite launched
  1995: System declared operational
  2000: Purposeful degradation of the civil signals stopped
- **Satellite Constellation**
  24 satellites in six orbital planes inclined at 55°; near-circular orbits with radius 26,560 km; orbital period: 11h 58m; ground track repeats each sidereal day
- **Reference Standards**
  Coordinate frame: WGS 84
  Time: UTC(USNO)
- **Signals**

Carrier Frequency (Wavelength)	L1: 1575.42 MHz (0.19029 m)
	L2: 1227.60 MHz (0.24421 m)
Multiple Access Scheme	Code division multiple access (CDMA)
PRN Codes	C/A-code on L1; P(Y)-code on L1 and L2
Code frequency (Mcps)	C/A-code: 1.023; P(Y)-code: 10.23

- **Performance Achievable**
  Real-time: Typically, absolute positioning error of several meters with a single receiver, decimeters in differential mode
  Post-processing: Millimeter-level relative positioning
- **Receiver Bazaar** (Courtesy of Dr. Frank van Diggelen, Global Locate)
  A broad survey of the GPS user equipment is offered on the next page. The receivers, OEM boards, and chips are designed for different user groups with different performance requirements. Hikers are generally satisfied with meter-level accuracy in absolute positioning available from an L1-only design. Geodesists demand millimeter-level accuracy in relative positioning and would require code and carrier phase measurements at both L1 and L2 frequencies. Cell phone users require their positions within tens of meters when making an emergency call and the device must produce a position wherever it is, even indoors.

The above categories and requirements are much as they were five years ago, at the time of the first edition of this book (with steady price decline across most GPS hardware as Moore's Law continues to prove accurate). However there are several developments worth noting:

- In 2004 the number of GPS receivers built into cell phones approximately equaled all other GPS implementations combined. In 2005 GPS-cell phones are expected to exceed all other GPS receivers. However, the only widespread cell phone application, so far, is public safety: In many phones the GPS does nothing

Complete Receivers

	Description	*Price range*
	Handheld receivers for hikers, backpackers, and sailors. Small in size with simple maps. Ruggedized for outdoor/marine use.	$70–$500
	GPS cell-phones GPS is embedded. Applications are emerging, but mostly GPS is for emergency calls only.	$200–$500
	GPS smart-phones and PDAs GPS is embedded. GPS can be used for emergency calls in smart-phones, but many applications are available for Pocket-PC or Palm OS. Applications range from basic maps (as in handheld GPS) to spoken turn-by-turn directions.	$400–$800
	Dedicated in-car navigation systems. Detailed street maps and spoken turn-by-turn directions	$400–$2000
	Marine navigation. Fixed mount, large screens with electronic charts	$400–$3000
	Aviation FAA certified, usually panel mounted, with maps	$1000–$13,000 (higher prices with integrated comm's)
	Survey and Mapping Often tripod mounted, exclusively Differential GPS, one meter to centimeter accuracy	$2500–$40,000

Modules

	Plug-in modules Integrated receiver and antenna. Interface to PC or PDA through Serial or USB cable, CF (Compact Flash) or SD (Secure Digital) card, or Bluetooth.	$100–$300
	OEM boards Receiver circuitry for customer integration	$30–$100
	Chip sets/ single chip GPS Lowest prices achieved by sharing the CPU of the host device	$5–$20

until you make an emergency call (9-1-1 in the US), and then it provides your position to the emergency response center.

- A new category of product: The GPS-enabled PDA and "smart-phone" merges cell-phone and PDA. With downloadable applications and relatively large screens, this category has few limits on applications and may become a substitute for the dedicated in-car navigation system.
- The standard hand-held receiver has generally maintained the same price-level it occupied five years ago, by adding improved functionality (better displays and maps) instead of price reductions, a notable rarity in consumer electronics.
- Location-based games have been talked about for years, but one game has emerged from the talk to become a worldwide phenomenon: Geocaching (GPS-based treasure hunting) had, by the end of 2004, over 260,000 registered participants in more than a hundred countries.
- Galileo: Five years ago this was the system of the future, and it still is. However, the next five years will no doubt see Galileo in operation. Galileo will transmit a signal on the same L1 frequency as GPS, and future GPS products will use both GPS and Galileo signals.
- GPS chip sets have evolved in the last five years to true single-chip receivers. Using RFCMOS technology, both RF and digital functions are now performed on the same chip. Size is less than half a square centimeter.

2.8 Summary

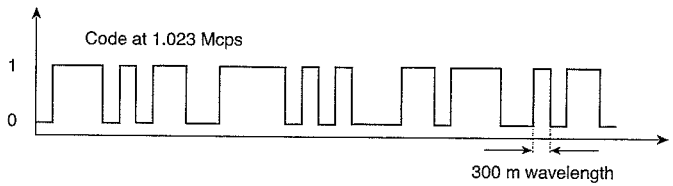
We covered a lot of ground in this chapter as we surveyed all aspects of GPS: policy, system architecture, signals, receivers, and various applications. An intuitive discussion of the signal structure and receiver design provided here should be adequate for the purposes of Chapters 5–7 (Part II) dealing with estimation of position, velocity, and time. The signals and receivers are given their due in Chapters 8–13 (Parts III and IV).

GPS is a great technological success story with far greater impact on the military and civil worlds than could have been foreseen by the designers of the system in the 1970s. As a military system, GPS has proven its worth as a "force enhancer." Civil applications unforeseen by the developers of the system are thriving and many more are on the way. Commerce in GPS equipment and services continues to grow rapidly. GPS is widely seen as the second most important gift of the DoD to the civil world. (There appears to be a consensus of opinion that the most important gift of the DoD to the civil world is the Internet, but there is no shame in being second to the Net.)

This success of GPS has also created expectations, indeed demands, which the system was not designed to meet. It is expected that the planned GPS modernization, when complete sometime between 2015 and 2020, would make determining precise position as easy as determining precise time is today. And knowledge of position would come to occupy the same important place in our daily lives as time does today.

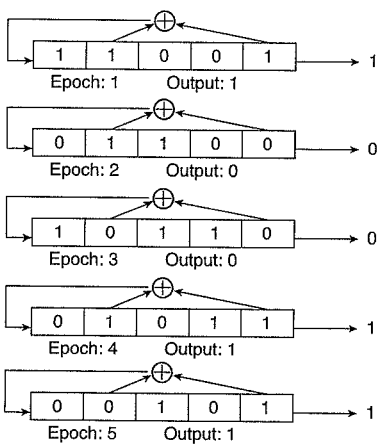
Homework Problems

2-1. A C/A-code is a series of binary *chips* that may look like:



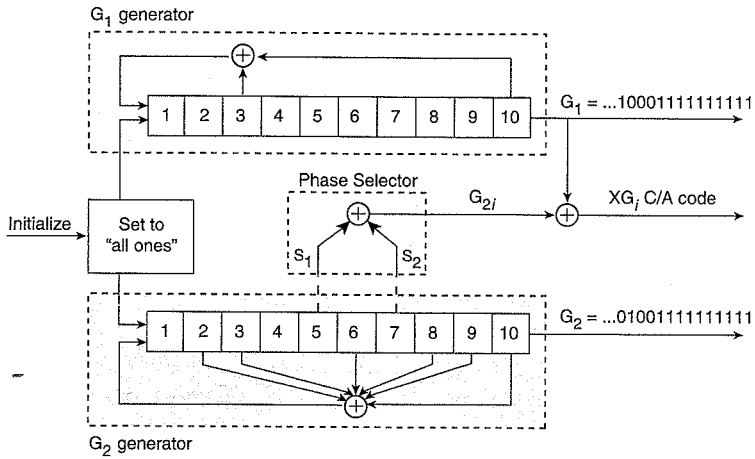
The C/A-codes can be generated with *bit shift registers*. Bit shift registers are arrays of bits (1's and 0's) of specified length. At each clock cycle (or *epoch*), all of the bits are moved one space to the right in the array. The rightmost bit is output, and a new leftmost bit is generated by an operation on the bits from the previous clock cycle.

The following is an illustration of a 5-bit long shift register in action.



Thus, the output sequence from this register, given that this register was initialized with 11001, is {1, 0, 0, 1, 1, ...}. Note that the sequence can go on forever, but eventually the output will repeat. The ⊕ symbol indicates "modulo 2 addition" (or "bitwise XOR"): $0 \oplus 0 = 0$; $0 \oplus 1 = 1$; $1 \oplus 0 = 1$; $1 \oplus 1 = 0$. You can add more than two bits: $1 \oplus 1 \oplus 1 = 1$; $1 \oplus 1 \oplus 1 \oplus 1 = 0$; etc.

The satellites are identified by their C/A-code or pseudorandom noise (PRN) sequence number. We often use 'PRN *k*' to identify both a C/A-code and the satellite transmitting it. To create the C/A-codes, GPS satellites use two 10-bit shift registers in the configuration shown below.



Note the phase selector in the middle of the diagram. S_1 and S_2 indicate which bits of the G_2 shift register are added to create the G_{2i} output at each epoch. S_1 and S_2 are different for the different satellites. For example, PRN 1 is generated by adding bits 2 and 6 from the G_2 shift register to form the G_{2i} bit; PRN 2 is formed by adding bits 3 and 7. The C/A-code output bit is XG_i. For code selection, see (CD) *Documents\IS-GPS-200D.pdf*, Table 3-I, pp. 7–8.

Briefly, the algorithm for creating the C/A-codes is:

i. Load both the G_1 and G_2 shift registers with 'all 1s'.
ii. Compute the sums from all the ⊕ operations to determine the output bit for the current epoch.
iii. Shift both registers one element to the right; load the leftmost elements of G_1 and G_2 with the appropriately calculated bits from just prior to the shift.
iv. Go back to step (ii).

(MATLAB Hint: Let **G** be a 1 × 10 matrix representing a shift register. A quick way to shift the register by one position is as follows: `G = [newBit G(1:9)]`, where `newBit` is the new leftmost element.)

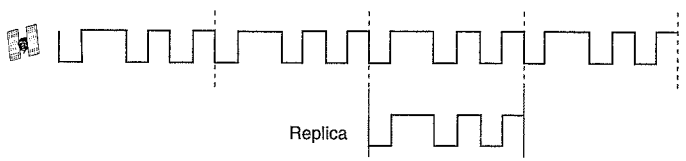
(a) Create the entire 1023-chip C/A-code for PRN 19 as a vector in MATLAB (or any other math program you'd like to use). Plot the first 16 and last 16 chips of the code. It'll be easier for you and your instructor to check the answer if you express it in hexadecimal. Verify that the first 16 chips of PRN 19, expressed in hexadecimal, are E6D6.

(b) Create the C/A-code output for PRN 19 from epochs 1024 to 2046 as a 1023-element vector. How does this vector compare with the one you generated in part (a)?

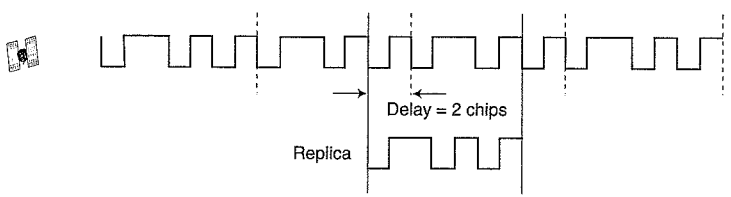
(c) Repeat part (a) for PRN 25 (we will use this C/A-code later).

(d) Repeat part (a) for PRN 5 (this C/A-code will also be used later).

2-2. A GPS receiver measures pseudorange to a satellite by examining the correlation of the received signal with a replica generated by the receiver. The receiver compares its C/A-code replica with a continuously repeating C/A-code from the satellite, as shown in the following figure, where the signal is "frozen" in time.



If the user moves away from the satellite, the repeating C/A-code's first bit has its arrival time effectively delayed with respect to the C/A-code replica, as shown below:



You will need code from Problem 2-1 to solve this problem. However, you should first convert the PRN sequences from {0,1} to {1,–1}: $0 \rightarrow 1, 1 \rightarrow -1$. This change is required because we want the non-matching parts of the PRN sequences to subtract from the result in the correlation process. (MATLAB Hint: You can transform your PRN sequences from Problem 2-1 as follows: `CA_Problem2= (CA_Problem1==0)*(1) + (CA_Problem1==1)*(-1);`)

Denoting the C/A-codes (or PRNs) associated with satellites k and l as $x^{(k)}(i)$ and $x^{(l)}(i)$, respectively, $i = 0, 1, 2, \ldots, 1022$, let us define their auto- and cross-correlation functions as follows. The normalized auto-correlation function of PRN k for shift n (= 0, ±1, ±2, ...) is defined as

$$R^{(k)}(n) = \frac{1}{1023}\sum_{i=0}^{1022} x^{(k)}(i)x^{(k)}(i+n)$$

and the normalized cross-correlation between PRNs k and l is defined for shift n as

$$R^{(k,l)}(n) = \frac{1}{1023}\sum_{i=0}^{1022} x^{(k)}(i)x^{(l)}(i+n)$$

where $x^{(\cdot)}(m + 1023) = x^{(\cdot)}(m)$. Note that both correlation functions are periodic, and

repeat as the shift exceeds 1023. (MATLAB Hint: A shifted version of a PRN can be created as follows: `CA_Replica = CA_Replica([end (1:end - 1)]);`)

(a) Plot $R^{(19)}(n)$, the auto-correlation function of PRN 19.

(b) Create a 1023-chip PRN sequence as PRN 19 delayed by 200 chips. Plot the normalized cyclic cross correlation function of PRN 19 with this delayed version. Is the peak of the correlation where you had expected it to be?

(c) Plot the cyclic cross-correlation function of PRN 19 and PRN 25. How does this plot compare with the one you made in part (a)?

(d) Plot the cyclic cross-correlation function of PRN 19 and PRN 5. How does this plot compare with the one you made in part (c)?

(e) Create three 1023-chip PRNs as follows. Define x_1 as PRN 19 delayed by 350 chips; x_2 is defined as PRN 25 delayed by 905 chips; and x_3 is PRN 5 delayed by 75 chips. Sum these three PRNs together and correlate the result with a replica of PRN 19. Is the peak of the correlation where you had expected it to be? (Hint: Addition here means simple addition: 1 + 1 = 2, and 1 – 1 = 0.)

(f) In order to simulate the effect of radio frequency noise on acquisition and tracking of a C/A-code, create a 1023-element vector *noise*, each element of which is normally distributed with zero mean and standard deviation of 4. Plot x_1, x_2, x_3, and *noise* in separate plots on a page. Scale the vertical axes of these plots so that they all have identical ranges. (MATLAB Hints: (1) `noise = 4*randn(1,1023);` (2) Use the command `subplot` to generate multiple plots on a page.)

(g) Now sum x_1, x_2, x_3, and *noise*, and correlate the result with a replica of PRN 19. Is there a correlation peak that stands out from the others? Is it where you'd expect it to be? Are you surprised by the ability of PRN 19 to survive the interference from *noise* and the other PRNs?

References

El-Rabbany, Ahmed (2002). *Introduction to GPS: The Global Positioning System*, Artech House.

Farrell, Jay A., and Matthew Barth (1999). *The Global Positioning System and Inertial Navigation*, McGraw Hill.

Forssell, Börje (1992). *Radionavigation Systems*, Prentice Hall.

FRP (2005) *2005 Federal Radionavigation Plan*, U.S. Departments of Defense, Homeland Security, and Transportation [CD *Documents\2005 FRP.pdf*]

FRS (2001) *2001 Federal Radionavigation Systems*, U.S. Departments of Defense and Transportation [CD *Documents\2001 FRS.pdf*]

Fisher, Steven, and Kamran Ghassemi (1999). Block IIF—The Next Generation, *Proc. IEEE*, vol. 87, no. 1, pp. 24–47.

Grewal, Mohinder S., Lawrence R. Weill, and Angus R. Andrews (2001). *Global Navigation Systems, Inertial Navigation, and Integration*, John Wiley & Sons.

Hoffmann-Wellenhof, B., H. Lichtennegger, and J. Collins (1997). *GPS – Theory and Practice*, 4th Edition, Springer Wein.

IEEE (1999). *Proceedings of the IEEE*, Special Issue on GPS, vol. 87, no. 1.

ION (1980). *Global Positioning System*, Vol. I, The Institute of Navigation.

IS-GPS-200D (2004). Interface Specification IS-GPS-200, Revision D, Navstar GPS Space Segment/ Navigation User Interfaces, Navstar GPS Joint Program Office [CD *Documents\IS-GPS-200D.pdf*].

Kaplan, Elliott D. (ed.) (1996). *Understanding GPS: Principles and Applications*, Artech House.

Leick, Alfred (2004). *GPS Satellite Surveying*, 3rd Edition, John Wiley & Sons.

NAPA-NRC (1995). *The Global Positioning System: A Shared National Asset, Recommendations for Technical Improvements and Enhancements*. National Academy Press.

Parkinson, Bradford W., James J. Spilker, Penina Axelrad, and Per Enge (eds.) (1996) *Global Positioning System: Theory and Applications, Vols. I and II*, American Institute of Aeronautics and Astronautics.

Seeber, Günter (2003). *Satellite Geodesy*, 2nd Edition, Walter de Gruyter.

SPS (2001). *Global Positioning System Standard Positioning Service Performance Standard*, U.S. Department of Defense [CD *Documents\SPS Performance 2001.pdf*]

Strang, Gilbert, and Kai Borre (1997). *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press.

Teunissen, Peter J. G., and Kleusberg, Alfred (eds.) (1998). *GPS for Geodesy*, 2nd Edition, Springer.

Tsui, James Bao-Yen (2004), *Fundamentals of Global Positioning Receivers: A Software Approach*, 2nd Edition, John Wiley & Sons.

US GPS Policy (1996). Presidential Decision Directive, 29 March 1996 [CD *Documents\US GPS Policy 1996.pdf*].

US PNT Policy (2004). National Security Policy Directive, 15 December 2004 [CD *Documents\US PNT Policy 2004.pdf*].

Van Dierendonck, A. J. (1997). Satellite Radio Navigation, in *Avionics Navigation Systems*, 2nd Edition, Myron Kayton and Walter R. Fried (eds.), pp. 178–282, John Wiley & Sons.

Volpe Report (2001). Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Positioning System, report prepared by John A. Volpe National Transportation Center for the U.S. Department of Transportation [CD *Documents\Volpe Report 2001.pdf*].

Wells, David (ed.) (1986). *Guide to GPS Positioning*, Canadian GPS Associates. (Download "Special Publication" from <http://plan.geomatics.ucalgary.ca/publications.php>.)

Xu, Guochang (2003). *GPS: Theory, Algorithms, and Applications*, Springer.

Chapter 3

Future Global Navigation Satellite Systems

GNSS, as we have noted previously, is the generic name for the class of systems of which GPS today is the sole working (indeed, thriving) example. GPS was designed in the 1970s, in an essentially pre-microprocessor era preoccupied with the Cold War. There have been important technological advances since, and a tectonic shift in the world political order. GLONASS, whose development began in the Soviet Union a few years behind GPS, has languished as a Russian system, but there are signs that it may yet be revived. Galileo, the European system, has the momentum and funds, and is planned to become operational in 2008. The United States wouldn't like GPS to be upstaged by the European newcomer and has ambitious plans to develop GPS III, a more capable system. The newly announced U.S. PNT Policy (2004) states that the United States "shall ensure that the utility of (GPS) civil services exceeds, or is at least equivalent to, those routinely provided by foreign space-based positioning, navigation, and timing services." It remains to be seen if this is a wishful schoolyard boast or a serious commitment.

It would be 2010, or later, before the full benefits of the modernized GPS, GLONASS, and Galileo are realized. Therefore, our goal in this chapter is simply to offer a brief overview of their current plans at a level consistent with our overview of GPS in Chapter 2.

We begin with a discussion in the next section of the performance metrics of a navigation system. These metrics provide a basis for comparing two GNSS systems and, in particular, for measuring the incremental benefits of a modernization program or a whole new system. Section 3.2 takes up frequency allocations for GPS, GLONASS, and Galileo. Before introducing the signals associated with the new GNSS systems, we describe in Section 3.3 parameters which characterize the efficacy of a spread spectrum ranging signal. With this background, we'll be ready to discuss the salient features of GPS modernization in Section 3.4, GLONASS in Section 3.5, and Galileo in Section 3.6. How a civil user would benefit from the three autonomous GNSS systems is discussed in Sections 3.7 and 3.8.

3.1 Performance Metrics

Accuracy is the simplest and most commonly used performance metric of a navigation system, but not the only one. We introduce below three criteria—accuracy, robustness, and integrity—relevant to different users of a GNSS, and discuss measures required to improve performance.

- **Accuracy.** We have quickly become accustomed to the astonishingly accurate estimates of position and time obtained from GPS. Low-end GPS receivers now routinely deliver better than ten-meter absolute positioning accuracy in autonomous mode worldwide (Figure 6.6), and two-meter accuracy in the United States where most receivers for outdoor use are now WAAS-enabled (Figure 6.7). (When EGNOS and MSAS come on line, two-meter positioning accuracy would become available in Europe and Asia.) Decimeter-, centimeter-, and even millimeter-level accuracies are now routine in relative positioning mode. We discuss the roles of satellite constellation size and measurement errors in determining the accuracy of the estimates of position and time in Chapter 6.
- **Robustness** (service availability and continuity). What good is accuracy if you can't count on the service to be available when you need it? A system must be able to withstand occasional hiccups of its own and casual or half-hearted attempts to defeat it. All radionavigation systems are vulnerable to radio frequency interference (RFI). A GNSS is especially vulnerable because the signals reaching a receiver from satellites in mid-earth orbits are extremely weak and the spread spectrum processing gain against interference is modest. Measures to enhance robustness are: frequency diversity, higher signal power, and higher processing gain. A larger satellite constellation and lower satellite failure rate will also enhance the system robustness and continuity of service.
- **Integrity.** For a safety-of-life application, what good is a service if you can't be sure whether a position estimate is accurate enough to meet your requirements? (Think of yourself as a pilot or passenger on a GNSS-guided airplane landing at an airport shrouded in fog.)

  System integrity refers to the assurance that the system is operating correctly and there will be a timely warning if any signal(s) were anomalous. Until mid-2005, when its network of monitor stations was expanded [Section 2.2.2], GPS users couldn't be certain that the signals were within specifications because a satellite may be out of view of all monitor stations for a couple of hours at a time. Galileo seized on this weakness and promoted plans to broadcast system integrity data in real time as an important advantage of their system. Galileo's plans for broadcasting integrity data remain ambitious, but the details are yet to be worked out.

  System integrity can be ensured through self-monitoring within each satellite, cross-links among satellites for constellation-level monitoring, and ground-based monitoring of the signals. The last would use the monitor stations of the Control Segment and any additional networks of reference stations fielded, for example, by the Coast Guard's Maritime Differential GPS or the Nationwide DGPS [Section 2.5]. But even such extensive monitoring can't guarantee, in general, that the position error of a user is no worse than he can tolerate. In other words, the operators of a system can only guarantee system integrity, not the integrity of the position estimates for *each* user. (WAAS [Section 2.5] is an exception, as discussed below.)

  With system integrity assured, the quality of a position estimate obtained by a user would depend upon factors outside the control of the system: number of satellites in his view and their geometry [Chapter 6], signal strengths [Chapters 10 and 13], and pseudorange measurement errors due to signal propagation uncertainties associated with the ionosphere, troposphere, and multipath [Chapter 5]. Given that the system integrity has been verified, how is a user to ascertain an error bound on a GNSS-provided position estimate? The commonly used approach consists of some form of consistency check among the measurements, often referred to as receiver autonomous integrity monitoring (RAIM). This approach requires that the measurement set be redundant — basically the larger the number of measurements available, the better. With measurements from GPS and Galileo combined, it would be possible to obtain a 'tight' upper bound on the position error without restrictive assumptions on the nature of the errors in the measurements [Misra and Bednarz (2004)].

  WAAS, developed by the U.S. Federal Aviation Administration (FAA) to

meet the requirements of precision approaches, is a success story in integrity monitoring, and a useful model for Galileo. While the WAAS error bounds are not as tight as its managers would like, they are tight enough to permit near-Category I precision approaches. But WAAS has the advantage of dealing with much simpler multipath scenarios than those on the ground; multipath due to the reflectors on aircraft can be modeled, measured, and bounded. The WAAS avionics compute a position error bound which includes the effects of geometry, noise and interference, troposphere, ionosphere, and multipath. In two years of operation, the entire performance database does not contain a single instance where the true position error exceeded the computed bound.

3.2 Frequency Allocations

A GNSS requires frequency allocations for its signals from the International Telecommunication Union (ITU), an international organization within the United Nations system where governments and private sector coordinate global telecom networks and services. Allocation of frequency bands is a tricky process which allows multiple users to coexist. In addition, the same frequencies may be allocated for different purposes in a country as long as they do not interfere with international allocations in other countries.

A GNSS requires allocations out of several bands of frequencies set aside for Radio Navigation Satellite Services (RNSS). The band 1559–1610 MHz, of which GPS L1 is a part, is allocated to RNSS. Both GLONASS and Galileo also have allocations in this band (see Figure 3.1). This frequency band is particularly attractive for safety-of-life services because it is allocated for Aeronautical Radio Navigation Services (ARNS) on a primary basis worldwide. In other words, no other users of this band are permitted to interfere with the GNSS signals.

The frequency band 1215–1350 MHz, which includes GPS L2, is allocated by the ITU to Radiolocation Services (ground radars) and RNSS on a co-primary basis. Portions of this band are also allocated to other services such as Earth Exploration Satellites and Space Research. GPS L2 operates in 1215–1240 MHz portion of this band on a primary basis but does not have protection rights over most Radiolocation Service systems. Therefore, there is no assurance that a GPS signal on L2 will not experience interference from a signal that's 'legal.' This risk may be acceptable to some users, and not to others.

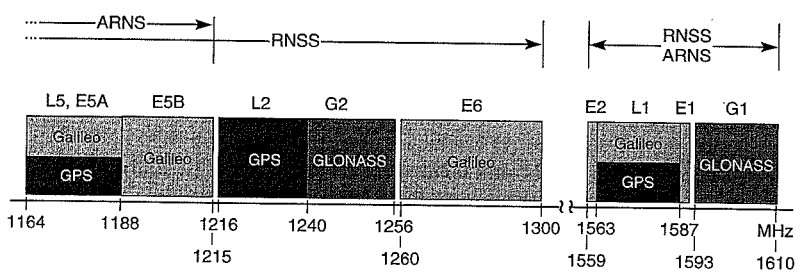


Figure 3.1 Frequency allocations for GPS, GLONASS, and Galileo.

In recent years, pressures have also grown for commercial exploitation of the electromagnetic spectrum, especially for mobile satellite communication. The GPS community was jolted when a satellite communication provider nearly succeeded in winning approval at the World Radiocommunication Conference 1997 (WRC-97) of a proposal to share a portion of the L1 frequency allocation for GPS. (WRCs are held every two to three years under the auspices of the ITU to hammer out changes in the International Radio Regulations.) The community was subsequently mobilized and not only defeated this attempt to grab a portion of the GPS frequencies but also won significant additional allocations at WRC-2000 for new signals for GPS and Galileo (see Figure 3.1).

The civil aviation community pushed for and obtained allocation for RNSS in the 960–1215 MHz band, which is currently allocated internationally to ARNS on a primary basis. This change paved the way for the third GPS civil signal. At the WRC-2000, the ITU authorizations were expanded to include L5. The GPS allocations now are (see Figure 3.1): L1 (1563.42–1587.42 MHz), L2 (1215.6–1239.6 MHz), and L5 (1164.45–1188.45 MHz).

Galileo has acquired frequency allocations for their navigation signals in three large stretches shown in Figure 3.1 as E5A+E5B (50 MHz), E6 (40 MHz), and E1+L1+E2 (32 MHz) ('E' is for Europe).

3.3 Spreading Codes and Ranging Signals

We introduced spread spectrum signals and pseudorandom noise (PRN) spreading codes in Section 2.3.1. In this section, we introduce some important concepts and vocabulary useful in characterizing the performance of these signals in navigation. Our approach is descriptive and illustrative. A careful treatment of these topics is found in Chapters 9 and 10.

In order to isolate a signal from a jumble of radio signals incident on an antenna, it's useful to introduce the concept of *orthogonality* of signals. We know from daily experience with radio and TV that two signals transmitted in non-overlapping frequency bands can be isolated and, therefore, can be thought of as being orthogonal to each other. Signals with different polarizations and the signals sent in different, pre-assigned time slots can also be thought of as orthogonal. Two signals modulating phase-shifted versions of the same carrier can also be orthogonal. We have already encountered an example in (2.1) of C/A- and P(Y)-codes modulating a carrier and a 90° phase-shifted version of it, respectively. These signals are said to be transmitted in phase quadrature. Signals modulating physically different carrier signals at the same frequency can also be made nearly orthogonal by careful selection of their spreading codes as discussed briefly below.

As noted in Section 2.3.1, a ranging signal has three components: (i) an RF carrier, (ii) a binary spread spectrum code, (iii) a binary navigation data message. The basic idea is to 'combine' the navigation message and the spreading code, and impress the resultant binary signal on the carrier using binary phase shift keying (BPSK). (See Figure 2.3 for the three components and Figure 2.4 for the BPSK-modulated carriers.)

The choice of carrier, though very important, is often a compromise between availability of certain frequency bands and their performance characteristics. The navigation data message is characterized by its bit rate and content. The navigation message can be sent out faster by raising the bit rate, but this would require a higher signal-to-noise ratio (SNR) to ensure that it is decoded correctly. The star of the show, however, is the spread spectrum code, or the

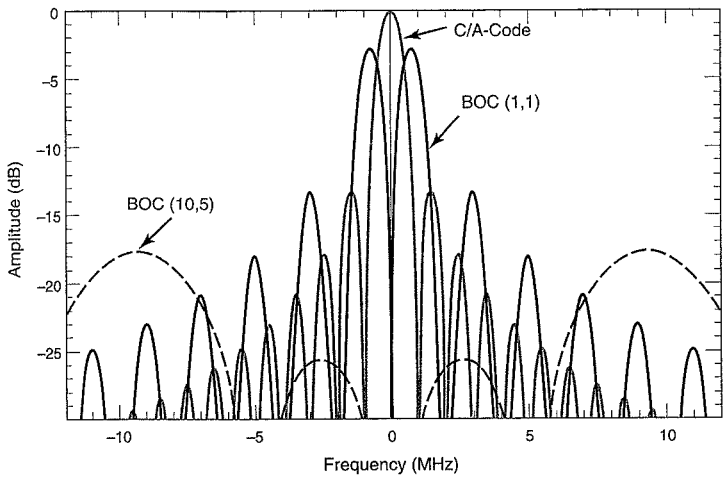


Figure 3.2 Power spectral densities of two new signals: BOC(1,1), to be broadcast at L1 by both Galileo and GPS III for civil use, and BOC(10,5), broadcast by GPS satellites starting in 2005 at both L1 and L2 for PPS users. The power spectral density of a C/A-code is shown for reference.

spreading code, so-named by communications engineers for 'hiding' a low-rate message in a code that spreads the signal power over a much larger frequency band than would have been required to transmit the message. The application of spread spectrum signals for ranging and navigation came later.

In a communication application, the attraction of spread spectrum signals is that a coded signal can be transmitted at such low power spectral density that snoopers wouldn't know it's there. The intended party would know the code, would 'de-spread' the signal to concentrate its power, and retrieve the message. Multiple users could use the same frequency band to transmit and receive messages if their codes could be isolated from the codes of others. Such codes are said to be orthogonal. They allow many users to transmit messages using the same frequencies. The technique is referred to as code division multiple access (CDMA).

In a navigation application, the navigation message is modulated on a spreading code, but carrying the navigation message is not the main function of the code. The structure of the code allows precise and unambiguous measurement of the arrival time of the signal from each of the satellites in view. Some newer signals come in pairs – one with navigation data and the other data-free (also called pilot signal). The latter allows for longer integration and, thus, offers higher sensitivity [Chapter 13].

PRN codes are characterized by the following parameters.

- Code clock rate or chipping rate (in chips/second). The faster the code, the wider the signal bandwidth (i.e., frequency band over which the signal energy is distributed [Section 8.5]), the sharper the auto-correlation peak [Section 9.2], and the more precise the estimate of the signal arrival time [Section 10.6]. Error due to signals reaching the antenna through indirect paths (e.g., after reflections) is lower, too [Section 10.7]. On the other hand, the higher the signal bandwidth, the higher the sampling rate required for digital processing, and the higher the power consumption in the receiver.
- Code length or period (in chips). The longer the codes, the lower the cross-correlations among them [Section 9.3], and the lower the interference among the signals. On the other hand, the longer the code, the greater is the computational requirement of signal acquisition. A short C/A-code (1023 chips) was deemed necessary in the 1970s for quick acquisition. The current technology can handle acquisition of longer codes by performing correlations in parallel on a large number of digital correlators.
- Code pulse shape and modulation. In order to transmit a sequence of bits corresponding to a code, a simple option is to use rectangular pulses with amplitudes ±1 and width equal to the chipping period (reciprocal of the code chipping rate). In this example, the rectangular pulse is our spreading symbol. GPS C/A- and P(Y)-codes use rectangular pulses. We introduce the following notation. C/A-code is an R-1 code, where 'R' stands for rectangular pulse and '1' represents the code clock rate of 1.023 Mcps. In this notation, a P-code is an R-10 code.

  The C/A- and P(Y)-codes place most of their power in the middle frequencies of their bands (see Figure 2.5). The outer frequencies in the band see rapidly diminishing action. In order to utilize a frequency band fully by cramming additional spread spectrum signals, engineers have devised modulations which divide the signal power mainly between two spectral lobes placed symmetrically about the center frequency. A class of these modulations is called binary offset carrier, represented as BOC(m,n), where parameters m and n determine the shape and placement of the spectral lobes [Section 9.10]. The spreading code of BOC(m,n) is the product of rectangular pulses corresponding to an R-n code multiplied by a square wave (subcarrier) of frequency $m \cdot 1.023$ MHz. The power spectral densities of a BOC(1,1) and BOC(10,5) signals, two signals of special interest to us in this chapter, are shown in Figure 3.2

  Full exploitation of the frequency band and spectral separation from the signals using R-n codes aside, the BOC codes have many attractive properties as navigation signals and have been chosen for the GPS M-codes and several of the Galileo codes. M-codes will offer better performance than the P(Y)-codes. We'll say more below.

3.4 GPS Modernization

GPS modernization was launched in the late 1990s soon after the system became operational and its full potential for military and civil applications became clear. The civil users lobbied hard for expanded benefits, pressing for changes in the system design and policies. Befitting their importance, the plans for GPS modernization to benefit the civil users were announced by Vice President Gore in 1998. The plans called for two new civil signals:

- A signal on L2 (a C/A-code signal, according to the initial plans, later changed to a more advanced design), and

- Another signal at an unspecified frequency to benefit civil aviation and other applications with safety-of-life considerations (later defined as L5, f_{L5} = 1176.45 MHz).

With GPS' primary mission as a "military warfighter support system," new measures were also deemed necessary to "protect its military use, prevent hostile use, and preserve peaceful civil use outside the area of military operations." In plain language, that means higher jam resistance for the encrypted military signals and ability to deny the civil signals locally. The U.S. program to achieve these objectives is called Navigation Warfare (Navwar).

It is planned to include the new signals in two stages by modifying the design of the Block IIR and Block IIF satellites to be launched in 2005 and beyond (see Figures 3.3 and 3.4). The number of ranging signals for civil users will increase from one to three. The military will get a pair of new, Navwar-compatible signals. There will be equally important changes behind the scene in the Control Segment to operate this system with significantly expanded capabilities.

In this section, we survey the considerations, criteria, and plans for GPS modernization. GPS is used by millions everyday and the first requirement is that any changes must ensure backward compatibility.

3.4.1 New Civil Signals and Their Benefits

Each satellite will broadcast two new signals for civil use in addition to the current C/A-coded signal on L1: a signal on L2, called L2C, and a signal on L5, called L5. The structure and characteristics of both are described briefly below and compared with the familiar C/A-code (see Figure 3.3). Additional details are given in Section 9.9, and the signals are described in

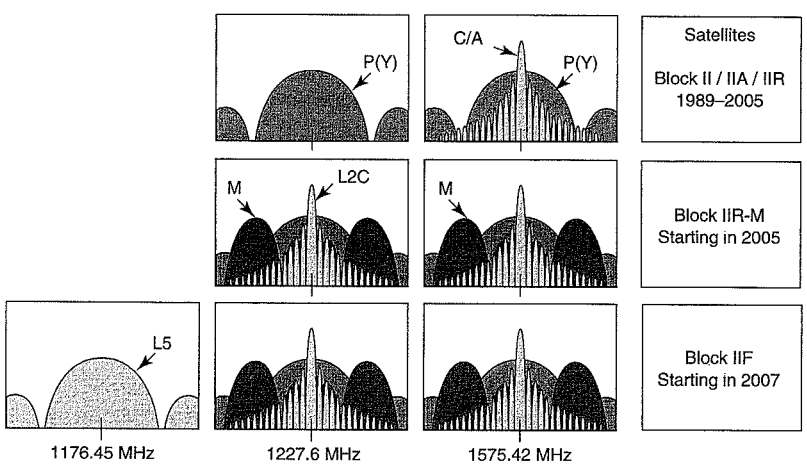


Figure 3.3 Power spectra of the current and planned GPS signals. The powers of the signals for the civil users, C/A-code on L1 and L2C on L2, are spread mainly over 2 MHz-wide frequency bands. The new civil signal at L5 will have a bandwidth of 20 MHz. M-codes, the new military signals are designed for greater security and flexibility. The P(Y)-code legacy military signals are expected to continue for a suitable period.

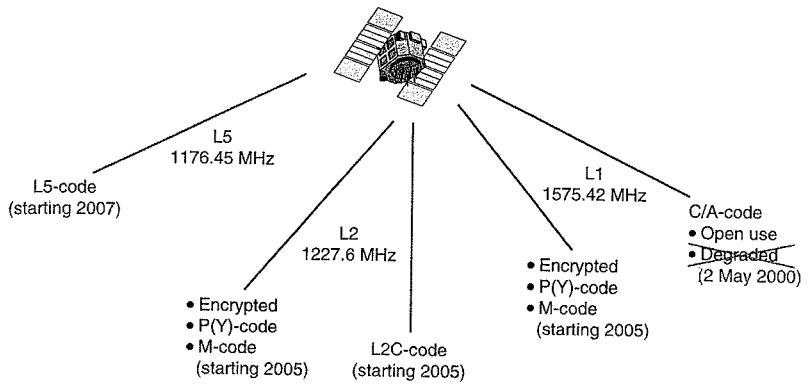


Figure 3.4 GPS signals circa 2015. Each satellite will transmit three signals for the civil users (a C/A-code on L1, L2C on L2, and a wide-band signal on L5), and four for military users (encrypted P(Y)- and M-codes on both L1 and L2).

exhaustive detail in the Interface Specifications [IS-GPS-200D and IS-GPS-705]. The broadcast of L2C started with the launch of the first Block IIR-M satellite in September 2005. The broadcast of L5 has to wait for the launch of the first Block IIF satellite in 2007.

L2C Signal

Instead of replicating C/A-code on L2, as was initially considered, a smarter signal structure has been adopted for L2C. The chipping rate, however, is the same (1.023 Mcps), and spectrally the L2C signals are similar to the C/A-codes. It's an R-1 code but with a twist. The main features are:

- Codes: Two PRN codes called CM (moderate-length code) and CL (long code) multiplexed chip by chip (i.e., a chip from CM followed by a chip from CL followed by the next chip from CM, and so on). It's a clever device to create in effect a data-free or pilot signal.
- Code clock rate: 1.023 Mcps, the same as for the C/A-codes. (The chipping rate is 511.5 kcps for CM and CL each.)
- Code length (period): CM: 10,230 chips (20 ms); CL: 767,250 chips (1.5 s).
- The L2C signal transmitted by the kth satellite can be represented, using the notation of (2.1) as:

$$s_{L2C}^{(k)}(t) = \sqrt{2P_{C2}}\; D_{C2}^{(k)}(t) CM^{(k)}(t) \cos(2\pi f_{L2} t + \theta_{L2}), \quad nT_{C2} < t \le (n+1/2)T_{C2}$$

$$s_{L2C}^{(k)}(t) = \sqrt{2P_{C2}}\; CL^{(k)}(t) \cos(2\pi f_{L2} t + \theta_{L2}), \quad (n+1/2)T_{C2} < t \le (n+1)T_{C2}$$

  where T_{C2} is the chip width ($\approx$ 1 microsecond). The L2 signal will also include the (old) P(Y)-code in phase quadrature with $s_{L2C}^{(k)}$, and a (new) M-code described below.
- Navigation data $D_{C2}^{(k)}(t)$ (25 bps, encoded with a rate-1/2 forward error correction

(FEC) code, resulting in a symbol rate of 50 symbols per second (sps)) will be carried by CM only. The data-free CL offers the option of longer integration time in environments where the signals are blocked. More on this when we discuss assisted GPS (AGPS) [Chapter 13].

- The moderate-length code (CM) is to be used for the initial acquisition, and the longer code (CL) to achieve better correlation properties, multipath mitigation, and higher resistance to interference than the C/A-code.

As noted previously, the signals transmitted on L2 do not enjoy the same institutional protection against RFI as those on L1 and, therefore, L2C is unlikely to be used for safety-of-life services. But the high-end scientific and commercial dual-frequency users will exploit the benefits of L2C. Low clock rate and excellent correlation properties are likely to make L2C an attractive option for small, low-power, and low-cost consumer applications.

L5 Signal

The planned L5 signal at the newly allocated frequency band centered at $f_{L5} = 1176.45$ MHz is a coup for civil users interested in precise positioning and safety-of-life applications. It's a longer and faster code than the C/A- and L2C-codes, and is to be transmitted with higher power. What's more, L5, like L1, is located in the ARNS band with institutional protections against RFI. The main features of the L5 signal are:

- Two signal components in phase quadrature: navigation data channel and data-free channel, sometimes referred to as I5 and Q5, respectively.
- Clock rate: 10.23 Mcps (bandwidth ≈ 20 MHz)
- Code length (period): 10,230 chips (1 ms) each for the in-phase and quadrature channels
- Spreading codes: PRN codes $g_I^{(k)}(t)$ and $g_Q^{(k)}(t)$, where the superscript refers to the kth satellite, of length 10,230 chips (1 ms). These codes are multiplied by binary sequences of length 10 and 20, respectively, to define product codes with period 10 ms and 20 ms, respectively.
- The in-phase component is modulated by PRN code $g_I^{(k)}$and navigation data (50 bps, but with FEC, the finals symbol rate is 100 sps)
- Representation of signal transmitted at L5:

$$s_{L5}^{(k)}(t) = \sqrt{2P_{L5}}\, D_{L5}^{(k)}(t)\, h_I(t)\, g_I^{(k)}(t) \cos\left(2\pi f_{L5} t + \theta_{L5}\right) + \sqrt{2P_{L5}}\, h_Q(t)\, g_Q^{(k)}(t) \sin\left(2\pi f_{L5} t + \theta_{L5}\right)$$

where $D_{L5}^{(k)}$ is the data navigation message (with a smarter organization than that used currently with C/A- and P(Y)-codes), $h_I(t)$ and $h_Q(t)$ are periodic sequences of rectangular pulses, each 1-ms long with amplitude ±1, and with period 10 ms and 20 ms, respectively.

The L5 will offer significantly better correlation properties than the C/A-code. Higher chipping rate means sharper autocorrelation peak. Longer codes give lower auto- and cross-correlation sidelobes. Longer integration time for the data-free channel means higher sensitivity.

Table 3.1 Summary of the GPS civil signals

Signal ID Center frequency	Frequency protection?	Minimum received power (dBW)	Code length Chipping rate	Nav message data rate	Planned signal availability: start–completion
L1 C/A 1575.42 MHz	yes (ARNS band)	–158.5	1023 chips 1.023 Mcps	50 bps	now
L2C 1227.40 MHz	not assured	–160.0	10,230 chips; 767,250 chips 1.023 Mcps	25 bps (50 sps)	2005–2013
L5 1176.45 MHz	yes (ARNS band)	–154.9	10,230 chips 10.023 Mcps	50 bps (100 sps)	2007–2014

3.4.2 New Military Signals

The U.S. military decided to live within the existing allocations for L1 and L2 for their new signals. The challenge, then, was to design new signal structures so that the military and civil signals will fit within the allocated bands but will have sufficient isolation to prevent mutual interference and to allow jamming of the civil signals locally, if deemed necessary.

The new military codes, called M-codes, are BOC(10,5) codes, introduced earlier. The power spectral density of these signals straddles the center of the band (see Figure 3.2). The two main lobes are located 6–9 MHz above and below the L1 and L2 band centers. The main features of the M-codes are:

- Co-existence with the current C/A- and P(Y)-codes and the planned L2C, and the ability to deny the civil signal without adverse consequences to the military.
- Better jamming resistance than Y-code signals, mainly because M-codes can be transmitted at much higher power without interfering with the C/A-code (or Y-code) receivers.
- More robust signal acquisition than for Y-code (direct acquisition, perhaps with the aid of some built-in acquisition signals.)
- Data message, now called MNAV, replaces the use of fixed format with repeated frames and subframes as in the current data message in favor of a message-based communication protocol which allows for new messages to be defined.
- New architecture for signal generation and transmission from the satellites. The Block IIF satellites will transmit two distinct M-code signals on L1 and L2 each.

The earth-coverage signal, generated by a new RF chain and transmitted from a new antenna mounted on the satellite body, will be received at a nominal power level of –158 dBW.

3.4.3 Control Segment Modernization

The Control Segment (CS) performs the vital but unglamorous behind-the-scene functions of operating GPS from day to day. The CS draws attention only when the system transmits out-of-spec signals, which, thankfully, is rare. And to keep it that way, the CS is going through a serious hardware and software upgrade. We have previously cited the expansion of the monitor station network for real-time data collection and processing for ephemeris and clock parameter prediction. Apparently, the CS modernization has not proceeded smoothly but, having tipped our hat, we'll move on.

3.4.4 GPS III

Recognizing that retrofitting features to an existing design can carry it only so far, GPS III is an attempt at redesign within the constraints of the frequency allocations, backward compatibility, and the recent agreement with Europe to transmit BOC(1,1) on L1. The basic idea is to take advantage of the newer technologies and exploit the present GPS infrastructure to meet the new military and civil needs and expectations economically. It's yet early in the planning process, and we basically have program objectives and a wish list, which include:

- Higher accuracy: Sub-meter positioning accuracy, 1–2 ns timing accuracy
- Higher integrity: assured accuracy in the face of normal degradation of the system or a planned attack; high degree of self-monitoring within a satellite and across a constellation; more secure ground-to-space and space-to-space links; improved monitoring and reporting via civil monitoring networks (e.g., WAAS and NDGPS).
- Higher availability and continuity: These considerations will determine the constellation architecture — number of satellites and their placement, satellite replacement strategy, and cross-link architecture.
- Support for specific military missions (e.g., small smart bomb delivery against a hard target, cruise missile guidance, computer/communication network synchronization), and high-power M-code signals broadcast with a directional spot beam several hundred kilometers in diameter 20 dB above the initial –158 dBW.
- Support for specific civil missions (e.g., Category IIIB/C instrument landing, precise automatic highway vehicle guidance and collision avoidance, precise construction equipment guidance).

3.4.5 Development Timetable

A launch in 2005 inaugurated the Block IIR-M series of satellites transmitting the new L2C signal and the M-codes. The next seven launches over two to three years would place the remaining IIR-M satellites in orbit. The first launch of Block IIF to inaugurate the L5 signal is currently planned for 2007 (see Figure 3.4 and Table 3.1). A full constellation of satellites

broadcasting the C/A, L2C, and L5 signals for the civil users, and P(Y)- and M-codes on both L1 and L2 for the military users, seems unlikely before 2015.

Ironically, the success of GPS is making it harder to make a case for the resources for GPS III. The satellites continue to perform well beyond their design lives, pushing off new launches. The emergence of Galileo as a serious competitor may lend urgency to the effort. The first GPS III satellite is planned to be launched in 2013, but the program is still going through birthing pains.

The recently released report on GPS by the Defense Science Board (DSB) Task Force [DSB Task Force (2005)] offers a broad-ranging review and recommendations. This report is expected to have a significant influence on the future direction of GPS.

3.5 GLONASS

GLONASS (pronounced Gluh' naas in Russian, an acronym for *Global'naya Navigatsionnaya Sputnikovaya Sistema*) has much in common with GPS in terms of its system architecture, origin as a military system, and even the terminology: C/A-code, P-code, Standard Positioning Service (SPS) and Precise Positioning Service (PPS). The GLONASS SPS based on the C/A-code was offered by the Soviet Union to the civil aviation community in 1988 at the height of *glasnost*. The P-code is intended to support the GLONASS PPS, of which little has been said in official pronouncements.

In the mid-1990s, GLONASS had looked very attractive to civil users. The system was being deployed smartly and approaching full constellation, and, best of all, its positioning accuracy was far better than that available from GPS (remember Selective Availability?). Unfortunately, the system declined and lost credibility with the user groups and receiver manufacturers. But there may now be new signs of hope. We offer a brief description below.

3.5.1 System Segments

Space segment: The constellation, consisting of 21 active satellites plus three on-orbit spares (see Figure 3.5), can be described compactly as a Walker 24/3/2 (see definition below) with an inclination angle of 64.8°. The orbital altitude is 19,100 km and period 11 h 15 min. The ground tracks repeat every eight days.

A *Walker T/P/F* represents a constellation of T satellites in circular orbits in P orbital planes. The orbital planes are spaced uniformly in right ascension of the ascending node (RAAN) [Section 4.3]. There are T/P satellites evenly distributed in each orbit. The relative phasing of the satellites between adjacent planes is F in units of $360°/T$. When a satellite in one plane is crossing the equator in northerly direction, the satellites nearest the equator in the adjacent planes are offset by $F \cdot 360°/T$ (the more easterly satellite leading the more westerly satellite). As an aside, we note here that the GPS constellation (Figure 4.14) with its uneven distribution of satellites in the orbital planes is not a Walker constellation, circular orbits and uniform spacing of the orbital planes notwithstanding.

The Soviet Union did things differently. They filled warehouses with prototype satellites designed for one-year service life, and launched them three at a time, as necessary. The result is that there are 100+ failed GLONASS satellites cluttering up mid-earth orbits (versus about 25 for GPS).

Control Segment: Performs the same functions as the GPS CS. The monitor stations were

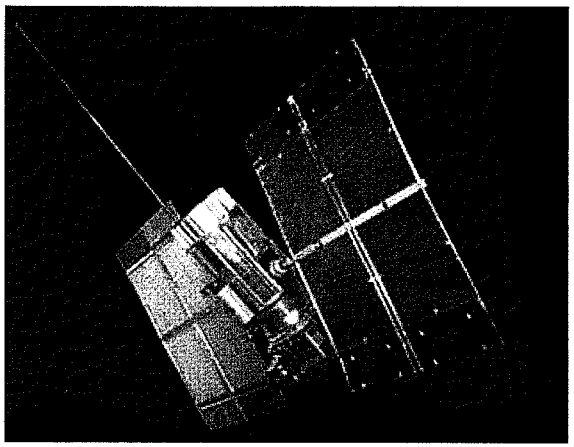


Figure 3.5 GLONASS satellite (Photo: Russian Federation Ministry of Defense).

located on the Soviet territory, and have now been limited to sites within Russia. The monitoring consisted of laser tracking in addition to the RF measurements of satellite transmissions.

With a full constellation, GLONASS SPS has the potential for offering PVT estimates comparable in accuracy with those available from GPS SPS of today.

User Segment: If it weren't for Javad, there wouldn't be any. (Indeed, Dr. Javad Ashjaee in different guises has been the main source of GLONASS receivers on the market since the early 1990s.)

3.5.2 Frequency Plan and Signal Structure

The earlier frequency allocations were: G1 (1598.0625–1607.0625 MHz) and G2 (1242.9375–1249.9375 MHz). Recently, GLONASS has picked up additional spectral real estate G3, but the plans for this frequency band have not been announced. (The GLONASS bands are often represented as L1 and L2, but we'd refer to them as G1 and G2 to distinguish them from the GPS bands.)

Like GPS, each satellite transmits three signals: On G1, a C/A-like, 511-chip long PRN code repeated with a period of 1 ms; and on both G1 and G2, a 511k-chip long PRN code with a period of 1 s. The chipping rates of the GLONASS SPS and PPS signals are half those of GPS. But the navigation data message is transmitted at the same rate as GPS — 50 bps.

Unlike GPS, which uses CDMA signaling scheme (i.e., each satellite transmits a unique PRN code on a common carrier frequency), GLONASS employs frequency division multiple access (FDMA) scheme: The same PRN is transmitted by each satellite, but at a different RF carrier frequency. The RF carriers at G1 and G2 are channelized, the channel spacing being 9/16 or 0.5625 MHz at G1, and 7/16 or 0.4375 MHz at G2. The carrier frequencies themselves are multiples of channel spacing.

At G1, the center frequency for the kth channel is

$$f_k = 1602 + k \cdot 0.5625 \text{ MHz} \quad k = -7, -6, \cdots, -1, 0, 1, \cdots, 6$$

The 24 satellites get by with 14 channels by assigning the same channel to satellites on the opposite sides of the earth (i.e., antipodal satellites). Difference in the carrier frequencies leads to low cross correlations between the FDMA signals, making them orthogonal in the terminology of Section 3.3.

Why did the GLONASS signal designers pick FDMA over CDMA? We can only speculate. A single tone jammer can take out at most one satellite signal in an FDMA system but all signals in a CDMA system.

3.5.3 Development Timetable

In mid-December 2005, the constellation consisted of 13 satellites, all of which were marked as healthy. As announced well in advance, a launch on 25 December 2005 placed three new satellites in orbit. The annual three-satellite launch late in the year has been a GLONASS ritual since 2000. (Two of the three satellites launched in 2000 and 2001 each were in service at the end of 2005, as were all nine satellites launched between 2002 and 2004.)

According to Russian officials, the system is planned to have 18 satellites in 2008 and full operational capability with 24 satellites in 2010–2011 [Revnivykh (2005)]. New, longer-life (seven years) satellites called GLONASS-M are now being deployed and newer satellites called GLONASS-K and GLONASS-MK are on the drawing boards.

Russian Presidential decrees galore over the past ten years notwithstanding, GLONASS hasn't received the necessary attention and resources to sustain it. With Galileo looking like a sure bet, it's unclear if a case can still be made for GLONASS. Stay tuned. [http://www.glonass-center.ru/]

3.6 Galileo

Galileo, a joint initiative of the European Union (EU) and the European Space Agency (ESA), is a GNSS of considerable political, strategic, and economic importance to Europe. Galileo has the extraordinary advantage of following in the path paved by GPS. A clean-slate start and newer technology would make it a formidable competitor to GPS III, which it would precede. Galileo may also turn a profit. The U.S. Government made no secret of its lack of enthusiasm for Galileo, seeing it at best as unnecessary and at worst as a security threat. The international civil user community, however, likes the idea of competition between providers of free navigation services and welcomes the new GNSS.

The political and business model adopted by Galileo is different from that of GPS. GPS is a dual-use system owned and operated by the U.S. Government and funded by the U.S. taxpayers. Galileo would be a civil system funded by a public-private partnership. The primary role for implementing and operating Galileo will go to a private consortium, which would seek revenues. Potential sources of revenue are: fees for encrypted, value-added services, taxes on chipsets manufactured or sold in Europe, and fees from equipment manufacturers who would license the IP from the concessionaire.

Unlike GPS, Galileo will have international participation and investment. The Canadian Space Agency, an associate member of ESA, is a participant. EU has also signed agreements with China and India that provide for investment and participation, though security and management issues remain to be worked out. Discussions with other countries are said to be in progress.

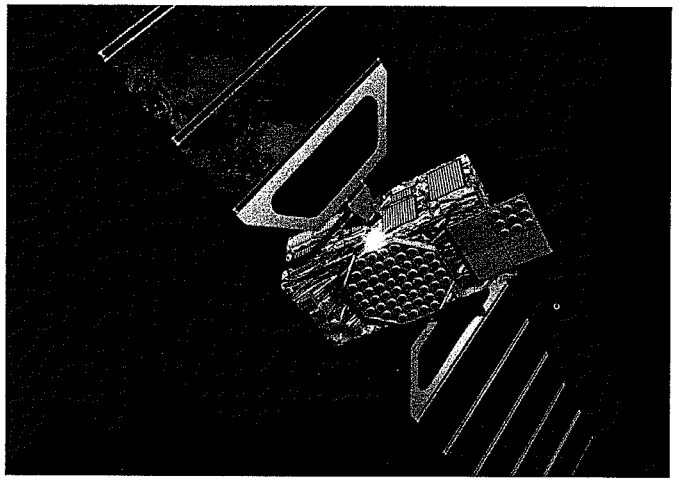


Figure 3.6 Artist's conception of a Galileo satellite (Photo: ESA)

Of course, the biggest difference between GPS and Galileo in 2005 is that GPS is being used by millions worldwide everyday while Galileo remains a gleam in the eyes of its planners and backers, a system "yet to (seriously) disturb the heavens or the ether" [Last (2004)]. The brief account below is drawn mostly from Hein and Wallner (2005)].

3.6.1 System Segments

Space segment: Planned as a Walker 27/3/1 with inclination angle of 56° (see Figures 3.6 and 3.7). The altitude is 23,222 km (orbital radius: 29,994 km). There would be a spare satellite (also transmitting) in each plane. The orbital period is 14 hrs 4 min, with ground track repeat every ten days.

Ground segment: Planned to comprise two control centers, five Tracking, Transmission and Control (TT&C) sites, ten uplink stations, 29 Galileo sensor stations, and an 'integrity processing facility.'

User Segment: Civil users would prefer to take advantage of all the signals available for free, with the usual calculation of incremental benefits versus incremental receiver cost. The signals associated with the GPS SPS (i.e., L1 C/A, L2C, and L5) and the Galileo Open Service (OS) (i.e., L1, E5A, and E5B, see below) will be used in different combinations for the different civil applications. We expect an explosive growth in dual-mode receivers when Galileo starts deploying operational satellites.

3.6.2 Services

Galileo plans to provide four types of navigation services and a search-and-rescue service. There will be an Open Service, patterned after the GPS SPS, and three fee-based services, access to which would be controlled via signal encryption. The fee-based services would offer higher performance in the form of greater assurance on availability and integrity. These value-

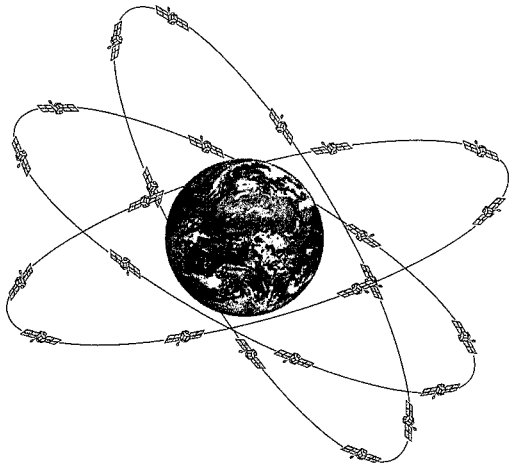


Figure 3.7 Galileo constellation (based on an ESA image).

added services would also translate reliability into some form of legal guarantees, a subject clearly outside our realm.

- Open Service (OS): Accessible to all without user fees; like GPS SPS, but claimed to be better.
- Commercial Service (CS): Fee-based service offering assured level of performance, including service availability.
- Safety-of-Life Service (SoL): Fee-based service aimed at transport applications with high integrity: authentication of signal, certification and guarantee of service, to comply with the requirements of the International Civil Aviation Organization (ICAO) and International Maritime Organization (IMO).
- Public Regulated Service (PRS): Fee-based service intended for government agencies (law enforcement, national security, and emergency services) and for military applications (Galileo-guided munition?). Controlled access with high integrity and availability, and interference-resistant signals.
- Search and rescue service (SAR): To support the 'search' part of search-and-rescue task; 98% probability of detecting a distress signal within ten minutes and 100 m accuracy; data downloaded to ground segment over dedicated UHF channels, with an acknowledgment to distress beacon.

How would Galileo's fee-based services fare? That's a good question.

3.6.3 Frequency Plan and Signal Structure

Galileo navigation signals are to be transmitted in four frequency bands (see Figure 3.1): E5A, E5B, E6, and E2-L1-E1. We'll refer to the last as L1 for short. All satellites will share the same frequency bands and utilize code division multiple access (CDMA) technique.

Each Galileo satellite is planned to transmit six navigation signals with navigation data: L1F, L1P, E6C, E6P, E5A, and E5B. Four of these signals (L1F, E6C, E5A, and E5B) will also have data-free versions transmitted in phase quadrature. The power spectral densities of these signals are shown in Figure 3.8. Note that this plot covers frequencies beyond the allocated band, but the out-of-band transmissions will conform to the ITU regulations.

Some combination of the Galileo signals will be available for each of the services listed above. Access to the full capabilities of a signal will be controlled by encryption:

- Codes and data encrypted (for CS and PRS): L1P, E6C, E6P
- Selected data fields encrypted (for CS): L1F, E5B
- No encryption: E5A

Table 3.2 summarizes the main features of L1F, E5A and E5B, the OS and SoL signals.

3.6.4 Development Timetable

Galileo's implementation is planned to proceed in three phases:

- Development and validation (2003–2005), to include building and deployment of two prototype satellites, called Galileo System Testbed Satellites, for in-orbit validation, and deployment of the first four operational satellites
- Deployment of space and ground segments (2006–2007)
- Commercial operation (starting in 2008).

The Galileo Joint Authority (GJU) was established in 2003 to manage the program until the end of its development phase in 2006. EC and ESA are the founding members of the GJU. Other organizations and countries, perhaps even the United States, may acquire a stake in it. A private sector concessionaire, a consortium of leading European aerospace and telecommunications companies, would take over from the GJU in 2006. The commercial concession-holder

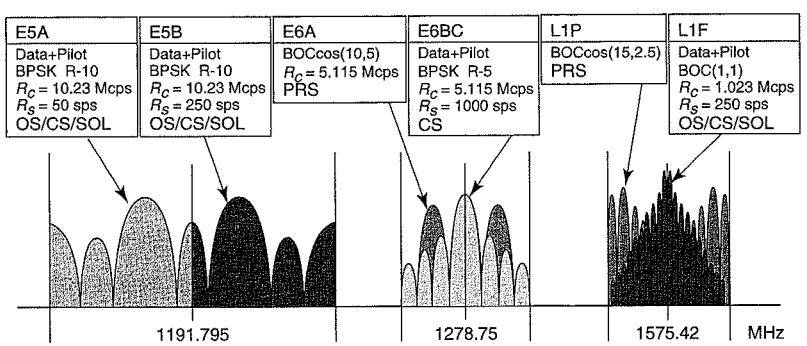


Figure 3.8 Power spectral densities of the planned Galileo signals (R_c: chipping rate, R_s: symbol rate) [adapted from Hein and Wallner (2005)].

Table 3.2 Signal characteristics for the Galileo Open Service (OS) signals

	L1F	*E5A*	*E5B*
Modulation	BOC(1,1)	R-10	R-10
Chipping rate (Mcps)	1.023	10.23	10.23
Channels	2 (data and pilot)	2 (data and pilot)	2 (data and pilot)
Data rate (bps/sps)	125/250	25/50	125/250
Encryption	selected data fields	none	selected data fields
Minimum received power at 10° elevation (dBW)	–157	–155	–155

will deploy the space and ground segments and run the Galileo operating company, supervised by the Galileo Supervisory Authority, a new EU public entity. The concessionaire will seek out revenue-producing applications and come up with private sector backing for the deployment phase.

The cost of development and fielding of the system is estimated at €3.5 billion, of which about €1.5 billion would be raised by the concessionaire. The yearly cost of operation and replenishment is estimated as €200 million.

ESA, responsible for managing the space and ground segments, awarded in 2003 contracts for two test bed satellites and four operational satellites. GIOVE-A, the first Galileo In-Orbit Validation Element (the new name for a test bed satellite), was launched in December 2005. It will transmit the Galileo signals to secure use of the frequencies allocated by the International Telecommunication Union (ITU). The in-orbit testing of the critical technologies for the future navigation payload will continue with GIOVE-B, soon to follow.

The US-EU Agreement on Interoperability of GPS and Galileo [US-EU Agreement (2004)] appears to have resolved the contentious issues after drawn-out, often bitter, negotiations. Now the European countries only have to negotiate among themselves to divide up the work and its rewards, and that may prove to be no easier. Europe also faces other challenges: Attract private investment capital for Galileo deployment, reconcile the economic and political interests of the foreign partners investing in Galileo, accommodate the interests of the newer members of EU, and build Galileo in a timely fashion.

Galileo fell short of its goal to launch four operational satellites in 2005. The current schedule calling for an operational system to be ready in 2008 appears too optimistic. It's more likely in 2010–2012. [http://www.galileoju.com, http://www.esa.int/esaNA/]

3.7 Compatibility and Interoperability of GPS, GLONASS, and Galileo

As the Europeans learned, RF compatibility between Galileo and GPS is easier to achieve than political compatibility. In concrete terms, compatibility with the U.S. Navigation Warfare policy [US PNT Policy (2004)] required that the Galileo signals be 'jammable' over an area of conflict without impact to the U.S. military signals. The good allies have now resolved their dispute over signals at L1: Galileo obliged by narrowing the bandwidth of the OS signal, changing it from BOC(2,2) to BOC(1,1), and moving the PRS signal farther away from the M-code slot (clear out of the GPS L1 band). GPS offered a boost to Galileo by agreeing to add to GPS III a BOC(1,1) signal, to be called L1C, for civil use at L1. Figure 3.9 shows the power spectral densities of the signals now planned for L1. (L1C seems far away and we haven't included it in Figures 3.3, 3.4, and 3.9 for simplicity.)

By RF compatibility between two signals or systems we simply mean that neither degrades the performance of the other in a significant way. For example, the C/A-, P(Y)-, and M-code GPS signals at L1 are designed to be compatible. It's easier for GPS and GLONASS to be compatible, or orthogonal in the terminology of Section 3.3, because their signals occupy different frequency bands (see Figure 3.1). Galileo signals overlay GPS signals in L1 and L5 bands and there is bound to be some impact, but it has been shown to be insignificant [Godet *et al.* (2002)].

Compatibility may be enough for military users, but civil users would go a step further and demand interoperability, i.e., the ability to combine signals from the two autonomous systems at the receiver level so that the performance is at least as good as the better of the two at that place and time. A user could then choose between the two independent, redundant services, or one robust integrated service. Below we look at what's required for interoperability beyond RF compatibility.

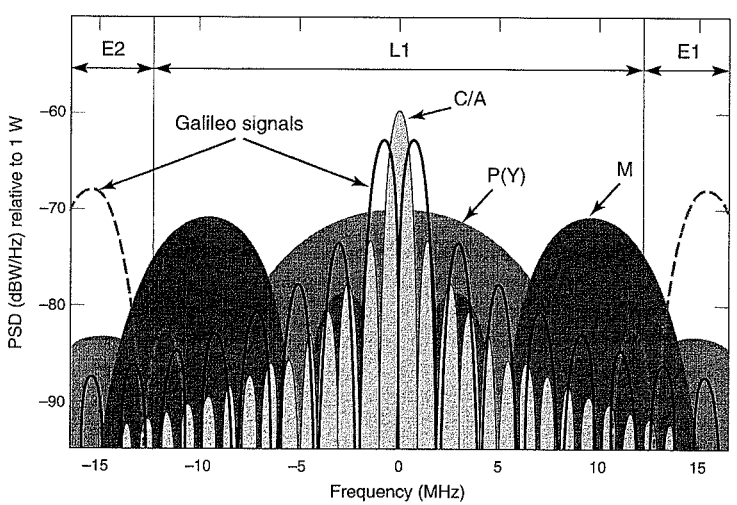


Figure 3.9 Power spectral densities of GPS and Galileo signals at L1 (f_{L1} = 1575.42 MHz).

As noted in Section 2.2.2, a GNSS requires a global coordinate frame in which to specify the positions of its satellites and users, and a time scale relative to which to specify the offsets of the satellite and user clocks. In order for GPS, GLONASS, and Galileo to be interoperable, it's required that transformations between their coordinate frames and time scales be known.

The Galileo Terrestrial Reference Frame (GTRF) is an independent realization of the International Terrestrial Reference System (ITRS) [Section 4.1.3]. Like WGS 84, GTRF shall be realized by defining the coordinates of a set of Galileo ground monitor stations. The difference between the two systems is expected to be at centimeter level, insignificant for most users of the two systems. The Soviet Union had defined its own autonomous coordinate frame called SGS 85, later renamed by the Russians as PZ-90, without regard to ITRS. In the early 1990s, it became clear that there were significant differences between WGS 84 and SGS 85. A transformation between the two was estimated empirically [Misra and Abbot (1994)].

The Galileo System Time (GST), like GPS Time (GPST), shall be an independently realized time scale maintained within a certain tolerance of the international time standard [Section 4.2]. According to the EU-US agreement of 2004, the offset between GST and GPST will be broadcast by both systems accurate to within 5 ns (2 sigma). The bias between the time scales of two GNSS systems can also be estimated by a receiver by 'sacrificing' a pseudorange measurement [Section 6.1]. That's how GPS+GLONASS receivers have dealt with their independent time scales.

Interoperability is not to be mistaken for optimality. GPS, GLONASS, and Galileo were not designed to optimize performance for any civil applications or to simplify the design of a common receiver. For example, Galileo chose not to provide a signal at L2 in spite of a petition in 2003 from the potential civil users who saw its value in combination with GPS L2C for assisted GPS (AGPS), potentially a huge market. Instead, Galileo appears to be steering AGPS to E5A. How this application evolves remains to be seen. For an interesting discussion of the issues associated with the use of GPS+Galileo for different civil applications, see Van Dierendonck (2005).

3.8 Performance of GPS+GLONASS+Galileo

We introduced the performance metrics of a GNSS earlier [Section 3.1] in anticipation of this discussion. More signals from more satellites of GNSS systems with autonomous control segments will offer improvements relative to each of the performance metrics: accuracy, robustness, and integrity.

A user will see more satellites. Figure 3.10 illustrates this point with a side-by-side comparison of two histograms of the number of satellites in view (elevation > 5°) of users worldwide from the baseline GPS constellation of 24 satellites and a hypothetical GNSS with Walker 48/6/1 constellation. Most users see six to ten GPS satellites and 14 to 19 satellites from the hypothetical GNSS. A GPS+Galileo constellation would be slightly larger than this hypothetical GNSS, and the GPS+GLONASS+Galileo constellation would be much larger.

Users facing sky blockages would benefit from a larger constellation in general. But there would be little improvement if the blockage is serious, as in an urban canyon — you can't get a position estimate from three satellites along a line in the sliver of the sky visible, and it

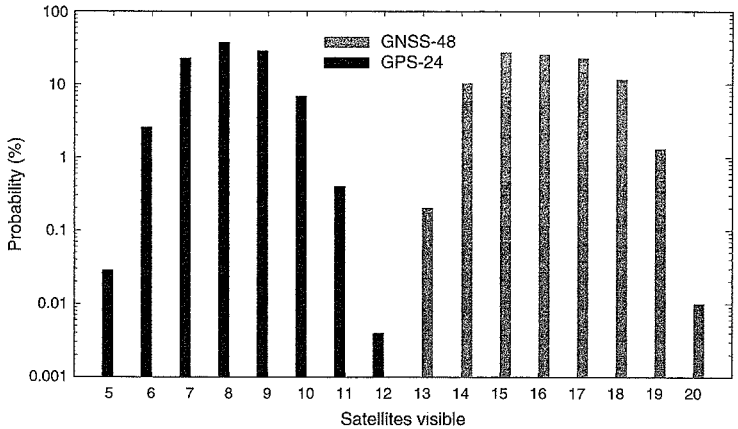


Figure 3.10 Histograms of the number of satellites in view (elevation > 5°) from the baseline GPS-24 and a Walker 48/6/1 representing the combined constellations of two GNSS systems.

doesn't get much better with eight satellites along the same line. Urban land navigation would continue to require augmentation of some sort.

A civil user would have access to many more signals. The number would increase from one from GPS today (not counting the 'quasi-authorized' use of the L2 carrier by geodetic-quality receivers) to three from GPS, three from Galileo, and one or more from GLONASS. The different applications would choose from the menu of signals to address their specific requirements. The frequency diversity would alleviate concerns over accidental interference and constitute a key step toward achieving robustness of service.

Position estimation from pseudoranges obtained from code phase measurements is limited to about one-meter accuracy. For higher accuracy, we have to go to carrier phase measurements, which can provide centimeter- and even millimeter-level accuracy in relative positioning mode. We should note here that the use of carrier phase measurements was not foreseen by the designers of GPS, but is now common for geodesy and applications in earth sciences, where the position estimates are not required in real time. Precise navigation with GPS carrier phase measurements has been explored, but the process appears fragile [Chapter 7]. The availability of additional signals from GPS, GLONASS, and Galileo would lead to *robust* centimeter- and decimeter-level position estimates in real time for navigation. Such order-of-magnitude improvement in real-time positioning accuracy would be a stunning and most visible gain from the multiple GNSS systems.

Galileo integrity monitoring architecture is still under discussion, but there are claims of guaranteed service in terms of accuracy and availability. At a minimum, we expect Galileo to provide integrity comparable to what's available to aircraft 200–300 feet above the ground from WAAS today, and from EGNOS and MSAS in the near future. With a larger number of satellites in view and each satellite offering more signals, new approaches become available for a user to generate tighter error bounds reliably through robust algorithms implemented in the receiver [Misra and Bednarz (2004)].

3.9 Summary

The GPS modernization program is well underway with the launch in 2005 of the first Block IIR-M satellite transmitting an additional signal for the civil users (L2C) and two new M-code signals for the military. The GPS III, the third-generation system, is on the drawing boards. The development of Galileo, the ambitious European system, is well along. And GLONASS, the Russian system, which appeared orphaned by the dissolution of the Soviet Union, shows signs of having been adopted by the Russian Federation.

GPS and Galileo are introducing new and more capable spread spectrum ranging signals with special features: BOC modulation, very long codes, convolutional data encoding, and pilot signals. These features allow fuller use of the allocated frequency bands with more accurate estimation of pseudoranges and robust demodulation of the navigation data with even lower signal-to-noise ratios.

The true consequences of two or more autonomous, interoperable GNSS systems can't be foreseen fully. Think of the modest expectations from GPS in the 1970s, and even the mid-1980s. In the next five to ten years, all GNSS users will see more signals from more satellites. It appears safe to say that most civil receivers will be multi-mode and the improvements in accuracy, robustness, and integrity of service will range from significant to dramatic. It's an exciting time to be associated with the development of these new GNSS systems and their applications.

Homework Problems

3-1. We formulated the problem of positioning with GPS pseudorange measurements in (2.5). (See also Figure 1.9 for an idealized case.) Consider now positioning with such measurements from two autonomous GNSS systems. Each GNSS uses its own, independently defined coordinate frame, but you know the transformation between the two. Each GNSS also uses its own clock as an independent time reference. The instantaneous offset between the two clocks changes with time and is unknown. How would you formulate the positioning problem?

3-2. Suppose the pseudorange measurement errors from the different satellites in (2.5) are independent and identically distributed with zero mean and a certain variance. If these errors could be cut in half, the position estimation error would be cut in half. (We actually prove this in Chapter 6.) But what if the errors remained the same but you had pseudorange measurements from twice as many satellites? An answer appealing to intuition will suffice.

References

Avila Rodriguez, J., M. Irsigler, G. Hein, and T. Pany (2004). Combined Galileo/GPS Frequency and Signal Performance Analysis, *Proc. ION GNSS 2004*, pp. 632–649.

Daly, P., and P. Misra (1996). GPS/GLONASS, in *The Global Positioning System — Theory and Applications*, Vol. 2, B. Parkinson, J. Spilker, Jr., P. Axelrad and P. Enge (eds.), American Institute of

Aeronautics and Astronautics (AIAA), pp. 243–271.

DSB Task Force (2005). Final Report of the Defense Science Board Task Force on "The Future of the Global Positioning System," U.S. Department of Defense [CD *Documents\DSB Task Force 2005.pdf*].

Godet, J., J.C. de Mateo, P. Erhard, and O. Nouvel (2002). Assessing the Radio Frequency Compatibility between GPS and Galileo, *Proc. ION GPS 2002*, pp. 1260–1269.

Hein, G. and S. Wallner (2005). Development and Design of Galileo, *Proc. ION Annual Meeting*, pp. 94–142.

IS-GPS-200D (2004). Interface Specification IS-GPS-200, Revision D, Navstar GPS Space Segment/ Navigation User Interfaces, Navstar GPS Joint Program Office [CD *Documents\IS-GPS-200D.pdf*].

IS-GPS-705 (2005). Interface Specification IS-GPS-705, Revision D, Navstar GPS Space Segment/ User Segment L5 Interfaces, Navstar GPS Joint Program Office [CD *Documents\IS-GPS-705.pdf*].

Lachapelle, G., M.E. Cannon, K. O'Keefe, and P. Alves (2002). How Will Galileo Improve Positioning Performance? *GPS World*, September 2002, pp. 38–48.

Last, David (2004). GPS and Galileo: Where Are We Headed? *ION Newsletter*, vol. 14, no. 1, p. 1.

McDonald, Keith (2005). GPS Modernization: Global Positioning System Planned Improvements, *Proc. ION Annual Meeting*, pp. 36–98.

Misra, P., and S. Bednarz (2004). Precision Approaches: Robust Integrity Monitoring using GPS+Galileo, *GPS World*, April 2004, pp. 42–49.

Misra, P., and R. Abbot (1994). SGS 85–WGS 84 Transformation, *manuscripta geodaetica*, vol. 19, pp. 300–308.

Revnivykh, S. (2005), GLONASS Update, Civil GPS Service Interface Committee (CGSIC), 45th Meeting, Long Beach, Calif. [http://www.navcen.uscg.gov/cgsic/meetings]

US PNT Policy (2004). National Security Policy Directive, 15 December 2004 [CD *Documents\US PNT Policy 2004.pdf*].

US-EU Agreement (2004). Agreement on the Promotion, Provision, and Use of Galileo and GPS Satellite-based Navigation Systems and Related Applications [CD *Documents\US-EU Agreement 2004.pdf*].

Van Dierendonck, A.J. (2005) GNSS User Assessment of the Plans and Benefits of GNSS Modernized Signals and Services, *Proc. ION Annual Meeting*, pp. 201–209.

Chapter 4

GPS Coordinate Frames, Time Reference, and Orbits

In this chapter, we expand on some ideas discussed briefly in Chapters 1 and 2. These ideas are associated with three areas central to the development and use of GPS:

- definition of coordinate systems in which to represent positions of the users and GPS satellites, and to analyze satellite motion,
- characterizations of an instant in time and a time interval,
- characterization of a satellite orbit and determination of the satellite position and velocity at an instant.

As noted in Chapter 1, definition of coordinate systems, timekeeping, and celestial mechanics have had complicated inter-relationships going back hundreds of years. Each blossomed in the seventeenth and eighteenth centuries with important contributions by the icons of the scientific community: Galileo, Kepler, Newton, Gauss, Huygens, and Euler, to name a few. Our objective in this chapter is really quite modest. We attempt a brisk survey of each of the three areas and introduce the main ideas relevant to GPS without straying too far beyond the concerns of a GPS engineer. An interested reader can pursue these topics further through the references listed at the end of the chapter.

The main purpose of Section 4.1 is to introduce the *World Geodetic System 1984* (WGS 84) which defines an earth-centered, earth-fixed (ECEF) Cartesian coordinate system in which the positions of the GPS users and the satellites are expressed. WGS 84 also defines an *ellipsoid of revolution* to serve as a model for the shape of the earth, and another surface, called the *geoid*, to model the mean sea level as the reference for measurement of height everywhere on earth. In order to appreciate the magnitude of the WGS 84 enterprise, we review briefly the subtleties associated with the definition and implementation of an ECEF coordinate system and definition of height. A student can go over these lightly in the first pass and return to them as necessary.

The main idea in Section 4.2 is the definition of *GPS Time* (GPST), a uniformly flowing time scale used by GPS to specify a time instant or a time interval. Now, there doesn't seem to be anything special about a uniformly flowing time scale. After all, who would want a time scale that speeds up and slows down, or has discontinuities? Actually, it turns out that finding a uniform time scale wasn't easy and, once it was discovered, we found it necessary to add discontinuities to it in the form of leap seconds. That's the current, internationally accepted civil time standard called *UTC*. We review briefly the history of timekeeping in Section 4.2. Again, we recommend a once-over-lightly approach.

The main purpose of Section 4.3 is to introduce the contents of the 50-bits-per-second (50-bps) navigation data message broadcast by each GPS satellite, with particular attention to a 16-parameter orbital set which allows a user to compute the position and velocity of a satellite at any instant. As in the previous sections in this chapter, we attempt to fill in the background. In order to illuminate the characterization of a satellite orbit, we return to Kepler's laws of planetary motion, which, in our view, all college graduates should be required to know. We discuss the ideal and perturbed Keplerian orbits and introduce the GPS orbital parameters and computation of satellite position.

Finally, Section 4.4 gives a brief description of the GPS satellite constellation and discusses satellite visibility or coverage analysis. A student may skip topics marked with an asterisk (*) on first reading without a significant loss of continuity.

4.1 Global Coordinate Systems

In this section we concentrate on the definition of global coordinate systems in which to represent the position of a point on the earth, and the position and velocity of a GPS satellite orbiting the earth. Precise definition of such coordinate systems is essential for obtaining precise position from a satellite navigation system. The problem would have been easier had the earth been a rigid, uniform sphere. It's not. It has been known since the days of Newton that the earth bulges at the Equator and is somewhat flattened at the poles. The density of the earth is not uniform either. These factors complicate everything we want to do in this section.

In order to compute the distance between a user and a GPS satellite, we'd like to express the positions of both in a common coordinate system. The position of a user is expressed conveniently in a coordinate system that is fixed to the earth and moves with it. In fact, it is a requirement: We want the coordinates of a stationary object to remain fixed. Such a coordinate system, however, is not suited to the analysis of satellite motion. The motion of a satellite is governed by the equations of motion, in particular, by Newton's second law relating force and acceleration, both expressed in an inertial reference system. An inertial coordinate system is defined as fixed in space, or in uniform motion without any acceleration. Clearly, a spinning, earth-fixed coordinate system suited to expressing positions of points on the earth will not do.

Our plan in this section is to define a Cartesian coordinate system fixed to the earth and another fixed in orientation relative to the so-called 'fixed' stars. Next we define a mathematically tractable surface to serve as a model for the shape of the earth, allowing us to represent a point by latitude, longitude, and height. This will take us to the definition of WGS 84, the main topic of this section.

4.1.1 Terrestrial and Inertial Reference Systems

Conventional Terrestrial Reference System (CTRS)

Consider a Cartesian coordinate system with its origin at the center of mass of the earth, the z-axis coinciding with the axis of rotation, and x-axis coinciding with the intersection of the Greenwich meridian with the equatorial plane. This would seem to be a good candidate for an earth-fixed coordinate system, if we can deal with a complication: The rotation axis is not fixed in relation to the solid earth. The pole of rotation wanders around on the surface of the earth in a roughly circular path, moving several meters over the course of a year. This phenomenon is called *polar motion.*

A change in the position of the pole produces a corresponding change in the definition of the equatorial plane and, therefore, a change in the latitude and longitude of all points on the earth. Considering that the wandering positions of the pole of rotation over the past century can be enclosed in a circle of 15-m radius, such changes are not large. But these changes cannot be disregarded if we require centimeter- or millimeter-level accuracy in specifying a position. No satisfactory models exist to fully predict the polar motion. The instantaneous pole position can be determined with an accuracy of 1–2 cm on the earth's surface by measurements of signals from space using, e.g., very long baseline interferometry (VLBI), satellite laser ranging (SLR) [Seeber (2003)], and GPS. The coordinates of the pole of rotation are computed daily and made available on the Internet by the International Earth Rotation Service (IERS), an international scientific organization dedicated to such matters [http://www.iers.org].

To get around the difficulty of the wandering pole, geodesists have defined an average position of the earth's pole of rotation between the years 1900 and 1905. This point, fixed to the earth's crust, is known as the *Conventional Terrestrial Pole* (CTP). Now we can define a Cartesian coordinate system fixed to the earth as follows:

- origin at the center of mass of the earth,
- z-axis through the CTP,
- x-axis passing though the intersection of the CTP's equatorial plane and a reference meridian. The y-axis is defined in the equatorial plane to complete a right-handed coordinate system.

The earth-centered, earth-fixed (ECEF) Cartesian coordinate system defined in this way is referred to as the *Conventional Terrestrial Reference System* (CTRS). It is illustrated in Figure 4.1 with axes marked as x_T, y_T, and z_T, where the subscript T stands for terrestrial.

The definition above of the CTRS is abstract. Who has seen the center of mass of the earth? And, how do we determine the coordinates of a point in this system? Actually, the CTRS is *implemented* or *realized* as a reference frame by adopting the coordinates of a set of points. The basic idea is quite straightforward. As a simple exercise, take a clean sheet of paper, mark two points on it, and assign them coordinates arbitrarily. You have implemented a coordinate frame: The location of the origin and the orientation of the x- and y-axes are uniquely determined. Can you pick the coordinates of three or more points in this way and still obtain a unique two-dimensional coordinate frame? The answer is yes, if the coordinates of the points selected are *consistent*. In practice, these coordinates would be determined from measurements (of angles, lengths, Doppler frequency shifts, etc.), and would have some error. We could still implement a coordinate frame as one that fits the data best in some sense (e.g., a least-squares fit). In fact, this is how the CTRS is implemented.

An ECEF coordinate frame is realized by establishing the coordinates of a globally distributed set of points. The more accurate these position coordinates, the smaller the residual

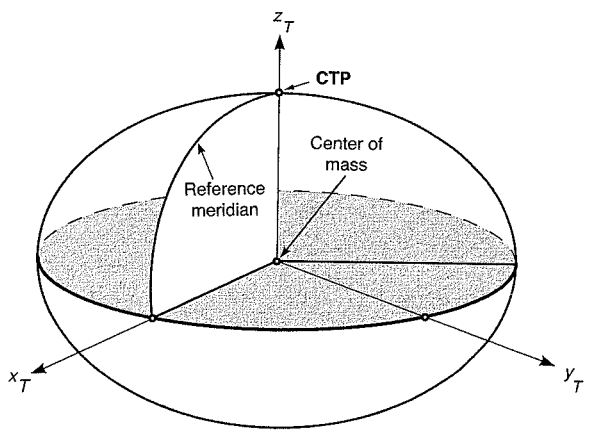


Figure 4.1 Conventional Terrestrial Reference System (CTRS).

error in the fit, and the more accurate the realization of the coordinate system. Note that in such realization of the CTRS, the zero meridian is no longer defined in terms of the famous brass line in the courtyard of the storied observatory at Greenwich, but on the basis of statistical calculations to reconcile the coordinates of many stations around the world. The reference meridian is often referred to as the *Mean Greenwich meridian*.

Realization of the CTRS is a large undertaking. A realization that is central to GPS is the *World Geodetic System 1984* (WGS 84) developed by the U.S. Department of Defense (DoD). GPS provides the position of a user in WGS 84 ECEF coordinate frame. Another important, and even more accurate, realization is the *International Terrestrial Reference Frame* (ITRF) defined by the scientific community. Why can't everyone use the same coordinate frame, say, ITRF, if it's more accurate? Because ITRF is a tool for research, and can be adapted and changed as it serves the needs of the scientific community. If you were the DoD, you too would want to control your maps and grids, and change and refine them as it suited your needs.

The Cartesian coordinates, while convenient for calculations, are cumbersome in daily use. Imagine giving your position coordinates in meters as (1,510,885, –4,463,460, 4,283,906). One can tell that you are in the northern hemisphere ($z > 0$). Those who remember their sine/cosine tables and can do some quick computations in their heads can tell that you are at mid-latitudes (40°–45°). It's hard to tell if you are on the earth, or above or below it, and by how much. And, of course, all three coordinates would change in general as you climb up a pole, or move to an upstairs room. Curvilinear coordinates (latitude, longitude, and height), which we introduce below, are better suited for everyday use.

Conventional Inertial Reference System (CIRS)

In order to formulate the problem of satellite motion around the earth in accordance with Newton's laws, we need an inertial (also called celestial or space-fixed) coordinate system in which to express the force, acceleration, velocity, and position vectors. An inertial reference system is defined to be stationary in space, or moving with a constant velocity (no acceleration).

The terrestrial reference system defined above, which spins with the earth, doesn't meet these requirements. While defining CTRS, our focus was on tying down three mutually orthogonal axes to the earth precisely. How the earth was oriented in space didn't matter. In defining CIRS, we aim to specify the orientation of the earth in space. We do so by defining three mutually orthogonal directions in inertial space, and relating them to CTRS via a (time-dependent) transformation.

An inertial coordinate system can be defined as follows:

- origin at the center of mass of the earth,
- z-axis along the axis of rotation,
- x-axis in the equatorial plane pointing toward the vernal equinox (i.e., the direction of intersection of the equatorial plane of the earth with the plane of the earth's orbit around the sun). The y-axis is defined to complete a right-handed system.

Strictly speaking, the above definition does not meet the requirements we gave earlier. The center of mass of the earth and, therefore, the above coordinate system, move around the sun with varying speed [Kepler's second law, Section 4.3]. The coordinate system, however, can be

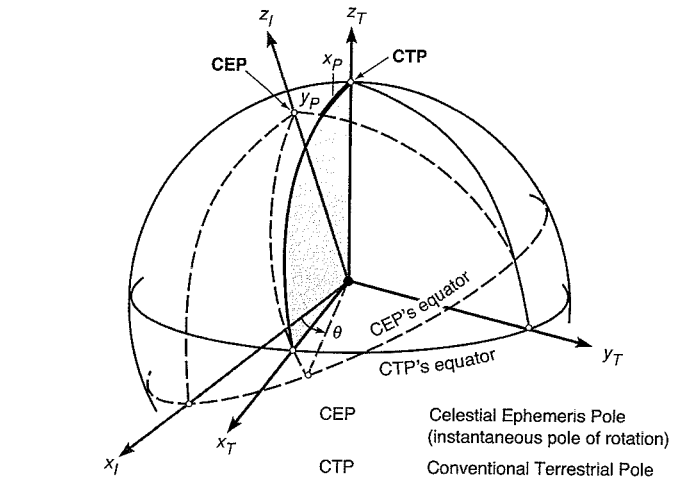


Figure 4.2 Inertial and terrestrial reference systems.

thought to be inertial over short periods. The above-defined CIRS is illustrated in Figure 4.2 with axes marked as x_I, y_I, and z_I, where the subscript I stands for inertial.

A trickier part of the definition of the space-fixed coordinate system relates to the fact that the axis of rotation of the earth is not fixed in space in relation to the distant stars. (Note that this is saying something different from the earlier remark that the axis of rotation is not fixed in relation to the earth's crust.) The motion of the earth's spin axis in space is a composite of periodic components referred to as *precession* and *nutation*. The gravitational attraction of the sun and the moon on the non-spherical earth causes the rotational axis of the earth to precess in space like a top (period: about 26,000 years!). In addition, the axis undergoes a small 'nodding' motion called nutation (principal period: 18.6 years). If the earth were spherical and homogeneous, it wouldn't precess or nutate. The good news is that precession and nutation are well understood, and can be traced accurately to any epoch [Bock (1998)].

Just as the CTRS is realized through the coordinates of a set of points on the earth, the CIRS is realized through a catalog of positions and proper motions of a set of fundamental stars and VLBI observations of quasars and other extra-galactic objects. Due to the precession and nutation of the earth's rotational axis, however, the coordinates of celestial objects change slowly, and must be referred to a specific epoch. By convention, the celestial reference system is defined as corresponding to the epoch J2000 (1 January 2000 12h, or Julian date 2,451,545.0), which we define in Section 4.2.1.

In order to describe the orientation of the earth in space and to relate the CIRS to CTRS, we have to account for two additional complications. First, there is the unpredictable polar

motion, mentioned above. Secondly, as we discuss in Section 4.2.1, the spin rate of the earth is not uniform. Five parameters, called earth orientation parameters (EOP), are required to specify the orientation of the earth in space precisely at an epoch (see Figure 4.2):

- two angles (x_p and y_p) to specify the direction of the rotation axis within the earth (position of the instantaneous pole of rotation relative to the CTRS) to account for the polar motion,
- two angles to characterize the direction of the earth's rotational axis in space to account for precession and nutation,
- an angle (θ) to describe the rotational motion of the earth in terms of the angle of the Mean Greenwich Meridian relative to the vernal equinox accounting for the non-uniform rotational rate.

Unlike precession and nutation, polar motion is not fully predictable, and the displacement of the instantaneous pole of rotation (x_p, y_p) from the CTP is determined empirically. Parameter θ specifies the angle between the Mean Greenwich Meridian and the direction of the vernal equinox, and is called *Greenwich Apparent Sidereal Time* (GAST). Once the precession and nutation have been accounted for precisely, we can relate CIRS to CTRS in terms of parameters of polar motion and the earth's rotation $\{x_p, y_p, \theta\}$.

As discussed in Appendix 4.A, two Cartesian coordinate frames with a common origin can be brought into coincidence by three rotations of the axes of either coordinate frame. Let us see in Figure 4.2 how we can rotate the coordinate frame (x_I, y_I, z_I) into (x_T, y_T, z_T). We rotate first about the z_I-axis by an amount θ, followed by a rotation about the new position of the x_I-axis by an amount $-y_p$, and finally a rotation about the new position of y_I by $-x_p$. The transformation between the CIRS and CTRS based on these three rotations can be written as

$$\mathbf{x}_T = \mathbf{R}_{2I}(-x_p)\mathbf{R}_{1I}(-y_p)\mathbf{R}_{3I}(\theta)\,\mathbf{x}_I \tag{4.1}$$

where $\mathbf{x}_T$ and $\mathbf{x}_I$ are the coordinates of a point represented in the CTRS and CIRS, respectively, and $\mathbf{R}_{1I}$, $\mathbf{R}_{2I}$, and $\mathbf{R}_{3I}$ are the rotation matrices corresponding to rotations about x_I-, y_I-, and z_I-axis, respectively. Note that we denote vectors by lower-case boldface letters, and matrices by upper-case boldface letters, with one exception (in Chapter 7).

We have dealt lightly with the definition of the inertial reference system and the transformation between the inertial and terrestrial reference systems. The reason is that the orbital parameters provided by the Control Segment to specify the position of a satellite take into account the precession, nutation, polar motion, and non-uniform spin rate of the earth implicitly. Apparently, this is done to make the parametric representation of an orbit simpler. We'll discuss this further in Section 4.3.

4.1.2 Geodetic Coordinates, Geoid, and Datums

We noted earlier the limitations of the Cartesian coordinates in everyday use. In mapping the surface of the earth, an alternative is to limit the information to horizontal position only, and express it as angular coordinates—latitude and longitude. For a point above or below the surface of the earth, we could include its height, defined appropriately. The surface of the earth, however, is irregular and changeable. What's needed is a model: a simple, smooth, easy-to-characterize geometrical surface relative to which we can represent the actual surface of the

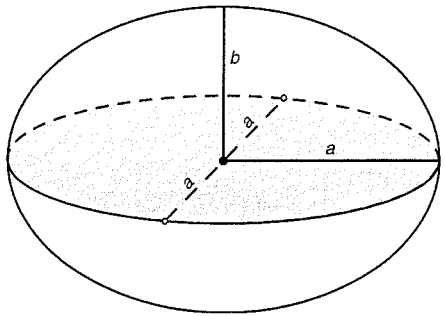


Figure 4.3 Ellipsoid of revolution (oblate ellipsoid).

earth. An early model was a sphere, and the cartographers and navigators expressed the coordinates of a point on this sphere in terms of geocentric latitude and longitude (see Figure 1.1).

Further refinement in modeling the shape of the earth had to wait until Newton, whose study of gravity led him to conclude that the earth was flattened slightly at the poles and bulged somewhat at the equator. It was a controversial theory, but Newton was vindicated in 1735 by the measurements obtained from an expedition to Peru to measure the length of a meridian degree near the equator, and another to Lapland to make similar measurements near the Arctic Circle. These measurements established the meridian to be an ellipse with about 20-km difference between the lengths of the semi-major and semi-minor axes. The figure of the earth is approximated as an ellipsoid of revolution generated by revolving an ellipse about its minor axis (see Figure 4.3). This figure is also referred to as an *oblate ellipsoid.* (The egg-like figure generated by revolving an ellipse about the major axis is called a prolate ellipsoid.)

Ellipsoid and Ellipsoidal Coordinates

For a global reference system it makes sense to define an ellipsoid in conjunction with an ECEF Cartesian coordinate system, with a common origin at the center of mass of the earth, and the axis of revolution of the ellipsoid coincident with the z-axis. Having specified the origin and the orientation of the ellipsoid, there are only two parameters left for full characterization: the lengths of the semi-major and semi-minor axes, denoted as a and b, respectively. The eccentricity is defined as $e^2 = (a^2 - b^2)/a^2$. In geodesy, it is more common to characterize the ellipsoid by specifying the semi-major axis and *flattening*, denoted as f and defined as $f = (a - b)/a$. The flattening and eccentricity are related by $e^2 = 2f - f^2$.

Now we can define the *geodetic coordinates* (also called *geographic* or *ellipsoidal* coordinates) of a point P as follows (see Figure 4.4):

- *geodetic latitude* (ϕ): the angle measured in the meridian plane through the point P between the equatorial (x-y) plane of the ellipsoid and the line perpendicular to the surface of the ellipsoid at P (measured positive north from the equator, negative south),
- *geodetic longitude* (λ): the angle measured in the equatorial plane between the reference meridian and the meridian plane through P (measured positive east from the zero meridian),
- *geodetic height* (h): measured along the normal to the ellipsoid through P.

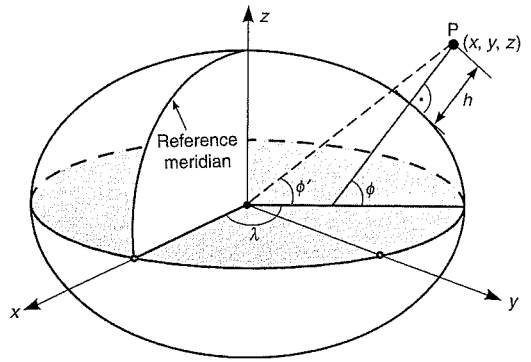


Figure 4.4 Cartesian (x, y, z) and ellipsoidal (ϕ, λ, h) coordinates. Geocentric latitude is denoted as ϕ'. We denote a right angle as ⦜.

Figure 4.4 shows the Cartesian and ellipsoidal coordinates of P. For simplicity, we have dropped the subscript T used earlier to denote the axes of the CTRS. Connecting all points on the ellipsoid with constant ϕ generates a closed curve called a *parallel.* All parallels are circles. Similarly, a closed curve defined by connecting the points with constant λ is called a *meridian.* All meridians are ellipses.

The ellipsoidal coordinates have their limitations as well. Take the definition of height. The geodetic or ellipsoidal height is defined relative to an abstraction (the reference ellipsoid) and has no physical meaning. As discussed below, the problem is solved by defining a hypothetical surface called the *geoid*, which essentially represents an extension of the idealized mean sea level over (actually, mostly under) the land surface of the earth. Another problem with the ellipsoidal coordinates is the lack of a constant distance relationship: 1° of longitude is about 110 km (60 nautical miles) at the equator, and 80 km at 45° latitude. When dealing with a small part of the surface of the earth, say, a region stretching tens of kilometers, as a land navigator or an engineer planning a construction project does, it would be easier to work with a properly defined local rectangular coordinate system.

A geocentric ellipsoid specifies a *global datum* or a reference surface to be used in defining 3-D coordinates of a point anywhere. Parameters a and f have been refined over the years. The International Ellipsoid (1924) was defined as: a = 6,378,388 m, f = 1/297. The best available values today are only slightly different: a = 6,378,137 m, and $f \approx$ 1/298.257 ($e \approx 0.082$). Given the coordinates (x, y, z) of a point in the ECEF Cartesian coordinate frame, we can obtain the ellipsoidal coordinates (ϕ, λ, h), and vice versa, as described in Appendix 4.A.

We have now touched upon all the main ideas required to describe WGS 84. A student in no hurry to grapple with the definition of height, or to understand why the latitude and longitude of a point outside the United States obtained from GPS may not match those shown on a local map, can now proceed to Section 4.1.3.

Definition of Height*

Historically, the height of a point relative to another has been measured as the separation between their horizontal planes defined by leveling instruments. When it came to defining the absolute height, a logical choice for the reference was the idealized mean sea level (MSL). By definition, a level surface is perpendicular to the plumb line or the gravity vector. This connection between gravity and measurement of heights has made the study of the earth's gravity field and its variations important to the geodesists. Having defined an accurate ECEF coordinate frame and an ellipsoid, in this section we examine the definition of a surface which serves as the global zero reference for measurement of height.

If the earth were spherical in shape and uniform in density, its gravitational field would be easy to describe. The gravitational field would be spherically symmetric and the gravitational force at any point would depend only upon its distance from the center of the earth. Actually, the earth is neither uniform in composition nor spherical in shape, and its gravitational field is a complicated function of latitude, longitude, and distance. Gravitational force is a vector, with magnitude and direction. In order to characterize the gravity field, then, we could specify the gravitational force vector for all points of interest. It's simpler to specify the *gravitational potential*, a scalar quantity, defined so that its spatial gradient (i.e., a vector specifying the rate of change along three orthogonal directions) equals the gravity vector at that point.

The locus of all points with the same value of the gravity potential is called an *equipotential surface*. An equipotential surface is a closed, smooth surface surrounding the earth to which the direction of gravity is perpendicular at each point. It's a level surface, by definition. This surface, however, is not 'regular' enough to be represented by a simple mathematical function. Instead, it is represented numerically with values specified over a grid. The equipotential surface best fitting the average sea level globally is called the *geoid*. Roughly speaking, the idea is to extend this surface representing the idealized mean sea level over the land portion of the globe. The geoid is a physically defined level surface which reflects the variations in the earth's gravity field caused by geological formations and topographic relief. The dimensions of the ellipsoid used to describe the figure of the earth are actually defined so as to fit the geoid best in a least-squares sense.

How to map the geoid? One way is to map it relative to the reference ellipsoid: At each point, define the *geoidal height* in terms of geoid-ellipsoid separation measured along a line orthogonal to the ellipsoid. By convention, the geoidal height is measured from the ellipsoid to the geoid and is denoted by N. Defining the heights measured from the ellipsoidal and geoidal surfaces, as h and H, respectively,

$$h = H + N.$$

The above relation is an approximation, as shown in Figure 4.5, but it is accurate enough for most practical purposes. The angle between the plumb line, which is perpendicular to the geoid (called the *vertical*) and the perpendicular to the ellipsoid (called the *normal*) is defined as the *deflection of the vertical*.

Height measured relative to the geoid is called *orthometric height*, also known as elevation or height above the mean sea level. The heights shown on topographic maps, for example, are orthometric heights. Calculation of the orthometric height of a point from GPS measurements is a two-step process:

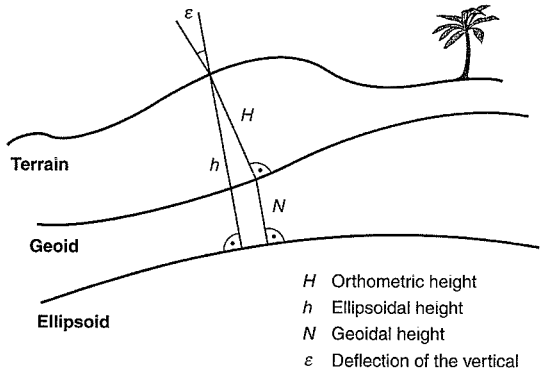


Figure 4.5 Geoid, geoidal height, and deflection of the vertical.

(i) determine the ellipsoidal coordinates (ϕ, λ, h) from GPS measurements,
(ii) determine the geoidal height from a data base, and subtract it from the ellipsoidal height h.

We, therefore, need an accurate mapping of the geoidal heights on a global level. This is a large undertaking. Analysis of the perturbations in satellite orbits gives large-scale features of the gravity field. Features with finer resolution, called gravity anomalies, are analyzed primarily by terrestrial measurements. The most accurate geoid model available at present is the WGS 84 geoid, described below.

Regional Datums and Map Projections*

Before the advent of satellite navigation and development of global datums, attention was restricted to mapping a country, a region, or a continent by terrestrial techniques. The process required two datums, or reference surfaces: a *horizontal datum*, an ellipsoid, relative to which we could define the latitude and longitude of a point; and a *vertical datum*, or geoid, defined as the surface of zero height. The regional datums proliferated as each country or group of countries defined an ellipsoid.

For mapping a region, it was convenient to define an ellipsoid which fitted the geoid best over the region of interest. There was no reason to constrain the ellipsoid to be geocentric, or to constrain the orientation of its axes (see Figure 4.6). Having defined an ellipsoid, the points on the surface of the earth can be projected to the ellipsoid and assigned latitude and longitude values. The next step is to represent the geographical features of the earth's surface on a map as rectangular coordinates.

Mapping a curved surface on a flat surface in terms of rectangular grid coordinates is a classical problem [Brown (1977)]. Hundreds of map projections have been proposed over the years. None is completely satisfactory in the sense that none is free of distortion. The choice is among projections that preserve areas, or distances, or angles. For example, an area-preserving projection preserves the shapes of small areas, and an angle-preserving transformation main-

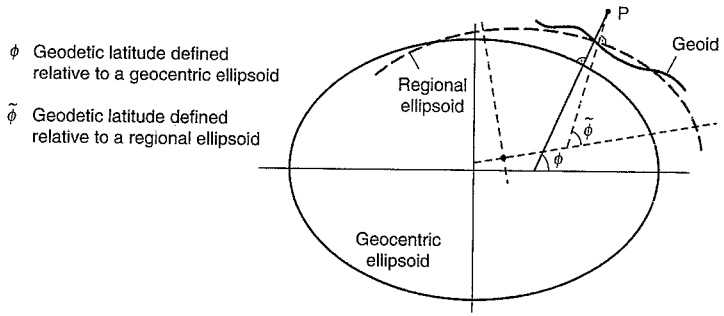


Figure 4.6 Global and regional datums.

tains the angle between two lines. The mapping authorities decide on the basis of the location on the ellipsoid which distortion to tolerate and which to avoid. Over the last two hundred years, hundreds of datums and dozens of projection methods have been used widely [Ashkenazi (1986), Featherstone and Langley (1997)].

The point to remember is that the latitude and longitude of a point on the earth are defined relative to a geodetic datum, and represented on a map by using a projection method. The latitude and longitude of a point, therefore, are not unique. In particular, your position estimate obtained from GPS outside the United States may not match the coordinates shown on a local map. You may be off by hundreds of meters (more than a kilometer in Japan). You would not have this problem within the United States as the *North American Datum of 1983* (NAD83), the current legal datum of the United States, uses coordinates that are virtually identical to those obtained from GPS for the purposes of most users. It is a different realization of the conventional terrestrial reference system. The definition and maintenance of NAD83 is the responsibility of the National Geodetic Survey, U.S. Department of Commerce. The U.S. Geological Survey (USGS) maps are based on NAD83.

The coordinates expressed in one geodetic datum can be transformed into another with a seven-parameter transformation [NIMA (1997), Featherstone and Langley (1997)]. The U.S. military receivers have over a hundred such transformations built into their software. The receivers intended for civil use also generally offer options. When in Australia, be sure to get your position from GPS transformed to the Australian Geodetic Datum in order to be compatible with the local maps. When in Japan, go with the Tokyo Datum.

4.1.3 World Geodetic System 1984 (WGS 84)

WGS 84 is a realization of the conventional terrestrial reference system (CTRS) developed by the Defense Mapping Agency (DMA), which became in 1996 a part of the National Imagery and Mapping Agency (NIMA), of the U.S. Department of Defense, and reorganized in 2004 as the National Geospatial-Intelligence Agency (NGA). WGS 84 is the 'official' geodetic system for all mapping, charting, navigation, and geodetic products to be used throughout the DoD. The development of this global datum was essential to the development of GPS. The widespread use of GPS is turning WGS 84 from a global datum into an international datum, a *de facto* world standard. Simplifying the mapping of the earth by unifying the diverse datums in

itself is an important contribution of GPS. The principal reference for this section is the NIMA Technical Report on WGS 84 [NIMA (1997)].

WGS 84 is actually more than a realization of the CTRS. It comprises a coherent set of global models and definitions:

- an ECEF Cartesian coordinate frame,
- an ellipsoid of revolution as a geometric model of the shape of the earth,
- a characterization of the earth's gravity field and geoid,
- a consistent set of fundamental constants.

We'll refer to the position of a point in WGS 84 as WGS 84 XYZ coordinates, or WGS 84 LLA (for latitude, longitude, and altitude) in accordance with the above models. When referring to the ECEF coordinate frame, or simply ECEF, we mean WGS 84 ECEF Cartesian coordinate frame.

Realization of the CTRS is a bootstrapping process of definition and refinement. WGS 84 is a refinement of the earlier versions: WGS 60, WGS 66, and WGS 72. The WGS 84 ECEF coordinate frame is based on the coordinates of a set of globally distributed points (satellite tracking stations) whose positions had been defined earlier using Doppler measurements from Transit. The coordinate frame was realized by fitting a model to these coordinates by least-squares technique. The residuals in this exercise give an indication of the consistency of the adopted coordinates. The definition of the origin and the coordinate axes was refined by GPS measurements and space observations using VLBI and SLR techniques [Seeber (2003)].

The positions of the GPS satellites computed from the parameter values specified in their navigation messages are expressed in WGS 84. The user positions, therefore, are obtained as WGS 84 coordinates. The satellite orbits are computed on the basis of code and carrier phase measurements at the monitor stations operated by the GPS Control Segment [Section 2.2.2]. Precise estimation of the monitor station coordinates in WGS 84 is the key step in GPS' implementation of WGS 84. In the 1980s, the accuracy of these coordinates was at 1–2 m level. The GPS implementation of WGS 84 has been refined three times since 1980, most recently in 2002 [Merrigan *et al.* (2002)], by obtaining more accurate coordinates of the monitor stations (now at centimeter level). The definition of WGS 84 itself has undergone refinements resulting in adjustment to the values of the fundamental constants (Table 4.1).

Table 4.1 WGS 84 fundamental parameters (revised in 1997)

Parameter	*Value*
Ellipsoid	
Semi-major axis (a)	6378137.0 m
Reciprocal flattening ($1/f$)	298.257223563
Earth's angular velocity (ω_E)	$7292115.0 \times 10^{-11}$ rad/sec
Earth's gravitational constant (GM)	3986004.418×10^8 m^3/s^2
Speed of light in a vacuum (c)	2.99792458×10^8 m/s

Plate tectonic motion causes stations on the earth's crust to move horizontally. Such movement can be 5–10 cm per year, or even larger, and coordinates of 'fixed' points can change over time. Therefore, for precise positioning, a time tag has to be associated with the coordinates of a station. An alternative is to estimate a velocity in addition to the position coordinates. It is basically a never-ending process of modeling physical processes, redefining positions, and readjusting the reference frames.

As noted previously, the scientific community maintains a reference frame of its own. The International Terrestrial Reference Frame (ITRF) was developed and is maintained by the International Earth Rotation Service (IERS). The current ITRF is defined on the basis of coordinates of about two hundred terrestrial stations determined on the basis of VLBI, SLR, lunar laser ranging (LLR), and GPS. The accuracy of the ITRF stations is estimated to be at centimeter level. Velocities of these stations due to geophysical forces are estimated, too. The IERS updates the coordinates of their stations, and defines a new realization annually. IERS maintains the ITRF with an annual series: ITRFyy, where 'yy' specifies the year. ITRF and WGS 84 are getting close. The current realization of WGS 84 and the current ITRF (ITRF2000) are essentially identical for the purposes of most users.

We discuss next the WGS 84 earth gravitational model. As noted previously, the gravity field is specified in terms of gravitational potential, a scalar quantity, defined so that its spatial gradient equals the gravity vector at that point.

The gravitational potential of a spherical, uniform earth would be GM/r, where GM is the earth's gravitational constant [Section 4.3.1]. The gravitational potential of the real, non-spherical, non-uniform earth is far more complex and is expressed in the form of a series expansion. The terms of the series expansion are called spherical harmonics. Each term is identified by a pair of indices called degree (n) and order (m). The gravitational potential function at a point at distance r from the earth's center with longitude λ and geocentric latitude ϕ' is modeled as [NIMA (1997)]

$$V(r,\phi',\lambda) = \frac{GM}{r}\left[1 + \sum_{n=2}^{n_{\max}} \sum_{m=0}^{n} \left(\frac{a}{r}\right)^n \overline{P}_{nm}(\sin\phi')\left(\overline{C}_{nm}\cos m\lambda + \overline{S}_{nm}\sin m\lambda\right)\right] \tag{4.2}$$

where a is the semi-major axis length of the WGS 84 ellipsoid, $\overline{C}_{nm}$ and $\overline{S}_{nm}$ are the normalized gravitational coefficients, and $\overline{P}_{nm}$ are Legendre functions and polynomials. The gravity field is completely described by specifying the coefficients ($\overline{C}_{nm}$ and $\overline{S}_{nm}$). The units of the gravity potential function are m^2/s^2. The definition of the geoid, the reference surface for measurement of height, is based on (4.2). Our purpose here is simply to offer a general idea of how the earth's gravity field is specified and we wouldn't stop to explain the various terms of (4.2). The interested reader is referred to Vaníček and Krakiwsky (1986) for a comprehensive discussion.

WGS 84 *Earth Gravitational Model 96* (EGM96), developed jointly by NIMA and NASA Goddard Space Flight Center (GSFC), consists of a table of estimates of the coefficients $\overline{C}_{nm}$ and $\overline{S}_{nm}$ to degree and order 360, altogether 130,676 coefficients determined empirically on the basis of satellite tracking and terrestrial gravity observations. High-accuracy orbit calculations typically account for EGM96 through degree and order 70. The coefficients of the terms of higher degree and order decay quickly.

The WGS 84 geoid based on EGM96 can be depicted as a contour chart showing the

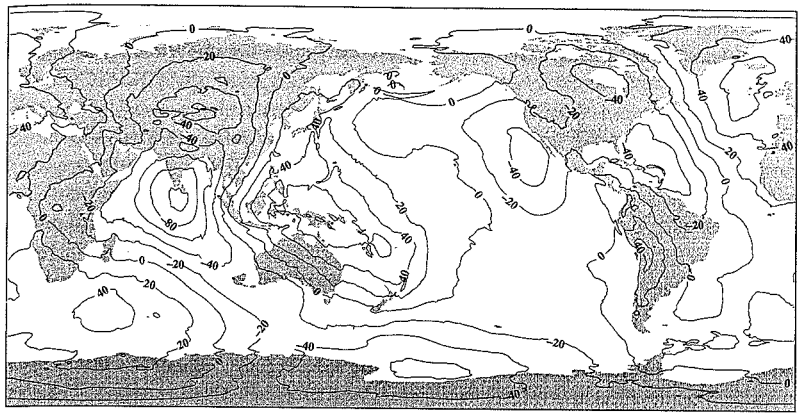


Figure 4.7 Contour plot of the geoidal height (EGM96). (Courtesy of Steve Kenyon and Don Schmidt, NIMA)

deviation of the geoid from the ellipsoid. A contour chart created from the geoidal heights calculated on 15′ × 15′ grid worldwide from WGS 84 parameters and EGM96 is available on the Internet (http://cddis.gsfc.nasa.gov/926/egm96/egm96.html). A simplified view is shown in Figure 4.7. The geoidal height globally ranges from –104 m at the southern tip of India to +75 m in New Guinea. The average geoidal height worldwide is approximately zero. Throughout the conterminous U.S. (CONUS), the geoidal height is negative (i.e., the geoid is below the ellipsoid). Therefore, in a coastal area, the ellipsoidal height obtained from GPS would be negative. The rms WGS 84 geoidal height worldwide is about 30 m. This is a global characterization of how well the ellipsoid fits the geoid.

4.2 Time References and GPS Time

The second, the unit of time, is one of seven independent standards, or basic units of measurements. The other six are: the meter, the kilogram, the ampere, the kelvin, the mole, and the candela. All other measurements (e.g., force, electric charge, and magnetic flux) can be expressed in terms of these seven basic units. Unlike other measured quantities, however, time is always changing. Time flows; sometimes it even flies. A few hundred years ago, we would have been happy to reckon time well enough so as not to miss a planting season. And today we measure time routinely with a precision of nanoseconds (10^{-9} s) and picoseconds (10^{-12} s).

Generation of precisely synchronized signals aboard spacecraft and measurement of their transmission time is at the heart of GPS. We must, therefore, understand the definition of a precise time scale, and measurement of time in accordance with it. We determine pseudorange to a satellite by multiplying the apparent transit time of a signal by the speed of light. Synchronization error of 1 μs in a satellite clock will introduce an error of 300 m in the pseudorange, with a corresponding error in the position estimate. If we require meter-level position estimates, the clock synchronization among the satellites must be maintained within a few nanoseconds.

We must distinguish between the concepts of an *instant* of time (an 'epoch') and an *interval* or duration of time between two events. Actually, there are no physical phenomena to suggest a beginning of time for a unique determination of an epoch. Such a definition has to be adopted by consensus. An epoch may be represented, for example, in terms of year, month, week, day, and time of day (hours, minutes, seconds, and fraction of a second). On the other hand, having reached such an agreement, the definition of a time interval appears easy: We would like to measure it with a scale that is uniform. As we'll see, that's not necessarily so. In fact, the world has settled on a civil time scale that's uniform except for occasional discontinuities.

The basic idea of timekeeping is simple: Observe a periodic process and count the periods (or cycles). Examples of periodic processes are: the apparent diurnal motion of the sun in the sky; revolutions of the earth around the sun; swings of a pendulum; and oscillations of a quartz crystal. A clock, then, is essentially a generator of periodic events (frequency source or resonator) and a counting mechanism (counter or integrator) for the events. The accuracy of such a clock to measure time would depend upon (i) error in the initial frequency setting (accuracy), and (ii) ability to maintain the rate of the periodic process (frequency stability).

The earth's diurnal rotation has been the basis for timekeeping since the beginning. In fact, the rotation rate of the earth has served as the source for definition of time interval of one second until recently. The earth's diurnal rotation has given rise to two time scales: *solar time* and *sidereal time*, which we discuss below. The need for greater precision and the discovery that the earth's rotation rate is not constant led to redefinition of the second in terms of the resonance frequency of the cesium atom. We discuss these definitions below, followed by an optional subsection on analysis of stability of periodic processes. Finally, we define the time scale of GPS, the main idea of this section.

4.2.1 Time Scales: Astronomical and Atomic

Solar and Sidereal Times

A complete revolution of the earth with respect to the sun (i.e., time between two successive transits of the sun across a local meridian plane) defines an *apparent solar day* (apparent here means *actual*). It has been known since the Greeks that no two apparent solar days are precisely the same length. There are two reasons for this. First, the earth turns slightly more than a complete rotation relative to the 'fixed' stars during an apparent solar day as it travels about 1/365th of the way in its orbit around the sun. The orbit is slightly elliptical, and the earth's orbital speed is not uniform [Kepler's second law, Section 4.3]. So, on some days the earth has to turn a little less, and on others a little more. Secondly, the earth's axis of rotation is not perpendicular to the plane of the orbit. (The angle between the plane of the earth's orbit, called the ecliptic, and the equatorial plane is about 23.5°.)

An observer at the Greenwich meridian equipped with a 'perfect' clock to keep time would find the sun at 12:00 noon wandering ±4° in the east-west direction in the course of a year. That is an error of ±16 minutes in a clock based on apparent solar time. Recognition of this fact led to definition of *mean solar time* which corresponds to the earth in a hypothetical circular orbit around the sun with the same period, and its axis of rotation perpendicular to the orbital plane. The mean solar time at the Greenwich meridian came to be called the *Greenwich Mean Time* (GMT) or Zulu time. Until 1960, a second was defined as 1/86400 of the mean solar day.

Another time scale based on the rotation of the earth is *sidereal time* (pronounced sigh-deer-ee-al). A *sidereal day*, or twenty-four hours of sidereal time, is defined as the time the earth takes to rotate once on its axis relative to the stars much farther away than the sun. The earth revolves once around the sun in a year (365.25 mean solar days), and there is one extra rotation of the earth to account for. Mean solar day is approximately four minutes longer than a sidereal day. The orbital period of a GPS satellite is one-half sidereal day. After two revolutions around the earth, each satellite rises at the same spot on the horizon relative to each observer, but four minutes earlier than the day before by his watch.

Actually, an apparent sidereal day is not of constant length either, leading the astronomers to define *mean sidereal time*. We'll not discuss it further. The basic relationship between mean solar time and mean sidereal time is stated below.

1 day of mean solar time = 24h = 86,400 mean solar seconds
$\approx 1 + 1/365.25 \approx 1.002737$ days of mean sidereal time

1 mean sidereal day = 23h 56m 4.0954s = 86164.09954 seconds of mean solar time

Sidereal time, like solar time, depends upon the observer's longitude and is defined as the hour angle of the vernal equinox (i.e., angle between the observer's meridian and the vernal equinox). Sidereal time at the reference meridian is specified *as Greenwich Apparent Sidereal Time* (GAST), as shown in Figure 4.2.

It would be impractical to have time varying continuously with longitude. Therefore, by international agreement (International Meridian Conference, 1884), the earth is divided into twenty-four standard time zones bounded by meridians at 15° intervals and centered on longitudes that are integral multiples of 15°, with some exceptions. Within each time zone, the civil time is the same: the mean solar time at the central meridian. The meridian of 180° longitude is called the *international date line*. The times in the different time zones are defined to differ from GMT (mostly) by integral number of hours.

Early in the twentieth century it became clear that mean solar time is a non-uniform time scale. One reason for the irregularity is the polar motion, discussed earlier [Section 4.1.1]. The movement of the pole changes by a minute (as in very small) amount the latitude and longitude of every point on the earth as determined by astronomical observations. The local time is tied to the longitude and changes accordingly. The second (as in number two) source of non-uniformity of the mean solar time is the irregularity in the spin rate of the earth.

By the mid-1930s, astronomical observations had made it clear that the earth's rate of rotation was not constant. In fact, the length of two consecutive days could vary by several milliseconds as a consequence of this non-uniformity of the rotation rate. There is a long-term trend of the earth slowing down. Superimposed on this secular change are variations that appear seasonal, and others that appear random. The secular variations are attributed mainly to tidal friction. The seasonal changes appear related to periodic physical processes on or within the earth which cause redistribution of its mass and moment of inertia. The random changes may be related to processes such as interaction of the wind with the earth's topographic features. The rotational speed of the earth remains unpredictable over the long run, and the phenomena affecting it are not fully understood. At present the earth appears to slow down by

about one second a year.

The non-uniformity of the spin rate notwithstanding, the earth's rotation rate continued to be the source of civil time until 1960 with astronomical observatories correcting mechanical clocks to conform to the observed mean solar time. *Universal time* (UT) is the general designation given to time scales based on the rotation rate of the earth.

UT0: Mean solar time at the prime meridian obtained from astronomical observations. UT0 (UT zero) is the time actually measured at an observatory, and is subject to the effects of both earth's irregular spin rate and polar motion. The spin rate affects the measurements equally, but the effect of the polar motion on UT0 depends upon the location of the observatory.

UT1: UT0 corrected for the observed effects of polar motion (up to 0.06 s). Such corrections are determined from measurements of UT0 from observatories around the world. UT1 is based on the true orientation of the earth in space. To determine latitude and longitude using a sextant and star observations, you need to know UT1. An error of 1 s in time can result in an error of about 500 m in position. UT1 is the time scale of astronomers and celestial navigators. A uniform time scale required by engineers it is not.

Increasing demands for precision required redefinition of the second. In 1960, a new time standard was defined in terms of the earth's orbit around the sun, which is not influenced by the unpredictable polar motion or the variation in the rotation rate. The *ephemeris second* was defined as the fraction 1/31556925.9747 of the year 1900. The time scale based on ephemeris time turned out to be a cumbersome time scale and was short-lived. It was replaced by much more precise and convenient atomic time.

Atomic Time

The current definition of the second, adopted in 1967 by international agreement, is based on the resonance frequency of the cesium atom. The fundamental time interval unit in the International System of Units (SI) is the *SI second*. (SI is the newest name for what began two hundred years ago as the Metric System, a recommended practical system of units of measurements.) The SI second is defined as "the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom." Why 9,192,631,770? Because it corresponded exactly with the previous definition of the second, the ephemeris second. The continuous time scale based on the above definition of atomic second is called *International Atomic Time* (TAI). (SI day = 86,400 SI seconds; Julian century = 36,525 SI days.)

TAI is a precise and uniform time scale which is not tied to the earth's rotation on its axis or its revolution around the sun. Cutting ties with the earth time (UT1), however, is unsettling. It is estimated that in about four thousand years the earth could lose about twelve hours, or half a day, and the sun would be high in the sky while local time based on a TAI clock indicates midnight [Allan, Ashby, and Hodge (1997)]. The result has been a compromise time scale called the *Coordinated Universal Time* (UTC). The definition of the UTC second is the same as that for atomic time, and is based on the cesium atom. UTC is now the scale for public time throughout the world. It is the new GMT.

UTC was set to agree with UT1 at 00 hours on January 1, 1958. At first, the two time scales were kept close by introducing 0.1-second steps in UTC, as needed. Since 1972, changes in the earth's spin rate have been accommodated by introducing leap seconds in UTC. UTC and TAI, therefore, differ by an integer number of seconds. In principle, a leap second can be either positive or negative, and is introduced to keep UT1 within 0.9 seconds of UTC. In practice, the leap seconds so far have all been positive, and a minute containing a leap second had 61 SI seconds. By international agreement a leap second can be introduced at the end of any month, but June and December are preferred. (Leap seconds are cumbersome and messy. There is now a movement to abolish them.)

The sequence of UTC second markers on 31 December 2005, when a leap second was added, proceeded as follows.

31 December 2005	23h 59m 59s
31 December 2005	23h 59m 60s
01 January 2006	00h 00m 00s

The difference between TAI and UTC, has now increased to 33 seconds: TAI – UTC = 33 s. (The previous leap second was added on 31 December 1998.)

The time scale and the Earth Orientation Parameters (x_p, y_p, θ), introduced in Section 4.1.1, are maintained through international cooperation. UTC is generated at the Bureau International des Poids et Mesures (BIPM), located near Paris [http://www.bipm.fr]. The leap second steps are determined by the International Earth Rotation Service (IERS), located at the Paris Observatory [http://www.iers.org].

The atomic standards, though extraordinarily accurate and precise, are not perfect. UTC is generated after the fact on the basis of the times kept by about 250 cesium clocks and hydrogen masers located at about 65 different laboratories located around the world. In the United States, UTC estimates are generated by the National Institute of Standards and Technology (NIST), Boulder, Colorado, and the United States Naval Observatory (USNO), Washington, D.C. Both institutions are charged with supplying time and frequency to the U.S. government and BIPM. Their UTC estimates are referred to as UTC(USNO) and UTC(NIST). Other countries have similar institutional arrangements to generate their national time standards, which serve as the basis for real-time estimates of UTC. It takes BIPM about a month to collect and process the data to generate TAI and UTC. A monthly bulletin from BIPM reports the time difference that existed between each of the contributing clocks and UTC.

Definition of Time Epoch*

In civil life, we usually specify a time epoch in terms of year, month, day, hour, minute, and second. In activities requiring long periods of observations or operations, as in case of GPS, the calculation of time differences is simpler if the time epoch is defined in terms of a day number, counted from a reference, and decimal fraction of a day. Such continuous count of the days also avoids the confusion associated with the early development of calendars which created jumps in dates to bring the calendars in line with the seasons. A fundamental concept in reckoning time is that of Julian Date (JD), which is used extensively by astronomers, celestial navigators, and geophysicists.

JD is defined by a continuous count of days and fraction of a day, starting from 12h UT on

1 January 4713 BC. Unlike the civil day, which begins at midnight, the Julian date is measured from noon to noon, making it easier for the astronomers to complete their nightly observations without changing the date. As an aside, note that Julian Day count has nothing to do with the Julian calendar, which was introduced by Julius Caesar (44 BC) and remained in force until the Gregorian calendar was introduced by Pope Gregory XIII in 1582. The JD count for a given Gregorian date can be found in various almanacs. For two epochs important to GPS:

JD of the Standard epoch of GPS:
1980 January 6 0h = JD 2,444,244.5

Standard Epoch of UT:
J2000.0 = 2000 January 1 12h UT
= JD 2,451,545.0

The Modified Julian Date (MJD), introduced in the 1950s, is defined to shift the origin so as not to have to deal with such large numbers, and to begin a day at midnight in conformity with the civil practice. The origin corresponds to midnight November 17, 1858, which corresponds to JD 2,400,000.5.

$$\text{MJD} = \text{JD} - 2{,}400{,}000.5$$

4.2.2 Stability Measures of Frequency Sources*

The performance of a frequency source is defined in terms of its accuracy and stability. Accuracy is a measure of how well a frequency source can be 'tuned' to a specified (or, advertised) frequency. The frequency departure from the specified value is usually given in terms of normalized or *relative frequency deviation*, also known as fractional frequency, defined below. Let f_0 and f be the specified and actual frequencies of a frequency source. The frequency deviation is $\Delta f = f - f_0$. Then the relative frequency deviation is defined as

$$F = \frac{f - f_0}{f_0} = \frac{\Delta f}{f_0} \tag{4.3}$$

An oscillator with a specified frequency of 1 MHz but running at 999,999 Hz has an error of one part in a million. This may also be written as accuracy of 10^{-6}, or an error of 0.0001%. It is left as an exercise to show that time kept on the basis of a frequency source with a normalized error of $\Delta f/f$ would result in an error of Δt over a period T such that

$$\frac{\Delta f}{f} = -\frac{\Delta t}{T} \tag{4.4}$$

Frequency stability is a measure of the ability of a frequency source to operate at the specified frequency over a period of interest. The stability of a clock is commonly specified in terms of relative frequency deviation, as defined above. Consider, for example, the challenge faced by John Harrison. In order to claim the full prize under the Longitude Act (1714) of the British Parliament, longitude had to be found and kept within one-half degree over a six-week voyage. This translates into cumulative timekeeping error of no more than two minutes over six weeks, or three seconds per day. The corresponding relative frequency deviation permissible was 3/86400 or 3.5×10^{-5} over a day. It was an extraordinary challenge two hundred years ago to achieve such stability with a mechanical device to be carried aboard a ship. The best frequency standards of today are approaching frequency stability of 1×10^{-15} over a day.

An ideal clock would be set perfectly to the desired frequency, and this frequency would remain stable for the life of the clock. In practice, no clock can be set perfectly, and there are systematic and random variations tied to the aging of the resonator and environmental effects such as vibrations and changes in temperature, pressure, and humidity. The frequency may be modeled as

$$f(t) = f_0 + \Delta f + (t - t_0)\dot{f} + \tilde{f}(t)$$

where f_0 is the nominal frequency; Δf is the frequency bias or offset; $\dot{f}$ is the frequency drift; $\tilde{f}$ is the random frequency error; and t_0 is the reference epoch. The time error in this clock at time t_1 (disregarding the sign) is

$$\Delta t(t_1) = \Delta t(t_0) + \frac{\Delta f}{f_0}(t_1 - t_0) + \frac{\dot{f}}{2f_0}(t_1 - t_0)^2 + \int_{t_0}^{t_1} \frac{\tilde{f}(t)}{f_0} dt \tag{4.5}$$

The first three terms in (4.5) represent systematic effects, which can be estimated by comparing the clock with a reference clock. The last term represents random frequency fluctuations, the size of which depends upon the environmental effects. The random frequency fluctuations are characterized by their variance and auto-correlation function. A widely used characterization of the timekeeping ability of a clock with respect to random processes is called *Allan variance*, which we describe next.

Suppose we have a time series of measurements y_i, $i = 1, 2, \ldots, N$, of relative frequency deviation of an oscillator due to random fluctuations measured over N successive time intervals of length τ. Then the Allan variance can be estimated as

$$\sigma_y^2(\tau) = \frac{1}{N}\sum_{i=i}^{N} \frac{1}{2}\left(y_{i+1} - y_i\right)^2$$

Why invent a new definition of variance for this time series? The answer is that the classical variance doesn't converge for some commonly observed oscillator noise processes. For example, for a random walk-type noise, commonly observed in oscillators (typically for $\tau > 1$ second for a crystal oscillator, $\tau > 10^3$ seconds for a rubidium standard, and, $\tau > 10^5$ seconds for a cesium standard) the classical variance estimate would continue to increase as N increases. The main thing to note is that the nature of the random process associated with the measurements of relative frequency deviations changes with the size of the averaging interval τ, and the Allan variance as defined above converges for all of them.

The square-root of Allan variance $\sigma_y(\tau)$ is called *Allan deviation*, or two-sample deviation. The rms time error of a clock after an interval τ is approximated as $\tau \cdot \sigma_y(\tau)$. In other words, if a clock were synchronized with a true time scale at the beginning of an interval of length τ, at the end of the interval it would have deviated (on average) by $\tau \cdot \sigma_y(\tau)$. That's really the main result we want to point out before proceeding to a survey of timekeeping abilities of the various kinds of oscillators and clocks. We'll be concerned with the short-term performance of oscillators when we discuss the role of noise in phase lock loops in Chapter 12.

4.2.3 Oscillators and Their Stability*

Most GPS receivers use quartz crystal oscillators, which were invented early in the twentieth century and were vital to the development of radio, television, and radar. The workings of a quartz oscillator are tied to the piezo-electric effect: An external voltage applied to the opposite faces of a quartz crystal cut in a prescribed way causes it to expand and contract. Conversely, mechanical vibrations of the crystal produce oscillatory voltage between metal electrodes applied to its surfaces. The size and cut of the crystal determine its resonant frequency, which becomes the source of time. Chances are that the quartz oscillator in your wristwatch has a resonant frequency of 32,768 Hz (2^{15} Hz). The digital circuits generate a pulse per second by counting 2^{15} oscillations.

The quartz crystal oscillators (XO) have short-term (≈ one second) stability of approximately 10^{-9}–10^{-11}, and long-term (≈ one day) stability two to three orders of magnitude worse. The crystal's resonance frequency varies with temperature and frequency stability requires that the temperature be controlled or compensated for. A temperature-compensated crystal oscillator (TCXO) uses the output of a temperature sensor (a thermistor) to generate correction voltage to compensate for the temperature change. In an oven-controlled crystal oscillator (OCXO), the crystal unit is maintained at a constant temperature in an oven. The OCXOs can provide three orders of magnitude improvement in stability, but require more power and are larger and more expensive.

Cesium and rubidium clocks can have short-term stability as good as a quartz oscillator, and long-term stability of 10^{-12}–10^{-13}. While the stability of a rubidium clock degrades over a longer term, a cesium clock can have stability of several parts in 10^{14} over tens of days. The most stable of the atomic clocks at present is the hydrogen maser, which can have stability on

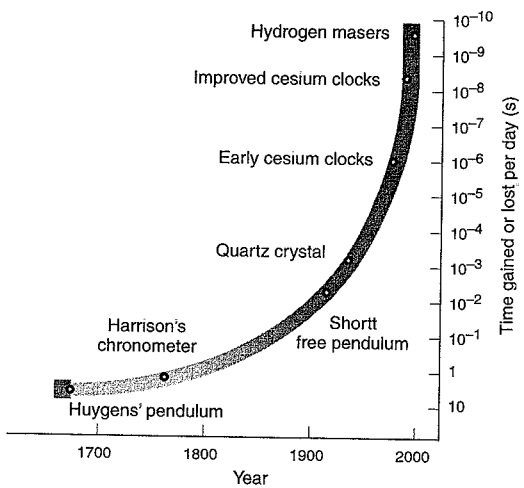


Figure 4.8 Evolution of timekeeping from Huygens' pendulum to atomic standard [adapted from Allan, Ashby, and Hodge (1998)].

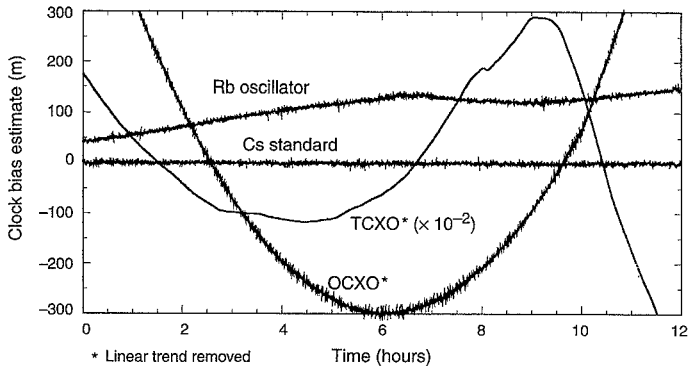


Figure 4.9 Stability of different kinds of clocks varies widely. Shown here are the changes in clock bias of four receiver clocks estimated from laboratory measurements relative to GPS Time.

the order of several parts in 10^{15}. Hydrogen masers are being considered for on-board clocks for Galileo and GPS III.

Today's portable atomic clocks are about the size of a cigarette pack and consume several watts of power. A promising new development with the potential to revolutionize the field of precise timing in portable applications such as GPS is chip-scale atomic clocks (CSAC). Such clocks would be manufactured using microelectronic manufacturing techniques, and would be about 1 cm^3 in size, consume 30 milliwatts of power, and offer stability of 10^{-11} over an hour or longer. Such atomic clocks exist in 2005 as 'physics packages,' demonstrating that the components can be assembled into a compact structure producing an output frequency tied to highly stable atomic transitions [http://www.darpa.mil/mto/csac].

Figure 4.8 shows a simple, idealized view of the increasing accuracy of clocks over the past three hundred years. As noted previously, of all physical quantities, time is now measured with the highest accuracy and precision. In fact, since 1983, the meter is defined in terms of the atomic second as the distance traveled by light in a vacuum during 1/299792458 of a second.

As an illustration of typical performance of clocks used with GPS receivers, Figure 4.9 shows the behavior of a cesium clock, a rubidium clock, an oven-controlled crystal oscillator (OCXO), and a temperature-compensated crystal oscillator (TCXO) as observed in a laboratory. Our purpose is simply to call attention to certain qualitative features. The clock bias estimates were computed from GPS pseudorange measurements with differential corrections obtained from a local reference station. The methodology for obtaining clock bias estimates is discussed in Chapter 6.

Figure 4.9 shows three-second samples of clock bias estimates in meters for each clock relative to GPS Time. The cesium clock (cost ≈ $20k) offers excellent stability: No frequency offset or drift is apparent. The rubidium clock ($2–4k) shows signs of slight frequency offset, and slight drift. The OCXO ($1k) had accumulated a small frequency offset (about one part in 10^8) so that during the test the clock bias changed by the equivalent of about 2 m/s on average. We have taken out this effect by fitting a straight line to the data and plotting the residuals. The parabolic shape implies a constant frequency drift rate. The TCXO ($100 or less) had accu-

mulated a frequency offset of about one part in 10^6, and its bias changed by the equivalent of about 200 m/s on average. Again, this trend was taken out by fitting a straight line to the data and plotting the residuals. These residuals, however, are two orders of magnitude larger than those for the OCXO, and swing widely by comparison implying changing drift rate.

Until recently, a GPS receiver typically didn't ask much of the clock beyond a modest requirement on the phase noise to allow the phase lock loops to track a signal smoothly [Section 12.3]. However, the newer applications dealing with more challenging signal environments (indoors or in the presence of radio frequency interference) are discovering the benefits of stable GPS receiver clocks. The GPS applications have served as an important impetus for a DARPA-sponsored effort to develop chip-scale atomic clocks referred to above. Such clocks have a great potential for promoting assisted GPS (AGPS) [Section 13.4] and raising the anti-jam margins.

4.2.4 GPS Time

Like UTC, GPS Time (GPST) is a composite time kept by a 'paper clock.' GPST is defined on the basis of measurements from a set of cesium and rubidium frequency standards in use at the monitor stations and aboard the satellites. But there are two important differences between GPST and UTC. First, GPST is defined in real time and, secondly, it is a continuous time scale (no leap seconds). GPST is 'steered' to remain within one microsecond (1 μs) of UTC(USNO) modulo one second. Actually, GPST has been maintained within about 10 ns of UTC(USNO) in recent years. As a result, GPST differs from UTC by a certain number of whole seconds plus a fraction of a microsecond. As of 1 January 2006, GPST is ahead of UTC by 14 seconds.

$$\text{GPST} - \text{UTC} \approx +14 \text{ s}$$

A time epoch in GPST is defined in terms of the week number and number of seconds of the week (or 'seconds into the week'). The GPS Week field in the navigation message is modulo 1024. The first GPS cycle of 1024 weeks began at the Standard Epoch of GPS: midnight of the transition between Saturday, 5 January 1980, and Sunday, 6 January 1980 (00:00:00 UTC, 6 January 1980, Julian Day 2,444,244.500). The first week-rollover occurred at midnight (GPST, not UTC) 21–22 August 1999. Apparently, the Control Segment dealt with the rollover without a hitch but some old receivers were tripped up by the event. The number of seconds of the week is measured since the previous midnight (GPS Time) marking the transition between Saturday and Sunday. There are 604,800 seconds in a GPS week.

The navigation data message broadcast by each satellite carries time stamps in accordance with the satellite clock every 6 s in the form of number of seconds into the week. These time stamps specify transmission times of certain bits, as discussed in Section 4.3.6. A satellite time scale can be related to GPST and UTC, as discussed below.

GPS satellites carry rubidium and cesium atomic frequency standards. Block II and IIA satellites carried two of each. Each Block IIR carries three rubidium standards. The performance of each clock on board is monitored by the Master Control Station (MCS), which selects one of the clocks to generate the signals. The bias in the satellite clock relative to GPST is modeled as a quadratic function of time, and the parameters of this model, estimated by MCS and uploaded to the satellites, are broadcast as a part of the navigation message. At t seconds (GPST), the satellite clock offset ($\delta t^s = t^s - t$) is computed from

$$\delta t^s(t) = a_{f0} + a_{f1}(t - t_{0c}) + a_{f2}(t - t_{0c})^2 + \Delta t_r \tag{4.6}$$

where t_{0c} is the reference epoch, a_{f0} is the clock offset (seconds), a_{f1} is the fractional frequency offset (seconds/second), and a_{f2} is the fractional frequency drift (seconds/second^2). The parameters a_{f0}, a_{f1}, and a_{f2} are sometimes identified as bias, drift, and aging parameters of the clock, respectively. Typically, parameter a_{f0} ranges between 1 μs and 1 ms; $a_{f1} \approx 10^{-11}$ s/s; and, for the cesium clocks, $a_{f2} = 0$ s/s^2. The term Δt_r is associated with relativistic correction, discussed below. With such correction terms computed and uploaded typically once a day, the satellite clocks are kept synchronized well within 5–10 ns.

In order to provide an estimate of UTC from GPS, the navigation message broadcast by each satellite includes estimates of the time difference between GPST and UTC(USNO) modulo one second, and its rate. The navigation message also includes the whole-second difference between the two time scales due to leap seconds. These parameters allow a receiver clock to calculate an accurate estimate of UTC(USNO). The current accuracy of such estimates is about 25 ns rms.

Clocks aboard the satellites are subject to *relativistic effects*. We offer only a brief account. A careful definition of the atomic second and the time scales based on it would have required us to state that these clocks are at rest on the geoid. According to Einstein's theories, a clock will tick at a different rate if it is in motion, or placed above or below the geoid. The GPS clocks are about 20,000 km above the geoid and moving at about 4 km/s. As we'll see below, ignoring relativistic effects would have led to intolerable errors in GPS-based position and time.

According to the Special Theory of Relativity, a clock aboard a satellite traveling at a constant speed would appear to lose time relative to a clock on the ground. According to the General Theory of Relativity, a clock in a satellite would run faster than one on the ground due to the difference in gravitational potential. The net effect on a clock aboard a satellite in a circular orbit around the earth with radius of 26,560 km would be to gain 38.4 μs per day. In order to compensate for this effect, the fundamental frequency of the nominally 10.23-MHz GPS satellite clocks is set 0.0045674 Hz lower. This is often referred to as 'factory offset.'

No further relativistic compensation would have been required if the GPS orbits were truly circular. Actually, the eccentricity of a GPS satellite orbit can be as high as 0.02. In an elliptical orbit, both the speed of the satellite and the gravitational potential change with the position of the satellite in its orbit [Section 4.3.3]. The orbital parameters are computed by the Control Segment and transmitted by each satellite. A receiver applies the relativistic correction to the satellite clock time in accordance with the GPS Interface Specification [IS-GPS-200D (2004)]. The magnitude of this time-dependent correction can range between zero and about 45 ns depending upon the position of the satellite in its orbit. For a readable account of these relativistic corrections, see Ashby (2002).

4.3 GPS Orbits and Satellite Position Determination

The ability to predict accurately the position of a satellite at the instant of signal transmission is vital to the operation of GPS. Indeed, the position of a satellite is determined within meters of its true location on the basis of the orbital parameters broadcast by each satellite as a part of its navigation message. What's truly impressive is that these orbital parameters are predicted

on the basis of measurements made 24–48 hours earlier. Such success in orbital prediction is based on knowledge accumulated over nearly five hundred years of study of the laws of celestial mechanics, and experience with artificial satellites since 1957 when the space age began with the launch of *Sputnik I* by the Soviet Union.

In this section, we introduce the parameter set used to describe the orbit of a GPS satellite over a time interval and to compute the satellite position at an epoch. But first we step back and describe a six-parameter set, called Keplerian elements, used to characterize an ideal elliptical orbit prescribed by Kepler's laws of planetary motion. We also discuss computation of the position and velocity of a satellite in terms of the Keplerian elements. A discussion of the perturbing forces on a satellite leads to the sixteen-parameter pseudo-Keplerian elements whose values are broadcast by each GPS satellite. We begin below with Kepler's laws, marked optional because we expect the student to be familiar with them.

4.3.1 Kepler's Laws*

Until Copernicus (1473–1543), the prevalent and politically safe view of the universe was geocentric: The earth was at the center of the universe, and the other heavenly bodies revolved around it. There was a price to be paid for a divergent viewpoint. But everything changed in the sixteenth and seventeenth centuries, mainly due to Tycho Brahe (1546–1601), Johannes Kepler (1571–1630), and Isaac Newton (1642–1727). Credit goes to Brahe for methodical observations of the positions of Mars and other planets recorded painstakingly over many years with instruments of his own design in a pre-telescope, pre-pendulous-clock era. Kepler's extraordinary accomplishment was to sift patiently through Brahe's measurements and to decipher a pattern in these data to show how the planets moved around the sun. The crowning achievement of Newton was to show why.

We begin with Kepler's three laws of planetary motion (see Figure 4.10).

- The orbit of each planet is an ellipse with the sun at one of the foci.
- Each planet revolves so that the line joining it to the sun sweeps out equal areas in equal lengths of time regardless of the position of the planet in its orbit.
- The squares of the periods of any two planets are in the same proportion as the cubes of their mean distances from the sun.

The first law is straightforward, though it surprised Kepler and his contemporaries, who expected the orbits to be circular, or generated in some way using circles (e.g., circles moving on circles). The second and third laws, though stunning in their simplicity and elegance, were far beyond the capacities of the science establishment of the day to appreciate. Galileo, a contemporary of Kepler, continued to believe in circular orbits and doubted the validity of Kepler's laws. He had about thirty years in which to change his mind, but apparently never did. The second law says that the speed of a planet is not constant—the farther a planet gets from the sun, the slower it moves. According to the third law, the square of the time taken by a planet to complete its orbit is proportional to the cube of its mean distance from the sun. It is left as an exercise to show that for elliptical orbits the mean distance from a focus to the planet is the same as the semi-major axis length. The orbital periods of two satellites with the same semi-major axis length are, therefore, equal, regardless of the shape of the orbit (see Figure 4.10).

Kepler showed that these laws also governed the motion of the earth's moon and the satellites of Jupiter. Kepler's laws apply equally well (to first order) to the GPS satellites revolving around the earth. Accustomed as we now are to computerized data files and user-friendly software packages for statistical analysis, Kepler's achievement in empirical data analysis nearly four hundred years ago is simply awe-inspiring. He reached his conclusions with unremitting labor over long multiplications and divisions armed only with the conviction that a theory must fit all the facts. Kepler died a poor man.

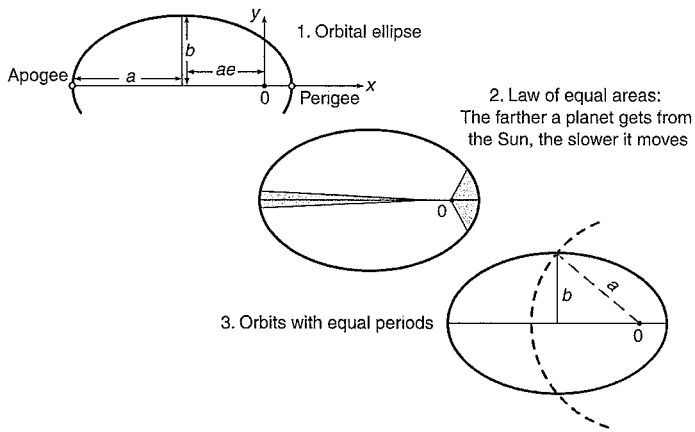


Figure 4.10 Kepler's laws of planetary motion illustrated.

In 1687, nearly fifty years after Kepler's death, Newton supplied the scientific foundation to Kepler's laws, showing that they could be derived from his theory of universal gravitation. It was the first unified theory: *Heavens followed the same laws of motion as objects on the earth*! Newton devised a new branch of mathematics, now called calculus, to prove that the planets move along elliptical paths around the sun, satisfying Kepler's laws. Newton also showed that the constant of proportionality in Kepler's third law was related to the masses of the sun and the planet and the gravitational constant.

The motion of a GPS satellite is governed by Newton's laws of motion. Recall the second law: Acceleration of the center of mass of a body is proportional to the force applied to it. At its simplest, we can model the earth and a satellite as point masses and analyze satellite motion under the force of gravitational attraction. Actually, there are other perturbing forces, but they are smaller than the gravitational force by orders of magnitude, and we will ignore them for now. The requirement of a precise orbit prediction, however, makes it necessary to account for these forces, and we will do so in a later section.

Let the masses of the earth and a satellite be represented as M and m, respectively, and the positions of these point masses in an inertial coordinate frame be represented by $\mathbf{r}_E$ and $\mathbf{r}_s$, respectively. The position of the satellite relative to the earth will be denoted by $\mathbf{r} = \mathbf{r}_s - \mathbf{r}_E$. Then, according to Newton's law of universal gravitation, the force of earth on the satellite is

$$\mathbf{F} = -\frac{GMm}{r^2}\frac{\mathbf{r}}{r} \tag{4.7}$$

The gravitational force of the satellite on the earth is –**F**. Now we apply Newton's second law to determine the equation of motion of the satellite relative to the earth.

$$M\ddot{\mathbf{r}}_E = \frac{GMm}{r^3}\mathbf{r}; \quad m\ddot{\mathbf{r}}_s = -\frac{GMm}{r^3}\mathbf{r},$$

where G is the universal gravitational constant (6.673×10^{-11} m^3 Kg^{-1} s^{-2}). From the above pair of equations we obtain

$$\ddot{\mathbf{r}} + \frac{G(M+m)}{r^3}\mathbf{r} = 0 \tag{4.8}$$

a second-order, nonlinear differential vector equation in the position vector of the satellite relative to the center of the earth. Considering that M, the mass of the earth $\approx 6 \times 10^{24}$ Kg, and m, the mass of the satellite $\approx$ 1000 Kg, a good approximation to (4.8) is

$$\ddot{\mathbf{r}} + \frac{GM}{r^3}\mathbf{r} = 0 \tag{4.9}$$

where $GM = 3{,}986{,}004.418 \times 10^8$ m^3/s^2 is the earth's gravitational constant.

The solution of (4.9) is found in any textbook on celestial mechanics or astrodynamics, and will not be presented. Suffice it to say that each of Kepler's laws can be derived from it. The first integration of (4.9) would introduce three constants: instantaneous velocity vector at a given instant. The second integration would introduce three more constants: the instantaneous position vector at the given instant. The orbit can, therefore, be completely described by six elements in the form of initial conditions:

$$\mathbf{r}_0 = (x_0, y_0, z_0),$$
$$\dot{\mathbf{r}}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0),$$

specified at some epoch t_0. Given these initial conditions, the differential equation can be integrated forward (or, backward) in time to determine the satellite position at a time of interest. In the next section, we discuss an alternate characterization of an ideal, elliptical satellite orbit using more intuitive and geometrical notions in the form of Keplerian parameters. The GPS' characterization of a satellite orbit, and the position of a satellite on it, is in the form of an expanded set of Keplerian-like parameters. We'll describe these in a later section.

It is worth repeating that (4.9) is the equation of motion for an idealized case: Both the earth and the satellite are assumed to be point masses (or spheres uniform in composition), and there are no other forces in play. This is the classic *two-body, central-force-field problem.* (The gravitational force on each body acts along the line joining the centers of mass of the two bodies.) The orbital plane is fixed in the inertial space. The orbital ellipse is fixed in the orbital plane. The position of a satellite at an instant can be determined if its position and velocity are known at another time instant.

4.3.2 Ideal Elliptical Orbits: Keplerian Elements

Under idealized conditions discussed above, the motion of a GPS satellite would be characterized by an elliptical orbit fixed in space with the earth at one of the foci. Such an orbit can be specified by six parameters. These parameters could be the six elements of satellite position and velocity vectors at a specified epoch. An alternate, geometrical characterization can be

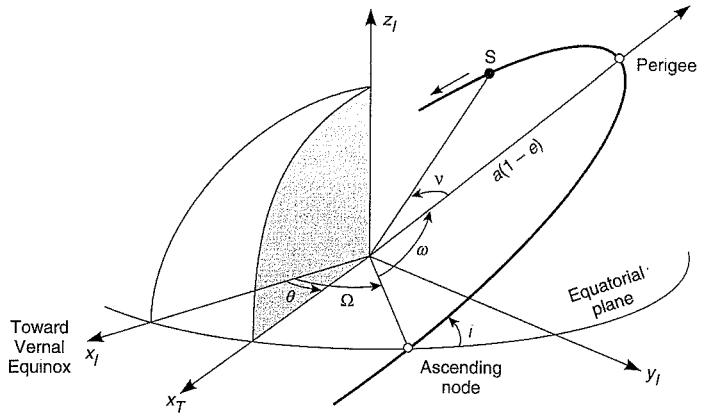


Figure 4.11 Characterization of an ideal orbit and the satellite position by Keplerian elements: {a, e, i, Ω, ω, and ν}.

given in terms of six Keplerian elements defined below (see Figure 4.11). Five of the elements describe the size and shape of the orbit and its orientation in space. The sixth element specifies the position of the satellite at a particular time instant or epoch. Given the six elements, the satellite position and velocity can be computed at any other epoch. [A good general reference for the material presented in this section is Bate, Mueller and White (1972).]

- First we characterize the orbital ellipse by two parameters which determine its size and shape:
  - *semi-major axis* (a),
  - *eccentricity* (e).
- Next, we specify the orientation of the orbital plane relative to the fixed stars (i.e., inertial reference frame) by two parameters:
  - *inclination* (i), defined as the angle between the orbital plane and the earth's equatorial plane,
  - *right ascension of the ascending node* or RAAN (Ω), defined as the angle measured in the earth's equatorial plane between a reference direction in space pointing to the vernal equinox (i.e., the direction determined by the intersection of the equatorial plane of the earth with the plane of the earth's orbit around the sun), and the ascending node, the point on the satellite's orbit where it crosses the equatorial plane, moving in the northerly direction. (The intersection of the orbital plane and the earth's equatorial plane defines the *line of nodes.*) Inclination is the angle between the equatorial plane and the half-plane containing the part of the orbit stretching from the ascending node to the descending node. The right ascension of the ascending node is measured in a counter-clockwise sense when viewed from the positive z-axis.
- We characterize the orientation of the ellipse in the orbital plane by a single pa-

rameter: *argument of perigee* (ω), defined as the angle in the plane of the orbit between the ascending node and the perigee, i.e., the point at which the satellite is closest to the center of the earth. (The point in the orbit at which the satellite is farthest from the center of the earth is called the apogee.). Argument of perigee is measured in the direction of motion of the satellite.

- Finally, we need an element to specify the position of the satellite in the orbit at a given epoch. There are several ways to do this. We start by introducing *true anomaly* (ν), defined as the angle measured in the orbital plane between the perigee and the satellite position at the specified instant. The sum of the true anomaly and the argument of perigee defines *argument of latitude*: It's the angle between the ascending node and the satellite position as measured in the orbital plane. It turns out to be more convenient to specify the true anomaly in terms of two other descriptors of the satellite position in its orbit, eccentric anomaly and mean anomaly, described in the next section. Note that for a circular orbit ($e = 0$), the argument of perigee and true anomaly are undefined. The satellite position, however, can be specified by argument of latitude.

4.3.3 Satellite Position and Velocity

The ideal orbital plane and the ideal elliptical orbit are specified completely by the five Keplerian parameters discussed in the last section (see Figure 4.11): a, e, i, Ω, and ω. The satellite position can be determined at all times if its true anomaly is given at an epoch. We now want to determine the satellite position and velocity expressed in a Cartesian coordinate system (inertial and terrestrial) in terms of these Keplerian elements.

Let's define an *orbital coordinate system*. The origin is located at the focus of the elliptical orbit corresponding to the position of the center of mass of the earth. The x- and y-axes are defined as along the major axis and parallel to the minor axis, respectively, as shown in Figure 4.12 as x_O-axis and y_O-axis (the subscript 'O' stands for orbital). The x_O-axis points toward the perigee. The z_O-axis is orthogonal to the orbital plane, and is not shown because there is no action along this axis in our ideal system. The satellite position in the orbit is shown by S. The corresponding true anomaly and position vector are denoted by ν and $\mathbf{r}$, respectively. Both $\mathbf{r}$ and ν are functions of time, but for simplicity we will not show this dependence explicitly. It is a left as an exercise to show that

$$\mathbf{r} = \frac{a(1-e^2)}{1+e\cos\nu}\begin{bmatrix}\cos\nu\\ \sin\nu\\ 0\end{bmatrix} \tag{4.10}$$

The magnitude of this vector

$$r = \|\mathbf{r}\| = \frac{a(1-e^2)}{1+e\cos\nu} \tag{4.11}$$

is known as the *orbit radius*. It is the distance between the earth's center and the satellite.

As noted previously, the true anomaly is more conveniently obtained in terms of two other parameters, which we now define. The *eccentric anomaly*, E, is defined, as shown in Figure 4.12, as the angle subtended at the center of the orbit between the perigee and the projection of the satellite position on a circle of radius a. Bear in mind that this center coincides with the

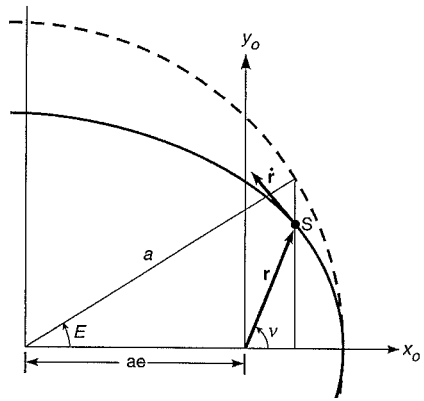


Figure 4.12 Position and velocity of a satellite in a Keplerian orbit.

center of a circular orbit ($e = 0$), and it's the point midway between the two foci for an elliptical orbit.

The satellite position vector in the orbital coordinate frame can be expressed in terms of a, e, and E as

$$\mathbf{r} = \begin{bmatrix} a\cos E - ae \\ a\sqrt{1-e^2}\sin E \\ 0 \end{bmatrix} \tag{4.12}$$

Therefore, an alternate expression for the orbit radius (4.11) is

$$r = a(1 - e\cos E) \tag{4.13}$$

From (4.10) and (4.12), the eccentric anomaly is related to the true anomaly by

$$\nu = \tan^{-1}\left(\frac{\sqrt{1-e^2}\sin E}{\cos E - e}\right) \tag{4.14}$$

In order to give another useful characterization of the satellite position, we first define the mean angular velocity of the satellite, called *mean motion* and denoted by n. If the period of the satellite is T,

$$n = \frac{2\pi}{T} = \sqrt{\frac{GM}{a^3}} \tag{4.15}$$

The average angular velocity of the satellite and, therefore, the orbital period, depend upon the semi-major axis only!

Now we are ready to define the *mean anomaly*, M. The mean anomaly is a non-geometrical quantity defined as the angle between the perigee and a fictitious satellite in a circular orbit with the same focus and same period as the satellite of interest but moving with a constant

speed. The constant speed is the mean motion of the satellite. The actual and fictitious satellites cross the apogee-perigee lines in phase, and the mean anomaly of the satellite is the true anomaly of this fictitious satellite. The mean anomaly at epoch t is given by

$$M = n(t - t_p)$$

where t_p is the time of the perigee crossing. Note that M is a linear function of time and, for a circular orbit, M is the same as the argument of latitude.

