REFERENCE

INERTIAL NAVIGATION PRIMER

VectorNav Library

VECTORNAV

CONTENTS

1 THEORY OF OPERATION

The navigation industry has a rich history of discovering new methods to meet clients' technical and economic demands. As manufacturing capabilities have developed over time, demand for both economical navigation systems and highly precise systems has risen. Today, there exist several solutions of varied price, reliability, and scale which offer advantages to different markets. Most significantly, the advancement of inertial sensors, satellite systems, and gyroscopes has expanded the viability of navigation systems across a multitude of applications.

1.1 INERTIAL SENSORS

Inertial sensors have been used for over a century to measure the motion of an object with respect to an inertial reference frame. When first created, inertial sensors consisted of large mechanical gyroscopes and accelerometers. However, since the early 1930s, numerous scientists, engineers, and institutions have been perfecting this technology, creating a wide variety of inertial sensors with various performance capabilities which has allowed them to be used today in more applications than ever before.

1.1.1 Grades of Inertial Sensors

The inertial sensor market spans an enormous range in terms of product price and performance between the highest-end inertial systems and the lowest. This wide variety can lead to great confusion for many customers when determining component selection and pricing. In addition, there are no agreed-upon definitions or standards of high-, medium-, and low-grade performance when it comes to inertial sensors, so what one expert considers high-end may be the low-end to another. However, in general, inertial sensors can be grouped into one of the following four performance categories:

- Navigation Grade
- Tactical Grade
- Industrial Grade
- Automotive/Consumer Grade

These performance categories are typically defined based on the in-run bias stability of the sensor, as the in-run bias stability plays such a large role in determining inertial navigation performance.

As shown in Figure 1.1, there are two main types of accelerometers that make up the different accelerometer performance categories: mechanical accelerometers and quartz/MEMS accelerometers. Quartz and MEMS accelerometers typically have an in-run bias stability ranging anywhere from 1000 µg down to 1 µg and span all four of the performance categories, while mechanical accelerometers can reach in-run bias stabilities less than 1 µg but are generally only used in navigation grade applications due to their large size and cost.

There are many different types of gyroscopes available on the market, which range over various levels of performance and include mechanical gyroscopes, fiber-optic gyroscopes (FOGs), ring laser gyroscopes (RLGs), and quartz/MEMS gyroscopes as illustrated in Figure 1.2. Quartz and MEMS gyroscopes are typically used in the consumer grade, industrial grade, and tactical grade markets, while fiber-optic gyroscopes span all four of the performance categories. Ring laser gyroscopes typically consist of in-run bias stabilities ranging anywhere from 1 °/hr down to less than 0.001 °/hr, encompassing the tactical and navigation grades. Mechanical gyroscopes make up the highest performing gyroscopes available on the market and can reach in-run bias stabilities of less than 0.0001 °/hr.

Accelerometers and gyroscopes can be used as individual inertial sensors, but most applications combine these sensors together into an inertial system. When a gyroscope is used in conjunction with an accelerometer, the performance of the gyroscope typically has the greater impact on the inertial navigation performance. Due to this, the

Accelerometer Performance Grades

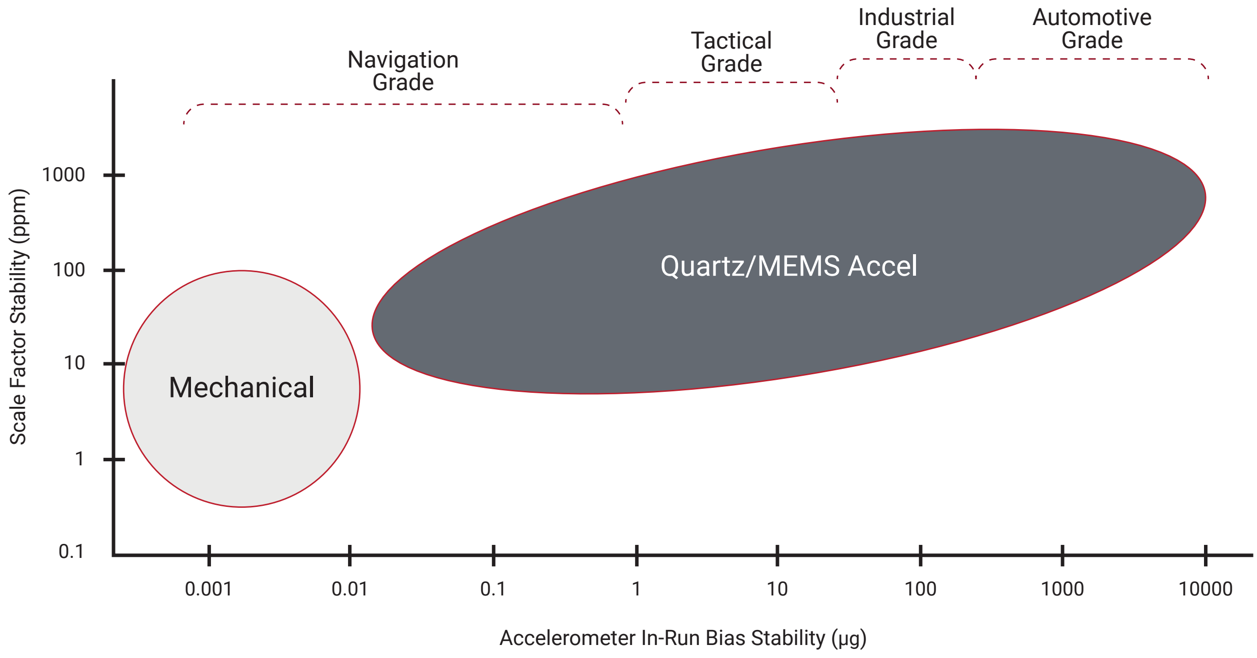


FIGURE 1.1

Gyroscope Performance Grades

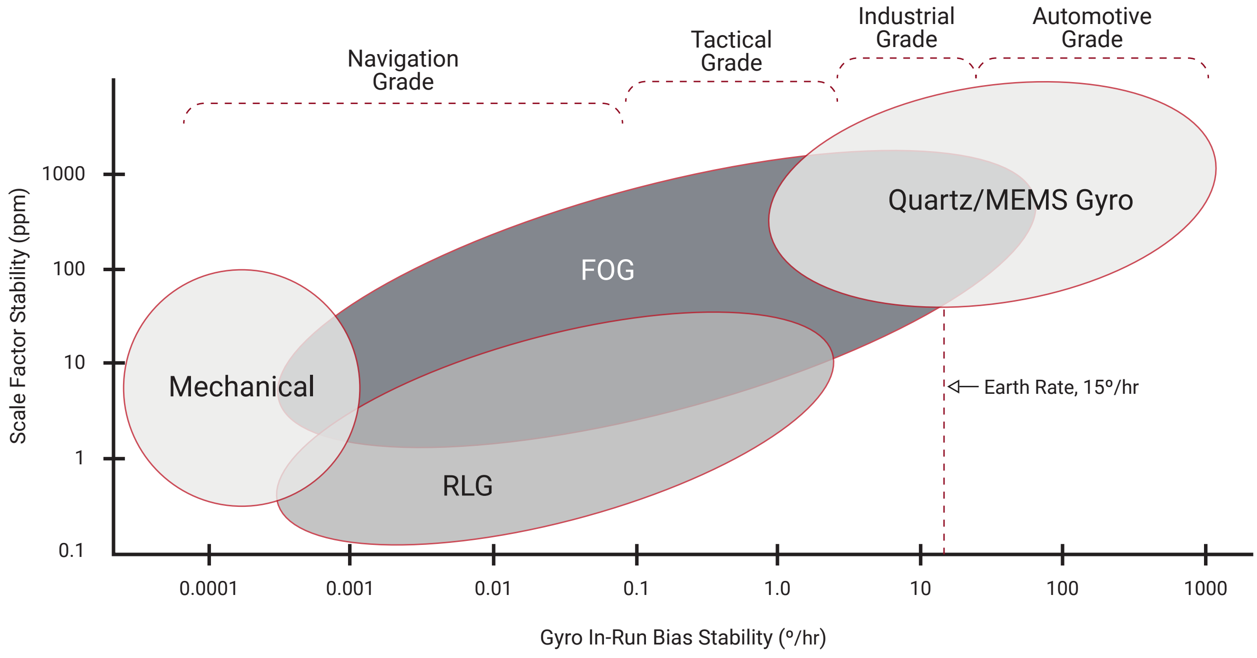


FIGURE 1.2

gyroscope in-run bias stability is often used as a short-hand measure of inertial system quality. Table 1.1 provides the typical gyroscope in-run bias stabilities associated with each of the different inertial sensor performance grades.

Performance Grades of Inertial Sensors

GRADE	COST	GYRO IN-RUN BIAS STABILITY	GNSS-DENIED NAVIGATION TIME	APPLICATIONS
Consumer	<$10	–	–	Smartphones
Industrial	$100-$1000	<10 °/hr	<1 min	UAVs
Tactical	$5,000-$50,000	<1 °/hr	<10 min	Smart Munitions
Navigation	>$100,000	<0.1 °/hr	Several hours	Military

TABLE 1.1

When determining which grade of inertial sensor is the best fit for a specific application, it is important to know what type of accuracy is required as well as the budget constraints for the project. As the performance grades of the inertial sensors increase, so does their associated cost, as shown in Table 1.1. The low cost of consumer grade inertial sensors makes them ideal for use in smartphones and tablets. Navigation grade inertial sensors, on the other hand, have a much higher accuracy, but also have a much higher cost making them practical only in the most mission-critical applications.

1.1.2 Inertial Sensing Nomenclature

As the inertial sensor market has grown, the use of inertial sensors and consequently the different nomenclature used to describe inertial sensors has expanded as well. The plethora of technical terms and acronyms used in company literature and information available on the internet makes it difficult to determine which combination of inertial sensors is best suited for a specific application. Furthermore, many of these terms are used interchangeably, so it is important to understand the components that make up each of these systems as well as the calculated navigation outputs that each provides. Some of the more common inertial sensing nomenclature is described in Table 1.2, note that the sensors listed in parentheses may or may not be included in the inertial system.

Inertial Sensing Nomenclature

ACRONYM	NAME	SENSORS	NAVIGATION OUTPUTS
IMU	Inertial Measurement Unit	Gyro + Accel + (Mag)	None
IRU	Inertial Reference Unit	Gyro + Accel + (Mag)	None
INS	Inertial Navigation System	Gyro + Accel + (Mag)	Position, Velocity, Attitude
VRU	Vertical Reference Unit	Gyro + Accel	Pitch, Roll & Heave
AHRS	Attitude Heading Reference System	Gyro + Accel + Mag	Attitude
MRU	Motion Reference Unit	Gyro + Accel + (Mag)	Varies

TABLE 1.2

When referring to the inertial systems listed in Table 1.2, it is common to describe them based on the number of total axes that the system measures. An individual inertial sensor can only sense a measurement along or about a single axis. To provide a three-dimensional solution, three individual inertial sensors must be mounted together into an orthogonal cluster known as a triad. This set of inertial sensors mounted in a triad is commonly referred to as a 3-axis inertial sensor, as the sensor is able to provide one measurement along each of the three axes. Similarly, an inertial system consisting of a 3-axis accelerometer and a 3-axis gyroscope is referred to as a 6-axis system as it provides two different measurements along each of the three axes for a total of six measurements. Table 1.3 defines a few of the more common sensor combinations.

1.2 GLOBAL NAVIGATION SATELLITE SYSTEM (GNSS)

A Global Navigation Satellite System (GNSS) is a satellite configuration, or constellation, that provides coded satellite signals which are processed by a GNSS receiver to calculate position, velocity, and time. GNSS is a passive system, meaning that there is no limit to the number of users allowed to utilize its technology, making it available for anyone to use around the world.

N-Axis Inertial Sensors

NUMBER OF AXES	ACCELEROMETER	GYROSCOPE	MAGNETOMETER	BAROMETER
6-Axis	3-Axis	3-Axis	–	–
9-Axis	3-Axis	3-Axis	3-Axis	–
10-Axis	3-Axis	3-Axis	3-Axis	1-Axis

TABLE 1.3

In 1978, the first satellite for a navigation system was launched by the United States. This led to a fully operational constellation of 24 satellites known as the NAVSTAR Global Positioning System in the early 1990s. Today this system is known simply as the Global Positioning System, or GPS, and contains 31 satellites in its constellation.

1.2.1 Constellations

Since the U.S. launched the first operational global navigation satellite system, several other nations have launched similar GNSS constellations. Some of these systems are currently available for use while others will be fully operational in the coming years, as shown in Table 1.4.

International GNSS Constellations

NAME	COUNTRY OF ORIGIN	FULLY OPERATIONAL	NUMBER OF SATELLITES	CARRIER FREQUENCIES
GPS	USA	1993	31	L1/L2/L5
GLONASS	Russia	1995	24+	G1/G2
Galileo	Europe	2020	30 (22 current)	E1/E5a/E5b
BeiDou	China	2020	30 (28 current)	B1/B2
QZSS	Japan	2024	7 (4 current)	L1/L2/L5

TABLE 1.4

1.2.2 Segments

GNSS operates through three different segments known as the Space Segment, the Ground Control Segment, and the User Segment, as shown in Figure 1.3a. The Space Segment consists of the satellites themselves placed into a specific constellation, as seen in Figure 1.3b. The Ground Control Segment utilizes Earth-based tracking stations around the world to manage the entire navigation system. Specific locations of these stations for the U.S.-based system, GPS, are shown in Figure 1.4. The User Segment is comprised of the GNSS receivers that can be used anywhere around the world.

The Ground Control Segment tracks and monitors errors and biases in the satellite's orbit, clock, and health. This information is sent through radio signals up to the Space Segment. It is through this process of the Ground Control Segment tracking the orbits of the Space Segment and uploading orbit corrections to them that the satellites are able to know to a high precision where they are located. The satellites then transmit that information back down to the User Segment where they are tracked, decoded, and utilized in determining a user's position, velocity, and time.

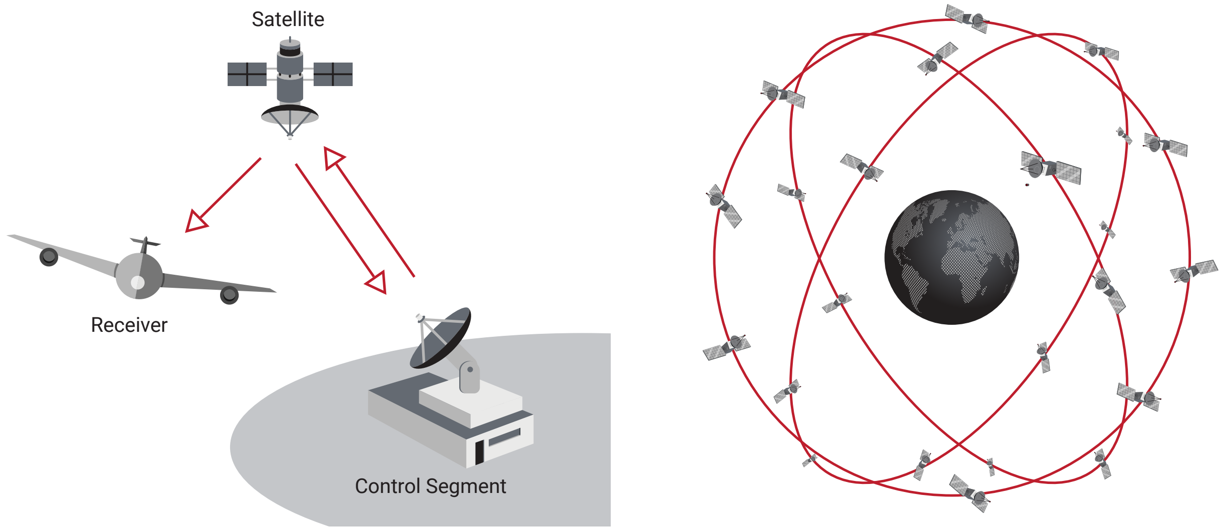
1.2.3 Navigation Message

The signals that the Ground Control Segment sends to the satellites which then get sent to the end user are known as the navigation message. The GPS navigation message contains four main parts: GPS time, satellite health, ephemeris, and the almanac. While this discussion is specific to the GPS constellation, the basic features exist across all GNSS constellations.

GPS Time

The GPS time in the navigation message is based on an atomic clock that is able to keep time to a high degree of accuracy. It is specified in terms of the week number and the seconds of the week. The week number is a counter that designates the number of weeks that have passed since January 6, 1980, or Week 0. However, this counter can only store values from 0 to 1,023, so once Week 1,024 was reached on August 21,1999 and then again on April 6, 2019, the week number was rolled back to 0. This rollover cycle of the week number will continue to repeat every 1,024 weeks. The seconds of the week is the number of seconds into the current week, starting Sunday at 12:00 A.M. GMT.

Global Navigation Satellite System (GNSS)



(a) GNSS Segments

(b) GNSS Satellite Constellation

FIGURE 1.3

GNSS Ground Control Stations

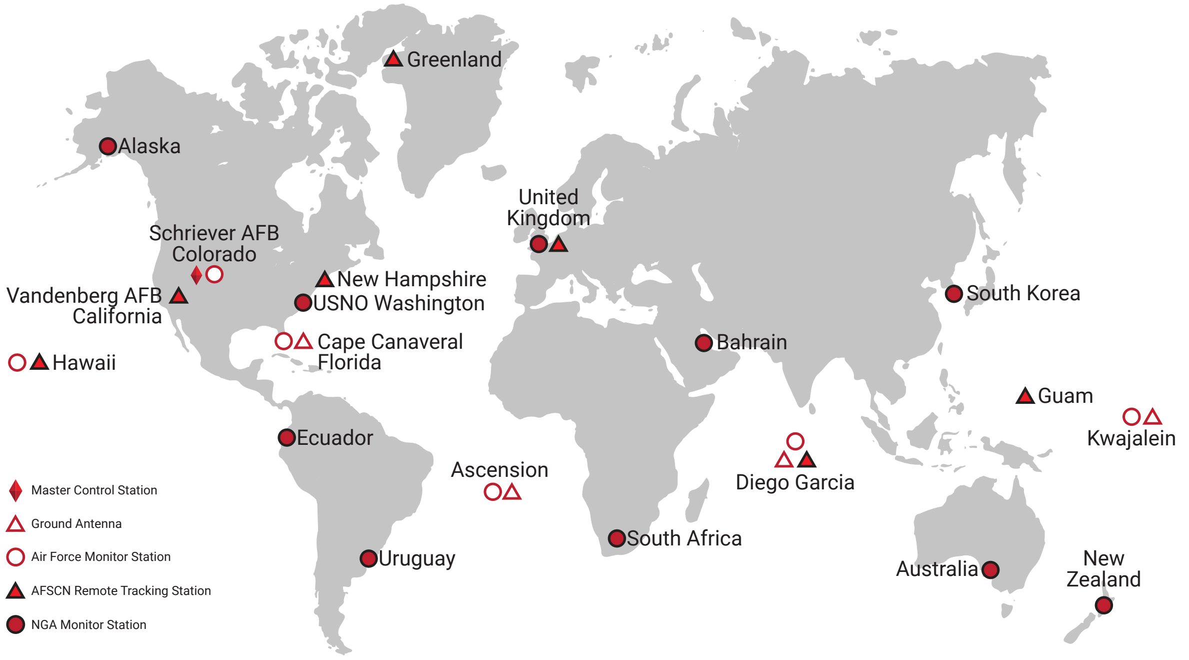


FIGURE 1.4

Satellite Health

The satellite health information conveys to a GPS receiver whether the satellite is healthy and the navigation data it is transmitting can be trusted. If a satellite is deemed to be ”healthy”, the navigation data transmitted by the satellite is considered usable. However, a satellite that is considered ”unhealthy” contains navigation data that is either partially or fully unusable.

Ephemeris

The ephemeris contains high-accuracy orbit data specific to the satellite that is transmitting the navigation message. This data is only considered useable for up to four hours from the time it was uploaded to the satellites by the Ground Control Segment. Therefore, the ephemeris is updated for each of the satellites every four hours by the Ground Control Segment. Fortunately, downloading the full ephemeris from a satellite only takes a GPS receiver approximately thirty seconds.

Almanac

The almanac is a collection of low-accuracy ephemerides for every satellite in the Space Segment constellation. This library of data is updated much less frequently than the ephemeris and takes a GPS receiver about 12.5 minutes to download. Since the almanac contains low accuracy ephemerides, receivers use this information mainly for determining which satellites will soon be visible on the horizon to track. The almanac also contains leap second information which is needed to convert GPS time to Coordinated Universal Time (UTC), as UTC lags GPS time by the number of leap seconds.

1.2.4 Pseudorange, Carrier Phase, Doppler

There are three raw observables that a GNSS receiver tracks: pseudorange, carrier phase, and Doppler.

Pseudorange

In order to determine the range from the satellite to the user, a GNSS receiver measures the time required for a signal to travel from a satellite down to the receiver. Since the signal is traveling at the speed of light, the product of the signal time of travel measured by the receiver (t) and the speed of light (c) equals the range ($r = t \cdot c$).

This measurement, however, relies on high-accuracy timing. Receivers use low-end clocks for timing, not atomic clocks, resulting in an unknown bias from the true GPS time. Due to this clock bias error, receivers are not measuring the true range to the satellite, but rather a pseudorange (ρ). The pseudorange is the basis for calculating a user's position and time.

Carrier Phase

A signal transmitted from a satellite contains a sinusoidal signal called the carrier wave. While the signal contains no information itself, it carries other signals containing information that have been modulated on top of it. The distance from a satellite to a receiver can be broken into an integer number of full wavelengths of the carrier signal plus a fractional wavelength. This fractional wavelength is known as the carrier phase and can be directly measured. Though a standalone receiver cannot estimate the integer number of wavelengths, the carrier phase can be used in a multi-receiver technique, known as RTK (see Section 1.5), to enable high-precision positioning.

Doppler

As a GNSS receiver is receiving and tracking a signal from a satellite, the frequency of the signal appears to shift due to the combined motion of the user and of the satellite orbiting around the Earth. This shift in frequency can be used to determine a relative speed. Multiple Doppler shift measurements are able to produce an actual velocity for the user.

1.2.5 Position, Velocity & Time (PVT)

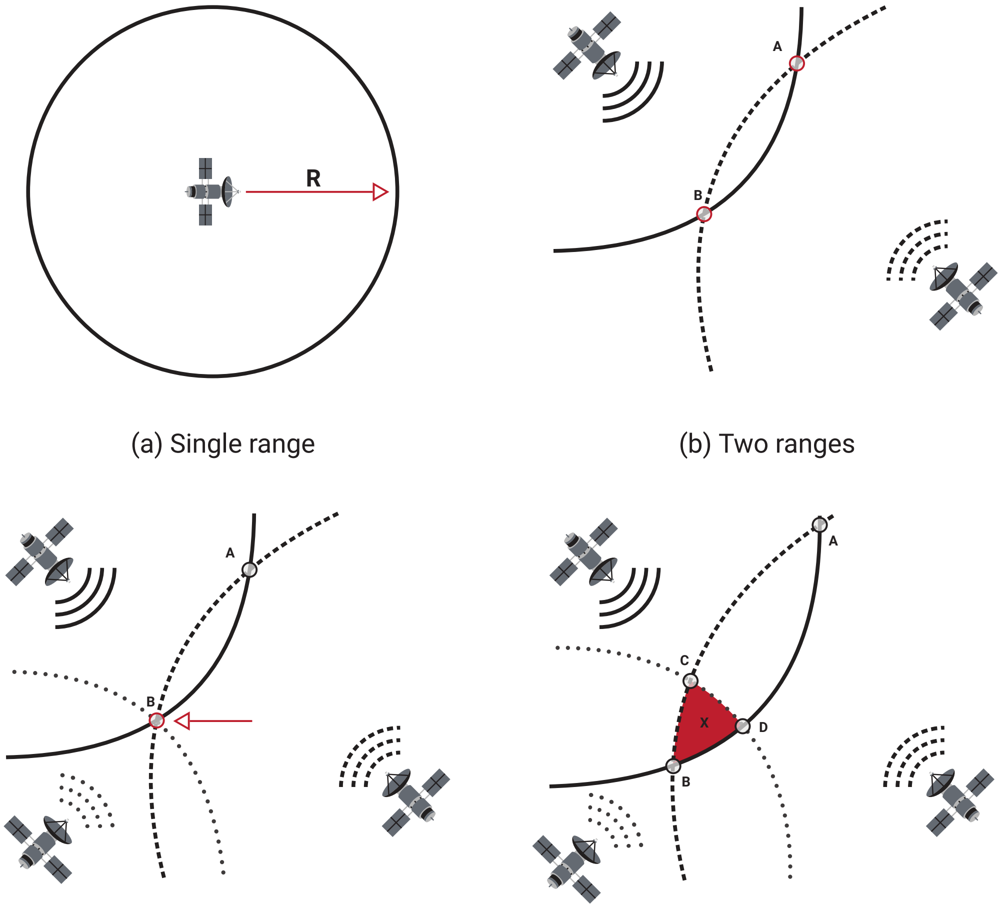
The information from the navigation message and the data in the raw observables can be used to determine the position, velocity, and time (PVT) of a GNSS receiver.