The mean and eccentric anomalies are related by *Kepler's equation*, which we state without proof [Bate, Mueller, and White (1971)].

$$M = E - e \sin E \tag{4.16}$$

Given E, we can easily solve for M. Given M, (4.16) can be solved for E, but not in closed form. Kepler's equation is solved by iterative techniques or a solution is expressed in a series form.

Having defined the eccentric and mean anomalies, we can now write an expression for the satellite velocity vector expressed in the orbital coordinate frame as follows by taking the first time derivatives in (4.12)

$$\dot{\mathbf{r}} = \frac{na}{(1 - e\cos E)} \begin{bmatrix} -\sin E \\ \sqrt{1-e^2}\cos E \\ 0 \end{bmatrix} \tag{4.17}$$

where we have used Kepler's equation (4.16) to get the rate of E

$$\dot{M} = \dot{E}(1 - e\cos E) = n \tag{4.18}$$

Differentiating (4.17) one more time yields the familiar equation of motion

$$\ddot{\mathbf{r}} = -\frac{GM}{r^3}\,\mathbf{r}$$

which is reassuring, and its derivation is left as an exercise.

The position and velocity vectors, expressed above in the orbital coordinate system (Figure 4.12), can be transformed into the ECEF coordinate system in two steps by appropriate rotations as follows. First, let us transform the position vector $\mathbf{r}$ given above in the orbital coordinate system into the inertial coordinate system defined in Section 4.1.1, and represent it as $\mathbf{r}_I$. From Figures 4.11 and 4.12, we can see how the inertial coordinate system, with axes marked as x_I, y_I, and z_I, may be rotated into the orbital coordinate system with axes x_O, y_O, and z_O (z_O is not shown). This step requires three rotations readily seen in Figure 4.11. The first rotation occurs about the z_I-axis by an angle Ω, followed by a rotation about the new position of the x_I-axis by angle i, and, finally, a rotation about the new position of the z_I-axis by an amount ω. The transformation can, therefore, be written as

$$\mathbf{r} = \mathbf{R}_3(\omega)\mathbf{R}_1(i)\mathbf{R}_3(\Omega)\mathbf{r}_I \tag{4.19}$$

Using the basic properties of the rotation matrices [Appendix 4.A], we can write the inverse transformation as

$$\mathbf{r}_I = \mathbf{R}_3(-\Omega)\mathbf{R}_1(-i)\mathbf{R}_3(-\omega)\mathbf{r} \tag{4.20}$$

This satellite position vector, expressed in the inertial coordinate system, satisfies the basic equation obtained from Newton's second law of motion (4.7). Each of the Keplerian elements can be written down as a function of $\mathbf{r}$ and $\dot{\mathbf{r}}$ at a specified time.

We discussed the transformation of vectors between terrestrial and inertial reference systems in Section 4.1.1 in the context of Figure 4.2. We now disregard polar motion. This conceptual simplification is necessary in order to be consistent with the Control Segment's estimation procedure which buries the effect of polar motion in the ephemeris parameters. In order to transform the satellite position vector into the ECEF coordinate system, we need only one rotation to account for the fact that the reference meridian rotates around the earth's spin axis. The rotation by Greenwich apparent sidereal time (see Figure 4.11) gives

$$\mathbf{r}_T = \mathbf{R}_3(\theta)\mathbf{r}_I \tag{4.21}$$

where the subscript T identifies the vector as expressed in the terrestrial reference system.

The velocity vector can be transformed similarly to obtain its representation in the ECEF [Homework Problem 4-12]. A key step is to obtain from (4.21)

$$\dot{\mathbf{r}}_T = \mathbf{R}_3(\theta)\dot{\mathbf{r}}_I + \dot{\mathbf{R}}_3(\theta)\mathbf{r}_I \tag{4.22}$$

4.3.4 Perturbed Keplerian Orbits*

We have dealt so far with the ideal Keplerian orbits for the two-body, central-force-field problem. Such orbits are fixed in space, and are characterized once and for all by specifying the values of five parameters. Kepler's laws would have described the motion of a GPS satellite if the earth had been spherical in shape and uniform in composition, and the only force acting on the satellite was the gravitational pull of the earth. The earth is not a uniform sphere, and there are other forces, beside gravity, at play. The net result of these perturbing forces is that the orbit of a GPS satellite changes with time, and has to be characterized with an appropriate set of time-dependent parameters.

Disregarding intervention in the form of rocket firings or any other intentional maneuvers, the main forces perturbing the motion of a GPS satellite are as follows.

Non-central gravitational force field. As noted previously, the shape of the earth resembles an ellipsoid with equatorial radius about 20 km longer than the polar radius. The density of the earth is not uniform either, and the gravitational force of the earth varies with latitude and longitude, in addition to the radial distance. The gravitational potential function is modeled as a spherical harmonic expansion which we have encountered already. For our immediate purpose, it is enough to examine a two-term gravity potential function from (4.2).

$$V(r, \phi', \lambda) \approx \frac{GM}{r}\left[1 - \frac{\sqrt{5}}{2}\left(\frac{a}{r}\right)^2 \bar{C}_{20}(1 - 3\sin^2\phi')\right] \tag{4.23}$$

The first term (GM/r) is the potential for a spherical earth of uniform density. Its gradient

$$\nabla\left(\frac{GM}{r}\right) = -\frac{GM}{r^3}\mathbf{r} \tag{4.24}$$

corresponds to the central force for ideal Keplerian motion, giving a centripetal acceleration ≈ 0.56 m/s^2 for GPS orbits. The second term with coefficient $\bar{C}_{20}$, the harmonic coefficient of degree two and order zero, essentially models the ellipsoidal shape of the earth, disregarding

anomalies due to variation in the density. Its gradient gives the major non-central component of the force on a satellite. From NIMA (1997), $\overline{C}_{20} \approx -0.5 \cdot 10^{-3}$. (The second term on the right-hand side of (4.23) is often referred to as the J_2 term, where $J_2 = -\sqrt{5}\,\overline{C}_{20} \approx 1.1 \cdot 10^{-3}$.) The inclination of a GPS satellite orbit is 55° and, therefore, $\phi' < 55°$. The corresponding acceleration is less than 5×10^{-5} m/s^2. The coefficients of the terms of higher degree and order are smaller by several orders of magnitude.

The equatorial bulge of the earth produces two effects. First, the non-radial component of the force on the satellite produces a torque which results in rotation of the orbital plane, and the line of nodes thus rotates in the inertial coordinate frame. This effect depends upon the inclination of the orbit, and is zero for a polar orbit and maximum for an equatorial orbit. It is a secular effect (i.e., grows with time) and amounts to a change of about –1.2° per month in RAAN for a GPS satellite.

The second effect of the equatorial bulge is to produce a twice-per-orbit harmonic perturbation (period ≈ 6 h). Each time the satellite approaches the equatorial plane, it experiences a greater gravitational force and speeds up. As the satellite gets farther away from the equatorial plane, it slows down. This results in rotation of the major axis in the orbital plane.

Gravitational fields of the sun and moon. The larger of these perturbing forces is the gravitational attraction of the moon. The sun, though much more massive is also much farther away. Both these forces have to be accounted for. The gravitational forces of the sun and the moon also produce tides, deforming the shape and, therefore, gravitational potential of the earth. This effect, however, is negligible for GPS orbits.

Solar radiation pressure. The photons striking a satellite exert a minute pressure on it. The acceleration due to such a force depends mainly upon the mass of the satellite and its surface area exposed to the sun. The solar radiation pressure is zero while a satellite is in the earth's shadow.

The equation of motion for a GPS satellite can now be written as [Beutler (1996)]

$$\ddot{\mathbf{r}} = -\frac{GM}{r^3}\,\mathbf{r} + \mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t) \tag{4.25}$$

where $\mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t)$ denotes the perturbation forces. Given that $GM/r^2 >> |\mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t)|$, the solution to unperturbed equations can serve as a good approximation in the neighborhood of an initial condition.

The effect of these perturbing forces is shown in Table 4.2. Note that the perturbing accelerations appear very small but they can add up to significant changes in a satellite orbit if not

Table 4.2 Forces acting on a GPS satellite and the resultant accelerations

Force	*Acceleration (m/s^2)*
Central gravitational force (GM/r^2)	0.56
Equatorial bulge (J_2 term)	5×10^{-5}
Lunar/solar gravity	5×10^{-6}
Solar radiation	10^{-7}

taken into account over an extended period. For example, the perturbation effect of the lunar/solar gravity can be about 25 m after one hour. Clearly, the accuracy required of GPS would not allow these effects to be ignored in orbital prediction by the Control Segment.

4.3.5 GPS Orbital Parameters

In the presence of the perturbing forces discussed above, we could regard the orbit of a GPS satellite, to first approximation, as Keplerian, and treat the perturbations as temporal variations in the six parameters introduced earlier. Instead, GPS accounts for these perturbations with an expanded orbital parameter set which retains a Keplerian look. This parameter set was chosen as a compromise to obtain the required accuracy without burdening a receiver with undue computational and storage requirements. This trade-off must be understood in the context of the computer capabilities in the early 1970s [Van Dierendonck *et al.* (1980)].

The expanded set of *quasi-Keplerian parameters* consists of 15 elements whose values are specified relative to a reference epoch. So, there are 16 ephemeris parameters in all. There is also an identifier parameter, called Issue of Data—Ephemeris (IODE), associated with each ephemeris parameter set. A change in the value of IODE between two successive transmissions indicates that the ephemeris parameter set has been updated.

The expanded set of ephemeris parameters includes two additional types:

- Three rate parameters to account for the linear changes with time: rate of the right ascension (Ω, or Ω-dot), rate of the inclination (i, or i-dot), and correction to the mean motion (Δn).
- Three pairs of amplitude terms (C_c, C_s) for sinusoidal corrections modeled as

  $$C_c \cos(2\Phi) + C_s \sin(2\Phi)$$

  where Φ is the argument of latitude (i.e., the angle of the satellite measured in the orbital plane from the equator [$\Phi = \omega + v$]). There is one pair each for argument of latitude, orbit radius, and inclination angle.

The ephemeris parameters broadcast by a GPS satellite, listed in Table 4.3, are described fully in the Interface Specification (IS), which also provides a step-by-step account of how to use them to compute the position and velocity of a satellite in WGS 84 ECEF coordinate frame [IS-GPS-200D (2004), Table 20-IV, p. 111]. A student who has understood our treatment in this section would have no trouble following the discussion in the IS. We only have to introduce one more term. GPS uses *longitude of the ascending node* (LAN) of the orbit plane rather than the right ascension of the ascending node (RAAN), introduced earlier, because it facilitates the transformation of the satellite position to the WGS 84 ECEF coordinate frame. LAN is measured in the equatorial plane from the reference meridian to the line of nodes (Figures 4.2 and 4.11):

$$\text{LAN}(t) = \text{RAAN} - \text{GAST}(t) \tag{4.26}$$

The IS doesn't specifically address the computation of satellite velocity, but the steps are similar to those for position computation [Homework Problem 4-12].

The homework problem set at the end of this chapter includes exercises which require a student to implement the processing steps laid out in the IS. The equations in the IS are gener-

Table 4.3 Ephemeris parameters in the GPS navigation message

Parameter	*Description*
t_{0e}	ephemeris reference time
$\sqrt{a}$	square root of the semi-major axis
e	eccentricity
i_0	inclination angle at the reference time
Ω_0	longitude of the ascending node at the beginning of the GPS week
ω	argument of perigee
M_0	mean anomaly at the reference time
Δn	correction to the computed mean motion
i (i-dot)	rate of change of inclination with time
Ω (Ω-dot)	rate of change of RAAN with time
C_{uc}, C_{us}	amplitudes of harmonic correction terms for the computed argument of latitude
C_{rc}, C_{rs}	amplitudes of harmonic correction terms for the computed orbit radius
C_{ic}, C_{is}	amplitudes of harmonic correction terms for the computed inclination angle

ally easy to follow but the English text can be rough. Here is an example: "Whenever the *effectivity* time of the leap second event, as indicated by WN_{LSF} and DN values, is in the 'past' (relative to the user's current time) ..." (italics added). Search for 'effectivity' brings out four other instances.

The orbital parameters are predicted by the Master Control Station on the basis of the code and carrier phase measurements at the monitor stations (see Figure 2.2). These measurements are used to predict the future orbits of the satellites. The quality of the predicted orbits improved in 2005 with the addition to the monitor station network of six stations operated by the National Geospatial-Intelligence Agency (NGA). Five more NGA stations will be added in the future.

The computed orbit, called an *ephemeris* (plural: *ephemerides*, pronounced efeh-méreh-deez), is a tabulation of the position and velocity vectors over time. The ephemeris parameters are estimated by a least-squares curve fit to the orbit, typically using overlapping data spans of four hours. The parameter sets covering several days are uploaded to the satellite, which broadcasts the appropriate set. The ephemeris parameters broadcast by a satellite currently change every two hours.

At present, the orbital data are uploaded to the satellites typically once a day. An upload provides orbital element data sets to be broadcast over the next fourteen days. Without fresh updates, the quality of the broadcast ephemeris would deteriorate with time. The period of autonomous operation for a satellite would be extended in the future to a much longer period. The crosslinks and ability to range among the satellites would make the operation even more

precise. In 2005, the quality of the broadcast ephemerides is such as to introduce an rms error of about 2 m in the pseudorange measurements, as discussed in Chapter 5.

Precise Ephemerides

Applications which require precise positions but have the luxury of time can access satellite tracking data from the GPS monitor stations or reference stations obtained before, during, and after the time interval of interest. These applications can use post-processed ephemerides. The NGA and the Naval Surface Warfare Center (NSWC) produce the 'official' post-processed precise ephemerides for the Control Segment to check the quality of the broadcast ephemerides based on the measurements from about a dozen monitor stations operated by the Air Force and NGA (Figure 2.2). The user community has access to precise ephemerides from several sources.

The International GNSS Service (IGS) now has a network of over 350 GPS reference stations distributed around the world and produces precise ephemerides (and satellite clock bias estimates) widely used by the geodetic community. These products have improved steadily in accuracy and timeliness. The current offerings are: Ultra-Rapid—Predicted (real time, error ~10 cm), Ultra-Rapid—Observed (latency: 3 hours, error < 5 cm), Rapid (17 hours, error < 5 cm), and Final (13 days, error < 5 cm) [http://igscb.jpl.nasa.gov/components/prods.html]. Precise ephemerides are also available from the analysis centers of IGS: National Geodetic Service (NGS), Jet Propulsion Laboratory (JPL), Center for Orbit Determination in Europe (CODE) at the University of Berne, NR Canada, and others.

Almanac

Each satellite broadcasts its own ephemeris. In addition, each satellite transmits in its navigation message a coarse version of the ephemerides of *all* satellites in the constellation in the form of an *almanac*. The almanac is a subset of the clock and ephemeris data with reduced precision. The purpose of the almanac is basically to allow a receiver to determine approximately when a satellite would rise above the horizon, given an approximate user position, so that the receiver can plan to initiate signal acquisition. The almanac parameters are, therefore, not required to be as accurate as the ephemeris parameters.

The almanac is updated much less frequently (currently, at least every six days) than the ephemeris. One-sigma range error for the almanac ephemeris can be up to 1–2 km. The satellite clock offset is specified using a linear model and its two parameters are given in the almanac. (Recall that the clock model used in the ephemeris is a quadratic in time, but in recent years the coefficient of the quadratic term appears to have been set to zero mostly.) The quasi-Keplerian parameter set to specify the orbits is limited to seven parameters: square root of semi-major axis, eccentricity, inclination, longitude of ascending node, rate of right ascension, argument of perigee, and mean anomaly, all specified at a common reference time. The satellite position is computed as with the larger parameter set in the ephemeris by setting the values of the parameters not specified in the almanac to zero.

4.3.6 GPS Navigation Data Message

GPS transmits navigation data message at 50 bits per second. Each bit is 20 milliseconds long. The message is formatted into frames of 1500 bits. It takes 30 seconds to transmit a frame. Each frame is organized into five subframes. Each subframe is 6 seconds long and contains

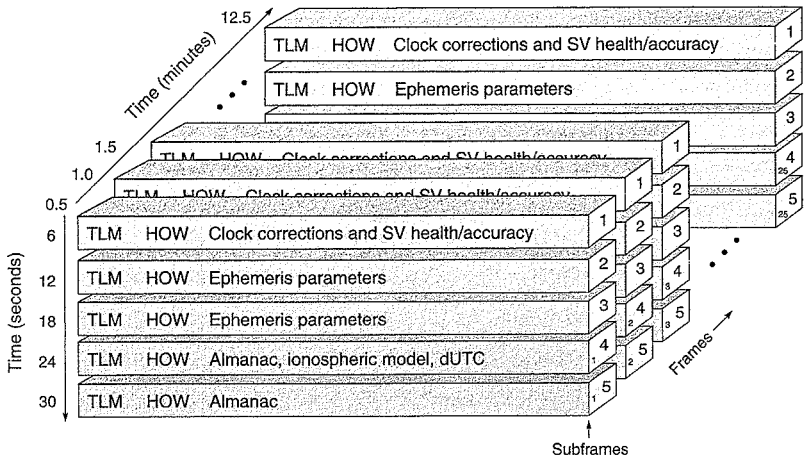


Figure 4.13 GPS navigation message organization: frames and subframes. Subframes 1, 2, and 3 repeat every 0.5 minutes; subframes 4 and 5 repeat every 12.5 minutes. Subframes 1, 2, and 3 are specific to the transmitting satellite; subframes 4 and 5 are common to all satellites. (Courtesy of Dr. Frank van Diggelen, Global Locate)

ten 30-bit words (see Figure 4.13). Subframes 1–3 typically repeat the same information from frame to frame. Subframes 4–5 of the consecutive frames, however, contain different 'pages' of the navigation message. It takes 25 frames (12.5 minutes) to transmit the complete navigation message. A unit of 25 frames is called a Master Frame. The information content of the various subframes is summarized below. A detailed description is given in the Interface Specification [IS-GPS-200D].

Subframe 1:	satellite clock corrections, health indicators, age of data
Subframes 2–3:	satellite ephemeris parameters
Subframe 4:	ionosphere model parameters, UTC data, almanac and health status data for satellites numbered 25 and higher
Subframe 5:	almanac and health status data for satellites numbered 1–24

The first two words of each six-second subframe have a special significance: Telemetry word (TLM) and Hand-over word (HOW). The TLM contains a fixed 8-bit synchronization pattern. The HOW word, discussed below, conveys the satellite time to a receiver.

Conveying Satellite Time to a Receiver: Z-Count

Each satellite keeps time in accordance with an atomic standard on board. Parameters used to relate the satellite time to GPST and UTC were discussed in Section 4.2.4. In the next chapter, we'll be concerned with the transit time of a signal from a satellite to the receiver, and we'll want to determine the time (per satellite clock) at the start of transmission of a navigation data bit. We have avoided all bit-and-byte-level discussion of the navigation message so far, referring the reader instead to the IS, but we make an exception here and give a brief account of the

parameters which give us the satellite time associated with the beginning of each subframe.

The satellite time is kept as number of 1.5-second epochs since the zero time point (currently, midnight between 21 August and 22 August 1999 [Section 4.2.4]) by a 29-bit number called Z-count. The 10 MSBs of the Z-count specify the week number (modulo 1024). The 19 LSBs specify the Time of Week (TOW) count in 1.5-second epochs. The first 17 bits of HOW, the second word in each subframe, are the 17 MSBs of TOW, specifying the time of week in 6-second epochs or the number of 6-second subframes since the beginning of the week and start of transmission of the next subframe. The Z-count increases by four between two consecutive subframes.

We'll resume this discussion of how to compute the transit time of a signal in Section 5.1.

4.4 GPS Satellite Constellation and Visibility Displays

A constellation is specified by defining the nominal orbits and positions of the satellite slots within each orbit, and tolerances for insertion and maintenance of the satellites within their assigned slots. The orbital characteristics of the baseline, 24-satellite GPS constellation are listed below [SPS (2001)].

- Semi-major axis: 26,560 km (± 50 km for Block II satellites)
  The orbital altitude above the earth corresponds roughly to three earth radii.
- Period: one-half mean sidereal day, or about 11 h 58 m.
  The earth spin period in space is about 23h 56m, so a satellite would repeat the same ground track after two revolutions. A stationary user would see the spatial distribution of the satellites above her repeat after 23h 56m. (The ground tracks appear to repeat from day to day, but not from month to month due to the slow westward drift, referred to above.)
- Eccentricity < 0.02.
- Six orbital planes, named A through F, all with inclination 55° (± 3°) relative to the equatorial plane. The right ascensions of the ascending node for the six orbital planes are separated by 60° in the equatorial plane.
- Four satellites per plane, distributed unevenly (see Figure 4.14). Two of the satellites are spaced between 30.0° and 32.1°. The remaining three inter-satellite spacings are between 92.38° and 130.98°. Relative spacing of the satellites and their ground tracks are maintained within their nominal values by adjusting their semi-major axes periodically.

The rationale for the uneven spacing of the satellites is to minimize the effect of single satellite failures out of the 24-satellite constellation. Apparently, a symmetric arrangement increases the chances of poor geometry when satellites fail. The baseline constellation geometry was obtained from a constrained optimization, and is claimed to offer the best global coverage defined as "PDOP < 10 for the best four satellites above 5° elevation with a single satellite failure." [Massatt and Zeitzew (1998)].

Each satellite is identified with a two-character code: A letter identifies the orbital plane (A through F), and a number identifies the satellite slot number in the plane (1 through 4). The assignment of slot numbers in an orbital plane is not progressive (see Figure 4.14 and Table

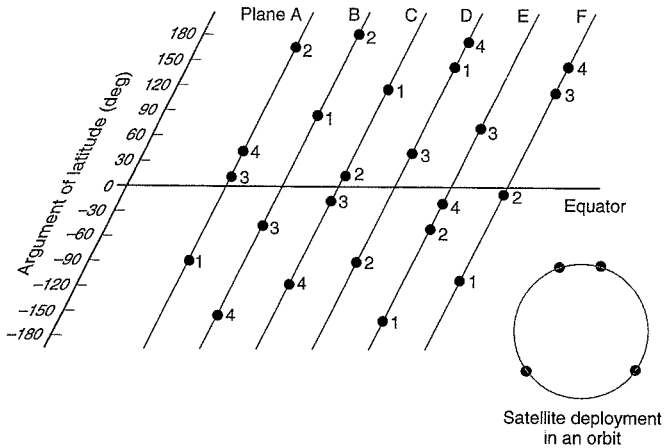


Figure 4.14 The baseline GPS constellation consists of six orbital planes with four satellites in each plane. Each satellite is identified with a two-character code: A letter identifies the orbital plane (A through F), and a number identifies the satellite slot number in the plane (1 through 4). [Adapted from Massatt and Zeitzew (1998)].

4.4): A3, A4, A2, A1; B1, B2, B4, B3; etc. There is no logic behind the numbering scheme. It just evolved to be what it is.

Each satellite has associated with it a unique PRN number corresponding with the PRN code transmitted by the satellite. There is also an SVN or NAVSTAR number assigned to each

Table 4.4 Orbital parameters for the baseline 24-satellite GPS constellation

Semi-major axis: 26,559.8 km; eccentricity: 0; inclination: 55°; argument of perigee: 0 [SPS (2001)]

Slot ID	*Right ascension (deg)*	*Mean anomaly (deg)*	*Slot ID*	*Right ascension (deg)*	*Mean anomaly (deg)*
A3	272.85	11.68	D1	92.85	135.23
A4	272.85	41.81	D4	92.85	167.36
A2	272.85	161.79	D2	92.85	265.45
A1	272.85	268.13	D3	92.85	35.16
B1	332.85	80.96	E1	152.85	197.05
B2	332.85	173.34	E2	152.85	302.60
B4	332.85	204.38	E4	152.85	333.69
B3	332.85	309.98	E3	152.85	66.07
C1	32.85	111.88	F1	212.85	238.89
C4	32.85	241.56	F2	212.85	345.23
C3	32.85	339.67	F3	212.85	105.21
C2	32.85	11.80	F4	212.85	135.35

satellite. Table 4.4 gives the orbital parameters for the baseline 24-satellite constellation corresponding to epoch 1993 July 1 0h. A graphical representation is given in Figure 4.14.

Several graphical devices and displays are used to help an observer determine which satellites would be visible from her position at a certain time, and how they would be distributed in the sky. Such analysis is commonly based on a recent almanac from a GPS satellite. A commonly used bar chart (see Figure 4.15) shows the rise and set times of the satellites for a location, typically over a twenty-four-hour period. The bars shift to the left by about four minutes from one day to the next.

An observer can determine from Figure 4.15 the number of satellites in view at a particular time but not their distribution in the sky. In order to see the changing 'geometry' of the satellites, we need a sky plot (or polar plot) shown in Figure 4.16. A sky plot shows the track of each satellite in the sky above an observer for the time period of interest. The outer circle corresponds to the observer's horizon, and the center corresponds to the zenith. The three inner circles correspond to different elevation angles. The position of a satellite is thus characterized by azimuth and elevation angle (both terms are defined in Appendix 4.A). We know intuitively that it would be better for position estimation if the satellites were 'well distributed' in the sky rather than all bunched up. We will revisit this notion in Chapter 6.

Satellite visibility plots like Figures 4.15 and 4.16 were mainstays for planning data collection in the early days of GPS while the constellation was sparse. Since 1995, however, the constellation has comprised 25-plus satellites and the users have come to expect that they will see 6–8 satellites unless the view of the sky is obstructed. Figure 4.16 can be customized to represent such obstructions and, with additional annotation (time, direction of satellite motion, etc.), can still be a valuable tool for planning data collections.

A constellation designer would be interested in a global characterization of the number of satellites in view. A plot of the ground tracks of the satellites is a useful graphical device for

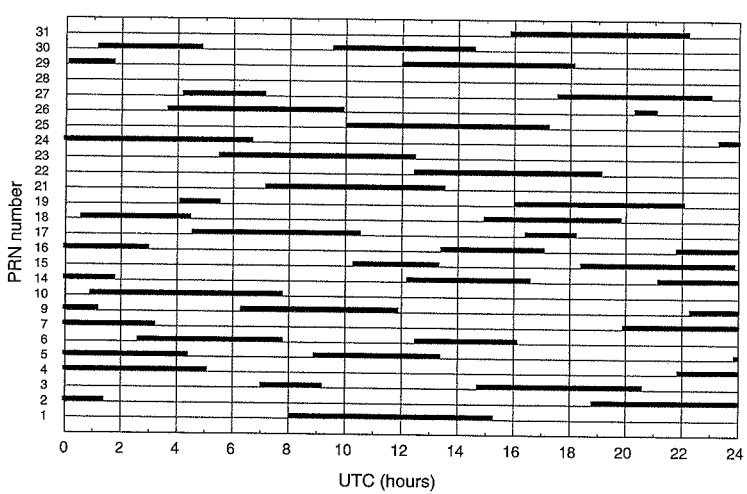


Figure 4.15 Satellite visibility (rise and set times) at Lexington, Massachusetts, plotted for a day in 1998.

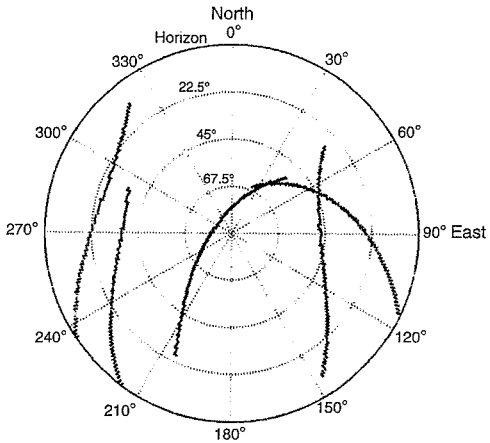


Figure 4.16 Sky view of satellite tracks.

this purpose. The *subsatellite point* is defined as the point at which the line from the satellite to the center of the earth intersects the earth's surface. An observer at a subsatellite point on the equator would see the satellite directly overhead, and nearly so at other latitudes (within 15 arc-minutes at mid-latitudes). A ground track is the locus of the subsatellite point as the satellite revolves in its orbit and the earth rotates. Because of uneven phasing of the satellites in each orbital plane, the longitudes of ascending node are all different, and there are 24 separate ground tracks. It is a messy picture we will skip.

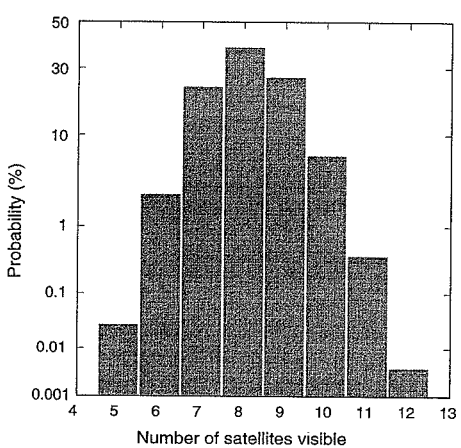


Figure 4.17 A histogram of the number of satellites in view of users worldwide (24-satellite baseline constellation, satellite elevation angle > 5°).

We present instead in Figure 4.17 a histogram of the number of satellites visible to users around the world with a clear view of the sky. This display is based on the baseline 24-satellite GPS constellation (Table 4.4), and we count only those satellites which are above 5° in elevation. According to this model, about 99.99% of GPS users around the world see six or more satellites, and 70% see eight or more. Figure 4.17 was generated in a simulation consisting of the following steps: (i) select a user location on the globe at random; (ii) select a time of day at random; (iii) compute satellite positions corresponding to the time selected and determine how many were in the user's view; and (iv) repeat steps (i) through (iii) a large number of times.

4.5 Summary

We have covered a lot of ground in this sprawling chapter in an attempt to fill in the background. The main ideas are:

- definition of WGS 84,
- GPS Time (GPST) reference and its relationship with other time references,
- GPS navigation message parameters and computation of satellite position and velocity at an epoch.

In order to specify the position of a point accurately, we first have to define a coordinate frame precisely. For GPS, the framework for specifying a position is provided by the World Geodetic System 1984 (WGS 84). WGS 84 comprises definitions of (i) an earth-centered, earth-fixed (ECEF) Cartesian coordinate frame in which the positions of the GPS satellites and the users are represented; (ii) an ellipsoid which serves as a geometric model for the shape of the earth, and (iii) the geoid which serves as the zero-level surface for measurement of height globally.

Measurement of time and time intervals is at the heart of GPS. The signals transmitted by GPS satellites are referenced to GPS Time. GPST is a composite time defined on the basis of the times kept by a set of atomic clocks at the monitor stations and in the satellites. Unlike UTC, the current civil standard, GPST makes no provisions for leap seconds to stay in sync with the earth's rotation. GPST, however, is steered to stay within 1 μs of UTC (modulo one second). GPST identifies a particular epoch in terms of two parameters: (i) the number of seconds that have elapsed since the previous Saturday/Sunday midnight, and (ii) a week number counted from the Standard Epoch of GPS (modulo 1024).

Each GPS satellite transmits a message at 50 bps. This navigation message provides the values of parameters and coefficients of models. Of particular interest is a sixteen-parameter set which allows a receiver to determine the satellite position and velocity at a time epoch.

Appendix 4.A Coordinate Conversion

4.A.1 Conversion between Geodetic and Cartesian Coordinates

We consider below the transformation of the coordinates of a point P from earth-centered, earth-fixed (ECEF) Cartesian coordinate frame (x, y, z) to the ellipsoidal coordinates (ϕ, λ, h), and vice versa. The center of the ellipsoid coincides with the origin of the ECEF Cartesian coordinate frame, and the minor axis (the axis of revolution) is coincident with the z-axis (Figure 4.A.1).

The transformations are made easier by defining distance N along the normal from P to the meridian ellipse between P′ and the z-axis. (A meridian ellipse is defined by connecting all points on the ellipsoid with λ = constant.) It is left as an exercise to show that

$$N = \frac{a^2}{(a^2 \cos^2 \phi + b^2 \sin^2 \phi)^{1/2}} = \frac{a}{(1 - e^2 \sin^2 \phi)^{1/2}} \tag{4.A.1}$$

Ellipsoidal to Cartesian. From Figure 4.A.1, the Cartesian coordinates (x, y, z) of a point with ellipsoidal coordinates (ϕ, λ, h) are given by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} (N + h) \cos\phi \cos\lambda \\ (N + h) \cos\phi \sin\lambda \\ (N(1 - e^2) + h) \sin\phi \end{bmatrix} \tag{4.A.2}$$

Cartesian to Ellipsoidal. An iterative scheme for the inverse transformation to obtain the ellipsoidal coordinates from the Cartesian coordinates is sketched below. The longitude is easy.

$$\tan\lambda = \frac{y}{x} \tag{4.A.3}$$

Getting the geodetic latitude and the geodetic height is a little tricky. Define

$$p = \sqrt{x^2 + y^2}$$
$$= (N + h)\cos\phi$$

from (4.A.2). Therefore,

$$h = \frac{p}{\cos\phi} - N \tag{4.A.4}$$

Also from (4.A.2),

$$z = [(1 - e^2)N + h]\sin\phi$$

and

$$\frac{z}{p} = \left(1 - e^2 \frac{N}{N + h}\right)\tan\phi$$

Therefore,

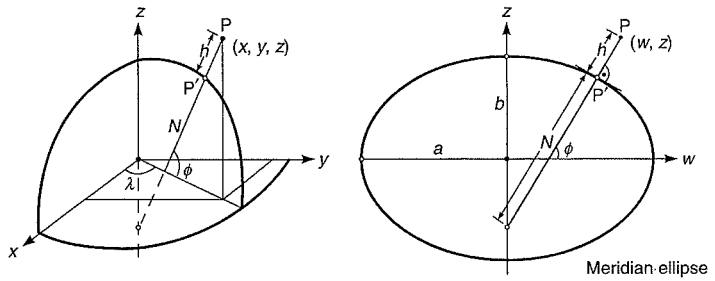


Figure 4.A.1 Cartesian and geodetic coordinates

$$\tan\phi = \frac{z}{p}\left(1 - e^2 \frac{N}{N + h}\right)^{-1} \tag{4.A.5}$$

Now we are all set. Obtain longitude from (4.A.3), and iterate between (4.A.4) and (4.A.5) to get ϕ and h. The implementation is left as an exercise.

4.A.2 Transformation between Cartesian Coordinate Frames

In this chapter, we saw the need to express the coordinates of a satellite in different Cartesian coordinate frames: inertial, orbital, and ECEF. In each case, the coordinate transformation was effected with elementary rotation matrices. The idea is very simple. Two Cartesian coordinate frames with a common origin can be brought into coincidence with three rotations of the axes of either coordinate frame. The coordinate frames in this book are all right-handed, and a positive rotation is defined as counterclockwise when looking toward the origin from the positive end of the axis about which the rotation takes place.

Let $\mathbf{x}_p = (x_p, y_p, z_p)^T$ represent a position vector expressed in a Cartesian coordinate frame, as shown in Figure 4.A.2. Consider another Cartesian coordinate frame obtained by rotating this coordinate frame about the z-axis (3-axis) by angle θ. The position vector $\mathbf{x}_p$ is represented in the new coordinate frame as $\mathbf{x}'_P = (x'_P, y'_P, z'_P)$

$$\mathbf{x}'_P = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}_P$$
$$= \mathbf{R}_3(\theta)\ \mathbf{x}_P$$

$\mathbf{R}_3(\theta)$ is a rotation matrix. We would obtain a similar expression for the position vector if we rotate the original coordinate frame about the x-axis (1-axis) or the y-axis (2-axis). The corresponding rotation matrices are

$$\mathbf{R}_1(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}$$

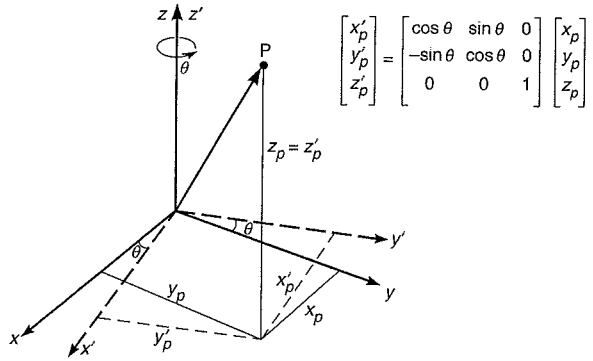


Figure 4.A.2 Rotation of a Cartesian coordinate frame.

and

$$\mathbf{R}_2(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$$

Each of these elementary rotation matrices is an orthogonal matrix: column (and row) vectors are of unit length, and are mutually orthogonal (i.e., the dot product of two column [or row] vectors is zero). Note also that

$$\mathbf{R}_i(-\theta) = \mathbf{R}_i^T(\theta) = \mathbf{R}_i^{-1}(\theta)$$

If a coordinate frame is rotated two or more times, the corresponding rotation matrix is obtained as the product of the elementary rotation matrices. It is important, however, to keep the order of rotation straight. The matrix product is not commutative, in general. Note, however, that

$$\mathbf{R}_i(\theta + \phi) = \mathbf{R}_i(\theta)\mathbf{R}_i(\phi) = \mathbf{R}_i(\phi)\mathbf{R}_i(\theta)$$

Consider a sequence of rotations. A position vector $\mathbf{x}$ is represented as $\mathbf{x}'$ in the coordinate frame obtained after the first rotation, $\mathbf{x}''$ after the second rotation, and $\mathbf{x}'''$ after the third rotation. The coordinate axes of these coordinate frames are similarly designated. It is left as an exercise to show that a rotation around the z-axis by an angle γ, followed by a rotation around the y'-axis by an angle β, followed by a rotation around the x''-axis by an angle α will result in a rotation matrix

$$\mathbf{R} = \begin{bmatrix} \cos\beta\cos\gamma & \cos\beta\sin\gamma & -\sin\beta \\ \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\cos\beta \\ \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\cos\beta \end{bmatrix}$$

and

$$\mathbf{x}'''_P = \mathbf{R}\,\mathbf{x}_P = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma)\,\mathbf{x}_P$$

ECEF to ENU

We will use the results given above to determine a transformation of the coordinates of a point (say, a satellite) from the ECEF coordinate frame to a local coordinate frame defined at the user position. Such a coordinate frame is called a *local-level system* (LLS) or *east-north-up* (ENU) system. We will define the ENU coordinate frame (ENU for short) at the user position. As expected, the coordinate axes of ENU are oriented so that 1-axis points east, 2-axis points north, and the 3-axis points upward (see Figure 4.A.3). Suppose the user's ECEF XYZ coordinates are represented as components of a vector $\mathbf{x}_0$, and the geodetic coordinates are (ϕ, λ, h). Let the satellite position in ECEF be $\mathbf{x}_s$. The user-to-satellite vector in ECEF is $(\mathbf{x}_s - \mathbf{x}_0)$. We want to represent this vector in ENU and determine the azimuth and zenith angle of the satellite (Figure 4.A.4).

In order to determine the transformation from the ECEF to ENU, we have to determine the rotations of the ECEF coordinate axes that would bring them into coincidence with the axes of ENU. Refer to Figure 4.A.3. The first rotation is about the z-axis by an angle $(\lambda + 90°)$. The second rotation is about the x-axis (which is now parallel to the east-axis after the first rotation) by an angle $(90° - \phi)$. The corresponding transformation of the position vector $(\mathbf{x}_s - \mathbf{x}_0)$ from ECEF to ENU is obtained from the elementary rotation matrices as follows. We use subscript L (for local) rather than the unwieldy ENU.

$$\mathbf{R}_L = \mathbf{R}_1(90° - \phi)\,\mathbf{R}_3(\lambda + 90°)$$
$$= \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{bmatrix}$$

Let $\mathbf{x}_L = (x_E, x_N, x_U)$ be the representation of $(\mathbf{x}_s - \mathbf{x}_0)$ in ENU, then

$$\mathbf{x}_L = \mathbf{R}_L(\mathbf{x}_s - \mathbf{x}_0)$$

The elementary rotations are not unique. It is left as an exercise to show that $\mathbf{R}_L$ could also have been obtained with a different set of elementary rotations.

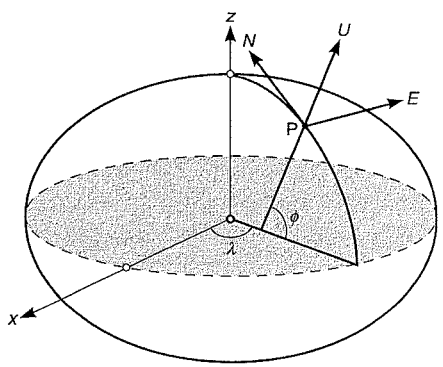


Figure 4.A.3 East-North-Up (ENU) coordinate system.

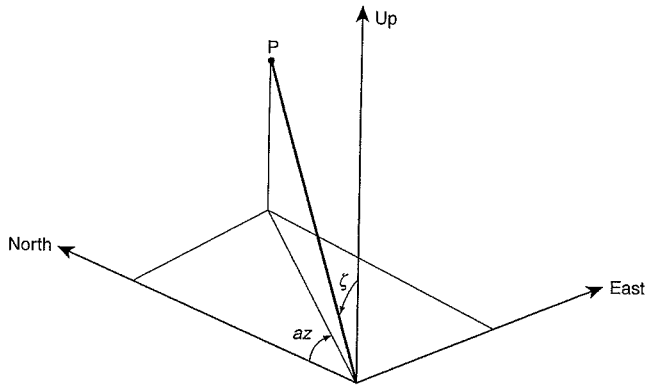


Figure 4.A.4 Azimuth (*az*) and zenith angle (ζ) defined in a local-level coordinate frame.

$$\mathbf{R}_L = \mathbf{R}_3(\pi)\mathbf{R}_2(\phi - \pi/2)\mathbf{R}_3(\lambda - \pi)$$

The transformation matrix $\mathbf{R}_L$ is orthogonal. Note also that each row vector is a representation of the unit vector along the axes of ENU frame expressed in the ECEF frame. (It is much easier to check this fact in the elementary rotation matrices.) Representing unit vectors along the east-, north-, and up-axes as **e**, **n**, and **u**, respectively,

$$\mathbf{e} = \begin{bmatrix} -\sin\lambda \\ \cos\lambda \\ 0 \end{bmatrix} \qquad \mathbf{n} = \begin{bmatrix} -\sin\phi\cos\lambda \\ -\sin\phi\sin\lambda \\ \cos\phi \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} \cos\phi\cos\lambda \\ \cos\phi\sin\lambda \\ \sin\phi \end{bmatrix}$$

We can now obtain the azimuth (*az*) and zenith angle (ζ) of the satellite, defined in the ENU coordinate frame (see Figure 4.A.4). The azimuth is measured from the north (0° to 360°), and is positive eastward (i.e., clockwise). The zenith angle is measured from the local vertical (0° to 90°). The elevation angle (*el*) is measured from the local horizontal (positive up), $el = (90° - \zeta)$.

$$\tan az = \frac{x_E}{x_N},$$

$$\sin el = \cos\zeta = \frac{x_U}{\sqrt{x_E^2 + x_N^2 + x_U^2}}$$

The position of a satellite with respect to a user can be specified with azimuth and elevation angle (or, azimuth and zenith angle).

Homework Problems

4-1. A seventeenth century navigator would have been pleased to estimate his longitude with an error of 1°. At your location, what would have been the resultant position error in meters? First, try the simple spherical model of the earth with radius of 6371 km. Next, use the WGS 84 ellipsoidal model (Table 4.1). Note that 6371 km is the approximate radius of a sphere whose volume equals that of the WGS 84 ellipsoid.

4-2. (Courtesy of Dr. Keith Alter) The Palo Alto Municipal Airport has a single runway. The locations defining the start and end of Runway 30 (rwy30) have been surveyed using GPS. The measured WGS 84 XYZ coordinates (in meters) are

$$\text{rwy30Start} = \begin{bmatrix} -2694685.473 \\ -4293642.366 \\ 3857878.924 \end{bmatrix} \qquad \text{rwy30End} = \begin{bmatrix} -2694892.460 \\ -4293083.225 \\ 3858353.437 \end{bmatrix}$$

These coordinates are believed to be accurate to within 1 cm. (You will find some of the MATLAB functions in folder (CD) *Navigation_Utilities* useful in solving this problem.)

(a) Determine the geodetic latitude, longitude, and altitude above the WGS 84 ellipsoid of the rwy30Start position. Provide the latitude and longitude in degrees, and the altitude in meters. Make sure you include enough significant digits to capture the accuracy of the XYZ coordinates.

(b) Determine the latitude and longitude of the rwy30Start position in degrees, arc-minutes, and arc-seconds.

(c) Given that in the vicinity of Palo Alto the geoid is 33 meters below the WGS 84 reference ellipsoid, what is the height of the rwy30Start position above the geoid? This height is referred to as the orthometric height or 'height above mean sea level (MSL).' The FAA-published height of rwy30Start is 3 feet MSL. Compare this with the height you just calculated.

(d) Define an east-north-up (ENU) coordinate system with rwy30Start as the origin (or reference point). What are the ENU coordinates of the rwy30End in meters?

(e) Calculate the length of Runway 30 in feet. Compare this with the published runway length of 2500 feet?

(f) Calculate the true heading of Runway 30 (90° is true east, 180° is true south, 270° is true west, and 360° is true north). Compare your result with the approximate runway true heading of 315° (magnetic heading of 300° from the runway designation, plus 15° magnetic variation). (Palo Alto is expected to re-designate this runway as Runway 31.)

(g) Calculate the runway gradient in the ENU coordinate frame assuming a constant slope from one end to the other (the upslope or downslope from rwy30Start to rwy-30End) in degrees. Compare this gradient with the published gradient of 0.0°.

4-3. Show that the time kept on the basis of a frequency source with a normalized error $\Delta f/f$ will accumulate an error of Δt over a period T such that

$$\frac{\Delta f}{f} = -\frac{\Delta t}{T}$$

4-4. Show that clocks with long-term stability of one part in 10^{13} would keep pace with each other within about six-millionths of a second over a year's time.

4-5. Convert the epoch 08:00:00 on 28 July 2001 Eastern Standard Time to: UTC, Julian Date, Modified Julian Date, TAI, and GPST (expressed as number of GPS weeks and GPS seconds into the week).

4-6. According to Kepler's third law, the orbital period of a planet depends only upon its mean distance from the sun. The average value of the radius vector for an elliptical orbit, however, depends upon how you define the average. Show that the average value is:

(i) a, if averaged over eccentric anomaly (or, weighing each point on the ellipse equally), as in (4.13), as Kepler apparently meant,

(ii) $b = \sqrt{1-e^2}$, if averaged over true anomaly in (4.11),

(iii) $a(1+e^2/2)$, if averaged over time (i.e., weighing each point on the ellipse by the reciprocal of the planet's speed at that point per Kepler's Second Law).

4-7. Consider two satellites S1 and S2 in circular orbits with periods of 8 hours and 16 hours, respectively. How high are the satellites above the earth's surface? (Use the spherical model of the earth.)

4-8. A satellite is circling the earth in a four-hour orbit with eccentricity 0.5.

(a) Find the smallest and largest *altitude* of the satellite above the earth. (Use the spherical model of the earth.) Does this seem to you a precarious orbit to use?

(b) Determine the angle swept by the line joining the satellite to the earth's center in the half-hour before its perigee?

(c) Determine the angle swept by the satellite in the half-hour after its apogee?

(d) Compare the angles from (b) and (c) with those for a circular four-hour orbit.

(e) Find the true anomaly of the satellite one hour after it passes its perigee.

4-9. Assuming the orbits of GPS satellites are circular, show that for a stationary user on the earth, range rate < 1 km/s and range acceleration < 0.2 m/s^2. Show that the corresponding Doppler frequency shift < 5 kHz, and Doppler frequency rate < 1 Hz/s. Hint: The range rate is zero for a satellite at the zenith and maximum (in magnitude) for a satellite rising or setting perpendicular to the horizon. For range acceleration, it's the opposite.

4-10. A receiver is located somewhere along the ground track of a GPS satellite. The SV is at an elevation of 60°. Define three mutually orthogonal directions at the SV position in its circular orbit as radial (along the line joining the SV position to the center of the earth), along-track (along the tangent to the orbit), and cross-track. Ephemeris errors are often specified in terms of their radial, along-track, and cross-track components. Consider an error of ε meters in the broadcast ephemeris along each direction, one at a time, and compute the corresponding range measurement error. Are the three components equally damaging in user position estimation?

4-11. Compute satellite position at a given epoch based on the parameter values broadcast in the navigation message and the algorithm described in the IS [(CD) *Documents\IS-GPS-200D.pdf*, Table 20-IV, pp. 98]. Write a function for which the input variables are the ephemeris parameters and time, and the output variable is the 3-D position of the satellite.

For this exercise, we'll work with the files (CD) *Data\Original\rcvr.dat* and *eph.dat*, which contain GPS data corresponding to a single epoch as described in Appendix A. (MATLAB Hint: Use `load rcvr.dat` and `load eph.dat` to load the files into two corresponding matrices called *rcvr* and *eph*.)

Matrix *rcvr* contains pseudorange measurements from all SVs and their common measurement time. From this information we can derive signal transmission time from each satellite as follows.

$$t^{(k)} = t_u - \rho^{(k)}/c$$

where $t^{(k)}$ is the time of transmission from the kth satellite, t_u is the time of signal reception, $\rho^{(k)}$ is the measured pseudorange to satellite k; and c is the speed of light. This equation gives us the epoch at which we want to compute the position of the kth satellite. (The signal transmission time obtained from the above equation is in accordance with the receiver clock. We need the time of transmission in GPS Time, which means taking into account the receiver clock bias and satellite clock bias. We will disregard this complication for now.) Verify that the position of PRN 6 (in meters) is:

$$\mathbf{x}^{(6)} = (-8{,}087{,}137.8 \quad -16{,}946{,}009.4 \quad 18{,}816{,}191.8)^T$$

4-12. As we'll see in Chapter 6, we need GPS satellite velocity expressed in ECEF in order to estimate the user velocity. The IS gives a step-by-step method for computing satellite position (see previous exercise) but not velocity. Determine satellite velocity in ECEF in terms of the broadcast ephemeris parameter values in files (CD) *Data\Original\rcvr.*

dat and *eph.dat* used in the previous exercise. (Hint: Equation (4.17) gives the satellite velocity in an inertial frame. Transform it into ECEF. Note that unlike the corresponding transformation for the instantaneous position, this transformation now has to account for a velocity term associated with the rotating ECEF.)

4-13. Determine the satellite clock bias for all SVs in the file *eph.dat*. Write a function that returns broadcast satellite clock bias (in meters) based on the time of transmission and the broadcast values of parameters a_{f0}, a_{f1}, and a_{f2}. See (CD) *Documents\IS-GPS-200D .pdf*, pp. 88–89. The approximate time of transmission computed in Problem 4-11 is adequate for our purpose here. (Answer: PRN 5 clock bias = 56,679.92 meters.)

4-14. Continuing with the data files *rcvr.dat* and *eph.dat*:

(a) These measurements were taken close to the position of rwy30Start given in Problem 4-2 above. Determine the elevation angle and the azimuth (see Appendix 4.A for the definitions) of each satellite. This would require you to establish an ENU coordinate system at the user location, and determine the ENU coordinates of each satellite given the XYZ position. Write a function that returns the satellite azimuth and elevation angle based on the estimated location of the user and the calculated location of the satellite. (Hints: You will find some of the MATLAB functions in the folder (CD) *Navigation _Utilities* useful. Use the MATLAB function *atan2* in computing azimuth. For PRN 6, elevation = 78.1°, azimuth = 30.9°.)

(b) Low-elevation satellites often provide very noisy range measurements, and users often set up an elevation mask angle (typically 5°–7.5°) and drop any satellites that are lower. Determine which satellites, if any, are below 7.5° in elevation.

4-15. Analyze the effect of the non-central perturbations on satellite motion by calculating satellite positions with and without the perturbation parameters. Download file (CD) *Data\Stanford\January_6_2000\e010600a.dat*. Select a satellite which was in view for a period of four hours or longer. Compute the SV position at fifteen-minute intervals starting on the hour using all or a subset of the ephemeris parameters.

(a) Compute the satellite positions using the full set of parameters broadcast by the satellite.

(b) Compute the satellite positions disregarding the harmonic correction terms.

(c) Compute the satellite positions disregarding both the secular and harmonic correction terms.

(d) Compare the satellite positions in (a)-(c) with the post-processed, precise ephemeris generated by IGS, (CD) *Data\Precise_ephemeris\Parsed\ ig010600.spp*, and plot the magnitude of the satellite position error as a function of time for each case. Are there any surprises?

References

Allan, David W., Neil Ashby, and Clifford C. Hodge (1997). *The Science of Timekeeping*, Application Note 1289, Hewlett-Packard Corporation.

Ashby, Neil (2002). Relativity and Global Positioning System, *Physics Today*, May 2002, pp. 41–47.

Ashkenazi, V. (1986). Coordinate Systems: How to Get Your Position Very Precise and Completely Wrong, *The Journal of Navigation*, vol. 39, no. 3, pp. 269–278.

Bate, Roger R., Donald D. Mueller, and Jerry E. White (1971). *Fundamentals of Astrodynamics*, Dover Publications.

Beutler, Gerhard *et al.* (1998). GPS Satellite Orbits, in *GPS for Geodesy*, 2nd edition, P. J. G. Teunissen and A. Kleusberg (eds.), Springer, Lecture Notes on Earth Sciences, pp. 43–106.

Blair, Byron E., (ed.) (1974). *Time and Frequency: Theory and Fundamentals*, U.S. Department of Commerce, National Bureau of Standards Monograph 140.

Bock, Yehuda (1996). Reference Systems, in *GPS for Geodesy*, A. Kleusberg and P.J.G. Teunissen (eds.), Springer, Lecture Notes on Earth Sciences, pp. 3–37.

Boucher, Claude, and Zuheir Altamimi (1996). International Terrestrial Reference Frame, *GPS World*, vol. 7, no. 9, pp. 71–74.

Brown, Lloyd A. (1977). *The Story of Maps*, Dover Publications.

DMA (1984). *Geodesy for the Layman*, Defense Mapping Agency, TR 80-003.

Featherstone, Will, and Richard B. Langley (1997). Coordinates and Datums and Maps! Oh My! *GPS World*, vol. 8, no. 1, pp. 34-41.

IS-GPS-200D (2004). Interface Specification IS-GPS-200, Revision D, Navstar GPS Space Segment/ Navigation User Interfaces, Navstar GPS Joint Program Office [CD *Documents\IS-GPS-200D.pdf*].

Kitching, John *et al.* (2005). Chip Scale Atomic Frequency References, *Proc. ION GNSS 2005*, pp. 1662–1669.

Massatt, Paul and Michael Zeitzew, (1998). GPS Constellation Design—Current and Projected, *Proc. ION National Technical Meeting*, pp. 435–445.

Merrigan, Michael J., Everett R. Swift, Robert F. Wong, and Joedy T. Saffel (2002). A Refinement to the World Geodetic System 1984 Reference Frame, *Proc. ION GPS-2002*, pp. 1519–1529.

Montenbruck, Oliver, and Eberhard Gill (2001). *Satellite Orbits: Models, Methods, and Applications*, Springer.

NIMA (1997). *Department of Defense World Geodetic System 1984*, National Imagery and Mapping Agency, TR8350.2, Third Edition. [CD *Documents\WGS 84.pdf*]

Seeber, Günter (2003). *Satellite Geodesy*, 2nd edition, Walter de Gruyter.

Snyder, John P. (1987). Map Projections—A Working Manual, U.S. Geological Survey Professional Paper 1395, U.S. Government Printing Office.

SPS (2001). *Global Positioning System Standard Positioning Service Signal Specification*, U.S. Department of Defense. [CD *Documents\SPS Performance 2001.pdf*]

Van Dierendonck, A.J., S.S. Russel, E.R. Kopitzke, and M. Birnbaum (1980). The GPS Navigation Message, in *Global Positioning System*, vol. I, The Institute of Navigation, pp. 53–73.

Vaníček, P., and E.J. Krakiwsky (1986). *Geodesy: The Concepts*, 2nd edition, Elsevier Science Publishers.

PART II

Estimation of Position, Velocity and Time

Accuracy of the estimates of position, velocity and time (PVT) obtained from GPS can vary widely with time, place and—most important—a user's resources. The SPS performance specifications (e.g., the oft-quoted x-meter horizontal accuracy better than)5% of the time, where the value of x was 100 in the days of SA and is about 10 now) are a :onservative estimate of the GPS performance available in real time with a minimum-capability receiver operating in an autonomous mode. A resourceful user can do much better.

Since WAAS became operational in 2003, most receivers sold in the United States, even he $100 pocket receivers, are WAAS-enabled and provide an accuracy of 1–5 m. Similar ac-:uracy will soon be available in Europe and Asia as EGNOS and MSAS come on-line. Better ıccuracy is available from local-area DGPS corrections, now widely available, for users within ens of kilometers from a reference station. A user without access to differential corrections :an compensate for the ionospheric delay, now the largest source of error, with a dual frequency (L1–L2) receiver.

Millimeter-to-centimeter-level positioning accuracy has been available for twenty years ıy post-processing carrier phase measurements. Centimeter-to-decimeter-level accuracy is ıow being achieved in real time by users on the move who can stand still for a minute or two, f necessary.

In Part II we take a close look at the measurements provided by GPS and the algorithms or processing these measurements to obtain PVT estimates. In Chapter 5, we discuss the measurements, error sources, and error mitigation. Chapter 6 deals with algorithms for estimation ıf PVT based on measurements of pseudoranges and pseudorange rates. In Chapter 7, we examine precise relative positioning based on carrier phase measurements.

Chapter 5

GPS Measurements and Error Sources

GPS provides two types of measurements. Code tracking provides estimate of apparent transit time of a signal. Carrier phase tracking provides measurement of the received carrier phase relative to the phase of a sinusoidal signal generated by the receiver clock. Both are biased (and, in case of carrier phase, ambiguous) estimates of the instantaneous user-satellite range. How these measurements are formed in a receiver is discussed in Chapter 12 and in a comprehensive paper by Ward (1996).

In this chapter, we examine the measurements and develop simple mathematical models to relate them to receiver-satellite ranges and range rates. This process requires us to think about the sources of errors in these models. Analysis of the errors and schemes for their alleviation take up the rest of the chapter. These models are used in the next chapter for estimation of position, velocity, and time (PVT).

In this chapter and the next two, we sometimes speak loosely of the user position or receiver position, but what we mean is the position of the receiver antenna. When dealing with centimeter- and millimeter-level position estimates in Chapter 7, we must be even more specific and refer to the electrical phase center of the antenna.

5.1 Measurement Models

5.1.1 Code Phase Measurements

A basic measurement made by a GPS receiver is the apparent transit time of the signal from a satellite to the receiver, defined as the difference between signal reception time, as determined by the receiver clock, and the transmission time at the satellite, as marked on the signal. It is measured as the amount of time shift (modulo 1 ms) required to align the C/A-code replica generated at the receiver with the signal received from the satellite. This measurement is biased due to the fact that the satellite and receiver clocks are not synchronized, and each keeps time independently. Each satellite generates its signals in accordance with a clock on board. The receiver generates a replica of each signal in accordance with its own clock. The corresponding biased range, or pseudorange, is defined as the transit time so measured multiplied by the speed of light in a vacuum.

There are three time scales to deal with: two of these are the times kept by the satellite and receiver clocks; the third is a common time reference, GPS Time (GPST), introduced in Section 4.2.4 as a composite time scale derived from the times kept by clocks at GPS monitor stations and aboard the satellites.

Let τ be the transit time associated with a specific code transition of the signal from a satellite received at time t per GPST. Let $t^s(t-\tau)$ be the corresponding emission time, and $t_u(t)$ be the arrival time measured by the user's receiver clock. The measured apparent range ρ, called pseudorange, is determined from the apparent transit time as

$$\rho(t) = c[t_u(t) - t^s(t-\tau)] \tag{5.1}$$

Both t and τ are unknown, and are to be estimated. The range to a satellite is about 20,000 km when overhead, and about 26,000 km when rising or setting. The signal transit time varies between about 70 ms and 90 ms.

In this section, we deal with measurements from a GPS satellite in a generic way, making no reference to the satellite ID or carrier frequency (L1 or L2). We employ superscript s to identify a term associated with a satellite, and subscript u to identify a term associated with the user or receiver, where needed for clarity. Additional notation will be introduced in later sections as needed to deal with dual-frequency measurements from multiple satellites.

The time scales of the receiver and the satellite clocks can be related to GPST as

$$t_u(t) = t + \delta t_u(t) \tag{5.2}$$

$$t^s(t-\tau) = (t-\tau) + \delta t^s(t-\tau) \tag{5.3}$$

where δt_u is the receiver clock bias and δt^s is the bias in the satellite clock, both measured relative to GPST, as shown in Figure 5.1. In our notation, both δt_u and δt^s reflect the amounts by which the satellite and receiver clocks are advanced in relation to GPST. The satellite clock bias (δt^s) is estimated by the Control Segment and specified in terms of the coefficients of a quadratic polynomial in time. The values of these coefficients are broadcast in the navigation message [Section 4.3.6].

Accounting for the clock biases, the measured pseudorange (5.1) can be written as

$$\begin{aligned}\rho(t) &= c[t + \delta t_u(t) - (t - \tau + \delta t^s(t-\tau))] + \varepsilon_\rho(t) \\ &= c\tau + c[\delta t_u(t) - \delta t^s(t-\tau)] + \varepsilon_\rho(t)\end{aligned} \tag{5.4}$$

Symbol ε is used throughout Part II to denote unmodeled effects, modeling errors, and measurement errors; subscripts, and occasionally overbar and tilde, are used to distinguish among the different scenarios. The transit time multiplied by the speed of light in a vacuum can be modeled as

$$c\tau = r(t, t-\tau) + I_\rho(t) + T_\rho(t) \tag{5.5}$$

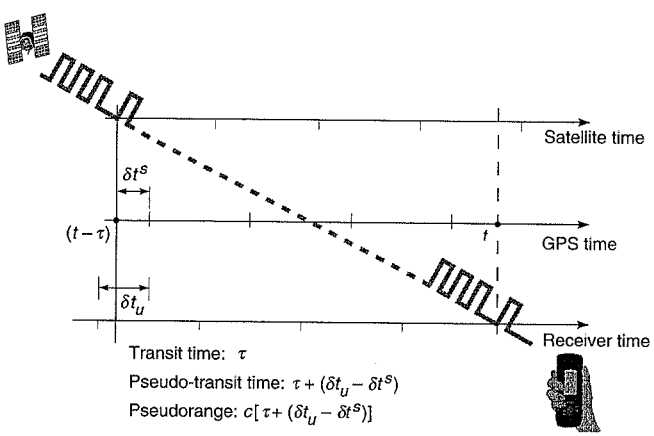


Figure 5.1 A conceptual view of the pseudorange measurements. The receiver and satellite clocks are unsynchronized. The apparent transit time (pseudo-transit time) is the difference between the signal reception time according to the receiver clock and the transmission time imprinted upon the signal in accordance with the satellite clock.

where $r(t, t-\tau)$ is the geometric (or, true) range between the user position at time t and the satellite position at $(t-\tau)$. I_ρ and T_ρ reflect the delays associated with the transmission of the signal through the ionosphere and the troposphere, respectively. Both terms are positive, and we will discuss both later in this chapter. For simplicity, we drop explicit reference to the measurement epoch t, and rewrite the model for the measured pseudorange as

$$\rho = r + c[\delta t_u - \delta t^s] + I_\rho + T_\rho + \varepsilon_\rho \tag{5.6}$$

Ideally, we would have liked to measure r, the true range to the satellite. What we have instead is ρ, the pseudorange, a biased and noisy measurement of r. How accurate an estimate of position, velocity, or time we obtain from these measurements would depend upon our ability to compensate for, or eliminate, the biases and errors. Equation (5.6) is basic, and we will return to it again and again.

The receiver clocks are generally basic quartz crystal oscillators, and tend to drift. The receiver manufacturers attempt to limit the deviation of the receiver clock from GPST, and schedule simultaneous measurements from all satellites typically at 1 Hz at epochs that are within ± 1 ms of the GPST seconds. One approach to maintaining the receiver clock within a certain range of GPST is to steer the receiver clock 'continuously.' Such steering is implemented in software. The second approach is to let the clock drift until it reaches a certain threshold (typically, 1 ms), and then reset it with a jump to return the bias to zero. How a receiver keeps track of the bias between its clock and GPST is discussed in Chapter 6. An example of pseudorange measurements with a receiver using the second approach is shown in Figure 5.2.

Figure 5.2 shows pseudorange measurements simultaneously from three satellites which rose about the same time but were in different orbits. One came overhead and stayed in view for almost seven hours. The other two stayed lower in the sky and were seen for shorter periods. The discontinuities common to all three sets of measurements are due to the resetting of the receiver clock. It is left as an exercise to determine if the receiver clock is running fast or slow, and to estimate its frequency offset from the nominal value of 10.23 MHz.

Constructing Pseudorange Measurements

The measurement of pseudoranges is scheduled in accordance with the receiver clock. In (5.1), determining $t_u(t)$ is straightforward, but $t^s(t-\tau)$ is a little tricky. We now resume the discussion we began toward the end of Section 4.3.6. The measurement epoch would typically occur in the middle of a C/A-code chip. The corresponding transmission time per satellite clock is reconstructed as follows in a process akin to determining time from an analog clock—by reading the hour hand, the minute hand, and the second hand.

Recall from Section 4.3.6 that the satellite time is transmitted in the navigation message in the form of Z-count, which increments in units of 1.5 seconds, and is specified at the beginning of each subframe. The Z-counts associated with two successive subframes differ by four (i.e., six seconds). In order to determine the satellite clock time associated with the current measurement epoch, we need the Z-count associated with the *current* subframe, plus the satellite time elapsed since the beginning of the subframe. This elapsed time can be measured in terms of its components as follows: the whole number of navigation message data bits transmitted in the current subframe, the whole number of C/A-code periods since the beginning of the current navigation message bit, the number of whole chips in the current C/A-code cycle, and the fraction of the current C/A-code chip (see Figure 5.3).

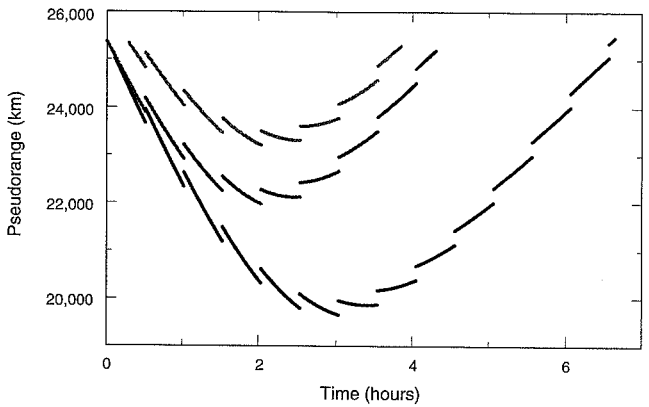


Figure 5.2 Pseudorange measurements from three satellites obtained with a stationary receiver. The variation in the pseudorange is mainly due to change in geometric range resulting from the satellite motion and rotation of the earth. The discontinuities are common to all measurements, and are due to 1-millisecond jumps in the receiver clock. (Some receivers 'steer' the receiver clock 'continuously' to control its offset from GPS Time.)

$$\begin{aligned} t^s(t-\tau) = {} & \text{Z-count} \times 1.5 \\ & + \text{number of navigation data bits transmitted} \times 20 \times 10^{-3} \\ & + \text{number of C/A-code repeats} \times 10^{-3} \\ & + \text{number of whole C/A-code chips}/(1.023 \times 10^6) \\ & + \text{fraction of a C/A-code chip}/(1.023 \times 10^6) \text{ seconds} \end{aligned}$$

The Z-count is read off the navigation data. The number of navigation data bits in the subframe transmitted and the number of C/A-code periods in the current navigation data bit are counted by the receiver software. As we'll see in Chapter 12, the delay lock loop (DLL) provides the last two terms.

5.1.2 Carrier Phase Measurements

A measurement much more precise than that of code phase is the phase of the carrier received from a satellite. The carrier phase measurement by a GPS receiver, also called carrier beat phase measurement, is the difference between the phases of the receiver-generated carrier signal and the carrier received from a satellite at the instant of the measurement. As we'll see, this measurement is an indirect and ambiguous measurement of the signal transit time. There are some subtleties associated with the carrier phase measurements. We begin with basic definitions and deal first with simple, idealized scenarios

We measure carrier phase in terms of the number of cycles generated or received since the starting point of an interval. The phase at time t is defined as

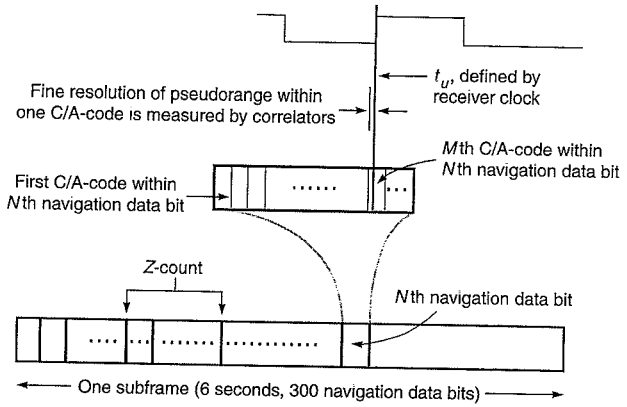


Figure 5.3 Pseudorange determination simplified. The transmission time per satellite clock is reconstructed in a process similar to reading time from an analog clock—by reading the hour hand, the minute hand, and the second hand.

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} f(s)\,ds$$

where $f(s)$ is the time-varying frequency and $\phi(t_0)$ is the initial phase offset. In this model, the timer used to measure the epochs t_0 and t is assumed to be perfect. If the interval $(t - t_0)$ is 'short' and the signal generator is 'highly stable,' we can write

$$\phi(t) = \phi(t_0) + f \cdot (t - t_0)$$

where f is the instantaneous frequency. But what if we were measuring time intervals by counting these very cycles generated by an (imperfect) oscillator with nominal frequency specified as f_0? In this case

$$\phi(t) = \phi(t_0) + f_0 \cdot (t - t_0)$$

This is how we'll model the phase of the carrier generated by the receiver clock, which is also our timer.

Consider first an idealized case of error-free measurements with perfect and synchronized satellite and receiver clocks, and no relative motion between the satellite and the user. In this model, the carrier phase measurement (i.e., the difference between the phases of the receiver generated carrier and the received carrier) would remain fixed at a fraction of a cycle, and the distance between a satellite and the receiver would be an unknown number of whole cycles plus the measured fractional cycle. This measurement, however, contains no information regarding the number of whole cycles, referred to as *integer ambiguity*. Now suppose that the carrier phase is tracked while the receiver or the satellite moves so that the distance between them grows by a wavelength. The corresponding carrier phase measurement would now be a full cycle plus the fractional-cycle phase measured before the movement began.

To measure carrier phase in GPS, a receiver acquires phase lock with the satellite signal, measures the initial fractional phase difference between the received and receiver-generated signals, and from then on tracks changes in this measurement, counting full cycles and keeping track of the fractional cycle at each measurement epoch.

An important idea of this section is that the phase of the received signal at any instant can be related to the phase at the satellite at the time of transmission by the transit time of the signal, and it's τ we are after. Staying with our idealized case of perfect and synchronized clocks, the carrier phase measurement in units of cycles can be written as

$$\phi(t) = \phi_u(t) - \phi^s(t - \tau) + N \tag{5.7}$$

where $\phi_u(t)$ is the phase of the receiver-generated signal; $\phi^s(t - \tau)$ is the phase of the signal received from the satellite at time t, or the phase of the signal at the satellite at time $(t - \tau)$; τ is the transit time of the signal; and N is the integer ambiguity. Estimation of N is referred to as *integer ambiguity resolution.*

Simplifying the above expression by writing

$$\phi^s(t - \tau) = \phi^s(t) - f \cdot \tau$$

we obtain

$$\begin{aligned} \phi(t) &= f \cdot \tau + N \\ &= \frac{r(t, t-\tau)}{\lambda} + N \end{aligned} \tag{5.8}$$

where f and λ are the carrier frequency and wavelength, respectively, and $r(t, t - \tau)$, as before, is the geometric range between the user position at time t and the satellite position at $(t - \tau)$. As in the previous section, we are dealing with the measurements at a receiver from a single satellite, and we identify terms with the receiver or the satellite, if needed for clarity.

Now we account for the clock biases, initial phase offsets, atmospheric propagation delays, and measurement errors. Dropping explicit reference to the measurement epoch, (5.8) can be rewritten as

$$\phi = \lambda^{-1}[r + I_\phi + T_\phi] + \frac{c}{\lambda}(\delta t_u - \delta t^s) + N + \varepsilon_\phi \tag{5.9}$$

where I_ϕ and T_ϕ are the ionospheric and tropospheric propagation delays in meters, respectively; and c is the speed of light in a vacuum. Note that the carrier phase measurement in (5.9) is in units of cycles. The integer term N could have gone on either side of the equation. We keep it on the right and leave its sign unspecified.

The change in carrier phase measurement over a time interval corresponds to changes in both the user-satellite range and receiver clock bias, and is referred to as integrated Doppler or *delta pseudorange*. The rate of change of the carrier phase measurement gives the *pseudorange rate*, which is made up of the actual range rate plus the receiver clock frequency bias.

Equation (5.9) appears similar to that for the pseudorange measurements based on code tracking (5.6). Both the code and carrier phase measurements are corrupted by the same error sources, but there is an important difference. Code tracking provides essentially unambiguous pseudoranges which, as we'll see, are coarse by comparison. The carrier phase measurements are extremely precise, but are encumbered with integer ambiguities. The integers remain con-

stant as long as the carrier tracking loop maintains lock. Any break in tracking, no matter how short, could change the integer values.

In order to take full advantage of the precision of these measurements to obtain accurate position estimates, we have to compensate for the various errors and resolve the integer ambiguities. As will be clear following the discussion of measurement errors in the next section, we cannot expect to determine the value of integer N in (5.9) in real time, and we'd have to look for alternative approaches to take advantage of the precision of these measurements.

One way to get at least a partial benefit of the precise carrier phase measurements without being bogged down by the integer ambiguities is via delta pseudoranges obtained from the change in carrier phase measurement over a time interval. From (5.9), the change in carrier phase measurements between time instants t_0 and t_1 is

$$\phi(t_1)-\phi(t_0)\approx\lambda^{-1}[r(t_1)-r(t_0)]+\frac{c}{\lambda}[\delta t_u(t_1)-\delta t_u(t_0)]+\tilde{\varepsilon}_\phi$$

The integer ambiguity term drops out if the carrier is tracked continuously between t_0 and t_1. The error in the above representation is related to the changes during the observation interval in the satellite clock bias and the ionospheric and tropospheric propagation delays. This idea is discussed further in Section 5.7.

5.1.3 An Instructive Model for the Code and Carrier Measurements

We can think of the code and carrier phase measurements as though obtained with a measuring tape. In this conceptual exercise suggested by van Diggelen (1997), one end of the tape measure is attached to the satellite and you are holding the other end, including the spool, allowing the tape to roll forward or back freely as you and the satellite move. The difference in the two types of measurements can be illustrated by the different markings on the tape measure, as shown in Figure 5.4.

For code measurements, the tape has coarse tick marks with resolution of one meter, and the tick marks are labeled from zero to 26,560,000 m to indicate the distance to a satellite. At a measurement epoch, you simply look at the tape and read off the range. Actually, the code measurements entail little conceptual difficulty and this analogy was hardly necessary. So, let's move on to the carrier phase measurements.

The tape used for the carrier phase measurements is peculiar: It has very fine tick marks in units of cycles with resolution of 0.01 cycle, but no labels (1 cycle ≈ 19 cm at L1, and 24 cm at L2). At a measurement epoch, when you look at the tape all you can make out is a partial cycle. You know that the range to the satellite is a certain (large) number of whole cycles plus the observed partial cycle. While this partial cycle can be measured very precisely, you have no idea of the number of whole cycles between you and the satellite. The situation repeats at the next measurement epoch.

How to take advantage of these very fine but ambiguous measurements? There is a way. The trick is not to take your eyes off the tape between measurement epochs. At the first measurement epoch, make a note of the partial cycle. From that point on, count the number of whole cycle marks as they go by. At each successive measurement epoch, record the change in range in terms of the number of whole cycles and a partial cycle. Each of these measurements, when combined with the initial unknown whole number of cycles, represents range to the satellite at that instant. A collection of such measurements over a 'short' period from all

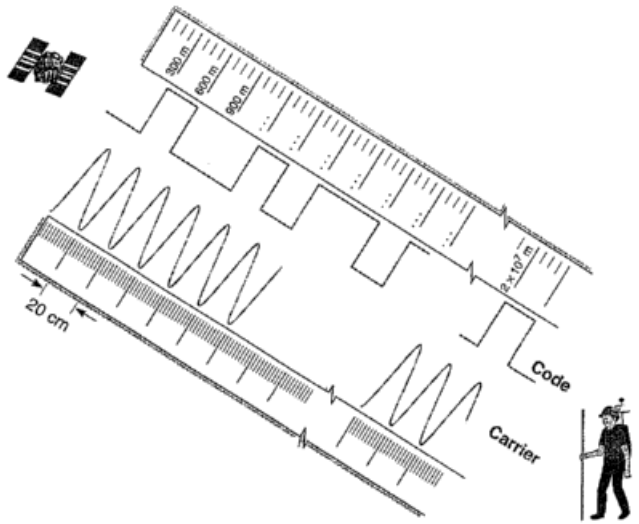


Figure 5.4 A conceptual representation of the code and carrier phase measurements with a tape measure.

satellites in view can be used to estimate the unknown whole cycles, and determine your position very precisely, as we'll discuss in Chapter 6. The problem is that if you took your eyes off the tape, or even blinked, you would miss the count of whole cycles, and will have to start all over again.

5.1.4 Error Sources and Models

Measurement errors are often categorized as *noise* or *bias*. Noise generally refers to a quickly varying error that averages out to zero over a 'short' time interval, where short is defined in relation to the integration time or smoothing time of a receiver [Chapter 11]. A bias tends to persist over a period of time. This distinction is helpful in understanding the effects of the measurement errors, and we'll return to this point in Section 5.5 after we have examined the errors in measurement of pseudoranges in the next section. The errors can be grouped as follows:

- errors in the parameter values broadcast by a satellite in its navigation message for which the Control Segment is responsible,
- uncertainties associated with the propagation medium which affect the travel time of the signal from a satellite to the receiver,
- receiver noise which affects the precision of a measurement, and interference from signals reflected from surfaces in the vicinity of the antenna.

We examine these error sources below. Our objective is to understand the nature and the size of the pseudorange error introduced by each source and analyze the efficacy of different approaches to error mitigation.

5.2. Control Segment Errors: Satellite Clock and Ephemeris

The ephemeris and clock parameter values broadcast by the satellites are computed by the Control Segment on the basis of measurements at GPS monitor stations. The current values of these parameters are obtained from a Kalman filter used to generate the states of the satellites (position and velocity) and their clocks (phase bias, frequency bias, and frequency drift rate). A prediction model is then used to generate the ephemeris and clock parameters to be uploaded to the satellites and broadcast by them in the 50 bps navigation message. There are errors associated with both the estimation of the current values of the parameters, and prediction of their values for the future. The prediction error grows with the age of data (AoD), defined as the time since the last parameter upload. Clearly, the more accurate are the models used to estimate and predict the ephemeris and clock parameters, and the more frequent the data uploads to the satellites, the lower the Control Segment errors.

The rms ranging error attributed to the clock and ephemeris parameter errors was limited to 6 m for the Precise Positioning Service (PPS) by specification. Actually, the frequency standards aboard the satellites have consistently exceeded their specifications, and the ephemeris prediction errors have been kept low by frequent uploads. As a result, Block II/IIA and IIR satellites beat the specification, limiting the Control Segment error in recent years to under 3 m rms. The specification has now been tightened to 3 m for Block IIF.

The ephemeris error is usually decomposed into components along three orthogonal directions defined relative to the satellite orbit: radial, along-track and cross-track (see Figure 5.5). In estimation of an orbit based on range measurements, the radial component of the ephemeris error tends to be the smallest. The along-track and cross-track components can be several times larger. This is fortunate because, as shown in Figure 5.5, the error in a pseudorange measurement is the projection of the satellite position error vector on the satellite-receiver line of sight, which depends mostly upon the radial component of the ephemeris error. The components of the along-track and cross-track errors along the line of sight are small.

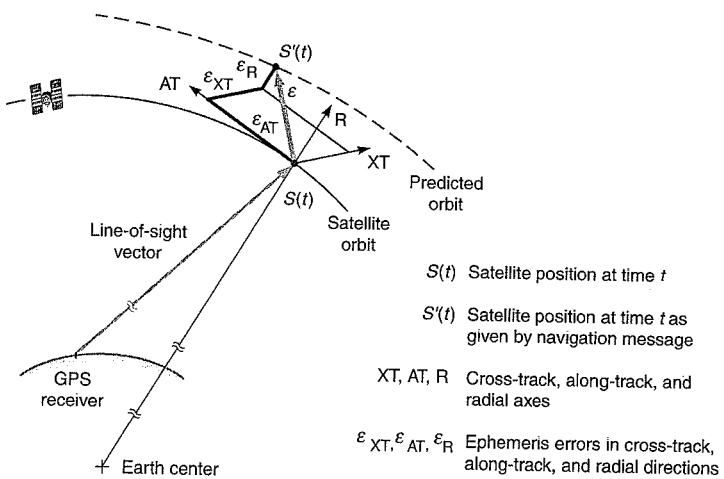


Figure 5.5 Ephemeris error components in the along-track (AT), cross-track (XT), and radial (R) directions. The radial component is the principal source of error in the range measurement. The components of the along-track and cross-track errors along the line of sight are small.

The range error due to the errors in the clock and ephemeris parameters is defined as the root-sum-square (rss) value of the clock error and the line-of-sight component of the ephemeris error. The size of this error is estimated and tracked by the Control Segment in real time within 1 m rms. With typical once-a-day data uploads, the current estimates of the rms range errors due to the ephemeris and clock parameters are about 1.5 m each. (The 3-D ephemeris error over a day is typically 3–5 m rms.)

The Control Segment monitors the growth in parameter errors by comparing the broadcast values to the best current estimates available. If the estimated range error for a satellite exceeds a threshold, a 'contingency data upload' is scheduled. (The threshold was revised in 1997 from 8 m to 5 m.) New initiatives would reduce the navigation parameter errors further [Malys *et al.* (1997)].

The Block IIR and IIF satellites are capable of operating in Autonav mode while out of contact with the Control Segment over extended periods, though this function currently appears to be inoperative. In Autonav mode, cross-link ranging among the satellites would be used to correct the ephemerides stored in the on-board computers. The Block IIF satellites are planned to maintain the clock and ephemeris errors below 3 m for up to sixty days out of contact with the Control Segment, not counting the error due to polar motion [Section 4.1] and changes in UT1 [Section 4.2], which cannot be estimated in Autonav mode. In normal mode of operation, the IIFs would limit AoD to three hours by uploading the ephemeris and clock data to all satellites via ultra-high-frequency cross links [Fisher and Ghassemi (1999)].

5.3 Signal Propagation Modeling Errors

5.3.1 Signal Refraction, Wave Propagation, and Dispersive Media

The GPS signals are affected by the medium through which they travel from the satellites to a receiver. The travel distance ranges from about 20,000 km when a satellite is overhead to about 26,000 km when it is rising or setting. All but the final 5% of the signal travel can be regarded as in a vacuum or free space, through which the electromagnetic signals travel with a constant speed c = 299,792,458 m/s, the well-known universal constant. Closer to the surface of the earth (Figure 5.6), at a height of about 1000 km, the signals enter an atmosphere of charged particles, called the ionosphere. Later, at a height of about 40 km, the signals encounter an electrically neutral gaseous atmosphere to be referred to as the troposphere.

The atmosphere changes the velocity (speed and direction) of propagation of radio signals. This phenomenon is referred to as *refraction*. The change in speed of propagation changes the signal transit time, which is the basic measurement from GPS. We take a slight detour in this section to review refraction and wave propagation from freshman physics.

The *refractive index* of a medium (n) is defined as the ratio of the speed of propagation of the signal in a vacuum (c) to the speed in the medium (v),

$$n = \frac{c}{v} \tag{5.10}$$

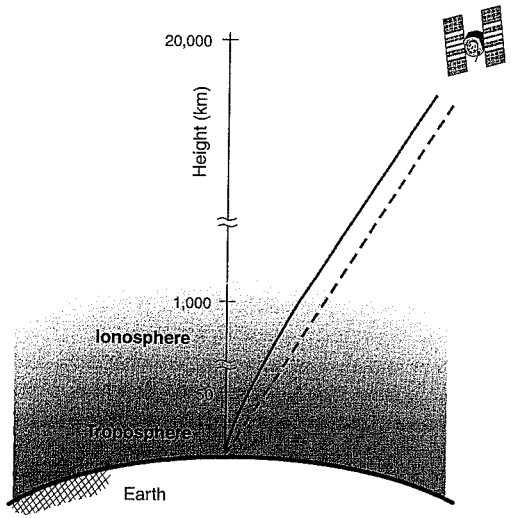


Figure 5.6 Refraction of GPS signals in the earth's atmosphere results in changes to both speed and direction. Increase in path length due to bending of the signal ray, greatly exaggerated above, is generally insignificant. The effect of the change in speed of propagation, however, can result in pseudorange measurement error of several meters or more.

Actually, the ionosphere and troposphere are not uniform in composition and the refractive index changes all along the path of a signal. Change in signal speed changes the travel time of the signal and, therefore, changes the apparent range to the satellite as computed from (5.1). Changing refractive index in a medium results in bending of the path of a signal ray (Snell's law), making the path longer than the geometrical straight-line path (Figure 5.6). The transit time along this curved path, however, is shorter than that for a straight-line path (Fermat's principle of least time). The effect of signal bending, however, is very small except at very low elevation angles and we will assume the signal paths to be straight lines.

Given the refractive index profile along the propagation path of a signal, the travel time

$$\tau = \frac{1}{c}\int_{\text{Satellite}}^{\text{Receiver}} n(l)\,dl \tag{5.11}$$

where the integration is along the signal propagation path, and $n(l)$ represents the (changing) refractive index of the medium. In the rest of this section, we will use 'R' for the receiver and 'S' for the satellite in the integration limits for simplicity, or ignore the limits altogether. We can write an expression for the excess delay in signal propagation due to refraction as

$$\Delta\tau = \frac{1}{c}\int_{S}^{R}[n(l)-1]dl \tag{5.12}$$

or, equivalently, effective increase in path length

$$\Delta\rho = \int_{S}^{R}[n(l)-1]dl\,. \tag{5.13}$$

If the refractive index of a medium depends upon the frequency of the signal, the medium is said to be *dispersive*. We know, for example, that a glass prism is dispersive for visible light, and so are water droplets hanging in the air. For radio signals in the L-band, the ionosphere is dispersive, but the troposphere is not.

In order to determine the excess travel time for the GPS signals as they propagate through the ionosphere and troposphere, we have to determine the refractive index along the propagation path. That is the main subject of Sections 5.3.2 and 5.3.3. But first we review wave propagation and set the stage for discussion of *code-carrier divergence*, an important phenomenon associated with the propagation of GPS signals through the ionosphere.

Let us consider a monochromatic sinusoidal wave moving in the x-direction.

$$s(x,t) = s_0\cos(\omega t - kx + \phi_0) \tag{5.14}$$

where s_0 is the amplitude of the wave; ω is the *circular frequency*; t is time; k is called the *wave number*, to be addressed presently; and ϕ_0 is the phase offset. Given below are the familiar, basic relationships among the parameters of wave motion: ω, k, frequency f, (temporal) period T, and wavelength (or, spatial period) λ.

$$\omega = 2\pi/T$$
$$k = 2\pi/\lambda$$
$$T = 1/f = 2\pi/\omega$$

Returning to (5.14), as t or x changes, so does the phase $(\omega t - kx + \phi_0)$ of the wave. The phase is constant if $(\omega t - kx)$ is constant, for example, if

$$x = \frac{\omega}{k}t = v_p t$$

where v_p is the *phase velocity* of the wave. (Actually, the correct term is phase speed, velocity being associated with a vector.)

$$v_p = \lambda f = \frac{\lambda}{T} = \frac{\omega}{k}$$

We can rewrite (5.14) as

$$s(x,t) = s_0\cos[\omega(t - x/v_p) + \phi_0] \tag{5.15}$$

As the wave moves, the whole sinusoidal pattern moves along the x-direction with speed v_p. The distance between the wave crests is the wavelength (or, spatial period) λ. The frequency f with which the wave crests pass a stationary observer is $f = v_p/\lambda$.

Now let us look at the propagation of a modulated signal through a dispersive medium. Consider a simple amplitude-modulated (AM) signal

$$s(t) = \cos(\omega_m t)\cos(\omega_c t) \tag{5.16}$$

where ω_c is the carrier frequency and ω_m is the modulation frequency, $\omega_m \ll \omega_c$. The signal is plotted in Figure 5.7(a). [In the context of a GPS SPS signal, ω_c and ω_m are akin to the carrier frequency ($\approx$ 1.5 GHz) and the chipping rate ($\approx$ 1 MHz) of the code, respectively.] The reason for this choice of the signal becomes clear when we invoke a simple trigonometric identity to rewrite (5.16) as

$$s(t) = \tfrac{1}{2}\left[\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t\right]$$

The amplitude-modulated signal (5.16) is simply the average of two sinusoidal signals of slightly different frequencies. We now consider transmission of this signal through a dispersive medium where each of the sinusoidal signals travels at a slightly different speed. Let the wave number corresponding to frequency ω_c be k, and the wave numbers corresponding to $(\omega_c + \omega_m)$ and $(\omega_c - \omega_m)$ be $(k + \Delta k)$ and $(k - \Delta k)$, respectively.

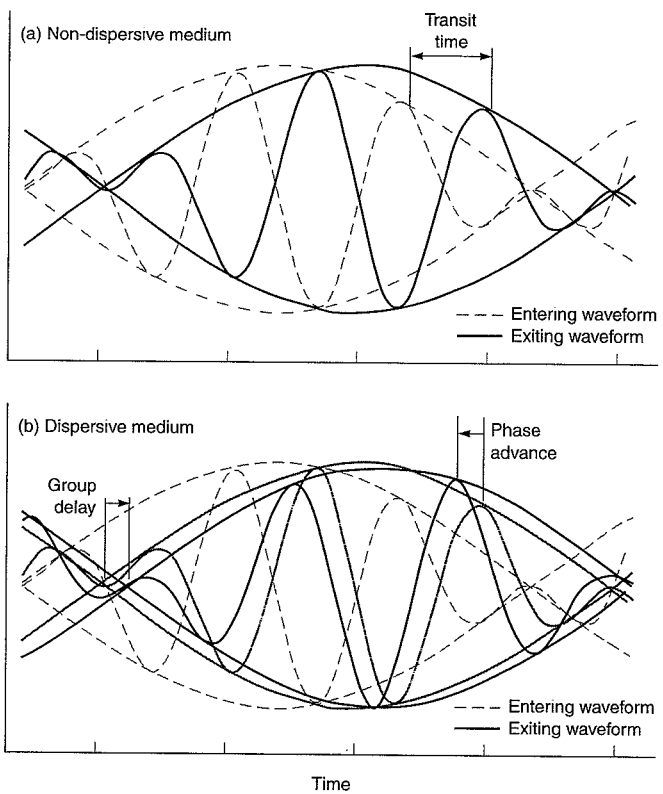


Figure 5.7 Propagation of a modulated signal in (a) free space, and (b) a dispersive medium which advances the carrier phase and delays the modulation function.

After the two frequency components have traveled a distance x in this dispersive medium

$$s(x,t) = \tfrac{1}{2}\left[\cos\big((\omega_c + \omega_m)t - (k + \Delta k)x\big) + \cos\big((\omega_c - \omega_m)t - (k - \Delta k)x\big)\right]$$

Again, by a simple trigonometric identity

$$\begin{aligned} s(x,t) &= \cos\big(\omega_m t - \Delta k \cdot x\big)\cos\big(\omega_c t - kx\big) \\ &= \cos\omega_m\left(t - \left(\frac{\Delta k}{\omega_m}\right)x\right)\cos\omega_c\left(t - \frac{k}{\omega_c}x\right) \end{aligned}$$

As ω_m gets smaller, $\Delta k/\omega_m$ approaches, in the limit, $dk/d\omega$, and

$$s(x,t) = \cos\omega_m\left(t - \frac{dk}{d\omega}x\right)\cos\omega_c\left(t - \frac{k}{\omega_c}x\right) \tag{5.17}$$

Equation (5.17) is an important result: In a dispersive medium, the carrier and the modulation travel at different speeds. The carrier (frequency ω_c) travels with phase velocity

$$v_p = \frac{\omega_c}{k},$$

and the modulation function $\cos(\omega_m t)$ travels with *group velocity*

$$v_g = \frac{d\omega}{dk}$$

where the derivative is evaluated at $\omega = \omega_c$. We can rewrite (5.17) as

$$s(x,t) = \cos\omega_m\left(t - \frac{x}{v_g}\right)\cos\omega_c\left(t - \frac{x}{v_p}\right) \tag{5.18}$$

Consider first the simple case of free space where $v_p = v_g$. We have a superposition of two waves of slightly different frequencies, $(\omega_c + \omega_m)$ and $(\omega_c - \omega_m)$, traveling at the same speed. The net result, shown in Figure 5.7(a) as a function of t for a fixed x, is a beat signal or a wave with a slowly pulsating intensity (or amplitude modulation). In other words, a wave whose frequency is ω_c, the average of the two frequencies, and which oscillates in strength with a frequency ω_m, one-half of the difference of the two frequencies. We see an envelope riding on top of the carrier waves, both propagating at the same speed.

Let's consider next the case of a dispersive medium in which the two waves of slightly different frequencies travel at slightly different speeds. The difference is that we have an amplitude-modulated signal with the modulation traveling at a different speed than the carrier. Figure 5.7(b) is a plot of (5.18), as before, as a function of t for a fixed x. As the waves travel along, the envelope riding on them travels slower ($v_g < v_p$). The important point for our immediate purpose is related to measurement of phases of the carrier and the modulation (i.e., the code) of GPS signals. As shown in Figure 5.7(b), the carrier phase would be measured too short and the code phase would be measured too long, as compared to measurements in Figure 5.7(a). This phenomenon is referred to as *code-carrier divergence*.

Earlier we defined the refractive index of a medium as $n = c/v$. For a dispersive medium, the phase and group velocities are different, and we define the phase and group refractive indices as

$$n_p = \frac{c}{v_p} \quad \text{and} \quad n_g = \frac{c}{v_g}.$$

It is left as an exercise to show from the definitions of phase and group refractive indices that they are related as

$$n_g = n_p + f\frac{dn_p}{df}. \tag{5.19}$$

We will use this relationship below when we examine transmission of the carrier and modulation of GPS signals through the ionosphere.

5.3.2 Ionospheric Delay

The ionosphere, extending from a height of about 50 km to about 1000 km above the earth, is a region of ionized gases (free electrons and ions). The ionization is caused by the sun's radiation, and the state of the ionosphere is determined primarily by the intensity of the solar activity. The ionosphere is composed of layers (named D, E, F1, and F2) at different heights, each with different rates of production and loss of free electrons. The peak electron density (number of electrons/m^3) occurs in height range 250–400 km (F2 layer).

The physical characteristics of the ionosphere change widely between day and night. As the sun rises, its ultraviolet (UV) radiation starts breaking up gas molecules (mainly H_2 and He at higher altitudes, and O_2 and N_2 lower down) into ions and free electrons. The electron density builds up, with a peak around 2 p.m. local time, and then starts declining. At night, there is no further ionization and the ions and electrons find each other and recombine, reducing the free-electron count. The electron densities change by one to two orders of magnitude between day and night.

There can be considerable variability from day to day, depending upon the solar activity and geomagnetic disturbances. There are significant changes with seasons and the phases of the eleven-year solar cycle. (In 2005 we are at the tail end of the solar cycle which began in 1995 and peaked in 2000–2001. The next solar max will occur around 2010.) There are also unpredictable short-term effects and localized anomalies (traveling ionospheric disturbances). In this section, we offer an overview of the ionospheric propagation effects on GPS signals. A comprehensive discussion is found in Klobuchar (1996).

The speed of propagation of radio signals in the ionosphere depends upon the number of free electrons in the path of a signal, defined as the *total electron content* (TEC): the number of electrons in a tube of 1 m^2 cross section extending from the receiver to the satellite.

$$TEC = \int_S^R n_e(l)dl \tag{5.20}$$

where $n_e(l)$ is the variable electron density along the signal path, and the integration is along the signal path from the satellite to the receiver. The path length through the ionosphere is shortest in the zenith direction and, therefore, TEC in the vertical direction (TECV) is the lowest. In the previous paragraphs, when we referred to the state of the ionosphere, we really meant a map of TECV. TEC is measured in units of TEC Units (TECU), defined as 10^{16} electrons/m^2. TECV typically varies between 1 and 150 TECU. For a given place and time, vertical TEC can vary 20–25% from its monthly average! The current models of the ionosphere do not provide an adequate representation of the day-to-day variability of TEC.

The ionosphere is generally well behaved in the temperate zones but can fluctuate near the equator and the magnetic poles. The region of highest ionospheric delay is within ±20° of the magnetic equator. Solar flares and the resulting magnetic storms can create large and quickly varying electron densities, especially in the polar regions, causing rapid fluctuation in the carrier phase (called scintillation) and in amplitude (called fading) of GPS signals. This phenomenon, though transient and infrequent at mid-latitudes, can present a difficulty in tracking the signal continuously in the polar and equatorial regions.

Phase Advance and Group Delay

Ionized gas is a dispersive medium for radio waves. The refractive index for a radio wave of frequency f (to first order) is

$$n_p \approx 1 - \frac{40.3\ n_e}{f^2} \tag{5.21}$$

where n_p is the phase refractive index and n_e is the electron density, defined earlier. The refractive index is slightly less than one, and phase velocity of the GPS carriers in the ionosphere exceeds that of light in a vacuum by enough to be significant for precise positioning. From the expression for the refractive index we can calculate the excess phase delay $\Delta\tau_p$ (in seconds) incurred by a signal as it propagates through the ionosphere (5.12).

$$\begin{aligned}\Delta\tau_p &= \frac{1}{c}\int_S^R (n_p(l) - 1)dl \\ &= -\frac{1}{c}\int \frac{40.3\, n_e(l)}{f^2}\, dl \\ &= -\frac{40.3 \cdot TEC}{cf^2}\end{aligned} \tag{5.22}$$

The phase delay is negative, i.e., the phase is advanced. The phase advance is in direct proportion to the number of electrons in the path of the signal. The excess phase delay in meters, which we have denoted as I_ϕ in (5.9), is

$$I_\phi = c \cdot \Delta\tau_p = -\frac{40.3 \cdot TEC}{f^2} \tag{5.23}$$

From (5.19), we can determine the group refractive index and the group delay in meters as

$$\begin{aligned} n_g &= 1 + \frac{40.3\, n_e}{f^2} \\ I_p &= \frac{40.3 \cdot TEC}{f^2} \end{aligned} \tag{5.24}$$

The ionospheric delay terms in measurements of pseudorange (5.6) and carrier phase (5.9) are equal in magnitude but opposite in sign.

$$I_\rho = -I_\phi = \frac{40.3 \cdot TEC}{f^2} \tag{5.25}$$

In later sections, we will represent the ionospheric group delay simply as I, and the phase delay as $-I$. Note that a change in TEC of 1 TECU corresponds to a change in ionospheric delay at L1 of about 16 cm.

Obliquity Factor

For the purpose of simple, geometrical modeling, the ionosphere may be thought of as a thin shell surrounding the earth (Figure 5.8). The signal path length through the ionosphere varies with the satellite position in the sky: the lower the satellite, the longer the path length and higher the TEC. Assuming that there are no lateral electron gradients, we can get a simple and compact characterization of TEC along a signal path in terms of the vertical TEC (TECV) and a multiplier to account for the longer path length. The multiplier is called *obliquity factor* (OF). We use both the elevation angle (*el*) and its complement, zenith angle (ζ), to represent the satellite position as it suits us.

The mean height of the ionospheric shell, or the mean ionospheric height (h_I), is usually taken in the range of 300–400 km. The *ionospheric pierce point* (IP) is defined as the point of intersection of the line of sight with the spherical shell at height h_I. Now we can relate TECV to TEC at zenith angle ζ in terms of path lengths through the thin ionospheric shell as

$$TEC(\zeta) = \frac{1}{\cos\zeta'} \cdot TECV \tag{5.26}$$

where ζ and ζ' are the zenith angles of the satellite at the user position and IP, respectively. The term $(\cos\zeta')^{-1}$ defines the obliquity factor, which we can write in terms of the zenith angle of the satellite at the user location as follows. From Figure 5.8, by law of sines

$$\frac{\sin\zeta}{(R_E + h_I)} = \frac{\sin\zeta'}{R_E} \tag{5.27}$$

where R_E is the average radius of the earth. The ionospheric obliquity factor for zenith angle ζ

$$OF_I(\zeta) = \left[1 - \left(\frac{R_E \sin\zeta}{R_E + h_I}\right)^2\right]^{-1/2} \tag{5.28}$$

The value of OF_I ranges from one for the zenith direction to about three for elevation angle of 5°. This *thin shell model* works surprisingly well.

The relationship (5.26) can be translated directly into group delay (or phase advance) for the GPS signals in accordance with (5.23) and (5.24). Denoting the ionospheric delay as a function of the zenith angle as $I(\zeta)$

$$\text{ionospheric delay}(\zeta) = \text{zenith delay} \times \text{obliquity factor}(\zeta)$$

$$I(\zeta) = I_z \cdot OF_I(\zeta)$$

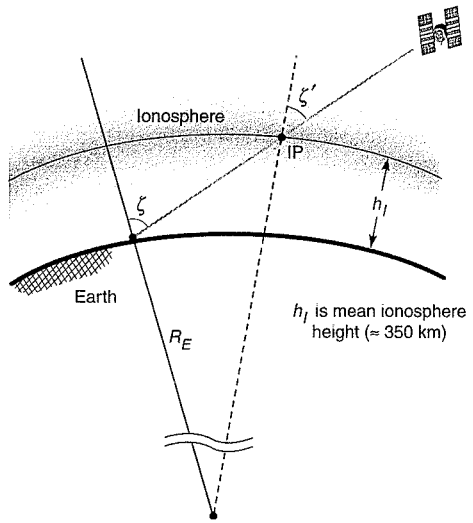


Figure 5.8 The propagation path length of a signal through the ionosphere increases with the zenith angle ζ. The increased path length is accounted for in terms of a multiplier of the zenith delay. The multiplier is called obliquity factor.

The ionospheric zenith delay I_z (i.e., the path delay in the zenith direction) typically varies at mid-latitudes from about 1–3 m at night to 5–15 m in the mid-afternoon. The zenith delay has been observed to be as high as 36 m near the equator at the peak of a solar cycle.

We can summarize our discussion so far of signal propagation through the ionosphere as follows.

- The carrier and its modulating signal (i.e., the code and navigation message) propagate at different speeds through the ionosphere: the code phase is delayed while the carrier phase is advanced by the same amount. In other words, the ionospheric refraction causes the code phase to be measured too long and the carrier phase to be measured too short by an equal amount.
- The phase advance and group delay are in direct proportion to TEC along the propagation path of a signal.
- The path length of a signal through the ionosphere depends upon the elevation angle of the satellite and is accounted for in the form of a multiplier called obliquity factor.

Delay Estimation with Dual-Frequency Measurements

A user equipped with a dual-frequency (L1-L2) GPS receiver can estimate the ionospheric group delay and phase advance from the measurements, and essentially eliminate the ionosphere as a source of measurement error. We rewrite the expression for pseudorange measurements (5.6), introducing additional notation to indicate L1 and L2 measurements.

$$\rho_q = r + c[\delta t_u - \delta t^s] + I_q + T + \varepsilon_{\rho_q}$$

where $q = L1$ or $L2$. Consolidating terms unrelated to the ionospheric effect and modeling the ionospheric delay to first order as varying inversely with the carrier frequency squared (5.25), we can rewrite the above equation simply as

$$\rho_q = \rho_{IF} + \frac{A}{f_q^2} \tag{5.29}$$

where ρ_{L1} and ρ_{L2} are the pseudorange measurements at L1 and L2, respectively; f_{L1} and f_{L2} are the corresponding carrier frequencies; ρ_{IF} is the *ionosphere-free* (IF) *pseudorange*, or the pseudorange measurement if the ionospheric effect were absent; and the unknown parameter A = 40.3 TEC. Both ρ_{IF} and A can be estimated from dual-frequency measurements (5.29). The estimate of ionospheric group delay at L1 is

$$I_{L1} = \frac{A}{f_{L1}^2} = \frac{f_{L2}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}(\rho_{L2} - \rho_{L1}) \tag{5.30}$$

The estimate of ionosphere-free pseudorange measurement is

$$\begin{aligned}\rho_{IF} &= \frac{f_{L1}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\rho_{L1} - \frac{f_{L2}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\rho_{L2} \\ &= 2.546\,\rho_{L1} - 1.546\,\rho_{L2}.\end{aligned} \tag{5.31}$$

There are two things to note about the ionosphere-free measurements (5.31). First, the errors due to satellite clock, ephemeris, and troposphere are present in these measurements in full measure, as in ρ_{L1} and ρ_{L2}. Second, the elimination of the ionospheric effect has been achieved at a price: The ionosphere-free pseudorange is significantly noisier than the pseudoranges measured at L1 and L2. If we model the multipath and receiver noise at L1 and L2 to be uncorrelated and having the same variance, the noise in the ionosphere-free pseudorange is

$$\sqrt{2.546^2 + 1.546^2} \approx 3$$

times larger than that in ρ_{L1} or ρ_{L2}. Actually, the assumption of uncorrelatedness is reasonable but the noise in the measurements at L2 is larger for the SPS receivers at present, a price for taking the forbidden measurements.

Estimates of the ionospheric delay based on L1-L2 pseudorange measurements in accordance with (5.30) are shown in Figure 5.9. The estimates appear noisy, but it is no surprise in view of the remarks above. The ionospheric delay estimates (5.30) are about twice as noisy as the individual pseudorange measurements. It's a trade-off: We get rid of a significant bias term but pick up additional noise in the process.

The carrier phase measurements are much less noisy, and we can attempt to estimate the ionospheric phase advance based on them. First, we rewrite the expression for carrier phase measurements (5.9) introducing notation to distinguish between L1 and L2 measurements.

$$\phi_q = \lambda_q^{-1}[r - I_q + T_\phi] + f_q(\delta t_u - \delta t^s) + N_q + \varepsilon_{\phi_q} \tag{5.32}$$

Equations similar to (5.30) and (5.31) can be written for the ionosphere-free carrier phase measurement, but involve integer ambiguities. It is left as an exercise to show that the phase advance at L1 is

$$I_{L1} = \frac{f_{L2}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\left[\lambda_{L1}(\phi_{L1} - N_{L1}) - \lambda_{L2}(\phi_{L2} - N_{L2})\right] \tag{5.33}$$

The estimate of ionospheric delay based on the code measurements (5.30) is unambiguous but noisy. The corresponding estimate based on carrier phase measurements (5.33) is precise but ambiguous in integer values.

We have noted previously that it's a challenge to estimate N_{L1} and N_{L2} in real time. So, what's the point of (5.33)? Well, as long as continuous carrier tracking is maintained, N_{L1} and N_{L2} remain fixed, and we can use (5.33) to estimate accurately change in the ionospheric delay between measurement epochs in real time. Such change is called *differential delay*, and its estimate obtained from (5.33) is good to centimeter level. We have plotted the estimates of differential delay in Figure 5.9 for comparison with the corresponding estimates of ionospheric group delay obtained from (5.30). The noisy code-based estimates of the ionospheric delay can be smoothed with carrier phase-based estimates of the differential delay. The basic technique is discussed in Section 5.7. The resulting smoothed estimates of the ionospheric delay would typically have decimeter-level accuracy after a satellite has risen to about 30° in elevation angle.

Dual-frequency GPS measurements have provided atmospheric physicists with a powerful tool to study the ionosphere. Actually, Transit with its dual-frequency signaling was the first satellite navigation system to be employed for monitoring the ionospheric dynamics. GPS offers a vast improvement over Transit. The ability to map TEC in real time on a global scale and

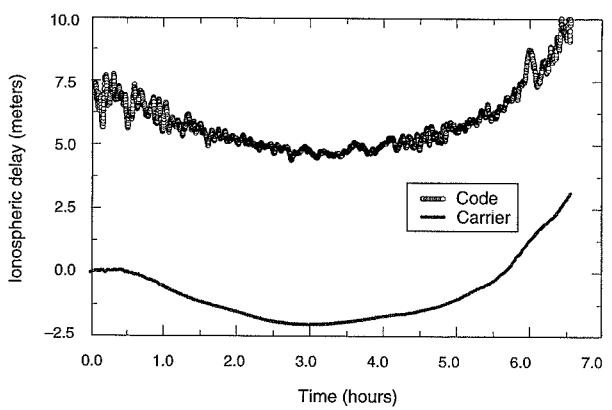


Figure 5.9 Ionospheric delay (L1) estimates obtained from code and carrier phase measurements at both L1 and L2. The code-based estimates are noisy. The carrier-based estimates are precise but ambiguous, and the plot starts arbitrarily at zero value. Changes in the ionospheric delay can be measured accurately with the carrier phase measurements while the carrier is tracked continuously.

to predict or identify disturbed regions is of great value to HF communications facilities and communication satellites, and for radar tracking of objects in space. The International GNSS Service (IGS) now offers global ionospheric maps giving TECV values over a grid, but not yet in real time.

We note in passing a calibration issue to account for differences in path lengths for the L1 and L2 signals, or inter-frequency biases, in a satellite. These parameters, also known as differential group delays or T_{GD} values, are specified in the satellite navigation message (Subframe 1). These correction terms are relevant to estimation of ionospheric delay from dual-frequency measurements, positioning with single frequency receivers, and time transfer [Wilson *et al.* (1999)].

Broadcast Model

Receivers limited to L1 measurements have recourse to an empirical model whose parameter values are broadcast by the satellites. This model, often referred to as the *Klobuchar model* in honor of its developer [Klobuchar (1996)], represents the zenith delay as a constant value at nighttime and a half-cosine function in daytime (Figure 5.10). The zenith ionospheric delay estimate $\hat{I}_{z,L1}$ at local time t is given by

$$\frac{\hat{I}_{z,L1}}{c} = \begin{cases} A_1 + A_2 \cos\left(\dfrac{2\pi(t - A_3)}{A_4}\right), & \text{if } |t - A_3| < A_4/4 \\ A_1 & \text{otherwise} \end{cases} \tag{5.34}$$

where A_1: nighttime value of the zenith delay (fixed at 5×10^{-9} s),
A_2: amplitude of the cosine function for daytime values,
A_3: phase corresponding to the peak of the cosine function (fixed at 50,400 s, or 14 h, local time), and
A_4: period of the cosine function ($\geq$ 72,000 s).

The values of parameters A_2 and A_4 are specified in the navigation message broadcast by each satellite in terms of four coefficients each of a polynomial function (Subframe 4, page 18). A set of eight coefficients is selected by the Master Control Station from a pool of 370 such sets associated with the different seasons and levels of solar activity. The parameters are typically updated every six days. The ionospheric delay is determined from the broadcast parameter values (A_2 and A_4) and the user's latitude, longitude, satellite azimuth and elevation, and local time. Step-by-step calculations of the zenith delay and obliquity factor are given in the GPS Interface Specification (IS) [IS-GPS-200D (2004)].

The Klobuchar model was developed on the basis of empirical data, but under constraints on the number of parameters to be used (eight, at most) and how often they could be updated (daily, at most). We offer a few remarks. First, the time t in (5.34) is the local time at the ionospheric pierce point (IP): The ionospheric action is at IP, not at the user location. Secondly, the position of the IP is expressed in terms of geomagnetic latitude, not geodetic. Why? Because the model fits the empirical data better when geomagnetic coordinates are used. (The computation uses the geodetic coordinates of the geomagnetic north pole: N78.3°, E291.0°.) The model represents the peak ionospheric delay as occurring at 14 h local time at all places. In fact, the time of the local peak can vary between 11 h and 17 h depending upon the season, latitude, and solar activity.

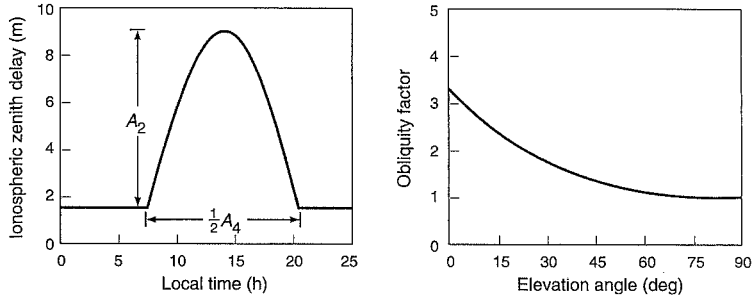


Figure 5.10 The Klobuchar ionospheric model. Parameter values A2 and A4 are selected by the Control Segment to reflect the prevailing ionospheric conditions and are broadcast by the satellites.

The obliquity factor for the broadcast model is expressed in terms of the satellite elevation angle *el* as

$$OF_I(el) = 1.0 + 16.0 \times (0.53 - el)^3 \tag{5.35}$$

where *el* is in units of semi-circles (1 semi-circle = 180° or π radians). This expression is an approximation of (5.28). The motivation for simplifying (5.28) is to make computations easier. In the pre-microprocessor 1970s, there was a lot of concern about keeping the computational demands on a receiver low.

The broadcast model is estimated to reduce the rms range measurement error due to uncompensated ionospheric delay by about 50% [Feess and Stephens (1987)]. At mid-latitudes, the remaining error in zenith delay can be up to 10 m during the day, and much worse during heightened solar activity.

5.3.3 Tropospheric Delay

The GPS signals are also refracted by the lower part of the earth's atmosphere composed of dry gases (mainly N_2 and O_2) and water vapor. Most of the water vapor is below 4 km, and all of it is below about 12 km, measured from sea level. The dry gases, however, are found in gradually thinning quantities at altitudes of hundreds of kilometers. About three-quarters of the gaseous mass is found in the troposphere, which stretches to about 16 km above the equator and about 9 km above the poles. The overall effect of the neutral atmosphere is, therefore, referred to as the tropospheric effect. In this section, we offer an overview of the tropospheric models. A comprehensive discussion is found in Spilker (1996).

Unlike the ionosphere, troposphere is non-dispersive for GPS frequencies (i.e., the refractive index doesn't depend upon the frequency of the signal). The refractive index of the atmospheric gases is only slightly larger than unity: $n \approx 1.0003$ at sea level, and much closer to unity at the upper end of the troposphere. The speed of propagation of GPS signals in the troposphere is lower than that in free space and, therefore, the apparent range to a satellite appears longer, typically by 2.5–25 m depending upon the satellite elevation angle. The phase and group velocities are the same, and the measurements of code and carrier at L1 and L2 frequencies all experience a common delay. This delay cannot be estimated from GPS mea-

surements, and a user has to resort to models to correct for it. In the notation of Section 5.1, $T_\rho = T_\phi$, and, therefore, we will drop the subscripts ρ and ϕ in later sections and denote the tropospheric delay simply as T. In this section, however, we need T to denote the temperature and switch to $\tilde{T}$ to denote the tropospheric delay. (T is an overworked symbol in this book, used in Part III to denote both the period of a C/A-code and the noise temperature.)

The extent of the tropospheric delay experienced by a signal depends upon the refractive index of the air mass along its path. The refractive index of an air mass depends upon its density, which can be expressed as the sum of the densities of the dry air constituents and water vapor. These densities are functions of the pressures of the dry gases and water vapor, and the temperature. The dry and wet constituents of the atmosphere affect propagation of an RF signal differently, and each is modeled separately. The composition of the dry gases depends upon latitude, season, and altitude, and is relatively stable. The water vapor content in the troposphere, however, is much harder to model; water vapor density varies with the local weather and can change quickly. Fortunately, most of the tropospheric delay ($\approx 90\%$) is due to the more predictable dry atmosphere.

Dry and Wet Delays

It is convenient to define *refractivity*, $N = (n-1) \times 10^6$, and express it as the sum of refractivities of the dry gases and the water vapor in the atmosphere

$$N = N_d + N_w, \tag{5.36}$$

where N_d and N_w are called the dry and wet refractivities, respectively. As in the derivation of (5.13), the excess path delay due to tropospheric refraction (in meters) can be written as

$$\begin{aligned}\tilde{T} &= 10^{-6}\int N(l)\,dl = 10^{-6}\int [N_d(l) + N_w(l)]\,dl \\ &= \tilde{T}_d + \tilde{T}_w,\end{aligned} \tag{5.37}$$

where $\tilde{T}_d$ and $\tilde{T}_w$ are called dry delay and wet delay, respectively, and the integration is along the signal path. As with signal propagation through the ionosphere, we will disregard the curvature of the signal path and focus on excess delay due to the reduced signal speed.

Models of the troposphere attempt to estimate the dry and wet refractivities along the signal path, and estimate the total delay from (5.37). The refractivity of a parcel of air depends upon its temperature and partial pressures of the dry gases and water vapor. Simple, approximate expressions (with empirically obtained coefficients) for the dry and wet refractivities of a parcel of air are

$$\begin{aligned}N_d &= 77.64\frac{P}{T}, \\ N_w &= 3.73 \cdot 10^5 \frac{e}{T^2},\end{aligned} \tag{5.38}$$

where P is the total pressure and e is the partial pressure of the water vapor, both in millibars; and T is the temperature in kelvin.

Knowledge of pressure, temperature, and humidity along the propagation path of a signal can determine the refractivity profile and the tropospheric delay (5.37) precisely. Such measurements are obtained for the nominally vertical direction by atmospheric researchers using

instrumented weather balloons (radiosondes). GPS users rarely have access to such measurements. A less onerous approach is to measure the meteorological conditions (pressure, temperature, and humidity) at the antenna location, and relate these measurements to P, T, and e along the signal path using gas laws or empirical models. Researchers requiring great positioning precision with GPS often record surface meteorological conditions. For example, GPS stations of a permanent tectonic monitoring network make provisions for meteorological instruments at each site.

In navigation applications, however, meteorological measurements are impractical. In fact, most GPS users base the estimation of tropospheric delay upon average meteorological conditions at their locations obtained from a model of the *standard atmosphere* for the day of the year and the user's latitude and altitude. (A standard atmosphere is basically a specification of pressure, temperature, and humidity profile with altitude, and variations in these parameters with latitude and seasons based on certain average conditions.)

We'll consider estimation of the tropospheric delay for a signal in two steps:

1. Estimation of the zenith delay ($\tilde{T}_z$) (i.e., delay associated with a signal from the zenith direction) in terms of the corresponding dry ($\tilde{T}_{z,d}$) and wet ($\tilde{T}_{z,w}$) delays
$$\tilde{T}_z = \tilde{T}_{z,d} + \tilde{T}_{z,w}$$
2. Definition of obliquity factor to scale the zenith delay as a function of the elevation angle (el) of the satellite. The atmospheric physicists refer to the tropospheric obliquity factors as mapping functions.
$$\tilde{T} = \tilde{T}_{z,d} \cdot m_d(el) + \tilde{T}_{z,w} \cdot m_w(el)$$
where we have introduced separate mapping functions m_d and m_w for the dry and wet components, respectively. Simple models often use a common mapping function ignoring differences in the atmospheric profiles of the dry gases and water vapor.

There is no dearth of tropospheric models. The models differ in their assumptions regarding changes in temperature and water vapor with altitude. We describe two simple models below. There is no shortage of mapping functions either. We discuss three below.

Tropospheric Models

The *Saastamoinen model* was derived using gas laws and simplifying assumptions regarding changes in pressure, temperature, and humidity with altitude. The zenith dry and wet delays are given as

$$\tilde{T}_{z,d} = 0.002277\,(1 + 0.0026\cos 2\phi + 0.00028\,H)\,P_0, \tag{5.39a}$$

$$\tilde{T}_{z,w} = 0.002277\left(\frac{1255}{T_0} + 0.05\right)e_0 \tag{5.39b}$$

where T_0 is the temperature (kelvin), P_0 is the total pressure and e_0 is the partial pressure due to water vapor (both in millibars), all determined at the antenna location by measurements or from models of standard atmosphere; ϕ is the latitude and H is the orthometric height of the

antenna (km). This model has been refined with additional correction terms which we disregard here.

The *Hopfield model* is based on a relationship between dry refractivity at height h to that at the surface. The relationship was derived empirically on the basis of extensive measurements. This model, referred to as a quartic model of refractivity profile, is

$$N_d(h) = N_{d0}\left(1 - \frac{h}{h_d}\right)^4, \tag{5.40}$$

where h denotes the height above the antenna, N_{d0} is the dry refractivity at the surface, and h_d ($\approx$ 43 km) is defined as the height above the antenna at which the dry refractivity is zero: $N_d(h_d) = 0$.

The Hopfield model for wet refractivity assumes a relationship similar to (5.40), though with less persuasive evidence.

$$N_w(h) = N_{w0}\left(1 - \frac{h}{h_w}\right)^4,$$

where N_{w0} is the wet refractivity at the surface and h_w = 12 km. From (5.37)

$$\begin{aligned}\tilde{T}_z &= 10^{-6}\int\left[N_d(h) + N_w(h)\right]dh \\ &= \tfrac{10^{-6}}{5}\left[N_{d0}\,h_d + N_{w0}\,h_w\right] \\ &= \tilde{T}_{z,d} + \tilde{T}_{z,w}\end{aligned}$$

Substituting expressions for the dry and wet refractivities from (5.38)

$$\tilde{T}_{z,d} = 77.6 \cdot 10^{-6}\,\frac{P_0}{T_0}\,\frac{h_d}{5} \tag{5.41a}$$

$$\tilde{T}_{z,w} = 0.373\,\frac{e_0}{T_0^2}\,\frac{h_w}{5} \tag{5.41b}$$

The value of $\tilde{T}_{z,d}$ is 2.3–2.6 m at sea level, and gets lower as the altitude increases: about 2 m at Denver, Colorado, the mile-high city; and about 1 m atop a Himalayan peak. The value of $\tilde{T}_{z,w}$ ranges from near-zero to 80 cm (millimeters in polar region, a few centimeters in deserts, and tens of centimeters in tropical areas).

The dry delay for the zenith direction ($\tilde{T}_{z,d}$) can be predicted with an accuracy of a few millimeters from accurate surface pressure measurements. The corresponding wet delay depends upon the distribution of water vapor along the signal path, and can be highly variable. (The mixing of the water vapor and dry air is a complicated process depending upon the local weather conditions, and this distribution can change quickly.) The models of wet delay ($\tilde{T}_{z,w}$) based on meteorological data at the surface are less accurate, with typical error of 1–2 cm. Use of average meteorological conditions rather than actual measurements introduces additional modeling errors in both the dry and wet delays, and the total zenith delay error can be 5–10 cm.

Mapping Functions

A number of mapping functions (or, tropospheric obliquity factors) have been proposed. The simplest model for both the dry and the wet components is $1/\sin el$. This model is consistent with a flat earth, and is a poor approximation for low-elevation satellites ($el < 15°$). An example of a more accurate model is

$$m(el) = \frac{1}{\sqrt{1 - \left(\cos el/1.001\right)^2}} \tag{5.42}$$

The following is a simple example of separate mapping functions for dry and wet delays.

$$m_d(el) = \frac{1}{\sin el + \dfrac{0.00143}{\tan el + 0.0445}}$$

$$m_w(el) = \frac{1}{\sin el + \dfrac{0.00035}{\tan el + 0.017}}$$

More sophisticated mapping functions are based on a truncated form of continued fraction.

$$m_i(el) = \frac{1 + \dfrac{a_i}{1 + \dfrac{b_i}{1 + c_i}}}{\sin(el) + \dfrac{a_i}{\sin(el) + \dfrac{b_i}{\sin(el) + c_i}}}, \tag{5.43}$$

where subscript $i = d$ or w, and the coefficients a_i, b_i, and c_i in the different models are empirically determined constants or functions of variables such as latitude, height, surface temperature and pressure, and day of the year.

The obliquity factor increases sharply as a satellite gets lower in the sky. At low elevation angles, the typical values are about two at 30°, four at 15°, six at 10°, and ten at 5°. The obliquity factors for the troposphere are much larger at low elevation angles than those for the ionosphere. The reason for this would be clear from a comparison of the corresponding path lengths as shown in Figures 5.8 and 5.11.

A 5–10 cm residual error in tropospheric zenith delay estimates based on average meteorological conditions grows into an error of 0.5–1 m at 5° elevation. Such error would not be of serious concern for meter-level navigation but, clearly, is too large for centimeter-level positioning. For an interesting discussion of the role of tropospheric errors in DGPS for landing airplanes, see Skidmore and van Graas (2004).

The atmospheric scientists are examining the use of GPS measurements to map the amount of water vapor in the air globally and assess its role in weather prediction and climate change. The profile of water vapor content in a column of air is measured at considerable expense using radiosondes. The U.S. Weather Service actually launches radiosondes from about eighty sites in the conterminous United States twice a day for this purpose. While GPS measurements

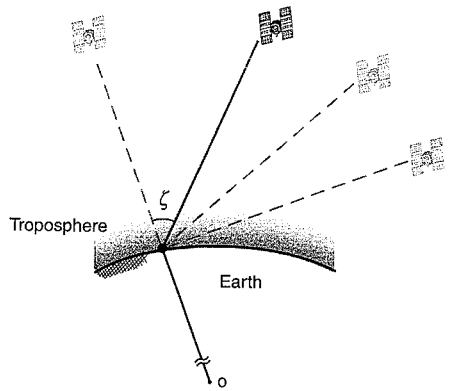


Figure 5.11 The path length of a signal through the troposphere increases significantly at low elevation angles. The obliquity factor is about two at 30° elevation angle, four at 15°, six at 10°, and ten at 5°. (Symbol ζ denotes zenith angle; $\zeta = 90° - el.$)

cannot give the profile of the water vapor content, a network of GPS stations can provide estimates of total water vapor content over large areas in real time continuously. It is, however, a challenge to compensate for the errors in the GPS measurements (satellite ephemeris and clock, ionosphere, tropospheric dry delay, multipath, and receiver noise) to estimate the zenith wet delay with the required accuracy in real time.

The main features of GPS signal propagation through the ionosphere and the troposphere are summarized in Table 5.1.

5.4 Measurement Errors

The errors in pseudorange measurements discussed so far are common to the code and carrier phase measurements (except for a sign change for the ionospheric delay). The signal or receiver design played no role. Now we examine a class of errors which actually depend upon signal power, code structure, and receiver/antenna design.

In this section, we address the question: How well can a receiver measure the code and carrier phase of a GPS signal reaching its antenna? We have to make a distinction between precision and accuracy of the measurements. Precision and accuracy are often used as synonyms, and we haven't been too careful about it in this book. Actually, there is a distinction. We know that accuracy is assessed in relation to the truth. Without being too pedantic about it, precision simply refers to the resolution or fineness of a measurement. A measurement or an observation can be extremely precise and completely wrong.

We look at two sources of error below. The first, receiver noise, essentially smudges the signal, affecting precision of the code or carrier phase measurement. The second, multipath, introduces interfering signals which actually change the phase being measured. Both effects are discussed in detail in Chapter 10.

Table 5.1 A summary of the features of ionosphere and troposphere relevant to GPS signal propagation

	Ionosphere	*Troposphere*
Variability	High (diurnal, seasonal, and solar cycles; solar flares)	Low (significant change only in the wet component, which is less than 10% of the delay)
Zenith delay	Several meters to several tens of meters	2.3–2.6 m at sea level
Obliquity factor (= 1 in zenith direction)	≈ 1.8 at 30° elevation angle; 2.5 at 15°; 3 at 5°	≈ 2 at 30° elevation angle; 4 at 15°; 10 at 5°
Modeling error for zenith delay	1–10 m, or more	5–10 cm (without meteorological data)
Dispersive for GPS frequencies?	Yes (therefore, the delay can be measured from dual-frequency measurements)	No

5.4.1 Receiver Noise

The code and carrier measurements are affected by random measurement noise, called receiver noise, which is a broad term covering the RF radiation sensed by the antenna in the band of interest unrelated to the signal; noise introduced by the antenna, amplifiers, cables, and the receiver; multi-access noise (i.e., interference from other GPS signals and GPS-like broadcasts from system augmentations); and signal quantization noise.

A receiver cannot follow changes in the signal waveform perfectly and invariably there are delays and distortions. In the absence of any interfering signals, a receiver sees a waveform which is the sum of the GPS signal and randomly fluctuating noise. The net result is that the fine structure of a signal can be masked by noise, especially if the signal-to-noise ratio is low. The measurement error due to receiver noise varies with the signal strength, which, in turn, varies with the satellite elevation angle.

According to a rule of thumb, it is not a challenge for the receiver technology to measure the phase of a reasonably strong sinusoidal signal with a precision of 1/2%–1% of a cycle. Empirical characterization of receiver noise is discussed by Nolan *et al.* (1992). The effect of receiver noise on the precision of the code and carrier phase measurements is analyzed in Section 10.6.

5.4.2 Multipath

Multipath refers to the phenomenon of a signal reaching an antenna via two or more paths. Typically, an antenna receives the direct (i.e., line-of-sight) signal and one or more of its reflections from structures in the vicinity and from the ground (see Figure 5.12). A reflected signal is a delayed and usually weaker version of the direct signal. The subsequent code and

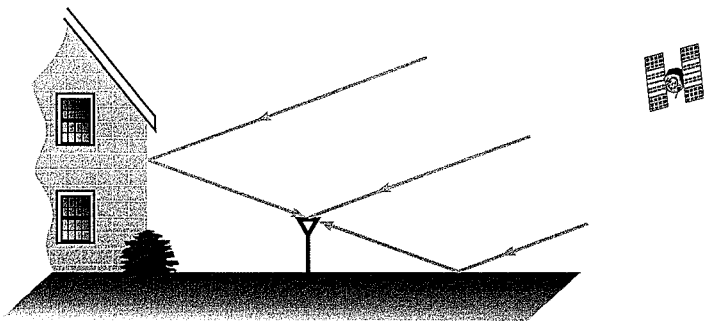


Figure 5.12 Multipath: The signal may reach an antenna via more than one path. A reflected signal is a delayed and usually weaker version of the direct signal.

carrier phase measurements are for the sum of the received signals. The range measurement error due to multipath depends upon the strength of the reflected signal and the delay between the direct and reflected signals. Multipath affects both code and carrier measurements, but the magnitudes of the error differ significantly. We offer a brief account below and an in-depth discussion in Section 10.7.

The primary defense against multipath is to locate the antenna away from reflectors, but that is not always practical. The effect of multipath can be reduced in antenna design process by lowering the contribution of some types of reflections (e.g., from the ground below the antenna). The effect can also be reduced in the signal processing step in a receiver, and several receiver manufacturers have developed and implemented proprietary techniques.

Actually, a measure of multipath immunity is built into the signal structure. A reflected signal which is delayed by more than 1.5 chips would be suppressed automatically in the correlation process in a receiver because the auto-correlation for the C/A-code is nearly zero for delays longer than 1.5 chips. Such delay corresponds to about 500 m of increased path length

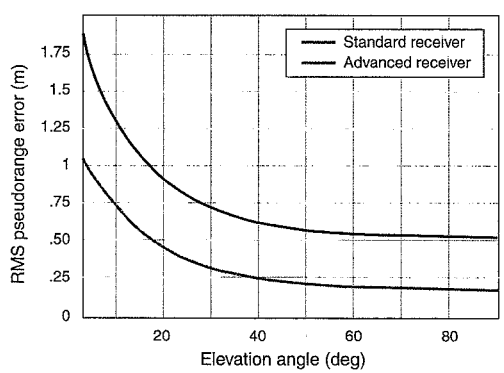


Figure 5.13 Code phase measurement error modeled as a function of elevation angle for two types of receivers.

for a C/A-code signal and 50 m for a P(Y)-code signal. One of the benefits of a higher chipping rate is greater multipath immunity. The impact of a reflected signal delayed by less than 1.5 chips would depend upon the amount of delay and the signal amplitude.

Typical multipath error in pseudorange measurements varies from 1 m in a benign environment to more than 5 m in highly reflective environment. The corresponding errors in the carrier phase measurements are typically two orders of magnitude smaller (1–5 cm). We know that the phase measurement error can be no worse than a cycle. Actually, it's no worse than a quarter cycle, as shown below.

Consider a simple model. An antenna receives two signals: a line-of-sight signal and a reflected signal with phase shift $\Delta\phi$ and amplitude attenuation α.

$$\text{received signal} = A\cos\phi + \alpha A\cos(\phi + \Delta\phi) \tag{5.44}$$

Application of straightforward but tedious trigonometric relationships allows us to determine the error in the carrier phase measurement as a result of multipath as

$$\delta\phi = \arctan\left(\frac{\sin\Delta\phi}{\alpha^{-1} + \cos\Delta\phi}\right) \tag{5.45}$$

For $\alpha < 1$, in the worst case, $\delta\phi = 90°$. In other words, the error in the carrier phase measurements due to multipath does not exceed a quarter cycle if the reflected signal has a smaller amplitude than the direct signal.

5.4.3 Measurement Error Models

The measurement errors due to receiver noise and multipath depend upon the satellite elevation angle. As a satellite gets lower in the sky, the received signal power decreases and multipath increases. The combined effect on pseudorange measurements has been analyzed by McGraw *et al.* (2000) in certain 'clean' environments as a function of the satellite elevation angle. We present a summary of their results in Figure 5.13 for two categories of receivers. The first type we call 'standard' receivers. The second type, which we call 'advanced' receivers, are characterized by features such as narrow correlator, discussed in Section 10.5, and choke-ring antenna [Schupler and Clark (1991)]. Measurements from very low satellites are often more trouble than they are worth, and an elevation angle cutoff of 5°–7.5° appears to offer a good trade-off between the loss of measurements and potential for large errors.

Typical rms errors for code and carrier phase measurements are shown in Table 5.2. Chances are that your measurements wouldn't be much better, and can be considerably worse. High-end receivers with special provision to deal with multipath would offer better measurements. Receivers intended for the consumer market can be considerably worse. Table 5.2 will serve as our simple model for measurement errors in illustrative examples in Chapters 6 and 7. For civil users, the L2 measurements are noisier than the L1 measurements, but we'll disregard this for simplicity.

5.5 User Range Error (URE)

We now return briefly to classification of a measurement error as noise or bias. Noise generally refers to a quickly varying error that averages out to zero over a 'short' time interval, where short is defined in relation to the integration time or smoothing time of a receiver [Chapters 11

Table 5.2 Typical receiver-related measurement errors (rms) in code and carrier phase measurements

	Receiver Noise	*Receiver Noise and Multipath (RNM)*
Code phase	0.25–0.50 m	0.5–1.0 m
Carrier phase	0.005–0.01 cycle ≈ 1–2 mm	0.025–0.05 cycle ≈ 0.5–1 cm

and 12]. Receiver noise meets this requirement. The range error due to multipath is quasi-sinusoidal and may also be thought of as noise. The errors due to signal propagation delays in the ionosphere and troposphere, corrected with models or not, may persist for tens of minutes, or longer. The satellite clock and ephemeris errors also change slowly and, like the propagation errors, would be classified as biases.

We adopt a simple model for pseudorange measurements from a single-frequency (L1) receiver after corrections have been applied based on the Klobuchar model for ionospheric delay and a reasonable model for tropospheric delay. In our model, the range error (RE) attributed to the Control Segment (CS) has an rms value of about 3 m ($\sigma_{RE/CS} \approx 3$ m). This number is based on empirical data. The rms residual range error due to atmospheric propagation models at mid-latitudes is about 5 m ($\sigma_{RE/P} \approx 5$ m). The ionospheric component of the propagation error can be highly variable depending upon the user location, satellite elevation angle, and state of the medium. The rms range error due to receiver noise and multipath (RNM), discussed above, is assumed to be 1 m ($\sigma_{RE/RNM} \approx 1$ m). The error due to multipath can easily be higher for many users. Our purpose here is to set up a simple model for later analyses.

We can now characterize the combined effect of these error sources on pseudorange measurements. The combined error is referred to as the user range error (URE), also known as the user equivalent range error (UERE). It is reasonable to model the errors due to the satellite clock and ephemeris, atmospheric propagation, multipath, and receiver noise to be uncorrelated, and the URE can be defined as the root-sum-square (rss) of these components, all expressed in units of length.

$$\sigma_{URE} = \sqrt{\sigma_{RE/CS}^2 + \sigma_{RE/P}^2 + \sigma_{RE/RNM}^2} \approx 6\,\text{m}$$

We should note in passing that while SA was active, $\sigma_{URE} \approx \sigma_{RE/CS} \approx 25$ m.

The error budget presented in Table 5.3 is noteworthy for its round numbers, reflecting our uncertainty about the actual values. A user would refine this model to take into account the knowledge of an actual measurement scenario. The atmospheric propagation error in the equatorial region or at high latitudes may double for a single-frequency receiver during high solar activity. On the other hand, $\sigma_{RE/P} \approx 0$ for a dual-frequency receiver in space. Similarly, a user with a low-end receiver located near reflectors may triple the size of the error due to multipath.

There is no standard terminology for characterization of ranging error, and this can be confusing. Sometimes total error, denoted as UERE, is divided into Signal-in-space URE

Table 5.3 Typical pseudorange measurement errors for a single-frequency (L1) receiver

Error Source	*RMS Range Error (RE)*
Satellite clock and ephemeris parameters	$\sigma_{RE/CS} \approx 3$ m
Atmospheric propagation modeling	$\sigma_{RE/P} \approx 5$ m
Receiver noise and multipath	$\sigma_{RE/RNM} \approx 1$ m
User range error (URE)	$\sigma_{URE} \approx 6$ m

(SIS URE or SISRE) and user equipment error (UEE). In our notation, SIS URE = $\sigma_{RE/CS}$ and UEE = $(\sigma_{RE/P}^2 + \sigma_{RE/RNM}^2)^{1/2}$.

5.6 Measurement Error: Empirical Data

To illustrate the size and character of the various errors, we have plotted in Figure 5.14 estimates of the error components in pseudorange measurements at L1. The measurements were taken in March 1997 from a satellite from the time it rose around noon until it set at about 4:30 p.m. local time. At its highest, the satellite elevation angle was 38°. In estimation of the errors, we took advantage of information generally not available to a user: We knew our antenna position precisely. We also sidestepped the receiver clock bias and drift issues by using an external cesium atomic standard. The measurements consisted of code and carrier phases at both L1 and L2 frequencies.

Figure 5.14(a) shows the dominant role of SA while it was active. The measurements are consistent with the commonly used model of the SA error as a zero-mean Gauss-Markov process with standard deviation of about 25 m. The correlation time of the error was about three to four minutes (i.e., error samples taken three to four minutes apart appear uncorrelated). SA is not missed.

Figure 5.14(b) shows the estimates of the ionospheric propagation delay obtained from the dual-frequency measurements and from the parameters of the broadcast model. The ionospheric delay profile for the entire satellite track was estimated separately from the dual frequency code and carrier measurements. The code-based estimates were then used effectively to resolve the integer ambiguity in the carrier phase measurements, and the resulting smooth estimate is shown. As noted earlier, the ionospheric effects on the code and carrier measurements are equal but in opposite directions. For simplicity, only the magnitude of the effect is shown. The ionospheric activity picks up at sunrise and peaks in the early afternoon. The plot covers the period of most intense ionospheric activity during the day, but the delays are small (< 5 m). In 1997, we were at the beginning of a new solar cycle and the ionosphere was relatively quiet. The broadcast model overcompensated for the ionospheric delay, and the remaining error is seen to be 3–6 m, depending upon the elevation angle. It is interesting to note that the propagation delay did not increase significantly at low elevation angles as the satellite set.

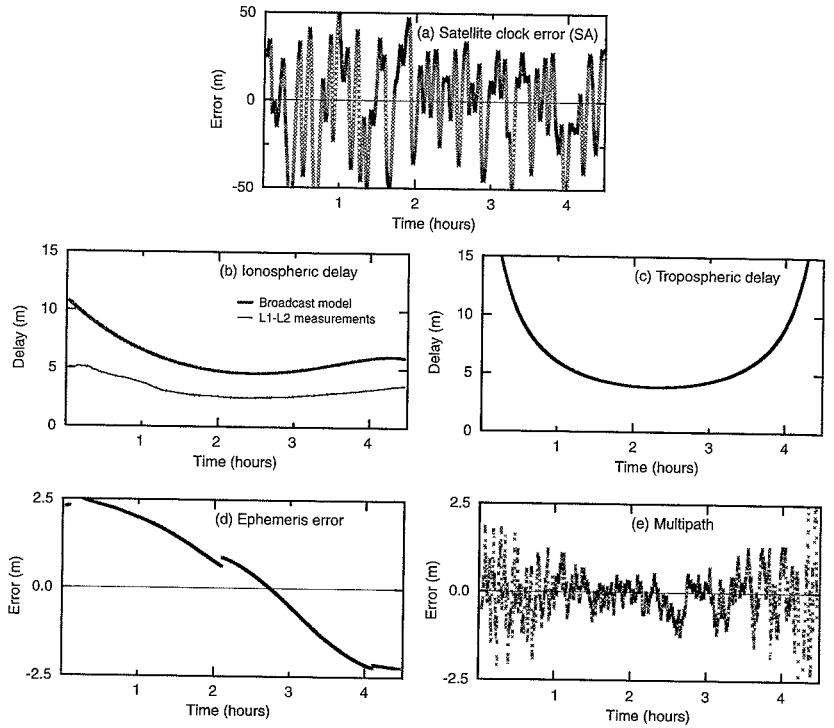


Figure 5.14 Pseudorange measurement errors due to the different error sources for a satellite from its rise time to set time (highest elevation angle: 38°). These measurements were taken in 1997 while Selective Availability (SA) was active.

The reason is simple: The satellite set in the north-east direction, away from the sun.

Figure 5.14(c) shows the tropospheric delay estimated using two different models. The first model required knowledge of the local meteorological conditions: pressure, temperature, and humidity measured at the antenna location. The second model, University of New Brunswick's UNB4 [Collins *et al.* (1996)], based the estimate on a definition of the standard atmosphere, given the season, latitude, and altitude. It is interesting that the two models gave virtually identical results. The role of the satellite elevation angle is clear: At low elevation angles, the signal traverses more of the troposphere, and the delay can be significantly larger than that at higher elevation angles. The delay ranges from about 4 m when the satellite was at its highest (38°) to about 15 m at low elevations (5°) as the satellite rose and set. If the satellite had come directly overhead, the corresponding delay would have been about 2.5 m.

Figure 5.14(d) shows the typical size of the error in broadcast ephemeris. Precise, post-fitted ephemerides are now readily available on the Internet from several sources, generally within two to four days. The broadcast ephemeris was compared with the post-fitted ephemeris obtained from the National Geodetic Survey (NGS). The pseudorange error was computed by projecting the ephemeris error vector on the line of sight. The discontinuities in the error

reflect the routine ephemeris updates at two-hour intervals. The observed ephemeris error did not exceed 2.5 m.

Figure 5.14(e) shows the remaining error in the measurements, identified as multipath and receiver noise. The antenna was mounted on the roof of MIT Lincoln laboratory in a relatively clean environment. The error is below 1 m, except for low satellite elevation angles, and shows quasi-sinusoidal oscillations with a period of several minutes, characteristic of multipath. The receiver did not have any special provisions to mitigate the multipath error. The code measurements, however, were smoothed using the carrier phase measurements, as discussed in Section 5.7, to achieve some reduction in the multipath error.

Our objective in introducing Figure 5.14 is simply to offer an appreciation for the nature and magnitude of the errors. As noted earlier, the errors show considerable variability. Uncompensated ionospheric and tropospheric delays would introduce a slowly changing bias in the measurements. We expect the receiver noise to be essentially uncorrelated from one measurement epoch to the next, but we cannot separate it from the multipath error. The combined effect has noise-like appearance, and can be mitigated to an extent by averaging, as discussed in the next section.

5.7 Combining Code and Carrier Measurements

While the carrier phase can be measured with great precision, the measurements include errors due to satellite clock and ephemeris parameter error, ionospheric and tropospheric refraction, and receiver clock bias just like the code-based pseudorange measurement. The error due to multipath and receiver noise in the carrier phase measurements, at centimeter level, is about one-hundredth of that in the pseudoranges. On the other hand, there is the extra complication due to the integer ambiguity. In this section, we examine the benefits of combining the noisy-but-unambiguous code measurements with the precise-but-ambiguous carrier phase measurements for absolute positioning in a single-receiver autonomous mode.

5.7.1 Single-Frequency Measurements

While carrier tracking is continuous, the carrier phase measurements offer precise and unambiguous measurements of change in pseudorange between two measurement epochs. In this section, we examine how to exploit the precision of these delta pseudoranges to smooth the code-based pseudoranges.

So far, we have represented the carrier phase measurement in units of cycles. When combining code and carrier measurements, it is more convenient to represent carrier phase measurements in units of length, like the code phase measurements. The measurements are

$$\rho(t) = r(t) + c[\delta t_u(t) - \delta t^s(t-\tau)] + I(t) + T(t) + \varepsilon_\rho(t). \tag{5.46}$$

$$\begin{aligned}\Phi(t) &= \lambda\phi(t) \\ &= r(t) + c[\delta t_u(t) - \delta t^s(t-\tau)] - I(t) + T(t) + \lambda N + \varepsilon_\Phi(t),\end{aligned} \tag{5.47}$$

Both measurements are from the same satellite and a reference to the satellite number has been dropped for simplicity. We follow loosely the simple and elegant treatment of Goad (1996). Let us define ionosphere-free (IF) pseudorange [Section 5.3.2] as

$$\rho_{IF}(t) = r(t) + c[\delta t_u(t) - \delta t^s(t-\tau)] + T(t)$$

We rewrite the code and carrier phase measurement equations as

$$\begin{aligned}\rho(t) &= \rho_{IF}(t) + I(t) + \varepsilon_\rho(t)\\ \Phi(t) &= \rho_{IF}(t) - I(t) + \lambda N + \varepsilon_\Phi(t)\end{aligned} \tag{5.48}$$

We have two equations in three unknowns (ρ_{IF}, I, N), and it is not clear so far that we have gained anything aside from the obvious notational simplicity. Figure 5.15 offers a conceptual view of the two measurements, illustrating the higher code measurement noise and the ambiguity of the carrier phase measurements. In view of the ambiguity, we have set the starting value $\Phi(t_0)$ arbitrarily to zero.

Let us examine the change in the code and carrier phase measurements between two measurement epochs t_{i-1} and t_i.

$$\begin{aligned}\Delta\rho(t_i) &= \rho(t_i) - \rho(t_{i-1}) = \Delta\rho_{IF}(t_i) + \Delta I(t_i) + \Delta\varepsilon_\rho(t_i)\\ \Delta\Phi(t_i) &= \Phi(t_i) - \Phi(t_{i-1}) = \Delta\rho_{IF}(t) - \Delta I(t_i) + \Delta\varepsilon_\Phi(t_i)\end{aligned} \tag{5.49}$$

where $\Delta\rho_{IF}$ is the change in the ionosphere-free pseudorange between the two measurement epochs. Similarly, ΔI is the corresponding change in the ionospheric delay, and $\Delta\varepsilon$ is the change in the error term.

How to combine these measurements so as to take advantage of the low-noise carrier phase measurements and unambiguous nature of the code measurements? We have two equations in two unknowns. The noise term in the first equation is at meter level, and in the second equation at centimeter level. We could weight the two equations appropriately and solve for the two unknowns: $\Delta\rho_{IF}(t_i)$ and $\Delta I(t_i)$. An alternative is to disregard ΔI, which would be near zero if the measurement epochs are close together, and use $\Delta\Phi(t_i)$ as an accurate estimate of $\Delta\rho_{IF}(t_i)$. If we can somehow get an accurate estimate of $\rho(t_0)$, we can construct the pseudorange profile using the carrier-derived delta pseudoranges (Figure 5.15).

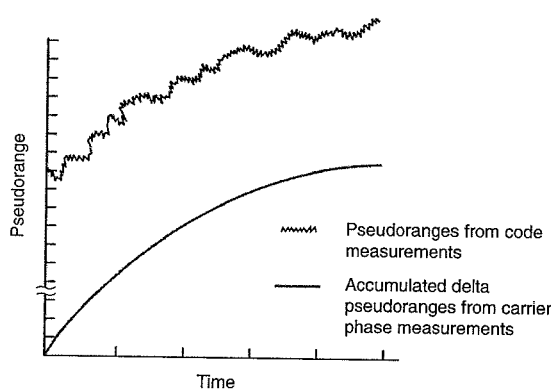


Figure 5.15 A conceptual view of the pseudoranges computed from the code and carrier phase measurements. The code-based measurements are noisy. The carrier-based estimates are precise but ambiguous, and the plot starts arbitrarily at zero value. The carrier phase measurements provide accurate estimates of delta pseudoranges, and can be used to smooth the code-based pseudoranges.

Actually, an estimate of $\rho(t_0)$, denoted as $\hat{\rho}(t_0)$, is available from each epoch as

$$\hat{\rho}(t_0)_i = \rho(t_i) - [\Phi(t_i) - \Phi(t_0)]$$

We can simply average these estimates over n epochs to obtain a smoothed estimate

$$\bar{\rho}(t_0) = \frac{1}{n}\sum_i \hat{\rho}(t_0)_i$$

and reconstruct the smoothed pseudorange profile as

$$\bar{\rho}(t_i) = \bar{\rho}(t_0) + [\Phi(t_i) - \Phi(t_0)]. \tag{5.50}$$

Carrier-smoothing of the code measurements is now routine in modern receivers and offers a modest improvement. The feature of 100-second smoothing was active in the receivers used to generate the plot of the multipath error in Figure 5.14(e).

An efficient implementation of the above idea is in terms of a recursive filter of length M as

$$\begin{aligned}\bar{\rho}(t_i) &= \frac{1}{M}\rho(t_i) + \frac{(M-1)}{M}[\bar{\rho}(t_{i-1}) + (\Phi(t_i) - \Phi(t_{i-1}))],\\ \bar{\rho}(t_1) &= \rho(t_1).\end{aligned} \tag{5.51}$$

The filter weighs the carrier phase measurements more heavily than the code phase measurements. Note that in deriving (5.50), we disregarded any change in the ionosphere between measurement epochs. While such change would generally be insignificant over a few seconds, we have to be careful that the change does not add up to become significant over M epochs. Actually, we will see double the effect of this change in filter (5.51) because the change is going in one direction for the carrier phase measurement and in the opposite direction for the code phase measurements. The smoothing filter (5.51) is the reason for our earlier discussion of *code-carrier divergence* [Section 5.3.1].

We know *a priori* that such smoothing of the code phase measurements cannot compensate for the errors due to satellite clock and ephemeris, ionosphere, and troposphere. We expect the receiver noise to be smoothed, and the code multipath to be smoothed to the extent of the filter length. The effect of such smoothing applied to measurements from a stationary receiver is shown in Figure 5.16. The raw and smoothed pseudorange errors are plotted for a satellite as it rises. (These measurements were taken in 1997 from a satellite with SA inactive.)

The measured raw pseudoranges in Figure 5.16, corrected using the broadcast ionospheric model and tropospheric model UNB4, are essentially unbiased with an rms error of about 4 m. The quasi-sinusoidal oscillations are characteristic of multipath. A 100-second smoothing filter reduces the rms error to 1.7 m. The residual error retains the character of multipath, and is similar to that in Figure 5.14(e). A 15-minute smoothing filter smoothes the multipath also, but a small bias becomes apparent. A 30-minute filter shows clear signs of code-carrier divergence in the middle of the time interval while the satellite is still low and rising, and the ionospheric propagation delay is changing. If the ionosphere had been more active, the divergence would have become manifest for a shorter filter length. In differential mode, however, the ionospheric propagation error is substantially eliminated, and such smoothing offers a robust technique to control measurement error, as discussed in Section 5.8.2.

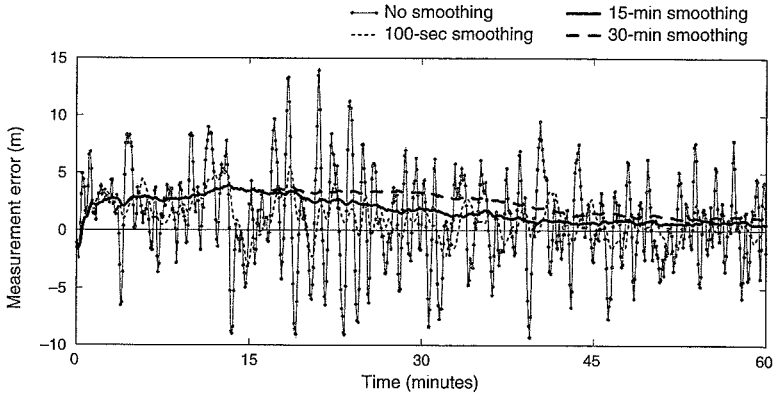


Figure 5.16 Carrier-smoothed pseudoranges with different filter lengths.

5.7.2 Dual-Frequency Measurements

Let us start by writing (5.48) for dual-frequency code and carrier phase measurements, using subscript q (= $L1$ or $L2$) to identify a measurement with L1 or L2. The ionospheric delay at L1 is denoted as I.

$$\begin{aligned}
\rho_{L1} &= \rho_{IF} + I + \varepsilon_{\rho_{L1}} \\
\rho_{L2} &= \rho_{IF} + \left(\frac{f_{L1}^2}{f_{L2}^2}\right) I + \varepsilon_{\rho_{L2}} \\
\Phi_{L1} &= \rho_{IF} - I + \lambda_{L1} N_{L1} + \varepsilon_{\Phi_{L1}} \\
\Phi_{L2} &= \rho_{IF} - \left(\frac{f_{L1}^2}{f_{L2}^2}\right) I + \lambda_{L2} N_{L2} + \varepsilon_{\Phi_{L2}}
\end{aligned} \tag{5.52}$$

Measurements from one epoch give four equations in four unknowns: (ρ_{IF}, I, N_{L1}, N_{L2}). If the measurements were error-free, we would be all set. Recall our model for the measurement noise: $\sigma(\varepsilon_{\rho_q}) \approx 1$ m, and $\sigma(\varepsilon_{\Phi_q}) \approx 1$ cm. The pseudorange measurements appear too coarse for estimation of the integers in a single snapshot of the measurements. On the bright side, N_{L1} and N_{L2} would not change from epoch to epoch as long as the carriers are tracked continuously. Each new epoch would bring four new measurements and introduce two new unknowns regardless of whether the receiver is stationary or in motion.

We can stack the measurements and process them sequentially or in batch, weighting code and carrier measurements appropriately, and expect the estimates of the integers to converge. In fact, this would be a good exercise using one of the data set on the enclosed CD. Once we have the integers, we can get precise ranges from the carrier phase measurements, and then position estimates, as discussed in Chapter 6.

Actually, the situation is trickier, but not hopeless, especially with the third civil frequency (L5) on the way. We'll pick up this discussion in Chapter 7 where we consider precise point positioning.

5.8 Error Mitigation: Differential GPS (DGPS)

The basic GPS measurements, described in Section 5.1, consist of biased and noisy estimates of ranges to the satellites. The principal source of bias is the unknown receiver clock offset relative to GPST. The remaining errors are:

- errors in modeling the satellite clocks and ephemerides,
- errors in modeling the ionospheric and tropospheric delays,
- errors in measuring the code and carrier phase due to multipath and receiver noise.

The carrier phase measurements involve an additional complication due to integer ambiguities of whole cycles. The multipath and receiver noise, however, are two orders of magnitude lower for the carrier phase than for the code measurements.

If a position is to be determined in real time using a single GPS receiver, the only option is to use code-based pseudoranges, perhaps smoothed by carrier phase. With $\sigma_{URE} \approx 6$ m, the horizontal and vertical positioning errors (95%) for an SPS user are about 10 m and 15 m, respectively. In order to obtain higher positioning accuracy, the user would have to reduce the measurement errors further. This requires a change in the mode of GPS usage from single-receiver autonomous positioning to differential GPS (DGPS), which we discuss next.

The basic idea behind DGPS is to take advantage of the fact that the errors associated with a satellite clock, ephemeris, and the atmospheric propagation are similar for users separated by tens, even hundreds of kilometers, and these errors vary 'slowly' with time. In other words, the errors exhibit spatial and temporal correlations. The closer the two users are to each other, and the closer the measurement epochs, the more similar are the errors listed above. The errors are said to get 'decorrelated' with increasing distance between the users and increasing time difference between their measurement epochs.

If the position of a GPS receiver is known, the combined effect of the errors can be estimated for each satellite. If these error estimates can be made available to the GPS users in the area, each can apply them to his measurements to mitigate the errors and improve the quality of the position estimates. That's the basic idea behind DGPS. This approach is usable with both code and carrier measurements. For navigation, such corrections will have to be made available in real time using a radio link. In practice, a user would receive and apply the corrections with some delay, called *latency*. The closer a user is to the reference station and the shorter the latency, the higher the benefit from the differential corrections.

5.8.1 Error Mitigation

Let us revisit the error sources discussed in Sections 5.2–5.4 to see how they would be mitigated in differential mode. It would be useful to review Figure 5.14 showing the nature of the errors introduced by these sources. Table 5.4 gives a summary of these errors and their mitigation in differential mode.

- *Satellite clock error.* The clock modeling error is small (2 m rms, and getting smaller), and changes slowly over hours. There is no fear of decorrelation with distance, and this error can be eliminated nearly completely in differential mode

with correction messages every hour.

- *Satellite ephemeris error.* Another small error (2 m rms, and getting smaller) which changes slowly over minutes. Recall that damage is done only by the line-of-sight component of the ephemeris error vector (Figure 5.5). The post-differential-correction residual error, therefore, would depend upon the separation between the lines of sight from the user and reference station. The satellites are 20,000 km away, and the angular separation at the satellite of the lines of sight from two points on the earth separated by 100 km is only 0.3°. A conservative bound on the residual range error is given as [Parkinson and Enge (1996)]

$$\varepsilon_R \leq \frac{d \cdot \delta r}{r}$$

where r is the range to a satellite, δr is the magnitude of the error in the ephemeris broadcast by the satellite, and d is the distance between the two receivers. For $d = 100$ km and $\delta r = 10$ m, the uncompensated range error would be less than 5 cm.

- *Ionospheric propagation delay.* Recall that the ionospheric delay depends upon the total electron content (TEC) along the signal path. Consider two receivers separated by 100 km. The corresponding ionospheric pierce points of the signals are also separated by about 100 km. The post-differential correction residual error, therefore, would depend upon the variability of TEC in the ionosphere. In fact, the ionosphere can show considerable variability both spatially and temporally resulting from solar activity and magnetic storm-induced traveling disturbances. For a satellite overhead, typical post-differential-correction residual error for two receivers separated by 100 km would be 0.1–0.2 m, but can be 1 m, or more, when the ionosphere is active.
- *Tropospheric propagation delay.* The tropospheric delay depends upon the air density profile along the signal path. Two receivers separated by several kilometers may be subject to different weather conditions. The water vapor content shows considerable variability both spatially and temporally. The post-differential-correction residual error is generally higher for the low-elevation satellites. With a 10-km separation between two receivers, the residual range error can be 0.1–0.2 m. For a longer distance or significant altitude difference, it would be preferable to correct for the tropospheric delay at the reference and user receivers separately. For a low satellite, the residual range error can be 2–7 mm per meter of altitude difference.
- *Multipath and receiver noise.* These errors are uncorrelated at the reference and user receivers, and cannot be corrected by DGPS. In fact, a user inherits the errors incurred at the reference station. Therefore, it is important to minimize these errors by careful siting and equipment selection at both the reference and user stations.

The design of the DGPS systems fielded in the 1990s was determined by SA, which introduced the largest and fastest-changing of the measurements errors. These systems typically compute new measurement corrections each five to ten seconds. In order to extend the life of a

Table 5.4 A summary of the errors in GPS measurements. The estimates for DGPS are based on the premise that the user's distance from the reference receiver is tens of kilometers and signal latency is tens of seconds.

Source	*Potential error size*	*Error mitigation and residual error*
Satellite clock model	Clock modeling error: 2 m (rms)	DGPS: 0.0 m
Satellite ephemeris prediction	Component of the ephemeris prediction error along the line of sight: 2 m (rms)	DGPS: 0.1 m (rms)
Ionospheric delay	Effect upon the code and the carrier is equal and opposite: The code is delayed while the carrier is advanced by the same amount.	Single-frequency receiver using broadcast model: 1–5 m
	Delay in zenith direction ≈ 2–10 m, depending upon user latitude, time of the day, and solar activity.	Dual-frequency receiver (compensates for the ionospheric delay but magnifies noise): 1 m (rms)
	Delay for a satellite at elevation angle *el* = zenith delay × obliquity factor (*el*)	DGPS: 0.2 m (rms)
	Obliquity factor: 1 at zenith; 1.8 at 30° elevation angle; and 3 at 5°.	
Tropospheric delay	Code and the carrier are both delayed by the same amount.	Models based on average meteorological conditions: 0.1–1 m
	Delay in zenith direction at sea level ≈ 2.3–2.5 m; lower at higher altitudes	
	Delay for satellite at elevation angle *el* = zenith delay × obliquity factor (*el*)	DGPS: 0.2 m (rms) plus altitude effect
	Obliquity factor: 1 at zenith; 2 at 30° elevation angle; 4 at 15°; and 10 at 5°	
Multipath	In a 'clean' environment:	Uncorrelated between antennas.
	Code: 0.5–1 m Carrier: 0.5–1 cm	Mitigation through antenna design and siting, receiver design, and carrier-smoothing of code measurements
Receiver noise	Code: 0.25–0.5 m (rms)	Uncorrelated between receivers
	Carrier phase: 1–2 mm (rms)	Mitigation through receiver design

correction message and cut down on the data traffic, the correction messages generally transmit both the error size and its rate of change observed at the reference station at the measurement epoch. With SA gone, the correction update interval can be increased from five or ten seconds to one minute, and the correction rate can be dropped from the message.