Trilateration

While a range measures the distance between a satellite and a user, this measurement on its own does not provide a user's position. However, if range measurements to multiple satellites are able to be determined, a method known as trilateration can then be used to estimate a user's position.

Trilateration uses a range measurement from the satellite to the user to create a region encompassing all of the possible positions of the user. In the case of 3D positioning, this possible region is a sphere of radius equal to the range measurement, centered at the location of the satellite, as seen in Figure 1.5a. Once an additional range measurement to another satellite has been determined, the possible position of the user can be reduced down to the circle where the spheres intersect, as seen in Figure 1.5b.

GNSS Trilateration



(a) Single range

(b) Two ranges

(c) Three ranges

(d) Three imperfect ranges

FIGURE 1.5

In order to determine the estimated position of the user down to a single point, a minimum of three range measurements to three different satellites must be calculated. These three range measurements provide three different spheres of possible positions the user could be and all intersect at one point, as shown in Figure 1.5c. In practice, each of those measurements is imperfect, due to a variety of errors, and an estimated position is calculated from the best fit of those measurements, as seen in Figure 1.5d.

While that is the standard definition of trilateration, in the case of GNSS, a receiver does not measure a true range to the satellite but rather a pseudorange due to the clock bias error. In fact, a fourth pseudorange measurement is needed to determine the estimated position and estimated clock bias simultaneously. Note that the delays due to the length of cable between an antenna and its receiver are also accounted for when estimating clock bias, so cable length does not impact the positioning results.

Pulse Per Second (PPS)

GNSS receivers use a lower accuracy clock that is then disciplined to GPS time using the specific timing messages in the signals sent from the satellites to the receivers. Once in sync with GPS time, the receiver can output a pulse per second signal, or PPS, on the top of every second in GPS time. Because the top of a second in GPS time is the same as the top of a second in UTC time, this output can be used in many different timing applications, with accuracy on the order of 10s of nanoseconds.

Time-to-First-Fix (TTFF)

The time-to-first-fix is the duration of time needed for a GNSS receiver to acquire signals from the satellites, perform trilateration, and obtain a position solution, sometimes referred to as a GNSS fix. This length of time depends upon how the GNSS receiver is started up. A receiver can be started using either a cold start, a warm start, or a hot start. As shown in Table 1.5, these three different types of start ups have different amounts of information available to the receiver to use in its process of acquiring a GNSS fix.

A cold start takes the longest amount of time to obtain a GNSS fix as the receiver possesses no information regarding where the satellites are located and must complete the 30-second download of ephemeris data. A warm start takes less time than a cold start because it already has valid almanac data, however, it is not much quicker, as it still must wait to obtain the ephemeris data. A hot start takes the least amount of time, typically just a few seconds, since the receiver already has valid almanac data, ephemeris data, and time.

Types of GNSS Receiver Starts

START-UP	STORED RECEIVER DATA	EXAMPLE
Cold Start	No information available	GNSS receiver is first turned on, or it has been a long time since the receiver was last turned on.
Warm Start	Receiver only has valid almanac data	Turning on the receiver when it has been turned off for one day.
Hot Start	Receiver has valid almanac data, ephemeris data, and time	Turning on the receiver when it has been turned off for less than four hours. Requires continuous backup power to maintain clocks.

TABLE 1.5

1.3 MEMS OPERATION

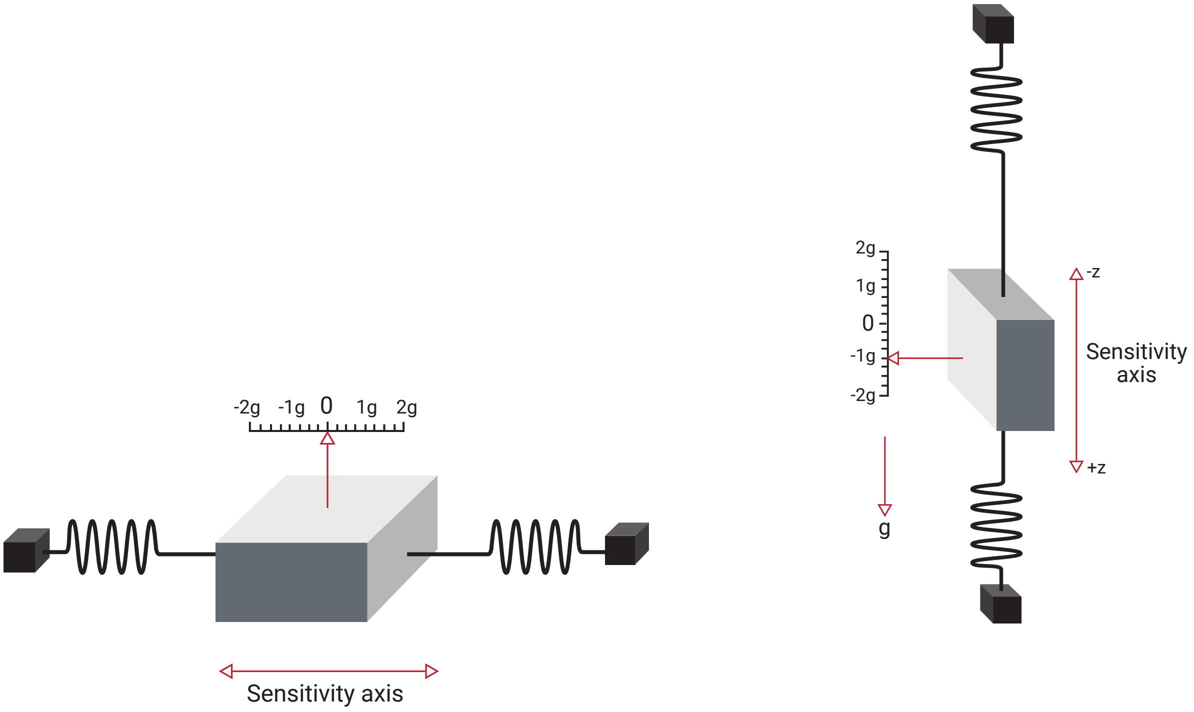
Up until the emergence of microelectromechanical systems (MEMS) technology, inertial sensors were high-cost, precision instruments, typically reserved for high-end applications. As MEMS technology has matured, low-cost solid-state chip level inertial sensors have become available as alternatives to the larger high-end inertial sensors. This addition of MEMS to the inertial sensing market has provided a wide variety of performance capabilities and allowed inertial sensing technology to be used in more applications than ever before.

1.3.1 MEMS Accelerometers

An accelerometer is the primary sensor responsible for measuring inertial acceleration, or the change in velocity over time, and can be found in a variety of different types, including mechanical accelerometers, quartz accelerometers, and MEMS accelerometers. A MEMS accelerometer is essentially a mass suspended by a spring, as illustrated in Figure 1.6a. The mass is known as the proof mass and the direction that the mass is allowed to move is known as the sensitivity axis.

When an accelerometer is subjected to a linear acceleration along the sensitivity axis, the acceleration causes the proof mass to shift to one side, with the amount of deflection proportional to the acceleration.

Simple Accelerometer Model



(a) Horizontal

(b) Vertical

FIGURE 1.6

Now consider that the accelerometer is rotated such that the sensitivity axis is aligned with the gravity vector, as shown in Figure 1.6b. In this case, gravity acts on the proof mass causing it to deflect downward. Due to this, the accelerometer measures both the linear acceleration due to motion as well as the pseudo-acceleration caused by gravity. The acceleration caused by gravity is referred to as a pseudo-acceleration as it does not actually result in a change in velocity or position.

In the coordinate frame shown in Figure 1.6b, the pseudo-acceleration caused by gravity is measured as a −1 g, as gravity has the same effect on the accelerometer as an acceleration due to motion in the negative z-axis. It is also important to note that during free fall, the springs in the accelerometer do not deflect, and consequently the sensor reports an acceleration of zero, though the actual acceleration is non-zero.

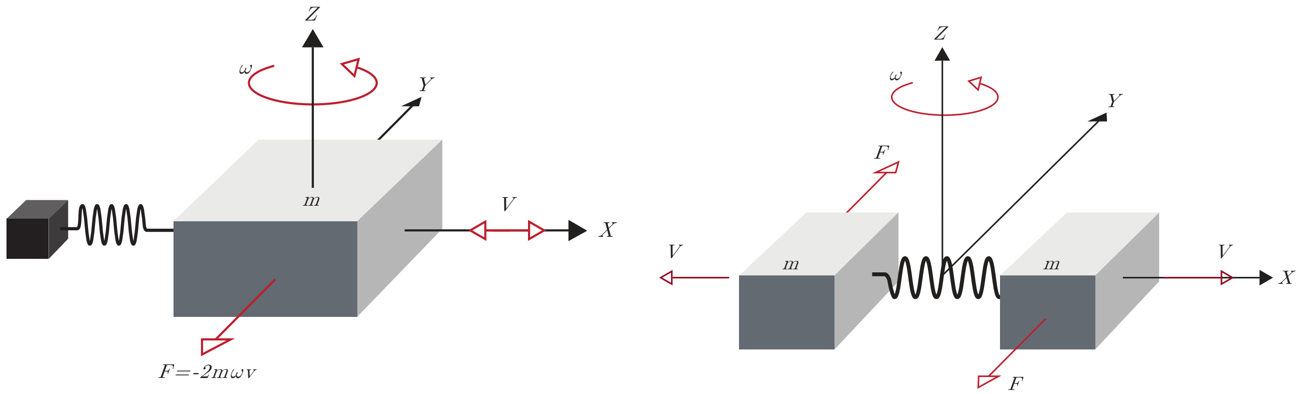
1.3.2 MEMS Gyroscopes

A gyroscope is an inertial sensor that measure an object's angular rate with respect to an inertial reference frame. MEMS gyroscopes measures the angular rate by applying the theory of the Coriolis effect, which refers to the force of inertia that acts on objects in motion in relation to a rotating frame. To better understand, consider a mass suspended on springs, as illustrated in Figure 1.7a. This mass has a driving force on the x-axis causing it to oscillate rapidly in the x-axis. While in motion an angular velocity, ω, is applied about the z-axis. This results in the mass experiencing a force in the y-axis as a result of the Coriolis force, and the resultant displacement is measured by a capacitive-sensing structure.

A quick derivation of this Coriolis force may provide further clarity. The position of the mass, m, in the body frame is given by Equation 1.1:

$$^{B}\boldsymbol{r} = \begin{bmatrix} x \\ y \end{bmatrix} \tag{1.1}$$

Simple Gyroscope Model



(a) Single Mass

(b) Tuning Fork Configuration

FIGURE 1.7

The inertial velocity of the mass in the body frame is then defined as the derivative of the position plus the tangential velocity due to rotation.

$$ {}^B\dot{\boldsymbol{r}} = \begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} + {}^B\boldsymbol{\omega} \times {}^B\boldsymbol{r} = \begin{bmatrix}\dot{x} - \omega y\\ \dot{y} + \omega x\end{bmatrix} \qquad (1.2) $$

The inertial acceleration of the mass in the body frame can be described as the derivative of the velocity plus the tangential acceleration due to rotation.

$$ {}^B\ddot{\boldsymbol{r}} = \begin{bmatrix}\ddot{x} - \omega\dot{y}\\ \ddot{y} + \omega\dot{x}\end{bmatrix} + {}^B\boldsymbol{\omega} \times {}^B\dot{\boldsymbol{r}} = \begin{bmatrix}\ddot{x} - 2\omega\dot{y} - \omega^2 x\\ \ddot{y} + 2\omega\dot{x} - \omega^2 y\end{bmatrix} \qquad (1.3) $$

The first element in Equation 1.3 represents the acceleration experienced by the driven axis, which is actively controlled by the gyroscope's electronics. The second element in Equation 1.3 represents the acceleration from the sensing axis of the gyroscope. From Newton's Second Law of Motion, the sum of the forces in the sensing direction is equal to the product of the mass of the block, m, and the acceleration in the sensing direction, $\ddot{r}_y$:

$$ F_y = m^B\ddot{r}_y = m(\ddot{y} + 2\omega\dot{x} - \omega^2 y) \qquad (1.4) $$

For illustrative purposes, if the mass starts from rest in the y-axis ($y = \dot{y} = \ddot{y} = 0$), the sum of the forces in the y-axis reduces down to only the Coriolis term, $F_y = 2m\omega\dot{x}$. Since the mass is driven in the x-axis at high frequency (10s of kHz), the value of $\dot{x}$ is significant and the Coriolis effect causes significant, oscillatory displacement in the y-axis proportional to the angular rate.

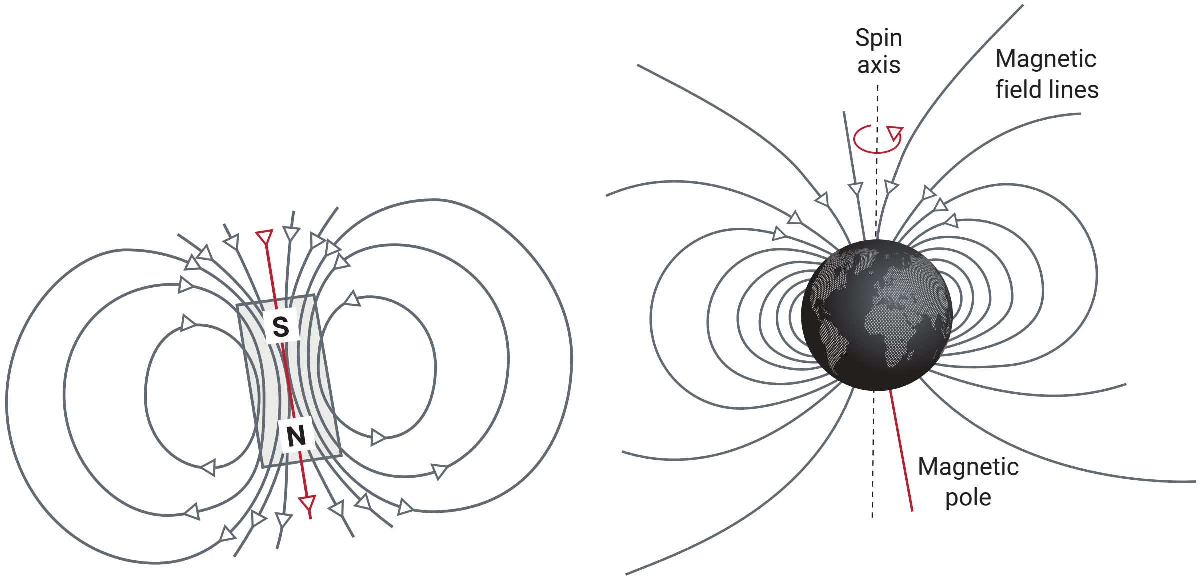
Typically, MEMS gyroscopes use a tuning fork configuration in which two masses are connected by a spring, as shown in Figure 1.7b. When an angular rate is applied, the Coriolis force on each mass acts in the opposite direction and the resulting change in capacitance is directly proportional to the angular velocity. However, when a linear acceleration is applied, the two masses moves in the same direction, resulting in no change in capacitance and a measured angular rate of zero. This configuration minimizes a gyroscope's sensitivity to linear acceleration from instances of shock, vibration, and tilt.

1.3.3 MEMS Magnetometers

A magnetometer is a type of sensor that measures the strength and direction of a magnetic field. While there are many different types of magnetometers, most MEMS magnetometers rely on magnetoresistance to measure the surrounding magnetic field. Magnetoresistive magnetometers are made up of permalloys that change resistance due to changes in magnetic fields. Typically, MEMS magnetometers are used to measure a local magnetic field which consists of a combination of Earth's magnetic field as well as any magnetic fields created by nearby objects.

As illustrated in Figure 1.8, Earth's magnetic field is a self-sustaining magnetic field that resembles a magnetic dipole with the geomagnetic poles slightly offset from the geographic North and South poles. This magnetic field is characterized by a strength and direction, which varies across the earth and can shift over time.

Dipole Approximation of Earth's Magnetic Field



(a) Standard dipole magnet

(b) Earth's magnetic field

FIGURE 1.8

The direction of Earth's magnetic field contains a horizontal component as well as a vertical component and is often described using the magnetic inclination and declination angles. Magnetic inclination describes the angle between Earth's magnetic field lines and a horizontal plane. At Earth's magnetic poles the magnetic field is vertical and has an inclination angle of 90°, whereas Earth's magnetic field is horizontal at the equator and has an inclination angle of 0°. The magnetic declination is used to account for the fact that the magnetic North Pole of the earth is not in the same location as True North or the geographic North Pole of the earth and is characterized as the angle between these two locations, relative to the point of measurement.

1.4 HIGH-PERFORMANCE GYROSCOPES

In guidance, navigation, and control (GNC) systems there is a sometimes need for high-performance or high-end gyroscopes. Such gyros provide unique capabilities in unaided navigation performance and heading determination through gyrocompassing that are impossible with today's MEMS gyro technology. The most prevalent of such high-end gyroscopes are optical gyros, used in applications with the most stringent performance requirements.

1.4.1 Optical Gyroscopes

Optical gyros rely on the Sagnac effect to measure angular rate. Because there are no mechanical moving parts, they generally lack sensitivity to vibration. There are primarily two types of optical gyroscopes: Ring Laser Gyros (RLGs) and Fiber Optic Gyros (FOGs).

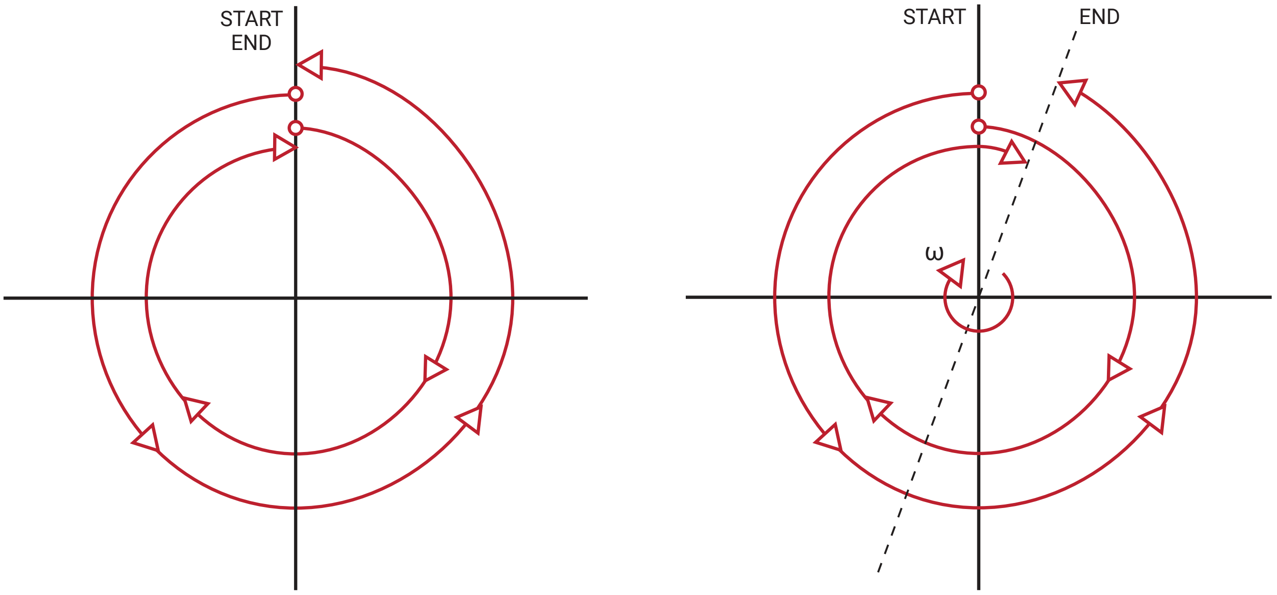
Sagnac Effect

Optical gyroscopes use beams of light travelling in opposite directions around a closed-loop ring to measure a system's angular rate. Obviously, those two light beams travel exactly the same distance in a non-rotating system, as seen in Figure 1.9a. But since light travels at a constant speed relative to an inertial frame of reference, if a system is subjected to an angular rate, one light beam travels a longer distance than the other, as observed in Figure 1.9b. When the two beams are brought together again, this gives rise to interference dependent on the amount of rotation—a phenomenon known as the Sagnac effect.

Ring Laser Gyroscope

A ring laser gyroscope (RLG) is a high-performance optical gyroscope that uses the Sagnac effect to detect rotation. As seen in Figure 1.10a, an RLG utilizes a closed-loop laser cavity, typically filled with helium-neon gas, to perform its measurements. The laser itself is integrated within the chamber, making the externally observed interference pattern directly proportional to the rotation angle. These gyroscopes are the highest-performance available, which combined their complexity, makes them the most expensive as well.

Sagnac Effect Gyro Measurement

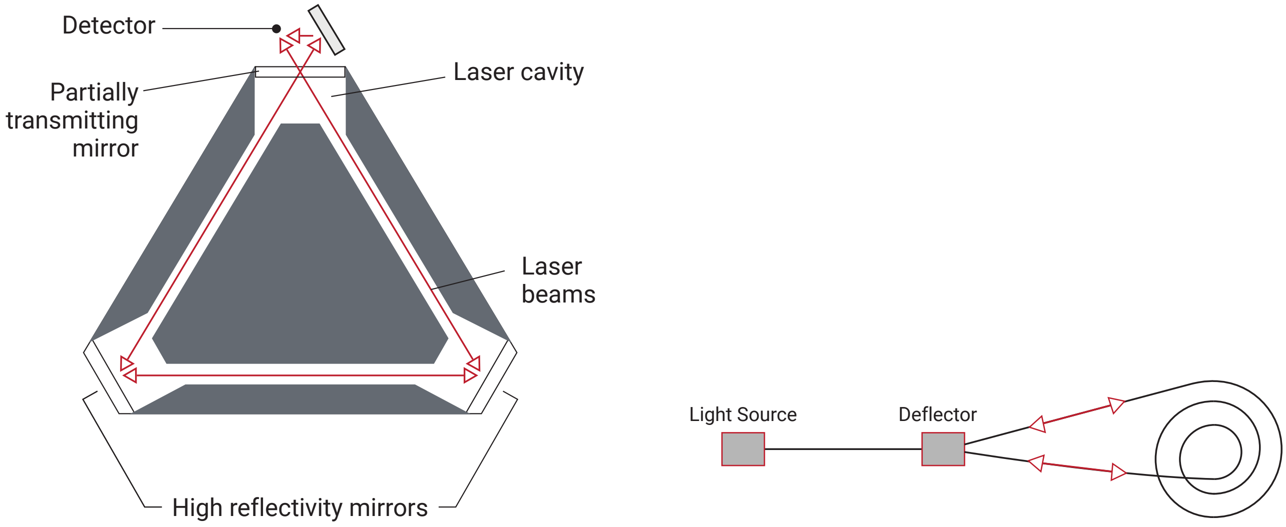


(a) Stationary

(b) Rotating

FIGURE 1.9

Optical Gyroscopes



(a) Ring Laser Gyro (RLG)

(b) Fiber Optic Gyro (FOG)

FIGURE 1.10

Fiber-Optic Gyroscope

A fiber-optic gyroscope (FOG) is a high-performance optical gyroscope that also implements the Sagnac effect in its calculations to detect rotation. The FOG uses a laser source, beam splitters, a detector, and fiber-optic coil as shown in Figure 1.10b. It utilizes a light source that splits into two wavelengths that travels through the optical fiber in opposite directions. Once the beams reach the detector, it can determine the rotation rate through the Sagnac effect. The sensitivity and performance of a FOG can vary depending on the coil diameter and number of turns it has, with performance directly correlated to the length of the fiber (some use >1 km of fiber in the coil). Fiber-optic gyros are a more recent technology than RLGs and take advantage of existing, lower-cost technologies, yielding much better pricing, though at somewhat reduced performance relative to the RLG.

1.4.2 Gyrocompassing

Gyrocompassing is the ability of a high-performance gyroscope to determine heading without external aiding. A gyrocompass detects True North by directly measuring Earth's angular rate as it spins on its axis, seen in Figure 1.11. Using an accelerometer to measure the direction of gravity, Earth's angular rate (Ω_E) can be decomposed into horizontal (ω_N) and vertical (ω_D) components, with the horizontal component pointing due North. The direction of that horizontal component with respect to the sensor axes provides the heading (ψ).

Achieving accurate heading via gyrocompassing requires a particularly low-noise sensor with superior bias stability. The Earth rotates at approximately 15 °/h, with the horizontal component equal to that times the cosine of latitude (Φ). At a 45° latitude, an error as small as 0.1 °/h in the angular rate measurement results in a 0.5° heading error. The size, weight, power, and cost (SWAP-C) of gyros capable of gyrocompassing is often prohibitive, but it remains the single most reliable method of heading determination, entirely self-contained to the inertial sensors.