In a common realization of local DGPS, a reference receiver is set up at a surveyed site, and the pseudorange error is computed for each satellite as the difference between the geometric range and the measured pseudorange. Such errors are computed and broadcast in the local area on a radio link as differential corrections [see Figure 2.10]. Users within a hundred kilometers can obtain one-meter-level positioning accuracy by applying these corrections to their measurements.

5.8.2 Local-Area DGPS and Relative Positioning

Let us examine pseudorange measurements at the user and reference stations. We are still dealing with measurements from a single satellite, and will use superscript s, as necessary, to identify satellite-related terms. We have to introduce additional notation to keep the measurements of the two receivers apart. Following established practice, we will use subscripts u and r to identify the user and reference receivers, respectively. With a small change in the notation of (5.6), the pseudorange measurements at the reference and user receivers are

$$\begin{aligned}\rho_u &= r_u + c(\delta t_u - \delta t^s) + I_u + T_u + \varepsilon_{\rho u} \\ \rho_r &= r_r + c(\delta t_r - \delta t^s) + I_r + T_r + \varepsilon_{\rho r}\end{aligned} \tag{5.53}$$

The geometric range to a satellite from the reference station is computed as

$$r_r = \left\| \mathbf{x}^s - \mathbf{x}_r \right\|$$

where $\mathbf{x}^s$ is the satellite position obtained from the navigation message and $\mathbf{x}_r$ is the surveyed position of the antenna at the reference station. Any error in $\mathbf{x}^s$ would generally be harmless, as we'll see below.

For simplicity, we have dropped any reference to measurement epochs in (5.53) but let us consider measurements taken at the two receivers within a minute or two of each other. The error in ρ_r, the pseudorange measurement at the reference station, is computed as

$$\begin{aligned}e_r &= r_r - \rho_r \\ &= -c(\delta t_r - \delta t^s) - I_r - T_r - \varepsilon_{\rho r},\end{aligned} \tag{5.54}$$

and broadcast as a differential correction. The differentially corrected pseudorange measurement of a GPS user in the area is

$$\begin{aligned}\tilde{\rho}_u &= \rho_u + e_r \\ &\approx r_u + c(\delta t_u - \delta t_r) + (I_u - I_r) + (T_u - T_r) + \varepsilon_{\rho u} - \varepsilon_{\rho r} \\ &\approx r_u + c(\delta t_u - \delta t_r) + \varepsilon_{\rho,ur}\end{aligned} \tag{5.55}$$

The error introduced by the satellite clock bias term (δt^s) would be similar at the two receivers. If the distance between the two receivers is not 'too large,' and the corrections are not 'too old,' we can conclude that (i) the ephemeris error would affect the measurements at the two receivers similarly (see Figure 5.5), and (ii) $I_u \approx I_r$ and $T_u \approx T_r$. With two receivers separated by 25 km, the differential ionospheric delay is typically at 10–20 cm level. This difference can add up to 1 m at a distance of 100 km. A significant distance and/or difference between the altitudes of the two receivers will require that a tropospheric model be used to correct the measurements separately at each receiver. However, there is no reprieve from the errors introduced at the reference and user receivers by multipath and receiver noise, included in $\varepsilon_{\rho,ur}$.

In the simple treatment above, the effect of receiver clock bias at the reference station was absorbed in the corrections and transmitted to the user, resulting in a change in the apparent bias of the user receiver clock. In practice, a reference station would attempt to limit the size of the differential corrections to be transmitted by accounting for and removing its clock bias. It is important for the user and the reference receivers to coordinate how the measurements are to be processed. If the reference receiver applies the broadcast ionospheric delay model before computing the corrections, the user must do the same. Similarly, if the reference receiver uses a tropospheric delay model, so must the user. Both must also use the same ephemeris parameters. When the ephemeris parameters are updated, the reference station broadcasts corrections for a period using both the old and the new parameter sets, each identified with a unique Issue of Data (IOD) parameter. Finally, special care is required in locating a reference station to minimize multipath.

Two basic practical questions associated with DGPS are: How large an area can a reference station serve? And how often must the corrections be provided? As noted earlier, there are no clear-cut answers. Basically, the shorter the distance and the more frequent the corrections, the higher the assurance of obtaining position estimates with meter-level error. Keep in mind though that the errors due to the ionosphere and troposphere generally change slowly over minutes.

For real-time users of DGPS, the resources required are establishment of a reference station and a radio link to transmit data to the users. A standard format for data communication developed by the Radio Technical Commission for Maritime Services (RTCM) is now in wide use [RTCM SC-104 (1994)]. RTCM has defined data messages and an interface between the data link receiver and the DGPS receiver. The U.S. Coast Guard's Maritime DGPS system broadcasts differential corrections in the RTCM SC-104 format from marine radiobeacons at a modest data rate of 100–200 bps. These signals are available in coastal regions and along inland waterways of the United States for the price of a radiobeacon receiver and a differential-capable GPS receiver. Maritime DGPS is being expanded into *Nationwide DGPS* (NDGPS) to cover most of the country [Section 2.5]. Commercial services offering DGPS corrections using different radio links are available worldwide.

We have dealt so far with implementation of local DGPS in which the reference station was fixed at a surveyed location. It was a requirement for calculation of the corrections to the measurements. Actually, the concept of DGPS is useful even if the position of the reference station is not known accurately. Consider a simple case where the reference station has an unknown error in its position estimate. Clearly, the differential corrections computed at the reference station would be in error, and so would the computed position of the user based on these corrections. As we'll see in Chapter 6, the subsequent error in the user's estimated position would be approximately the same as the error in the reference station position. The relative position vector between the antennas of the two receivers would be largely unaffected. In other words, the user-computed position is relative to the position ascribed to the reference station [Hatch (1992)].

Errors in the carrier phase measurements due to satellite clock and ephemeris errors and ionospheric and tropospheric refraction can also be removed substantially in differential mode. The differentially corrected carrier phase measurements, however, differ from the similarly corrected pseudorange measurements in two important ways. On the positive side, the remaining error in the carrier phase measurements due to multipath (at centimeter level) and receiver noise (at millimeter level) is typically two orders of magnitude lower than that in the code-phase measurements. But there is a price: unlike the code phase measurements, the carrier phase measurements involve an inherent ambiguity. The basic idea of DGPS can be implemented using carrier phase measurements in the same way as with the pseudorange measurements: compute the corrections to be applied to the carrier phase measurements at the reference station, and transmit them to the users [Frodge *et al.* (1994), Lapucha *et al.* (1996)].

Relative Positioning

In an alternate implementation of DGPS, the reference station could broadcast its time-tagged measurements of pseudoranges rather than the computed differential corrections. In this case, the user receiver would form differences of its own measurements with those at the reference receiver satellite by satellite, and estimate its position relative to the reference receiver. This approach is referred to as *relative positioning*. Of course, knowledge of the reference receiver location would allow the user to determine his absolute position, if required. In some applications, for example, formation flying, the relative position vector is of main interest. Clearly, practical details of data transmission delays would have to be accounted for in generating real-time position and velocity estimates.

Differencing time-matched measurements from a satellite at the user and reference receivers results in cancellation of the same error terms as discussed earlier in this section. From (5.53), this difference can be written using a shorthand notation of a double subscript as

$$\begin{aligned}\rho_{ru} &= \rho_r - \rho_u \\ &= (r_r - r_u) + c(\delta t_r - \delta t_u) + (I_r - I_u) + (T_r - T_u) + (\varepsilon_{\rho r} - \varepsilon_{\rho u}) \\ &\approx r_{ru} + c(\delta t_r - \delta t_u) + \varepsilon_{\rho,ru}\,, \end{aligned} \tag{5.56}$$

where $r_{ru} = r_r - r_u$. We have a nonlinear equation in four unknowns: three elements of the relative position vector (hidden in the r_{ru} term) and the difference in the receiver clock biases. With measurements available from four or more satellites at the two receivers, the user can solve for the relative position vector. We'll use the double-subscript notation introduced in (5.56) throughout Part II to denote the difference between measurements at two receivers and between parameters in their models: $(\bullet)_{ru} = (\bullet)_r - (\bullet)_u$.

It is instructive to compare (5.56) with (5.6). We have essentially eliminated the ionospheric and tropospheric effects and ephemeris error in our new model. Rather than estimating $\mathbf{x}_u$ (absolute positioning), we are now content with estimating $(\mathbf{x}_r - \mathbf{x}_u)$ (relative positioning), which we can accomplish with higher accuracy because we only have to contend with errors due to receiver noise and multipath. Equation (5.56) represents a re-parameterization of the problem of positioning with code phase measurements (5.6).

The errors due to receiver noise and multipath are much lower in carrier phase measurements than in code phase measurements. Of course, the problem with carrier phase measurements has to do with the integer ambiguities. We can obtain much more precise estimates of relative position vector using carrier phase measurements if we can deal with the integer ambiguities. We'll develop this approach fully in Chapter 7. For now, we content ourselves with obtaining a partial benefit of the carrier phase measurements without getting tangled up in the integer ambiguity terms by smoothing the pseudorange differences (5.56) using the corresponding carrier phase differences.

The time-matched carrier phase measurements from a satellite in units of cycles measured at the user and reference receivers are

$$\begin{aligned}\phi_u &= \lambda^{-1}[r_u - I_u + T_u] + f \cdot (\delta t_u - \delta t^s) + N_u + \varepsilon_{\phi u} \\ \phi_r &= \lambda^{-1}[r_r - I_r + T_r] + f \cdot (\delta t_r - \delta t^s) + N_r + \varepsilon_{\phi r}\end{aligned} \tag{5.57}$$

The difference can be written using our double-subscript notation as

$$\begin{aligned}\phi_{ru} &= \phi_r - \phi_u \\ &\approx \lambda^{-1} r_{ru} + f \cdot (\delta t_r - \delta t_u) + N_{ru} + \varepsilon_{\phi,ru}\,, \end{aligned} \tag{5.58}$$

where $N_{ru} = N_r - N_u$ is the integer ambiguity associated with the differenced measurements. Equation (5.58) is the counterpart of (5.56). The main differences are the presence of an ambiguity (N_{ru}) and much lower measurement error ($\varepsilon_{\phi,ru}$) due to multipath and receiver noise in (5.58).

The basic idea is similar to that of carrier-smoothed code for single-receiver measurements discussed earlier in Section 5.7.1. We can take advantage of the much higher accuracy and precision of the pseudorange differences between the user and the reference station obtained from carrier phase to smooth the corresponding differences obtained from the code measurements. The integer ambiguities drop out. The error due to multipath and receiver noise at the two receivers in the single-difference code-based pseudoranges is smoothed. The smoothing filter can be written as before (5.51)

$$\begin{aligned}\bar{\rho}_{ru}(t_i) &= \frac{1}{M}\rho_{ru}(t_i) + \frac{(M-1)}{M}[\bar{\rho}_{ru}(t_{i-1}) + \lambda(\phi_{ru}(t_i) - \phi_{ru}(t_{i-1}))], \\ \bar{\rho}_{ru}(t_1) &= \rho_{ru}(t_1)\,, \end{aligned} \tag{5.59}$$

where $\bar{\rho}_{ru}$ is the smoothed pseudorange difference. Unlike the discussion on code smoothing in the previous subsection, code-carrier divergence poses much less of a threat to the filter above. The reason is that the ionospheric effects have been substantially eliminated in both the code and carrier single differences. The filter can be allowed to be long enough to smooth the oscillations in the pseudorange measurements due to multipath.

Figure 5.17 illustrates the effect of smoothing on L1 code measurement differences between two stations about 25 km apart. The differences of 'raw' code measurements are nearly unbiased with an rms error of 0.8 m. The 100-second smoothing reduces the rms error to 0.5 m, but the residual effect of multipath is apparent. The 15-minute smoothing filter effectively smoothes out the multipath. The results for a 30-minute filter are virtually identical to those for the 15-minute filter. Each reduces the rms error to about 0.1 m, and there is no sign of code-carrier divergence.

The architecture proposed for the FAA's Local Area Augmentation System (LAAS) to be

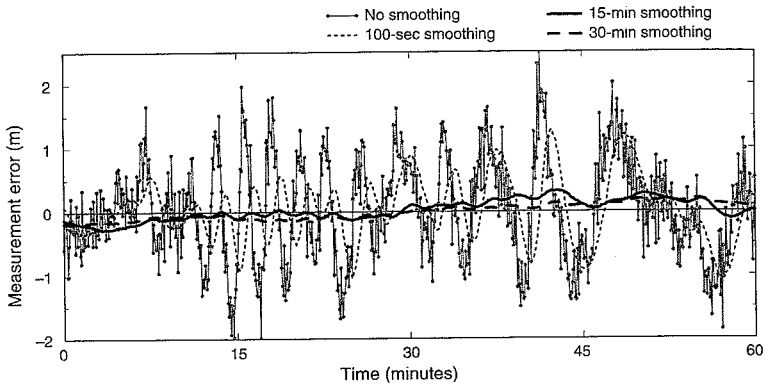


Figure 5.17 Carrier-smoothed differentially corrected pseudoranges obtained with different filter lengths.

used to guide aircraft during approach and landing operations under poor visibility is based on DGPS carrier-smoothed code measurements. The challenge lies in ensuring the accuracy of navigational guidance every time. The LAAS reference stations would be set up on airports and their measurements would be transmitted to aircraft on a VHF communications channel with a data rate of about 2 kbps. A comprehensive description of LAAS and the requirements of civil aviation is given by Enge (1999).

An idea of considerable interest in recent years has been to provide GPS-like signals from pseudo-satellites, called pseudolites, located on the ground, a ship, or aircraft. The pseudolites can be used in addition to deliver differential corrections for the GPS satellites by modulating the corrections onto the signals. Problems associated with the signal and receiver design to cope with the so-called near-far problem and the issues of multipath remain active areas of research [Elrod and Van Dierendonck (1996)].

5.8.3 Wide-Area DGPS

It would require a large number of autonomous DGPS systems to cover a continent-wide area. An alternative is wide-area DGPS (WADGPS) [Kee (1996)] implemented as a centralized system which provides differential corrections to users in a form so as to be usable over a large geographical area.

A set of reference stations is deployed over the region of interest and measurements from each are processed centrally to decompose the errors into their constituents: satellite clock, ephemeris, and ionosphere. The corrections are broadcast for each error source separately from geostationary satellites or from a network of FM stations so that each user can apply the correction terms appropriately depending upon her location. A WADGPS is said to broadcast differential correction vectors, in contrast to the scalar corrections in a local DGPS where all the errors are lumped together.

Commercial wide-area DGPS services are already operational and in wide use in offshore explorations, seismic surveys, and agriculture. A comprehensive description of the most ambitious wide-area DGPS to date, the FAA's *Wide Area Augmentation System* (WAAS), is given by

Enge *et al.* (1996). What makes WAAS special is the uncompromising requirement of safety in civil aviation. A schematic representation of WAAS is given in Figure 5.18. Dual-frequency (L1-L2) measurements from about 25 WAAS Reference Stations distributed over the conterminous United States are processed at the Master Station to estimate differential corrections and error bounds on them. The differential corrections are decomposed into three components: a fast-changing component due to the clock error and two slow-changing components due to the ephemeris error and ionospheric propagation delays for a set of points corresponding to a latitude/longitude grid.

The differential corrections are coded as a 250-bps navigation message of GPS/SPS-like signals transmitted at L1 frequency from geostationary satellites. WAAS actually generates such signals on the ground and transmits them to geostationary satellites, which essentially serve as 'bent pipes' transmitting the signals back to the earth over their area of coverage. A GPS receiver would require software modification in order to receive this additional ranging signal and to demodulate the navigation message for the differential corrections.

WAAS was commissioned in 2003. It offers positioning accuracy of 1.5–2 m in both horizontal and vertical dimensions, adequate for lateral and vertical (glide path) guidance for instrument approaches. The system also provides indications to the users where GPS or GPS/WAAS is unusable due to system errors or other effects.

WAAS-like GPS augmentations are also being developed by the European countries and Japan for deployment over regions of interest to them. These systems are called the *European Geostationary Navigation Overlay System* (EGNOS) and the *Multifunction Transportation Satellite (MTSAT)-based Satellite Augmentation System* (MSAS), respectively. Both are expected to be ready for operational use starting in 2006. The generic name for such augmentations is Satellite-Based Augmentation System (SBAS). International agreements have been negotiated so that WAAS, EGNOS, and MSAS together will provide a seamless coverage of the globe.

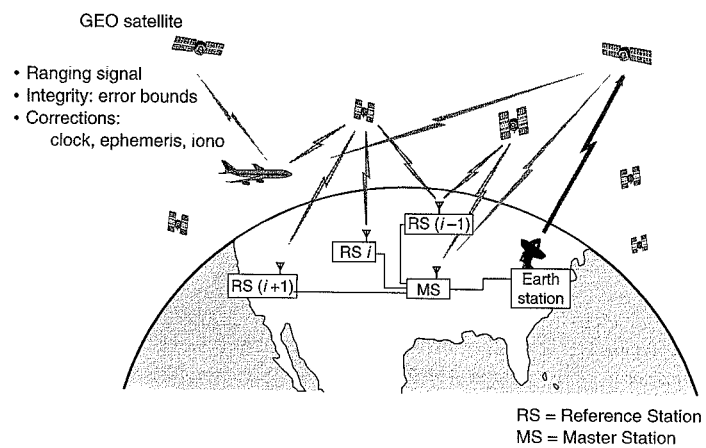


Figure 5.18 Wide Area Augmentation System (WAAS) schematic.

Another noteworthy wide-area DGPS system is NASA's Global Differential GPS (GDGPS) developed by the Jet Propulsion Laboratory (JPL) and capable of providing decimeter-level positioning accuracy to dual-frequency receiver users worldwide. The dual-frequency receivers, commonly used in low-earth orbiter missions, science applications, and high-accuracy commercial activities, have the advantage of eliminating the ionosphere as a source of error. GDGPS, therefore, focuses on providing accurate, real-time estimates of GPS ephemeris and clock parameters. Such estimates can be obtained from a relatively sparse network of global reference stations. Another innovation of GDGPS is to transmit the corrections over the Internet [http://www.gdgps.net]. (While GDGPS may offer high accuracy for gravity mapping over Greenland, it wouldn't land any airplanes in fog-bound BOS or SFO because, unlike WAAS, it wasn't designed to meet the civil aviation's requirements for integrity of each position estimates and continuity of service.)

5.9 Summary

After all kinds of preliminaries and generalities of the earlier chapters, we finally got down to brass tacks. In this chapter, we examined the GPS measurements and errors therein. Our ability to get the most out of GPS depends upon how well we understand what's being measured and how the various sources of error corrupt these measurements.

The measurement of code phases gives pseudoranges to the satellites, i.e., ranges with a common bias. Measurement of carrier phases gives pseudoranges with an additional complication of ambiguity of whole cycles. The bias and ambiguity aside, both measurements suffer from a common set of errors. The main sources of these errors are:

- Errors in satellite position and clock bias. Steady improvements in clock technology and orbital prediction algorithms, and a tight control of the constellation by the Master Control Station, have reduced these errors to 2–3 m level, an astonishing achievement.
- Errors in modeling the signal propagation delay through the ionosphere and troposphere. A signal travels through the earth's atmosphere with slightly varying speeds depending upon the composition of the medium. A dual-frequency receiver has the advantage of accounting almost fully for propagation delay through the ionosphere. When it comes to the troposphere, a dual-frequency receiver offers no advantage. Models can, however, be used to correct for most of the tropospheric delay.
- Errors due to receiver noise and multipath (i.e., antenna receiving the direct signal and one or more reflections).

The errors in satellite position and clock bias, and propagation delays are highly correlated both spatially and temporally. In other words, these errors would be substantially similar for two users located tens of kilometers apart and taking measurements within tens of seconds of each other. This observation offers a basis for augmenting the GPS signals with differential corrections to mitigate the error sources. Many differential GPS services are now in wide use worldwide offering meter-level positioning.

Homework Problems

5-1. Determine from Figure 5.2 if the receiver clock is running faster or slower relative to GPST, and estimate its frequency offset from the nominal value of 10 MHz.

5-2. Derive the relationship (5.19) between the phase and group refractive indices.

5-3. The problem of code-carrier divergence [Section 5.7.1] can be largely overcome with dual-frequency measurements as follows. From pseudorange measurements ρ_{L1} and ρ_{L2}, and carrier phase measurements ϕ_{L1} and ϕ_{L2}, 'create' two new measurements

$$\rho^* = \frac{f_{L1}}{f_{L1}+f_{L2}}\rho_{L1} + \frac{f_{L2}}{f_{L1}+f_{L2}}\rho_{L2}$$

$$\Phi^* = (\phi_{L1} - \phi_{L2})\frac{c}{f_{L1}-f_{L2}}$$

(a) Show that the wavelength of the signal $(\phi_{L1} - \phi_{L2})$ is $c/(f_{L1} - f_{L2})$, where c is the speed of light.

(b) Show that the ionospheric error in the two derived measurements is identical and, therefore, you can smooth the pseudorange measurements ρ^* by the carrier phase measurement Φ^* to your heart's content (to first order!).

5-4. Determine the ionospheric zenith delay from dual-frequency measurements over a 24-hour period, and observe the variability as a function of local time. Use the data set: (CD) *Data\Pigeon_Point\September_18_2000\Parsed\pp091800.rio*. The solar cycle was close to its peak on September 18, 2000, but there was no unusual solar activity on this day.

(a) Generate a scatter plot of ionospheric zenith delay versus local time at 15-minute intervals as follows. (i) Compute the slant delay from the dual frequency code measurements (5.30) from each satellite; (ii) determine the corresponding satellite position; (iii) use the Pigeon Point position file to determine the satellite elevation angle and the ionospheric obliquity factor; (iv) compute zenith delay.

(b) Compare the results from (a) with those obtained from the broadcast model. The broadcast model is described in the IS (CD) *Documents\IS-GPS200D.pdf*, pp. 125–128, and you'll find the needed coefficients in the header of file (CD) *Data\Pigeon_Point\September_18_2000\pp091800.nav*.

5-5. Repeat Problem 5-4, with data from July 15, 2000, a day of intense solar storm: (CD) *Data\Pigeon_Point\July_15_2000\Parsed*. Comment on the differences in the plots for the two days. (Caution: Watch for GPS week rollover.)

5-6. The CORS site at Durmid Hill provides meteorological data (pressure, temperature, and relative humidity) recorded ten minutes apart. These data can be used to estimate the tropospheric zenith delay.

(a) Use a model to estimate the tropospheric zenith delay (wet and dry) for a 24-hour period from the meteorological measurements from Durmid Hill in file (CD) *Data\Durmid_Hill\July_15_2000\dh071500.met*. Partial pressure due to water vapor, if required by your model, can be obtained from relative humidity and temperature as follows:

$$e = 6.108 \times RH \times \exp[(17.15T - 4684)/(T - 38.45)]$$

where e = water vapor pressure (mbars), RH = relative humidity (fraction), T = surface temperature (K = °C + 273.16). Did you expect a day-night variation?

(b) Use a mapping function to estimate the delay experienced by a signal at an arbitrary elevation angle. Plot tropospheric delay versus elevation angle.

5-7. To see the seasonal dependence of the tropospheric delay, repeat Problem 5-6(a) for a winter day: (CD) *Data\Durmid_Hill\January_6_2000\dh010600.met*. Is the difference roughly what you expected?

5-8. Inadvertent radio frequency interference (RFI) was experienced while we were collecting measurement files (CD) *Data\Stanford\RFI*. The symptoms can be seen in a plot of the number of tracked satellites against time. Can you tell when the RFI started? What was the effect on the measurements?

References

Braasch, Michael S., and A.J. Van Dierendonck (1999). GPS Receiver Architectures and Measurements, *Proc. IEEE*, vol. 87, no. 1, pp. 48–64.

Collins, Paul, Richard B. Langley, and J. LaMance (1996). Limiting Factors in Tropospheric Propagation Delay Error Modelling for GPS Airborne Navigation, *Proc. ION 52nd Annual Meeting*, pp. 519–528.

Counselman III, Charles C. (1999). Multipath-Rejecting GPS Antennas, *Proc. IEEE*, vol. 87, pp. 86–91.

Elrod, Bryant, and A.J. Van Dierendonck (1996). Pseudolites, in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 51–80.

Enge, Per, T. Walter, S. Pullen, C. Kee, Y. Chao, and Y. Tsai (1996). Wide Area Augmentation of the Global Positioning System, *Proc. IEEE*, vol. 84, no. 8, pp. 1063–1088.

Enge, Per (1999). Local Area Augmentation of GPS for the Precision Approach of Aircraft, *Proc. IEEE*, vol. 87, no. 1, pp. 111–132.

Feess, W.A., and S.G. Stephens (1987). Evaluation of GPS Ionospheric Model, *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-23, no. 3, pp. 332–338.

Fisher, Steven C., and Kamran Ghassemi (1999). GPS IIF – The Next Generation, *Proc. IEEE*, vol. 87, no. 1, pp. 24–47.

Goad, Clyde C. (1996). Single-Site GPS Models, in *GPS for Geodesy*, A. Kleusberg and P. Teunissen (eds.), Lecture Notes on Earth Sciences, Springer, pp. 219–237.

Hatch, Ron (1982). The Synergism of GPS Code and Carrier Measurements, *Proc. Third International Symposium on Satellite Doppler Positioning*, DMA, Las Cruces, NM, pp. 1213–1232.

Hatch, Ron (1992). Current Issues in Kinematic Navigation, *Proc. ION GPS-92*, pp. 1063–1068.

Hatch, Ronald R. (1996). Promise of a Third Frequency, *GPS World*, vol. 7, no. 5, pp. 55–58.

IS-GPS-200D (2004). Interface Specification IS-GPS-200, Revision D, Navstar GPS Space Segment/Navigation User Interfaces, Navstar GPS Joint Program Office [CD Documents\IS-GPS-200D.pdf]

Kee, Changdon (1996). Wide Area Differential GPS, in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 81–115.

Klobuchar, John A. (1996). Ionospheric Effects on GPS, in *Global Positioning System: Theory and Applications, Vol. I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 485–515.

Langley, Richard B. (2000). GPS, the Ionosphere, and the Solar Maximum, *GPS World*, vol. 11, no. 7, pp. 44–49.

Lapucha, Dariusz, Richard Barker, and Ziwen Liu (1996). High-Rate Precise Real-Time Positioning using Differential Carrier Phase, *Navigation*, vol. 43, no. 3, pp. 295–305.

Leick, Alfred, Ali Oufrid, Paulo Segantine, and Inseong Song (1994). Centimeter Navigation and Surveying with On-the-Fly Ambiguity Resolution, *Proc. ION 49th Annual Meeting*, pp. 495–504.

Malys, Stephen, M. Larezos, S. Gottschalk, S. Mobbs, B. Winn, W. Feess, M. Menn, E. Swift. M. Merrigan, and W. Mathon (1997). The GPS Accuracy Improvement Initiative, *Proc. ION GPS-97*, pp. 375–384.

McGraw, Gary A., Tim Murphy, Mats Brenner, Sam Pullen, A.J. Van Dierendonck (2000). Development of LAAS Accuracy Models, *Proc. ION GPS 2000*, pp. 1212–1223.

Misra, Pratap, Brian P. Burke, and Michael M. Pratt (1999). GPS Performance in Navigation, *Proc. IEEE*, vol. 87, no. 1, pp. 65–85.

Nolan, John, Sergei Gourevitch, and Jon Ladd (1992). Geodetic Processing Using Full Dual-Band Observables, *Proc. ION GPS-92*, pp. 1033–1041.

Parkinson, Bradford W., and Per Enge (1996). Differential GPS, in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 3–50.

Parkinson, Bradford W. (1996). GPS Error Analysis, in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 469–483.

RTCM SC-104 (1994). *RTCM Recommended Standards for Differential Navstar GPS Service*, version 2.1, Radio Technical Commission for Maritime Services.

Skidmore, Trent A., and Frank van Graas (2004). An Investigation of Tropospheric Errors on Differential GNSS Accuracy and Integrity, *Proc. ION GNSS-2004*, pp. 2752–2760.

Spilker, James J. (1996). Tropospheric Effects on GPS, in *Global Positioning System: Theory and Applications, Vol. I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 517–546.

Schupler, B.R., and T.A. Clark (1991). How Different Antennas Affect the GPS Observables, *GPS World*, vol. 2, no. 10, pp. 32–36.

SPS (2001). *Global Positioning System Standard Positioning Service Performance Standard*, U.S. Department of Defense [CD Documents\SPS Performance 2001.pdf].

van Diggelen, Frank (1997). GPS and GPS+GLONASS RTK, *Proc. ION GPS-97*, pp. 139–144.

Ward, Philip (1995). The Natural Measurements of a GPS Receiver, *Proc. ION 51st Annual Meeting*, pp. 67–85.

Weill, Lawrence (1997). Conquering Multipath: The GPS Accuracy Battle, *GPS World*, vol. 8, no. 4, pp. 59–66.

Wilson, Brian D., Colleen H. Yinger, William A. Feess, and Capt. Chris Shank (1999). The Broadcast Interfrequency Biases, *GPS World*, vol. 10, no. 9, pp. 56–66.

Chapter 6
PVT Estimation

In this chapter we consider estimation of position, velocity, and time (PVT) in real time based on measurements of pseudoranges and pseudorange rates. With such measurements available from four or more satellites, PVT can be estimated instantaneously. The problem is formulated in a general way and the treatment applies to both GPS measurements in autonomous mode (i.e., without any augmentation) and differentially corrected measurements.

As discussed in the previous chapter, the pseudoranges and pseudorange rates provided by a GPS receiver are noisy and biased measurements of user-satellite ranges and range rates, respectively. From these measurements, we would like to estimate PVT in an optimal way in some sense, or at least smartly. In this chapter, we use the least-squares approach to parameter estimation.

We use several basic concepts from probability theory. The student is expected to be familiar with histograms, probability density functions, cumulative distribution functions, multivariate Gaussian (or, normal) distributions. There is no dearth of excellent introductory

textbooks on probability theory [Brown and Hwang (1997), Helstrom (1991), and Stark and Woods (1994)]. The student is also expected to be comfortable with vector and matrix notation, matrix operations, and basic concepts of linear algebra. We recommend a review of the introductory sections in Strang and Borre (1997) dealing with linear algebra and least-squares solution of a system of linear equations.

In many navigation applications, it would be advantageous to base the current estimates of PVT on a time series of measurements, perhaps in combination with a model of the user dynamics. A Kalman filter implementation [Brown and Hwang (1997)] is commonly used to integrate measurements from GPS and an inertial measurement unit (IMU). Such a filter would also take into account any processing delays associated with the measurements to provide the current position and velocity estimates. In this chapter, we focus instead on estimates obtained from a snapshot of instantaneous GPS measurements. This approach simplifies the discussion of the principles of satellite navigation and allows for a cleaner characterization of the role of satellite geometry and measurement errors on the quality of the estimates.

The three main sections of this chapter deal with estimation of position, velocity, and time, in that order. The appendix presents some basic concepts from estimation theory relevant to this chapter.

6.1 Position Estimation with Pseudoranges

Let us start with the basic pseudorange measurement equation (5.6), rewritten below. We dealt in Chapter 5 with measurements from a single satellite and now need additional notation to distinguish among the measurements from K satellites in view. Following established practice, a superscript is used to identify the satellite. Parentheses are used to distinguish a superscript from an exponent. The pseudorange measurement from the kth satellite at epoch t (GPS time) can be modeled as

$$\rho^{(k)}(t) = r^{(k)}(t, t-\tau) + c[\delta t_u(t) - \delta t^{(k)}(t-\tau)] + I^{(k)}(t) + T^{(k)}(t) + \varepsilon_\rho^{(k)}(t) \qquad (6.1)$$

where $k = 1, 2, \ldots, K$. As discussed in Section 5.1, $r^{(k)}(t, t-\tau)$ is the actual distance between the receiver antenna at signal reception time t and the satellite antenna at signal transmission time $(t-\tau)$; $\delta t_u(t)$ and $\delta t^{(k)}(t-\tau)$ are the receiver and satellite clock offsets, respectively, relative to GPST; $I^{(k)}(t)$ and $T^{(k)}(t)$ are the ionospheric and tropospheric propagation delays, respectively; $\varepsilon_\rho^{(k)}(t)$ accounts for modeling errors (e.g., satellite clock modeling error and orbit prediction error) and unmodeled effects (e.g., receiver noise and multipath). Recall that we use ε to denote measurement errors, employing a subscript, superscript, overbar, or tilde to distinguish among the different scenarios we encounter. For simplicity, we now drop explicit reference to time in (6.1).

A user would correct each measured pseudorange for the known errors using parameter values in the navigation message from the satellite. The main corrections available to a civil user are: (i) satellite clock offset relative to GPST, (ii) relativity effect, and (iii) ionospheric delay using the parameter values for the Klobuchar model. Step-by-step implementation of these corrections is given in the GPS Interface Specification [IS-GPS-200D (2004)]. A resourceful user can go further in correcting the measured pseudoranges as follows.

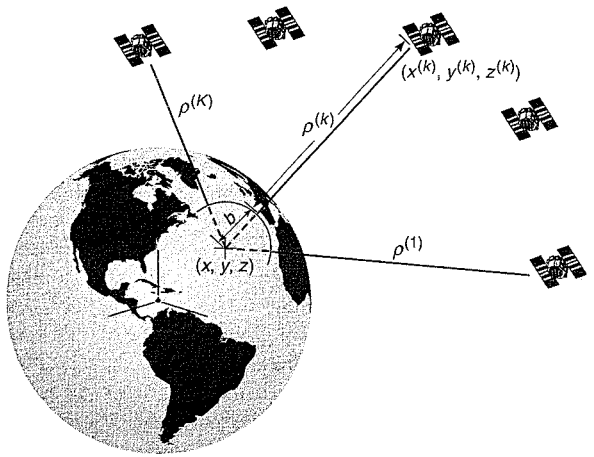


Figure 6.1 The ranges to GPS satellites measured by a receiver have a common bias and are called pseudoranges.

- The code measurements may be smoothed using the carrier phase measurements to reduce the effects of multipath [Section 5.7.1].
- The ionospheric delay can be eliminated as a source of error by a user equipped with an L1-L2 receiver [Section 5.3.2].
- The tropospheric delays can be accounted for in the software using a model [Section 5.3.3].
- The satellite position and clock errors and signal propagation delays may be mitigated if the user has access to corrections from a local- or wide-area DGPS service [Section 5.8].

We denote by $\rho_c^{(k)}$ the pseudorange obtained after accounting for the satellite clock offset and compensating for the remaining errors in the measurements to the extent it is practical for a user. We rewrite (6.1) for the 'corrected' pseudorange measurements simply as

$$\rho_c^{(k)} = r^{(k)} + c \cdot \delta t_u + \tilde{\varepsilon}_\rho^{(k)} \qquad (6.2)$$

where the error term $\tilde{\varepsilon}_\rho^{(k)}$ denotes the combined effect of the residual errors. The standard deviation of $\tilde{\varepsilon}_\rho^{(k)}$ can range typically from less than 1 m for differentially corrected measurements to about 6 m for measurements from a receiver in autonomous mode.

Let vectors $\mathbf{x} = (x, y, z)$ and $\mathbf{x}^{(k)} = (x^{(k)}, y^{(k)}, z^{(k)})$, for $k = 1, 2, \ldots, K$, represent the position of the user at the time of the measurement and the position of the kth satellite at the time of signal transmission. The user-to-satellite geometric range (Figure 6.1) is

$$r^{(k)} = \sqrt{(x^{(k)}-x)^2 + (y^{(k)}-y)^2 + (z^{(k)}-z)^2} = \left\|\mathbf{x}^{(k)} - \mathbf{x}\right\|$$

We can write (6.2) as

$$\rho_c^{(k)} = \left\| \mathbf{x}^{(k)} - \mathbf{x} \right\| + b + \tilde{\varepsilon}_\rho^{(k)} \tag{6.3}$$

where we have replaced the user clock bias term $c \cdot \delta t_u$ with a simpler b with units of meters. We use boldface, lower-case characters to denote vectors and boldface upper case to denote matrices; $\|\bullet\|$ denotes the magnitude of a vector.

Now, a subtle point. In computation of geometric range in (6.3), positions of the satellite and the user both must be expressed in the same coordinate frame, which could be either inertial or the revolving earth-centered, earth-fixed (ECEF) coordinate frame [Section 4.1]. The user position is best expressed in the ECEF frame of epoch t, the instant of signal reception and, therefore, we will express the satellite position in the ECEF frame of epoch t as well. The ephemeris parameters in the navigation message give us $\tilde{\mathbf{x}}^{(k)}$, the satellite position at $(t - \tau)$, the time of signal transmission, expressed in the ECEF frame of $(t - \tau)$. During the time it takes for the signal to propagate from the satellite to the user antenna, the earth and the ECEF coordinate frame would have rotated about the z-axis by $\omega_E \tau$, where ω_E is the rotation rate of the earth. So, we have to express the satellite position at $(t - \tau)$ in the ECEF frame of time t, and we have already discussed how this can be accomplished through a rotation matrix [Appendix 4.A].

$$\mathbf{x}^{(k)} = \begin{bmatrix} \cos\omega_E\tau & \sin\omega_E\tau & 0 \\ -\sin\omega_E\tau & \cos\omega_E\tau & 0 \\ 0 & 0 & 1 \end{bmatrix} \tilde{\mathbf{x}}^{(k)} \tag{6.4}$$

In 70–90 ms of signal transit time, the earth's rotation amounts to about 7×10^{-6} radians, and we can simplify above by using approximations: $\cos \omega_E \tau \approx 1$ and $\sin \omega_E \tau \approx \omega_E \tau$. The rotation angle may appear small but disregarding it can introduce an error of 10–20 m in the range computation in (6.3), and a corresponding east-west error in position estimate.

We are now ready to discuss position estimation.

6.1.1 Linear Model for Position Estimation

We have pseudorange measurements from K satellites, each modeled as a nonlinear equation (6.3). Each equation involves four unknowns: b and three components of $\mathbf{x}$. Clearly, four equations are required at a minimum ($K \geq 4$) to solve for the four unknowns. In other words, concurrent pseudorange measurements are required from at least four satellites in order to estimate the user's instantaneous position. The receiver clock bias is considered a nuisance parameter in the sense that in general a navigator has no need for its estimate. In fact, as noted in Section 5.1.1, the estimate is used within the receiver to schedule measurements. Once the receiver clock bias is known, the receiver can synchronize to GPST or UTC, and can generate, say, a one-pulse-per-second (1PPS) signal to control other equipment.

A simple approach to solving the K equations (6.3) is to linearize them about an approximate user position, and solve iteratively. The idea is to start with rough estimates of the user position and clock bias, and refine them in stages so that the estimates fit the measurements better. The approach described below is generally referred to as *Newton-Raphson method*. Let $\mathbf{x}_0 = (x_0, y_0, z_0)$ and b_0 be the first guesses of the user position and receiver clock bias, respectively. From (6.3), the corrected pseudorange measurement from satellite k is $\rho_c^{(k)}$. Let $\rho_0^{(k)}$ be

the corresponding approximation based on the initial guesses $\mathbf{x}_0$ and b_0.

$$\rho_0^{(k)} = \left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\| + b_0 \tag{6.5}$$

Let the true position and true clock bias be represented as $\mathbf{x} = \mathbf{x}_0 + \delta\mathbf{x}$ and $b = b_0 + \delta b$, where $\delta\mathbf{x}$ and δb are the unknown corrections to be applied to our initial estimates. We now develop a system of linear equations in which $\delta\mathbf{x}$ and δb are the unknowns to be determined.

$$\begin{aligned} \delta\rho^{(k)} &= \rho_c^{(k)} - \rho_0^{(k)} \\ &= \left\| \mathbf{x}^{(k)} - \mathbf{x}_0 - \delta\mathbf{x} \right\| - \left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\| + (b - b_0) + \tilde{\varepsilon}_\rho^{(k)} \\ &\approx -\frac{(\mathbf{x}^{(k)} - \mathbf{x}_0)}{\left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\|} \cdot \delta\mathbf{x} + \delta b + \tilde{\varepsilon}_\rho^{(k)} \\ &= -\mathbf{1}^{(k)} \cdot \delta\mathbf{x} + \delta b + \tilde{\varepsilon}_\rho^{(k)} \end{aligned} \tag{6.6}$$

where we have used a Taylor series approximation of a vector norm. In (6.6), $\mathbf{1}^{(k)}$ is the estimated line-of-sight unit vector directed from the initial estimate of the user position to satellite k, and $\mathbf{a} \cdot \mathbf{b}$ represents the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$. The elements of $\mathbf{1}^{(k)}$ are direction cosines of the vector drawn from the estimated receiver location to the satellite.

$$\mathbf{1}^{(k)} = \frac{1}{\left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\|} \left(x^{(k)} - x_0, y^{(k)} - y_0, z^{(k)} - z_0 \right)^T \tag{6.7}$$

A student may find it easier to derive (6.6) using long-hand notation as

$$\begin{aligned} &\sqrt{\left(x^{(k)} - x_0 - \delta x\right)^2 + \left(y^{(k)} - y_0 - \delta y\right)^2 + \left(z^{(k)} - z_0 - \delta z\right)^2} \\ &\approx \sqrt{\left(x^{(k)} - x_0\right)^2 + \left(y^{(k)} - y_0\right)^2 + \left(z^{(k)} - z_0\right)^2} \\ &\quad - \frac{\left(x^{(k)} - x_0\right)\delta x + \left(y^{(k)} - y_0\right)\delta y + \left(z^{(k)} - z_0\right)\delta z}{\sqrt{\left(x^{(k)} - x_0\right)^2 + \left(y^{(k)} - y_0\right)^2 + \left(z^{(k)} - z_0\right)^2}} \end{aligned}$$

which is written more compactly as

$$\left\| \mathbf{x}^{(k)} - \mathbf{x}_0 - \delta\mathbf{x} \right\| \approx \left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\| - \frac{(\mathbf{x}^{(k)} - \mathbf{x}_0)}{\left\| \mathbf{x}^{(k)} - \mathbf{x}_0 \right\|} \cdot \delta\mathbf{x}$$

The set of K linear equations (6.6) can be written in matrix notation as

$$\delta\boldsymbol{\rho} = \begin{bmatrix} \delta\rho^{(1)} \\ \delta\rho^{(2)} \\ \vdots \\ \delta\rho^{(K)} \end{bmatrix} = \begin{bmatrix} (-\mathbf{1}^{(1)})^T & 1 \\ (-\mathbf{1}^{(2)})^T & 1 \\ \vdots & \\ (-\mathbf{1}^{(K)})^T & 1 \end{bmatrix} \begin{bmatrix} \delta\mathbf{x} \\ \delta b \end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_\rho \tag{6.8}$$

We now have K (≥ 4) linear equations in four unknowns: $\delta\mathbf{x}$ and δb. We can write (6.8) more compactly as

$$\delta\boldsymbol{\rho} = \mathbf{G}\begin{bmatrix}\delta\mathbf{x}\\ \delta b\end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_\rho \tag{6.9}$$

where

$$\mathbf{G} = \begin{bmatrix}(-\mathbf{l}^{(1)})^T & 1\\ (-\mathbf{l}^{(2)})^T & 1\\ \vdots & \\ (-\mathbf{l}^{(K)})^T & 1\end{bmatrix} \tag{6.10}$$

is a $(K \times 4)$ matrix characterizing the user-satellite geometry. We will refer to $\mathbf{G}$ as the *geometry matrix*.

If $K = 4$, we can generally solve the four equations for four unknowns directly.

$$\begin{bmatrix}\delta\mathbf{x}\\ \delta b\end{bmatrix} = \mathbf{G}^{-1}\delta\boldsymbol{\rho}$$

A problem would arise if the equations are linearly dependent, making $\mathbf{G}$ rank-deficient. This could happen, for example, if the elevation angles of all satellites measured from the user position are the same. In this case the tips of the unit line-of-sight vectors would all lie in a plane, and rank($\mathbf{G}$) = 3. Similar problem would arise if the line-of-sight vectors were all contained in a plane. In this case, the satellites would all seem to line up in the sky view plot (Figure 4.16). With the GPS constellation of 24 or more satellites, such situations are rare, and change quickly.

If the sky is not obstructed, most likely $K > 4$, we have an over-determined system of equations. $\mathbf{G}$ would be full-rank, in general. We can now look for a solution which fits the measurements best in some sense [Appendix 6.A]. It is common to use the criterion of least-squares: The best solution $(\delta\hat{\mathbf{x}}, \delta\hat{b})$ is that which minimizes the sum of squared residuals (i.e., discrepancy in the equations when we plug in the estimated values for the unknowns):

$$\min \left\| \delta\boldsymbol{\rho} - \mathbf{G}\begin{bmatrix}\delta\hat{\mathbf{x}}\\ \delta\hat{b}\end{bmatrix} \right\|^2$$

We can use tools from calculus to find the stationary point of the above expression. Alternately, we can use concepts from linear algebra (orthogonality principle [Strang and Borre (1997)]) to find the least-squares solution for the corrections to our initial estimates from (6.9) as

$$\begin{bmatrix}\delta\hat{\mathbf{x}}\\ \delta\hat{b}\end{bmatrix} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\delta\boldsymbol{\rho} \tag{6.11}$$

The new, improved estimates of the user position and clock bias are

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{x}_0 + \delta\hat{\mathbf{x}}\\ \hat{b} &= b_0 + \delta\hat{b}\end{aligned} \tag{6.12}$$

The observation equations may now be linearized about these new estimates of the user position and clock bias, and the solution may be iterated until the change in the estimates is sufficiently small. Each iteration leads to new estimates of (i) GPS Time t at the time of signal reception, (ii) for each satellite, signal transit time τ, transmission time $(t - \tau)$, the corresponding satellite position, and the transformation (6.4), and (iii) geometry matrix $\mathbf{G}$. The estimates converge quickly, generally in two to four iterations, even if starting with initial estimates $\mathbf{x}_0 = \mathbf{0}$ and $b_0 = 0$.

What can we say about the quality of our least-squares estimates? The answer is: not much. A 'large' value of the sum of squared residuals would suggest that the estimates don't fit the measurements or the errors are large. 'Small' residuals might make us feel smug, but it's best to be skeptical. We expect the quality of the position estimates to depend upon the quality of the measurements, and we haven't said anything so far about the nature or the size of the errors in our measurements.

There was an implicit assumption in our least-squares solution above that all measurements are of equal quality. In practice, this assumption is almost never true. Measurements from low-elevation satellites would generally have larger errors than those from high-elevation satellites. There is no difficulty in accounting for the unequal quality of the measurements: We could simply weight the different measurement residuals appropriately, and weighting based on satellite elevation angle is often used [McGraw *et al.* (2000)]. It is left as an exercise [Homework Problem 6-1] to show that if we use weights $\{w_1, w_2, \ldots, w_k\}$ on the measurements, the weighted-least-squares solution of (6.11) is

$$\begin{bmatrix}\delta\hat{\mathbf{x}}\\ \delta\hat{b}\end{bmatrix} = (\mathbf{G}^T\mathbf{W}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}\,\delta\boldsymbol{\rho} \tag{6.13}$$

where $\mathbf{W}$ is a diagonal matrix with diagonal elements $\{w_1, w_2, \ldots, w_k\}$.

Actually, the measured pseudoranges are not only of unequal quality, they are also correlated [Tiberius *et al.* (1999)]. If the ionospheric and tropospheric delay models are used, the error in the zenith delay enters in all measurements in a proportional way, introducing correlations. The differentially corrected measurements would also be correlated. Roughly speaking, the stronger the (positive) correlations among the pseudorange errors, the lower the position error. In the limit, an error common to all measurements shows up in the clock bias estimate and is harmless to position estimation. This should be clear from (6.8). [We saw in Figure 5.14 how the measurement errors from a satellite were correlated over time (temporal correlation). Here we are talking about measurements at an instant from different satellites. For clarity, we should refer to correlations among such measurement errors as spatial correlation.]

Ideally, we would like a characterization of the errors in the pseudorange measurements in terms of a mean vector and a covariance matrix. With such structure, we can look for a minimum mean-square error estimate of the position and determine the variance of the position error. Such characterization of the measurement error, however, is very difficult, in general. For simplicity, we treat the errors in the measurements from different satellites as zero-mean, uncorrelated, and identically distributed. (In other words, we resort to estimation techniques which would be optimal if the measurements indeed conformed to this model.)

Our model for the measurement error in (6.2) is

$$\begin{aligned}E(\tilde{\boldsymbol{\varepsilon}}_\rho) &= \mathbf{0}\\ \mathrm{Cov}(\tilde{\boldsymbol{\varepsilon}}_\rho) &= E(\tilde{\boldsymbol{\varepsilon}}_\rho\tilde{\boldsymbol{\varepsilon}}_\rho^T) = \sigma_{URE}^2\mathbf{I}\end{aligned} \tag{6.14}$$

where $E(\bullet)$ denotes the mean or expected value, Cov(•) denotes covariance, $\mathbf{I}$ is the identity matrix, and σ_{URE} is the common standard deviation of the user range error (URE) for each of the satellites, which we will often denote in the sequel simply as σ. Strictly speaking, this assumption is unjustified in most cases. (Actually, the model (6.14) fitted the pseudorange measurements while SA was active. The SA-induced errors were generated at each satellite independently and were uncorrelated across satellites. And, with $\sigma_{SA} \approx 25$ m, the SA error was also much larger than the other errors and dominated them. While deactivation of SA has made the users happy, there is a downside for the analysts.)

Model (6.14) gives useful results, in general, as we'll see below [Section 6.1.5], but we have to be careful not to push the model too far. It's important to understand that the problem is not that we don't know how to estimate position 'optimally' from pseudorange measurements with correlated errors. The problem is that an accurate characterization of the correlations is very difficult and, rather than impose an uncertain, complicated structure on the problem, we prefer to simplify the structure and use the results with appropriate caution.

6.1.2 RMS Positioning Error

Our approach to derivation of a position estimate in the previous section was algebraic, formal, and entirely devoid of insight. We attempt to make up for it in this section. Let's return to the question: How good is a GPS-based position estimate? We expect that the larger the errors in the measurements, the worse the position estimate. The set of equations (6.8) is linear in the errors, and we can say that if the measurement errors double, so would the position error. We also expect that our position estimate would improve as the number of satellites in view increases, but the exact relationship appears unclear. Given the measurement errors, how much would the position estimate improve if the number of pseudorange measurements available went up from four to eight? Actually, the number of measurements does not tell us enough because a measurement brings additional information only if the corresponding equation is linearly independent of the previous equations. This leads us to conclude that the distribution of the satellites in the sky relative to a user is important, but how? We examine this question next.

Satellite Geometry

In order to gain some insight into the role of the user-satellite geometry, let us examine a simple two-dimensional example. A user measures his distance from a pair of stations S1 and S2 at known locations (Figure 6.2). If the range measurements were perfect, the user could determine his position exactly as lying at the intersection of two circles centered at S1 and S2, with the measured ranges as their radii. The measurements, however, are imperfect, and have an uncertainty of up to, say, $\pm\varepsilon$ in each case. Figure 6.2 illustrates how this measurement uncertainty translates into position uncertainty in the three cases with different geometries.

In Figure 6.2(a), the two stations are separated by an acute angle as seen from the user position. In Figure 6.2(b), this angle is 90°. In Figure 6.2(c), the angle is obtuse. While the quality of the range measurements is the same in the three cases, clearly the quality of the position estimates is not. The region of uncertainty, shown as shaded area, is smallest in case (b), and considerably larger in cases (a) and (c). The area of uncertainty is largest for case (c) because it combines the position uncertainty with position ambiguity (the two circles intersect at two points which are so close that an observer may not be able to rule one out). By the end

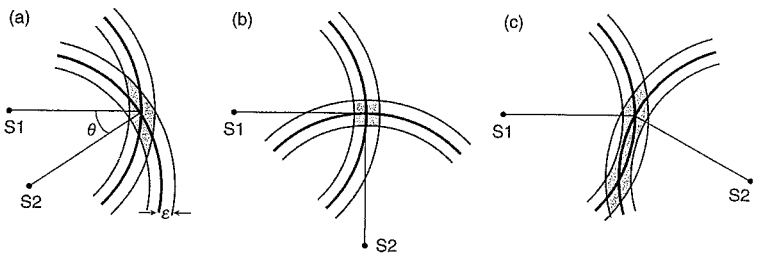


Figure 6.2 A simple 2-D example of position estimation with range measurements. The quality of the position estimate depends upon both the quality (ε) and geometry (θ) of the range measurements. The shaded region represents the uncertainty in the position estimate.

of this section, we will be ready to characterize the role of the measurement geometry in this problem.

Position estimation with GPS is analogous to the previous example.

We characterize the quality of the estimates in terms of the user-satellite geometry matrix $\mathbf{G}$. In our model, the pseudorange measurement errors are zero-mean and, therefore, errors in the estimates of the position coordinates and clock bias are zero-mean as well. Such estimates are said to be unbiased. We can now write down the mean and covariance of the errors in the position and clock bias estimates. Defining $\Delta\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$ and $\Delta b = \hat{b} - b$, where $\mathbf{x}$ represents the position coordinates and $\hat{\mathbf{x}}$ its estimate from (6.12); and b is the clock bias and $\hat{b}$ its estimate,

$$E(\Delta\mathbf{x}) = E(\hat{\mathbf{x}} - \mathbf{x}) = 0$$
$$E(\Delta b) = E(\hat{b} - b) = 0 \tag{6.15}$$

It is shown in Appendix 6.A that

$$\mathrm{Cov}\begin{bmatrix}\Delta\mathbf{x}\\ \Delta b\end{bmatrix} = \mathrm{Cov}\begin{bmatrix}\hat{\mathbf{x}}\\ \hat{b}\end{bmatrix} = \sigma^2\left(\mathbf{G}^T\mathbf{G}\right)^{-1} \tag{6.16}$$

We can now analyze these estimates $(\hat{\mathbf{x}}, \hat{b})$ component by component. Let σ_x^2, σ_y^2, and σ_z^2 denote the variances of the x-, y-, and z-component of the position error, respectively. Similarly, let σ_b^2 be the variance of the clock bias estimate. For simplicity, let

$$\mathbf{H} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1} \tag{6.17}$$

Denoting the ith entry on the diagonal of matrix $\mathbf{H}$ as H_{ii}, it follows from (6.16) that

$$\sigma_x^2 = \sigma^2 H_{11},\ \sigma_y^2 = \sigma^2 H_{22},\ \sigma_z^2 = \sigma^2 H_{33},\ \sigma_b^2 = \sigma^2 H_{44} \tag{6.18}$$

and

$$\begin{aligned}\text{RMS position error} &= \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}\\ &= \sigma\sqrt{H_{11} + H_{22} + H_{33}}\end{aligned} \tag{6.19}$$

Dilution of Precision (DOP)

Equations (6.18) and (6.19) show that the different components of the position estimates depend upon two factors: (i) variance of the user range error (σ^2), and (ii) a term which depends entirely on the user-satellite geometry (i.e., elements of matrix $\mathbf{H}$ (6.17) which depend entirely on matrix $\mathbf{G}$). The DOP parameters are defined on the basis of the above equations to characterize the contribution of the user-satellite geometry.

$$\text{Position dilution of precision (PDOP)} = \sqrt{H_{11} + H_{22} + H_{33}} \tag{6.20a}$$

$$\text{Time dilution of precision (TDOP)} = \sqrt{H_{44}} \tag{6.20b}$$

$$\text{Geometric dilution of precision (GDOP)} = \sqrt{H_{11} + H_{22} + H_{33} + H_{44}} \tag{6.20c}$$

The quality of the estimates obtained from a single snapshot of the measurements can be described simply (in units of length) as

$$\text{RMS (3-D position estimation error)} = \sigma \cdot \text{PDOP} \tag{6.21a}$$

$$\text{RMS (Clock bias estimation error)} = \sigma \cdot \text{TDOP} \tag{6.21b}$$

$$\text{RMS (3-D position and clock bias estimation error)} = \sigma \cdot \text{GDOP} \tag{6.21c}$$

We have now estimated the user position in the ECEF Cartesian coordinate frame and have expressions for the rms errors in its x-, y-, and z-components. The ECEF coordinates, however, are not easy for a user to relate to, and the user position is generally transformed to geodetic coordinates: latitude, longitude, and height. When it comes to assessment of a position error, it is generally more meaningful to a user to think in terms of horizontal and vertical components of the error defined relative to the local east-north-up (ENU) coordinate frame. As described in Appendix 4.A, a position vector can be transformed from ECEF to ENU with a (3×3) orthonormal matrix $\mathbf{R}_L$. The position error vector $\Delta\mathbf{x}$ defined in the ECEF coordinate frame can be represented in the ENU coordinate frame as

$$\Delta\mathbf{x}_L = \mathbf{R}_L \Delta\mathbf{x}$$

where $\Delta\mathbf{x}_L = (\Delta x_E, \Delta y_N, \Delta z_U)^T$. We use subscript L (for local) instead of *ENU* to keep the notation simple. We also define

$$\begin{bmatrix} \Delta\mathbf{x}_L \\ \Delta b \end{bmatrix} = \begin{bmatrix} \mathbf{R}_L & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta b \end{bmatrix} = \tilde{\mathbf{R}}_L \begin{bmatrix} \Delta\mathbf{x} \\ \Delta b \end{bmatrix} \tag{6.22}$$

The covariance matrix of $(\Delta\mathbf{x}_L, \Delta b)^T$ can now be obtained from that of $(\Delta\mathbf{x}, \Delta b)^T$ from (6.16).

$$\begin{aligned} \text{Cov}\begin{bmatrix} \Delta\mathbf{x}_L \\ \Delta b \end{bmatrix} &= \tilde{\mathbf{R}}_L \text{Cov}\begin{bmatrix} \Delta\mathbf{x} \\ \Delta b \end{bmatrix} \tilde{\mathbf{R}}_L^T \\ &= \sigma^2 \tilde{\mathbf{R}}_L (\mathbf{G}^T\mathbf{G})^{-1} \tilde{\mathbf{R}}_L^T \\ &= \sigma^2 (\tilde{\mathbf{R}}_L \mathbf{G}^T \mathbf{G} \tilde{\mathbf{R}}_L^T)^{-1} \\ &= \sigma^2 [\tilde{\mathbf{G}}^T \tilde{\mathbf{G}}]^{-1} \end{aligned} \tag{6.23}$$

where $\tilde{\mathbf{G}} = \mathbf{G}\tilde{\mathbf{R}}_L^T$ has a structure similar to $\mathbf{G}$. It is a $(K \times 4)$ matrix with each row made up of

three elements of direction cosine vectors represented in the local ENU coordinate frame, and an entry of 1 in the last column. It is left as an exercise to show that the unit vector of direction cosines in the ENU coordinate frame can be written as

$$\mathbf{1}^{(k)} = \left(\cos el^{(k)} \sin az^{(k)} \quad \cos el^{(k)} \cos az^{(k)} \quad \sin el^{(k)}\right)^T \tag{6.24}$$

where $az^{(k)}$ and $el^{(k)}$ are the azimuth and elevation angle of satellite k as observed from the user position [Appendix 4.A, Figure 4.A.4]. For simplicity of notation, let $\tilde{\mathbf{H}} = (\tilde{\mathbf{G}}^T\tilde{\mathbf{G}})^{-1}$, then

$$\text{Cov}\begin{bmatrix} \Delta\mathbf{x}_L \\ \Delta b \end{bmatrix} = \sigma^2 \tilde{\mathbf{H}}$$

and

$$\sigma_E^2 = \sigma^2 \tilde{H}_{11};\ \ \sigma_N^2 = \sigma^2 \tilde{H}_{22};\ \ \sigma_U^2 = \sigma^2 \tilde{H}_{33};\ \ \sigma_b^2 = \sigma^2 \tilde{H}_{44} \tag{6.25}$$

where σ_E, σ_N, σ_U are the standard deviations of the east, north, and up (vertical) components of the position error, respectively. The diagonal elements of $\tilde{\mathbf{H}}$, denoted as $\tilde{H}_{ii}$ above, correspond to DOP parameters: East DOP (EDOP), North DOP (NDOP), Vertical DOP (VDOP), and the previously defined TDOP.

$$\tilde{\mathbf{H}} = \begin{bmatrix} \text{EDOP}^2 & \bullet & \bullet & \bullet \\ \bullet & \text{NDOP}^2 & \bullet & \bullet \\ \bullet & \bullet & \text{VDOP}^2 & \bullet \\ \bullet & \bullet & \bullet & \text{TDOP}^2 \end{bmatrix} \tag{6.26}$$

where we have left the off-diagonal terms unspecified. Defining $\text{HDOP} = \sqrt{\tilde{H}_{11} + \tilde{H}_{22}}$ and $\text{VDOP} = \sqrt{\tilde{H}_{33}}$ we obtain

$$\text{RMS horizontal error} = \sqrt{\sigma_E^2 + \sigma_N^2} = \sigma \cdot \text{HDOP} \tag{6.27a}$$

$$\text{RMS vertical error} = \sigma_U = \sigma \cdot \text{VDOP} \tag{6.27b}$$

$$\text{RMS 3-D error} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2} = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2} = \sigma \cdot \text{PDOP} \tag{6.27c}$$

The DOPs provide a simple characterization of the user-satellite geometry. The more favorable the geometry, the lower the DOP. The lower the DOP and σ, the better the quality of the position estimate, in general. If a receiver is limited in the number of satellites it can track simultaneously, the user pays a price with larger rms position error due to higher DOP values. The early GPS receivers were limited to tracking only four satellites at a time, and PDOP or GDOP was a common basis for the selection of the 'best' four satellites. The newer receivers, even the inexpensive ones for the consumer market, now typically track all satellites in view.

With four satellites, PDOP is lowest if three satellites are evenly distributed in azimuth near the horizon and the fourth satellite is at zenith. As a rule, a low HDOP requires good azimuthal coverage with satellites that are not too high above the horizon. A low VDOP (or, PDOP) requires a satellite near zenith and good coverage in azimuth with satellites that are close to the horizon. Actually, VDOP would be lower if a receiver could track satellites below the horizon, as would be case with space-borne receivers [Spilker (1996)].

The clean characterization of position error in (6.27) in terms of the geometry and measurement error is based on our model (6.14). The pseudorange measurement errors are uncorrelated and have a common variance. If a user had a good model to estimate the covariance matrix Σ of the measurement error, we would have formulated the problem as one of weighted least squares, with Σ^{-1} as the weighting matrix. An expression like (6.27) could still be written, but the new definition of DOP parameters, reflecting the structures of both the geometry matrix and measurement error covariance matrix, would no longer be simple or intuitive. Actually, as noted earlier, an adequate characterization of the correlation structure is difficult, and is generally sidestepped in favor of using (6.14) for broad guidance on the quality of the position estimates.

The role of DOP in GPS positioning is often misunderstood. A lower DOP value does not automatically mean a lower position error, and vice versa. To illustrate this point, Figure 6.3 gives a scatter plot of HDOP and the corresponding horizontal position error computed from measurements taken three minutes apart over a three-month period early in 1993. SA was active during this period and our data model (i.e., uncorrelated measurements with equal variance) actually fitted the data. GPS only had a partial satellite constellation during this period and the HDOP values at times were high. The main point of this figure is to illustrate that the higher the HDOP value, the bigger the scatter in position error, and you are as likely to obtain a low-error position estimate as one with high error. If we group the points by HDOP ranges and compute the rms error for each range bin, shown as diamonds, we see that the relationship between the rms position error and HDOP is in fact linear. This is an empirical validation of the relationship (6.27a).

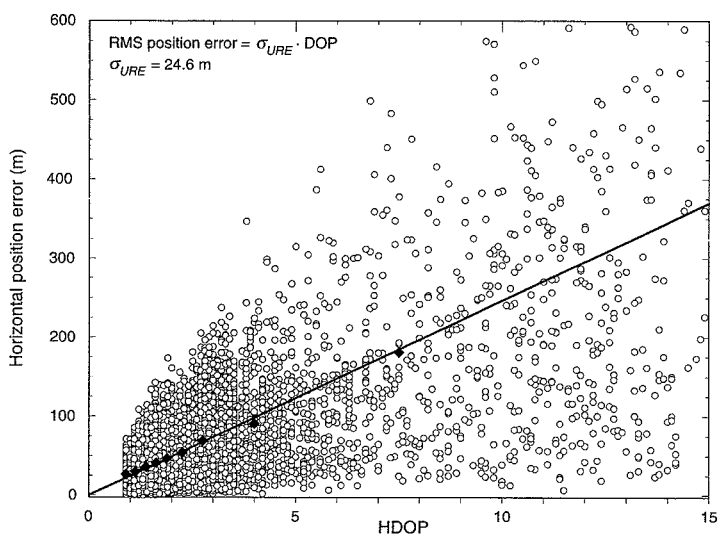


Figure 6.3 Scatter plot of the horizontal position error versus HDOP for measurements taken at three-minute intervals over a three-month period in 1993 while SA was active. The relationship between the rms horizontal position error and HDOP is linear. The slope of the line is σ_{URE}.

$$\text{RMS horizontal position error} = \sigma_{URE} \cdot \text{HDOP}$$

The slope of the line fitted to the computed rms values of position error in the different bins is σ_{URE}, the standard deviation of the user range error. This scatter plot thus gives us a way to estimate σ_{URE}. The estimate is 24.6 m, which is close to our model of σ_{URE} = 25 m while SA was active.

Now we return to the question with which we began this section: How good is a GPS-derived position estimate? The answer is that the position error depends upon the measurement geometry and pseudorange measurement errors. Given a simple model of the measurement error (6.14), the best we can do is to characterize the rms position error as in (6.27) in terms of a DOP parameter and the rms measurement error. In order to be able to say more about the position error, we would need to know more about the measurement errors.

If the pseudorange errors could be modeled as Gaussian, the error in x-, y-, and z-component of the position estimate would be Gaussian as well. We would know not just the position error statistics (i.e., mean and standard deviation), but the error distribution. We could determine the median horizontal error, and the 95th percentile of the vertical error distribution. We could even determine the 99.999th percentile of the position error, but that would be going against our dictum of pushing a simple model too far.

Distribution of DOPs

The satellites are in motion and DOP values change with the changing user-satellite geometry. To illustrate the pattern of such change, DOP values measured over a twenty-four-hour period from a roof-top antenna at MIT Lincoln Laboratory, Lexington, Massachusetts, early in 1997 are shown in Figure 6.4. The GPS constellation comprised 25 satellites, and the receiver

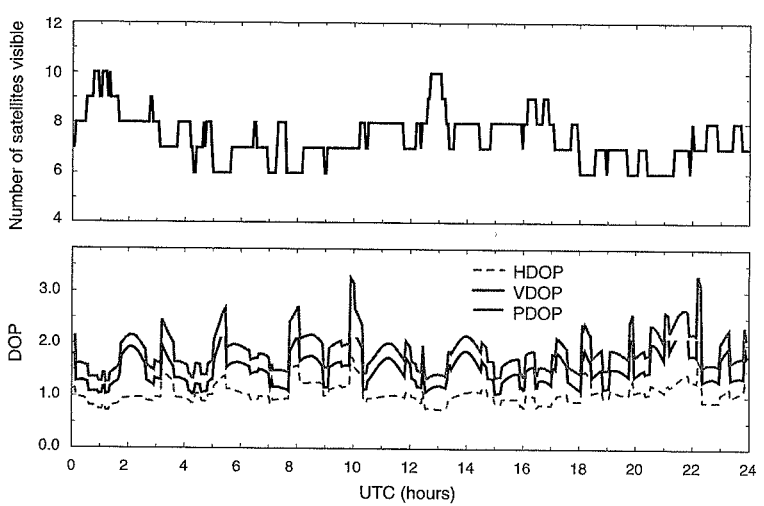


Figure 6.4 Number of satellites in view and the corresponding DOPs observed at Lexington, Massachusetts, early in 1997 with a 25-satellite constellation.

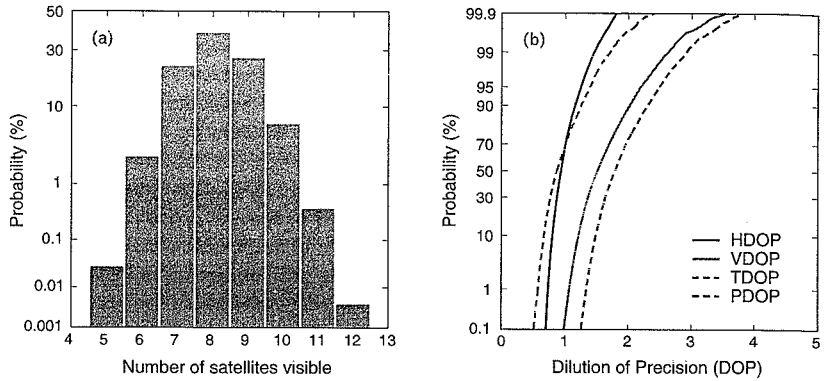


Figure 6.5 (a) Histogram of the number of satellites in view of users worldwide with an unobstructed view of the sky (elevation angle >5°) for the 24-satellite baseline constellation. (b) The corresponding cumulative distribution functions of HDOP, VDOP, PDOP, and TDOP values.

tracked all satellites in view. As is typical, the HDOP values are consistently smaller than the VDOP values. The distribution of HDOP would be roughly similar at other locations. VDOP, however, tends to be larger at higher latitudes (north or south). The reason is that the inclination of the GPS orbits is 55°, and at higher latitudes there are no satellites overhead.

In view of (6.27), we can characterize the distribution of the positioning error in terms of probability distributions of the DOP values. Such a global characterization is offered in Figure 6.5 for the baseline 24-satellite GPS constellation on the basis of a simulation.

Figure 6.5(a) is a histogram of the number of satellites in view at elevation angles higher than 5° for users distributed worldwide. Figure 6.5(b) gives the corresponding cumulative distribution functions of HDOP, VDOP, TDOP, and PDOP. Note in Figure 6.5(a) that with the baseline 24-satellite GPS constellation, about 90% of the users with an unobstructed view of the sky see seven to nine satellites, and nearly all see five or more satellites. The quality of these position estimates would depend upon the corresponding DOPs. The HDOP and TDOP values are typically close to one, and generally do not exceed two; VDOP tends to be larger than HDOP. With the full GPS constellation, the median and 99th percentile of the HDOP distribution are 1 and 1.6, respectively; the corresponding values for VDOP are 1.6 and 3.7, respectively. We should note in passing that (i) the GPS constellation has had more than 25 working satellites continuously since 1995, and (ii) an unobstructed view of the sky is a luxury for most navigators while on land.

What's DOP Good For?

In the 1980s, there was only a partial satellite constellation and GPS coverage was spotty. The receivers were also less capable, generally limited to tracking four satellites. The early researchers, therefore, had to schedule their data collection experiments on the basis of predicted DOPs in order to get useful measurements. The best time for GPS observations was invariably between 2 a.m. and 4 a.m., regardless of the user location, or that's how the old-timers remember it. In any case, that's how DOP came to acquire its considerable importance in the GPS user community.

Now with 24-plus satellites in the constellation and most receivers tracking all satellites in view, HDOP is seldom greater than one, unless parts of the sky are obstructed from view. The urban users and hikers who commonly face such situations have to be advised of poor satellite geometry and its consequences for positioning, but it seems pointless to quote DOP to them. The concept of DOP, however, is a valuable tool in design and analysis of a satellite constellation, and as a diagnostic tool to indicate receiver malfunction.

6.1.3 Price of an Inexpensive Receiver Clock

Clocks are at the heart of GPS. As discussed in Section 4.2.4, the satellites carry on board atomic frequency standards which are extraordinarily stable. The widespread use of GPS, however, is made possible by the fact that the receivers can work with inexpensive quartz oscillators not much better than those common in wristwatches. Geodetic-quality receivers may use temperature-compensated crystal oscillators (TCXO), which are a bit larger and use more power. The use of atomic standards in GPS receivers is rare, generally limited to monitoring stations and a few reference stations used for research purposes.

In our discussion so far, it has been understood implicitly that the price of using an inexpensive receiver clock is simply that we cannot estimate 3-D position from three satellites, and need a minimum of four satellites in view. Actually, there is more to it. There is a price in terms of the quality of the position estimates. If the clock bias were known and we could measure ranges rather than pseudoranges to the satellites, we would obtain better position estimates.

The vertical error in GPS-based position estimates tends to be larger than the horizontal error. This can be inferred from the plots of HDOP and VDOP values in Figures 6.4 and 6.5. The discrepancy between the quality of the horizontal and vertical position estimates is often explained by an appealing intuitive argument as follows. The horizontal position estimates are better because the satellites are on all sides of a user. The vertical position estimates suffer by comparison because all the satellites are above the user, and none below.

Indeed, if the earth were a transparent bubble allowing each user to see all GPS satellites, a simple and rough argument based on symmetry would lead us to conclude that the error in each of the north, east, and vertical directions would be similar. The blockage of satellites below a user's horizon introduces asymmetry which puts the estimation of vertical position at a disadvantage. In fact, the consequence of this geometrical asymmetry would be relatively small if the receiver clock bias were known *a priori*. It is left as an exercise to show that the VDOPs would be reduced significantly, and would be generally lower than the corresponding HDOPs, if the receiver could measure true ranges to the satellites rather than pseudoranges [Misra (1996)]. Our earlier definition of DOPs was for a 4-D estimation problem dealing with three unknowns of the user position and the receiver clock bias. If the receiver clock bias is known, the problem reduces to 3-D estimation, and the definition of DOPs would be amended appropriately, following the same lines of treatment as given earlier in this section.

It may appear at first that the idea of a receiver clock with a known bias is strictly academic. Not true. Note that we didn't require that the bias remain constant, only that it be known accurately. It would be enough if the clock behavior could be predicted, given the recent past. This would require that the clock bias change smoothly, without any corners or jumps. Actually, this is a rather stringent requirement as the timekeeping characteristics of a clock can change with temperature, pressure, and acceleration. An ingenious, alternate approach to 'clock aiding' to improve DGPS navigation accuracy is described by Hwang *et al.* (2005).

6.1.4 Other Performance Measures and Specifications

In the previous section, we characterized the rms position error for a user in terms of the standard deviation of the pseudorange measurement error (σ), and a DOP parameter (which depends upon the satellite geometry). The measurement errors change over time. The rms value of the measurement errors would depend upon the mode of usage (autonomous, differential, or aided). The satellite geometry changes as the satellites rise, move across the sky, and set. Since the GPS constellation essentially repeats its ground tracks from one day to the next (actually, after 23h 56m), the user would see the same geometries repeat. The performance of GPS, therefore, changes with place and time, and is said to be dynamic.

Our discussion so far has focused on *absolute accuracy*: How well a position estimate matches what we know to be the right answer. If we didn't know the right answer, we may want to evaluate the position estimates in terms of their repeatability. We could call this *repeatable accuracy*. (A fisherman may simply want to be led back to the same spot without regard to the position coordinates). There are still other ways to think of accuracy, but we'll limit ourselves to absolute accuracy.

In order to characterize the global performance of GPS, we have to take into account the position errors of users at all places at all times. Different users employ receivers of different capabilities and some users operate in areas with significant sky blockage or are hampered by severe multipath. The satellite constellation changes, too, as satellites are added and removed or temporarily sidelined for maintenance. For global specifications, it is necessary to limit the scenarios and define average performance predicated upon a number of conditions [SPS (2001)]: 24-satellite constellation; clear view of the sky for elevation angles higher than 5°; at least four satellites in view with PDOP ≤ 6; and performance averaged over twenty-four hours. The error sources can also be limited only to those for which the Control Segment can be held responsible. The resulting so-called signal-in-space (SIS) accuracy specifications, cited in Table 2.1, are to be understood in the context of these conditions.

There are several ways to characterize positioning accuracy. Given a mathematical model of the measurements (6.9) and a characterization of the measurement error (6.14), we determined mean error, standard deviation, and rms error in 1-D, 2-D, and 3-D earlier (6.27). We can also determine the various percentiles of the error distributions if we assume further that the errors follow a Gaussian distribution. Alternatively, the error statistics can be computed empirically on the basis of a sample of position estimates. Let Δx_j, Δy_j, and Δz_j, $j = 1, 2, \ldots, n$, be the errors in the east, north, and up components of the jth position estimate sample. RMS horizontal (2-D), vertical, and 3-D errors are defined as

$$\text{RMS vertical error} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\Delta z_j^2} \tag{6.28a}$$

$$\text{2-D rms error} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\Delta x_j^2 + \Delta y_j^2\right)} \tag{6.28b}$$

$$\text{3-D rms error} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\Delta x_j^2 + \Delta y_j^2 + \Delta z_j^2\right)} \tag{6.28c}$$

The more extensive the sample, the more faith we can have in the measure of the quality of a position estimate.

The rms error is sometimes represented as synonymous with standard deviation of the error. That's incorrect. The two terms are interchangeable only if the mean position error along each component is zero (i.e., the estimates of position coordinates are unbiased). For GPS position estimates, mean error equals zero if averaged over a long time interval. Position estimates averaged over seconds or minutes are generally biased because the pseudorange measurement errors due to satellite clock, orbit, and propagation effects change slowly.

If the position estimates are modeled as unbiased, the above expressions for rms error can be written in terms of standard deviations of the east, north, and vertical components of position, introduced earlier as σ_E, σ_N, and σ_U, respectively, as follows.

$$\text{2-D rms error} = \sqrt{\sigma_E^2 + \sigma_N^2} \tag{6.29a}$$

$$\text{RMS vertical error} = \sigma_U \tag{6.29b}$$

$$\text{3-D rms error} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2} \tag{6.29c}$$

The 50th and 95th percentiles of the horizontal, vertical, and 3-D errors are used commonly, and are easily computed. The median (50th percentile) horizontal error, called *circular error probable* (CEP), defines the radius of a circle centered at the true position which would contain the position estimate with probability of 0.5. *Spherical error probable* (SEP) is defined similarly as median 3-D error. The 95th percentiles are often written as horizontal error (95%) and vertical error (95%) as shorthand. They are also sometimes referred to as error values at 95% confidence level. (Actually, confidence level has a different, well-defined meaning in statistics, and we don't recommend this usage)

Another measure of the quality of the position estimates in common use is *2drms*, defined as twice distance rms in 2-D (or, twice the rms horizontal error, 2 × 2-D rms). It's a confusing nomenclature which now appears entrenched. To make matters worse, 2drms is often used incorrectly as equivalent to the 95th percentile of the horizontal error.

The system requirements and performance of positioning systems are often expressed by different, often inconsistent, parameters. For example, the original GPS positioning specifications were: 100 m (2drms) for the SPS users, and 16 m (SEP) for the PPS users. That's all. Different user groups choose to specify their requirements in terms of 2drms, CEP, SEP, and 95th percentile. Given one measure of performance, we can determine another. Such 'translation' of the performance measures often requires simplifying assumptions which, though not strictly true, serve as a useful model if not pushed too far. We follow van Diggelen (1998).

Our model for GPS position estimates is:

- The position estimates follow a multivariate normal distribution centered at the true position.
- Errors in the east, north, and up (E, N, and U) components are uncorrelated. (It's a reasonable assumption for low DOPs.) The standard deviation of the north and east errors are roughly equal: $\sigma_E \approx \sigma_N$. The standard deviation of the vertical error is roughly twice the rms horizontal error: $\sigma_U \approx 2(\sigma_E^2 + \sigma_N^2)^{1/2}$.

With the above assumptions, the various performance measures can be related as follows using the basic properties of the normal and chi-squared distributions.

(i)	RMS vertical error	$\approx 0.5 \times$ vertical error (95%)	(6.30a)
		$\approx$ 2drms	(6.30b)
		$\approx 0.9 \times$ 3-D rms error	(6.30c)
(ii)	2-D rms error	$= 0.5 \times$ 2drms error	(6.31a)
		$\approx 0.6 \times$ horizontal error (95%)	(6.31b)
		$\approx 1.2 \times$ CEP	(6.31c)
(iii)	3-D rms error	$\approx 2.2 \times$ 2-D rms error	(6.32a)
		$\approx 1.2 \times$ horizontal error (95%)	(6.32b)
		$\approx 1.3 \times$ SEP	(6.32c)

6.1.5 Empirical Positioning Results

In this section, we examine the typical positioning results obtained from GPS. The objective is to offer a broad appreciation for the variability in the position error in different modes (autonomous and differential). The measurements were taken with stationary rooftop antennas at surveyed locations at Stanford University and MIT Lincoln Laboratory and the receiver tracking all satellites in view. Each of the error plots of Figures 6.6 and 6.7 below presents horizontal position error computed from pseudorange measurements taken one minute apart over a 24-hour period early in 2004.

SPS Position Estimates

Representative positioning results from GPS Standard Positioning Service (SPS) are given as scatter plots in Figure 6.6. Figure 6.6(a) presents positioning results from an inexpensive and popular receiver designed for outdoorsmen. As expected, the user's manual isn't too specific about how the receiver treats the ionospheric or tropospheric errors. The receiver computes and displays 3-D position estimates as latitude, longitude, and height, and can record these data internally. The results are plotted in Figure 6.6(a) as north error and east error. The pixellated look of the scatter plot comes from the limited precision of the internal recording: 0.00001 degrees for latitude and longitude (and 0.1 feet in height). The median error is about 4 meters and the 95th percentile point is below 7.5 meters. The performance is impressive, and more than adequate for a hiker. The performance far exceeds the specifications [Table 2.1].

Figure 6.6(b) shows SPS results obtained with a more capable receiver which recorded code and carrier phase measurements and the navigation messages. The C/A-code pseudorange measurements were corrected for propagation errors using the models discussed previously. The ionospheric delay was estimated from the broadcast parameter values of the Klobuchar model [Section 5.3.2]. The tropospheric delay was estimated from the model UNB4 [Collins *et al.* (1996)]. We didn't, however, use the carrier phase measurements to smooth the pseudoranges [Section 5.7.1]. The performance is significantly better than that for the pocket receiver discussed above. The median error is 2 meters, and the 95th percentile point is 4 meters.

It is interesting to note in Figure 6.6 that the east and north position errors show no obvious correlation, thus validating our model in Section 6.1.4. Though not shown, there is no correlation apparent between the vertical and horizontal position errors either.

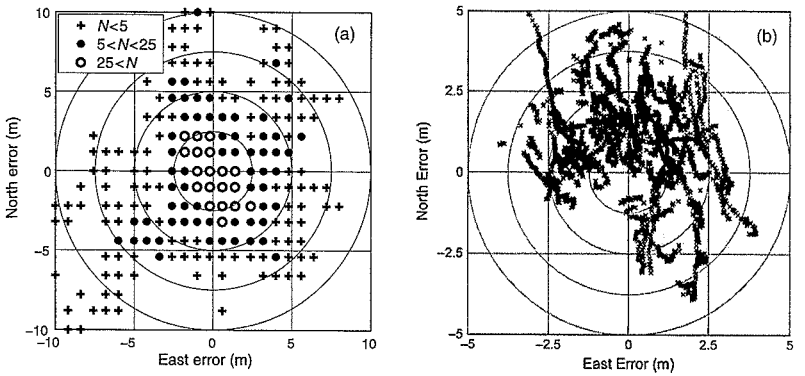


Figure 6.6 Scatter plots of horizontal position error in position estimates obtained from GPS Standard Positioning Service (SPS). (a) Quantized position estimates from an inexpensive receiver for hikers (one-minute samples over 24 hours, N represents the number of outcomes corresponding to each bin). (b) A more capable receiver which gave access to the pseudorange measurements and navigation data (one-second samples over 24 hours), courtesy of Dr. Todd Walter, Stanford University.

DGPS Position Estimates

Representative positioning results from wide-area and local-area differential GPS are provided in Figures 6.7. The wide-area corrections were obtained from the Wide Area Augmentation System (WAAS) [Section 5.8.3] for the data set used in Figure 6.6(b). The corrections were applied to the pseudorange measurements, and the resulting positioning errors are shown. The median and 95th percentile errors are now 0.7 meters and 1.7 meters, a significant improvement over SPS.

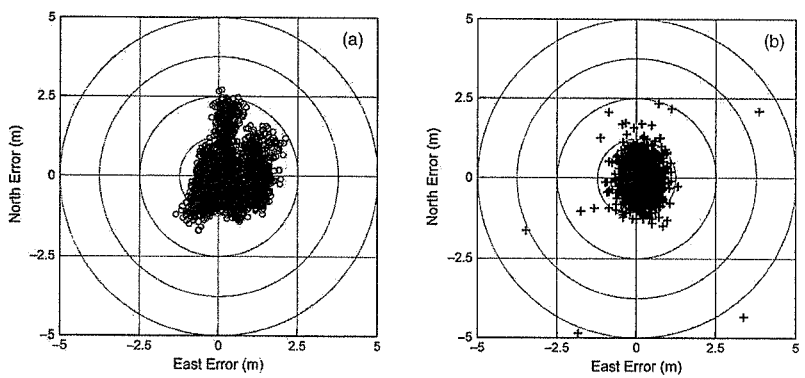


Figure 6.7. Scatter plots of horizontal position error in DGPS position estimates using pseudorange corrections obtained from (a) WAAS (1-second samples over 24 hours), courtesy of Dr. Todd Walter, Stanford University. (b) Local-area DGPS with the reference station 25 km away (1-minute samples over 24 hours).

A local-area differential GPS can do better yet. Figure 6.7(b) shows the positioning results using corrections obtained from a CORS site [Section 2.5] located about 25 km away. In this experiment, the pseudorange measurements at both ends were smoothed using carrier phase measurements [Section 4.7.1]. The 95th percentile point is less than 1 meter.

6.2 Position and Velocity from Pseudorange Rates

6.2.1 Velocity Estimation

The relative motion of a satellite and the user results in changes in the observed frequency of the satellite signal. This Doppler shift is measured routinely in the carrier tracking loop of a GPS receiver [Chapter 12]. Given the satellite velocity, the Doppler shift can be used to estimate the user velocity. The Doppler shift, or equivalently, the range rate [Section 1.3.2], can be written as a projection of the relative velocity vector on the satellite line-of-sight vector. The measurement, however, is biased by the receiver clock bias rate (i.e., frequency offset), and what we actually measure is the pseudorange rate.

The delta pseudoranges obtained from carrier phase measurements are proportional to the average pseudorange rates or the line-of-sight velocity of the user relative to the satellite over the time interval. The model for pseudorange rates can be obtained by differentiating (6.1). It is left as an exercise to show that

$$\begin{aligned}\dot{\rho}^{(k)} &= \dot{r}^{(k)} + (\dot{b} - \dot{b}^{(k)}) + \dot{I}^{(k)} + \dot{T}^{(k)} \\ &= (\mathbf{v}^{(k)} - \mathbf{v}) \cdot \mathbf{1}^{(k)} + \dot{b} + \varepsilon_{\dot{\rho}}^{(k)},\end{aligned} \tag{6.33}$$

where $\mathbf{v}^{(k)}$ is the satellite velocity vector, obtained from the navigation message broadcast by the satellite (Homework Problem 4-12); $\mathbf{v}$ is the user velocity vector, to be estimated. Both $\mathbf{v}^{(k)}$ and $\mathbf{v}$ are expressed in the ECEF coordinate frame. The user-to-satellite line-of-sight unit vector $\mathbf{1}^{(k)}$ is determined from an estimate of the user position; $\dot{b}$ and $\dot{b}^{(k)}$ are the rates of change in the receiver and satellite clocks (m/s), respectively, and $\varepsilon_{\dot{\rho}}^{(k)}$ denotes the combined error due to changes during the measurement interval in the satellite clock, ionosphere, and troposphere. Note that the velocity of an object attached to the earth is zero in the ECEF coordinate frame.

The principal source of error in (6.33) throughout the 1990s was the satellite clock frequency dithering ($\dot{b}^{(k)}$) due to SA. Now with SA gone, this term is negligible. The errors due to changes during the measurement interval in the ionospheric and tropospheric delays and in multipath are also generally small. For moderate user speeds, the main source of error in velocity estimation is the error in the predicted satellite position and velocity $\mathbf{v}^{(k)}$. Any error in the predicted satellite position (especially the along-track and cross-track components) shows up in the line-of-sight vectors, and is magnified in the dot product with the satellite velocity vector.

Problems can arise if the user dynamics are high. The delta ranges give only an average velocity over a time interval. High accelerations and jerks would clearly be problematic. Such dynamics can also create an additional error source by raising the receiver clock phase noise [Section 4.2.3]. For these reasons, the DoD has avoided specifying the velocity estimation performance. (According to an old document [JPO (1991)], the PPS performance specifications for a constant-velocity scenario are: 0.1 m/s rms in any direction; 0.2 m/s 2drms. These are overly conservative, as we'll see below.)

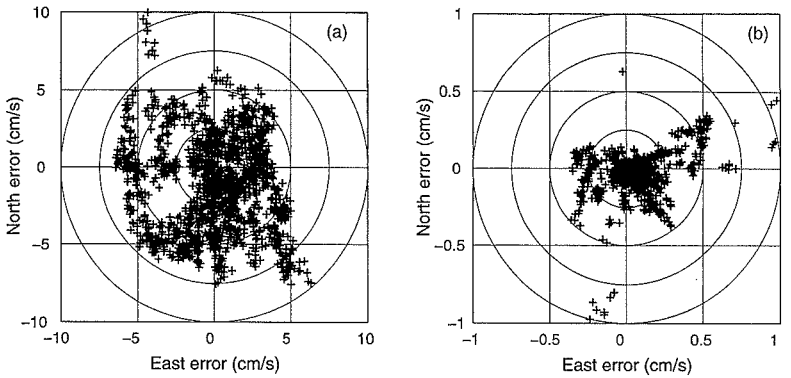


Figure 6.8. Scatter plots of horizontal error in GPS velocity estimates computed from pseudorange rate (Doppler frequency) measurements taken one minute apart over a 24-hour period from a pair of stationary antennas about 25 km apart. (a) Error in velocity estimates; (b) error in relative velocity estimates. (Note: 1 m/s ≈ 2 knots.)

Equation (6.28) is linear in user velocity components, and can be rewritten as

$$(\dot{\rho}^{(k)} - \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)}) = -\mathbf{1}^{(k)} \cdot \mathbf{v} + \dot{b} + \varepsilon_{\dot{\rho}}^{(k)}$$

Denoting $(\dot{\rho}^{(k)} - \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)})$ as $\tilde{\dot{\rho}}^{(k)}$, the combined set of measurements from the K satellites can be written as a set of equations compactly in matrix notation as

$$\tilde{\dot{\boldsymbol{\rho}}} = \mathbf{G}\begin{bmatrix}\mathbf{v} \\ \dot{b}\end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_{\dot{\rho}}, \tag{6.34}$$

where matrix $\mathbf{G}$ characterizes the user-satellite geometry, as defined previously (6.10). The problem of estimation of user velocity based on pseudorange rates is identical in structure to that of estimation of user position from pseudoranges (6.9). A least-squares solution and the DOP parameters can be defined, as before, and related to the rms error in these estimates.

Velocity estimates based on pseudorange rates are presented in Figure 6.8. We recorded pseudorange rates measured one minute apart over a day at a pair of receivers with stationary antennas about 25 km apart. Figure 6.8(a) gives a scatter plot of the horizontal velocity estimates for one of the receivers. The 50th and the 95th percentiles of the horizontal velocity error are 0.03 m/s and 0.08 m/s, respectively. The effect of errors in the predicted satellite position and velocity, discussed above, is mitigated in relative velocity computations, and results in a dramatic improvement. This is seen in Figure 6.8(b), which presents the corresponding estimates of relative velocity between the two receivers.

The estimates of the vertical component of velocity are not as good as those for horizontal velocity. The reasons are similar to those given earlier in this section in the discussion of the accuracy of the horizontal and vertical position estimates from conventional processing.

6.2.2 Position Estimation

As an aside of some historical interest, we discuss Doppler positioning [Section 1.2.3] with GPS. If a user is stationary at an unknown location, we can rewrite (6.33) as

$$\begin{aligned}\dot{\rho}^{(k)} &= \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)} + b + \varepsilon_{\dot{\phi}}^{(k)}, \\ &= \mathbf{v}^{(k)} \cdot \frac{(\mathbf{x}^{(k)} - \mathbf{x})}{\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\|} + b + \varepsilon_{\dot{\phi}}^{(k)}\end{aligned} \tag{6.35}$$

In principle, this set of K equations ($K \geq 4$) can be solved for the four unknowns: $(\mathbf{x}, b)$. As before, we could resort to linearization about the initial estimates of position and receiver clock rate, and set up a linear system of equations to solve iteratively. Actually, the information content in the above formulation regarding the user position is weak: the line-of-sight vector is not very sensitive to small changes in user position. As noted previously, the angle at a satellite between the lines of sight from two users separated by 100 km is less than 0.3°.

The DOP parameters defined for the system of equations (6.35) would be large (in thousands), in general. This is another way to characterize the information content in the system of equations (6.35) for position estimation. The DOPs can be brought down by stacking the measurements taken over a period of time. It would take measurements at well-separated epochs to get a good position estimate.

While Doppler positioning has little to recommend to the GPS users, it was the basis for positioning with Transit, the first operational satellite navigation system [Section 1.3.2]. A Transit user saw only one satellite at a time for ten to twenty minutes, took multiple measurements, and solved for his 2-D position and the difference of the clock bias rates of the satellite and receiver clocks (i.e., frequency offset between the satellite and receiver clock). Unlike GPS, the clocks in Transit satellites did not require precise synchronization, but the frequencies of the Transit satellites and user clocks were required to be stable over ten to twenty minutes. In other words, any frequency instability during a satellite pass contributed to position error.

6.3 Time Transfer

The problem of keeping precise time and synchronizing clocks that are separated by considerable distances is an old one. What's changed over time is that the requirements of new applications for accuracy and precision have become more and more demanding. In Section 4.2, we discussed time scales and the evolution of timekeeping devices. In this section, we discuss the distribution of time, or time transfer, in which GPS now plays an increasingly important role. In particular, we discuss two applications: direct time distribution from GPS and comparison of remote clocks.

The coming of the telegraph in the 19th century and its use in distribution of time was vital to smooth operation of railroads across a continent. Early in the 20th century, telegraph was superseded by radio for time transfer. Radio broadcast time over land and oceans and in the air with a millisecond-level accuracy. Then came radionavigation systems with transmission of their signals carefully synchronized with reference to a precise time standard. Omega, Loran-C, and Transit were all capable of time transfer with an accuracy level of about 1 ms [Klepczynski (1983)]. With the current techniques, GPS can distribute time with an accuracy

of about 30 ns, and can compare remote clocks with an accuracy of about 5 ns.

Precise synchronization of the clocks aboard the satellites is at the heart of GPS. Synchronization of a satellite clock with UTC(USNO) at nanosecond level is made possible by use of atomic frequency standards aboard the satellites and in the monitor stations of the Control Segment. GPS provides access to this network of synchronized atomic clocks for dissemination of precise time and frequency data worldwide.

A user interested in obtaining precise time would typically require a one-pulse-per-second (1PPS) signal synchronized with UTC, the civil standard. Actually, we'll look for a 1PPS signal synchronized to UTC(USNO), introduced earlier in Section 4.2 as UTC realized in real time at the U.S. Naval Observatory (USNO). In order to generate such a signal from a GPS receiver, we have to deal with four time scales:

- time kept by a satellite clock (t^s),
- GPS time (GPST) (t_{GPS}), defined by the Control Segment on the basis of a set of atomic standards aboard the satellites and in monitor stations,
- UTC(USNO) (t_{UTC}), the U.S. national standard defined by the U.S. Naval Observatory,
- time kept by a user's receiver clock (t_u).

Starting with t_u, we want to generate t_{UTC}. We do so by estimating and accounting for the offset between these two time scales in steps as follows.

Recall that the Control Segment determines the bias in each satellite clock (δt^s) on the basis of measurements from the monitor stations.

$$\delta t^s = t^s - t_{GPS} \tag{6.36}$$

The bias is modeled as a quadratic function over a time interval. The parameters $\{a_{f0}, a_{f1}, a_{f2}\}$ of this model are computed and uploaded to the satellite, which broadcasts them in the navigation message (Subframe 1) to the users. At time t_{GPS}

$$\delta t^s = a_{f0} + a_{f1}(t_{GPS} - t_{0c}) + a_{f2}(t_{GPS} - t_{0c})^2 + \Delta t_r \tag{6.37}$$

where t_{0c} is the reference time for the model in GPST, and Δt_r is the relativistic term [Section 4.2.4]. The rms error in estimation of δt^s is currently estimated to be about 5 ns [$\sigma(\delta t^s) \approx$ 5 ns].

The U.S. Naval Observatory monitors the offset of GPST relative to UTC(USNO) as

$$\delta t_{UTC} = t_{GPS} - t_{UTC} \tag{6.38}$$

and provides this information to the Control Segment. The bias between GPST and UTC(USNO) can be computed at any instant (defined in GPST) from

$$\delta t_{UTC} = A_0 + A_1(t_{GPS} - t_{0U}) + \Delta t_{LS} \tag{6.39}$$

where the first two terms specify the bias modulo 1 second, and Δt_{LS} is the number of leap seconds added to UTC since 1980 (Δt_{LS} = 14 seconds as of 1 January 2006). Values of parameters $\{A_0, A_1, t_{0U}, \Delta t_{LS}\}$ are broadcast by each satellite in its navigation message (page 18 of Subframe 4). The rms error in estimation of δt_{UTC} is currently estimated to be about 10 ns [$\sigma(\delta t_{UTC}) \approx 10$ ns].

Finally, in order to generate from the receiver clock a 1PPS signal synchronized to UTC(USNO), we have to account for the receiver clock bias relative to GPST

$$\delta t_u = t_u - t_{GPS} \tag{6.40}$$

We dealt with estimation of receiver clock bias earlier in this chapter, given pseudorange measurements from four or more satellites. A navigation receiver routinely computes δt_u in order to schedule measurements, to time tag position estimates, and to time-align the measurements for precise relative positioning [Section 7.3.1]. Having estimated δt_u, the receiver can determine and display UTC(USNO). RMS error in estimation of the receiver clock bias (δt_u) is given by [Section 7.1.2]

$$\sigma(\delta t_u) = \sigma_{URE} \cdot TDOP$$

With a 24-satellite constellation, TDOP is typically 1–1.5 and $\sigma_{URE} \approx 6$ m. (Note that the error term δt^s is already included in URE.) A navigation receiver can, therefore, estimate the receiver clock bias with an rms error of about 25 ns [$\sigma(\delta t_u) \approx 25$ ns].

The steps for generating t_{UTC} are now clear from the following equation.

$$t_{UTC} = t_u - \delta t_u - \delta t_{UTC} \tag{6.41}$$

We can assess the total error in direct time distribution from GPS.

$$\sigma(t) = \sqrt{\sigma^2(\delta t_u) + \sigma^2(\delta t_{UTC})} \approx 25 \text{ ns}$$

Telecommunications applications typically require synchronization of multiple nodes with an accuracy of 100 ns, or better. Such synchronization can be achieved by setting up a GPS antenna at a fixed, surveyed location at each node, and determining time independently. With antenna position known, a receiver can determine precise time by tracking a single satellite, as discussed below. Such applications generally use specialized, single-channel timing receivers

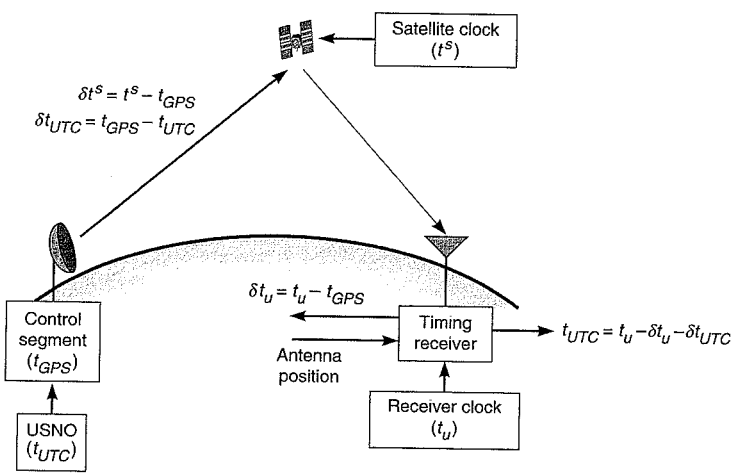


Figure 6.9 Time transfer: obtaining precise time from a GPS satellite.

which produce a 1PPS clock signal synchronized with UTC.

The measured pseudorange from the kth satellite is given in accordance with (6.1) as

$$\rho^{(k)}(t) = r^{(k)}(t, t-\tau) + c[\delta t_u(t) - \delta t^{(k)}(t-\tau)] + I^{(k)}(t) + T^{(k)}(t) + \varepsilon_\rho^{(k)}(t) \tag{6.42}$$

Given the antenna position, the geometric range term $r^{(k)}(t, t-\tau)$ is known. The satellite clock bias term $\delta t^{(k)}$ is determined from the parameter values in the navigation message in accordance with (6.37). The ionospheric and tropospheric delay terms are estimated and accounted for as discussed in Section 5.3. The receiver clock bias term δt_u also includes signal delays due to the antenna, preamplifier, cable, and receiver hardware. The timing receivers are calibrated to isolate these signal delays from clock bias. Having estimated the actual receiver clock bias, the receiver produces time markers (e.g., 1PPS signals) in synchronism with UTC in accordance with (6.41) to time external events. The concept is illustrated in Figure 6.9.

We close out this section by describing a specialized technique which offers a time transfer accuracy of 5–10 ns. The technique is called *common view*, and the principal users of this technique are the international timing centers interested in comparing clocks at one site with those at another. The best previous technique entailed transporting a clock to the other timing center for such comparison. As the name common view suggests, the technique consists of taking measurements from a common GPS satellite simultaneously at two or more sites [Lewandowski *et al.* (1999), Beard and White (1999)]. A regular schedule of data collection and distribution is coordinated by BIPM for the international timing centers. The measurements are exchanged and processed so as to compensate for or cancel errors to the extent possible. The basic idea is similar to that behind DGPS. The luxury of time allows the use of precise, post-processed ephemerides. Corrections for the propagation delays are similarly addressed with greater accuracy. Greater attention is also paid to mitigation of multipath and maintenance of antenna and cables in a stable environment. The concept is illustrated in Figure 6.10 for a case of two timing centers A and B comparing their clocks.

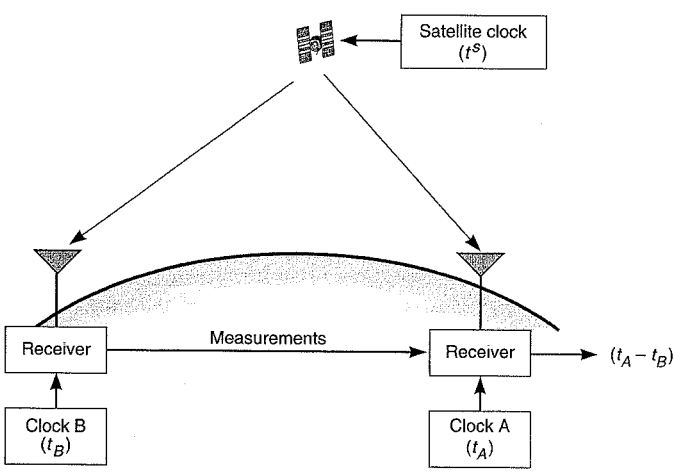


Figure 6.10 Common-view time transfer.

6.4 Summary

In this chapter, we have focused on a couple of bottom-line issues:

- How to estimate position, velocity, and time, given the measurements of pseudoranges and pseudorange rates?
- What determines the quality of the PVT estimates, and how can it be improved?

The main ideas in this chapter are associated with least-squares estimation of position and velocity, and with various measures of positioning accuracy. Other important ideas relate to models of measurement errors and dilution of precision (DOP) parameters to characterize the role of satellite geometry in the quality of the position and velocity estimates. We also examined empirical data on position and velocity estimates obtained in autonomous and differential modes.

A user equipped with a basic GPS receiver can expect to estimate her horizontal position typically within 10 m of the true location. The typical velocity error would be less than 0.05 m/s and time error less than 50 ns. Differential corrections to the measurements are now widely available. The position and time estimation errors cited above can be reduced by two-thirds, or more, in differential mode.

Appendix 6.A Parameter Estimation

In this appendix we present several basic results from estimation theory. Our focus is narrow and we limit ourselves to the requirements of Chapters 6 and 7.

6.A.1 Least-Squares Problem

Suppose n parameters $(x_1, x_2, \ldots, x_n)$, denoted as n-vector $\mathbf{x}$, are to be estimated from m ($>n$) noisy measurements $(y_1, y_2, \ldots, y_m)$, which we represent by m-vector $\mathbf{y}$. Each measurement is related linearly to the parameters of interest. In vector-matrix notation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} \tag{6.A.1}$$

$\mathbf{A}$ is an $m \times n$ matrix called the design matrix. For simplicity we consider only the case where $\mathbf{A}$ is of full rank. The m-vector $\boldsymbol{\varepsilon}$ represents the unknown errors in the measurements.

A commonly used approach to 'solve' the over-determined set of linear equations (6.A.1) is to determine $\mathbf{x}$ which fits the measurements 'best' in some sense. The lack of fit is represented by the *residual vector* $\mathbf{r} = (\mathbf{y} - \mathbf{A}\mathbf{x})$. The least-squares approach, which dates back to Gauss, consists of minimizing the sum of the squares of the components of this residual vector. In other words, the least-squares estimator, which we represent as $\hat{\mathbf{x}}$, minimizes the cost function

$$\|\mathbf{r}\|^2 = (\mathbf{y} - \mathbf{A}\hat{\mathbf{x}})^T(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}) = \|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|^2 \tag{6.A.2}$$

Using basic results from calculus, we can determine the stationary point of this function to show that

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{6.A.3}$$

How good is this estimate? Well, we can write an expression for the error as

$$\begin{aligned}\tilde{\mathbf{x}} &= \hat{\mathbf{x}} - \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} - \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}) - \mathbf{x} \\ &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\varepsilon}\end{aligned} \tag{6.A.4}$$

It shouldn't come as a surprise that the position error depends upon $\boldsymbol{\varepsilon}$ and its relationship with the column vectors of $\mathbf{A}$.

We can also write an expression for the residual vector as

$$\begin{aligned}\mathbf{r} &= \mathbf{y} - \mathbf{A}\hat{\mathbf{x}} = -\mathbf{A}\tilde{\mathbf{x}} + \boldsymbol{\varepsilon} \\ &= \left[\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\right]\boldsymbol{\varepsilon}\end{aligned} \tag{6.A.5}$$

We recognize the projection matrix on the right, but not knowing *anything* about $\boldsymbol{\varepsilon}$, we can't say much about the residuals. The next step is to characterize the nature of $\boldsymbol{\varepsilon}$ probabilistically, and see how it helps us assess the parameter estimates. We do this below.

The simplest meaningful characterization of $\boldsymbol{\varepsilon}$ is through second-order statistics. Suppose $\boldsymbol{\varepsilon}$ has mean value zero and covariance matrix $\mathbf{R}$

$$\begin{aligned}E[\boldsymbol{\varepsilon}] &= \mathbf{0} \\ \mathrm{Cov}[\boldsymbol{\varepsilon}] &= E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \mathbf{R}\end{aligned} \tag{6.A.6}$$

where $E[\bullet]$ denotes the expected value. What can we say about our least-squares estimator (6.A.3) now?

$$E[\tilde{\mathbf{x}}] = E[\hat{\mathbf{x}} - \mathbf{x}] = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T E[\boldsymbol{\varepsilon}] = \mathbf{0} \tag{6.A.7}$$

The estimator is unbiased. We can also determine its covariance $\mathbf{P}$.

$$\mathbf{P} = E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{R}\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1} \tag{6.A.8}$$

This expression becomes much simpler if the components of $\boldsymbol{\varepsilon}$ are uncorrelated and each has identical variance σ^2, i.e., $\mathbf{R} = \sigma^2\mathbf{I}$,

$$\mathbf{P} = \sigma^2(\mathbf{A}^T\mathbf{A})^{-1} \tag{6.A.9}$$

We can also write expressions for the mean and covariance of the residual vector, but we'll leave it as an exercise.

6.A.2 Generalized Least-Squares Problem

The least-squares cost function (6.A.2) tacitly assumes that all measurements are of equal quality. If some measurements were known to be 'purer' than others, we could incorporate this knowledge in the cost function through a weighting matrix $\mathbf{W}$. We'll look at the general case of a symmetric, positive-definite weighting matrix $\mathbf{W}$. The new cost function is

$$\|\mathbf{r}\|_{\mathbf{W}}^2 = (\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{W}(\mathbf{y} - \mathbf{A}\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{\mathbf{W}}^2 \tag{6.A.10}$$

It is left as an exercise to show that the corresponding weighted least-squares estimator is

$$\hat{\mathbf{x}}_{\mathbf{W}} = (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{y} \tag{6.A.11}$$

If, as assumed earlier, $\boldsymbol{\varepsilon}$ has mean value zero and covariance matrix $\mathbf{R}$, the error in our estimate $\hat{\mathbf{x}}_W$ has mean zero and covariance matrix

$$\mathbf{P}_W = (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{A}(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1} \tag{6.A.12}$$

The string of matrices on the right is so long we don't even know where to look for insight. So, let's take a different tack.

6.A.3 Best Linear Unbiased Estimator (BLUE)

We now look at an alternate approach to estimation of $\mathbf{x}$ in our observation model (6.A.1), given the second-order characterization of error $\boldsymbol{\varepsilon}$ (6.A.6). We want to determine an estimator which has the smallest error variance. We limit ourselves to linear, unbiased estimators and define a new cost function, the mean-square error

$$E[(\hat{\mathbf{x}}-\mathbf{x})^T(\hat{\mathbf{x}}-\mathbf{x})] \tag{6.A.13}$$

We now look for an $(n \times m)$ matrix $\mathbf{B}$ which would generate the estimator we are after.

$$\hat{\mathbf{x}} = \mathbf{B}\mathbf{y} \tag{6.A.14}$$

The estimator is required to be unbiased.

$$E(\hat{\mathbf{x}}-\mathbf{x}) = \mathbf{B}E(\mathbf{y})-\mathbf{x} = (\mathbf{B}\mathbf{A}-\mathbf{I})\mathbf{x} \tag{6.A.15}$$

Therefore, it is required that

$$\mathbf{B}\mathbf{A} = \mathbf{I} \tag{6.A.16}$$

The error covariance is

$$E[(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T] = E(\mathbf{B}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{B}^T) = \mathbf{B}\mathbf{R}\mathbf{B}^T \tag{6.A.17}$$

The mean-square error

$$E[(\hat{\mathbf{x}}-\mathbf{x})^T(\hat{\mathbf{x}}-\mathbf{x})] = \text{trace}\,(\mathbf{B}\mathbf{R}\mathbf{B}^T) \tag{6.A.18}$$

So, we are looking for matrix $\mathbf{B}$ which minimizes the trace (i.e., the sum of the diagonal terms) of $\mathbf{B}\mathbf{R}\mathbf{B}^T$ and satisfies the constraint (6.A.16). Without going into the derivation, the answer is

$$\mathbf{B} = (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{R}^{-1} \tag{6.A.19}$$

Obviously, (6.A.16) is satisfied, and the covariance of the error (6.A.17) is

$$\mathbf{P} = (\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A})^{-1} \tag{6.A.20}$$

The expression for $\mathbf{B}$ should look familiar. If we use $\mathbf{R}^{-1}$ as the weighting matrix $\mathbf{W}$ for the generalized least-squares solution (6.A.11), we get BLUE! When you replace $\mathbf{W}$ by $\mathbf{R}^{-1}$ in (6.A.12), it reduces the error covariance matrix (6.A.12) to the much simpler (6.A.20).

The results above required only the second-order statistics for the measurement error. If the error could be characterized as Gaussian, the best linear unbiased estimator would in fact be the minimum mean-square error estimator. An interested reader can learn more about this subject from Sorenson (1980).

Homework Problems

6-1. Derive the weighted least-squares solution (6.13) for position estimation from the system of linear equations (6.9).

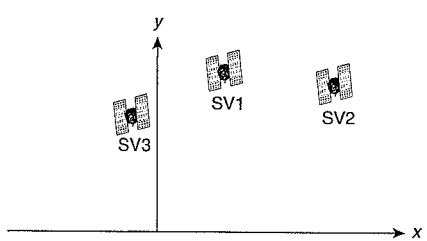
6-2. Consider the 2-D position estimation problem shown in Figure 6.2. A user measures his ranges from two stations whose coordinates are known. The errors in the range measurements are uncorrelated and distributed identically with zero mean and variance σ^2. Obtain an expression for the position estimate, and show that

$$\text{RMS position error} = \frac{\sigma\sqrt{2}}{|\sin\theta|}$$

where θ is the angle between the two stations as observed at the user location.

6-3. Consider another idealized 2-D position estimation problem, now with pseudorange measurements. The satellite positions are shown on the next page. The user is believed to be near the origin. Using the transmitted ephemeris information, the satellite positions at the time of signal transmission are calculated as: $\mathbf{x}^{(1)} = [3.00, 9.53]^T$, $\mathbf{x}^{(2)} = [8.41, 8.65]^T$, and $\mathbf{x}^{(3)} = [-1.23, 6.99]^T$ in units of length. The measured pseudoranges are: $\rho^{(1)} = 1.19$, $\rho^{(2)} = 3.71$, and $\rho^{(3)} = -1.71$. (Note that the pseudorange to satellite 3 is negative, yet it is a valid pseudorange.) Set up linearized navigation equations as in Section 6.1.1.

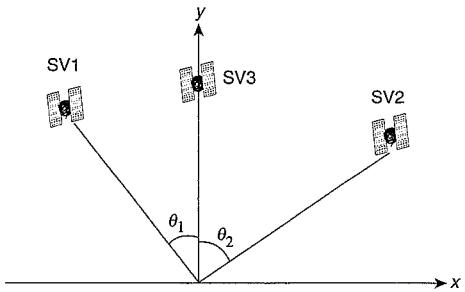
(a) Estimate the user position and receiver clock bias starting with the initial guesses of (0, 0) for the position coordinates and zero for the clock bias.



(b) Explain how the answers would change if $\rho^{(1)} = 101.19$, $\rho^{(2)} = 103.71$, $\rho^{(3)} = 98.29$.

(c) Now suppose that the user is on a train. The train tracks run along the y-axis, so the x-coordinate of the user must be 0. Compute the least-squares estimates for y and b.

6-4. Consider 2-D position estimation with pseudoranges from the three-satellite configuration below.



(a) Determine the geometry matrix **G** defined in (6.10).

(b) Determine HDOP as a function of θ_1 and θ_2? (Do not simplify the trig functions.)

(c) What is HDOP for $\theta_1 = \theta_2 = 45°$? What is HDOP for $\theta_1 = \theta_2 = 90°$?

(d) Compare HDOP with that from Problem 6-1. Does it help to know the receiver clock bias?

6-5. Determine HDOP and VDOP for the following user-satellite geometry. There are four GPS satellites in view. One is located directly overhead. The three remaining satellites are distributed 120° apart in azimuth, each at elevation angle α. Show that as α decreases from 30° to 0°, VDOP decreases by about 50% (from 2.3 to 1.2) but HDOP remains essentially unchanged (at 1.2). Unlike HDOP, VDOP continues to drop as α becomes negative (i.e., the three satellites drop below the horizon). What is the value of α that minimizes VDOP? This insight is of no value to earth-bound users who can't see satellites below the horizon, but can be exploited in navigation of spacecraft in low earth orbits.

6-6. In this chapter we focused on instantaneous position estimation when four or more GPS satellites were in view. A stationary user with a restricted view of the sky can estimate his position if at least two satellites remain in view for a period of time. Show that pseudorange measurements are required at a minimum from two satellites at three epochs, or from three satellites at two epochs. What happens if the measurement epochs are close together?

6-7. Suppose you have only three satellites in view. Obviously, you cannot estimate your 3-D position. You can try for a 2-D position estimate if you knew the vertical position.

(a) Develop a mathematical model for 2-D position estimation with GPS, given the vertical position as altitude in the local ENU coordinate frame.

(b) Suppose the 'known' vertical position is in error by α meters. How would this affect the quality of the horizontal position estimate?

6-8. A baroaltimeter is a trusted old friend of the pilots. It is not as accurate as GPS, but when set properly, it allows the pilots to maintain vertical separation among aircraft to the required accuracy. Let's simplify the problem and say that you have a baroaltimeter-like instrument that measures height above the WGS 84 ellipsoid, and these measurements are unbiased with a standard deviation of 24 m. The rms error in the GPS pseudorange measurements is 6 m. Make any other reasonable simplifying assumptions, if necessary. Develop a mathematical model for estimation of your position based on the measurements from both GPS and this hypothetical altimeter. Is the horizontal position error affected by the availability of these measurements? What about the vertical error?

6-9. Derive relationships (6.31b) and (6.31c) among the three performance measures of the horizontal positioning accuracy: CEP, 2-D rms, and the 95th percentile of the error distribution. Model the distribution of the north and east components of the horizontal position error as bivariate normal, and assume that the two components are uncorrelated and distributed identically with zero mean and variance σ^2.

6-10. Consider a local DGPS implementation in which there is a significant error (say, 1 km) in the estimated position of the reference antenna. The differential corrections computed at the reference station would, therefore, be in error, and a user applying these corrections would get her position wrong. Show that the error in the user's position would be approximately the same as that in the reference antenna position, and the estimated relative position vector between the reference and the user would be essentially unaffected by this error. In other words, the user position is accurate relative to the position ascribed to the reference, and any error in the reference position would be harmless if the user were proceeding to a rendezvous with the reference station.

6-11. Let us now return to the single-epoch data sets (CD) *Data\Original\rcvr.dat* and *eph.dat* to set up the linearized navigation equations and solve for user position and clock bias. These equations can be solved iteratively as described in Section 6.1. The required corrections for the satellite clock bias and relativity are detailed in the IS. We'll skip the ionospheric corrections because we do not have access to the parameter values of the Klobuchar model for this data set. Tropospheric correction based on standard atmosphere model is optional. Use rwy30Start from Problem 4-2 as the initial position to begin your iteration. Initialize your algorithm with a clock bias of zero. Terminate the iteration when the change in the estimates is suitably small.

What is your estimate of the user clock bias b? Does your estimate of the user clock bias in seconds offer insight as to why the reported receiver clock time at this epoch (Column 1 of the *rcvr* matrix) is 440992.00173454 seconds? (Hints: Your initial iteration should give (in meters) $\delta x \approx -5710$, $\delta y \approx 1080$, $\delta z \approx -2610$, $\delta b \approx 519{,}450$. Your position estimate (WGS 84 coordinates, in meters) should be:

$$\mathbf{x} \approx (-2{,}700{,}400 \quad -4{,}292{,}560 \quad 3{,}855{,}270)^T.$$

6-12. We are now ready to work with multiple-epoch data files and analyze the variability in the position estimates. Download receiver and ephemeris files (CD) *Data\Stanford \September_18_2000\r091800a.dat* and *e091800a.dat*. The column formats of these files are identical to those for the files used in Problem 6-11. (See Appendix A for a description.)

Column 1 of the ephemeris file is the *time tag* corresponding to the most recent update of the ephemeris parameter values. Column 1 of the receiver file contains the measurement time tag. When solving for user position and clock bias, you must use the most recent ephemeris information for each PRN.

(a) Compute position and clock bias estimates for each epoch when four or more satellites are in view. Express the positions in the ENU coordinate frame centered at the position estimate from Problem 6-11. Plot east, north, and up positions versus time. Calculate the mean and standard deviation of each component.

(b) Plot the user clock bias estimates $\hat{b}$ versus time. What does this plot tell you about the timekeeping ability of the receiver clock?

(c) Calculate the range estimate for PRN 6 ($\rho^{(6)} + \hat{b}$), and plot it epoch by epoch for a two-hour period in the middle of the satellite pass. Plot the elevation angle of PRN 6 for the same time period. Does the time of the minimum range to the satellite line up with the time of the maximum elevation angle?

6-13. SA now is history, but in order to see its damaging effect, repeat Problem 6-12(a) and (b) with data files (CD) *Data\Stanford\January_6_2000.*

6-14. Use the data files from Problem 6-13.

(a) Create a sky plot [Section 4.4] for a subset of the SVs over a four-hour period. Represent the SV positions in polar coordinates where the azimuth represents the angle and the elevation represents the distance from the origin.

(b) Plot HDOP and VDOP versus time. (Tip: Do not forget to rotate the rows of the **G** matrix into the ENU frame at each epoch, but don't do this until *after* you have a valid XYZ position solution.)

6-15. (Courtesy of Dr. Guttorm R. Opshaug) The files in (CD) *Data\Stranda\July_30_2000* were logged in Stranda, Norway.

(a) Plot the number of satellites in view versus time.

(b) Generate a sky plot as in Problem 6-14(a).

(c) Find the approximate LLA position of Stranda from a map and convert it to WGS 84 XYZ coordinates. This would serve as the initial estimate for positioning routines developed in the previous problems.

(d) Compute the position estimate epoch by epoch. Calculate the receiver position in an ENU coordinate frame centered at the position determined in part (c) above. Plot east, north, and up position estimates versus time and compute the mean and standard deviation for each component.

(e) Plot HDOP and VDOP versus time.

(f) How do the position estimates obtained from these data logged at latitude 62° N compare with those in Problems 6-12 and 6-14 for data logged at 37° N?

References

Axelrad, P., and R.G. Brown (1996). GPS Navigation Algorithms, in *Global Positioning System: Theory and Applications, Vol. I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 409–434.

Beard, R.L., and J.D. White (1999). GPS Application to Time Transfer and Dissemination, *GPS Solutions*, vol. 3, no. 1, pp. 17–25.

Brown, Robert Grover, and Patrick Y.C. Hwang (1997). *Introduction to Random Signals and Applied Kalman Filtering*, 3rd edition, John Wiley & Sons.

Collins, Paul, Richard B. Langley, and J. LaMance (1996). Limiting Factors in Tropospheric Propagation Delay Error Modelling for GPS Airborne Navigation, *Proc. ION 52nd Annual Meeting*, pp. 519–528.

Helstrom, Carl W. (1991). *Probability and Stochastic Processes for Engineers*, 2nd edition, Macmillan Publishing Co.

Hwang, Patrick Y., Gary A. McGraw, Bernard A. Schnaufer, and David A. Anderson (2005). Improving DGPS Accuracy with Clock Aiding Over Communication Links, *Proc. ION GNSS-2005*, pp. 1961–1970.

IS-GPS-200D (2004). Interface Specification IS-GPS-200, Revision D, Navstar GPS Space Segment/ Navigation User Interfaces, Navstar GPS Joint Program Office [CD Documents\IS-GPS-200D.pdf]

Klepczynski, William J. (1983). Modern Navigation Systems and Their Relation to Timekeeping, *Proc. IEEE*, vol. 71. no. 10, pp. 1193–1198.

Leva, Joseph L., Maarten Uijt de Haag, and Karen L. Van Dyke (1996). Performance of Standalone GPS, in *Understanding GPS*, E. D. Kaplan (ed.), Artech House, pp. 237–320.

Lewandowski, Wlodzimierz, Jacques Azoubib, and William J. Klepczynski (1999). GPS: Primary Tool for Time Transfer, *Proc. IEEE*, vol. 87, no. 1, pp. 163–172.

Misra, Pratap (1996). The Role of the Clock in a GPS Receiver, *GPS World*, vol. 7, no. 4, pp. 60–66.

Serrano, Luis, Don Kim, and Richard B. Langley (2004). A Single GPS Receiver as a real-Time, Accurate Velocity and Acceleration Sensor, *Proc. ION GNSS 2004*, pp.2021–2034.

Sorenson, Harold W. (1980). *Parameter Estimation: Principles and Problems*, Marcel Dekker.

Spilker, J.J. (1996). Satellite Constellation and Geometric Dilution of Precision, in *Global Positioning System: Theory and Applications, Vol. I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 177–208.

SPS (2001). *Global Positioning System Standard Positioning Service Performance Standard*, U.S. Department of Defense [CD Documents\SPS Performance 2001.pdf].

Stark, Henry, and John W. Woods (1994). *Probability, Random Processes, and Estimation Theory for Engineers*, 2nd edition, Prentice Hall.

Strang, Gilbert, and Kai Borre (1997). *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press, Wellesley, MA.

Tiberius, Christian, Niels Jonkman, and Frank Kenselaar (1999). The Stochastics of GPS Observables, *GPS World*, vol. 10, no. 2, pp. 49–54.

van Diggelen, Frank (1998). GPS Accuracy: Lies, Damn Lies, and Statistics, *GPS World*, vol. 9, no. 1, pp. 41–45.

van Graas, Frank and Andrey Soloviev (2004). Precise velocity estimation using a stand-alone GPS receiver, *Navigation*, vol. 51, no. 4, pp. 283–292.

Chapter 7

Precise Positioning with Carrier Phase

In Section 5.5, we discussed the precision and accuracy of the code and carrier phase measurements, accounting for receiver noise and multipath. The precision of code phase measurements is not much better than 0.5–1.0 m and, therefore, the positioning accuracy achievable from code measurements can be no better than meter level. The carrier phase can be measured with a precision of 0.01–0.05 cycle (2 mm–1 cm). Precise positioning, which we interpret in this chapter to mean centimeter-level positioning, requires carrier phase measurements.

We return to relative positioning, introduced in Section 5.8.2, but now using the carrier phase measurements. Recall that the basic idea was to difference time-matched measurements at two points to eliminate their common-mode errors, and re-parametrize the problem to use these difference measurements to estimate the relative position vector between the two points. Precise positioning with GPS in the last twenty years has come to mean precise relative positioning. This would change. Availability of predicted satellite orbits with centimeter-level errors is now paving the way for precise point positioning in near-real time.

GPS was designed for positioning using code phase measurements. Positioning with centimeter-level accuracy was not among the capabilities foreseen by the designers of the system. The techniques of precise relative positioning with GPS carrier phase measurements are based on the earlier work of radio astronomers on very long baseline interferometry (VLBI) used to measure relative positions of two antennas observing radio signals from extragalactic sources [Gipson *et al.* (1994), Seeber (2003)]. Precise positioning with GPS was first demonstrated by Counselman and colleagues at MIT and Draper Laboratory in the late 1970s [Counselman *et al.* (1979), Counselman and Gourevitch (1981)]. The fundamental ideas of single and double differences and wide laning with GPS carrier phase measurements discussed in this chapter are due to Counselman.

We begin below with a simple, idealized case of one-dimensional relative positioning based on carrier phase. The purpose is to discuss some of the subtleties of carrier phase measurements in the context of a simple model. In the following two sections, we reexamine the carrier phase measurements and set up the relative positioning problem in its full generality. The problem of precise relative positioning with carrier phase turns out to be a problem of estimation of integer ambiguities, which we discuss in Sections 7.4 and 7.5. Finally, in Section 7.6 we discuss precise point positioning with carrier phase measurements from a single receiver in autonomous mode.

7.1 Carrier Phase and Integer Ambiguity Resolution: A Simple Model

Let us examine an idealized case of precise positioning in one dimension. As shown in Figure 7.1(a), we have two receivers with their antennas denoted as A and B, both tracking the carrier from a lone satellite. The purpose of this exercise is to measure the distance d between the two antennas accurately. Actually, now we have to be more precise: d is the distance between the phase centers of the two antennas [Schupler and Clark (2001)]. Angle θ_0 is known. We follow an admirable account of Hwang (1991).

A plane wave front from the satellite, representing points of constant carrier phase, reaches

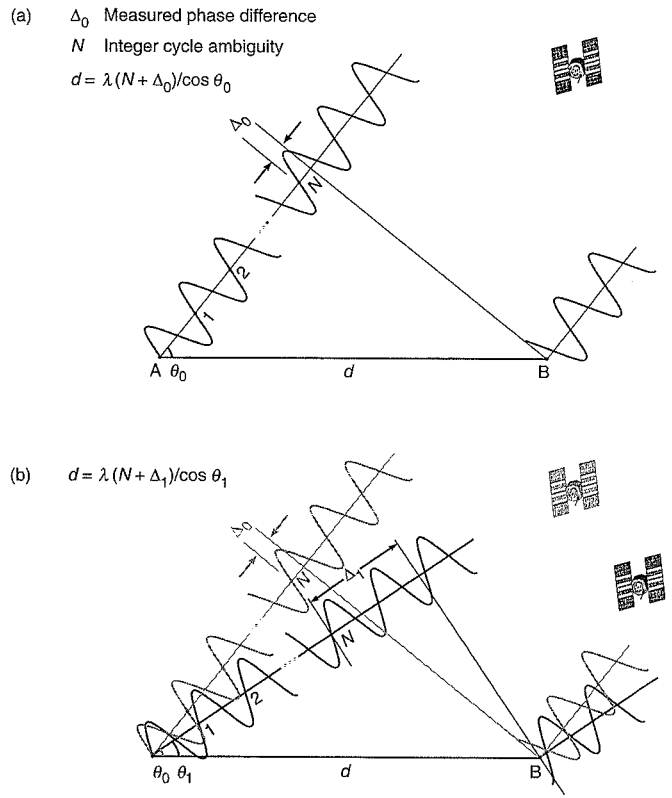


Figure 7.1 Integer ambiguity resolution in an idealized, one-dimensional case. (a) Initial measurement geometry and integer ambiguity. The choice for integer ambiguity resolution is between waiting until there is a significant change in the satellite position, as shown in (b), or performing an antenna swap.

antenna B first. It takes several full cycles and a partial cycle Δ_0 for the plane front to reach antenna A. We have frozen the picture at the instant t_0 when we begin our measurements, choosing this instant carefully to make the figure and this discussion simpler. Clearly, the difference between the carrier phase measurements at the two antennas is a certain number of whole cycles and partial cycle Δ_0. Given the nature of the sinusoidal signals, however, when we first compare the phase measurements at A and B, we can only discern a partial cycle Δ_0. Denoting the carrier phase measurements at the two antennas as ϕ_A and ϕ_B, the difference is

$$\begin{aligned}\phi_{AB}(t_0) &= \phi_A(t_0) - \phi_B(t_0) \\ &= \Delta_0 + N\end{aligned}$$

where N is an unknown integer. The measurements have an integer ambiguity. If we can determine the value of N, we can determine d precisely from the simple geometrical relationship

$$d\cos\theta_0 = \lambda(\Delta_0 + N) \tag{7.1a}$$

It's clear from the above equation that in order to take full advantage of the carrier phase measurements we have to resolve the integer ambiguity. If the relative positions of the two receivers and the satellite remain fixed, the observed carrier phase difference (Δ_0) would not change, and we would have no way to estimate the value of N. But as the satellite moves, the satellite-receiver geometry changes, the signal path lengths to the two antennas change, and the difference of the carrier phase measurements changes.

Now consider the carrier phase measurements at a later time t_1 while both receivers track the carrier continuously. The satellite moves from elevation θ_0 to θ_1 (Figure 7.1(b)) and the observed carrier phase difference at the two antennas changes from Δ_0 to Δ_1. Recognizing that the integer ambiguity has not changed, the new carrier phase difference is in fact $(N+\Delta_1)$. Since the antennas have not moved

$$d\cos\theta_1 = \lambda(\Delta_1 + N) \tag{7.1b}$$

In order to simplify the notation, we introduce a new variable $d' = d/\lambda$. Now we have two equations (7.1a) and (7.1b) in two unknowns, N and d', both in units of cycles.

$$\begin{aligned} d'\cos\theta_0 - N &= \Delta_0 \\ d'\cos\theta_1 - N &= \Delta_1 \end{aligned} \tag{7.2}$$

The problem is solved, but only in principle. What happens if we take the two sets of measurements back to back so that θ_1 is not much different from θ_0? We expect trouble because the two equations would be nearly identical. To see the consequences, we write the solution of (7.2) in matrix notation as

$$\begin{bmatrix} d' \\ N \end{bmatrix} = \frac{1}{(\cos\theta_1 - \cos\theta_0)} \begin{bmatrix} -1 & 1 \\ -\cos\theta_1 & \cos\theta_0 \end{bmatrix} \begin{bmatrix} \Delta_0 \\ \Delta_1 \end{bmatrix} \tag{7.3}$$

Now the difficulty becomes clearer. Small errors in the measurements would get blown up into big errors in the estimates by the multiplier $(\cos\theta_1 - \cos\theta_0)^{-1}$. If we can wait for the satellite to move 'appreciably,' we can get decent estimates from (7.3). We'll refer to such change in receiver-satellite geometry as providing *geometric diversity*.

We can define a dilution of precision (DOP) parameter associated with estimation of d' and N just as we had defined the various DOPs associated with estimation of position and clock bias in Chapter 6. Modeling the error in the carrier phase measurements at t_0 and t_1 to be uncorrelated with zero mean and a common variance $\sigma^2(\Delta)$, we can write the covariance of our estimates of d' and N. It is left as an exercise to show that the DOP associated with the estimation of N, which we label integer-ambiguity-resolution DOP (IDOP), is

$$\text{IDOP} = \frac{\sqrt{\cos^2\theta_0 + \cos^2\theta_1}}{\left|(\cos\theta_1 - \cos\theta_0)\right|} \tag{7.4}$$

and the standard deviation of N is related to the standard deviation of the measurements by

$$\sigma(N) = \text{IDOP} \cdot \sigma(\Delta)$$

Now we have the complete solution. The integer ambiguity can be resolved with carrier phase measurements at two time instants, but the change in the satellite geometry between the two epochs is crucial to the quality of the estimates. Note also that we have imposed no constraint so far that the estimate of N be an integer, and it wouldn't be, in general.

If the two antennas are located close to each other, as in our idealized case, we do not have to wait for the satellite to move in order to resolve the integer ambiguity. We can effect a change in the signal path lengths and estimate N by an ingenious scheme called *antenna swap*, proposed by Remondi (1985). Having obtained the measurement Δ_0, we simply switch the positions of the antennas A and B. Antenna A (along with the receiver it is connected to) is moved to the spot of antenna B, which is moved to the spot vacated by antenna A. Carrier tracking is maintained throughout by both receivers.

Let us suppose that the satellite has not moved significantly during this maneuver. The antenna swap adds path length $d'\cos\theta_0$ to the signal reaching antenna B, and reduces the path length of the signal to antenna A by the same amount. Suppose the measured carrier phase difference changes from the partial cycle Δ_0 before the swap to Δ_0' after the swap. The net change in the carrier phase measurement is

$$\Delta_0' - \Delta_0 = -2d'\cos\theta_0$$

which gives us directly what we are after.

$$d' = \frac{(\Delta_0 - \Delta_0')}{2\cos\theta_0}$$

We can now determine the integer ambiguity from (7.1a)

$$\begin{aligned} N &= d'\cos\theta_0 - \Delta_0 \\ &= -\left(\frac{\Delta_0' + \Delta_0}{2}\right) \end{aligned}$$

Once the ambiguity is resolved, the good news is that we would know the carrier phase difference between the two receivers as long as both track the carrier continuously. The antennas are now free to move along the line joining them and the distance between the two can be determined with precision instantaneously. The bad news is that it's not hard to lose carrier track. Low SNR (due to low satellite elevation angle, for example) and momentary signal blockage are the two common causes.

In this section, we dealt with an idealized model: 1-D positioning and option to wait for the satellite geometry to change and/or ability to swap antennas. Going from 1-D to 3-D positioning would not be a problem, in principle. We'll simply need additional satellites to deal with the additional unknowns. Navigation applications, however, require position estimates in real time and, in general, a user would not have the luxury of waiting for the satellites to move. Antenna swap would seldom be practical for a navigator. And there is no getting away from biases and errors in the measurements.

In our model, we compared the phases of the signal at the two antennas directly. In practice, the carrier phase measurements at each receiver would be made relative to the signal generated by the receiver clock, introducing a time-varying bias in the measurements. If the two antennas are separated by a significant distance, the ionospheric and tropospheric propagation effects would introduce different biases in the measurements at the two antennas. There are

also multipath and receiver noise effects at each antenna to contend with. We'll look at these real-life complications below. We should note in passing that the electrical phase center of an antenna [Schupler and Clark (2001)], to which the measurements are referred, is generally not identical to its geometrical center. The phase center can also vary with the direction of arrival (azimuth and elevation) of a signal and such variation can range from under a millimeter to 1–2 cm, depending upon antenna design. We'll treat the phase center variation as measurement noise.

Before leaving this section, let us briefly consider the inverse problem. If we knew the distance d between the two antennas, we can estimate θ accurately. This is essentially the problem of *attitude determination*. In general, the problem would be formulated in three dimensions and will involve four or more antennas placed in a pattern with carefully measured spacing. The measurement errors discussed in the previous paragraph would be minimized by using a common receiver to process the carrier phase measurements from all antennas. Precise attitude determination for automatic control of spacecraft, aircraft, ships, and all kinds of agricultural and industrial equipment has emerged as an important application of GPS [Cohen (1996)].

7.2 Carrier Phase Measurements and Precise Positioning

7.2.1 Carrier Phase Measurements

Let us start by rewriting (5.9) for the carrier phase measurement in units of cycles as

$$\phi = \lambda^{-1}[r - I + T] + f \cdot (\delta t_u - \delta t^s) + N + \varepsilon_\phi \tag{7.5}$$

where λ and f are the carrier wavelength and frequency, respectively. The geometric range between the satellite and the receiver is r, and I and T are the ionospheric advance and tropospheric delay, respectively, all expressed in units of length. The only difference between (5.9) and (7.5) relates to the notation for the ionospheric and tropospheric terms. The satellite and receiver clock biases (in seconds) are denoted as δt^s and δt_u, respectively. N is the integer ambiguity, the estimation of which will occupy us throughout this chapter. Recall that we use ε with different subscripts, superscripts, overbar, and tilde to denote the measurement or modeling errors. In this section, we work with the minimal notation without regard to the satellite ID, receiver ID, and carrier frequency (f_{L1} or f_{L2}). This will change in later sections.

It is instructive to review the salient features of the carrier and code phase measurements discussed in Chapter 5. Let us rewrite the familiar code measurement equation (5.6) as

$$\rho = r + I + T + c(\delta t_u - \delta t^s) + \varepsilon_\rho \tag{7.6}$$

Two main differences between the code and carrier phase measurements are discussed below and summarized in Table 7.1. Code tracking provides essentially unambiguous pseudoranges. As discussed in the previous section, the carrier phase measurements are encumbered with integer ambiguities which have to be resolved before the measurements can be used for precise positioning or navigation. The integers remain fixed as long as the carrier tracking loop maintains lock. Momentary loss of phase lock can result in a discontinuity in the integer cycle count even though the fractional part of the phase is measured continuously. Such discontinuity in the integer cycle count is called a *cycle slip*. Cycle slips were quite frequent in the early receivers of the 1980s which were limited in both processing power and memory, and had to

Table 7.1 Characteristics of the code and carrier phase measurements

	Code Measurements	*Carrier Phase Measurements*
Ambiguity	Unambiguous	Ambiguous
Error sources	Satellite clock and ephemeris Ionospheric and tropospheric refraction Receiver noise and multipath	Same
Typical measurement accuracy (high-end receivers)	$\sigma(\varepsilon_\rho) \approx 0.5$ m	$\sigma(\varepsilon_\phi) \approx 0.025$ cycle (5 mm)

resort to crude approximations in carrier tracking and recovery. The modern receivers have more sophisticated implementations of the phase lock loops, with fewer instances of cycle slips.

We have discussed models of measurement errors earlier in Section 5.4.3 (Table 5.2). We now specialize these models to high-end receivers used with due care for precise positioning (Table 7.1). Carrier phase can be measured with great precision ($\approx$ 0.005 cycle or 1 mm). Code phase measurements are coarse ($\approx$ 0.25 m) by comparison. The carrier phase measurements are also much more accurate. In the absence of significant multipath, our model for the standard deviations of the measurement errors for this chapter is: $\sigma(\varepsilon_\rho) \approx 0.5$ m, and $\sigma(\varepsilon_\phi) \approx 0.025$ cycle (5 mm).

Let's take a simple geometrical view of positioning with carrier phase measurements. The measurement from each satellite is ambiguous in whole cycles, and what we are left to work with in each case is a fractional cycle. The user location is the point in space where plane wave fronts from the different satellites meet in the observed phase relationship. Let us consider an idealized 2-D case: All error terms in (7.5) are zero and the clocks are all synchronized. Wave fronts from each satellite separated by one wavelength ($\approx$ 20 cm) are shown in Figure 7.2 in the neighborhood of interest. We have frozen the picture at the instant of the measurements. The wave fronts are drawn so that the fractional cycle part of the carrier phase at each point along a wave front matches the measurement from that satellite. If we had measurements from only two satellites, the position could be any of the points at which the two sets of wave fronts intersect. Measurements from a third satellite resolve this uncertainty. There is only one point in the area of search which conforms to the measurements from all three satellite: The user is located where the wave fronts from the three satellites intersect at a point. We'll return to Figure 7.2 briefly in Section 7.5.

7.2.2 Precise Relative Positioning and Navigation

The surveyors refer to the relative position vector to be measured as the baseline vector or, simply, *baseline*. In the pre-satellite era, traditional surveying methods were based on measurements of distances and angles. In the 1960s, such measurements were taken with electronic distance meters (EDM) and theodolites. The process consisted of selecting a starting point located in the middle of the area of interest, which could be a region, a country, or a continent, and determining its positional coordinates from astronomical observations. A network of sur-

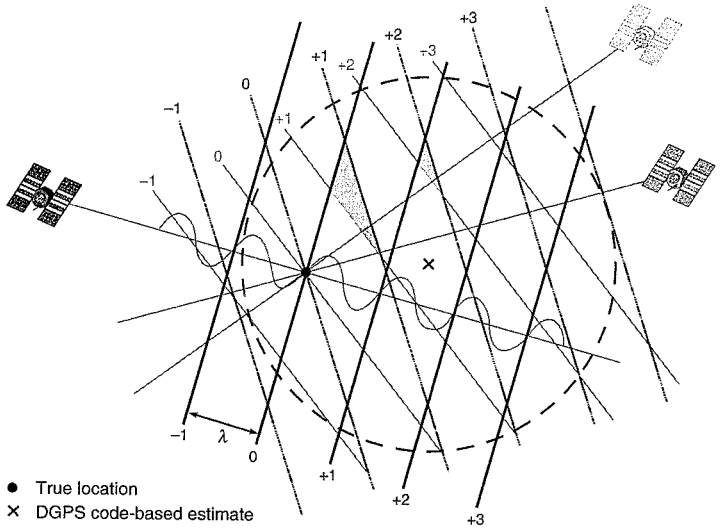


Figure 7.2 Positioning with carrier phase measurements. The user is located at a point where the wave fronts from the different satellites meet in the observed phase relationship.

vey control stations covering the area of interest would then be set up. These stations were required to be inter-visible and were often located on hilltops. The baselines were typically limited to tens of kilometers. The accuracy of relative positioning generally decreases approximately in proportion to the baseline length, and the positioning error is often normalized by the baseline length. A position vector with an error of 1 cm over a baseline of 10 km is said to be accurate to one part per million (or, 1 ppm).

In 1980, the United States was covered with a network of markers located most carefully using the traditional surveying methods with lengths measured between them with an accuracy of about 5 ppm with typical baselines of tens of kilometers. These so-called first-order markers served as the basis for scientific studies requiring exact geodetic data. There was a denser network of second-order survey markers, connected with the national network and covering high-value land areas, that was accurate to about one part in 10^5. A network of denser third-order markers accurate to about one part in 10^4 served as controls for local developments and to mark private property boundaries. GPS has now changed everything.

The potential of GPS for precise relative positioning was demonstrated in the early 1980s. It was shown that GPS L1 carrier phase measurements offered 1 ppm accuracy at night, and 2–5 ppm in daytime, over tens of kilometers. What's more, GPS did not require inter-visibility of the sites, an onerous requirement of the traditional methods. Availability of dual-frequency measurements improved the accuracy by an order of magnitude. With subsequent improvements in receivers and algorithms, accuracy to about one part in 10^9 (or, 1 mm over a 1000–km baseline) appears achievable.

Precise relative positioning using GPS requires simultaneous carrier phase observations at two points. The receiver with its antenna at the reference point is referred to as the *reference receiver*. The second receiver, whose antenna position is to be determined, is called the *mobile receiver* or *rover*. When the antennas at the two points are stationary during the observations, the process is called static relative positioning or *static survey*. In the 1980s, a GPS survey required data collection from antennas held stationary for an hour, or more, to make sure that the data would be adequate for integer ambiguity resolution in subsequent post-processing. The measurements were often limited to L1-only receivers, and the computational methods for the integer ambiguity resolution relied upon geometric diversity (i.e., change in satellite geometry) over the observation period, as in our example in Section 7.1. It was recognized quickly that field productivity would be improved if the rover could be free to rove, and its track could be determined relative to a fixed reference receiver. That's *kinematic survey*.

By static survey we really meant static initialization. In kinematic surveys, if the position estimates are not required in real time, static initialization may still work. The rover could start from the reference site and initialization may be carried out in one of two ways: (i) by placing the two antennas at the ends of a known baseline and taking measurements for a few seconds, or (ii) by an antenna swap. The rover is now free to move, and its position would be known within centimeters as long as carrier tracking is maintained at both receivers.

The next step in making surveys more efficient was to compute positions of the rover in real time in the field [Talbot (1993)]. That's *real-time kinematic* (RTK) mode. In RTK mode, the measurements at the reference receiver are transmitted to the rover on a radio link. A key feature required by RTK is the ability to estimate the integer ambiguities while the rover is in motion. That's *on-the-fly* (OTF) *initialization*, a principal focus of this chapter. (On-the-fly is an American expression meaning 'while on the move.') Of course, the initialization must be completed in real time.

Major receiver manufacturers now offer RTK solution packages consisting of a pair of receivers, a radio link, and software. Surveying aside, RTK techniques have found a number of industrial applications in earth moving, dredging, mining, construction and agriculture. For typical baselines of several kilometers, integer ambiguity resolution in thirty to sixty seconds is common, and the answers are generally right. The performance of an RTK system is measured by (i) initialization time, and (ii) reliability (or, correctness) of the integer estimates. There is an obvious trade-off between getting an answer quickly and getting it right.

RTK can also serve as a model for precise navigation if ambiguity resolution can be accomplished reliably and nearly instantaneously, ideally with measurements from a single epoch. The requirements of a navigation system are inherently different from those of surveying or geodesy. A navigation system is required to be robust because the consequences of failure may be severe. A general, robust navigation system based on carrier phase measurements for fast-moving platforms is yet to be demonstrated. We should, however, take note of an ingenious system developed at Stanford University which is being used to land unmanned air vehicles (UAVs) which fly over pseudolites on the ground, producing the requisite geometric diversity to resolve the integer ambiguities [Cohen *et al.* (1994)].

7.3 Elimination of Nuisance Parameters

We now formulate the problem of positioning with carrier phase measurements in all its complexity. In view of our previous discussion, we'll deal with measurements at two receivers from multiple satellites, each broadcasting at two frequencies. Additional notation is needed to

distinguish among these measurements. Let the carrier phase measurements from satellite k at the user receiver be modeled as

$$\phi_u^{(k)} = \lambda^{-1}[r_u^{(k)} - I_u^{(k)} + T_u^{(k)}] + f \cdot (\delta t_u - \delta t^{(k)}) + N_u^{(k)} + \varepsilon_{\phi,u}^{(k)} \tag{7.7}$$

The terms of (7.7) are now familiar to us. The only difference between (7.7) and (7.5) is that we have introduced a superscript to identify the satellite, and a subscript to identify the receiver. For now, we leave the carrier frequency unspecified.

Ideally, we would like clean equations involving only the measurements and position coordinates (or, geometric ranges). The unknown terms in (7.7) related to the ionosphere, troposphere, satellite and receiver clocks, and the integer ambiguities are simply nuisance parameters, meaning that we have no direct interest in them but need their values in order to relate the measurements to the parameters of interest. We could estimate the values of the nuisance parameters, but the estimates would not be error-free. An alternative is to eliminate these nuisance parameters from the equations and re-parametrize the problem as one of relative positioning. This is no surprise as we have been hinting at it relentlessly since the beginning of this chapter.

Suppose there is another receiver in the area that is also measuring carrier phases. For our immediate purposes, we suppose that this receiver is fixed at a known position. We'll call this the reference receiver, and denote its measurements from satellite k as

$$\phi_r^{(k)} = \lambda^{-1}[r_r^{(k)} - I_r^{(k)} + T_r^{(k)}] + f \cdot (\delta t_r - \delta t^{(k)}) + N_r^{(k)} + \varepsilon_{\phi,r}^{(k)} \tag{7.8}$$

Our objective in the rest of this section is to use (7.7) and (7.8) to eliminate the nuisance parameters and to obtain 'clean' equations relating the carrier phase measurements to the relative position vector between the reference and user antenna phase centers.

7.3.1 Single Difference

Let us form a between-receiver, single-difference carrier phase measurement, called simply *single difference*, by taking a difference of the measurements at the user and reference receivers at the same epoch. We resort again to the notation of double subscripts introduced in Section 4.8.2 to denote the difference between the measurements at two receivers and between parameter values associated with their models.

$$\begin{aligned}
\phi_{ur}^{(k)} &= \phi_u^{(k)} - \phi_r^{(k)} \\
&= \lambda^{-1}[(r_u^{(k)} - r_r^{(k)}) - (I_u^{(k)} - I_r^{(k)}) + (T_u^{(k)} - T_r^{(k)})] + f \cdot (\delta t_u - \delta t_r) \\
&\quad + (N_u^{(k)} - N_r^{(k)}) + (\varepsilon_{\phi,u}^{(k)} - \varepsilon_{\phi,r}^{(k)}) \\
&= \lambda^{-1}[r_{ur}^{(k)} - I_{ur}^{(k)} + T_{ur}^{(k)}] + f \cdot \delta t_{ur} + N_{ur}^{(k)} + \varepsilon_{\phi,ur}^{(k)}
\end{aligned} \tag{7.9}$$

where $(\bullet)_{ur} = (\bullet)_u - (\bullet)_r$. The single-difference ambiguities ($N_{ur}^{(k)}$) are still integers which may be positive or negative. The measurement noise in the single difference is larger by $\sqrt{2}$ than that in the measurements at either receiver. The between-receiver, single-difference phase measurements are sometimes represented as $\Delta\phi_{ur}^{(k)}$. In this notation, each term of (7.9) would be preceded by a 'Δ'. Our notation is on the minimal side.

By forming single differences we have cast the problem into one of relative positioning while eliminating the common-mode errors between measurements at the two receivers. We have one fewer parameter to deal with. The satellite clock error term $\delta t^{(k)}$ is common to the two measurements and cancels out in the difference. The satellite ephemeris error, to the extent it is common to the two sets of measurements, is also gone. The ionospheric and tropospheric terms in (7.9) are the differences of the corresponding delays at the two sites. The sizes of these terms and the residual ephemeris error would depend mainly upon the distance between the user and reference stations, or the baseline length.

We introduced the idea of relative positioning in Section 5.8.2 as an alternate implementation of DGPS. Actually, there is a subtle difference between the two. The differential corrections vary slowly. The useful life of a differential correction was several seconds while SA was active, and is several minutes now. In this section, we are dealing with actual measurements, which change much faster. The range rate can be up to 800 m/s and, therefore, the range can change by about four wavelengths in 1 ms. Clearly, when comparing carrier phase measurements from two receivers, we cannot disregard a difference of 1 ms in their measurement epochs.

We have to account for any significant difference in the measurement epochs. Real-time implementations invariably entail delays in data transmissions, typically one to two seconds. It is, therefore, required that the measurements at the reference station be extrapolated to match the epoch of the user receiver measurements. Estimating the receiver clock bias at each station to within 1 μs is not a challenge and, with an appropriate extrapolation technique, can reduce the error due to latency and epoch mismatch to under 1 cm.

In order to estimate the position with centimeter-level accuracy, the error terms in (7.9) must also be reduced to centimeter level. That's our *mantra* for this chapter. The carrier wavelength at L1 is about 19 cm, and an error of 10 cm (slightly larger than one-half cycle) could correspond to a change of one cycle in the integer ambiguity estimate. Actually, as we'll see later, our tolerance for measurement errors for estimation of integer ambiguities is less than one-quarter cycle. On the bright side, once the ambiguities are resolved correctly, the high-precision carrier phase measurements effectively turn into high-accuracy pseudorange measurements, and lead to position estimates of commensurate quality.

If the distance between the user and the reference station is 'short,' the residual ionospheric, tropospheric, and ephemeris errors in (7.9) would be small in comparison with the typical errors due to receiver noise and multipath. In fact, this appears to be a good definition of a *short baseline*. Note, however, that a 100-km baseline may qualify as short according to this criterion when the ionosphere is quiescent, and a 25-km baseline may not when it's not. It would be smart to correct for the tropospheric delay in the measurements at each location separately.

Our model for single-difference measurements for a short baseline is

$$\phi_{ur}^{(k)} = \lambda^{-1} r_{ur}^{(k)} + f \cdot \delta t_{ur} + N_{ur}^{(k)} + \varepsilon_{\phi,ur}^{(k)} \tag{7.10}$$

We can write an expression similar to (7.10) for single difference of code measurements. In fact, we already have in (5.56).

$$\rho_{ur}^{(k)} = r_{ur}^{(k)} + c \cdot \delta t_{ur} + \varepsilon_{\rho,ur}^{(k)} \tag{7.11a}$$

As an example, we present single difference code and carrier phase measurements from an experiment with a baseline of about 150 m. The measurements were taken at one-second intervals over a couple of hours at the two ends of the baseline using dual-frequency, geodetic-

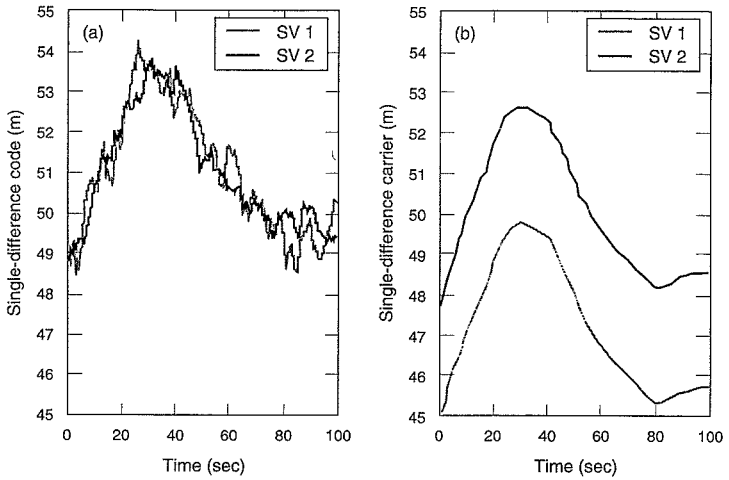


Figure 7.3 Code and carrier phase single differences from a 150-meter baseline. (Courtesy of Dr. Jaewoo Jung, Trimble Navigation)

quality receivers. The single differences over a 100-second time interval are plotted in Figure 7.3 for measurements at L1 from two of the satellites. For an easy side-by-side comparison, the carrier phase measurements have been converted to units of length by multiplying by the wavelength.

$$\Phi_{ur}^{(k)} = \lambda \cdot \phi_{ur}^{(k)} = r_{ur}^{(k)} + c \cdot \delta t_{ur} + \lambda \cdot N_{ur}^{(k)} + \lambda \cdot \varepsilon_{\phi,ur}^{(k)} \tag{7.11b}$$

It is interesting to note that the pattern of variation with time is common to all four plots. This is the pattern of change in the difference of the two receiver clock biases. As expected, the code measurements are noisy, with about 1-m-level random error. We'll return to this data set below.

Estimation of Position and Change in Position: The Role of Geometric Diversity

Let us now consider position estimation based on (7.10). The relative position vector to be estimated, $\mathbf{x}_{ur} = \mathbf{x}_u - \mathbf{x}_r$, is hiding in the range difference term. As in Chapter 6, we use boldface characters (Roman and Greek) to denote vectors and matrices. The following relationship is obtained from Figure 7.4, noting that the baseline is shorter than the distance to a satellite by orders of magnitude.

$$r_{ur}^{(k)} = r_u^{(k)} - r_r^{(k)} = -\mathbf{1}_r^{(k)} \cdot \mathbf{x}_{ur} \tag{7.12}$$

where $\mathbf{1}_r^{(k)}$ is the unit vector pointing to satellite k from the reference receiver position. The above approximation to the single-difference range has the benefit of simplicity and is adequate for short baselines (say, shorter than 10 km). A more careful user may define the line-of-sight unit vector from the mid-point between the reference and user receivers. Long baselines would require a slightly different treatment.

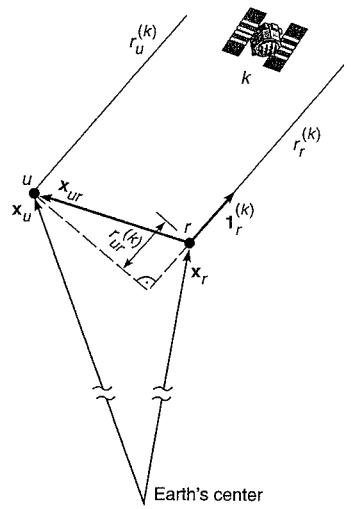


Figure 7.4 Geometry of the single-difference measurements.

The single differences formed from a snapshot of measurements from all K satellites in view at the user and reference stations can be written in vector-matrix notation as below.

$$\boldsymbol{\phi}_{ur} = \lambda^{-1} \begin{bmatrix} (-\mathbf{1}_r^{(1)})^T \\ (-\mathbf{1}_r^{(2)})^T \\ \vdots \\ (-\mathbf{1}_r^{(K)})^T \end{bmatrix} \mathbf{x}_{ur} + f \cdot \delta t_{ur} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} N_{ur}^{(1)} \\ N_{ur}^{(2)} \\ \vdots \\ N_{ur}^{(K)} \end{bmatrix} + \boldsymbol{\varepsilon}_{\phi,ur} \tag{7.13}$$

To simplify the notation in (7.13) further, we write the carrier phase as $\Phi_{ur}^{(k)} = \lambda \phi_{ur}^{(k)}$ and receiver clock delay as $b_{ur} = c \cdot \delta t_{ur}$, both in units of length. Note that the clock bias and the integers are not determined uniquely by (7.13). We can raise or lower all integer ambiguities by a common amount and the equations would still be satisfied if we compensate for this change in the clock bias. We, therefore, redefine the unknowns as below.

$$\boldsymbol{\Phi}_{ur} = \begin{bmatrix} \left(-\mathbf{1}_r^{(1)}\right)^T & 1 & 0 & \cdots & 0 \\ \left(-\mathbf{1}_r^{(2)}\right)^T & 1 & & & \\ \vdots & \vdots & & \mathbf{I} & \\ \left(-\mathbf{1}_r^{(K)}\right)^T & 1 & & & \end{bmatrix} \begin{bmatrix} \mathbf{x}_{ur} \\ b_{ur} + \lambda N_{ur}^{(1)} \\ \lambda\left(N_{ur}^{(2)} - N_{ur}^{(1)}\right) \\ \vdots \\ \lambda\left(N_{ur}^{(K)} - N_{ur}^{(1)}\right) \end{bmatrix} + \boldsymbol{\varepsilon}_{\Phi,ur}$$

where $\mathbf{I}$ is a $(K-1) \times (K-1)$ identity matrix. Rearranging terms,

$$\mathbf{\Phi}_{ur} = \begin{bmatrix} \left(-\mathbf{1}_r^{(1)}\right)^T & 1 \\ \left(-\mathbf{1}_r^{(2)}\right)^T & 1 \\ \vdots & \vdots \\ \left(-\mathbf{1}_r^{(K)}\right)^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_{ur} \\ b_{ur} + \lambda N_{ur}^{(1)} \end{bmatrix} + \begin{bmatrix} 0 \\ \lambda\left(N_{ur}^{(2)} - N_{ur}^{(1)}\right) \\ \vdots \\ \lambda\left(N_{ur}^{(K)} - N_{ur}^{(1)}\right) \end{bmatrix} + \boldsymbol{\varepsilon}_{\Phi,ur} \tag{7.14}$$

The reason for rearranging terms should now be clear. The first matrix on the right-hand side is the familiar geometry matrix **G** defined in (6.10), which played an important role in determining the quality of position estimates in the previous chapter. The second term, a vector whose elements are the integers of interest, will remain fixed as long as the carrier is tracked continuously.

Consider now the first differences measured at two time epochs, t_0 and t_1. We assume that the carrier tracking was continuous between the two time epochs and, therefore,

$$\mathbf{\Phi}_{ur}(t_1) - \mathbf{\Phi}_{ur}(t_0) = \mathbf{G}(t_1)\begin{bmatrix} \mathbf{x}_{ur}(t_1) \\ b_{ur}(t_1) \end{bmatrix} - \mathbf{G}(t_0)\begin{bmatrix} \mathbf{x}_{ur}(t_0) \\ b_{ur}(t_0) \end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_{\Phi,ur} \tag{7.15}$$

We have formed between-receiver, between-measurement-epochs difference to make an important point. Denoting $\mathbf{x}_{ur}(t_1) = \mathbf{x}_{ur}(t_0) + \delta\mathbf{x}_{ur}$, and $b_{ur}(t_1) = b_{ur}(t_0) + \delta b_{ur}$, the above equation can be written as

$$\mathbf{\Phi}_{ur}(t_1) - \mathbf{\Phi}_{ur}(t_0) = \mathbf{G}(t_1)\begin{bmatrix} \delta\mathbf{x}_{ur} \\ \delta b_{ur} \end{bmatrix} + \left(\mathbf{G}(t_1) - \mathbf{G}(t_0)\right)\begin{bmatrix} \mathbf{x}_{ur}(t_0) \\ b_{ur}(t_0) \end{bmatrix} + \tilde{\boldsymbol{\varepsilon}}_{\Phi,ur} \tag{7.16}$$

This is an insightful result. In view of our experience with position estimation based on (6.9), we can conclude that:

(i) estimation of $\delta\mathbf{x}_{ur}$ and δb_{ur}, the changes in relative position and relative clock bias, is 'tied to' the geometry matrix at epoch t_1, and wouldn't be a problem, in general;
(ii) estimation of $\mathbf{x}_{ur}(t_0)$, the absolute position, is tied to change in the geometry matrix, $[\mathbf{G}(t_1) - \mathbf{G}(t_0)]$, and would be problematic if such change is not significant;
(iii) $b_{ur}(t_0)$ is actually 'absent' from (7.16) and can't be estimated at all.

We should make two points. First, we have encountered similar results in Section 6.2 while discussing estimation of position and velocity from measurements of Doppler shift or pseudorange rates. Secondly, the GPS satellites are far away and there is little change in user-satellite geometry over tens of seconds. Geometric diversity, however, can be introduced in industrial and agricultural applications through pseudolites [Cohen *et al.* (1994)].

7.3.2 Double Difference

We now return to the single difference (7.10) and attend to the relative receiver clock bias term δt_{ur}, a nuisance parameter, which is common to the single-difference measurements from all satellites at each epoch. This term can, therefore, be eliminated by forming between-receiver, between-satellite double-difference measurements, called simply *double differences*, as follows. Let's form single differences for carrier phase measurements from another satellite denoted as l. For a short baseline as defined above,

$$\begin{aligned} \phi_{ur}^{(l)} &= \phi_u^{(l)} - \phi_r^{(l)} \\ &= \lambda^{-1} r_{ur}^{(l)} + f \cdot \delta t_{ur} + N_{ur}^{(l)} + \varepsilon_{\phi,ur}^{(l)} \end{aligned} \tag{7.17}$$

The receiver clock error term in (7.17) is the same as in (7.10). Clearly, if we take the difference of the single differences for the satellites k and l, we would be rid of all satellite and receiver clock error terms from our measurements. We write this difference using a double-superscript notation to denote the difference between measurements from two satellites and between parameter values of their models. Subtracting (7.17) from (7.10), we obtain

$$\begin{aligned} \phi_{ur}^{(kl)} &= \phi_{ur}^{(k)} - \phi_{ur}^{(l)} \\ &= \lambda^{-1} r_{ur}^{(kl)} + N_{ur}^{(kl)} + \varepsilon_{\phi,ur}^{(kl)} \end{aligned} \tag{7.18}$$

where $(\bullet)_{ur}^{(kl)} = (\bullet)_{ur}^{(k)} - (\bullet)_{ur}^{(l)}$. In particular,

$$\phi_{ur}^{(kl)} = \left(\phi_u^{(k)} - \phi_r^{(k)}\right) - \left(\phi_u^{(l)} - \phi_r^{(l)}\right)$$

It is left as an exercise to show that we could have obtained (7.18) by first forming between-satellite single differences at the user and reference receivers, and then taking the difference of these. The double-difference phase measurements are sometimes represented as $\nabla\Delta\phi_{ur}^{(kl)}$. In this notation, each term of (7.18) would be preceded by a '$\nabla\Delta$'. We'll, however, stay with our minimalist notation.

For an example of double differences, we return to the measurements from our experiment with the 150-m baseline. Figure 7.5 gives plots of the code and carrier double differences formed from the single differences for the two satellites shown in Figure 7.3. Given the shortness of the baseline, there is little change in 100 seconds. The difference between the two plots is entirely due to the integer ambiguity and noise. The scale is too coarse to see the millimeter-level noise in the carrier phase measurements. We'll return to this example later for integer ambiguity resolution.

We now return to (7.18) and relate the range double-difference term to the relative position vector $\mathbf{x}_{ur}$. The corresponding measurement geometry is shown in Figure 7.6. From (7.12)

$$\begin{aligned} r_{ur}^{(kl)} &= (r_u^{(k)} - r_r^{(k)}) - (r_u^{(l)} - r_r^{(l)}) \\ &= -(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}) \cdot \mathbf{x}_{ur} \end{aligned} \tag{7.19}$$

In order to estimate $\mathbf{x}_{ur}$, we'll have to estimate $N_{ur}^{(kl)}$, the integer ambiguities associated with pairs of satellites and receivers.

With K satellites in view, there are $(K-1)$ double differences which can be written in vector-matrix notation as below. Designating satellite number 1 as the reference,

$$\begin{bmatrix} \phi_{ur}^{(21)} \\ \phi_{ur}^{(31)} \\ \vdots \\ \phi_{ur}^{(K1)} \end{bmatrix} = \lambda^{-1} \begin{bmatrix} -\left(\mathbf{1}_r^{(2)} - \mathbf{1}_r^{(1)}\right)^T \\ -\left(\mathbf{1}_r^{(3)} - \mathbf{1}_r^{(1)}\right)^T \\ \vdots \\ -\left(\mathbf{1}_r^{(K)} - \mathbf{1}_r^{(1)}\right)^T \end{bmatrix} \mathbf{x}_{ur} + \begin{bmatrix} N_{ur}^{(21)} \\ N_{ur}^{(31)} \\ \vdots \\ N_{ur}^{(K1)} \end{bmatrix} + \begin{bmatrix} \varepsilon_{\phi,ur}^{(21)} \\ \varepsilon_{\phi,ur}^{(31)} \\ \vdots \\ \varepsilon_{\phi,ur}^{(K1)} \end{bmatrix} \tag{7.20}$$

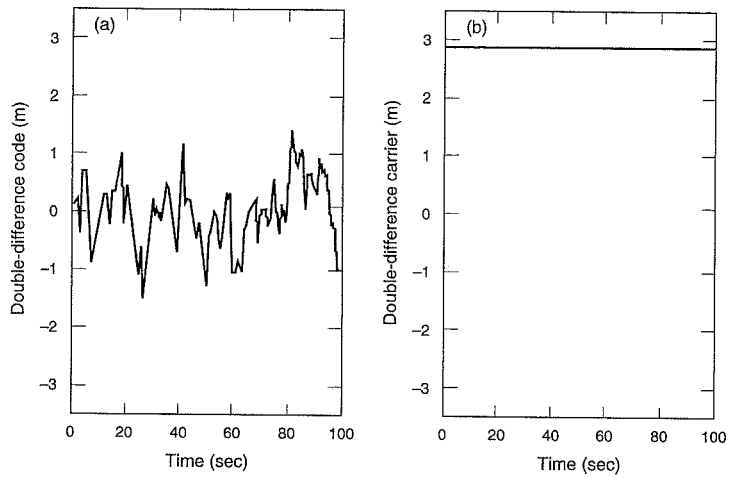


Figure 7.5 Code and carrier phase double differences from the 150-meter baseline. (Courtesy of Dr. Jaewoo Jung, Trimble Navigation)

Subscript *ur* is common to all terms and we'll drop it for simplicity if there is no chance of confusion.

Double differencing appears to have served a modest purpose when compared with single differencing. While single differencing changed the structure of the problem, double differencing essentially eliminated an unknown (receiver clock bias) at the expense of one equation, and redefined the ambiguities. In practice, double differencing also serves to eliminate certain hard-to-estimate error terms common to the measurements taken at the ends of a short baseline. An example is the frequency-dependent terms reflecting differential path delays in the satellites (T_{GD}) and the receiver, introduced in Section 5.3.2. The manufacturers of geodetic-quality receivers often 'optimize' receiver design for precise positioning using double differences. We'll return to the double differences in Section 7.4.

7.3.3 Triple Difference

The integer ambiguities in the double-difference carrier phase equation (7.20) are themselves nuisance parameters. So, why not try to eliminate them just as we eliminated the other nuisance parameters. Actually, we did this in the context of single differences above in Section 7.3.1. The integers remain fixed while both receivers maintain carrier lock and if we form differences of the double differences between two measurement epochs denoted as t_{i+1} and t_i,

$$\begin{aligned}\delta\phi_{ur}^{(kl)}(i) &= \phi_{ur}^{(kl)}(t_{i+1}) - \phi_{ur}^{(kl)}(t_i) \\ &= \lambda^{-1}\delta r_{ur}^{(kl)}(i) + \delta\varepsilon_{\phi,ur}^{(kl)}(i)\end{aligned} \tag{7.21}$$

where $\delta(\bullet)(i) = (\bullet)(t_{i+1}) - (\bullet)(t_i)$. Dropping the subscript *ur* and reference to measurement epoch, we rewrite (7.21) as

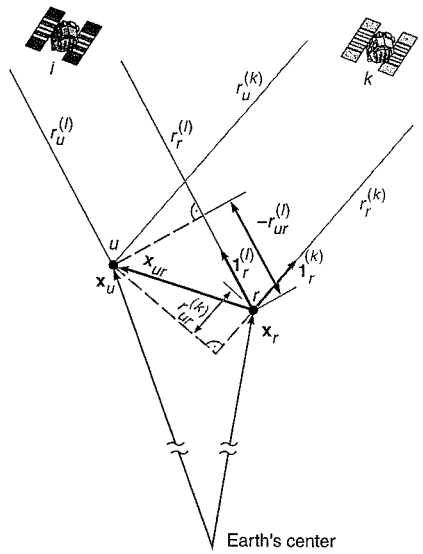


Figure 7.6 Geometry of the double-difference measurements.

$$\delta\phi^{(kl)} = \lambda^{-1}\delta r^{(kl)} + \delta\varepsilon_{\phi}^{(kl)} \tag{7.22}$$

If the user and the reference station are stationary, we can write the vector of triple differences formed from all satellites in view as

$$\delta\boldsymbol{\phi} = \lambda^{-1}\begin{bmatrix} -\delta\left(\mathbf{1}_r^{(2)} - \mathbf{1}_r^{(1)}\right)^T \\ -\delta\left(\mathbf{1}_r^{(3)} - \mathbf{1}_r^{(1)}\right)^T \\ \vdots \\ -\delta\left(\mathbf{1}_r^{(K)} - \mathbf{1}_r^{(1)}\right)^T \end{bmatrix}\mathbf{x} + \tilde{\boldsymbol{\varepsilon}} \tag{7.23}$$

We have now achieved what we had set out to do. We have clean equations involving only the carrier phase measurements and geometric ranges. All nuisance parameters are gone, but at a price. We lost geometrical leverage. We have seen the undifferenced version of (7.23) before when we discussed position estimation with delta pseudoranges or pseudorange rates [Section 6.2.2]. Equation (7.23) can be solved for the relative position vector by linearizing about an initial estimate, as before. We can also define a dilution of precision (DOP) parameter, as before, and will find the DOPs to be large numbers. The position estimates obtained from (7.23) tend to be less accurate than those from the double differences, in general. The triple differences, however, are useful in identifying discontinuities in carrier tracking resulting in loss of cycle count or cycle slip.

It is left as an exercise to show that for a user in motion we can derive equations similar to (7.16), leading to a similar conclusion. The triple differences can be used to keep track of

changes in user position, but it will take a significant change in geometry to be able to estimate the user's starting position [van Graas and Lee (1995)].

7.3.4 Integer Ambiguity Resolution and Position Estimation

We now turn to the problem of integer ambiguity resolution and position estimation for short baselines using the double differences represented as

$$\phi_{ur}^{(kl)}(t_i) = \lambda^{-1} r_{ur}^{(kl)}(t_i) + N_{ur}^{(kl)} + \varepsilon_{\phi,ur}^{(kl)}(t_i) \tag{7.24}$$

where we have included the time argument to emphasize that the integer ambiguities remain constant while the carrier phase tracking is maintained for both satellites at both receivers. As noted previously, the possibility of loss of lock makes the carrier phase measurements fragile, and their use for precise navigation is contingent upon our ability to resolve the integer ambiguities quickly and reliably. For simplicity, we'll now drop any explicit reference to the measurement epoch. We'll also drop the subscript *ur*, which is common to all terms. Besides, we need to distinguish between L1 and L2 measurements and will use the subscript q = L1 or L2 to denote the carrier frequency.

Equation (7.24) can be written in our new notation as

$$\phi_q^{(kl)} = \lambda_q^{-1} r^{(kl)} + N_q^{(kl)} + \varepsilon_{\phi,q}^{(kl)} \tag{7.25}$$

The code phase measurements can be written similarly except for the ambiguity terms. The single, double, and triple differences can be formed with the code measurements as well to eliminate the nuisance parameters. In particular, double-difference code phase measurement akin to (7.25) can be written as

$$\rho_q^{(kl)} = r^{(kl)} + \varepsilon_{\rho,q}^{(kl)} \tag{7.26}$$

With K satellites in view, we can form K single differences for the code and carrier each at L1, and K more at L2. At each frequency, we can generate a total of $K(K-1)$ double differences, but only $(K-1)$ of these are linearly independent. All other double differences can be formed as linear combinations of a linearly independent set and, therefore, contain no additional information. A linearly independent set of $(K-1)$ double differences can be selected in many ways. Recognizing that the errors in measurements from the reference satellite enter in each double difference, we often pick the highest satellite as the reference because the propagation and multipath errors associated with it are generally the smallest.

Equations (7.25) and (7.26) give our model for the double-difference carrier and code phase measurements for a short baseline. In the absence of significant multipath and differential ionospheric and tropospheric errors, we can characterize the measurement error terms based on our model for measurement accuracy [Section 5.4] and the noise amplification in the differencing process. The standard deviations of the errors in the carrier and code phase double differences are

$$\sigma\left(\varepsilon_{\phi,q}^{(kl)}\right) \approx 0.05 \text{ cycle } (\approx 1 \text{ cm}) \tag{7.27a}$$

$$\sigma\left(\varepsilon_{\rho,q}^{(kl)}\right) \approx 1 \text{ m} \tag{7.27b}$$

q = L1 or L2. These numbers are consistent with the data from our experiment (Figure 7.5).

A large number of applications requiring real-time position estimates fall in the category of short baselines (say, 10 km, or shorter). Long baselines extending hundreds, even thousands, of kilometers are of interest to geophysicists. The answers, however, are not required in real time, in general, and one can post-process a long stretch of data using precise ephemerides and accounting explicitly for differential ionospheric and tropospheric delays [Bock (1996)]. We now focus below on integer ambiguity resolution in the short-baseline model (7.25) and (7.26) exclusively.

7.4 Resolving Ambiguities One at a Time

7.4.1 Using Code Measurements to Estimate Integers

A simple approach to ambiguity resolution is suggested by the equation set (7.25) and (7.26): Use the code measurements, which are unambiguous, to estimate the integers in the carrier phase measurements. We expect the success of this approach to depend upon the accuracy of the code measurements. This approach deals with measurements from one pair of satellites at a time and is often referred to as ambiguity resolution in the measurement domain. The approach is also called *geometry-free approach*, though the value-laden word 'free' may be misleading insofar as we only free ourselves of whatever advantage may accrue from redundant measurements. But this approach is certainly useful as a conceptual exercise. We follow a lucid treatment by Hatch (1996).

In this section, we deal with the double-difference measurements from a specific pair of satellites, and will drop the superscript (kl) for simplicity. Recall that the subscript denotes the carrier frequency (L1 or L2). Let us look at the double-difference code and carrier phase measurements at L1.

$$\begin{aligned} \phi_{L1} &= \frac{r}{\lambda_{L1}} + N_{L1} + \varepsilon_{\phi_{L1}} \\ \rho_{L1} &= r + \varepsilon_{\rho_{L1}} \end{aligned} \tag{7.28}$$

We can form an estimate of N_{L1} as

$$\hat{N}_{L1} = \left[\phi_{L1} - \frac{\rho_{L1}}{\lambda_{L1}}\right]_{roundoff} \tag{7.29}$$

How good is this estimate? From our error model (7.27), $\sigma(\phi_{L1}) \approx 0.05$ cycle, and $\sigma(\rho_{L1}) \approx 1$ m. The L1 wavelength is about 0.2 m. Therefore, the standard deviation of the estimate of N_{L1} is about 5 cycles: $\sigma(\hat{N}_{L1}) \approx 5$ cycles. That's a lot of error. On the bright side, N_{L1} does not change as long as there is no loss of lock or cycle slips and, in principle, we can reduce the uncertainty in its estimate by averaging over a sequence of estimates and rounding off to the nearest integer. The measurement errors, however, are highly correlated over time and it will take a long stretch of clean data to get a good estimate. We'll have to average uncorrelated measurements from over 100 epochs to reduce the integer estimation error to one-half cycle. In order for us to have faith in an estimate, the uncertainty would have to be reduced to much less than that.

We now return to our measurements from the 150-meter baseline to see how well we can estimate the integers. It's a rather easy problem with a very short baseline and a relatively

clean antenna environment. The double-difference code and carrier phase measurements (Figure 7.5) appear consistent with our model for the measurement errors. Can you guess the value of the unknown integer for this satellite pair from Figure 7.5? The correct answer is 15. The estimates computed epoch by epoch in accordance with (7.29) are shown in Figure 7.7. Actually, only the estimation error is shown. The real-valued estimates have been connected with lines. The rounded-off integers are shown as dots. There are no surprises. Even in this easy case, we have to average the estimates over tens of seconds to approach the right answer. The sample standard deviation of the error in integer estimates is 3.75 cycles. Our simple model had predicted a standard deviation of about 5 cycles. That's pretty good agreement.

Now we are ready for a methodical treatment of integer ambiguity resolution by combining the code and carrier phase measurements from a dual-frequency receiver. Following Yang, Goad, and Schaffrin (1994), the double-difference code and carrier phase measurements in units of length at epoch t_i are written as

$$\begin{aligned}
\rho_{L1}(i) &= r(i) + \varepsilon_{\rho_{L1}}(i) \\
\Phi_{L1}(i) &= r(i) + \lambda_{L1}N_{L1} + \varepsilon_{\Phi_{L1}}(i) \\
\rho_{L2}(i) &= r(i) + \varepsilon_{\rho_{L2}}(i) \\
\Phi_{L2}(i) &= r(i) + \lambda_{L2}N_{L2} + \varepsilon_{\Phi_{L2}}(i)
\end{aligned} \tag{7.30}$$

or, in matrix notation, as

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & \lambda_{L1} & 0 \\ 1 & 0 & 0 \\ 1 & 0 & \lambda_{L2} \end{bmatrix} \begin{bmatrix} r \\ N_{L1} \\ N_{L2} \end{bmatrix} = \begin{bmatrix} \rho_{L1} \\ \Phi_{L1} \\ \rho_{L2} \\ \Phi_{L2} \end{bmatrix} \tag{7.31}$$

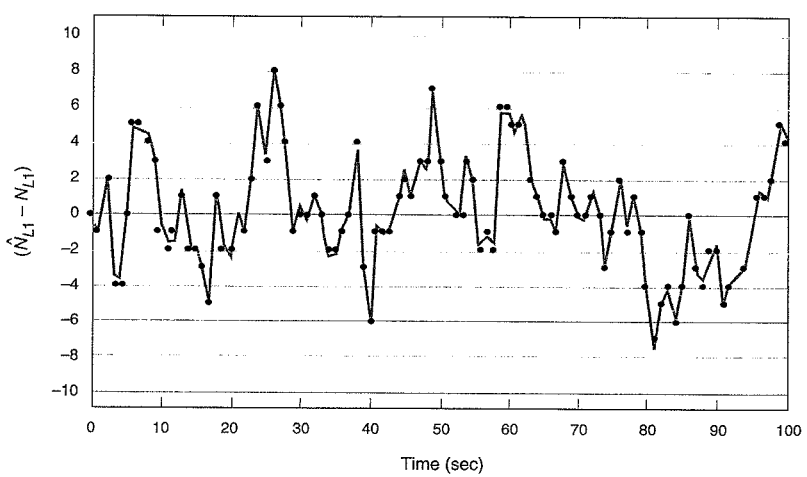


Figure 7.7 Integer estimates for L1 double-difference measurements from the 150-meter baseline. (Courtesy of Dr. Jaewoo Jung, Trimble Navigation)

Measurements at a single epoch give four equations in three unknowns. The measurements, however, are not error-free. The quality of the code and carrier measurements is also different. On the basis of our error model for the code and carrier phase measurements, we set up a diagonal weight matrix

$$\mathbf{W} = \begin{bmatrix} 1 & & & \\ & 1/0.01^2 & & \\ & & 1 & \\ & & & 1/0.01^2 \end{bmatrix}$$

and look for a weighted least-squares solution for the integers. We can accumulate measurements from multiple epochs while the carrier is tracked continuously. Each new set of measurements provides four new equations and introduces one new unknown (r). We'll have an over-determined set of equations to solve for N_{L1} and N_{L2}. Ideally, the unconstrained estimates would be close to integers, and can be rounded off to provide the integer estimates $\hat{N}_{L1}$ and $\hat{N}_{L2}$.

How to tell if the integer estimates are good? For a short baseline, we had assumed that the ionospheric differential delay in (7.30) is negligible. The validity of this assumption and the correctness of the integers can both be checked out on the basis of (5.33) by verifying that

$$I_{L1} = \frac{f_{L2}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\left[\left(\Phi_{L1} - \lambda_{L1}\hat{N}_{L1}\right) - \left(\Phi_{L2} - \lambda_{L2}\hat{N}_{L2}\right)\right] \approx 0 \tag{7.32}$$

Alternatively, we can form a simple test for the integer estimates based directly on (7.29)

$$60\hat{N}_{L1} - 77\hat{N}_{L2} = \left[60\frac{\Phi_{L1}}{\lambda_{L1}} - 77\frac{\Phi_{L2}}{\lambda_{L2}}\right]_{roundoff} \tag{7.33}$$

Recall that $f_{L2}/f_{L1} = \lambda_{L1}/\lambda_{L2} = 60/77$.

In principle, this approach to integer ambiguity resolution is applicable whether the user is stationary or in motion. In practice, this approach would require long periods of clean data. An approach that deals with measurements from one satellite pair at a time has an obvious shortcoming: It fails to take advantage of the fact that the measurements from all satellites form a consistent set. There is no benefit derived from the redundant measurements. Below, we discuss alternate approaches which aim to estimate the entire ambiguity vector on the basis of the full set of measurements.

Once the integers are known, estimation of the relative position vector is straightforward, as discussed below, following the same approach as given in Section 6.1.1 for estimation of position and clock bias from the pseudorange measurements.

7.4.2 Dual-Frequency Measurements: Wide Laning

We saw in (7.29) that the uncertainty in integer estimation depends upon the carrier wavelength. The longer the wavelength, the better the estimate. The L2 wavelength is only slightly longer than the L1 wavelength, and the estimate of N_{L2} wouldn't be much better than that for N_{L1}. But, given the measurements at L1 and L2, we can 'create' a signal with significantly lon-

ger wavelength as follows. Define a carrier phase measurement

$$\phi_{L12} = \phi_{L1} - \phi_{L2} = r\left(\frac{1}{\lambda_{L1}} - \frac{1}{\lambda_{L2}}\right) + (N_{L1} - N_{L2}) + \varepsilon_{\phi_{L12}}$$
$$= r\left(\frac{f_{L1} - f_{L2}}{c}\right) + (N_{L1} - N_{L2}) + \varepsilon_{\phi_{L12}}$$
$$= \frac{r}{\lambda_{L12}} + N_{L12} + \varepsilon_{\phi_{L12}} \tag{7.34}$$

where $\lambda_{L12} = c/(f_{L1} - f_{L2}) = 0.862$ m is the wavelength of our newly defined measurements ϕ_{L12}, called *wide-lane measurements*. The corresponding frequency is $f_{L12} = (f_{L1} - f_{L2}) = 347.82$ MHz, and $N_{L12} = (N_{L1} - N_{L2})$ represents the integer ambiguity in the wide-lane double differences.

We can form an estimate of N_{L12} as

$$\hat{N}_{L12} = \left[\phi_{L12} - \frac{\rho_{L1}}{\lambda_{L12}}\right]_{roundoff} \tag{7.35}$$

Proceeding as before, we determine that $\sigma(\hat{N}_{L12}) \approx 1.2$ cycles. In principle, uncorrelated measurements from ten epochs would reduce the standard deviation of the estimate to less than one-half cycle. Clearly, the job of estimating the wide-lane ambiguity is easier than that of estimating N_{L1} or N_{L2}. This benefit accrues from the availability of measurements at multiple frequencies, and we'll refer to it as *frequency diversity*.

Returning to our measurements from the 150-m baseline for the last time, Figure 7.8 gives a plot of the error in the wide-lane integer estimates generated in accordance with (7.35) second by second. Again, there are no surprises. We get the right answer about half the time. The standard deviation of the estimation error is about 0.8 cycles, as compared to 1.2 cycles predicted by our simple model. As noted earlier, this was a rather easy problem but the data set served a useful purpose.

We have now shown that N_{L12} is easier to estimate than N_{L1} or N_{L2}. So, why not forget about N_{L1} and N_{L2}, and simply estimate N_{L12} and obtain the relative position vector from it? Yes, you can, if you are happy with the position accuracy, which we discuss below. As we'll see, the wide-lane measurements are significantly noisier than the L1 or L2 measurements.

Once the ambiguities are resolved, we'll represent carrier phase in units of length for position estimation. As before, denoting $\Phi_q = \lambda_q \phi_q$, where q = L1, L2, or L12, we solve

$$r = (\Phi_q - \lambda_q N_q)$$

for the position estimate. It is left as an exercise to show that

$$\Phi_{L12} = \frac{f_{L1}}{(f_{L1} - f_{L2})}\Phi_{L1} - \frac{f_{L2}}{(f_{L1} - f_{L2})}\Phi_{L2}$$
$$= \frac{154}{34}\Phi_{L1} - \frac{120}{34}\Phi_{L2} \tag{7.36}$$

The large coefficients amplify the noise in Φ_{L12} to about six times the noise level in Φ_{L1} or Φ_{L2}.

$$\sqrt{\left(\frac{154}{34}\right)^2 + \left(\frac{120}{34}\right)^2} = 5.7$$

This is the downside of the wide-lane measurements. But we can define other linear combinations of the carrier phase measurements with different properties. Consider, for example, the so-called *narrow-lane measurements* (or, *Ln* measurements)

$$\phi_{Ln} = \phi_{L1} + \phi_{L2} \tag{7.37}$$

The frequency of ϕ_{Ln} is $(f_{L1} + f_{L2})$, and wavelength $\lambda_{Ln} = 10.70$ cm. We know that the narrow wavelength would make it harder to resolve the corresponding integer ambiguity, but there is a compensation: $\Phi_{Ln} = \lambda_{Ln}\phi_{Ln}$ can be written as

$$\Phi_{Ln} = \frac{154}{274}\Phi_{L1} + \frac{120}{274}\Phi_{L2} \tag{7.38}$$

The small coefficients make the narrow-lane measurements less noisy than those at L1 and L2. The position estimates based on the narrow-lane measurements would be more precise than those from the wide-lane measurements. Other useful linear combinations of the dual-frequency measurements are considered in the Homework Problems 7-4 and 7-5.

Having estimated N_{L12} correctly, we can estimate N_{L1} and N_{L2} as follows [Dedes and Goad (1994)]. From the measurement equations

$$\phi_{L1} = \frac{r}{\lambda_{L1}} + N_{L1} + \tilde{\varepsilon}_{\phi_{L1}}$$
$$\phi_{L2} = \frac{r}{\lambda_{L2}} + N_{L2} + \tilde{\varepsilon}_{\phi_{L2}}$$

we obtain

$$N_{L1} - \frac{\lambda_{L2}}{\lambda_{L1}}N_{L2} = \phi_{L1} - \frac{\lambda_{L2}}{\lambda_{L1}}\phi_{L2} + \varepsilon \tag{7.39}$$

And, we already have

$$N_{L1} - N_{L2} = N_{L12} \tag{7.40}$$

From (7.39) and (7.40), we can solve for N_{L1} and N_{L2}.

$$\hat{N}_{L1} = \left(\frac{\lambda_{L2}}{\lambda_{L1}} - 1\right)^{-1}\left(\frac{\lambda_{L2}}{\lambda_{L1}}N_{L12} - \phi_{L1} + \frac{\lambda_{L2}}{\lambda_{L1}}\phi_{L2}\right) \tag{7.41}$$

The standard deviation of this estimate is

$$\sigma(\hat{N}_{L1}) = \left(\frac{\lambda_{L2}}{\lambda_{L1}} - 1\right)^{-1}\sqrt{2.65}\,\sigma(\varepsilon_{\phi_{L1}})$$
$$\approx 6\,\sigma(\varepsilon_{\phi_{L1}}) \tag{7.42}$$

Having estimated the wide-lane integer ambiguity correctly, the success in estimation of the integer ambiguities at L1 and L2 depends upon the quality of the measurements. Standard

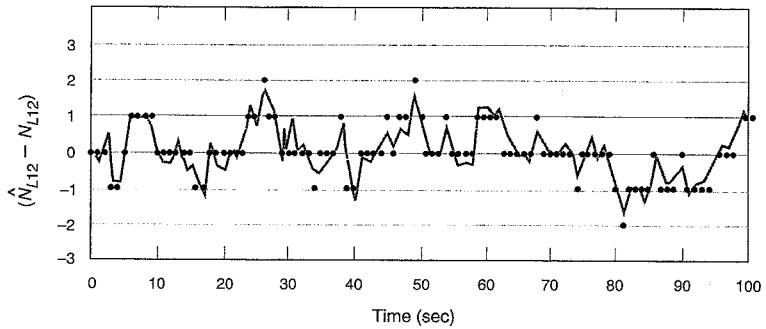


Figure 7.8 Integer estimates for the wide-lane double-difference measurements from the 150-meter baseline. (Courtesy of Dr. Jaewoo Jung, Trimble Navigation)

deviation of 0.05 cycle of a double-difference measurement at L1 or L2 would be magnified to 0.3 cycle in estimation of N_{L1} or N_{L2}, and we may have to resort to averaging the estimates over time to reduce the uncertainty.

7.4.3 Three-Frequency Measurements: L1, L2, and L5

The planned GPS modernization [Section 3.3] includes, among other enhancements, addition of a third civil signal at L5 centered at 1176.45 MHz with a higher power and much longer PRN code than the C/A-codes, and chipping rate of 10.23 Mcps. Our immediate concern, however, is limited to the carrier phase measurements. We discuss below the potential benefits of the three-frequency carrier phase measurements to become available in the next few years. The three frequencies are: 1575.42 MHz (L1), 1227.6 MHz (L2), and 1176.45 MHz (L5).

As before, our model for the double-difference carrier and code phase measurements from a short baseline is:

$$\begin{aligned} \phi_q &= \frac{r}{\lambda_q} + N_q + \varepsilon_{\phi_q} \\ \rho_q &= r + \varepsilon_{\rho_q} \end{aligned} \qquad (7.43)$$

where q = L1, L2, or L5; $\sigma(\phi_q) = 0.05$ cycle; and $\sigma(\rho_q) = 1$ m. With three civil frequencies, a receiver can generate three beat-frequency signals.

$\phi_{L12} = \phi_{L1} - \phi_{L2}$	wide-lane (WL) signal
$\phi_{L15} = \phi_{L1} - \phi_{L5}$	medium-lane (ML) signal
$\phi_{L25} = \phi_{L2} - \phi_{L5}$	extra-wide-lane (EWL) signal

We have already discussed the wide-lane signal above. The medium-lane signal has a wavelength of 0.751 m. The extra-wide-lane signal has a wavelength of 5.861 m. Table 7.2 gives a summary of the signal characteristics.

In the analysis below, we essentially follow the approach laid out in the previous section. The basic idea is to resolve the integer ambiguities in steps starting with the easiest case of

Table 7.2. Carriers and beat-frequency signals to be available when GPS Modernization is completed

Carrier Signal	*Frequency (MHz)*	*Wavelength (m)*
L1	$f_{L1} = 1575.42$	$\lambda_{L1} = 0.190$
L2	$f_{L2} = 1227.60$	$\lambda_{L2} = 0.244$
L5	$f_{L5} = 1176.45$	$\lambda_{L5} = 0.255$
L1–L5	$f_{L15} = 398.97$	$\lambda_{L15} = 0.751$ (medium lane)
L1–L2	$f_{L12} = 347.82$	$\lambda_{L12} = 0.862$ (wide lane)
L2–L5	$f_{L25} = 51.15$	$\lambda_{L25} = 5.861$ (extra-wide lane)

extra-wide-lane signal. Estimation of the extra-wide-lane integer gives us leverage to estimate the wide-lane integer, which, in turn, helps with estimation of the L1 integer. We refine the estimate of double-difference range (r) at each step, using the notation $\hat{r}(i)$ for the estimate corresponding to the ith step.

Step 1. Start with the best initial estimate of the double-difference range available: double-difference pseudorange from measurements at L5.

$$\hat{r}(1) = \rho_{L5}, \quad \sigma\left[\hat{r}(1)\right] \approx 1\,\text{m}$$

Step 2. Estimate the integer ambiguity for the extra-wide-lane signal using the estimate of double-difference range from Step 1 as

$$\hat{N}_{L25} = \left[\left(\phi_{L2} - \phi_{L5}\right) - \frac{\hat{r}(1)}{\lambda_{L25}} \right]_{roundoff} \qquad (7.44)$$

$\lambda_{L25} = 5.86$ m and, therefore, $\sigma(\hat{N}_{L25}) \approx 0.2$ cycles. Not bad! It appears that we can trust the integer estimate obtained with measurements from a single epoch. Estimating the extra-wide-lane integer seems easy and, having estimated it correctly (i.e., $\hat{N}_{L25} = N_{L25}$), we can refine our estimate of the double-difference range as

$$\hat{r}(2) = \lambda_{L25}\left(\phi_{L2} - \phi_{L5} - \hat{N}_{L25}\right), \quad \sigma\left[\hat{r}(2)\right] \approx 40\,\text{cm}$$

Range estimates with sub-meter error would lead to position estimates with sub-meter error. We could stop here and be happy. If not, proceed to Step 3.

Step 3. Use the refined estimate of the double-difference range obtained in Step 2 to estimate the wide-lane integer N_{L12} as

$$\begin{aligned} \hat{N}_{L12} &= \left[\left(\phi_{L1} - \phi_{L2}\right) - \frac{\hat{r}(2)}{\lambda_{L12}} \right]_{roundoff} \\ \sigma\left(\hat{N}_{L12}\right) &\approx 0.5 \text{ cycles} \end{aligned} \qquad (7.45)$$

Now, that's not such good news. In order to feel comfortable with this estimate of wide-lane integer, we have to reduce its standard deviation to half as much and that means averaging, as discussed previously in Section 7.4.2. Once we have successfully resolved the wide-lane integer ambiguity, we can refine further the estimate of double-difference range.

$$\hat{r}(3) = \lambda_{L12}\left(\phi_{L1} - \phi_{L2} - \hat{N}_{L12}\right), \quad \sigma\left(\hat{r}(3)\right) \approx 6\,\text{cm}$$

Such estimates of double-difference ranges would lead to a position estimate with error of several centimeters. Again, if we can find happiness in such an estimate, we can stop. If not, we proceed to the next step.

Step 4. Use the estimate of the double-difference range from Step 3 to estimate the integer ambiguity at L1.

$$\hat{N}_{L12} = \left[\left(\phi_{L1} - \phi_{L2}\right) - \frac{\hat{r}(2)}{\lambda_{L12}}\right]_{roundoff}$$
$$\sigma\left(\hat{N}_{L12}\right) \approx 0.3\,\text{cycles} \tag{7.46}$$

That's the end of the line. Having successfully resolved the L1 integer ambiguity, we now have range estimates with standard deviation of 5 mm, and can get position estimates of commensurate quality.

So, what would the third frequency (L5) buy us for precise positioning? Well, we can summarize the above discussion. We would generate the extra-wide-lane measurements, and the corresponding integer would be easy to estimate if the errors can be controlled. Having resolved the extra-wide-lane ambiguity, the problem of estimating the wide-lane integer would be considerably easier than before, though still not a sure thing. The next step of estimating integers corresponding to L1 and L2 would remain unchanged. Note that the analysis above is based on our model for the measurement data quality (Table 7.1). If the measurements are worse, so would be the results.

7.5 Resolving Ambiguities as a Set

We now deal exclusively with double-difference carrier phase measurements currently available, and switch back to units of cycles. The carrier phase measurements at L1 from K satellites give $(K-1)$ double-difference equations at each epoch. A dual-frequency receiver would provide another set of $(K-1)$ equations from the measurements at L2. For simplicity of notation, we assign number 1 to the reference satellite ($l = 1$), and write a set of $(K-1)$ double-difference equations for measurements at carrier frequency q at epoch t_i as

$$\begin{aligned}
\phi_q^{(21)}(i) &= \lambda_q^{-1} r^{(21)}(i) + N_q^{(21)} + \varepsilon_{\phi,q}^{(21)}(i) \\
\phi_q^{(31)}(i) &= \lambda_q^{-1} r^{(31)}(i) + N_q^{(31)} + \varepsilon_{\phi,q}^{(31)}(i) \\
&\vdots \\
\phi_q^{(K1)}(i) &= \lambda_q^{-1} r^{(K1)}(i) + N_q^{(K1)} + \varepsilon_{\phi,q}^{(K1)}(i)
\end{aligned} \tag{7.47}$$

The equation set (7.47) is simply an expanded view of what we wrote in shorthand notation in (7.25) for the double-difference measurements at an epoch. The next epoch (t_{i+1}) brings another set of $(K-1)$ equations at each frequency. While the phase tracking is continuous, three new unknowns (new position coordinates) are introduced at each epoch if the user is in motion, and none if the user is stationary. With dual-frequency measurements, we have a choice to work with the wide-lane measurements (L12), alone or in conjunction with L1 or L2 measurements.

In principle, estimation of the integers in (7.47) poses no special problem if there is no time pressure and the unmodeled errors are not unduly large. Consider static initialization. There are $[3 + (K-1)]$ unknowns for single frequency measurements and $[3 + 2(K-1)]$ unknowns for dual-frequency measurements and, over time, many more equations. We expect the estimation process to be helped by redundant measurements, good satellite geometry, dual-frequency measurements, and significant change in satellite geometry over the observation period. Unmodeled errors would hurt. The errors generally grow with the baseline length, but significant multipath at either station can pose a challenge.

Equation (7.47) is linear in the integers but nonlinear in the position coordinates. In order to estimate these parameters, we follow the same approach as in Section 6.1.1, i.e., to form an over-determined system of linear equations and solve it using the least-squares criterion. We take this up next.

7.5.1 Linear Model for Position Estimation

Let us start with a generic double–difference equation for carrier phase (7.25)

$$\phi^{(kl)} = \lambda^{-1} r^{(kl)} + N^{(kl)} + \varepsilon_\phi^{(kl)} \tag{7.48}$$

The position of the stationary reference station $\mathbf{x}_r$ is known, and we are interested in estimating $\mathbf{x}_{ur}$, the position of the user relative to the reference station.

$$\mathbf{x}_{ur} = \mathbf{x}_u - \mathbf{x}_r$$

where $\mathbf{x}_u$ is the user position. Let an initial estimate of the relative position vector be $\mathbf{x}_0$. (In practice, $\mathbf{x}_0 = \mathbf{0}$ would serve as a good starting point.) We can write

$$\mathbf{x}_{ur} = \mathbf{x}_0 + \delta\mathbf{x}$$

where $\delta\mathbf{x}$ is the unknown correction required to our initial guess. The benefit of this approach is that we can solve for $\delta\mathbf{x}$ from a linear system of equations as follows. From (7.19)

$$\begin{aligned}
r^{(kl)} &= -\left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right) \cdot \mathbf{x}_{ur} \\
&= -\left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right) \cdot \mathbf{x}_0 - \left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right) \cdot \delta\mathbf{x} \\
&= r_0^{(kl)} - \left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right) \cdot \delta\mathbf{x}
\end{aligned} \tag{7.49}$$

Substituting this expression for $r^{(kl)}$ in (7.48)

$$\phi^{(kl)} = \lambda^{-1} r_0^{(kl)} - \lambda^{-1}\left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right) \cdot \delta\mathbf{x} + N^{(kl)} + \varepsilon_\phi$$

Regrouping terms and defining $y^{(kl)} = \phi^{(kl)} - \lambda^{-1} r_0^{(kl)}$ and $\mathbf{g}^{(kl)} = -\lambda^{-1}\left(\mathbf{1}_r^{(k)} - \mathbf{1}_r^{(l)}\right)$,

$$y^{(kl)} = \mathbf{g}^{(kl)} \cdot \delta\mathbf{x} + N^{(kl)} + \varepsilon_\phi \tag{7.50}$$

is the linear equation derived from (7.48).

We can combine all such linear equations obtained from single- or dual-frequency mea-

surements at measurement epoch t_i into a generic vector-matrix representation

$$\mathbf{y}(i) = \mathbf{G}(i)\,\delta\mathbf{x} + \mathbf{N} + \boldsymbol{\varepsilon}_\phi(i) \tag{7.51}$$

where $\mathbf{y}(i)$ denotes the difference between the measured and computed carrier phase double differences for the initial position estimate; $\mathbf{G}(i)$ is the observation matrix characterizing the double-difference user-reference station-satellite geometry; $\delta\mathbf{x}$ is the error in the initial position estimate; and $\mathbf{N}$ is the vector of double-difference integer ambiguities to be estimated. If there are $\tilde{K}$ double-difference measurements, $\mathbf{G}(i)$ is a $(\tilde{K} \times 3)$ matrix and $\mathbf{N}$ is a $\tilde{K}$-vector. [$\tilde{K} = (K - 1)$ for single-frequency measurements, and $\tilde{K} = 2(K - 1)$ for dual-frequency measurements, where K is the number of satellites in view.]

Considering a simple case of a stationary user, the next measurement epoch brings additional $\tilde{K}$ equations

$$\mathbf{y}(i+1) = \mathbf{G}(i+1)\,\delta\mathbf{x} + \mathbf{N} + \boldsymbol{\varepsilon}_\phi(i+1) \tag{7.52}$$

We can combine (7.51) and (7.52) as

$$\begin{bmatrix} \mathbf{y}(i) \\ \mathbf{y}(i+1) \end{bmatrix} = \begin{bmatrix} \mathbf{G}(i) \\ \mathbf{G}(i+1) \end{bmatrix} \delta\mathbf{x} + \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} \mathbf{N} + \begin{bmatrix} \boldsymbol{\varepsilon}_\phi(i) \\ \boldsymbol{\varepsilon}_\phi(i+1) \end{bmatrix} \tag{7.53}$$

For a general case of measurements, perhaps from multiple epochs, we can write (7.53) as

$$\mathbf{y} = \mathbf{G}\,\delta\mathbf{x} + \mathbf{A}\mathbf{N} + \boldsymbol{\varepsilon} \tag{7.54}$$

Using the least-squares criterion, we look for a real-valued 3-vector $\delta\mathbf{x}$ and a $\tilde{K}$-vector of integers $\mathbf{N}$ which minimize the cost function

$$c(\delta\mathbf{x}, \mathbf{N}) = \| \mathbf{y} - \mathbf{G}\delta\mathbf{x} - \mathbf{A}\mathbf{N} \|^2 \tag{7.55}$$

The cost function is simply the sum of the lengths of residual vectors squared. In a kinematic case, $\delta\mathbf{x}$ would change from one measurement epoch to the next. There is no difficulty, however, in formulating the problem as above. Note also that if we have a basis for assigning different weights to the measurements, we can formulate (7.55) as a weighted least-squares problem. We'll revisit this issue below.

Minimization of (7.55) would be straightforward were it not for the constraint that each element of $\mathbf{N}$ be an integer. We have an *integer least-squares* problem [Hassibi and Boyd (1998)]. We could disregard the constraint to make the problem routine and, in fact, this approach is commonly used. We discuss this approach in Section 7.5.2. An alternative is to limit the estimates to a set of integers, and then 'search' this set for the 'best' solution. In principle, we can obtain the solution with measurements from a single epoch. We discuss this approach below in Section 7.5.3.

Given the estimates of the integers, it would be straightforward to obtain precise position estimates. In order to ensure that the position estimates are precise *and* accurate, we have to ensure that the integer ambiguities are estimated correctly. This raises an important issue: How to tell if the integer estimates are correct? The step of ascertaining the correctness of the integer estimates is called *validation*. Validation of the estimates in real time is a hard problem which has received less attention than the integer estimation problem [Joosten and Tiberius (2000)]. We'll discuss it further below in the context of an example.

7.5.2 Float Solution

Given a time series of carrier phase measurements, we can develop an algorithm to process the measurements in batch or sequential mode to estimate the ambiguities as though they were real-valued (or, floating-point) numbers. Such estimates are called *float solutions*. Rounding the float solutions gives the fixed solutions or integer estimates. If the float solutions appear to be approaching integer values, it would offer an indication that the estimates are good. Actually, convergence of the float estimates to integers tends to be slow because measurements taken within several seconds of each other tend to be highly correlated, and rounding off float estimates to obtain integers can be an error-prone process.

Unconstrained solution of (7.55) offers no basic challenge and there is no dearth of integer ambiguity algorithms based on float solutions [Hwang (1991), Hatch and Euler (1994), Han and Rizos (1997)]. Proprietary variations of the basic technique have been incorporated in a number of commercial RTK packages. These packages generally work well in surveying and industrial/agricultural applications where it is acceptable to wait for a few seconds on occasion for the results. Occasional mistakes (i.e., wrong integers) caught within a few seconds also have to be acceptable. Neither the integer ambiguity resolution nor validation of the integers is a problem if the measurements are given over a long stretch without any break in signal tracking. The challenge is to make both steps work reliably in a navigation application involving a fast-moving vehicle. Even with dual-frequency (L1-L2) measurements, the problem of instantaneous (or, single-epoch) ambiguity resolution and validation of the results remains a challenge.

7.5.3 Search Techniques

A search for the best solution essentially consists of the following steps: Define the volume to be searched; set up a grid within this volume; define a cost function; and evaluate the cost function at each grid point. The solution corresponds to the grid point with the lowest value of the cost function.

The earliest work on integer ambiguity resolution is that of Counselman and Gourevitch (1981). Their search scheme, called *Ambiguity Function Method* (AFM), was designed to work in the presence of cycle slips, the bane of early receivers. The basic idea behind AFM is to search a volume, say a cube, in which the position is known to lie. The cube could be centered at a position estimate obtained from double-difference pseudorange measurements. We set up a grid in this cube and step across all candidates to verify conformance with (7.47). In the absence of errors, at the true position the difference between the double-differences of the computed ranges and the carrier phase measurements should be an integer for each pair of satellites. The search volume has to be large enough to contain the true location and small enough to be searched in a reasonable amount of time.

An alternative to searching in the 3-D position space as in AFM is to set up the search in the space of integers. Again, we can define a window to be searched on the basis of the code measurements, which give an initial position estimate. The corresponding estimates of the integers are obtained from (7.47). An estimate of the error bound on the initial position estimate identifies the range of uncertainty in each of the integers. Now we have a grid of integer values to be searched. The basic idea is illustrated in the idealized 2-D example shown in Figure 7.2. We have drawn a circle centered at the code-based position estimate that we believe encloses

the true position. With three satellites and an uncertainty of ±2 wavelengths in each range, there are 4^3 combinations of the three integers to be checked as potential solutions. For a cost function, we could define a measure of the extent to which the wave fronts do not intersect at a point, as shown by shaded areas. Such a search would take us to the right answer.

The search space can easily become too large for real-time solution. Suppose that the error in a code-based position estimate is known to be less than 1 m. With a wavelength of about 20 cm, it follows that the uncertainty in each integer can be ±5 cycles. If ten integers are to be estimated, we have to evaluate a cost function for more than 10^{10} integer vectors. This set of candidate integer vectors is simply too large for real time search. For the wide-lane measurements with a wavelength of 0.86 m, the uncertainty in the above example is reduced to three cycles per integer, or evaluation of a cost function for 3^{10} integer vectors. Below we look at an approach which attempts to reduce the size of the search space. As before, an asterisk (*) marks optional sections.

Constraints on the Integers*

Let us consider a simple case: single-epoch carrier phase measurements at L1 from K satellites. From (7.51), the set of linear equations for position estimation can be written as

$$\mathbf{y} = \mathbf{G}\,\delta\mathbf{x} + \mathbf{N} + \boldsymbol{\varepsilon} \tag{7.56}$$

Our plan is to determine the 'best' $\mathbf{N}$ by searching over a grid of integer vectors. The steps are: (i) select an $\mathbf{N}$; (ii) compute $\delta\hat{\mathbf{x}}$ from (7.56) as the least squares solution; (iii) compute cost function $C(\delta\hat{\mathbf{x}}, \mathbf{N})$ from (7.55). The optimal $\mathbf{N}$ would minimize the cost function.

We have $(K-1)$ double-difference equations, and it appears that we have $(K+2)$ unknowns: three position coordinates and $(K-1)$ integer ambiguities. If it were not for the constraint that the ambiguities be integers, we would have an under-determined system of equations with an infinite set of solutions. In fact, as was pointed out by Hatch (1990), the number of independent unknowns in (7.56) or (7.53) is actually three. Given the values of any three of the ambiguities, the user position can be established with precision, and the remaining ambiguities resolved. Alternately, given the 3-D position, all $(K-1)$ integer ambiguities are resolved automatically. $(K-4)$ of the equations are redundant. These remarks are to be understood in the context of search for the integers: Not all combinations of integers are admissible, and we don't have to try out each grid point in a rectangular box. The trick is to find a way to identify the admissible integer vectors.

Actually, we have known all along that measurements from four satellites allow us to estimate the position and, given K (> 4) satellites, $(K-4)$ are redundant. The $(K-1)$ integer ambiguities are, therefore, constrained to three degrees of freedom. It is not immediately clear how to leverage this insight into solving (7.56) for the integer ambiguities while taking full advantage of the redundant measurements. The argument of the paragraph above remains valid if we had dual-frequency measurements, perhaps from multiple measurement epochs.

An insightful approach to exploit the constraints on the integer ambiguities was suggested by Hatch (1990). The basic steps are as follows.

- Divide the satellites in view into two groups: a primary group of four satellites and a secondary group of $(K-4)$ satellites.
- Given an initial position estimate and its associated uncertainty, generate a set of candidate integer values associated with the primary satellites, and generate the set of corresponding position estimates.
- Check each position estimate against measurements from the secondary satellites. With the 'correct' position estimate, the difference between the measured and computed carrier phase measurements would be close to an integer for each satellite pair.

The difficulty is that by estimating the position from a subset of four satellites, this approach fails to take full advantage of the redundant measurements for position estimation. We'll return to this point later in this section.

We review below two interesting methods using the search approach: *Least-squares Ambiguity Decorrelation Adjustment* (LAMBDA) method of Teunissen [Teunissen (1996), Teunissen *et al.* (1997)], and *Local Minima Search* (LMS) algorithm of Pratt [Pratt *et al.* (1997)]. LAMBDA came before LMS, but it is easier to introduce LMS first. Both algorithms search for the integer ambiguities among a set of candidates. This set, however, is greatly reduced in size in both cases using different approaches. These algorithms require some facility with concepts of linear algebra. A student unsure of his grasp of these concepts would do well to review the relevant chapters in Strang and Borre (1997). We discuss the basic concepts only. The details of implementation are found in the references.

Local Minima Search (LMS) Algorithm*

We start by rewriting the set of linear equations (7.56) for measurements from a single epoch. The unknowns are: a correction to an initial position estimate (3-vector) and a vector of integer ambiguities.

$$\mathbf{y} = \mathbf{G}\,\delta\mathbf{x} + \mathbf{N} + \boldsymbol{\varepsilon} \tag{7.57}$$

With K satellites in view, this set consists of $(K-1)$ equations corresponding to the measurements at L1 and $(K-1)$ more for measurements at L2. Actually, given the dual-frequency measurements, we would prefer to form wide-lane measurements, and write $(K-1)$ equations for the double-difference wide-lane measurements. The size of this set of equations is not important to our discussion below. The efficacy of the method, however, would depend upon K and whether dual-frequency measurements are available. As before, suppose we have $\tilde{K}$ double-difference measurements (i.e., $\mathbf{G}$ is a $\tilde{K} \times 3$ matrix and $\mathbf{N}$ is a $\tilde{K}$-vector).

We attempt to solve (7.57) in three steps as follows: (i) obtain a least-square estimate of $\delta\mathbf{x}$ in terms of $\mathbf{y}$ and $\mathbf{N}$; (ii) substitute this estimate back in (7.57); and (iii) search for $\mathbf{N}$ (over a predefined grid) to minimize the resultant residuals. First, the least-squares estimate of $\delta\mathbf{x}$

$$\delta\hat{\mathbf{x}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T(\mathbf{y} - \mathbf{N}) \tag{7.58}$$

Plugging this estimate back in (7.57) gives the vector of residuals

$$(\mathbf{y} - \mathbf{N}) - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T(\mathbf{y} - \mathbf{N}) \tag{7.59}$$

The cost function to be minimized is the sum of the squares of the residuals, or the squared magnitude of the above vector. To simplify the notation, define

$$\mathbf{P} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$$

P is a projection matrix. It projects a vector onto the range space of **G** (i.e., space spanned by the columns of **G**). Like any projection matrix, **P** is symmetric and idempotent (i.e., $\mathbf{P}^T = \mathbf{P}$ and $\mathbf{PP} = \mathbf{P}$). $(\mathbf{I} - \mathbf{P})$, which we denote as **Q**, is also a projection matrix. It projects a vector onto the subspace orthogonal to the range space of **G**. The reason for introducing **P** and **Q** would become clear when we write the vector of residuals (7.59) as $\mathbf{Q}(\mathbf{y} - \mathbf{N})$ and the cost function based on the residuals as

$$c(\mathbf{N}) = (\mathbf{y} - \mathbf{N})^T \mathbf{Q} (\mathbf{y} - \mathbf{N}) \tag{7.60}$$

The final step is to select an integer vector **N** so as to minimize this cost function or, equivalently, solve $\mathbf{QN} = \mathbf{Qy}$ in the least-squares sense.

In principle, we can perform the minimization of (7.60) by an exhaustive search over the candidate integer vectors **N**. There are two problems. First, as noted previously, the set of integers to be searched is generally too large. The second problem is that (7.60) is lumpy function with many local minima. Suppose that there is a local minimum at **N**. By definition, the cost function is higher for all neighboring vectors generated by changing a single element of **N** by one. The neighboring vectors of **N** can be represented as

$$\mathbf{N} + (0, \ldots, 0, 1, 0, \ldots, 0)^T \text{ or } \mathbf{N} + (0, \ldots, 0, -1, 0, \ldots, 0)^T.$$

Evaluating (7.60) at these vectors, we obtain a set of inequalities as necessary conditions that the cost function has a local minimum at **N**. The inequalities are represented compactly as

$$\mathbf{Qy} - \tfrac{1}{2}\mathbf{d} < \mathbf{QN} < \mathbf{Qy} + \tfrac{1}{2}\mathbf{d} \tag{7.61}$$

where elements of vector **d** are the terms on the main diagonal of **Q** (i.e., $d_i = Q_{ii}$). The inequalities are to be interpreted element by element. For projection matrix **Q**, $0 \le Q_{ii} \le 1$.

Now we are in a position to illustrate the lumpy nature of the cost function. If (7.60) has a local minimum at **N**, and **M** is in the null space of **Q**, or nearly so (i.e., $\mathbf{QM} \approx \mathbf{0}$), then $\mathbf{N} + \mathbf{M}$ also satisfies (7.61) and, therefore, is also a local minimum. If we can find an integer representation for the null space of **Q**, each point when added to a solution of $\mathbf{QN} = \mathbf{Qy}$ would also give a local minimum of (7.60).

The null space of **Q** is three-dimensional. To see this, note that matrix **G** is $(\tilde{K} \times 3)$ and has rank three. Therefore, the range space of **G** is three-dimensional, and the subspace orthogonal to it is $(\tilde{K} - 3)$-dimensional. Since **Q** is a projection onto the subspace orthogonal to the range space of **G**, the range space of **Q** is $(\tilde{K} - 3)$-dimensional. Therefore, the null space of **Q** is three-dimensional.

Now we are beginning to approach a mathematical representation of the insight of Hatch (1990) that the integer ambiguities are constrained to three degrees of freedom. The mathematical statement is that the candidate integer vectors all lie on a 'thick' three-dimensional plane regardless of the dimensionality of the integer space being searched.

The outline of an approach is now clear. We don't have to evaluate the cost function for each candidate **N** from a rectangular grid. The search can be limited to integer vectors which correspond to the local minima of (7.60). The search proceeds from a local minimum to another local minimum, hence the name of the algorithm.

To solve $\mathbf{QN} = \mathbf{Qy}$, we perform an LU-decomposition of **Q** with partial pivoting to get

$$\mathbf{UN} = \mathbf{L}^{-1}\mathbf{Qy} \tag{7.62}$$

Since the rank of **Q** is $(\tilde{K} - 3)$, the lower rightmost 3×3 block of **U** will be all zeroes, and the last three elements of **N** are unconstrained by the above equation. Again, due to the rank deficiency, the forward elimination (i.e., formation of $\mathbf{L}^{-1}$) stops with three rows remaining. The LMS algorithm searches the last three elements of **N** exhaustively.

The search is thus effectively limited to an integer grid inside a three-dimensional rectangular box, regardless of the size of the integer vector. The remaining elements of **N** are determined by back substitution using the above triangular system. At each step in the back substitution, the real-valued estimate is replaced by its nearest integer. If an integer vector so constructed fails to satisfy the necessary condition for a local minimum (7.61), it is dropped from further consideration.

Pratt *et al.* (1997) offer an insightful presentation of their empirical results using measurements over extended periods. Some of these are presented in Figures 7.9 and 7.10 for 2-km and 9-km baselines, respectively. Both figures present analysis results for measurements taken at five-second intervals over a one-hour period. Both antennas were stationary, and the positions of their phase centers had been surveyed carefully. The objective of the experiment was to test the LMS algorithm with single-epoch measurements. Each of the four panels presents a different piece of information, discussed below.

It is easiest to start with Figure 7.9, panel (d) at the bottom, which shows the 3-D position error. In most cases, the position error is at centimeter level, as expected, when the integers are determined correctly. In four cases, however, the position error is considerably larger. Obviously, in these cases the integer estimates were wrong. This brings up the validation problem, mentioned earlier. Panels (b) and (c) give plots of two statistics which could be used for validation of the integer estimates. Panel (b) is a plot of the rms residuals associated with (7.57) when the estimated values of $\delta\hat{\mathbf{x}}$ and $\hat{\mathbf{N}}$ are plugged in. With the right answers, we expect the residuals to be low. Indeed, the four cases associated with high position errors correspond to relatively large values of the rms residuals. The rms residuals, therefore, could have served as a validation statistic for these estimates. Panel (c) is a plot of the search ratio, defined as the ratio of the cost function at the global minimum and its value at the next deepest minimum. The rationale is that if the global minimum found is much lower than the value for any other candidate in the search region, it is more likely to be associated with the right answer. If two or more candidates give roughly similar values of the cost function, we may doubt the answer. Indeed, the search ratio is close to one for the four cases in question.

Finally, for some of the double-difference measurements, Panel (a) shows the measurement residuals computed using the surveyed positions of the two antennas. The residuals show oscillations which appear consistent in size and frequency with multipath error, and the search failures are associated with the multipath peaks. This plot shows the difficulty of the problem: Error of a quarter-cycle in one or more measurements can make it hard to estimate the integer ambiguities with measurements from a single epoch.

From Figure 7.9, Panels (b) and (c), it appears that it would be easy to detect a failed search for the integers. Unfortunately, that is not the case, as seen in Figure 7.10, which presents results from the 9-km baseline in four panels as in the previous case. With a longer baseline, we expect additional modeling errors due to the ionospheric and tropospheric differences at the two sites. From Panels (b) and (c), it is no longer clear that any combination of the rms residuals and search ratio would yield an effective statistic to detect search failures. Setting a threshold low on these statistics would raise false alarms to intolerable levels. And the

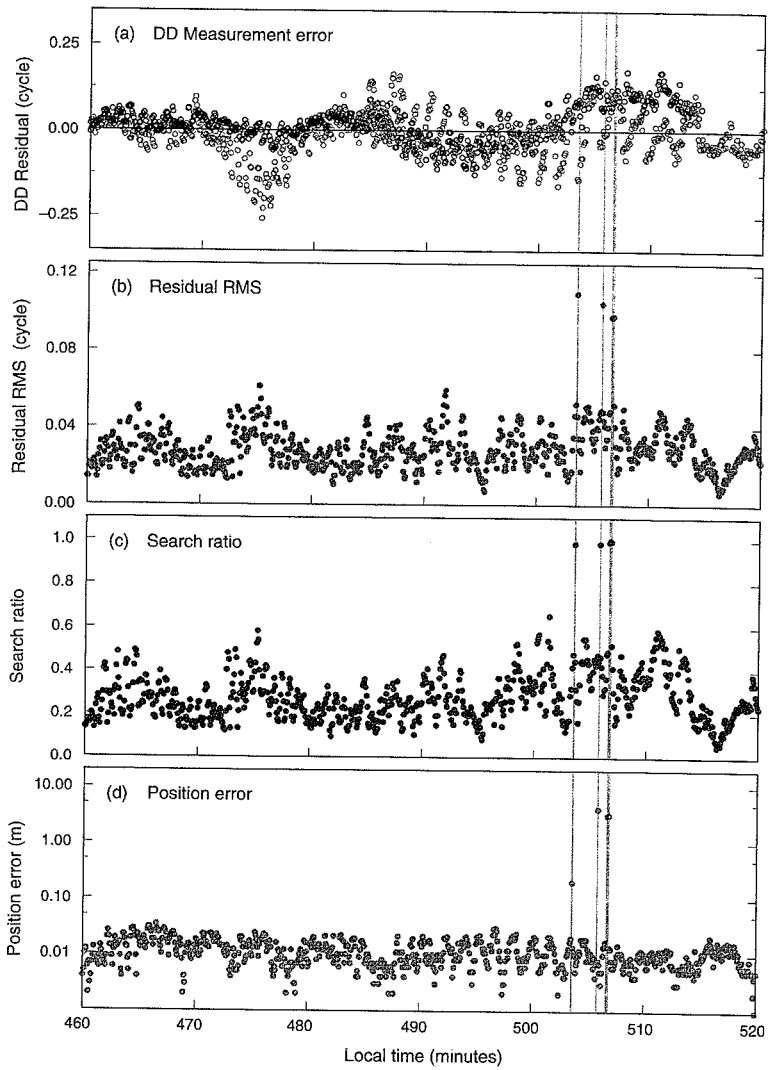


Figure 7.9 Single-epoch integer ambiguity resolution with LMS algorithm for a 2-kilometer baseline.

measurement residuals in Panel (a) are downright depressing. Apparently, the ionospheric and tropospheric differences are significant, and there is multipath.

To repeat our *mantra*, the key to success in integer ambiguity resolution lies in mitigation of the measurement and modeling errors. Multipath has to be kept in check by antenna design and placement, and receiver design. Modeling errors may be mitigated by fielding multiple

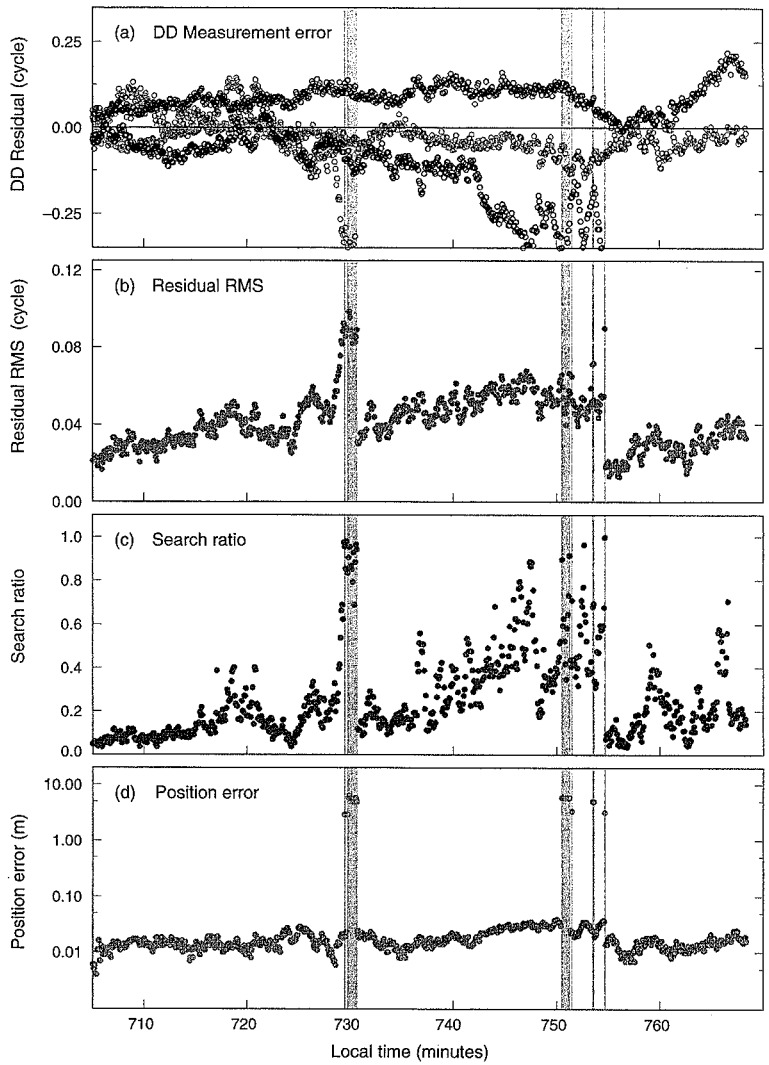


Figure 7.10 Single-epoch integer ambiguity resolution for a 9-kilometer baseline with LMS algorithm.

reference stations over an area to assess the gradients in the measurement residuals due to the ephemeris error and ionospheric and tropospheric differences, and to correct the measurements at the test site for these effects [Raquet and Lachapelle (2000)].

The final method for integer ambiguity resolution in this chapter is the *Least-Squares Ambiguity Decorrelation Adjustment* (LAMBDA) method of Teunissen (1996). This algorithm

reduces the search space for the integers by exploiting the correlation structure of the float solutions. We have so far sidestepped any discussion of correlations among the measurements and among the float estimates of the integers. These issues are addressed below.

Correlations among the Double-Difference Measurements*

We have assumed the measurements (both code and carrier) from the satellites in view to be uncorrelated. In particular, the covariance of the carrier phase measurements at an instant is modeled as

$$\boldsymbol{\Sigma}_{\phi} = \sigma_{\phi}^2 \mathbf{I} \tag{7.63}$$

where, from our error model, $\sigma_{\phi} = 0.025$ cycle, and $\mathbf{I}$ is an identity matrix. As discussed in Section 6.1.2, we make this simplifying assumption in the absence of a simple, truer model [Tiberius *et al.* (1999)]. In this section, we examine the correlations among the single and double differences.

The single differences corresponding to a pair of satellites can be written in matrix notation as

$$\begin{bmatrix} \phi_{ur}^{(k)} \\ \phi_{ur}^{(l)} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \phi_{u}^{(k)} \\ \phi_{r}^{(k)} \\ \phi_{u}^{(l)} \\ \phi_{r}^{(k)} \end{bmatrix}$$

The covariance matrix of this pair of single differences is

$$\begin{aligned} \boldsymbol{\Sigma}_{sd} &= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{\phi}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\phi}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\phi}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\phi}^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \\ &= 2\sigma_{\phi}^2 \mathbf{I} \end{aligned} \tag{7.64}$$

We, therefore, conclude that if the measurements are uncorrelated, so are their single differences. The common variance of the single differences is twice that for the carrier phase measurements, as we had noted previously.

Now, on to the double differences. Taking satellite l as the reference, we can write a pair of double differences corresponding to satellites k, l, and m as

$$\begin{bmatrix} \phi_{ur}^{(kl)} \\ \phi_{ur}^{(ml)} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \phi_{ur}^{(k)} \\ \phi_{ur}^{(l)} \\ \phi_{ur}^{(m)} \end{bmatrix}$$

The covariance matrix of this pair of double differences is

$$\boldsymbol{\Sigma}_{dd} = 2\sigma_{\phi}^2 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{7.65}$$

The double differences are correlated even if the original measurements are not. For K satellites, the $(K-1) \times (K-1)$ covariance matrix would have 2's on the main diagonal and 1's in all off-diagonal positions. The triple differences are also correlated. It is left as an exercise to show that in our three-satellite problem above the covariance matrix of the triple differences is

$$\boldsymbol{\Sigma}_{td} = 4\sigma_{\phi}^2 \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \tag{7.66}$$

The reason for this detour should now be clear. The cost functions (7.55) and (7.60) assumed implicitly that the double differences were uncorrelated. Now we know better, and exploit this additional structure, as discussed below.

LAMBDA Method*

We reformulate the problem as one of estimating $\delta\mathbf{x}$, a 3-vector of real numbers, and $\mathbf{N}$, a $\tilde{K}$-vector of integers, which are solutions of

$$\mathbf{y} = \mathbf{G}\,\delta\mathbf{x} + \mathbf{A}\mathbf{N} + \boldsymbol{\varepsilon} \tag{7.67}$$

given that the covariance of $\boldsymbol{\varepsilon}$ is Σ_{dd} from (7.65). In other words, find $\delta\mathbf{x}$ and $\mathbf{N}$ which minimize a revised cost function

$$\begin{aligned} C_{\mathbf{W}}(\delta\mathbf{x}, \mathbf{N}) &= \|\mathbf{y} - \mathbf{G}\delta\mathbf{x} - \mathbf{A}\mathbf{N}\|_{\mathbf{W}}^2 \\ &= (\mathbf{y} - \mathbf{G}\,\delta\mathbf{x} - \mathbf{A}\mathbf{N})^T \mathbf{W} (\mathbf{y} - \mathbf{G}\,\delta\mathbf{x} - \mathbf{A}\mathbf{N}) \end{aligned} \tag{7.68}$$

which uses the inverse of the noise covariance matrix $\mathbf{W} = \boldsymbol{\Sigma}_{dd}^{-1}$ to give different weights to the contributions of the residuals. The algorithm implementation comprises three steps.

Step 1. Obtain float solutions: Disregard the constraint on the ambiguities and obtain solutions for $\delta\mathbf{x}$ and $\mathbf{N}$ which minimize the cost function (7.68). We have discussed the float solutions before. One difference here is that we now use the weighted least-squares criterion to account for the correlations among the double differences. A more important difference is that in the earlier discussion the integer ambiguities were obtained by rounding off the float solutions to the nearest integers. In LAMBDA this is only the first step, which ends with the float solution for the position and ambiguities and their covariance matrix written in block form below.

$$\begin{aligned} \text{Float solution} &= \begin{bmatrix} \delta\hat{\mathbf{x}} \\ \hat{\mathbf{N}} \end{bmatrix} \\ \text{Cov}\begin{bmatrix} \delta\hat{\mathbf{x}} \\ \hat{\mathbf{N}} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\delta\hat{\mathbf{x}}} & \boldsymbol{\Sigma}_{\delta\hat{\mathbf{x}},\hat{\mathbf{N}}} \\ \boldsymbol{\Sigma}_{\delta\hat{\mathbf{x}},\hat{\mathbf{N}}} & \boldsymbol{\Sigma}_{\hat{\mathbf{N}}} \end{bmatrix} \end{aligned} \tag{7.69}$$

where $\delta\hat{\mathbf{x}}$ and $\hat{\mathbf{N}}$ are the float solutions; $\boldsymbol{\Sigma}_{\delta\hat{\mathbf{x}}}$ and $\boldsymbol{\Sigma}_{\hat{\mathbf{N}}}$ are the corresponding covariance matrices; $\boldsymbol{\Sigma}_{\delta\hat{\mathbf{x}},\hat{\mathbf{N}}}$ gives the cross-correlations between the two.

Step 2. Find the integer vector $\mathbf{N}$ which minimizes the cost function

$$c(\mathbf{N}) = \left(\mathbf{N} - \hat{\mathbf{N}}\right)^T \mathbf{W}_{\mathbf{N}} \left(\mathbf{N} - \hat{\mathbf{N}}\right) \tag{7.70}$$

where $\hat{\mathbf{N}}$ is the float solution from Step 1 and the weight matrix $\mathbf{W}_{\mathbf{N}}$ is the inverse of its covariance matrix: $\mathbf{W}_{\mathbf{N}} = \boldsymbol{\Sigma}_{\hat{\mathbf{N}}}^{-1}$.

Step 2 is at the heart of LAMBDA method. The measure of distance (or, 'nearness') of an integer vector to $\hat{\mathbf{N}}$ is given by (7.70). The contour of points with a constant value of the cost function is an ellipse in two dimensions and an ellipsoid in higher dimensions, centered at $\hat{\mathbf{N}}$. The search space is delimited by selecting the size of the ellipsoid to be searched via a parameter value $c > 0$. The inequality

$$\left(\mathbf{N} - \hat{\mathbf{N}}\right)^T \mathbf{W}_\mathbf{N} \left(\mathbf{N} - \hat{\mathbf{N}}\right) \leq c \tag{7.71}$$

defines the integer vectors $\mathbf{N}$ which are candidates for the solution. The search space consists of the integer grid points inside an ellipsoid. Clearly, this search space must be large enough to contain the right answer and small enough to be searched quickly. (In the earlier discussions, with the exception of LMS, the search space was a grid inside a rectangular box.)

In practice, the constant-cost ellipsoids can be very elongated, longer by orders of magnitude in one direction than in another. This is specially the case when the measurements are limited to a single epoch or only a few epochs. The result is that points which appear much farther away from $\hat{\mathbf{N}}$ may have lower values of the cost function than those which appear nearby. Brute-force search, therefore, would be inefficient. What's needed is a change of variable which would turn the elongated ellipsoid into a sphere so that the search can be limited to the neighbors of $\hat{\mathbf{N}}$.

If the weight matrix $\mathbf{W}_\mathbf{N}$ in (7.70) is diagonal, the minimization of the cost function is trivial. The best estimate of an integer ambiguity is the corresponding float estimate rounded off to the nearest integer. A diagonal $\mathbf{W}_\mathbf{N}$ would mean that the integer ambiguity estimates in the float solution are all uncorrelated. In general, $\mathbf{W}_\mathbf{N}$ would not be diagonal, and the objective of Step 2 is to introduce a change of variables so that the resultant correlation matrix is diagonal.

$\mathbf{W}_\mathbf{N}$ is a positive semi-definite matrix and it would appear that diagonalizing it would not be a problem. We can use the matrix of its eigenvectors to transform the variables. Actually, this approach will not work here because the transformation will not preserve the integer nature of the ambiguities. We have to restrict the transformations to those that take integers into integers. Actually, the inverse transformation must do the same, too, so that we can find the solution of the original problem. The required transformation $\mathbf{Z}$ must satisfy the following conditions.

- $\mathbf{Z}$ must have integer entries,
- $\mathbf{Z}$ must be invertible,
- $\mathbf{Z}^{-1}$ must have integer entries.

These conditions ensure that there is a one-to-one relationship between integers in the original and transformed spaces. It is left as an exercise to show that it follows from the above conditions that $\mathbf{Z}$ and $\mathbf{Z}^{-1}$ are volume-preserving transformations, i.e., $|\det(\mathbf{Z})| = 1$.

Consider a hypothetical transformation $\mathbf{Z}$ in this restricted class of transformations which diagonalizes $\mathbf{W}_\mathbf{N}$. Let

$$\mathbf{M} = \mathbf{Z}\mathbf{N} \quad \text{and} \quad \hat{\mathbf{M}} = \mathbf{Z}\hat{\mathbf{N}}$$

The cost function in the transformed space is

$$\left(\mathbf{M} - \hat{\mathbf{M}}\right)^T \left(\mathbf{Z}^{-T}\mathbf{W}_\mathbf{N}\mathbf{Z}^{-1}\right)\left(\mathbf{M} - \hat{\mathbf{M}}\right) \tag{7.72}$$

Since $\mathbf{Z}^{-T}\mathbf{W}_\mathbf{N}\mathbf{Z}^{-1}$ is diagonal, we find the solution for $\mathbf{M}$ right away by rounding off each element of $\hat{\mathbf{M}}$. We now transform the problem back and find $\mathbf{N}$ from

$$\mathbf{N} = \mathbf{Z}^{-1}\mathbf{M} \tag{7.73}$$

LAMBDA would be a simple algorithm if $\mathbf{W}_\mathbf{N}$ could be diagonalized using our restricted class of transformations. Unfortunately, that's almost never the case and the integer ambiguities are not decorrelated fully. LAMBDA involves many subtle steps to transform $\mathbf{W}_\mathbf{N}$ into a matrix that is as nearly diagonal as possible [Teunissen (1996), Teunissen *et al.* (1997)].

Step 3. Obtain 'fixed' solution $\delta\mathbf{x}$ from (7.67) after fixing the integer ambiguities to $\mathbf{N}$ found in Step 2.

A simple, 2-D example from Joosten and Tiberius (2000) is reproduced in Figure 7.11.

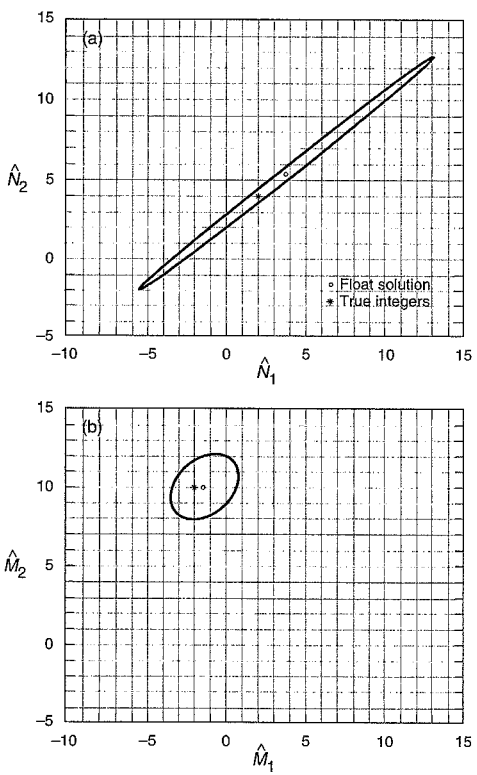


Figure 7.11 Search ellipses for integer ambiguities in a simple, 2-D problem to illustrate the LAMBDA method. (a) The float estimates of the ambiguities are highly correlated, and simple rounding leads to incorrect integers. (b) In the transformed space the ambiguity estimates are nearly uncorrelated, and simple rounding gives true integers. [Adapted from Joosten and Tiberius (2000)]

The float solution is $\hat{\mathbf{N}} = (3.875, 5.400)^T$, and the ambiguities are highly correlated as indicated by the elongated shape of the search ellipse. A simple rounding would give a wrong answer because the true integers are known to be $\mathbf{N} = (2, 4)^T$. Now introduce transformation

$$\mathbf{Z} = \begin{bmatrix} 1 & -1 \\ -3 & 4 \end{bmatrix}$$

The float solution transforms to $\hat{\mathbf{M}} = (-1.525, 9.975)^T$. The ambiguities are much less correlated in the transformed space as seen in the rounded shape of the search ellipse. The benefit is that simple rounding of $\hat{\mathbf{M}}$ gives the right answer for the integers in the transformed space: $\mathbf{M} = (-2, 10)^T$. Inverse transformation takes $\mathbf{M}$ into $\mathbf{N}$, the right answer to the original problem.

7.6 Precise Point Positioning

A main point of this and the previous chapter has been that in order to obtain centimeter-level position estimates the errors in the measurements have to be reduced to centimeter level. In this chapter, we found that 'short' baselines could be estimated with centimeter-level accuracy from carrier phase measurements through double differences, which essentially cancelled the common errors at the two receivers. We change gears now and switch from relative positioning to point positioning (or, absolute positioning).

We devoted most of Chapter 6 to the subject of point positioning. The difference is that we were content with position estimates with meter-level errors, but wanted the results in real time. The pseudorange measurements met our requirements and, as shown in Table 5.4, DGPS could cut the pseudorange errors from several meters to one meter, or less, with the corresponding improvement in the position accuracy.

For centimeter-level positioning with measurements from a single receiver, we need carrier phase measurements, and must find ways to correct for the errors which cancel out effortlessly in double differencing. This approach is referred to as precise point positioning (PPP). The benefit of PPP is freedom from simultaneous measurements at two sites and limitation on the length of the baseline. We follow a lucid account in Kouba and Héroux (2001) and Héroux *et al.* (2004).

Centimeter-level point positioning introduces some unusual complications. The coordinates of a 'fixed' point in the earth-centered, earth-fixed coordinate frame are in fact changing all the time, if we look closely. The 'solid' earth goes through elastic deformations due to the same gravitational forces that cause ocean tides. This phenomenon, referred to as solid earth tides, results in displacements consisting of a latitude-dependent permanent displacement (at 10-cm level at mid-latitudes) and a periodic part with semidiurnal and diurnal periods of changing amplitudes (at centimeter level). Ocean loading (or, loading due to the ocean tides) also produces displacements with diurnal and semi-diurnal components. Such displacements, however, are an order of magnitude smaller than those due to solid earth tides, but can be at centimeter-level in coastal regions. (The displacements due to atmospheric loading and snow buildup are generally less than one centimeter.)

How do we then assign precise coordinates to a point? The answer is that procedures for this are established by general agreement. The keepers of International Terrestrial Reference Frame (ITRF) have adopted a convention for the geodesists to follow. We wouldn't pursue

this further except to note that for differential positioning over short baselines (<100 km), both stations would have nearly identical displacements, and the relative position can be measured with centimeter-level accuracy without concern for the solid earth tides or ocean loading.

7.6.1 Measurement Models

Let us start with the basic models for the code and carrier phase measurements in our familiar notation without identifying the satellite or carrier frequency.

$$\rho = r + I + T + c(\delta t_u - \delta t^s) + \varepsilon_\rho \tag{7.74a}$$

$$\phi = \lambda^{-1}[r - I + T] + \frac{c}{\lambda}(\delta t_u - \delta t^s) + N + \varepsilon_\phi \tag{7.74b}$$

We have already listed the sources of error in these measurements in Table 5.4. We now expand this list and examine the contribution of each source in the context of PPP for estimation in 'near real time,' meaning precise point positioning in kinematic mode after a certain initialization period.

- **Satellite ephemeris and clock errors.** The error in the ephemeris and clock parameters broadcast by the GPS satellites is about 2–3 m (rms). This error is reduced by two orders of magnitude by post-processing measurements from a wide-area or global network of receivers. The highest quality orbit products (ephemeris error < 5 cm, clock error < 0.1 ns) are available now from the International GNSS Service (IGS) with a latency of one to two weeks [http://igscb.jpl.nasa.gov]. The predicted ephemeris and clock data available from IGS in real time are impressive as well: ephemeris error < 10 cm, clock error < 5 ns! If it weren't for these precise estimates, we wouldn't be talking about PPP.
- **Ionospheric delay.** It's best to eliminate it with dual-frequency measurements. The residual error would be at centimeter level [Section 5.3.2].
- **Tropospheric delay.** In a change from our previous approach, we now aim to estimate the tropospheric delay T, and replace it with $T_z \cdot m$, where T_z is the zenith delay and m is the elevation-dependent mapping function [Section 5.3.3]. We now have an additional parameter to estimate.
- **Multipath and receiver noise.** No relief here, but at least we don't have to deal with the multipath issues at the reference station.

We now expand the above list to include centimeter-level errors which would have cancelled in differencing for relative positioning. In PPP, they have to be accounted for explicitly.

- **Phase wind-up correction.** The satellites transmit right-hand circularly polarized radio waves and, therefore, the observed carrier phase depends upon the relative orientation of the satellite and receiver antennas. A rotation of either antenna around its bore (vertical) axis will change the carrier phase by up to one cycle. The satellite antennas undergo rotations as their solar panels are being oriented toward the sun, mostly slowly, but sometimes completing a rotation in less than thirty minutes.

- **Satellite antenna offsets.** The force models used for satellite orbit modeling refer to its center of mass. The IGS precise satellite coordinates and clock biases refer to the satellite center of mass. The pseudorange measurements, however, refer to satellite antenna phase center. If not accounted for, this would introduce decimeter-level errors. What's needed is the phase center offset for each satellite and the orientation of the offset vector in space as the satellite orbits the earth.

We form the ionosphere-free (IF) combination of the code and carrier phase measurements from a dual-frequency receiver. From (5.31), the ionosphere-free pseudorange measurement ρ_{IF} is

$$\rho_{IF} = \frac{f_{L1}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\rho_{L1} - \frac{f_{L2}^2}{\left(f_{L1}^2 - f_{L2}^2\right)}\rho_{L2}$$
$$= 2.546\rho_{L1} - 1.546\rho_{L2}$$

which we model as

$$\rho_{IF} = r + c\cdot\delta t_u + T_z\cdot m(el) + \varepsilon_\rho \tag{7.75a}$$

Similarly, we write the ionosphere-free combination of the carrier phase measurements as

$$\Phi_{IF} = r + c\cdot\delta t_u + T_z\cdot m(el) + \lambda_{IF}\cdot N_{IF} + \varepsilon_\Phi \tag{7.75b}$$

where N_{IF} is a combination of the integer ambiguities at L1 and L2, and no longer an integer, and λ_{IF} is the carrier wavelength corresponding to this IF combination.

We now have four types of parameters to estimate in (7.75): station position (x, y, z), receiver clock bias (δt_u), troposphere zenith path delay T_z, and carrier phase ambiguity for each satellite in view (N_{IF}). The problem would be easier if the position is fixed (static mode) than if it's changing (kinematic mode). The receiver clock bias would change from epoch to epoch. The zenith path delay would change slowly, on the order of a few centimeters/hour. The carrier phase ambiguities will remain constant as long as the carrier tracking is continuous.

Equation (7.75) is linear in δt_u, T_z, and N_{IF}, but nonlinear in (x, y, z). The approach to solve for the unknown parameters is the same as in Section 6.1.1: Linearize (7.75) about initial estimates, solve the linear simultaneous equations to determine corrections to these estimates, linearize (7.75) about the new estimates, and iterate. The convergence time would depend upon the number of satellites in view and their geometry, user dynamics, and quality of the measurements. Convergence of the estimates in static mode can take fifteen to thirty minutes, and the results have been shown to have centimeter-level accuracy. A challenge is to reduce the initialization period.

We note in passing that PPP with a single-frequency receiver is now receiving much attention. The current techniques offer decimeter-level positioning with convergence times similar to those for the dual-frequency data.

7.6.2 Online Positioning Services

It started with the JPL's Auto-Gipsy service in 1998. Now a half-dozen centers offer automated GPS data analysis service using e-mail and ftp (file transfer program), mostly for static positioning. The users make their raw, geodetic quality measurements collected for an hour, preferably longer, available on the Internet (files in RINEX format, see Appendix A). The data

are processed on the computers of the service provider and the results file is e-mailed to the user. Free access to the sophisticated positioning software is a great benefit, and it's a painless way to obtain centimeter-level position estimates.

The online services offered by JPL [http://milhouse.jpl.nasa.gov/ag/agfaq.html] and National Resources Canada (NRCan) [http://www.geod.nrcan.gc.ca/ppp_e.php] use the PPP approach outlined above. The U.S. National Geodetic Service's OPUS [http://www.ngs.noaa.gov/OPUS/index.html] and Scripps Orbit and Permanent Array Center's (SOPAC's) SCOUT [http://sopac.ucsd.edu/cgi-bin/SCOUT.cgi] use the traditional approach of double differencing the user data and multiple base station data from their reference networks. Similar services are available in Europe.

7.7 Summary

The carrier phase can be measured with millimeter-level precision but these measurements are corrupted by the same error sources that affect the code measurements: satellite clock and ephemeris error, uncertainties of the propagation medium, and receiver noise and multipath. The main differences between the code and carrier phase measurements are that the receiver noise and multipath corrupt the carrier phase typically at centimeter level and code phase at meter level, and the carrier phase measurements are ambiguous in whole cycles. If the errors can be mitigated and the ambiguities can be resolved, the carrier phase measurements can be turned into accurate pseudorange measurements and, then, into accurate position estimates.

The errors due to the satellite clock and ephemeris and the propagation medium for two users within tens of kilometers of each other are highly correlated, and can be mitigated by differencing their time-matched measurements. Such single differencing also re-parametrizes the problem from one of point positioning to relative positioning.

In this chapter, we have focused mainly on integer ambiguity resolution for accurate relative positioning. There is no basic challenge if the user is not in a hurry and the measurement errors can be kept in check. The estimation of the integers is helped by a change in the user-satellite geometry, referred to as geometric diversity. Actually, there is little change in such geometry over a few seconds, and the benefits of geometric diversity are being exploited in commercial systems through pseudolites. Important benefits also accrue from measurements at more than one frequency, referred to as frequency diversity. The benefits of frequency diversity provided the impetus for the major receiver manufacturers to develop proprietary techniques to measure the L2 signal, which had previously been off-limits to civil users. Additional benefits of frequency diversity would become available in a few years with the addition of coded civil signals at L2 and L5.

Centimeter-level relative positioning with carrier phase is now routine in surveying, geodesy, geophysics, and some industrial applications. The challenge lies in estimating the integers quickly and correctly for precise positioning in real time, ideally with measurements from a single epoch. Availability of precise satellite orbit and clock data in real time has led to centimeter-level point positioning in near-real time with measurements from a single receiver.

Homework Problems

7-1. Consider the estimation of the integer ambiguity and baseline length from (7.2) for the idealized relative positioning problem depicted in Figure 7.1. Derive the expression for the integer-ambiguity-resolution dilution of precision (IDOP) given by (7.4) by modeling the carrier phase measurement errors as uncorrelated and distributed identically with zero mean and variance σ^2.

7-2. Derive the expression for the single-difference range $r_{ur}^{(k)}$ given by (7.12) and shown in Figure 7.4.

7-3. To illustrate the amplification of measurement noise in wide laning, derive expression (7.36) starting with the definition of the wide-lane measurements in cycles as: $\phi_{L12} = \phi_{L1} - \phi_{L2}$.

7-4. Rewrite the double-difference carrier phase measurements (7.18) to include the residual ionospheric bias term I_q (in meters), q = L1 or L2, as follows.

$$\phi_q = \frac{r}{\lambda_q} + N_q - \frac{I_q}{\lambda_q} + \varepsilon_{\phi_q}$$

The wide-lane measurements (7.34) now appear as follows.

$$\phi_{L12} = \frac{r}{\lambda_{L12}} + N_{L12} + \frac{17}{60}\frac{I_{L1}}{\lambda_{L1}} + \varepsilon_{\phi_{L12}}$$

The wide-lane combination reduces the impact of the ionospheric bias expressed in cycles, making it easier to estimate N_{L12} than N_{L1} or N_{L2}. Show that the impact of the ionospheric bias is actually amplified when expressed in meters for the subsequent estimation of position. This is yet another reason why we'd prefer to leverage the knowledge of the wide-lane integer to estimate N_{L1} and N_{L2} for position estimation.

7-5. Consider a general linear combination of the carrier phase measurements

$$\phi_\alpha = \alpha_1 \phi_{L1} + \alpha_2 \phi_{L2}$$

(a) Show that the corresponding frequency and wavelength are

$$f_\alpha = \alpha_1 f_{L1} + \alpha_2 f_{L2}$$

$$\lambda_\alpha = \left(\frac{\alpha_1}{\lambda_{L1}} + \frac{\alpha_2}{\lambda_{L2}} \right)^{-1}$$

(b) Accounting for the ionospheric double difference explicitly, show that this measurement can be modeled as

$$\phi_\alpha = \frac{r}{\lambda_\alpha} + N_\alpha - \left(\alpha_1 + \alpha_2 \frac{\lambda_{L2}}{\lambda_{L1}} \right) \frac{I_{L1}}{\lambda_{L1}} + \varepsilon_{\phi_\alpha}$$

where

$$\left(\alpha_1 + \alpha_2 \frac{\lambda_{L2}}{\lambda_{L1}} \right)$$

is the ionospheric scale factor and $N_\alpha = \alpha_1 N_{L1} + \alpha_2 N_{L2}$ is the new ambiguity term.

(c) We can select

$$\alpha_1 = 1 \quad \text{and} \quad \alpha_2 = -\frac{\lambda_{L1}}{\lambda_{L2}} = -\frac{60}{77}$$

to eliminate the first-order effect of the ionosphere, but the ambiguities would no longer be integers. How much noise amplification would result from this ionosphere-free combination?

(d) We could select $\alpha_1 = 77$ and $\alpha_2 = -60$ to eliminate the ionospheric effect and preserve the integer nature of the ambiguity. Show that the corresponding wavelength, however, is a mere 0.006 m!

7-6. Take a different view of the problem depicted in Figure 7.1. Given the initial carrier phase measurement Δ_0 at epoch t_0, and accounting for the integer ambiguity, the correct answer for the baseline length is one of the following:

$$\lambda(\Delta_0 + N)/\cos\theta_0, \quad N = 0, 1, 2, \ldots$$

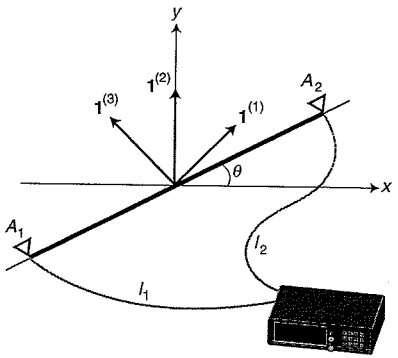
As the satellite moves to elevation θ_i and the carrier phase measurement changes to Δ_i, all the above values change, except for the right answer for the baseline length. This constancy is what distinguishes the right answer from the other candidates. If the measurements were perfect, we would know the right answer quickly. It would help to draw a picture similar to Figure 7.1, marking the candidate baseline lengths. (To draw the 'right' picture is to solve the problem.)

As usual, there is some uncertainty associated with the measurements, which translates into uncertainty about each of the candidates for the baseline length. So, the right answer wouldn't stand out right away. The bigger the uncertainty in the measurements, the longer you have to wait for the right answer to emerge. Determine the change in satellite elevation required for you to be able to identify the correct answer given that $\theta_0 = 45°$, $\lambda = \lambda_{L1}$, and the uncertainty in the carrier phase measurements is no more than 0.01 cycle. What if $\theta_0 = 10°$?

7-7. Consider an idealized attitude estimation problem. Two antennas lie on a plane at the ends of a baseline, which is at an unknown angle θ from the x-axis. The distance between the antenna phase centers is exactly 16 cm. There is an extra complication that the cables connecting the two antennas to a common receiver may be of unequal lengths. Three satellites are in view with the following line-of-sight vectors:

$$\mathbf{1}^{(1)} = (0.707,\ 0.707)^T$$
$$\mathbf{1}^{(2)} = (0,\ 1)^T$$
$$\mathbf{1}^{(3)} = (-0.707,\ 0.707)^T$$

The single-difference carrier phase measurements at L1 from the three satellites in centimeters are $\Phi = (9.1404, -3.0001, 8.5475)^T$. Estimate the angle θ, the integer ambiguities, and the line bias between the cables.



7-8. Show that the covariance matrix of the triple differences is given by (7.66).

7-9. Show that the transformation matrix $\mathbf{Z}$ and its inverse, required in the LAMBDA method, are volume-preserving transformations, i.e., $|\det(\mathbf{Z})| = 1$. (Hint: This derivation only requires that $\mathbf{Z}$ and its inverse have integer entries.)

References

Bock, Yehuda (1996). Medium Distance GPS Measurements, in *GPS for Geodesy*, A. Kleusberg and P. Teunissen (eds.), Lecture Notes in Earth Sciences, Springer, pp. 337–378.

Cohen, Clark E., Boris S. Pervan, David G. Lawrence, H. Stewart Cobb, J. David Powell, and Bradford W. Parkinson (1994). Real-Time Flight Testing Using Integrity Beacons for GPS Category III Precision Landing, *Navigation*, vol. 41, no. 2, pp. 145–157.

Cohen, Clark E. (1996). Attitude Determination, in *Global Positioning System: Theory and Applications, Vol. II*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 519–538.

Counselman III, C.C., I.I. Shapiro, R.L. Greenspan, and D.B. Cox, Jr. (1979). Backpack VLBI Terminal with Subcentimeter Capability, *Proc. Radio Interferometric Techniques for Geodesy*, NASA Conference Publication, vol. 2115, pp. 409–413.

Counselman III, Charles C. and Sergei Gourevitch (1981). Miniature Interferometer Terminals for Earth Surveying: Ambiguity and Multipath with Global Positioning System, *IEEE Transactions on Geosciences and Remote Sensing*, vol. GE-19, no. 4, pp. 244–252.

Dedes, George, and Clyde Goad (1994). Real-time cm-level GPS Positioning of Cutting Blade and Earth Moving Equipment, *Proc. ION National Technical Meeting*, pp. 587–593.

Gipson, John, Jim Ryan, and Tom Clark (1994). GPS Results are Consistent with VLBI, *Proc. ION GPS-94*, pp. 383–390.

Goad, Clyde C. (1996). Short Distance GPS Models, in *GPS for Geodesy*, Alfred Kleusberg and Peter J. G. Teunissen (eds.), Lecture Notes in Earth Sciences, Springer, pp. 239–262.

Han, Shaowei, and Chris Rizos (1997). Comparing GPS Ambiguity Resolution Techniques, *GPS World*, vol 8, no. 10, pp. 54–61.

Hassibi, Arash, and Stephen Boyd (1998). Integer Parameter Estimation in Linear Models with Applications to GPS, *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2938–2952.

Hatch, Ron (1982). The Synergism of GPS Code and Carrier Measurements, *Proc. Third International Symposium on Satellite Doppler Positioning*, DMA, Las Cruces, NM, pp. 1213–1232.

Hatch, Ron (1990). Instantaneous Ambiguity Resolution, *Proc. Kinematic Systems in Geodesy, Surveying, and Remote Sensing*, KIS Symposium, Banff, Canada, pp. 299–308.

Hatch, Ron, and Hans-Juergen Euler (1994). Comparison of Several AROF Kinematic Techniques, *Proc. ION GPS-94*, pp. 363–370.

Hatch, Ronald R. (1996). Promise of a Third Frequency, *GPS World*, vol. 7, no. 5, pp. 55–58.

Hatch, Ron, Jaewoo Jung, Per Enge, and Boris Pervan (2000). Civilian GPS: The Benefits of Three Frequencies, *GPS Solutions*, vol. 3, no. 4, pp. 1–9.

Héroux, P, Y. Gao, J. Kouba, F. Lahaye, Y. Mireault, P. Collins, G. Macleod, P. Tétreault, and K. Chen (2004). Products and Applications for Precise Point Positioning—Moving Toward Real-Time, *Proc. ION GNSS 2004*, pp. 1832–1843.

Hwang, Patrick Y.C. (1991). Kinematic GPS for Differential Positioning: Resolving Integer Ambiguities on the Fly, *Navigation*, vol. 38, no.1, pp. 1–15.

Joosten, Peter, and Christian Tiberius (2000). Fixing Ambiguities: Are you Sure They're Right? *GPS World*, vol. 11, no. 5, pp. 46–51.

Kouba, Jan and Pierre Héroux (2001). Precise Point Positioning Using IGS Orbit and Clock Products, *GPS Solutions*, vol. 5, no. 2, pp. 12–28.

Lachapelle, G., M.E. Cannon, and G. Lu (1992). Ambiguity Resolution On the Fly—A Comparison of P Code and High Performance C/A Code Receiver Technologies, *Proc. ION GPS-92*, pp. 1025–1032.

Langley, Richard B. (1998). RTK GPS, *GPS World*, vol. 9, no. 9, pp. 70–76.

Pratt, M., B. Burke, and P. Misra (1997), Single-Epoch Integer Ambiguity Resolution with GPS-GLONASS L1 Data, *Proc. ION 53rd Annual Meeting*, pp. 691–699.

Pratt, M., B. Burke, and P. Misra (1997). Single-Epoch Integer Ambiguity Resolution with GPS L1-L2 Carrier Phase Measurements, *Proc. ION GPS-97*, pp. 1737–1746.

Raquet, John, and Gérard Lachapelle (2000). Development and Testing of a Kinematic Carrier Phase Ambiguity Resolution Method Using a Reference Receiver Network, *Navigation*, vol. 46, no. 4, pp. 283–296.

Remondi, Benjamin W. (1985). Performing Centimeter-Level Surveys in Seconds with GPS Carrier Phase: Initial Results, *Navigation*, vol. 32, no. 4, pp. 386–400.

Schupler, Bruce R., and Thomas A. Clark (2001). Characterizing the Behavior of Geodetic GPS Antennas, *GPS World*, vol. 12, no. 2, pp. 48–54.

Seeber, Günter (2003). *Satellite Geodesy: Foundations, Methods, and Applications*, 2nd edition, Walter de Gruyter.

Strang, Gilbert, and Kai Borre (1997). *Linear Algebra, Geodesy, and GPS*, Wellesley-Cambridge Press.

Talbot, Nicholas (1993). Centimeters in the Field, A User's Perspective of Real-Time Kinematic Positioning in a Production Environment, *Proc. ION GPS-93*, pp. 1049–1057.

Teunissen, Peter J.G. (1996). GPS Carrier Phase Ambiguity Fixing Concepts, in *GPS for Geodesy*, A. Kleusberg and P. Teunissen (eds.), Lecture Notes in Earth Sciences, Springer, pp. 263–335.

Teunissen, P.J.G., P.J. De Jonge, and C.C.J.M. Tiberius (1997). Performance of the LAMBDA Method for Fast GPS Ambiguity Resolution, *Navigation*, vol. 44, no. 3, pp. 373–383.

Tiberius, Christian, Niels Jonkman, and Frank Kenselaar (1999). The Stochastics of GPS Observables, *GPS World*, vol. 10, no. 2, pp. 49–54.

van Graas, Frank, David W. Diggle, and Richard M. Hueschen (1994). Interferometric GPS Flight Reference/Autoland System: Flight Test Results, *Navigation*, vol. 41, no. 1, pp. 57–82.

van Graas, Frank, and Shane-Woi Lee (1995). High Accuracy Differential Positioning for Satellite–based Systems without Using Code-Phase Measurements, *Navigation*, vol. 42, no. 4, pp. 605–618.

Wübbena, Gerhard (1989). The GPS Adjustment Software Package GEONAP, Concepts and Models, *Proc. 5th International Geodetic Symposium on Satellite Positioning*, Las Cruces, NM, pp. 452–461.

Yang, Ming, Clyde Goad, and Burkhard Schaffrin (1994). Real-Time On-the-Fly Ambiguity Resolution Over Short Baselines in the Presence of Anti-spoofing, *Proc. ION GPS-94*, pp. 519–525.

PART III

GPS Signals

The GPS signals travel 20,000 kilometers from a medium earth orbit and arrive at the earth's surface with a power density of only 10^{-14} to 10^{-13} watts/m^2. Even so, these electromagnetic whispers:

- Enable high-precision ranging even in the presence of natural noise and modest amounts of man-made interference. Indeed, the precision of the pseudorange measurements is approximately 0.5 meters for the civilian code. This remarkable ability derives from the auto-correlation properties of the spread spectrum codes that modulate the signal from each satellite.
- Distinguish the direct (desired) signal from reflected signals. This handy property also derives from the auto-correlation properties of the codes.
- Allow the satellites to simultaneously use the same transmission frequencies. The satellites do not offset their carrier frequencies. Nor do they use a time sharing scheme, where each satellite must transmit only during a prescribed time slot. As we shall discover, this multiple-access property derives from the cross-correlation properties of the GPS codes.
- Carry all the ancillary data required for position fixing. As described in Chapter 4, each satellite sends an estimate of its ephemeris and clock offset relative to GPS time.

The GPS signal and user receivers act in concert to achieve these goals. Both are carefully crafted and well worth our study. To enable this study, Chapter 8 provides a review of the signals and systems tools that we will need in the remainder of this book. It includes an introduction to convolution, Fourier series, Fourier transforms, and Laplace transforms. If the reader is familiar with these bedrock techniques, than please feel free to skip Chapter 8 or simply use it for reference. Chapters 9 and 10 discuss the GPS signal design in earnest. Chapter 9 uses the tools from Chapter 8 to provide time and frequency domain characterizations of the GPS signal. It also introduces the auto-correlation and cross-correlation functions that give the signals properties listed above, and it provides a random sequence analysis of the GPS codes. Chapter 10 analyzes the power in the GPS signals and the competing background noise. The ratio of these two powers is called the signal-to-noise ratio (SNR) and is a critical system parameter. In fact, the fundamental ranging performance of GPS is determined by SNR, bandwidth of the signal, and averaging time used by the receiver. To enable this analysis, Chapter 10 introduces the delay lock loop used by almost all GPS receivers.

Chapter 8

Signals and Linear Systems

A GPS receiver is a *signal processing* engine. The front end of the receiver amplifies the GPS signals and filters the received signals to remove unwanted signals that may interfere with the GPS signals. The front end also shifts the frequency of the GPS signals downward to more manageable intermediate frequencies. The next stage of the receiver correlates the incoming signals with internally generated replicas of the pseudo-random noise codes. This digital signal processing stage searches for the individual satellite signals amongst the forest of signals from other satellites, natural noise and man-made radio frequency interference. Once the signals are found, the receiver continuously estimates the arrival time of the GPS pseudo-random noise (PRN) codes and the underlying carrier signals.

Happily, most of these processes can be modeled as linear time-invariant (LTI) systems acting on input signals to provide output signals. In fact, the GPS receiver is well modeled as a cascade of linear time-invariant (LTI) subsystems, each with a prescribed task. The GPS signals are the electrical voltages and currents that propagate through this cascade, with each subsystem processing the signals in transit. To facilitate our study of GPS receivers and signals, this chapter introduces the beautiful theory of linear systems and signals.

Section 8.1 describes the most important building block signals that we need. Chief among these are *imaginary exponential* functions and *complex exponential* functions, which we shall call sinusoids for brevity. Singularity functions are also important and we will study the unit step, unit ramp, unit pulse and impulse. Section 8.1 also defines signal energy and power.

Convolution enables us to solve for the output of an LTI system, given the input. It is the subject of Section 8.2. Convolution decomposes the input signal into a weighted sum of pulses. For accuracy, convolution uses very narrow pulses—in fact, the pulses become impulse functions in the limit. Each impulse gives rise to an impulse response. Hence, the input is modeled as a weighted sum of impulses, with the output of an LTI system equaling a weighted sum of impulse responses.

Convolution is fairly easy to visualize, but impulse responses are not always easy to measure or derive. For this reason, we shift our attention to sinusoids. Sinusoids form the basis for most of our analysis. They are *basis functions* of the space of functions of interest to us. Amongst sinusoids, the imaginary exponentials comprise the Fourier basis and the complex exponentials comprise the Laplace basis. Neither is universally most powerful. One basis makes short work of one problem, but is awkward in another setting.

Sinusoids are important to us for three reasons. First, the response of a linear system to an isolated sinusoid is generally easy to measure in practice. Electrical engineering laboratories are full of instruments that measure sinusoidal responses. These instruments include oscilloscopes that enable us to measure the response to one sinusoid at a time. They also include spectrum analyzers and network analyzers that sweep through large sets of sinusoids automatically. We will show some measurements from these powerful instruments in the chapters that follow.

Second, sinusoids in LTI systems are easy to analyze because they are *eigenfunctions* of such systems. If the input to an LTI system is a sinusoid (either an imaginary or complex exponential), then the output is also a sinusoid. The amplitude may be greater or smaller. The sinusoid may be shifted in time, but the form of the signal is undisturbed. Moreover, the *transfer function* of the system describes the output amplitude and time shift relative to the input amplitude and time shift, and the transfer function is easy to measure or derive. This proof is offered in Section 8.3, which discusses transfer functions and basis functions.

Third, sinusoids can be combined to model very non-sinusoidal signals. All signals of interest to us can be adequately approximated by a weighted sum of sinusoids. We find the weights or coefficients needed to model the input. Then the output is equal to the same weighted sum, except that each elemental sinusoid is scaled by the associated value of the transfer function.

The Fourier series, the subject of Section 8.4, is a preferred tool for the analysis of periodic signals. Our main application of the Fourier series will be to study the periodic signal that arises when the GPS receiver samples the incoming GPS signal. Sampling converts the signal from an analog signal in the front end of the receiver to the digital signal that is processed by the later stages of the receiver. This analog to digital conversion is further studied in Chapters 11 and 13, but we set the stage with an example in Section 8.4.

The Fourier transform is perfect for the study of modulation and many filtering problems. After we introduce the Fourier transform in Section 8.5, we study modulation and filters as examples that are particularly relevant to GPS.

Random signals are studied in Section 8.6. They contrast the deterministic signals considered in the remainder of this chapter. Deterministic signals are described as explicit functions of time. A random signal depends on time, but also on chance. It should be viewed as a member of a great collection of possible functions of time. Depending on some underlying random event, one of the possible functions of time is drawn from the collection. Random signals populate GPS. The *noise* generated in the front end of a GPS receiver is a random signal. Man-made signals that come from terrestrial transmitters may interfere with GPS operation. These are also well described as random signals. We will also discover that the spread-spectrum codes that enable precise ranging are well modeled as sequences of random events (coin flips). In this case, the GPS signals themselves are modeled as random signals. Happily, the transform techniques that we develop in this chapter also work well with random signals.

Finally, the Laplace transform is more convenient than the Fourier transform for certain input signals including the step function, ramp function, and parabolic function. It is also more convenient for finding the complete solution to linear differential equations subject to nonzero initial conditions. After defining the Laplace transform at the beginning of Section 8.7, we include subsections that demonstrate these strengths.

This chapter reviews the tools needed in Parts III and IV. Chapter 9 will use the impulse

function and convolution to describe the GPS signals as a function of time (time-domain analysis). It will use the Fourier transform to describe the frequency content of the GPS signals. As mentioned earlier, Chapter 11 uses the Fourier series and transform to study analog to digital conversion. Chapter 12 uses the Laplace transform to analyze the delay lock loops and carrier tracking loops included in every GPS receiver.

This chapter is not intended as a general introduction to signals and systems. A student with no prior background in this subject has to be prepared to work hard and may wish to consult any of several excellent books. We recommend the following for expanded and alternate treatments of the topics covered here: Pursley (2002), Ziemer and Tranter (1995), Papoulis (1962), Scott (1987), McGillem and Cooper (1984), and Franklin, Powell, and Emami-Naeini (1994). Readers familiar with the basics of convolution and transform theory may skip this chapter with no great loss of continuity.

8.1 Overview

8.1.1 Linear Time Invariant Systems

Most GPS subsystems can be treated as linear time-invariant (LTI) systems. Figure 8.1 shows a single-input, single-output LTI system. For our LTI systems, two pieces of information are needed to find the output, $y(t)$, for any given input, $x(t)$. First, we need the linear differential equation that describes the system under study.

$$b_N \frac{d^N y(t)}{dt^N} + b_{N-1} \frac{d^{N-1} y(t)}{dt^{N-1}} + \cdots + b_0 y(t) =$$
$$a_M \frac{d^M x(t)}{dt^M} + a_{M-1} \frac{d^{M-1} x(t)}{dt^{M-1}} + \cdots + a_0 x(t) \qquad (8.1)$$

Second, we need a set of the initial conditions. In most of our work on GPS, we can ignore the response due to non-zero initial conditions, so we will not pay too much attention to these initial conditions. However, Section 8.7.4 does describe a powerful technique to incorporate them, should they be important.

Linear differential equations, like (8.1), are used to describe an enormous variety of systems across engineering. These include translational mechanical systems where the inputs and outputs might be forces and accelerations. They include rotational systems where the inputs and outputs might be torques and rotational rates. They also include thermal systems and liquid systems. However, the systems in this book are electrical, and $x(t)$ and $y(t)$ are voltages or currents that appear throughout the GPS system.

For the time being, we compact (8.1) as follows

$$x(t) \xrightarrow[H]{} y(t) \qquad (8.2)$$

This simple notation strives to communicate that system H maps the input, $x(t)$, into the output, $y(t)$.

Our system is time invariant, because

$$x(t) \xrightarrow[H]{} y(t) \Rightarrow x(t-\tau) \xrightarrow[H]{} y(t-\tau) \qquad (8.3)$$

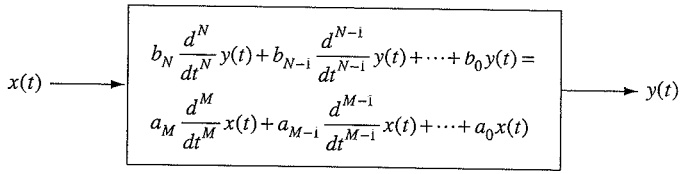


Figure 8.1 Block diagram of a single-input, single-output linear time-invariant (LTI) system. The input and output are signals denoted $x(t)$ and $y(t)$, respectively. This simple block diagram represents any of a wide variety of systems that can be described by a linear differential equation.

If the input, $x(t)$, yields $y(t)$, then a delayed version of $x(t)$ will simply yield a delayed version of $y(t)$. This means that the coefficients a_n and b_n in (8.1) do not vary with time.

LTI systems have some more nice properties. First, if the input is scaled by a constant, then the output will simply be scaled by the same constant.

$$x(t) \xrightarrow[H]{} y(t) \Rightarrow kx(t) \xrightarrow[H]{} ky(t)$$

If $x(t)$ produces $y(t)$, then $kx(t)$ will produce $ky(t)$. Second, the superposition of two inputs will produce the superposition of the corresponding outputs. More formally, we write

$$\left.\begin{aligned} x_1(t) &\xrightarrow[H]{} y_1(t) \\ x_2(t) &\xrightarrow[H]{} y_2(t) \end{aligned}\right\} \Rightarrow x_1(t) + x_2(t) \xrightarrow[H]{} y_1(t) + y_2(t)$$

Finally, we may put these properties together as follows

$$k_1 x_1(t-\tau_1) + k_2 x_2(t-\tau_2) \xrightarrow[H]{} k_1 y_1(t-\tau_1) + k_2 y_2(t-\tau_2)$$

As we shall see in Section 8.2, these properties will allow us to find the output of a linear system given the input. First, we need to spend some time with the signals that will interest us. Their character also influences our strategy.

8.1.2 Sinusoids

We begin with the ubiquitous cosine wave.

$$x(t) = A\cos(2\pi f_0 t + \theta)$$

This function is shown in Figure 8.2. As shown, it has three parameters. The amplitude, A, scales the cosine wave in the vertical direction. The phase, θ, shifts the waveform earlier or later along the time axis. In fact, $A\sin(2\pi f_0 t) = A\cos(2\pi f_0 t - \pi/2)$. Finally, the cosine wave is characterized by a frequency, f_0, because it is *periodic*. It repeats every T_0 seconds with the frequency counting the number of periods (or cycles) that occur in a given amount of time. Hence, we have $f_0 = 1/T_0$ and the units are cycles per second, or hertz (Hz). From time to time, frequency is expressed as radians per second rather than cycles per second. The radian frequency is given by $\omega_0 = 2\pi f_0$ because there are 2π radians per cycle. In this book, we will try to use f rather than ω.

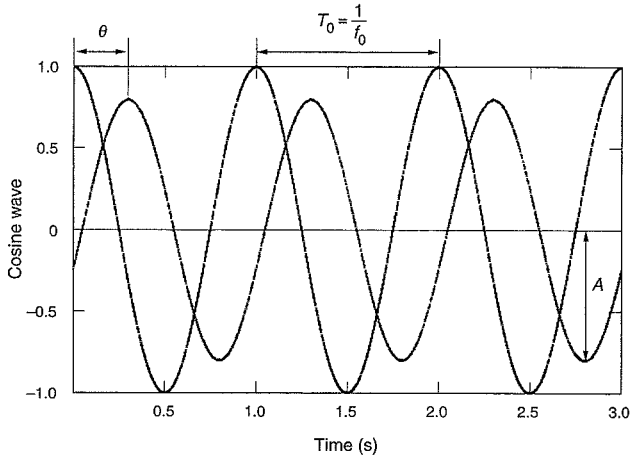


Figure 8.2 Cosine waves.

The cosine wave has a central role in electrical engineering. Power engineers use it to model the waveforms that carry power throughout our civilization. For example, the voltage that comes from North American wall outlets is a cosine wave with an amplitude of $120\sqrt{2}$ volts and a frequency of 60 Hz. In Europe, the amplitude is $240\sqrt{2}$ volts and the frequency is 50 Hz.

GPS engineers use the cosine wave to model the radio frequency (RF) signal that carries the ranging codes and navigation data from the satellite to the user. Indeed, we are a small subset of radio engineers who all rely on cosine waves to serve as electro-magnetic vehicles for information. The frequency of a radio carrier places the signal in the radio spectrum. The GPS signals at L1 have a frequency of 1575.42×10^6 hertz or 1575.42 Megahertz (MHz). The L2 signals and L5 signals have carrier frequencies of 1227.68 and 1176.45 MHz, respectively. All of these signals belong to the so-called L band portion of the radio spectrum.

When used as a radio frequency carrier, the cosine waves can equally well be characterized by their wavelength. If we froze time, we would discover that each cycle had a physical length of c/f_0 where $c \approx 3 \times 10^8$ m/s is the speed of light. With this conversion, we find that the GPS L1 carrier has a wavelength of $3\times10^8/1575.42\times10^6 \approx 0.19$ meters or 19 cm. The L2 carrier has a wavelength of $3\times10^8/1227.68\times10^6 \approx 0.24$ m, and the L5 carrier has a wavelength of 25 cm. We will have much more to say about the GPS carriers in Chapters 9 and 10.

Next, we consider an *imaginary exponential.*

$$x(t) = A\exp j(2\pi f_0 t + \theta)$$

As shown, it has the same three parameters as the cosine wave: amplitude, A, phase, θ, and frequency, f_0. In fact, the connection is even deeper. From Euler's great formula, we find

$$A\exp j(2\pi f_0 t + \theta) = A(\cos(2\pi f_0 t + \theta) + j\sin(2\pi f_0 t + \theta))$$

$$\begin{aligned} A\cos(2\pi f_0 t + \theta) &= \frac{A}{2}\left(\exp j(2\pi f_0 t + \theta) + \exp(-j(2\pi f_0 t + \theta))\right) \\ &= \mathrm{Re}\left\{A\exp j(2\pi f_0 t + \theta)\right\} \\ A\sin(2\pi f_0 t + \theta) &= \frac{A}{2j}\left(\exp j(2\pi f_0 t + \theta) - \exp(-j(2\pi f_0 t + \theta))\right) \\ &= \mathrm{Im}\left\{A\exp j(2\pi f_0 t + \theta)\right\} \end{aligned} \tag{8.4}$$

We will leverage this connection in Section 8.3 and in later chapters. We will find that imaginary exponentials are key to our study of Fourier series and Fourier transforms. They will also permeate our study of the GPS signal and receivers.

Finally, we take a look at the *complex exponential* function, $A\exp(j\theta)\exp(s_0 t)$. This function is central to our study of Laplace transforms. More specifically, it is vital to our study of the code and carrier tracking operations within the GPS receiver. Like the cosine wave and imaginary exponential, the complex exponential has an amplitude, A, and phase, θ. However, the frequency, s_0, is now a complex number.

$$\begin{aligned} s_0 &= \sigma_0 + j2\pi f_0 \\ &= \sigma_0 + j\omega_0 \end{aligned} \tag{8.5}$$

Consequently, we may write

$$\begin{aligned} A\exp(j\theta)\exp(s_0 t) &= A\exp j\theta\exp(\sigma_0 + j2\pi f_0)t \\ &= A\exp j\theta\exp(\sigma_0 t)\exp(j2\pi f_0 t) \\ &= A\exp j\theta\exp(\sigma_0 t)\left(\cos(2\pi f_0 t) + j\sin(2\pi f_0 t)\right) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Re}\left\{A\exp(j\theta)\exp(s_0 t)\right\} &= A\exp(\sigma_0 t)\cos(2\pi f_0 t + \theta) \\ \mathrm{Im}\left\{A\exp(j\theta)\exp(s_0 t)\right\} &= A\exp(\sigma_0 t)\sin(2\pi f_0 t + \theta) \end{aligned} \tag{8.6}$$

Examples of $\mathrm{Re}\left\{\exp(s_0 t)\right\}$ are sketched in Figure 8.3. At the origin, $\sigma_0 = 0$ and $f_0 = 0$, so the signals are constant functions of time. Their exponential envelope is constant, $\exp(0t) = 1$, and their underlying sinusoids have zero frequency, $\exp(j2\pi 0t) = 1$. As we move up the $\mathrm{Im}\{s_0\}$ axis, the exponential envelope remains constant, but the sinusoidal variation appears. Close to the origin, the frequency is low. As we move away from the origin, the frequency becomes higher.

The horizontal axis plots the real part of the complex frequency, $\mathrm{Re}\{s_0\} = \sigma_0$. This parameter controls the exponential envelope of the signal. Negative values of σ_0 cause the function to decrease with time, and positive values give increasing functions.

Needless to say, signals that increase exponentially with time are worrisome. Electrical systems break down if the voltage is too high. They burn out if the currents are beyond tolerance. Mechanical systems have similar frailties. Consequently, *stability* is a major concern of all engineers, including navigation engineers. We desire bounded-input bounded-output (BIBO) stability. If the input is bounded in magnitude, then the output should also be bounded. For much more on stability, the reader is referred to Franklin, Powell, and Emami-Naeini (2002).

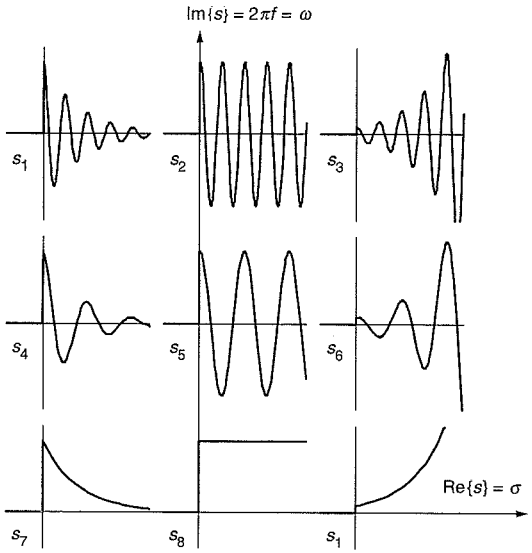


Figure 8.3 Plots of $\exp(s_i t)$ for nine values of s_i.

8.1.3 Singularity Functions: Unit Step, Unit Pulse, and Impulse

Unlike sinusoids, singularity functions are unruly. They are discontinuous or have discontinuous derivatives. Charter members of this class are the unit step, ramp, and parabola functions. The unit step is given by

$$u(t)=\begin{cases}1 & t\geq 0\\ 0 & \text{otherwise}\end{cases}$$

The unit ramp and unit parabola functions are given by

$$tu(t)=\begin{cases}t & t\geq 0\\ 0 & \text{otherwise}\end{cases}$$

$$t^2u(t)=\begin{cases}t^2 & t\geq 0\\ 0 & \text{otherwise}\end{cases} \qquad (8.7)$$

The step function is very important to navigation. If our roving receiver takes a step to the left, we want the position estimate to also step crisply to the left. We want the reported position to converge to the true position without any long delay. At the same time, we do not want the reported position to step wildly beyond the true position, nor do we want any steady state errors. As such, the step response is a critical navigation stress test.

The unit ramp and parabola have similar importance. A ramp in position (or range) corresponds to a step change in user *velocity*. A parabola in position (or range) corresponds to a step change in user *acceleration*. We have the same three criteria: rapid convergence to the new value of velocity or acceleration; small overshoot; and small steady state errors. Taken together, the step response, ramp response, and parabolic response form a trilogy of stress tests to evaluate the dynamic performance of a navigation system.

The unit pulse is another singularity function that will serve us well.

$$p(t)=u(t+1/2)-u(t-1/2)$$
$$=\begin{cases}1 & \frac{-1}{2}\leq t\leq\frac{1}{2}\\ 0 & \text{otherwise}\end{cases} \qquad (8.8)$$

Two such pulses, $p(t)$ and $Ap(\frac{t-\tau}{T_C})$, are shown in Figure 8.4. The first has unity amplitude, unity duration, and is centered on $t = 0$. The second has amplitude, A, duration, T_C seconds, and is shifted by τ seconds.

The unit pulse also finds great utility in the work ahead. First, it will be key to our development of the convolution integral in the next section, where we will leverage its connection to the impulse function. Second, we will use the unit pulse when describing the GPS signal. GPS modulates the RF carriers with a pseudo-random noise (PRN) sequence, and the unit pulse well describes the individual *chips* in this sequence. We will have much more to say about these sequences in Chapter 9.

The impulse function is the final singularity function that we need to consider. Heuristically, it can be defined as a unit pulse that is becoming increasingly spike-like.

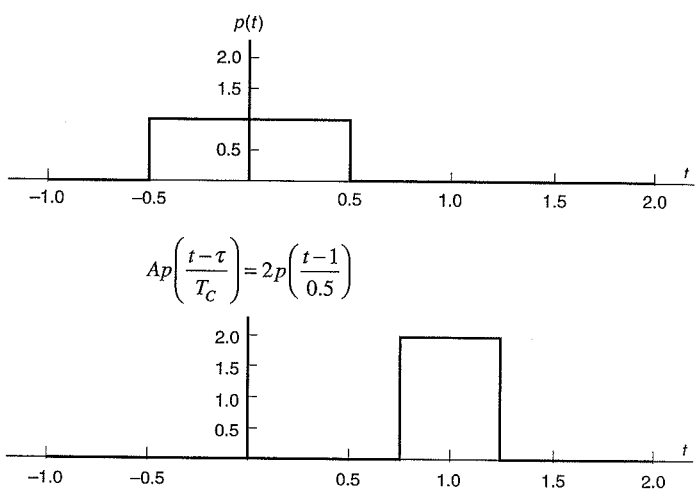


Figure 8.4 Unit pulse and a unit pulse that has been shifted, compressed, and scaled.

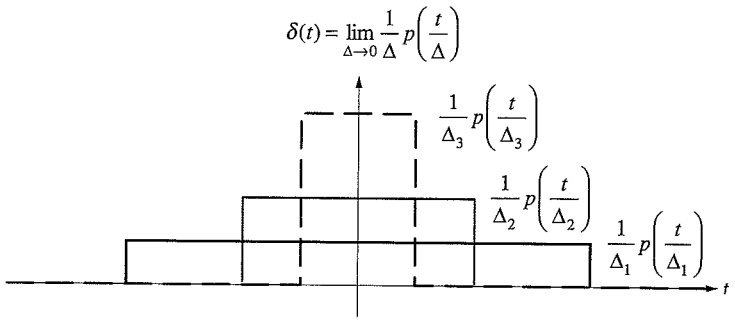


Figure 8.5 Construction of the unit impulse.

$$\delta(t) = \lim_{\Delta\to 0} \frac{1}{\Delta} p\left(\frac{t}{\Delta}\right)$$
$$\delta(t-\tau) = \lim_{\Delta\to 0} \frac{1}{\Delta} p\left(\frac{t-\tau}{\Delta}\right) \tag{8.9}$$

As shown in Figure 8.5, the approximation for $\delta(t)$ maintains unity area as it becomes infinitesimally narrow. In the limit, the unit impulse function has zero width, infinite amplitude, and unit area. It is always equal to zero except at the origin, where it is undefined. This lack of definition troubles mathematicians and they do not like to think of $\delta(t)$ as an ordinary function that assigns specific values to given values of t. They call the unit impulse function a generalized function or distribution. The unit impulse is best defined by the following rule.

$$\int_{t_0}^{t_2} x(t)\delta(t)dt = x(0) \tag{8.10}$$

As shown, $\delta(t)$ simply assigns the value $x(0)$ to this integral, where $x(t)$ simply needs to be continuous at $t = 0$ with $t_0 < 0 < t_2$. If we substitute variables, then we can also write

$$\int_{t_0}^{t_2} x(t)\delta(t-t_1)dt = x(t_1) \tag{8.11}$$

where $t_0 < t_1 < t_2$. As shown, $\delta(t-t_1)$ *sifts* through $x(t)$ and simply returns the value of x at t_1.

Unit impulses will provide significant utility in the work to come, but they must be used with care. They are best kept inside an integral. After all, their defining property is the sifting property, which speaks to their action when they appear inside an integral. We have no straightforward definition for $\delta(t)$ in any other context. The homework problems showcase one pitfall that results from careless use of the unit impulse. More material can be found in Papoulis (1962).

8.1.4 Signal Power and Energy

This book contemplates signals that may exist somewhere inside the GPS receiver or the radio transmitter carried by the satellite. From time to time, we will be interested in the amount of energy contained in the signal. At other times, we will be interested in the amount of power carried by one of these signals. Certainly, we will have this interest when we estimate how much power is captured by a GPS receiver after the signal has traveled from orbit to the surface of the earth.

If the signal is a voltage $x(t) = v(t)$, then the amount of energy dissipated in a resistance R from time t_1 to t_2 is

$$\mathcal{E} = \int_{t_1}^{t_2} \frac{|v(t)|^2}{R} dt \quad \text{joules}$$

If the signal is a current, then $x(t) = i(t)$ and

$$\mathcal{E} = \int_{t_1}^{t_2} R|i(t)|^2 dt \quad \text{joules}$$

In either case, the energy lost is proportional to the square of the signal. For simplicity, radio engineers frequently assume that $R = 1$ ohm. This way, we can write

$$\mathcal{E} = \int_{t_1}^{t_2} |x(t)|^2 dt \quad \text{joules} \tag{8.12}$$

Signals that have finite energy even when the time interval becomes infinite are called *energy signals*. We will discover that GPS has some very important energy signals. Lead amongst these is the pulse signal described in Section 8.1.3. It has a total energy of

$$\begin{aligned}\mathcal{E} &= \int_{-\infty}^{\infty} \left| Ap\left(\frac{t-\tau}{T_C}\right)\right|^2 dt \\ &= \int_{\tau-T_C/2}^{\tau+T_C/2} A^2 dt \\ &= A^2 T_C \quad \text{joules}\end{aligned}$$

However, not all signals of import to GPS are energy signals. For example, the sinusoidal signals discussed in Section 8.1.2 have infinite energy. These signals are characterized by their *average* power, where power is the amount of energy transferred per unit time. The time average of signal power from t_1 to t_2 is

$$P = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} |x(t)|^2 dt \quad \text{watts}$$

This definition continues to assume that $x(t)$ is a voltage across or current through a 1-ohm resistor. If the signal carries finite power even over an infinite time interval, then we call it a *power signal*. Formally, a power signal has the following property.

$$P = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 dt < \infty$$

For examples, the cosine wave and the imaginary exponential both have finite average power.

$$P = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \left|A\cos(2\pi f_0 t + \theta)\right|^2 dt$$
$$= \frac{A^2}{2} \quad \text{Watts}$$
$$P = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \left|A\exp j(2\pi f_0 t + \theta)\right|^2 dt$$
$$= A^2 \quad \text{Watts} \tag{8.13}$$

The unit step function also has finite power, but the unit ramp and parabola do not.

These definitions shall take both physical and mathematical significance in GPS. GPS receivers attempt to collect as much signal energy as possible before making any estimates of pseudorange or before decoding the navigation messages. Greater energy means more accurate pseudorange estimates and more reliable decisions about the value of a navigation data bit. Since energy is the integral of power, we are acutely sensitive to the power in the received GPS signal. Chapter 10 estimates this critical system parameter.

Mathematically, energy signals are well handled by Fourier transforms. Power signals also have Fourier transforms, but these transforms include impulse functions or higher order generalized functions. These can be awkward to work with, so we turn to the Laplace transform for the analysis of the unit step function, the ramp, and the parabolic function.

8.2 Convolution

8.2.1 Superposition of Impulse Responses

The superposition property is now used to find the response of a linear system to an arbitrary input. This line of attack begins by approximating the input, $x(t)$, with the following piecewise constant function.

$$x(t) \approx \sum_{n=-\infty}^{\infty} x(n\Delta)p\left(\frac{t-n\Delta}{\Delta}\right) \tag{8.14}$$

The sensibility of (8.14) can be seen in Figure 8.6. As shown, the approximation is rather rough if $x(t)$ changes appreciably during the time interval Δ. To ensure an accurate approximation, we simply take the limit as Δ becomes small.

$$x(t) = \lim_{\Delta\to 0} \sum_{n=-\infty}^{\infty} x(n\Delta)p\left(\frac{t-n\Delta}{\Delta}\right)$$
$$x(t) = \lim_{\Delta\to 0} \sum_{n=-\infty}^{\infty} x(n\Delta)\frac{1}{\Delta}p\left(\frac{t-n\Delta}{\Delta}\right)\Delta$$
$$x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)\,d\tau \tag{8.15}$$

The final integral results from a limiting process that causes $n\Delta$ to tend towards τ, Δ to approach $d\tau$, and $p(t/\Delta)/\Delta$ to approach $\delta(t)$.

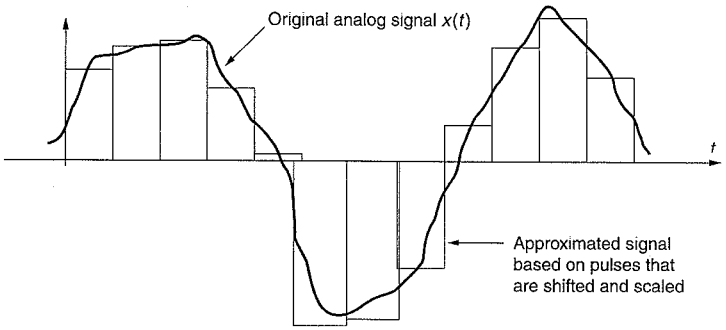


Figure 8.6 Continuous analog signal, $x(t)$, and an approximation based on pulses that are shifted and scaled.

Now, we define the *impulse response* as the output of a linear system when the input is a unit impulse

$$\delta(t) \xrightarrow[H]{} h(t)$$

Since our system is linear and time invariant, it has the following properties

$$k\delta(t) \xrightarrow[H]{} kh(t)$$
$$k\delta(t-\tau) \xrightarrow[H]{} kh(t-\tau)$$

If the input to our linear system, $x(t)$, can be expressed as a weighted sum of time-shifted impulses, our output, $y(t)$, will be a weighted sum of time-shifted, impulse responses as follows.

$$y(t) = \lim_{\Delta\to 0} \sum_{n=-\infty}^{\infty} x(n\Delta)h(t-n\Delta)\Delta$$
$$= \int_{-\infty}^{\infty} x(\beta)h(t-\beta)\,d\beta$$
$$= \int_{-\infty}^{\infty} x(t-\beta)h(\beta)\,d\beta$$
$$= x(t) * h(t)$$

The convolution integral, $x(t)*h(t)$, has an enormous role in the study of signals and systems. The output of a time-invariant, linear system is given by the convolution of the input with the impulse response of the linear system. There is one caveat. Convolution computes the so-called zero-state response of the linear system. In other words, it ignores any initial energy stored in the system. In general, we would need to add the zero-input response, which accounts for the initial energy storage. The complete response, $y(t)$, is equal to the zero-state response plus the zero-input response. Most GPS problems concern themselves with the zero-state response. Even so, we do return to the zero-input response in Section 8.7.4.

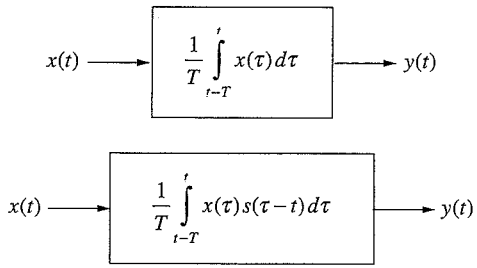


Figure 8.7 Block diagram of a moving-average filter and a weighted moving-average filter.

All realizable systems are *causal* or *non-anticipatory*. In other words, they do not respond in anticipation of an input that has not yet begun. Consequently, they have impulse responses, which are zero for negative time, $h(t) = 0$ for $t < 0$. System causality allows us to further develop the convolution integral.

$$\begin{aligned} y(t) &= \int_0^{\infty} x(t-\beta)h(\beta)d\beta \\ &= \int_{-\infty}^{t} x(\alpha)h(t-\alpha)d\alpha \end{aligned} \tag{8.16}$$

If the input is also zero for negative time, then the convolution becomes

$$\begin{aligned} y(t) &= \int_0^{t} x(\alpha)h(t-\alpha)d\alpha \\ &= \int_0^{t} x(t-\beta)h(\beta)d\beta \end{aligned} \tag{8.17}$$

8.2.2 Example: Moving Averages

We now work some examples using convolution. Consider the systems shown in Figure 8.7. These systems compute moving averages and are closely related to the correlators that are used in GPS receivers. All moving averages will tend to smooth the input signal and suppress details.

The top system in Figure 8.7 is called a box-car moving average because it averages the last T seconds of data and all the input data is weighted evenly. The impulse response for this system is

$$\begin{aligned} h_1(t) &= \frac{1}{T}\int_{t-T}^{t} \delta(\tau)d\tau = \begin{cases} 0 & t<0 \\ \frac{1}{T} & 0 \le t < T \\ 0 & T \le t \end{cases} \\ &= \frac{1}{T}(u(t)-u(t-T)) \end{aligned} \tag{8.18}$$

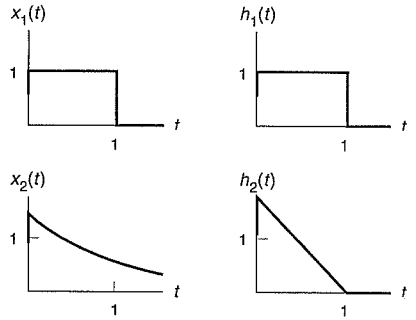


Figure 8.8 A variety of input signals and impulse responses.

where $u(t)$ is the unit step function. Take a moment to examine this result. Even for this simple example, finding the impulse response requires rather careful consideration of whether the impulse falls within the limits of integration. The resulting $h_1(t)$ is plotted in the upper right hand corner of Figure 8.8.

As another example, consider the weighted moving average filter shown on the bottom of Figure 8.7. This structure is often used to de-emphasize the input data as they become older. This might be a reasonable strategy if the newer data is likely to be more accurate than older data.

Any number of de-weighting functions, $s(t)$, can be considered, and the impulse response can be found as follows.

$$\begin{aligned} h_2(t) &= \frac{1}{T}\int_{t-T}^{t} \delta(\tau)s(\tau-t)d\tau \\ &= \begin{cases} 0, & t<0 \\ \frac{1}{T}s(-t), & 0 \le t < T \\ 0, & T \le t \end{cases} \end{aligned} \tag{8.19}$$

Once again, the lion's share of the work involves determining whether the impulse function falls within the limits of integration. If yes, then the sifting property readily returns $s(-t)$. The impulse response is simply the time reverse of the weighting function.

For both of the examples in Figure 8,7, we found the impulse response, more or less, by inspection. In general, the impulse response is harder to come by. Please see Gabel and Roberts (1973) for a more complete set of strategies for finding impulse responses from differential equations that describe system dynamics.

We now consider the two-by-two matrix of problems defined by Figure 8.8. Two inputs are shown in the left hand column, with two impulse responses shown in the right hand column. We wish to find the outputs using convolution. Rather than work all four problems, we will work one of the more challenging ones.

Consider the convolution of the exponential input, $x_2(t)$, with the ramp impulse response,

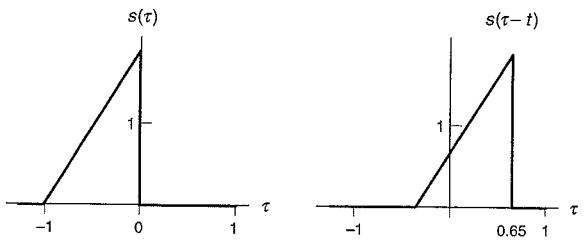


Figure 8.9 Weighting function and shifted weighting function.

$h_2(t)$. The convolution is given by

$$\begin{aligned} y_{2,2}(t) &= x_2(t) * h_2(t) \\ &= \int_0^\infty x_2(\tau) h_2(t-\tau)\, d\tau \\ &= \int_{\max(0,t-T)}^{t} \sqrt{2}\exp(-\tau/T)\frac{\sqrt{3}}{T}\left(1-\frac{t}{T}+\frac{\tau}{T}\right)d\tau \\ &= \begin{cases} 0 & t<0 \\ \sqrt{6}(-2\exp(-t/T)+2-t/T) & 0\le t<T \\ \sqrt{6}\exp(-t/T)(\exp 1-2) & T\le t \end{cases} \end{aligned} \tag{8.20}$$

Once again, care must be taken to identify the proper limits of integration as t varies. Sketches, such as the those shown in Figures 8.9 and 8.10, are generally very helpful. Our answer, $y_{2,2}(t) = x_2(t) * h_2(t)$ is shown in the lower right hand corner of Figure 8.11. It is a smoothed

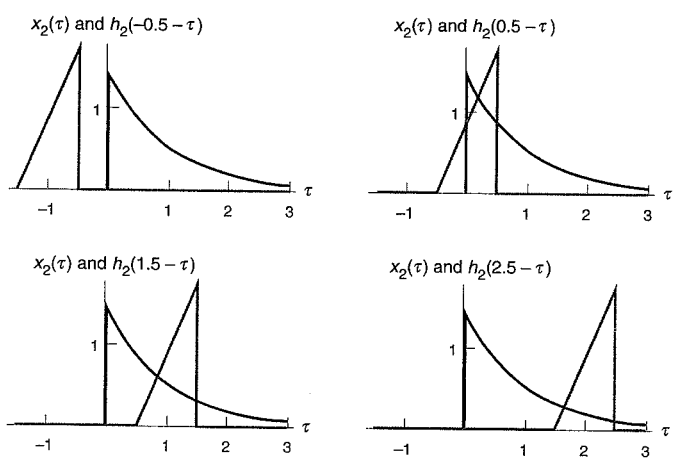


Figure 8.10 Sketches to aid the calculation of a convolution.

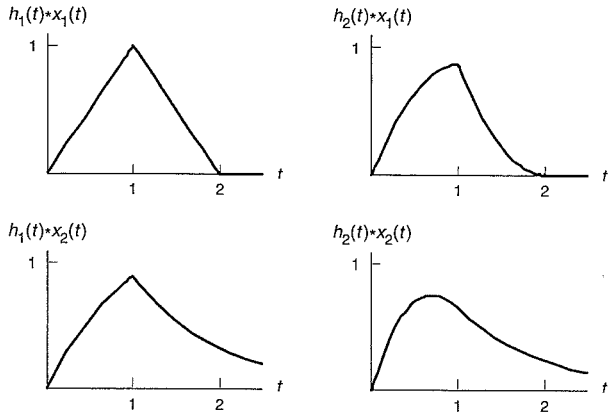


Figure 8.11 Responses to the inputs shown on the left side of Figure 8.8.

version of the input, as we might expect. Figure 8.11 also shows $x_1 * h_1$, $x_1 * h_2$, and $x_2 * h_1$. We ask you to validate these three additional results in the homework. In all cases, the output is a smoothed version of the input and the details have been attenuated.

8.3 Transfer Functions and Basis Functions

We now use convolution to find the response of a linear system to a sinusoid. We find that the response is simply another sinusoid with the same frequency, but possibly different amplitude and phase. This complex scaling is given by the *transfer function*, defined below, evaluated at the frequency of the input sinusoid. This result will draw us forward and we use sinusoids as basis functions in the remainder of the chapter. In other words, we will model any signal of interest to us as the weighted sums of sinusoids at different frequencies. The response will be the sum of the individual responses, where each individual response is simply scaled by the transfer function.

8.3.1 Response to a Single Imaginary Exponential

Consider the input $x(t) = A\exp(j2\pi f_0 t)$. Convolution finds the zero-state response as follows

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(\tau)x(t-\tau)\, d\tau \\ &= A\int_{-\infty}^{\infty} h(\tau)\exp(j2\pi f_0 t)\exp(-j2\pi f_0\tau)\, d\tau \\ &= A\exp(j2\pi f_0 t)\int_{-\infty}^{\infty} h(\tau)\exp(-j2\pi f_0\tau)\, d\tau \\ &= A\exp(j2\pi f_0 t)H(f_0) \end{aligned} \tag{8.21}$$

The output is simply a scaled version of the input, with the scaling factor equal to $H(f_0)$, where f_0 is the frequency of the input. $H(f)$ is called the transfer function because it transfers knowl-

edge of the input to knowledge of the output. The transfer function is a complex function of f, and it is the *Fourier transform* of the impulse response.

$$H(f)=\int_{-\infty}^{\infty}h(\tau)\exp(-2\pi f\tau)\,d\tau \tag{8.22}$$

We shall have much more to say about Fourier transforms in Section 8.5. For the time being, we simply wish to set the stage with some preliminary results.

First, the transfer function is a complex function of frequency, f. As such, it may be expressed in rectangular or polar form.

$$\begin{aligned}H(f)&=\mathrm{Re}\{H(f)\}+j\,\mathrm{Im}\{H(f)\}\\&=|H(f)|\exp(j\angle H(f))\\|H(f)|&=\sqrt{(\mathrm{Re}\{H(f)\})^2+(\mathrm{Im}\{H(f)\})^2}\\&=\sqrt{H(f)H^*(f)}\\H^*(f)&=\mathrm{Re}\{H(f)\}-j\,\mathrm{Im}\{H(f)\}\\\angle H(f)&=\tan^{-1}\left(\frac{\mathrm{Im}\{H(f)\}}{\mathrm{Re}\{H(f)\}}\right)\end{aligned} \tag{8.23}$$

$|H(f)|$ and $\angle H(f)$ are the amplitude and phase of the transfer function, respectively.

Second, transfer functions of physically realizable systems have some nice symmetry properties. The amplitude is even, $|H(-f)|=|H(f)|$, and the phase is odd, $\angle H(-f)=-\angle H(f)$. These symmetries follow because the impulse response, $h(t)$, of a realizable system must be real, and so we write

$$\begin{aligned}H(-f)&=\int_{-\infty}^{\infty}h(t)\exp(j2\pi ft)\,dt\\&=H^*(f)\\|H(-f)|&=\sqrt{H(-f)H^*(-f)}\\&=\sqrt{H^*(f)H(f)}\\&=|H(f)|\end{aligned} \tag{8.24}$$

For the angle, we proceed as follows.

$$\begin{aligned}\angle H(-f)&=\tan^{-1}\left(\frac{\mathrm{Im}\{H(-f)\}}{\mathrm{Re}\{H(-f)\}}\right)\\&=\tan^{-1}\left(\frac{-\mathrm{Im}\{H(f)\}}{\mathrm{Re}\{H(f)\}}\right)\\&=-\angle H(f)\end{aligned} \tag{8.25}$$

Third, the transfer function has a very nice connection to the original differential equation (8.1). Since

$$A\exp(2\pi f_0t)\xrightarrow[H]{}AH(f_0)\exp(2\pi f_0t)$$

we may write

$$\begin{aligned}&\left(b_N\frac{d^N}{dt^N}AH(f_0)\exp(2\pi f_0t)+\cdots b_0AH(f_0)\exp(2\pi f_0t)\right)\\&\qquad\left(a_M\frac{d^M}{dt^M}A\exp(2\pi f_0t)+\cdots a_0A\exp(2\pi f_0t)\right)\\&AH(f_0)\exp(2\pi f_0t)\left(b_N(2\pi f_0)^N+\cdots+b_0\right)=\\&\qquad A\exp(2\pi f_0t)\left(a_M(2\pi f_0)^M+\cdots+a_0\right)\end{aligned}$$

This implies that

$$H(f)=\frac{\left(a_M(j2\pi f)^M+\cdots+a_0\right)}{\left(b_N(j2\pi f)^N+\cdots+b_0\right)}$$

As shown, the transfer function can be determined from the differential equation describing the system. If that knowledge is not available from theory, then it can be readily measured in the laboratory.

8.3.2 Response to a Single Cosine Wave

Consider the input, $x(t)=A\cos(2\pi f_0t)$. Euler's great formula allows us to write the cosine wave as the sum of two imaginary exponentials.

$$\begin{aligned}x(t)&=A\cos(2\pi f_0t+\theta)\\&=\frac{A}{2}\left(\exp j(2\pi f_0t+\theta)+\exp-j(2\pi f_0t+\theta)\right)\end{aligned}$$

The response for each of these imaginary exponentials is known from the last subsection.

$$\begin{aligned}y(t)&=\frac{A}{2}\left(H(f_0)\exp j(2\pi f_0t+\theta)+H(-f_0)\exp-j(2\pi f_0t+\theta)\right)\\&=\frac{A}{2}\left(|H(f_0)|\exp(\angle H(f_0))\exp j(2\pi f_0t+\theta)+|H(-f_0)|\exp(\angle H(-f_0))\exp-j(2\pi f_0t+\theta)\right)\\&=\frac{A}{2}|H(f_0)|\left(\exp j(2\pi f_0t+\theta+\angle H(f_0))+\exp-j(2\pi f_0t+\theta+\angle H(f_0))\right)\\&=A|H(f_0)|\cos(2\pi f_0t+\theta+\angle H(f_0))\end{aligned}$$

So the zero-state response to a cosine wave is a cosine wave with scaled magnitude and shifted phase. The magnitude is scaled by the amplitude of the transfer function, $|H(f_0)|$, at the input frequency, f_0. The phase shift is the angle of the transfer function at the same frequency, $\angle H(f_0)$. In general, the zero-input response needs to be added to the zero-state response to account for non-zero initial conditions. However, if the system is stable, then the zero-input response will decay away and we will be left with a scaled and shifted version of our input.

8.3.3 Response to a Single Complex Exponential

If the input is $x(t) = A\exp(s_0 t)$, where $s_0 = \sigma_0 + j2\pi f_0$, then the results are just as sweet. We find

$$y(t) = AH_L(s_0)\exp(s_0 t) \tag{8.26}$$

where $H_L(s)$ is the *Laplace transform* of the impulse response.

$$\begin{aligned} H_L(s) &= \int_{-\infty}^{\infty} h(\tau)\exp(-s\tau)d\tau \\ &= \frac{\left(a_M s^M + \cdots a_0\right)}{\left(b_N s^N + \cdots b_0\right)} \end{aligned} \tag{8.27}$$

For most cases of interest, $H_L(j2\pi f) = H(f)$, but we will have more to say on this later. In Section 8.7, we will return to the subject of Laplace transforms in earnest.

8.3.4 Vector Representation of Signals

Few GPS signals are well represented as a single pure sinusoid. However, we plan to use these simple functions as building block functions for the GPS signals. In the language of mathematicians, we will use sinusoids as basis functions over the space of functions most interesting to us. In short, we plan to represent any $x(t)$ as a sum of sinusoids. Please know that sinusoids are not the only functions that find such service, so we will begin by discussing basis functions in general.

We seek a representation of the following sort.

$$x(t) = \sum_{n=-N}^{n=N} x_n \phi_n(t)$$

As shown, the basis functions, $\phi_n(t)$, are weighted by the coefficients, x_n. We would like these coefficients to be easy to find. Specifically, we want them to be independent, meaning that the value of x_n does not depend on x_k for $n \neq k$. This nifty condition is obtained when the basis functions are *orthogonal* over the time interval of interest.

$$\int_{t_0}^{t_1} \phi_n^*(t)\phi_k(t)dt = \begin{cases} 0 & n \neq k \\ \lambda_n & n = k \end{cases}$$

$\phi_n^*(t)$ is the complex conjugate of $\phi_n(t)$ and λ_n is real. In fact, λ_n is the energy in $\phi_n(t)$. If $\lambda_n = 1$ for all n, then the basis is *orthonormal*.

With an orthogonal basis, the coefficients, x_n, may be found as follows

$$\begin{aligned} \int_{t_0}^{t_1} \phi_n^*(t)x(t)dt &= \int_{t_0}^{t_1} \phi_n^*(t)\left(\sum_{k=-N}^{k=N} x_k\phi_k(t)\right)dt \\ &= \sum_{k=-N}^{k=N} x_k \int_{t_0}^{t_1} \phi_n^*(t)\phi_k(t)dt \\ &= x_n\lambda_n \end{aligned} \tag{8.28}$$

and so

$$x_n = \frac{1}{\lambda_n}\int_{t_0}^{t_1} \phi_n^*(t)x(t)dt$$

Our function, $x(t)$, can be fully represented by the coefficients $\{x_n\}_{n=-\infty}^{n=\infty}$ and can be reconstructed from these coefficients. These coefficients measure the correlation of $x(t)$ with a set of basis functions. Correlation integrates the product of the two functions and essentially measures the similarity between the two functions. If $x(t)$ is well described by $\phi_n(t)$, then $|x_n|$ will be large. If $x(t)$ has no similarity to such a waveform then the coefficient will be small.

Even if we use a small number of basis functions, correlation provides the best coefficients. More precisely, correlation provides the coefficients, x_n, that minimize the following measure of error:

$$MSE_N = \frac{1}{T_0}\int_{t_0}^{t_0+T_0}\left(x(t) - \sum_{-N}^{N} x_n\phi_n(t)\right)^2 dt$$

In other words, the mean square error (MSE) is minimized.

8.4 Fourier Series

8.4.1 Definition and Discussion

The Fourier basis is given by

$$\left\{\phi_n(t) = \exp(j2\pi nt/T_0)\right\}_{n=-\infty}^{\infty} \tag{8.29}$$

This basis is orthogonal over any interval $t_0 \le t \le t_0 + T_0$, and the Fourier series for $x(t)$ is

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{\infty} x_n \exp(j2\pi nt/T_0) \\ x_n &= \frac{1}{T_0}\int_{t_0}^{t_0+T_0} x(t)\exp(-j2\pi nt/T_0) \end{aligned} \tag{8.30}$$

This is the exponential Fourier series. Alternate forms of the Fourier series exist, but this is the only one we will need in this book. If $x(t)$ is aperiodic, then the Fourier series converges to $x(t)$ over the interval of integration, $t_0 \le t \le t_0 + T_0$. If $x(t)$ is periodic, like the functions in Figure 8.12, and has period T_0, then the Fourier representation is valid for all t.

We hasten to mention that not all $x(t)$ have satisfactory Fourier representations. The reader is referred to McGillem and Cooper (1984) for a nice discussion of the Dirichlet conditions. All of the $x(t)$ of interest to us will have satisfactory Fourier series.

The Fourier series is handy for two reasons. First, if $x(t)$ is an input to an LTI system, then the zero-state response of that system is

$$y(t) = \sum_{n=-\infty}^{\infty} x_n H(j2\pi nf_0 t)\exp(j2\pi nf_0 t)$$

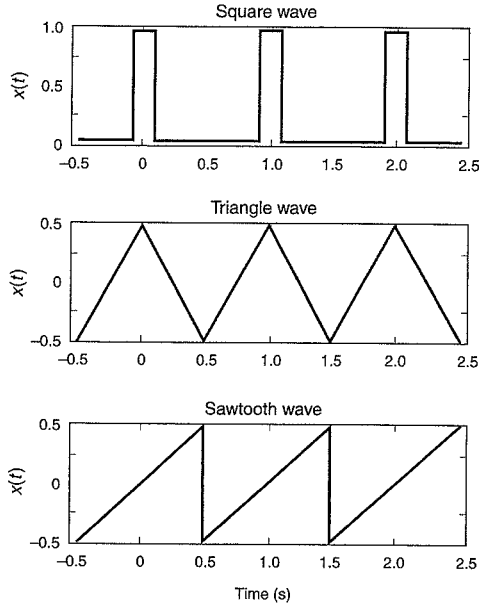


Figure 8.12 Examples of periodic waveforms: the square wave, the triangle wave, and the sawtooth wave.

This result is powerful because $H(j2\pi nf_0 t)$ can be readily determined from measurements or the differential equation that describes the system.

Second, the coefficients, x_n, are also easy to find. We need only correlate $x(t)$ with the member functions from the basis.

$$x_n = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x(t)\exp(-j2\pi nf_0 t)\,dt$$

With the Fourier expansion, $x(t)$ is represented with an infinite sum of sinusoids. This sum will eventually converge to $x(t)$ at all points where $x(t)$ is continuous. If $x(t)$ is discontinuous at t_0, then the series will converge to $(x(t_0^-)+x(t_0^+))/2$.

The Fourier coefficient, x_n, measures the correlation of $x(t)$ with the complex exponential $\exp(j2\pi nf_0 t)$. Correlation integrates the product of the two functions and essentially measures the similarity between the two functions. If $x(t)$ is real and well described by $\cos(2\pi nf_0 t)$, then the real part of x_n will be large. If a real $x(t)$ is well described by a sine wave, $\sin(2\pi nf_0 t)$, then the imaginary part of x_n will be large. If $x(t)$ has no similarity to either waveform then both the real and imaginary part of x_n will be small.

8.4.2 Example: Square Wave or Sampling Waveform

The shift from time to Fourier coefficients is illustrated by the square wave shown at the top of Figure 8.12. In the time domain, we have a periodic train of rectangular pulses. Each individual pulse has duration T_C seconds, and the period is $T_0 = 1/f_0$ seconds. We seek the Fourier series for this waveform.

$$\begin{aligned}
x_n &= \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x(t)\exp(-j2\pi nf_0 t)\,dt \\
&= \frac{1}{T_0}\int_{-T_C/2}^{T_C/2} \cos(2\pi nf_0 t) - j\sin(2\pi nf_0 t)\,dt \\
&= \frac{1}{T_0}\int_{-T_C/2}^{T_C/2} \cos(2\pi nf_0 t)\,dt \\
&= \frac{1}{T_0}\frac{\sin(2\pi nf_0 t)}{2\pi nf_0}\Bigg|_{-T_C/2}^{T_C/2} \\
&= \frac{T_C}{T_0}\frac{\sin(\pi nT_C/T_0)}{\pi nT_C/T_0} \\
&= \frac{T_C}{T_0}\operatorname{sinc}(\pi nT_C/T_0)
\end{aligned} \tag{8.31}$$

Among other things, this equation defines the sinc function, which we will encounter many times in Parts III and IV.

$$\operatorname{sinc}(x) \triangleq \frac{\sin x}{x}$$

With these coefficients, we can write the Fourier series.

$$\begin{aligned}
x(t) &= \sum_{n=-\infty}^{\infty} x_n \exp(j2\pi nt/T_0) \\
&= \sum_{n=-\infty}^{\infty} x_n \cos(2\pi nt/T_0) \\
&= \frac{T_C}{T_0} + \sum_{n=1}^{\infty} 2x_n \cos(2\pi nt/T_0)
\end{aligned} \tag{8.32}$$

These equations follow since $x_n = x_{-n}$, $\sin 0 = 0$, and $\cos x = \cos(-x)$.

As shown, our example signal, $x(t)$, can be written as an infinite sum of cosine waves with frequency $nf_0 = n/T_0$. The sine waves are not necessary because $x(t)$ has even symmetry around the origin $[x(t) = x(-t)]$, whereas the sine waves have odd symmetry ($\sin -t = -\sin t$). The amplitude of the cosine coefficients are given by (8.31), and these coefficients are plotted in Figure 8.13. As shown, there is a significant term at zero frequency. This is due to the non-zero mean of $x(t)$. The nearby coefficients are also quite strong. Apparently, they contribute meaningfully to the reconstruction of $x(t)$. Three partial reconstructions are shown in Figure 8.14. The $N = 1$ reconstruction uses only x_0 and x_1; it only begins to approximate $x(t)$. The $N = 3$ reconstruction uses x_0, x_1, x_2 and x_3. As shown, our square wave is beginning to appear. The bottom half of Figure 8.14 uses the first twenty terms. By now our square wave is quite

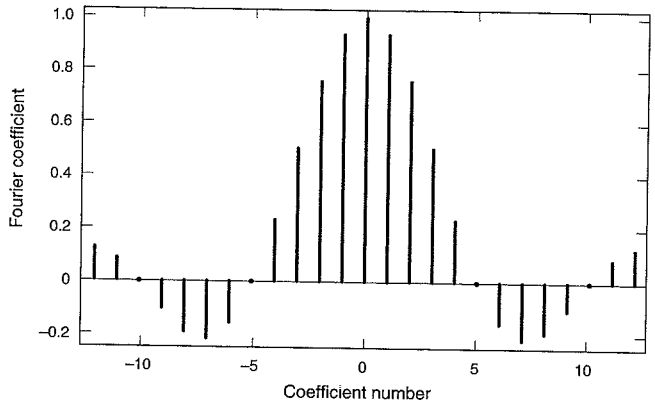


Figure 8.13 Fourier coefficients for the square wave shown in the top trace of Figure 8.12.

clear. The homework asks you to find the Fourier series for the triangle and sawtooth waveforms shown in Figure 8.12.