Gyrocompassing

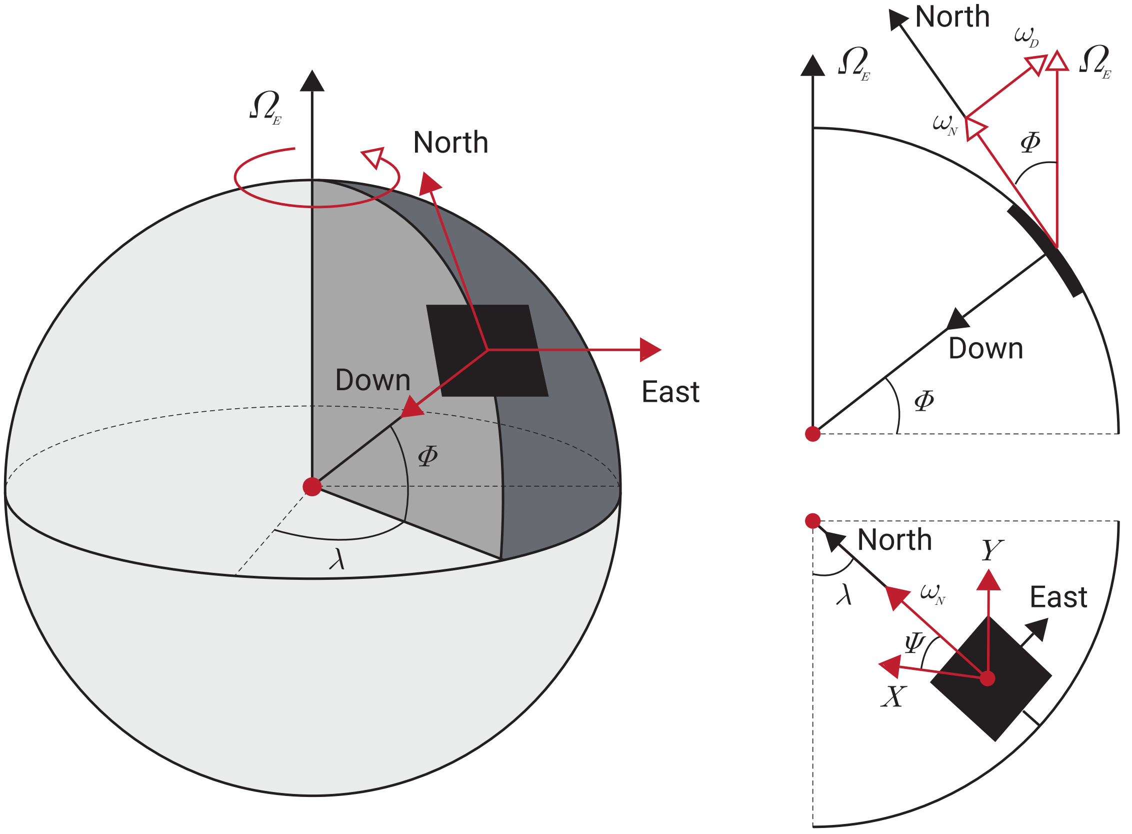


FIGURE 1.11

1.4.3 Unaided Inertial Navigation

In addition to gyrocompassing, the main advantage of high-end gyroscopes is their performance in unaided inertial navigation, also known as dead-reckoning. As described in Section 3.3, the performance of the gyro typically dominates the position errors when performing a pure integration of inertial sensors. While they obviously provide improved navigation performance in aided situations (eg. GNSS-aided), the improvement in performance relative to price is not usually warranted.

1.5 ADVANCED GNSS

While trilateration from pseudoranges can be used to calculate a user's position as discussed in Section 1.2, this method on its own is typically only able to provide a positioning accuracy on the order of several of meters (see Section 3.4). This section describes various advanced GNSS techniques that can be employed to achieve a higher positioning accuracy, and the signals that make them possible.

1.5.1 Signals

In order to understand the advanced GNSS positioning techniques, it is helpful to know more about the signals used in these methods. The following discussion focuses on the GPS constellation, but similar signals exist across all GNSS constellations.

L1, L2, L5 frequencies

The Global Positioning System (GPS) currently operates on three carrier frequencies known as L1, L2, and L5, as shown in Table 1.6. They are located on the L-band in the radar section of the electromagnetic spectrum and were chosen due to their ability to penetrate most atmospheric obstructions such as clouds, fog, or rain. In addition, ionospheric delay is much smaller for the frequency range of 1 GHz to 10 GHz, and frequencies below 2 GHz can be received without need for a beam antenna. Note that the L5 band is a recent development that is currently available on only a handful of satellites in the GPS constellation.

GPS Frequencies

NAME	FREQUENCY	APPLICATION
L1	1575.42 MHz	Civilian navigation
L2	1227.60 MHz	Military, but some civilian use
L5	1176.45 MHz	Precision guidance

TABLE 1.6

Carrier, Code (C/A and P(Y)), Navigation

The GPS signal from each satellite is composed of three components called the carrier, the pseudo-random noise (PRN) code, and the navigation message.

The PRN code signal gives each satellite a unique identifier which allows all satellites on the band to transmit on the same frequency without jamming. Coarse/acquisition (C/A) code is generated at 1.023 Mbps on the L1 band, while Precise code (P-code) is generated at 10.23 Mbps on L1 and L2. P-code is encrypted with W-code to become P(Y)-code, which acts as an anti-spoofing measure and is often used in military applications.

In an attempt to increase the security of military GPS systems, a code known as M-code has also been developed for the L1 and L2 bands. M-code provides additional protection against jamming due to its ability to be transmitted at a higher power. It is expected to eventually replace P(Y)-code, although uptake into the market has been slow.

The navigation message is a binary-coded representation of the GPS time, satellite health, ephemeris, and almanac data which the receiver uses to determine its position. Due to its slow speed of 50 bytes per second, the full message, broken into 25 frames, takes about 12.5 min to receive. Every frame, which each takes 30 s to transmit, includes the full GPS time and individual satellite health and ephemeris, allowing for a receiver to achieve a position solution before receiving the entire navigation message.

The code and navigation message are modulated onto a sinusoidal carrier signal as shown in Figure 1.12. The high frequency of the carrier signal allows transmission over the necessary distances and through adverse weather conditions on the way from the satellite to the receiver. The receiver then demodulates this signal to extract the original information.

1.5.2 Differential GPS (DGPS)

In order to overcome some inherent limitations of a single GPS receiver, multiple GPS receivers may be used in a technique known as Differential GPS, or DGPS.

Modulated GPS Signals

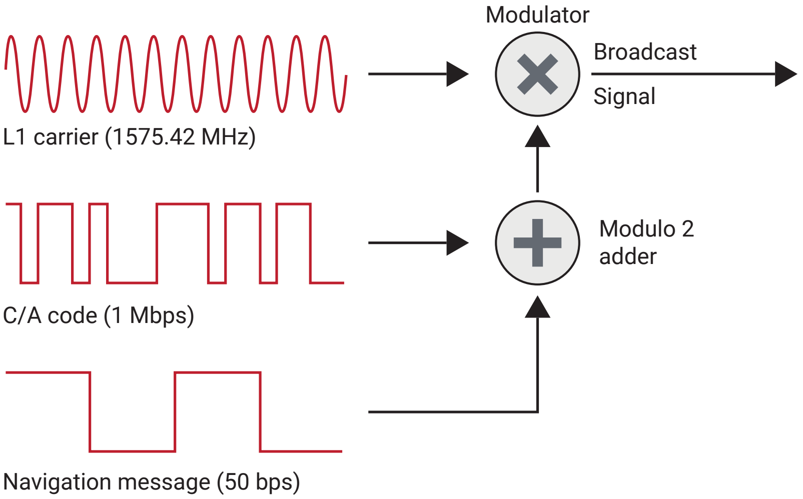


FIGURE 1.12

Local-Area Augmentation System

DGPS requires some form of radio link between the receivers, and that one receiver have a well-known GPS location to serve as the *base* station. The signal transmitted from a single satellite to both the base station and *rover* receiver experiences the same satellite clock errors, orbit errors, and atmospheric propagation errors if the distance between them is below roughly 20 km. The base station can then determine an estimate of how much error is present compared to its already known location and transmit a pseudorange correction to a local receiver. Local-area DGPS systems can achieve positioning accuracy down to the one-meter level.

Satellite Based Augmentation System (SBAS)

SBAS is a DGPS system where errors present at location-established reference stations are transmitted to a central location to compute differential corrections. These corrections are then broadcast over a region by geostationary satellites on the same L1 frequency as GPS, which receivers can track and use to obtain a positional accuracy of one to two meters. The map in Figure 1.13 illustrates the locations of SBAS systems in use, covering most of the Northern hemisphere. More information on each of these SBAS systems can be found in Table 1.7.

SBAS Systems Worldwide

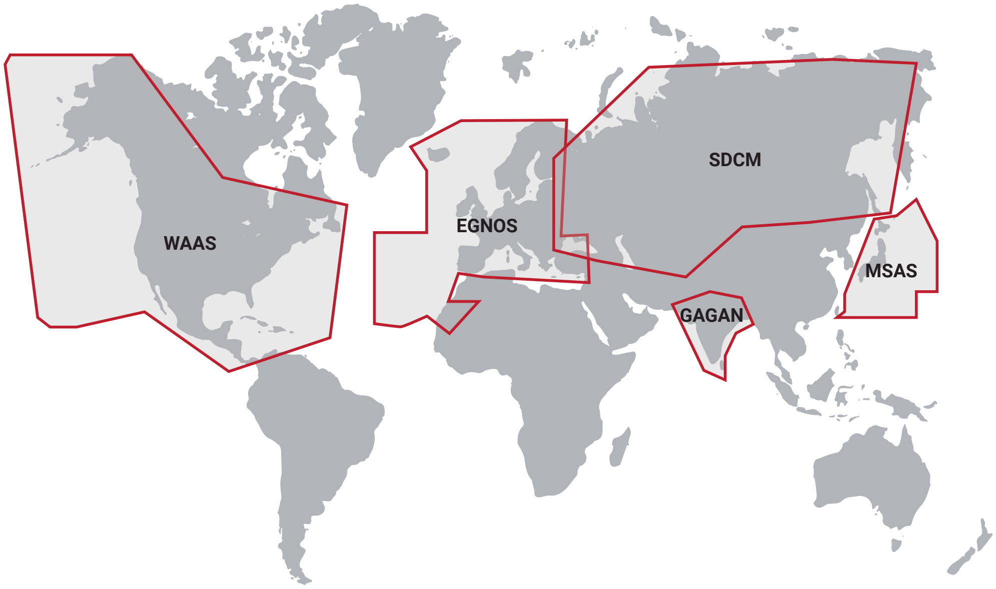


FIGURE 1.13

SBAS Systems Worldwide

REGION	NAME	ACRONYM	SATELLITES
North America	Wide-Area Augmentation System	WAAS	2
Europe	European Geostationary Navigation Overlay Service	EGNOS	3
Russia	System for Differential Corrections and Monitoring	SDCM	4
India	GPS-Aided GEO Augmented Navigation System	GAGAN	3
Japan	Multi-functional Satellite Augmentation System	MSAS	2

TABLE 1.7

1.5.3 Real-Time Kinematic (RTK) & Post-Processed Kinematic (PPK) Positioning

To achieve more precise location results, a multiple-receiver system called real-time kinematic (RTK) positioning was developed. Like differential GPS, RTK compares measurements between a base receiver and a rover receiver, but RTK relies on the carrier phase observables rather than the pseudoranges. Post-processed kinematic (PPK) positioning utilizes the exact same techniques and algorithms, but is computed offline on a PC, negating the need for computations to run in real time on the receiver or for a reliable real-time radio link for corrections.

Double-Differencing

Carrier phase measurements are incredibly precise measures of the partial wave tracked between a satellite and a receiver. The difficulty is that the integer number of full wavelengths between the satellite and receiver is impossible to determine in a standalone GNSS configuration, a problem known as the integer ambiguity. Using double-differencing algorithms between two receivers, a related form of the integer ambiguity problem can be solved, yielding relative position accuracy to within two centimeters.

Double-differencing requires two receivers and two satellites. By differencing the carrier phase data between two satellites on a single receiver, most receiver errors (eg. receiver clock bias, including delays due to cable length) can be eliminated. Meanwhile, differencing the measurements from a single satellite between two different receivers eliminates the satellite-related errors (eg. orbit errors, atmospheric delays). Differencing these two differences—*double-differencing*—yields a result where most error sources have been eliminated. This process is summarized graphically in Figure 1.14.

Single vs. Dual Frequency

An important consideration for RTK or PPK systems is whether the receivers are tracking a single frequency (L1) or dual frequencies (L1/L2). When tracking only a single frequency, the integer ambiguity problem yields a large number of feasible solutions, making it difficult for a reliable RTK fix to be achieved and maintained. Utilizing an INS in combination with GNSS allows for much more robust tracking of the correct RTK fix in a single-frequency system, though initial acquisition can still take 10s of seconds to minutes. When two or more frequencies (with different wavelengths) are utilized, the intersection of feasible solutions to the integer ambiguities on each frequency is often a single point, leading to a near-instantaneous, reliable RTK fix.

1.5.4 Precise Point Positioning (PPP)

Like SBAS providing a base-station free differential correction, many private companies have developed satellite-based corrections services that can achieve real-time, centimeter-level accuracy without the added infrastructure of an RTK system. Precise point positioning (PPP) combines precise clocks, orbit locations, ephemeris, and atmospheric models with proprietary software algorithms to increase the ability of a receiver to determine its location. This information is transmitted to a receiver either from a geostationary satellite or over the internet, and requires a subscription and additional software specific to the company used. These L-band corrections are not transmitted on the same exact frequency as any GNSS signal, but are nearby in the frequency spectrum, so GNSS antennas can be designed to track both simultaneously. Using this additional data, accuracy down to 3 cm (<10 cm typical) is achievable globally without need for a base station or radio link. However, it is important to note is that PPP systems may take 20-40 minutes to converge on such a high-accuracy solution. Notable L-band correction services utilizing PPP systems include TerraStar, OmniStar, StarFire, and Veripos.

1.6 ATTITUDE & HEADING REFERENCE SYSTEM (AHRS)

An attitude and heading reference system (AHRS) uses an inertial measurement unit (IMU) consisting of microelectromechanical system (MEMS) inertial sensors to measure the angular rate, acceleration, and Earth's magnetic field. These measurements can then be used to derive an estimate of the object's attitude.

RTK Double-Differencing

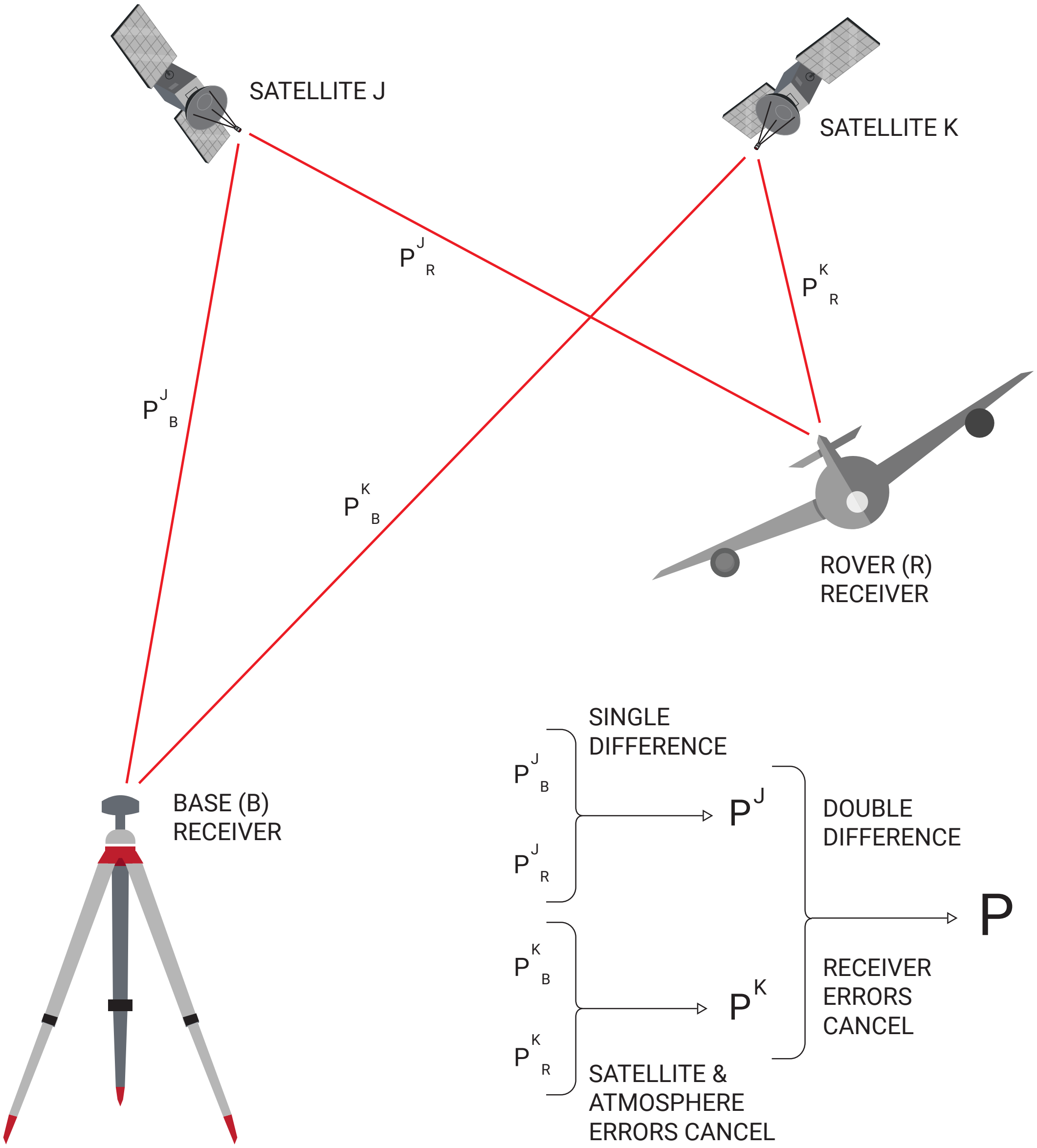


FIGURE 1.14

1.6.1 System Contributions

An AHRS typically includes a 3-axis gyroscope, a 3-axis accelerometer, and a 3-axis magnetometer to determine an estimate of a system's orientation. Each of these sensors contribute different measurements to the combined system and each exhibit unique limitations.

Gyroscope

A gyroscope provides an AHRS with a measurement of the system's angular rate. These angular rate measurements are then integrated to determine an estimate of the system's attitude. However, in order to determine the current attitude, the initial attitude of the system must also be known. Over time, this calculated attitude drifts unboundedly from the true attitude of the system due to the inherent noise and bias properties of the gyroscope itself.

Accelerometer

An accelerometer supplies an AHRS with a measure of the system's acceleration and is assumed to be measuring gravity alone. This assumption allows the accelerometer to calculate the pitch and roll angles from the direction of the gravity vector, as illustrated in Figure 1.15. However, any biases or other errors in the accelerometer measurements cause errors in the calculation of the pitch and roll angles. In addition, since the accelerometer is assumed to be measuring gravity alone, any added dynamic motion also causes an error in the calculation of the system's pitch & roll.

Accelerometer Pitch and Roll

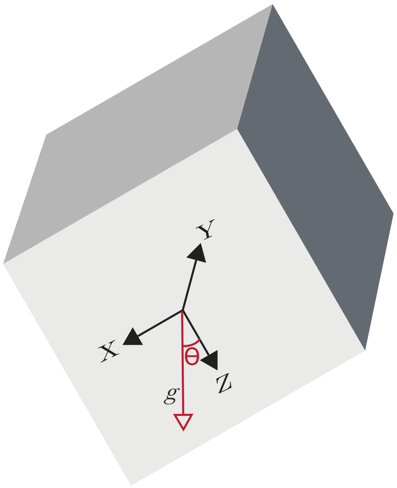


FIGURE 1.15

Magnetometer

Since the accelerometer can only measure pitch & roll, a magnetometer provides an AHRS with a measurement of yaw by comparing the measurement of the magnetic field surrounding the system to Earth's magnetic field, just like a traditional magnetic compass. In most AHRS units, the magnetometer measurements have no impact on the pitch and roll angle estimates.

While seemingly straightforward, using a magnetometer to accurately estimate the heading can actually prove to be quite challenging. Earth's magnetic field is weak, so large metal structures, high power cables, or any other magnetic disturbances can distort Earth's magnetic field and cause errors in the estimated heading angle. Disturbances caused by objects to which the AHRS is fixed (eg. the vehicle) can be compensated using a calibration known as hard & soft iron (HSI) calibration, but only when those disturbances do not vary over time. Advanced filtering techniques can be used to mitigate the impact of external disturbances in the environment, but their effectiveness varies by manufacturer and application.

Additionally, the magnetic North Pole of the earth is not in the same location as True North or the geographic North Pole of the earth. If the heading angle with respect to True North is desired, the declination angle between these two poles must be factored into the heading determination.

1.6.2 System Fusion

In an AHRS, the measurements from the gyroscope, accelerometer, and magnetometer are combined to provide an estimate of a system's orientation, often using a Kalman filter. This estimation technique uses these raw measurements to derive an optimized estimate of the attitude, given the assumptions outlined for each individual sensor.

The Kalman filter estimates the gyro bias, or drift error of the gyroscope, in addition to the attitude. The gyro bias can then be used to compensate the raw gyroscope measurements and aid in preventing the drift of the gyroscope over time. By combining the data from each of these sensors into a Kalman filter, a drift-free, high-rate orientation solution for the system can be obtained.

AHRS Component Diagram

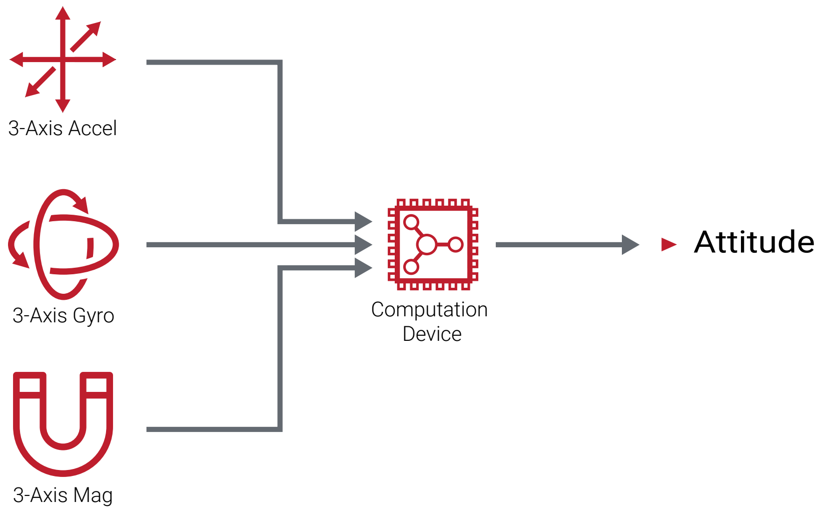


FIGURE 1.16

1.6.3 Challenges of AHRS

While many of the limitations a gyroscope, accelerometer, and magnetometer face on their own can be mitigated by combining them together, there are still a few challenges that come with using a AHRS that can cause errors in the system's attitude estimate. These challenges include transient and AC disturbances on the accelerometer and magnetometer, sustained dynamic accelerations, and internal and external magnetic disturbances.

Transient or Oscillating Disturbances

Any type of transient or AC disturbance that induces an acceleration or a magnetic disturbance for a short period of time can be almost completely mitigated through proper tuning and reliance on the integrated gyro through those time constants. For industrial grade sensors, a ”short” period of time corresponds to roughly a duration of <1 s, or at an oscillation faster than 1 Hz, with higher-grade sensors able to filter out longer time constants and vice versa.

Sustained Acceleration

Sustained dynamic accelerations can cause a problem in the estimation of the pitch and roll angles as the assumption that the accelerometer is measuring gravity alone is constantly being violated. The most common case where this becomes a significant problem for an AHRS is when an aircraft is operating in a banked turn. When this occurs, the accelerometer measures gravity plus a long-term acceleration due to the centripetal force created by traveling along a curved path. This results in a measurement vector that acts perpendicular to the wings of the aircraft and cause the AHRS to estimate a roll angle of zero while the aircraft is in fact in a banked turn and thus has significant roll relative to the horizon, as shown in Figure 1.17.

Sustained dynamic accelerations can also be caused by the starting and stopping of a system, such as an aircraft during takeoff and landing or a vehicle at a stoplight. This type of motion causes problems in estimating the pitch angle of the system. Unfortunately, during periods of sustained dynamic accelerations the gyroscope cannot be used to ride out the motion as its inherent drift means it cannot be trusted over longer periods of time.

Finally, ballistic flight, free-fall, or orbital dynamics leave the accelerometer measuring zero, providing an AHRS filter no information regarding the orientation of the sensor. This is especially problematic for ballistic flight, when the AHRS may confuse wind-resistance for gravity.