The square wave and its Fourier series will find application later in Chapter 11 when we discuss sampling. Indeed, the Fourier series is a perfect tool for the sampling application because the sampling waveform is our example square wave, $x(t)$. It is periodic. However, we will need other tools for waveforms that are not periodic. After all, the Fourier series only

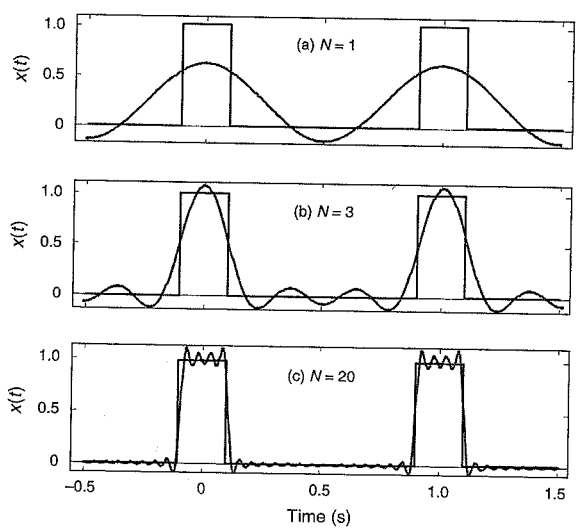


Figure 8.14 Square wave and Fourier approximations for $N = 1$, 3, and 20.

works well when we are reconstructing a waveform that is periodic, or when we are only concerned with the quality of the reconstruction over a fixed interval $t_0 \le t \le t_0 + T_0$. For non-periodic waveforms, we often turn to the Fourier and Laplace transforms.

8.4.3 Parseval's Theorem

The Fourier coefficients can be used to compute the power in the signal.

$$P = \frac{1}{T_0}\int_{t_0}^{t_0+T_0} x^2(t)\,dt$$

$$P = \frac{1}{T_0}\int_{t_0}^{t_0+T_0} x(t)\sum_{n=-\infty}^{\infty} x_n \exp(j2\pi nt/T_0)\,dt$$

$$= \sum_{n=-\infty}^{\infty} \frac{x_n}{T_0}\int_{t_0}^{t_0+T_0} x(t)\exp(j2\pi nt/T_0)\,dt \tag{8.33}$$

If $x(t)$ is real, then

$$\frac{1}{T_0}\int_{t_0}^{t_0+T_0} x(t)\exp(j2\pi nt/T_0)\,dt = x_n^*$$

and so

$$P = \sum_{n=-\infty}^{\infty} x_n x_n^*$$

$$= \sum_{n=-\infty}^{\infty} |x_n|^2 \tag{8.34}$$

So the total power can be found by summing the magnitude squared of all of the Fourier coefficients. Moreover, each term in this sum is the power in the nth component. The equation sums the power in each of the components that contribute to $x(t)$. Since the underlying basis functions are orthogonal, the sum of the component powers is equal to the total power in $x(t)$. This theorem can be articulated a number of different ways, but it is usually referred to as Parseval's theorem. Moreover, it is valid for any orthogonal basis—not just the Fourier series. We will discover a similar result for Fourier transforms.

8.5 Fourier Transforms

8.5.1 Derivation from Fourier Series

The Fourier transform provides a spectral representation for non-periodic signals. Even so, our derivation begins with a signal, $x(t)$, with period, T_0. The Fourier series for this signal is

$$x(t) = \sum_{n=-\infty}^{\infty}\left(\frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x(t)\exp(-j2\pi nf_0 t)\,dt\right)\exp(j2\pi nf_0 t) \tag{8.35}$$

Now let the period, T_0, tend toward infinity. The frequency spacing, f_0, becomes a differential,

df. The number of harmonics, n, grows without limit and nf_0 becomes a continuous variable, f. Our sum (8.35) becomes

$$x(t) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)\, dt \right) \exp(j2\pi ft)\, df \tag{8.36}$$

The inner integral is a function of f only because we have integrated over time. It is the Fourier transform of $x(t)$.

$$\begin{aligned} X(f) &= \mathcal{F}\{x(t)\} \\ &= \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)\, dt \end{aligned} \tag{8.37}$$

Recall that we first saw this equation in Section 8.3.1 when discussing transfer functions. If $h(t)$ is the impulse response of a linear time invariant system, then $H(f) = \mathcal{F}\{h(t)\}$ is the transfer function of that LTI system. The transfer function is all that is needed to determine the zero-state response to an imaginary exponential. In Section 8.5.6, we shall discover that it provides the response to a much greater variety of inputs.

The Fourier transform, $X(f)$ has the same properties we listed for the transfer function, $H(f)$, in Section 8.3.1. It can be expressed in polar or rectangular form.

$$\begin{aligned} X(f) &= \mathrm{Re}\{X(f)\} + j\mathrm{Im}\{X(f)\} \\ &= |X(f)| \exp(j\angle X(f)) \end{aligned} \tag{8.38}$$

$|X(f)|$ and $\angle X(f)$ are the amplitude and phase of $X(f)$, respectively. If $x(t)$ is real, then $|X(f)|$ and $\angle X(f)$ have even and odd symmetry, respectively.

Once again, we should be sure to mention that $X(f)$ does not always exist. If $x(t)$ is absolutely integrable, then $X(f)$ exists. In other words, the following condition is sufficient for the existence of $X(f)$.

$$\int_{-\infty}^{\infty} |x(t)|\, dt < \infty \tag{8.39}$$

This condition is sufficient, but not necessary. Many functions that are not absolutely integrable have Fourier transforms. However, these latter functions have Fourier transforms that include impulse functions, which are not always easy to work with. In these cases, the Laplace transform may be preferred.

Like the Fourier series, the Fourier transform measures the frequency content of $x(t)$ by correlating $x(t)$ with complex exponentials, $\exp(-j2\pi ft)$, for all values of f. Large values of $|X(f_0)|$ indicate that $x(t)$ is similar to an imaginary exponential with frequency f_0. If $x(t)$ is real and similar to $\cos(2\pi f_0 t)$, then $\mathrm{Re}\{X(f_0)\}$ will be large. If $x(t)$ is real and similar to $\sin(2\pi f_0 t)$, then $\mathrm{Im}\{X(f_0)\}$ will be large.

The inverse Fourier transform, $\mathcal{F}^{-1}\{x(t)\}$, allows $x(t)$ to be recovered completely and unambiguously from $X(f)$.

$$\begin{aligned} x(t) &= \mathcal{F}^{-1}\{X(f)\} \\ &= \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft)\, df \end{aligned} \tag{8.40}$$

8.5.2 Energy Spectrum

The energy spectrum is $|X(f)|^2$, and derives its name as follows.

$$\begin{aligned} \mathcal{E} &= \int_{-\infty}^{\infty} x^2(t)\, dt \\ &= \int_{-\infty}^{\infty} x(t) \left(\int_{-\infty}^{\infty} X(f) \exp(j2\pi ft) df \right) dt \\ &= \int_{-\infty}^{\infty} X(f) \left(\int_{-\infty}^{\infty} x(t) \exp(j2\pi ft) dt \right) df \\ &= \int_{-\infty}^{\infty} X(f) X(-f)\, df \end{aligned} \tag{8.41}$$

If $x(t)$ is real, then $X(-f) = X^*(f)$ and so

$$\mathcal{E} = \int_{-\infty}^{\infty} |X(f)|^2 df \tag{8.42}$$

This result is another form of Parseval's theorem, which tells us that the total energy in $x(t)$ can be found by integrating over either time or frequency. It also tells us that $|X(f)|^2$ measures the energy contained within every df of frequency. So the energy in the band $f_L < f < f_U$ is

$$\mathcal{E}_{L,U} = 2\int_{f_L}^{f_U} |X(f)|^2 df \tag{8.43}$$

This result holds for all real $x(t)$. Thus $|X(f)|^2$ is called the energy spectrum.

Frequently, engineers plot the energy spectrum using the decibel scale introduced in Chapter 2. This logarithmic treatment compresses the dynamic range of the spectrum and allows contemplation of very small and very large values of spectrum on one plot. The energy spectrum expressed in decibels is

$$|X(f)|^2_{dB} = 10\log_{10}|X(f)|^2 = 20\log_{10}|X(f)|$$

8.5.3 Fourier Transform Properties

A host of nice properties can be readily derived from (8.37) and (8.40). The most important of these are listed in Table 8.1.

Among these properties, linearity is particularly easy to prove. We write

$$\begin{aligned} \mathcal{F}\{k_1 x(t) + k_2 y(t)\} &= \int_{-\infty}^{\infty} (k_1 x(t) + k_2 y(t)) \exp(-j2\pi ft)\, dt \\ &= k_1 \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)\, dt \\ &\quad + k_2 \int_{-\infty}^{\infty} y(t) \exp(-j2\pi ft)\, dt \\ &= k_1 X(f) + k_2 Y(f) \end{aligned} \tag{8.44}$$

Table 8.1 Properties of Fourier transforms

Property/operation	*Time function*	*Fourier transform*
Linearity	$k_1x(t)+k_2y(t)$	$k_1X(f)+k_2Y(f)$
Time delay	$x(t-t_0)$	$X(f)\exp(-j2\pi ft_0)$
Time scaling	$x(kt)$	$\frac{1}{\lvert k\rvert}X\left(\frac{f}{k}\right)$
Differentiation	$\frac{dx}{dt}$	$j2\pi fX(f)$
n differentiations	$\frac{d^nx}{dt^n}$	$(j2\pi f)^nX(f)$
Symmetry	$\mathcal{F}\{x(t)\}=X(f)$	$\mathcal{F}\{X(t)\}=x(-f)$
Frequency translation	$x(t)\exp(j2\pi f_0t)$	$X(f-f_0)$
Modulation	$x(t)\cos(j2\pi f_0t)$	$\frac{1}{2}X(f-f_0)+\frac{1}{2}X(f+f_0)$
Modulation	$x(t)\sin(j2\pi f_0t)$	$\frac{1}{2j}X(f-f_0)-\frac{1}{2j}X(f+f_0)$
Convolution	$x(t)*h(t)$	$X(f)H(f)$
Multiplication	$x(t)\times h(t)$	$X(f)*H(f)$
Integration	$\int_{-\infty}^{t}x(\varphi)d\varphi$	$\frac{X(f)}{j2\pi f}+\frac{1}{2}X(0)\delta(f)$

The differentiation property may be proven as follows.

$$\frac{dx(t)}{dt}=\frac{d}{dt}\int_{-\infty}^{\infty}X(f)\exp(j2\pi ft)df$$
$$=\int_{-\infty}^{\infty}X(f)j2\pi f\exp(j2\pi ft)df$$

which means that

$$\mathcal{F}\left\{\frac{dx(t)}{dt}\right\}=j2\pi fX(f) \tag{8.45}$$

The symmetry property may be proven as follows.

$$x(t)=\int_{-\infty}^{\infty}X(f)\exp(j2\pi ft)df$$
$$x(-t)=\int_{-\infty}^{\infty}X(f)\exp(-j2\pi ft)df \tag{8.46}$$

We now replace the variable t with f to obtain

$$x(-f)=\int_{-\infty}^{\infty}X(t)\exp(-j2\pi ft)dt$$
$$=\mathcal{F}\{X(t)\} \tag{8.47}$$

The proof of the time scaling and time delay properties are homework problems. The frequency translation and modulation properties are developed in Section 8.5.5, while the convolution property is developed in Section 8.5.6. The proof of the integration property is relegated to the homework because it poses a nice challenge and is not needed for any of our immediate objectives.

8.5.4 Transforms of Key Functions

The unit pulse will be a particularly important signal in this book, so let us find its Fourier transform. We begin as follows

$$P(f)=\mathcal{F}\{p(t)\}$$
$$=\int_{-\infty}^{\infty}p(t)\exp(-j2\pi ft)\,dt$$
$$=\int_{-1/2}^{1/2}\cos(2\pi ft)\,dt$$
$$=\left.\frac{\sin 2\pi ft}{2\pi f}\right|_{-1/2}^{1/2}$$
$$=\mathrm{sinc}(\pi f)$$
$$\lvert P(f)\rvert=\lvert\mathrm{sinc}(\pi f)\rvert \tag{8.48}$$

The integral of the absolute value of the pulse waveform is finite, so we know from (8.39) that the Fourier transform exists.

If we scale and shift the pulse waveform, it is still absolutely integrable, so we know that its Fourier transform will still exist.

$$P(f)=\mathcal{F}\left\{Ap\left(\frac{t-\tau}{T_C}\right)\right\}$$
$$=A\int_{\tau-T_C/2}^{\tau+T_C/2}\exp(-j2\pi ft)dt \tag{8.49}$$

If we substitute $\beta=t-\tau$, then $t=\beta+\tau$ and

$$\mathcal{F}\left\{Ap\left(\frac{t-\tau}{T_C}\right)\right\} = A\int_{-T_C/2}^{T_C/2} \exp(-j2\pi f(\beta+\tau))\,d\beta$$

$$= AT_C \exp(-j2\pi f\tau)\frac{\sin \pi f T_C}{\pi f T_C}$$

$$= AT_C \exp(-j2\pi f\tau)\,\mathrm{sinc}(\pi f T_C)$$

$$= AT_C \exp(-j2\pi f\tau)\,P(fT_C)$$

$$\left|\mathcal{F}\left\{Ap\left(\frac{t-\tau}{T_C}\right)\right\}\right| = \left|AT_C P(fT_C)\right| \tag{8.50}$$

The amplitude spectra for two important pulses are shown in Figure 8.15. The top pulse has the same duration, 1 μs, as the chips of the C/A code, the civil pseudo-random noise (PRN) code used by GPS. The bottom pulse has the same duration, 0.1 μs, as the chips used within the P(Y) military signals. Narrower pulses have greater spectral width, typically referred to as greater bandwidth. We will return to this subject in Section 8.5.9.

Transform properties are used to build one transform from another. Let's begin with the unit impulse for some more examples. The Fourier transform of $\delta(t)$ or $\delta(t-t_0)$ can be found from the sifting property used to define the impulse function.

$$\mathcal{F}\{\delta(t)\} = \int_{-\infty}^{\infty} \delta(t)\exp(-j2\pi ft)\,dt = 1$$

$$\mathcal{F}\{\delta(t-t_0)\} = \int_{-\infty}^{\infty} \delta(t-t_0)\exp(-j2\pi ft)\,dt$$

$$= \exp(-j2\pi ft_0)$$

We now work forward by using the symmetry property that dictates

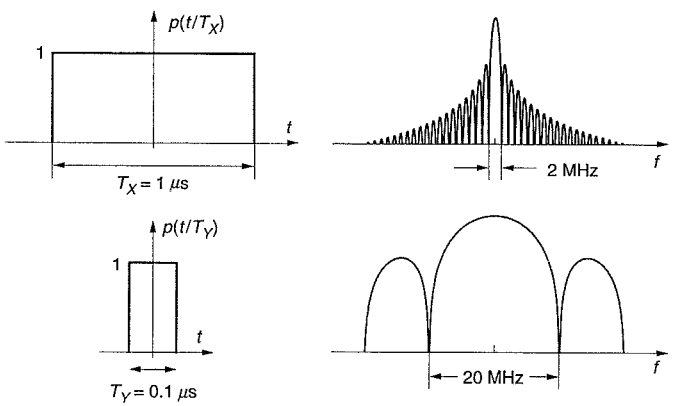


Figure 8.15 Pulses and their amplitude spectra.

$$\mathcal{F}\{1\} = \delta(f)$$

$$\mathcal{F}\{\exp(j2\pi f_0 t)\} = \delta(f-f_0)$$

From our last result, we can find two more allied transforms

$$\mathcal{F}\{\cos 2\pi f_0 t\} = \mathcal{F}\left\{\frac{1}{2}\exp(j2\pi f_0 t) + \frac{1}{2}\exp(-j2\pi f_0 t)\right\}$$

$$= \frac{1}{2}\delta(f-f_0) + \frac{1}{2}\delta(f+f_0)$$

$$\mathcal{F}\{\sin 2\pi f_0 t\} = \mathcal{F}\left\{\frac{1}{2j}\exp(j2\pi f_0 t) - \frac{1}{2j}\exp(-j2\pi f_0 t)\right\}$$

$$= \frac{1}{2j}\delta(f-f_0) - \frac{1}{2j}\delta(f+f_0)$$

These last four pairs are especially sensible. If the signal is simply a constant, then it correlates only with a sinusoid with zero frequency. We say that a constant signal has all of its power at zero frequency. A sinusoid with frequency f_0 correlates only with other such sinusoids and has no power at other frequencies. The complex exponential, $\exp(j2\pi f_0 t)$, has power only at the positive frequency f_0. However, the cosine, $\cos(2\pi f_0 t)$, and sine waves, $\sin(2\pi f_0 t)$, have power at $-f_0$ and $+f_0$. Euler's formula dictates that we need negative frequencies for the appropriate mathematical bookkeeping.

All of our Fourier transform pairs are summarized in Table 8.2. The table includes the signum function, the unit step, the ramp and the comb function, but we reserve the proofs for the homework problems.

8.5.5 Modulation

Most radio systems, including GPS, modulate a radio frequency (RF) carrier with an information bearing message signal. The carrier is the electromagnetic vehicle for carrying information from a transmitter to a receiver. A simple example of a modulated signal is

$$s(t) = x(t)\cos(2\pi f_0 t)$$

where $x(t)$ is the message, and the message spectrum $X(f)$ concentrates all of its significant energy near zero frequency. For this reason, radio engineers sometimes call $x(t)$ the *baseband signal*. As we shall see, the RF carrier, $\cos(2\pi f_0 t)$, shifts the message spectrum up to a radio frequency and creates a bandpass signal. Modulation places the message bearing signal in an appropriate portion of the radio spectrum. Other modulation schemes are possible, but many, including GPS, are reasonably well described by this simple multiplicative model.

The relevant Fourier transforms are given in Table 8.1. Of these, the frequency translation property can be proven as follows.

$$\mathcal{F}\{x(t)\exp(j2\pi f_0 t)\} = \int_{-\infty}^{\infty} x(t)\exp(j2\pi f_0 t)\exp(-j2\pi ft)\,dt$$

$$= \int_{-\infty}^{\infty} x(t)\exp(-j2\pi(f-f_0)t)\,dt$$

$$= X(f-f_0)$$

From this, we can write

Table 8.2 Fourier transform pairs

Time function	*Fourier transform*
Unit pulse $p(t)$	$\operatorname{sinc}(\pi f)$
Shifted and scaled pulse $Ap\left(\frac{t-\tau}{T_C}\right)$	$AT\exp(-j2\pi f\tau)\operatorname{sinc}(\pi fT)$
Unit impulse function $\delta(t)$	1
$\delta(t-t_0)$	$\exp(-j2\pi ft_0)$
1	$\delta(f)$
$\exp(j2\pi f_0t)$	$\delta(f-f_0)$
$\cos(2\pi f_0t)$	$\frac{1}{2}\delta(f-f_0)+\frac{1}{2}\delta(f+f_0)$
$\sin(2\pi f_0t)$	$\frac{1}{2j}\delta(f-f_0)-\frac{1}{2j}\delta(f+f_0)$
$\operatorname{sgn}t=\begin{cases}1 & t\geq 0\\ -1 & t<0\end{cases}$	$\frac{1}{j\pi f}$
Step function $u(t)$	$\frac{1}{2}\delta(f)+\frac{1}{j2\pi f}$
Ramp function $tu(t)$	$j\pi\delta'(f)-\frac{1}{(2\pi f)^2}$
Comb function $\sum_{i=-\infty}^{\infty}\delta(t-iT_s)$	$\frac{1}{T_s}\sum_{i=-\infty}^{\infty}\delta\left(f-\frac{i}{T_s}\right)$

$$\mathcal{F}\left\{x(t)\cos(2\pi f_0t)\right\}=\frac{1}{2}\int_{-\infty}^{\infty}x(t)\left(\exp(j2\pi f_0t)+\exp(-j2\pi f_0t)\right)\exp(-j2\pi ft)\,dt$$
$$=\frac{1}{2}X(f-f_0)+\frac{1}{2}X(f+f_0)$$

Modulating $x(t)$ with a cosine wave shifts half of the spectrum up to f_0 and the other half down to $-f_0$. Once again, negative frequencies are needed for mathematical bookkeeping.

Let's entertain a simple example. Suppose that the baseband signal is a single pulse, $x(t)=Ap(t/T_C)$. As we know from (8.50), the Fourier transform of this baseband signal is

$$\mathcal{F}\left\{Ap(t/T_C)\right\}=AT_C\operatorname{sinc}(\pi fT_C)$$

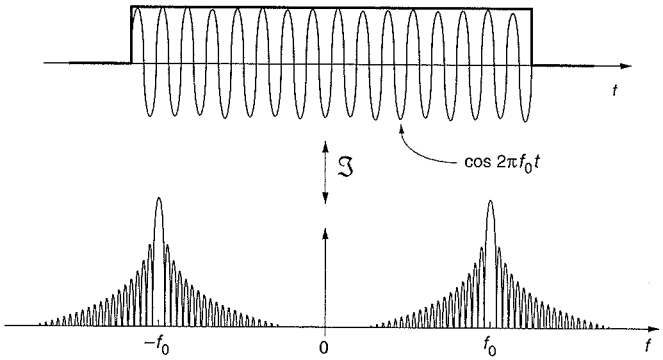


Figure 8.16 Impact of modulation on pulse spectrum.

The Fourier transform of the RF signal is

$$\mathcal{F}\left\{Ap(t/T_C)\cos(2\pi f_0t)\right\}=\frac{AT_C}{2}\left(\operatorname{sinc}(\pi(f-f_0)T_C)+\operatorname{sinc}(\pi(f+f_0)T_C)\right)$$

These results are shown in Figure 8.16. Even though these results derive from a simple model, Figure 8.16 is a decent approximation of the GPS spectrum. We will develop the details of the GPS signal and spectrum in Chapter 9.

8.5.6 Convolution Revisited

Recall our work on convolution. If we apply the signal, $x(t)$, to a linear system with impulse response, $h(t)$, then the output, $y(t)$, is the convolution, $x(t)*h(t)$. Table 8.1 contains an additional result of importance. If $\mathcal{F}\left\{x(t)\right\}=X(f)$ and $\mathcal{F}\left\{h(t)\right\}=H(f)$, then $\mathcal{F}\left\{x(t)*h(t)\right\}=X(f)H(f)$. This means

$$\begin{aligned}y(t)&=x(t)*h(t)\\&=\mathcal{F}^{-1}\left\{X(f)H(f)\right\}\end{aligned}\tag{8.51}$$

This remarkable theorem provides an alternate path to find the zero-state response of a linear system to a known input. It is so important that it deserves a proof. For this proof, we will use Fourier transforms, but bear in mind that the result also holds for Laplace transforms.

The proof follows.

$$\begin{aligned}y(t)&=\int_{-\infty}^{\infty}h(\beta)x(t-\beta)\,d\beta\\Y(f)&=\mathcal{F}\left\{\int_{-\infty}^{\infty}h(\beta)x(t-\beta)\,d\beta\right\}\\&=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}h(\beta)x(t-\beta)\,d\beta\right)\exp(-j2\pi ft)\,dt\\&=\int_{-\infty}^{\infty}h(\beta)\left(\int_{-\infty}^{\infty}x(t-\beta)\exp(-j2\pi ft)\,dt\right)d\beta\end{aligned}\tag{8.52}$$

We substitute $\gamma = t - \beta$ to obtain

$$\begin{aligned} Y(f) &= \int_{-\infty}^{\infty} h(\beta)\left[\int_{-\infty}^{\infty} x(\gamma)\exp(-j2\pi f(\gamma+\beta))\,d\gamma\right]d\beta \\ &= \int_{-\infty}^{\infty} h(\beta)\exp(-j2\pi f\beta)\,d\beta \int_{-\infty}^{\infty} x(\gamma)\exp(-j2\pi f\gamma)\,d\gamma \\ &= X(f)H(f) \end{aligned} \tag{8.53}$$

Recall our earlier result from Section 8.3.1.

$$A\exp(j2\pi f_0 t) \xrightarrow[H]{} AH(f_0)\exp(j2\pi f_0 t)$$

The transfer function determines how strongly the system will respond to a sinusoid with frequency f_0, and our most recent result generalizes this idea. The transform, $X(f)$, finds the energy content of any $x(t)$ as a function of frequency. The transfer function, $H(f)$, will reweight that energy content with the output containing the energy content of $x(t)$ reshaped by $H(f)$. In fact, the output energy spectra are related to the input energy spectra, $|X(f)|^2$, as follows.

$$|Y(f)|^2 = |X(f)|^2 |H(f)|^2 \tag{8.54}$$

This result enables our discussion of filters. The transfer functions for low-pass filters are contoured to allow low frequency signals to pass unfettered, but attenuate high frequencies. High-pass filters step on low frequencies and allow high frequency signals to pass. Bandpass filters attenuate all frequencies other than those in the selected pass band.

8.5.7 Ideal Filters

An ideal low-pass filter might have the spectrum shown as the dashed line in Figure 8.17. This filter passes all frequencies within $-1/T < f < 1/T$, and simply stops all frequencies outside of

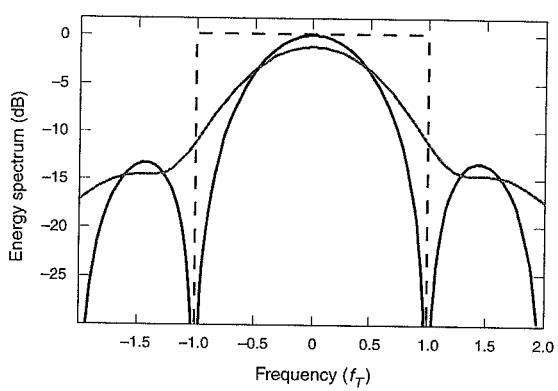


Figure 8.17 Spectrum of the ideal filter compared to spectra of moving averages. The dashed curve is the ideal filter, the black curve is for the unweighted boxcar average, and the gray curve is for an average with linear weighting.

this passband. However, such a filter is challenging to build. At a minimum, it would require us to appreciably delay the signal passing through the filter. The ideal filter in Figure 8.17 is specified in the frequency domain, so we may write

$$H_{\text{ideal}} = \begin{cases} 0 & f < -1/T \\ 1 & -1/T \le f < 1/T \\ 0 & f \ge 1/T \end{cases}$$

The corresponding impulse response is

$$\begin{aligned} h(t) &= \int_{-\infty}^{\infty} H(f)\exp(j2\pi ft)\,df \\ &= \int_{-1/T}^{1/T} \exp(j2\pi ft)\,df \\ &= \frac{\exp(j2\pi ft)}{j2\pi t}\Bigg|_{-1/T}^{1/T} \\ &= \frac{2}{T}\text{sinc}(2\pi t/T) \end{aligned} \tag{8.55}$$

This impulse response is non-zero for $t < 0$. It starts before the impulse is applied. We say such filters are non-causal. Even though it seems unlikely, approximations to non-causal filters can be implemented. However, the implementation requires us to store the input data in a buffer and process it after a delay. In this way, we are able to process data that arrived in the past. With greater delays, we can better approximate the ideal filter. For this book, we will avoid this complication and stick with causal filters like those described in the next subsection.

8.5.8 Moving Averages and Butterworth Filters

Recall our discussion of moving averages. These linear systems attenuate details and preserve average behavior. In other words, they pass low frequency components in $x(t)$ and attenuate high frequency components in $x(t)$. We should be able to see this in their transfer functions. The unweighted moving average with averaging time T is a linear system with the following impulse response, transfer function, and energy spectrum.

$$\begin{aligned} h_1(t) &= \frac{1}{T}p\left(\frac{t - T/2}{T}\right) \\ H_1(f) &= \frac{\sin(\pi Tf)}{\pi Tf}\exp(-j\pi fT) \\ |H_1(f)|^2 &= \text{sinc}^2(\pi Tf) \end{aligned} \tag{8.56}$$

This energy spectra, $|H_1(f)|^2$, is plotted as the black curve in Figure 8.17. Like the ideal filter, it passes frequency components near the origin and attenuates high frequencies.

Just for fun, let us also consider the moving average with triangular weighting. In this case, we have

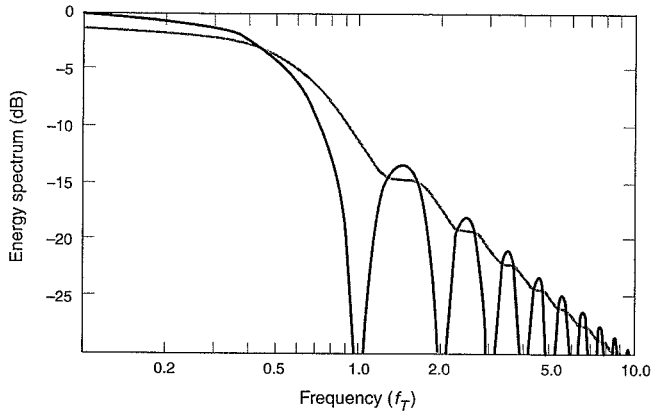


Figure 8.18 Spectrum of moving averages plotted versus the logarithm of frequency. The black curve is for the unweighted boxcar average and the gray curve is for an average with linear weighting. Both frequency responses roll off at 20 dB per decade change in frequency.

$$h_2(t) = \sqrt{3}\left(\frac{1}{T} - \frac{t}{T^2}\right) p\left(\frac{t - T/2}{T}\right)$$

$$H_2(f) = \sqrt{3}\int_0^T \left(\frac{1}{T} - \frac{t}{T^2}\right) \exp(-j2\pi ft)\, dt$$

$$= \frac{\sqrt{3}(1 - j2\pi fT - \exp(-j2\pi fT))}{4f^2\pi^2T^2} \qquad (8.57)$$

This energy spectra, $|H_2(f)|^2$, is plotted as the gray curve in Figure 8.17. Both averages are low-pass filters. They pass frequencies below $f = 0.2/T$ with little attenuation. They attenuate high frequencies.

Figure 8.18 provides another look at $|H_1(f)|^2$ and $|H_2(f)|^2$. It plots the energy spectra versus logarithmic frequency from $fT = 0.1$ to $fT = 10$. This alternate view is important because it allows us to quantify the ability of the filter to attenuate competing signals in more distant portions of the frequency spectrum. These signals may be much stronger than the desired signal. Indeed, GPS signals are very weak after traveling from medium earth orbit to the surface of the earth. They are much weaker than signals that originate terrestrially. In such cases, filters are needed to pass the desired signal and greatly attenuate competing signals.

As shown in Figure 8.18, the unweighted and weighted moving averages have rolloffs of 20 dB per decade of frequency. In other words, they have 20 dB more attenuation at $f = 10/T$ than they have at $f = 1/T$ and this trend continues for higher frequencies. A rolloff of 20 dB per decade is seldom satisfactory for real radio receivers.

Butterworth filters provide more rapid rolloffs. A low-pass Butterworth filter of order n has the following energy spectrum,

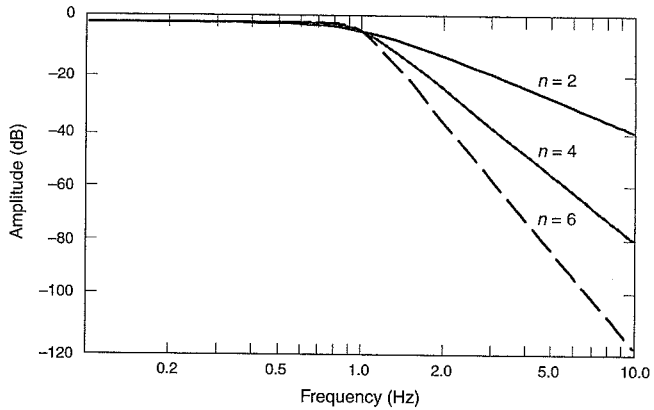


Figure 8.19 Butterworth low-pass filter of order 2, 4, and 6. These low-pass filters roll off at $20n$ dB per decade, where n is the order of the filter.

$$|H(f)|^2_{LP} = \frac{1}{1 + (f/\alpha)^{2n}} \qquad (8.58)$$

where α is a free design parameter. This characteristic is shown in Figure 8.19 for $n = 2, 4, 6$. The left hand scale has dramatically changed from Figure 8.18 for the moving averages, and the roll-off rates are $20n$ dB/decade. With increasing order, the spectrum approaches the ideal filter characteristic described in the last subsection. The rolloff rate can be derived if we approximate the transfer function for large values of frequency.

$$|H(f)|^2 \approx \frac{\alpha^{2n}}{f^{2n}} \qquad f >> \alpha$$

$$10\log_{10}\left[|H(f)|^2\right] \approx 20n\log_{10}\alpha - 20n\log_{10} f \qquad (8.59)$$

As we increase n, we increase the attenuation of signals outside of the pass band. However, we also increase the complexity of the filter. In earlier days, radio filters were discrete components and high order filters had more components. Increasing the parts count increased cost and also made filter tuning more difficult. Today, the most selective filters in a GPS receiver are realized with technologies that do not require tuning.

8.5.9 Bandwidth of Signals and Filters

Most signals and filters have bandwidth. The bandwidth of a signal measures the breadth of its spectral occupation. The definition of bandwidth is straightforward when $|X(f)|$ or $|H(f)|$ are non-zero only within a finite interval, say from $-f_0$ to f_0 or from $f_0 - f_1$ to $f_0 + f_1$. Sadly, this clarity only exists in idealized cases. With real-life signals, the energy spectrum rolls off with no clear boundary. In these cases, there are several ways to define bandwidth. For example,

the null-to-null bandwidth can be identified in Figure 8.15 for the rectangular chip waveform used by GPS or Figure 8.17 for the unweighted moving average. It simply measures the difference of the two null frequencies closest to zero frequency for the baseband case, or closest to the RF frequency, f_0, in the RF case. In either case, the null-to-null bandwidth is $B_{2,\text{null}} = 2/T$. This definition is certainly handy and evocative when the spectrum contains nulls, and we will use this definition often. The null-to-null bandwidth is said to be two-sided because it reaches from the lower null to the upper null.

Figures 8.17 and 8.18 show that the average with triangular weighting does not have these same sharp nulls. Figure 8.19 shows that Butterworth filters do not have sharp nulls either. In these cases, we cannot use null-to-null bandwidth to measure spectral selectivity. However, other reasonable measures of bandwidth exist. For example, bandwidth can be defined as the difference between the frequencies where the energy spectra drops to 1/2 of the maximum value. Such a bandwidth is called the –3 dB bandwidth because $10\log_{10}(1/2) = -3$.

For example, the –3 dB bandwidth for the Butterworth filters shown in Figure 8.19 is α. To prove this result, we define $B_{1,-3}$ as the frequency where $|H(f)|^2$ falls to 1/2 of its value at zero frequency. For our Butterworth filter, this means

$$\left|H(f = B_{1,-3})\right|^2_{LP} = \frac{1}{1+(B_{1,-3}/\alpha)^{2n}} = \frac{1}{2} \tag{8.60}$$

We find the following.

$$\begin{aligned}(B_{1,-3}/\alpha)^{2n} &= 1\\ B_{1,-3} &= \alpha\end{aligned} \tag{8.61}$$

In contrast to our null-to-null bandwidth, this bandwidth is one-sided. It measures bandwidth from zero frequency up to the 1/2 power point. One-sided bandwidths are the norm when considering low-pass filters. In Chapter 12, we will study the delay lock loops and phase lock loops that appear inside GPS receivers. These feedback systems contain low-pass filters, and are characterized by the one-sided bandwidths of these loops.

As their name suggests, bandpass filters pass a slice of frequency centered at f_0. They are used in the front end of GPS receivers. The earliest filters in the processing chain pass signals in the neighborhood of the GPS carrier frequency, $f_0 = f_{L1} = 1575.42 \times 10^6$ Hz. Two-sided bandwidth is the norm when considering bandpass filters. It measures bandwidth from the 1/2 power point below f_0 to the 1/2 power point above f_0.

From time to time we need a measure of bandwidth based on fractional signal energy. Consider a unit pulse of width T. In this case, an infinite bandwidth would be needed to contain 100% of the signal energy. However, the bandwidth that contains β% of the energy, $B_{1,\beta}$, could be found as follows.

$$\begin{aligned}\int_{-B_{1,\beta}}^{B_{1,\beta}} |P(f)|^2\,df &= \frac{\beta}{100}\int_{-\infty}^{\infty} |P(f)|^2\,df\\ &= \frac{\beta}{100}\int_{-\infty}^{\infty} p^2(t)dt \quad = \frac{\beta}{100}\int_{-T/2}^{T/2} 1\,dt\\ &= \frac{\beta T}{100}\end{aligned} \tag{8.62}$$

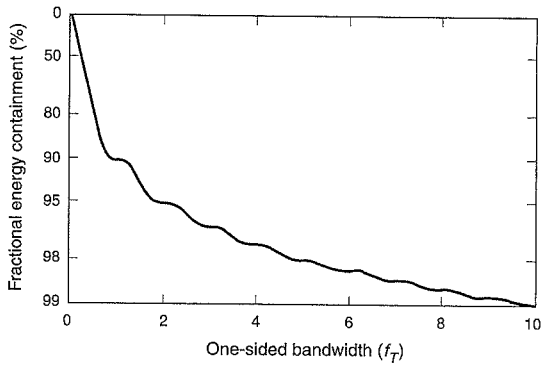


Figure 8.20 Fractional energy containment for a unit pulse.

Energy containment versus bandwidth is shown in Figure 8.20 for the unit pulse. A bandwidth of zero contains none of the signal energy. However, a bandwidth of $2/T$ contains approximately 95% of the energy, and a bandwidth of approximately $10/T$ contains 99% of the energy. We summarize as follows.

$$\begin{aligned}B_{1,95} &\approx 2/T\\ B_{1,99} &\approx 10/T\end{aligned} \tag{8.63}$$

This trend makes qualitative sense. As we open our frequency window, we capture more of the signal energy.

We can also define a two-sided bandwidth based on energy containment. The two-sided β% bandwidth would be found from the following.

$$\int_{-B_{2,\beta}/2}^{B_{2,\beta}/2} |P(f)|^2\,df = \frac{\beta T}{100} \tag{8.64}$$

For our unit pulse of width T, we find

$$\begin{aligned}B_{2,95} &= 2B_{1,95} = 4/T\\ B_{2,99} &= 2B_{1,99} \approx 20/T\end{aligned} \tag{8.65}$$

8.6 Random Signals

So far, we have concerned ourselves with deterministic signals, which can be described as explicit functions of time. When we input one of these signals to an LTI system, either convolution or the Fourier transform can provide the output (zero-state response). However, not all signals of interest to us can be described deterministically.

For example, the noise generated in the front end of the GPS receiver is due to thermal agitation of electrons in the circuits that are used to filter and amplify the received signals. It certainly cannot be predicted ahead of time. After filtering, this noise resembles the irregular

traces shown in Figure 8.21. This radio noise never repeats itself and cannot be represented by any simple function of time, $n(t)$. In fact, the actual noise trace depends on chance and so we write $n(t,\zeta)$, where ζ is the underlying random event.

As another example, the collection $n(t,\theta) = A\cos(2\pi ft + \theta)$ is a random (or stochastic) signal if θ is allowed to vary randomly across the interval from $[0, 2\pi)$. If θ can take an infinite number of values, then the collection has an infinite number of members. Three of these members are also shown in Figure 8.21. These waves resemble man-made tone interference. Tone interference can be troublesome to GPS, if the frequency, f, falls within the GPS portions of the radio spectrum and the amplitude is high compared to the GPS signal amplitude.

We cannot write a single $n(t)$ for either of our examples. Even so, the white noise traces in Figure 8.21 resemble each other, and the tone interference traces certainly share a structure. The job of this section is to introduce the tools that capture that structure and allow us to analyze random signals, which we shall simply call noise.

By the way, we also find some utility in treating the GPS signals as random signals. Specifically, we gain insight from modeling the GPS spread spectrum codes as totally random sequences. This idea is introduced in Chapter 9.

A frighteningly complete characterization of random signals would provide a probability density function that describes the possible values of $n(t,\zeta)$ at all times, $t_K > t_{K-1} > t_{K-2} > \cdots t_1$. For $K = 1$, we define $n_1 = n(t_1,\zeta)$ and write

$$f_{N_1}(n_1)\,dn_1 = Pr(n_1 - dn_1 < N_1 \le n_1) \tag{8.66}$$

For $K = 2$, it gets worse.

$$f_{N_1,N_2}(n_1,n_2)\,dn_1\,dn_2 = Pr(n_1 - dn_1 < N_1 \le n_1 \text{ and } n_2 - dn_2 < N_2 \le n_2) \tag{8.67}$$

For large K, we have a real challenge, and it is clear that we need something more compact.

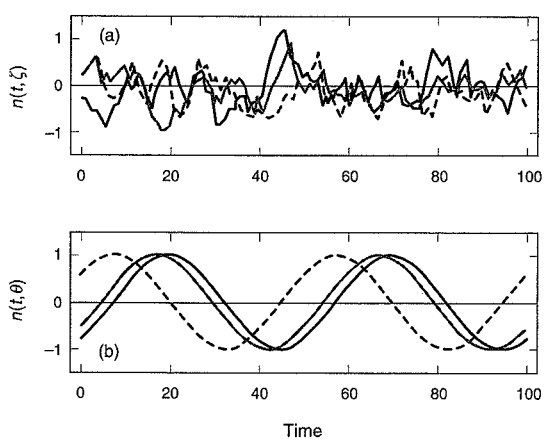


Figure 8.21 (a) Three traces of filtered white noise and (b) three traces of tone interference.

8.6.1 Moments

Fortunately, few GPS analyses need to consider the entire probability density function, and none of those appear in this book! All of our work will be well served by using a few noise moments. These moments are the mean, the variance, and the covariance as follows.

$$\begin{aligned}
m_N(t) &= E\{n(t,\zeta)\} = \overline{n(t,\zeta)} \\
\sigma_N^2(t) &= E\{(n(t,\zeta) - \overline{n(t,\zeta)})^2\} \\
&= E\{n^2(t,\zeta)\} - (\overline{n(t,\zeta)})^2 \\
\mu_N(t,t+\tau) &= E\{(n(t,\zeta) - \overline{n(t,\zeta)})(n(t+\tau,\zeta) - \overline{n(t+\tau,\zeta)})\} \\
&= E\{(n(t,\zeta)n(t+\tau,\zeta)\} - \overline{n(t,\zeta)n(t+\tau,\zeta)}
\end{aligned} \tag{8.68}$$

In all cases, expectation is with respect to the underlying random event, ζ, and so the mean, variance, and covariance no longer depend on ζ. Within these functions, $E\{n(t,\zeta)\}$ and $E\{n^2(t,\zeta)\}$ are known as the first and second moments. $E\{(n(t,\zeta)n(t+\tau,\zeta)\}$ is the auto-correlation function of the noise or $R_N(t_1,t_2)$.

$$\begin{aligned}
R_N(t_1,t_2) &= E\{(n(t,\zeta)n(t+\tau,\zeta)\} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} n_1 n_2 f_{N1,N2}(n_1,t_1;n_2,t_2)\,dn_1 dn_2
\end{aligned} \tag{8.69}$$

Fortunately, the noise processes that we encounter in this book are stationary. In other words, the noise statistics do not change as a function of time. Consequently, we may write

$$\begin{aligned}
m_N(t) &= m_N \\
\sigma_N^2(t) &= \sigma_N^2 = E\{(n(t,\zeta) - m_N)^2\} \\
\mu_N(t,t+\tau) &= \mu_N(\tau) \\
&= R_N(\tau) - m_N^2 \\
R_N(\tau) &= E\{n(t,\zeta)n(t+\tau,\zeta)\}
\end{aligned} \tag{8.70}$$

For our stationary processes, the mean and variance are constants. The covariance and auto-correlation function are only functions of the time between the two noise samples. They do not depend on the absolute time.

In addition, our noise processes are *ergodic*. This means that we may average over time to get the statistical averages shown above.

$$\begin{aligned}
m_N &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} n(t,\zeta)\,dt \\
\sigma_N^2 &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} (n(t,\zeta) - m_N)^2\,dt \\
\mu_N(\tau) &= R_N(\tau) - m_N^2 \\
R_N(\tau) &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} n(t,\zeta)n(t+\tau,\zeta)\,dt
\end{aligned} \tag{8.71}$$

For ergodic processes, all statistical averages may be replaced by time averages. This is key because it connects our noise characterization to measurements that can be made in the laboratory.

Now we seek a spectral characterization for random noise. After all, the Fourier transform has helped us understand and manipulate deterministic signals. For example, the energy spectrum of $x(t)$ is $|X(f)|^2$, and $|X(f_0)|^2\,df$ is the signal energy located within df hertz of f_0. In addition, if we input $x(t)$ to an LTI system with transfer function $H(f)$, then the zero-state output is $\mathcal{F}^{-1}\{X(f)H(f)\}$. We seek similar results for random signals.

Initially, we run into a snag. Since $n(t,\zeta)$ wanders on indefinitely, it has infinite energy, is not absolutely integrable, and has no Fourier transform. We circumvent this difficulty by defining a truncated version of $n(t,\zeta)$.

$$n_T(t,\zeta)=\begin{cases} n(t,\zeta) & |t|\le T/2 \\ 0 & \text{otherwise} \end{cases} \tag{8.72}$$

The Fourier transform of the truncated function is

$$N_T(f,\zeta)=\int_{-T/2}^{T/2} n(t,\zeta)\exp(-j2\pi ft)\,dt \tag{8.73}$$

Since $n(t,\zeta)$ is random, the corresponding energy spectrum, $|N_T(f,\zeta)|^2$, is also random. We divide this energy by time, T, such that the resulting quantity is a power spectrum. The resulting quantity, $|N_T(f,\zeta)|^2/T$, is known as a periodogram. However, we still have one periodogram for each sample realization of $n(t,\zeta)$. Hence, we take the expected value with respect to the underlying random variables. As T grows, the average periodogram is defined as the power spectral density.

$$S_N(f)=\lim_{T\to\infty}\frac{E\left\{|N_T(f,\zeta)|^2\right\}}{T} \tag{8.74}$$

This function gives the power spectral density of $n(t)$ as a function of frequency. Specifically, $|S_N(f_0)|df$ is the noise power located within df hertz of f_0.

The power spectral density also has a remarkable relationship to the autocorrelation function defined earlier.

$$S_N(f)=\mathcal{F}\{R_N(\tau)\}=\int_{-\infty}^{\infty} R_N(\tau)\exp(-j2\pi f\tau)d\tau \tag{8.75}$$

This relationship is called the Wiener-Khinchine relation and seems to make sense. After all, if the random signal is highly variable, then the expected correlation of $n(t,\zeta)$ and $n(t+\tau,\zeta)$ should be small. In other words, $E\{n(t,\zeta)n(t+\tau,\zeta)\}$ is a rapidly decreasing function of τ, and the autocorrelation function is peaked. A peaked autocorrelation function will lead to a broader power spectral density, $S_N(f)$, and power at high frequencies will be needed to rapidly change $n(t,\zeta)$. Elegant proofs of the Wiener-Khinchine relation are provided in Brown and Hwang (1992) and Ziemer and Tranter (1990). We provide examples in Sections 8.6.2 and 8.6.4.

8.6.2 Tone Interference

As an example, consider the random process that we sketched in Figure 8.21. Recall that this process is given by $n(t,\theta)=A\cos(2\pi ft+\theta)$ where θ will be distributed uniformly across $[0, 2\pi)$. In other words,

$$f_\Theta(\theta)=\begin{cases}\dfrac{1}{2\pi} & |\theta|\le\pi \\ 0 & \text{otherwise}\end{cases} \tag{8.76}$$

The first and second moments are

$$\begin{aligned} E\{n(t,\theta)\} &= \int_{-\infty}^{\infty} A\cos(2\pi ft+\theta)f_\Theta(\theta)\,d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} A\cos(2\pi ft+\theta)\,d\theta \\ &= 0 \\ E\{n^2(t,\theta)\} &= \int_{-\infty}^{\infty} A^2\cos^2(2\pi ft+\theta)f_\Theta(\theta)\,d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} A^2\cos^2(2\pi ft+\theta)\,d\theta \\ &= \frac{A^2}{2} \end{aligned} \tag{8.77}$$

For this process, we get the same answers if we average over time.

$$\begin{aligned} &\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} A\cos(2\pi ft+\theta)\,dt = 0 \\ &\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} A^2\cos^2(2\pi ft+\theta)\,dt = \frac{A^2}{2} \end{aligned} \tag{8.78}$$

Since the time averages give the same answers as the statistical averages, we suspect that our tone interference is an ergodic process. Even though we cannot prove ergodicity by simply examining the first two moments, our suspicion is correct.

Next, we seek the autocorrelation function for this process.

$$\begin{aligned} R_N(\tau) &= E\{n(t,\theta)n(t+\tau,\theta)\} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} A^2\cos(2\pi ft+\theta)\cos(2\pi f(t+\tau)+\theta)\,d\theta \\ &= \frac{A^2}{4\pi}\int_{-\pi}^{\pi}(\cos(2\pi f\tau)+\cos(2\pi f(2t+\tau)+2\theta))\,d\theta \\ &= \frac{A^2}{2}\cos(2\pi f\tau) \end{aligned} \tag{8.79}$$

An average over time would yield the same result.

Our tone interference example can also be used to illustrate the Wiener-Khinchine rela-

tionship. Since $n(t,\theta) = A\cos(2\pi f_0 t + \theta)$, we may write

$$\begin{aligned} n_T(t,\theta) &= Ap(t/T)\cos(2\pi f_0 t + \theta) \\ &= Ap(t/T)\cos\left(2\pi f_0\left(t + \frac{\theta}{2\pi f_0}\right)\right) \end{aligned} \tag{8.80}$$

Using the time delay property of Fourier transforms together with the Fourier transform for $\cos(2\pi f_0 t)$, we may write

$$\mathcal{F}\left\{\cos(2\pi f_0 t + \theta)\right\} = \frac{1}{2}\delta(f - f_0)\exp(j\theta) + \frac{1}{2}\delta(f + f_0)\exp(-j\theta) \tag{8.81}$$

We also recall that $\mathcal{F}\left\{Ap(t/T)\right\} = AT\,\mathrm{sinc}(\pi f T)$ and so we use the multiplication property of Fourier transforms as follows.

$$\begin{aligned} N_T(f,\theta) &= (AT\,\mathrm{sinc}(\pi f T)) * \left(\frac{1}{2}\delta(f - f_0)\exp(j\theta) + \frac{1}{2}\delta(f + f_0)\exp(-j\theta)\right) \\ &= \frac{AT}{2}\left(\exp(j\theta)\,\mathrm{sinc}(\pi(f - f_0)T) + \exp(-j\theta)\,\mathrm{sinc}(\pi(f + f_0)T)\right) \end{aligned} \tag{8.82}$$

The corresponding energy spectral density for this particular noise sample, $n(t,\theta)$, is given by

$$\begin{aligned} \left|N_T(f,\theta)\right|^2 = \left(\frac{AT}{2}\right)^2 &\left(\mathrm{sinc}^2(\pi(f - f_0)T) + \mathrm{sinc}^2(\pi(f + f_0)T)\right. \\ &\left. + 2\cos(2\theta)\,\mathrm{sinc}(\pi(f - f_0)T)\,\mathrm{sinc}(\pi(f + f_0)T)\right) \end{aligned} \tag{8.83}$$

We now average over the underlying random variable. In this case, that random variable is the phase, θ.

$$E_\theta\left\{\left|N_T(f,\theta)\right|^2\right\} = \left(\frac{AT}{2}\right)^2\left(\mathrm{sinc}^2(\pi(f - f_0)T) + \mathrm{sinc}^2(\pi(f + f_0)T)\right) \tag{8.84}$$

This last result follows because

$$E_\theta\{\cos(2\theta)\} = 0 \tag{8.85}$$

We can now write

$$S_N(f) = \lim_{T\to\infty}\left(\frac{A}{2}\right)^2 T\left(\mathrm{sinc}^2(\pi(f - f_0)T) + \mathrm{sinc}^2(\pi(f + f_0)T)\right) \tag{8.86}$$

Since $\delta(t) = \lim_{T\to\infty} T\,\mathrm{sinc}^2(\pi T t)$, we arrive at the following.

$$S_N(f) = \left(\frac{A}{2}\right)^2\left(\delta(f - f_0) + \delta(f + f_0)\right) \tag{8.87}$$

Finally, we can confirm that $R_N(\tau)$ from (8.79) is the inverse Fourier transform of this power spectral density.

8.6.3 Noise in Linear Systems

Recall that we have an additional goal: understand how noise propagates through linear systems. For deterministic signals we have

$$\begin{aligned} x(t) \xrightarrow[H]{} y(t) &= \mathcal{F}^{-1}\left\{X(f)H(f)\right\} \\ Y(f) &= H(f)X(f) \end{aligned} \tag{8.88}$$

If $x(t,\zeta)$ is a random signal input to an LTI system and $y(t,\zeta)$ is the corresponding output, then we certainly have the following.

$$\begin{aligned} x(t,\zeta) \xrightarrow[H]{} y(t,\zeta) &= \mathcal{F}^{-1}\left\{X(f,\zeta)H(f)\right\} \\ Y(f,\zeta) &= H(f)X(f,\zeta) \end{aligned} \tag{8.89}$$

However, we also have something much more useful.

$$S_Y(f) = \left|H(f)\right|^2 S_X(f) \tag{8.90}$$

$S_X(f)$ is the power spectral density of $x(t)$. This random signal is input to a filter with transfer function $H(f)$. $S_Y(f)$ is the power spectral density of the noise at the output of the linear system, which means that $\left|S_Y(f_0)\right|df = \left|H(f_0)\right|^2 S_X(f_0)\,df$ is the noise power located within df hertz of f_0. Reasonably, $S_Y(f)$ depends on the input power spectral density shaped by the frequency response of the linear system.

From $S_Y(f)$, we can determine the autocorrelation function of the output process.

$$\begin{aligned} R_Y(\tau) &= \mathcal{F}^{-1}\left\{S_Y(f)\right\} \\ &= \int_{-\infty}^{\infty}\left|H(f)\right|^2 S_X(f)\exp(j2\pi f\tau)\,df \end{aligned} \tag{8.91}$$

In addition, we can find the total power or variance of the output process.

$$\begin{aligned} R_Y(\tau = 0) &= E\{y^2(t)\} \\ &= \sigma_Y^2 \\ &= \int_{-\infty}^{\infty}\left|H(f)\right|^2 S_X(f)\,df \end{aligned} \tag{8.92}$$

This result is reasonable because it integrates over all possible values of $\left|H(f)\right|^2 S_X(f)\,df$ to determine the total power in $y(t)$.

In summary, these results show that the same transfer functions that served so well with deterministic signals are equally powerful when considering random signals. They give the power spectral density and autocorrelation function of the output given the autocorrelation function or power spectral density of the input. We leave the proof to Ziemer and Tranter (1990) or Pursley (2002) and focus on another application—white noise.

8.6.4 White Noise and Noise-Equivalent Bandwidth

White noise is the most important random signal in this book. White noise is so named, because it has equal power at all frequencies. Like white light, it sums energy from all colors of the rainbow, and so its power spectral density is a constant function of frequency.

$$|S_X(f)|^2 = \frac{N_0}{2} \text{ watts/hertz}$$

This power spectral density (PSD) gives the amount of noise power per unit hertz of spectrum. In one Hz of bandwidth, we would have $N_0/2$ watts of noise power.

If white noise passes through a filter with transfer function, $H(f)$, then the output noise will have the following PSD.

$$|S_Y(f)|^2 = \frac{N_0}{2}|H(f)|^2 \text{ watts/hertz}$$

If we use (8.92), then we can find the total power in the output noise.

$$\begin{aligned} P &= \sigma_Y^2 \\ &= \int_{-\infty}^{\infty} |S_Y(f)|^2 \, df \quad \text{watts} \\ &= \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2 \, df \quad \text{watts} \end{aligned} \tag{8.93}$$

For white noise, Parseval's theorem allows the following development.

$$\begin{aligned} P &= \sigma_Y^2 \\ &= \frac{N_0}{2}\int_{-\infty}^{\infty} h^2(t)\,dt \quad \text{watts} \end{aligned} \tag{8.94}$$

We will get good use from both (8.93) and (8.94) in the chapters to come.

Consider an ideal low-pass filter with a one-sided bandwidth equal to B_1. If this ideal filter has a zero frequency value of $|H(0)|$, then the output power is

$$\begin{aligned} P &= \frac{N_0}{2}\int_{-B_1}^{B_1} |H(f)|^2 \, df \quad \text{watts} \\ &= \frac{N_0}{2} 2B_1 |H(0)|^2 \quad \text{watts} \\ &= N_0 B_1 |H(0)|^2 \quad \text{watts} \end{aligned}$$

This is a neat result and sensible too. The noise power goes up with increasing power spectral density, $N_0/2$, bandwidth, B_1, and filter gain, $|H(0)|^2$.

However, not all filters are readily approximated as ideal filters. In this case, we define the noise-equivalent bandwidth as the bandwidth of the ideal filter that would pass the same noise power as the filter under study. If we denote this new bandwidth as $B_{1,n}$, then we may write

$$\begin{aligned} P_{\text{non-ideal}} &= \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2 \, df = N_0 B_{1,n} |H(0)|^2 \quad \text{watts} \\ B_{1,n} &= \frac{1}{2|H(0)|^2}\int_{-\infty}^{\infty} |H(f)|^2 \, df \quad \text{hertz} \end{aligned} \tag{8.95}$$

In Chapter 12, we will use one-sided, noise-equivalent bandwidths to characterize the phase lock loops and delay lock tracking loops in the GPS receiver.

8.7 Laplace Transforms

8.7.1 Definition and Discussion

From time to time Laplace transforms are more convenient than Fourier transforms. They have simpler transforms for some key waveforms including the unit step, ramp and parabola waveforms. The Fourier transforms for these waveforms do exist, but they include impulse functions that are sometimes awkward to work with. In addition, the Laplace transform more readily incorporates initial conditions when seeking the zero-input solution to linear differential equations. Once again, the Fourier transform can be used to solve for the zero-input solution as well as the zero-state solution, but it can be more painful.

The one-sided Laplace transform of $x(t)$ is given by

$$\begin{aligned} X_L(s) &= \mathcal{L}\{x(t)\} \\ &= \mathcal{F}\{x(t)\exp(-\sigma t)u(t)\} \\ &= \int_0^{\infty} x(t)\exp(-\sigma t)\exp(-j2\pi f t)\,dt \\ &= \int_0^{\infty} x(t)\exp(-(\sigma + j2\pi f)t)\,dt \\ &= \int_0^{\infty} x(t)\exp(-st)\,dt \end{aligned} \tag{8.96}$$

This definition makes sense for all values of s for which the integral converges. As shown, the Laplace transform can be thought of as the Fourier transform of $x(t)$ multiplied by $\exp(-\sigma t)$ and $u(t)$. Both of these new functions are introduced for a purpose.

First, multiplication by $\exp(-\sigma t)$ improves the convergence properties of the integral. We will choose σ large enough to guarantee that $x(t)\exp(-\sigma t)$ is absolutely integrable. In other words,

$$\int_{t_0}^{t_0+t} |x(t)\exp(-\sigma t)|\,dt < \infty$$

for any t_0. If $X_L(s) = \mathcal{L}\{x(t)\}$ exists for some $s_0 = \sigma_0 + j2\pi f_0$, then it will exist for all s such that $\text{Re}\{s\} < \text{Re}\{s_0\}$. The greatest lower bound on $\text{Re}\{s_0\}$ is called the *abscissa of convergence* for $x(t)$.

The inclusion of $\exp(-\sigma t)$ tames many functions, $x(t)$, that are otherwise troublesome. For example, all polynomial functions, $x(t) = t^n$, have Laplace transforms. These polynomial func-

tions include the unit step, $u(t)$, ramp, $tu(t)$, and parabola, $t^2u(t)$, which are very important to the study of navigation systems. These functions have Fourier transforms as well, but the Fourier transforms include impulse functions in frequency. These latter functions are sometimes tricky to work with. However, even the Laplace transform cannot tame $\exp(t^2)$. Fortunately, we have no use for this last function in our study of GPS.

Second, multiplication by $u(t)$ gives the problem a starting point in time, namely $t = 0$. Control and circuit engineers frequently use the one-sided Laplace transform because many problems seek the system response to an input with a specified starting time. GPS engineers are happy for that capability, and we will employ the Laplace transform in Chapter 12 when we seek the step response to the phase lock loops used in many GPS receivers. We will also discover that the Laplace transform is able to incorporate initial conditions that may exist at that starting time. Any previous inputs are neatly captured in the initial conditions. We will explore this handy feature in Section 8.7.4, where we use Laplace transforms to solve linear differential equations. Please understand that the Fourier transform can also be used to solve linear differential equations, but the Laplace transform is simpler to use.

Like the Fourier series and the Fourier transform, the Laplace transform, $X_L(s) = \mathcal{L}\{x(t)\}$, correlates $x(t)$ with a set of basis functions, but the Laplace basis functions are complex exponentials not imaginary exponentials. As always, correlation measures the similarity between the two functions. If $x(t)$ is well described by a damped cosine wave (or sine wave) with complex frequency s_0, then either the real or imaginary part of the correlation, $X_L(s_0)$, will be large. If $x(t)$ has no similarity to such waveforms then $|X_L(s_0)|$ will be small.

Like the Fourier transform, the Laplace transform is a complex function. The Fourier transform maps f onto the set of complex numbers, while the Laplace transform maps complex frequency, $s = \sigma + j2\pi f$, onto the set of complex numbers. Consequently, the Laplace transform also takes polar or rectangular form.

The Laplace inversion integral is

$$x(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} X_L(s)\exp(st)\,ds \tag{8.97}$$

We provide this result only for the record because we will use tables to invert Laplace transforms in this book.

8.7.2 Key Properties and Transforms

Key Laplace transform properties are shown in Table 8.3, and a list of useful Laplace transforms is provided in Table 8.4. The proofs for most of the properties are similar to the proofs provided for the Fourier properties, so we do not include them. We do derive the transforms for the unit step, $u(t)$, ramp, $tu(t)$, and parabola, $t^2u(t)$, because these functions are so central to our work in Chapter 12. We leave the rest of the proofs for homework problems.

The Laplace transform for the step function is found as follows.

$$\begin{aligned} U_L(s) &= \int_0^\infty u(t)\exp(-st)\,dt \\ &= \int_0^\infty \exp(-st)\,dt = \frac{\exp(-st)}{-s}\Big|_0^\infty = \frac{1}{s} \end{aligned} \tag{8.98}$$

The transforms for the ramp and parabola are also reasonably straight forward.

$$\begin{aligned} \mathcal{L}\{tu(t)\} &= \int_0^\infty t\exp(-st)\,dt \\ &= -\frac{t\exp(-st)}{s} - \frac{\exp(-st)}{s^2}\Big|_0^\infty = \frac{1}{s^2} \\ \mathcal{L}\{t^2u(t)\} &= \int_0^\infty kt^2\exp(-st)\,dt \\ &= \int_0^\infty 2t\frac{\exp(-st)}{-s}\,dt + \frac{t^2\exp(-st)}{s}\Big|_0^\infty \\ &= \frac{2}{s^2}\frac{\exp(-st)}{-s}\Big|_0^\infty = \frac{2}{s^3} \end{aligned}$$

These equations make liberal use of integration by parts.

The Laplace transforms of the unit step, ramp, and parabola functions are simple polynomial functions and are very nice to work with. In contrast, the Fourier transforms of these functions contain delta functions and derivatives of delta functions. Some of the homework problems are designed to make this difference clear.

8.7.3 Example: Moving Averages Revisited

Laplace transforms in hand, we now return to the moving averages that we studied in Section 8.2.2. We will apply the ramp function, $tu(t)$, to the boxcar average, $h(t) = (u(t) - u(t-T))/T$. For fun, try to solve this problem using Fourier transforms. Compared to the Fourier solution, the Laplace approach is elegant in its compactness. Indeed, the problem is solved as follows.

$$\begin{aligned} X_L(s) &= \frac{1}{s^2} \\ H_L(s) &= \frac{1}{T}\left(\frac{1}{s} - \frac{1}{s}e^{-sT}\right) \\ Y_L(s) &= X(s)H(s) = \frac{1}{T}\left(\frac{1}{s^3} - \frac{1}{s^3}e^{-sT}\right) \end{aligned} \tag{8.99}$$

The inverse Laplace transforms are readily found using Tables 8.3 and 8.4.

$$\begin{aligned} y(t) &= \mathcal{L}^{-1}\{Y(s)\} \\ &= \frac{1}{T}\left(\frac{t^2}{2}u(t) - \frac{(t-T)^2}{2}u(t-T)\right) \end{aligned} \tag{8.100}$$

We can rewrite the final expression for $y(t)$ as follows

Table 8.3 Properties of Laplace transforms

Property/operation	*Time function*	*Laplace transform*
Linearity	$k_1x(t)+k_2y(t)$	$k_1X_L(s)+k_2Y_L(s)$
Time delay	$x(t-t_0)$	$X_L(s)\exp(-st_0)$
Time scaling	$x(kt)$	$\frac{1}{\lvert k\rvert}X_L\left(\frac{s}{k}\right)$
Differentiation	$\frac{dx}{dt}$	$sX_L(s)-x(0-)$
n differentiations	$\frac{d^nx}{dt^n}$	$s^nX_L(s)-s^{n-1}x(0-)$ $-s^{n-2}x^{(1)}(0-)-\cdots x^{(n-1)}(0-)$
Frequency translation	$x(t)\exp(s_0t)$	$X_L(s-s_0)$
Modulation	$x(t)\cos(j2\pi f_0t)$	$\frac{1}{2}X_L(s-s_0)+\frac{1}{2}X_L(s+s_0)$
Modulation	$x(t)\sin(j2\pi f_0t)$	$\frac{1}{2j}X_L(s-s_0)-\frac{1}{2j}X_L(s+s_0)$
Convolution	$x(t)*h(t)$	$X_L(s)\,H_L(s)$
Multiplication	$x(t)\times h(t)$	$\frac{1}{j2\pi}\int_{c-j\infty}^{c+j\infty}X_L(s-\lambda)H_L(\lambda)d\lambda$
Integration	$\int_0^t x(\varphi)d\varphi$	$\frac{X_L(s)}{s}$

$$y(t)=\begin{cases}0 & t<0\\ \dfrac{t^2}{2T} & 0<t<T\\ t-\dfrac{T}{2} & T<t\end{cases} \tag{8.101}$$

and we are done.

Table 8.4 Laplace transform pairs

Time Function	*Laplace Transform*
Unit impulse function $\delta(t)$	1
Unit step $u(t)$	$\frac{1}{s}$
Unit ramp $tu(t)$	$\frac{1}{s^2}$
Parabola $t^2u(t)$	$\frac{2}{s^3}$
Pulse $Ap\left(\frac{t}{T}\right)$	$\frac{A}{s}(1-\exp(-sT))$
Damped exponential $\exp(-kt)\,u(t)$	$\frac{1}{s+k}$
$\cos(at)u(t)$	$\frac{s}{s^2+a^2}$
$\sin(at)u(t)$	$\frac{a}{s^2+a^2}$
$(\exp(-at)-\exp(-bt))\,u(t)$	$\frac{b-a}{(s+a)(s+b)}$
$(1-at)\exp(-at)\,u(t)$	$\frac{s}{(s+a)^2}$
$(b\exp(-bt)-a\exp(-at))\,u(t)$	$\frac{(b-a)s}{(s+a)(s+b)}$
$\exp(-at)\cos(bt)u(t)$	$\frac{s+a}{(s+a)^2+b^2}$
$\exp(-at)\sin(bt)u(t)$	$\frac{b}{(s+a)^2+b^2}$
$1-\exp(-at)\left(\cos(bt)+\frac{a}{b}\sin(bt)\right)u(t)$	$\frac{a^2+b^2}{s((s+a)^2+b^2)}$

8.7.4 Solving Linear Differential Equations

Laplace transforms are adept at solving the linear differential equations that we introduced in Section 8.1.1. They provide the complete solution, not just the zero-state solution, because they sweep up any initial conditions automatically. This convenience stems from the following results for the Laplace transforms of $\dot{x}(t)$ and $\int_0^t x(\lambda)d\lambda$.

For $\mathcal{L}\{\dot{x}(t)\}$, we begin by recalling the definition of the Laplace transform.

$$\mathcal{L}\{x(t)\} = \int_0^\infty x(t)\exp(-st)dt$$

Hence

$$\mathcal{L}\{\dot{x}(t)\} = \int_0^\infty \dot{x}(t)\exp(-st)dt$$

We invoke integration by parts to yield

$$\begin{aligned}\mathcal{L}\{\dot{x}(t)\} &= x(t)\exp(-st)\Big|_0^\infty + s\int_0^\infty x(t)\exp(-st)dt \\ &= sX_L(s) - x(0)\end{aligned} \tag{8.102}$$

The last equation follows because $\lim_{t\to\infty} x(t)\exp(-st) = 0$ for $X(s)$ to exist. This result neatly includes any initial conditions, $x(0)$, and makes Laplace transforms particularly convenient for solving linear differential equations.

Higher order differentials provide similar results. Specifically,

$$\mathcal{L}\{\ddot{x}(t)\} = s^2X_L(s) - sx(0-) - \dot{x}(0-)$$

$$\mathcal{L}\{\dddot{x}(t)\} = s^3X_L(s) - s^2x(0-) - s\dot{x}(0-) - \ddot{x}(0-)$$

The Laplace transform of $\int_0^t x(\lambda)d\lambda$ is found as follows.

$$\begin{aligned}\mathcal{L}\left\{\int_0^t x(\lambda)d\lambda\right\} &= \mathcal{L}\left\{\int_0^\infty x(\lambda)u(t-\lambda)d\lambda\right\} \\ &= X_L(s)U_L(s) \\ &= \frac{X_L(s)}{s}\end{aligned} \tag{8.103}$$

Let's work an example to demonstrate the power of these results. Consider the following differential equation where $x(t)$ is the input and $y(t)$ is the output.

$$x(t) = \ddot{y}(t) + A\,\dot{y}(t) + y(t)$$

This example system is assumed to have initial conditions, $y(0-) = y_0$ and $\dot{y}(0-) = \dot{y}_0$.

This second-order differential equation describes a wealth of physical systems. If we choose $A = \sqrt{2}$, then it is a good model for the n = 2 Butterworth low-pass filter studied in Section 8.5.8. Recall that Section 8.5.8 provided the frequency response of this filter. We now wish to find the time response to a specified input and initial condition. To this end, we take the Laplace transform of the entire equation to find

$$X_L(s) = (s^2Y(s) - sy_0 - \dot{y}_0) + \sqrt{2}(sY(s) - y_0) + Y(s)$$

We can solve this equation for the output $Y(s)$.

$$Y_L(s) = \frac{X(s)}{s^2 + \sqrt{2}s + 1} + \frac{sy_0 + \dot{y}_0 + \sqrt{2}y_0}{s^2 + \sqrt{2}s + 1} \tag{8.104}$$

This expression is particularly neat because it separates the response due to the input $X(s)$ from the response due to initial conditions. The first term gives the zero-state response as $X_L(s)H_L(s)$, where

$$\begin{aligned}H_L(s) &= \frac{Y_L(s)}{X_L(s)} \\ &= \frac{1}{s^2 + \sqrt{2}s + 1}\end{aligned} \tag{8.105}$$

The transfer function provides the zero-state response because it is derived from convolution. If the initial conditions, y_0 and $\dot{y}_0$, are equal to zero, then it also provides the complete response. If the initial conditions are non-zero, then we need to include the second term with the zero-input response. If $x(t) = X_L(s) = 0$, then the second term provides the complete response.

Let's consider a step input, $x(t) = ku(t)$. In this case, $X_L(s) = k/s$ and

$$Y_L(s) = \frac{k}{s(s^2 + \sqrt{2}s + 1)} + \frac{sy_0 + \dot{y}_0 + \sqrt{2}y_0}{s^2 + \sqrt{2}s + 1} \tag{8.106}$$

Now we come to an interesting juncture in this problem. We seek $\mathcal{L}^{-1}\{Y_L(s)\}$. To make this easy, we will break our expression for $Y_L(s)$ into a sum, where each term in the sum can be found in our table of canned Laplace transforms. We can express $Y(s)$ as a weighted sum of the following terms.

$$\mathcal{L}^{-1}\left\{\frac{1}{s}\right\} = u(t)$$

$$\mathcal{L}^{-1}\left\{\frac{s+a}{(s+a)^2 + b^2}\right\} = \exp(-at)\cos(bt)u(t)$$

$$\mathcal{L}^{-1}\left\{\frac{b}{(s+a)^2 + b^2}\right\} = \exp(-at)\sin(bt)u(t)$$

These inverse transforms have been taken from Table 8.4.

$\mathcal{L}^{-1}\{Y_L(s)\}$ will be the sum of the individual inverse transforms. However, the challenge is to find the coefficients associated with each of these terms. The partial fraction expansion is

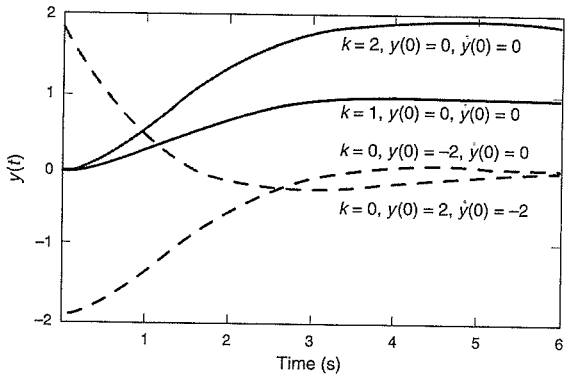


Figure 8.22 Zero-input and zero-state responses of a second-order Butterworth low-pass filter.

the traditional tool for finding the appropriate coefficients. This tool is elegantly described in Franklin, Powell and Emami-Naeni (2001) or McGillem and Cooper (1984), and we will only provide the results for our $Y_L(s)$.

$$Y_L(s) = k\left(\frac{1}{s} - \frac{s+\frac{1}{\sqrt{2}}}{s^2+\sqrt{2}s+1} - \frac{\frac{1}{\sqrt{2}}}{s^2+\sqrt{2}s+1}\right) + \frac{y_0\left(s+\frac{1}{\sqrt{2}}\right)}{s^2+\sqrt{2}s+1} + \frac{\frac{1}{\sqrt{2}}\left(\sqrt{2}\,\dot{y}_0+y_0\right)}{s^2+\sqrt{2}s+1} \tag{8.107}$$

We invite the reader to verify that this equation agrees with (8.106).

Now, we can find $y(t)$.

$$y(t) = \left(k-(y_0+k)\exp(-t/\sqrt{2})\cos(t/\sqrt{2}) + (y_0+\sqrt{2}\,\dot{y}_0-k)\exp(-t/\sqrt{2})\sin(t/\sqrt{2})\right)u(t) \tag{8.108}$$

Some results are plotted in Figures 8.22 and 8.23. In Figure 8.22, the solid lines are zero-state responses for $k = 1$ and $k = 2$. They start obediently at $y_0 = 0$ and $\dot{y}_0 = 0$ and grow smoothly to final values of 1 and 2. The dashed lines have $k = 0$ and so they are zero-input responses. One of these zero-input responses starts with $y_0 = -2$ and $\dot{y}_0 = 0$, and the other starts with $y_0 = 2$ and $\dot{y}_0 = -2$. They follow different paths, but both die out with time. The zero-input response of any stable system eventually dies out. In Figure 8.23, both curves are complete responses. They have $k = 1$ and initial conditions as shown.

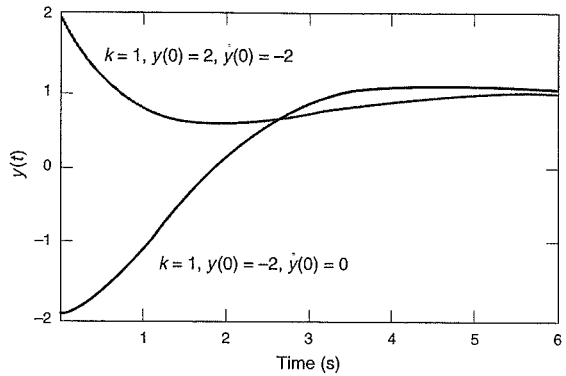


Figure 8.23 Complete responses of a second-order Butterworth low-pass filter.

8.7.5 Characteristic Equation

In the example above, $H_L(s)$ was found to have a denominator polynomial equal to $s^2+\sqrt{2}s+1$. Moreover, this same polynomial appeared in the denominator of the second term for $Y(s)$ that gave the zero-input response. This polynomial is called the characteristic equation of this system. The roots of the characteristic equation are called the poles of $H_L(s)$. This name is evocative. At a pole, the magnitude of $H_L(s)$ becomes very large (infinite) like a tent pole. In contrast, the roots of the numerator of $H_L(s)$ are called the zeroes.

For all the systems analyzed in this book, the characteristic equation can be written as the product of linear terms and quadratic terms. The linear terms may be written as $s + p_i$, where the ith pole in the system is located at $s = -p_i$. This pole must be real and the corresponding time function is $\exp(-p_i t)$. If $p_i > 0$: the poles are in the left hand plane of Figure 8.3, and the exponential decreases with time. If $p_i < 0$, then the poles are in the right hand plane, and the exponential increases with time. In this latter case, the system is unstable.

The characteristic equation may also contain quadratic terms with complex conjugate roots. For our recent example, the characteristic equation is a quadratic with roots at

$$s_{1,2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a} = -\frac{1}{\sqrt{2}} \pm \frac{j}{\sqrt{2}}$$

These roots are in the left hand plane. They have $\mathrm{Re}\{s_{1,2}\} < 0$ and the corresponding time function includes an exponential function that decreases with time. If these roots were in the right hand plane, then the exponential function would explode with time, and the system would be unstable.

If the poles are on the boundary between the left and right hand planes of Figure 8.3, then we have $s_{1,2} = \pm j2\pi f_0$. These poles fall on the imaginary axis. If the system input is a sinusoid with frequency f_0, then the output will become unbounded. However, if the input has no energy at f_0, then the output is bounded.

8.7.6 Connection Between the Laplace and Fourier Transforms

If all the poles of $X_L(s)$ are in the left hand plane, then $X(f) = X_L(j2\pi f)$, where $X(f)$ is the Fourier transform and $X_L(s)$ is the Laplace transform. Let us consider $x(t) = \exp(-\alpha t)u(t)$. The Laplace transform is given by

$$\begin{aligned} X_L(s) &= \mathcal{L}\{\exp(-\alpha t)u(t)\} \\ &= \int_0^\infty \exp(-\alpha t)\exp(-st)\,dt \quad = \frac{\exp(-(\alpha+s)t)}{-s}\Big|_0^\infty \\ &= \frac{1}{\alpha+s} \end{aligned}$$

The Fourier transform is given by

$$\begin{aligned} X(f) &= \mathcal{F}\{\exp(-\alpha t)u(t)\} \\ &= \int_0^\infty \exp(-\alpha t)\exp(-j2\pi ft)\,dt \quad = \frac{\exp(-(\alpha+j2\pi f)t)}{-(\alpha+j2\pi f)}\Big|_0^\infty \\ &= \frac{1}{\alpha+j2\pi f} \quad = X_L(j2\pi f) \end{aligned}$$

However, if $X_L(s)$ has poles on the $j2\pi f$ axis, then the story is not so simple. The Fourier transform exists, but includes impulse functions not needed by the Laplace transform. In this case, $X(f) \neq X_L(j2\pi f)$. A nice discussion of this case can be found in McGillem and Cooper (1984). Finally, if $X_L(s)$ has poles in the right half plane, then the Laplace transform may continue to exist, but the Fourier transform does not.

8.7.7 Initial and Final Value Theorems

The initial and final value theorems allow us to determine $x(t = 0^+)$ and $x(\infty)$ directly from $X_L(s)$.

$$x(0^+) = \lim_{s\to\infty} sX_L(s) \tag{8.109}$$

$$\lim_{t\to\infty} x(t) = \lim_{s\to 0} sX_L(s) \tag{8.110}$$

The final value theorem will see important use in our discussion of phase lock loops in Chapter 12, so we will take the time to prove it. We begin with (8.102)—the Laplace transform for $\dot{x}(t)$.

$$\begin{aligned} \mathcal{L}\{\dot{x}\} &= sX_L(s) - x(0) \\ \int_0^\infty \dot{x}\exp(-st)\,dt &= sX_L(s) - x(0) \\ \lim_{s\to 0}\int_0^\infty \dot{x}\exp(-st)\,dt &= \lim_{s\to 0}(sX_L(s) - x(0)) \end{aligned} \tag{8.111}$$

If $sX_L(s)$ has all of its singularities in the left half of the s plane, then $\lim_{s\to 0} sX_L(s)$ exists and we may write

$$\begin{aligned} \int_0^\infty \dot{x}\,dt &= \lim_{s\to 0}(sX_L(s) - x(0)) \\ \lim_{t\to\infty} x(t) - x(0) &= \lim_{s\to 0} sX_L(s) - x(0) \\ \lim_{t\to\infty} x(t) &= \lim_{s\to 0} sX_L(s) \end{aligned}$$

As mentioned earlier, the final value theorem will find key use in Chapter 12. For the time being, we will exercise it on some simple examples.

$$\begin{aligned} \lim_{t\to\infty} u(t) &= \lim_{s\to 0} s\frac{1}{s} = 1 \\ \lim_{t\to\infty} tu(t) &= \lim_{s\to 0} s\frac{1}{s^2} = \infty \\ \lim_{t\to\infty} t^2u(t) &= \lim_{s\to 0} s\frac{2}{s^3} = \infty \end{aligned} \tag{8.112}$$

Happily, these results agree with the correct answers for the step, ramp, and parabola. For fun, use the initial and final value theorems to validate the initial and final values of $y(t)$ for the Butterworth complete response derived in Section 8.7.4.

8.8 Summary

We have reviewed the fundamental theory of linear systems and signals in this chapter. We began by introducing the linear, time-invariant systems that lie underneath our entire study. We also provided the linear differential equation that connects the system input to the system output. The convolution integral was derived, and we found that the zero-state response of any LTI could be modeled as the input convolved with the system's impulse response. Much good came from this result. As an immediate example, we used convolution to study moving averages.

Sinusoidal signals were then introduced as basis functions. We found that a Fourier series can be used to describe non-sinusoidal, periodic signals as the weighted sum of sinusoids. More specifically, the Fourier basis functions are imaginary exponentials. Each component sinusoid can be separately propagated through any linear system. The response to the individual sinusoids can be determined using transfer functions, which are generally easy to measure or derive. The zero-state response of the linear system is then the sum of the responses to the component sinusoids.

Fourier transforms were derived from Fourier series but have increased scope. They can handle non-periodic signals like pulses as well as periodic signals. Fourier series describe periodic signals with imaginary exponential functions with the discrete frequencies of f_0, $2f_0$, $3f_0$, $4f_0$, ... , where $T_0 = 1/f_0$ is the period of the original signal. In contrast, Fourier transforms use a continuum of frequencies to describe non-periodic signals. These frequency domain descriptions are akin to the digital readout on your car radio. The radio seeks signal energy at the

commanded frequency. Some signals occupy rather narrow slices of frequencies, and others, like GPS, are intentionally spread out over large swaths of frequency.

Fourier transforms are also handy for characterizing random signals. This is a relief because GPS has no shortage of random signals. These include the natural noise found in the front end of all GPS receivers and man-made interference that can trouble GPS operations.

Laplace transforms are sometimes more convenient than Fourier transforms. They provide simple transforms for an important family of signals that have awkward Fourier transforms. These include the step, ramp, and parabolic functions that are very important to navigation system analysis.

All of these tools are important in the chapters to come. Chapter 9 uses convolution and Fourier transforms to detail the GPS signals. It also uses random codes to provide a first analysis of code performance. Chapter 10 uses white noise to quantify the ranging precision of GPS. Chapter 11 uses Fourier series to analyze the analog to digital converter (ADC) used by all GPS receivers and uses white noise to analyze signal acquisition. Chapter 12 uses Laplace transforms to model the delay lock loop (DLL) and phase lock loop (PLL). These devices continuously estimate the key signal parameters that become the pseudorange and carrier phase described earlier in this book. The DLL tracks the arrival time of the GPS pseudo random noise (PRN) code, and the PLL tracks the underlying carrier signal. Chapter 12 also analyzes the performance of these key tracking loops in the presence of white noise. It also uses random signal models for clock noise to fully understand the selection of the PLL bandwidth. Finally, Chapter 13 returns us to random signals. They are used to analyze error growth in inertial navigation systems.

Homework Problems

8-1. Find $x_1 * h_1$, $x_1 * h_2$, and $x_2 * h_1$, where these functions are defined in Figure 8.8. Compare your answers to the sketches given in Figure 8.11.

8-2. Prove that the steady state response of a linear system to a complex exponential is a scaled and shifted complex exponential. In other words, prove that

$$A\exp(s_0 t) \xrightarrow[H]{} AH(s_0)\exp(s_0 t)$$

8-3. Find the Fourier series representation for:
(a) the triangle wave shown in Figure 8.12.
(b) the sawtooth wave shown in Figure 8.12.

8-4. Prove the time scaling and time delay properties of Fourier transforms given in Table 8.1.

8-5. Prove the Fourier transform relationship for the comb function given in Table 8.2.

8-6. Any realizable low-pass characteristic can be converted to a high-pass filter that is also realizable. In (8.58) we replace

$$\frac{f}{\alpha} \to \frac{f_{-3}}{f}$$

where f_{-3} is the frequency at which the high-pass filter output is attenuated by 3 dB. Find the high-pass energy spectrum corresponding to the Butterworth low-pass filter described by (8.58). Plot these high-pass energy spectra for $n = 2, 4, 6$. Find the corresponding rolloff rates.

8-7. We can also build bandpass filters and band-stop filters. Based on (8.58), the required transformations are

$$\frac{f}{\alpha} \to \frac{f_0}{B_{2,-3}}\left(\frac{f}{f_0} - \frac{f_0}{f}\right)$$

$$\frac{f}{\alpha} \to \frac{B_{2,-3}}{f_0\left(\frac{f}{f_0} - \frac{f_0}{f}\right)} \qquad (8.113)$$

For these transformations, $B_{2,-3}$ are the two-sided, –3 dB bandwidths and we define

$$f_0 = \sqrt{f_1 f_2}$$

where f_1 and f_2 are the desired band edges. Find the bandpass and band-stop energy spectrum corresponding to the Butterworth low-pass filter described by (8.58). Plot these bandpass and band-stop energy spectra for $n = 2, 4, 6$. Find the corresponding rolloff rates.

8-8. Prove the initial value theorem given by (8.109).

8-9. Find the Fourier transform for
(a) the nth derivative of the delta function, $\mathcal{F}\left\{\frac{d^n\delta(t)}{dt^n}\right\}$.
(b) t^n.
(c) $|t|$.

8-10. Taken together, this problem and the next explore the treacheries of impulse functions. Consider the signum function.

$$\text{sgn}(t) = \begin{cases} -1 & t < 0 \\ +1 & t \geq 0 \end{cases}$$

(a) Find $\mathcal{F}\{\text{sgn}(t)\}$ using $\frac{d}{dt}\text{sgn}(t) = 2\delta(t)$ and the derivative property of Fourier transforms,

$$\mathcal{F}\{\dot{x}(t)\} = j2\pi f X(f)$$

(b) Find $\mathcal{F}\{\text{sgn}(t)\}$ based on the Fourier integral and integration by parts. Hint:

$$\mathcal{F}\{\text{sgn}(t)\} = \lim_{\alpha \to 0} \int_{-\infty}^{\infty} e^{-\alpha|t|}\,\text{sgn}(t)\exp(-j2\pi f t)\,dt$$

(c) Do your answers agree?

8-11. Consider the unit step

$$u(t)=\begin{cases}0 & t<0\\ 1 & t\geq 0\end{cases}$$

(a) Find $U(f)=\mathcal{F}\{u(t)\}$ using $\frac{d}{dt}u(t)=\delta(t)$ and the derivative property of Fourier transforms,

$$\mathcal{F}\{\dot{x}(t)\}=j2\pi fX(f)$$

(b) Find $U(f)$ using

$$u(t)=\frac{1}{2}+\frac{1}{2}\text{sgn}(t)$$

and the linearity property of Fourier transforms.

(c) Do your answers to parts *a* and *b* agree? Which answer do you trust and why?

8-12. Find

$$\mathcal{F}\left\{\int_{-\infty}^{t}x(\beta)d\beta\right\}$$

Hint:

$$\begin{aligned}y(t)&=\int_{-\infty}^{t}x(\beta)d\beta\\&=\int_{-\infty}^{\infty}x(\beta)u(t-\beta)d\beta\\&=x(t)*u(t)\end{aligned}$$

8-13. Find the Fourier transforms for
(a) the unit ramp function, $r(t)=tu(t)$.
(b) the unit parabola, $t^2u(t)$.

8-14. Use the Fourier transform to find the output of an unweighted moving average when:
(a) the input is the unit step.
(b) the input is the unit ramp.
Compare your solutions to the answers when Laplace transforms are used.

References

Baher, H. (1990). *Analog and Digital Signal Processing*, John Wiley.

Franklin G., J.D. Powell, and A. Emami-Naeini (1994), *Feedback Control of Dynamic Systems* (3rd edition), Addison-Wesley.

Gabel, R.A. and R.A. Roberts (1973). *Signals and Linear Systems* (1st edition), John Wiley.

McGillem, Clare and George Cooper (1984). *Continuous and Discrete Signal and System Analysis*, Holt, Rinehart and Winston.

Papoulis, A. (1962). *The Fourier Integral and Its Applications*, McGraw Hill.

Pursley, M. (2002). *Random Processes in Linear Systems*, Prentice Hall.

Ziemer, R.E. and W.H. Tranter (1995). *Principles of Communications: Systems, Modulation, and Noise* (4th edition), John Wiley.

Chapter 9

GPS Signals

GPS relies on spread spectrum signaling to achieve some truly remarkable capabilities. In general, spread spectrum signals occupy a large amount of radio spectrum relative to that required by a traditional radio system. To send 50 bits per second of data, most systems would require bandwidths of approximately 10 to 250 Hz. In contrast, GPS sends 50 bits per second of navigation data, but occupies over 2,000,000 Hz of spectrum. This bandwidth expansion is not careless or wasteful. It results from the introduction of a carefully designed structure into the signal. Specifically, GPS superposes a fast code onto each transmission. These codes, also known as sequences, are chosen for their auto- and cross-correlation properties. These properties enable precise ranging and facilitates the rejection of signal reflections. They also allow simultaneous use of the same transmission frequencies by all satellites, and help to reject unwanted signals from other radio sources (radio frequency interference or RFI).

Section 9.1 describes the civil signal broadcast by GPS on the L1 frequency. We begin with the civil signal in deference to its uncontested primacy with respect to applications. In 2005, this signal, by itself, supports the vast majority of GNSS applications worldwide. The military signals, discussed later, support many important applications, but the total number of military users is miniscule compared to the civil population. Section 9.1 provides time domain representations for the civil signal and approximates the amplitude spectrum of the GPS signal. Section 9.1 also describes the underlying radio carrier and reviews the use of binary phase shift keying (BPSK). Indeed, BPSK is used to modulate the carrier with the spread-spectrum code and the navigation data.

Sections 9.2 and 9.3 dig into the spread spectrum codes. The GPS codes serve so well because they have special auto-correlation and cross-correlation functions. Sections 9.2 and 9.3 define the auto-correlation and cross-correlation function, respectively, and discuss desired properties for these functions. We find that good properties for the auto-correlation function enable precise ranging and signal acquisition. Both of these are further analyzed in Chapters 10 and 11, respectively, but we establish a qualitative connection in this chapter. Favorable cross-correlation properties enable channel sharing. With uniformly low cross-correlation, the signals from the different GPS satellites can be distinguished by their codes alone. They do not need to use different transmission frequencies and they can also transmit simultaneously. In the lingo of the communications community, GPS (and Galileo) use code division multiple access (CDMA) rather than time division multiple access (TDMA) or frequency division multiple access (FDMA).

Sections 9.2 and 9.3 also introduce a very powerful tool for the analysis of spread spectrum ranging systems—random sequences. A random sequence can be thought of as a sequence of coin flips. The GPS codes are not generated by flipping coins. Even so, the GPS codes are sometimes called pseudorandom noise (PRN) sequences, and we find that this moniker is very appropriate.

Sections 9.4 and 9.5 turn our attention away from random sequences and toward the actual codes used by GPS. Section 9.4 introduces maximal length linear shift register sequences because these are used as building blocks for all GPS codes. Section 9.5 constructs the C/A-codes used for the GPS civil signals.

Section 9.6 deepens our study of random codes by computing the power spectral density, PSD, of GPS signals. Recall from Chapter 8 that the PSD is used to characterize the spectral content of random signals. In fact, Section 9.7 contains two analyses. The first analysis computes the PSD for signals that are modulated with long codes, which are well modeled as random. This analysis is on target for all GNSS codes except the C/A-codes. Second, Section 9.6 computes the PSD for the GPS civil signals with their C/A-codes. The C/A-codes are rather short and thus not perfectly modeled as random. For this reason, the second analysis models the C/A-codes as deterministic, but treats the navigation data bits as random.

Section 9.7 discusses another nice property of spread-spectrum, which is the ability to attenuate radio frequency interference (RFI). Radio frequency interference is a major heartache for GPS because the GPS signals come from space. After they have traveled from medium earth orbit to the user, they are weak compared to man-made signals generated on the surface of the earth. Chapter 13 analyzes techniques to harden GPS receivers against RFI, but Section 9.7 analyzes the intrinsic ability of the GPS signals to provide a processing gain against narrowband RFI.

The remainder of the chapter is a survey of the GNSS signals other than the GPS civil signal at L1. Specifically, military signals are broadcast by GPS on both the L1 and L2 frequencies. These are discussed in Section 9.8. The new civilian codes to be radiated by GPS on L2 and L5 are introduced in Section 9.9. Binary offset carrier (BOC) signals are introduced in Section 9.10, because these will be used by GPS starting in 2005 and Galileo in the near future. Finally, Section 9.11 is a brief summary that qualitatively compares the signals that we have described in this chapter.

By the way, this entire chapter makes good use of the signal theory fundamentals taught in Chapter 8.

9.1 Civil Signal on L1

9.1.1 Time Domain Description

In mid-2005, each GPS satellite broadcasted the following navigation signals

$$\begin{aligned} s_{L1}(t) &= \sqrt{2P_{C1}}D(t)x(t)\cos(2\pi f_{L1}t+\theta_{L1}) \\ &\quad +\sqrt{2P_{Y1}}D(t)y(t)\sin(2\pi f_{L1}t+\theta_{L1}) \\ s_{L2}(t) &= \sqrt{2P_{Y2}}D(t)y(t)\sin(2\pi f_{L2}t+\theta_{L2}) \end{aligned} \tag{9,1}$$

As shown, each satellite sends three rather similar signal components. The first, with amplitude $\sqrt{2P_{C1}}$, is the basis for the vast majority of civil applications and will be the object of our attention in this section. For compactness, we refer to it as the civil signal, even though the military also uses this signal. This signal is modulated with the C/A-code, $x(t)$. As shown, the second and third signals have amplitudes $\sqrt{2P_{Y1}}$ and $\sqrt{2P_{Y2}}$ and are modulated by the P(Y) codes, $y(t)$. The military is the main beneficiary of these latter signals, and so we refer to them as the military signals.

Any of the three signals in (9.1) is the product of four terms: an amplitude, $\sqrt{2P}$; the navigation data, $D(t)$; a spread spectrum code, $x(t)$ or $y(t)$; and the radio frequency (RF) carrier, $\cos(2\pi ft+\theta)$ or $\sin(2\pi ft+\theta)$. This hierarchy is shown in Figure 9.1.

As we discussed in Chapter 8, power measures the time rate of the energy carried by the signal and has units of watts or joules/second. Average power is given by the time average of

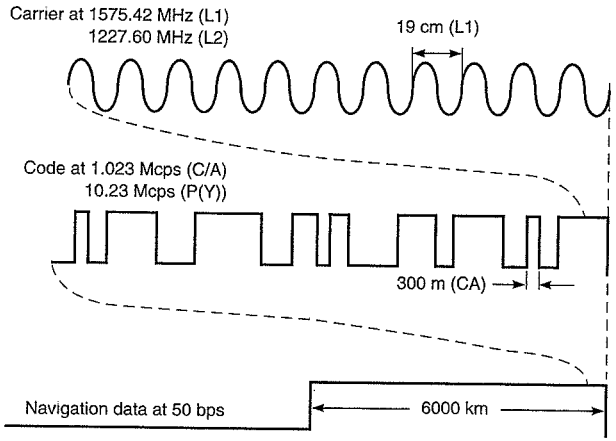


Figure 9.1 GPS signals including carrier, code, and navigation data.

the signal squared. If we average over T seconds with $T >> 1/f_{L1}$, then the average power is as follows

$$\begin{aligned}\frac{1}{T}\int_0^T s^2(t)dt &= \frac{1}{T}\int_0^T 2P_{C1}D^2(t)x^2(t)\cos^2(2\pi f_{L1}t+\theta_{L1})dt \\ &= \frac{1}{T}\int_0^T 2P_{C1}\cos^2(2\pi f_{L1}t+\theta_{L1})dt \\ &= P_{C1}\end{aligned} \tag{9.2}$$

This equation uses the civil signal as an example, but identical relationships hold for the military signals. In this chapter, we will use P to denote the average power in a generic signal. We will use P_{C1} when referring specifically to the civil GPS signal and P_{Y1} and P_{Y2} when referring to the power in the military GPS signals. In Chapter 10, we will begin to use C when referring to the received signal power after adding the power gain due to the receiving antenna and subtracting any receiver implementation losses.

Equation (9.2) leverages the fact that the navigation data, $D(t)$, and the codes, $x(t)$, are sequences of +1's and –1's, and so $D^2(t) = x^2(t) = 1$. The navigation message has been described in Chapter 4 and the codes will be described later in this chapter. In the next few paragraphs, we pay attention to the RF carriers.

In 2005, GPS uses three such carriers, and any one of these is shown at the top of Figure 9.1. The carrier is completely specified by its frequency, f, and phase, θ. Phase is discussed in Chapter 7, because it carries pseudorange information. It will be further discussed in Chapter 12, where we discuss how the receiver recovers the phase information.

Frequency is the number of sinusoidal cycles that occur in a second. Reasonably enough, the units of frequency are cycles per second or hertz (Hz). In the year 2005, the GPS navigation frequencies are

$$\begin{aligned}f_{L1} &= 1575.42\times10^6 \text{ hertz} = 1575.42 \text{ MegaHz} = 1575.42 \text{ MHz} \\ f_{L2} &= 1227.60\times10^6 \text{ hertz} = 1227.60 \text{ MegaHz} = 1227.60 \text{ MHz}\end{aligned} \tag{9.3}$$

As shown in Figure 9.2, these frequencies place GPS in the UHF (Ultra-High Frequency) portion of the radio spectrum. In contrast, FM (Frequency Modulation) broadcast radio sends its signals to our cars and homes at frequencies in the VHF (Very High Frequency) band. FM broadcast radio frequencies are more than ten times lower than those used by GPS. AM broadcast radio uses frequencies from 680×10^3 to 1260×10^3 Hz. These AM frequencies fall in the MF (Medium Frequency) band and are more than 1000 times lower than the GPS frequencies. In Chapter 10, we shall discover that the GPS frequencies were chosen, in part, to allow the use of small omni-directional antennas by the users.

The GPS signals all reside in radio frequency bands that are designated to include navigation signals from satellites. For example, the L1 signal at 1575.42 MHz resides in the middle of an Aviation Radio Navigation Satellite (ARNS) band that extends from 1559 to 1610 MHz. GPS is not alone in this band. The Russian system, GLONASS, resides in the upper portion of this band, and the European system, Galileo, also plans to use this band.

As shown in (9.1), two signals use the same f_{L1} carrier frequency. The civilian signal modulates $\cos(2\pi f_{L1}t+\theta_{L1})$ and one of the military signals modulates $\sin(2\pi f_{L1}t+\theta_{L1})$. These signals are said to be in phase quadrature because the cosine wave leads the sine wave by 90°. Equivalently, the sine wave lags the cosine wave by 90°. Sometimes, the component with $\cos(2\pi f_{L1}t+\theta_{L1})$ is called the inphase signal and the component with $\sin(2\pi f_{L1}t+\theta_{L1})$ is called the quadrature signal. The receiver can readily distinguish these signals because of this phase shift and the strong differences in the civil and military codes, $x(t)$ and $y(t)$.

RF carriers can also be specified by their wavelength

$$\begin{aligned}\lambda \text{ (meters/cycle)} &= \frac{c \text{ (meters/second)}}{f \text{ (cycles/second)}} \\ &\approx \frac{3\times10^8}{f}\end{aligned} \tag{9.4}$$

If we could stop time and examine the so-frozen RF carrier in space, then the wavelength would be the distance from one peak to the next. As shown above, λ is inversely proportional to frequency, and the GPS wavelengths are $\lambda_{L1} \approx 19$ cm and $\lambda_{L2} \approx 24$ cm.

Both the data and the codes use binary phase shift keying (BPSK) to modulate the transmitted carrier. As shown in Figures 9.1 and 9.3, they are composed of sequences of rectangular pulses with amplitude +1 or –1. These rectangular pulses are well described by $p(t)$ introduced in Chapter 8. Recall that the pulse

$$Ap\left(\frac{t-\tau}{T}\right)$$

has amplitude, A, duration, T, and is delayed by τ seconds. The navigation data and the spread spectrum codes are sequences of these elemental pulses. For the navigation data, each +1 or –1 is called a bit, and the bit stream carries the information required from the satellite for position fixing. This data only requires 50 bits per second (bps) which means that the duration of each

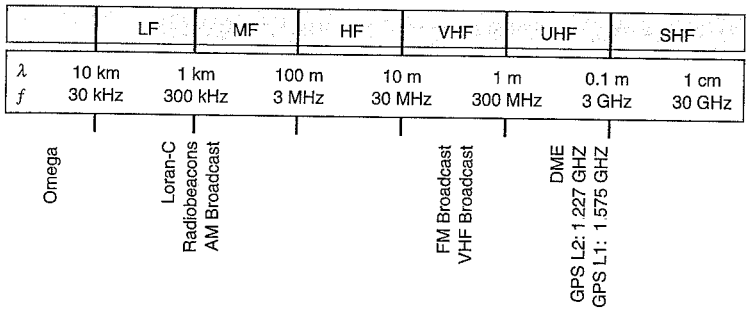


Figure 9.2 Radio spectrum.

pulse is $T_B = 20$ milliseconds. For the codes, each elemental pulse +1 or −1 is called a chip, with a chip being much shorter than a bit, $T_C << T_B$. In either case, the pulses flip the polarity of the waveform. Equivalently, they change the phase of the underlying carrier by 180°. These operations are shown in Figure 9.3.

As mentioned earlier, the civilian code, $x(t)$, is known as the Coarse/Acquisition or C/A-code. Each C/A-code is periodic, repeating a 1023-chip pattern. One period of this code can be written as

$$x_1(t) = \sum_{n=0}^{N-1} x_n p\left(\frac{t-nT_C}{T_C}\right)$$
$$N = 1023 \text{ chips} \tag{9.5}$$

In this equation, $p(t)$ is the elemental chip waveform with unit width, unit height and centered at the origin. In this equation, it is modified to have duration T_C and delay equal to nT_C. The amplitude of the nth pulse is modulated by x_n, which is the nth element in the sequence for this satellite. This code element modulates the sign of the individual chips or pulses. For the C/A-code, the chip duration is 1/1.023 microseconds, which we often approximate as 1 microsecond for brevity. The chipping rate is 1.023×10^6 chips per second or 1.023 Mcps. The C/A-code chipping rate is 20,460 times greater than the data rate for the navigation message. Since the C/A-code has length 1023, it repeats every millisecond, and the C/A-code is sent 20 times during each navigation bit. The code chips and navigation bits are clock synchronous as shown in Figure 9.3.

The military codes, $y(t)$, are called P(Y) codes. These codes are ten times faster than the C/A-codes, and much longer. Specifically, they have chipping rates of 10.23 Mcps, and each satellite repeats a unique, one-week section of a single P(Y)-code that is approximately 38 weeks long. The P(Y) codes are further discussed in Section 9.8.

Very different time scales characterize the navigation data, code, and carrier. The C/A-codes repeat 20 times for each navigation bit, and each repeat contains 1023 chips. The carrier at f_{L1} has 1540 cycles per C/A chip and 31,508,400 cycles per navigation bit. These very different time scales cannot be replicated in Figures 9.1 or 9.3, which are simply illustrative.

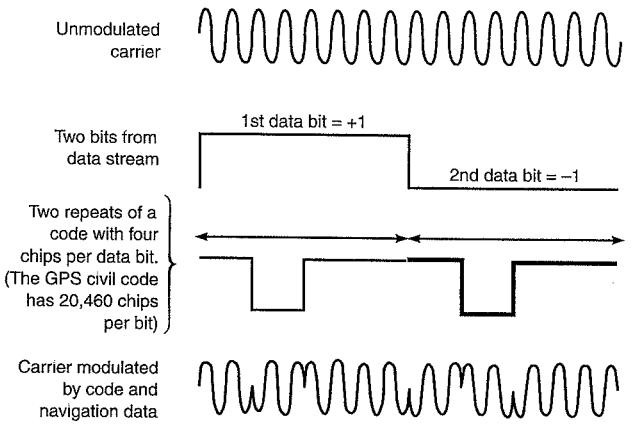


Figure 9.3 Binary phase shift keying (BPSK).

Using unit impulse functions, we can write one period of the C/A-code as a convolution.

$$x_1(t) = p\left(\frac{t}{T_C}\right) \star \sum_{n=0}^{N-1} x_n \delta(t-nT_C) \tag{9.6}$$

where $\star$ denotes convolution. This result is proven as follows.

$$\begin{aligned} p\left(\frac{t}{T_C}\right) \star \sum_{n=0}^{N-1} x_n \delta(t-nT_C) &= \int_{-\infty}^{\infty} p\left(\frac{t-\beta}{T_C}\right) \sum_{n=0}^{N-1} x_n \delta(\beta-nT_C)\, d\beta \\ &= \sum_{n=0}^{N-1} x_n \int_{-\infty}^{\infty} p\left(\frac{t-\beta}{T_C}\right) \delta(\beta-nT_C)\, d\beta \\ &= \sum_{n=0}^{N-1} x_n p\left(\frac{t-nT_C}{T_C}\right) \end{aligned} \tag{9.7}$$

This alternate expression for the GPS signal enables us to analyze the frequency content of the C/A-code signal.

9.1.2 Amplitude Spectrum

We now use (9.6) to derive the Fourier transform of one period of the C/A-code signal. For simplicity, we ignore the data modulation and set the carrier phase offset to zero. If we denote $X_1(f) = \mathcal{F}\{x_1(t)\}$, then the Fourier transform of the simplified signal is

$$\mathcal{F}\left\{\sqrt{2P_{C1}}\, x_1(t)\cos(2\pi f_{L1}t)\right\} = \sqrt{\frac{P_{C1}}{2}} X_1(f-f_{L1}) + \sqrt{\frac{P_{C1}}{2}} X_1(f+f_{L1}) \tag{9.8}$$

This result is a straightforward application of the modulation theorems provided in Table 8.1. It indicates that the spectrum of a GPS signal is the spectrum of $x_1(t)$ shifted up to the carrier frequency f_{L1} and down to $-f_{L1}$.

We find $X_1(f)$ as follows.

$$\begin{aligned}
\mathcal{F}\{x_1(t)\} &= \mathcal{F}\left\{p\left(\frac{t}{T_C}\right) * \sum_{n=0}^{N-1} x_n \delta(t-nT_C)\right\} \\
&= \mathcal{F}\left\{p\left(\frac{t}{T_C}\right)\right\}\mathcal{F}\left\{\sum_{n=0}^{N-1} x_n \delta(t-nT_C)\right\} \\
&= T_C P(fT_C)\int_{-\infty}^{\infty}\sum_{n=0}^{N-1} x_n \delta(t-nT_C)\exp(-j2\pi ft)dt \\
&= T_C P(fT_C)\sum_{n=0}^{N-1} x_n \exp(-j2\pi fnT_C) \\
&= T_C\sqrt{N}P(fT_C)\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x_n \exp(-j2\pi fnT_C) \\
&= T_C\sqrt{N}\,\mathrm{sinc}(\pi fT_C)X_{\mathrm{code}}(f)
\end{aligned} \tag{9.9}$$

$P(f)$ is the Fourier transform of the rectangular chip waveform given in Table 8.1. As shown in (9.9), the Fourier transform for the chip is $P(fT_C) = \mathrm{sinc}(\pi fT_C)$, and the corresponding amplitude spectrum is shown in Figure 8.15. It takes a maximum value of unity at $f = 0$ and falls to 0 at $1/T_C$. Hence, the null-to-null bandwidth is $2/T_C$, which is 2.046 MHz for the C/A-code and 20.46 MHz for the P(Y)-code. The main lobe falls between the first nulls, and the sidelobes fall outside of the first nulls.

We call $X_{\mathrm{code}}(f)$ the code transform because it depends only on the code $\{x_n\}_{n=0}^{N}$.

$$X_{\mathrm{code}}(f) = \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x_n \exp(-j2\pi fnT_C) \tag{9.10}$$

Combining (9.8) and (9.9) yields

$$\begin{aligned}
\mathcal{F}\left\{\sqrt{2P_{C1}}x_1(t)\cos(2\pi f_{L1}t)\right\} &= \sqrt{\frac{P_{C1}N}{2}}T_C\,\mathrm{sinc}\left(\pi(f-f_{L1})T_C\right)X_{\mathrm{code}}(f-f_{L1}) \\
&+ \sqrt{\frac{P_{C1}N}{2}}T_C\,\mathrm{sinc}\left(\pi(f+f_{L1})T_C\right)X_{\mathrm{code}}(f+f_{L1})
\end{aligned} \tag{9.11}$$

The amplitude of this hard-fought result is depicted three times in Figure 9.4: once for the C/A-code at f_{L1}, once for the P(Y)-code at f_{L1}, and finally for the P(Y)-code at f_{L2}. The figure only shows positive frequencies and the horizontal axis is not to scale. At f_{L1} = 1575.42 MHz, the spectrum for the C/A-code chip and P(Y)-code chip are superposed. As shown, the P(Y)-code bandwidth is 10 times the C/A-code bandwidth. At f_{L2} = 1227.60 MHz, the P(Y) code spectrum appears by itself.

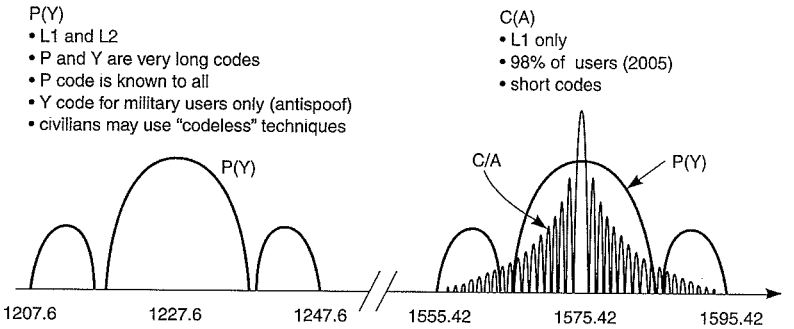


Figure 9.4 Amplitude spectrum of GPS signal in mid-2005 with f_{L1} = 1575.42 MHz and f_{L2} = 1227.60 MHz. Only positive frequencies are shown.

The sinc function from the chip waveform dominates the scenery in Figure 9.4. The influence of the code transform, $X_{\mathrm{code}}(f)$, is suppressed in this figure because it superposes a rather fine structure on top of the dominant sinc functions. We have also ignored the navigation message, $D(t)$, and the carrier phase in this development. These also have subtle effects compared to the chip waveform and the carrier modulation. In addition, we have only considered one period of the spread spectrum code, $x_1(t)$. We return to all of these rather more subtle effects in Section 9.6.

For fun, Figure 9.5 shows a spectrum measurement made with a large dish antenna at Stanford University. This antenna has a diameter of 46 meters and can create a very narrow beam in space. When this beam is trained on a single satellite, the signals from that satellite are greatly amplified and appear above the background noise. Most GPS antennas are very small and do not form beams. These antennas do not amplify the GPS signal so dramatically and hence the GPS signal is swamped by measurement noise. We will have much more to say about this subject in Chapter 10, when we quantify the strength of the background noise in the GPS band, and describe how the correlation process finds the GPS signal in all that noise. By the way, bandpass filters in the measurement system also influence the spectrum measurement shown in Figure 9.5. They cause the measured spectrum to roll off rapidly outside of ±20 MHz; the GPS signal spectrum does not roll off that quickly.

9.2 Auto-Correlation

As noted earlier, the GPS codes were selected for their auto- and cross-correlation properties. We now introduce these functions. In general, correlation measures the similarity of two waveforms or sequences. Auto-correlation measures the similarity between any waveform and time shifts of itself. Cross-correlation compares a given waveform with all time shifts of a second waveform.

The time-average auto-correlation function for the code from the kth satellite, $x^{(k)}(t)$, is [Pursley (1977), Sarwate and Pursley (1980)]

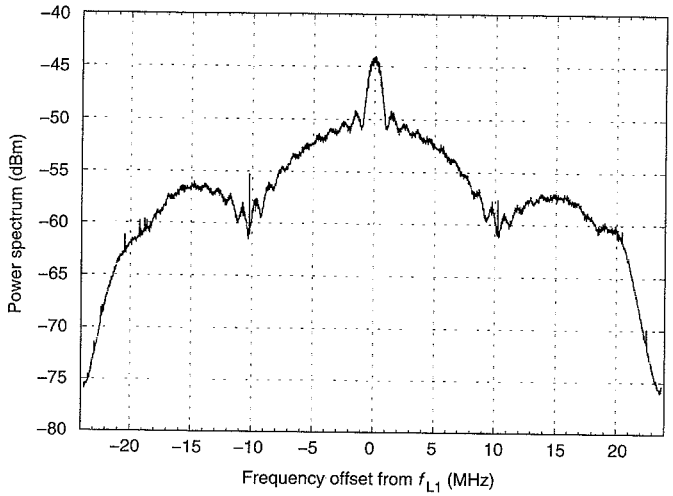


Figure 9.5 Measured amplitude spectrum showing the C/A and P(Y) codes from PRN 31 in February of 2003. The spectrum was measured using the 46-m-diameter dish antenna at Stanford University. Each division of the horizontal axis is 5 MHz and the first P(Y) code nulls are apparent at ±10 MHz. Most of the C/A-code energy is packed into ±1 MHz.

$$R(\tau)=\frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x^{(k)}(t)x^{(k)}(t-\tau)\,dt \tag{9.12}$$

As shown in this equation and Figure 9.6, auto-correlation multiplies $x^{(k)}(t)$ by a time-shifted replica of itself and integrates the product. If $x^{(k)}(t)$ resembles $x^{(k)}(t-\tau)$, then $R^{(k)}(\tau)$ will be large. This function correlates two deterministic functions and is similar to the correlation function for random signals introduced in Section 8.6. It assumes that the codes $x^{(k)}(t)$ and $x^{(k)}(t-\tau)$ repeat indefinitely. In this case, (9.12) is called the periodic or circular auto-correlation function. Equation (9.12) uses T_{code} to denote the period of the code. Recall that T_{code} = 1 ms for the C/A-code. From this point on, we will simplify our notation from $x^{(k)}(t)$ to $x(t)$ unless we really need to identify the satellite associated with the code.

As described in Section 9.1, we can write a single period of the civil code, $x_1(t)$, as follows.

$$x_1(t)=\sum_{n=0}^{N-1} x_n p\left(\frac{t-nT_C}{T_C}\right) \tag{9.13}$$

As before, N is the number of chips in each repeat of the code, and T_C is the duration of a single chip. Every T_{code} seconds, the code generator sends N chips of duration T_C, so $T_{\text{code}}=NT_C$. Finally, x_n is the nth element in the sequence for the kth satellite.

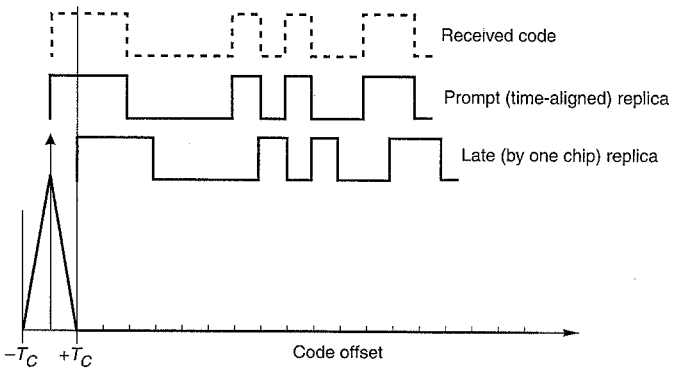


Figure 9.6 Correlation of the received code with time-shifted replicas. The correlation function that is shown is close to the ideal for ranging signals. It has a peak that is sharp and unique.

At first, we limit our attention to time shifts that are integer multiples of the chip duration $\tau=\ell T_C$. In this case, we can combine (9.12) and (9.13) as follows.

$$\begin{aligned}R(\tau=iT_C)&=\frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x(t)x(t-\ell T_C)\,dt\\&=\frac{T_C}{T_{\text{code}}}\sum_{n=0}^{N-i} x_n x_{n+i}\\&=\frac{T_C}{T_{\text{code}}}(\#\text{ agreements}-\#\text{ disagreements})\\&=\frac{1}{N}(\#\text{ agreements}-\#\text{ disagreements})\end{aligned} \tag{9.14}$$

The auto-correlation can be computed by simply summing the products of the underlying +1's and –1's. Note that the resulting sum simply counts the number of times the shifted waveforms agree in polarity minus the number of times they disagree.

The auto-correlation, $R(\tau)$, takes its maximum value when $\tau=0$. At $\tau=0$, all the chips are aligned. In other words, they all agree and there are no disagreements. Consequently,

$$\begin{aligned}R(\tau)\le R(0)&=\frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x^2(t)\,dt\\&=\frac{1}{N}(\#\text{ agreements}-\#\text{ disagreements})\\&=1\end{aligned} \tag{9.15}$$

Finally, we should consider what happens when $\tau\neq iT_C$. To this end, consider Figure 9.7, where the time delay, τ, falls between iT_C and $(i+1)T_C$.

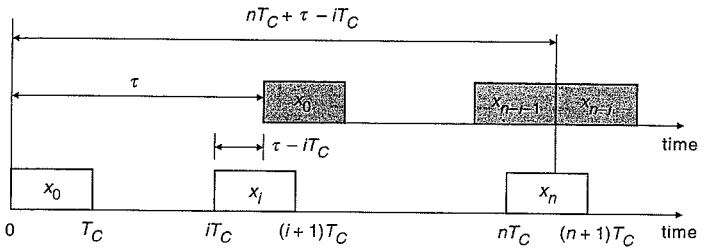


Figure 9.7 Time line to compute correlation when the two sequences are misaligned. The unshaded chips belong to the replica code created by the receiver. The shaded chips belong to the received code.

$$
\begin{aligned}
R(\tau) &= \frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x(t)\,x(t-\tau)\,dt \\
&= \frac{1}{T_{\text{code}}}\sum_{n=0}^{N-i}\int_{nT_C}^{(n+1)T_C} x(t)\,x(t-\tau)\,dt \\
&= \frac{1}{T_{\text{code}}}\sum_{n=0}^{N-i}\left(\int_{nT_C}^{(n+1)T_C-(\tau-iT_C)} x(t)\,x(t-\tau)\,dt + \int_{(n+1)T_C-(\tau-iT_C)}^{(n+1)T_C} x(t)\,x(t-\tau)\,dt\right) \\
&= \frac{1}{T_{\text{code}}}\sum_{n=0}^{N-i} x_n x_{n+i}\left(T_C-(\tau-iT_C)\right) + x_n x_{n+i+1}\left(\tau-iT_C\right) \\
&= R(iT_C)\left(i+1-\frac{\tau}{T_C}\right) + R\left((i+1)T_C\right)\left(\frac{\tau}{T_C}-i\right)
\end{aligned}
\tag{9.16}
$$

The correlation, $R(iT_C < \tau < (i+1)T_C)$, falls on the straight line that connects $R(iT_C)$ and $R((i+1)T_C)$.

9.2.1 Random Sequences

The GPS codes are sometimes called pseudo-random noise (PRN) sequences or pseudo-random codes. This name is apt and worth exploring. Contemplate a sequence of encoded chips where the sign (or polarity) of each chip is determined by a coin toss. If the coin toss produces a head, then the sign is +1; and if a tail shows, we send a –1. Clearly, any number of codes can be constructed as outcomes of such random experiments and we could determine the auto-correlation function for any one of them after construction.

Even though the GPS codes are not constructed by tossing coins, random codes have tremendous utility. They can be used to estimate system performance without flipping a single coin. Specifically, we can compute the average correlation function for an ensemble of random codes. In addition, the likely deviation of any individual code from this average can be determined. In this way, random codes provide a powerful design guide that allows us to estimate

the performance of spread-spectrum signaling without designing any actual sequences (or flipping any coins). Consideration of random codes allows the determination of the chipping rate and code lengths needed to meet certain design criteria. After these basic parameters have been determined, the design of specific sequences can begin.

Our random codes are constructed by choosing every chip randomly and independently of all other chips. In other words, the outcome of the coin toss for a given chip has no impact on any other chip in that sequence or on any chip in another satellite's sequence. We seek the average auto-correlation function for all of these random sequences.

Our search begins with the following result from (9.16).

$$
R(\tau) = R(iT_C)\left(i+1-\frac{\tau}{T_C}\right) + R\left((i+1)T_C\right)\left(\frac{\tau}{T_C}-i\right) \tag{9.17}
$$

We take the expected value with respect to randomly chosen chips.

$$
\begin{aligned}
E\{R(\tau)\} &= \bar{R}(\tau) \\
&= \bar{R}(iT_C)\left(i+1-\frac{\tau}{T_C}\right) + \bar{R}\left((i+1)T_C\right)\left(\frac{\tau}{T_C}-i\right)
\end{aligned}
\tag{9.18}
$$

With this result, we only need to focus on the average auto-correlation function when the time shift between codes is equal to an integer number of chips, $\tau = iT_C$. In this case, we write

$$
\begin{aligned}
E\left\{R(iT_C)\right\} &= E\left\{\frac{T_C}{T_{\text{code}}}\sum_{m=0}^{N-i} x_m x_{m+i}\right\} = \frac{T_C}{T_{\text{code}}}E\left\{\sum_{m=0}^{N-1} x_m x_{m+i}\right\} \\
&= \frac{T_C}{T_{\text{code}}}\sum_{m=0}^{N-1} E\left\{x_m x_{m+i}\right\} = \begin{cases} 1 \text{ if } i=0 \\ 0 \text{ if } i\neq 0 \end{cases}
\end{aligned}
\tag{9.19}
$$

because

$$
E\{x_m x_{m+i}\} = \begin{cases} 1 \text{ if } i=0 \\ 0 \text{ if } i\neq 0 \end{cases} \tag{9.20}
$$

Combining (9.18) and (9.19), we may write

$$
\bar{R}(\tau) = \begin{cases} \dfrac{\tau}{T_C}+1 & \text{if } -T_C < \tau < 0 \\ \dfrac{-\tau}{T_C}+1 & \text{if } 0 < \tau < T_C \\ 0 & \text{otherwise} \end{cases} \tag{9.21}
$$

As shown in (9.21) and Figure 9.8, the average of the auto-correlation functions has a sharp and distinct peak. The maximum value occurs when the relative shift is zero, and the average is zero for all other shifts. As we shall discover, these are very nice properties.

However, this desirable average will be small solace if the individual codes within the ensemble tend to have auto-correlation functions that are very different from the average. To characterize this possibility, we compute the variance, which is the mean squared deviation of

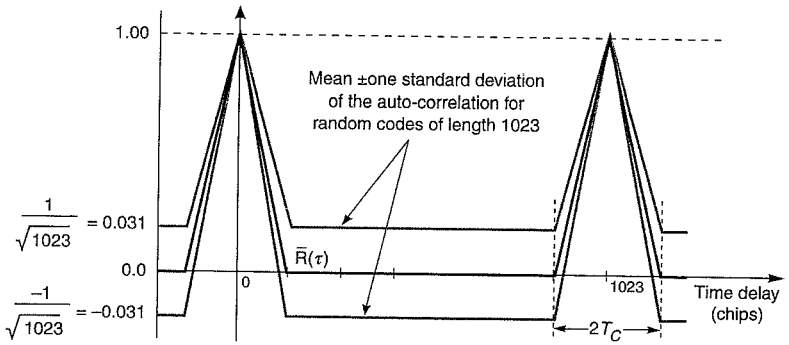


Figure 9.8 Auto-correlation properties of random codes that have the same length as the C/A-codes ($N = 1023$). The dark line is the mean auto-correlation function, and the two gray lines are the mean ±one standard deviation.

an individual correlation function away from the average. The variance is given by

$$E\left\{\left(R(iT_C) - E\left\{R(iT_C)\right\}\right)^2\right\} = E\left\{\left(R(iT_C)\right)^2\right\} - \left(E\left\{R(iT_C)\right\}\right)^2 \tag{9.22}$$

The second of these terms is simply the mean squared. Since we have already calculated the mean, we pursue the first term.

$$\begin{aligned} E\left\{\left(R(iT_C)\right)^2\right\} &= \frac{T_C^2}{T_{\text{code}}^2} E\left\{\sum_{m=0}^{N-i} x_m x_{m+i} \sum_{n=0}^{N-i} x_n x_{n+i}\right\} \\ &= \frac{1}{N^2} \sum_{m=0}^{N-i} \sum_{n=0}^{N-i} E\left\{x_m x_{m+i} x_n x_{n+i}\right\} \\ &= \begin{cases} \dfrac{1}{N} & \text{if } i \neq 0 \\ 1 & \text{if } i = 0 \end{cases} \end{aligned} \tag{9.23}$$

This result follows because the chips are independent and

$$E\left\{x_m x_{m+i} x_n x_{n+i}\right\} = \begin{cases} 1 & m = n \text{ or } i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{9.24}$$

Finally, the standard deviation is given by

$$\sqrt{E\left\{\left(R(iT_C) - E\{R(iT_C)\}\right)^2\right\}} = \begin{cases} \dfrac{1}{\sqrt{N}} & \text{if } i \neq 0 \\ 0 & \text{if } i = 0 \end{cases} \tag{9.25}$$

Figure 9.8 depicts the auto-correlation function for a set of random codes with $N = 1023$. More specifically, the dark line shows $\bar{R}(\tau)$, and the gray lines show $\bar{R}(\tau)$ plus and minus one standard deviation. These gray lines would enclose most, but certainly not all, auto-correlation sidelobes for random codes.

Often, we use decibels to describe these sidelobe levels. In this case, we have

$$\begin{aligned} 20\log_{10}\frac{1}{\sqrt{N}} &= 10\log_{10}\frac{1}{N} \\ 20\log_{10}\frac{1}{\sqrt{1023}} &= 10\log_{10}\frac{1}{1023} \\ &\approx 10\log_{10} 10^{-3} \\ &= -30 \text{ dB} \end{aligned} \tag{9.26}$$

We take $20\log_{10} 1/\sqrt{N}$ because the standard deviation represents voltage, not power. For random codes of length 1023, the sidelobe levels are ±0.031 or –30 dB relative to the main peak. As we shall discover, random codes give a reasonable indication of the sidelobes for the deterministic C/A-codes actually used by GPS. Moreover, they are so much easier to find.

9.2.2 Ranging Precision

A sharp auto-correlation peak is a wonderful thing. For GPS, the nice peak shown in Figures 9.6 and 9.8 enables precise range measurements. This feature is the greatest value of spread spectrum signaling to GPS. After all, GPS is a ranging system. Recall that the C/A-code chip has a duration of $T_C \approx 1$ microsecond. Since the speed of light is $c = 3 \times 10^8$ m/s, a chip duration of one microsecond corresponds to a chip length of 300 meters. We can measure the arrival time with a precision of approximately 0.1%. This rule of thumb is applicable only in environments without obstructions or reflections. For the C/A-code, this corresponds to a range precision of 0.3 meters, which is not bad. If we did not use spread spectrum, we could still range on the navigation data bits used to send the navigation message. However, the navigation data rate is only 50 bps, so the data bits are 20 milliseconds in duration or 6000 kilometers in length. With the same measurement precision of 0.1%, we would have a ranging precision of only six kilometers. Needless to say, this latter precision would be disappointing for someone trying to navigate in an unfamiliar city, or worse yet, trying to land an airplane.

Higher chipping rates give better ranging precision. As shown in Figures 9.6 and 9.8, our precious correlation peak has a base width of two chip durations, $2T_C$. Increasing the chipping rate, $1/T_C$ shortens the chip and narrows the correlation peak. The arrival time of a crisp event is easier to estimate than an event with a slow onset.

These qualitative arguments show that ranging precision is proportional to chip width. The next chapter substantiates this claim, and we borrow against the future with this result from Chapter 10.

$$\sigma_{\Delta\tau} \leq cT_C \sqrt{\frac{1}{4TC/N_0}} \text{ meters} \tag{9.27}$$

This nifty result gives the standard deviation of the ranging error when the GPS signal is troubled by measurement noise alone. In this equation, c is the speed of light, and T is the averaging time used by the receiver. C/N_0 is the signal power (watts) divided by the measurement noise power density (watts/hertz). Ranging precision improves with increased averaging time, T, and increased signal-to-noise ratio, C/N_0. However, it is most strongly tied to T_C, the chip duration. As we suggested, ranging precision is proportional to the chip duration.

A sharp auto-correlation peak also helps ward off the deleterious effects of signal reflections. It helps the receiver distinguish the signal arriving directly from the satellite from any signals that have traveled paths that include a reflection. The latter (and later) signals are called multipath signals. They confound the range measurements by creating correlation peaks that shadow the main peak created by the direct signal. If the shadow peak merges with the peak from the direct signal, then large ranging errors can result. If the peak due to the reflected signals is distinct from the direct peak, then the receiver can distinguish the two peaks and no harm results. Consequently, spread spectrum signals are once again helpful. They create narrow correlation peaks and enable the receiver to distinguish a greater fraction of shadow peaks from main peaks. As before, higher chipping rates are better. The multipath performance of spread spectrum signals will be further discussed in Chapter 10.

Even though spread spectrum signaling enables high precision ranging, it does not, by itself, guarantee high accuracy. It allows us to precisely estimate the arrival time of a chip transition, but does not remove the bias errors discussed in Chapter 5. Recall that these biases are due to uncertainties in the signal delay due to the ionosphere and the troposphere, and errors in the satellite clock and ephemeris data. Usually, these bias errors determine the accuracy of the range measurement.

9.2.3 Signal Acquisition

Well-designed sequences also help a receiver achieve initial synchronization with the incoming codes. This initial synchronization process is called signal acquisition and is so important that we devote most of Chapter 11 to this subject. During signal acquisition, the receiver tests any number of starting times for the incoming code. After all, a cold receiver has no idea when the code begins. It must search for the beginning of the code. During this process, a receiver sweeps through the correlation function and seeks strong correlation peaks. Strong sidelobes pose a risk. In the presence of measurement noise, a receiver could confuse a strong sidelobe for the main lobe.

Longer codes give lower auto-correlation sidelobes. For random codes, the standard deviation of the sidelobes is given by $1/\sqrt{N}$. On average, the sidelobes for an $N = 1023$ code will be approximately 30 dB below the main peak. If the code is lengthened to $N = 10{,}230$, then the sidelobes will fall to 40 dB below the main peak and the likelihood of acquiring a false peak is reduced.

Code length is eventually limited by the duration or complexity of the signal acquisition process. If the code is short, then the receiver usually correlates a replica of the whole code with the incoming signal. The time required for each such test is equal to the duration of the code. If the code is long, then the receiver may conduct many tests in parallel, where each test correlates with a different piece of the incoming code. However, the number of correlators increases. A tremendous amount of imagination has been applied to the problem of signal acquisition [J. Holmes (1990)], and many ingenious algorithms exist.

9.3 Cross-Correlation and Channel Sharing

Now let $x^{(k)}(t)$ and $x^{(\ell)}(t)$ denote the codes from the kth satellite and the ℓth satellite, respectively. In this case, the cross-correlation function is of interest.

$$
\begin{aligned}
R^{(k,\ell)}(\tau) &= \frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x^{(k)}(t)\,x^{(\ell)}(t-\tau)\,dt \\
R^{(k,\ell)}(\tau = iT_C) &= \frac{1}{T_{\text{code}}}\int_0^{T_{\text{code}}} x^{(k)}(t)\,x^{(\ell)}(t-iT_C)\,dt \\
&= \frac{T_C}{T_{\text{code}}}\sum_{m=0}^{N-i} x_m^{(k)} x_{m+i}^{(\ell)} \\
&= \frac{1}{N}(\#\text{ agreements} - \#\text{ disagreements})
\end{aligned}
\tag{9.28}
$$

For $R^{(k\neq\ell)}(\tau)$, the agreements and disagreements are counted between the codes for the two different satellites. If $k = \ell$, then $R^{(k,k)}(\tau) = R(\tau)$ and we have the auto-correlation function discussed in Section 9.2.

Once again, random sequences can be used to ballpark the correlation performance of spread spectrum codes. In this case, we find the mean and standard deviation of the cross-correlation between two random sequences.

$$
\begin{aligned}
E\left\{R^{(k,\ell)}(\tau = iT_C)\right\} &= \frac{T_C}{T_{\text{code}}}\sum_{n=0}^{N-i} E\left\{x_n^{(k)} x_{n+i}^{(\ell)}\right\} = 0 \\
E\left\{\left(R^{(k,\ell)}(iT_C)\right)^2\right\} &= \frac{T_C^2}{T_{\text{code}}^2}\sum_{m=0}^{N-i}\sum_{n=0}^{N-i} E\left\{x_m^{(k)} x_{m+i}^{(\ell)} x_n^{(k)} x_{n+i}^{(\ell)}\right\} = \frac{1}{N}
\end{aligned}
\tag{9.29}
$$

The corresponding standard deviation is given by

$$
\sqrt{E\left\{\left(R^{(k,\ell)}(iT_C) - E\left\{R^{(k,\ell)}(iT_C)\right\}\right)^2\right\}} = \frac{1}{\sqrt{N}} \tag{9.30}
$$

Accordingly, we expect cross-correlation sidelobes at approximately the same level as the auto-correlation sidelobes. Recall that the auto-correlation sidelobes for an $N = 1023$ code are expected to be around 30 dB below the auto-correlation peak. The cross-correlation function will have no peak, and the sidelobes are also expected to be 30 dB below the peak auto-correlation for the desired satellite code.

As we shall discover, these easily generated results for random codes give a good indication of the sidelobes for actual codes used by GPS. In addition, the cross-correlation for $iT_C < \tau < (i+1)T_C$ still falls on the straight line that connects $R^{(k,\ell)}(iT_C)$ and $R^{(k,\ell)}((i+1)T_C)$.

These remarkable cross-correlation properties enable channel sharing. With uniformly low cross-correlation, the signals from the different GPS satellites can be distinguished by their codes, so the satellites broadcast simultaneously without any time sharing or offsets in their transmission frequencies. In the lingo of the communication community, GPS uses code division multiple access (CDMA) rather than time division multiple access (TDMA) or frequency division multiple access (FDMA). Even though the transmitted frequencies are nearly identical, the received GPS signals do usually have slightly different frequencies because the Doppler shifts differ. This slight frequency offset also helps distinguish the satellites. By the way, GLONASS uses FDMA, but Galileo will use CDMA. Indeed, GPS and Galileo are premier applications of CDMA, where the spread spectrum codes enable the receiver to distinguish between the satellites.

Low cross-correlation sidelobes are particularly important for GPS operations in obstructed environments. If a signal passes through foliage or a wall, then it will be weakened and the cross-correlation sidelobe from an unobstructed satellite would be particularly troublesome. When the received signal powers are different, our $1/N$ result needs to be scaled by the power ratio. If the receiver power for the desired satellite is $P^{(k)}$ and the received power for the competing signal is $P^{(m)}$, then we need $NP^{(k)}/P^{(m)}$ to be large. These environments are discussed in Chapter 13, where we find that signal attenuations of 20 dB or more are not uncommon in cities. Consequently, we would like the cross-correlation sidelobes to be 30 dB weaker than the main peak, so that a 10 dB margin would remain. For this reason, long sequences are favored for urban environments.

9.4 Maximal Length Linear Shift Register Sequences

We now turn our attention away from random codes and toward the deterministic machines that generate the actual GNSS codes. The linear shift register is the principal building block for all GNSS codes. The so-called maximal-length linear shift register sequences are combined to generate the C/A-codes used by GPS. Thankfully, these sequences are also known by a shorter name as m-sequences. In any event, they can be generated using shift registers like those shown in Figures 9.9 and 9.10. The outputs from two such shift registers can be combined to generate all of the C/A-codes used by GPS. The C/A-codes repeat after 1023 chips, but the examples shown in Figures 9.9 and 9.10 repeat after only $N = 31$ chips. We use such a short sequence simply because it is more manageable as an example.

Each of our example registers have five cells, which are filled with –1's and +1's. Since each cell can take two values, there are $2^5 = 32$ possible states for each shift register. At each time step, the current elements are combined as shown to create a new input element, and all elements are shifted to the right. Using the notation from Figures 9.8 and 9.9, the new elements are generated from the following feedback equations

$$\begin{aligned} u_{\text{new}} &= u_n u_{n+2} \\ v_{\text{new}} &= v_n v_{n+2} v_{n+3} v_{n+4} \end{aligned} \tag{9.31}$$

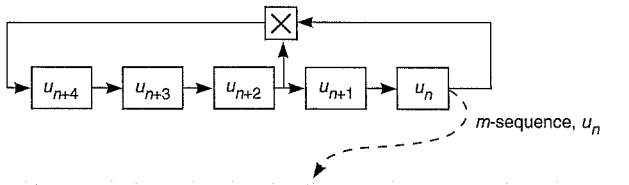
By the way, we need not use –1's and +1's together with multipliers. We can get the same results if we replace –1 with 1 and +1 with 0, and replace multiplication with the exclusive-OR operation defined as follows.

$$\begin{aligned} 0 \oplus 0 &= 0 \\ 0 \oplus 1 &= 1 \\ 1 \oplus 0 &= 1 \\ 1 \oplus 1 &= 0 \end{aligned} \tag{9.32}$$

In this case, the feedback equations become

$$\begin{aligned} u_{\text{new}} &= u_n \oplus u_{n+2} \\ v_{\text{new}} &= v_n \oplus v_{n+2} \oplus v_{n+3} \oplus v_{n+4} \end{aligned} \tag{9.33}$$

Multipliers are used in Figure 9.9 and exclusive-OR operations are used in Figure 9.10. The latter formulation shows the new input to be a linear combination of the current values. For this reason, the registers shown in the figures are called linear feedback shift registers.



{–1, 1, 1, 1, 1, –1, 1, 1, –1, 1, –1, –1, –1, 1, 1, –1, –1, –1, –1, –1, 1, 1, 1, –1, –1, 1, –1, –1, –1, 1, –1, 1}

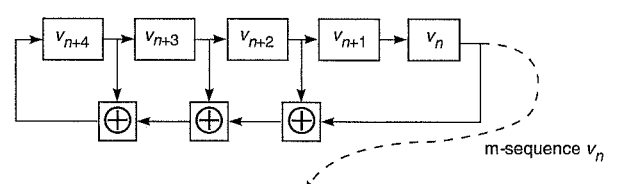
or equivalently

{1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0}

Figure 9.9 A maximal-length linear shift register using multiplication and +1's and –1's. This finite state machine generates a sequence of length 31.

Whether we use multiplications or exclusive-OR operations, these feedback equations are special. They generate sequences that repeat after exactly 31 time steps. Of the 32 possible states, the all-zero state (or all +1's) is to be avoided because it would feedback a 0 and result in an all-zero output. The zero output would continue in perpetuity, creating a very dull sequence with miserable auto-correlation properties. If all 31 non-zero states are exercised, the resulting sequence has maximum length. After all, the sequence must begin to repeat whenever the shift register returns to a state that has been previously exercised. This property motivates the name maximum-length linear shift register sequences, which in turn motivates the shorthand m-sequences. By the way, m-sequences do require careful selection of the feedback taps. For fun, move the taps in Figures 9.9 or 9.10 around a few times, compute the resulting sequences and see if they are still maximum length.

In addition to maximal length, m-sequences have many other amazing properties [MacWilliams and Sloane (1976)], and we will review a few of these. First, m-sequences have auto-correlation functions that are virtually perfect for range measurements. This nice feature is



{1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0}

or equivalently

{–1, 1, 1, 1, 1, –1, –1, 1, 1, –1, 1, 1, –1, –1, –1, –1, –1, 1, –1, –1, –1, 1, 1, 1, –1, 1, –1, 1, –1, –1, 1}

Figure 9.10 Another maximal-length linear shift register using exclusive-OR operations and 0's and 1's. This finite state machine also generates a sequence of length 31. To do so, every non-zero state is exercised.

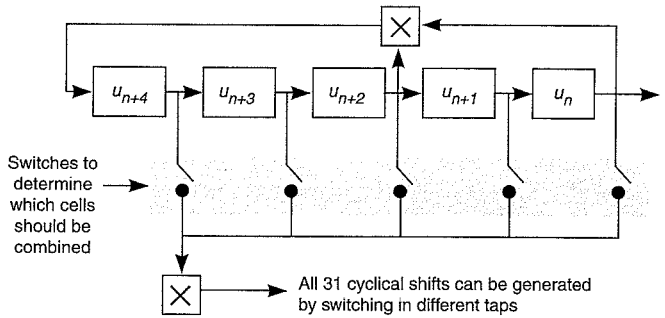


Figure 9.11 All cyclical shifts of a given m-sequence can be generated by combining the contents of the five cells. If at least one switch is closed, then there are 31 possible switch settings, and these generate the 31 shifts of this length-31 m-sequence.

explored in the homework. Second, the sum of an m-sequence and a cyclically shifted version of itself is another cyclical shift of that same sequence. Third, all $N-1$ cyclical shifts of an m-sequence can be generated by combining the contents of the cells in the shift register. This property is depicted in Figure 9.11 where a set of switches tap into the shift register, and the outputs from the chosen taps are multiplied together. As we open and close these switches, we generate all possible shifts of the m-sequence. As described in the next section, this last property will come in handy when generating the Gold sequences used by GPS.

For any length $N = 2^n - 1$, there is usually more than one m-sequence. For example, the m-sequence generated by Figure 9.9 is not a cyclical shift of the m-sequence generated by Figure 9.10. In some cases, two distinct m-sequences with equal length can be used to create an entire family of sequences that have both good auto-correlation properties and good cross-correlation properties—these are the Gold codes used by GPS and described in the next section. However, the two m-sequences must be chosen carefully, and there is no guarantee that two randomly chosen m-sequences will result in a Gold family.

9.5 Gold Codes of Length 31 and 1023

The C/A-codes used by GPS are called Gold codes or Gold sequences [Gold (1967)]. The C/A-codes repeat after 1023 chips, but Gold's original recipe can be used to construct codes of any length $N = 2^n - 1$, where $n \neq 0 \bmod 4$. As such, they exist for $N = \{7, 31, 63, 127, 511, 1023, 2047, 8191, \ldots\}$, but codes of lengths 15 and 255 cannot be constructed. In this section, we will certainly pay attention to the length-1023 codes used for GPS, but we will continue to discuss the length-31 codes, because they are more manageable as examples.

9.5.1 Construction

A carefully chosen pair of m-sequences of length N can be used to generate a family of $N+2$ Gold sequences [Sarwate and Pursely (1980)]. For example, two carefully chosen m-sequences of length 31 can be used to construct a family of 33 Gold sequences each with length 31. As another example, two carefully chosen m-sequences of length 1023 can be used to generate

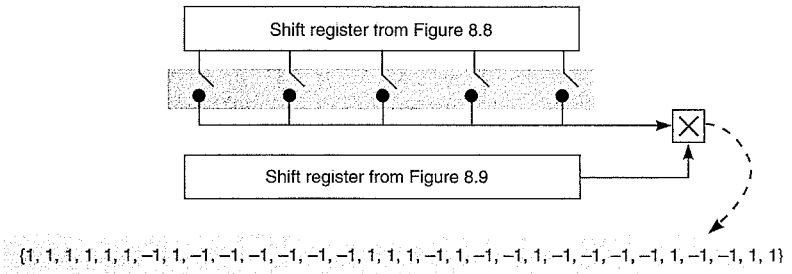


Figure 9.12 Combining outputs from two maximum-length shift registers to generate a length-31 Gold sequence. Not every pair of length-31 m-sequences can be used.

1025 sequences with length 1023. Since we only have a few dozen GPS satellites, the latter family easily provides one unique sequence per satellite. All of these Gold sequences have good auto-correlation functions. In addition, any pair of sequences drawn from within the family of $N+2$ members will also have uniformly low cross-correlation. As we have discussed, this latter property allows the GPS receiver to distinguish the signals from the different satellites.

The first two members of a Gold family are the original two m-sequences. The remaining N members are created by adding one of the m-sequences to the N cyclical shifts of the other. After all, the sequences have length N and so there are N distinct shifts of either one. If u and v are the two m-sequences that are capable of generating a Gold family, then we denote the Gold family of length N as

$$\begin{aligned} \underline{u} &= \{u_n\}_{n=0}^{N-1} \\ \underline{v} &= \{v_n\}_{n=0}^{N-1} \\ &\left\{\underline{u} \oplus S^m \underline{v}\right\}_{m=0}^{N-1} \end{aligned} \tag{9.34}$$

$S^m \underline{v}$ is a cyclical shift of the sequence v by m places.

$$\underline{u} \oplus S^m \underline{v} = \{u_n \oplus v_{n-m}\}_{n=0}^{N-1} \tag{9.35}$$

This family may be pre-computed and stored in a computer for transmission. Alternatively, the shift register machinery shown in Figure 9.12 may be used to generate an entire family of $N = 31$ Gold sequences. The first two Gold codes are the m-sequences from the two shift registers acting alone. Moving the connection taps to the top shift register generates the remaining 31 sequences. As the switches are opened and closed, we generate all of the cyclical shifts of the first m-sequence. These are added to the second m-sequence as prescribed by (9.34).

9.5.2 Correlation Functions

When the time shift is constrained to be an integer multiple of a chip width, then the auto-correlation function for any Gold sequence is equal to one of the following four values [Sarwate

and Pursley (1980)].

$$R^{(k)}(\tau = iT_C) \in \left\{1, \frac{-1}{N}, \frac{-\beta(n)}{N}, \frac{\beta(n)-2}{N}\right\}$$
$$\beta(n) = 1 + 2^{\lfloor (n+2)/2 \rfloor} \tag{9.36}$$

where $N = 2^n - 1$ is the length of the code and $\lfloor a \rfloor$ is the largest integer less than a. Unity auto-correlation only occurs for zero shift; the other values are the auto-correlation sidelobes.

The cross-correlation functions for all Gold codes only take three values

$$R^{(k,\ell)}(\tau = iT_C) \in \left\{\frac{-1}{N}, \frac{-\beta(n)}{N}, \frac{\beta(n)-2}{N}\right\} \quad \text{for all } k \neq \ell \tag{9.37}$$

If the time shifts are constrained to be integer multiples of chip durations, they take the same three values that the auto-correlation functions take for non-zero time shifts.

The sidelobe values for the $N = 31 = 2^5 - 1$ and $N = 1023 = 2^{10} - 1$ codes are {−1/31, −9/31, 7/31} and {−1/1023, −65/1023, 63/1023}, respectively, because

$$\beta(5) = 1 + 2^3 = 9$$
$$\beta(10) = 1 + 2^6 = 65 \tag{9.38}$$

Figure 9.13 shows the entire auto-correlation function for one of the $N = 31$ code generated by Figure 9.12. As described in Section 9.2, $R(\tau)$ falls on the straight line that connects $R(iT_C)$ and $R[(i+1)T_C]$ when $iT_C < \tau < (i+1)T_C$. This piecewise linear behavior is shown in the figure. The auto-correlation function for the length $N = 1023$ code was the subject of a homework problem in Chapter 2, and a portion of the overall function is depicted in Figure 9.14. The cross-correlation sidelobes are depicted in Figure 9.15 for the length-1023 codes used by GPS.

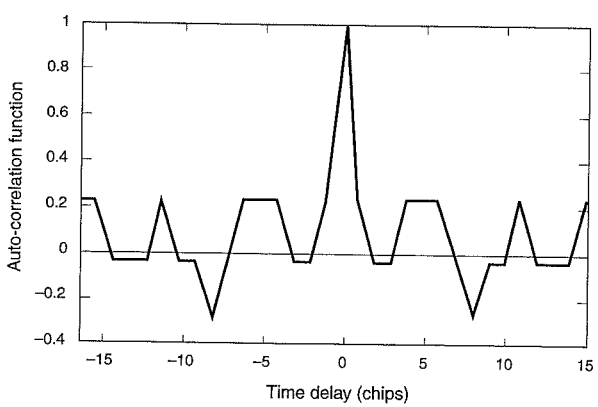


Figure 9.13 Auto-correlation function for an $N = 31$ Gold sequence. Like the auto-correlation function for all Gold codes, this function takes only four values. For this length-31 code, those values are {1, −1/31, −9/31, 7/31}.

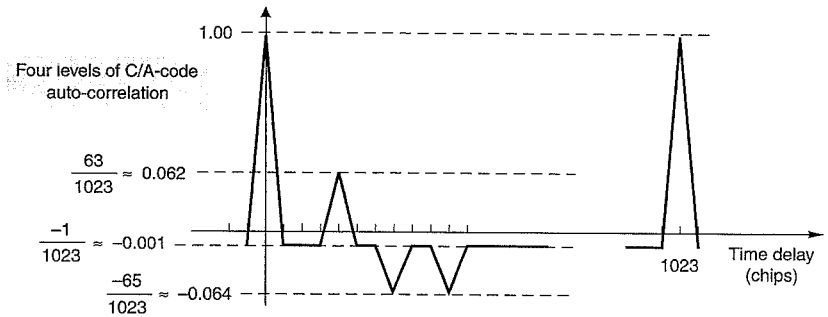


Figure 9.14 Auto-correlation for the C/A-codes. This auto-correlation function takes only the four values that are shown above, but there are many more sidelobes than shown. In addition, this figure assumes that the receiver has infinite front-end bandwidth.

If we use decibels to describe these sidelobe levels, then we have

$$10\log_{10}\left|\frac{63}{1023}\right|^2 = 20\log_{10}\left|\frac{63}{1023}\right|$$
$$\approx -24 \text{ decibels} = -24 \text{ dB}$$
$$20\log_{10}\left|\frac{-65}{1023}\right| \approx -24 \text{ dB}$$
$$20\log_{10}\left|\frac{-1}{1023}\right| \approx -60 \text{ dB} \tag{9.39}$$

The strongest C/A-code sidelobes are −24 dB relative to the main auto-correlation peak. Since

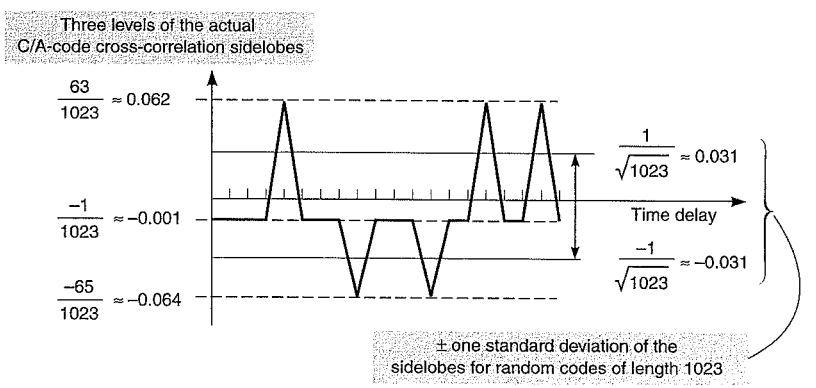


Figure 9.15 Cross-correlation function for the C/A-codes. These functions take only the three values shown above, but there are many more sidelobes than shown. As shown, random codes predict the sidelobe values with fair accuracy.

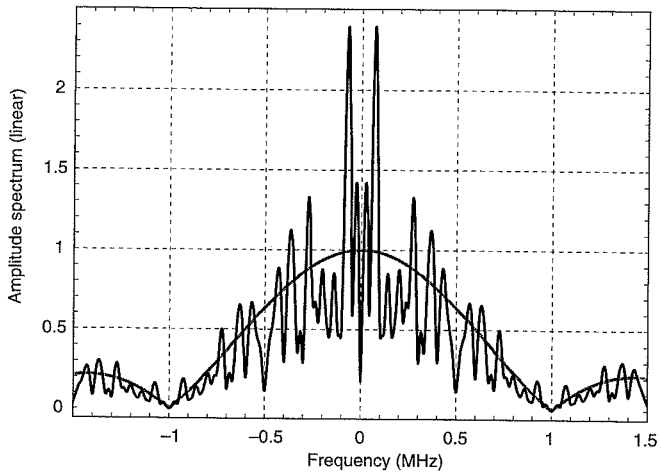


Figure 9.16 Amplitude spectrum of a length-31 Gold code signal showcasing the impact of the code spectrum, (9.40), on a linear scale.

these are the worst case sidelobes, they should be somewhat stronger than the expected sidelobe levels of –30 dB predicted by our random sequence analysis.

9.5.3 Code Spectrum

From (9.9), the amplitude spectrum of one period of a spread-spectrum code is given by

$$\left|X_1(f)\right| = T_C \left|\text{sinc}(\pi T_C f)\right| \sqrt{N} \left|X_{\text{code}}(f)\right|$$

$$X_{\text{code}}(f) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n \exp(-j2\pi f n T_C) \tag{9.40}$$

We can now plot $\left|X_1(f)\right|$ since Figure 9.12 gives us $\{x_n\}_{n=0}^{N-1}$ for one of the length-31 Gold codes. The smooth curve in Figures 9.16 and 9.17 is the familiar main lobe of the amplitude spectrum for the rectangular chip waveform, $\left|\text{sinc}(\pi f T_C)\right|$. The jagged curve is the amplitude of the product, $\left|\text{sinc}(\pi f T_C) X_{\text{code}}(f)\right|$. Figure 9.16 is a plot of these functions on a linear scale, and Figure 9.17 gives the exact same functions in decibels. The code spectrum superposes a fine structure on top of the characteristic $\left|\text{sinc}(\pi f T_C)\right|$ function. In the next section, we study the spectrum for random codes.

9.6 Power Spectral Density

So far, we have focused on the C/A-codes used with the civil signal on L1. Soon we turn our attention to the P(Y) codes and the new codes for GPS and Galileo. These codes are longer than the C/A-codes. For example, the P(Y) codes are one week in duration. The new civil codes for GPS have periods of 10,230 chips, 767,250 chips, 102,300 chips and 204,600 chips. So the shortest of these is ten times longer than the C/A-code.

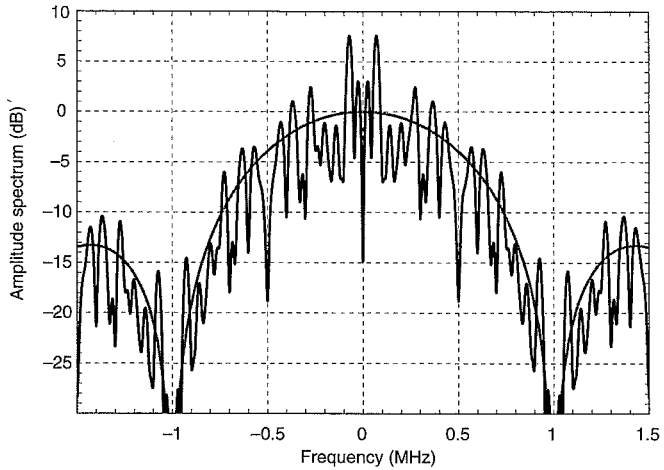


Figure 9.17 Amplitude spectrum of a length-31 Gold code signal showcasing the impact of the code spectrum, (9.40), in decibels.

Longer codes have two nice features. Most importantly, they are better at attenuating narrowband radio frequency interference (RFI). We shall have more to say about RFI in Section 9.7. Second, longer codes are easier to analyze because they are well modeled as random codes.

When we invoke random codes, we convert the GPS signal from a deterministic signal into a random signal. As such, the GPS signal acquires a power spectral density (PSD). This section computes the PSD for two cases. First, we examine long codes where the chips are chosen randomly. As you shall see, this analysis is easy. Second, we compute the PSD for C/A-codes. Sadly, this analysis is hard. After all, the C/A-codes are short and not well modeled as random codes. In addition, they repeat 20 times for each data bit. For our C/A-code analysis, we model the codes as deterministic, but treat the navigation data bits as random.

9.6.1 Long Codes

Recall our definition of power spectral density from Chapter 8.

$$S_N(f) = \lim_{T\to\infty} \frac{E\left\{\left|N_T(f,\zeta)\right|^2\right\}}{T} \tag{9.41}$$

In this equation, $N_T(f,\zeta) = \mathcal{F}\{n_T(t,\zeta)\}$, where $n_T(t,\zeta)$ is a truncated version of the random signal. As shown, the PSD is based on the energy spectrum, $\left|N_T(f,\zeta)\right|^2$, discussed in Chapter 8. Since $n(t,\zeta)$ has infinite energy, we consider a T-second segment of the signal, $n_T(t,\zeta)$. This truncated signal, $n_T(t,\zeta)$, has finite energy and is drawn randomly from a set of truncated signals. Hence, its energy spectrum, $\left|N_T(f,\zeta)\right|^2$, will also depend on the underlying random event ζ. As shown, we average over the underlying random variable (or variables), and the resulting expectation, $E\{\left|N_T(f,\zeta)\right|^2\}$ no longer depends on ζ. However, we still have an energy

spectrum that will grow without bound as T grows. Consequently, we divide by T to convert energy into power. After all, power measures the time rate of energy transfer.

For our long codes, we write

$$y_T(t,\zeta) = \sum_{n=0}^{N-1} y_n p\left(\frac{t-nT_C}{T_C}\right)$$
$$= p\left(\frac{t}{T_C}\right) * \sum_{n=0}^{N-1} \delta(t-nT_C)$$
$$S_Y(f) = \lim_{T\to\infty} \frac{E\left\{\left|Y_T(f,\zeta)\right|^2\right\}}{T} \tag{9.42}$$

As $T \to \infty$, N also approaches infinity because $T = NT_C$.

From (9.9), we have our expression for $Y_T(f,\zeta)$.

$$\left|Y_T(f,\zeta)\right|^2 = T_C^2 N \operatorname{sinc}^2(\pi f T_C)\left|Y_{\text{code}}(f)\right|^2 \tag{9.43}$$

The expectation is given by

$$E\left\{\left|Y_T(f,\zeta)\right|^2\right\} = T_C^2 N \operatorname{sinc}^2(\pi f T_C) E\left\{\left|Y_{\text{code}}(f)\right|^2\right\} \tag{9.44}$$

This follows, because $Y_{\text{code}}(f)$ is the only term that contains random variables—the chips in the random code. The expectation of $\left|Y_{\text{code}}(f)\right|^2$ is found as follows.

$$E\left\{\left|Y_{\text{code}}(f)\right|^2\right\} = \frac{1}{N} E\left\{\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} y_n \exp(-j2\pi f n T_C) y_m \exp(+j2\pi f m T_C)\right\}$$
$$= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} E\{y_n y_m\}\exp\left(-j2\pi f(n-m)T_C\right)$$
$$= \frac{1}{N}\sum_{n=0}^{N-1}\exp\left(-j2\pi f(n-n)T_C\right)$$
$$= 1 \tag{9.45}$$

Now we can write

$$E\left\{\left|Y_T(f,\zeta)\right|^2\right\} = T_C^2 N \operatorname{sinc}^2(\pi f T_C)$$
$$S_Y(f) = T_C \operatorname{sinc}^2(\pi f T_C) \tag{9.46}$$

A similar derivation can be used when a long code modulates an RF carrier. The results are summarized as follows.

$$y_T(f,\zeta) = \sum_{n=0}^{N-1} y_n p\left(\frac{t-nT_C}{T_C}\right)\cos(2\pi f_L t)$$
$$\left|Y_T(f,\zeta)\right|^2 = \frac{T_C^2 N}{2}\operatorname{sinc}^2\left(\pi(f-f_L)T_C\right)\left|Y_{\text{code}}(f-f_L)\right|^2$$
$$+\frac{T_C^2 N}{2}\operatorname{sinc}^2\left(\pi(f+f_L)T_C\right)\left|Y_{\text{code}}(f+f_L)\right|^2$$
$$S_Y(f) = \frac{T_C}{2}\left(\operatorname{sinc}^2\left(\pi(f-f_L)T_C\right)+\operatorname{sinc}^2\left(\pi(f+f_L)T_C\right)\right) \tag{9.47}$$

9.6.2 C/A-Codes*

The last section derived a compact and therefore agreeable expression for the power spectral density of direct-sequence spread-spectrum signals when the code is long and therefore well approximated by a random code. Sadly, this result does not apply to our work horse—the C/A-code. The C/A-code is so short that it repeats twenty times for each navigation bit, and so $20NT_C = T_B$. As we shall discover, this repetition gives rise to a line spectrum.

Consider an infinite string of data bits, $D_m = \pm 1$, each with duration $T_B = 20NT_C$, where $T_C \approx 1$ microsecond is the C/A-code chip duration and $N = 1023$ is the length of the C/A-code. In this case, we begin as follows.

$$x_1(t) = \sum_{n=0}^{N-1} x_n p\left(\frac{t-nT_C}{T_C}\right)$$
$$x_T(t,\zeta) = \sum_{m=0}^{M-1} D_m \sum_{\ell=0}^{19} x_1(t-\ell N T_C - mT_B)$$
$$= x_1(t) \star \left[\sum_{m=0}^{M-1} D_m \sum_{\ell=0}^{19} \delta(t-\ell N T_C - mT_B)\right]$$
$$= x_1(t) \star \left[\sum_{m=0}^{M-1} D_m p\left(\frac{t-\alpha_m}{T_B}\right)\sum_{\ell=-\infty}^{\infty} \delta(t-\ell N T_C)\right] \tag{9.48}$$

In this equation, $x_1(t)$ is a single period of the C/A-code. Twenty such periods are included in each navigation bit. We will model the code chips as deterministic. The overall signal is a stream of pulses corresponding to the navigation data bits, $p((t-\alpha_m)/T_B)$. Each pulse is shifted in time by $\alpha_m = ((m+0.5)T_B + 0.5T_C)$ and the polarity of each pulse is controlled by the navigation data, D_m. We will treat the data bits, D_m, as random. For our C/A-code analysis, M goes to infinity because $T = MT_B$.

We can find the corresponding Fourier transform if we employ the following transforms and properties.

$$\mathcal{F}\{x_1(t)*z(t)\} = X_1(f)Z(f)$$

$$\mathcal{F}\left\{\sum_{\ell=-\infty}^{\infty}\delta(t-\ell NT_C)\right\} = \frac{1}{NT_C}\sum_{\ell=-\infty}^{\infty}\delta\left(f-\ell/NT_C\right)$$

$$\mathcal{F}\left\{p\left(\frac{t-\alpha_m}{T_B}\right)\right\} = T_B\,\mathrm{sinc}(\pi f T_B)\exp(-j2\pi f\alpha_m)$$

The first of these transforms is in Table 8.1; the second is in Table 8.2; and the third can be derived from results in Tables 8.1 and 8.2. Using these results and some effort, we find the following.

$$X_T(f,\zeta) = \frac{T_B}{NT_C}X_1(f)\sum_{\ell=-\infty}^{\infty}\mathrm{sinc}\left(\pi T_B\left(f-\frac{\ell}{NT_C}\right)\right) \times\sum_{m=0}^{M-1}D_m\exp\left(-j2\pi\alpha_m\left(f-\frac{\ell}{NT_C}\right)\right)$$

Since the supports for the different $\mathrm{sinc}(\pi T_B(f-\ell/NT_C)$ functions are disjoint, we may write

$$\left|X_T(f,\zeta)\right|^2 = \left(\frac{T_B}{NT_C}\right)^2\left|X_1(f)\right|^2\sum_{\ell=-\infty}^{\infty}\left|\mathrm{sinc}\left(\pi T_B\left(f-\frac{\ell}{NT_C}\right)\right)\right|^2 \times\left|\sum_{m=0}^{M-1}D_m\exp\left(-j2\pi\alpha_m\left(f-\frac{\ell}{NT_C}\right)\right)\right|^2$$

This result is still pretty messy. However, we now take the expectation with respect to the random navigation data bits. More specifically, we leverage the following result.

$$E\{D_mD_k\} = \begin{cases}1 & \text{if } m=k\\ 0 & \text{if } m\neq k\end{cases} \tag{9.49}$$

With this help and some more effort, we may finally write

$$E\left\{\left|X_T(f,\zeta)\right|^2\right\} = \frac{MT_B^2}{(NT_C)^2}\left|X_1(f)\right|^2\sum_{\ell=-\infty}^{\infty}\mathrm{sinc}^2\left(\pi T_B\left(f-\frac{\ell}{NT_C}\right)\right) \tag{9.50}$$

$$S_{C/A}(f) = \frac{T_B}{(NT_C)^2}\left|X_1(f)\right|^2\sum_{\ell=-\infty}^{\infty}\mathrm{sinc}^2\left(\pi T_B\left(f-\frac{\ell}{NT_C}\right)\right) \tag{9.51}$$

where

$$\left|X_1(f)\right|^2 = T_C^2\,N\,\mathrm{sinc}^2(\pi f T_C)\left|X_{\mathrm{code}}(f)\right|^2 \tag{9.52}$$

As shown, $S_{C/A}(f)$ is quite a complicated beast for the C/A-code. It is the product of three functions of frequency. The first of these is $\mathrm{sinc}^2(\pi f T_C)$, which is due to the rectangular chip waveform used by GPS and can never be greater than unity. The second is $\left|X_{\mathrm{code}}(f)\right|^2$, which

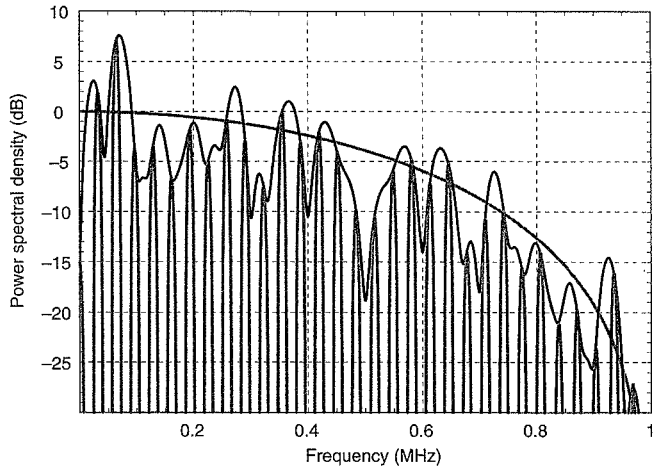


Figure 9.18 Power spectral density of two periods of the length-31 Gold code clocked at 1 Mcps. This figure showcases the impact of the code transform and line structure.

is plotted in Figures 9.16 and 9.17 for a length-31 Gold code. For the C/A-code, this term can be as large as 10 dB. The third function is the most wondrous. It arises from the 20 repeats of the C/A-code per navigation bit and consists of a comb of sinc functions. Each tooth of the comb is a sinc function with a null-to-null bandwidth of $2/T_B$ Hz, where T_B is the navigation bit duration. For GPS, the null-to-null bandwidth of an individual tooth is 100 Hz. These teeth are separated by $1/NT_C$ Hz, which is 1000 Hz for the GPS C/A-code. As such, there are 2000 such teeth within the ±2 MHz bandwidth of the C/A-code signal.

Figure 9.18 shows the power spectral density for our favorite example—a length-31 Gold code. The factor, $\mathrm{sinc}^2(\pi f T_C)$, for the rectangular chip waveforms is the smoothest gray curve. The product, $\mathrm{sinc}^2(\pi f T_C)\left|X_{\mathrm{code}}(f)\right|^2$, is the black curve. The product of all three correction factors, including the comb of sinc functions, is the rapidly varying gray curve. Since the code has length $N = 31$ and is clocked at 1 Mcps, the teeth are separated by $10^6/31 \approx 32{,}258$ Hz.

9.7 Narrowband Radio Frequency Interference

For many, the *raison d'être* of spread spectrum is its ability to combat radio frequency interference (RFI). RFI refers to any man-made signals that fall in the GPS portion of the radio spectrum. Most often, these signals are unintentional. However, GPS is a military system, and so we must also face the prospect of intentional jamming of the GPS band by an adversary during hostilities. In any event, RFI is worrisome for GPS because signals from terrestrial sources can be much stronger than the GPS signals that travel all the way from medium earth orbit to the user. Indeed, the power in a satellite signal decays with the distance squared, and so signals from these distant sources are at a significant disadvantage.

RFI comes in many flavors. For example, it can be pulsed or continuous. GPS is very

tolerant of pulsed RFI provided the pulses are short compared to the 20 ms duration of a navigation data bit. On the other hand, GPS has difficulty coping with continuous RFI. Continuous RFI can be further classified by its bandwidth. Broadband RFI has a bandwidth equal to or greater than the GPS bandwidth (2 MHz for the C/A-code). Narrowband RFI has a narrow bandwidth relative to GPS. Much of Chapter 13 is devoted to RFI. The present section simply describes the basic action that enables a spread spectrum system to attenuate narrowband RFI.

Consider the following GPS signal received in the presence of a narrowband interference signal.

$$\sqrt{P}\,D(t)x(t)+\sqrt{2P_J}\cos(2\pi f_J t+\theta_J) \tag{9.53}$$

Notice that we have ignored the underlying carrier signal for both the GPS signal and jammer ($f_J << f_{L1}$). Indeed, our analysis considers baseband signals only. This is done for simplicity, but the basic action of spread spectrum against narrowband RFI is faithfully captured by this simpler analysis.

The power in the baseband GPS signal is given by P, and the power in the interference signal is given by P_J. As received, the signal to interference ratio is simply P/P_J. For GPS, the received interfering power can be many times greater than the desired satellite signal power. Fortunately, the correlation process dramatically changes this balance of power.

9.7.1 Spreading the Interference

We begin with an intuitive analysis based on the power spectral densities shown in Figure 9.19. The top half of Figure 9.19 shows the spectrum of narrowband RFI superposed on the main lobe of the GPS signal spectrum. The desired signal spreads its power, P, across a bandwidth of $B_{code} = 1/T_C$ Hz. We have drawn the main lobe with $|X_{code}(f)| \approx 1$, and we show no line

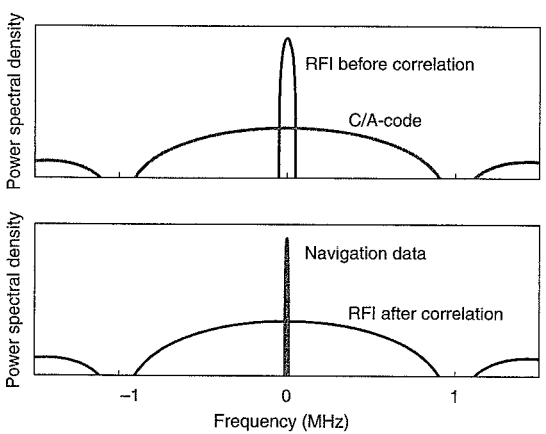


Figure 9.19 Processing gain of spread spectrum signals against narrowband radio frequency interference (RFI). The top half of the figure shows the spectrum of the GPS C/A-code and narrowband RFI prior to correlation with a code replica in the receiver. The bottom half of the figure shows the GPS and RFI spectra after correlation with an aligned replica.

spectrum. So we have assumed long, virtually random codes. In addition, f_J is in the middle of the band, $f_J << 1/T_C$. The RFI power, P_J falls in a bandwidth that is narrow compared to the code bandwidth. As mentioned above, P_J may be appreciably greater than P.

When the receiver correlates the incoming signals with replica codes, the spectrum of the narrowband jammer spreads as shown in the bottom half of the figure. After all, the receiver multiplies the incoming tone by a spread spectrum code. This action is akin to the modulation process in the GPS satellite, where the navigation data is multiplied by the satellite's signature sequence and the spectrum is spread. Now the RFI power, P_J, is spread across a bandwidth of $B_{code} = 1/T_C$ Hz. The correlation action has the opposite effect on the GPS signal. If the replica code is aligned with the incoming code, the receiver wipes off the incoming code and the GPS spectrum collapses to that of the navigation data. We say the incoming signal is de-spread, and the GPS power, P, is now concentrated in a bandwidth of only $B_{data} = 1/T_B = 50$ Hz.

In this de-spread bandwidth, GPS finally enjoys a significant power advantage over the radio frequency interference. To extract the navigation message, the receiver need only process the signals contained within the 50 Hz bandwidth of the navigation message, where the GPS power is now concentrated. A filter attenuates power outside of this de-spread bandwidth. This action removes most of the jammer power because we only find a fraction of the RFI power in this de-spread bandwidth. The remaining RFI power is only $P_J B_{data}/B_{code}$. Processing gain, PG, is the improvement in the signal-to-interference ratio due to this action.

$$PG = \frac{B_{code}}{B_{data}} = \frac{T_B}{T_C} \tag{9.54}$$

This analysis was based on a smooth spectrum for the GPS signal, without the wrinkles introduces by short codes or the C/A-code lines. The next section treats these complications, but can be omitted on a first reading.

9.7.2 Impact of Code and Line Spectra*

Our interest in this section is the same as the last. As received, the signal to interference ratio is P/P_J. We hope that correlation improves this ratio. Hence we will examine the impact of correlation on the sum given by (9.53). If the replica code is aligned with the received code, then

$$\begin{aligned} S+J &= x(t)\star\left(\sqrt{P}\,D(t)x(t)+\sqrt{2P_J}\cos(2\pi f_J t+\theta_J)\right) \\ S &= x(t)\star\sqrt{P}\,D(t)x(t) \\ J &= x(t)\star\sqrt{2P_J}\cos(2\pi f_J t+\theta_J) \end{aligned} \tag{9.55}$$

When aligned, the signal correlation is given by

$$\begin{aligned} S &= \frac{1}{T}\int_0^T \sqrt{P}\,D(t)x(t)x(t)\,dt \\ &= \sqrt{P}D \\ |S|^2 &= P \end{aligned} \tag{9.56}$$

If navigation bits are being demodulated, then we will integrate the incoming signal for $T = T_B = 20$ ms since this is the navigation bit duration. We will use $T = T_B$ in what follows, and $D = \pm 1$ is the navigation bit during that period.

The treatment of J is somewhat more involved.

$$\begin{aligned}J &= \frac{1}{T_B}\int_0^{T_B}\sqrt{2P_J}\cos(2\pi f_J t+\theta_J)x(t)\,dt \\ &= \frac{\sqrt{2P_J}}{2T_B}\left(\int_0^{T_B}\exp(j(2\pi f_J t+\theta_J))x(t)\,dt+\int_0^{T_B}\exp(-j(2\pi f_J t+\theta_J))x(t)\,dt\right) \\ &= \frac{\sqrt{2P_J}}{2T_B}\left(\exp(j\theta_J)\int_0^{T_B}x(t)\exp(j2\pi f_J t)\,dt\right. \\ &\qquad \left.+\exp(-j\theta_J)\int_0^{T_B}x(t)\exp(-j2\pi f_J t)\,dt\right)\end{aligned} \tag{9.57}$$

This equation makes use of Euler's formula. Even so, it is beginning to look messy. Happily, we recognize the integrals in the last line as $X_{T_B}(f_J)$ and $X_{T_B}(-f_J)$. They are the Fourier transforms of a T_B second piece of the code evaluated at f_J and $-f_J$. Since $x_{T_B}(t)$ is real, we may also write $X_{T_B}(-f_J)=X_{T_B}^*(f_J)$. All of this allows us to write

$$J = \frac{\sqrt{2P_J}}{2T_B}\left(\exp(j\theta_J)X_{T_B}^*(f_J)+\exp(-j\theta_J)X_{T_B}(f_J)\right) \tag{9.58}$$

We are interested in the magnitude of the interference squared.

$$\begin{aligned}|J|^2 &= \frac{P_J}{2T_B^2}\left(\exp(j\theta_J)X_{T_B}^*(f_J)+\exp(-j\theta_J)X_{T_B}(f_J)\right) \\ &\qquad\cdot\left(\exp(-j\theta_J)X_{T_B}(f_J)+\exp(j\theta_J)X_{T_B}^*(f_J)\right) \\ &= \frac{P_J}{2T_B^2}\left(2\left|X_{T_B}(f_J)\right|^2+\exp(j2\theta_J)X_{T_B}^*(f_J)X_{T_B}^*(f_J)\right. \\ &\qquad \left.+\exp(-2j\theta_J)X_{T_B}(f_J)X_{T_B}(f_J)\right)\end{aligned} \tag{9.59}$$

As shown, this expression is still a function of the interference phase, θ_J, relative to the baseband signal. This is annoying because we have no idea what that phase may be. After all, it is determined by the transmission time of the interfering transmitter and the length of the path from the interference source to the GPS receiver. We have no real knowledge of these variables. For this reason, we simply average over all possible phases.

$$\begin{aligned}E_\theta\left\{|J|^2\right\} &= \frac{1}{2\pi}\int_0^{2\pi}|J|^2\,d\theta \\ &= \frac{P_J}{T_B^2}\left|X_{T_B}(f_J)\right|^2\end{aligned} \tag{9.60}$$

If the codes or data are random, then we write

$$E_\theta\left\{|J|^2\right\} = \frac{P_J}{T_B^2}\left|X_{T_B}(f_J,\zeta)\right|^2 \tag{9.61}$$

where ζ represents the underlying random chips or bits. We then average over these random variables.

$$E_\zeta E_\theta\left\{|J|^2\right\} = \frac{P_J}{T_B^2}E_\zeta\left\{\left|X_{T_B}(f_J)\right|^2\right\} \tag{9.62}$$

This is the sweet result we have sought. Now we can compute the post-correlation signal to interference ratio and the processing gain.

$$\frac{|S|^2}{E\left\{|J|^2\right\}} = \frac{P}{\frac{P_J}{T_B^2}E\left\{\left|X_{T_B}(f_J)\right|^2\right\}}$$

$$PG = \frac{T_B^2}{E\left\{\left|X_{T_B}(f_J)\right|^2\right\}} \tag{9.63}$$

If the code is well modeled as a random code, then (9.46) provides the following.

$$PG \approx \frac{T_B}{T_C\,\mathrm{sinc}^2(\pi f T_C)} \tag{9.64}$$

If, in addition, f_J is near the band center, then $T_C f_J << 1$ and $\left|\mathrm{sinc}(\pi T_C f_J)\right|^2 \approx 1$. Hence

$$PG \approx \frac{T_B}{T_C} = \frac{B_{\text{code}}}{B_{\text{data}}} \tag{9.65}$$

Happily, this result agrees with (9.54).

For the C/A-codes, the results are more complicated. From (9.51) and (9.63), we find

$$PG = \frac{N}{\left|\mathrm{sinc}(\pi f_J T_C)\right|^2\left|X_{\text{code}}(f_J)\right|^2\sum_{\ell=-\infty}^{\infty}\left|\mathrm{sinc}\left(\pi T_B\left(f_J-\frac{\ell}{NT_C}\right)\right)\right|^2} \tag{9.66}$$

We use $M = 1$, because we are trying to demodulate one navigation bit. Using decibels, we write

$$\begin{aligned}PG_{\text{dB}} &= 10\log_{10}N-10\log_{10}\left(\left|\mathrm{sinc}(\pi f_J T_C)\right|^2\right)-10\log_{10}\left(\left|X_{\text{code}}(f_J)\right|^2\right) \\ &\quad -10\log_{10}\left(\sum_{\ell=-\infty}^{\infty}\left|\mathrm{sinc}\left(\pi T_B\left(f_J-\frac{\ell}{NT_C}\right)\right)\right|^2\right)\end{aligned} \tag{9.67}$$

As shown, processing gain is complicated for the C/A-code. It begins with the familiar gain of $10\log_{10}N = 30$ dB, but this term is followed by three correction terms corresponding to the main pieces of the PSD for the C/A-code. The first of these is $\left|\mathrm{sinc}(\pi f_J T_C)\right|^2$, which stems from the rectangular chip waveforms and can never be greater than unity. If interference is near the band center, $f_J << 1/T_C$, then this term is approximately unity (or 0 dB) and can be ignored. However, the correlator is a filter that will reject frequencies outside of those used by the GPS signals. Consequently, this term becomes smaller as the interference frequency moves away from the band center. If f_J is at the band edge or outside the band, processing gain improves.

The second correction term is $X_{\text{code}}(f_J)$, which comes from the C/A-code and can be as large as 10 dB. Hence, the processing gain against an interfering signal with the worst case frequency, f_J, is reduced to 20 dB.

The third correction arises from the twenty repeats of the C/A-code per navigation bit, and this is the comb of sinc functions. Each tooth of the comb is a sinc function with a null-to-null bandwidth of $2/T_B$ Hz, where T_B is the navigation bit duration. For GPS, the null-to-null bandwidth of an individual tooth is 100 Hz. These teeth are separated by $1/NT_C$ Hz, which is 1000 Hz for the GPS C/A-code. If f_J falls on a tooth, then the processing gain is given by

$$PG_{\text{dB}} = 10\log_{10} N - 10\log_{10}\left(\left|\text{sinc}(\pi f_J T_C)\right|^2\right) - 10\log_{10}\left(\left|X_{\text{code}}(f_J)\right|^2\right) \qquad (9.68)$$

If f_J also is near the band center and not near any of the hot spots given by $X_{\text{code}}(f)$, then the processing gain for the C/A-code is approximately $10\log_{10}N$ or 30 dB. This is disappointing because it means that we did not get any credit for the twenty repeats of the code. In other words, we may have hoped for a processing gain of $10\log_{10}20N$, but we did not get it.

If f_J is between the teeth, then narrowband RFI has a much smaller impact. Unfortunately, the teeth move with the Doppler frequency of the satellite signal, and so narrowband RFI will probably fall on top of a tooth for one satellite or another. For this reason, the spread spectrum codes discussed in Sections 9.8 and 9.9 either are longer than a navigation bit or repeat once per navigation bit.

We conclude this section with a sobering note. As we shall discover in Chapters 10 and 13, the received GPS signals are very weak, and terrestrial signals can easily overwhelm the processing gain provided by the spread spectrum codes. For this reason, much of Chapter 13 is devoted to additional RFI countermeasures.

9.8 P(Y) Codes on L1 and L2

As we have discussed, GPS currently broadcasts P(Y) codes on f_{L1} and f_{L2}. This section provides a brief description based on Spilker (1996), and the reader should refer to that authoritative work for more detail.

The designation P(Y) refers to two codes, where the P code is known to the public and the Y code is not. The P code was broadcast for the early years of GPS, but the Y code has long since replaced the P code. The secrecy of the Y code prevents ill-intentioned individuals from spoofing authorized users of GPS. Spoofing is distinct from jamming. Jamming simply knocks the GPS receiver out of commission by overwhelming the GPS signal with strong inband signals. This is troublesome of course, but the receiver can detect jamming and warn the user that no position fix is available. Spoofing is more subtle and more dangerous. In this case, the adversary sends a signal that resembles the GPS signal. The goal is to introduce an error without detection. If the receiver employs reasonability checks on the GPS measurements, then spoofing of the P code or even the C/A-code is already difficult. Since the Y code is unknown to the attacker, spoofing of the Y code is a formidable task.

Relative to the C/A-codes, the P and Y codes are fast. As we have already observed, the P and Y codes chip at 10.23 Mcps, and the null-to-null bandwidth is 20.46 MHz. This ten-fold increase in chipping rate brings the benefits described in Sections 9.2 and 9.7: ranging precision and a 10 dB increase in the processing gain against RFI.

P codes are also very long. A multiplicity of maximal length shift registers are used to create a single sequence that is approximately 38 weeks long (even at 10 Mcps!). P codes are not Gold codes. Rather, this single 38-week long sequence is chopped into 37 non-overlapping sequences each with a period of one week. Thus P codes are available for all the satellites in the GPS constellation and some number of ground transmitters (pseudo-satellites or pseudolites), and all codes have very low correlation sidelobes by virtue of their length.

Needless to say, the receiver needs some kind of help while synchronizing its P code replica to the received P code. That help can come from the C/A-code and data in the navigation message. The receiver synchronizes to the short C/A-code and demodulates the navigation message. Within this information, it finds data that describes the current contents of one of the shift registers used to generate the P code. With this help, P code synchronization is accomplished.

9.9 New Civil Signals for GPS

New signals are coming for GPS. Civil signals will be added at f_{L2} = 1227.60 MHz and f_{L5} = 1176.45 MHz. These are shown in Figure 9.20 along with the incumbent C/A and P(Y) signals. The new civilian signal for f_{L2} is described in the next subsection, and the new civilian signal for f_{L5} is described in the subsequent subsection. Finally, Section 9.9.3 describes a nifty feature of both signals—data free signal components. Interestingly, the two new signals will realize this feature in very different ways, but with a common goal—to achieve better performance in weak signal-to-noise ratio environments.

New military signals will be added at f_{L1} and f_{L2}. We will have little to say about these signals, except that they are binary offset carrier (BOC) signals, which are described briefly in Section 9.10.

9.9.1 New Civil Signal at L2

New civil and military signals are planned for f_{L2} = 1227.60 MHz. This brief description is based on the fine paper by Fontana *et al.* (2001). With the new signals, the L2 signal will become

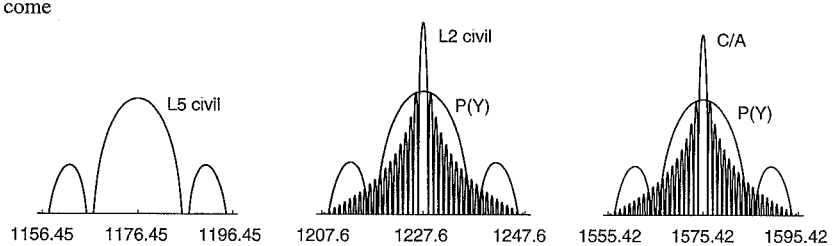


Figure 9.20 GPS plans to add new civilian signals at L2 and L5. New signals from the European satellite navigation system, Galileo, will be added as spectral neighbors.

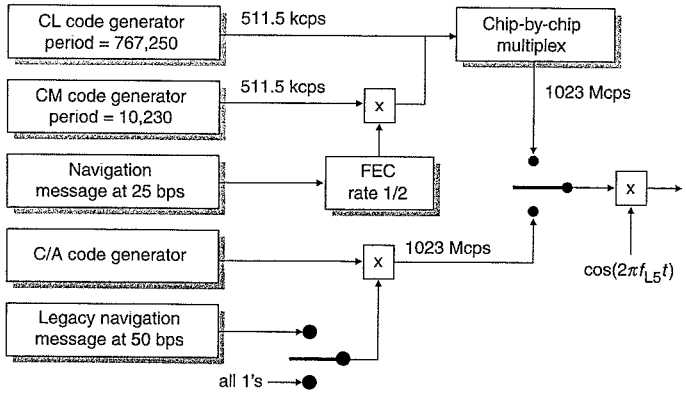


Figure 9.21 Block diagram for generating the new civil signal at L2.

$$s_{L2}(t) = \sqrt{2P_{Y2}}D(t)y(t)\sin(2\pi f_{L2}t + \theta_{L2}) + \sqrt{2P_{C2}}D_{C2}(t)z(t)\cos(2\pi f_{L2}t + \theta_{L2}) + \text{new military signal using the M-code} \quad (9.69)$$

The first line in (9.69) contains the incumbent military signal with P(Y) code modulation, which also appears in (9.1). The second line contains the new civil signal at L2. The third line is for the new military signal at L2, which we describe briefly in Section 9.10.

The new civil signal is further detailed in Figure 9.21, which is a functional block diagram of the signal generator. The new code, $z(t)$, called L2C, and the L2C-codes have the same 1.023×10^6 chipping rate as the C/A-codes and so the null-to-null bandwidth is still 2.046 MHz. Since the chipping rate is the same as for the C/A-code, the new civil signal on L2 has the same ranging precision and multipath performance as the C/A-code.

However, the L2C-codes are significantly longer than the C/A-codes. In fact, two codes, called CM (for medium-length code) and CL (for long code), are time multiplexed to form the RC code. The CM code has a length of 10,230 chips, and so it is ten times longer than the C/A-code. The CL code is even longer with a period of 767,250 chips. Neither of these codes can be a Gold code because $10{,}230 \neq 2^n - 1$ and $767{,}250 \neq 2^n - 1$ for any n. However, the CM and CL codes are constructed using linear shift registers, and so they do have a kinship to the C/A-codes. As discussed in Sections 9.2.2 and 9.3, increased code length decreases the autocorrelation and cross-correlation sidelobes relative to the C/A-code. As discussed in Section 9.7, the increased length also gives better performance against narrowband RFI because long codes are more nearly random.

As shown in Figure 9.22, the two codes alternate, with the CM code controlling every second chip that is broadcast. In this way, the two codes are time multiplexed. The CL code is not modulated by navigation data. This data-free signal is very helpful for operation in low signal-to-noise ratio environments as we shall discuss in Section 9.9.3. The CM code, in contrast, is

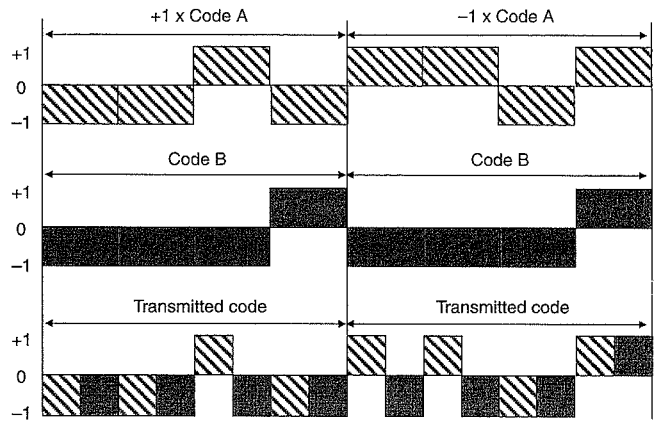


Figure 9.22 Chip splitting to achieve a data-free signal component. Code A has four chips, {−1, −1, +1, −1} and is modulated by data {+1, −1}. Code B also has four chips, {−1, −1, −1, +1}, but is not modulated by the data. The transmitted code splits each chip in half with the first half of the chip from code A and the second half of the chip from Code B.

modulated by navigation data. This data is encoded for forward error correction (FEC) and this feature is also discussed in Section 9.9.3.

Future GPS satellites will be able to broadcast the L2C code or the C/A-code. If the C/A-code is broadcast, the satellite will be able to include the navigation data or switch to a data-free C/A-code. The switches in Figure 9.21 allow for these contingencies.

9.9.2 New Civil Signal at L5

A new civil signal is also planned for $f_{L5} = 1176.45$ MHz. This brief description is based on the authoritative paper by Spilker and Van Dierendonck (2002), and the reader is referred to that paper for the details.

The new L5 signal for one satellite is

$$s_{L5}(t) = \sqrt{2P_{I5}}D_{I5}(t)h_I(t)g_I(t)\cos(2\pi f_{L5}t + \theta_{L5}) + \sqrt{2P_{Q5}}h_Q(t)g_Q(t)\sin(2\pi f_{L5}t + \theta_{L5}) \quad (9.70)$$

No military signal is planned for L5! Consequently, the inphase, $\cos(2\pi f_{L5}t + \theta_{L5})$, and quadrature, $\sin(2\pi f_{L5}t + \theta_{L5})$, components are both given over to civil signals. This rich signal is further detailed in Figure 9.23, which is a functional block diagram of the signal generator. As shown, the inphase signal is modulated with navigation data, $D_{L5}(t)$, and a spreading code, $h_I(t)g_I(t)$. The quadrature component is modulated with the $h_Q(t)g_Q(t)$ code, but no navigation data is applied. Like the new L2 signal, the L5 signal will also provide a data-free signal component, and this feature is further discussed in Section 9.9.3. Both codes, $h_I(t)g_I(t)$ and $h_Q(t)g_Q(t)$, will be sent at rates of 10.23 Mcps, and so the null-to-null bandwidth will be 20.46 MHz. This ten-fold increase in chipping rate will bring the ranging and multipath

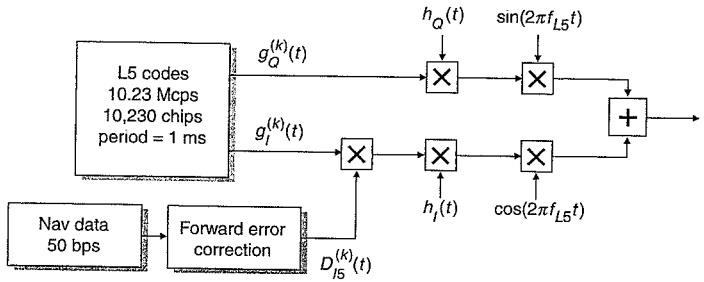


Figure 9.23 Block diagram for generating the new GPS signal at L5. Each satellite will have a unique pair of codes, $g_Q^{(k)}(t)$ and $g_I^{(k)}(t)$, but the Neumann-Hoffman codes, $h_Q(t)$ and $h_I(t)$, are the same for all satellites.

benefits described in Section 9.2. It also brings a 10 dB increase in the processing gain against RFI.

The new L5 codes are also longer than the C/A-codes. By themselves, the codes, $g_I(t)$ and $g_Q(t)$, are 10,230 chips in length. At their chipping rate of 10.23 Mcps, these codes repeat every millisecond. Since the new codes are ten times longer than the C/A-codes, their auto-correlation and cross-correlation sidelobes are approximately 10 dB lower than the sidelobes for the C/A-code. Like the C/A-codes and the new L2 codes, the L5 codes are also built using linear feedback shift registers.

As shown in Figure 9.23, the codes, $g_I(t)$ and $g_Q(t)$, are multiplied by Neumann-Hoffmann (N-H) codes, $h_I(t)$ and $h_Q(t)$. The code, $h_I(t)$, is only ten chips long, and the entire code is shown in the top half of Figure 9.24. The duration of each $h_I(t)$ chip is 1 ms, and so one chip corresponds to one period of the $g_I(t)$ code. Every $h_I(t)$ chip multiplies an entire $g_I(t)$ code.

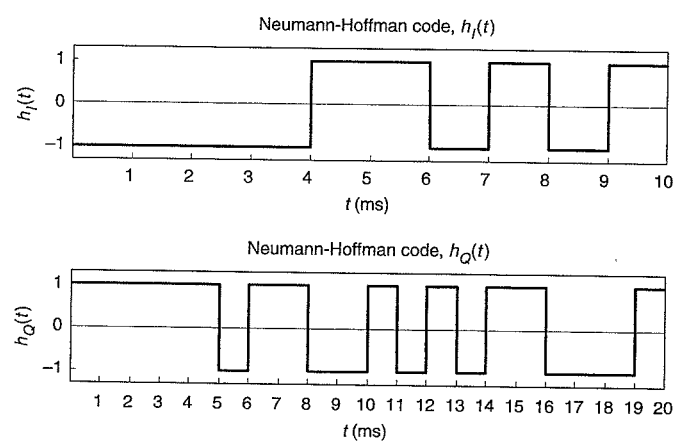


Figure 9.24 Neumann-Hoffman code. The $h_I(t)$ code in the top trace modulates the inphase signal component for L5. The $h_Q(t)$ code in the bottom trace modulates the quadrature component for L5.

This action extends the length of the inphase code from 10,230 chips to 102,300 chips. With a chipping rate of 10.23 Mcps, the $h_I(t)g_I(t)$ code period is 10 ms long and is equal to the duration of one L5 navigation bit. Similarly, $h_Q(t)$ is twenty chips long and each chip is 1 ms long. Thus, $h_Q(t)$ extends the length of the quadrature code from 10,230 chips to 204,600 chips, and $h_Q(t)g_Q(t)$ has a period of 20 ms.

The N-H codes are chosen for their auto-correlation properties, and so they are similar to m-sequences. N-H codes are chosen over m-sequences, because $10 \neq 2^n - 1$ and $20 \neq 2^n - 1$ for any n. Their differences are further explored in Homework Problem 9-5.

If the codes, $g_I(t)$ and $g_Q(t)$, were used without the N-H codes, the spectral lines would be 100 Hz apart. The product codes, $h_I(t)g_I(t)$ and $h_Q(t)g_Q(t)$, squeeze this separation down to 10 Hz. At 10 Hz, they are so close that the spectrum becomes continuous because they are closer than the navigation data bandwidth of 50 Hz. In this case, the line structure may be ignored. However, the use of product codes does not garner the $1/\sqrt{N}$ reduction in correlation sidelobes. The product structure is simply not random, and so the advantage found with random sequences should not be expected.

9.9.3 Navigation Data and Data-Free Transmission

As described in Sections 9.9.1 and 9.9.2, both new signals have data-free components. Since the new L2 signal must co-exist with the existing military signal, it time multiplexes the CM and CL codes. As shown in Figure 9.22, the CL code (represented by a very short Code B) is not modulated with data, and the CM code (represented by a very short Code A) is modulated with data. The new L5 signal need not co-exist with a military signal, and so the data-free signal is in phase quadrature with the signal that has data modulation.

Data-free signal components are very useful in low signal-to-noise ratio (SNR) environments. Since the navigation bits are not known *a priori*, today's GPS receiver must square the received signal to strip the navigation bits. This squaring action is fully described in Chapter 12. It achieves the intended purpose of removing the data modulation, but it also introduces squaring loss. Squaring loss arises simply because the received noise is also squared. Unfortunately, squaring loss is greatest in low SNR environments where the receiver can least afford it. The new signals will obviate squaring loss.

If the receiver is ranging based on data-free signals, then it must obtain the navigation data from another source. Fortunately, GPS receivers, for city use, are often integrated with cell phones; and so the cell signal can include the GPS navigation message. Alternatively, the receiver can predict future values of ephemeris and clock based on earlier navigation data received in more favorable signal environments. These ideas belong to a concept called assisted GPS (AGPS) that will be discussed in Chapter 13.

The provision of data-free signals means that the full satellite power is not available for the signal components that do send navigation data. This loss is compensated by using forward error correction or FEC. FEC adds redundant bits to the transmitted message. These are akin to parity bits. While parity bits are used to detect bit errors, FEC can be used to detect and correct bit errors. The error detection capability increases with the fraction of redundant bits in the transmission. The navigation messages at L2 and L5 will use rate 1/2 codes. In other words, they will contain two encoded symbols for every raw data bit. The L2 message will send raw navigation data at 25 bps, and so the FEC-encoded rate will be 50 symbols per second. The L5 message will send navigation data at 50 bps, and so the encoded rate will be 100 symbols/sec-

ond. FEC will reduce the signal power needed for reliable data demodulation by 5 dB. In this way, FEC successfully compensates for the power split between the signal component with data and without data.

9.10 Binary Offset Carrier Signals*

Binary offset carrier (BOC) signals are the newest flavor of GNSS signals [Betz (2001), Hegarty and Tran (2002)], and they are easy to describe. So far, all of our GNSS signals have used a chip waveform equal to a rectangular pulse, $p(t/T_C)$. BOC signals simply replace this rectangular pulse with the following chip waveform.

$$\sum_{m=0}^{M-1}(-1)^m p\left(\frac{t-mT_S}{T_S}\right) \tag{9.71}$$

This waveform is shown in the top trace of Figure 9.25 for $M = 4$. As shown, the chip waveform is chopped into M shorter rectangles each with duration $T_S = T_C/M$. These *subchips* have alternating sign.

The Fourier transform is also easy to find. For long sequences with $NT_C = T_B$, we can approximate the spectrum of the existing GPS signal as follows.

$$\mathcal{F}\left\{\sqrt{2P}\,x_1(t)\cos(2\pi f_{L1}t)\right\} \approx \sqrt{\frac{NP}{2}}\mathcal{F}\left\{p(t/T_C)\right\}\Big|_{f=f-f_L} + \sqrt{\frac{NP}{2}}\mathcal{F}\left\{p(t/T_C)\right\}\Big|_{f=f+f_L} \tag{9.72}$$

With GNSS sequences, the chip waveform really dominates the spectral landscape.

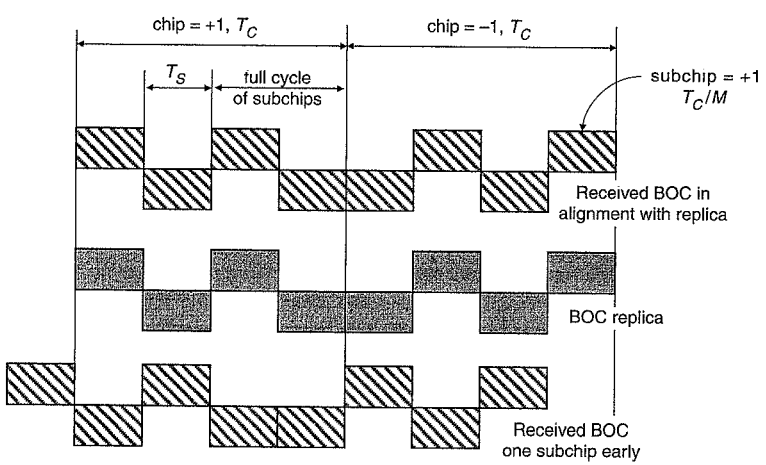


Figure 9.25 Diagram to help derive the auto-correlation function for binary offset carrier signals. This example uses $M = 4$ subchips per chip.

For our new chip waveform, we simply plan to replace $\mathcal{F}\left\{p(t/T_C)\right\}$ with the Fourier transform of (9.71).

$$\mathcal{F}\left\{\sum_{m=0}^{M-1}(-1)^m p\left(\frac{t-mT_S}{T_S}\right)\right\} = \int_{-\infty}^{\infty}\sum_{m=0}^{M-1}(-1)^m p\left(\frac{t-mT_S}{T_S}\right)\exp(-j2\pi ft)\,dt$$
$$= \sum_{m=0}^{M-1}(-1)^m \int_{mT_S-T_S/2}^{mT_S+T_S/2}\exp(-j2\pi ft)\,dt \tag{9.73}$$

If we substitute, $\tau = t - mT_S$, then we find

$$\mathcal{F}\left\{\sum_{m=0}^{M-1}(-1)^m p\left(\frac{t-mT_S}{T_S}\right)\right\} = \sum_{m=0}^{M-1}(-1)^m \int_{-T_S/2}^{+T_S/2}\exp(-j2\pi f(\tau+mT_S))\,d\tau$$
$$= \sum_{m=0}^{M-1}(-1)^m \exp(-j2\pi fmT_S)\int_{-T_S/2}^{+T_S/2}\cos(2\pi f\tau)\,d\tau$$
$$= \sum_{m=0}^{M-1}(-1)^m \exp(-j2\pi fmT_S)T_S\,\mathrm{sinc}(\pi fT_S)$$
$$= T_S\,\mathrm{sinc}(\pi fT_S)\sum_{m=0}^{M-1}(-1)^m \exp(-j2\pi fmT_S) \tag{9.74}$$

The amplitude spectrum of this BOC chip is compared to $\left|\mathcal{F}\left\{p(t/T_C)\right\}\right|$ in Figures 9.26 and 9.27. These figures also introduce the accepted notation for BOC signals. The notation BOC(α,β) has the following meaning. β is the chipping rate normalized to 1.023 Mcps. α is

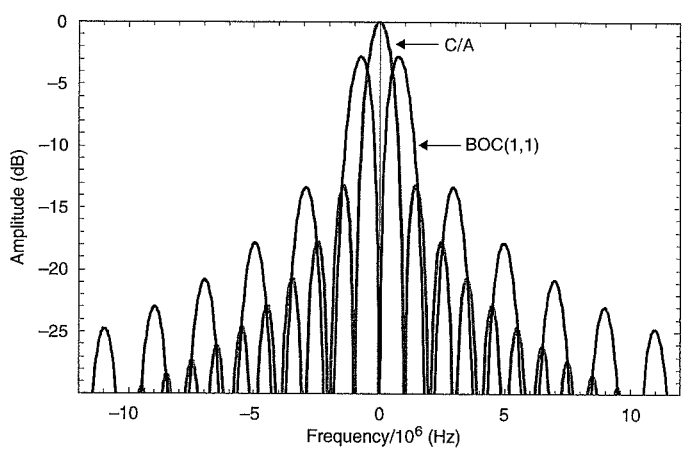


Figure 9.26 Amplitude spectrum for the C/A-code and a binary offset carrier, BOC(1,1). The BOC signal has a chipping rate of 1×10^6 and a subcarrier frequency of 1 MHz, which yields two subchips per chip ($M = 2$).

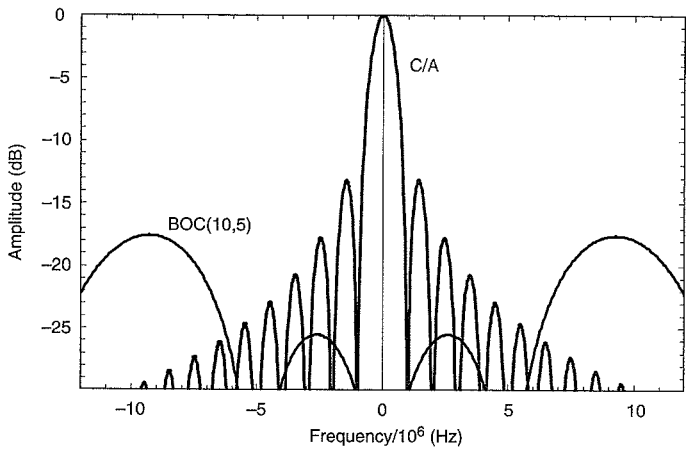


Figure 9.27 Amplitude spectrum for the C/A-code and a binary offset carrier, BOC(10,5). The BOC signal has a chipping rate of 5×10^6 and a subcarrier frequency of 10 MHz, which yields four subchips per chip ($M = 4$).

the subcarrier frequency normalized to 1.023 Mcps, where the subcarrier frequency is $1/2T_S$. Consequently, the number of subchips per chip is $M = 2\alpha/\beta$.

For example, BOC(1,1) has a chipping rate of 1.023 Mcps and a subcarrier frequency of 1.023×10^6 cycles per second (cps). A full cycle includes two subchips and so BOC(1,1) sends two subchips per chip ($M = 2$). As another example, BOC(10,5) has a chipping rate of 5.115 Mcps and a sub-chipping rate of 10.23×10^6 cps. Each cycle contains two subchips as shown in Figure 9.25, and so this signal sends four subchips per chip.

Figure 9.26 shows the spectrum for the C/A-code and BOC(1,1). As shown, BOC(1,1) has no energy at the band center, but pushes its power out to ±1 MHz. Happily, the subcarrier frequency, α, identifies the location of the spectral peak for BOC signals.

The new military signals for GPS are BOC(10,5), and these are compared to the C/A-code in Figure 9.27. As shown, the new military signals place their power at approximately ±10 MHz relative to the C/A-code at band center. This spectral separation is a key motivation for BOC. During a military conflict, the U.S. military could send a strong radio signal from a ground based transmitter to jam GPS users in the battle area. The spectrum of this jamming signal would be similar to the spectrum of the C/A-code and would deny all use of the civil signal. However, it would not interfere with the operation of the new military signals. This capability is called selective denial. Since the jammer is ground based, it would not interfere with civil operations well outside of the battle area.

The auto-correlation functions for BOC signals differ significantly from those for the C/A-codes. We can derive their auto-correlation functions in short order. Consider Figure 9.25. The replica code is shown in the middle. An aligned code is shown in the top trace, and the bottom code is one subchip early.

Let $R_{\text{BOC}}(\tau)$ denote the auto-correlation function for the BOC signal. Further let $R_{\text{code}}(\tau)$ denote the auto-correlation function for the underlying code absent BOC subchips. In other words, it is the auto-correlation function for traditional rectangular chips. With the aligned code, we can readily write

$$\begin{aligned} R_{\text{BOC}}(\tau = 0) &= \frac{1}{T_{\text{code}}} \int_0^{T_{\text{code}}} x^2(t)\,dt \\ &= \frac{T_S}{T_{\text{code}}}(\#\text{ agreements} - \#\text{ disagreements}) \\ &= \frac{T_S}{T_{\text{code}}} MN \\ &= 1 \end{aligned} \tag{9.75}$$

In this equation, $T_{\text{code}} = MNT_S$ because we have M subchips per chip and N chips per code period.

If the code shifts by an integer number of whole chips (not subchips), then

$$\begin{aligned} R_{\text{BOC}}(\tau = iT_C + 0T_S) &= \frac{T_C}{T_{\text{code}}} \sum_{n=0}^{N-i} x_n x_{n+i} \\ &= \frac{T_C}{T_{\text{code}}}(\#\text{ agreements} - \#\text{ disagreements}) \\ &= R_{\text{code}}(\tau = iT_C) \end{aligned} \tag{9.76}$$

When $\tau = iT_C + 0T_S$, we do not need to worry about the subchips, and $R_{\text{code}}(\tau = iT_C)$ is the auto-correlation function of the code irrespective of the chip waveform. The agreements and disagreements are between chips and not subchips because the subchips are all aligned.

However, if the code shifts by i whole chips and m subchips, $\tau = iT_C + mT_S$, then the subchips come into play and the auto-correlation is a weighted sum of the $R_{\text{code}}(\tau = iT_C)$ and $R_{\text{code}}(\tau = (i+1)T_C)$. For $m = 1$ and $M = 4$, we find

$$\begin{aligned} R_{\text{BOC}}(\tau = iT_C + T_S) &= \frac{4T_S}{T_{\text{code}}} \sum_{n=0}^{N-i} x_n x_{n+i} \\ &= -\frac{3}{4} R_{\text{code}}(iT_C) - \frac{1}{4} R_{\text{code}}((i+1)T_C) \end{aligned} \tag{9.77}$$

The weights are –3/4 and –1/4 because three subchips fall on top of the chip with shift i and only one falls on the chip with shift $i + 1$. The minus signs arise because of the alternating signs in the BOC subchips. For $m = 2$ and $M = 4$, we find

$$R_{\text{BOC}}(\tau = iT_C + 2T_S) = \frac{2}{4} R_{\text{code}}(iT_C) + \frac{2}{4} R_{\text{code}}((i+1)T_C) \tag{9.78}$$

In general, we write

$$R_{\text{BOC}}(\tau = iT_C + mT_S) = (-1)^m \left(\frac{M-m}{M} R_{\text{code}}(iT_C) + \frac{m}{M} R_{\text{code}}((i+1)T_C) \right) \tag{9.79}$$

This expression can be simplified based on the approximation

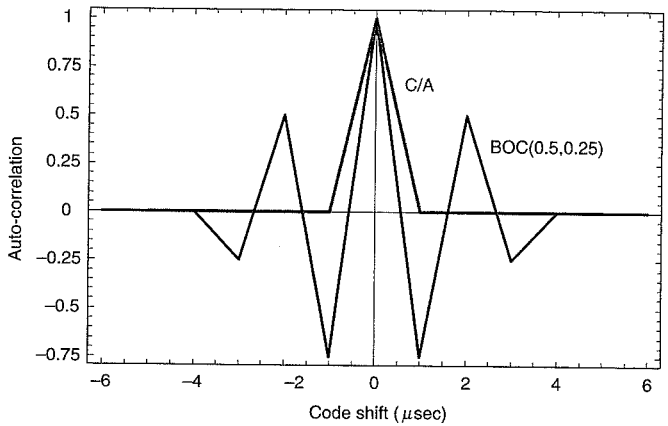


Figure 9.28 Auto-correlation function chip waveform for the C/A-code and a binary offset carrier, BOC(0.5,0.25). The BOC signal has a chipping rate of 10^6 cps and a subcarrier frequency of 0.5 MHz, which yields four subchips per chip ($M = 4$).

$$\begin{aligned} R_{\text{code}}(0) &= 1 \\ R_{\text{code}}(iT_C) &\approx 0 \quad \text{for } i \neq 0 \end{aligned} \tag{9.80}$$

We may now write

$$\begin{aligned} R_{\text{BOC}}(\tau = mT_S) &\approx (-1)^m \left(\frac{|M-m|}{M} \right) \quad \text{for } m = \{-M, -(M-1), \cdots M\} \\ &\approx 0 \quad \text{otherwise} \end{aligned} \tag{9.81}$$

Notice that this result is for $\tau = mT_S$. For intermediate values of τ, the values at $\tau = mT_S$ are connected with a straight line.

Figure 9.28 compares two auto-correlation functions. The gray curve is $\overline{R}(\tau)$ for the C/A-code, and the black curve is $R_{\text{BOC}}(\tau)$ for the BOC(0.5, 0.25) signal. These two signals have an important kinship. The C/A-code employs a chipping rate of 1 Mcps and no subchips. The BOC(0.5, 0.25) code employs a chipping rate of 0.25 Mcps. However, it sends four subchips per chip and so the subchipping rate is the same as the C/A-code's chipping rate. In this way, the two codes are quite similar.

In spite of their kinship, the BOC signal has a sharper auto-correlation peak. We know this to be a good quality. However, the BOC(0.5,0.25) curve also has false peaks at $\tau = \pm 1$ μsec, and these peaks are almost as strong as the main peak. If we were to confuse these pretenders for the main peak, we would suffer a ranging error of 300 m for our BOC(0.5, 0.25) example. The reader is referred to Betz (2001) for more on this interesting challenge.

9.11 Summary

This chapter defined and analyzed the signals used by GPS today and in the future. Most of the chapter concerned itself with the civil GPS signal at L1 because this signal is the basis for the vast majority of GNSS applications to date. However, we also provided brief treatments of the P(Y) codes used by the military, the planned new civil signals for the GPS L2 and L5 frequencies, and the BOC signals that will be used by the new GPS military signals and Galileo.

The chapter began with a focus on the civil signal broadcast on the GPS L1 frequency. It taught some radio signal basics including frequency, wavelength, and binary phase shift keying. We used convolution and the unit impulse function to provide an alternate description of the L1 signal. This alternate description enabled us to approximate the amplitude spectrum of the GPS signal.

We then turned our attention to the spread-spectrum codes used by all global navigation satellite systems (GNSS). These sequences are designed to provide special auto-correlation and cross-correlation properties, and so we discussed these functions at some length. We used random codes to deepen our understanding of spread spectrum signaling. We discussed the virtues of spread spectrum for ranging precision and signal acquisition. We also discussed channel sharing via code division multiple access (CDMA).

We introduced three correlation functions that will have tremendous utility going forward and so we review them here. First, the auto-correlation function for deterministic sequences like the pseudorandom codes actually used by GPS is

$$\begin{aligned} R(\tau) &= R(iT_C)\left(i+1-\frac{\tau}{T_C} \right) + R((i+1)T_C)\left(\frac{\tau}{T_C} - i \right) \\ R(\tau = iT_C) &= \frac{T_C}{T_{\text{code}}} \sum_{n=0}^{N-1} x_n x_{n+i} \end{aligned} \tag{9.82}$$

Second, the average auto-correlation function when the codes are random is

$$\overline{R}(\tau) = \begin{cases} \dfrac{\tau}{T_C} + 1 & \text{if } -T_C < \tau < 0 \\ \dfrac{-\tau}{T_C} + 1 & \text{if } 0 < \tau < T_C \\ 0 & \text{otherwise} \end{cases} \tag{9.83}$$

Third, the cross-correlation function is

$$\begin{aligned} R^{(k,\ell)}(\tau) &= R^{(k,\ell)}(iT_C)\left(i+1-\frac{\tau}{T_C} \right) + R^{(k,\ell)}((i+1)T_C)\left(\frac{\tau}{T_C} - i \right) \\ R^{(k,\ell)}(\tau = iT_C) &= \frac{T_C}{T_{\text{code}}} \sum_{n=0}^{N-1} x_n^{(k)} x_{n+i}^{(\ell)} \end{aligned} \tag{9.84}$$

By the way, Chapter 11 will introduce the fourth member of this mighty family—the ambiguity function, which will be denoted $\tilde{R}(\Delta\tau, \Delta f_D)$.

Table 9.1. Summary of the present and future GPS codes

	C/A-code on L1	*P(Y) code on L1 and L2*	*Civil signal for GPS L2*	*Civil signal for GPS L5*
Power relative to C/A-code on L1	0 dB	–3 dB	0 dB	+4 dB
Chipping rate and null-to-null bandwidth	1.023 Mcps 2.046 MHz	10.23 Mcps 20.46 MHz	1.023 Mcps 2.046 MHz	10.23 MHz 20.46 Mcps
RFI resistance relative to C/A code on L1	0 dB	>7 dB	0 dB	>14 dB
Code length	1023 chips	P: 1 week	CM: 10,230 chips CL: 767,250 chips	I: 102,300 chips Q: 204,600 chips
Data free component	no	no	yes chip splitting	yes inphase and quadrature
Error correction	no	no	yes	yes

After introducing auto-correlation and cross-correlation, we turned away from random sequences and considered the deterministic machines used to generate the actual GNSS codes. Specifically, we studied linear shift register sequences. We paid particular attention to maximal length linear shift register sequences (or *m*-sequences), because these are building blocks for the codes used by GPS. We also described Gold codes of length 1023 because these include the C/A-codes used by GPS. We also discussed length 31 Gold codes because these are much more manageable as examples. We presented results on both the auto-correlation and cross-correlation of these Gold codes.

The chapter then computed the power spectral densities of the GPS signals and connected these spectrum calculations to the processing gain that spread spectrum signals provide against narrowband radio frequency interference (RFI). We found that the analysis of the C/A-code is quite involved because it repeats twenty times for each navigation bit. This repetition introduces a line spectrum that requires some thought. If a spread spectrum code repeats exactly once for each navigation bit, then it is appreciably easier to analyze. In both cases, we derived expressions for the processing gain against RFI and quantified one of the key strengths of spread spectrum signals.

The remainder of the chapter surveyed the other GNSS signals—present and future. Specifically, we expanded our study of spread spectrum to include the P(Y) codes, used primarily by the military, on the L1 and L2 frequencies. We also looked at the new codes proposed for civil use on L2 and L5. Finally, we discussed binary offset carriers because this modulation technique is proposed for the new military signals for GPS, Galileo, and may someday replace the C/A-codes on the GPS L1 transmissions.

Table 9.1 offers a summary comparison of all of these codes. Notice that the new civil signal at L5 will have at least 14 dB more strength against RFI than the C/A-codes. This strength derives from two sources. First, the chipping rate is ten times faster, and so the processing gain will increase by 10 dB. Second, the received signal is at least 4 dB stronger.

Homework Problems

9-1. Review Problems 2-1 and 2-2 from Chapter 2. If you did not do these problems already, this would be an excellent time to try them! The next two problems explore code lengths other than those used by GPS.

9-2. Generate *m*-sequences of length 7, 15 or 63 (or all three!). Show the linear shift registers used in each case. Verify that the sequences have maximum length and that the auto-correlation function has only four values for time shifts equal to integer multiplies of a chip duration. Shift the feedback taps until the resulting sequence does not have maximum length. Compute the auto-correlation function for this new sequence and comment. Hint: Suitable generator polynomials for *m*-sequences can be found in Sarwate and Pursley (1980).

9-3. Generate Gold sequences of length 7 or 63. Draw the code generators that combine your selected *m*-sequences. Predict the values of the auto- and cross-correlation sidelobes. Compute example auto- and cross-correlation functions and check for the predicted sidelobes. Attempt to create Gold sequences by using *m*-sequences that are not maximally connected. Show the resulting auto-correlation functions and comment. Hint: Suitable pairs of *m*-sequences are called maximally connected, and such pairs are specified in Sarwate and Pursley (1980).

9-4. The auto- and cross-correlation functions used in this chapter are called periodic or circular because the underlying codes repeat infinitely. In contrast, aperiodic or linear correlation functions can be defined as follows.

$$R(\tau) = \frac{1}{T_{\text{code}}} \int_{\tau}^{T_{\text{code}}} x(t)x(t-\tau)\,dt$$

This integral differs from (9.12), because the lower limit of integration has changed—this change makes all the difference. Aperiodic correlation treats each code as only one period. Compute the auto-correlation and cross-correlation of two Gold sequences considered in problem 9-3 and comment on the differences. When considering how the C/A-codes are used by GPS, is the aperiodic correlation function more or less relevant than the periodic correlation function?

9-5. Consider the N-H sequences, $h_I(t)$ and $h_Q(t)$, shown in Figure 9.24.
(a) Compute the periodic auto-correlation function.
(b) Compute the aperiodic auto-correlation function.
(c) When considering how the N-H codes are used by GPS, is the aperiodic correlation function more or less relevant than the periodic correlation function?

References

Betz, J. (2001). Binary Offset Carrier Modulations for Radionavigation, *Navigation*, vol. 48, no. 4, pp. 227–246.

Fontana, R., W. Cheung, P. Novak, and T. Stansell (2001). The New L2 Civil Signal, *ION GPS-2001*, pp. 617–631.

Gold, R. (1967). Optimal Binary Sequences for Spread Spectrum Multiplexing, *IEEE Trans. on Information Theory*, vol. IT-13, pp. 619–621.

Hegarty, C. and M. Tran (2002). Compatibility of the New Military GPS Signals with Civil Aviation Receivers, *Proc. ION Annual Meeting*.

Holmes, J. (1990). *Coherent Spread Spectrum Systems*, Krieger Publishing Company, Malabar, Florida.

MacWilliams, F.J. and N.J.A. Sloane (1976). Pseudo-random Sequences and Arrays, *Proc. IEEE*, vol. 64, pp. 1715–1729.

Pursley, M. (1977). Performance Evaluation for Phase-coded Spread-spectrum, Multiple-access Communication—Part I: System Analysis, *IEEE Trans. on Communications Theory*, vol. COM-25, pp. 795–799.

Sarwate, D.V. and M.B. Pursley (1980). Cross-correlation Properties of Pseudorandom and Related Sequences, *Proc. IEEE*, vol. 68, no. 5, pp. 593–618.

Spilker, J. (1996). GPS Signal Structure and Theoretical Performance, Chapter 3 from *Global Positioning System: Theory and Applications, Volume I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics.

Spilker, J. and A.J. Van Dierendonck (2001). Proposed New L5 Civil GPS Codes, *Journal of the Institute of Navigation*, vol. 48, no. 3, pp. 135–144.

Chapter 10

Signal-to-Noise Ratio and Ranging Precision

Absent obstructions or reflections, today's GPS receiver can measure pseudoranges to a satellite with a precision of 0.5 meters or better. In the last chapter, we attributed this ranging prowess to the auto-correlation functions of the C/A-codes and more specifically to the slope of the main peak of that function. In this chapter, we will substantiate that claim with a quantitative analysis of ranging precision in the presence of white noise. We will discover that the ranging precision does indeed depend on the slope of the correlation function, which in turn depends on the bandwidth of the signal.

As we shall see, the ranging performance also depends on the signal-to-noise ratio, C/N_0, and the averaging time used by the receiver. C/N_0 is the ratio of the power in the received signal to the power spectral density of the competing noise. Signal power, C, has units of watts or joules/sec. Power spectral density, N_0, has units of watts/hertz. Hence, C/N_0 has units of hertz.

Table 10.1 Received signal power density: path loss and transmitter antenna gain

	Satellite at low elevation ($el = 5°$, $\alpha = \pm 13.9°$)	Satellite at moderate elevation ($el = 40°$, $\alpha = \pm 10.6°$)	Satellite at zenith ($el = 90°$, $\alpha = \pm 0°$)
Power at the satellite antenna input	14.3 dBW (27 watts). This power is derived from the GPS specifications. However, typical GPS satellites broadcast 2 to 4 dB more power.		
Range (km)	25,240	22,020	20,190
Path loss ($1/4\pi R^2$)	–159.0 dB	–157.8 dB	–157.1 dB
Satellite antenna gain	12.1 dB	12.9 dB	10.2 dB
Effective isotropic radiated power (EIRP) toward the earth	26.4 dBW (437 watts)	27.2 dBW (525 watts)	24.5 dBW (282 watts)
Atmospheric loss	0.5 dB	0.5 dB	0.5 dB
Received power density	4.9×10^{-14} W/m^2 = –133.1 dBW/m^2	7.8×10^{-14} W/m^2 = –131.1 dBW/m^2	4.9×10^{-14} W/m^2 = –133.1 dBW/m^2

The power available for the C/A-code on the satellite is approximately 27 watts, but the power collected by a typical receiver on the earth's surface is only 10^{-16} watts or so. When received, the GPS signal is swamped by the noise in the front end of the receiver. In this chapter, we develop estimates of the received signal power and the received noise power spectral density. The ratio of these two powers is key to ranging precision.

We proceed as follows. Sections 10.1 and 10.2 begin by analyzing the power in the GPS signal. Section 10.1 shows that the majority of the signal attenuation is due simply to loss introduced by the long path from the satellites to the surface of the earth. In this section, we also show how the transmitting antennas on the satellites are used to ameliorate some of this loss. Section 10.2 discusses the role of the receiver's antenna and estimates the received signal power as a function of satellite elevation relative to the receiver.

Sections 10.3 and 10.4 turn our attention to the power in the competing noise. Section 10.3 introduces the language used to discuss noise and shows how to compute the aggregate noise for a system with multiple subsystems—this latter result is called Friis' formula. The work in Section 10.3 is applicable to any radio system, and Section 10.4 applies these general results to GPS.

Sections 10.5 and 10.6 analyze the performance of GPS ranging when additive white noise is the only disturbance. They teach the advantages of spread spectrum signaling when the goal is high precision ranging. To do so, they introduce the delay lock loop (DLL), which correlates the incoming signal with replicas of the signal generated by the receiver. As described in Section 9.8.1, this process is called *de-spreading*. As the name suggests, de-spreading concentrates the GPS signal power in a narrow bandwidth, where the GPS signal is stronger than the noise. Section 10.6 contains much of the detailed analysis and can be omitted on a first reading. Section 10.7 shows how spread spectrum signals also enable precise ranging in the presence of reflected signals called *multipath*. Finally, Section 10.8 is a brief summary.

10.1 Signal Path Loss and Transmit Antenna Gain

In the year 2005, a GPS satellite dedicates approximately 27 watts of power for the C/A-code signal on L1 [Aparicio *et al.* (1996)]. In decibels, this value is equal to $10\log_{10}27 = 14.3$ dBW and appears at the top of Table 10.1, which summarizes the analysis of this section.

The transmitted signal power is limited by cost. Ultimately, the signal power that serves our earthbound population of users comes from the solar panels that are carried by the satellite. When the satellite is in eclipse, the signal power comes from onboard batteries that were charged when the satellite was in sunlight. Increased signal power demands larger solar arrays and batteries, and the cost to launch is a fast function of the satellite weight. GPS signal power is also limited by the need to coexist with other GNSS systems operating in the same band. Recall that GPS uses code division multiple access to identify satellites, but this code division capability can be overwhelmed if one satellite is much stronger than another.

If this C/A signal power were broadcast in all directions, then the *power spatial density* at radius R meters would be $1/4\pi R^2$ times the radiated power. This term is the so-called *path loss* or *spreading loss*. As shown in Figure 10.1, this term accounts for the spreading of the total energy over the surface area of the sphere centered on the satellite. If the satellite antenna truly broadcast its energy uniformly in all directions, then the received power density at the surface of the earth would be

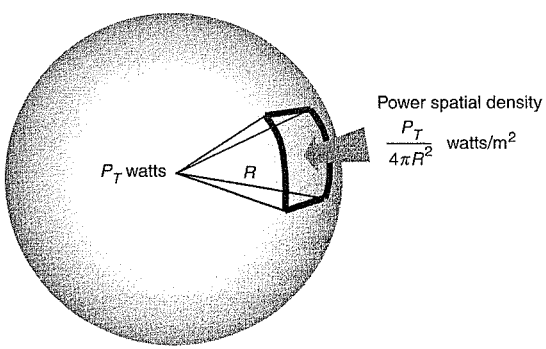


Figure 10.1 Power spatial density of the GPS signal. A power spatial density has units of watts per unit area.

$$\mathcal{P} = \frac{P_T}{4\pi R^2} \text{ watts/m}^2$$

$$\mathcal{P}_{dB} = P_{T,dB} - 11 - 20\log_{10} R \text{ dBW/m}^2$$

$$P_T = 27 \text{ watts}$$

R = distance from satellite to user in meters (10.1)

Do not confuse P_T and $\mathcal{P}$. P_T is a power measured in watts, while $\mathcal{P}$ is a power spatial density measured in watts/m^2. $P_{T,dB}$ and $\mathcal{P}_{dB}$ are the same quantities measured in dBW and dBW/m^2, respectively.

The altitude of the satellite is approximately 20,190 kilometers, and so spreading loss will be appreciable. As shown in Figure 10.2, distance from the satellite to a user depends on the user location on the earth. In fact, the satellite to user range can be calculated as a function of the satellite's elevation angle, *el*, at the user.

$$R = -R_E \sin el + \sqrt{R_E^2\left(\sin^2 el - 1\right) + R_{SV}^2}$$

$$R_E \approx 6371 \times 10^3 \text{ m}$$

$$R_{SV} \approx 26{,}560 \times 10^3 \text{ m}$$

el = satellite elevation angle at the user (10.2)

Even when the satellite is directly over the head of the user, the range is 20,190 km and the corresponding path loss is –157.1 dB! When the satellite is only 5° above the user's horizon, this range increases to 25,240 km and the path loss increases to –159 dB. These losses correspond to a power attenuation of approximately 1.6×10^{-16}. No wonder the received signal is so weak!

Some of this power can be recovered because the satellite can focus its energy towards the earth. This benefit is quantified relative to the power density from an antenna that radiates uniformly in all directions. Antenna gain gives amplification of the power density in a given direction relative to that predicted by $1/4\pi R^2$ for an omni-directional (or isotropic) antenna.

In general, this gain depends on the size and design of the satellite antenna, but we do not

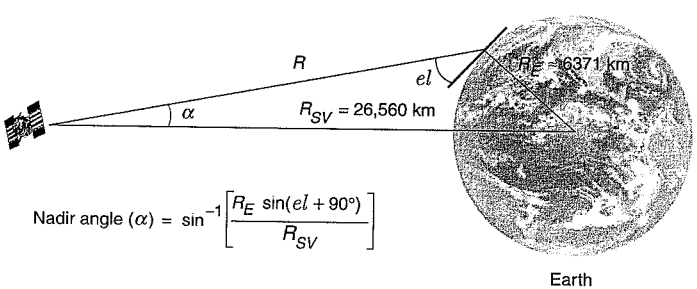


Figure 10.2 Satellite/earth geometry.

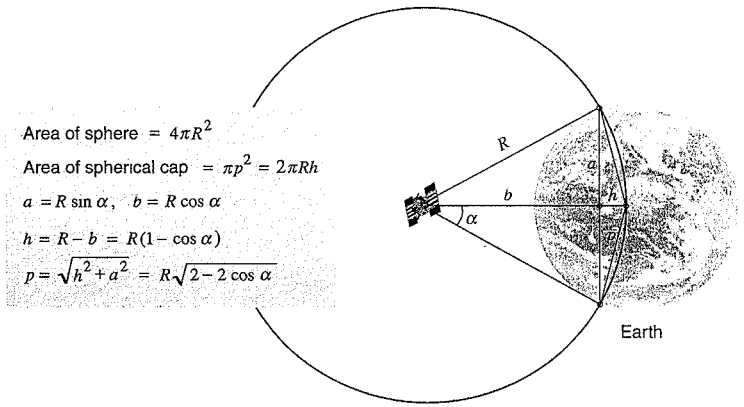


Figure 10.3 Using a spherical cap to *approximate* the gain of the GPS satellite antenna. (This drawing is not to scale.)

consider those complexities here. Rather, we estimate the satellite antenna gain by considering Figure 10.3, which shows a sphere centered on a GPS satellite. We assume that the satellite is able to concentrate its radiated power within the solid angle 2α measured away from the vector that connects the satellite to the center of the earth. The angle α is called the nadir angle. As shown in Figure 10.2, the nadir angle can be related to the elevation angle, *el*, of the satellite relative to the user's horizon.

With such focusing, the concentration factor (transmitter antenna gain) is given by the ratio of the area of the sphere to the area of the spherical cap created by the angle α. That ratio is given by

$$G_T(\alpha) = \frac{4\pi R^2}{\pi p^2} = \frac{2}{1 - \cos\alpha} \tag{10.3}$$

The details are given in Figure 10.3.

The earth subtends an angle ±13.9° as seen from a GPS satellite, but the GPS beam has a somewhat wider spread of ±21.3° for the L1 signal. Consequently, the gain may be approximated as $10\log_{10} G_T(21.3°) = 14.7$ dB.

The actual antenna gain for the GPS transmit antenna is smaller than this approximation for two reasons. First, there is additional loss in the antenna itself that suppresses the radiated power. Second, the gain is tailored to compensate for the greater distance to the users at the edge of the earth (as seen by the satellite). The satellite antenna gain is approximately 2 dB stronger at the edge of coverage ($\alpha = 13.9°$) than along the so-called bore sight ($\alpha = 0°$).

For a satellite at low elevation, the antenna gain is around 12.1 dB, and the effective radiated power in the direction of such a user is 437 watts. For a satellite at 40° elevation, the gain is approximately 12.9 dB and the effective radiated power in this direction is effectively 525 watts. Finally, for a satellite at zenith, the gain is approximately equal to 10.2 dB and the effective radiated power is 282 watts.

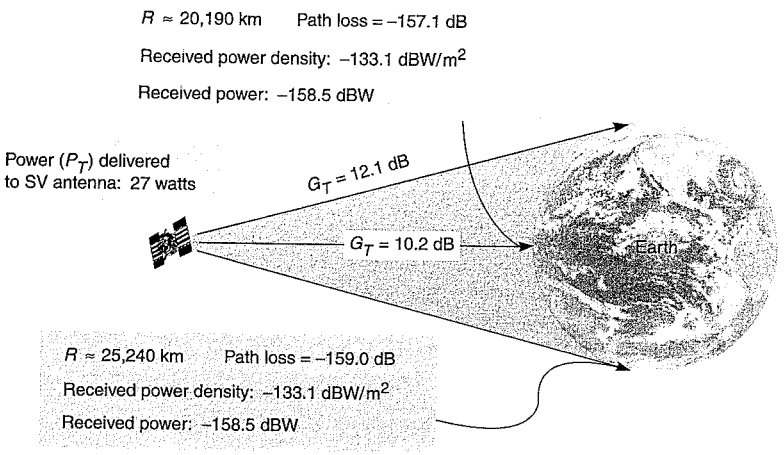


Figure 10.4 Summary of power density of received GPS signal. The actual values are typically higher.

When the factors described above are combined, the received power density is given by

$$\mathcal{P} = \frac{P_T G_T}{4\pi R^2 L_A} \text{ watts/m}^2$$

$$\mathcal{P}_{\text{dB}} = P_{T,\text{dB}} + G_{T,\text{dB}} - 20\log_{10} R - 11 - L_{A,\text{dB}} \text{ dBW/m}^2 \qquad (10.4)$$

Equation (10.4) includes a term, L_A, to model power loss as the signal propagates through the atmosphere. As shown, we use a value of 0.5 dB in Table 10.1 for this loss.

Our work so far is summarized in Table 10.1 and Figure 10.4. Radio engineers refer to such calculations as *link budgets*.

10.2 Received Signal Power and Receiver Antenna Gain

The power in the incident signal field is captured by the receiver's antenna. In fact, the received power is equal to the power density of the incident field times the effective area of the receive antenna. The effective area, A_E, measures the antenna's ability to capture the power in a field incident from a certain direction, and the gain, G_R, measures the antenna's ability to focus transmitted power in a certain direction. Remarkably, the effective area of any antenna is related to its gain. An antenna's ability to capture power from a certain direction is proportional to its ability to send power in that same direction. This reciprocity is skillfully explored in the classic textbook by Jordan and Balmain (1968), and is given by

$$A_E = G_R \lambda^2 / 4\pi \qquad (10.5)$$

In this equation, λ is the wavelength of the signal and is given by $\lambda = c/f$, where c is the speed of light and f is the frequency of the signal.

Table 10.2 Received signal power for C/A-code signal

	Satellite at Low Elevation ($el = 5°$, $\alpha = \pm 13.9°$)	*Satellite at Moderate Elevation* ($el = 40°$, $\alpha = \pm 10.6°$)	*Satellite at Zenith* ($el = 90°$, $\alpha = \pm 0°$)
Received power density (PD_S)	4.9×10^{-14} W/m² = –133.1 dBW/m²	7.8×10^{-14} W/m² = –131.1 dBW/m²	4.9×10^{-14} W/m² = –133.1 dBW/m²
Effective area of an omni-directional receive antenna ($\lambda^2/4\pi$)	2.87×10^{-3} m² –25.4 dBm²		
Receive power available from an isotropic antenna	–158.5 dBW	–156.5 dBW	–158.5 dB
Gain of a typical patch receive antenna (G_R) relative to isotropic antenna	–4 dBic	+2 dBic	+4 dBic
C/A-code received power available to a typical receive antenna	–162.5 dBW	–154.5 dBW	–154.5 dBW

An isotropic antenna is equally sensitive to signals coming from any direction. Such an antenna has unit gain, $G_R = 1$, for all azimuth and elevation angles and $A_E = \lambda^2/4\pi$. For such an antenna, the received signal power is given by

$$P_R = \frac{P_T G_T}{L_A}\left(\frac{\lambda}{4\pi R}\right)^2 \text{ watts} \qquad (10.6)$$

For GPS, these powers are shown in Table 10.2, which shows that the received power is a mild function of the satellite's elevation angle. This relationship is also plotted in Figure 10.5 for all elevation angles from 5° to zenith.

In fact, the antennas used by GPS receivers are not isotropic because they only need to receive signals from above the user's horizon. Above the horizon, they must provide nearly full sky coverage, because GPS satellites are sprinkled across the heavens, and good performance requires that satellites in all directions be received. Otherwise, the geometric dilution of precision will grow and positioning accuracy will deteriorate (see Section 6.1.2). Like isotropic antennas, the gain does not vary with azimuth because GPS satellites are found at all points of the compass. Unlike isotropic antennas, the gain does vary with elevation angle, and this variation is captured in the elevation pattern.

Such an elevation pattern is shown in Figure 10.6 for a patch antenna that measures approximately 10 cm × 10 cm × 1 cm, and is relatively low cost. Patch antennas can be smaller, and so they are used in a wide variety of applications. As shown, gain, G_R, decreases slowly

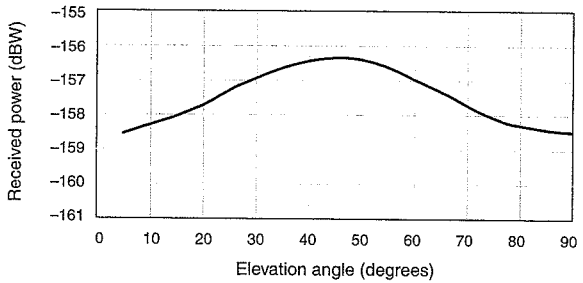


Figure 10.5 Received C/A-code signal power available from an isotropic antenna as a function of elevation angle for a user on the surface of the earth. This curve is from the GPS signal specifications. Typical receiver powers are up to 4 dB higher because the transmitters broadcast more power. The receiver's antenna gain also modifies these results.

from approximately +4 dBic at zenith to −4 dBic at an elevation angle of 5°. The notation dBic means gain relative to an isotropic antenna. Gains from Figure 10.6 are used to modify the receiver signal powers shown in Table 10.2.

GPS engineers use the symbol C to denote the received signal power inclusive of the gain from the receiver antenna, G_R, and any losses from the receiver, L_R. With these added factors, we have

$$C = \frac{P_R G_R}{L_R} \tag{10.7}$$

If we substitute for P_R, then

$$C = \frac{P_T G_T G_R}{L_A L_R}\left(\frac{\lambda}{4\pi R}\right)^2 = \frac{P_T G_T G_R}{L_A L_R}\left(\frac{c}{4\pi R f}\right)^2 \text{ watts} \tag{10.8}$$

Table 10.2 gives numerical values for this power in the GPS civil signal at the L1 frequency.

The frequency dependence in (10.8) was a critical consideration when choosing the GPS carrier frequencies. The receiving antenna must be small and able to capture signals coming from all directions. Hence, it cannot greatly concentrate or amplify the signal coming from any given direction. Since $G_R \approx 1$, the frequency cannot be too high. Otherwise, the received signal would be too weak. Satellite systems that only need to track one satellite have greater flexibility when choosing their carrier frequencies. Frequently, these systems can employ receiving antennas with higher directionality and gain, and can therefore afford higher frequencies.

Of course, a host of other factors affected the frequency selection for GPS. The GPS signal needed to fall within the available radio spectrum that could be used for space-to-earth signaling. In addition, the frequency could not be too low—otherwise ionospheric delays become too unpredictable.

Most civilian GPS receive antennas generate patterns similar to the one shown in Figure 10.6. However, the exceptions are both important and interesting. For example, some antennas

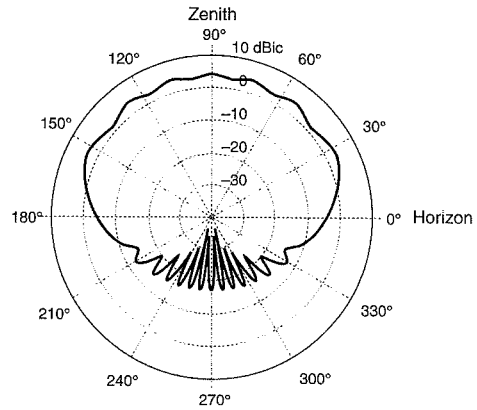


Figure 10.6 Commercial L1 antenna gain patterns. Predicted pattern for a standard patch antenna mounted on a four-wavelength-diameter circular ground plane (courtesy of Frank Bauregger, Novariant, Inc.).

place beams on the individual GPS satellites, and thus strengthen the GPS signals. Some adapt their patterns to attenuate undesired signals that may interfere with GPS. Others discriminate against signals that have polarizations other than the one used by GPS. Needless to say, these adaptive antennas are significantly more sophisticated than the simple antenna characterized by Figure 10.6. Even so, they are used in military applications for protection against intentional interference. We have more to say about these adaptive antennas in Chapter 13.

Other antennas are designed to reject multipath at fixed sites like differential GPS reference stations, where multipath is particularly worrisome [Counselman (1999) and dBSystems (2000)]. These multipath-limiting antennas (MLAs) are taller than the patch antenna described above. Their height precludes their use on most mobile platforms, but also enables a much more rapid roll off of gain with elevation angle. This rapid roll off allows the antenna to attenuate signals that are reflected from below the ground. This is further discussed in Section 10.7.3.

As shown in Table 10.2, the received GPS signals are certainly weak, but the competing signals are also usually weak. The competitors include:

- thermal noise generated in the receiver
- natural noise from sources outside of the receiver
- reflected signals
- signals from other GPS satellites
- man-made signals from systems other than GPS.

Of these, the first two are considered in the next four sections. More specifically, Section 10.3 provides a general introduction to noise. Section 10.4 applies the findings from Section 10.3 to GPS. Sections 10.5 and 10.6 introduce the delay lock loop and analyze its performance in the presence of noise. Section 10.7 moves on to describe the impact of signal reflections on

the same delay lock loop. As described in Chapter 9, the multiple-access interference between GPS satellites is minimized by careful selection of the spread spectrum sequences. More thorough analyses of this inter-satellite interference are contained in Sarwate and Pursley (1980). Interference from other man-made radio sources is the focus of Chapter 13.

10.3 Noise

Noise competes with a man-made signal at the receiver in almost all radio systems. In this chapter, we concern ourselves with *white noise*. Recall from Chapter 8 that white noise combines noise components from all frequencies with equal strength, and so it has a power spectral density that is a constant $N_0/2$ W/Hz. White noise is an excellent model for the natural noise that is received along with GPS signals, because natural noise has a constant power spectral density across the GPS band. However, white noise may not be a good model for man-made signals that may find their way into the GPS band. We return to this subject in Chapter 13.

Our power *spectral* density, $N_0/2$, should not be confused with the power *spatial* density, $\mathcal{P}$, used in the last two sections. Power spectral density measures the power per unit of bandwidth—it has units of watts per hertz. Power spatial density measures the amount of power in a unit of area—it has units of watts per meter^2.

The power spectral density is denoted $N_0/2$ rather than N_0 for convenience. Consider Figure 10.7. The top half shows $|H(f)|^2$ for a low-pass filter, where $H(f)$ is the transfer function for the filter. The bottom half shows $|H(f)|^2$ for a bandpass filter. White noise is input to both

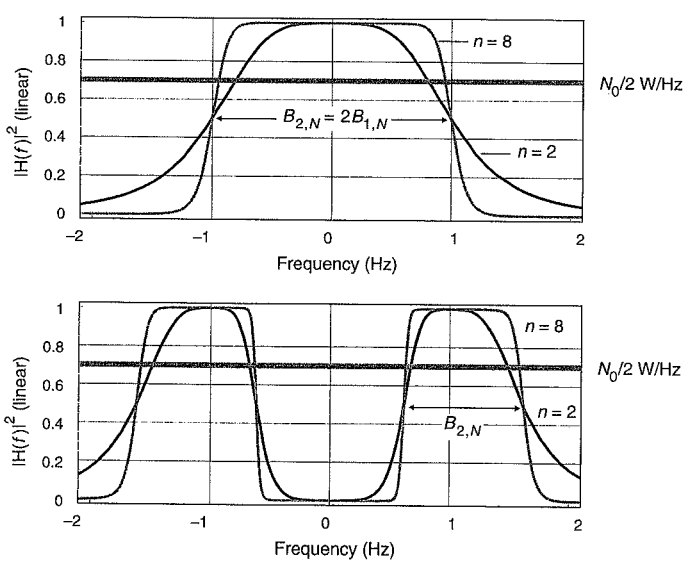


Figure 10.7 Noise power spectral density. Power spectral density (PSD) has units of watts per hertz. The notation $N_0/2$ preserves the relationship $P_N = N_0 B_N$ for low-pass filters and bandpass filters. The parameter n refers to the order of a Butterworth filter, which is described in Section 8.5.8.

filters, and we wish to calculate the noise power at the output, P_N.

We usually characterize low-pass filters with one-sided bandwidths because it is sensible to measure the frequency response from zero frequency up to some frequency where the filter response is weak. Since we are concerned with noise power, it also makes sense to consider the noise equivalent bandwidth described in Chapter 8. If we use the one-sided noise equivalent bandwidth, $B_{1,N}$, then the power in the output noise is given by the following simple relationship.

$$\begin{aligned} P_N &= 2B_{1,N}\frac{N_0}{2} \text{ watts} \\ &= B_{1,N}N_0 \end{aligned} \tag{10.9}$$

For the bandpass filter, we tend to rely on two-sided bandwidths because it is sensible to *span* the frequencies with a strong response. The two-sided noise equivalent bandwidth, $B_{2,N}$, is shown in Figure 10.7. The power in the output noise is given by

$$\begin{aligned} P_N &= 2B_{2,N}\frac{N_0}{2} \text{ watts} \\ &= B_{2,N}N_0 \end{aligned}$$

In both cases, we end up with the simple relationship, $P_N = B_N N_0$.

In this chapter and all of the subsequent ones, we will discover that the ratio of the signal power, C, to the total noise power spectral density is a key system parameter. We will denote this ratio as C/N_0. As described in Section 10.3.1, *noise figure* computes the degradation of C/N_0 as the signal passes through receiver components that add their own noise to the system. In the chapters that follow, we shall also discover that the bandwidth of a GPS receiver is wider for the sections nearest the antenna and becomes narrower as the processing develops. For any given noise-equivalent bandwidth, B_N, the signal power to noise power ratio, C/P_N, is

$$\frac{C}{P_N} = \frac{C}{B_N N_0} \tag{10.10}$$

10.3.1 Noise Temperature and Noise Figure

In a given bandwidth, B_N, noise power is also related to an equivalent noise temperature, T_{eq}, as follows:

$$\begin{aligned} &P_N = kT_{eq}B_N \text{ watts} \\ &k = 1.38\times10^{-23} \text{ J/K} = \text{Boltzmann's constant} \\ &T_{eq} = \text{equivalent temperature} \\ &N_0 = kT_{eq} \text{ watts/Hz} \end{aligned} \tag{10.11}$$

At first, the use of a temperature to describe noise power spectral density may not seem intuitive. However, this use finds its sensibility when considering thermal noise, where the equivalent temperature is simply the physical temperature of the device generating the noise. At any temperature above absolute zero (0 K), thermal noise is created by the inevitable motion of charge carriers within any conductor or semiconductor. The resulting noise power is proportional to the physical temperature of the device.

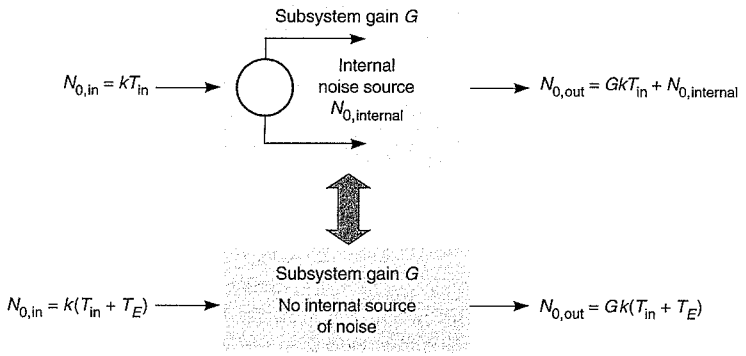


Figure 10.8 Two equivalent models for internal noise.

However, Equation (10.11) is also used for noise sources that have little dependence on the actual physical temperature—they are not thermal noise sources at all. For example, *antenna temperature* is not the physical temperature of the antenna, but simply the temperature of a thermal noise source that would provide the same noise power. For many radio systems, the antenna temperature models noise from the sky and from the warm earth. For GPS, the antenna temperature tends to be below the ambient temperature. For radio systems that operate at lower frequencies, the antenna temperature can be much higher than the ambient temperature. From this point forward, we will drop the adjective 'equivalent' when describing noise temperatures and understand that the noise source may or may not be thermal.

For many systems, including GPS, the total system noise is a blend of noise from the antenna and noise from subsystems within the receiver. A generic subsystem is shown in Figures 10.8 and 10.9. The noise at the subsystem output is the sum of internal noise and amplified input noise, where the internal noise may be due to any noise producing mechanism within the subsystem. To simplify analysis, this internal noise can be described as a second noise temperature at the input. This second noise temperature is said to be referenced to the input and is denoted T_E for effective input temperature. Then the total noise power density at the output is simply $Gk(T_{in} + T_E)$, where G is the power gain of the subsystem.

As mentioned above, noise figure, F, computes the degradation of C/N_0 as the signal passes through receiver components that add their own noise to the system.

$$\begin{aligned} F(T_{in}) &= \frac{(C/N_0)_{in}}{(C/N_0)_{out}} \\ &= \frac{CGk(T_{in}+T_E)}{GCkT_{in}} \\ &= 1+\frac{T_E}{T_{in}} \geq 1 \\ T_E &= \left(F(T_{in})-1\right)T_{in} \end{aligned} \tag{10.12}$$

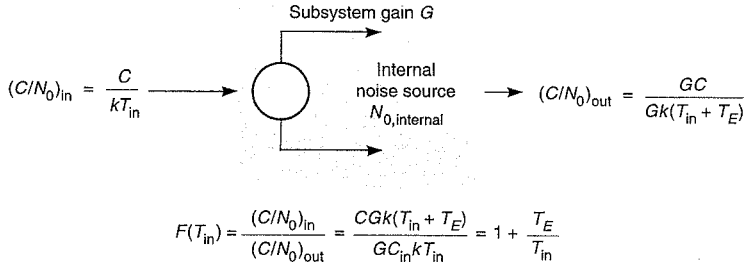


Figure 10.9 Definition of noise figure and relation to noise temperature.

T_E is the effective temperature of the internal noise source. If there is no internal noise, then $F = 1$ and the signal-to-noise ratio at the output of the subsystem is equal to the signal-to-noise ratio at the input. If there is internal noise, then the noise figure is greater than unity and the signal-to-noise ratio is degraded at the output.

There is no one-to-one relationship between noise figure and effective temperature unless the input temperature is known or specified. When writing a noise figure specification for a radio device or subsystem, the manufacturer does not know T_{in} because it depends on the application. Consequently, noise figure specification for components are often based on the assumption that the input noise temperature is room temperature or $T_0 = 290$ K. This protocol yields the following one-to-one relationship between $F(T_0)$ and T_E.

$$\begin{aligned} F(T_0) &= 1+\frac{T_E}{T_0} \\ T_E &= \left(F(T_0)-1\right)T_0 \end{aligned} \tag{10.13}$$

This relationship allows radio components to be characterized either by their noise figure or effective temperature. However, it must be used with care because

$$\frac{(C/N_0)_{in}}{(C/N_0)_{out}} = F(T_{in}) \neq F(T_0) \tag{10.14}$$

10.3.2 Noise in a Cascade of Subsystems

Any radio receiver is a cascade of subsystems or components. For noise analysis, all of these components are characterized by their power gain and noise figure. As we shall discover, the components found nearest to the receiving antenna determine the noise performance of the receiver.

Some of the subsystems amplify power. Some attenuate the signal and are called *passive*. The passive subsystems include cables and connectors between subsystems. The gain for these

elements is less than one, $G < 1$. Passive elements dissipate the lost power as heat and therefore introduce thermal noise. Conveniently, the noise figure for a passive element is equal to the power loss, L, which is the inverse of gain [Tsui (1995) and Vizmuller (1995)].

$$F = L = \frac{1}{G} > 1$$
$$T_E = \left(\frac{1-G}{G}\right)290 \tag{10.15}$$

Homework Problem 10-2 asks you to prove this relationship. Passive elements inevitably degrade the signal-to-noise ratio. The receiver design goal is to keep the loss small—especially for the components near the antenna. After all, if the gain is 0.9, then the effective temperature of the device is only 290/9 ≈ 32 K. However, if the gain is 0.1, then the effective temperature rises to 9 × 290 ≈ 2610 K.

Our task in this section is to find the aggregate noise figure of a receiving system from the specifications for the individual components. Consider Figure 10.10, which shows a cascade of subsystems. The input contains signal power, C, and noise power spectral density, kT_A, where T_A is the effective noise temperature of the external noise received by the antenna. Each subsystem is characterized by: a gain or loss, G_i or L_i; and a noise figure or effective temperature, F_i or $T_{E,i}$. If the element is passive, then $F_i = L_i$.

As shown in Figure 10.10, the signal power at the end of the chain is $G_1G_2G_3C$, and the noise density at the end of the cascade is

$$N_{0,3} = kG_1G_2G_3(T_A + T_R)$$
$$T_R = T_{E,1} + \frac{T_{E,2}}{G_1} + \frac{T_{E,3}}{G_1G_2} \tag{10.16}$$

This result is called *Friis' formula*. T_R is the effective temperature of all of the noise sources that are internal to the receiver. $T_A + T_R$ sums the external noise with the internal noise.

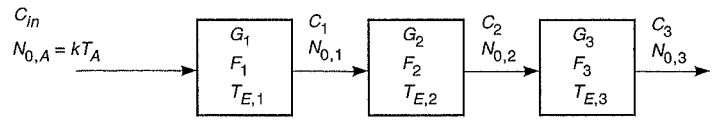
The noise figure of the entire receiver may also be written using the noise figures for the individual subsystems as follows.

$$F_R = F_1 + \frac{(F_2 - 1)}{G_1 + \frac{(F_3 - 1)}{G_1G_2}} \tag{10.17}$$

With these results, C/N_0 may be written in a way that includes the effect of the external noise and the noise that is internal to the radio system.

$$\left(\frac{C}{N_0}\right)_{\text{out}} = \frac{C}{k\left(T_A + T_{E,1} + \frac{T_{E,2}}{G_1} + \frac{T_{E,3}}{G_1G_2}\right)}$$
$$= \frac{C}{k(T_A + T_R)} \tag{10.18}$$

These formulas reveal much that is of interest. The antenna temperature and the temperature of the first subsystem contribute directly to the overall effective temperature. However, the noise contributions of sections later in the cascade may or may not be significant depending on the gains of the preceding subsystems. If G_1 is large, then the relative importance of the internal noise from subsystems 2 and 3 will be diminished. If G_2 is the first large gain, then the noise from subsystems 1 and 2 will be important, but the noise from section 3 will be attenuated. In most radio receivers, gain is distributed throughout the cascade. Hence, the noise performance of the entire radio is dominated by the noise performance of the first few sections encountered by the received signal. This makes sense because noise in these early sections will be amplified by all of the following sections.



$$C_3 = G_1G_2G_3C_{in}$$
$$N_{0,1} = G_1k(T_A + T_{E,1})$$
$$N_{0,2} = G_1G_2k(T_A + T_{E,1}) + G_2kT_{E,2}$$
$$N_{0,3} = G_1G_2G_3k(T_A + T_{E,1}) + G_2G_3kT_{E,2} + G_3kT_{E,3}$$
$$= G_1G_2G_3k\left(T_A + T_{E,1} + \frac{T_{E,2}}{G_1} + \frac{T_{E,3}}{G_1G_2}\right)$$
$$= G_1G_2G_3k(T_A + T_R)$$

Figure 10.10 Noise analysis for a cascade of subsystems.

We can also write C/N_0 as a function of the noise figure of the overall receiving chain, F_R.

$$\left(\frac{C}{N_0}\right)_{\text{out}} = \frac{1}{F_R(T_A)}\left(\frac{C}{N_0}\right)_{\text{in}}$$
$$= \frac{C}{kT_AF_R(T_A)} \tag{10.19}$$

If we include (10.7) and use decibels, we may write

$$\left(\frac{C}{N_0}\right)_{\text{out}} = P_{R,dB} + G_{R,dB} - L_{R,dB} - 10\log_{10}(kT_A) - F_{R,dB}(T_A) \quad \text{dB-Hz} \tag{10.20}$$

If we use a noise figure that is referenced to an input temperature other than the temperature of our antenna, then we must write

$$\left(\frac{C}{N_0}\right)_{\text{out}} = \frac{C}{kT_0F_R(T_0) + k(T_A - T_0)} \tag{10.21}$$

Note the correction term, $k(T_A - T_0)$, in the last equation if $T_A \neq T_0$.

Table 10.3 Typical noise characterization of the components in the front end of a GPS receiver

	Cable and filter that precede LNA	*Low-noise amplifier (LNA)*	*Cable that follows LNA*
Gain G	0.8 = –1 dB	100 = 20 dB	0.1 = –10 dB
Loss $L = 1/G$	1.26 = 1 dB	0.01 = –20 dB	10 = 10 dB
Noise figure F	1.26 = 1 dB (equal to loss)	2 = 3 dB (from manufacturer's specification)	10 = 10 dB (equal to loss)
Effective temperature $T_E = (F-1)290$ K	75.4 K	290 K	2610 K

10.4 Noise Analysis of a GPS Receiver

We now apply the general results from Section 10.3 to GPS. Table 10.3 and Figure 10.11 will guide our analysis. The figure shows external noise entering the receiver through the antenna and the first few stages of a generic GPS receiver. The noise generated in these early stages blends with the noise from the antenna to create the overall noise floor for this receiver. The antenna temperature is approximately 75–100 K due to noise received from the sky plus ground radiation. Table 10.3 summarizes the noise contribution from the components in the receiver front end, and is typical for a GPS receiver.

The first and third receiver elements listed in the table are passive components. Their noise figures will be equal to their losses and their effective temperatures are given by (10.15).

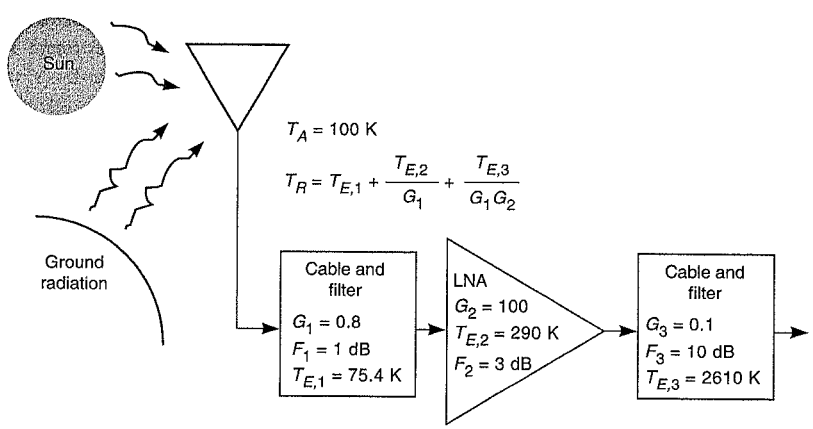


Figure 10.11 Noise analysis for the front end of a GPS receiver. For GPS, the noise is a blend of external noise and noise generated inside the receiver.

The first component is a filter to remove signals outside the GPS band with a cable that connects the antenna to the low-noise amplifier (LNA). This filter is simple and designed for low loss. Moreover, the cable also has low loss because the LNA is located next to the antenna and so the cable is short. For these reasons, we model the loss as 1 dB. The noise figure and temperature are correspondingly small.

To mitigate the noise temperature of the second cable, the LNA is designed for high gain and low noise figure. The table gives a typical value for a GPS LNA. As shown, we assume that the noise figure is 3 dB and the gain is 100 or 20 dB.

The effective temperature of the entire GPS cascade, including the antenna temperature is

$$T_A + T_R(F_2, G_1) = T_A + \left(\frac{1}{G_1} - 1\right)290 + \frac{(F_2 - 1)290}{G_1} + \frac{\left(\frac{1}{G_3} - 1\right)290}{G_1 G_2}$$

$$\approx T_A + 290\left(\frac{F_2}{G_1} - 1\right)$$

$$N_0 = 10\log_{10} k(T_A + T_R(F_2, G_1)) \quad \text{dBW/Hz} \tag{10.22}$$

$$N_0(F_2 = 3\,\text{dB}, G_1 = 0.8, T_A = 100\,\text{K}) \approx -201.3 \quad \text{dBW/Hz}$$
$$N_0(F_2 = 4\,\text{dB}, G_1 = 0.8, T_A = 100\,\text{K}) \approx -200.0 \quad \text{dBW/Hz} \tag{10.23}$$

As shown, the receiver noise floor depends on three critical parameters—the antenna tem-

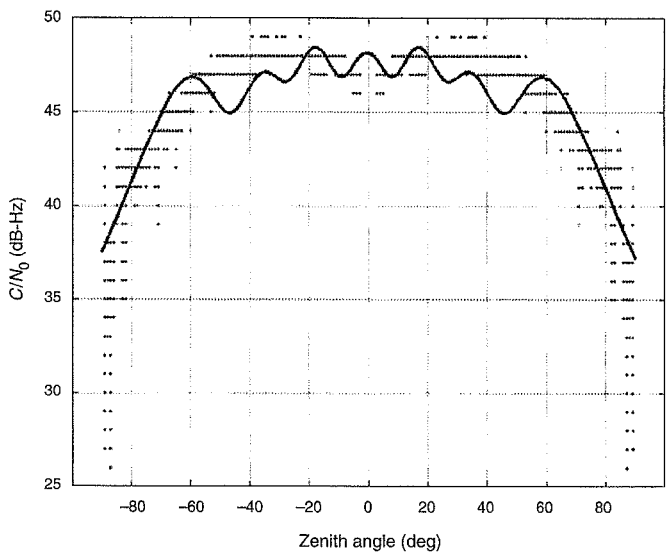


Figure 10.12 Measured C/N_0 versus elevation angle (courtesy of Frank Bauregger, Novariant, Inc.).

Table 10.4 Signal-to-noise ratios as a function of bandwidth

	5° elevation	*Zenith*
Received power in C/A-code signal (PC)	–162.5 dBW	–154.5 dBW
Noise power density (N_0) for a 3 dB LNA noise figure	–201 dBW/Hz	–201 dBW/Hz
C/N_0	38.5 dB Hz (approximately 7000)	46.5 dB Hz (approximately 45,000)
$C/P_N = \frac{C}{N_0 BW} = \frac{C}{N_0 20\times10^6}$ (20 MHz bandwidth)	–34.5 dB (approximately 1/3000)	–26.5 dB (approximately 1/450)
$C/P_N = \frac{C}{N_0 BW} = \frac{C}{N_0 2\times10^6}$ (2 MHz bandwidth)	–24.5 dB (approximately 1/300)	–16.5 dB (approximately 1/45)

perature (external noise), the gain of the cable and filter that precede the LNA, and the noise figure of the LNA. Thankfully, the loss in the second cable is not very important provided the LNA has reasonably high gain.

We now turn our attention to C/N_0 for GPS. Recall from Table 10.2, that the effective carrier power, C, in the civil signal varies from –162.5 dBW to –154.5 dBW for satellites at low elevation and zenith, respectively. As discussed above, the white noise power density is approximately –201 dBW/Hz, and so C/N_0 ranges from 38.5 up to 46.5 dB-Hz. This situation is shown in Figure 10.12, which plots measured C/N_0 versus zenith angle. The plot includes the impact of the receive antenna gain to show how this gain influences C/N_0. The signal power is approximately 7000 to 45,000 times stronger than the noise power in a 1-Hz bandwidth.

As we mentioned earlier, the bandwidth of a GPS receiver is wider for the sections nearest the antenna and becomes narrower as the processing develops. In fact, the earliest filters in the receiver front end have bandwidths of tens of megahertz. In a 20-MHz bandwidth, the noise power, $P_N = N_0 \times 20 \times 10^6$, is some 450 to 7000 times stronger than the signal power. In decibels, the GPS power is 26.5 to 34.5 dB weaker than the noise power. Radio engineers say that the GPS signal is below the noise floor.

As the signal travels deeper into the receiver, it eventually reaches the delay lock loops described in the next section. The delay lock loops contain correlators that dramatically despread the signal bandwidth. After de-spreading, the GPS signal power is packed into a null-to-null bandwidth of 100 hertz. In this bandwidth, the GPS signal is approximately 21.5 dB above the noise floor. Table 10.4 lists signal-to-noise ratios as a function of bandwidth.

Figure 10.13 also depicts the effect of de-spreading the GPS signal. Both traces show the GPS noise floor of –201 dBW/Hz. However, the units have been converted to dBW/MHz.

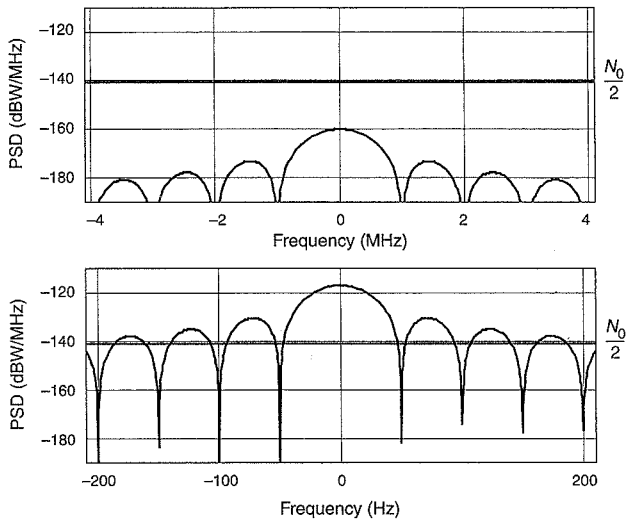


Figure 10.13 Power spectral densities for the GPS signal and background noise. The top trace compares the PSD for the background noise (–140 dBW/MHz) to the PSD for the received GPS signal before de-spreading. The bottom trace compares the noise PSD to the PSD for the GPS signal after de-spreading. Note the change in frequency scale.

Thus the noise floor has increased by 60 dB to a level of –141 dBW/MHz, because the noise power in 1 MHz is 10^6 times the power in 1 Hz. The top and bottom traces use results from Section 9.6 to show the power spectral density of the GPS signal before and after de-spreading. Before de-spreading, the GPS signal is much weaker than the noise. After de-spreading, it is 18.5 to 26.5 dB above the noise floor.

10.5 Delay Lock Loops and Ranging Precision

A GPS receiver tracks the distinct and sharp peak of the code auto-correlation function. The resulting pseudorange estimate is unambiguous because the main peak is noticeably larger than any side peaks. The pseudorange estimate is precise because the peak is narrow and the correlation measurements are very sensitive to the location of this sharp event. Indeed, the correlation peak is a triangle with a base width of 600 meters. As mentioned in Chapter 9, the receiver can resolve the arrival time to approximately 0.1% of this base width or 0.6 meters.

In this section, we quantify the performance of spread spectrum codes for ranging. To simplify our analysis, we think of the signal as consisting of code alone. We dispense with the complications of the carrier and data message. We also introduce the basic *delay lock loop* (or DLL). After signal acquisition, most GPS receivers use a DLL to track the signal from each GPS satellite [Spilker (1996), Van Dierendonck (1996), Holmes (1990)]. This fascinating structure is shown in Figures 10.14 and 10.15. The two figures are similar. Let's begin with Figure 10.14 and transition our attention to Figure 10.15 as the analysis matures.

As shown, the delay lock loop correlates the received signal with a slightly early replica

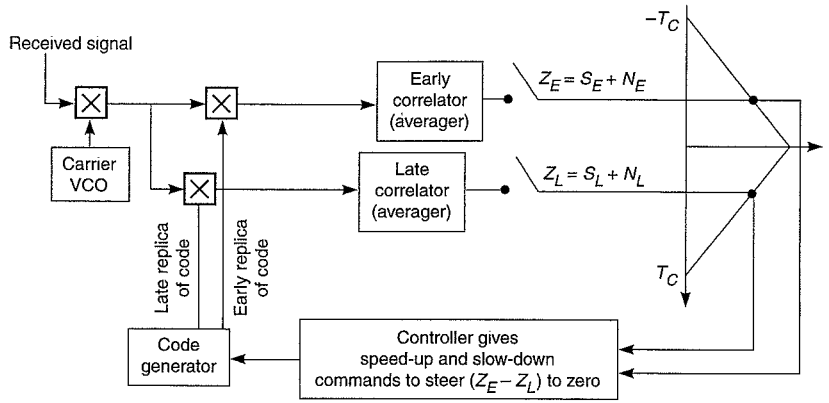


Figure 10.14 GPS delay lock loop to track code phase.

of the signal and a slightly late replica of the signal. When locked to the received signal, the *early correlator* samples the peak of the correlation function on the rising edge, and the *late correlator* samples the falling edge. Modern receivers sample the rising and falling edge of the peak simultaneously. Some early receivers used only one correlator and moved it back and forth between the rising and falling edge. This time-shared strategy is called a *tau-dither* loop. However, such a loop is not further considered in this book because it does not perform as well as strategies that sample both edges simultaneously.

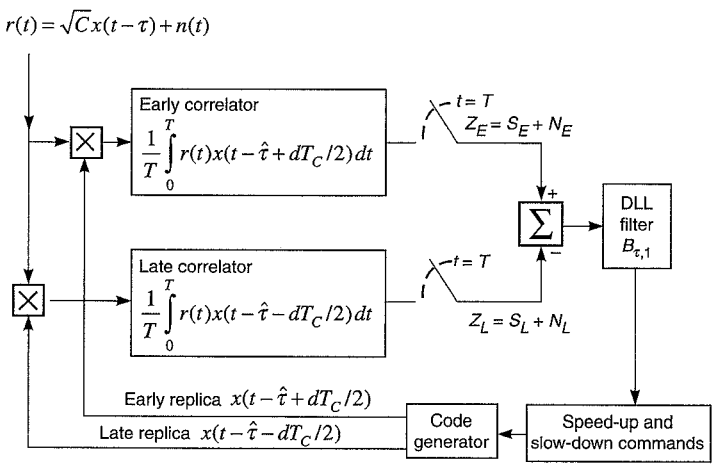


Figure 10.15 Delay lock loop (DLL) for white noise analysis of ranging precision. The function $n(t)$ is additive white noise (AWN).

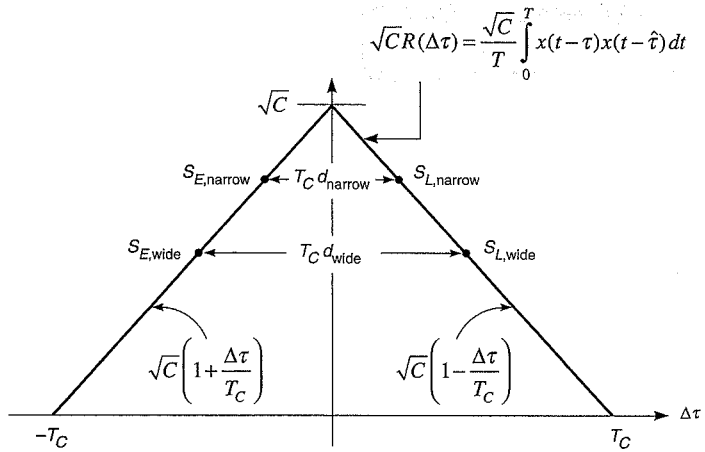


Figure 10.16 Correlation peak showing early and late samples for two sampling systems: wide ($d_{wide} = 1$) and narrow ($d_{narrow} = 0.5$).

A popular control strategy, called *null seeking*, attempts to equalize the early and late correlator values. In other words, a null seeking loop will shift the time of the replica code so that the difference between the early and late correlator values is zero. Such a loop is widely used because the difference in the correlator values is not sensitive to nuisance effects that may cause the absolute correlator values to change. For example, if the receiver moves under foliage, the absolute values will decrease, but this foliage attenuation will not be due to an authentic change in pseudorange. The null seeking strategy is robust to this nuisance effect because both correlator values will decrease and so their difference will remain approximately constant. Consequently, the null seeker will not confuse the signal attenuation with an effect that needs to be tracked.

The fixed time between the early and late correlator samples is called the *correlator spacing*. As Figure 10.16 shows, a wide correlator spacing is approximately one chip width or $d = 1$. The second spacing shown in the figure is around $d = 0.5$, and even narrower spacings are used. We shall discover that the correlator spacing is a powerful design parameter. Remember, d is the correlator spacing measured in chips, and dT_C is the corresponding time spacing.

The model in Figure 10.16 assumes that the received signal has been processed by the front end of the GPS receiver and is given by $\sqrt{C}x(t-\tau)+n(t)$, where $x(t)$ is the satellite code. C is the power in the signal including the receiver's antenna gain and any implementation loss. The signal amplitude is not $\sqrt{2C}$ because the signal does not include an RF carrier. In fact, our model is called a *baseband* model for the signal and receiver because it ignores the underlying carrier. We do not need the carrier to study the ranging prowess of spread spectrum signals. Do not worry—the carrier will reappear in Chapters 11 and 12.

Our input includes an additive disturbance, $n(t)$. As we know, GPS suffers from several such disturbances. These include natural radio noise, reflected signals, signals from the other GPS satellites, and other man-made interference. The arrival time of the satellite code, τ, is

our unknown or estimandum—we wish to accurately estimate this variable in spite of the strong additive disturbance, $n(t)$.

The early and late samples shown in Figure 10.16 are denoted

$$\begin{aligned} Z_E &= S_E + N_E \\ Z_L &= S_L + N_L \end{aligned} \tag{10.24}$$

where S_E and N_E are the signal and noise components of the early sample. S_L and N_L are the signal and noise contributions to the sample on the falling edge. The measurement noise contains the effects of natural noise, man-made interference, and multipath.

The signal components of the early and late samples are given by

$$\begin{aligned} S_E &= \frac{1}{T}\int_0^T \sqrt{C}x(t-\tau)\,x(t-(\hat{\tau}-dT_C/2))\,dt \\ &= \sqrt{C}R(\Delta\tau - dT_C/2) \\ S_L &= \frac{1}{T}\int_0^T \sqrt{C}x(t-\tau)\,x(t-(\hat{\tau}+dT_C/2))\,dt \\ &= \sqrt{C}R(\Delta\tau + dT_C/2) \end{aligned} \tag{10.25}$$

In this equation, $\Delta\tau = \tau - \hat{\tau}$ is the error in the propagation time estimate, and $R(\tau)$ is the autocorrelation function given in (9.82). T is the averaging time.

Equation (10.25) ignores the pre-correlation smoothing of the received signal due to front end filtering. The filters that precede the correlator tend to smooth or round the corners in the correlation peak, but our current discussion can neglect this complication. Thus we make the *infinite pre-correlator bandwidth* assumption.

Null tracking is enabled by subtracting the late sample from the early sample. The resulting difference is called the *discriminator function*, and is given by

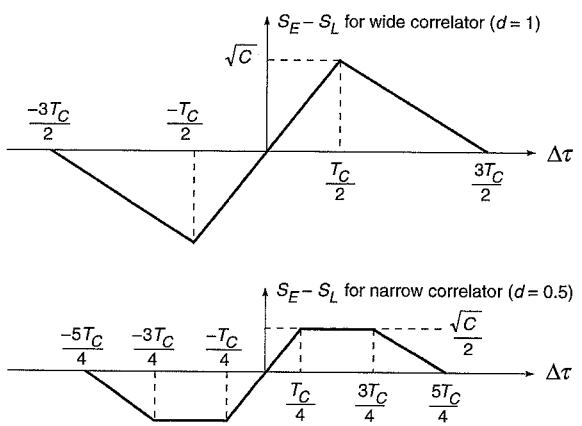


Figure 10.17 Discriminator functions for narrow and wide correlation peak sampling.

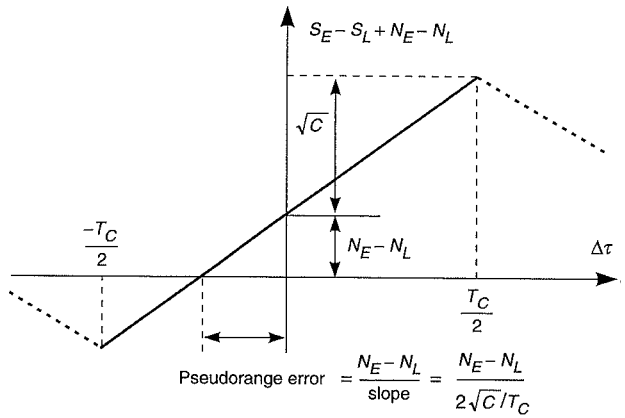


Figure 10.18 Noise impact on discriminator function for wide correlation peak sampling.

$$\begin{aligned} L_\tau &= Z_E - Z_L \\ &= S_E - S_L + N_E - N_L \end{aligned} \tag{10.26}$$

The signal component of this discriminator, $S_E - S_L$, is shown in Figure 10.17. Note that the shape of the discriminator is a function of the correlator spacing. The top trace is for $d = 1$ and the bottom trace is for $d = 0.5$. However, the slope of the discriminator near the origin is not affected by the spacing.

The estimated arrival time is the time at which the discriminator is zero. In the absence of $n(t)$, this zero crossing occurs at the true arrival time of the signal. However, disturbances push this zero crossing around, and this struggle is shown in Figure 10.18. As shown, the noise causes an error in the time estimate equal to

$$\begin{aligned} \Delta\tau &\approx \frac{N_E - N_L}{\text{slope}(L_\tau)\big|_{\Delta\tau=0}} \\ \text{slope}(L_\tau)\big|_{\Delta\tau=0} &= \frac{\partial L_\tau}{\partial\Delta\tau}\bigg|_{\Delta\tau=0} = \frac{\partial S_E}{\partial\Delta\tau}\bigg|_{\Delta\tau=0} - \frac{\partial S_L}{\partial\Delta\tau}\bigg|_{\Delta\tau=0} \\ &= \frac{\sqrt{C}}{T_C} - \frac{-\sqrt{C}}{T_C} = \frac{2\sqrt{C}}{T_C} \\ \Delta\tau &\approx \frac{T_C(N_E - N_L)}{2\sqrt{C}} \end{aligned} \tag{10.27}$$

This expression is valid for any additive disturbance, $n(t)$, with one proviso. The disturbance cannot be so large that it pushes the zero crossing out of the linear portion of the discriminator function shown in Figures 10.17 and 10.18.

The next section develops this equation for the more specific case where $n(t)$ is additive white noise (AWN) with a power spectral density equal to $N_0/2$ watts/hertz. We find that

white noise causes the time error, $\Delta\tau$, to be a random variable with zero mean and standard deviation equal to

$$\begin{aligned}\sigma_{\Delta\tau} &= \sqrt{E\left\{\Delta\tau^2\right\}} \\ &\approx \frac{T_C\sqrt{\text{var}\{N_E - N_L\}}}{2\sqrt{C}} \\ &= T_C\sqrt{\frac{d}{4TC/N_0}} \quad \text{seconds} \\ &= cT_C\sqrt{\frac{d}{4TC/N_0}} \quad \text{meters}\end{aligned} \tag{10.28}$$

In this equation, T is still the averaging time, and c is the speed of light ($c \approx 3 \times 10^8$ m/s). When expressed in meters, this error is called the pseudorange error.

These equations may be enjoyed without suffering the proof provided in the next optional section. They reveal many important features of the GPS signal design. First, error decreases as the power in the signal, C, increases relative to the power spectral density of the noise, $N_0/2$. Indeed, the ratio, C/N_0 will play a major role in all that follows.

Equations (10.27) and (10.28) also reveal the utility of spread-spectrum signaling. With all other parameters held constant, ranging performance improves by reducing the chip width, T_C. Reducing the chip width increases the chipping rate, $1/T_C$, and the bandwidth. Thus ranging performance is improved by increasing the chipping rate—even when the signal power and noise power spectral density are held constant. Spread spectrum signaling provides a *processing gain* proportional to signal bandwidth when the goal is to provide precise ranging measurements. This processing gain also mitigates multipath, competing signals from other GPS satellites, and other man-made interference. In Section 10.7, we shall discuss the anti-multipath properties of the GPS signal, but first we complete our noise analysis.

Figure 10.19 plots DLL error from (10.28) versus C/N_0 in dB-Hz. The averaging time, T, is 10 seconds, and the correlator spacing is equal to one chip width, $d = 1$. Three different chip widths are used. The curve for the P(Y)-code uses $T_C = 0.1$ microseconds, and the C/A-code curve assumes $T_C = 1$ microsecond. The final curve assumes that $T_C = 10$ microseconds. Hopefully, the virtue of spread spectrum signaling for ranging is clear.

Smaller values of correlator spacing improve DLL performance relative to what we see in Figure 10.19. Equation (10.28) suggests that we should make d as small as possible, and many receiver manufacturers use $d = 0.1$. Values below 0.1 do not offer any additional benefit because the receiver includes filters that round the correlation peak shown in Figure 10.16. These filters are discussed in Chapter 11 and are needed to remove radio frequency interference (RFI) from other man-made radio systems that occupy adjacent frequency bands. They round the corners of the correlation peak and reduce the slope of L_τ for very small correlator spacings. From (10.27), we see that a reduction in slope increases the error.

The DLL filter shown in Figure 10.15 plays a major role in determining the performance of the delay lock loop because it sets the averaging time. Long averaging times are achieved with narrow bandwidths, $B_{\tau,\text{I}}$; short averaging times correspond to large $B_{\tau,\text{I}}$. In general, this *averaging time* must be chosen to balance two considerations. Longer averaging times attenuate the effect of noise, and we call this reduction *integration gain*. However, longer averaging times can be problematic if the receiver is moving. If the averaging time is too long, then the correlator will smear together measurements made at different locations and the *dynamic performance* will suffer. In other words, the loop bandwidth is chosen to balance noise performance against authentic signal dynamics. In Chapter 12, we will develop expressions for the pseudorange error that depend on the bandwidth of the loop filter, $B_{\tau,\text{I}}$, shown in Figure 10.15. This future analysis converts (10.28) into the following.

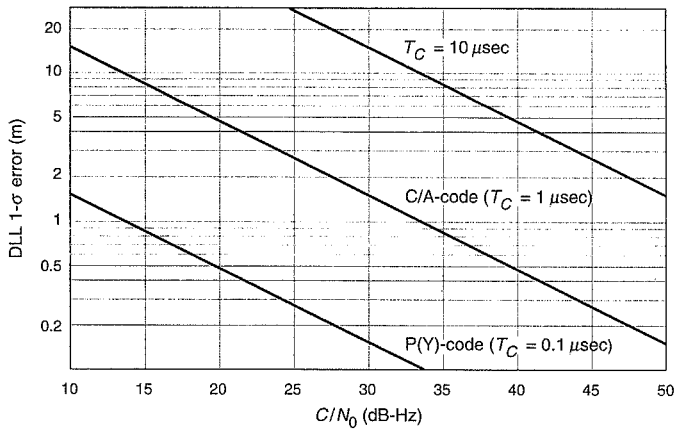


Figure 10.19 Delay lock loop noise performance in the presence of additive white noise. The averaging time, T, is ten seconds.

$$\sigma_{\Delta\tau} = cT_C\sqrt{\frac{dB_{\tau,\text{I}}}{2(C/N_0)}} \quad \text{meters}$$

As shown, the new expression replaces $1/2T$ with $B_{\tau,\text{I}}$. Such a replacement is reasonable because the bandwidth of a *box car* averager, like our correlator, is $B_{\tau,\text{I}} = 1/2T$. The longer we average, the narrower the bandwidth. We will take a closer look at the loop filter in Chapter 12.

10.6 Ranging Precision in the Presence of White Noise*

This optional section details the calculation of (10.28) above. For this analysis, $n(t)$ has a constant power spectral density $N_0/2$ watts/hertz for all GPS frequencies. This white noise model covers thermal noise in the receiver, natural noise received by the antenna, and any man-made noise that is broadband compared to the GPS signal. However, it is not appropriate for reflected signals or interference from signals that have a smaller bandwidth than the GPS receiver front end.

To conduct our noise analysis, we will need the Fourier transform pair developed in Section 8.6.

$$S_n(f) = \mathcal{F}\{R_n(\tau)\}$$
$$R_n(\tau) = E\{n(t)n(t-\tau)\} = \mathcal{F}^{-1}\{S_n(f)\} \tag{10.29}$$

For stationary processes, the power spectral density is the Fourier transform of the noise auto-correlation function. The auto-correlation function only depends on the time between the noise samples, τ, and is the inverse Fourier transform of the power spectral density.

For white noise, these become

$$S_n(f) = \frac{N_0}{2}$$
$$R_n(\tau) = \mathcal{F}^{-1}\left\{\frac{N_0}{2}\right\} = \frac{N_0}{2}\delta(\tau) \tag{10.30}$$

With these tools, our white noise analysis is fairly straightforward. Recall that the time error, $\Delta\tau$, is equal to $N_E - N_L$ divided by the slope of the discriminator function. In the present analysis, N_E and N_L are due to white noise and must be characterized by their mean and variance (or standard deviation). Correspondingly, the time (or pseudorange) error is also a random variable. The average pseudorange error is equal to zero because the noise samples have zero mean. However, individual values of N_E and N_L certainly deviate from zero. This fluctuation causes the ranging error to vary randomly. The variance of the pseudorange error is given by

$$\sigma_{\Delta\tau}^2 = E\{\Delta\tau^2\} \approx E\left\{\left(\frac{N_E - N_L}{\text{slope}(L_\tau)\big|_{\Delta\tau=0}}\right)^2\right\}$$
$$= \frac{E\{N_E^2 - 2N_E N_L + N_L^2\}}{\left(\frac{2\sqrt{C}}{T_C}\right)^2} \tag{10.31}$$

In pursuit of the numerator, we compute

$$\begin{aligned}
\text{var}\{N_E\} &= \text{var}\{N_L\} \\
&= E\{N_E^2\} = E\{N_L^2\} \\
E\{N_E^2\} &= E\left\{\frac{1}{T}\int_0^T n(t)\,x(t-\tau-dT_C/2)dt\,\frac{1}{T}\int_0^T n(s)x(s-\tau-dT_C/2)\,ds\right\} \\
&= \frac{1}{T^2}E\left\{\int_0^T\int_0^T n(t)n(s)\,x(t-\tau-dT_C/2)x(s-\tau-dT_C/2)\,dt\,ds\right\} \\
&= \frac{1}{T^2}\int_0^T\int_0^T E\{n(t)n(s)\}x(t-\tau-dT_C/2)x(s-\tau-dT_C/2)\,dt\,ds \\
&= \frac{1}{T^2}\int_0^T\int_0^T \frac{N_0}{2}\delta(t-s)\,x(t-\tau-dT_C/2)\,x(s-\tau-dT_C/2)\,dt\,ds
\end{aligned} \tag{10.32}$$

Next, we use the sifting property of the unit impulse function to provide

$$\begin{aligned}
E\{N_E^2\} &= \frac{N_0}{2T^2}\int_0^T x(t-\tau-dT_C/2)\,x(t-\tau-dT_C/2)\,dt \\
&= \frac{N_0}{2T^2}\int_0^T x^2(t-\tau-dT_C/2)\,dt \\
&= \frac{N_0}{2T}
\end{aligned} \tag{10.33}$$

Noise is reduced by longer averaging times—a sensible result.

However, we still have not found any influence from the correlator spacing, d. To this end, we compute the other term in the numerator of (10.31).

$$\begin{aligned}
E\{N_E N_L\} &= \frac{1}{T^2}E\left\{\int_0^T n(t)\,x(t-\tau-dT_C/2)\,dt\int_0^T n(s)\,x(s-\tau+dT_C/2)\,ds\right\} \\
&= \frac{1}{T^2}\int_0^T\int_0^T E\{n(t)n(s)\}x(t-\tau-dT_C/2)\,x(s-\tau+dT_C/2)\,dt\,ds \\
&= \frac{1}{T^2}\int_0^T\int_0^T \frac{N_0}{2}\delta(t-s)\,x(t-\tau-dT_C/2)\,x(s-\tau+dT_C/2)\,dt\,ds \\
&= \frac{N_0}{2T^2}\int_0^T x(t-\tau-dT_C/2)\,x(t-\tau+dT_C/2)\,dt \\
&= \frac{N_0}{2T}R(dT_C) \\
&\approx \frac{N_0}{2T}(1-d)
\end{aligned} \tag{10.34}$$

Now finally, we can put the entire puzzle together.

$$\begin{aligned}
E\{\Delta\tau^2\} &\approx \frac{E\{N_E^2 - 2N_E N_L + N_L^2\}}{\left(2\sqrt{C}/T_C\right)^2} \\
&= \frac{\frac{N_0}{T} - \frac{N_0}{T}(1-d)}{\left(2\sqrt{C}/T_C\right)^2} \\
&= \frac{dT_C^2}{4TC/N_0} \quad \text{seconds}^2 \\
\sigma_{\Delta\tau} &= cT_C\sqrt{\frac{d}{4TC/N_0}} \quad \text{meters}
\end{aligned} \tag{10.35}$$

This powerful result is discussed at the end of the last section, and we do not repeat that discussion here. Rather, we return only to the discussion of correlator spacing, d. Decreasing spacing improves performance, but not because of any improvement in the discriminator. As discussed in the last section, the discriminator slope does not change with correlator spacing.

The error reduction exists because the two noise samples begin to cancel as the correlator spacing is reduced. After all, the samples are becoming more nearly simultaneous as the spacing is reduced, and so the early and late noise samples are becoming more correlated [Van Dierendonck *et al.* (1994)].

10.7 Ranging Precision in the Presence of Signal Reflections (Multipath)

As discussed in Section 5.2, multipath arises when multiple paths exist from the satellite to the user antenna. The primary path is usually a direct, unobstructed path from the satellite to the antenna, while the secondary paths usually include a reflection off a nearby object or the ground. These reflections confound the receiver by distorting the correlation peak. After all, our analysis so far has assumed that this peak is a pristine triangle. If additional signals arrive, they will contribute secondary peaks and the early and late correlator samples may not be centered on the true arrival time of the direct ray.

In general, the impact of multipath depends on:

- the amplitude of the reflected signal relative to the direct
- the delay of the reflected signal relative to the direct
- the phase of the reflected signal relative to the direct
- the rate of change of the relative phase

Let's begin by exploring the phase relationship. The carriers for a direct ray and two reflected rays are shown in Figure 10.20. The top carrier is for a reflection that arrives in phase with the direct wave and therefore causes constructive interference—the direct ray is strengthened by the reflection. The bottom carrier arrives out of phase with the carrier. It causes destructive interference and weakens the direct ray. In this case, the signal is said to fade.

Figure 10.20 introduces vector diagrams for constructive and destructive interference. The length of each vector is the amplitude of the corresponding carrier. The angle between

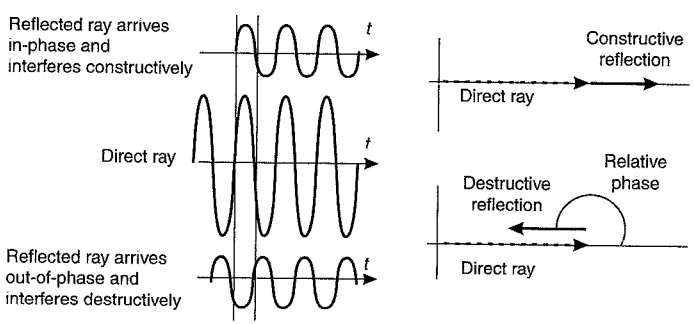


Figure 10.20 Reflected carriers arriving in-phase and out-of-phase, relative to the carrier from the direct ray.

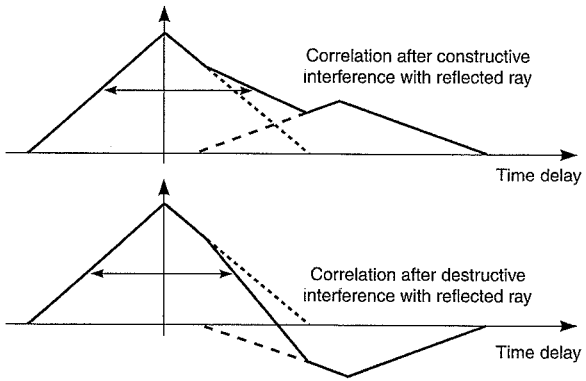


Figure 10.21 Correlation peak with constructive interference from a reflected ray, and with destructive interference from a reflected ray.

two vectors gives the relative phase of the two carriers. The direct ray is shown as the dashed vector and has larger amplitude than either of the reflections. The solid vectors for the reflections are added to the direct ray. Constructive interference has the same angle, and so the length of the sum is greater than the length of the direct ray by itself. In contrast, the angle for destructive interference is 180°, and so the sum is shorter than the direct ray. In general, the phase of the reflection varies and assumes all possible angles relative to the direct ray. Hence, the true vector picture would show a reflection vector with angle and magnitude that continuously varies relative to the direct ray.

The reflected wave always arrives after the direct wave and creates a delayed correlation peak as shown in Figure 10.21. The ratio of the direct peak amplitude to the delayed peak amplitude is $\sqrt{C/P_M}$, where C is the power in the direct signal and P_M is the power in the reflected signal. The late peak will also be shifted in time by $\Delta\tau_M$ seconds, where $\Delta\tau_M$ is the delay of the multipath relative to the direct. If the interference is constructive, then the late peak will be added to the earlier peak. If the reflection interferes destructively, then the late peak is subtracted from the early peak.

The relative delay of the multipath, $\Delta\tau_M$, plays a major role in determining its effect. If the delay is long compared to a chip width, T_C, then the auto-correlation properties of the code will suppress the effect—this regime is explored in the next subsection. If the delay is smaller than a chip width, then narrow correlator spacings are helpful—this effect is quantified in Section 10.7.2. Finally, special antennas can be used to mitigate multipath provided the user is stationary. This technique is explored in Section 10.7.3.

10.7.1 Long-Delay Multipath

If the delay is long compared to a chip width (approximately 300 meters or 1 microsecond for the C/A-code), then multipath will not cause any pseudorange errors. This situation is detailed in Figure 10.22. No errors exist when the rising edge of the delayed peak does not touch the late correlator sample. This condition holds when

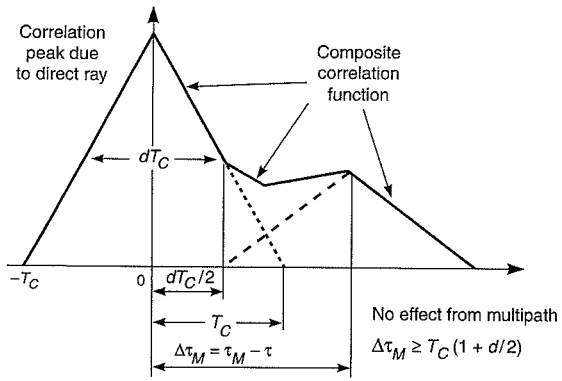


Figure 10.22 Spread spectrum signaling rejects reflections with large delays.

$$\Delta\tau_M - T_C \geq \frac{dT_C}{2}$$
$$\Delta\tau_M \geq T_C(1+d/2)$$
$$\Delta\tau_M(d=1) \geq 1.5T_C$$
$$\Delta\tau_M(d=0.1) \geq 1.05T_C \qquad (10.36)$$

As shown, spread spectrum signaling may be helpful in multipath environments. As the chip width, T_C, is reduced, multipath vulnerability is reduced. If a wide correlator is used, $d = 1$, with the C/A-code ($T_C = 1.0$ microseconds), then errors are obviated for all differential path lengths over 450 meters. If the P(Y)-code is used ($T_C = 0.10$ microseconds), then the receiver can tolerate all differential path lengths greater than 45 meters.

Figure 10.22 and (10.36) reveal another virtue of narrow correlator spacing. The C/A-code user can reduce vulnerability by using a narrow correlator spacing. If a narrow correlator, $d = 0.1$, is used with the C/A-code, then the errors are obviated for all differential path lengths greater than 315 meters.

If the multipath delay is less than $T_C(1+d/2)$, then the correlator pairs will usually move and pseudorange errors will result. As shown in Figure 10.22, constructive interference tends to move the correlation pair slightly to the right, and the measured pseudorange is longer than it should be. Destructive interference moves the correlation pair to the left. Although the reflected ray necessarily arrives after the direct ray, destructive interference causes the pseudorange to be measured short!

10.7.2 Short-Delay Multipath

Short-delay multipath has been analyzed by Braasch (1996), Van Dierendonck *et al.* (1992) and Enge (1999), and some typical results are shown in Figures 10.23 and 10.24. These unusual figures plot multipath error bounds. The top and bottom traces are the upper and lower bounds on multipath error for multipath that is 12 dB weaker than the direct ray, $C/P_M = 16$.

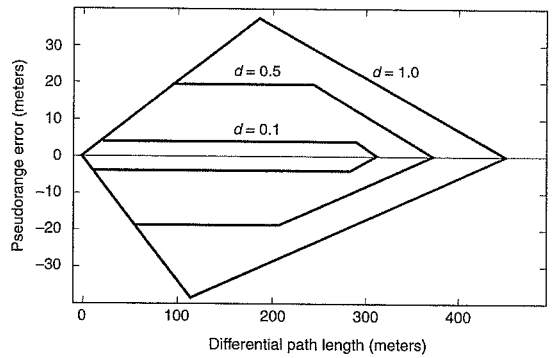


Figure 10.23 Bounds on C/A-code pseudorange error due to multipath. The actual error varies between the indicated upper and lower bounds as the relative phase changes. The upper bound corresponds to constructive interference, and the lower bound corresponds to destructive interference. The amplitude of the multipath is 12 dB below the amplitude for the direct ray.

The upper trace is for constructive interference where the pseudorange is measured long and the error is positive. The lower trace is for destructive interference where the measured pseudorange is short. As the relative phase varies from 0° (constructive interference) to 180° (destructive interference), the multipath error swings between its upper and lower bounds. If this variation in phase is adequately rapid in time, then the multipath error can be attenuated using carrier smoothing—as described in Section 4.7.

If the differential path length is small, then the error bounds are independent of the correlator spacing. For these short delays, the error grows linearly with differential path length and grows with multipath amplitude. As the relative path length increases, the role of correlator spacing asserts itself. Smaller correlator spacings have two good effects. As shown in Figure

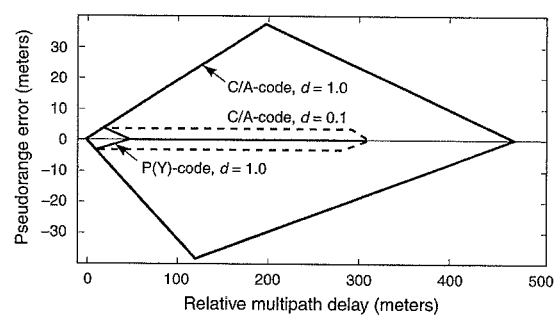


Figure 10.24 Bounds on P(Y)-code and C/A-code pseudorange error due to multipath. The actual error varies between the indicated upper and lower bounds as the relative phase changes. The upper bound corresponds to constructive interference and the lower bound corresponds to destructive interference. The amplitude of the multipath is 12 dB below the amplitude for the direct ray.

10.23, they cause the error bounds to be smaller. As described by (10.36), they also cause the error bounds to decrease to zero for smaller delays.

Figure 10.24 compares the multipath error bounds for the C/A and P(Y)-codes. In both cases, the correlator spacing is one code chip width. So the correlator spacing for the C/A-code is one microsecond, and the spacing for the P(Y)-code is 0.1 microsecond. In the presence of multipath, the P-code receiver enjoys a significant advantage. However, Figure 10.24 also shows that much of this advantage can be recovered by using narrow correlators with the C/A-code.

Narrow correlators do require that the front end of the receiver have a wide bandwidth relative to the C/A-code chipping rate. In other words, the bandwidth of the circuits that precede the correlators must be large compared to the 1 Mcps chipping rate. Typical narrow correlator receivers have pre-correlator bandwidths of 16 MHz. Narrower bandwidths tend to round the corners of the correlation peak and mute the advantage of narrow spacings. On the other hand, pre-correlator bandwidths cannot be too wide. Otherwise, they will not attenuate signals from competing radio systems that operate in adjacent radio bands. This trade-off discourages the use of narrow correlators in conjunction with P(Y)-code receivers. Pre-correlator bandwidths several times greater than the P(Y) chipping rate would be unduly vulnerable to interference.

10.7.3 Multipath-Limiting Antennas

Two other techniques are available to combat multipath with a relatively short delay. If the relative phase between direct and reflected changes rapidly, then the receiver can simply average the pseudorange measurements. If the averaging time is long compared to the time required for the phase to change by half a wavelength, then the averaged multipath will be attenuated and the net effect will be small.

If the delay and phase rate are both small, then a special antenna may be helpful. The gain pattern for such a multipath-resistant antenna is shown in Figure 10.25 (dBSystems, 2000).

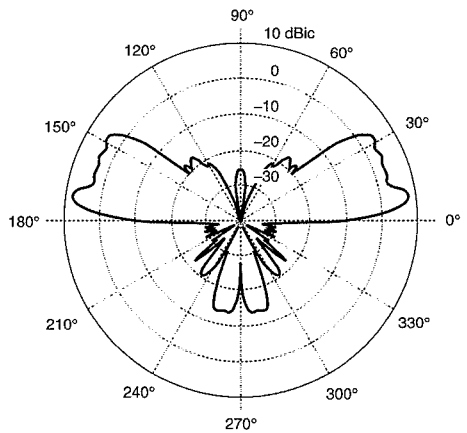


Figure 10.25 Measured pattern from a multipath-limiting antenna (courtesy of dB Systems).

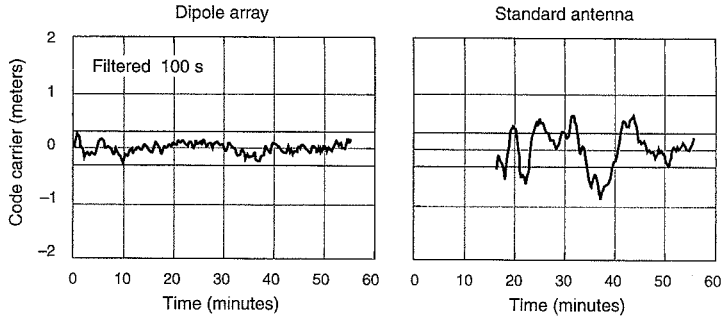


Figure 10.26 Code minus carrier measurements showing reduced multipath effect for the multipath-limiting antenna (courtesy of Professor Frank van Graas, Ohio University).

Contrast the gain pattern shown in Figure 10.25 to the gain pattern for a typical GPS patch antenna shown in Figure 10.6. As shown, the gain for the multipath-resistant antenna decreases much more rapidly when the satellite falls below 10°. Consider a satellite at low elevation (approximately 10°) and assume that the signal arrives on the direct path from the satellite, but is also reflected from below the antenna. GPS signals can, in fact, be reflected from below ground—they travel through the ground and then reflect off a layer of moist earth. After reflection, the multipath signal arrives at the antenna from an elevation angle that is the negative of the satellite elevation angle. If the satellite is at 10° elevation, then the reflected signal will arrive from an angle of –10°. Such reflections are troublesome at high-quality differential GPS stations. Moreover, receiver processing can do little to ameliorate this effect—the antenna alone must provide protection.

For the patch antenna, a reflection from –10° will be approximately 10 dB weaker than the direct signal from +10°. For the multipath-resistant antenna, a reflection from –10° will be approximately 30 dB weaker than the direct signal. This attenuation pays off with a much smaller multipath error in the code phase measurement. This promise is substantiated by Figure 10.26, which shows code minus carrier measurements for the multipath antenna on the left and for a standard patch antenna on the right. Over a short time period, code minus carrier measurements estimate the impact of multipath on code phase error because the code measurements are significantly more sensitive to multipath than the carrier measurements.

The multipath antenna characterized in Figure 10.25 has little gain above 35°. A second antenna is placed on top of the multipath-resistant antenna to fill in the response above 35°. Other multipath-rejecting antennas have been designed and some are based on a single antenna rather than a pair [Counselman (1999)].

10.8 Summary

The GPS signal is weak by the time it travels from orbit to the surface of the earth. The received power is proportional to $1/R^2$, where R is the distance traveled. This attenuation is partially offset by the gain of the satellite transmit antenna. Indeed, this antenna focuses most of its transmitted power on the earth and the corresponding gain is a little greater than ten. How-

ever, most receiver antennas cannot employ such focusing. They have a hemispherical gain pattern, so they can receive GPS satellite signals coming from all skyward directions. Consequently, they do not amplify the received signals. All told, the received GPS signal power is approximately 10^{-16} watts.

In the front end of the GPS receiver, natural noise is much stronger than the GPS signal. Friis formula accounts for the noise that is both external to the receiver as well as the noise that is developed in the front end of the receiver. Happily, the GPS signal power becomes stronger than the competing noise after the correlation process de-spreads the GPS signal.

We analyzed GPS ranging performance in the presence of white noise. Ranging performance improves with C/N_0 and averaging time. It is also sensitive to the chipping rate of the spread spectrum codes. Faster codes are better, and narrow correlator spacings also help.

Signal reflections, known as multipath, are a nuisance. For long delays, reflections are attenuated by the spread spectrum codes, but short-delay multipath can be more troublesome. This scourge is treated by using narrow correlator spacings, averaging, special antennas and careful siting of the receiver antenna. Indeed, the antenna and proper siting can treat multipath problems that no amount of subsequent processing by the receiver will be able to cure.

Homework Problems

10-1. GPS satellites direct their power toward the earth, but some power escapes around the edge of the earth and reaches geostationary satellites on the other side. What is the approximate GPS signal power at a geostationary satellite? Can this signal serve any purpose? What receiver processing could be done on the geostationary satellite to increase the utility of the GPS signal?

10-2. Derive the noise figure for a lossy element like a cable. (If desired, refer to Tsui (1995) or Vizmuller (1995) for some guidance.)

10-3. Compare the cascaded noise figures for the following two systems. System A is a cable followed by an amplifier followed by a receiver. System B uses the same elements, but in the following order: amplifier, cable and receiver. The cable has 5 dB of loss. The amplifier has a gain of 20 dB and a noise figure of 2 dB. The receiver has a gain of 50 dB and a noise figure of 7 dB.

10-4. Consider a cable followed by an amplifier. The cable loss is 1 dB for every 100 meters, and the amplifier noise figure is 7 dB. The source temperature is 290 K. How long can the cable be before the output signal-to-noise ratio is 5% of the input signal-to-noise ratio?

10-5. Calculate the error bounds due to multipath shown in Figures 10.23 and 10.24. (If desired, refer to Enge (1999) for some guidance.)

10-6. Calculate error bounds for the GPS carrier phase due to multipath. Comment on the relationship between these errors and the code phase errors.

References

Aparicio, M., P. Brodie, L. Doyle, J. Rajan, and P. Torrione (1996). GPS Satellite and Payload, in *Global Positioning System: Theory and Applications I*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), AIAA, pp. 209–244.

Braasch, Michael S. (1996). Multipath Effects, in *Global Positioning System: Theory and Applications I*, B. Parkinson, J. Spilker, P. Axelrad and P. Enge (eds.), AIAA, pp. 547–568.

Counselman, Charles C., III (1999). Multipath-Rejecting GPS Antennas, *Proceedings of the IEEE*, vol. 87, no. 1, pp. 86–91.

dBSystems (2000), personal communication.

Enge, Per (1999). Local Area Augmentation of GPS for the Precision Approach of Aircraft, *Proceedings of the IEEE*, vol. 87, no. 1, pp. 111–132.

Holmes, J.K. (1990). *Coherent Spread Spectrum Systems*, Krieger.

Jordan, E.C. and K.G. Balmain (1968). *Electromagnetic Waves and Rotating Systems* (2nd edition), Prentice Hall.

Sarwate, D.V. and M.B. Pursley (1980). Cross-correlation Properties of Pseudorandom and Related Sequences, *Proceedings of the IEEE*, vol. 68, no. 5, pp. 593–619.

Spilker, James J. (1996). Fundamentals of Signal Tracking Theory, in *Global Positioning System: Theory and Applications I*, B. Parkinson, J. Spilker, P. Axelrad and P. Enge (eds.), AIAA, pp. 245–326.

Tsui, James B.-Y. (1995). *Digital Techniques for Wideband Receivers*, Artech House.

Van Dierendonck, A.J. (1996). GPS receivers, in *Global Positioning System: Theory and Applications I*, B. Parkinson, J. Spilker, P. Axelrad and P. Enge (eds.), AIAA, pp. 329–408.

Van Dierendonck, A.J., P. Fenton, and T. Ford (1992). Theory and Performance of Narrow Correlator Spacing in a GPS Receiver, *Navigation*, vol. 39, no. 3, pp. 265–283.

Vizmuller, P. (1995). *RF Design Guide: Systems, Circuits and Equations*, Artech House.

Ward, Philip W. (1996). Effects of RF Interference on GPS Satellite Signal Receiver Tracking, in *Understanding GPS Principles and Applications*, Elliott D. Kaplan (ed.), Artech House, pp. 209–236.

Ziemer, R.E. and W.H. Tranter (1995). *Principles of Communications: Systems, Modulation, and Noise* (4th edition), John Wiley.

PART IV
Receivers

The received GPS signal is woefully ill-suited for computer processing. Even though we might like to, we cannot simply connect the GPS antenna to a fast computer. First, the signal power must be increased by approximately ten orders of magnitude in power, or 100 dB. Second, natural noise and man-made radio-frequency interference (RFI) must be removed to the extent possible. When the signal arrives, these competing signals are much stronger than the desired GPS signal. Third, the received carrier frequency is 1.5 billion cycles per second and most computers would have a difficult time coping with such a rapid variation. Consequently, we must down convert the frequency to something more manageable. Finally, we must convert from an analog signal to a digital signal. In short, signal conditioning is required, and one means to accomplish such processing is described in Chapter 10.

After conditioning, the signal has been amplified to a level that ignites the analog to digital (A/D) converter and creates numbers inside the digital portion of the receiver. Even though the signal is now well suited for digital processing, we have much work left to do. We must estimate the arrival time, τ, because it contains the basic range and clock information required to compute user position and clock offset. We must also estimate the Doppler shift, f_D, because it contains the pseudorange rate information used to compute the user velocity and clock frequency. If we desire the ultimate in precision, the carrier phase, θ, must also be estimated and tracked.

The estimation of our key triplet, $\{\tau, f_D, \theta\}$, proceeds in two stages. The first stage is a global search for approximate values of $\{\tau, f_D\}$. This process, known as signal acquisition, is described in Chapter 11. The second stage is a local search for accurate estimates of $\{\tau, f_D\}$ that may include estimation of the carrier phase, θ. This process is called signal tracking because it is continuous and the estimates are updated as the receiver and satellites move. Signal tracking is described in Chapter 12.

Needless to say, the signal acquisition and tracking processes suffer from interference if strong competing signals are present or the satellite signals are blocked. After all, the GPS signals travel 20,000 kilometers from medium earth orbit, and the received signal powers are approximately 10^{-16} watts. Signals from terrestrial sources are generally much stronger. If they fall in the GPS portion of the radio spectrum, then the GPS signal acquisition and tracking situation can quickly become bleak. Physical obstructions that block the GPS signals are also appreciable challenges. Fortunately, a growing portfolio of countermeasures exist to mitigate the deleterious effect of GPS signal obstructions or competing radio signals. These are described in Chapter 13.

Chapter 11

Signal Conditioning and Acquisition

A GPS signal experiences many changes on its journey from its satellite to earth. Primarily, as described in Chapter 10, the signal becomes much weaker. Indeed, the transmitted signal has an effective earthward power of 500 watts or so, but the received signal carries only 10^{-16} watts. The resulting whisper is swamped by noise. Recall from Table 10.4 that the noise in the same bandwidth as the received signal can be sixty times stronger than the GPS signal. This attenuation is faithfully captured in the mathematical models for the sig-

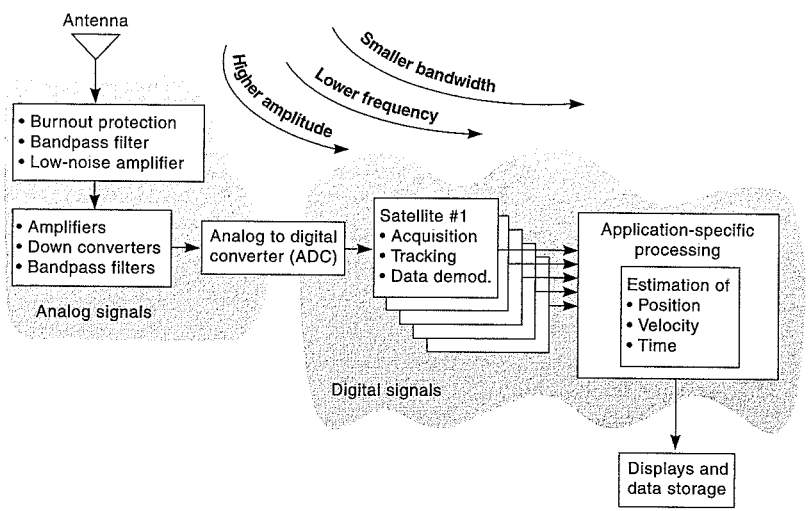


Figure 11.1 GPS receiver block diagram. Software radios strive to minimize cost by hosting the satellite-specific processing on the same processor as the application-specific processing.

nals. At the satellite, a GPS signal is well modeled as

$$s(t) = \sqrt{2P_{\text{tmt}}}\, D(t)x(t)\cos(2\pi f_L t + \theta_{\text{tmt}}) \tag{11.1}$$

At the receiver, we find

$$r(t) = \sqrt{2P_{\text{rcv}}}\, D(t-\tau)x(t-\tau)\cos\left(2\pi(f_L + f_D)t + \theta_{\text{rcv}}\right) + n(t) \tag{11.2}$$

where $P_{\text{tmt}} >> P_{\text{rcv}}$. The subscripts "tmt" and "rcv" refer to "transmitted" and "received," respectively. The rest of the notation should now be familiar. The noise, $n(t)$, is not the only competitor, however. Signals from the other GPS satellites also overlap in frequency and time.

Even so, the receiver manages to capture this tiny signal and extract the position, velocity and time information that we seek. Specifically, the receiver estimates τ for position and time, f_D for velocity, and θ for precise position. To do so, the receiver is configured as a multi-stage system as shown in Figure 11.1. The first stage of the receiver *conditions* the signal so that it is suitable for computer processing. This *front end* removes interfering signals in adjacent frequency bands by filtering the received signal. It also amplifies the power of the signal by approximately 10^{10}, and it reduces the carrier frequency of 1575.42 MHz by a factor between 100 and 1000.

During conditioning, the signal is an analog voltage. After adequate conditioning, the received signal is converted to a sequence of digital numbers by an analog-to-digital converter (ADC). This digital signal (or sequence) takes a finite number of values and only changes values at discrete times.

In general, receiver designs strive to minimize the cost and complexity of the conditioning section of the receiver. They wish to convert the signal to digital form as soon as possible. The digital components are not affected by temperature or humidity, and the unit-to-unit variation is negligible. In contrast, analog components are sensitive to the environment. In addition, some of the digital components can be reprogrammed if a design change is desired. The analog components would need to be replaced.

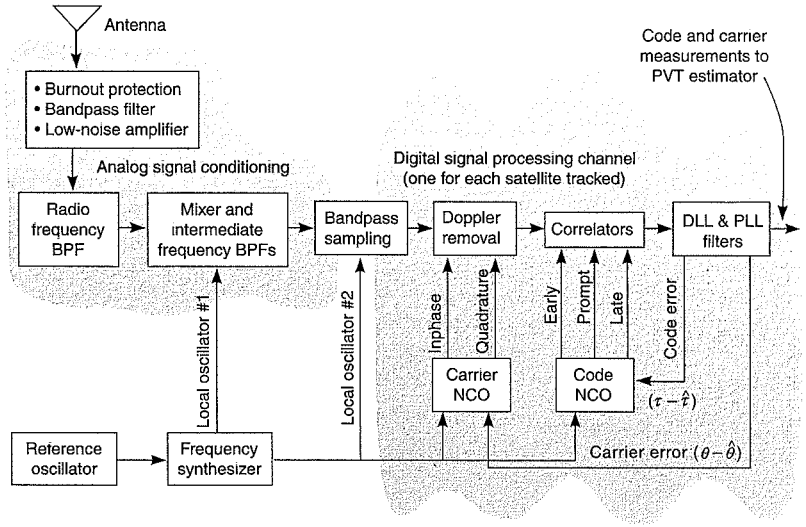


Figure 11.2 Expanded block diagram for the signal processing portions of the receiver emphasized in Chapters 11 and 12.

As shown in Figure 11.1 and Figure 11.2, the digital portion of the receiver contains a bank of estimators—one for each satellite. Each estimator contains the means to track the key signal parameters for its satellite. A carrier tracking loop, detailed in Chapter 12 and shown in Figure 11.2, uses feedback to estimate the Doppler frequency and phase of the received carrier. A delay lock loop, introduced in Chapter 9 and further developed in Chapter 12, uses feedback to track the arrival time of the spread spectrum code.

This chapter and the next focus on the functions of the receiver shown in Figure 11.2. Section 11.1 discusses the remainder of the signal conditioning chain. As mentioned earlier, this chain amplifies the signal. It also reduces the bandwidth and carrier frequency of the signal. Bandpass sampling is discussed because many receivers use this technique to minimize the complexity and cost of signal conditioning. Section 11.2 discusses *signal acquisition*, where a receiver acquires rudimentary knowledge of the Doppler frequency and code arrival time for each signal. This section will also introduce inphase and quadrature sampling, Doppler removal, and the ambiguity function. Section 11.3 analyzes the probability that the receiver will fail to acquire the Doppler frequency and code arrival time. This probability is plotted as a function of signal-to-noise ratio, C/N_0, introduced in Chapter 10. Some of the details of the noise analysis are relegated to Appendices 11.A and 11.B.

For clarity, this chapter will deal mostly with one satellite signal. From time to time, we will also ignore the noise, $n(t)$, so that we can focus on the sensibility of the signal processing. Chapters 11 and 12 do not further discuss receive antennas or low noise amplifiers (LNA), because these were covered in Chapter 10. Nor do they say more on the estimation of position, velocity, or time (PVT), because these were covered in Chapters 5, 6, and 7.

11.1 Signal Conditioning

The front end of a GPS receiver conditions the incoming signal for digital processing. Figures 11.2 and 11.4 show the front end of a generic GPS receiver. Importantly, many such approaches exist. This section discusses the following aspects of this process: frequency down-conversion, intermediate frequencies, image frequencies, and signal sampling.

11.1.1 Frequency Down Conversion

Frequency down conversion is based on the trigonometric identity

$$(A\cos\alpha)(B\cos\beta)=\frac{AB}{2}\left(\cos(\alpha+\beta)+\cos(\alpha-\beta)\right) \tag{11.3}$$

In our case, $A\cos\alpha$ is the received signal, and $B\cos\beta$ is the receiver generated reference signal.

$$\begin{aligned} A\cos\alpha &= \sqrt{2C}x(t-\tau)D(t-\tau)\cos\left(2\pi(f_L+f_D)t+\theta\right) \\ B\cos\beta &= \sqrt{2}\cos\left(2\pi(f_L-f_{IF})t+\theta_{IF}\right) \end{aligned} \tag{11.4}$$

The power in the received signal is denoted C. This notation follows the protocol introduced in Chapter 10, where C is the received power, P_{rcv}, increased by any antenna gain and decreased by any implementation losses. The received signal is at the GPS transmit frequency of f_L where this may be f_{L1}, f_{L2}, or f_{L5}. It is also Doppler shifted by f_D hertz and phase shifted by

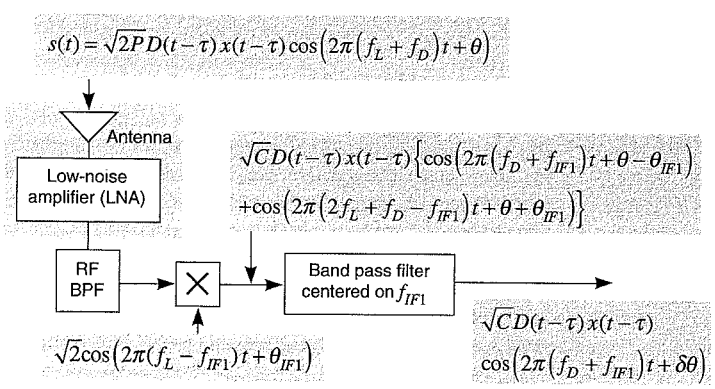


Figure 11.3 Mixers and intermediate frequency stages for a generic signal from one satellite.