If an AHRS receives real-time velocity measurements of the system, the sustained dynamic acceleration can be estimated and compensated for in the attitude estimation.

Magnetic Disturbances

Magnetic disturbances, which can be internal or external to the system, also pose a problem to an AHRS and cause the magnetometer to measure a biased and distorted magnetic field. Internal magnetic disturbances are a result of

Measured Acceleration in a Coordinated Turn

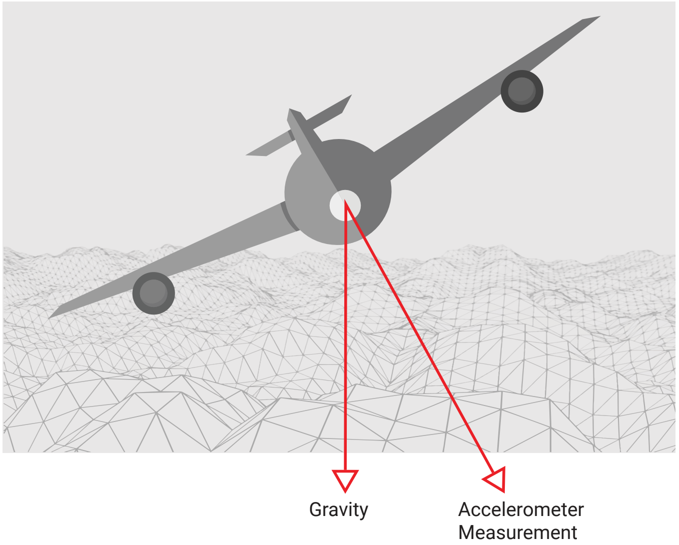


FIGURE 1.17

the magnetic signature of the system that the AHRS is rigidly attached to. They can be non-variable disturbances, such as a steel plate, or variable disturbances, such as motors or multi-rotors. External magnetic disturbances are caused by anything in the environment surrounding the system such as batteries, electronics, cars, rebar in concrete, and other ferrous materials. These magnetic disturbances lead to increased errors in the magnetometer measurements, causing errors in the estimates of the heading angle. To account for any non-variable magnetic disturbances internal to a system, a hard and soft iron (HSI) calibration can be performed on the system.

Drift in the "Drift-Free" Solution

Errors that exist in the accelerometer and magnetometer attitude solution, either due to sensor biases or to violations of the operating assumptions for each, cannot be avoided in the AHRS solution over longer periods of time. In fact, those errors can cause bounded drifting of what is otherwise considered a ”drift-free” attitude solution from the AHRS.

One simple illustration of this can be revealed through a static-dynamic-static test. This test is broken up into three parts in which the system is stationary during the first part of the test, experiences dynamic motion during a short second part, and finally returns to a stationary state in the third part of the test. During the stationary periods, the system's attitude is ultimately derived from the (possibly erroneous) accelerometer and magnetometer measurements. However, during the brief dynamic section, the gyroscope measurements dominate the AHRS response.

An example of a static-dynamic-static test is shown in Figure 1.18, in which the yaw measurements are tracked as a vehicle proceeds through a turn. In this scenario, the magnetic signature of the vehicle has not been compensated using an HSI calibration, so the magnetic heading measurements are inaccurate throughout the test.

During the initial stationary section of the test, the magnetometer measurements determine the vehicle's heading. Once the vehicle begins to drive through the turn, the gyroscope accurately tracks the *change* in heading, even though the initial heading was in error. After the turn is completed, the vehicle returns to a stationary state. Over time, even for a well-tuned AHRS, the magnetometer exerts itself as possible drift in the gyro prevents the AHRS from continuing to trust its integrated solution. Since the magnetometer is still impacted by the magnetic signature of the vehicle, the heading reported by the AHRS drifts until it settles into the new (still-erroneous) heading reported by the magnetometer.

1.7 GNSS-AIDED INERTIAL NAVIGATION SYSTEM (GNSS/INS)

The Global Navigation Satellite System (GNSS) is a satellite configuration, or constellation, that provides satellite signals to a GNSS receiver which can be used to calculate position, velocity, and time. An inertial navigation system (INS) uses an inertial measurement unit (IMU) consisting of microelectromechanical system (MEMS) inertial sensors to measure the system's angular rate and acceleration. Measurements from each of these two systems can be combined using advanced Kalman filtering estimation techniques to form a GNSS-aided INS system (GNSS/INS).

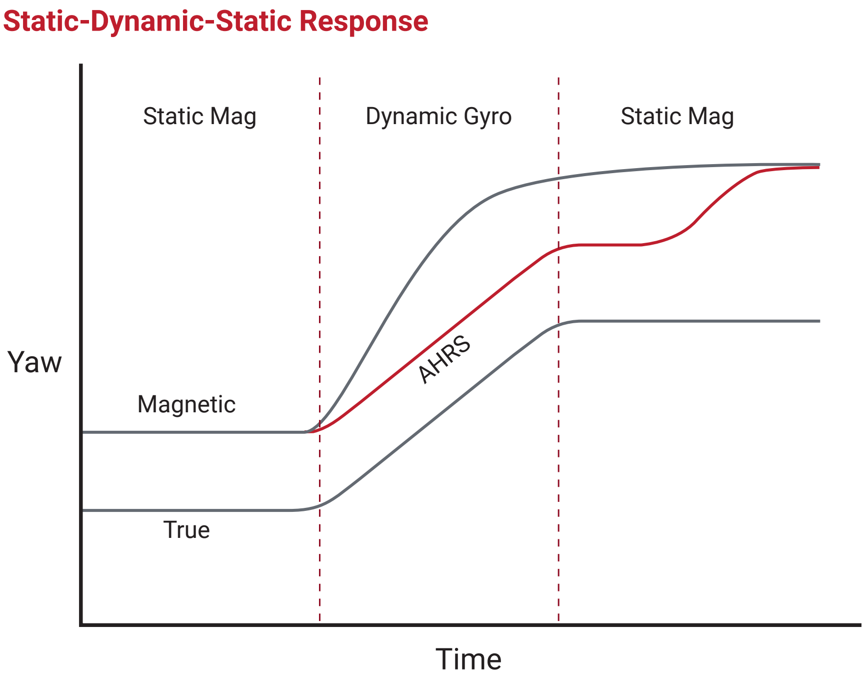


FIGURE 1.18

This combined system is able to provide position, velocity, and attitude estimates of higher accuracies and with better dynamic performance than a standalone GNSS or INS system can provide.

1.7.1 System Contributions

A GNSS/INS system typically includes a 3-axis gyroscope, a 3-axis accelerometer, a GNSS receiver, and sometimes a 3-axis magnetometer to estimate a navigation solution. Each of these sensors contribute different measurements to the GNSS/INS system.

Gyroscope & Magnetometer

Both the gyroscope and magnetometer provide a GNSS/INS system with the same contributions that they provide to an AHRS. The gyroscope angular rate measurements are integrated for a high-update rate attitude solution, while the magnetometer (if used) provides a heading reference similar to a magnetic compass. More information regarding the contribution of these sensors can be found in Section 1.6.

Accelerometer

An accelerometer in a GNSS/INS system measures both the system's linear acceleration due to motion and the pseudo-acceleration caused by gravity. In order to obtain the system's linear acceleration due to motion, the pseudo-acceleration caused by gravity must be subtracted from the accelerometer measurement using estimates of the system's attitude. The resulting linear acceleration measurement can then be integrated once to obtain the system's velocity and twice to obtain the system's position. However, these calculations are heavily dependent on the INS maintaining an accurate attitude estimate, as any error in the attitude causes an error in the calculated acceleration, consequently causing errors in the integrated position and velocity.

GNSS Receiver

A GNSS receiver uses the navigation message sent from the GNSS satellites and tracks the pseudorange and Doppler raw observable measurements to provide a GNSS/INS system with the receiver's position, velocity, and time (PVT). This drift-free PVT solution is used to stabilize the solutions offered by the integrals of the accelerometer and gyroscope.

1.7.2 System Fusion

Both the INS and GNSS can track the position and velocity of the system. An INS typically has reduced errors in the short-term, but larger, unbounded errors over extended periods of time. In contrast, GNSS tends to be noisier in the short-term, but can provide more stability over longer periods of time. When the two systems are integrated together, the GNSS measurements are able to regulate the INS errors and prevent their unbounded growth. On the other hand, an INS can provide a navigation solution at high output rates, while a GNSS navigation solution is typically only updated at rates between 1 Hz and 10 Hz. Combining the measurements from these two systems allows the

INS solution to bridge the gap between GNSS updates. A GNSS/INS system often uses a Kalman filter to track an optimal estimate of the system's position, velocity, attitude, gyro bias, and accelerometer bias.

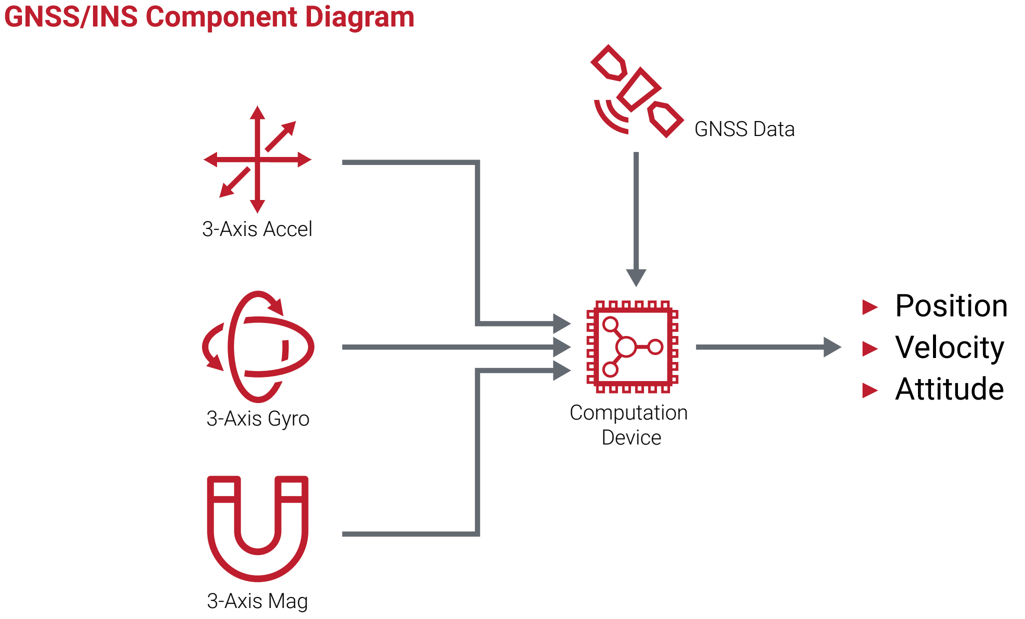


FIGURE 1.19

High-Accuracy Pitch & Roll

Unlike the AHRS filter, no assumption regarding the accelerometer measuring only gravity is made. Pitch and roll are still determined by knowing the direction of gravity, but the GNSS measurements make it possible to account for the impact of dynamic motion on the accelerometer readings. Combined with the ability to track accelerometer bias, the dynamic accuracy of pitch and roll in a GNSS/INS system is typically 1-2 orders of magnitude better than that of an AHRS.

Dynamic Alignment

Under sufficient dynamic motion, a GNSS/INS determines heading through a process known as dynamic alignment. The system correlates the acceleration measurements from the accelerometer with the position and velocity measurements from the GNSS receiver and is able to accurately derive the heading through this comparison.

For example, consider an accelerometer that measures that a system is accelerating in the negative y-axis of the vehicle, while the GNSS reports the system is accelerating West, as shown in Figure 1.20. Correlating these two measurements together yields that the negative y-axis must be aligned to West, and so the system must be pointing North.

Some systems—primarily legacy systems—require a specific pattern of motion to achieve dynamic alignment. But all that is required for most modern systems is horizontal acceleration of any type, such as accelerating down a runway at takeoff, driving around the block, or flying a figure-eight. In fact, most smaller vehicles simply need to get up to a decent speed to trigger dynamic alignment; the small fluctuations of a car at highway speed or a Cessna in light turbulence is enough for the Kalman filter to observe heading.

Note that the process of dynamic alignment is not the same as assuming heading is in the same direction as the velocity vector. It is a measure of the true heading of the vehicle, completely independent from course over ground (COG).

Coupling Architecture

When combining a GNSS and INS system together, there are a few different integration architectures that can be used to couple the measurements from each of the two systems. These different approaches are commonly referred to as loosely-coupled, tightly-coupled, and ultra-tightly-coupled, and are shown in Figure 1.21.

The loosely-coupled GNSS/INS system architecture is the most common integration approach. As shown in Figure 1.21a, this type of integration combines the GNSS navigation solution, consisting of the position, velocity, and time, with the INS navigation solution using an extended Kalman filter. The filter uses the INS measurements to predict the position, velocity, and attitude of the combined system. GNSS measurements are then used to update

Dynamic Alignment

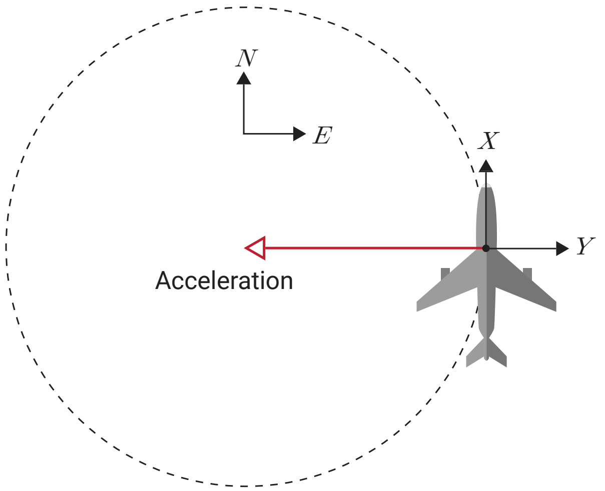
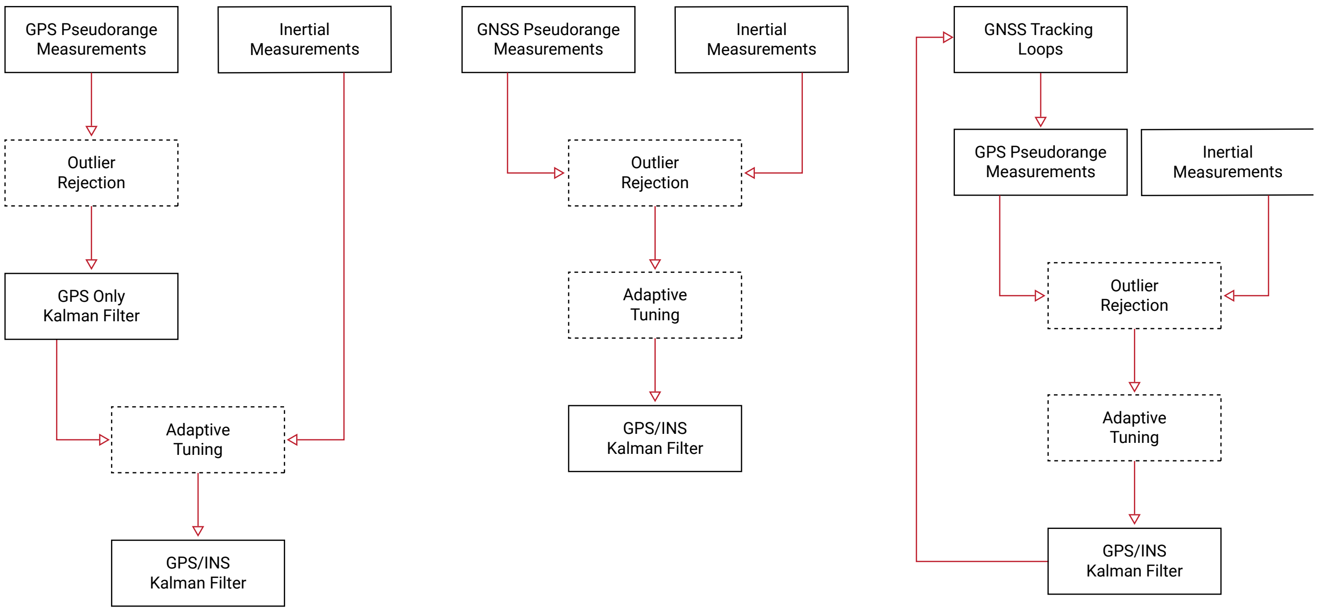


FIGURE 1.20

GNSS/INS Coupling



(a) Loosely-Coupled

(b) Tightly-Coupled

(c) Ultra-Tightly-Coupled

FIGURE 1.21

this prediction and estimate the gyro bias and accelerometer bias in the INS. These estimated biases are used to compensate the raw gyroscope and accelerometer measurements in the INS and improve its integration accuracy. In this approach, a GNSS receiver must have at least four satellites in view to calculate the receiver's position and velocity to send to the extended Kalman filter. If fewer than four satellites are in view of the receiver, the combined system will experience a GNSS outage and default to an INS.

As seen in Figure 1.21b, the tightly-coupled approach for the GNSS/INS system architecture is more closely integrated than that of the loosely-coupled design. This approach does not use the full navigation solution computed by the GNSS, but rather utilizes the raw GNSS pseudorange and Doppler measurements. As shown in Figure 1.21b, the raw GNSS measurements are combined with the INS navigation solution containing the integrated position, velocity, and attitude measurements into an extended Kalman filter. Since this approach uses the raw GNSS pseudorange and Doppler measurements rather than the full PVT solution, a single satellite can provide a useful GNSS update to the system. Due to this, the tightly coupled approach is most useful in applications that only have a partial view of the sky or that are susceptible to multipath error, such as urban canyons.

While the tightly-coupled approach has a potential advantage in restricted visibility environments, there is generally no benefit in clear-sky conditions. Furthermore, the outlier rejection and adaptive tuning algorithms employed (or not) by the GNSS receiver and the INS determine whether there is truly any advantage to tightly-coupled even in an urban canyon. If both a loosely-coupled and tightly-coupled filter were to naively factor in every GNSS measurement, the results of the two would be identical. While it is possible to create superior outlier rejection algorithms in a tightly-coupled scenario, in practice many tightly-coupled systems fall far short of loosely-coupled systems in head-to-head evaluation.

The ultra-tightly-coupled GNSS/INS system architecture is the most closely integrated approach, as seen in Figure 1.21c. Rather than having the GNSS and INS function as independent systems, the INS is used to help drive the tracking loops of the GNSS receiver, which track the carrier signals transmitted from the GNSS satellites. As satellites and the receiver move relative to each other, the INS provides high-rate feedback to maintain a tracking lock, even with a narrower tracking bandwidth than used in a standalone receiver. This narrower tracking bandwidth increases system accuracy and makes the receiver much less likely to track a multi-path signal rather than the true, direct signal from the satellite. However, the ultra-tightly-coupled approach is not as widely used in industry, as such feedback loops introduce new system instabilities and eliminate the redundancy that otherwise independent GNSS and INS systems provide in loosely- or tightly-couple systems.

1.7.3 Challenges of GNSS/INS

While many of the limitations GNSS and INS face as standalone systems can be mitigated by combining them together, there are still a few challenges that come with using a GNSS-aided INS system, including losing the heading information in static or low dynamic situations, the fact that GNSS errors are non-Gaussian and non-zero mean, and the possibility of GNSS outages.

Static or Low-Dynamic Situations

A GNSS/INS system loses observability of heading during low-dynamic or static situations, where dynamic alignment becomes impossible. During short duration periods of low dynamics, the INS can maintain an accurate, though continuously degrading, heading (on the order of 1 min for industrial grade). Most GNSS/INS systems fall back on an integrated magnetometer to continue stabilizing heading, though the issues with magnetic heading experienced in an AHRS system come into play.

GNSS Errors

Another challenge a GNSS/INS system faces is that the nature of GNSS measurement errors are non-Gaussian and non-zero mean. Non-Gaussian errors have a distribution that does not resemble that of a bell-curve shape, while non-zero mean errors contain a distribution with a mean that is not equal to zero, similar to Figure 1.22. A critical assumption used to derive optimality of a Kalman filter is that any errors in the system are Gaussian and zero-mean. Since GNSS errors violate this assumption, extra care must be taken when tuning a GNSS/INS Kalman filter to achieve the best performance.

GNSS Outages & Blockages

GNSS outages also pose a problem to a GNSS/INS system and can occur from a signal blockage or a signal interference. A GNSS signal blockage can be caused by anything from buildings to foliage that prevents the signal transmitted by the GNSS satellites from reaching the GNSS receiver, as illustrated in Figure 1.23. Signal interference is caused by a disturbance and can be intentional, such as in the case of jamming or spoofing, or unintentional, such as radio broadcasting signals that create disturbances on the signal. When GNSS outages occur, the GNSS/INS system defaults to an INS, which relies only on the IMU sensors to derive a navigation solution. Depending on the

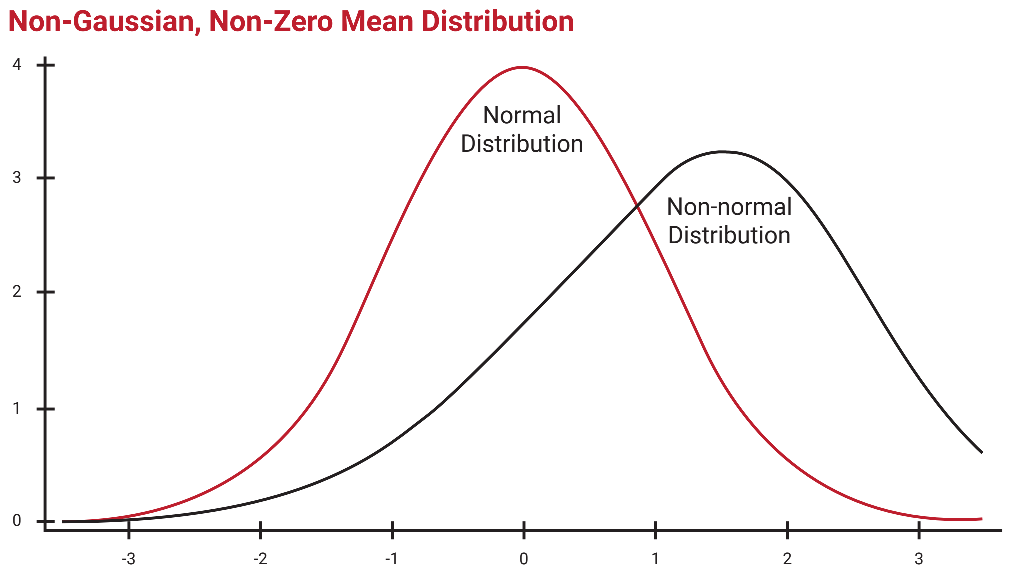


FIGURE 1.22

classification of the IMU sensors, using an INS alone to determine the navigation solution could lead to a large drift of the estimate over a short period of time.

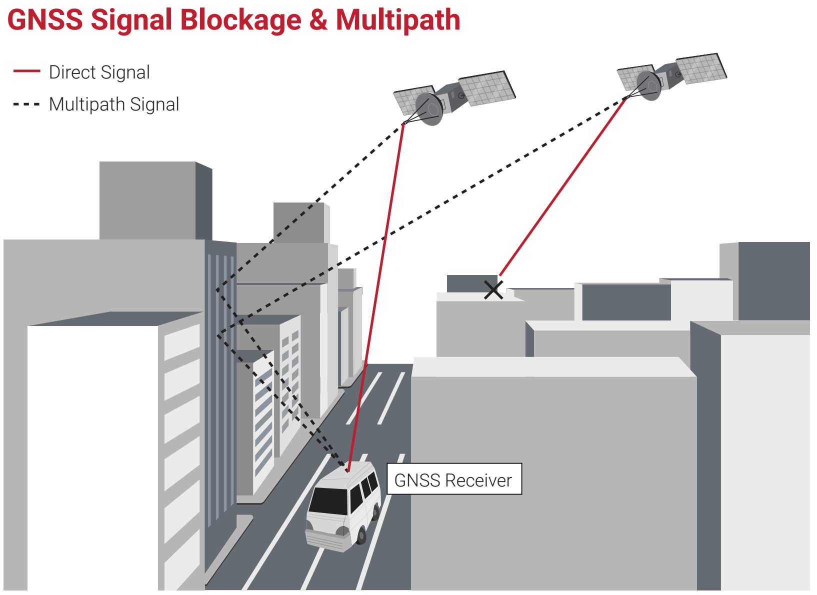


FIGURE 1.23

1.8 GNSS COMPASS/INS (DUAL GNSS/INS)

A GNSS Compass/INS, or Dual GNSS/INS, consists of two onboard GNSS receivers, forming a GNSS compass, and an onboard inertial navigation system (INS). The position and velocity measurements from the onboard GNSS modules are coupled with the inertial sensor measurements to provide estimates of a system's position, velocity, and attitude with higher accuracy and better dynamic performance than a standalone GNSS or INS system can provide.