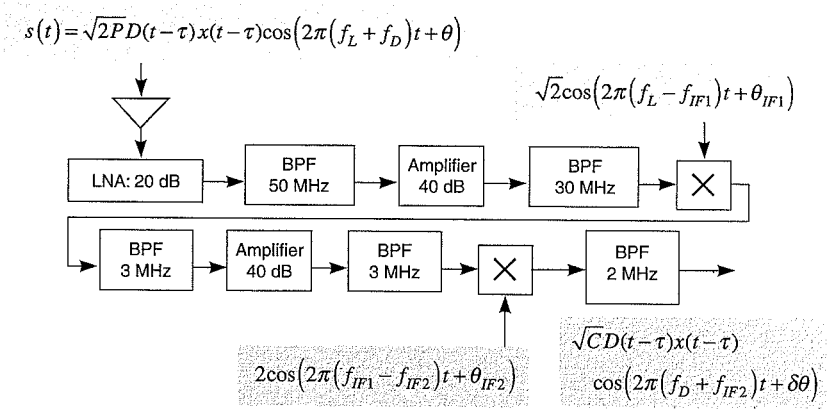


Figure 11.4 Multiple mixers and intermediate frequency stages are often employed.

θ. The locally generated signal has frequency equal to f_L-f_{IF}, where IF stands for intermediate frequency. It also has phase shift, θ_{IF}, that is random relative to the phase of the received GPS signal. Finally, it has an amplitude that is much greater than the received signal. We set this amplitude to $\sqrt{2}$ without loss of generality to ease the subsequent manipulations. After all, the noise and signal are multiplied by the same factor, and so we can choose this coefficient for convenience.

As shown in (11.3), the multiplication of the received signal and reference signal generates two terms. One is called the sum term because it has a frequency equal to the sum of the two input frequencies. The other is called the difference term because it has a frequency equal to the difference of the input frequencies.

$$\begin{aligned} \frac{AB}{2}\cos(\alpha+\beta) &= \sqrt{C}\,x(t-\tau)D(t-\tau)\cos\left(2\pi(2f_L-f_{IF}+f_D)t+\theta+\theta_{IF}\right) \\ \frac{AB}{2}\cos(\alpha-\beta) &= \sqrt{C}\,x(t-\tau)D(t-\tau)\cos\left(2\pi(f_{IF}+f_D)t+\theta-\theta_{IF}\right) \end{aligned} \tag{11.5}$$

The Doppler shift, f_D, is much lower than f_L, and f_{IF} is usually chosen to be much lower than f_L. Consequently, $2f_L >> f_{IF}$ and a filter readily removes the sum term. This filter is a bandpass filter (BPF) and only passes signals near f_{IF}. As shown, the resulting signal, $(AB/2)\cos(\alpha-\beta)$, is equal to the received signal translated down in frequency. This process of multiplying and filtering is called *mixing*.

Why use an intermediate frequency? Why not mix the received signal with a reference sinusoid at frequency f_{L1} and recover the codes and data signals without using the intermediate frequency steps shown in Figure 11.3 or 11.4? Such a process is called direct conversion to baseband and is possible in principle. However, direct conversion requires extremely careful design of the receiver hardware.

The receiver hardware must be designed with greater care because strong and weak sig-

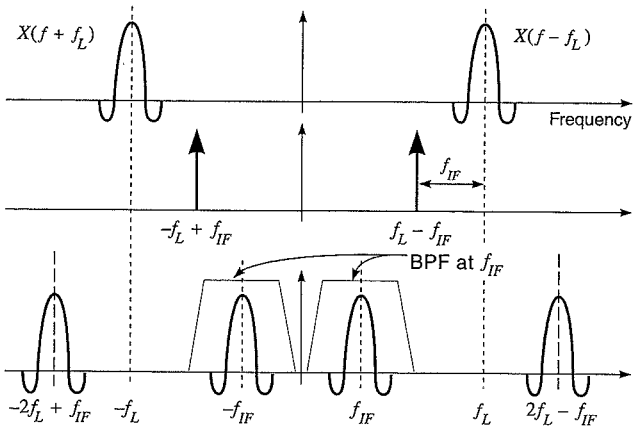


Figure 11.5 Fourier transforms of received signal before and after down conversion.

nals co-exist at one frequency. With direct conversion, the receiver front end performs all of its functions at either f_{L1} = 1575.42 MHz or the baseband frequency. As mentioned earlier, the front end must amplify the received signal by 10 orders of magnitude. If this required gain is evenly allocated between the two frequencies, then a gain of 10^5 is required at each frequency. In both cases, the output signal will be over one hundred thousand times more powerful than the input signal. If the strong output leaks back to the input, then the amplification would be applied to the stray signal, and oscillation could result in the receiver front end. In this case, our receiver would no longer function as an amplifier, but rather as an oscillator! For this reason, most receivers use one or more intermediate frequency stages, and the amplification is distributed over these multiple frequencies. The receiver in Figure 11.4 has two intermediate frequencies with amplification at the carrier frequency and both intermediate frequencies.

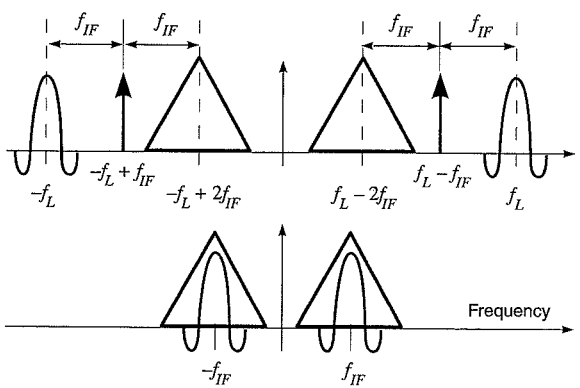


Figure 11.6 Image frequencies must be removed prior to down conversion.

11.1.2 Image Frequencies

The mixing process described above can be analyzed using Fourier transforms. The modulation property of Fourier transforms is especially helpful. As listed in Table 8.1, this property prescribes that

$$\mathcal{F}\{a(t)\cos(2\pi f_0 t)\} = \frac{A(f-f_0)}{2} + \frac{A(f+f_0)}{2} \tag{11.6}$$

In the satellite, multiplication by a sinusoid, at frequency f_0, splits $A(f)$ into two pieces, with one half moving up in frequency by f_0 and the other half moving down by an equal amount. In the receiver, mixing splits each of the two input pieces into two more pieces. As shown in Figure 11.5, the receiver splits its input spectrum into four pieces located at $\{-2f_{L1}+f_{IF}, -f_{IF}, f_{IF}, 2f_{L1}-f_{IF}\}$. Of these, the bandpass filter only passes the signals at $\{-f_{IF}, f_{IF}\}$.

Image frequencies are frequencies other than f_{L1} that are also moved to f_{IF} by the mixing process. As shown in Figure 11.6, these are located at $\{-f_{L1}+2f_{IF}, f_{L1}+2f_{IF}\}$. These frequencies are generally well outside of the spectral band reserved for GPS and may contain signals much stronger than the GPS signal. After all, the GPS signals are likely to be much weaker than signals generated on the surface of the earth. For this reason, the receiver front end must greatly attenuate any signals at the image frequencies before down conversion. This function is a critical requirement for the front end filters that operate at f_{L1}.

11.1.3 Sampling

Within the conditioner, the received signal is analog. It takes a continuum of values and is defined for all instants of time. Such a waveform is shown in Figure 11.7. After conditioning is complete, the received signal is converted from analog to digital. Our model for this conversion is shown in Figure 11.8, where the analog signal is multiplied by a sampling waveform. The resulting signal is discrete in time, but continuous in amplitude.

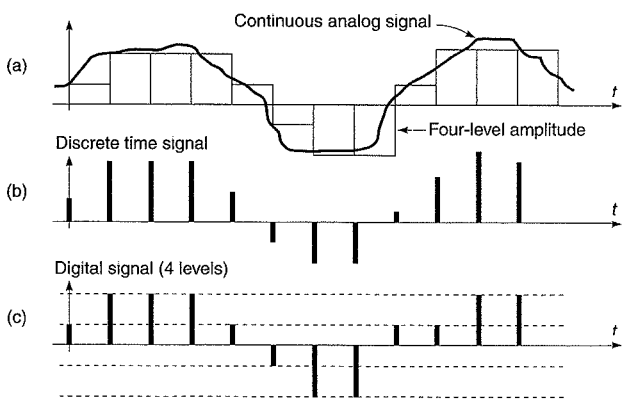


Figure 11.7 Types of signal: (a) continuous analog, (b) discrete time, and (c) digital.

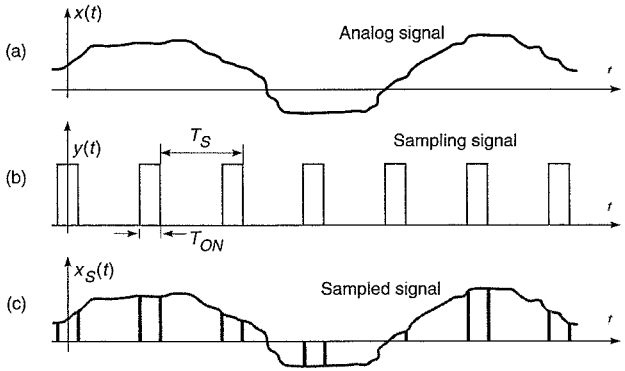


Figure 11.8 (a) Analog signal, (b) sampling signal, and (c) sampled signal.

Analog to digital converters also discretize the amplitude of the input signal. For example, one-bit converters simply retain the sign of the input signal. These are simple, but do incur a slight performance loss in natural noise. The SNR loss can be larger when one-bit A/Ds are used in the presence of narrowband radio frequency interference. We will have more to say on this topic in Chapter 13.

Many GPS receivers use a two-bit A/D and these have four levels of quantization as shown in Figure 11.7. Such receivers do not suffer much SNR loss relative to a receiver without quantization, but they do require automatic gain control (AGC). AGC is placed in front of the A/D and controls the amplitude of the incoming signal. It ensures that the signal amplitude is spread amongst the quantization levels of the A/D. We refer the reader to Van Dierendonck (1996) and Spilker and Natalie (1996) for more on this interesting subject.

The sampling waveform, $y(t)$, shown in Figure 11.8 has a high sample rate compared to the rate of change of the signal to be sampled, $x(t)$. If the sample rate is high enough, then the samples can be used to faithfully reconstruct $x(t)$. This notion is captured in the uniform sampling theorem for low-pass signals [Ziemer and Tranter (1995)]. If $x(t)$ has no energy above some frequency f_{UP}, then it can be reconstructed from samples uniformly spaced in time with a sample rate higher than $2f_{UP}$ samples per second. Such an approach based on the maximum frequency of the signal is called baseband sampling.

Bandpass sampling is an alternative to baseband sampling that is finding increased use in GPS receivers. It combines the sampling task with the down conversion task and is motivated by the uniform sampling theorem for bandpass signals [Ziemer and Tranter (1995)]. If $x(t)$ has upper frequency limit, f_{UP}, and bandwidth, B, then it can be reconstructed from uniform samples at a rate of $2f_{UP}/m$ where $m = \lfloor f_{UP}/B \rfloor$ and $\lfloor \cdot \rfloor$ is the largest integer not exceeding the argument. The required sample rate is now more or less dominated by the bandwidth of the signal and not its highest frequency component.

Our analysis of bandpass sampling begins with the sampling waveform, $y(t)$, shown in Figure 11.8(b). This signal may be written as

$$y(t) = \sum_{j=-\infty}^{\infty} p\left(\frac{t - jT_S}{T_{ON}}\right) \tag{11.7}$$

where $p(t)$ is the elementary rectangular pulse waveform defined in Section 8.1.3, and $y(t)$ is the pulse train analyzed in Section 8.4.2. T_S is the time between samples, and T_{ON} is the on-time of the sampling waveform. In Section 8.4.2, we found that the Fourier series is given by

$$\begin{aligned} y(t) &= \frac{T_{ON}}{T_S}\left(1 + \sum_{n=1}^{\infty} \frac{2\sin(\pi n f_S T_{ON})}{\pi n f_S T_{ON}} \cos(2\pi n f_S t)\right) \\ &= \frac{T_{ON}}{T_S}\left(1 + \sum_{n=1}^{\infty} 2\,\mathrm{sinc}(\pi n f_S T_{ON}) \cos(2\pi n f_S t)\right) \end{aligned} \tag{11.8}$$

The Fourier transform for our sampling waveform can now be readily found since $y(t)$ is a sum of cosine waves.

$$Y(f) = \frac{T_{ON}}{T_S}\left(\delta(f) + \sum_{n=1}^{\infty} \mathrm{sinc}(\pi n f_S T_{ON})\big(\delta(f - nf_S) + \delta(f + nf_S)\big)\right) \tag{11.9}$$

This function is depicted in Figure 11.9 along with the original sampling function. As shown, $Y(f)$ consists of a comb of delta functions. The amplitude of this comb is modulated by the sinc function mentioned earlier. The first nulls of the modulating sinc function occur at $f = \pm 1/T_{ON}$. In this work, we will assume that T_{ON} is small and ignore the slow variation due to the sinc function. In practice, T_{ON} need only be small compared to the period of the intermediate frequency.

Contrast the comb function shown in Figure 11.9 to the pair of delta functions shown in Figure 11.5 for the sinusoidal local oscillator described earlier in this chapter. Recall that the cosine wave generated two aliases for the original spectrum. By doing so, it created the spectrum shown in the bottom trace of Figure 11.5. Our current sampling signal, $y(t)$, generates a pair of aliases for every impulse shown in the bottom trace of Figure 11.9. This seems messy, but if the sampling frequency is well chosen, we can achieve a very nice effect.

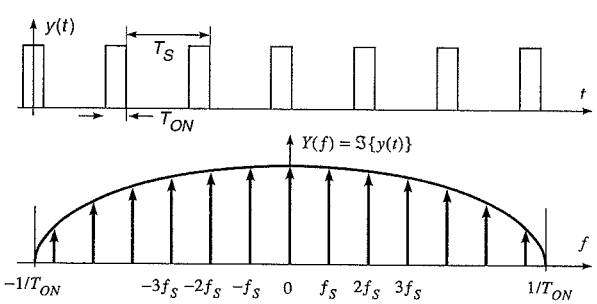


Figure 11.9 Sampling waveform and its Fourier transform.

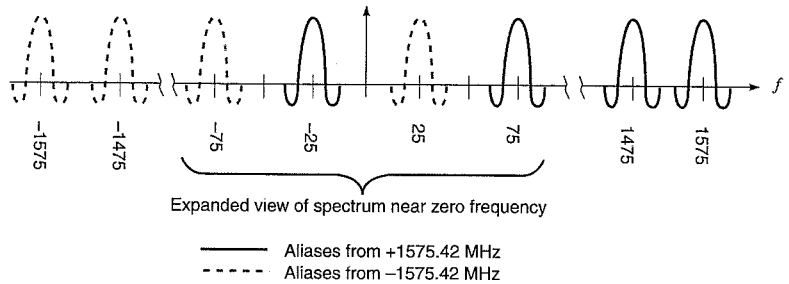


Figure 11.10 Spectra of aliased signals created by bandpass sampling.

As an example, consider a sampling rate of 100 million samples per second (f_S = 100 Msps). The GPS spectrum centered at 1575 MHz will be shifted up and down to $1575 \pm n100$ MHz, and the spectrum at -1575 will be shifted to $-1575 \pm n100$ MHz. The former will create aliases at {... –1625, –1525, ... –25, 75, 175, ... 1475, 1575, ...} MHz, and the latter will create aliases at {... –1675, –1575, ... –75, 25, 125, ... 1425, 1525, 1625, ...} MHz. The lowest frequency aliases are of greatest interest to us. They are sampled waveforms that have been frequency converted all the way down to ±25 MHz! These convenient signals are shown nearest the origin in Figure 11.10. These are the signals that we desire for further processing by the receiver, but we must separate them from the raft of image signals generated by bandpass sampling.

The images closest to 1575 are shown in Figure 11.11. As shown, any signal near 1525 MHz or 1625 MHz will be aliased down to 25 MHz when a sampling frequency of 100 Msps is used. Any signal near 1475 MHz or 1675 MHz will be aliased down to –25 MHz. These aliases must be attenuated before sampling; otherwise they will wreak havoc with the desired GPS signal. Figure 11.12 depicts the same action in the time domain. As shown, signals at two

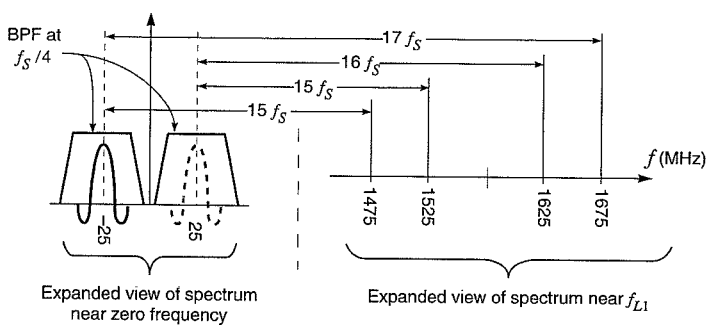


Figure 11.11 Multiplicity of image frequencies near f_{L1} and $-f_{L1}$ when sampling is used with a sampling rate of 100 Megasamples per second.

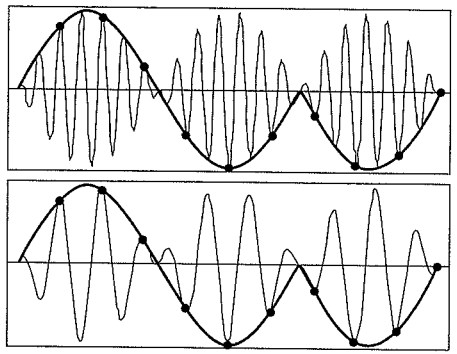


Figure 11.12 Bandpass sampling of a desired signal and an alias. Signals at two different frequencies can produce the same set of samples.

different frequencies produce the same set of samples. An anti-aliasing filter must be used to remove these images prior to sampling.

The image ladder shown in Figure 11.13 neatly summarizes the discussion above [Akos (1997), Poppe (1998)]. Frequency increases as we climb the zigzag ladder. The left hand corners in the zigzag are at integer multiples of the sample frequency, nf_S, with the bottom left hand corner at 0 Hz. The right hand corners are at $(n + 0.5)f_S$, and the bottom right hand corner is at $0.5f_s$ which is 50 MHz for our example. All input frequencies fall on a zig or a zag someplace.

The sample frequency, f_S, is chosen such that the GPS signal falls midway between the vertical bars. The GPS frequency band is centered at 1575 MHz, which is in the middle of the rung that connects 1550 to 1600 MHz. The overall band begins at 1565 and ends at 1585 MHz, and it is marked with the heavy solid line.

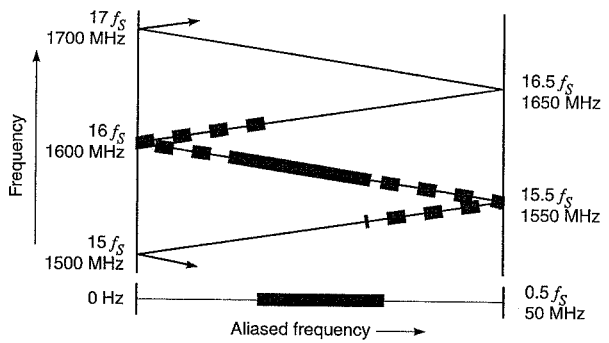


Figure 11.13 Image ladder when bandpass sampling is used with a sampling rate of 100 million samples per second for bandpass sampling [from Poppe (1999)].

Bandpass sampling aliases the frequencies of all input signals, and they fall down the ladder to the bottom rung. So the GPS signal at 1575 MHz falls straight down to 25 MHz. However, the image signals centered at 1625, 1675 and so forth also come to rest at this location. This nifty visualization gives this tool its name—image ladder.

Our new tool also makes it easy to visualize the filter needed to suppress the images. The ideal filter passes 1565 to 1585 MHz, and stops 1615 to 1635 MHz and 1515 to 1535 MHz. Any signals in these two stop-bands would fall on top of the aliased GPS signal at the bottom of the ladder. The dashed region in Figure 11.13 is the transition band, where the filter attenuation grows from the pass-band to the stop-band. To suppress these images, the filter has increasing attenuation in these regions.

In summary, the image ladder shows four frequency regions. Starting on the bottom rung, the heavy line shows the desired alias for the GPS input signal. The light line that zigs and zags upwards shows the frequencies that should be rejected by a filter to suppress images. As the input frequency continues to increase, the heavy dashed line shows frequencies in a transition region where the filter starts to pass frequencies that do not interfere with the GPS signal. The heavy solid line covers the desired GPS band. The input filter should pass signals in this region with no significant attenuation. As frequency continues to increase, we re-enter the transition band and eventually return to the stop band.

11.2 Signal Acquisition

Signal conditioning is complete. The signal has been amplified to a level that ignites the A/D converter and creates numbers inside the digital portion of the receiver. The signal from one satellite is

$$s(t) = \sqrt{C}\, D(t-\tau)\, x(t-\tau) \cos\left(2\pi(f_{IF} + f_D)t + \delta\theta\right) + n(t) \qquad (11.10)$$

The front end of the receiver has greatly amplified the received signal. However, (11.10) does not show this amplification factor because both the noise and the signal are amplified by the same factor, and only their ratio is critical to the analyses that follow. In addition, we still write the signal as a continuous function of time, even though it is now a discrete time signal. After analog to digital conversion, the signal is in fact a sequence that is only defined for a discrete set of times. We choose to ignore this because we assume that the sample rate is high enough to faithfully represent $s(t)$. Moreover, our purpose is to offer insight into the signal processing steps, and this insight comes readily from the continuous function.

Even though the signal is now well suited for processing, we have much work left to do. We must estimate the arrival time, τ, because it contains the basic range and time information required to compute user position and clock offset. We must also estimate the Doppler shift, f_D, because it contains the pseudorange rate information used to compute the user velocity and clock frequency. If we desire the ultimate in precision, the carrier phase offset $\delta\theta$ must also be estimated and tracked.

The estimation of our key triplet, (τ, f_D, $\delta\theta$), proceeds in two stages [Van Trees (1968)]. The first stage is a global search for approximate values of (τ, f_D), and is known as signal acquisition. The second stage is a local search for accurate estimates of (τ, f_D) that may include estimation of the carrier phase offset, $\delta\theta$. If the local search estimates the carrier phase offset, then it is called coherent signal tracking. If the carrier phase offset is ignored, then the search is called noncoherent tracking. The remainder of this chapter describes signal acquisition; tracking is discussed in Chapter 12.

Signal acquisition can be time consuming, and the receiver takes advantage of any available clues to minimize this time. A *cold start* is required when the receiver is first turned on and has no *a priori* information. A *warm start* is possible with a modest amount of *a priori* information. If the receiver has completed signal acquisition and is tracking signals, then *signal reacquisition* refers to the recovery of one signal after the satellite is momentarily blocked.

A cold start proceeds without any *a priori* information and may take several minutes. With no estimate of its position or the current time, the receiver does not know which satellites might be found in the sky. In this case, it must start searching for satellites chosen randomly. A warm start is possible if the receiver knows: its position to within a few hundred miles; time to within ten minutes or so; and possesses a recent satellite almanac. With this information, the receiver knows which satellites to search for and will typically search for a satellite high in the sky. High satellites are preferred because obstructions are unlikely and C/N_0 will probably be high.

After a cold or warm start, signal tracking begins. However, an appreciable amount of time may still be required to acquire all the navigation data. As described in Chapter 4, a complete frame of navigation data requires 30 seconds to send. At a minimum, subframes 1, 2 and 3 are required to solve for position, and they last 18 seconds.

Signal reacquisition refers to the recovery of a signal after only one satellite is momentarily blocked. Since the other satellites are continuously tracked, reacquisition typically takes only a few seconds because the receiver has a very good estimate of position and clock offset. Moreover, the delay lock loop for the lost signal may still have a reasonable estimate of the code delay, τ, for that satellite. The receiver only needs to search the code shifts and Doppler frequencies, (τ, f_D), that are neighbors of the last estimates for the lost signal.

Sections 11.2 and 11.3 describe the global search for $\{\tau, f_D\}$ and our discussion is based on Figures 11.14 through 11.18. Section 11.2.1 discusses inphase and quadrature processing. Section 11.2.2 introduces the ambiguity function. This fascinating creature is a generalization of the correlation function and provides a crisp visualization of the acquisition process. For fun, Section 11.2.3 plots the ambiguity function for the length-31 Gold codes that were introduced in Chapter 9. These codes only have 31 chips and they are little brothers of the length-1023 codes used by GPS. Section 11.2.4 derives the ambiguity function for random codes. Section 11.2.5 describes the search area that is traversed during signal acquisition. This area is two dimensional with code delay, τ, along one axis, and Doppler shift, f_D, along the other axis. The acquisition algorithm must find the cell in this area that corresponds to the best global estimates of τ and f_D. Section 11.3 addresses the role of noise in signal acquisition and culminates in a plot of the probability of acquisition failure as a function of C/N_0. The details of the analysis are relegated to Appendices 11.A and 11.B.

We do not discuss the myriad of techniques that leverage *a priori* information to aid signal acquisition. Finally, we limit our discussion to serial search and do not discuss parallel search or transform-based signal acquisition.

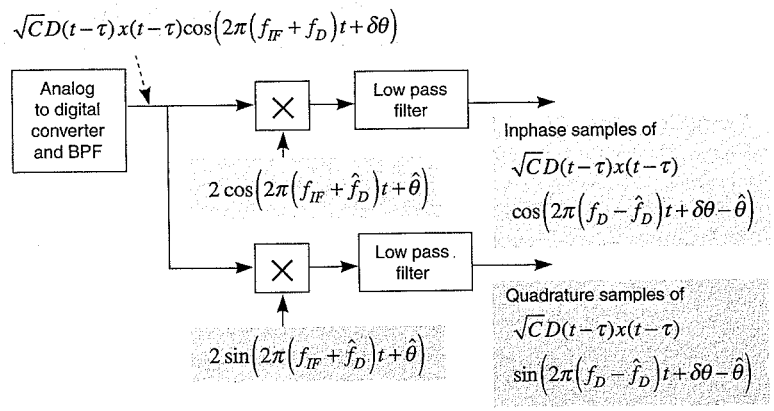


Figure 11.14 Inphase and quadrature sampling combined with Doppler removal. The signal from only one satellite is shown and the noise component is suppressed.

11.2.1 Inphase and Quadrature Processing and Doppler Removal

The GPS receiver employs two reference signals.

$$\sqrt{2}\cos\left(2\pi(f_{IF}+\hat{f}_D)t+\hat{\theta}\right)$$
$$\sqrt{2}\sin\left(2\pi(f_{IF}+\hat{f}_D)t+\hat{\theta}\right) \tag{11.11}$$

These are called the inphase and quadrature reference signals, respectively. Inphase and quadrature processing is shown in Figure 11.14. The inphase channel multiplies (11.10) by the inphase reference signal and low-pass filters the product. The quadrature channel multiplies (11.10) by the quadrature reference signal and low-pass filters the product.

After low-pass filtering, the outputs from the inphase and quadrature channels are

$$\sqrt{C}\,D(t-\tau)x(t-\tau)\cos(2\pi\Delta f_D t+\Delta\theta)$$
$$\sqrt{C}\,D(t-\tau)x(t-\tau)\sin(2\pi\Delta f_D t+\Delta\theta) \tag{11.12}$$

where

$$\Delta f_D = f_D - \hat{f}_D$$
$$\Delta\theta = \delta\theta - \hat{\theta} \tag{11.13}$$

This process is sometimes called *carrier wipeoff* because the signal is no longer modulated by the carrier frequency or any of the intermediate frequencies. The frequency of the resulting signal is the difference between the true Doppler, f_D, and the receiver's best estimate of Doppler, $\hat{f}_D$. The phase of the signal is the difference between the input phase, $\delta\theta$, and the receiver's best estimate of phase, $\hat{\theta}$.

Why do we require two reference signals? Why not multiply with either the inphase or quadrature reference, but not both? Inphase and quadrature (I/Q) processing enables the receiver to estimate f_D without a good estimate of carrier phase. In other words, the error, $\Delta f_D = f_D - \hat{f}_D$, can be maintained as a small number without any real control over $\Delta\theta$. I/Q processing also enables the receiver to distinguish positive Doppler shifts from negative shifts.

To appreciate these strengths, contemplate a receiver that uses only the inphase reference signal. The product signal, after filtering, would be

$$\sqrt{C}\,D(t-\tau)x(t-\tau)\cos(2\pi\Delta f_D t+\Delta\theta) \tag{11.14}$$

If the Doppler shift is equal to zero, then the product signal is equal to

$$\sqrt{C}\,D(t-\tau)x(t-\tau)\cos(\Delta\theta) \tag{11.15}$$

This signal will fade in amplitude anytime

$$\Delta\theta \approx \left\{\pm\frac{\pi}{2},\pm\frac{3\pi}{2},\pm\frac{5\pi}{2},\cdots\right\} \tag{11.16}$$

Prior to signal acquisition, we have no control over this phase difference, and such amplitude fades would hinder the acquisition of the GPS signals.

If the phase offset, $\Delta\theta$, is equal to zero, the product signal becomes

$$\sqrt{C}\,x(t-\tau)D(t-\tau)\cos(2\pi\Delta f_D t) \tag{11.17}$$

In this case, a closing pseudorange cannot be distinguished from an opening pseudorange because $\cos(-x) = \cos x$. A closing pseudorange refers to a pseudorange where the Doppler is positive, because the satellite is approaching. An opening pseudorange has a negative Doppler because the satellite is moving away from the user.

In summary, I/Q processing obviates problematic amplitude fading and enables the receiver to distinguish between a closing range and an opening range.

Code wipeoff usually follows the carrier wipeoff process described above. A pair of correlators is shown in Figure 11.15—one for the inphase channel and one for the quadrature channel. As shown, the signal contributions to the correlator outputs are given by

$$S_I(\Delta\tau,\Delta f_D,\Delta\theta)=\frac{\sqrt{C}\,D}{T_{CO}}\int_0^{T_{CO}} x(t-\tau)x(t-\hat{\tau})\cos(2\pi\Delta f_D t+\Delta\theta)\,dt$$
$$S_Q(\Delta\tau,\Delta f_D,\Delta\theta)=\frac{\sqrt{C}\,D}{T_{CO}}\int_0^{T_{CO}} x(t-\tau)x(t-\hat{\tau})\sin(2\pi\Delta f_D t+\Delta\theta)\,dt \tag{11.18}$$

For this equation, $\Delta\tau$ is the error in the code delay, and so $\Delta\tau = \tau - \hat{\tau}$.

This equation averages over T_{CO} seconds. These averages are called coherent averages because they preserve information about the carrier phase, $\Delta\theta$. As such, they contrast noncoherent averages that destroy this information and will be described shortly.

The coherent averaging time, T_{CO}, needs to be shorter than the 20 ms duration of a data bit, and we assume that $D = +1$ or -1 during the entire coherent averaging time. We must limit T_{CO} to minimize the probability of integrating across a data bit boundary. If we do cross a boundary and the data bit changes sign, then signal energy will be partially lost and acquisition performance will suffer. This particular danger will be further addressed in Chapter 13 when we talk about assisted GPS (AGPS).

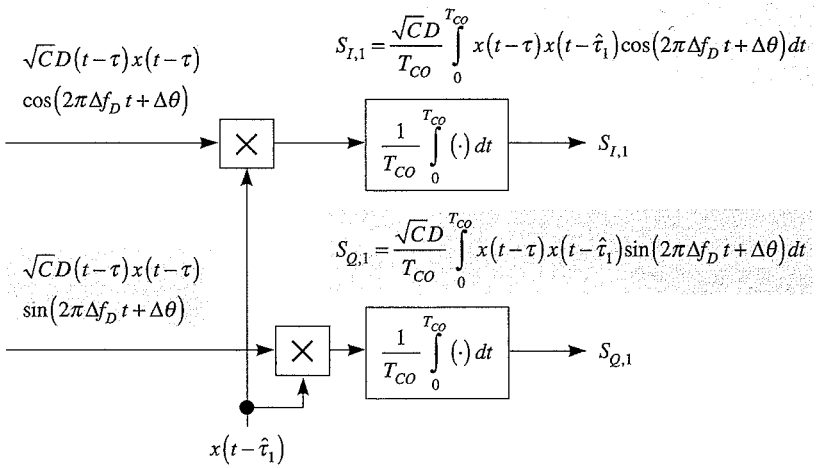


Figure 11.15 Inphase and quadrature correlators. The signal from only one satellite is shown and the effect of noise is not shown.

11.2.2 Ambiguity Function

Figure 11.16 compacts our notation for I/Q processing. As shown, the inphase and quadrature reference signals are now the real and imaginary parts of a complex reference signal.

$$\sqrt{2}\exp\left(-j\left(2\pi(f_{IF}+\hat{f}_D)t+\hat{\theta}\right)\right) \tag{11.19}$$

The corresponding outputs from the low-pass filter are also shown in Figure 11.16 as the real and imaginary parts of

$$\sqrt{C}\,D(t-\tau)x(t-\tau)\exp\left(j(2\pi\Delta f_D t+\Delta\theta)\right) \tag{11.20}$$

As shown in Figure 11.16, complex notation streamlines our notation for the outputs of the correlators. With this notation, (11.18) becomes

$$\begin{aligned}\tilde{S} &= S_I + jS_Q \\ &= \sqrt{C}\,D\exp(j\Delta\theta)\,\tilde{R}(\Delta\tau,\Delta f_D) \\ \tilde{R}(\Delta\tau,\Delta f_D) &= \frac{1}{T_{CO}}\int_0^{T_{CO}} x(t-\tau)\,x(t-\hat{\tau})\exp(j2\pi\Delta f_D t)\,dt\end{aligned} \tag{11.21}$$

This output, $\tilde{S}$, showcases the ambiguity function, $\tilde{R}(\Delta\tau,\Delta f_D)$, which is related to the correlation function and will be central to our work in the remainder of this chapter. Relative to the correlation function, the ambiguity function incorporates the kernel $\exp(j2\pi\Delta f_D t)$. Consequently, it depends on the Doppler error, Δf_D, as well as the code phase error, $\Delta\tau$—our two variables of immediate interest. Henceforth, we shall abbreviate our notation for this important function, $\tilde{R}=\tilde{R}(\Delta\tau,\Delta f_D)$.

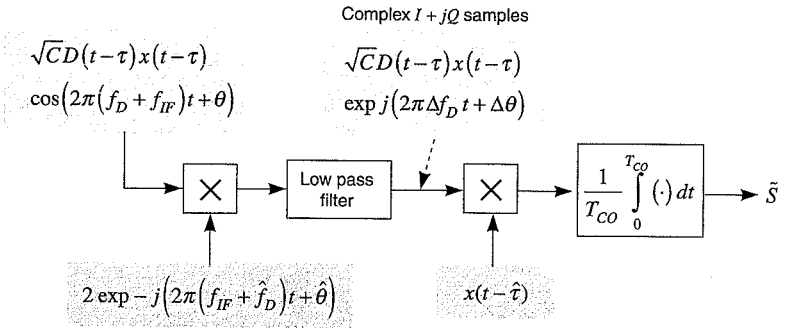


Figure 11.16 Inphase and quadrature correlation using complex notation. The signal contribution from only one satellite is shown, and the noise component is suppressed.

To estimate $(\Delta\tau,\Delta f_D)$, the receiver takes the magnitude of $\tilde{S}$ to strip the two nuisance parameters—the data bit, D, and the carrier phase offset, $\Delta\theta$. Unless treated, these parameters would obscure $(\Delta\tau,\Delta f_D)$.

$$\begin{aligned}\left|\tilde{S}\right|^2 &= S_I^2+S_Q^2 \\ &= C\left|\tilde{R}\right|^2\end{aligned} \tag{11.22}$$

This squaring operation provides a measurement that can be used to estimate $(\Delta\tau,\Delta f_D)$. However, this same squaring operation introduces a penalty. Recall that the correlator output contains signal plus noise, and the squaring operation squares the noise as well as the signal. We will return to this subject in Section 11.3 when we discuss the noise performance of signal acquisition. First, however, we will compute the ambiguity function for our length-31 Gold code and for random codes.

11.2.3 Ambiguity Function of a Length-31 Gold Code

In this subsection, we present plots of the ambiguity function for the length-31 Gold codes discussed in Chapter 9. Recall that we use these short codes as manageable stand-ins for the length-1023 codes used by GPS. The magnitude of the ambiguity function for a length-31 code is shown in Figure 11.17. Note that it has a main peak at $(\Delta\tau=0,\ \Delta f_D=0)$ surrounded by a bumpy surface. The peak at $(\Delta\tau=31T_C,\ \Delta f_D=0)$ is due to the periodicity of the code.

The auto-correlation function for the length-31 code is seen as a slice of the ambiguity function along the $\Delta f_D=0$ axis. Recall from Chapter 9 that the maximum magnitude for non-zero shifts is 9/31. Homework Problem 11-1 asks to you to compute the ambiguity function along the $\Delta\tau=0$ axis.

The Gold codes continue to amaze. As we know from Chapter 9, they have auto-correlation functions with one main peak and small sidelobes. As we search across the two-dimensional space, $(\Delta\tau,\Delta f_D)$, shown in Figure 11.18, we do find peaks with magnitude larger than 9/31, but they are not much larger. The main lobe remains distinct from all sidelobes even when we introduce Doppler offsets! By the way, the ambiguity function between different codes also has this nice property; no large new peaks are introduced.

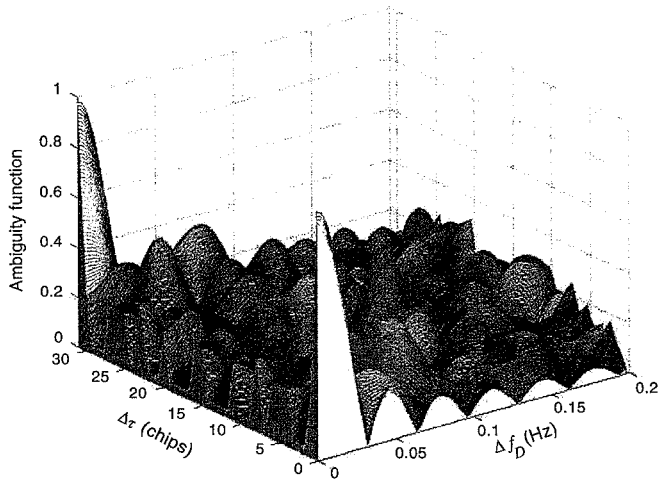


Figure 11.17 Ambiguity function for a length-31 Gold code, with $T_C = 1$ s and $T_{CO} = 31$ s for simplicity.

The ambiguity functions for the length-1023 codes used by GPS are qualitatively the same as the length-31 results shown in Figure 11.17. As shown in Figure 9, the $\Delta f_D = 0$ sidelobes for the length-1023 codes take the values {−1/1023, −65/1023, 63/1023}. Once again, the search over both dimensions, $(\Delta\tau, \Delta f_D)$, does find peaks with magnitude larger than 65/1023, but they are not much larger.

11.2.4 Ambiguity Function for Random Codes

Random codes enable a very compact (and popular) expression for the ambiguity function. Recall from Section 9.2 that these codes are created by choosing the chip polarity randomly. They are created as if the sign of each chip is chosen by flipping a coin. The probability of sending a chip with positive sign is 1/2, and the probability of sending a chip with negative sign is also 1/2. In addition, the sign for any given chip is independent of the sign for all other chips.

In Sections 9.3 and 9.4, we used random codes to compute the average auto-correlation, $\bar{R}(\Delta\tau)$, and average cross-correlation functions. We found that these average functions gave reasonable approximations of these functions for the actual GPS codes. Random codes enable a similar development for the ambiguity function. The proof follows the work that we did in Section 9.2, and so we relegate it to Homework Problem 11-2. The result is as follows.

$$E\{\tilde{R}\} = \bar{R}(\Delta\tau)\exp(j\pi\Delta f_D T_{CO})\,\mathrm{sinc}(\pi\Delta f_D T_{CO}) \tag{11.23}$$

$\bar{R}(\Delta\tau)$ is given in (9.83) and is the average auto-correlation function derived using random codes.

With this result, we may write

$$\begin{aligned} E\{\tilde{S}\} &= \sqrt{C}\,D\exp(j\Delta\theta)E\{\tilde{R}\} \\ &= \sqrt{C}\,D\exp(j\Delta\theta)\bar{R}(\Delta\tau)\exp(j\pi\Delta f_D T_{CO})\,\mathrm{sinc}(\pi\Delta f_D T_{CO}) \\ &= \sqrt{C}\,D\exp(j\Delta\theta_1)\bar{R}(\Delta\tau)\,\mathrm{sinc}(\pi\Delta f_D T_{CO}) \\ \Delta\theta_1 &= \Delta\theta + \pi\Delta f_D T_{CO} \end{aligned} \tag{11.24}$$

This approximation is popular because all three of our estimanda, $(\Delta\tau, \Delta f_D, \Delta\theta_1)$, appear independently with no integrals.

As usual, we are concerned about the variance of these results. After all, the convenience of the average results will be small solace if the individual codes have ambiguity functions that differ greatly from the average. Once again, we simply present the result for the variance and reserve the (arduous) proof for Homework Problem 11-3.

$$\mathrm{var}\{\tilde{R}\} = E\{|\tilde{R}|^2\} - |E\{\tilde{R}\}|^2$$

$$= \begin{cases} \left(\dfrac{\tau}{T_C}\right)^2 \dfrac{1}{N} & -T_C < \tau < +T_C \\ \left(\dfrac{\tau - iT_C}{T_C}\right)^2 \dfrac{1}{N} + \left(\dfrac{(i+1)T_C - \tau}{T_C}\right)^2 \dfrac{1}{N} & iT_C < \tau < (i+1)T_C\,;\, i \neq \{-1, 0\} \end{cases} \tag{11.25}$$

As shown, the approximations $E\{\tilde{S}\}$ and $E\{\tilde{R}\}$ from Equations (11.23) and (11.24) become sharp as the sequence length, N, grows. We will use these results in the statistical analysis of Section 11.3, where it will simplify our work greatly to replace $\tilde{S}$ for length-1023 Gold codes with $E\{\tilde{S}\}$ for random codes of the same length.

11.2.5 Search Area

During signal acquisition, the receiver conducts a search over the $(\Delta\tau, \Delta f_D)$ space shown in Figure 11.18. The search dwells at each possible value of $(\Delta\tau, \Delta f_D)$ long enough to determine

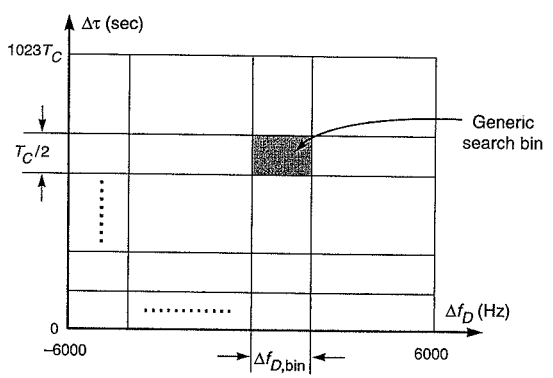


Figure 11.18 Signal search area covers Doppler frequency, Δf_D, and code phase, $\Delta\tau$. The total number of cells in the search space M is equal to the number of code phase bins times the number of Doppler bins.

whether a peak is present. Absent a peak, the search continues to the next candidate value of $(\Delta\tau, \Delta f_D)$. If a peak is detected, then the receiver transitions to the tracking mode described in Chapter 12.

The search over the code error, $\Delta\tau$, is broken into 0.5 chip steps because the main peak of $\tilde{R}(\Delta\tau, \Delta f_D)$ is only one chip wide along the τ axis. This means that we will use 2046 steps to cover the length of the C/A-code. Typically, the search over Doppler error, Δf_D, is broken into search bins that are approximately 500 hertz wide, $\Delta f_{D,\text{bin}} = 500$ Hz. The maximum Doppler range is approximately ±6000 Hz, so a bin width of 500 hertz implies that 24 bins cover the Doppler range. All told, we may need $M = 2046 \times 24 = 49{,}104$ tests to find the true correlation peak in this field of code and Doppler errors.

In fact, the Doppler bin width, $\Delta f_{D,\text{bin}}$, depends on the coherent averaging time, T_{CO}, and this connection is shown in (11.24). In general, we would like the coherent averaging time to be as long as possible, so we can accumulate as much signal energy as possible. However, longer T_{CO} means that we have to use smaller Doppler bins because the sinc function shown in (11.24) becomes narrower. After all, the sinc function in (11.24) takes the value zero when $\pi\Delta f_D T_{CO} = \pi$. To stay away from this condition, we might set $T_{CO} = 1/(2\Delta f_D)$. With $T_{CO} = 1$ millisecond, we can use $\Delta f_{D,\text{bin}} = 500$ Hz. These are typical values. If we extend the coherent averaging time to $T_{CO} = 2$ milliseconds, then we must use $\Delta f_{D,\text{bin}} = 250$ Hz. Some receivers use much longer values of T_{CO} to find very weak GPS signals. Since they must use smaller Doppler bins, they use outside assistance to limit the total Doppler uncertainty. This idea belongs to a family of techniques called Assisted GPS, which is further discussed in Chapter 13.

11.3 Statistical Analysis of Signal Acquisition

This section seeks the probability that noise prevents the receiver from successfully acquiring the signal. It concludes with the acquisition threshold, which is the C/N_0 required for reliable signal acquisition. As described above, Figure 11.18 divides the overall search space into M cells. We search across these cells for a strong auto-correlation peak. The vast majority of the cells are empty because they contain no peak. We say these cells have *noise only*. One or two cells contain the peak or a portion of the peak. For simplicity, our analysis assumes that only one cell contains the peak. This cell is said to contain *signal plus noise*.

The goal of acquisition is to find the one signal-plus-noise cell amongst the great multiplicity of noise-only cells. To this end, we evaluate a carefully chosen acquisition test statistic in each cell. These decision metrics, $\{L_m\}_{m=1}^{M}$, measure the likelihood that the signal is contained in the associated cell. In other words, they measure the likelihood that the estimates, $(\hat{\tau}_m, \hat{f}_{D,m})$, are the best estimates of the true code phase and Doppler shift. We say that L_m measures the likelihood that the signal is in bin m.

For the example described in Section 11.2.5, $M = 49{,}104$ because we scan 2046 possible values of code phase and 24 possible values of Doppler frequency. Acquisition succeeds if the decision metric for the signal-plus-noise cell is greater than the decision metrics for all the noise-only cells.

For simplicity, we assume, without loss of generality, that the signal is hidden in the first cell. In other words, $\hat{\tau}_1$ and $\hat{f}_{D,1}$ are the best guesses of the true code phase and Doppler frequency. The acquisition algorithm fails if $L_m > L_1$ for any $m \neq 1$. In such a case, we would conclude that $\hat{\tau}_m$ and $\hat{f}_{D,m}$ are closer to the true code shift and Doppler frequency than $\hat{\tau}_1$ and $\hat{f}_{D,1}$.

The remainder of our statistical analysis is divided into four sections. Section 11.3.1 introduces the union bound and our noise model. Section 11.3.2 analyzes a fictitious acquisition algorithm that assumes that the carrier phase is known, $\Delta\theta \approx 0$. Such phase coherence cannot exist prior to signal acquisition, and so the results are optimistic. However, the analysis is simple and hopefully instructive. Section 11.3.3 analyzes a more realistic noncoherent algorithm, and Section 11.3.4 discusses our numerical results. Some details of our statistical analysis are collected in Appendices 11.A and 11.B.

11.3.1 Union Bound

As mentioned above, acquisition fails if the metric from any of the noise-only cells is larger than the decision metric from the one signal-plus-noise cell. More formally, the probability of failure is the probability of a union of events. If any of these events occurs, then the algorithm fails. This union of events is shown in Figure 11.19, which is a Venn diagram, and

$$Pr(\text{acquisition failure}) = Pr\left(\bigcup_{m=2}^{M}\{L_m > L_1\}\right) \tag{11.26}$$

L_1 is the decision metric for the one and only signal-plus-noise cell. As shown in the figure, the union is somewhat smaller than the sum of the individual areas. This makes sense, because the sum of the individual areas includes the areas where events overlap. The sum includes events where two or more noise-only cells report higher metrics than the channel with the signal. The union bound simply overbounds the area of the union with the sum of the individual areas, and so we write

$$Pr(\text{acquisition failure}) \leq \sum_{m=2}^{M} Pr(L_m > L_1) \tag{11.29}$$

As mentioned earlier, the signal is hidden in the first cell and completely absent from all other cells. Moreover, the noise power is equal in all cells. Consequently, the pairwise failure probabilities are all equal.

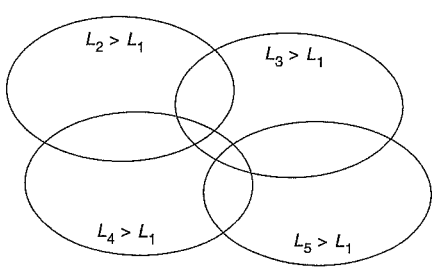


Figure 11.19 Venn diagram showing the sensibility of the union bound for $M = 5$. As shown, the area of the union of the four events is smaller than the sum of the individual areas. So the probability of the union is smaller than the sum of the probabilities.

$$Pr(L_m > L_1) = Pr(\text{pairwise failure}) \tag{11.28}$$

This allows us to define L_0 as the decision metric for a generic noise-only bin, and write

$$\begin{aligned} Pr(\text{acquisition failure}) &\le (M-1)Pr(\text{pairwise failure}) \\ &\le (M-1)Pr(L_1 < L_0) \\ &\le (M-1)Pr(L_1 - L_0 < 0) \end{aligned} \tag{11.29}$$

In other words, we only need to focus on the pairwise failure probability, $Pr(L_1 < L_0)$, which gives the probability that the decision metric for a generic noise-only bin exceeds the metric for the signal-plus-noise bin. The probability of acquisition failure is equal to the pairwise failure probability multiplied by $M-1$, where M is the number of cells.

The next section develops an expression for the pairwise error probability based on a fictitious coherent acquisition algorithm. After that, we treat the realistic noncoherent case. In both cases, we assume that the received noise, $n(t)$, is additive white Gaussian noise with a power spectral density of $N_0/2$ W/Hz. This means that our complex correlator outputs are now given by

$$\begin{aligned} \tilde{Z}_m &= \tilde{S}_m + \tilde{\eta}_m \\ &= S_{I,m} + jS_{Q,m} + \eta_{I,m} + j\eta_{Q,m} \end{aligned}$$

$\tilde{S}_m$ is given by (11.21) with $\hat{\tau} = \hat{\tau}_m$ and $\Delta f_D = f_D - \hat{f}_{D,m}$. In addition,

$$\tilde{\eta}_m = \frac{\sqrt{2}}{T_{CO}} \int_0^{T_{CO}} n(t)x(t-\hat{\tau}_m)\exp\left(j2\pi(f+\hat{f}_{D,m})t + \hat{\theta}\right)dt$$

11.3.2 Coherent Analysis

For our coherent signal acquisition algorithm (albeit fictitious), we will decide which channel contains the signal by seeking the largest value for the real part of the measured correlation. We may ignore the imaginary part of the measured correlation because $\Delta\theta \approx 0$, and so the imaginary part of the signal correlation is equal to zero, $\text{Im}\{\tilde{S}\} = 0$. In short,

$$L_m = \text{Re}\{\tilde{Z}_m\} = Z_{I,m} = S_{I,m} + \eta_{I,m}$$

where the subscript I stands for inphase, and $\eta_{I,m}$ is noise due to the white noise, $n(t)$, that is received in addition to the signal.

L_0 continues to denote a generic channel without signal, and so we may write

$$\begin{aligned} L_0 &= S_{I,0} + \eta_{I,0} \\ &= \sqrt{C}\cos(\Delta\theta_1)\bar{R}(\Delta\tau_0)\text{sinc}(\pi\Delta f_{D,0}T_{CO}) + \eta_{I,0} \\ &\approx \eta_{I,0} \end{aligned} \tag{11.30}$$

This simplification follows because either $|\Delta\tau_0| > T_C$ and so $\bar{R}(\Delta\tau_0) \approx 0$, or $|\Delta f_{D,0}| > 1/2T_{CO}$, and so $\text{sinc}(\pi\Delta f_{D,0}T_{CO}) \approx 0$.

L_1 denotes the channel with signal plus noise.

$$\begin{aligned} L_1 &= S_{I,1} + \eta_{I,1} \\ &\approx \sqrt{C} + \eta_{I,1} \end{aligned} \tag{11.31}$$

In this case, $\Delta\tau_1 \approx 0$ and $\Delta f_{D,m} \approx 0$, and so $\bar{R}(\Delta\tau_0) \approx 1$ and $\text{sinc}(\pi\Delta f_{D,1}T_{CO}) \approx 1$. Clearly, this is an approximation because the code delay and Doppler estimates will not be perfect, but the resulting error is not large if the search area is sampled densely.

At this point, we seek the pairwise probability

$$Pr(\text{pairwise failure}) = Pr(L_1 - L_0 < 0) \tag{11.32}$$

Since the input noise, $n(t)$, is Gaussian, the metrics $L_0 = \eta_{I,0}$ and $L_1 = \sqrt{C} + \eta_{I,1}$ have Gaussian distributions as well. The Gaussian distribution is a welcome visitor to any analysis because it can be completely characterized by its mean and variance. In fact, our sought after pairwise error probability is

$$Pr(\text{pairwise failure}) = Q\left(\frac{E\{L_1 - L_0\}}{\sqrt{\text{var}\{L_1 - L_0\}}}\right) \tag{11.33}$$

This result uses the complement of the cumulative distribution function for the Gaussian random variable.

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-t^2/2)\,dt \tag{11.34}$$

All of this means that if we can find the mean and variance of $L_1 - L_0$, we will be able to write the pairwise error probability for the coherent acquisition algorithm. The following moments are developed in Appendix 11.A.

$$\begin{aligned} E\{L_1 - L_0\} &= \sqrt{C} \\ \text{var}\{L_1 - L_0\} &= 2\sigma^2 \\ &= \frac{N_0}{T_{CO}} \end{aligned} \tag{11.35}$$

Recall that $N_0/2$ is the power spectral density for the received noise, $n(t)$, and T_{CO} is the coherent averaging time.

With these moments, we may write

$$\begin{aligned} Pr(\text{pairwise failure}) &= Q\left(\sqrt{\frac{C}{2\sigma^2}}\right) \\ &= Q\left(\sqrt{\frac{CT_{CO}}{N_0}}\right) \end{aligned} \tag{11.36}$$

This result is the pairwise probability of acquisition failure for a coherent algorithm. It is plotted in Figure 11.20 as the solid black curve. The union bound for the probability of acquisition failure is

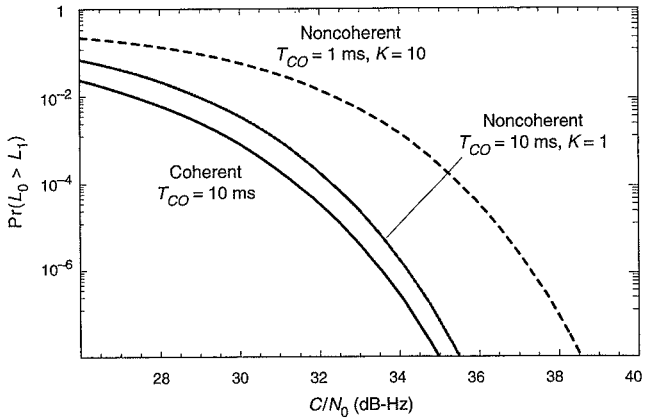


Figure 11.20 Pairwise probability of acquisition failure. To estimate the probability of acquisition failure, these probabilities must be multiplied by $M-1$, where M is the total number of acquisition tests.

$$Pr(\text{acquisition failure}) \le (M-1)Q\left(\sqrt{\frac{CT_{CO}}{N_0}}\right) \tag{11.37}$$

11.3.3 Noncoherent Analysis

We now turn our attention to the noncoherent algorithm that compensates for the absence of any estimate of the carrier phase. As described above, we will not detail every step in this analysis because it is similar to the more manageable coherent analysis. For the details, we refer the reader to Helstrom (1968), Helstrom (1995) and Van Trees (1968). We also provide some challenging homework problems that ask you to prove several of the key results contained in this section and the associated Appendix 11.B.

The noncoherent algorithm is shown in Figure 11.21. It takes the magnitude of the correlator outputs to remove the influence of the unknown carrier phase and navigation data bits. Consequently, we must consider the complex correlations. We assume that $\tilde{Z}_0$ is the complex correlation for any one of $M-1$ cells with incorrect estimates, and $\tilde{Z}_1$ is the complex correlation for the one and only cell with the correct estimates of code phase and Doppler shift. We call $\tilde{Z}_0$ the *noise-only* channel and $\tilde{Z}_1$ is the channel with *signal plus noise*.

$$\begin{aligned}
\left|\tilde{Z}_0\right|^2 &= \left|\tilde{\eta}_0\right|^2 \\
\left|\tilde{Z}_1\right|^2 &= \left|\tilde{S}_1 + \tilde{\eta}_1\right|^2 \\
&= \left|\sqrt{C}\,D\exp(j\Delta\theta_1)\tilde{R}(\Delta\tau_1, \Delta f_{D,1}) + \tilde{\eta}_1\right|^2 \\
&\approx \left|\sqrt{C}\,D\exp(j\Delta\theta_1)\overline{R}(\Delta\tau_1)\mathrm{sinc}(\pi\Delta f_{D,1}T_{CO}) + \tilde{\eta}_1\right|^2
\end{aligned} \tag{11.38}$$

If we choose to use these decision metrics, then

$$\begin{aligned}
L_0 &= \left|\tilde{Z}_0\right|^2 \\
L_1 &= \left|\tilde{Z}_1\right|^2
\end{aligned} \tag{11.39}$$

In this case, we have the following exact expression from Ziemer and Tranter (1995).

$$\begin{aligned}
Pr(\text{pairwise failure}) &= Pr(L_1 - L_0 < 0) \\
&= \frac{1}{2}\exp\left(-\frac{C}{4\sigma^2}\right) \\
&= \frac{1}{2}\exp\left(-\frac{CT_{CO}}{2N_0}\right)
\end{aligned} \tag{11.40}$$

The corresponding probability of acquisition failure is

$$Pr(\text{acquisition failure}) = \frac{M-1}{2}\exp\left(-\frac{CT_{CO}}{2N_0}\right) \tag{11.41}$$

More likely, we average K such magnitudes together before making a decision about which channel contains the signal. This embellishment is also shown in Figure 11.21. The integral over T_{CO} is called the *predetection* filter or coherent average, while the average of K samples is called noncoherent averaging.

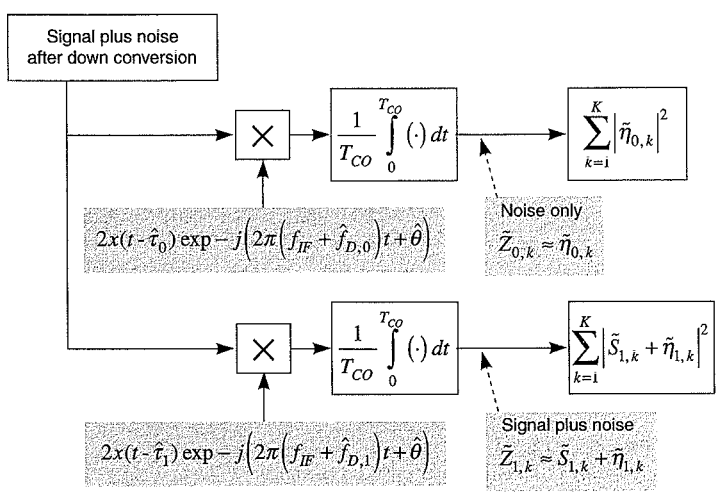


Figure 11.21 Two complex correlators configured for search over code phase and Doppler frequency. The top correlator does not have accurate estimates of code delay and Doppler frequency. Hence, the signal correlation is weak and we call this correlator the *noise-only* channel. In contrast, the bottom correlator has accurate estimates of both parameters. The signal correlation is strong and we call this correlator the *signal-plus-noise* channel.

When we include noncoherent averaging, our decision statistics are

$$\begin{aligned} L_0 &= \sum_{k=1}^{K} \left|\tilde{Z}_{0,k}\right|^2 = \sum_{k=1}^{K} \left|\tilde{\eta}_{0,k}\right|^2 \\ L_1 &= \sum_{k=1}^{K} \left|\tilde{Z}_{1,k}\right|^2 = \sum_{k=1}^{K} \left|\tilde{S}_{1,k} + \tilde{\eta}_{1,k}\right|^2 \\ &= \sum_{k=1}^{K} \left|\sqrt{C}\, D \exp(j\Delta\theta) + \tilde{\eta}_{1,k}\right|^2 \\ &= \sum_{k=1}^{K} \left|\sqrt{C} \exp(j\Delta\theta) + \tilde{\eta}_{1,k}\right|^2 \end{aligned} \tag{11.42}$$

The second to last line leverages our assumption that the code phase and Doppler shift estimates for bin 1 are close to the truth. Consequently, $\Delta\tau_1 \approx 0$ and $\Delta f_{D,m} \approx 0$, and so $\bar{R}(\Delta\tau_0) \approx 1$ and $\text{sinc}(\pi \Delta f_{D,1} T_{CO}) \approx 1$. The last line simply assumes that the data bit, D, does not change during the coherent averaging time.

For $K > 1$, we must use the following Gaussian approximation

$$\begin{aligned} Pr(\text{pairwise failure}) &= Pr(L_1 - L_0 < 0) \\ &\approx Q\left(\sqrt{\frac{(E\{L_1 - L_0\})^2}{\text{var}\{L_1\} + \text{var}\{L_0\}}}\right) \end{aligned} \tag{11.43}$$

This result holds because L_1 and L_0 are uncorrelated. Using the moments listed in Appendix 11.B, this result can be developed as follows.

$$\begin{aligned} Pr(\text{pairwise failure}) &\approx Q\left(\sqrt{\frac{(KC + 2K\sigma^2 - 2K\sigma^2)^2}{4K(C\sigma^2 + \sigma^4) + 4K\sigma^4}}\right) \\ &= Q\left(\sqrt{\frac{KC}{2\sigma^2}\left(\frac{1}{2 + \frac{4\sigma^2}{P}}\right)}\right) \\ &= Q\left(\sqrt{\frac{KCT_{CO}}{2N_0}\left(\frac{1}{1 + \frac{1}{\frac{CT_{CO}}{N_0}}}\right)}\right) \end{aligned} \tag{11.44}$$

In these equations, $\sigma^2 = N_0 / 2T_{CO}$.

The corresponding probability of acquisition failure is

$$Pr(\text{acquisition failure}) = (M-1)\, Q\left(\sqrt{\frac{KCT_{CO}}{2N_0}\left(\frac{1}{1 + \frac{1}{\frac{CT_{CO}}{N_0}}}\right)}\right) \tag{11.45}$$

These results for the pairwise error probability, $Pr(L_1 < L_0)$, are plotted in Figure 11.20 and discussed in the next section.

11.3.4 Discussion

Our coherent and noncoherent results are shown in Figure 11.20, which contains three curves corresponding to (11.36), (11.40) and (11.44). All three plot the pairwise error probability versus C/N_0, so they must be multiplied by $M - 1$ to estimate the probability of acquisition failure. All three assess signal acquisition based on 10 ms of data.

The coherent curve, shown as the black curve in Figure 11.20, shows that the probability of pairwise failure falls to a probability of 10^{-8} at $C/N_0 = 35$ dB-Hz. We call $C/N_0 = 35$ dB-Hz the *acquisition threshold*. The gray curve is the exact noncoherent result for $K = 1$ and $T_{CO} = 10$ ms from (11.40). As shown, the acquisition threshold grows to approximately $C/N_0 = 36$ dB-Hz. Apparently, the magnitude operation does not introduce much performance loss provided T_{CO} is still 10 ms. However, the dashed curve is the noncoherent approximation from (11.44) with $K = 10$ and $T_{CO} = 1$ ms. In this case, the acquisition threshold has almost reached $C/N_0 = 38$ dB-Hz. So, performance for the noncoherent acquisition algorithm with $K = 10$ and $T_{CO} = 1$ ms is 3 dB weaker than the coherent algorithm. Some of this loss is due to the Gaussian approximation. However, some is due to the use of noncoherent averaging rather than coherent averaging.

11.4 Summary

This chapter has introduced the workings of a GPS receiver. We have addressed the need to condition the signal before converting it into the digital numbers processed by a computer. Conditioning includes amplification, down conversion, and filtering. We also described bandpass sampling because this innovation mixes the signal to a lower frequency while converting the signal into the digital sequence required by the computer.

This chapter discussed signal acquisition in some detail. This process provides coarse estimates of the code delay and Doppler shift. Acquisition ballparks these estimanda, while signal tracking, described in the next chapter, refines the estimates. Our discussion on signal acquisition included a rather detailed noise analysis that estimated the signal-to-noise ratio required for reliable estimation of the code phase and Doppler frequency.

Appendix 11.A Moments for the Coherent Metrics

The mean of L_0 is given by

$$\begin{aligned} E\{L_0\} &= E\{\eta_{I,0}\} \\ &= E\left\{\frac{\sqrt{2}}{T_{CO}}\int_0^{T_{CO}} n(t)\,x(t-\hat{\tau}_0)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)dt\right\} \\ &= \frac{\sqrt{2}}{T_{CO}}\int_0^{T_{CO}} E\{n(t)\}\,x(t-\hat{\tau}_0)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)dt \\ &= 0 \end{aligned} \tag{11.48}$$

The mean of L_1 is given by

$$\begin{aligned} E\{L_1\} &= E\{S_{I,1}+\eta_{I,1}\} \\ &= \sqrt{C} \end{aligned} \tag{11.49}$$

The variances are a little more challenging. Since $E\{\eta_{I,0}\}=0$, we may write

$$\begin{aligned} \mathrm{var}\{L_0\} &= E\{(\eta_{I,0}-E\{\eta_{I,0}\})^2\} = E\{\eta_{I,0}^2\} \\ &= \frac{2}{T_{CO}^2}\int_0^{T_{CO}}\int_0^{T_{CO}} E\{n(t)\,n(s)\}\,x(t-\hat{\tau}_0)\,x(s-\hat{\tau}_0) \\ &\quad\times\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})s+\hat{\theta}\right)dt\,ds \\ &= \frac{2}{T_{CO}^2}\int_0^{T_{CO}}\int_0^{T_{CO}} \frac{N_0}{2}\delta(t-s)\,x(t-\hat{\tau}_0)\,x(s-\hat{\tau}_0) \\ &\quad\times\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})s+\hat{\theta}\right)dt\,ds \\ &= \frac{N_0}{T_{CO}^2}\int_0^{T_{CO}} x(t-\hat{\tau}_0)\,x(t-\hat{\tau}_0) \\ &\quad\times\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)dt \\ &= \frac{N_0}{T_{CO}^2}\int_0^{T_{CO}} \cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)\cos\left(2\pi(f_{IF}+\hat{f}_{D,0})t+\hat{\theta}\right)dt \end{aligned} \tag{11.50}$$

Since the carrier frequency, f_{IF}, varies much more quickly than T_{CO}, we may write

$$\mathrm{var}\{L_0\} = \frac{N_0}{2T_{CO}} \tag{11.49}$$

Recall that the baseband analysis in Chapter 10 found the exact same result (10.33). In both cases, noise decreases with increasing averaging time—a sensible result.

With a similar proof, we find that the variance of L_1 is also equal to $N_0/2T_{CO}$. Consequently, we write

$$\begin{aligned} \sigma^2 &= \mathrm{var}\{L_0\} = \mathrm{var}\{\eta_{I,0}\} \\ &= \mathrm{var}\{L_1\} = \mathrm{var}\{\eta_{I,1}\} \\ &= \frac{N_0}{2T_{CO}} \end{aligned} \tag{11.50}$$

The difference of the two metrics, $L_1 - L_0$, is also a Gaussian random variable because any linear combination of Gaussian random variables will follow a Gaussian distribution. Consequently, we are interested in the mean and variance of $L_1 - L_0$. The mean is straightforward.

$$E\{L_1 - L_0\} = E\{L_1\} - E\{L_0\} = \sqrt{C} \tag{11.51}$$

The variance is given by

$$\begin{aligned} E\{(L_1 - L_0 - E\{L_1 - L_0\})^2\} &= E\{(L_1 - \sqrt{C} - L_0)^2\} \\ &= E\{(L_1 - \sqrt{C})^2\} - 2E\{(L_1 - \sqrt{C})L_0\} + E\{L_0^2\} \\ &= \mathrm{var}\{L_1\} + \mathrm{var}\{L_0\} \\ &= \frac{N_0}{T_{CO}} \\ &= 2\sigma^2 \end{aligned} \tag{11.52}$$

This result follows because L_1 and L_0 are uncorrelated.

$$\begin{aligned} E\{(L_1 - \sqrt{C})L_0\} &= \frac{2}{T_{CO}^2}\int_0^{T_{CO}}\int_0^{T_{CO}} E\{n(t)\,n(s)\}\,x(t-\hat{\tau}_0)\,x(s-\hat{\tau}_1) \\ &\quad\times\cos\left(2\pi(f+\hat{f}_{D,0})t+\hat{\theta}\right)\cos\left(2\pi(f+\hat{f}_{D,1})s+\hat{\theta}\right)dt\,ds \\ &= 0 \end{aligned} \tag{11.53}$$

Appendix 11.B Densities and Moments for the Noncoherent Metrics

11.B.1 Noise Only

If $K = 1$, then our noise-only probability density function (pdf) for our decision statistic, L_0, is

$$p_0(L_0, K=1) = \frac{1}{2\sigma^2}\exp\left(-\frac{L_0}{2\sigma^2}\right) \tag{11.54}$$

From time to time we are interested in the pdf of the normalized decision statistic, $\ell = L/\sigma^2$.

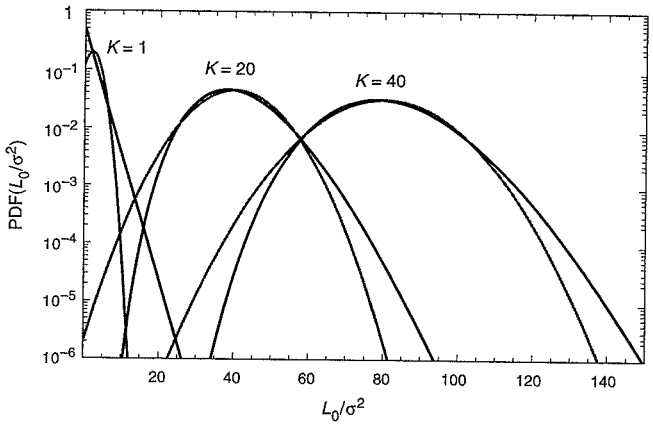


Figure 11.22 Probability density functions and their Gaussian approximations (in gray) for acquisition test statistic in the presence of noise only. K denotes the number of noncoherent samples.

$$p(\ell) = \sigma^2 p(L)\Big|_{L=\sigma^2\ell}$$

$$p_0(\ell_0, K=1) = \frac{1}{2}\exp\left(-\frac{\ell_0}{2}\right) \tag{11.55}$$

As mentioned in Section 11.3.3, acquisition algorithms generally average K measurements. In this case, the pdf for L_0 is [Van Trees (1968) and Helstrom (1968)]

$$p_0(L_0) = \frac{L_0^{K-1}}{(2\sigma^2)^K (K-1)!}\exp\left(-\frac{L_0}{2\sigma^2}\right) \tag{11.56}$$

The corresponding moments are

$$\begin{aligned} E\{L_0\} &= 2K\sigma^2 \\ E\{L_0^2\} &= 4K(1+K)\sigma^4 \\ \text{var}\{L_0\} &= 4K\sigma^4 \end{aligned} \tag{11.57}$$

For the normalized decision statistic, $\ell_0 = L_0/\sigma^2$, the pdf (11.56) becomes

$$p_0(\ell_0) = \frac{(\ell_0/2)^{K-1}}{2(K-1)!}\exp\left(-\frac{\ell_0}{2}\right) \tag{11.58}$$

The associated moments are

$$\begin{aligned} E\{\ell_0\} &= 2K \\ E\{\ell_0^2\} &= 4K(1+K) \\ \text{var}\{\ell_0\} &= 4K \end{aligned} \tag{11.59}$$

Figure 11.22 plots our normalized noise-only pdf for $K = \{1, 20, 40\}$, along with the

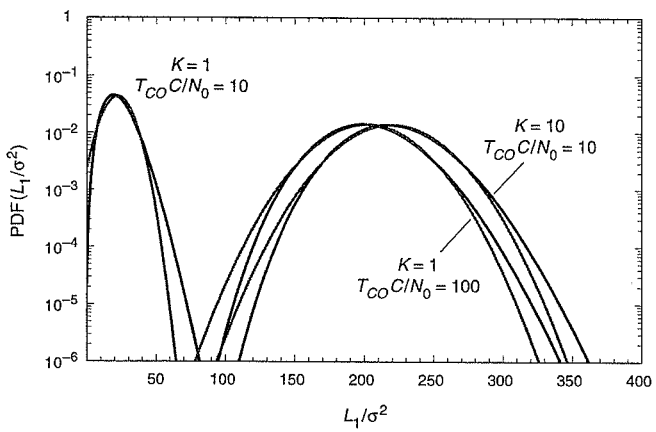


Figure 11.23 Probability density functions and their Gaussian approximations (in gray) for the acquisition test statistic in the presence of signal plus noise. K denotes the number of noncoherent samples, C is the signal power, T_{CO} is the coherent integration time, and N_0 is the noise power spectral density.

Gaussian approximations, $p_G(2K, 4K)$. Notice that the Gaussian approximation for $K = 1$ is pretty poor. Among other things, the Gaussian pdf is non-zero for negative values of ℓ_0. Since ℓ_0 is the sum of squared magnitudes, it cannot be negative. In addition, the Gaussian tail falls off too quickly for positive values of ℓ_0. However, these shortcomings are less important for $K = \{20, 40\}$. As we add together more random variables, the pdf apparently becomes more Gaussian. This trend is slow, but supported by the Central Limit Theorem that claims that the pdf of a sum of random variables with similar variances will tend towards the Gaussian.

11.B.2 Signal Plus Noise

If $K = 1$, then the signal plus noise probability density function (pdf) is

$$\begin{aligned} p_1(L_1, K=1) &= \frac{1}{2\sigma^2}\exp\left(-\frac{L_1+C}{2\sigma^2}\right) I_0\left(\frac{\sqrt{L_1 C}}{\sigma^2}\right) \\ p_1(\ell_1, K=1) &= \frac{1}{2}\exp\left(-\frac{\ell_1+S}{2}\right) I_0\left(\sqrt{\ell_1 S}\right) \end{aligned} \tag{11.60}$$

The second pdf is based on $\ell = L/\sigma^2$ and $S = C/\sigma^2$. I_0 is the modified Bessel function of order 0.

For arbitrary K, our signal-plus-noise pdf is [Helstrom (1968) and Helstrom (1995)]

$$p_1(L_1) = \frac{1}{2\sigma^2}\left(\frac{L_1}{KC}\right)^{\frac{K-1}{2}}\exp\left(-\frac{L_1+KC}{2\sigma^2}\right) I_{K-1}\left(\frac{\sqrt{L_1 KC}}{\sigma^2}\right) \tag{11.61}$$

where I_{K-1} is a modified (or hyperbolic) Bessel function of order $K-1$.

L_1 has the following mean and variance.

$$E\{L_1\} = K(C + 2\sigma^2)$$
$$\text{var}\{L_1\} = 4K(C\sigma^2 + \sigma^4) \tag{11.62}$$

If we normalize as before, $\ell_1 = L_1/\sigma^2$, then we have

$$p_1(\ell_1) = \sigma^2 p_1(L_1)\Big|_{L_1=\sigma^2\ell_1}$$
$$p_1(\ell_1) = \frac{1}{2}\left(\frac{\ell_1}{S}\right)^{\frac{K-1}{2}} \exp\left(-\frac{\ell_1 + S}{2}\right) I_{K-1}\left(\sqrt{\ell_1 S}\right) \tag{11.63}$$

where

$$S = \frac{KC}{\sigma^2}$$
$$E\{\ell_1\} = K\left(\frac{C}{\sigma^2} + 2\right)$$
$$\text{var}\{\ell_1\} = 4K\left(\frac{C}{\sigma^2} + 1\right) \tag{11.64}$$

Figure 11.23 plots $p_1(\ell_1)$ along with the following Gaussian approximations.

$$p_G\left(K\left(\frac{C}{\sigma^2} + 2\right), 4K\left(\frac{C}{\sigma^2} + 1\right)\right) \tag{11.65}$$

When $K = 1$ and the signal-to-noise ratio is weak, $CT_{CO}/N_0 = C/2\sigma^2 = 10$, the Gaussian approximation is okay but not great. If either the number of samples increases, $K = 10$ or the signal-to-noise ratio increases, $CT_{CO}/N_0 = C/2\sigma^2 = 100$, the approximation improves.

Homework Problems

11-1. Compute the ambiguity function along the $\Delta\tau = 0$ axis. Hint: The answer appears along the $\Delta\tau = 0$ axis in Figure 11.17.

11-2. Prove (11.23). In other words, find the expected value of the ambiguity function, $E\{\tilde{R}\}$, when random codes are used. Hint: Use Figure 9.7 and review the development of (9.16).

11-3. Prove (11.25). In other words, find the variance of the ambiguity function

$$\text{var}\{\tilde{R}\} = E\{|\tilde{R}|^2\} - |E\{\tilde{R}\}|^2$$

when random codes are used. Hint: Incorporate your answer from Problem 11-2, and continue to use Figure 9.7 and the development of (9.16).

11-4. Compute the mean and variance for the noncoherent acquisition decision statistic in the presence of noise only when $K = 1$. Hint: There are two ways to proceed. One approach finds

$$\int_0^\infty L_0^\ell p_0(L_0)\, dL_0$$

for $\ell = \{1, 2\}$ where $p_0(L_0)$ is the probability density function given in (11.54). The second approach uses the following.

$$|\tilde{Z}_0|^2 = |\tilde{\eta}_0|^2 = \eta_{I,0}^2 + \eta_{Q,0}^2$$
$$|\tilde{Z}_0|^4 = \eta_{I,0}^4 + 2\eta_{I,0}^2\eta_{Q,0}^2 + \eta_{Q,0}^4 \tag{11.66}$$

The second approach also uses the following result for the fourth moment of a Gaussian random variable. If x is Gaussian with zero mean and variance σ^2, then $E\{x^4\} = 3\sigma^4$. Either approach should find that the answers are $2\sigma^2$ and $4\sigma^4$.

11-5. What happens to the mean and variance computed in the last problem when $K > 1$ and the random variables are independent?

11-6. Compute the mean and variance for the noncoherent acquisition decision statistic in the presence of signal plus noise when $K = 1$. Hint: there are two ways to proceed. One approach finds

$$\int_0^\infty L_1^\ell p_1(L_1)\, dL_1$$

for $\ell = \{1, 2\}$ and $p_1(L_1)$ is the probability density function given in (11.60). The second approach uses the following.

$$\begin{aligned}|\tilde{Z}_1|^2 &= C + 2S_{I,1}\eta_{I,1} + 2S_{Q,1}\eta_{Q,1} + \eta_{I,1}^2 + \eta_{Q,1}^2\\ |\tilde{Z}_1|^4 &= C^2 + 4CS_{I,1}\eta_{I,1} + 4CS_{Q,1}\eta_{Q,1} + 2C\eta_{I,1}^2 + 2C\eta_{Q,1}^2\\ &\quad + 4S_{I,1}^2\eta_{I,1}^2 + 8S_{I,1}S_{Q,1}\eta_{I,1}\eta_{Q,1} + 4S_{I,1}\eta_{I,1}^3 + 4S_{I,1}\eta_{I,1}\eta_{Q,1}^2\\ &\quad + 4S_{Q,1}^2\eta_{Q,1}^2 + 4S_{Q,1}\eta_{I,1}^2\eta_{Q,1} + 4S_{Q,1}\eta_{Q,1}^3\\ &\quad + \eta_{I,1}^4 + 2\eta_{I,1}^2\eta_{Q,1}^2 + \eta_{Q,1}^4\end{aligned} \tag{11.67}$$

The second approach also uses the following result for the fourth moment of a Gaussian random variable. If x is Gaussian with zero mean and variance σ^2, then $E\{x^4\} = 3\sigma^4$. Either approach should find that the answers are $C + 2\sigma^2$ and $4C\sigma^2 + 4\sigma^4$.

11-7. What happens to the mean and variance computed in the last problem when $K > 1$?

11-8. Prove that (11.56) is correct. In other words, prove that the probability density function (pdf) of the sum of the squares of K Gaussian random variables is given by (11.56),

when the underlying Gaussian random variables have zero mean. Hint: The solution relies on some nice results from the theory of probability density functions and characteristic functions. The characteristic function is the Fourier transform of the pdf, and the characteristic function of the sum of independent random variables is the product of the individual characteristic functions. The pdf of the sum is then the inverse Fourier transform of the product of the characteristic functions. For much more on this technique see Helstrom (1968) and Helstrom (1995).

11-9. Prove that the magnitude of a complex correlation with noise only, $|\tilde{Z}_0|$, follows the Rayleigh pdf given by

$$p_{\text{Rayleigh}}\left(|\tilde{Z}_0|\right) = \frac{|\tilde{Z}_0|}{\sigma^2}\exp\left(\frac{|\tilde{Z}_0|^2}{2\sigma^2}\right)u\left(|\tilde{Z}_0|\right) \tag{11.68}$$

Derive the $K = 1$ pdf for noise only (11.54) from the Rayleigh pdf. Hint: Remember that the $K = 1$ pdfs given in Appendix 11.B are for $|\tilde{Z}_0|^2$, not $|\tilde{Z}_0|$.

11-10. Prove that (11.61) is correct. In other words, prove that the pdf of the sum of the squares of K Gaussian random variables is given by (11.61), when the underlying Gaussian random variables have non-zero mean. Hint: See Problem 11-8.

11-11. Prove that the magnitude of a complex correlation with signal plus noise, $|\tilde{Z}_1|$, follows the Rician pdf given by

$$p_{\text{Rician}}\left(|\tilde{Z}_1|\right) = \frac{|\tilde{Z}_1|}{\sigma^2}\exp\left(\frac{-\left(|\tilde{Z}_1|+|\tilde{S}_1|\right)^2}{2\sigma^2}\right)I_0\left(\frac{|\tilde{S}_1|}{|\tilde{Z}_1|}\right)u\left(|\tilde{Z}_1|\right) \tag{11.69}$$

Derive the $K = 1$ pdf for signal plus noise (11.60) from the Rician pdf. Hint: Remember that the $K = 1$ pdfs given in Appendix B are for $|\tilde{Z}_1|^2$, not $|\tilde{Z}_1|$.

11-12. Prove that (11.40) is correct. In other words, derive our stated result for the $K = 1$ pairwise error probability for the noncoherent case. Hint: For $K = 1$, we have

$$Pr\left(|\tilde{Z}_1|^2 - |\tilde{Z}_0|^2 < 0\right) = Pr\left(|\tilde{Z}_1| - |\tilde{Z}_0| < 0\right) \tag{11.70}$$

In other words, it does not matter whether we square the magnitude or not when we are only considering one noncoherent sample. In this case, we may use the Rayleigh and Rician pdfs. For more help, refer to Ziemer and Tranter (1995).

References

Akos, D. (1997). Software Radio Approach to Global Navigation Satellite System Receiver Design, Ph.D. dissertation, Ohio University.

Helstrom, C. (1968). *Statistical Theory of Signal Detection*, Pergamon Press (2nd edition), pp. 217–218.

Middleton, D. (1987). *Introduction to Statistical Communication Theory*, Pennisula Publishing, Los Altos, California.

Poppe, M. (1998). Personal communication on image ladders.

Spilker, J. and F. Natali (1996). Interference Effects and Mitigation Techniques, Chapter 20 in *Global Positioning System: Theory and Applications*, B.W. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), AIAA, pp. 717–772.

Van Dierendonck, A.J. (1996). GPS Receivers, Chapter 8 in *Global Positioning System: Theory and Applications*, B.W. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), AIAA, pp. 329–408.

Van Trees, H.L. (1968). *Detection, Estimation and Modulation Theory, Part I*, John Wiley and Sons.

Ziemer, R.E. and W.H. Tranter (2002). *Principles of Communications: Systems, Modulation and Noise* (5th edition), John Wiley and Sons.

Chapter 12
Signal Tracking

This chapter introduces the delay lock loop (DLL) and phase lock loop (PLL). An appendix discusses the frequency lock loop (FLL). As discussed in Chapter 11, signal acquisition provides rough estimates of the code phase, τ, and the Doppler frequency, f_D. The DLL refines this initial estimate of code phase and tracks changes into the future. To do so, the DLL generates a replica of the PRN code being transmitted by a satellite and keeps it aligned with the received code. This estimate of code phase is combined with the

other terms, described in Section 5.1, to construct the whole pseudorange. The FLL refines the initial estimate of the Doppler frequency and tracks change into the future. To this end, the FLL generates a sinusoidal carrier and keeps the frequency matched to the frequency of the received carrier. Like the FLL, the PLL is a carrier tracking loop, but it also provides estimates of the carrier phase, θ. As such, it must adjust the frequency and phase of the replica sinusoid to match the phase of the incoming carrier. This aggregation of code and carrier operations is called *signal tracking*.

The DLL, PLL and FLL can all be modeled as control systems that use feedback to control the behavior of some plant. This common structure is explained in Section 12.1, which introduces the discriminator function, the plant, feedback, and the loop filter.

Section 12.2 focuses on the delay lock loop (DLL) that tracks the code. The DLL uses a nonlinear discriminator function that extracts an unambiguous estimate of the code tracking error even before the data bits or carrier phase are known. Even though the discriminator is nonlinear, a linear model is valid for normal operating conditions when the replica code is nearly aligned with the received code. In fact, Section 12.2 develops a linear model for the entire DLL.

This linear model is used to analyze the noise and dynamic performance of the DLL. Good noise performance requires the DLL to attenuate the effect of measurement noise. Good dynamic performance requires the DLL to accurately track changes in the code delay due to *system dynamics*. System dynamics include clock drifts and the line-of-sight components of user and satellite motion. All of these dynamics must be tracked.

Section 12.2 also explores rate-aiding, where rate measurements from the PLL or FLL are used to help the DLL. We find that rate-aiding is very helpful because it enables excellent noise performance without sacrificing dynamic performance. Normally, noise performance and dynamic performance conflict.

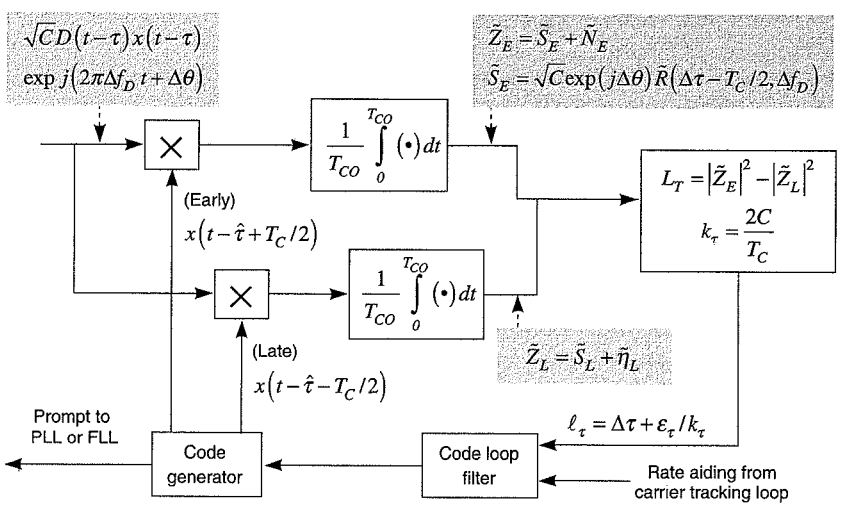


Figure 12.1 Delay lock loop.

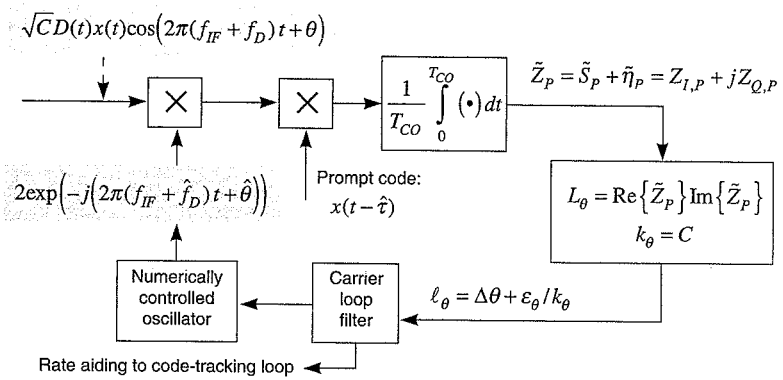


Figure 12.2 Phase lock loop.

Section 12.3 focuses on the phase lock loop (PLL). Once again, a nonlinear discriminator is used. This device, also known as a *phase detector*, provides estimates of the carrier tracking error in spite of unknown data modulation. These detectors are also linear within their normal operating ranges. So we can once again develop a linear model. This linear model is used to explore the role of averaging within the PLL. Averaging can be used to trade between noise performance and dynamic performance. It reduces the impact of noise on the signal parameter estimates, but may obscure authentic dynamics.

Section 12.4 is a summary that reminds us that this chapter is just a bare introduction to signal tracking. This fascinating topic has many variations. An appendix describes the frequency lock loop, which is a valuable adjunct to the phase lock loop.

This entire chapter owes much to the work of Van Dierendonck (1996), Ward (1996) and Spilker (1996). It also makes use of Laplace transforms to characterize dynamic performance and random signals to analyze noise performance. If needed, refer to Chapter 8 for a review of these topics.

12.1 Overview of a Signal Tracker

Signal tracking commences after the acquisition algorithm has significantly reduced the Doppler error and the code error. The correlators used for signal acquisition are reconfigured to perform tracking. The new configurations are shown in Figure 12.1 for code tracking and Figure 12.2 for carrier phase tracking. These two figures show the delay lock loop (DLL) and phase lock loop (PLL), respectively. The frequency lock loop is discussed in the appendix and is shown in Figure 12.A.1.

All three trackers are sophisticated systems and it will take a while to appreciate their complexities. However, they can all be modeled as control systems like the one shown in Figure 12.3. All three take measured correlations as input. All three have discriminators to strip out the desired error signal. All three use feedback to control the behavior of some *plant* as described below.

As such, our trackers are akin to the automatic speed control used in automobiles. For

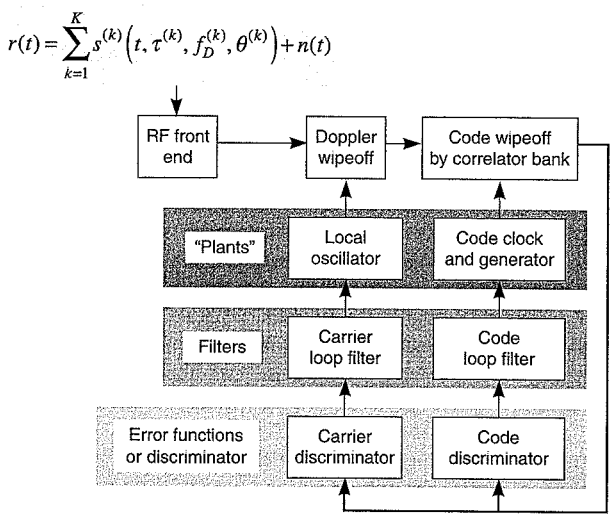


Figure 12.3 GPS receiver showing code and carrier tracking loops.

speed control, the automobile is the plant, and we wish to control its speed so that it is approximately equal to some driver-specified speed while road conditions change. To this end, the actual speed of the car is measured and fed back to the controller. The controller computes an error signal that is the difference between the measured speed and the desired speed. This error signal is then provided to the controller that acts to minimize the error. If the car is going faster than desired, then the controller reduces fuel flow to the engine. If the car is going too slow, then fuel flow is increased. Automatic speed control and the loops used in a GPS receiver are surprisingly similar.

For our DLL, the code generator is the plant that we wish to control. The DLL strives to control the code generator so that the replica code is aligned in time with the received code. It does so by issuing speed-up and slow-down commands to the clock that controls the speed of the replica code generator. The controller does not command abrupt step changes in the time of the replica code. Rather, it simply speeds up or slows down the clock controlling the generator. If the code replica is early, the DLL commands the code generator to slow down. If the code replica is late, the DLL commands the code generator to speed up. Eventually, these commands align the replica code with the received code.

For our PLL, the numerically controlled oscillator (NCO) is the plant that we wish to control. The NCO is akin to the code generator, except that it produces the replica carrier rather than the replica code. The PLL endeavors to synchronize the frequency and phase of the replica to the frequency and phase of the received carrier. To do so, the PLL detects the phase difference and issues commands to the NCO to minimize this difference. Once again, the commands do not precipitate abrupt changes in the output of the NCO. They simply speed up or slow down the NCO to maintain synchronization.

12.1.1 Correlators

All of our trackers take the measured correlations as input. Recall our notation for the complex correlation developed in Chapter 11. The real part of the complex correlation comes from the inphase correlator, and the imaginary part comes from the quadrature correlator. The correlation outputs at time $t = nN_{CO}$ are given by

$$\begin{aligned}\tilde{Z}_n &= \tilde{S}_n + \tilde{\eta}_n \\ &= S_{I,n} + jS_{Q,n} + \eta_{I,n} + j\eta_{Q,n} \\ Z_{I,n} &= \mathrm{Re}\left\{\tilde{Z}_n\right\} \\ &= S_{I,n} + \eta_{I,n} \\ Z_{Q,n} &= \mathrm{Im}\left\{\tilde{Z}_n\right\} \\ &= S_{Q,n} + \eta_{Q,n}\end{aligned} \tag{12.1}$$

These equations use the notation of Section 11.3. The signal contributions to the inphase and quadrature samples are given by $S_{I,n}$ and $S_{Q,n}$, and the noise contributions are given by $\eta_{I,n}$ and $\eta_{Q,n}$.

If we suppress the sample number, n, then the signal contribution is

$$\begin{aligned}\tilde{S} &= S_I + jS_Q \\ &= \sqrt{C}D\exp(j\Delta\theta)\frac{1}{T_{CO}}\int_0^{T_{CO}} x(t-\tau)x(t-\hat{\tau})\exp(j2\pi\Delta f_D t)\,dt \\ &= \sqrt{C}D\exp(j\Delta\theta)\tilde{R}(\Delta\tau,\Delta f_D) \\ \tilde{R}(\Delta\tau,\Delta f_D) &= \frac{1}{T_{CO}}\int_0^{T_{CO}} x(t-\tau)x(t-\hat{\tau})\exp(j2\pi\Delta f_D t)\,dt\end{aligned} \tag{12.2}$$

After signal acquisition, we may assume that the frequency estimate is reasonably accurate, $\Delta f_D \approx 0$. In this case, we may write

$$\tilde{S} = \sqrt{C}D\exp(j\Delta\theta)R(\Delta\tau) \tag{12.3}$$

In this equation, $R(\cdot)$ is the baseband auto-correlation function introduced in Chapter 9, and not the complex-valued ambiguity function in (12.2).

If the frequency estimate is accurate and the code delay is accurate, then $\Delta\tau \approx 0$ and we say the correlator is prompt. (We use subscripts E, L, and P for early, late, and prompt.)

$$\begin{aligned}\tilde{Z}_P &= \tilde{S}_P + \tilde{\eta}_P \\ &\approx \sqrt{C}D\exp(j\Delta\theta) + \tilde{\eta}_P\end{aligned} \tag{12.4}$$

The prompt correlator, $\tilde{Z}_P$, has a central role in the phase lock loop and frequency lock loop.

If the frequency estimate is accurate and the code delay is early or late by $\pm d/2$ chips, then the correlations are, reasonably enough, called early and late.

$$\begin{aligned}\tilde{Z}_E &= \tilde{S}_E + \tilde{\eta}_E \\ &= \sqrt{C}D\exp(j\Delta\theta)R(-dT_C/2) + \tilde{\eta}_E \\ \tilde{Z}_L &= \tilde{S}_L + \tilde{\eta}_L \\ &= \sqrt{C}D\exp(j\Delta\theta)R(+dT_C/2) + \tilde{\eta}_L\end{aligned} \tag{12.5}$$

The noise contributions to the measured correlations were characterized in Appendix 11.A. We model the received noise as additive white noise with power spectral density $N_0/2$. The means and variances are

$$\begin{aligned}E\{\eta_{I,n}\} &= E\{\eta_{Q,n}\} = 0 \\ \text{var}\{\eta_{I,n}\} &= \text{var}\{\eta_{Q,n}\} = \frac{N_0}{2T_{CO}}\end{aligned} \tag{12.6}$$

In addition, the inphase and quadrature samples are uncorrelated and noise samples at different times are also uncorrelated.

$$\begin{aligned}E\{\eta_{I,n}\eta_{Q,m}\} &= 0 \quad \text{for all } m \text{ and } n \\ E\{\eta_{I,n}\eta_{I,m}\} &= 0 \quad \text{for all } m \neq n \\ E\{\eta_{Q,n}\eta_{Q,m}\} &= 0 \quad \text{for all } m \neq n\end{aligned} \tag{12.7}$$

These second-order moments are key to the noise performance results given in the remainder of this chapter.

12.1.2 Discriminators

Discriminators constitute a particularly imaginative portion of the signal trackers studied in this chapter. They implement a nonlinearity that has been carefully chosen to expose the parameter to be estimated while they suppress the effect of other unknown parameters. The DLL must accurately estimate code delay before the PLL begins to track carrier phase because the DLL provides the prompt correlation measurement required by the PLL and FLL. Consequently, the DLL discriminator must be sensitive to code delay error, but insensitive to the unknown carrier phase and data bits. Similarly, the FLL operates when the signal-to-noise ratio is too weak for PLL operation. Its discriminator must be sensitive to frequency error, but insensitive to unknown carrier phase and data bits. Finally, the PLL must track the carrier phase in the presence of an unknown string of navigation bits, so its discriminator must be sensitive to phase error, but insensitive to the data modulation.

By design, discriminators are approximately linear functions of the error to be estimated. We cite three examples below, leaving the details for later sections.

A common DLL discriminator is

$$\begin{aligned}L_\tau &= \left|\tilde{Z}_E\right|^2 - \left|\tilde{Z}_L\right|^2 \\ &\approx k_\tau \Delta\tau + \varepsilon_\tau\end{aligned} \tag{12.8}$$

In this equation, ε_τ is due to measurement noise. We will provide details on this term and the linearization in Section 12.2.

The discriminator can be normalized so that it is equal to the estimandum of interest, $\Delta\tau$ plus normalized noise, ε_τ/k_τ.

$$\begin{aligned}\ell_\tau &= L_\tau/k_\tau \\ &= \Delta\tau + \varepsilon_\tau/k_\tau\end{aligned} \tag{12.9}$$

Phase lock loops often use the well known Costas discriminator.

$$\begin{aligned}L_\theta &= \text{Re}\left\{\tilde{Z}_P\right\}\text{Im}\left\{\tilde{Z}_P\right\} \\ &\approx k_\theta \Delta\theta + \varepsilon_\theta \\ \ell_\theta &= L_\theta/k_\theta \\ &= \Delta\theta + \varepsilon_\theta/k_\theta\end{aligned} \tag{12.10}$$

In this equation, ε_θ is due to noise and will be further detailed in Section 12.3.

Finally, frequency lock loops often use the following discriminator

$$\begin{aligned}L_f &= \text{Re}\left\{Z_{P,n-1}\right\}\text{Im}\left\{Z_{P,n}\right\} - \text{Re}\left\{Z_{P,n}\right\}\text{Im}\left\{Z_{P,n-1}\right\} \\ &\approx k_f \Delta f_D + \varepsilon_f \\ \ell_f &= L_f/k_f \\ &= \Delta f_D + \varepsilon_f/k_f\end{aligned} \tag{12.11}$$

These three discriminators are nonlinear functions. Even so, all can be modeled as a gain times $\Delta\tau$, $\Delta\theta$, or Δf_D, where k_τ, k_θ, and k_f are called *detector gains*. During normal operation, the errors are small and this linear model is applicable. This linearization enables us to create a linear model for the entire tracking loop, which we describe in the next section. However, we must be careful. The gain is usually a function of signal strength. As the signal strength changes, the model parameters will change. We will ignore this subtlety. The interested reader is referred to Van Dierendonck (1996) or Van Dierendonck and Socci (1982) for more on this effect.

12.1.3 Linear Models

The linear model shown in Figure 12.4 is applicable to the DLL, PLL and FLL. As shown, this figure uses Laplace transforms for the functions of interest.

$$\begin{aligned}T(s) &= \mathcal{L}\{\tau(t)\} \\ \Theta(s) &= \mathcal{L}\{\theta(t)\} \\ \varepsilon_T(s) &= \mathcal{L}\{\varepsilon_\tau(t)\} \\ \varepsilon_\Theta(s) &= \mathcal{L}\{\varepsilon_\theta(t)\}\end{aligned} \tag{12.12}$$

In the above, $T(s)$ is the Laplace transform of the code delay, $\tau(t)$. $\Theta(s)$ is the transform of the carrier phase. $\varepsilon_T(s)$ and $\varepsilon_\Theta(s)$ are the measurement noise terms associated with the code delay and carrier phase measurements, respectively.

The time delay of the replica code is $\hat{T}(s)$ in Figure 12.4. To accomplish alignment, the DLL measures the delay of the replica relative to the received code. This difference is the error signal, $\Delta(s)$, and is similar to the difference between the measured speed of the car and the desired speed of the car. Based on this error, the controller issues commands, $U(s)$, to the code

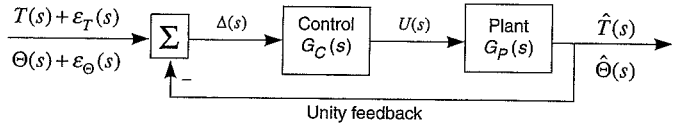


Figure 12.4 Using feedback to track a noisy measurement of code delay or carrier phase.

generator that cause the replica code to shift earlier and later in time. For the PLL, the phase of the replica would be $\hat{\Theta}(s)$ in Figure 12.4. The PLL detects the replica phase relative to the received phase, $\Delta(s)$. It issues commands, $U(s)$, to the NCO to minimize this difference.

The model input is assumed to be the parameter to be estimated plus noise. For our two most pressing examples, these inputs are given by

$$\begin{matrix} T(s)+\varepsilon_T(s)/k_\tau \\ \Theta(s)+\varepsilon_\Theta(s)/k_\theta \end{matrix} \tag{12.13}$$

where these functions were introduced in (12.12).

We seek the transfer function that gives the estimate, $\hat{T}(s)$ or $\hat{\Theta}(s)$, as a function of the input, $T(s)$ or $\Theta(s)$. For simplicity, we will use $T(s)$ and $\hat{T}(s)$ in what follows, but please remember that the analysis could apply equally well to carrier phase, $\Theta(s)=\mathcal{L}\{\theta(t)\}$, or Doppler frequency, $F_D(s)=\mathcal{L}\{f_D(t)\}$.

In the absence of noise, $\varepsilon_T(s)=0$, the output is given by

$$\hat{T}(s)=G_C(s)G_P(s)\Delta(s)$$

$G_P(s)$ represents the plant (NCO or code generator), and $G_C(s)$ is the controller. The error signal is

$$\Delta(s)=T(s)-\hat{T}(s)$$

Consequently, we can write

$$\begin{aligned} \hat{T}(s) &= G_C(s)G_P(s)\left(T(s)-\hat{T}(s)\right) \\ \frac{\hat{T}(s)}{T(s)} &= H(s) \\ &= \frac{G_C(s)G_P(s)}{1+G_C(s)G_P(s)} \end{aligned} \tag{12.14}$$

$H(s)$ is the closed loop transfer function of our DLL. Given $T(s)$ and $H(s)$, we can find $\hat{T}(s)$.

We would also like a transfer function that gives the error, $\Delta(s)$. This transfer function is derived as follows.

$$\begin{aligned} \Delta(s) &= T(s)-\hat{T}(s) \\ &= T(s)\left(1-\frac{\hat{T}(s)}{T(s)}\right) \\ \frac{\Delta(s)}{T(s)} &= 1-H(s) \\ &= \frac{1}{1+G_C(s)G_P(s)} \end{aligned} \tag{12.15}$$

With this new function, we can compute how the error depends on the code delay to be tracked. Both $H(s)$ and $1-H(s)$ will be important in what follows.

From time to time we will prefer the Fourier Transform over the Laplace transform. Our trackers are stable, so we simply replace s with $j2\pi f$ to obtain

$$\begin{aligned} H(s=j2\pi f) &= \frac{G_C(j2\pi f)G_P(j2\pi f)}{1+G_C(j2\pi f)G_P(j2\pi f)} \\ 1-H(s=j2\pi f) &= \frac{1}{1+G_C(j2\pi f)G_P(j2\pi f)} \end{aligned} \tag{12.16}$$

In the sections that follow, we will detail these linear models for the DLL and PLL. The resulting models will be used to evaluate dynamic performance and noise performance of the loop. Our analysis will divide and conquer. Good dynamic performance requires that the DLL accurately tracks changes in the code delay due to clock dynamics and antenna motion. It will be evaluated in the absence of noise. Good noise performance requires that the DLL reject measurement errors. Noise performance will be measured as the response when the input is noise only; no system dynamics will be included.

12.2 Delay Lock Loops

The delay lock loop was first introduced in Chapter 10. At that time, we were studying the GPS signal and our goal was to show that ranging precision improved with signal bandwidth, signal-to-noise ratio, and averaging time. In other words, we were studying the GPS signal rather than the DLL. For that reason, we studied the baseband DLL and ignored the impact of the underlying carrier signal on the receiver.

In this section, we dig deeper into the DLL by considering the effect of the sinusoidal carrier. This embellishment is important and will require us to study the coherent DLL and the noncoherent DLL. Both of these share some basic features with the baseband DLL. All three attempt to maintain a correlator on the leading edge of the correlation peak and attempt to maintain a second correlator on the falling edge. As noted in Section 12.1.1, these correlators are called the early and late correlators.

12.2.1 Coherent Delay Lock Loop

The coherent DLL is most similar to the baseband DLL, and can be used when the carrier phase is known, $\Delta\theta=0$, and the data bit, D, is known.

The coherent DLL acts to null the following discriminator function

$$L_\tau = \left(\mathrm{Re}\{\tilde{Z}_E\} - \mathrm{Re}\{\tilde{Z}_L\}\right) D$$
$$= \left(\mathrm{Re}\{\tilde{S}_E\} - \mathrm{Re}\{\tilde{S}_L\}\right) D + \varepsilon_\tau \qquad (12.17)$$

The noise contributions are combined in ε_τ as before in (12.8). If the early and late correlators are spaced by one chip width, T_C, and the carrier phase is known ($\Delta\theta \approx 0$), then this discriminator may be linearized as follows.

$$L_\tau(\Delta\tau) = \sqrt{C}R(\Delta\tau - T_C/2) - \sqrt{C}R(\Delta\tau + T_C/2) + \varepsilon_\tau$$
$$\approx L_\tau(\Delta\tau = 0) + \left.\frac{\delta L_\tau}{\delta\Delta\tau}\right|_{\Delta\tau=0} \Delta\tau + \varepsilon_\tau$$
$$= \frac{2\sqrt{C}}{T_C}\Delta\tau + \varepsilon_\tau$$
$$\ell_\tau = \Delta\tau + \varepsilon_\tau \frac{T_C}{2\sqrt{C}} \qquad (12.18)$$

The discriminator is linear provided that $\Delta\tau$ is a small fraction of a chip. As was to be expected, the coherent DLL is sensitive to the quality of the carrier phase estimate. The carrier tracking loop can provide such an estimate, but it is the most fragile estimate in a GPS receiver. If this estimate is lost or erroneous, then code tracking performance suffers. In addition, the code delay estimate is needed before carrier tracking can begin. For these reasons, we now turn our attention to a noncoherent DLL.

12.2.2 Early Power Minus Late Power Delay Lock Loop

To avoid dependence on carrier phase tracking, many receivers use a noncoherent DLL, and several appropriate discriminator functions have been devised. Amongst these, the early power minus later power discriminator is given by

$$L_\tau = \left|\tilde{Z}_E\right|^2 - \left|\tilde{Z}_L\right|^2$$
$$= \left|\tilde{S}_E + \tilde{N}_E\right|^2 - \left|\tilde{S}_L + \tilde{N}_L\right|^2$$
$$= \left(\left|\tilde{S}_E\right|^2 + \left|\tilde{N}_E\right|^2 + 2S_{I,E}N_{I,E} + 2S_{Q,E}N_{Q,E}\right)$$
$$-\left(\left|\tilde{S}_L\right|^2 + \left|\tilde{N}_L\right|^2 + 2S_{I,L}N_{I,L} + 2S_{Q,L}N_{Q,L}\right)$$
$$= \left|\tilde{S}_E\right|^2 - \left|\tilde{S}_L\right|^2 + \varepsilon_\tau \qquad (12.19)$$

where

$$S_{I,E} = \mathrm{Re}\{\tilde{S}_E\}$$
$$N_{I,E} = \mathrm{Re}\{\tilde{N}_E\}$$
$$S_{Q,E} = \mathrm{Im}\{\tilde{S}_E\}$$
$$N_{Q,E} = \mathrm{Im}\{\tilde{N}_E\}$$

Once again, the noise contributions are collected in a single term, ε_τ. Even though we use the same symbol, the noise term in (12.19) is not equal to the noise term in (12.18).

If we limit our attention to small values of $\Delta\tau$, this discriminator can be linearized as follows.

$$L_\tau = CR^2(\Delta\tau - T_C/2) - CR^2(\Delta\tau + T_C/2) + \varepsilon_\tau$$
$$\approx L_\tau(\Delta\tau = 0) + \left.\frac{\delta L_\tau}{\delta\Delta\tau}\right|_{\Delta\tau=0} \Delta\tau + \varepsilon_\tau$$
$$= \frac{2C}{T_C}\Delta\tau + \varepsilon_\tau$$
$$\ell_\tau = \Delta\tau + \varepsilon_\tau \frac{T_C}{2C} \qquad (12.20)$$

This approximation is valid for a small but useful range of $\Delta\tau$. The normalized discriminator, ℓ_τ, is equal to the delay error, $\Delta\tau$, plus normalized noise, $\varepsilon_\tau T_C/2C$. Notice that our detector gain for the noncoherent discriminator, $k_\tau = 2C/T_C$, differs from the gain for the coherent discriminator, $k_\tau = 2\sqrt{C}/T_C$.

This DLL uses a pair of complex correlators and a discriminator to measure the difference between the received code and the estimated code. This early power minus late power discriminator eliminates the worrisome nuisance parameters, $\{D, \Delta\theta\}$, by taking the magnitude of the complex correlator outputs. To do so, it uses twice as many correlators as the coherent DLL, and other noncoherent discriminators, such as the dot-product discriminator described by Van Dierendonck (1996), are more efficient. In any event, the early power minus late power discriminator is linear in the operating range of interest to GPS receivers. In the next section, we create a linear model for the entire DLL by combining this linearized discriminator with models for the loop filter, code clock, and code generator.

12.2.3 Linear Models for the Delay Lock Loop

Two linear models for an unaided DLL are shown in Figure 12.5. The top model is an analog prototype for a first-order loop. It takes only one input—noisy measurements of the code tracking error from the discriminator. The code clock and generator are modeled together as an integrator because they take speed-up or slow-down commands from the controller. The output code delay is the integral of these frequency commands.

The bottom diagram in Figure 12.5 is an approximation to our analog prototype that replaces the integrator with a digital implementation. This approximation simply adds a scaled version of the current input to the value stored in the accumulator. Today, loop filters are implemented digitally, so we provide the digital DLL for reference. For the time being, our analysis will use the analog prototype.

The two diagrams in Figure 12.6 are rate-aided DLLs. These structures take an additional input—very accurate estimates of the pseudorange rate from the PLL or FLL. This input is derived from Doppler measurements and makes a big difference. As we shall discover, the rate-aided DLL improves the trade between dynamic performance and noise performance.

Our unaided, first-order DLL issues a command that is proportional to the relative delay between the observed code and the replica code. For this reason, it is called a proportional controller. The closed-loop transfer function is found from (12.14)

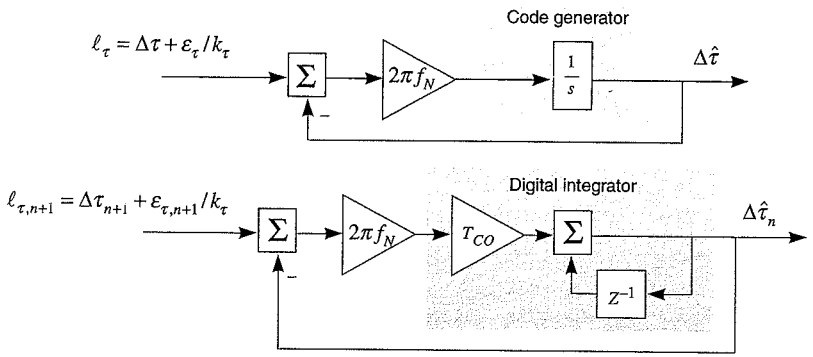


Figure 12.5 Block diagrams of a linearized delay lock loop, including code generator and loop filter. The top diagram shows an analog prototype for a first-order loop and includes a continuous time model for the integrator (1/s). The bottom diagram replaces this analog integrator with a digital approximation. Neither loop enjoys the benefits of rate-aiding.

$$\frac{\hat{T}(s)}{T(s)} = H(s) = \frac{2\pi f_N \frac{1}{s}}{1 + 2\pi f_N \frac{1}{s}} = \frac{2\pi f_N}{s + 2\pi f_N} \tag{12.21}$$

The frequency response is

$$H(s = j2\pi f) = \frac{f_N}{jf + f_N}$$
$$\left|H(j2\pi f)\right|^2 = \frac{f_N^2}{f^2 + f_N^2} \tag{12.22}$$

The frequency response can be used to derive another characterization of the DLL—the noise equivalent bandwidth. As described in Chapter 8, the noise equivalent bandwidth is the bandwidth of the boxcar filter that passes the same amount of noise as the filter under study. In general, the noise equivalent bandwidth is given by

$$B_{N,1} = \frac{1}{\left|H(j0)\right|^2} \int_0^\infty \left|H(j2\pi f)\right|^2 df \tag{12.23}$$

For our first-order filter, the noise equivalent bandwidth becomes

$$B_{\tau,1} = \int_0^\infty \frac{f_N^2}{f^2 + f_N^2} df$$
$$= \frac{\pi f_N}{2} \tag{12.24}$$

This single parameter captures the entire design freedom available from a first-order filter. As described in the next few sections, it enables a limited trade between dynamic performance and noise performance.

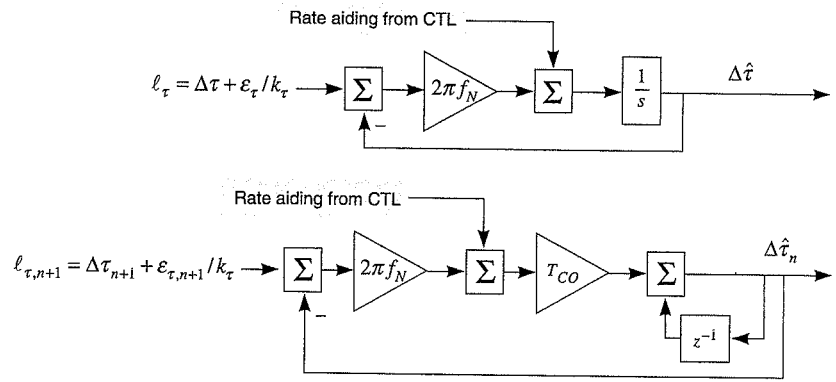


Figure 12.6 Block diagrams of a linearized delay lock loop, including rate aiding. Once again, the top diagram shows an analog model for the integrator (1/s), and the bottom diagram replaces this analog integrator with a digital approximation.

12.2.4 Step Response of the Unaided Delay Lock Loop

In this section, we turn off the input noise and determine the response of our DLL to a step change in pseudorange. We can find the step response of our loop from (8.25) and (12.21).

$$\begin{aligned} y(t) &= \mathcal{L}^{-1}\left\{H(s)U(s)\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{2\pi f_N}{s + 2\pi f_N}\frac{1}{s}\right\} \\ &= 1 - \exp(-2\pi f_N t) \\ &= 1 - \exp(-4B_{\tau,1} t) \end{aligned} \tag{12.25}$$

In Figure 12.7, we have plotted $y(t)$ for unaided DLLs with noise equivalent bandwidths of 5, 15, and 25 hertz. Large values of $B_{N,1}$ cause the DLL to be very nimble—it closely follows the input step. Small values of $B_{N,1}$ cause the DLL to be sluggish. As we shall discover, smaller bandwidths also provide much greater attenuation of the noise input. So, noise performance and dynamic performance are traded against each other.

Fortunately, the DLL can use information from the carrier tracking loop (PLL or FLL) to improve this trade-off. Specifically, it takes estimates of the pseudorange rate from the carrier loop. These rate estimates are very accurate because they are directly based on the measured Doppler shift of the GPS signals.

12.2.5 Rate-Aided Delay Lock Loop

The rate-aided DLL is shown in Figure 12.6. The top half of the figure shows an analog prototype and the bottom half shows the corresponding digital implementation. The digital block diagram provides the following nifty result.

$$\Delta\hat{\tau}_{n+1} = \Delta\hat{\tau}_n + T_{CO}\left(\Delta\hat{\tau}_n + 2\pi f_N(\Delta\tau_{n+1} + \varepsilon_{\tau,n+1}/k_\tau - \Delta\hat{\tau}_n)\right) \tag{12.26}$$

As shown, the new estimate of code phase error is essentially equal to the old estimate propagated forward in time using the rate estimate from the carrier loop, $\Delta\hat{\tau}_n$, summed with the current DLL measurement minus the old estimate, $\Delta\tau_{n+1} + \varepsilon_{\tau,n+1}/k_\tau - \Delta\hat{\tau}_n$. Typically, the rate estimates dominate this mix because they are so accurate. The DLL measurement is only used to remove any long term biases in the rate measurements from the carrier loop.

The noise response of the DLL is calculated in the next section. As we shall discover, noise decreases as bandwidth decreases. Rate-aiding allows the noise bandwidth to be significantly smaller. A smaller bandwidth is tolerable because any system dynamics are captured by the rate estimates from the PLL or FLL. The bandwidth of rate-aided DLLs can be as small as 0.005 Hz. Smaller bandwidths are not generally used because the ionosphere eventually causes the received code delay to diverge from the carrier phase. This divergence is described in Section 5.3.1.

12.2.6 Performance in Additive White Noise

In this section, we turn off any dynamics and determine the response of our DLL to noise only. To this end, we follow the elegant derivation in Van Dierendonck (1996). If we set dynamics to zero, then $\Delta\tau_n = 0$ and $\Delta\dot{\tau}_n = 0$. In addition, we set $\Delta\hat{\dot{\tau}}_n = 0$ because the carrier loop estimates rate so accurately. With these assumptions, we may simplify (12.26) as follows.

$$\begin{aligned}\Delta\hat{\tau}_{n+1} &= \Delta\hat{\tau}_n + 2\pi f_N T_{CO}(\varepsilon_{\tau,n+1}/k_\tau - \Delta\hat{\tau}_n) \\ &= \Delta\hat{\tau}_n(1 - 2\pi f_N T_{CO}) + \varepsilon_{\tau,n+1} 2\pi f_N T_{CO}/k_\tau\end{aligned} \tag{12.27}$$

As before, $\Delta\hat{\tau}_n$ is our best estimate of the code tracking error at time step n, and $\Delta\hat{\tau}_{n+1}$ is the best estimate at time step $n + 1$. It is our best previous estimate updated by the most recent measurement. Recall that the detector gain is $k_\tau = 2C/T_C$ for our early power minus later power discriminator.

The variance of the tracker's estimate is

$$\begin{aligned}\mathrm{var}(\Delta\hat{\tau}_{n+1}) &= E\left\{(\Delta\hat{\tau}_{n+1} - \Delta\tau_{n+1})^2\right\} \\ &= E\{\Delta\hat{\tau}_{n+1}^2 - 2\Delta\hat{\tau}_{n+1}\Delta\tau_{n+1} + \Delta\tau_{n+1}^2\} \\ &= E\{\Delta\hat{\tau}_{n+1}^2\} \\ &= E\{\Delta\hat{\tau}_n^2\}(1 - 2\pi f_N T_{CO})^2 + E\{\varepsilon_{\tau,n+1}^2\}(2\pi f_N T_{CO}/k_\tau)^2\end{aligned} \tag{12.28}$$

The derivation of (12.28) relies on the following observations. First, our measurement noise has zero mean and is uncorrelated from sample to sample.

$$\begin{aligned}E\{\varepsilon_{\tau,n}\} &= 0 \quad \text{for all } n \\ E\{\varepsilon_{\tau,m}\varepsilon_{\tau,n}\} &= 0 \quad \text{for all } m \neq n\end{aligned} \tag{12.29}$$

Hence, $E\{\Delta\hat{\tau}_n\} = 0$ because the measurement noise has zero mean and this noise analysis excludes any change in code error due to system dynamics, ($\Delta\tau_n = 0$). Second, our current sample of noise, $\varepsilon_{\tau,n+1}$, is uncorrelated with our previous estimate of the code's arrival time, $\Delta\hat{\tau}_n$. After all, the previous estimate was formed before the current noise sample arrived.

In time, the noise variance will reach a steady state value because our DLL is stable. Consequently,

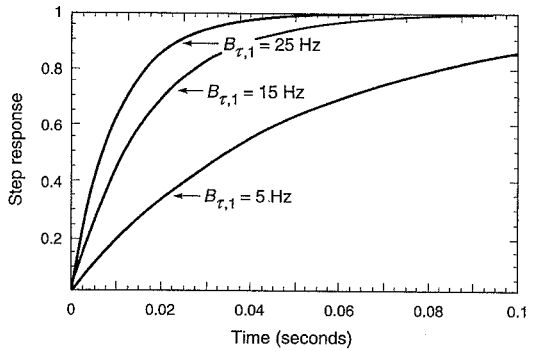


Figure 12.7 Step responses of unaided first-order delay lock loops.

$$E\left\{\Delta\hat{\tau}_{n+1}^2\right\} = E\left\{\Delta\hat{\tau}_n^2\right\} = \mathrm{var}(\Delta\hat{\tau}) \tag{12.30}$$

With this result, we may proceed as follows.

$$\begin{aligned}\mathrm{var}(\Delta\hat{\tau}) &= \mathrm{var}(\Delta\hat{\tau})(1 - 2\pi f_N T_{CO})^2 + \mathrm{var}(\varepsilon_\tau)(2\pi f_N T_{CO}/k_\tau)^2 \\ &= \frac{\mathrm{var}(\varepsilon_\tau)(2\pi f_N T_{CO}/k_\tau)^2}{4\pi f_N T_{CO} - (2\pi f_N T_{CO})^2} \\ &\approx \frac{\mathrm{var}(\varepsilon_\tau) 2\pi f_N T_{CO}}{2k_\tau^2} \\ &= \frac{2\mathrm{var}(\varepsilon_\tau) B_{\tau,1} T_{CO}}{k_\tau^2}\end{aligned} \tag{12.31}$$

The approximation is valid provided $2B_{\tau,1}T_{CO} \ll 1$, which normally holds.

Finally, we need to know $\mathrm{var}(\varepsilon_\tau)$. Fortunately, we can find this result using the ideas from Chapter 10 for the analysis of baseband DLLs. One of the homework problems asks you to confirm the following result for the early power minus late power discriminator.

$$\mathrm{var}(\varepsilon_\tau) = \left(\frac{2N_0^2}{T_{CO}^2} + \frac{CN_0}{T_{CO}}\right)\mathrm{sec}^2 \tag{12.32}$$

This equation gives the noise variance when the early power minus late power discriminator is used with a coherent averaging time of T_{CO} seconds, and the power spectral density of our noise is $N_0/2$.

Finally, we can combine (12.31) and (12.32).

$$\mathrm{var}\left\{\Delta\hat{\tau}\right\} = \frac{B_{\tau,1}T_C^2}{2C/N_0}\left(1 + \frac{2}{T_{CO}C/N_0}\right)\mathrm{sec}^2 \tag{12.33}$$

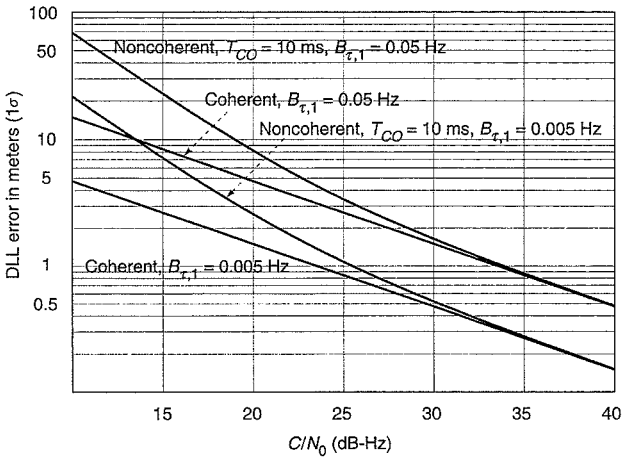


Figure 12.8 Performance of delay lock loop in the presence of white noise, including the effect of squaring loss.

Squaring loss is the term inside the parenthesis, and it multiplies our earlier result from Section 10.5 for the baseband delay lock loop, which is also the error variance for the coherent DLL. All noncoherent discriminators suffer from squaring loss. Noncoherent discriminators are able to remove the nuisance parameters, carrier phase and data bits, but they do so at a cost. The nonlinear operation also amplifies the impact of the received noise.

Using bandwidths of 0.005 Hz and 0.05 Hz in (12.33), we can plot code tracking error in Figure 12.8, which shows the $1-\sigma$ error as a function of signal-to-noise ratio (C/N_0). The coherent curves ignore squaring loss and the noncoherent curves include squaring loss. As shown, squaring loss is not significant until the signal-to-noise ratio falls below 25 dB-Hz. Recall from Table 10.4, most GPS operations are conducted with signal-to-noise ratios above 30 dB-Hz and so squaring loss in the noncoherent DLL is not usually a major consideration.

If we consider correlator spacings, d, other than one chip, then the following result can be derived.

$$\mathrm{var}\left\{\Delta\hat{\tau}\right\} = \frac{B_{\tau,1} d T_C^2}{2C/N_0}\left(1+\frac{2}{T_{CO}C/N_0}\right)\sec^2 \tag{12.34}$$

This result is akin to the narrow correlator result for the baseband DLL developed in Section 10.6. The discussion on correlator spacings in Section 10.6 applies here as well.

12.3 Phase Lock Loops

Like the DLL, the phase lock loop (PLL) is well modeled as a control system that uses feedback to control the behavior of some plant, but now the numerically controlled oscillator (NCO) is the plant to be controlled. As shown in Figure 12.2, our NCO outputs the following signal

$$2\exp\left(-j\left(2\pi(f_{IF}+\hat{f}_D)t+\hat{\theta}\right)\right) \tag{12.35}$$

The PLL strives to control the NCO so that the Doppler estimate, $\hat{f}_D$, for the signal is accurate. It also has the goal of steering the phase of the replica signal, $\hat{\theta}$, to the phase of the received signal, θ. To accomplish these goals, the PLL measures the difference between the received frequency and phase and the replica frequency and phase. The frequency difference is similar to the difference between the measured speed of the car and the desired speed of the car. The phase difference is akin to the difference between the measured location of the car and the desired location of the car. Based on these errors, the controller commands the NCO to increase or decrease its frequency. In either case, the PLL does not command abrupt changes in the replica phase. Rather, it simply controls the frequency. In addition to tracking frequency and phase, the PLL also delivers estimates of the navigation data bits.

The frequency lock loop (FLL) is a close relative of the PLL. It too tracks the carrier. However, it only tracks the received frequency and makes no attempt to track phase. The FLL is often used for coarse frequency tracking as the receiver transitions from signal acquisition to carrier phase tracking. GPS receivers also revert to frequency-only tracking if they lose the ability to track carrier phase. Indeed, the FLL can track in more challenging signal environments. For these reasons, the FLL is described in Appendix 12.A.

12.3.1 Coherent Phase Lock Loop

If the navigation data bits are known, then the PLL could use the following discriminator

$$\begin{aligned} L_\theta &= \mathrm{Im}\left\{\tilde{Z}_P\right\} D \\ &= \sqrt{C} R(\Delta\tau)\sin(\Delta\theta)+\varepsilon_\theta \\ L_\theta(\Delta\tau\approx 0) &\approx \sqrt{C}\Delta\theta+\varepsilon_\theta \\ \ell_\theta(\Delta\tau\approx 0) &\approx \Delta\theta+\varepsilon_\theta/k_\theta \\ k_\theta &= \sqrt{C} \end{aligned} \tag{12.36}$$

This equation assumes the prompt correlator is accurate, $\Delta\tau\approx 0$. It also uses the small angle approximation $\sin(\Delta\theta)\approx\Delta\theta$, which is valid when the angle error is small.

12.3.2 Costas Phase Lock Loop

The navigation data bits are not usually known *a priori*. Consequently, we explore the well-known Costas discriminator.

$$\begin{aligned} L_\theta &= \mathrm{Re}\left\{Z_P\right\}\mathrm{Im}\left\{Z_P\right\} \\ &= \mathrm{Re}\left\{S_P\right\}\mathrm{Im}\left\{S_P\right\}+\varepsilon_\theta \\ &= CD^2R^2(\Delta\tau)\sin(\Delta\theta)\cos(\Delta\theta)+\varepsilon_\theta \\ &= CR^2(\Delta\tau)\frac{\sin(2\Delta\theta)}{2}+\varepsilon_\theta \end{aligned}$$

$$\begin{aligned} &\approx L_\theta(\Delta\theta=0)+\left.\frac{\delta L_\theta}{\delta\Delta\theta}\right|_{\Delta\theta=0}\Delta\theta+\varepsilon_\theta \\ &= C\Delta\theta+\varepsilon_\theta \\ \ell_\theta &= \frac{L_\theta}{C}=\Delta\theta+\frac{\varepsilon_\theta}{C} \\ k_\theta &= C \end{aligned} \tag{12.37}$$

This development collects the noise components in the single term ε_θ. Even though we use the same symbol, the noise term in (12.37) is not equal to the noise term in (12.36). The Costas discriminator eliminates the need to know D *a priori*. However, it incurs squaring loss—just like the early power minus late power DLL. We will analyze the white noise performance of the Costas loop in Section 12.3.6.

12.3.3 Navigation Data Recovery

Once the PLL is locked, the navigation data bits can be readily recovered. The signal contribution to the inphase and quadrature samples from the prompt correlator are as follows.

$$\begin{aligned} S_{I,P} &= \sqrt{C}DR(\Delta\tau)\cos(\Delta\theta) \\ S_{Q,P} &= \sqrt{C}DR(\Delta\tau)\sin(\Delta\theta) \end{aligned} \tag{12.38}$$

If the carrier phase and code delay are reasonably well known, then $\Delta\tau \approx 0$ and $\Delta\theta \approx 0$, and the inphase correlation will show the data bit, D.

12.3.4 Step Response of the Second-Order Loop

The PLL faces exactly the same challenge as the DLL. It too must strike a balance between noise performance and dynamic performance. Unlike the DLL, it does not usually have access to error rate measurements. After all, the PLL provides those estimates to the DLL! Hence, the PLL most often develops rate estimates autonomously. If the loop estimates both phase and phase rate, then it is called a second-order loop. As shown in Figures 12.9 and 12.10, our second-order loop has two design parameters: the undamped natural frequency, f_N, and the damping ratio, ζ.

The transfer function for our second-order PLL is

$$\frac{\hat{\Theta}(s)}{L_\theta(s)} = H_\theta(s) = \frac{4\pi\zeta f_N s+(2\pi f_N)^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2} \tag{12.39}$$

The frequency response is

$$\begin{aligned} H_\theta(j2\pi f) &= \frac{j2(2\pi)^2\zeta f_N f+(2\pi f_N)^2}{-(2\pi f)^2+j2(2\pi)^2\zeta f_N f+(2\pi f_N)^2} \\ &= \frac{j2\zeta f_N f+f_N^2}{-f^2+j2\zeta f_N f+f_N^2} \end{aligned} \tag{12.40}$$

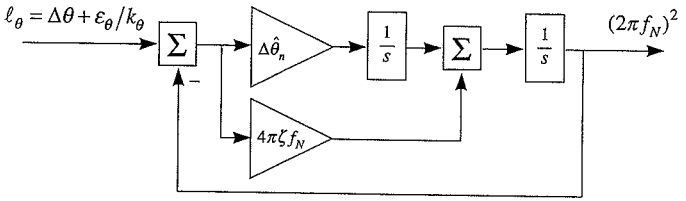


Figure 12.9 An analog model for a second-order phase lock loop.

Application of (12.23) reveals that the noise equivalent bandwidth is

$$B_{\theta,1} = \frac{\pi f_N}{4\zeta}(4\zeta^2+1) \tag{12.41}$$

The step response for our second-order PLL can be found from the inverse Laplace transform as follows.

$$y(t) = \mathcal{L}^{-1}\left\{H_\theta(s)U(s)\right\} = \mathcal{L}^{-1}\left\{\frac{4\pi\zeta f_N s+(2\pi f_N)^2}{s^2+4\pi\zeta f_N s+(2\pi f_N^2)}\frac{1}{s}\right\} \tag{12.42}$$

We proceed as follows.

$$\begin{aligned} y(t) &= \mathcal{L}^{-1}\left\{\frac{4\pi\zeta f_N s}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\frac{1}{s}+\frac{(2\pi f_N)^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\frac{1}{s}\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{4\pi\zeta f_N}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}+\frac{(2\pi f_N)^2}{s(s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\right\} \end{aligned} \tag{12.43}$$

The inverse Laplace transform for each of these terms can be found in Table 8.4.

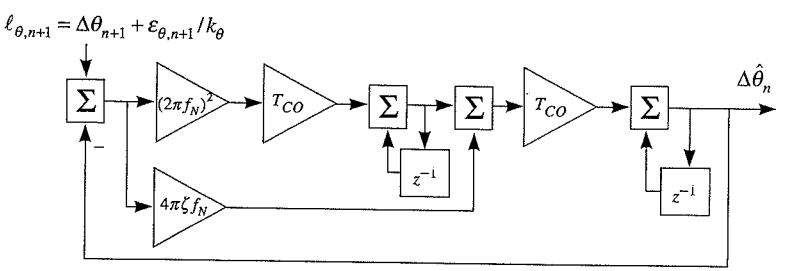


Figure 12.10 A digital model for a second-order phase lock loop.

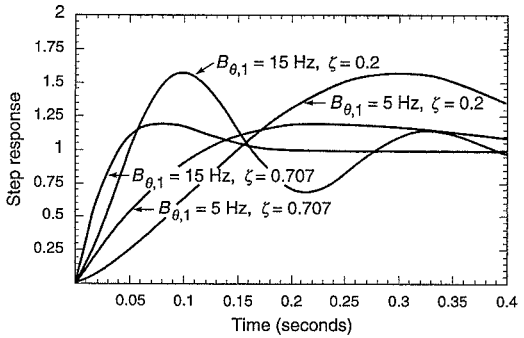


Figure 12.11 Second-order step-response noise bandwidths of 5 and 15 Hz and damping coefficients of 0.2 and 0.707.

$$y(t)=\left(1-\frac{\exp-2\pi\zeta f_N t}{\sqrt{1-\zeta^2}}\left(\sin\left(2\pi f_N\sqrt{1-\zeta^2}\,t+\phi\right)-2\zeta\sin\left(2\pi f_N\sqrt{1-\zeta^2}\,t\right)\right)\right)u(t)$$

$$\phi=\tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \tag{12.44}$$

Some of these step responses are shown in Figure 12.11. The larger noise bandwidths provide a more nimble loop with rapid rise times. The damping ratio appears well named because small values of ζ lead to step responses that oscillate before they settle down to the desired value. GPS receivers typically use $\zeta \approx 1/\sqrt{2}$ because the step response rises quickly with little overshoot. When $\zeta \approx 1/\sqrt{2}$, the noise equivalent bandwidth becomes $B_{\theta,1} = 1.06\pi f_N$. We will use this popular value in what follows.

12.3.5 Steady State Error for the Second-Order Loop

Now we investigate the steady state response to step, ramp and parabolic inputs. The tracking error is the difference between the input signal and the estimate

$$\begin{aligned}\Delta(s)&=\Theta(s)-\hat{\Theta}(s)\\&=\Theta(s)\left(1-\frac{\hat{\Theta}(s)}{\Theta(s)}\right)\\&=\Theta(s)\left(1-H(s)\right)\end{aligned} \tag{12.45}$$

At this point, we apply a step input and compute the steady state error using the final value theorem described in Section 8.7.7.

$$\begin{aligned}\lim_{t\to\infty}\Delta(t)&=\lim_{s\to 0}sU(s)(1-H(s))\\&=\lim_{s\to 0}\left(1-\frac{4\pi\zeta f_N s+(2\pi f_N)^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\right)\\&=\lim_{s\to 0}\left(\frac{s^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\right)=0\end{aligned} \tag{12.46}$$

In the fullness of time, our second-order PLL will converge to the correct answer.

If we apply a ramp input, then

$$\lim_{t\to\infty}\Delta(t)=\lim_{s\to 0}\frac{s}{s^2}\left(\frac{s^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\right)=0 \tag{12.47}$$

Once again, the second-order PLL is able to cope and no steady state error results. However, it is not able to keep up with a parabolic input.

$$\lim_{t\to\infty}\Delta(t)=\lim_{s\to 0}\frac{2s}{s^3}\left(\frac{s^2}{s^2+4\pi\zeta f_N s+(2\pi f_N)^2}\right)=\frac{2}{(2\pi f_N)^2} \tag{12.48}$$

This latter result is important and appears in a number of guises in the literature. Accordingly, we summarize the foregoing as follows. A rather general input may be written as follows.

$$\Delta\theta(t)=\left(k_P+k_V t+\frac{k_A}{2}t^2\right)u(t)\quad\text{radians} \tag{12.49}$$

In this case, the steady state response from a second-order loop will contain an error equal to

$$\begin{aligned}\lim_{t\to\infty}\Delta(t)&=\frac{k_A}{(2\pi f_N)^2}\quad\text{radians}\\&\approx\frac{k_A}{4B_{\theta,1}^2}\quad\text{radians}\\&=\frac{k_A}{4B_{\theta,1}^2}\frac{360}{2\pi}\quad\text{degrees}\end{aligned} \tag{12.50}$$

As an example, consider an acceleration of $k_G g$ gravities, where g = 9.8 m/s^2. The corresponding steady state error in a second-order loop is

$$\lim_{t\to\infty}\Delta(t)=\frac{k_G g\cdot 360}{4\lambda B_{\theta,1}^2}\quad\text{degrees} \tag{12.51}$$

Phase lock loops are likely to lose lock when the 1 – σ phase error is greater than 15°. Hence, the bandwidth required to track a 1g acceleration is approximately 17.6 Hz. If the maximum acceleration is only 0.25g, then the requirement falls to around 8.8 Hz.

Please bear in mind that these results are for a second-order PLL. High dynamic environments often motivate designers to use third-order loops that have no steady state error due to

constant acceleration. However, third-order loops do have sensitivity to jerk, and their stability is a concern. Please see Van Dierendonck (1996) and Ward (1996) for more on third-order loops.

12.3.6 Performance of the Costas Loop in Additive White Noise

With some effort, noise bandwidth can be used to characterize the white noise performance of our second-order PLL. The variance of the error in the phase estimate after smoothing by the second-order loop is

$$\text{var}(\Delta\hat{\theta}) = \frac{2\text{var}(\varepsilon_\theta)B_{\theta,1}T_{CO}}{k_\theta^2}\ \text{rad}^2 \tag{12.52}$$

By the way, this equation is not easy to derive, but the analysis in Section 12.2.6 can be used as a guide for the adventurous. We will employ (12.52) but first we need an expression for $\text{var}(\varepsilon_\theta)$.

$$\text{var}(\varepsilon_\theta) = \frac{CN_0}{2T_{CO}}\left(1+\frac{1}{2T_{CO}C/N_0}\right) \tag{12.53}$$

As usual, we have used $N_0/2$ to denote the power spectral density of the white noise. We have left the non-trivial proof of (12.53) for homework.

We find the aggregate noise performance by combining (12.52) and (12.53) as follows.

$$\text{var}\left(\Delta\hat{\theta}\right) = \frac{B_{\theta,1}}{C/N_0}\left(1+\frac{1}{2T_{CO}C/N_0}\right)\ \text{rad}^2 \tag{12.54}$$

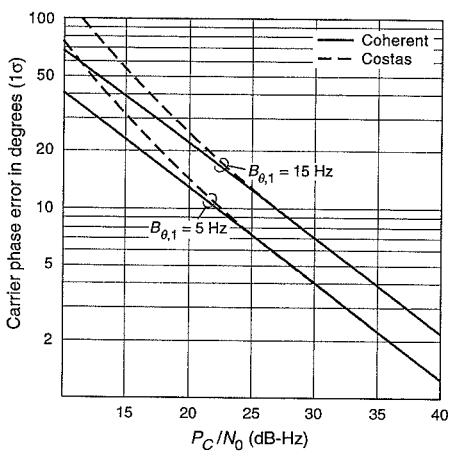


Figure 12.12 Performance of phase lock loop in the presence of white noise. The Costas curves show squaring loss relative to the coherent curves.

The term inside the parenthesis is squaring loss for the carrier tracking loop that uses the Costas discriminator. If squaring loss is omitted, then we have the phase error variance for the coherent PLL.

Figure 12.12 shows the standard deviation, 1σ, of the phase error, $\Delta\hat{\theta}$, as a function of signal-to-noise ratio, C/N_0. The coherent curves ignore squaring loss and the noncoherent curve includes squaring loss. Once again, squaring loss is not significant until the signal-to-noise ratio is below 25 dB-Hz.

12.3.7 Choosing Loop Bandwidth

The noise equivalent bandwidth of the PLL is chosen to balance dynamic performance against noise performance. Dynamic performance is partially captured by our steady state error analysis, and (12.51) is plotted in Figures 12.13 and 12.14 for accelerations of $0.25g$ and $1g$. As shown, higher accelerations require higher bandwidths. Noise performance is quantified by (12.54), and this relationship is also plotted in Figures 12.13 and 12.14. In this case, we plot the 1σ error versus bandwidth for signal-to-noise ratios of 20 dB-Hz, 30 dB-Hz and 40 dB-Hz.

As shown, the ideal bandwidth for a signal-to-noise ratio of 30 dB-Hz and a maximum acceleration of $1g$ is slightly over 20 Hz. This operating point is fragile because the 1σ phase error (also known as jitter) is around 10°, and cycle slips are increasingly likely as the jitter increases above 10°. In fact, 15° is considered to be the maximum allowable 1σ phase error before cycle slips simply rule out the use of a phase lock loop. At this point, the receiver should revert to frequency-only tracking. If the maximum acceleration drops to $0.25g$, then the optimum bandwidth drops. Please remember that these bandwidths are for the second-order loop and third-order loops offer an alternate approach to coping with high dynamics.

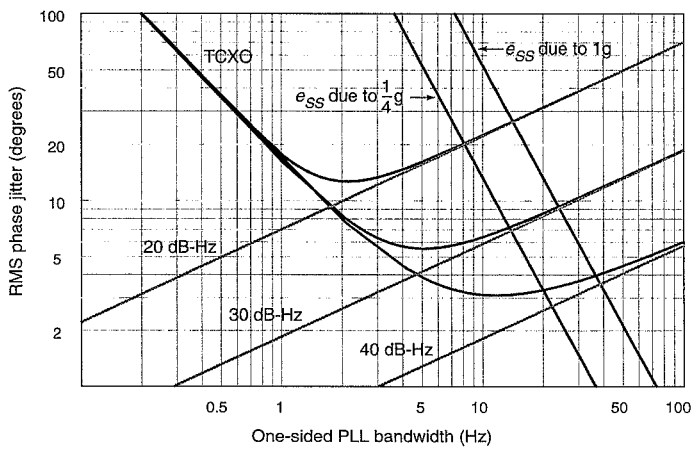


Figure 12.13 Standard deviation of the phase error as a function of loop bandwidth when a temperature-compensated crystal oscillator is used. Narrow bandwidths suffer from poor dynamic performance and high bandwidths suffer poor noise performance.

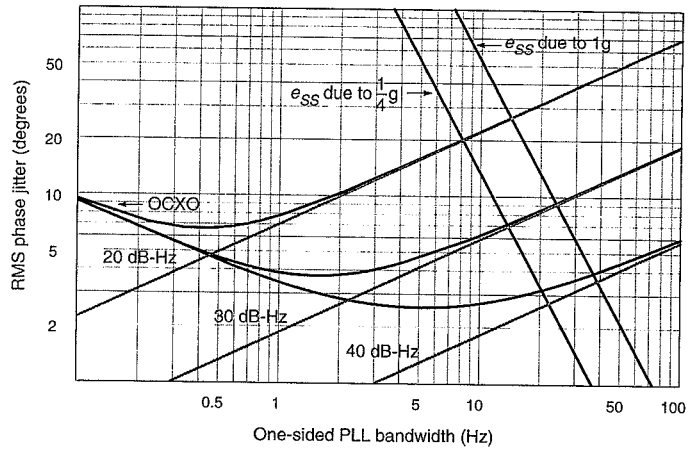


Figure 12.14 Standard deviation of the phase error as a function of loop bandwidth when an ovenized crystal oscillator is used. Narrow bandwidths suffer from poor dynamic performance and high bandwidths suffer poor noise performance.

If the user is stationary or an inertial navigation system is used to estimate user motion, then the PLL bandwidth is determined by clock dynamics. The satellite and receiver clocks inject authentic dynamics into the GPS measurements and they must be tracked.

The telling relationship between the clock dynamics and the PLL frequency response is shown in Figure 12.15. The figure shows the amplitude response for two second-order PLLs. One PLL has $B_{\theta,1}$ = 1 Hz and the other has $B_{\theta,1}$ = 10 Hz. For $\zeta = 1/\sqrt{2}$, this response is given by

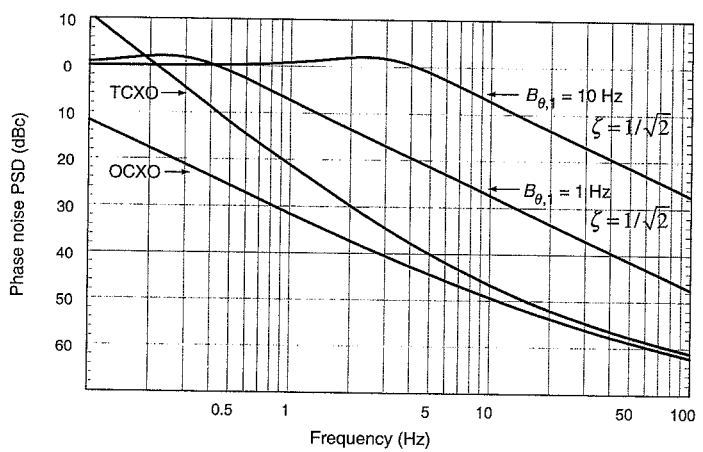


Figure 12.15 Second-order frequency response and clock noise power spectral density.

$$\left|H_\theta(f)\right|^2 = \frac{f_N^4 + 2f_N^2 f^2}{f^4 + f_N^4} \tag{12.55}$$

This result can be derived from the second-order transfer function given in (12.40).

The TCXO and OCXO curves show the power of clock noise as a function of frequency. The TCXO is a temperature compensated crystal oscillator, and the OCXO is an oven controlled crystal oscillator. In general, OCXOs are more expensive than TCXOs and achieve better performance. Their long term drift is smaller and their short term performance is better. The TCXO and OCXO curves in Figure 12.15 are examples of power spectral densities, introduced in Chapter 8. In spite of its name, clock noise is not noise at all to the receiver—it must be tracked. In contrast, thermal noise, analyzed in Section 12.3.6, is an outside disturbance and should be rejected.

The clock power spectral densities are given by

$$\begin{aligned} G_{\text{clock}}(f) &= \sum_{i=0}^{4} \frac{h_i}{f^i} \\ &= h_0 + \frac{h_1}{f} + \frac{h_2}{f^2} + \frac{h_3}{f^3} + \frac{h_4}{f^4} \end{aligned} \tag{12.56}$$

All of these terms have names. The h_0 term is called white phase noise because the power spectral density is independent of frequency. The h_1/f term is called flicker phase noise, the h_2/f^2 term yields a random phase walk, the h_3/f^3 term is flicker frequency noise, and the h_4/f^4 term is associated with a random frequency walk.

The magnitude of these noise terms depends on the quality of the clock. Santiago (2004) has measured the coefficients for a TCXO and an OCXO. His results are

$$\begin{aligned} G_{\text{TCXO}}(f) &= 5.0\times10^{-8} + \frac{6.2\times10^{-5}}{f} + \frac{9.6\times10^{-4}}{f^2} + \frac{6.0\times10^{-3}}{f^3} + \frac{6.0\times10^{-4}}{f^4} \\ G_{\text{OCXO}}(f) &= 5.5\times10^{-8} + \frac{5.0\times10^{-5}}{f} + \frac{6.5\times10^{-4}}{f^2} + \frac{9.0\times10^{-7}}{f^3} + \frac{1.0\times10^{-7}}{f^4} \end{aligned} \tag{12.57}$$

These are the curves shown in Figure 12.15.

As shown in Figure 12.15, the narrow bandwidth PLL (1 Hz) attenuates an appreciable amount of the power in both oscillators. The PLL with $B_{\theta,1}$ = 10 Hz tracks more of the clock noise and cycle slips would be rare. We can quantify this effect by calculating the clock power that is suppressed by the filters. The lost power is given by

$$\begin{aligned} \sigma^2_{\text{clock}} &= \int_0^\infty \left|1 - H_\theta(f)\right|^2 G_{\text{clock}}(f)\, df \\ &= \int_0^\infty \frac{f^4}{f^4 + f_N^4} G_{\text{clock}}(f)\, df \end{aligned}$$

$$\left|1 - H_\theta(f)\right|^2 = \frac{f^4}{f^4 + f_N^4} \tag{12.58}$$

Happily, this integral can be solved in closed form [Spilker (1977)], and the result is

$$\begin{aligned}\sigma^2_{\text{clock}} &= \int_0^\infty \frac{f^4}{f^4+f_N^4}\sum_{i=0}^{4}\frac{h_i}{f^i}df \\ &\approx \int_0^\infty \frac{f^4}{f^4+f_N^4}\sum_{i=2}^{4}\frac{h_i}{f^i}df \\ &= \frac{(2\pi)^3 h_4}{(1.2B_{\theta,1})^3}\frac{\pi}{6}+\frac{(2\pi)^2 h_3}{(1.2B_{\theta,1})^2}\frac{\pi}{3\sqrt{3}}+\frac{2\pi h_2}{1.2B_{\theta,1}}\frac{\pi}{3}\ \text{rad}^2\end{aligned} \tag{12.59}$$

The two low-order terms do not contribute meaningfully to the overall integral.

If we combine this error with the error due to external noise, then the aggregate error is as follows.

$$\sqrt{\sigma^2_{\text{clock}}+\text{var}(\Delta\hat{\theta})}\ \text{radians} \tag{12.60}$$

The second term under the radical comes from (12.54) for external noise. This aggregate error is plotted in Figures 12.13 for the TCXO and 12.14 for the OCXO. At high bandwidths, these new curves merge with the curves for external noise only. This makes sense because high bandwidths can certainly capture all of the meaningful clock dynamics, but they also admit more external noise. At low bandwidths, the aggregate noise increases again. The external noise has been attenuated, but authentic clock behavior has also been suppressed.

If the user is stationary, then lower PLL bandwidths can be used. If a TCXO is used, then the bandwidth can be between 5 and 12 Hz. If an OCXO is used, then the bandwidth can be a few hertz.

If the user motion is removed by inertial measurements, then the situation is not so straightforward. If the user motion causes the clock to vibrate, then the clock power spectral density will be greater than those assumed in this analysis. The clock is also sensitive to acceleration. If the receiver is vibrating or accelerating, a more detailed analysis is required and the reader is referred to Hegarty (1997) and Ward (1996).

12.4 Summary

This chapter has explored a key aspect of a GPS receiver—the tracking loops. These loops swing into action after the signal acquisition algorithms have roughly located a GPS signal in the $\{\tau, f_D\}$ search space. The tracking loops refine these rough estimates and continuously track changes in these parameters from that point forward.

The delay lock loop tracks the code phase with correlators that straddle the correlation peak. We analyzed the simplest strategy that places one correlator on each side of the correlation peak and steers the code phase estimate to null the difference in these two measurements. Once tracking, the DLL also provides a prompt correlator that is used by the phase lock loop and the frequency lock loop.

Many DLL variations exist, and the literature is replete with imaginative designs including strobe correlators, double delta discriminators, etc. The literature is also rich with designs that are based on a Fast Fourier Transform (FFT) of the digitized signal. These designs differ markedly in appearance from what we have described here. Their dynamic performance and white noise performance is similar, but they can have large advantages against narrowband radio frequency interference.

The phase lock loop leverages the prompt correlator measurements to develop a metric that can be used to track the carrier frequency and phase. Like the DLL, the PLL is a null seeking servo. It strives to minimize the value of the discriminator function.

The DLL and PLL are a team. For each satellite, the DLL provides prompt correlation measurements to the PLL. Once tracking, the PLL returns the favor and provides rate estimates to the DLL. The DLL uses these to estimate system dynamics. This enables the DLL to use narrower bandwidths and thus reduce the impact of external noise.

The frequency lock loop is briefly described in the appendix. It too is a valuable member of the team. Prior to phase lock, it can track the satellite signal frequency. If the signal is blocked or interference exists in the environment, the FLL may be able to track when the PLL cannot. In this mode, the FLL can still provide rate estimates to the DLL.

Appendix 12.A Frequency Lock Loop

This appendix analyzes the white noise performance of the frequency lock loop (FLL) that uses a cross product discriminator to develop the frequency error signal. This analysis is based on the nice work of Van Dierendonck and Socci (1982).

The FLL is shown in Figure 12.A.1. Our present considerations assume that the delay lock loop has developed a reasonably accurate estimate of the code delay and the prompt correlator sits near the peak of the auto-correlation function. Signal acquisition has provided a rough estimate of the Doppler frequency. The FLL needs to refine this rough estimate and track future changes in this parameter.

The prompt correlator outputs at time $t = nT_{CO}$ are given by

$$\begin{aligned}\tilde{Z}_{P,n} &= \tilde{S}_{P,n}+\tilde{\eta}_{P,n} \\ &= S_{P,I,n}+jS_{P,Q,n}+\eta_{P,I,n}+j\eta_{P,Q,n} \\ S_{I,P,n} &= \sqrt{C}D\cos(\Delta\theta_n) \\ S_{Q,P,n} &= \sqrt{C}D\sin(\Delta\theta_n)\end{aligned} \tag{12.61}$$

The cross product discriminator essentially implements the following trigonometric identity.

$$\sin\alpha\cos\beta-\cos\alpha\sin\beta=\sin(\alpha-\beta)\approx\alpha-\beta \tag{12.62}$$

Specifically, the FLL uses the following discriminator.

$$L_f=\sum_{n=2}^{N}\text{Re}\left\{\tilde{Z}_{P,n-1}\right\}\text{Im}\left\{\tilde{Z}_{P,n}\right\}-\text{Re}\left\{\tilde{Z}_{P,n}\right\}\text{Im}\left\{\tilde{Z}_{P,n-1}\right\} \tag{12.63}$$

The sensibility of this choice can be seen by ignoring noise and considering one sample of signal only.

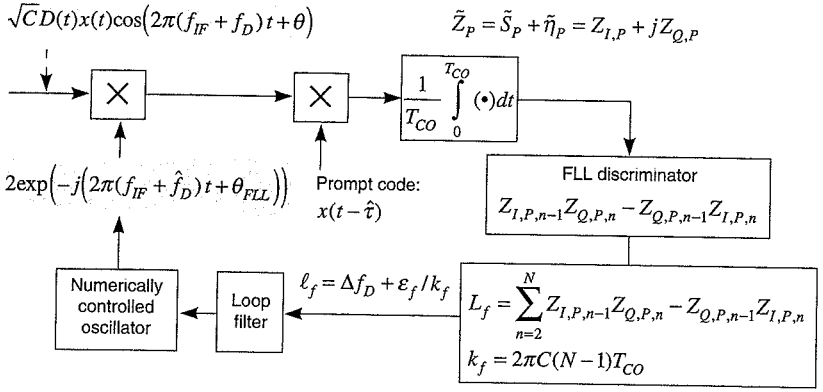


Figure 12.A.1 Frequency lock loop (FLL).

$$\begin{aligned}S_{P,I,n-1}S_{P,Q,n}-S_{P,I,n}S_{P,Q,n-1}&=C\left(\cos(\Delta\theta_{n-1})\sin(\Delta\theta_n)-\cos(\Delta\theta_n)\sin(\Delta\theta_{n-1})\right)\\&=C\sin(\Delta\theta_n-\Delta\theta_{n-1})\\&=C\sin(2\pi\Delta f_D T_{CO})\\&\approx 2\pi CT_{CO}\Delta f_D\end{aligned}\tag{12.64}$$

We have used the following connection between frequency and phase.

$$\Delta\theta_n-\Delta\theta_{n-1}=2\pi\Delta f_D T_{CO}$$

If we average N such signal-only samples, then

$$\sum_{n=2}^{N}S_{P,I,n-1}S_{P,Q,n}-S_{P,I,n}S_{P,Q,n-1}\approx 2\pi C(N-1)T_{CO}\Delta f_D\tag{12.65}$$

This implies that the detector gain for this FLL is $k_f=2\pi C(N-1)T_{CO}$. The sum over N constitutes a noncoherent sum. In the presence of noise, this averaging provides a performance improvement, but this noncoherent integration gain is not as strong as the coherent integration gain, T_{CO}.

With a great deal of work, we can show that the white noise performance of the frequency lock loop is given by

$$\begin{aligned}\text{var}\{\Delta f\}&=\frac{B_{f,1}N}{2\pi^2(N-1)^2T_{CO}^2C/N_0}\left(1+\frac{N-1}{2(C/N_0)T_{CO}}\right)\\\text{var}\{\Delta f\}&=\frac{B_{f,1}N^3}{2\pi^2(N-1)^2(T_B)^2C/N_0}\left(1+\frac{N(N-1)}{2(C/N_0)T_B}\right)\end{aligned}\tag{12.66}$$

These equations assume that $B_{f,1}$ is the one-sided noise bandwidth of the FLL loop filter. T_{CO} is the coherent integration time and T_B is the duration of a GPS bit (20 ms). The derivation of this white noise result is particularly grueling, so it is relegated to a file called FLL.pdf, which can be found on the CD.

Figure 12.A.2 shows some results from (12.66). In both of the equations above, the second term inside the brackets can be ignored for the moderate to high signal-to-noise ratios, where GPS receivers typically operate. However, the second term is important near loop threshold—at low signal-to-noise ratios that can readily be induced by radio frequency interference or signal blockage. The first equation uses N as the number of noncoherent samples and T_{CO} as the coherent integration time. For a fixed loop bandwidth, $B_{f,1}$, the high SNR variance is inversely proportional to C/N_0, N and T_{CO}^2. The low SNR variance is inversely proportional to $(C/N_0)^2$ and T_{CO}^3. It is more or less independent of N.

The second equation above recognizes that $NT_{CO}=T_B$, where T_B is the duration of a data bit. We use this relationship to remove the dependence on T_{CO}. After all, if we specify N and T_B, then T_{CO} is known. In this case, the high SNR variance is proportional to N and inversely proportional to C/N_0 and T_B^2. The low SNR variance is proportional to N^3, and inversely proportional to $(C/N_0)^2$ and T_B^3.

In all cases, coherent averaging is more effective than noncoherent averaging. Sometimes, this is dramatically true. Clearly, the coherent averaging time should be as large as possible. When the signal has been acquired and the locations of the data bit boundaries are well known, then large values of T_{CO} can be used. Typically, $T_{CO}=0.5T_B$ and $N=2$. However, the bit boundaries are not known during signal acquisition, and large values of T_{CO} are risky. After all, if a bit transition occurs in the middle of the coherent average, then the average will probably be worthless. Hence, smaller values of T_{CO} are used and larger values of N result. Typically, $T_{CO}=0.1T_B$ and $N=10$. For GPS, this means that the coherent average is only 2 ms during signal acquisition, but extends to 10 ms or more during tracking.

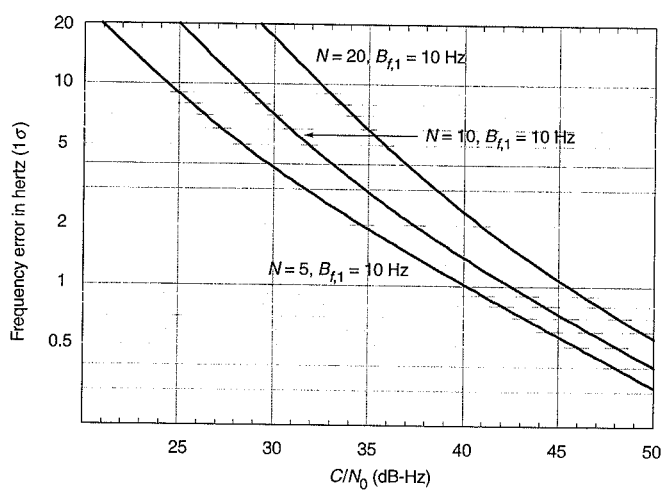


Figure 12.A.2 Performance of frequency lock loop in the presence of white noise.