1.8.1 System Contributions

Typically, a GNSS Compass/INS includes a 3-axis gyroscope, a 3-axis accelerometer, two GNSS receivers, and sometimes a 3-axis magnetometer to determine an estimate of the system's position, velocity, and attitude. Each of these devices provide similar contributions to a GNSS Compass/INS that they provide to a GNSS/INS system, which can be found in Section 1.7. A GNSS Compass/INS also contains an additional feature that utilizes the two separate onboard GNSS receivers to accurately estimate the system's heading: the powerful technique known as GNSS compassing.

Dual GNSS/INS Component Diagram

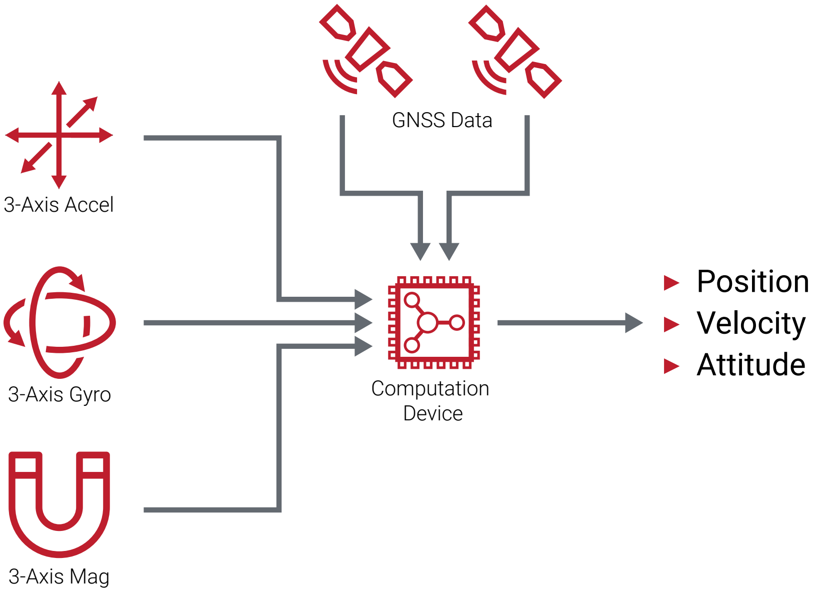


FIGURE 1.24

GNSS Compass

The GNSS compass technique uses a form of the real-time kinematic positioning (RTK) technique known as moving-baseline RTK to determine a system's heading. Traditional RTK is a precise positioning method that compares the carrier phase measurements between two antennas, a reference antenna which is typically stationary at a known location and a rover antenna that is able to move freely. This comparison of the carrier phase measurements allows RTK to determine a relative position between the two antennas to very high accuracy. For more information on carrier phase measurements and traditional RTK, see Section 1.2 and Section 1.5.

Moving-baseline RTK is a particular form of traditional RTK that allows both the reference antenna and the rover antenna to move freely. A GNSS compass is unique in that the two antennas are rigidly mounted with respect to each other with a fixed distance between the two antennas, known as the compass baseline. Ideally, both antennas should also be mounted in the same orientation, as the RF phase center of an antenna is not always located in the center of the antenna. The RF phase center of an antenna is the point where the antenna can receive signals, so aligning the two antennas in the same orientation ensures an accurate compass baseline measurement.

Similar to traditional RTK, moving-baseline RTK compares the carrier phase measurements from the GNSS signals between two antennas, allowing the GNSS Compass to determine the relative positioning of the two antennas in an inertial frame of reference to millimeter-level accuracy. If the position of the two antennas relative to each other is also known in the sensor's frame of reference, then the heading angle can be calculated in real-time with a high degree of accuracy. It is important to note that this heading measurement is derived directly from differencing the two GNSS receiver measurements at a single point in time, it does not require motion as is the case for dynamic alignment.

1.8.2 System Fusion

A GNSS Compass/INS operates similar to a GNSS/INS system, though it differs from a single-antenna system in that is has the capability to accurately estimate the heading in both static and dynamic conditions using the GNSS compass. This allows the system to accurately estimate the heading with respect to true North, without any reliance on magnetic sensors. In practice, a GNSS Compass/INS relies on dynamic alignment when available, but instead of falling back on the magnetic compass like a single-antenna GNSS/INS, it falls back on the GNSS compass solution.

The accuracy of the heading estimate derived from the GNSS compass is dependent on the quality of the GNSS signal, the baseline distance between the two antennas, and the accuracy of this baseline distance measurement. As shown in Equation 1.5, the error in the heading estimate, θ_{err}, is inversely proportional to the antenna baseline L. Due to this, it is important that the compass baseline between the two antennas is measured as accurately as possible.

$$\theta_{err} = \frac{P_{err}}{L} \tag{1.5}$$

Longer baseline distances provide higher accuracy GNSS compass heading estimates as the position error from the moving-baseline RTK, P_{err} , remains nearly constant over varying baseline distances. However, longer baseline distances do increase the start-up time required for the GNSS compass to lock onto the correct heading estimate.

The relative positioning from the moving-baseline RTK can also provide the GNSS Compass with a measurement of the pitch and roll angles. However, because the vertical channel of GNSS is half as accurate as the horizontal channel, these measurements typically are not able to provide accurate estimates of pitch and roll and are consequently not commonly used in a GNSS Compass/INS.

GNSS Compass Heading Accuracy vs Baseline Length

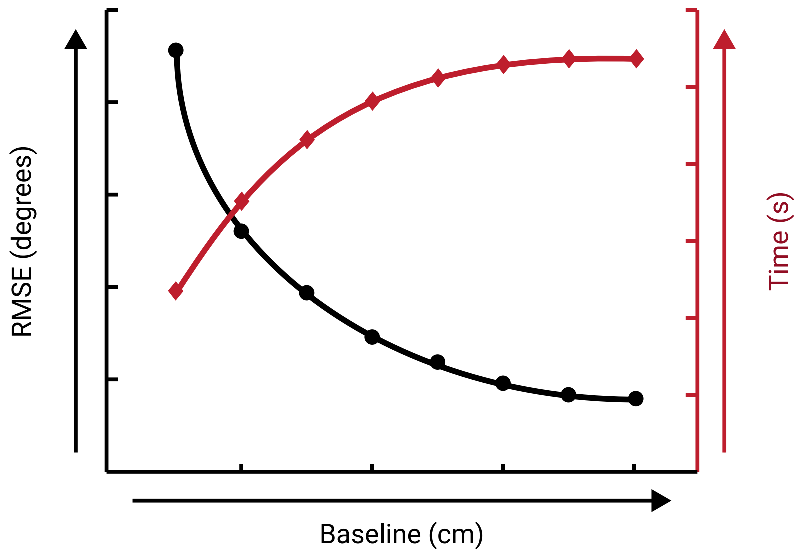


FIGURE 1.25

1.8.3 Challenges of GNSS Compass

A GNSS Compass/INS system experiences all of the same challenges as a GNSS-aided INS system, which can be found in Section 1.7. However, the addition of the GNSS compass requires a GNSS Compass/INS system to have much better satellite signal conditions than a single antenna system. Due to this, there are a few added challenges that must be overcome when using a GNSS Compass INS including:

- Maintaining a direct line-of-sight to the satellites
- Observing six common satellites between the two antennas
- Reducing multipath interference

A GNSS Compass/INS system requires that both GNSS antennas are mounted with a clear view of the sky, allowing the antennas to have a direct line-of-sight to the satellites. In addition, in order for the GNSS compass to estimate the heading using moving-baseline RTK, the two antennas must observe **at least six of the same satellites**. If there are any type of obstructions, such as foliage or buildings, that prevent the antennas from seeing six of the same satellites, the GNSS compass is unable to estimate the system's heading.

The GNSS Compass/INS is also more sensitive to multipath interference than a single antenna GNSS/INS system. Multipath interference occurs when GNSS satellite signals reflect off of solid objects such as buildings and terrain, resulting in the signal taking multiple paths to reach the antenna, as shown in Figure 1.23. These reflected signals are delayed as compared to direct line-of-sight signals, causing errors in the carrier phase measurements, which are used in the GNSS compass. The most common cause of multipath interference is due to signals reflecting up from the ground to the bottom of the GNSS antennas. To mitigate this, it is recommended that ground planes be placed underneath the antennas to block these signals from reflecting up from the ground.

1.9 HEADING DETERMINATION

The heading, also referred to as the yaw or azimuth, is the rotation of a system about the vertical axis of the inertial reference frame (aligned to gravity). A variety of techniques for determining a system's heading utilizing inertial sensors have been discussed in this chapter, each with their own pros and cons, and are summarized here.

1.9.1 Magnetometer (Magnetic Compass)

A detailed description of using a magnetometer for heading determination is given in Section 1.6.

Theory of Operation

A magnetometer is used to measure the magnetic field surrounding a system. This magnetic field measurement can be compared to models of Earth's magnetic field to determine the heading of a system with respect to magnetic North. Though the geographical location of magnetic North is different than that of true North, the heading can be found with respect to true North by taking into account the declination angle between the two locations.

Limitations

Using a magnetometer to accurately estimate a system's heading can prove to be quite challenging. Earth's magnetic field is weak, making magnetometers highly susceptible to magnetic disturbances, which are caused by any ferrous materials or electric currents near the magnetometer. These disturbances will bias and distort the background magnetic field, leading to increased errors in the heading estimate. Earth's magnetic field can also shift as much as 2° from one day to the next. Due to this, even in the most ideal magnetic environments, a magnetometer can only provide a heading accuracy of 1° to 2° over an extended period of time.

1.9.2 GNSS/INS (Dynamic Alignment)

A detailed description of how a GNSS/INS system achieves heading determination via dynamic alignment is given in Section 1.7.

Theory of Operation

A combined GNSS/INS system determines the heading of a system through the correlation of measurements from the two systems, a process known as dynamic alignment. The accelerometer measurements from the INS solution are compared to the position and velocity measurements from the GNSS solution to determine the heading of the system to high levels of accuracy, depending on the quality of the GNSS and INS. Note this is not the same as assuming that heading equals the course over ground reported by the GNSS.

Limitations

Though a GNSS/INS system can provide an accurate and reliable estimate of the system's heading, there are a few conditions that can cause the loss of heading observability when using this system. If the combined system experiences any low dynamic or static situations, the horizontal acceleration of the system will be nearly zero, making the comparison of the INS and GNSS measurements impossible and causing the heading observability of the system to be lost. The loss of heading observability can also be caused by GNSS outages which occur from signal blockages or signal interference.

1.9.3 GNSS Compassing

A detailed description of using a GNSS compass for heading determination is given in Section 1.8.

Theory of Operation

The GNSS compassing technique uses a form of real-time kinematic (RTK) positioning known as moving-baseline RTK to determine a system's heading. Moving-baseline RTK compares the carrier phase measurements between two GNSS antennas that are fixed relative to each other at a given distance. Through this comparison, the GNSS compass can determine a relative positioning of each antenna to millimeter-level accuracy and estimate the system's heading with an accuracy that is inversely proportional to the separation distance between the two antennas.

Limitations

While GNSS compassing can provide an accurate and reliable heading estimate, there are a few limitations of using this technique. In order for the GNSS compass to estimate a system's heading using moving-baseline RTK, the two antennas must have a clear view of the sky and observe at least six of the same satellites. The GNSS compass also has a high sensitivity to multipath and has difficulty providing an accurate heading in applications subject to multipath conditions, such as urban canyons.

1.9.4 Gyrocompassing

A detailed description of the types of gyros capable of utilizing gyrocompassing for heading determination is given in Section 1.4.

Theory of Operation

Gyrocompassing is a technique that determines a system’s heading without reliance on GNSS or magnetic field measurements. This method uses a gyroscope to measure the angular rate of the earth in conjunction with measurements of the gravity vector to detect the direction of North and determine the system’s heading.

Limitations

Although gyrocompassing can provide a reliable heading estimate, independent of both magnetometers and GNSS, this technique is limited to using high performance gyroscopes such as fiber-optic gyroscopes (FOGs) and ring laser gyroscopes (RLGs). The size, weight, power, and cost of gyros capable of gyrocompassing is prohibitive for most applications.

2 MATH FUNDAMENTALS

Whether the goal is to understand the workings of inertial systems or just use them within a larger system, it is important to understand a few basic mathematical concepts governing their behavior. This chapter covers everything from how attitude is defined and represented, to the basics of a Kalman filter, which underlies most navigation algorithms.

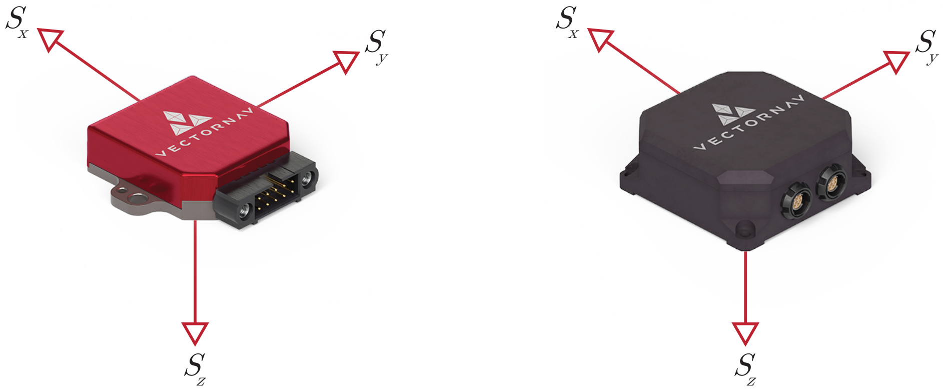
2.1 REFERENCE FRAMES

In order to describe the position, velocity, and orientation of an object of interest, the object's motion must be compared to a type of standard known as a reference frame. A reference frame is comprised of an origin that defines a position in space and three orthogonal unit vectors or axes that make up a right-handed system. These three axes are typically denoted by x-y-z or via subscripts 1-2-3. There are numerous reference frames that can be used to measure an object's motion depending on the type of application and desired results, such as a sensor frame, a body frame, an Earth-centered, Earth-fixed (ECEF) frame, or a local North-East-Down (NED) frame. Note that in most of the ensuing discussion, the location of the origin is often neglected as it is irrelevant to defining the attitude of a system.

2.1.1 Sensor Frame

The sensor frame is a type of reference frame that is fixed to the sensor. On VectorNav sensors, the sensor frame is aligned as shown in Figure 2.1. The individual sensor element measurement axes within the instrument are aligned to the sensor frame as part of the calibration process. When reviewing a sensor's datasheet, the "Misalignment" or "Alignment Error" specification (see Section 3.1) provides an indication of how closely the measurement axes are aligned with the indicated sensor frame.

VectorNav Sensor Frame



(a) VN-100-CR

(b) VN-110

FIGURE 2.1

2.1.2 Body Frame

In most applications, the position, velocity, and orientation of a system are found using sensors mounted to a vehicle or platform. The platform can have its own reference frame known as the body frame, or sometimes called the vehicle frame. This type of reference frame consists of an origin that is typically placed at the platform's center of gravity and three orthogonal axes that comprise a right-handed system. These axes are usually configured to the body in such a way that the x-axis is pointing forward, the y-axis is pointing to the right, and the z-axis is pointing down as shown in Figure 2.2. In some applications, the sensor frame cannot be perfectly aligned to the body frame and will require a reference frame rotation to align the two reference frames.

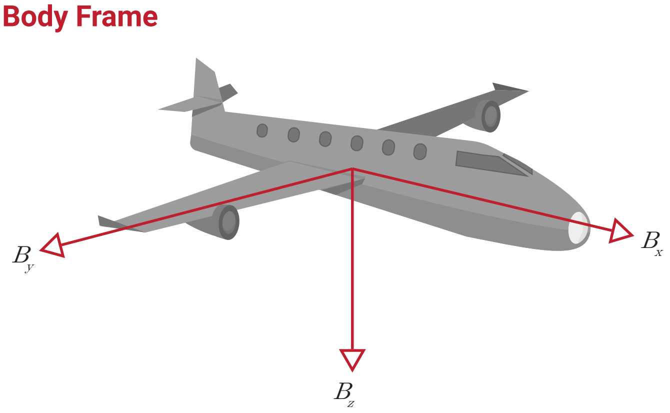


FIGURE 2.2

2.1.3 Earth-Centered, Earth-Fixed (ECEF) Frame

The Earth-centered, Earth-fixed (ECEF) frame is a global reference frame with its origin at the center of the earth and three orthogonal axes fixed to the earth. As shown in Figure 2.3, the E_z axis points through the North Pole, the E_x axis points through the intersection of the IERS Reference Meridian (IRM) and the equator, and the E_y axis completes the right-handed system. This reference frame rotates with Earth at an angular velocity of approximately 15 °/hr (360° over 24 hr).

Within the ECEF reference frame, there are two primary coordinate systems used when describing a system's position: cartesian and geodetic. The cartesian coordinate system uses the E_x, E_y, and E_z axes to represent an object's position directly. The geodetic coordinate system is essentially a set of polar coordinates, but one that accounts for the first-order effect of Earth being an ellipsoid rather than a sphere, and describes an object's ECEF position in terms latitude (ϕ), longitude (λ), and altitude (h).

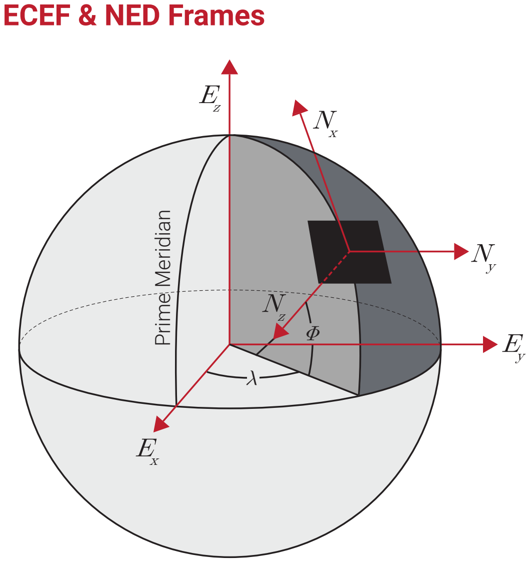


FIGURE 2.3

2.1.4 North-East-Down (NED) Frame

The North-East-Down (NED) frame is a local reference frame that is defined by its ECEF coordinates. Often this frame is fixed to the vehicle or platform and moves with the body frame. The NED frame is defined such that the North and East axes form a plane tangent to Earth's surface at its present position, assuming a WGS84 ellipsoid model of the earth. As shown in Figure 2.3, the NED frame contains three orthogonal axes in which the N_x axis points to

True North, the N_z axis points towards the interior of the earth and the N_y axis completes the right-handed system pointing east.

Similar to the NED frame, there is also an East-North-Up (ENU) frame that can be placed locally on a vehicle or platform and moves around with the system. The ENU frame differs from the NED frame in the direction of the three orthogonal axes. The N_y axis points to True North, the N_z axis points away from the interior of the earth, and the N_x axis completes the right-handed system pointing east.

2.1.5 Earth-Centered Inertial (ECI) Frame

The earth-centered inertial (ECI) frame is a global reference frame that has its origin at the center of the earth. This reference frame does not rotate with Earth and serves as an inertial reference frame for satellites orbiting Earth. Due to this, the ECI frame is used primarily in space applications.

2.2 DATUMS & WORLD MODELS

Like reference frames, datums and world models provide needed references for navigation. A datum is a reference used for taking a measurement and is used in many applications including navigation and land surveying as well as for drafting purposes such as geometric dimensioning and tolerancing. World models provide references for the gravitational and magnetic fields that can be used across multiple science and engineering disciplines, including navigation.

2.2.1 Gravity & Altitude

Various models and datums are used to describe the shape of the earth and its gravitational potential field. These not only define the value of gravity used in navigation equations, but define the measure of altitude. The WGS84 model described below defines altitude relative to a simple ellipsoid model of the earth, whereas the EGM96 model can be used to define altitude relative to Mean Sea Level (MSL), which is an equipotential surface of the EGM96 gravitational model. The altitude measured relative to MSL can be as much as 100 m off from the altitude measured relative to WGS84.

World Geodetic System 1984 (WGS84)

The World Geodetic System of 1984 (WGS84) is a global datum and standard, whose definition of the ellipsoid shape of the earth is widely used in many different fields, including navigation. The size and shape of the earth described by WGS84 defines the transformation between cartesian ECEF coordinates and latitude-longitude-altitude (LLA) coordinates. Since GPS uses it for reporting LLA, its use is ubiquitous in navigation.

Earth Gravitational Model 1996 (EGM96)

The Earth Gravitational Model of 1996 (EGM96) is a world model and reference frame used as a gravity model of the earth. The model uses spherical harmonics to define Earth's gravitational potential complete to degree and order of 360. In addition to providing the value of g, this model defines the geoid: the shape that the surface of the oceans would take under the influence of Earth's gravity and rotation alone. The geoid is an equipotential surface which by definition means that every point in a specific region of space is at the same gravitational potential. Such an equipotential surface is used to define Mean Sea Level.

2.2.2 Magnetic Field Models

Earth's magnetic field varies drastically across the planet. While a simple dipole model provides a reasonable first order approximation, more detailed models are required to achieve heading accuracy in single-digit degrees. These models provide the magnetic field as a 3D vector in the ECEF frame for every point on Earth, from which can be calculated the declination angle (offset between Magnetic North and True North), and the inclination angle (offset of total field vector from horizontal).

World Magnetic Model (WMM)

The World Magnetic Model (WMM) is a spherical harmonics based model of Earth's magnetic field and provides the declination angle, which is the difference between Magnetic North and True North, for a specific location. It is the standard model for navigation systems, such as an attitude and heading reference systems (AHRS), and is used by many government and international agencies including the U.S. Department of Defense, the U.K. Ministry of Defense, NATO, and the International Hydrographic Organization (IHO). The WMM is a predictive model as Earth's magnetic field changes over time. Due to this, the WMM is updated every five years to represent any changes in the magnetic field that may occur.

International Geomagnetic Reference Field (IGRF)

The International Geomagnetic Reference Field (IGRF) is a backward-looking, corrected magnetic model of Earth that combines parameters of the World Magnetic Model with measured magnetic field data from surveys, observatories, and satellites around the world. This model is used largely by the scientific community and has similar accuracy to that of the World Magnetic Model.

2.3 ATTITUDE REPRESENTATIONS

An attitude representation is often defined as a set of coordinates that describe the orientation of a given reference frame with respect to a second reference frame. Defining the rotational orientation of a rigid body requires a minimum of three parameters, however, many attitude representations utilize more than three parameters in defining the orientation. While there are numerous attitude representations that can be used to define the orientation of a system, some of the more common representations include direction cosine matrices (DCM), Euler angles, and quaternions.