Homework Problems

12-1. Plot the discriminator for the coherent DLL (12.17) and the early power minus late power discriminator (12.19) versus $\Delta\tau$ for ±2 chips. Comment on the linear ranges and the approximations in (12.18) and (12.20).

12-2. Confirm that the coherent DLL, described by (12.17) and (12.18) has the same noise performance as the baseband DLL given by (10.28). Hint: The basic analysis strategy used in Section 10.6 to derive (10.28) also works for the analysis of the coherent DLL.

12-3. Validate (12.32). Hint: This problem is significantly more challenging than the previous problem because the early power minus late power DLL includes the magnitude operation to strip the nuisance parameters, and this nonlinear operation also acts on the noise. However, the basic strategy described in Section 10.6 is still effective.

12-4. Plot the discriminators $L_\theta = \mathrm{Im}\left\{\tilde{Z}_P\right\} D$ and $L_\theta = \mathrm{Re}\left\{\tilde{Z}_P\right\} \mathrm{Im}\left\{\tilde{Z}_P\right\}$ versus phase error, $\Delta\tau$, and comment on their linear ranges. Ignore noise.

12-5. Validate (12.53). Develop a similar expression for the coherent PLL. This is a nice starting point because it does not multiply one noise term by another, and thus does not suffer squaring loss.

12-6. Develop an expression for the discriminator given in (12.63) as a function of the frequency error. Approximate the slope of the discriminator at zero error, and comment on the linear range of a frequency lock loop (FLL) that uses this discriminator.

References

Akos, D. (1997). A Software Radio Approach to Global Navigation Satellite System Receiver Design, Ph.D. dissertation, Ohio University.

Akos, D., P.-L. Normark, J-T Lee, and K. Gromov (2000). Low Power Global Navigation Satellite System (GNSS) Signal Detection and Processing, *Proceedings of ION GPS 2000*, pp. 784–791.

Hegarty, C. (1997). Analytical Derivation of Maximum Tolerable In-Band Interference Levels for Aviation Applications of GNSS, *Navigation*, vol. 44, no. 1, Spring 1997, pp. 25–34.

Spilker, J. (1977). *Digital Communications by Satellite*, Prentice Hall.

Spilker, J. (1996). Fundamentals of Signal Tracking, Chapter 7 of *Global Positioning System: Theory and Applications, Volume 1*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), AIAA, pp. 245–328.

Van Dierendonck, A.J. (1995). GPS Receivers: What All Those Terms Mean, *GPS World*.

Van Dierendonck, A.J. (1996). GPS Receivers, Chapter 8 of *Global Positioning System: Theory and Applications, Volume 1*, B. Parkinson, J. Spilker, P. Axelrad, and P. Enge (eds.), AIAA, pp. 329–408.

Van Dierendonck, A.J. and G. Socci (1982). AFC Loop Noise Performance for MSR Receivers, Stanford Telecommunications Report STI-TR-22001.

Van Trees, H. (1968). *Detection, Estimation and Modulation Theory, Part 1*, John Wiley.

Ward, P. (1996). Satellite Signal Acquisition and Tracking, Chapter 5 in *Understanding GPS: Principles and Applications*, Elliott Kaplan (ed.), Artech House Publishers, pp. 119–208.

Chapter 13

Coping with Radio Frequency Interference and Signal Obstructions

This chapter introduces the two greatest technical challenges that face satellite navigation: radio frequency interference (RFI) and signal obstructions. As described in Chapter 10, a GPS signal travels 20,000 kilometers from medium earth orbit, and so the received signal power is approximately 10^{-16} watts. This power is comparable to the natural noise contained in a 1 MHz bandwidth at the GPS L1 or L2 frequency. If man-made interference is added to natural noise, then the situation quickly becomes bleak. Similarly, if the GPS user is downtown or indoors, then signal obstructions can be a real challenge.

This chapter contains six essays designed to teach the basics. The first essay, Section 13.1, introduces the numbers. Specifically, it provides nominal values of signal-to-noise ratio, C/N_0. It goes on to present signal-to-noise ratios measured in the heavily obstructed environment of downtown Tokyo. Then it approximates the power needed by a terrestrial source of radio frequency interference to significantly reduce C/N_0.

The second essay, Section 13.2, develops some simple techniques for predicting the received power from a terrestrial source of RFI. It describes some of the complications that distinguish the terrestrial path from the satellite-to-earth path. We also develop a technique to predict the RFI power based on a two-path propagation model. For such a model, the received power follows a $1/R^4$ law rather than the $1/R^2$ law in force for the satellite-to-earth path.

The remaining essays are devoted to techniques that mitigate RFI and signal obstructions. For these, Figure 13.1 serves as a guide [Ward (1996)]. Section 13.3 is devoted to the role of the antenna in mitigating RFI. Military applications of GPS include adaptive antennas that either steer a null towards interference or steer a beam towards the GPS satellites. Others discriminate based on the polarization of the interfering wave. Rather than describe this entire technology, we develop one beautiful example—a ring nulling antenna developed by Baueregger (2003).

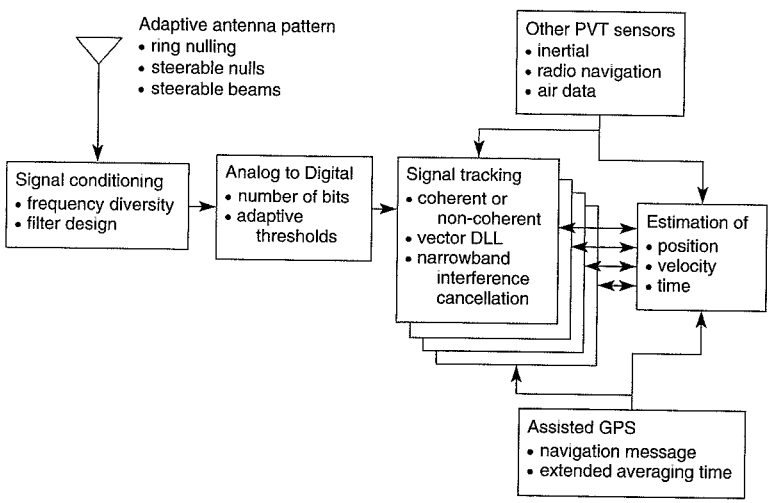


Figure 13.1 Summary of techniques to mitigate radio frequency interference (RFI) or signal obstructions.

Section 13.4 introduces assisted GPS, or AGPS, which is primarily designed to support GPS receivers in cell phones. Such phones will serve many interesting consumer applications, and they will also serve an important public safety purpose. GPS-enabled phones will automatically report the GPS position when a caller dials the emergency code (911 in the United States). If the caller is involved in an accident, injury or stress may prevent him from giving an accurate description of his own location. If the caller is a good Samaritan, he may not be familiar with the local area. However, the emergencies may occur in urban areas or even indoors where signal obstructions are a real problem. AGPS provides needed assistance. It enables the GPS phone to provide a rough position estimate based on the short snippets of noisy data likely to be available downtown and indoors. To do so, AGPS sends assistance data from a receiver at an electromagnetically quiet site to the roving receiver in the uncontrolled environment.

Section 13.5 turns our attention to inertial navigation, which is a natural complement to GPS. Inertial sensors are self contained and thus not threatened by RFI. On the other hand, they suffer from an error that grows with time. Absent RFI, GPS can be used to calibrate the inertial sensors and mitigate the error growth. When RFI is present, inertial navigation can allow the user to coast through the outage. A large literature exists to describe all of the possible ways to combine GPS and inertial navigation. Rather than review this entire field, we teach the basic workings of an inertial navigation system (INS). By so doing, we hope to convey the essential power of the combined use of GPS and INS.

The fifth essay, Section 13.6, returns our attention to tone interference. In Chapter 9, we found that RFI with bandwidths less than 1 kHz can have an impact that is 10 dB greater than the impact of wideband RFI with the same power. However, the spectral structure that makes narrowband RFI so dangerous also enables a special collection of mitigation techniques. These techniques are not effective against wideband RFI, but are wonderfully effective against narrowband RFI. Once again, the literature is replete with analyses of these signal processing algorithms. We make no attempt to summarize this extensive literature. Instead, Section 13.6 focuses on one simple example—adaptive analog-to-digital conversion. We choose this example because it is relatively simple, it clearly leverages the narrowband nature of the RFI, and it can be incorporated in modern GPS receivers.

Finally, we should list some important ideas that are not covered in this chapter. First, a variety of techniques exist to detect the presence of RFI and/or signal obstructions. Some receivers monitor the total signal-plus-noise power. The GPS signal does not contribute noticeably to the power measured in the front end of the receiver, and so any power increase indicates the presence of RFI. Additionally, some receivers monitor the scatter of the inphase and quadrature (I/Q) samples. Absent RFI, the I/Q scatter should be tightly clustered around two points representing the two possible polarities of the binary data in the navigation message. As the signal-to-noise ratio increases, the clusters will move farther apart. If RFI is present or a signal is obstructed, then the clusters will move closer together or overlap. Finally, receivers monitor the failure rate of the parity check on the GPS navigation data. A cluster of parity failures indicates RFI or signal blockage.

Second, the simplest cure to RFI and signal obstructions is to carry a backup navigation system, and the prudent navigator does just that. Example backups include Loran-C for mariners and distance measuring equipment (DME) for aviators. Consumer applications may well augment GPS with range measurements to television stations or to cell phone base stations [Rabinowitz and Spilker (2003)]. Indeed, the literature is filled with such proposals.

Third, the new GPS and Galileo signals, described in Chapter 3, will certainly help combat accidental RFI. This newfound strength will accrue in a number of ways. If one frequency is lost to RFI, say L1, then the receiver can continue to navigate using L2 or L5. In addition, the new L5 signal will have a chipping rate that is ten times faster than the chipping rate for the C/A-code. This provides an additional 10 dB of processing gain against interference. Finally, vector delay lock loops may provide a particularly powerful way to aggregate the signal power from all of the visible satellites [Sennott and Senffner (1994) and Spilker (1996)].

However, the new signals will not, by themselves, eliminate the threat of malevolent RFI. Malevolent jammers will need to radiate more power, but small lightweight transmitters could still disrupt GPS operations over large areas. Moreover, signal blockages are still a nuisance because all the signals from a given satellite would be blocked. For these reasons, future GNSS receivers for critical applications will still need some combination of advanced antennas, assistance data, inertial navigation, and tone cancellation.

13.1 Overview

13.1.1 Nominal Signal-to-Noise Ratios

Mercifully, the GPS portion of the radio spectrum is usually devoid of man-made interference. Figures 13.2 and 13.3 are *spectrum surveys* [P. Enge *et al.* (2004)]. In other words, they are measurements of received power density as a function of frequency. Figure 13.2 shows measurements taken in Yosemite Park, and Figure 13.3 shows measurements taken in downtown San Jose, California.

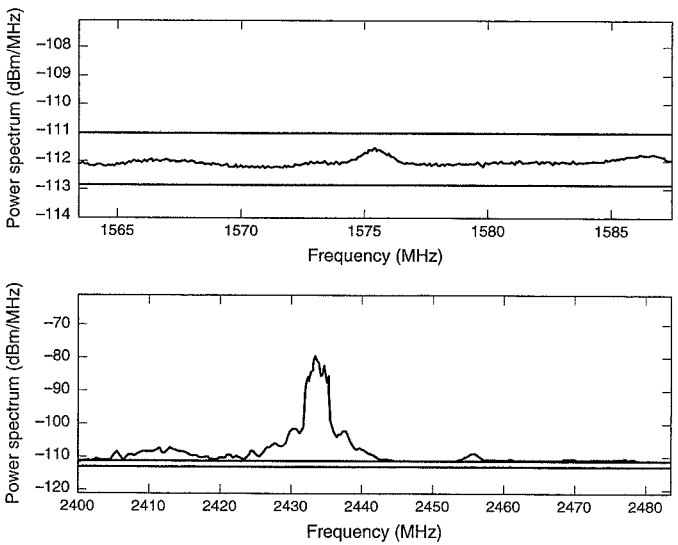


Figure 13.2 GPS and ISM spectrum surveys from a rural site (courtesy of Juyong Do, Stanford University).

In both cases, the top trace covers a 20 MHz span from 1565 to 1585 MHz, which is part of the Aeronautical Radio Navigation Service (ARNS) band that extends from 1559 to 1610 MHz. This ARNS band is primarily reserved for signals coming from the navigation satellite systems—GPS, GLONASS, and Galileo. The GPS L1 signal occupies the middle of the ARNS band, and the top trace in Figure 13.2 shows a slight bump centered at f_{L1} = 1575.42 MHz. This bump is the aggregation of the power received from all of the GPS satellites in view from this unobstructed site in Yosemite park.

The bottom traces in both figures cover an 80 MHz span from 2400 to 2480 MHz, which is an Industrial, Scientific and Medical (ISM) band. This band houses signals for cordless telephones, the *wi-fi* network that wirelessly connects our laptop computers, and unintentional emissions from microwave ovens. These signals are terrestrial, and so the signals in the ISM band are much stronger than the signals in the ARNS band. To capture this difference, the top and bottom traces in Figure 13.2 use different vertical scales.

To quantify the difference between the two bands, both figures also show theoretical power densities for natural noise as two horizontal lines. One line is for an antenna pointed skyward and the other is for an antenna pointed toward the earth. The skyward prediction is 2 dB lower than the earthward prediction because the sky is *cooler* than the warm earth with respect to radio frequency noise (Section 10.4).

These spectrum surveys show that the ARNS and ISM bands constitute dramatically different radio environments. As it needs to be, the ARNS band is much quieter than the ISM band. The ARNS band houses weak satellite signals that serve many critical applications, including aircraft landing, ambulance guidance, and harbor entrance. Hence, the radio regulatory agencies have acted to protect this band from terrestrial radio signals. The top traces indicate

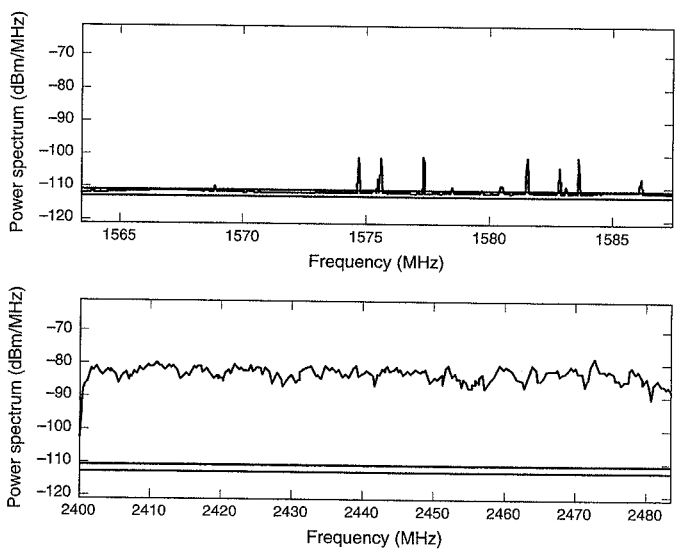


Figure 13.3 GPS and ISM spectrum surveys from an urban site (courtesy of Juyong Do, Stanford University).

that their efforts have been conspicuously successful. A few weak man-made signals are apparent in the San Jose survey, but even these only occupy a narrow bandwidth and are not much stronger than the natural noise background. The ARNS band is almost pristine.

In contrast, the ISM band is an electromagnetic circus. In San Jose, man-made signals in this band raise the *noise* floor by almost 30 dB (1000 times more power). Even in Yosemite Park, man-made signals appear 30 dB above the noise floor. However, this band also serves its purpose. After all, it is designed to serve a wide variety of terrestrial radio users. These users access the band knowing that they will find other strong signals in the band. The measured congestion indicates that great utility is achieved. However, we would be reluctant to place safety critical signals in this band—they need assured access to the spectrum.

Most of the time, GPS signals enjoy electromagnetic quiet and a clear path from the satellite to the user's antenna. In this case, the received carrier power-to-noise-density ratio, C/N_0, may be estimated from (10.8) and (10.18), which we repeat with slight modifications.

$$\begin{aligned} C &= G_R P \quad \text{watts} \\ &= \frac{P_T G_T G_R}{L_A}\left(\frac{\lambda}{4\pi R}\right)^2 \quad \text{watts} \\ N_0 &= k(T_A + T_R) \quad \text{W/Hz} \end{aligned} \tag{13.1}$$

In these equations, $P_T G_T$ is the equivalent power radiated by a GPS satellite, where G_T is the gain due to the satellite antenna. The GPS wavelength is λ and the distance from the satellite to the user is R. The received GPS signal power, P, is only 10^{-16} watts or –160 dBW or –130 dBm. This power is amplified by the gain of the user's receiving antenna, G_R, to yield the effective carrier power, C. Even in the absence of RFI, the GPS signal must still compete with natural noise received by the antenna, T_A, and noise generated in the front end of the receiver, T_R. The noise density, N_0, is equal to Boltzmann's constant, k, times the sum of these two noise temperatures, and is typically around –201 dBW/Hz. The C/N_0 estimated by (13.1) does not include implementation losses such as the quantization loss associated with the analog to digital converter.

As discussed in Chapter 10, signal-to-noise ratio is a function of elevation angle. At an angle of 5°, we have $C/N_0 \approx 35$ dB-Hz. At zenith, $C/N_0 \approx 50$ dB-Hz, where these numbers are for a typical patch antenna. This state of affairs is captured in Table 13.1, which presents C/N_0 measurements taken in Japan [Agarwal *et al.* (2002)]. Under open sky, the strongest signal was always 50 dB-Hz or stronger. The second strongest signal ranged from 45 to 50 dB-Hz, and the weaker signals ranged from 25 to 50 dB-Hz.

13.1.2 Signal Obstructions

Signal obstructions pose our first challenge. More and more GPS users find themselves in environments that block satellite signals. For example, cell phone users want to be able to automatically communicate their position when they call to report an emergency. Moreover, they want this capability even if they are downtown or indoors.

If an obstruction attenuates the signal, then the numerator of our signal-to-noise ratio suffers, and the received signal power is reduced by $G_{\text{block}} < 1$. With signal obstructions, our signal-to-noise ratio becomes

Table 13.1 Measured C/N_0 in Tokyo (in dB-Hz)

	Clear sky	*Downtown*	*Inside a hotel*
Strongest signal	50	40–50	10–25
Second strongest signal	45–50	35–45	< 20
Other received signals	25–50	10–40	< 20

$$\frac{C}{N_0} = \frac{G_{\text{block}} G_R P}{N_0}$$

Table 13.1 shows the impact of obstructions on C/N_0. In downtown Tokyo, the C/N_0 for the strongest signal is still 40 to 50 dB-Hz, but now the second strongest signal has dropped to 35 to 45 dB-Hz. The weaker signals are really suffering with C/N_0 ranging from 10 to 40 dB-Hz. Inside a Tokyo hotel, the measured C/N_0 is 10 to 25 dB-Hz for the strongest signal and weaker than 20 dB-Hz for all other signals.

13.1.3 Radio Frequency Interference

Radio frequency interference (RFI) poses our second challenge. As introduced in Section 9.8, RFI refers to man-made signals other than GPS that exist in the GPS portion of the radio spectrum. As discussed above, the radio regulatory agencies around the world work hard to protect the GPS spectrum, and most terrestrial radio transmissions sharply limit the amount of energy they send in the GPS bands. In other words, they limit their out-of-band-emissions (OOBE). However, OOBE can result when radio transmitters age and their components drift out of tolerance. At other times, ill-designed equipment is fielded and results in RFI. Even properly operating equipment can result in interference if it is close to a GPS receiver. Television signals are so powerful that their harmonics can interfere with GPS.

Sadly, some RFI is malevolent, and military users of GPS must face the prospect of intentional jamming. In these cases, the adversary radiates strong signals in the GPS band. As we shall discover, they can interfere with ill-prepared GPS-based operations for tens or even hundreds of kilometers. Needless to say, the military has developed a host of countermeasures to mitigate such an electromagnetic attack. In this chapter, we will concentrate on countermeasures developed for the civil community. However, most of these find their origins in military programs.

If RFI is weak compared to the received natural noise, then it can be neglected. If the received RFI is strong, then its impact depends on whether it is pulsed or continuous in time.

Pulsed interference is sometimes very well tolerated by properly designed GPS receivers. This type of RFI is roughly categorized by the *pulse duration* and *duty cycle*. Pulse duration is the time width of an individual pulse of RFI. Duty cycle is the percentage of time that the interfering pulses are on. If the pulse duration is small compared to a GPS data bit (20 ms) and the duty cycle is less than 10%, then pulsed RFI is generally not problematic even if it is very

powerful. The receiver clips the amplitude of such pulses. Once clipped, these pulses punch a hole in the received signal. However, if the hole is small compared to the total duration of a data bit, then the receiver can still make reliable bit decisions.

Continuous interference is generally more troublesome than pulsed RFI and is our sole focus in the remainder of this chapter. This variety of RFI is sometimes known as continuous wave or CW interference. It includes wideband RFI, narrowband RFI, and tone interference. All of these waveforms are continuous in time, but have different spectral characteristics.

Wideband Interference

Wideband RFI has a bandwidth greater than the bandwidth of the victim signal. So RFI that is wideband relative to the C/A-code has a bandwidth of 2 MHz or more. To be wideband to the P(Y) code, the RFI bandwidth must be 20 MHz or more. In these cases we simply add the RFI power spectral density, denoted $J_0/2$, to the PSD for natural noise. The ratio C/N_0 becomes $C/(N_0 + J_0)$.

$$\frac{C}{N_0 + J_0} = \frac{G_R P}{N_0 + J_0} \quad \text{Hz}$$

$$J_0 = \frac{J_{0,T} G_{J,T} G_{J,R}}{L_{J,A}} \left(\frac{\lambda}{4\pi R_J} \right)^2 \quad \text{W/Hz} \tag{13.2}$$

In this equation, the RFI source radiates an interference power spectral density of $J_{0,T}$ W/Hz. The RFI transmitter has an antenna gain of $G_{J,T}$ toward the GPS victim receiver. The gain of the GPS receive antenna is $G_{J,R}$ in the direction of the jammer. The RFI wavelength, λ, is the same as the GPS wavelength because the two signals occupy the same portion of the radio spectrum. The distance from the RFI source to the GPS victim receiver is R_J. Our formula (13.2) assumes that the RFI power obeys the same $1/R^2$ propagation law that applies to the GPS signals as they travel through space. As we shall discover in Section 13.2, this approximation is very crude, especially for RFI that follows a complicated overland path from its source to the GPS receiver. For the time being, $L_{J,A}$ is our attempt to account for departures from the $1/R^2$ law.

We wish to 'ballpark' the values of $J_{0,T}$ that will be harmful to GPS. To this end, we set $G_{J,T} = G_{J,R} = L_{J,A} = 1$. We seek the value of $J_{0,T}$ that will raise the natural noise floor by a factor of γ. Thus, we wish to solve

$$J_0 = \gamma N_0$$

$$J_{0,T} = \gamma N_0 \left(\frac{4\pi R_J}{\lambda} \right)^2 \quad \text{watts/Hz} \tag{13.3}$$

We may rewrite these formulas using dB.

$$J_{0,dB} = \gamma_{dB} + N_{0,dB}$$

$$J_{0,T,dB} = \gamma_{dB} + N_{0,dB} + 10\log_{10}\left(\frac{4\pi R_J}{\lambda} \right)^2 \quad \text{watts/Hz} \tag{13.4}$$

Recall from Chapter 10 that the natural noise floor for a GPS receiver is approximately –200 dBW/Hz or –140 dBW/MHz.

Like most satellite radio systems, GPS is designed to operate close to its natural noise floor. It is simply too expensive to launch transmitters with reserve power into medium earth orbits. Hence we wish to sharply limit the growth in the noise floor due to RFI. To protect low-lying satellites, we might limit that growth to a factor of 2, which means $\gamma_{dB} = 3$. Solving (13.4) for this value of γ_{dB}, we find that we can tolerate an RFI source with –80 dBW/MHz at 10 meters, –60 dBW/MHz at 100 meters, and –40 dBW/MHz at 1000 meters. A radiated J_0 of –80 dBW/MHz corresponds to 10^{-8} watts/MHz or only 0.01 micro-watts per megahertz of bandwidth.

A level of 10^{-8} watts/MHz certainly sounds low, and it is. Laptop computers are allowed to radiate 10^{-7} watts/MHz in any band, and most consumer electronic equipment is allowed to radiate at this level. Fortunately, the average laptop or hair dryer seldom radiates at its allowed level in the GPS band!

Tone Interference

As discussed in Section 9.7, tone interference concentrates all of its power at one frequency and can be modeled as a sine wave.

$$J(t) = \sqrt{2P_J}\cos(2\pi f_J t + \theta_J) \tag{13.5}$$

The power in this interfering sine wave is P_J. The frequency, f_J, is close to the GPS frequency, $f_J \approx f_L$.

If we continue to rely on our temporary assumption that the radio power follows the $1/R^2$ propagation law, then

$$P_J = \frac{P_{J,T} G_{J,T} G_{J,R}}{L_{J,A}} \left(\frac{\lambda}{4\pi R_J} \right)^2 \quad \text{watts} \tag{13.6}$$

This formula is the same as (13.2) with one difference. It assumes that the RFI source radiates a tone with power equal to $P_{J,T}$ watts. In contrast, (13.2) assumes that the RFI source radiates an interference power spectral density of $J_{0,T}$ watts/hertz.

Our key ratio, C/N_0, has served us well. It is the single most powerful characterization of the signal environment and enables us to predict acquisition and tracking performance. Equation (13.2) embellishes this ratio to account for wideband RFI. Happily, we can also modify C/N_0 to account for tone interference.

Prior to correlation, the GPS signal energy is spread over a bandwidth of several megahertz around f_L, and the power of the RFI tone is concentrated at one frequency, $f_J = f_L$. After correlation, the GPS power is concentrated in a narrowband, and the tone power is now spread. As discussed in Section 9.7, correlation spreads the tone power such that only P_J/PG watts appear in the *de-spread* bandwidth. PG is called the processing gain and is defined as follows.

$$PG = \frac{B_{\text{code}}}{B_{\text{despread}}} \tag{13.7}$$

Processing gain is proportional to the chipping rate of the spread spectrum code. It is inversely proportional to the de-spread bandwidth. In Section 9.7, we considered the impact of RFI on the demodulation of navigation bits. Hence, our de-spread bandwidth was the noise equivalent bandwidth of the navigation message or 50 hertz. In this section, we wish to approximate the

impact of a tone interference as a power spectral density. Hence, we consider $B_{\text{despread}} = 1$ Hz. After all, N_0 is the noise power in 1 hertz of bandwidth. For the C/A code, we now write

$$\begin{aligned} PG &= \frac{10^6}{1} \\ PG_{\text{dB}} &= 10\log_{10}(10^6) = 60 \text{ dB} \\ J_0 &\approx P_J/\text{PG} \quad \text{W/Hz} \\ J_{0,\text{dB}} &\approx P_{\text{J,dB}} - 60 \text{ dB W/Hz} \end{aligned} \tag{13.8}$$

When the tone dominates the thermal noise, we may write

$$\frac{C}{N_0 + J_0} \approx \frac{C}{P_J/PG} \tag{13.9}$$

Recall from Chapter 12, the phase lock loop begins to suffer large tracking errors and the prospect of cycle slips if C/N_0 drops below 30 dB-Hz or so. If we denote this threshold value as $(C/N_0)^*$, we may solve for the maximum tolerable power in a received tone.

$$\begin{aligned} \left(\frac{C}{P_J/PG}\right) &= \left(\frac{C}{N_0}\right)^* \\ \left(\frac{C}{P_J}\right)_{\text{dB}} &= \left(\frac{C}{N_0}\right)^*_{\text{dB}} - PG_{\text{dB}} \\ \left(\frac{C}{P_J}\right)_{\text{dB}} &= 30 - 60 \\ \left(\frac{C}{P_J}\right)_{\text{dB}} &= -30 \end{aligned} \tag{13.10}$$

In other words, we can tolerate a tone jammer with 30 dB more power than the C/A signal. Since, the P(Y) codes are ten times faster than the C/A code, they tolerate jammers that are 40 dB more powerful than the P(Y) code signal.

Bear in mind that this approximation is rough. As discussed in Section 9.8.2, short codes like the C/A code can depart sharply from this approximation. Moreover, the analog-to-digital converter can also introduce appreciable implementation loss in the presence of tone interference. This will be further discussed in Section 13.6. In addition, (13.10) ignores thermal noise. Homework Problem 13-2 asks you to plot the tolerable tone power versus the nominal signal-to-noise ratio. If C/N_0 = 40 dB-Hz without the tone and the receiver requires 30 dB-Hz to maintain track, then the receiver can tolerate a tone that is 29.5 dB stronger than the GPS signal. If the nominal signal-to-noise ratio drops to 36 or 32 dB-Hz, then the tolerable tone power drops to 28.7 and 25.7 dB, respectively.

As shown, correlation provides a processing gain against tone interference. Even so, a tone jammer can impact geographical areas as large as the wideband jammer analyzed above. However, a greater variety of mitigation techniques are available to combat tone jammers; we discuss one such technique in Section 13.6.

13.2 Terrestrial Radio Propagation

Our RFI study must begin with an understanding of how much RFI power might arrive at a GPS antenna. As such, this section is akin to Chapter 10, where we derived a link budget for the GPS signals, but now we derive a link budget for a signal from a terrestrial source. Equations (13.2) and (13.6) assumed that the RFI power follows a $1/R^2$ propagation law. This assumption is decent for signals from space to earth. However, the terrestrial propagation is much more complicated. We now consider a two-path propagation that embraces some of this complexity.

13.2.1 Two-Path Model: Exact Expression

The present analysis is excerpted from lecture notes prepared by Professor Donald Cox at Stanford University. It considers the two-path terrestrial propagation model shown in the top half of Figure 13.4. The RFI source is located at height h_T and the GPS victim antenna is at height h_R. The ground causes a reflection with relative amplitude α. The horizontal distance from the jammer to the victim antenna is given by R. The direct path without a reflection has length R_D, and the path that includes the reflection has length R_M. Both these ranges are slightly longer than R and are exactly given by

$$\begin{aligned} R_D &= \sqrt{R^2 + (h_T - h_R)^2} \\ R_M &= \sqrt{R^2 + (h_T + h_R)^2} \end{aligned} \tag{13.11}$$

The interfering signal is now given by the following sum.

$$J(t) = \sqrt{2P_J}\cos\left(2\pi f_J t - kR_D\right) + \alpha\sqrt{2P_J}\cos\left(2\pi f_J t - kR_M\right) \tag{13.12}$$

As shown, the tone that arrives at the GPS receiver has two components—one for each path. The amplitude of the second tone is attenuated by α. Otherwise, the two amplitudes follow the slow $1/R^2$ dependence that we are familiar with. However, the phases are much more subtle

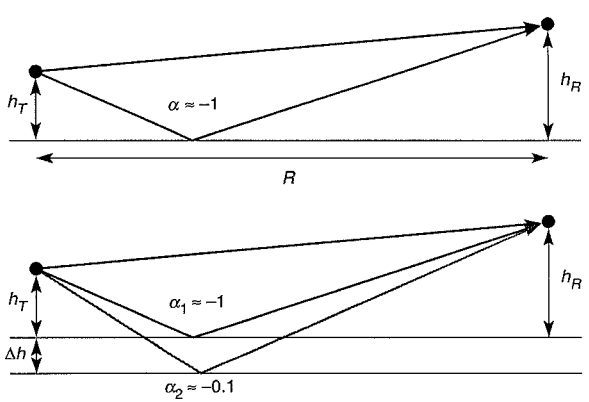


Figure 13.4 Two-path and three-path propagation models for terrestrial interference.

and must be handled with care. They change by 2π radians every 19 centimeters—we really do need to keep track of the exact propagation distances when computing phase. To that end, Equation (13.12) introduces the so-called wave number, $k = 2\pi/\lambda$, to streamline our bookkeeping for the analysis that follows.

The interfering signal from (13.12) has the following average power.

$$\begin{aligned} P_{J,2} &= \frac{1}{T}\int_0^T 2P_J\left(\cos\left(2\pi f_J t - kR_D\right) + \alpha\cos\left(2\pi f_J t - kR_M\right)\right)^2 dt \\ &= \frac{2P_J}{T}\int_0^T \cos^2\left(2\pi f_J t - kR_D\right) + 2\alpha\cos\left(2\pi f_J t - kR_D\right)\cos\left(2\pi f_J t - kR_M\right) \\ &\quad +\alpha^2\cos^2\left(2\pi f_J t - kR_M\right)dt \end{aligned} \tag{13.13}$$

We now assume that the averaging time, T, is long compared to the period of this carrier signal. This very reasonable assumption allows us to write

$$\begin{aligned} P_{J,2} &= \frac{P_J}{T}\int_0^T (1+\alpha^2) + 2\alpha\cos\left(k(R_M - R_D)\right)dt \\ &= P_J\left((1+\alpha^2) + 2\alpha\cos\left(k(R_M - R_D)\right)\right) \\ &= P_{J,T}G_{J,T}G_{J,R}\left(\frac{\lambda}{4\pi R}\right)^2\left((1+\alpha^2) + 2\alpha\cos\left(k(R_M - R_D)\right)\right) \end{aligned} \tag{13.14}$$

This relationship gives the scalloped curve in Figure 13.5, where we assume that $h_T = 1$, $h_R = 2$, and $\alpha = -1$. We also assume that the antenna gains are 0 dB ($G_{J,T} = G_{J,R} = 1$). Finally, we assume that the radiated interference power is 1 mW ($P_{J,T} = 10^{-3}$ W). At short ranges, the received power scallops around the $1/R^2$ expression provided by free-space propagation. Absent any truly detailed information on the reflection environment, we might as well use the free space expression for these short ranges. At longer ranges, the received interference power is well modeled by a $1/R^4$ dependence that we now derive.

13.2.2 Two-Path Model: Approximation for Long Range

We seek the constant of proportionality for the $1/R^4$ approximation valid at long ranges. Since we will be concerned with ranges that are long compared to the antenna heights, we begin with Taylor Series approximations as follows

$$\begin{aligned} R_D &= \sqrt{R^2 + (h_T - h_R)^2} \\ &\approx R\left(1 + \frac{(h_T - h_R)^2}{2R^2}\right) \\ R_M &\approx R\left(1 + \frac{(h_T + h_R)^2}{2R^2}\right) \end{aligned} \tag{13.15}$$

In both cases, we employ the first three terms of the Taylor series expansion.

If we use these approximations, (13.14) becomes

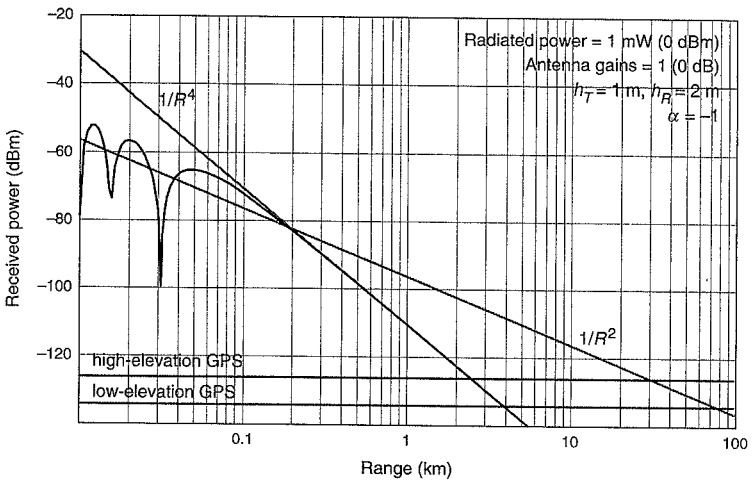


Figure 13.5 Two-path propagation law for L1 RFI.

$$\begin{aligned} P_{J,2} &\approx P_J\left(1 + \alpha^2 + 2\alpha\cos\left(\frac{2kh_Th_R}{R}\right)\right) \\ &= P_J\left(1 + \alpha^2 + 2\alpha - 4\alpha\sin^2\left(\frac{kh_Th_R}{R}\right)\right) \\ &= P_J\left((1+\alpha)^2 - 4\alpha\sin^2\left(\frac{kh_Th_R}{R}\right)\right) \end{aligned} \tag{13.16}$$

At this point, the small angle approximation for the sine function is reasonable because

$$R >> kh_Th_R$$

With this approximation, we may write

$$P_{J,2} = P_J\left((1+\alpha)^2 - 4\alpha\left(\frac{kh_Th_R}{R}\right)^2\right)$$

Now we assume that $\alpha = -1$, incorporate (13.6), and $k = 2\pi/\lambda$.

$$\begin{aligned} P_{J,2} &= P_{J,T}G_{J,T}G_{J,R}\left(\frac{\lambda}{4\pi R}\right)^2 4\left(\frac{2\pi h_Th_R}{\lambda R}\right)^2 \\ &= \frac{P_{J,T}G_{J,T}G_{J,R}h_T^2h_R^2}{R^4} \quad \text{watts} \end{aligned} \tag{13.17}$$

We express this result using decibels

$$P_{J,2,dB} = P_{J,T,dB} + G_{J,T,dB} + G_{J,R,dB} + 20\log_{10} h_T + 20\log_{10} h_R - 40\log_{10} R \quad \text{dBW} \tag{13.18}$$

This is the expression for the $1/R^4$ line shown in Figure 13.5.

If we equate our $1/R^2$ approximation to our $1/R^4$ approximation, we find that they are equal when

$$R^* = 2kh_T h_R \quad \text{m} \tag{13.19}$$

Beyond R^*, the $1/R^4$ approximation is better than the $1/R^2$ approximation. At shorter ranges, the $1/R^2$ approximation is better. Notice that R^* only depends on the heights of the transmit and receive antennas, and the wave number, k.

13.2.3 Three-Path Model

Consider the three-path model shown in the bottom half of Figure 13.4. The direct path length is still denoted R_D. The two paths with reflections have lengths $R_{M,1}$ and $R_{M,2}$. The reflection coefficients for the two paths are α_1 and α_2. For this case, the average received interference power is given by

$$P_{J,3} = P_{J,T}G_{J,T}G_{J,R}\left(\frac{\lambda}{4\pi R}\right)^2 \left(1 + \alpha_1^2 + \alpha_2^2 + 2\alpha_1 \cos(k(R_{M,1} - R_D)) + 2\alpha_2 \cos(k(R_{M,2} - R_D)) + 2\alpha_1\alpha_2 \cos(k(R_{M,2} - R_{M,1}))\right) \tag{13.20}$$

Homework Problem 13-3 asks you to derive this expression.

This equation gives the scalloped curve in Figure 13.6, where we continue to assume that $h_T = 1$ m, $h_R = 2$ m, $G_{J,T} = 1$, $G_{J,R} = 1$, and $P_{J,T} = 10^{-3}$ W. The second path contains a strong reflection with $\alpha_1 = -1$. However, the third path contains a reflection from a boundary that is 0.5 meter below the first reflection. The reflection for the third path is weak with a reflection coefficient of only $\alpha_2 = -0.1$.

At ranges less than 100 meters, the three path behavior shown in Figure 13.6 does not differ dramatically from the two path behavior shown in Figure 13.5. Both curves scallop around the same $1/R^2$ approximation. At ranges between 100 m and 1000 m, the third path does not change the result markedly from the two path results. However, for ranges greater than 1000 m, the results are quite different. As previously noted, the two path behavior follows the $1/R^4$ approximation. However, the third path, albeit weak, causes the received power to follow a different $1/R^2$ asymptote.

Apparently, predicting received power for terrestrial radio signals is tough! We added a third path with a weak reflection coefficient. This small change significantly changed the received power at all ranges beyond 1000 m. Moreover, actual terrestrial radio paths are plagued by more vagaries than we have considered here, including terrain, buildings, foliage, and atmospheric variations. For all these reasons, RFI power levels are difficult to predict, and the results herein are rough approximations.

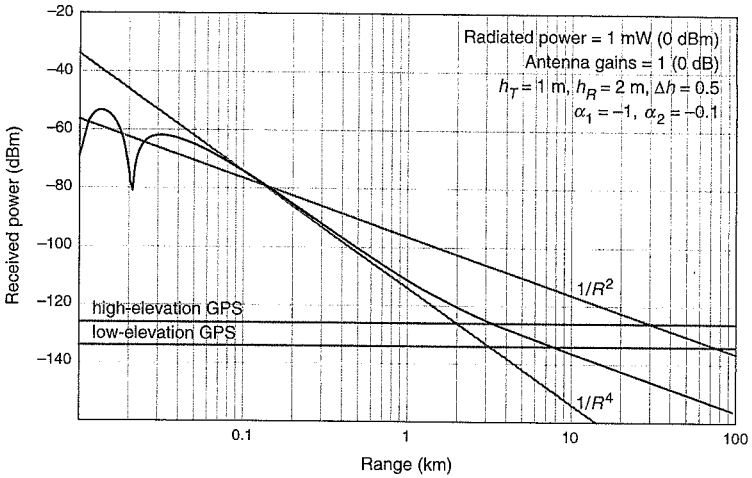


Figure 13.6 Three-path propagation law for L1 RFI.

13.3 Antennas

The GPS antenna is our first line of defense against RFI. After all, the received RFI power is proportional to $G_{J,R}$ which is the gain of the GPS receiver towards the RFI source. Some military users employ adaptive antennas that either steer a null towards interference or steer a beam towards the GPS satellites. Others discriminate based on the polarization of the interfering wave. Rather than describe this entire technology, we develop a simplified example of a nulling antenna and then describe one beautiful implementation—a ring nulling antenna developed by Bauregger (2003). More comprehensive treatments of this subject are contained in Compton (1988), Applebaum (1976), Widrow (1967), Rounds (2004), and Rosen and Braasch (1998).

13.3.1 Two-Element Nulling Antenna

To understand antenna nulling, consider Figure 13.7, which shows two antennas separated by a distance d meters. The signal arriving at the right hand antenna is given by

$$\sqrt{2P}D(t)x(t)\cos(2\pi(f_L + f_D)t + \theta_S) + \sqrt{2P_J}\cos(2\pi f_J t + \theta_J) \tag{13.21}$$

The left-hand term is our familiar GPS signal modulated by the code, $x(t)$, and navigation data, $D(t)$. Tone interference appears on the right. We neglect natural noise and the signals from the other GPS satellites, so that we can focus on the game at hand.

These signals pass through the signal conditioning chain depicted in Figures 11.3 and 11.4. Both are amplified by a common gain, so we will ignore these scaling factors. Both signals are also down converted by the mixers in the front end of the GPS receiver and multiplied by inphase and quadrature reference signals. Recall that Figure 11.16 depicts this process for a generic GPS receiver. The resulting signal from the right-hand antenna is given by

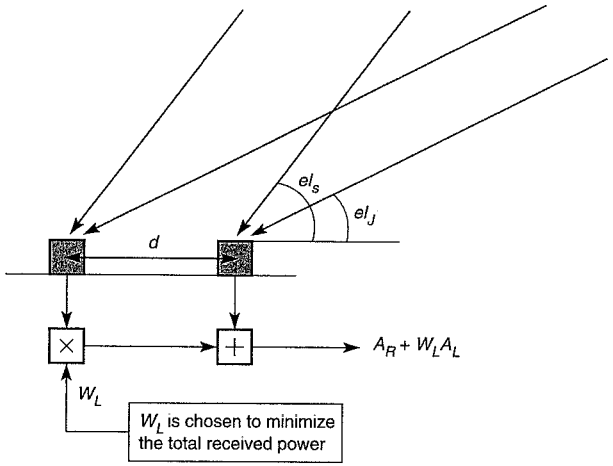


Figure 13.7 Geometry for a one-dimensional analysis of a simple two-element array used for spatial nulling.

$$A_R(t) = \sqrt{P}D(t)x(t)\exp j(2\pi\Delta f_D t + \Delta\theta_S) + \sqrt{P_J}\exp j(2\pi\Delta f_J t + \Delta\theta_J) \tag{13.22}$$

In this equation, $\Delta f_D = f_D - \hat{f}_D$ and $\Delta f_J = f_J - f_L - \hat{f}_D$ are the frequency offsets of the incoming carrier signals and the receiver's replica of the carrier. $\Delta\theta_S$ and $\Delta\theta_J$ are the corresponding phase offsets.

Of course, the left-hand antenna provides similar signals, with one key exception. There is an additional phase shift due to the increased path length from the interference source and satellite to the left-hand side of Figure 13.7. This additional path length is $d\cos el_J$ for the interference source and $d\cos el_S$ for the satellite signal. We hope to leverage the difference between these two phases. Consequently, we write

$$A_L(t) = \sqrt{P}D(t)x(t)\exp(jkd\cos el_S)\exp j(2\pi\Delta f_D t + \Delta\theta_S) + \sqrt{P_J}\exp(jkd\cos el_J)\exp j(2\pi\Delta f_J t + \Delta\theta_J) \tag{13.23}$$

In this equation, we have multiplied the additional path length measured in meters by $k = 2\pi/\lambda$ to get the incremental distance measured in radians.

Our essential nulling strategy is shown in Figure 13.7. We intend to form a linear combination of the signals from the two antennas. Specifically, we plan to multiply $A_L(t)$ with a complex weight and then add it to $A_R(t)$. The following weight removes all interference from the weighted sum $A_R(t) + W_L A_L(t)$.

$$W_L = -\exp(-jkd\cos el_J)$$

In practice, of course, the angle to the interference source is not necessarily known. So the receiver must adapt W_L based on the available measurements. Fortunately, such a criterion exists. The receiver simply adjusts W_L to minimize the total power in the aggregate of received signals. Since the GPS signal is below the natural noise floor, it will not contribute to the total measured power. Noise arrives from all directions, so it will not rule the roost. Interference is the strongest variable contribution to the total measured power. So we choose W_L to minimize the total received power.

With such a selection of W_L, the following signal will result.

$$A_R(t) + W_L A_L(t) \approx \sqrt{P}D(t)x(t)\exp j(2\pi\Delta f_D t + \Delta\theta_S) \times(1 - \exp(jkd(\cos el_S - \cos el_J))) \tag{13.24}$$

The interference is gone, but notice that the signal power has also been modified. The following new term now modifies the amplitude of the signal.

$$1 - \exp(jkd(\cos el_S - \cos el_J)) \tag{13.25}$$

Reasonably, this term predicts that if the satellite signal and interference source are in the same direction, then the satellite signal will be canceled as well. More generally, if $\cos el_S \approx \cos el_J$, then we should expect signal attenuation.

This new amplitude factor is plotted in Figure 13.8 for $el_J = 45°$ and three values of $d = \{0.25\lambda, 0.5\lambda, 0.75\lambda\}$. Figure 13.8 is a polar plot and the length of the vector from the origin to the curve is the amplitude for a satellite at the given angle. The dark solid curve is based on $d = \lambda/4$ and yields the broadest null around the assumed interference direction of $el_J = 45°$. Satellite signals anywhere within the angular neighborhood of the interference source are attenuated. If we choose $d = \lambda/2$, then we get the solid gray curve with a sharper null. If we go as far as $d = 3\lambda/4$, then the null becomes even sharper, but a parasitic null appears over at $el = 135°$.

In practice, adaptive antennas often include seven or nine elements rather than just two. This increase enables sharper nulls. However, these antennas are designed for military applica-

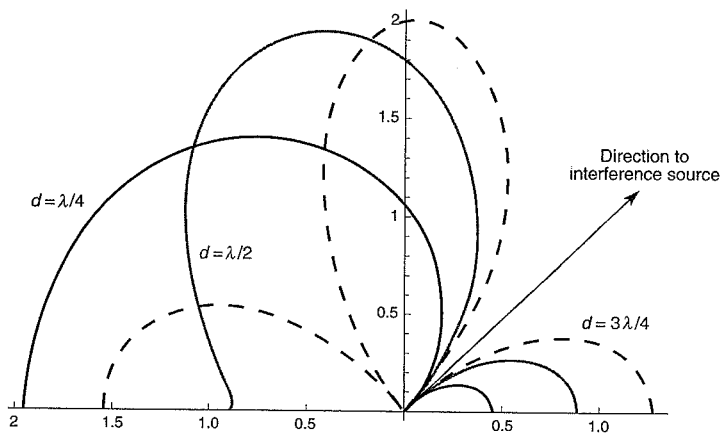


Figure 13.8 Polar plot of Equation (13.25) when the elevation angle of the interference source is 45°.

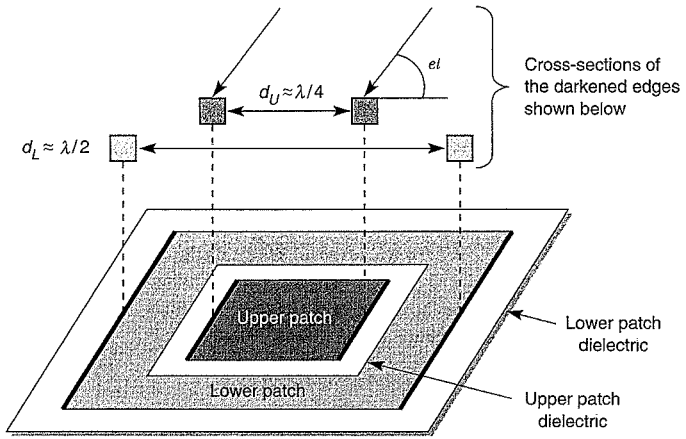


Figure 13.9 Stacked patch antenna.

tions and thus may seek to cancel more than one interference source. In general, an antenna with M elements can produce $M - 1$ nulls. This topic is well developed in Rounds (2004) and the references listed therein. Rather than following this route, we will turn our attention to a special nulling antenna developed for civil aviation.

13.3.2 Ring Nulling Antenna

We now discuss the ring nulling antenna developed by Bauregger (2003). Two views of this antenna are shown in Figure 13.9. As shown in the bottom view, this antenna simply stacks two GPS patch antennas. The outer patch is approximately one-half wavelength along each edge and the inner patch is approximately one-quarter wavelength along each edge. The stacked patches sit on a conducting ground plane, which, in the intended application, is the aircraft fuselage. The lower patch is separated from the fuselage by a thin layer of insulating dielectric. The upper patch is separated from the lower patch by a second layer of insulating dielectric. The lower patch is fed by wires that perforate the lower dielectric. The feeds for the upper patch perforate the lower and upper dielectrics and the lower patch. Notice that this antenna is well suited for aircraft use in one important respect—it is flat and thus offers little air drag.

A one-dimensional model for the stacked patch is shown in the top view. Every edge of a patch antenna can be modeled as a single antenna element, and so the top view models the bold edges shown in the bottom view. In the top view, these edges are perpendicular to the plane of the figure, and we consider a signal arriving from a source which is in the plane of the figure.

The left hand element of the upper patch is $d_L \cos el$ further from the source than the right hand element. Similarly, the left hand element of the lower patch is $d_U \cos el$ further from the source than its mate. No complex weights are applied to the signals from the edges, and so the signal amplitude is multiplied by the following pattern factor.

$$1 + \exp(jkd \cos el) \tag{13.26}$$

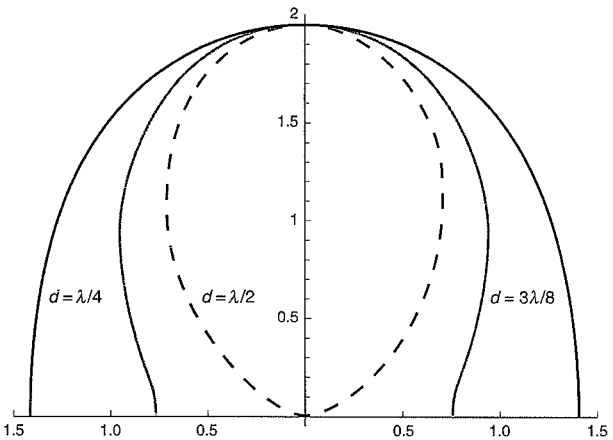


Figure 13.10 Polar plot of Equation (13.26) for a stacked patch antenna.

This factor is plotted in Figure 13.10 for $d = \{\lambda/4, 3\lambda/8, \lambda/2\}$. When $d = \lambda/4$, this amplitude gain is a smooth function of elevation angle with a little more gain at zenith than any other angle. When $d = 3\lambda/8$, the antenna noticeably attenuates signals coming from the horizon, but the overall pattern is still pretty smooth. However when $d = \lambda/2$, the antenna sharply cuts off any signals coming from the horizon. This cutoff makes sense. The signal coming from the horizon will travel exactly $\lambda/2$ from the first edge of the patch to the second edge. The antenna adds the signals from the two edges without any additional phase change or adjustment. Hence, the signal should perfectly cancel because the relative phase is $\lambda/2$ or 180°—the signals are opposite in polarity due to the additional distance traveled. Hence, the $d = \lambda/2$ patch is called a half-wavelength trap or half-wave trap.

RFI is probably coming from below the aircraft and must diffract around the fuselage to reach the sensitive GPS antenna mounted on the top of the airplane. The half-wave trap would greatly attenuate RFI coming from below. However, it also attenuates low-lying satellites. To compensate, Bauregger uses both antennas—he stacks a quarter-wave patch, $d = \lambda/4$, on top of the half-wave trap and switches between the two. When low-lying satellites are required, the quarter-wave patch is used. When RFI is present, the half-wave trap is used.

Needless to say, the design of this antenna involves many complications that we cannot fully explore in this short essay. For example, PIN diodes were used to switch between the two patches. These diodes were carefully placed to minimize coupling between the two antennas. In general, the pattern of any antenna changes when another antenna is placed close by.

In addition, the pattern for any antenna can change dramatically when placed on an aircraft. Figure 13.11 predicts the pattern for the stacked patches mounted on top of a Piper Caravan (a relatively small airplane). In fact, Figure 13.11 is the roll pattern for the Piper Caravan. It shows the signal sensitivity to a source rotated around the airplane's direction of travel. As shown, the pattern for both antennas is relatively smooth in the upper hemisphere. This is as desired to receive GPS satellite signals. For the quarter-wave patch, the signals below the aircraft are attenuated by 10 dB or more. For the half-wave trap, the signals are attenuated by

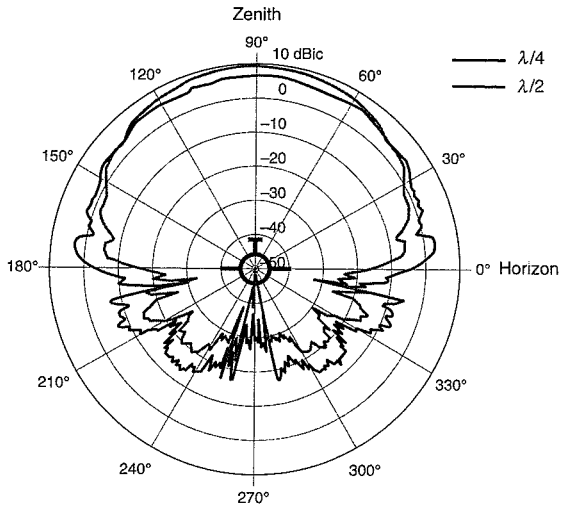


Figure 13.11 Theoretical radiation patterns (courtesy of Frank Bauregger, Novariant, Inc.).

roughly 20 dB. Switching to the half-wave trap patch would bring another 10 dB of RFI rejection.

Figure 13.12 shows flight results for the ring nuller. For these trials, a GPS signal source served as a simulated jammer and was placed near the runway close to the aircraft's touchdown point. The aircraft flew toward the signal source and the figure shows the received power of a satellite signal (PRN1) relative to the power from the jammer. One trace shows the relative power degradation as time lapses and the aircraft moves toward the jammer. This is as expected. The power received from the satellite remains more or less constant, but the signal from the ground is getting stronger. The second trace shows the relative power for the stacked patch antenna. At first, the stacked patch is using the wide-looking patch, and the relative power decreases at more or less the same slope as the standard GPS antenna. At the indicated time, the stacked antenna switches from the quarter-wave patch to the half-wave trap. At this point, the relative power jumps by 20 dB. The ring nuller provides quite a bit of protection from ground-sourced RFI.

13.4 Assisted GPS

Assisted GPS (AGPS) expands GPS coverage downtown and indoors. Fifty percent of the new GPS receivers shipped in 2004 were inside cell phones. In some areas, these phones automatically provide a GPS position fix whenever someone makes an emergency (911) call. With this automation, distraught callers do not have to provide an accurate description of their location. Indeed, they may not be able to. GPS-enabled phones will also serve a family of more lighthearted applications called location-based services or LBS. For example, friends nearby could be announced by your phone. If you allow it, location-specific advertising could be delivered to your phone.

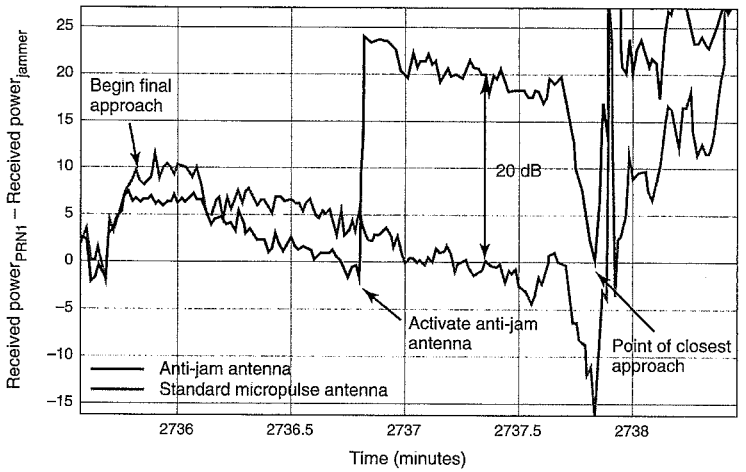


Figure 13.12 Flight results for a ring nulling antenna (courtesy of Frank Bauregger, Novariant, Inc.).

Most cell phone users are in the city or inside buildings. To serve these people, AGPS enables position fixing based on the short segments of noisy data likely to be available in these obstructed environments. Specifically, an assisted GPS receiver can estimate position to within 70 meters or so based on one second of data where the C/N_0 for some of the satellites is only 15 dB-Hz. The assistance comes from a GPS receiver at an electromagnetically quiet site with a clear view of the sky. The assistance data is sent over the cell phone channel to assist the GPS receiver within the receiving cell phone. In this way, a receiver at a controlled site sends valuable data to a roving receiver struggling with an uncontrolled environment.

Architecturally, assisted GPS resembles differential GPS, where reference receivers are connected in real time to roving users [Taylor and Sennott (1984); Enge, Fan, and Tiwari (2001)]. However, AGPS and DGPS have very different objectives. Differential GPS aims to improve the accuracy of the rover's position fix from ten meters to about one meter. In contrast, assisted GPS aims to improve the robustness and coverage of the rover's position fix.

AGPS provides two varieties of assistance. First, AGPS supplants the navigation message broadcast by the satellites. This assistance helps the mobile user cope with breaks in the received signal. Recall that each GPS satellite broadcasts its navigation message at only 50 bits per second, and so 30 seconds are required to send the essential navigation data. Needless to say, a clear 30 second window is not likely in an urban canyon, and so a second source for the navigation data is welcome. This assistance is the subject of Section 13.4.1.

Second, AGPS also enables position fixing in the miserable C/N_0 environments suggested by Table 13.1. In such dismal circumstances, AGPS must aggregate all of the signal energy contained within the received data snippet. These innovations are introduced in Section 13.4.2.

Even though it is effective, AGPS has limitations with respect to accuracy and building penetration. Accuracy better than 70 meters is tough to achieve downtown or indoors due to multipath. Total indoor coverage cannot be expected in all buildings—especially those with steel construction. In these cases, satellite signals cannot do the whole job, and terrestrial sig-

nals are required [Rabinowitz and Spilker (2003), Soliman (2000)]. These terrestrial radio systems may be designed to cover metropolitan areas or one floor at a time.

13.4.1 Supplant the GPS Navigation Message

To combat data breaks, AGPS supplants the orbit and time information usually derived from the navigation message.

AGPS Orbits

A local AGPS reference receiver can certainly extract the Keplerian parameters from the navigation messages for the satellites currently in view. This information can be broadcast to the AGPS rovers over the cell phone channel. Alternatively, a worldwide network of reference receivers can be used to provide so-called long-term orbits. In this case, each satellite is tracked through its entire orbit. The observed orbits are used to adjust parameters in an orbit prediction model, which generates estimates of the future orbits for several days. These predicted orbits suffer error growths of only one meter or so per day. Major cellular network operators use such long-term-orbits as a backup for live orbits. Personal digital assistants (PDA) with embedded GPS use long-term-orbits delivered over the Internet to achieve AGPS performance outside cell phone coverage [Lundgren and van Diggelen (2005)].

AGPS Time

Recall from Section 4.3.6 that the GPS navigation message contains the Handover Word (HOW), which gives the Z-count of the first bit in the current subframe. The receiver needs this time information to build the measured pseudorange. As described in Section 5.1, the receiver builds the whole pseudorange measurement from a number of pieces that vary from coarse to fine. These are: the Z-count associated with the first bit in the current subframe, plus the whole number of subsequent navigation bits, plus the whole number of C/A-codes within the current navigation bit, plus the number of whole C/A-code chips within the current code, plus the fraction of the current chip. If the satellite is occasionally obstructed, then this time-keeping function breaks down for the unassisted user. To be helpful, AGPS needs to develop a time stamp of its own to help the receiver identify the navigation bit. To time tag a navigation bit, AGPS must deliver time with an accuracy of about 5 msec. With this help, the user receiver need only recover the nearby navigation bit edges and count C/A-codes within the current navigation bit.

Time is also needed to interpret the Keplerian parameters that describe the satellite orbit and location within that orbit. These parameters typically change hourly. However, the mapping from the Keplerian parameters to the current satellite position requires accurate time because the satellites are moving at about 3860 m/s along their orbits. When the ephemerides are sent by the satellite, the Z-count is used to establish how far along the orbit the satellites have moved. Absent the Z-count, AGPS must provide time to enable this mapping. As discussed in Homework Problem 13-4, the required time accuracy is 10 ms.

Some cell phone networks are synchronized to GPS and thus have a sharp sense of time. In these cases, the AGPS time stamps are simply broadcast over the cell phone links. Unfortunately, most cellular phone networks do not know time to better than one second accuracy. In one second, the range to a rising satellite will decrease by up to 800 meters, and the range to a setting satellite will increase by a similar amount. If a one second timing error is ignored,

the average position error will be roughly 400 m and larger with bad satellite geometry. This time error yields range errors that are different for each satellite. However, these range errors are deterministic functions of the underlying error in the time stamp. Hence, the time stamp error can be treated as one extra state in the vector of navigation estimanda. The augmented state comprises three spatial unknowns, the common time bias, and the time stamp error [van Diggelen (2002)].

13.4.2 Support Two-Second Integration Times

To cope with low signal-to-noise ratios, AGPS handsets need to aggregate all of the signal energy contained in the received data snippet. AGPS base stations send data to support this integration.

As we know, the receiver must estimate the code shift, τ, and the Doppler frequency, f_D. To do so, the receiver correlates the received signal with a set of replica signals, where each replica is used to measure the likelihood of one possible combination of τ and f_D. This process is described in Section 11.3.5 and Figure 11.21. Without assistance, the replicas must span the area shown in Figure 11.18, which shows that the signal may fall at any code shift between 0 and 1023 chips and that the Doppler offset will fall somewhere between ±6500 Hz. The overall search area is broken into cells for each possible $(\hat{\tau}, \hat{f}_D)$. The correlation from cell $(\hat{\tau}, \hat{f}_D)$ measures the likelihood that the received signal has that particular code shift and Doppler frequency.

When C/N_0 is strong, a total integration time, KT_{CO}, of only 10 ms is required for reasonable estimation of τ and f_D. Indeed, Figure 11.20 shows a pairwise probability of acquisition failure of 10^{-8} when C/N_0 = 36 dB-Hz. AGPS supports a total integration time of up to two seconds to cope with the low values of C/N_0 reported in Table 13.1. However, the coherent integration time, T_{CO}, and noncoherent integration, K, are subject to certain limitations. The coherent integration time is limited by navigation bit reversals and the width of the frequency bin shown in Figure 11.18. Noncoherent integration is limited by changes in the signal frequency. We treat these topics in turn.

Coherent Integration Time

Recall (11.21) that gives the coherent signal correlation.

$$\tilde{S}(\Delta\tau_n, \Delta f_{D,n}) = \frac{\sqrt{C}D}{T_{CO}} \exp(j2\pi\Delta\theta) \int_0^{T_{CO}} x(t-\tau)x(t-\hat{\tau}_n)\exp(j2\pi\Delta f_{D,n}t)\,dt \tag{13.27}$$

Navigation data bits, D, must be provided to extend the coherent averaging time T_{CO} beyond 20 milliseconds. Left untreated, data bit reversals would wash out the coherent average. This particular danger is shown in Figure 13.13 [D. Akos *et al.* (2000)]. As shown, the coherent average over time interval 1 is untroubled by any navigation bit reversals, and so the correlation process results in a distinct peak that unambiguously identifies the true code shift. In contrast, a navigation bit reversal occurs inside time interval 2, and the correlation peak deteriorates. If a bit reversal occurs in the middle of the averaging interval, then the average after the bit reversal is subtracted from the average before the bit reversal, and the signal energy will not accumulate in the desired fashion.

To treat this problem, the navigation bits are needed to wipe off the navigation message

from the received GPS signal. The navigation bits can be obtained from at least three sources. First, the cell base station can broadcast them. Second, the receiver can predict them based on past navigation messages [S. Soliman *et al.* (2000)]. After all, the navigation message is quite repetitive, with only occasional changes in many of the message fields. Third, the receiver can guess at all possible combinations of navigation bits. In any event, these navigation bits must be correctly shifted in time to within one or two milliseconds. Otherwise, they will not fall on top of the received navigation bits. This registration can be based on time provided by the network or navigation bit timing recovery (BTR) by the receiver. In either case, time must be recovered to within approximately two milliseconds.

As discussed in Section 11.3.5, the coherent integration time is also limited by the size of the frequency bins, $f_{D,\text{bin}}$, used in the search area shown in Figure 11.18. If $\Delta\theta$ in (13.27) varies by more than a quarter cycle, the inphase signal correlation will rotate into the quadrature component, and the quadrature component will rotate into the inphase component. If $\Delta\theta$ varies by more than one half cycle, both components will suffer sign reversals. These sign reversals will have the same effect as sign reversals introduced by bit transitions in the navigation message—they will wash out the coherent average.

The phase, $\Delta\theta$, will vary most quickly when the true signal frequency is halfway between $\hat{f}_{D,n}$ and $\hat{f}_{D,n+1}$. If we denote $f_{D,\text{bin}} = \hat{f}_{D,n+1} - \hat{f}_{D,n}$, then we may write

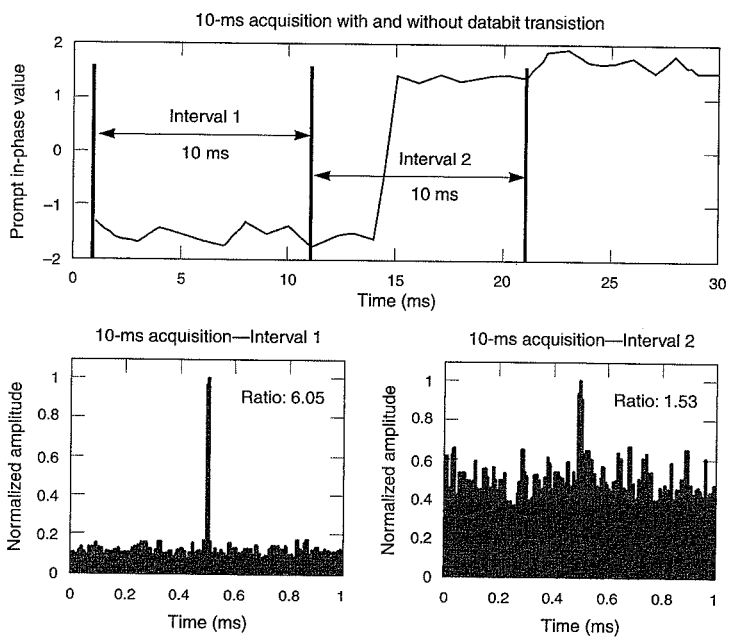


Figure 13.13 Impact of navigation bit reversals on coherent averaging (from D. Akos *et al.* (2000)).

$$\Delta\theta(t) = \frac{f_{D,\text{bin}}\, t}{2} \tag{13.28}$$

This equation simply captures the phase change for a frequency offset that is midway between two frequencies in the search grid. If this phase drift accumulates for one coherent integration time, T_{CO}, then the total phase change will be

$$\Delta\theta(T_{CO}) = \frac{f_{D,\text{bin}} T_{CO}}{2} \tag{13.29}$$

We desire this phase change to be less than one quarter of a cycle over the coherent integration time. Otherwise, the inphase and quadrature correlations will begin to average toward zero. This concern dictates the following constraint.

$$\frac{f_{D,\text{bin}} T_{CO}}{2} < \frac{1}{4}$$

$$f_{D,\text{bin}} < \frac{1}{2T_{CO}} \tag{13.30}$$

As shown, the coherent integration time and frequency bin size have an inverse relationship. Long coherent integration times dictate small bins. Typical values of T_{CO} are given in Table 13.2 along with the corresponding values of $f_{D,\text{bin}}$.

Noncoherent Integration

Recall from (11.38) and (11.42) that noncoherent integration is implemented as follows.

$$L_1 = \sum_{k=1}^{K} \left| \tilde{S}(\Delta\tau_1, \Delta f_{1,D}) + \tilde{\eta}_{1,k} \right|^2$$

$$= \sum_{k=1}^{K} \left| \sqrt{C} D \exp(j\Delta\theta_1) \tilde{R}(\Delta\tau_1, \Delta f_{1,D}) + \tilde{\eta}_{1,k} \right|^2 \tag{13.31}$$

The noncoherent test metric is formed by taking the magnitude squared of K coherent averages. The K squares are then summed. This yields a total integration time of KT_{CO} seconds.

K can be chosen without worrying about the phase, $\Delta\theta$, because the magnitude operation shown in (13.31) has stripped this dependence. However, frequency drift limits KT_{CO} [Chansarkar and Garin (2000)]. If the signal moves out of bin $(\hat{\tau}, \hat{f}_D)$ into another bin, then no amount of coherent or noncoherent averaging will be helpful. This constraint can be articulated as follows.

$$\dot{f}_D K T_{CO} < \frac{f_{D,\text{bin}}}{2} \tag{13.32}$$

In this equation, $\dot{f}_D$ is the rate of change of the Doppler frequency or so-called Doppler rate. This inequality yields

$$K < \frac{f_{D,\text{bin}}}{2\dot{f}_D T_{CO}} \tag{13.33}$$

Table 13.2 Coherent and noncoherent integration times for assisted GPS for Doppler rates of 5 Hz/s and 0.5 Hz/s*

Coherent integration time (ms) T_{CO}	*Approximate Doppler bin width (Hz)* $f_{D,\text{bin}} = \frac{1}{2T_{CO}}$	*Maximum Doppler rate (Hz/s)* f_D	*Number of noncoherent terms* $K = \frac{f_{D,\text{bin}}}{2T_{CO} f_D}$	*Total integration time (sec)* KT_{CO}
20	25	5	125	2.5
50	10	5	20	1.0
100	5	5	5	0.5
200	2.5	5	1	0.2
200	2.5	0.5	12	2.4
500	1.0	0.5	2	1.0

* This table assumes that the maximum tolerable phase change during the coherent integration time, T_{CO}, is 0.25 cycles, and that the maximum tolerable frequency change during the total integration time, KT_{CO}, is 0.5 frequency bins. The table does not report parameter combinations that yield total integration times greater than 2.5 seconds because longer data windows may be unlikely in urban environments.

Unlike T_{CO}, K is proportional to $f_{D,\text{bin}}$. Long coherent integration times favor small frequency bins, whereas large K favors large frequency bins.

To choose K, we must estimate the Doppler rate, f_D. Table 13.3 shows typical values for Doppler and Doppler rate. As shown, Doppler and Doppler rate can be attributed to the satellite motion, satellite clock, user clock and the user platform motion. Satellite motion introduces a velocity and acceleration along the user's line of sight to the satellite. Homework Problem 13-5 asks you to verify the following results from the table: the maximum Doppler and Doppler rate due to satellite motion are approximately 5000 Hz and 1 Hz/s, respectively.

The satellite Doppler is certainly large, but it is readily predicted from the satellite ephemeris. Indeed, the AGPS base station may predict the Doppler and Doppler rate and send this information to the rover. The only significant uncertainties that persist are due to initial uncertainties in user position. If the user position is known to within ±10 kilometers, then the residual error in satellite Doppler is only ±8 hertz. This maximum error occurs for a satellite at zenith. Homework Problem 13-6 asks you to prove this result.

User motion also contributes a Doppler offset and Doppler rate. Consider a slow-moving user with a velocity of only 10 knots and acceleration of 0.1 g. This modest velocity corresponds to 5 m/s, which gives a Doppler offset of 25 Hz; and the low acceleration yields a Doppler rate of 5 Hz/s. A fast-moving user may have a velocity of 100 knots and an acceleration of 1 g, which yields a Doppler offset of 250 Hz and Doppler rate of 50 Hz/s. We ask you to substantiate these findings in Homework Problem 13-7. Clearly, AGPS favors the slow-moving user—like a strolling cell phone user. Some AGPS researchers advocate the use of inexpensive inertial sensors to estimate the user platform motion and thus increase K.

Table 13.3 Doppler and Doppler rates due to satellite and user platform motion, satellite clock offset and drift, and user clock offset and drift

	Line-of-sight velocity (m/s)	*Doppler shift (Hz)*	*Line-of-sight acceleration (m/s^2)*	*Doppler rate (Hz/s)*
Satellite motion	800	5000	0.175	1
User platform motion with low dynamics: 10 knots and 0.1 g acceleration	5	25	1	5
User platform motion with moderate dynamics: 100 knots and 1g acceleration	50	250	10	50
GPS satellite clock offset and drift		<<1		<<1
User clock offset and drift for a free-running crystal oscillator		1575		0.5
User clock offset and drift when the user clock is derived from a CDMA cell signal synchronized to GPS		<<1		<<1
User clock offset and drift when the user clock is derived from a non-CDMA cell signal that is temperature controlled, but not synchronized to GPS		15		<<1

As shown in Table 13.3, the satellite clock is very stable and does not contribute meaningfully to the overall Doppler or Doppler rate budget. However, the user clock may be worrisome. If the mobile user has a free running clock, then the frequency stability may only be 1 part per million, $\Delta f/f \approx 10^{-6}$. The resulting Doppler shift is equal to the carrier frequency multiplied by the stability, which is approximately 1575 hertz. The frequency drift for the inexpensive oscillators used in cell phones can be 0.5 Hz/s or more depending on the temperature variations encountered by the roving user.

Fortunately, the frequency of the user clock can be locked to the received carrier signal from the cell base station [Taylor and Sennott (1984)]. In this case, the user clock performance approximates that of the more stable clock at the base station. In fact, cell phone networks that use code division multiple access (CDMA) are synchronized to GPS time. If the user oscillator is synchronized to the CDMA cell signal, then it will have negligible offset and drift. Most non-CDMA networks are not synchronized to GPS, but the stability of the cell site clock is still quite good. Moreover, the cell base stations are temperature controlled. In this case, the

user clock Doppler offset will be approximately 15 Hz, and the Doppler rate should be quite small.

Probability of Acquisition Failure

Table 13.2 shows reasonable values of T_{CO} and K based on (13.30) and (13.33). Two sets of results are shown. One set corresponds to an unknown Doppler rate of 5 Hz/s and the second assumes that the Doppler rate is estimated or constrained in some fashion to 0.5 Hz/s. Within these two sets, large values of T_{CO} yield small values of KT_{CO} because coherent averaging sharply curtails the size of the frequency bin, $f_{D,\text{bin}}$. If the frequency bin is small, then KT_{CO} must be small—otherwise the signal can readily move from one bin to the next.

Figures 13.14 and 13.15 show the pairwise failure probability for our two cases: 5 Hz/s and 0.5 Hz/s. As shown, the longer averaging time afforded by AGPS certainly improves performance relative to Figure 11.20, which used a total integration time of 10 ms. Indeed, the C/N_0 required for a probability of pairwise failure of 10^{-8} drops from around 36 dB-Hz to 14–18 dB-Hz, depending on how T_{CO} and K are chosen. Amongst the AGPS results, the smaller Doppler rate uncertainty gives noticeably better performance. In both cases, performance is best for the longest possible total integration time, KT_{CO}, rather than the longest possible coherent integration time, T_{CO}.

Beware—the probabilities in Figures 13.14 and 13.15 are only pairwise error probabilities. They must be multiplied by M to bound the probability of acquisition failure. Recall that M is the number of cells in the search grid over $(\hat{\tau}, \hat{f}_D)$. Hence, small values of $f_{D,\text{bin}}$ will have larger values of M. For example, if we use $f_{D,\text{bin}} = 25$ Hz and our total Doppler uncertainty is 250 Hz, then $M = 2 \times 1023 \times 100 = 204{,}600$. If we reduce the Doppler uncertainly to 25 Hz, then $M = 20{,}460$. For this reason, AGPS sends estimates of the Doppler shifts so the receiver can focus its efforts on the subset of cells that contain the signal. Such side information is welcome because the receiver can cut down the number of test cells and focus on the likely

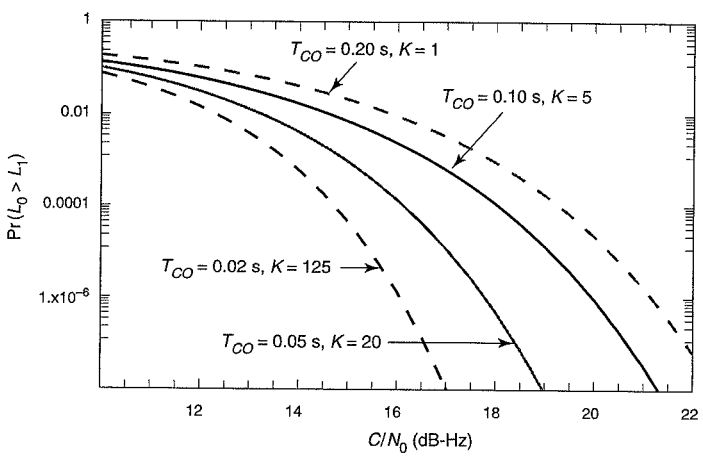


Figure 13.14 Pairwise probability of acquisition failure for coherent and noncoherent averaging times used by assisted GPS. The parameters shown assume a residual Doppler rate of 5 Hz/sec.

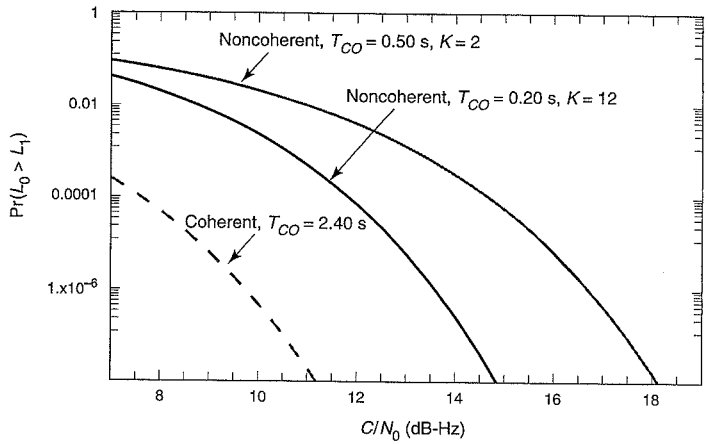


Figure 13.15 Pairwise probability of acquisition failure for coherent and noncoherent averaging times used by assisted GPS. The parameters shown assume a residual Doppler rate of 0.5 Hz/sec.

cells. Even with this side information, a fast algorithm or special correlation hardware is still required to execute all of these tests, and many AGPS innovations are directed at providing this capability either in hardware or software [Agarwal *et al.* (2002), Akos *et al.* (2000), Chansarkar and Garin (2000), Moeglin and Krasner (1999), and van Diggelen (2001)].

13.5 Inertial Aiding

13.5.1 Basics

Inertial sensors are akin to the combination of a speedometer and compass. As shown in Figure 13.16, a speedometer provides a measurement of velocity in the so-called body frame. In other words, the speedometer measures velocity in the direction that the vehicle is pointed. A compass is needed to determine velocity in an external navigation frame. This action is shown in Figure 13.17, which is a dead reckoning navigation computer for a speedometer and compass. As shown, the speedometer measurements are resolved into the navigation frame using the heading measurements from the compass.

Once resolved, we have

$$\begin{aligned} \frac{dN}{dt} &= S(t)\cos\theta(t) \\ \frac{dE}{dt} &= S(t)\sin\theta(t) \end{aligned} \tag{13.34}$$

This equation denotes the speedometer measurements as $S(t)$ and the compass measurements as $\theta(t)$. Please do not confuse this use of $\theta(t)$ for direction with our previous use for carrier phase. The operation shown in (13.34) estimates our velocity in the north and east directions.

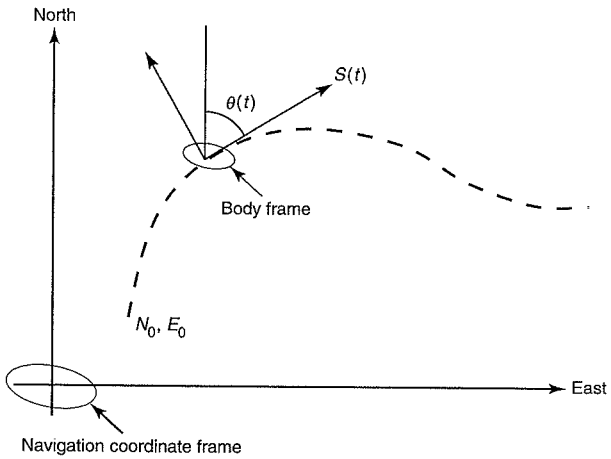


Figure 13.16 Speedometer and compass measurements. Speedometer measurements are taken in the body frame, and the compass measurements are used to rotate those measurements into the navigation frame.

We now seek our north position, $N(t)$, and our east position, $E(t)$. To this end, we implement the following equations.

$$N(t) = \int_0^t S(\tau)\cos\theta(\tau)\,d\tau + N_0$$
$$E(t) = \int_0^t S(\tau)\sin\theta(\tau)\,d\tau + E_0 \qquad (13.35)$$

We integrate our north velocity to obtain north position. Of course, we must add the integration constant corresponding to our estimated starting position in north. The same basic procedure is used to estimate our position in east.

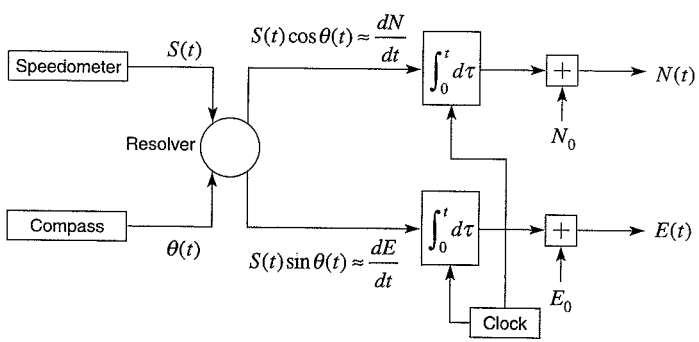


Figure 13.17 Block diagram showing integration of speed and compass measurements.

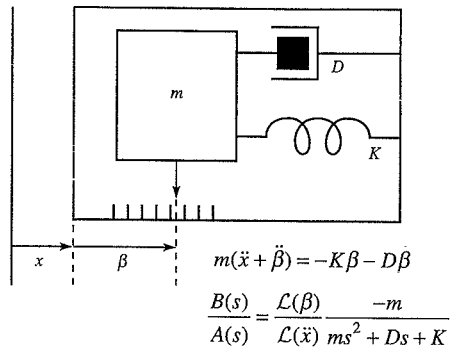


Figure 13.18 Notional accelerometer.

Inertial sensors also dead reckon, based on measurements made by accelerometers and gyroscopes. A simple accelerometer is shown in Figure 13.18. As shown, it places a proof mass inside a case. Consider the x-axis only, and assume that this axis is locally level. In other words, assume that x is perpendicular to the gravity vector. In this case, we may write the following based on Newton's second law.

$$m(\ddot{x} + \ddot{\beta}) = -D\dot{\beta} - K\beta$$
$$m\ddot{x} = -m\ddot{\beta} - D\dot{\beta} - K\beta \qquad (13.36)$$

The mass of the proof mass is m, the spring constant is K, and the damping coefficient is D. The position of the case in inertial space is x. The position of the proof mass relative to the case is β. Hence, the position and acceleration of the proof mass in inertial space is $x+\beta$ and $\ddot{x}+\ddot{\beta}$, respectively.

When the proof mass stops moving relative to the case

$$\beta = \ddot{\beta} = 0$$
$$m\ddot{x} = -K\beta \qquad (13.37)$$

For convenience we set $m = 1$ and denote $a = \ddot{x}$. With these conventions, we may write

$$a = -K\beta \qquad (13.38)$$

As shown, accelerometers estimate acceleration based on a measurement of the restoring force, $-K\beta$, needed to keep the proof mass stationary. This restoring force is proportional to the displacement, β, shown in Figure 13.18.

If the x-axis is not level, then our measurement will include a component due to gravity, g. If the tilt relative to the local horizontal is ϕ, then

$$m(\ddot{x} + \ddot{\beta}) = -D\dot{\beta} - K\beta - mg\sin\phi$$
$$m\ddot{x} = -m\ddot{\beta} - D\dot{\beta} - K\beta - mg\sin\phi \qquad (13.39)$$

When the proof mass stops moving relative to the case

$$m\ddot{x} = -K\beta - mg\sin\phi$$
$$= f - mg\sin\phi \tag{13.40}$$

Once again, we set $m = 1$ and $a = \ddot{x}$.

$$a = -K\beta - g\sin\phi \tag{13.41}$$

As shown, accelerometers don't really measure acceleration; their measurement includes gravity, and so their measurement is usually referred to as specific force. The gravity components need to be removed to estimate the vehicle's acceleration. Consequently, the inertial navigation community has devoted a great deal of effort to accurately estimate the worldwide gravity field.

Once corrected, the accelerations must be resolved into a known frame. This problem is akin to resolving the speedometer measurements into the navigation frame. Gyroscopes serve the purpose of the compass mentioned earlier in this section. However, these devices usually measure angular rate rather than angle. Hence these angle rate measurements must also be integrated before the acceleration measurements can be resolved into a known frame. Once resolved, the acceleration measurements are integrated twice, rather than once, to yield estimates of position.

Historically, inertial navigation systems have used two basic approaches to carry out this coordinate transformation. A gimballed implementation mechanically maintains the orientation of a platform. The accelerometers and gyros retain a fixed attitude in spite of changes in the vehicle attitude. In contrast, a strapdown implementation measures the accelerations and angular rates in the vehicle frame. These measurements are mathematically rotated into the navigation frame in a computer algorithm. In this case, the sensors do not move relative to the vehicle; they are strapped down.

When inertial navigation systems were first developed, digital computers were not available and computation was difficult. Furthermore, early gyroscopes could not cope with high turning rates. The gimballed platform reduced the need for complex coordinate transformations, and reduced the instrument turning rates relative to the vehicle turning rates. Today, computation is easy and gyroscopes can tolerate high dynamics without large performance penalties. Thus, strapdown mechanizations are generally preferred. Relative to gimballed mechanizations, they are small, light and mechanically simple.

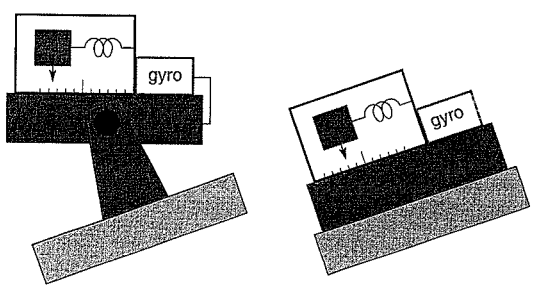


Figure 13.19 Gimballed and strapdown mechanizations of a one-dimensional inertial navigation system.

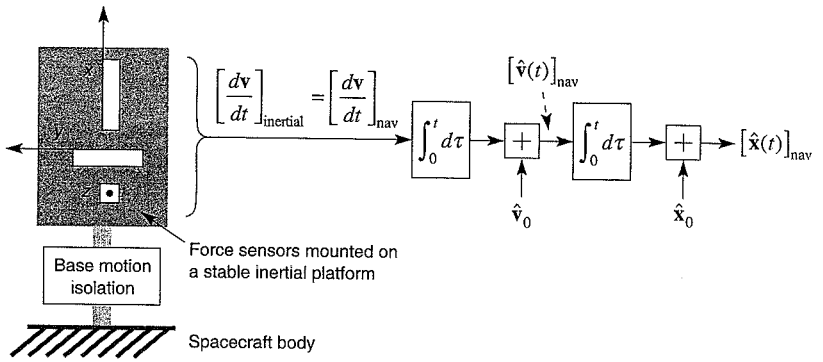


Figure 13.20 Gimballed mechanization of position and velocity estimation based on inertial measurements in remote space. No gravity compensation is required and the navigation frame is the inertial frame. The force sensors are mounted on a gyro-stabilized platform, which is always oriented in the inertial frame. Consequently, the coordinate conversion matrix is the identity matrix.

A gimballed implementation in one dimension is shown on the left hand side of Figure 13.19. A gyro is mounted with the accelerometer on the gimballed platform, and the platform tilt is controlled by a motor. The gyro measures attitude or attitude rate, and the motor acts to null any output from the gyro. This action stabilizes the platform in the desired frame. A gimballed accelerometer triad is shown in Figure 13.20. In this idealized case with no gravity, a triad of gyros stabilize the platform in three dimensions. Once stabilized, the measurements can be integrated twice to yield position in the navigation frame.

$$\mathbf{v}(t) = \int_0^t \mathbf{a}(\tau)\,d\tau + \mathbf{v}_0$$
$$\mathbf{x}(t) = \int_0^t \mathbf{v}(\tau)\,d\tau + \mathbf{x}_0 \tag{13.42}$$

In this equation, $\mathbf{a}$ is the triad of acceleration measurements, $\mathbf{v}$ is the corresponding velocity vector, and $\mathbf{x}$ is the corresponding position estimate in three dimensions. Notice the similarity to (13.35). As before, the initial conditions must be well known. In this case, we need both $\mathbf{v}_0$ and $\mathbf{x}_0$.

A strapdown implementation is shown on the right of Figure 13.19 and in Figure 13.21. A gyro measures angular rates. These measurements are used to mathematically rotate the measurement coordinates into the desired frame in the navigation computer. Once rotated, we proceed to integrate twice, as shown in Figure 13.21.

Figure 13.22 shows two of the embellishments that an earthbound inertial navigation system (INS) requires relative to Figures 13.20 and 13.21. For terrestrial navigation, the INS must model and remove the effect of gravity. Recall that the accelerometer really measures specific force, and so the effect of gravity must be removed. In addition, the Coriolis effect must be removed for a terrestrial INS.

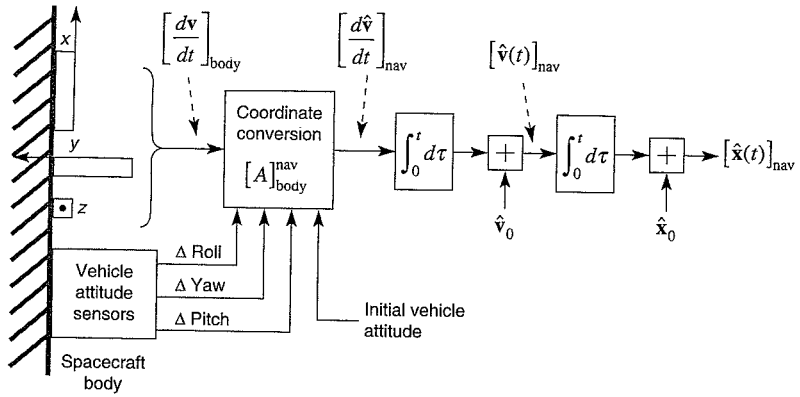


Figure 13.21 Strapdown mechanization of position and velocity estimation based on inertial measurements in remote space. No gravity compensation is required and the navigation frame is the inertial frame. The force sensors are mounted to the spacecraft, so vehicle attitude sensors are required to transform the force measurements into the inertial (navigation) frame.

13.5.2 Gyros and the Sagnac Effect

Many technologies have been used to measure angular rates including mechanical spinning mass gyros, fiber optic gyros (FOG), ring laser gyros (RLG), vibrating quartz rings, and vibrating tuning forks. The RLG is the predominant technology used today by airliners and the military.

The RLG and FOG employ the Sagnac effect to measure angular rate. Two light beams travel in opposite directions around a closed path. The RLG employs mirrors to establish the closed path. As shown in Figure 13.23, the FOG launches two laser beams inside a coil of optical fibre. The solid wave travels clockwise and the dashed wave travels counter-clockwise. Both travel around the coil until they reach a phase comparator that measures the phase of one

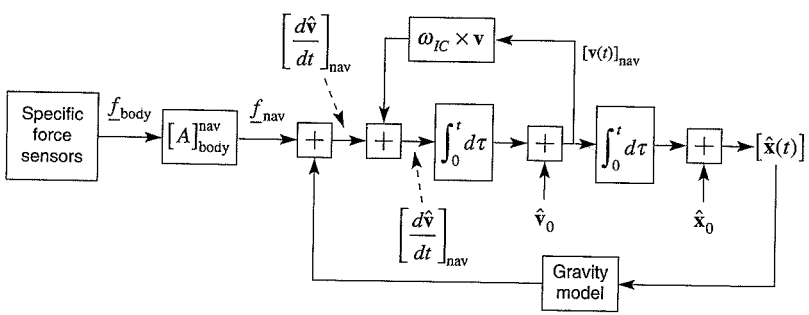


Figure 13.22 Position and velocity estimation based on inertial measurements on earth. The inertial navigation system must compensate for gravity and the Coriolis effect.

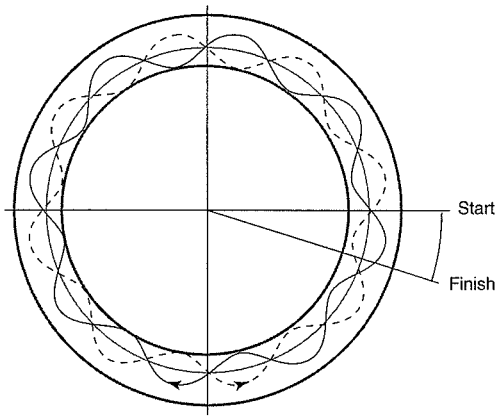


Figure 13.23 Fiber optic gyros (FOG) and the Sagnac effect. The FOG contains two counter-rotating light waves, with the solid wave going clockwise, and the dashed wave going counter-clockwise. While the waves are traveling, the detector moves from start to finish, and so the solid wave makes a longer journey than the dashed wave.

wave relative to the other. The phase comparator moves from start to finish while the wave is moving around the circle.

Assume the FOG is rotating in the plane shown with an angular rate of ω radians/sec. If the coil is rotating clockwise, then the counter-clockwise wave will travel a shorter path. In fact, this wave will travel $(2\pi - \omega t_1)R$ meters before it reaches the phase comparator. In this equation, R is the radius of the coil and t_1 is the travel time for the clockwise wave. The clockwise wave will have to travel a longer path. It will travel $(2\pi + \omega t_2)R$ meters before it reaches the phase comparator.

The travel times are equal to the path length divided by the speed of light, c.

$$t_1 = \frac{(2\pi - \omega t_1)R}{c}$$
$$t_2 = \frac{(2\pi + \omega t_2)R}{c} \tag{13.43}$$

These equations can be solved for the travel times.

$$t_1 = \frac{2\pi R}{c + \omega R}$$
$$t_2 = \frac{2\pi R}{c - \omega R} \tag{13.44}$$

The difference in their travel times is

$$\Delta t = \frac{4\pi\omega R^2}{c^2 - \omega^2 R^2} \text{ seconds} \tag{13.45}$$

The phase comparator will measure the corresponding phase difference

$$\Delta\phi = \frac{8\pi^2 f\omega R^2}{c^2 - \omega^2 R^2} \text{ radians} \tag{13.46}$$

where f is the frequency of the light wave.

This expression can be linearized since ωR is very small compared to c.

$$\begin{aligned}\Delta\phi(\omega) &\approx \Delta\phi(0) + \left.\frac{\delta\Delta\phi}{\delta\omega}\right|_{\omega=0}\omega \\ &= \frac{8\pi^2 f R^2 \omega}{c^2} \\ &= \frac{8\pi^2 R^2 \omega}{c\lambda} \text{ radians}\end{aligned} \tag{13.47}$$

Usually, a FOG contains many windings of the fibre optic cable to make $\Delta\phi$ more sensitive to ω. If n is the number of windings, then

$$\Delta\phi \approx \frac{8\pi^2 n R^2 \omega}{c\lambda} \text{ radians} \tag{13.48}$$

In 2005, RLGs sell for tens of thousands of dollars, and FOGs sell for thousands of dollars. As such, they are not used in automobiles or inexpensive airplanes. These markets are served by quartz solid state devices that contain vibrating elements that deflect during rotation. The deflections are measured and used to estimate angular rate. As we shall discover, accuracy improves with price. Before GPS, inertial systems were used to provide navigation for long periods of time. With GPS, their role has shifted. Many are used to cover short term GPS outages, and these applications are well served by the less expensive FOGs and solid state inertial sensors.

13.5.3 Combining GPS and Inertial Measurements

GPS and INS are complementary. The INS is self-contained and GPS is externally referenced. Since INS does not require external signals, it is not vulnerable to RFI. On the other hand, the INS error grows with time because it has no external anchors. We shall explore this error growth in the next two sections. In contrast, GPS is vulnerable to RFI, but does not have error growth. The fundamental synergy is clear. The INS enables the integrated receiver to coast through RFI outages. When available, GPS removes the error growth and calibrates the inertial sensors.

As a bonus, an INS can usually output position fixes at a much higher rate than a GPS receiver. Few GPS receivers output position estimates more quickly than ten per second. In contrast, an INS may output a hundred position estimates per second. Thus, an INS can also fill in between GPS measurements when high output rates are required.

Many GPS/INS receiver architectures have been developed to respond to this attractive match between GPS and INS. Figure 13.24 is a top-level summary. Figure 13.24(a) shows the simplest architecture, which fuses the two systems in the position domain. When GPS is available, it contributes to the position estimate and calibrates the INS. When GPS is not available, the INS is used to coast through any outages suffered by GPS.

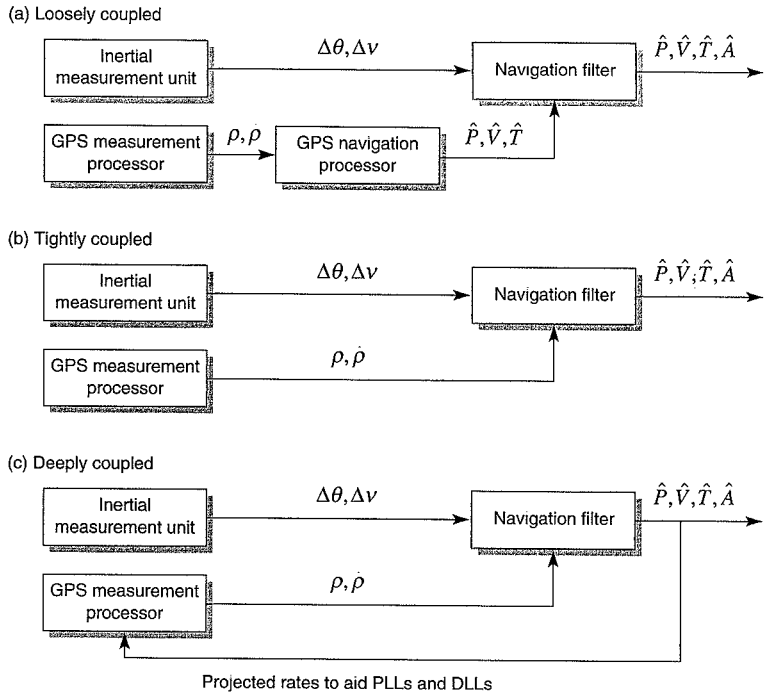


Figure 13.24 Inertial measurements to aid GPS tracking (courtesy of Dr. Jennifer Gautier, Stanford University).

Figure 13.24(b) shows a slightly more sophisticated architecture where GPS provides pseudorange measurements rather than position estimates to the joint navigation filter. When available, GPS contributes to the position estimate and calibrates the INS. If some of the GPS measurements are not available, then the surviving GPS pseudorange measurements still contribute to the position estimate and help to calibrate the INS. In this way, the strong satellite signals are still helpful even if weak ones have disappeared.

Figure 13.24(c) shows the most sophisticated architecture. As usual, GPS contributes to the position estimate and calibrates the INS when satellite signals are available. If GPS is struggling, then the combined position fix is used to aid the signal tracking by the GPS loops. The joint position and velocity estimates are used to predict the velocities along the lines of sight to the satellites. These predicted velocities are used to rate-aid the delay lock loop or even the phase lock loop. In other words, the INS helps the GPS tracking loops fend off RFI by enabling the use of narrower tracking bandwidths.

All of the architectures described by Figure 13.24 have been treated extensively in the literature [Albans (2004), Gautier (2003), Gebre (2001), Greenspan (1996), Soloviev *et al.* (2004)]. We will not repeat these elegant discussions here. Rather, our goal is to provide a basic, but non-trivial, INS error analysis. By so doing, we hope to make the more complicated

literature accessible to the reader. The analysis in the next two sections is excerpted from lecture notes written by Professor J. David Powell of Stanford University.

13.5.4 Error Growth in One Dimension without Tilt

We consider a one-dimensional positioning problem. Our true position, velocity and acceleration are x, $v = \dot{x}$, and $a = \ddot{x}$, respectively. Our true initial position and velocity are x_0 and v_0. For simplicity, we set $m = 1$.

Recall that our accelerometer measures the restoring force, $-K\beta$, to estimate acceleration. Sadly, this measurement is not perfect and suffers the following errors.

$$\begin{aligned}\delta a(t) &= \hat{a}(t) - a(t) \\ &= ka(t) + b + \left(\eta_{A,W}(t) + \eta_{A,GM}(t)\right)\end{aligned} \tag{13.49}$$

The true acceleration is $a(t)$ and the measured acceleration is $\hat{a}(t)$. As shown, the error, $\delta a(t)$, includes a term that is proportional to the acceleration, $ka(t)$. Please do not confuse this constant of proportionality, k, for the spring constant, K, in Figure 13.18, or the wave number in Section 13.2. Our measurement error also includes a bias, b, white noise, $\eta_{A,W}$, and Gauss-Markov noise, $\eta_{A,GM}$. We will have more to say about all of these error terms in what follows.

Notice that we do not include any tilt error in our error model. In other words, we assume that the accelerometer really is oriented along the true x-axis and so we do not have large errors due to the gravity field. This assumption distinguishes this section from the next, where we will include the tilt error resulting from imperfect gyro measurements.

The acceleration measurements are integrated to yield estimates of velocity, $\hat{v}(t)$. The resulting velocity estimates are integrated to estimate position, $\hat{x}(t)$.

$$\begin{aligned}\hat{v}(t) &= \int_0^t \hat{a}(\tau)d\tau + \hat{v}_0 \\ \hat{x}(t) &= \int_0^t \hat{v}(\tau)d\tau + \hat{x}_0\end{aligned} \tag{13.50}$$

Our error analysis simply propagates the errors contained in (13.49) through the estimator equations, (13.50). We proceed as follows.

$$\begin{aligned}\delta v &= \hat{v}(t) - v(t) \\ &= \int_0^t \delta a(\tau)d\tau + \delta v_0 \\ &= \int_0^t ka(\tau) + b + \left(\eta_{A,W}(\tau) + \eta_{A,GM}(\tau)\right)d\tau + \delta v_0 \\ &= k(v(t) - v_0) + bt + \int_0^t \left(\eta_{A,W}(\tau) + \eta_{A,GM}(\tau)\right)d\tau + \delta v_0\end{aligned} \tag{13.51}$$

In these equations, $\delta v_0 = \hat{v}_0 - v_0$ gives the error in the estimate of the initial velocity.

We continue this procedure to find the error in position.

$$\begin{aligned}\delta x &= \hat{x}(t) - x(t) \\ &= \int_0^t \delta v(\tau)d\tau + \delta x_0 \\ &= k(x(t) - x_0) - kv_0 t + \frac{bt^2}{2} + \int_0^t \int_0^t \left(\eta_{A,W}(\tau) + \eta_{A,GM}(\tau)\right)d\tau\, ds + \delta v_0 t + \delta x_0\end{aligned} \tag{13.52}$$

Consider the following example. An inertial navigation system must provide an accuracy of one kilometer at the end of a one-hour mission. It is providing guidance for a vehicle that travels 500 km in the hour. For this example, the tolerable error in the initial velocity, δv_0, may be estimated as follows.

$$\begin{aligned}\delta v_0 t &< 1000 \text{ m} \\ \delta v_0 &< 0.28 \text{ m/s}\end{aligned} \tag{13.53}$$

The tolerable error in the scale factor may be estimated as follows.

$$\begin{aligned}k(x(t) - x_0) &< 1000 \text{ m} \\ k &< \frac{1000}{x(t) - x_0} \\ &= \frac{1}{500}\end{aligned} \tag{13.54}$$

The tolerable error for the bias may be estimated as follows.

$$\begin{aligned}\frac{bt^2}{2} &< 1000 \text{ m} \\ b &< \frac{2000}{t^2} \\ &= 1.5 \times 10^{-4} \text{m/s}^2\end{aligned} \tag{13.55}$$

Notice that our bias tolerance is inversely proportional to the mission duration squared, t^2.

The noise terms are also important. In fact, the architectures described in Section 13.5.3 all use GPS to calibrate the initial conditions and bias errors. The noise terms may be the only significant terms that remain after such calibration. Over many missions, the noise terms should have a near-zero mean. However, they will depart from that mean on any given mission. Hence, we are interested in the standard deviation of these errors.

Accelerometer measurement noise can be reasonably well characterized as the sum of white noise and a Gauss-Markov random process. Equation (13.52) shows that this noise is integrated twice before it manifests itself in the position domain. The double integrator is a linear system, and so we recall (8.93).

$$\begin{aligned}\sigma_{A,W}^2 &= Q\int_{-\infty}^{\infty} |H(f)|^2 df \\ &= Q\int_{-\infty}^{\infty} h^2(\tau)d\tau\end{aligned} \tag{13.56}$$

These expressions give the variance of the noise at the output of a linear system when white noise appears at the input. The power spectral density of the input noise is Q watts/hertz. The

transfer function of the linear system is $H(f)$ and the corresponding impulse response is $h(t)$.

We begin with the impulse response for a single integrator.

$$\begin{aligned} h_1(t) &= \int_0^t \delta(\tau)\,d\tau \\ &= \begin{cases} 1 & t>0 \\ 0 & \text{else} \end{cases} \\ &= u(t) \end{aligned} \tag{13.57}$$

where $u(t)$ is the familiar step function. For the double integrator, we write

$$\begin{aligned} h_2(t) &= \int_0^t \int_0^s \delta(\tau)\,d\tau ds \\ &= tu(t) \end{aligned} \tag{13.58}$$

Employing (13.56), we find

$$\begin{aligned} \sigma^2_{\mathrm{A,W}} &= Q\int_{-\infty}^t h_2^2(\tau)\,d\tau \\ &= \frac{Qt^3}{3} \\ \sigma_{\mathrm{A,W}} &= \sqrt{\frac{Qt^3}{3}} \end{aligned} \tag{13.59}$$

As shown, the standard deviation due to white noise grows with $t^{3/2}$. Typical values for $\sqrt{Q}$ are given in Table 13.4 [Gebre-Egziabher (2001)], where $\sqrt{Q}$ has units of $\mathrm{m/s^2/\sqrt{Hz}}$. After all, Q is an acceleration power spectral density and so it has units of $\mathrm{m^2/s^4/Hz}$. The table gives typical values for $\sqrt{Q}$ for a tactical grade accelerometer and an automotive grade accelerometer. The *tactical grade* accelerometer is used to guide short-range missiles and costs thousands of dollars. The *automotive grade* accelerometer is used for automotive stability augmentation systems and to trigger the deployment of air bags. It costs tens of dollars.

Our second noise term, $\eta_{\mathrm{A,GM}}$, is due to Gauss-Markov noise. Unlike white noise, samples of Gauss Markov noise are correlated with the following auto-correlation function and power spectral density.

$$\begin{aligned} R_{GM}(\tau) &= \sigma^2 \exp(-\beta|\tau|) \\ S_{GM}(f) &= \frac{2\sigma^2\beta}{(2\pi f)^2+\beta^2} \end{aligned} \tag{13.60}$$

The variance is given by $\sigma^2 = R_{GM}(0)$. The auto-correlation decreases exponentially with time and is parameterized by a time constant, $1/\beta$. Typical values for these parameters are shown in Table 13.4 for our tactical and automotive grade sensors.

The output of the double integrator has the following variance.

$$\sigma^2_{\mathrm{A,GM}} = \frac{2\sigma^2}{\beta}\left(\frac{t^3}{3}-\frac{t^2}{2\beta}+\frac{1}{\beta^3}\right)-\frac{2\sigma^2}{\beta^3}\exp(-\beta t)\left(t+\frac{1}{\beta}\right) \tag{13.61}$$

Table 13.4 Typical parameters for the white noise and Gauss-Markov noise added to measurements for tactical grade and automotive grade inertial sensors

		Tactical INS	*Automotive INS*
Accelerometer	$\sqrt{Q}$ $(\mathrm{m/sec^2/\sqrt{Hz}})$	5×10^{-3}	1×10^{-2}
	$\sigma_{G\text{-}M}$ $(\mathrm{m/sec^2})$	5×10^{-4}	1.2×10^{-2}
	$\frac{1}{\beta}=\tau$ (sec)	60	100
Rate Gyro	$\sqrt{Q}$ $(\mathrm{rad/sec/\sqrt{Hz}})$	8.2×10^{-9}	2.4×10^{-7}
	$\sigma_{G\text{-}M}$ (rad/sec)	1.7×10^{-6}	8.7×10^{-4}
	$\frac{1}{\beta}=\tau$ (sec)	100	300

This result is derived in Homework Problem 13-9.

As shown in (13.59) and (13.61), the standard deviation for both noise terms grow with $t^{3/2}$ The 'acc-only' curves in Figure 13.25 plot the following quantity as a function of time

$$\sqrt{\sigma^2_{\mathrm{A,W}}+\sigma^2_{\mathrm{A,GM}}}$$

These curves are for the accelerometer parameters listed in Table 13.4. As shown, the automotive accelerometer noise terms cause the position error to grow to over 100 m in 200 seconds.

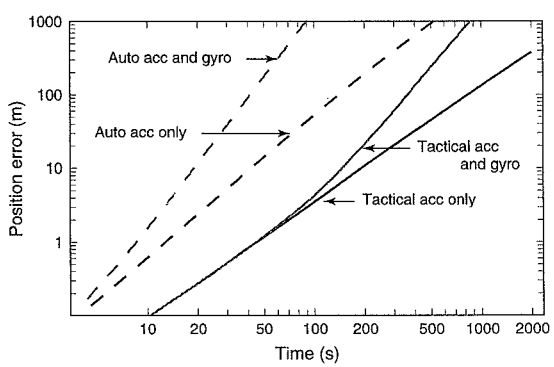


Figure 13.25 One-dimensional error growth for tactical and automotive inertial navigation systems.

In contrast, the position error from the tactical accelerometer noise grows to only 10 meters in 200 seconds. In the next section, we add the impact of tilt errors caused by imperfect gyro measurements.

13.5.5 Error Growth in One Dimension with Tilt

To account for tilt, we combine (13.41) and (13.49) to yield an embellished expression for the error in the measured acceleration.

$$\delta a = ka(t) + b + \left(\eta_{A,W}(t) + \eta_{A,GM}(t)\right) + g\sin\phi - \hat{g}\sin\hat{\phi} \tag{13.62}$$

This section focuses on the last two terms on the right since they are new relative to the analysis in the last section. In these terms, $\hat{\phi} = \phi + \delta\phi$ is the tilt estimate from the gyros, and the gravity estimate is $\hat{g} = g + \delta g$. The tilt error, $\delta\phi$, causes the accelerometer to unknowingly sense a component of the gravity field, and this error is our focus. The tilt error also causes an error in the rotation from the body frame to the navigation frame, but we ignore this error for simplicity.

For small values of tilt error, the following approximation holds.

$$(g\sin\phi - \hat{g}\sin\hat{\phi}) \approx -g\delta\phi\cos\phi - \delta g\sin\phi \tag{13.63}$$

The first term on the right is the object of our attention. The second term shows the impact of errors in the gravity model, δg. This error grows with tilt, but we assume that the accelerometer is nearly level, $\phi \approx 0$. In this case,

$$\delta a \approx ka(t) + b + \left(\eta_{A,W}(t) + \eta_{A,GM}(t)\right) - g\delta\phi \tag{13.64}$$

Tilt error rate may be modeled as follows.

$$\dot{\delta\phi} = \varepsilon + \left(\eta_{\phi,W}(t) + \eta_{\phi,GM}(t)\right) \tag{13.65}$$

In this equation, ε is the gyro drift rate. Like the accelerometer, the gyro measurement includes noise that is well modeled as white noise plus a Gauss Markov noise process (Gebre (2002)). We integrate to find

$$\delta\phi = \varepsilon t + \int_0^t \left(\eta_{\phi,W}(\tau) + \eta_{\phi,GM}(\tau)\right)d\tau + \delta\phi_0 \tag{13.66}$$

If we propagate these errors through our navigation equations, we find that the one-dimensional position error may be written as follows.

$$\delta x = \delta x(\phi = 0) + \frac{g\varepsilon t^3}{6} + \frac{g\phi_0 t^2}{2} + g\int_0^t\int_0^s\int_0^r \left(\eta_{\phi,W}(\tau) + \eta_{\phi,GM}(\tau)\right)d\tau\,dr\,ds \tag{13.67}$$

where $\delta x(\delta\phi = 0)$ is the result of our earlier analysis without tilt error. Note that the gyro adds terms that grow with t^2 and t^3.

We return to the navigation system that requires one-kilometer accuracy at the end of a one-hour mission. Recall that the vehicle was traveling at 500 km/hr. The tolerable error for the gyro drift, ε, is determined as follows.

$$\begin{aligned}
\frac{g\varepsilon t^3}{6} &< 1000 \text{ m} \\
\varepsilon &< \frac{6000}{gt^3} \\
\varepsilon &< 1.2\times10^{-8} \text{ rad/sec} \\
\varepsilon &< 0.0025 \text{ deg/hr}
\end{aligned} \tag{13.68}$$

This is an extremely small drift and illustrates the high sensitivity of inertial performance to gyro drift. Absent GPS or other suitable calibration, the associated cost dominates the cost of inertial navigation systems designed to autonomously provide high accuracy for long missions.

With GPS, the gyro drift and initial tilt can be calibrated along with the accelerometer biases. In this case, the noise terms are the most interesting. To find the variance at the output of a triple integration, we need the impulse response of a triple integration. This impulse response is

$$h_3(t) = \frac{t^2}{2}u(t) \tag{13.69}$$

Using (13.56), we find

$$\begin{aligned}
\sigma^2_{\phi,W} &= Q\int_{-\infty}^{t} h_3^2(\tau)\,d\tau \\
&= \frac{Qt^5}{20}
\end{aligned} \tag{13.70}$$

The variance due to the Gauss-Markov noise process is found using the path outlined in Homework Problem 13-10.

$$\sigma^2_{\phi,GM} = \frac{\sigma^2}{\beta^6}\left(-2 + \frac{t^3\beta^3}{3} - \frac{t^4\beta^4}{4} + \frac{t^5\beta^5}{10} + \exp(-\beta t)(2 + 2\beta t + t^2\beta^2)\right) \tag{13.71}$$

With the addition of tilt error, the noise term now grows with $t^{5/2}$. This faster function is shown for the tactical and automotive grade sensors as the 'acc and gyro' curves in Figure 13.25. As shown, the errors grow appreciably faster. Even so, the automotive grade sensors provide decimeter level accuracy for a few seconds. This suggests that they can be used to extend the time constant used by the GPS phase lock loop. With a longer time constant, the PLL filter has a narrower bandwidth and can better reject radio frequency interference. The tactical grade sensors can provide 100 meter accuracy for tens of seconds, and so they can be used to coast through complete GPS outages provided they do not last too long.

13.6 Tone Interference and Adaptive A/D Converters

The final essay in this chapter returns our attention to tone interference introduced in Sections 9.8 and 13.1. Recall our baseband model for an interfering tone from (9.53).

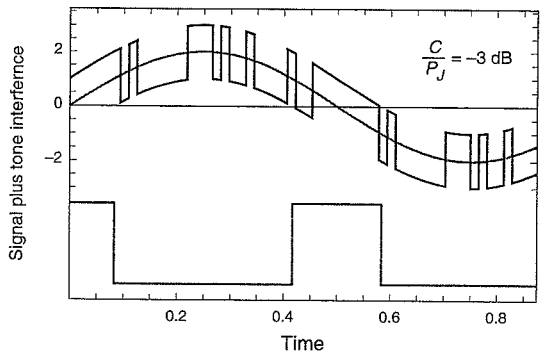


Figure 13.26 One-bit analog-to-digital converter (ADC) performs well when the GPS signal and the interfering tone have nearly equal power.

$$r(t) = \sqrt{C}D(t)x(t) + \sqrt{2P_J}\cos(2\pi f_J t + \theta_J) \tag{13.72}$$

The GPS signal appears in the top line and has power C. The interfering tone appears in the bottom line and has power P_J. This sum is shown in Figure 13.26 for $C/P_J = 1/2$ or $10\log_{10} C/P_J = -3$ dB.

Tone interference can be more or less painful than wideband interference. As detailed in Section 9.8.2, it can be more damaging to a C/A code receiver, if the tone frequency falls on one of the C/A code spectral lines. Fortunately, a number of RFI rejection techniques exploit the spectral structure of narrowband interference to good effect. For example, some GPS receivers contain adaptive filters that track the center frequency of the interference, P_J, and place a notch filter at that frequency. Of course, the notch filter also removes some of the GPS signal

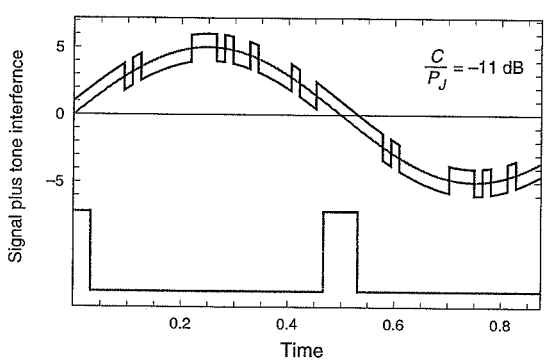


Figure 13.27 One-bit ADC is almost entirely captured by an interfering tone that is 11 dB stronger than the GPS signal.

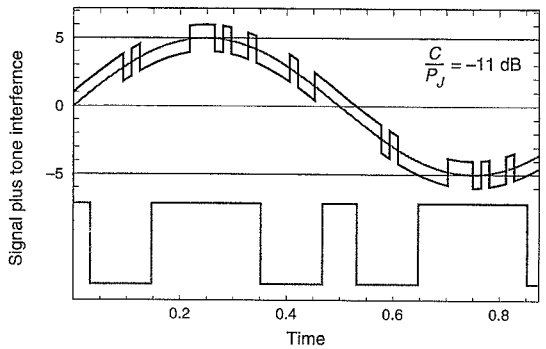


Figure 13.28 Two-bit ADC enables the GPS code to control the ADC output for a greater fraction of time. However, the ADC slicing levels must be carefully controlled.

power, but this loss is affordable if the bandwidth of the interference is narrow compared to the bandwidth of the GPS signal.

In this section, we consider another technique for mitigating narrowband interference—adaptive analog to digital (A/D) conversion. Consider Figure 13.26 again, where $P_J = 2C$. If the A/D converter outputs only one bit, then the output only contains the sign of the signal at the input. In other words, it slices the input along the zero axis. If the signal plus tone is above the axis, the A/D converter outputs a plus one. If the signal plus tone has negative sign, then the A/D converter outputs a negative one. In Figure 13.26, the signal and tone contend for control of the sign. When the tone is near zero, the signal is in control. When the tone is near peak amplitude, the interference is in control. This struggle is depicted by the heavy gray square wave at the bottom of the figure. This square wave is high when the signal determines the sign of the A/D output, and low when the tone interference is in control.

Figure 13.27 depicts the same struggle, but now the tone interference amplitude is stronger, $10\log_{10} C/P_J = -11$ dB. The tone is gaining more control over this one-bit A/D converter. As this trend continues, we say the tone captures the A/D converter.

Two-bit A/D converters can make capture much more difficult. Figure 13.28 shows the same signal plus tone shown in Figure 13.27, with $10\log_{10} C/P_J = -11$ dB. However, a two-bit A/D converter has four possible output states. It still slices the signal along the zero axis, and thus detects sign. However, it places two more slicing levels to subdivide the overall signal plus noise into four regions. The two-bit A/D converter shown in Figure 13.28 places the two new slicing levels at the peak of the sine wave interference. Now the GPS code still controls the output when the tone interference is anywhere near its own peak. Compare the gray square wave in Figure 13.28 to the square wave in Figure 13.27. Two-bit conversion greatly increases the fraction of time that the GPS signal is in control.

The slicing level must track the peak of the tone interference. In other words, the slicing level must adapt to any changes in the tone amplitude. This adaptation is implemented with an algorithm that tracks the fraction of time that the outer levels are exceeded.

Figure 13.29 sums up the advantage of adaptive A/D conversion by comparing the one-bit A/D converter to the two-bit A/D converter. It shows the fraction of time that the GPS

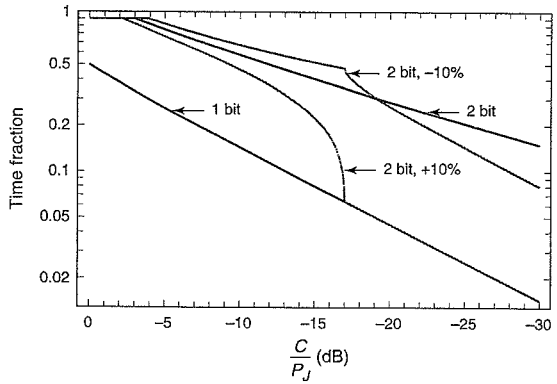


Figure 13.29 Fraction of time that the GPS code controls the output of the ADC as a function of the interference-to-signal ratio in decibels.

code controls the A/D output as a function of $10\log_{10}C/P_J$. As shown, when C/P_J is strong, the converter is given over to the signal. However, as $10\log_{10}C/P_J$ weakens, the one-bit A/D converter quickly gives control over to the tone interference. The two-bit A/D converter loses control much more slowly. The figure also has the curves for the two-bit A/D converter if the slicing level is set ten percent too high or too low.

13.7 Summary

This chapter contains six essays on GPS operation in radio frequency interference and in obstructed signal environments. The first essay characterized the nominal environment by focusing on the signal-to-noise ratios, C/N_0, that are found under open sky. Using this key metric, it also described the impact of signal obstructions and radio frequency interference. Specifically, it presented a table of signal-to-noise ratios measured in downtown Tokyo. It also estimated how signal-to-noise ratio depends on the power received from terrestrial sources of radio frequency interference. All these numbers are approximate, but serve to frame the rest of our discussion.

The second essay considered the basics of terrestrial radio propagation. It taught that signals from terrestrial sources can easily be much more powerful than signals from distant GPS satellites. It also showed that RFI levels are hard to predict when the signal comes from a terrestrial source.

The third essay described the first line of defense—the antenna! It introduced the notion of adaptive antennas, and focused on a ring nulling antenna that attenuates signals from the horizon. This essay reviewed the basic theory for a specific ring nuller and provided some flight results.

The fourth essay focused on Assisted GPS (AGPS), which is used to support GPS operation indoors and downtown. AGPS has been deployed to help distraught callers automatically report their position when they call 911. In time, it will become the basis for a spate of consumer services called location-based services. AGPS is similar to differential GPS because it sends information from a reference receiver at a controlled site to aid the roving GPS receiver. However, the goal of AGPS is to help the roving receiver estimate its position based on the short, noisy snippets of data likely to be available indoors and downtown. The assistance supplants the navigation message that normally comes from the satellites and sends data to support longer integration by the rover.

The fifth essay focused on inertial sensors that measure translational acceleration and rotational rates. These inertial sensors are powerful complements to GPS because they are self-contained. For this reason, they are not susceptible to radio frequency interference, but they do suffer from error growth with time. Fortunately, GPS can be used to calibrate the biases and drifts associated with inertial measurements. Our essay included a one-dimensional error analysis to teach the bare basics of this marriage.

The sixth essay focused on tone interference. GPS receivers employ a variety of techniques that mitigate the spectral structure of narrowband interference. Rather than teach all of these, we focused on one of the simplest—adaptive analog-to-digital conversion.

Please understand that these essays provide a first glimpse into six areas each of which is or could be the subject of several textbooks. Hence, we provide a longer than usual list of references.

Homework Problems

13-1. Consider a terrestrial radio transmitter located R meters from a GPS receiver. The transmitter radiates white noise with a power spectral density of –70 dBW/MHz and may interfere with GPS. The transmitter antenna gain in the direction of the GPS receiver is 0 dB, and the GPS antenna gain in the direction of the transmitter is –4 dB. Compute the ratio (in decibels) of the received power density from the terrestrial transmitter to the effective power density of natural noise at the input to the GPS receiver antenna. Plot the ratio as a function of range and consider the following two cases:

- The path loss for the interference is $1/4\pi R^2$
- The path loss for the interference is $1/4\pi R^4$

Comment on the results.

13-2. Generalize (13.10) to include the combined effect of tone interference and natural noise. Plot the tolerable tone power relative to GPS signal power, $(C/P_J)_{dB}$ as a function of the nominal signal-to-noise ratio, $(C/N_0)_{dB}$. Assume that the GPS receiver requires a 30 dB-Hz environment or better.

13-3. Derive (13.20). In other words, analyze a three path propagation model. The transmitter and receiver are h_T and h_R meters above the ground. The first path is direct, the second path reflects off the ground, and the third path reflects from Δh meters below the ground. The ground and below-ground reflection coefficients are α_1 and α_2 respectively. Hints: If the signal follows a direct path and two reflected paths, then the received interference can be modeled as

$$J(t) = \sqrt{2P_J}\cos(2\pi(f_L + f_J)t - kR_D) + \alpha_1\sqrt{2P_J}\cos(2\pi(f_L + f_J)t - kR_{M,1}) + \alpha_2\sqrt{2P_J}\cos(2\pi(f_L + f_J)t - kR_{M,2})$$

Now the path lengths are given by

$$R_D = \sqrt{R^2 + (h_T - h_R)^2}$$
$$R_{M,1} = \sqrt{R^2 + (h_T + h_R)^2}$$
$$R_{M,2} = \sqrt{R^2 + (h_T + h_R + 2\Delta h)^2}$$

13-4. As discussed in Section 13.4.1, AGPS supplants the GPS navigation message and provides an alternate means for the roving user to compute the locations of the GPS satellites. To do so, it must also send the present time to the rover. While the Keplerian parameters vary slowly with time, the true anomalies are fast functions of time. After all, the satellites are moving at 3860 m/s along track. The required time accuracy can be computed by considering a satellite on the user's horizon. The pseudorange error for such a low-lying satellite is most sensitive to errors in the time assistance. Show that the pseudorange error due to the time error is

$$\delta R \approx V_{SV}\Delta T\frac{R_E}{R_{SV}} \text{ meters}$$

In this equation, $V_{SV} \approx 3860$ m/s is the intrack velocity of the satellite, ΔT is the time error, $R_E = 6370$ km is the radius of the earth, and $R_{SV} = 26{,}560$ km is the distance to the satellite from the earth's center. Determine the time accuracy required to obtain $\delta R < 10$ m.

13-5. Verify the results for satellite Doppler and Doppler rate given in Table 13.3.

13-6. Determine the Doppler error and Doppler rate error that results for an AGPS rover as a function of the error in the *a priori* estimate of the rover's position.

13-7. Verify the results for the Doppler and Doppler rate due to the low and high dynamic users in Table 13.3.

13-8. Equation (13.48) may be generalized for FOG geometries other than the circle shown in Figure 13.23. For a path of any shape, the phase difference may be approximated as follows.

$$\Delta\phi \approx \frac{8\pi nA\omega}{c\lambda} \text{ radians}$$

where A is the area enclosed by the counter-rotating light beams. (a) Prove this result for a rectangular geometry. (b) Prove this result for an arbitrarily shaped path.

13-9. Derive (13.61), which gives the variance for the noise at the output of a double integrator when the input noise is a Gauss-Markov random process with auto-correlation function given by (13.60). Hint: The variance for the output noise of a linear system is given by

$$\sigma_Z^2 = \int_{-\infty}^{\infty} f(\lambda)R_X(\lambda)\,d\lambda$$
$$f(\lambda) = \int_{\max(0,\lambda)}^{\min(t,t+\lambda)} h(\tau - \lambda)h(\tau)\,d\tau \qquad (13.73)$$

In these equations, $R_X(\tau)$ is the auto-correlation function of the input noise, and $h(\tau)$ denotes the impulse response of the linear system.

13-10. Derive (13.71), which gives the variance for the noise at the output of a triple integrator when the input noise is a Gauss-Markov random process with auto-correlation function given by (13.60). Hint: The variance for the output noise of a linear system is given by (13.73).

13-11. Derive and plot the functions shown in Figure 13.29.

References

- **General**

Enge P., D. Akos, J. Do, J. Simoneau, L.W. Pearson and V. Seetharam (2004). Measurements of the Man-Made Spectrum Noise Floor, National Aeronautics and Space Administration, NASA/CR-2004-213551.

Rabinowitz, M. and J. Spilker (2003). The Rosum TV Position Technology, *Proc. ION*, 59th Annual Meeting of the Institute of Navigation, pp. 528–541.

Rounds, S. (2004a). Jamming Protection of GPS Receivers: Part I - Receiver Enhancements, *GPS World*, January 2004, pp. 54–59.

RTCA (2002). Assessment of Radio Frequency Interference Relevant to the GNSS, Report RTCA DO-235.

Sennott, J. and Senffner, D. (1994). Navigation Receiver with Coupled Signal-Tracking Channels, U.S. Patent 5,343,209, Aug. 30, 1994.

Spilker, J. (1996). Fundamentals of Signal Tracking Theory, Volume I of *Global Positioning System: Theory and Applications*, B. Parkinson, J. Spilker, P. Axelrad and P. Enge (eds.), American Institute of Astronautics and Aeronautics, pp. 245–328.

Ward, P. (1996). Effects of RF Interference on GPS Satellite Signal Receiver Tracking, *Understanding GPS Principles and Applications*, Elliott Kaplan (ed.), Artech House, pp. 209–236.

- **Adaptive Antennas**

Applebaum, A. (1976). Adaptive Arrays, *IEEE Transactions on Antennas and Propagation*, vol. AP-24, no. 5, pp. 585–598.

Bauregger, F. (2003). Novel Anti-Jam Antennas for Airborne GPS Navigation, Ph.D. dissertation, Stanford University, Department of Aeronautics and Astronautics.

Compton, R. (1988). *Adaptive Antennas: Concepts and Performance*, Prentice Hall.

Kim, U-S, D. DeLorenzo, J. Gautier, P. Enge, and J. Orr (2004). Phase Effects Analysis of Patch Antenna CRPAs for JPALS, *Proc. ION GNSS 2004*, pp. 1531–1538.

Rounds, S. (2004b). Jamming Protection of GPS Receivers: Part II—Antenna Enhancements, *GPS World*, February 2004, pp. 38–45.

Widrow, B. *et al.* (1967). Adaptive Antenna Systems, *Proc. IEEE*, vol. 55, no. 12, pp. 2143–2159.

Rosen, M. and M. Braasch (1998). Low-Cost GPS Interference Mitigation Using Single Aperture Cancellation Techniques, *Proc. ION National Technical Meeting*, pp. 47–58.

- **Assisted GPS**

Agarwal, N., J. Basch, P. Bechmann, P. Bharti, S. Casadei, A. Chou, P. Enge, W. Fong, N. Hathi, W. Mann, A. Sahai, J. Stone, J. Tsitsiklis and B. Van Roy (2002). Urban GPS: Algorithms for GPS Operation Indoors and Downtown, *GPS Solutions*, vol. 6, pp. 149–160.

Akos, D., P-L Normark, J-T Lee, and K. Gromov (2000). Low Power Global Navigation Satellite System Signal Detection and Processing, *Proc. ION GPS 2000*, pp. 784–791.

Chansarkar, M. and L. Garin (2000). Acquisition of GPS Signals at Very Low Signal to Noise Ratio, *Proc. ION National Technical Meeting*, pp. 731–737.

Enge, P., R. Fan and A. Tiwari (2001). GPS Reference Network's New Role: Providing Continuity and Coverage, *GPS World*, vol. 12, no. 7, July 2001, pp. 38–45.

Garin, L., M. Chansarkar, S. Miocinovic, C. Norman, and D. Hilgenberg (1999). Wireless Assisted GPS-SiRF Architecture and Field Results, *Proc. ION GPS 1999*, pp. 489–497.

Lundgren, D. and F. van Diggelen (2005). Long-Term Orbit Technology for Cell Phone PDAs, *GPS World*, vol. 16, no. 10, pp. 32–36.

Moeglein, M. and N. Krasner (1998). An Introduction to SnapTrack™ Server-Aided GPS Technology, *Proc. ION GPS 1998*, pp. 333–342.

Soliman, S., P. Agashe, I. Fernandez, A. Vayonos, P. Gaal and M. Oljaca (2000). gpsOne™: A Hybrid Position Location System, *Proc. IEEE Sixth International Symposium on Spread Spectrum Techniques and Applications*, pp. 330–335.

Syrjarinnne, J. (2001). Wireless Assisted GPS: Keeping Time With Mobiles, *GPS World*, vol. 12, no. 1, pp. 22–31.

Taylor, R. and J. Sennott (1984). Navigation System and Method, U.S. Patent No. 4,445,118, April 24, 1984.

van Diggelen, F. and C. Abraham (2001). Indoor GPS: The No-Chip Challenge, *GPS World*, vol 12, no. 9, pp. 50–58.

van Diggelen, F. (2002). Method and Apparatus for Processing of Satellite Signals Without Time of Day Information, U.S. Patent No. 6,417,801, July 9, 2002.

- **Inertial Sensors and Navigation**

Albans, S. (2004). Design and Performance of a Robust GPS/INS Attitude System for Automobile Applications, Ph.D. dissertation, Stanford University, Department of Aeronautics and Astronautics.

Gautier, J. (2003). GPS/INS Generalized Evaluation Tool (GIGET) for the Design and Testing of Integrated Navigation Systems, Ph.D. dissertation, Stanford University, Department of Aeronautics and Astronautics.

Gebre-Egziabher, D. (2001). Design and Performance Analysis of a Low-Cost Aided Dead Reckoning Navigator, Ph.D. dissertation, Stanford University, Department of Aeronautics and Astronautics.

Greenspan, R. (1996). GPS and Inertial Integration, Chapter 7 of *Global Positioning System: Theory and Applications*, B. Parkinson, J. Spilker, P. Axelrad and P. Enge (eds.), American Institute of Astronautics and Aeronautics, pp. 187–218.

Soloview, A., F. van Graas, S. Gunawardena (2004). Implementation of Deeply Integrated GPS/Low-Cost IMU for Reacquisition and Tracking of Low CNR GPS Signals, *Proc. ION National Technical Meeting*, pp. 923–935.

- **Mitigation of Narrowband Interference**

Amoroso, F. (1983). Adaptive A/D Converter to Suppress CW Interference in Spread Spectrum Communications, *IEEE Transactions on Communications*, vol. 31, no. 10, pp. 1117–1123.

Appendix A GPS Data Sets on the CD

Dr. Guttorm R. Opshaug and Dr. Keith Alter

In this appendix, we describe the data sets provided on the Windows-formatted CD-ROM which accompanies this textbook. Several of these data sets are used in the problems in Chapters 4–7. Others are provided to give the student options to try out her ideas. The CD contains:

- GPS data sets
- Text files describing data formats
- MATLAB scripts and functions
- GPS-related documents

A.1 Introduction

GPS applications can be divided loosely into two types on the basis of how the data are used.

- Real-time applications (e.g., navigation and timing) require that the GPS measurements be processed in real time to derive the information of interest and to communicate this information to the user in a prearranged form. In some cases, the data may be processed immediately and the results used internally to update the receiver display (for a hiker or boater). In other cases, the measurements and/or processed data may be exported via an electronic connection (e.g., a serial communications cable) from the receiver to another device (e.g., the autopilot in an aircraft).
- Post-processing applications (e.g., precise position monitoring networks and textbooks such as this) require that the measurements be exported to a device for storage in an electronic file to be analyzed at a later time.

Whether obtained directly from a GPS receiver or from a file, the data are always formatted in a prearranged manner. The data are either in *binary* or *text* format (text is sometimes called 'ASCII'). Binary data are a prearranged series of bits, and tend to be more compact than the data sent as text. Binary data are often preferable when speed of transmission is of concern, or the amount of data storage is limited. Text data are sent as letters and numbers. While this format is less compact than the binary, the text format lends itself to easy visual inspection. All data sets on the CD are in text format.

Binary or text, a user of GPS data must understand the data format: the order in which the satellite data are sent; locations of the line breaks, tabs, and spaces (characters which separate data fields are called *delimiters*); and the units of numerical data (e.g., meters, degrees per second). How the data are organized is discussed in a document generally available from the receiver manufacturer or the provider of the data. This document is called a *specification* (sometimes referred to as 'the decoder ring').

We introduce below some of the data formats commonly used in the GPS community.

A.2 Common Data Formats

A.2.1 Real-Time Data Formats

Modern GPS receivers generally have the capability to output data electronically. In particular, some receivers provide data to a personal or portable computer. Often you can plug such receivers directly into a PC's serial port (often called *COM* port). Some GPS receivers are imbedded on cards, which are plugged into expansion slots inside the computer. In any case, the user must carefully configure his computer in accordance with the receiver manufacturer's specifications for receiver-to-computer communications to work properly. The user must also make certain that the GPS receiver itself is properly supplied with power and is properly set up to receive the satellite signals. In addition, careful configuration of the communication interface and the receiver itself is often required for proper operation. During development and/or use of the data output from a GPS receiver, it is a good idea to have a test program available to verify that the receiver is in fact sending data.

Many GPS receivers output data in formats that are specific to the manufacturer or even to a particular receiver model. As long as the format is specified, a user can create software to decode and utilize the data. In rare cases, the data formats are proprietary to the company that created them, and a user who wants to read such data must contact the company for specifications and authorization.

An example of a custom binary format for data export from a GPS receiver might look something like this (0x## refers to a two-digit hexadecimal (base 16) number, or *byte*, which has decimal equivalent values from 0 to 255):

0x10	(start or end)
0x26	(identifier for the data message type)
0x3C	(total number of data bytes)
0x4A	(first data byte)
⋮	(more data bytes...)
0xF1	(last data byte)
0x12	(checksum)
0x10	(start or end)
0x03	(end of transmission)

A series of bytes with information contained between two delimiter bytes (typically 0x10) is called a *packet*. A receiver might send out a continuous stream of these packets, and obviously the data bytes contained in these packets would change as the GPS measurements change.

We should make a note of two standard formats in common use in real-time applications.

- *NMEA 0183*
  This standard was created by the National Marine Electronics Association (NMEA) and is described in their document *NMEA 0183 Interface Standard* available from NMEA, P.O. Box 3435, New Bern, NC 28564-3435, USA. The receiver output consists of a set of standard messages in text format. Generally, this format is used by GPS receivers that output processed position, velocity, time, and other data via a serial connection. Messages are prefaced by a six-character identifier starting with a '$' symbol, such as '$GPGGA' for GPS fix data, and have text data fields separated by commas. See website <http://www.nmea.org/0183.htm>.

- *RTCM SC-104*
  This standard was developed by the Radio Technical Commission for Maritime Services (RTCM); it is described in detail in *RTCM Paper 134-89/SC 104-68*, available from RTCM, 1800 Diagonal Road, Suite 600, Alexandria, VA 22314-2480, USA. RTCM SC-104 format is widely used for transmission of differential GPS correction data in binary format. There are about two dozen message types providing code and carrier phase corrections. See website <http://www.rtcm.org>.

A.2.2 Post-Processed Data Formats

Scientific investigators often require access to GPS data recorded hours, weeks, and even years earlier. These data are usually stored in an electronic file. Sometimes this file is a direct copy of the binary or text output from a GPS receiver. In most cases, however, the file contains data that have been reorganized. The purpose of the reorganization is to reduce the size of the data record and to put it in a standard format. The current *de facto* standard for storage of GPS data is the *Receiver Independent Exchange* (RINEX) format. Generally, the RINEX files are made up of a header part followed by multiple blocks of data. A complete description is given in (CD) *Documents\RINEX.txt*.

The enclosed CD provides several GPS data sets in RINEX format from three Continuously Operating Reference Stations (CORS) in California: Durmid Hill, Pigeon Point, and Vandenberg. (The National Geodetic Survey (NGS) coordinates a network of CORS sites throughout the United States which contribute their GPS carrier phase and code range measurements to a central facility. These data are converted into RINEX format and placed in files available on the Internet: <http://www.ngs.noaa.gov/CORS/>.)

For our purposes, there are three types of RINEX files to deal with:

- observation data (*.obs* files),
- navigation message (*.nav* files),
- meteorological data (*.met* files)

These files are named *xxmmddyy.met*, *xxmmddyy.nav*, and *xxmmddyy.obs*. The prefix *xx* specifies the location: *dh* for Durmid Hill; *pp* for Pigeon Point; and *vb* for Vandenberg. The character string *mmddyy* in the file name refers to the month, day and year when the data were collected. All three CORS sites provide *.obs* and *.nav* files; Durmid Hill also provides *.met* files.

In order to use these data, the student would have to write code to read the RINEX files and extract the required data. To get the student started, we have extracted the relevant data from two of the RINEX files and reformatted them so that can be loaded directly into MATLAB. We encourage the student to use the additional files for other GPS projects (e.g., differential positioning at Stanford using data from the three CORS sites to analyze the effects of distance and latency).

Post-processing applications can take advantage of precise, post-processed ephemerides. The most common data format for dissemination of these data sets is a text format called SP3. SP3 files consist of a header followed by multiple blocks of data. A complete description of the data format is given in the file (CD) *Documents\SP3.txt*.

Several SP3 files are on the CD. The files are named *igmmddyy.sp3*. The prefix *ig* refers to the International GPS Service (IGS), an international scientific organization responsible for generating these data files. As before, *mmddyy* specifies the month, day, and year of the GPS data.

A.3 Formats of Data Files on the CD

A.3.1 Format A

Data in format A can be downloaded into MATLAB directly. These data were all logged at Stanford University. The raw GPS measurements are in files named *rmmddyyx.dat* and ephemeris parameter values are in files named *emmddyyx.dat*. The data collection period for these files ranges from ten minutes to 24 hours. To make the 24-hour data sets manageable, we have divided them into three separate blocks of about eight hours each. These files have eight-character names ending in letters *a*, *b* or *c* signifying the beginning, middle, and end of the original 24-hour data set.

The files *rcvr.dat* and *eph.dat* contain data corresponding to a single 'snapshot' of GPS measurements. These files are intended to help the student implement navigation algorithms in code while working with minimal data sets. The file *rcvr.dat* is an 8×7 matrix to be referred to below as *rcvr*. Each row corresponds to one of the eight satellites in view at the measurement *epoch* (an epoch is simply a term that refers to a single point in time). The seven columns provide the following data:

```
Column 1: rcvr_tow; - receiver time of week (s)
Column 2: svid;     - satellite PRN number (1-32)
Column 3: pr;       - pseudorange (m)
Column 4: cycles;   - number of carrier cycles
Column 5: phase;    - fraction of carrier cycle (1/2048 cycle)
Column 6: slp_dtct; - 0: no cycle slip detected; non-0: cycle slip
Column 7: snr_dbhz; - signal to noise ratio (dB-Hz)
```

For each satellite, use the following equation to construct the carrier phase measurement $\phi^{(k)}$ (in cycles) from the receiver data:

$$\phi^{(k)} = rcvr(k,4) + \frac{rcvr(k,5)}{2048}$$

Make sure that you check for cycle slips (column 6) before using a measurement in your navigation algorithm. (MATLAB Hint: Use 'load rcvr.dat' to load the file *rcvr.dat* into a MATLAB matrix called *rcvr*.)

The file *eph.dat* is an 8×24 matrix containing the ephemeris data used to determine the position of each satellite at any time. Each row corresponds to one of the satellites in view. The columns are described below.


```
Column 1: rcvr_tow; - receiver time of week (s)
Column 2: svid;     - satellite PRN number (1-32)
Column 3: toc;      - reference time of clock parameters (s)
Column 4: toe;      - reference time of ephemeris parameters (s)
Column 5: af0;      - clock correction coefficient (s)
Column 6: af1;      - clock correction coefficient (s/s)
Column 7: af2;      - clock correction coefficient (s/s/s)
Column 8: ura;      - user range accuracy (m)

Column 9: e;        - eccentricity
Column 10: sqrta;   - square root of semi-major axis a (m^1/2)
Column 11: dn;      - mean motion correction (rad/s)
Column 12: m0;      - mean anomaly at reference time (rad)
Column 13: w;       - argument of perigee (rad)
Column 14: omg0;    - Longitude of ascending node (rad)
Column 15: i0;      - inclination angle at reference time (rad)
Column 16: odot;    - rate of right ascension (rad/s)

Column 17: idot;    - rate of inclination angle (rad/s)
Column 18: cus;     - argument of latitude correction, sine (rad)
Column 19: cuc;     - argument of latitude correction, cosine (rad)
Column 20: cis;     - inclination correction, sine (rad)
Column 21: cic;     - inclination correction, cosine (rad)
Column 22: crs;     - radius correction, sine (m)
Column 23: crc;     - radius correction, cosine (m)
Column 24: iod;     - issue of data number
```

Multi-epoch data are organized as follows: *rmmddyyx.dat* files stack receiver data in the format described above, one row for each SV for each epoch; *emmddyyx.dat* adds a 24-column ephemeris data set each time the parameters are updated.

A.3.2 Format B (Parsed RINEX)

For the two sets of the CORS (RINEX) files required in Problems 5.4–5.5, we have extracted the relevant data and reformatted them so that they can be downloaded into MATLAB. These files have been renamed *xxmmddyy.rio* and *xxmmddyy.rin*, where the extensions *.rio* and *.rin* refer to observation and navigation files, respectively.

The receiver observation files (*.rio*) are formatted as follows:

```
Column 1: GPS_week; - GPS week number
Column 2: rcvr_tow; - receiver time of week (s)
Column 3: svid;     - satellite PRN number (1-32)
Column 4: C1;       - pseudorange from C/A-code on L1 (m)
Column 5: L1;       - L1 carrier cycles (number of L1 cycles)
Column 6: L2;       - L2 carrier cycles (number of L2 cycles)
```

```
Column 7: P1;        - pseudorange from P-code on L1 (m)
Column 8: P2;        - pseudorange from P-code on L2 (m)
Column 9: D1;        - Doppler on L1 (Hz)
Column 10: D2;       - Doppler on L2 (Hz)
```

The navigation data files (*.rin*) are formatted as follows:

```
Column 1: svid;      - satellite PRN number (1-32)
Column 2: m0;        - mean anomaly at reference time (rad)
Column 3: dn;        - mean motion correction (rad/s)
Column 4: e;         - eccentricity
Column 5: sqrta;     - square root of semi-major axis a (m^1/2)
Column 6: omg0;      - longitude of ascending node (rad)
Column 7: i0;        - inclination angle at reference time (rad)
Column 8: w;         - argument of perigee (rad)
Column 9: odot;      - rate of right ascension (rad/s)
Column 10: idot;     - rate of inclination angle (rad/s)
Column 11: cuc;      - argument of latitude correction, cosine (rad)
Column 12: cus;      - argument of latitude correction, sine (rad)
Column 13: crc;      - radius correction, cosine (m)
Column 14: crs;      - radius correction, sine (m)
Column 15: cic;      - inclination correction, cosine (rad)
Column 16: cis;      - inclination correction, sine (rad)
Column 17: toe;      - reference time of ephemeris parameters (s)
Column 18: iode;     - issue of data ephemeris
Column 19: GPS_week; - GPS week number
Column 20: toc;      - reference time of clock parameters (s)
Column 21: af0;      - clock correction coefficient (s)
Column 22: af1;      - clock correction coefficient (s/s)
Column 23: af2;      - clock correction coefficient (s/s/s)
Column 24: wdot;     - argument of perigee rate (rad/s)
```

A.3.3 Format C (Parsed SP3)

Problem 4-15 requires post-processed precise ephemeris. The commonly used format for such data is SP3, introduced earlier. In order to simplify the problem, we have extracted the relevant data and reformatted them so that the data can be downloaded into MATLAB. This parsed version of the original file is named *ig010600.spp*, and is formatted as follows:

```
Column 1: GPS_tow;   - GPS time of week (s)
Column 2: svid;      - satellite PRN number (1-32)
Column 3: x;         - satellite position x-component (m)
Column 4: y;         - satellite position y-component (m)
Column 5: z;         - satellite position z-component (m)
```

The satellite positions are given in WGS 84 earth-centered, earth-fixed (ECEF) coordinates.

A.3.4 ALMANACX.DAT Format

Two files on the CD, *almanac1.dat* and *almanac2.dat*, contain coarse position information on the satellites. The almanac data allow a GPS receiver to determine SV locations within kilometers. (Recall that the ephemeris data specify SV positions with meter-level accuracy.) When a receiver starts up after having been off-line for a period of time, the almanac retained in the receiver memory allows the receiver to generate a 'ballpark' estimate of satellite position, narrowing SV search and reducing the time to get back on-line.

Almanacx.dat contains blocks of almanac information for each SV, and the blocks are formatted in the following way:

```
Row 1: svid;         - satellite PRN number (1-32)
Row 2: health;       - satellite health status flag
Row 3: e;            - eccentricity
Row 4: toa;          - reference time of almanac parameters (s)
Row 5: i0;           - inclination angle at reference time (rad)
Row 6: odot;         - rate of right ascension (rad/s)
Row 7: sqrta;        - square root of semi-major axis (m^1/2)
Row 8: omg0;         - longitude of ascending node (rad)
Row 9: w;            - argument of perigee (rad)
Row 10: m0;          - mean anomaly at reference time (rad)
Row 11. af0;         - clock correction coefficient (s)
Row 12: af1;         - clock correction coefficient (s/s)
Row 13: GPS_week;    - GPS week number
```

A.4 Data Files on the CD

A.4.1 GPS Measurement Data Sets (Format A)

rcvr.dat *eph.dat*	Data from a single epoch logged at Stanford University on March 29, 1999. Selective Availability (SA) status: active. Folder: *Data\Original*
r091400.dat *e091400.dat*	Data logged at Stanford University on September 14, 2000. Data collection period: 10 minutes; SA status: off; sampling interval: 1 sec. Folder: *Data\Stanford\RF1*
r073000.dat *e073000.dat*	Data logged in Stranda, Norway, on July 30, 2000. Data collection period: 5 hours; SA status: off; sampling interval: 1 sec. Folder: *Data\Stranda\July_30_2000*
r010600a.dat *r010600b.dat* *r010600c.dat* *e010600a.dat* *e010600b.dat* *e010600c.dat*	Data logged at a location in Silicon Valley on January 6, 2000. Data collected over a 24-hour period are divided into three segments and stored in separate files (XXXa.dat-XXXc.dat). SA status: active; sampling interval: 5 sec. Folder: *Data\ Stanford\January_6_2000*
r091800a.dat *r091800b.dat* *r091800c.dat* *e091800a.dat*	Data logged at Stanford University on September 18, 2000. Data collected over a 24-hour period are divided into three segments and stored in separate files (XXXa.dat-XXXc.dat). SA status: off; sampling interval: 5 sec.

e091800b.dat Folder: *Data\ Stanford\September_18_2000*
e091800c.dat

A.4.2 GPS Measurement Data Sets (Format B)

These data sets are from the CORS site at Pigeon Point.

pp071500.rin Folder: *Data\Pigeon_Point\July_15_2000\Parsed*
pp071500.rio SA status: off
pp091800.rin Folder: *Data\Pigeon_Point\September_18_2000\Parsed*
pp091800.rio SA status: off

A.4.3 GPS Measurement Data Sets (RINEX Format)

These data files were downloaded from the website <http://www.ngs.noaa.gov/CORS/>. Each file below contains data logged over 24 hours starting at midnight UTC. Sampling interval is 30 sec.

Durmid Hill

Durmid Hill is located approximately 600 km south of Stanford University. For site information and antenna location see file (CD) *Data\Durmid_Hill\Durmid_Hill.pos.*

dh010600.met Folder:*Data\Durmid_Hill\January_6_2000*
dh010600.nav SA status: active
dh010600.obs
dh071500.met Folder: *Data\Durmid_Hill\July_15_2000*
dh071500.nav SA status: off
dh071500.obs
dh091800.met Folder: *Data\Durmid_Hill\September_18_2000*
dh091800.nav SA status: off
dh091800.obs

Pigeon Point

Pigeon Point is located approximately 40 km west of Stanford University. For site information and antenna location see file (CD) *Data\Pigeon_Point\Pigeon_Point.pos.*

pp091897.nav Folder: *Data\Pigeon_Point\September_18_1997*
pp091897.obs SA status: active
pp091898.nav Folder: *Data\Pigeon_Point\September_18_1998*
pp091898.obs SA status: active
pp091899.nav Folder: *Data\Pigeon_Point\September_18_1999*
pp091899.obs SA status: active
pp010600.nav Folder: *Data\Pigeon_Point\January_6_2000*
pp010600.obs SA status: active
pp071400.nav Folder: *Data\Pigeon_Point\July_14_2000*
pp071400.obs SA status: off
pp071500.nav Folder: *Data\Pigeon_Point\July_15_2000*
pp071500.obs SA status: off
pp071600.nav Folder: *Data\Pigeon_Point\July_16_2000*
pp071600.obs SA status: off

pp091800.nav Folder: *Data\Pigeon_Point\September_18_2000*
pp091800.obs SA status: off

Vandenberg

Vandenberg is located approximately 400 km south of Stanford University. For site information and antenna location see file *Data\Vandenberg\Vandenberg.pos.*

vb010600.nav Folder: *Data\Vandenberg\January_6_2000*
vb010600.obs SA status: active
vb071500.nav Folder: *Data\Vandenberg\July_15_2000*
vb071500.obs SA status: off
vb091800.nav Folder: *Data\Vandenberg\September_18_2000*
vb091800.obs SA status: off

A.4.4 Almanac Data (ALMANACX.DAT Format)

Folder: *Data\Almanac*

almanac1.dat Sample almanac data for all satellites: almanac1.dat is at the beginning of the day on January 6, 2000; almanac2.dat is the end of the day on January 6, 2000.
almanac2.dat

A.4.5 Precise Ephemeris Data (Format C)

This data set was obtained from the IGS website <http://igscb.jpl.nasa.gov/>.

ig010600.spp January 6 2000.

A.4.6 IGS Precise Ephemeris Data (SP3 Format)

These data sets were obtained from the IGS website <http://igscb.jpl.nasa.gov/>.
Folder: *Data\Precise_ephemeris*. Each file covers a 24-hour period starting at midnight UTC.

ig091897.sp3 September 18, 1997
ig091898.sp3 September 18, 1998
ig091899.sp3 September 18, 1999
ig010600.sp3 January 6, 2000
ig071400.sp3 July 14, 2000
ig071500.sp3 July 15, 2000
ig071600.sp3 July 16, 2000
ig091800.sp3 September 18, 2000

A.4.7 MATLAB Functions and Scripts

Folder: *Navigation_utilities*

d2dms.m Conversion from degrees to degrees, minutes and seconds.
dms2d.m Conversion from degrees, minutes and seconds to degrees.
enu2wgsxyz.m East-North-Up to Earth-Centered-Earth-Fixed XYZ position conversion.

enu2wgslla.m	East-North-Up to Longitude-Latitude-Altitude position conversion.
Wgslla2enu.m	Longitude-Latitude-Altitude to East-North-Up position conversion.
Wgslla2xyz.m	Longitude-Latitude-Altitude to ECEF XYZ position conversion.
Wgsxyz2enu.m	ECEF XYZ to East-North-Up position conversion.
Wgsxyz2lla.m	ECEF XYZ to Longitude-Latitude-Altitude position conversion.
rot.m	Generates rotation matrix.
Rxyz2enu.m	Generate rotation matrix between ECEF XYZ and ENU coordinates.
Rotenu2xyz.m	Rotates an ENU vector into ECEF XYZ coordinates
Rotxyz2enu.m	Rotates an ECEF XYZ vector into ENU coordinates.

Many MATLAB scripts useful to a GPS newcomer are available from Professor Kai Borre's website: <http://www.gps.auc.dk/~borre/easy>. We also recommend the website affiliated with the GPS Toolbox column in the quarterly journal *GPS Solutions* for algorithms and source code: <http://www.ngs.noaa.gov/gps-toolbox/index.html>.

A.4.8 GPS-Related Documents

Folder: *Documents*

The following documents, all in the public domain, are provided for ready reference.

RINEX.txt	Document describing the RINEX standard. <ftp://igscb.jpl.nasa.gov/igscb/data/format/rinex210.txt>
	For data preprocessing software, including conversion of raw data into RINEX <http://www.unavco.org/facility/software/teqc.html>
SP3.txt	Document describing the SP3 standard by P. Spofford and B. Remondi. <http://www.ngs.noaa.gov/GPS/SP3_format.html>
DSB Task Force 2005.pdf	Defense Science Board Task Force on the Future of the Global Positioning. System, U.S. DoD, 2005
EU-US Agreement 2004.pdf	Agreement on the Promotion, Provision, and Use of Galileo and GPS Satellite-based Navigation Systems, 2004
FRP 2005.pdf	2005 Federal Radionavigation Plan
FRS 2001.pdf	2001 Federal Radionavigation Systems
IS-GPS-200D.pdf	GPS Interface Specification IS-GPS-200, Revision D, 2004
IS-GPS-705.pdf	Navstar GPS Space Segment/User Segment L5 Interface, 2005
SPS Performance 2001.pdf	GPS Standard Positioning Service Performance Standard, U.S. DoD, 2001
US GPS Policy 1996.pdf	Presidential Decision Directives on U.S. GPS Policy, 1996
US PNT Policy 2004.pdf	U.S. Space-based Positioning, Navigation, and Timing Policy, 2004
Volpe Report 2001.pdf	Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Positioning System, U.S. DOT, 2001
WGS 84.pdf	NIMA Technical Report on World Geodetic System 1984

Index

The second edition of this widely praised book offers a comprehensive introduction to GPS: the system, signals, receivers, measurements, and algorithms for estimation of position, velocity, and time. It is intended as a textbook for a senior- or graduate-level engineering course and a self-study guide for practicing engineers.

The book is divided into four parts. Part I introduces the basic framework for a global navigation satellite system, including coordinate frames, time references, and satellite orbits, and provides an overview of GPS, GLONASS, and Galileo. Part II describes the fruits of GPS: estimation of position, velocity, and time. Part III discusses the ingenious structure of the GPS signals. Part IV introduces the signal processing steps required to extract the necessary measurements from these signals, and explores the challenges posed by signal blockage and RFI.

To give the student a hands-on experience with GPS, this book includes a CD with a number of GPS data sets from several sites. A set of homework problems requires the student to write simple MATLAB code to analyze these data.

Praise for the First Edition—

"... an excellent graduate level textbook on GPS crafted by two master teachers. The authors have succeeded in providing broad coverage of the workings of GPS without missing important details, and have done so in a coherent and readable form." —Professor Penina Axelrad, University of Colorado at Boulder

"This is a superb book... The authors are to be complimented for being up to date in a rapidly changing GPS environment." —Dr. Myron Kayton, Consultant

"This is the book to use to educate those people, more and more in demand, that not only know how to use GPS equipment and the measurements produced by them, but also have a fundamental understanding of how GPS technology and GPS sensors proliferating everywhere in society actually function."
—Professor Martin Vermeer, Helsinki University of Technology

Pratap Misra, Ph.D., is a Senior Staff Member at Lincoln Laboratory, Massachusetts Institute of Technology. He is a Fellow of the Institute of Navigation.

Per Enge, Ph.D., is a Professor of Aeronautics and Astronautics at Stanford University, and Director of the Stanford Center for Position, Navigation and Time. Winner of the Institute of Navigation's Kepler award in 2000 for "sustained and significant contributions to satellite navigation," he is a Fellow of the Institute of Navigation, a Fellow of the IEEE, and a member of the National Academy of Engineering.

ISBN 0-9709544-1-7

9 780970 954411