2.3.1 Direction Cosine Matrix

The direction cosine matrix (DCM) is one of the many ways to mathematically represent an object's orientation and utilizes nine parameters. Each of these parameters are referred to as the direction cosine values between an initial reference frame and a second reference frame. Consider two reference frames $\mathcal{B}$ and $\mathcal{N}$, where $\mathcal{B}$ represents a body frame and $\mathcal{N}$ is defined as the inertial reference frame. Since $\mathcal{B}$ and $\mathcal{N}$ are reference frames, both are defined by a set of three orthonormal vectors that make up a right-handed system. These vectors can be expressed as $\{\hat{\boldsymbol{n}}\}$ and $\{\hat{\boldsymbol{b}}\}$, as described in Equation 2.1.

$$\{\hat{\boldsymbol{n}}\} = \begin{Bmatrix} \hat{\boldsymbol{n}}_1 \\ \hat{\boldsymbol{n}}_2 \\ \hat{\boldsymbol{n}}_3 \end{Bmatrix} \quad \{\hat{\boldsymbol{b}}\} = \begin{Bmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{b}}_2 \\ \hat{\boldsymbol{b}}_3 \end{Bmatrix} \tag{2.1}$$

$\mathcal{B}$ and $\mathcal{N}$ Reference Frames

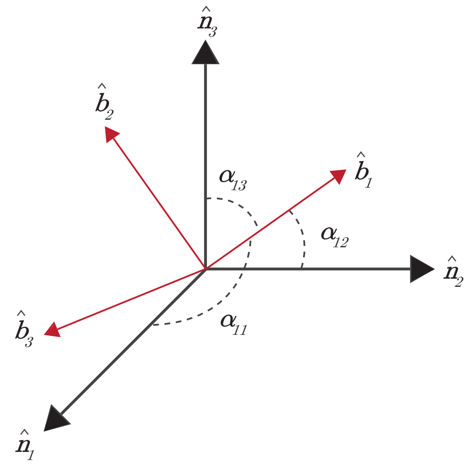


FIGURE 2.4

As shown in Figure 2.4, the vector $\hat{\boldsymbol{b}}_1$ creates an angle α with each of the vectors $\hat{\boldsymbol{n}}_1$, $\hat{\boldsymbol{n}}_2$, and $\hat{\boldsymbol{n}}_3$. In order to determine the direction cosine values of $\hat{\boldsymbol{b}}_1$ with respect to the $\mathcal{N}$ reference frame, the cosine of each of these angles is taken. The same follows for the other two vectors $\hat{\boldsymbol{b}}_2$ and $\hat{\boldsymbol{b}}_3$. Each of the unit vectors of $\{\hat{\boldsymbol{b}}\}$ can be represented in terms of the unit vectors of $\{\hat{\boldsymbol{n}}\}$ as shown in Equation 2.2.

$$\begin{aligned} \hat{\boldsymbol{b}}_1 &= \cos\,\alpha_{11}\,\hat{\boldsymbol{n}}_1 + \cos\,\alpha_{12}\,\hat{\boldsymbol{n}}_2 + \cos\,\alpha_{13}\,\hat{\boldsymbol{n}}_3 \\ \hat{\boldsymbol{b}}_2 &= \cos\,\alpha_{21}\,\hat{\boldsymbol{n}}_1 + \cos\,\alpha_{22}\,\hat{\boldsymbol{n}}_2 + \cos\,\alpha_{23}\,\hat{\boldsymbol{n}}_3 \\ \hat{\boldsymbol{b}}_3 &= \cos\,\alpha_{31}\,\hat{\boldsymbol{n}}_1 + \cos\,\alpha_{32}\,\hat{\boldsymbol{n}}_2 + \cos\,\alpha_{33}\,\hat{\boldsymbol{n}}_3 \end{aligned} \tag{2.2}$$

These three equations can also be expressed in matrix form as shown in Equation 2.3.

$$\{\hat{\boldsymbol{b}}\} = \begin{bmatrix} \cos\,\alpha_{11} & \cos\,\alpha_{12} & \cos\,\alpha_{13} \\ \cos\,\alpha_{21} & \cos\,\alpha_{22} & \cos\,\alpha_{23} \\ \cos\,\alpha_{31} & \cos\,\alpha_{32} & \cos\,\alpha_{33} \end{bmatrix} \{\hat{\boldsymbol{n}}\} = [C_{\mathcal{BN}}]\,\{\hat{\boldsymbol{n}}\} \tag{2.3}$$

The quantity $C_{\mathcal{BN}}$ is referred to as the DCM of $\mathcal{B}$ with respect to $\mathcal{N}$. Similarly, the unit vectors of $\{\hat{\boldsymbol{n}}\}$ can be projected onto the unit vectors of $\{\hat{\boldsymbol{b}}\}$ using Equation 2.4.

$$\{\hat{\boldsymbol{n}}\} = \begin{bmatrix} \cos\,\alpha_{11} & \cos\,\alpha_{21} & \cos\,\alpha_{31} \\ \cos\,\alpha_{12} & \cos\,\alpha_{22} & \cos\,\alpha_{32} \\ \cos\,\alpha_{13} & \cos\,\alpha_{23} & \cos\,\alpha_{33} \end{bmatrix} \{\hat{\boldsymbol{b}}\} = [C_{\mathcal{NB}}]\,\{\hat{\boldsymbol{b}}\} = [C_{\mathcal{BN}}]^{\intercal}\,\{\hat{\boldsymbol{b}}\} \tag{2.4}$$

This simple method can also be used to transform vectors from one reference frame into another. Consider a vector $\boldsymbol{v}$ that has its components in the $\mathcal{N}$ reference frame but needs to be described in the $\mathcal{B}$ reference frame. This can be accomplished using Equation 2.5. Likewise, a vector which has its components in the $\mathcal{B}$ reference frame can be transformed into the $\mathcal{N}$ reference frame with Equation 2.6.

$$^{\mathcal{B}}\boldsymbol{v} = [C_{\mathcal{BN}}]\,^{\mathcal{N}}\boldsymbol{v} \tag{2.5}$$

$$^{\mathcal{N}}\boldsymbol{v} = [C_{\mathcal{BN}}]^{\intercal}\,^{\mathcal{B}}\boldsymbol{v} \tag{2.6}$$

There are a few properties of the DCM that make it a unique method for representing an object's attitude. For example, the norm of each row and column of the DCM must be equal to +1. In addition, the DCM is orthogonal and more specifically, orthonormal. This means that the product of $[C_{\mathcal{BN}}]$ and the transpose of $[C_{\mathcal{BN}}]$ results in the identity matrix as shown in Equation 2.7.

$$[C_{\mathcal{BN}}]\,[C_{\mathcal{BN}}]^{\intercal} = [C_{\mathcal{BN}}]^{\intercal}\,[C_{\mathcal{BN}}] = [I_{3\times 3}] \tag{2.7}$$

Furthermore, this means that the transpose of $[C]$ is equal to its inverse:

$$[C_{\mathcal{BN}}]^{\intercal} = [C_{\mathcal{BN}}]^{-1} = [C_{\mathcal{NB}}] \tag{2.8}$$

One final DCM property to note is that of the determinant. For a right-handed system, the determinant of the DCM must be equal to +1.

$$\det\,[C_{\mathcal{BN}}] = +1 \tag{2.9}$$

While the direction cosine matrix is an important and commonly used method for representing an object's attitude, there is one major drawback to this approach. The DCM utilizes nine parameters to describe an orientation. Since only three parameters are required, six of the DCM values are redundant. Due to this, the DCM is hardly ever used to keep track of attitude in real time and is instead used primarily to project vectors to different reference frames.

2.3.2 Euler Angles (Yaw-Pitch-Roll)

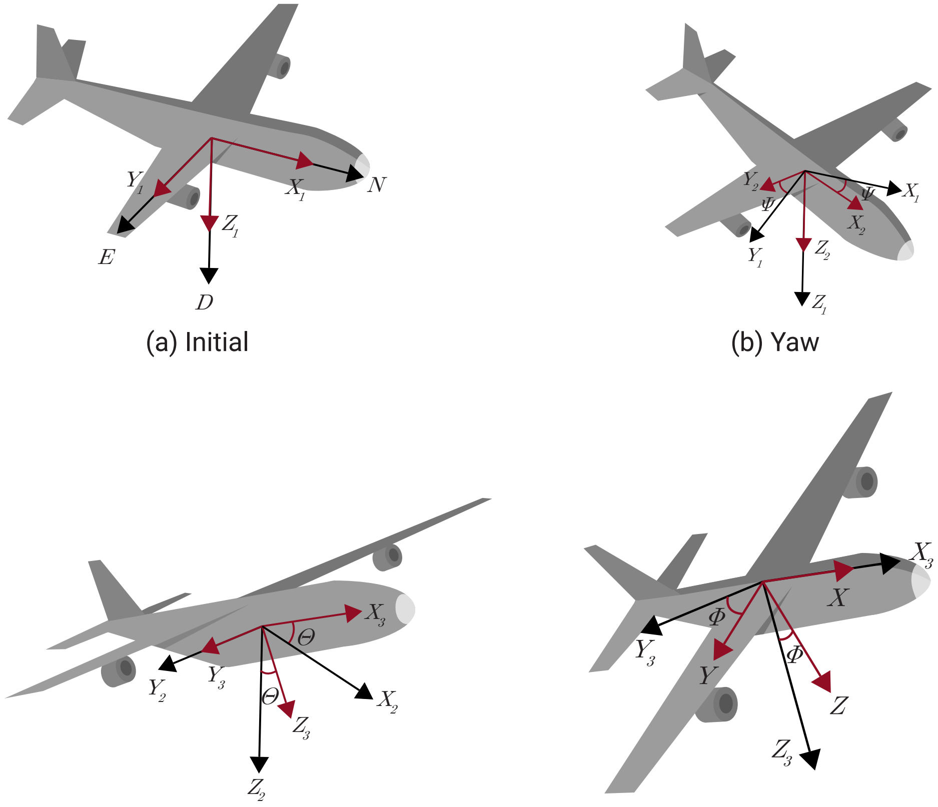
The orientation of a rigid body can also be described by three successive rotations about a set of intermediary axes, which transform the body from an inertial reference frame into the body frame. These three rotations are the most frequently used method for representing an object's attitude and are known as the Euler angles.

There are many different combinations of Euler angles, however, the (3-2-1) set of Euler angles corresponding to yaw-pitch-roll (ψ-θ-ϕ) is considered to be the standard, especially in terrestrial applications. These rotations are applied sequentially in a particular order, with each rotation specified about the specified body frame axis as it exists following the previous rotations. Figure 2.5a shows the body frame and NED frame initially aligned. Figure 2.5b shows the yaw rotation around the Z_1-axis. This is followed in Figure 2.5c by the pitch rotation about the new Y_2-axis. Finally, there is a roll rotation about the new X_3-axis in Figure 2.5d to achieve the final orientation of the aircraft.

The order of these rotations is important, as a (3-2-1) set of Euler angles corresponding to yaw-pitch-roll can result in a much different orientation than applying those same angles in a (1-2-3) sequence of roll-pitch-yaw. As an example, Figure 2.6 shows the difference between applying a −90° pitch, followed by a 45° roll (Fig. 2.6a), and applying a 45° roll, followed by a −90° pitch (Fig. 2.6b). The end orientation of the aircraft is very different! It is important to note that this order does not have as much of an effect when the angles are small. This makes Euler angles particularly easy to visualize and therefore are the most commonly used attitude representation.

While this seems to be an ideal way to represent attitude there is one major drawback to this method: every set of Euler angles has at least one geometric singularity, often referred to as gimbal lock. This means that at a specific orientation, two of the axes become ambiguous. For example, in the standard set of (3-2-1) Euler angles, this singularity exists when the pitch is $\pm$90°. In this case, the yaw and roll perform the same operation, as a yaw angle of 20° and a roll angle of 0° results in an orientation identical to that of a yaw angle of 0° and a roll angle of 20°. As a result, Euler angles must be used with caution, especially in applications that deal with angles close to the singularity points.

Example of a 3-2-1 Euler Angle Set



(a) Initial

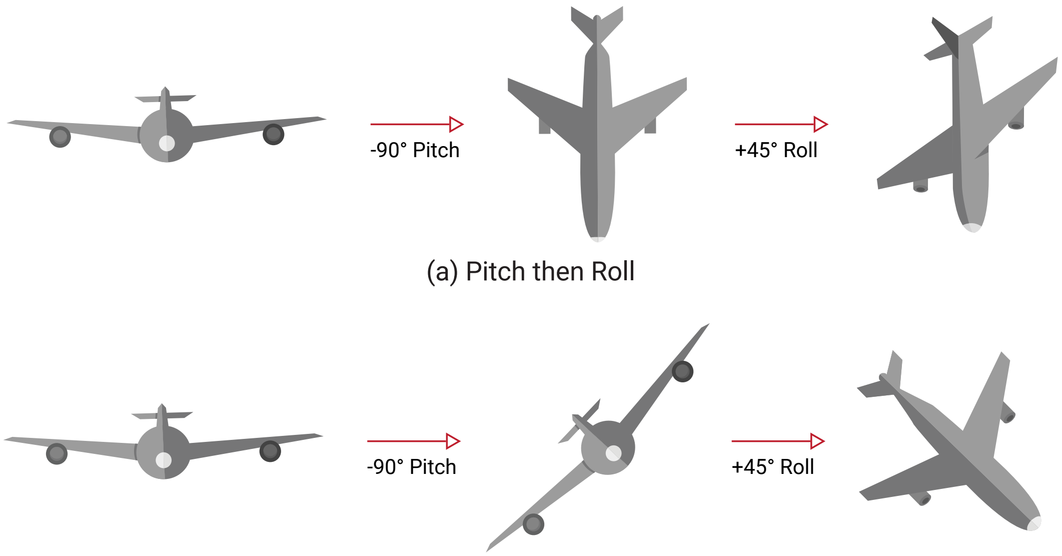
(b) Yaw

(c) Pitch

(d) Roll

FIGURE 2.5

Euler Angle Order Example



(a) Pitch then Roll

(b) Roll then Pitch

FIGURE 2.6

2.3.3 Euler's Principal Rotation Theorem

Euler's principal rotation theorem states that any rigid body or reference frame can be taken from an initial orientation to a final orientation via a single rigid body rotation about a principal axis, $\hat{e}$, through a principal angle, ϕ. Simply put, any arbitrary orientation can be described by a single unit vector and a single angle, as shown in Figure 2.7.

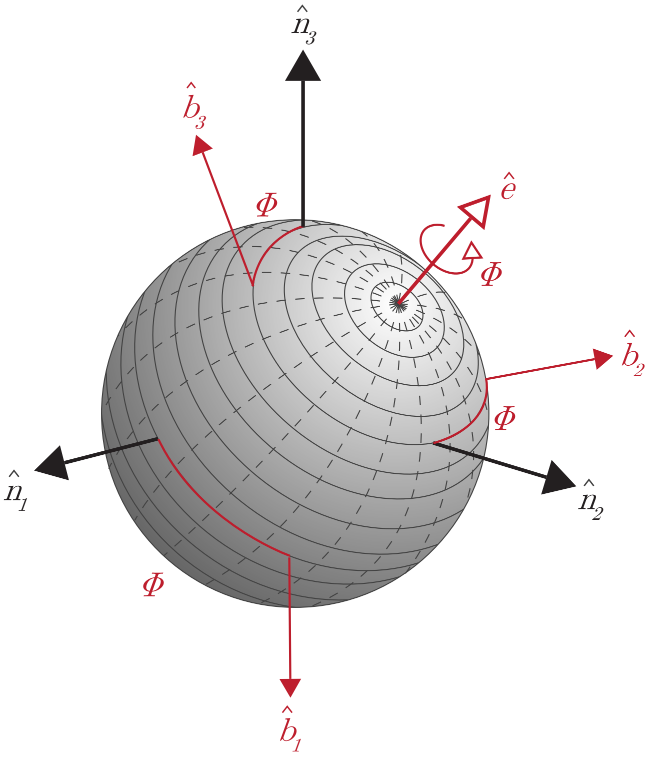

Euler's Principal Rotation Theorem

FIGURE 2.7

The principal axis defines the direction of rotation while the principal angle describes the amount of rotation about this axis from the initial attitude to the final attitude. This principal rotation theorem is valid for any rotation and the principal angle is useful as a scalar measure of the difference between two attitudes, such as the error between the measured attitude and the truth. Furthermore, the principal axis and principal angle form the basis for defining many other attitude representations.

2.3.4 Quaternion

A quaternion is an attitude representation that uses a normalized four-dimensional vector to describe a three-dimensional orientation. This approach is based upon Euler's principal rotation and consists of a scalar term q_s and a vector term q_v as shown in Equation 2.10.

$$\begin{aligned} \boldsymbol{q}_v &= \hat{e}\,\text{sin}\,\frac{\phi}{2} = \begin{bmatrix} e_1\,\text{sin}\,\frac{\phi}{2} \\ e_2\,\text{sin}\,\frac{\phi}{2} \\ e_3\,\text{sin}\,\frac{\phi}{2} \end{bmatrix} \\ q_s &= \text{cos}\,\frac{\phi}{2} \end{aligned} \tag{2.10}$$

Typically, the quaternion is represented by a single four-element vector. It is important to note that there is no agreed upon order for that four-element vector, so it is important to be clear whether a system or equation is assuming the scalar term last (Eq. 2.11) or first (Eq. 2.12). VectorNav sensors and documentation utilize the scalar term last representation.

$$\boldsymbol{q} = \begin{bmatrix} \boldsymbol{q}_v \\ q_s \end{bmatrix} \tag{2.11}$$

$$\boldsymbol{q}' = \begin{bmatrix} q_s \\ \boldsymbol{q}_v \end{bmatrix} \tag{2.12}$$

The quaternion is advantageous as it circumvents the singularity problem of Euler angles by adding an additional parameter, creating an over-defined attitude representation. However, this does create an over-defined attitude representation and requires the constraint that the norm of the quaternion be equal to one. Note also that there is an ambiguity when representing an attitude by a quaternion, because $\boldsymbol{q} \equiv -\boldsymbol{q}$.

2.4 ATTITUDE TRANSFORMATIONS

Since there is not a "standard" attitude representation, the technique chosen is highly dependent upon the specific application. However, the optimal method for a specific application may be different from the desired final representation of the orientation. Therefore, a conversion or transformation between the different attitude representations is needed. Below are some of the more common transformations used that are based on the VectorNav quaternion notation in which q_4 is the scalar term.

2.4.1 Quaternion to/from Direction Cosine Matrix

The elements of the DCM can be determined from the associated quaternion using Equation 2.13:

$$C = \begin{bmatrix} q_4^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_3q_4) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_4) & q_4^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_4) & 2(q_2q_3 - q_1q_4) & q_4^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \tag{2.13}$$

There are a variety of ways to extract the quaternion from the DCM defined in Equation 2.13, though several of them contain divide by zero singularities for certain attitudes. A numerically stable method for calculating the quaternion starts with calculating the squares of each quaternion term:

$$\begin{aligned} q_1^2 &= \frac{1}{4}(1 + C_{11} - C_{22} - C_{33}) \\ q_2^2 &= \frac{1}{4}(1 - C_{11} + C_{22} - C_{33}) \\ q_3^2 &= \frac{1}{4}(1 - C_{11} - C_{22} + C_{33}) \\ q_4^2 &= \frac{1}{4}(1 + C_{11} + C_{22} + C_{33}) \end{aligned} \tag{2.14}$$

Taking the square root of the maximum value amongst those terms provides the particular value for that term. The remaining terms can be computed using the appropriate formula from Equation 2.15.

$$\boldsymbol{q} = \frac{1}{4q_1}\begin{bmatrix} 4q_1^2 \\ C_{12} + C_{21} \\ C_{31} + C_{13} \\ C_{23} - C_{32} \end{bmatrix} = \frac{1}{4q_2}\begin{bmatrix} C_{12} + C_{21} \\ 4q_2^2 \\ C_{23} + C_{32} \\ C_{31} - C_{13} \end{bmatrix} = \frac{1}{4q_3}\begin{bmatrix} C_{31} + C_{13} \\ C_{23} + C_{32} \\ 4q_3^2 \\ C_{12} - C_{21} \end{bmatrix} = \frac{1}{4q_4}\begin{bmatrix} C_{23} - C_{32} \\ C_{31} - C_{13} \\ C_{12} - C_{21} \\ 4q_4^2 \end{bmatrix} \tag{2.15}$$

While the quaternion has more elements than the minimum required number of parameters for representing attitude, it offers the advantage that when moving from the DCM to the quaternion and back, only algebraic operations—no trigonometric operations—are required for conversion.

2.4.2 Euler Angles to/from Direction Cosine Matrix

The elements of the DCM can be determined from the associated Euler angles, though the precise equation depends on the particular Euler angle sequence (i.e. order of the rotations). The DCM for any Euler angle sequence can be constructed from the individual axis rotations presented in Equation 2.16, where the subscripts 1, 2, & 3 denote the axis about which the rotation is made (not the order of rotation).

$$R_1(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix} \quad R_2(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \quad R_3(\psi) = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.16}$$

For the standard (3-2-1) set of Euler angles corresponding to yaw-pitch-roll (ψ-θ-ϕ), the DCM is determined using Equation 2.17 (cos ϕ and sin ϕ have been abbreviated cϕ and sϕ, respectively). As seen in that equation, the individual rotation matrices R are combined according to the order of the Euler angle sequence, starting on the right and moving left.

$$C_{(3-2-1)} = R_1(\phi)R_2(\theta)R_3(\psi) = \begin{bmatrix} \mathrm{c}\theta\mathrm{c}\psi & \mathrm{c}\theta\mathrm{s}\psi & -\mathrm{s}\theta \\ \mathrm{s}\phi\mathrm{s}\theta\mathrm{c}\psi - \mathrm{c}\phi\mathrm{s}\psi & \mathrm{s}\phi\mathrm{s}\theta\mathrm{s}\psi + \mathrm{c}\phi\mathrm{c}\psi & \mathrm{s}\phi\mathrm{c}\theta \\ \mathrm{c}\phi\mathrm{s}\theta\mathrm{c}\psi + \mathrm{s}\phi\mathrm{s}\psi & \mathrm{c}\phi\mathrm{s}\theta\mathrm{s}\psi - \mathrm{s}\phi\mathrm{c}\psi & \mathrm{c}\phi\mathrm{c}\theta \end{bmatrix} \tag{2.17}$$

Extracting Euler angles from the DCM varies depending on the particular Euler angle sequence. The (3-2-1) set of Euler angles corresponding to yaw-pitch-roll (ψ-θ-ϕ) is determined using the Equation 2.18.

$$\begin{aligned} \psi &= \tan^{-1}\frac{C_{12}}{C_{11}} \\ \theta &= -\sin^{-1} C_{13} \\ \phi &= \tan^{-1}\frac{C_{23}}{C_{33}} \end{aligned} \tag{2.18}$$

For the special case where the attitude consists entirely of small-angle rotations, where small is defined as <5°, the DCM only differs from the identity matrix by small quantities, as seen in Equation 2.19. By removing any trigonometric operations in the transformation, this equation is useful both for high-rate control loops and deriving sensitivities to attitude for various filters. Great care must be taken when using this approach to ensure the small angle assumption is not violated.

$$C \approx \begin{bmatrix} 1 & \psi & -\theta \\ -\psi & 1 & \phi \\ \theta & -\phi & 1 \end{bmatrix} \tag{2.19}$$

2.4.3 Quaternion to/from Euler

A set of Euler angles is most easily determined from the quaternion through a series of two steps utilizing the transformations above. The quaternion are first transformed into a DCM using Equation 2.13. This DCM is then converted into a set of Euler angles with the transformation in Equation 2.18.

Similarly, the quaternion is most easily computed from a set of Euler angles using a two-step process. First, the set of Euler angles is transformed into a DCM using Equation 2.17. Equations 2.14 and 2.15 are then used to convert the DCM into the associated quaternion.

2.4.4 Attitude Kinematics

The attitude kinematic differential equation relates the time derivative of the attitude representation with the associated angular rate, ω (eg. as measured by a gyroscope). Each attitude representation can be written in terms of a kinematic differential equation.

The time rate of change of the DCM ($[\dot{C}]$) is given by Equation 2.20. The time rate of change of the quaternion ($\dot{q}$) is given by Equation 2.21. Finally, the time rate of change of yaw-pitch-roll ($\dot{\psi}$-$\dot{\theta}$-$\dot{\phi}$) is given by Equation 2.22. Note that this equation reveals that yaw rate, roll rate, and pitch rate are *not* equal to the angular rate measured by a gyro. Furthermore, Equation 2.22 reveals an additional singularity associated with Euler angles, in this case when $\cos\theta = 0$.

$$[\dot{C}] = -\begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix} [C] \tag{2.20}$$

$$\begin{pmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \\ \dot{q}_4 \end{pmatrix} = \frac{1}{2}\begin{bmatrix} q_1 & q_4 & -q_3 & q_2 \\ q_2 & q_3 & q_4 & -q_1 \\ q_3 & -q_2 & q_1 & q_4 \\ q_4 & -q_1 & -q_2 & -q_3 \end{bmatrix} \begin{pmatrix} 0 \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \tag{2.21}$$

$$\begin{pmatrix} \dot{\psi} \\ \dot{\theta} \\ \dot{\phi} \end{pmatrix} = \frac{1}{\cos\theta}\begin{bmatrix} 0 & \sin\phi & \cos\phi \\ 0 & \cos\phi\cos\theta & -\sin\phi\cos\theta \\ \cos\theta & \sin\phi\sin\theta & \cos\phi\sin\theta \end{bmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \tag{2.22}$$

2.5 ATTITUDE AND POSITION INTEGRATION

Inertial navigation systems use the angular rate and acceleration measurements from gyroscopes and accelerometers to determine the position, velocity, and attitude of a system by integrating this data over time. The non-linearities inherent to attitude require these integrations to occur at very high rates. To minimize the computational requirements of the user system, most inertial navigation systems output what are known as coning and sculling integrals which are integrated internally and can then be used at lower rates for full state integration.

2.5.1 Coning and Sculling Integrals

The coning and sculling integrals are integration processes that properly account for the coning and sculling motion and are valid despite the non-linearities inherent to real-world motion. Typically, the coning and sculling integrals are performed at higher rates, which allows the integration of the velocity and angular rate outputs to be performed at much lower speeds, thus reducing the amount of bandwidth needed to process the data. The coning integral provides a principal rotation vector known as Delta-Theta, $\Delta\boldsymbol{\theta}$, while the sculling integral generates a Delta-Velocity, $\Delta\boldsymbol{v}$, over specified amount of time, Δt.

These techniques have the advantage of providing the change in orientation and change in velocity over an arbitrary amount of time with higher accuracy as compared to averaging the accelerations or angular rates over longer time steps. In addition, the coning and sculling integrals provide the benefit of lower computational complexity as compared to other algorithms, such as the quaternion attitude update.

2.5.2 Attitude Integration

The Delta-Theta output from the coning integral is easily combined with quaternions to produce a continuously updated attitude estimate. An updated quaternion value ($\boldsymbol{q}_{k+1}$) is computed from the previous quaternion value ($\boldsymbol{q}_k$) using Equation 2.23. This equation assumes the scalar term of the quaternion is q_4 and that $\Delta\boldsymbol{\theta}$ is provided in radians.

$$\boldsymbol{q}_{k+1} = \begin{bmatrix} \cos\gamma[I_{3\times3}] - [\boldsymbol{\Psi}^{\times}] & \boldsymbol{\Psi} \\ -\boldsymbol{\Psi}^{\intercal} & \cos\gamma \end{bmatrix} \boldsymbol{q}_k \tag{2.23}$$

where

$$\gamma = \frac{\|\Delta\boldsymbol{\theta}\|}{2} \qquad \boldsymbol{\Psi} = \begin{cases} \frac{\sin\gamma}{2\gamma}\Delta\boldsymbol{\theta} & \gamma \geq 1\text{e}{-5} \\ \frac{1}{2}\Delta\boldsymbol{\theta} & \gamma < 1\text{e}{-5} \end{cases} \qquad [\boldsymbol{\Psi}^{\times}] = \begin{bmatrix} 0 & -\Psi_3 & \Psi_2 \\ \Psi_3 & 0 & -\Psi_1 \\ -\Psi_2 & \Psi_1 & 0 \end{bmatrix}$$

Because no computation can be achieved with perfect numerical precision, it is recommended that the updated quaternion is normalized per Equation 2.24 to ensure that this updated quaternion value remains unit length.

$$\hat{q}_{k+1} = \frac{\boldsymbol{q}_{k+1}}{\|\boldsymbol{q}_{k+1}\|} \tag{2.24}$$

Once the quaternion has been calculated from the Delta-Theta, this orientation can then be converted into the desired attitude representation. For more information about the quaternion and different attitude representations, refer to Sections 2.3 and 2.4.

2.5.3 Position and Velocity Integration

Information about an object's position can be obtained by integrating the velocity solution over a discrete period of time. Given a Delta-Velocity output in the body frame (${}^{B}\Delta\boldsymbol{v}$), the attitude at the start of the integration step is used to transform it into the inertial frame (typically NED):

$${}^{I}\Delta\boldsymbol{v} = [C]_k^{\intercal}\,{}^{B}\Delta\boldsymbol{v} \tag{2.25}$$

The inertial frame Delta-Velocity must then be corrected for gravity ($\boldsymbol{g}$) and the Coriolis term arising from Earth's angular rate ($\omega_{\oplus}$) and the current velocity estimate ($\boldsymbol{v}_k$), each in the inertial frame.

$$\begin{aligned} \Delta\boldsymbol{v}_{g/cor} &= \Delta t\left(\boldsymbol{g} - 2\boldsymbol{\omega}_{\oplus} \times \boldsymbol{v}_k\right) \\ \Delta\boldsymbol{v}_c &= {}^{I}\Delta\boldsymbol{v} + \Delta\boldsymbol{v}_{g/cor} \end{aligned} \tag{2.26}$$

VectorNav sensors can be configured to output the term $\Delta\boldsymbol{v}_c$ directly, utilizing the onboard Kalman filter attitude estimates, eliminating these steps. Once the corrected Delta-Velocity is available, the position and velocity can be easily updated via Equation 2.27.

$$\begin{aligned}\boldsymbol{v}_{k+1} &= \boldsymbol{v}_k + \Delta\boldsymbol{v}_c \\ \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + \Delta t\boldsymbol{v}_k + \frac{\Delta t}{2}\Delta\boldsymbol{v}_c\end{aligned} \tag{2.27}$$

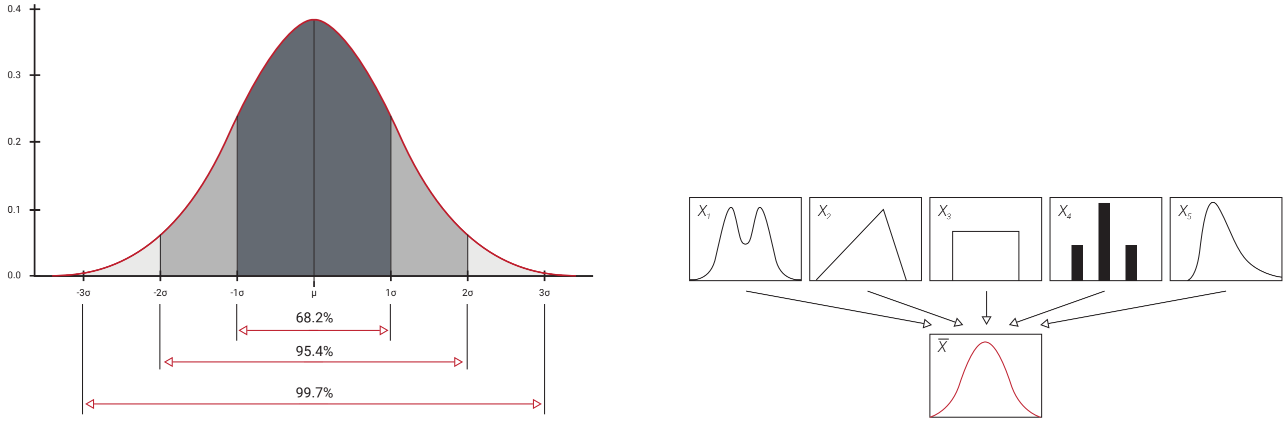
2.6 FILTERING BASICS

Sensors are prone to physical and electrical interference from their environment which can obscure input measurements. Noise, or the unwanted disturbances to the true input values, can often be statistically modeled and diminished with specific filters.

2.6.1 Gaussian Noise

In probability theory, a distribution termed the Gaussian distribution has found many applications in describing real-world events. This distribution is also known as a normal distribution or bell curve due to its shape, as seen in Figure 2.8a. The central limit theorem states that as independent random variables are summed, the result tends to approach a Gaussian distribution, even if the variables themselves are non-Gaussian, as illustrated in Figure 2.8b. As variables (or error sources) are combined, the result ends up resembling the normal distribution, making it a reasonable approximation for most real-world systems.

Gaussian (Normal) Distributions



(a) Distribution

(b) Central Limit Theorem

FIGURE 2.8

Values less than one standard deviation (1-σ), from the mean account for 68.2 % of all samples while two and three standard deviations (2-σ, 3-σ) contain 95.4 % and 99.7 % of all samples, respectively. Gaussian noise is noise that occurs within a Gaussian or normal distribution. It is assumed that Gaussian noise is zero-mean, with uncorrelated samples independent of previous values.

2.6.2 Standard Deviation, RMS, and Variance

When working with data sets, it can be useful to know the variation of a set of measurements. Variance, standard deviation, and the root-mean-square (RMS) are different ways to quantify this variation. The variance is a measure of the variation of a set of measurements with respect to themselves and can be described using Equation 2.28, where σ^2 is the variance, N is the total number of measurements in a data set, x_i is the i^{th} measurement in the data set, and μ is the mean of the set of measurements.

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.28}$$

Standard deviation is another way to describe the variation of a set of measurements with respect to themselves and is given by Equation 2.29.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.29}$$

From these equations, it can be seen that the standard deviation is simply the square root of the variance. When describing the variation in a set of measurements with respect to themselves, it is typically easier to visualize how the standard deviation relates to the original data set (since they have the same units) and is much more commonly used to describe the variation than the variance.

The root-mean-square (RMS) quantity is often a term that is used interchangeably with standard deviation, although these two characteristics have quite different meanings. RMS describes the variation of measurements with respect to the true value, rather than with respect to their mean, and can be found using Equation 2.30, where x_t is the true value that the measurements should read. If a quantity is unbiased—has zero-mean error—then RMS and standard deviation are indeed equivalent.

$$\text{RMS} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - x_t)^2} \tag{2.30}$$

2.6.3 Digital Filters

As microprocessor speeds have increased, it has become possible to process signals in software and apply digital filters to modify the input measurements. Digital filtering can be used in place of analog filtering to collect the desired data while removing noise. While this lowers the required circuit board component count, it does require additional software to handle the algorithms involved. Digital filters also have the advantage of being able to be adjusted within software, and can achieve filtering that would be difficult using analog components.

IIR Filter vs FIR Filter

The two main types of digital filters are known as infinite-impulse response (IIR) filters and finite-impulse response (FIR) filters. As shown in the most basic IIR filter in Equation 2.31, an IIR filter uses feedback in the form of the previous filtered result in filtering a measurement, so any anomaly that occurs in the data will always have some component left in the current value.

$$Y_k = \alpha Y_{k-1} + (1-\alpha)X_k \tag{2.31}$$

In this equation, the previous filtered result, Y_{k-1}, is the feedback term, X_k is the current measurement, and α is a multiplier less than 1. Larger values of α produce a smoother filtered result, while smaller values provide a more responsive (but noisier) output. IIR filters are easier and more efficient to implement in real-time since not much data buffering is required, but are less capable than FIR filters.

An FIR filter, the most common of which is referred to as a moving average filter or boxcar filter, uses previous measurements, but not previous outputs, in filtering a new measurement. These filters are always stable and can implement any type of filter response imaginable with enough previous values. A simple N-element low-pass boxcar filter can be implemented using Equation 2.32, where N is the number of samples used in the filter.

$$Y_k = \frac{1}{N}(X_k + X_{k-1} + ... + X_{k-N+1}) \tag{2.32}$$

Coefficients can also be applied to affect the frequency response of the filter. Additionally, FIR filters cause a linear phase shift for all frequencies, a useful property that can't be accomplished with analog or IIR filters. While this type of filter is generally considered more powerful, it does require a buffer to store previous values and can be more complex and time-consuming to implement than IIR filters.

Low-Pass, High-Pass, Band-Pass and Notch Filters

Depending on the frequency response desired, filters can be designed to pass or attenuate a range of different frequencies. Filters designed to pass low frequency signals while rejecting high frequency signals are called low-pass filters. High-pass filters reject low frequency signals and pass high frequency ones. Band-pass filters only allow signals in a certain range and reject all others, while notch or band-stop filters reject signals in a certain range and pass values outside of this range. The effect of each of these filters is shown in Figure 2.9.

From Figure 2.9, it can be seen that as the signal is swept from a low frequency to a higher frequency, the amplitude of the output response varies depending on the type of filter it passes through. Examples of filter use include applying a high-pass filter to a gyroscope to remove bias and a low-pass filter to an accelerometer to remove vibrations. Band pass filters can remove noise in signal transmission applications and band stop filters can remove specific troublesome frequencies.

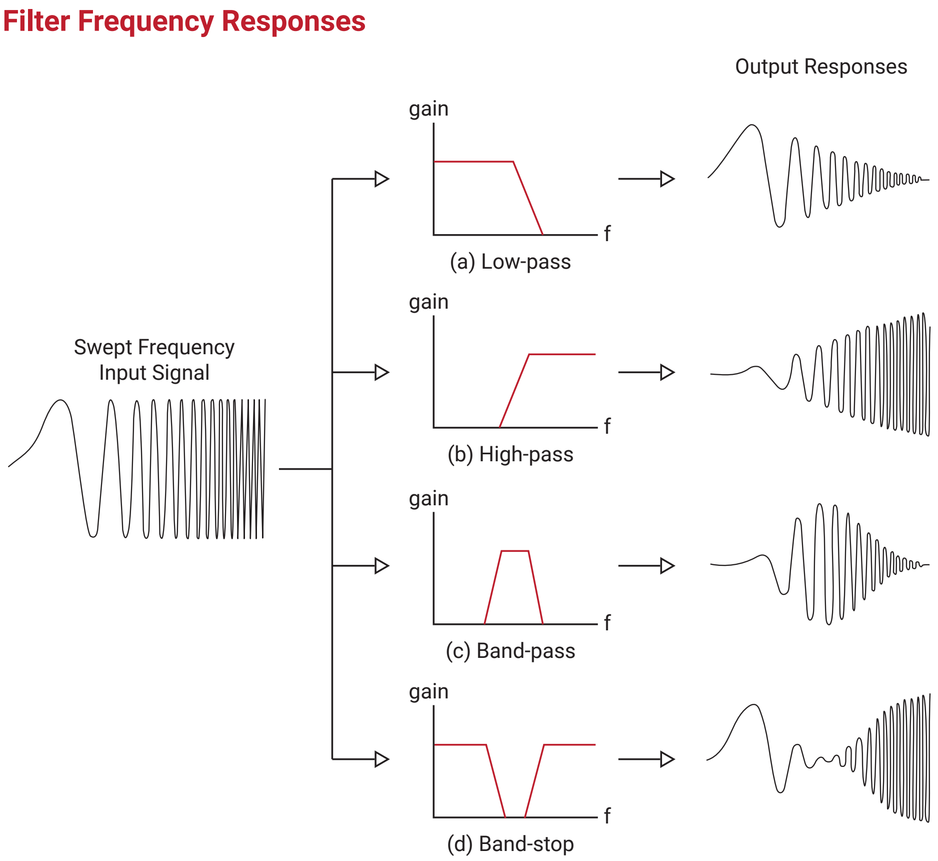


FIGURE 2.9

Phase Lag vs Smoothing

When a digital filter is applied to a measurement, a delay known as phase lag is introduced between the input signal and the filtered signal. As shown in Figure 2.10, the filtered output is smoother than the original signal and contains less change between successive data points, but there is a delayed response when sudden state changes occur in the original signal. Finding the best digital filter for an application requires making an acceptable trade-off between smoothness and phase lag.

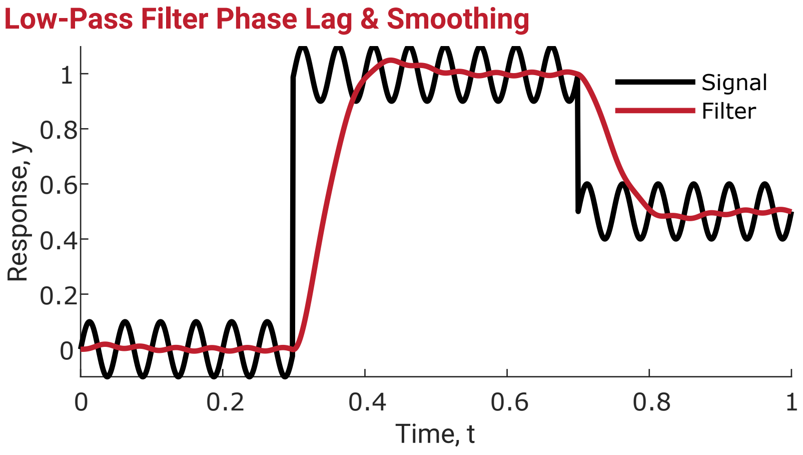


FIGURE 2.10

2.6.4 Complementary Filters

In certain situations, data from multiple sources can be combined using different filter types to determine a state, a simple method known as complementary filtering. For example, a gyroscope works best with high-pass filtering (HPF) to remove the gyro bias from the data, while an accelerometer is best when a low-pass filter (LPF) removes vibrations and other high frequency effects. As illustrated in Figure 2.11, the pitch/roll of a system can be determined by combining the filtered results of these two sensors.

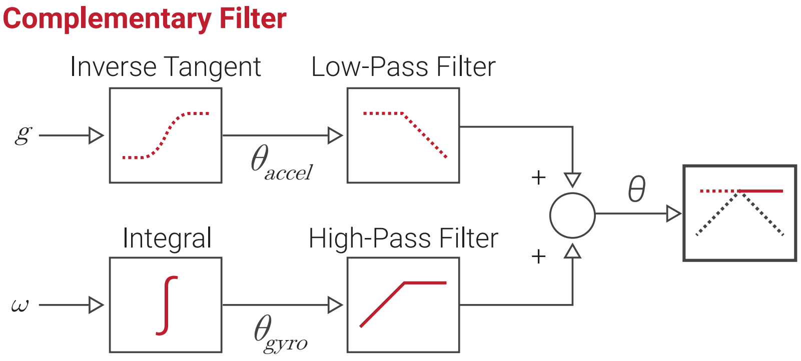


FIGURE 2.11

Mathematically, this is accomplished using Equation 2.33, in which θ is calculated the pitch or roll angle, ω is the angular rate measured by the gyro, and θ_{accel} is the pitch/roll angle derived from the accelerometer data (see Section 1.6).

$$\theta_{k+1} = \alpha(\theta_k + \omega \Delta t) + (1 - \alpha)\theta_{accel} \tag{2.33}$$

The weight, α, can be applied to adjust the contribution from each sensor and gives the system a complementary nature. The simplicity of the complementary filter is both its greatest strength and its greatest weakness. As discussed in Section 2.8, more advanced filters, like Kalman filters, can achieve substantially improved performance, but at the cost of computational complexity.

2.7 LEAST SQUARES

Least squares estimation is a batch estimation technique used to find a model that closely represents a collection of data and allows for the optimal determination of values or states within a system. This estimation technique can be applied to both linear and nonlinear system and is utilized in many different applications.

2.7.1 Linear Least Squares

Many real-world applications often contain an assortment of sensors that can be used to determine various parameters of interest in the system. These parameters of interest are typically referred to as states and can be anything needing to be tracked in a system, such as the position of a spacecraft or the level of saltwater in an aquarium tank.

Each of the states in a system are stored in a vector known as the state vector, $\boldsymbol{x}$. The assortment of sensors in a system is used to provide insight into what is actually happening in the system and how the system is changing over time. Each sensor yields a measurement which is stored in a vector known as the measurement vector, $\tilde{\boldsymbol{y}}$. However, these measurements often only provide information about a state indirectly and often require some type of conversion before being compared to the state vector. The measurement model matrix, H, describes this relationship between the measured values and the state values and is used to map the state vector, $\boldsymbol{x}$, into the measurement vector, $\tilde{\boldsymbol{y}}$, as shown in Equation 2.34, which also takes into account any noise, $\boldsymbol{\nu}$, in the measurement.

$$\tilde{\boldsymbol{y}} = H\boldsymbol{x} + \boldsymbol{\nu} \tag{2.34}$$

Ideally, when estimating a particular state, the error between the true value and the estimated value of the state should be minimized. However, in a real-world system, the true value of a state is never actually known due to various error sources, such as measurement errors and modeling errors. As a result, linear least squares instead seeks to minimize the residual error, or the error between the actual measurements, $\tilde{\boldsymbol{y}}$, and the measurements predicted from the measurement model and the estimated value of the state, $\hat{\boldsymbol{x}}$, as shown in the cost function of Equation 2.35. In this case, the optimal estimate of the state vector for a particular system is found using Equation 2.36.

$$\begin{aligned} J &= \frac{1}{2}\sum \boldsymbol{e}^{\intercal}\boldsymbol{e} \\ \boldsymbol{e} &= \tilde{\boldsymbol{y}} - H\hat{\boldsymbol{x}} \end{aligned} \tag{2.35}$$

$$\hat{\boldsymbol{x}} = (H^{\intercal}H)^{-1}H^{\intercal}\tilde{\boldsymbol{y}} \tag{2.36}$$

To visualize the linear least squares estimation process, consider the collection of data points plotted in Figure 2.12. While no line exists that can connect all of the data points together, there are a variety of lines that can be used to provide different fits of the data. As shown in detail in Appendix A.3, linear least squares determines the line that provides the best fit for this set of data. Though commonly used in curve fitting applications, linear least squares can be used in a variety of other applications as well including identifying the best model for a particular system or determining specific parameters of interest in a system.

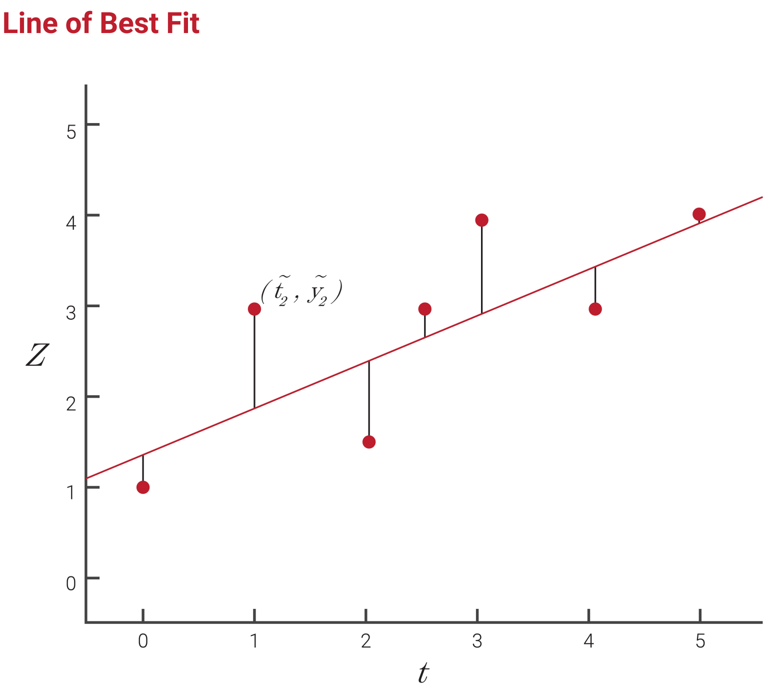

Line of Best Fit

FIGURE 2.12

2.7.2 Weighted Least Squares

The linear least squares solution determines the optimal estimate for each of the estimated state values by minimizing the residual error while weighing each of the measurements equally. However, many applications utilize numerous sensors that have varying performance specifications and uncertainties. In this case, weighing each measurement equally is not as useful. A technique known as weighted least squares adds an appropriate weight to each measurement to account for the uncertainty in each of the measurements. The linear least squares solution then becomes:

$$\hat{\boldsymbol{x}} = (H^\intercal W H)^{-1} H^\intercal W \tilde{\boldsymbol{y}} \tag{2.37}$$

where W is a symmetric, positive-definite matrix that contains the appropriate weights for each measurement. While any user-defined weights can be used in W, setting this matrix is set equal to the inverse of the measurement covariance matrix, R, yields optimal results.

$$\hat{\boldsymbol{x}} = (H^\intercal R^{-1} H)^{-1} H^\intercal R^{-1} \tilde{\boldsymbol{y}} \tag{2.38}$$

Note that in this case, the $(H^\intercal R^{-1} H)^{-1}$ term in the weighted least squares solution is then equal to the state covariance matrix, P.

$$P = (H^\intercal R^{-1} H)^{-1} \tag{2.39}$$

For more information about the measurement covariance matrix and the state covariance matrix, refer to Section 2.8.

2.7.3 Nonlinear Least Squares

While linear least squares can be used in various applications, some systems cannot be described by a linear model. For these nonlinear systems, the linear least squares solution can be extended to a nonlinear least squares solution, also known as the Gaussian Least Squares Differential Correction (GLSDC). The nonlinear least squares estimation process uses a model of the form:

$$\tilde{\boldsymbol{y}} = \boldsymbol{h}(\boldsymbol{x}) \tag{2.40}$$

where $\boldsymbol{h}(\boldsymbol{x})$ represents the equations of a nonlinear system. An optimal estimate for a nonlinear system can then be found by iterating the nonlinear least squares solution, using Equation 2.41.

$$\begin{aligned}\hat{\boldsymbol{x}}_{k+1} &= \hat{\boldsymbol{x}}_k + (H_k^{\intercal} H_k)^{-1} H_k^{\intercal} \left(\tilde{\boldsymbol{y}} - \boldsymbol{h}(\hat{\boldsymbol{x}}_k)\right) \\ H_k &= \frac{\delta \boldsymbol{h}}{\delta \hat{\boldsymbol{x}}_k}\end{aligned} \tag{2.41}$$

where the H matrix is known as the Jacobian matrix. Weighted versions of this calculation follow the same formulation as the linear case. Though this iterative process requires more computation than the linear least squares estimation process, nonlinear least squares provides the advantage of optimizing a wide range of real-world systems.

2.8 KALMAN FILTER

Originally developed in 1960 by R.E. Kalman, the Kalman filter provides an optimal estimate of how a system is going to change given noisy measurements and imperfect knowledge about the system. The Kalman filter is similar to least squares in many ways, but is a sequential estimation process, rather than a batch one. The standard Kalman filter is designed mainly for use in linear systems and is widely used in many different industries, including numerous navigation applications.

2.8.1 State Vector and State Covariance Matrix

As discussed in Section 2.7, many applications contain a number of different sensors that can be used to determine various parameters of interest in a system, commonly referred to as states. Unfortunately, a sensor is typically not able to provide a direct measurement of a state, but rather yields a noisy measurement containing some uncertainty that indirectly provides some information about one or more states.

The standard Kalman filter was developed to combine and process all available information about a linear dynamic system, both its dynamics model and any measurements, into an optimal estimate of the state. This state estimate is optimal in the sense that the variance is minimized, assuming that the measurement errors in the system are Gaussian and zero-mean. To help explain how a standard Kalman filter works, consider an aircraft that needs to track its position and velocity as it flies to its destination. The estimated position (p) and velocity (v) form the estimated state vector, $\hat{\boldsymbol{x}}$:

$$\boldsymbol{x} = \begin{bmatrix} p \\ v \end{bmatrix} \tag{2.42}$$

The key to a Kalman filter is that in addition to maintaining an estimate of the state, it also tracks an estimate of the uncertainty associated with that state—recognizing that the position and velocity estimates are never perfectly known. This uncertainty can be represented by a matrix known as the state covariance matrix, P. The state covariance matrix consists of the variances associated with each of the state estimates as well as the correlation between the errors in the state estimates. The covariance matrix can be graphically represented by n-D ellipsoids (where n is the number of states), where a particular ellipsoid maps to the 1-σ, 2-σ, or 3-σ uncertainty bounds.

In this particular system, σ_p^2 and σ_v^2 in Equation 2.43 are the variances associated with each of the state estimates and γ_{pv} is a value between 0 and 1, representing the correlation between the position and velocity errors.

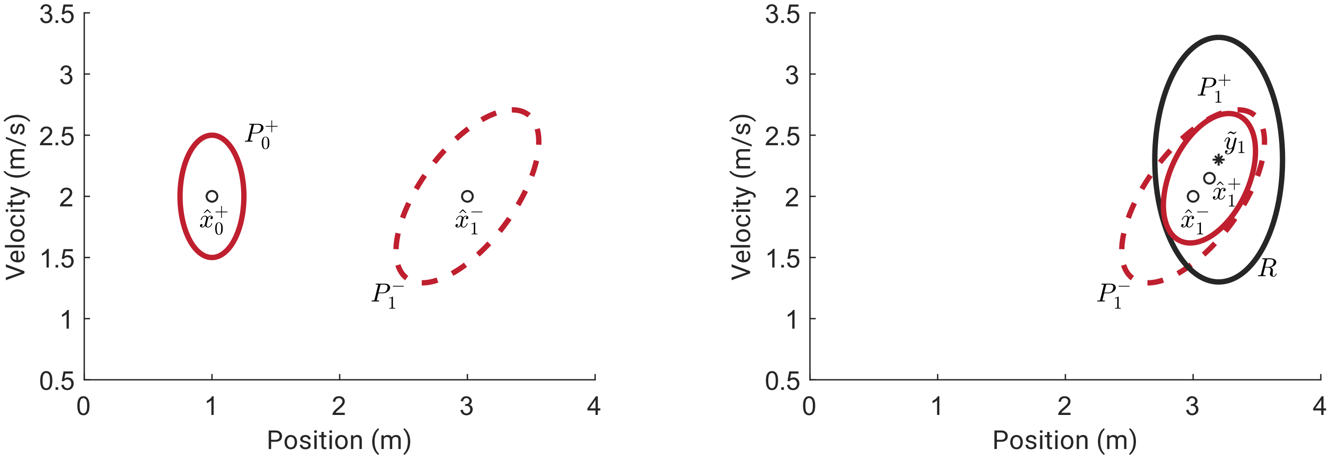
$$P = \begin{bmatrix} \sigma_p^2 & \gamma_{pv}\sigma_p\sigma_v \\ \gamma_{pv}\sigma_p\sigma_v & \sigma_v^2 \end{bmatrix} \tag{2.43}$$

2.8.2 Two-Step Process

The Kalman filter is designed to maintain an optimal estimate of the state vector, given the state covariance matrix, the system dynamic model, and noisy measurements ($\tilde{y}$) with their own associated measurement covariance matrix (R). This is achieved through a continuous two-step process: (a) propagate the state and covariance through the dynamic model from one time step to the next, and (b) process a measurement update at each time step where one exists.

The standard Kalman filter is ideal for real-time applications as it has the advantage of operating on a sequential basis, taking measurements and processing them as they are available, as opposed to a batch estimation technique like Least Squares estimation which requires that all of the data be available before any processing can be done. As a result, this estimation technique is frequently used in a wide range of different systems, including many navigation applications. For example, an attitude and heading reference system (AHRS) as well as a GNSS/INS system both rely on a Kalman filter for estimation. An AHRS predicts its next state estimates by integrating its current acceleration and angular rate measurements, this prediction is then updated by the next set of acceleration and angular rate measurements received from the system. Similarly, a GNSS/INS system utilizes the inertial navigation system to predict its next state estimate and subsequently updates this prediction using measurements from the GNSS receiver.

Kalman Filter Steps



(a) Propagate

(b) Update

FIGURE 2.13

2.8.3 Propagate Step

During the propagate step, a mathematical model is applied over a specified period of time to predict the next state vector and state covariance matrix of the system. This mathematical model is known as the state transition matrix, Φ, and is used to define how the state vector changes over time. In navigation applications, this matrix is generally a function of the system dynamics. For the case of this example, the changes in position and velocity over time can be defined by:

$$\begin{aligned} p_k &= p_{k-1} + v_{k-1}\Delta t \\ v_k &= v_{k-1} \end{aligned} \tag{2.44}$$

These equations can then be mapped into the state transition matrix, as shown in Equation 2.45.

$$\Phi = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \tag{2.45}$$

When propagating the state vector there may be unknown changes or disturbances that are not modeled in the state transition matrix, such as a crosswind that pushes an aircraft off course. These untracked changes can cause errors in the prediction of the state values and are typically modeled as noise with some uncertainty. The system noise covariance matrix, Q, is used to account for this uncertainty. To complete the propagation step, the next state vector and state covariance matrix can be predicted using Equations 2.46 and 2.47.

$$\boldsymbol{x}_k^- = \Phi_{k-1}\boldsymbol{x}_{k-1}^+ \tag{2.46}$$

$$P_k^- = \Phi_{k-1}P_{k-1}^+\Phi_{k-1}^\intercal + Q_{k-1} \tag{2.47}$$

Note that the negative superscript in $\boldsymbol{x}_k^-$ and P_k^- is used to indicate that these parameters correspond to the propagate step at time t_k and are the estimates of the state vector and state covariance matrix prior to the measurement update step. A positive superscript as in the $\boldsymbol{x}_{k-1}^+$ and P_{k-1}^+ indicate that those parameters are estimates of the state vector and state covariance matrix after the update step at time t_{k-1}.

As shown in Figure 2.13a, during propagation the uncertainties in the state covariance matrix grow due to the added uncertainty from the system noise covariance matrix. Based on the strong correlation between velocity and position in this particular example, the shape of the 1-σ ellipse also skewed to account for the fact that an uncertainty in velocity at t_0 maps to an uncertainty in position at t_1. Indeed, the system dynamics (represented by Φ) can generally transform the uncertainty in arbitrary ways.

2.8.4 Update Step

The update step follows the propagation step and combines the prediction of the state vector and state covariance matrix with any available measurement data from the suite of sensors on board a system to provide an optimal estimate of these parameters. The Kalman filter assumes these measurements are valid at an instantaneous time of validity, so timing is critical.

As described in Section 2.7, each of the measurements taken from a sensor is stored in a measurement vector, $\tilde{\boldsymbol{y}}$, and typically requires a conversion prior to being compared to a state. The measurement model matrix, H, is used to map the state vector, $\boldsymbol{x}$, into the measurement vector, $\tilde{\boldsymbol{y}}$, as shown in Equation 2.48.

$$\tilde{\boldsymbol{y}} = H\boldsymbol{x} + \boldsymbol{\nu} \tag{2.48}$$

Since all sensors produce a measurement with at least some uncertainty (e.g. from noise), the true measured value could actually be a range of possible readings. This uncertainty in each of the measurements is represented by the measurement noise covariance matrix, R.

The combination of the measurement model H and the measurement covariance R defines the *observability* of the system: how well individual states can be estimated (if at all) given those measurements. Low observability translates to large state covariances, and in some cases divergence of the filter if full-state observability is lost for an extended time period.

Once the measurement vector, measurement model matrix, and measurement noise covariance matrix have been formulated, the propagated state vector and state covariance matrix can be combined with this measurement information to provide an updated estimate of the states and state covariance. When combining the measurements with the predictions, a matrix known as the Kalman gain, K, is used to weight the measurement information by comparing the uncertainty of the measurement vector (R) with the current uncertainty of the state vector (P). The Kalman gain is designed such that it is optimal for the system to minimize the variance of the state estimates and can be found using:

$$K_k = P_k^- H_k^\intercal (H_k P_k^- H_k^\intercal + R_k)^{-1} \tag{2.49}$$

To complete the update step, the estimates of the state vector and state covariance matrix are both updated such that:

$$\begin{aligned} \boldsymbol{x}_k^+ &= \boldsymbol{x}_k^- + K_k(\tilde{\boldsymbol{y}}_k - H_k\boldsymbol{x}_k^-) \\ P_k^+ &= (I - K_k H_k)P_k^- \end{aligned} \tag{2.50}$$

In the aircraft example considered previously, suppose that a sensor is available to measure both position and velocity directly (eg. GPS). The measurement model matrix would then be:

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.51}$$

Given the measurement uncertainty R around a noisy measurement $\tilde{\boldsymbol{y}}$, shown in Figure 2.13b, the state and covariance are updated using Equation 2.50. By incorporating the available measurement data with the prediction of the state vector during the update step, the state covariances in the state covariance matrix shrink as a result of this additional insight into how the system is changing over time.

2.8.5 Tuning a Kalman Filter

Nominally, the terms Q and R represent Gaussian noise that are well-modelled or measured—after all, that is the basis for optimality of a Kalman filter. However, when designing a Kalman filter for a real system, there will always be modelling errors in the propagate step and non-noise-based measurement errors (eg. scale factor errors). To account for these errors and ensure best performance, tuning of the Kalman filter may be required.

Tuning a Kalman filter involves artificially or arbitrarily inflating the system noise covariance matrix, Q, as well as the measurement noise covariance matrix, R, and in some cases adjusting the initial system covariance matrix, P_0, as well. These parameters are often referred to as the tuning knobs for the system and must be large enough to maintain enough uncertainty to account for the unmodeled or incorrectly modeled errors, but not too large as unstable behavior could result from the system. The tuning of a Kalman filter is a delicate balancing act that is more of an art than a science. Due to this, tuning is best left to the professionals and is typically performed by the designer of the Kalman filter rather than the end user.

2.9 NONLINEAR KALMAN FILTER

The standard Kalman filter is designed mainly for use in linear systems, however, versions of this estimation process have been developed for nonlinear systems, including the extended Kalman filter and the unscented Kalman filter. Since many real-world systems cannot be described by linear models, these nonlinear estimation techniques play a large role in numerous real-world applications.

2.9.1 Extended Kalman Filter

While the standard Kalman filter is a powerful estimation tool, its algorithms begin to break down when the system being estimated is nonlinear. Fortunately, a version of the standard Kalman filter, known as the extended Kalman filter (EKF), has been extended to nonlinear systems and relies on linearization in estimating these nonlinear systems. Linearization operates on the principle that at a small section around a selected operating point a nonlinear function can be approximated as a linear function. This linearized function can be derived from the nonlinear function using the first-order terms in a Taylor series expansion shown in Equation 2.52.

$$\boldsymbol{g}(x) \approx \boldsymbol{g}(a) + \left.\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{x=a} (x-a) \tag{2.52}$$

Using this method of linearization, an EKF will follow the same propagate and update process as the standard Kalman filter, but with a few modifications to the standard equations. During the propagate step, rather than using Equation 5 in Section 2.8, the state vector is instead estimated by evaluating the nonlinear system model equations at the most recent state estimate as shown in Equation 2.53. Additionally, in the state covariance matrix propagation, the state transition matrix, Φ, is replaced with a matrix F, which is a Jacobian matrix containing the first-order partial derivatives of the nonlinear system model equations.

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k, \boldsymbol{u}_k) \qquad F = \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}} \tag{2.53}$$

In the update step, the expected measurement vector is derived using the nonlinear measurement model equations, evaluated at the most recent state estimate as provided in Equation 2.54. The measurement model matrix in each of the update equations is also replaced with the H Jacobian matrix containing the first-order partial derivatives of the nonlinear measurement model equations.

$$\boldsymbol{y}_k = h(\boldsymbol{x}_k) \qquad H = \left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}} \tag{2.54}$$

Though the EKF can be a powerful tool in estimating states in a nonlinear system, there are some limitations of its use. An EKF is designed in such a way to optimally update the state vector and state covariance matrix, assuming that the state covariance matrix is within the linear region of the linearization. However, if the uncertainties in the state covariance matrix grow to be larger than the size of this linear region, then the state covariance matrix can no longer accurately reflect the actual error in the system and divergence can occur. Typically, an EKF is best suited for applications with enough measurements to keep state uncertainties relatively low.

2.9.2 Unscented Kalman Filter

While the EKF works well for a majority of nonlinear systems, there are some cases where an EKF is not well suited, such as if the system is very nonlinear or poorly observable. In these particular systems, the unscented Kalman filter (UKF) can provide a more reliable estimation. Most navigation systems do not fall in this category, but the UKF is still seen in some systems.

The UKF estimates a nonlinear system by carefully selecting a number of points, known as sigma points, that adequately describe the state vector and associated uncertainty. These sigma points are then propagated through the nonlinear equations to estimate the next state vector and related uncertainty.

Though this estimation process is not as prone to divergence, a UKF does require quite a bit of computational efficiency to calculate the sigma points and propagate them through the nonlinear system. This is especially true for systems that have a large state vector, requiring large numbers of sigma points to be calculated and propagated.

2.10 FEEDBACK CONTROLS

There are two types of controls for dynamic systems: open-loop control and closed-loop (feedback) control. An open-loop system uses only a model of the system without the support of measuring the system response. For example, a conveyor belt that should move at a constant speed may be controlled by setting a constant voltage on the motor which should map to a particular speed given the typical motor and friction of the system. Of course, if the conveyor belt is overloaded, it will move slower than desired and the open-loop controller has no mechanism for correcting it. A closed-loop controller, on the other hand, feeds the measured response of the system back into its control calculations, providing the ability to accurately track the desired output even under varying conditions. For instance, an encoder measuring the turn rate of the conveyor belt would allow a closed-loop controller to maintain the desired speed regardless of loading. Generally, a closed-loop controller measures or estimates an error value, $e(t)$—the difference between the desired state and the measured state—and derives a control input based on that error signal.

2.10.1 Proportional-Integral-Derivative Controller (PID Controller)

Control systems can be implemented with many different control algorithms, but the vast majority can be characterized as a proportional-integral-derivative (PID) controller. There are three types of control provided by a PID controller based on its namesake: proportional control, integral control, and derivative control. A feedback correction term is then derived using separate gain values from each of the different types of feedback control used in the controller, as illustrated in Figure 2.14. Entire fields of study are devoted to solving for the optimal gains for each of these terms across a wide range of systems.

PID Controller

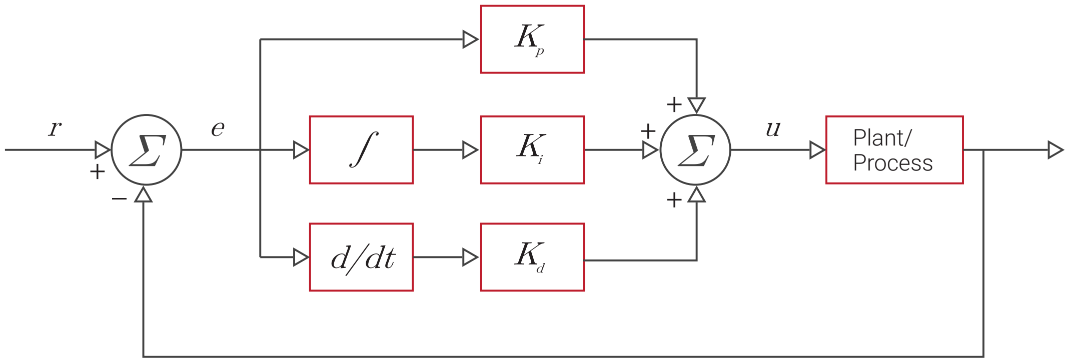


FIGURE 2.14

Proportional

The proportional term (P) applies a multiplier known as the gain, K_p, to the value of the direct difference between the measured state and the desired state, as shown in Equation 2.55. Proportional control is the easiest type of feedback control to implement, though it leads to steady-state errors in a system if used alone.

$$u(t) = K_p * e(t) \tag{2.55}$$

Integral

The integral term (I) integrates the error between the desired state and the measured state over time and scales this by a gain value, K_I, as shown in Equation 2.56. When used in conjunction with proportional control, this integration acts as a low-pass filter and eliminates the steady-state errors in the system. However, integration of the error can also wind up the controller and lead the system to overshoot the desired value.

$$u(t) = K_I \int_0^t e(\tau)\mathrm{d}\tau \tag{2.56}$$

Derivative

The derivative term (D) acts on the current rate of change of the error and is scaled by a gain value, K_D, as shown in Equation 2.57. This allows the controller to anticipate the future trend of the error. The more rapid the rate of change of the error, the more impact the derivative controller has on the system. When properly tuned, the derivative term prevents the system from overshooting the desired state.

$$u(t) = K_D \frac{de(t)}{dt} \tag{2.57}$$

Each of these feedback correction terms are then summed together and used to calculate an input to the system being controlled, commonly referred to as the plant, to drive the dynamics of the system to the desired state. The individual gain values from each of the different control terms can be tuned to achieve the desired response of the system. Typically, this tuning is performed by the designer of the control system through testing and experimentation.

To visualize how a PID controller works, consider the adaptive cruise control system found commonly in present-day automobiles. Sensors on the automobile take measurements to assess the current conditions of the vehicle and the surrounding environment, such as how fast the automobile is travelling or if another vehicle is approaching. The PID controller then uses this information to derive corrections that will either speed up or slow down the vehicle to achieve the desired speed without overshooting this value.

2.10.2 Latency

In the case of feedback controls, latency is the time delay between when a real-world event occurs and when this data is fed back into the controller. This time delay can cause performance degradation in a system and can even lead to a system becoming unstable or uncontrollable if the latency is longer in duration than the system's time constants. Latency can be at least partially mitigated with appropriate time stamping of the data which informs a controller when the real-world event actually occurred and is commonly referred to as the time of validity of the data. Additionally, lowering the latency in a system, such as by utilizing a real-time operating system, improves the ability to control the system using feedback controls.

Real-Time Operating System (RTOS)

A real-time operating system (RTOS) is a type of operating system (OS) whose purpose is to serve applications needing to process data in real-time with minimal latency. An RTOS has well-defined, fixed time constraints and provides finer controls over priorities and switching behavior between processes. This type of OS also has specialized replacement functions that are re-entrant, meaning that if a process is stopped for whatever reason (e.g. to handle a higher-priority event), the system can re-enter that function and continue where it left off. To minimize latency errors, an RTOS is a crucial component in feedback control systems, which are designed to have fast, low latency feedback responses to ensure proper performance without degradation.

In contrast, the timing on a non-RTOS systems like Windows or standard builds of Linux varies dramatically depending on everything else happening on the system. When timestamping data arriving over a serial port from a sensor, it is not uncommon for the timestamps to vary by 10s or even 100s of milliseconds, despite the data physically arriving at the serial port at a constant rate.

3 SPECIFICATIONS & ERROR BUDGETS

To determine the expected performance of a sensor within a specific application and environment, users should consult defined engineering specifications, calibration methods, and system error budgets. These typical performance expectations, as well as their respective confidence levels, are given for a wide variety of operating conditions. Environmental factors, such as temperature, velocity, or GNSS capabilities, may place real limitations on the performance of a system but can often be accurately predicted and accounted for.

3.1 IMU SPECIFICATIONS

Physical limitations constrain the operation of IMUs in certain applications. A wide variety of standard specifications, and the performance which they describe, should be reviewed to ensure the selection of an appropriate IMU to meet application requirements.

3.1.1 Common Specifications

Range

The range is defined as the minimum and maximum input values a sensor can measure. Anything outside of the range will not be measured or outputted by the sensor.

Resolution

The most precise unit of measure that can be outputted over a sensor's range is known as the resolution. Generally, this specification is not particularly important as most sensors used today have high resolution.

Bandwidth

The bandwidth is the maximum frequency to which a sensor or system will respond. This frequency is typically defined at the point where the response has dropped to half power, or the −3 dB point on a Bode magnitude plot. Bandwidth not only defines the frequencies a sensor can measure, it is inversely related to the time constant of the analog sensor response. Higher bandwidth sensors react more quickly to a given stimulus, and this rise time is the first component in calculating a total system latency.

Sample Rate

Sample rate is the number of samples output by a sensor per second. While sometimes used interchangeably with bandwidth, the sample rate differs as it can be any specified rate while the bandwidth is dependent upon the sensor or system response.

3.1.2 Noise

Noise is classified as any random variation of a measured output when a sensor is subjected to constant input at constant conditions and is usually characterized by either a standard deviation value or a root-mean-square (RMS) value.

Noise Density

Averaging a noisy signal and outputting it at a lower rate, a process known as downsampling, reduces the measured noise of a sensor output. So a single noise standard deviation is insufficient to properly characterize the inherent noisiness of a sensor. While some datasheets will specify that standard deviation at a particular sample rate, the more common measure is noise density which provides the noise divided by the square root of the sampling rate. For example, the noise density for a gyroscope can be represented as $^\circ/\mathrm{s}/\sqrt{\mathrm{Hz}}$ or $^\circ/\mathrm{hr}/\sqrt{\mathrm{Hz}}$. By multiplying the noise density (ND) by the square root of the sampling rate (SR), the noise standard deviation (σ) at that rate can be recovered ($\sigma = \mathrm{ND}\sqrt{\mathrm{SR}}$). Sometimes noise density is specified as a power spectral density, which is simply the noise density squared. In the case of a gyroscope, this yields units of $(^\circ/\mathrm{s})^2$/Hz.

Random Walk

If a noisy output signal from a sensor is integrated, for example integrating an angular rate signal to determine an angle, the integration will drift over time due to the noise. This drift is called random walk, as it will appear that the integration is taking random steps from one sample to the next. The two main types of random walk for inertial sensors are referred to as angle random walk (ARW), which is applicable to gyroscopes, and velocity random walk (VRW), which is applicable to accelerometers. The specification for random walk is typically given in units of $^\circ/\sqrt{\text{s}}$ or $^\circ/\sqrt{\text{hr}}$ for gyroscopes, and m/s/$\sqrt{\text{s}}$ or m/s/$\sqrt{\text{hr}}$ for accelerometers. By multiplying the random walk by the square root of time, the standard deviation of the drift due to noise can be recovered.

Unit Conversions

Given three different ways to specify noise (standard deviation at rate, noise density, random walk) and multiple different sets of units used to specify each, it is important to understand how to convert all of them into a common form to get an accurate comparison of different sensors. This mostly requires comfort in performing unit conversions. Appendix A.2 works through a number of examples, but here are a few relationships that are particularly useful.

First, it is important to realize that Hertz (Hz) is defined as the inverse of seconds, which means that a noise density specification of X°/s/$\sqrt{\text{Hz}}$ is exactly equivalent to an angle random walk specification of $X^\circ/\sqrt{\text{s}}$ with no conversion necessary.

Second, when working with the square root of time, converting between hours and seconds is a factor of 60 instead of 3600: $60\sqrt{\text{s}} = \sqrt{\text{hr}}$. So an angle random walk specification of $X^\circ/\sqrt{\text{s}}$ is equivalent to a specification of $(60{\cdot}X)\,^\circ/\sqrt{\text{hr}}$.

Finally, the units used for specifying accelerometers often switch between the SI m/s^2 and the more common milli-g (mg) (or even micro-g, μg). It is useful to remember that 1 mg $\approx$ 0.01m/s^2. So a noise density specification of X mg/$\sqrt{\text{Hz}}$ is roughly equal to a velocity random walk specification of $(0.01X)$ m/s/$\sqrt{\text{s}}$.

3.1.3 Bias

The bias is a constant offset of the output value from the input value. There are many different types of bias parameters that can be measured, including in-run bias stability, turn-on bias stability or repeatability, and bias over temperature.

In-Run Bias Stability OR Bias Instability

The in-run bias stability, or often called the bias instability, is a measure of how the bias will drift during operation over time at a constant temperature. This parameter also represents the best possible accuracy with which a sensor's bias can be estimated. Due to this, in-run bias stability is generally the most critical specification as it gives a floor to how accurately a bias can be measured.

Turn-on Bias Stability OR Bias Repeatability

When a sensor is started up, there is an initial bias present that can fluctuate in value from one turn-on to the next due to thermal, physical, mechanical, and electrical variations between measurements. This change in initial bias at constant conditions (eg. temperature) over the lifetime of the sensor is known as the turn-on bias stability, or is sometimes referred to as bias repeatability. While this initial bias cannot be calibrated during production due to its varying nature, an aided inertial navigation system (eg. GNSS-aided) can estimate this bias after each startup and account for it in the outputted measurements. The turn-on bias stability is most relevant for unaided inertial navigation systems or those performing gyrocompassing.

Bias Temperature Sensitivity

As a sensor is operated in a range of temperatures, the bias may respond differently to each of these temperatures. This parameter is known as bias temperature sensitivity and can be calibrated for after each of these biases are measured over the temperature range. However, this bias can only be measured within the limits of the in-run bias stability.

3.1.4 Scale Factor

Scale factor is a multiplier on a signal that is comprised of a ratio of the output to the input over the measurement range. This factor typically varies over temperature and must be calibrated for over the operational temperature range.

Scale Factor Error (ppm or %)

The scale factor will not be perfectly calibrated and will have some error in the estimated ratio. This error is categorized as one of two equivalent values, either as a parts per million error (ppm), or as an error percentage. As shown in Figure 3.1, the scale factor error causes the output reported to be different from the true output. For example, if the z-axis of an accelerometer only measures gravity (9.81 m/s^2), then the bias-corrected sensor output should be

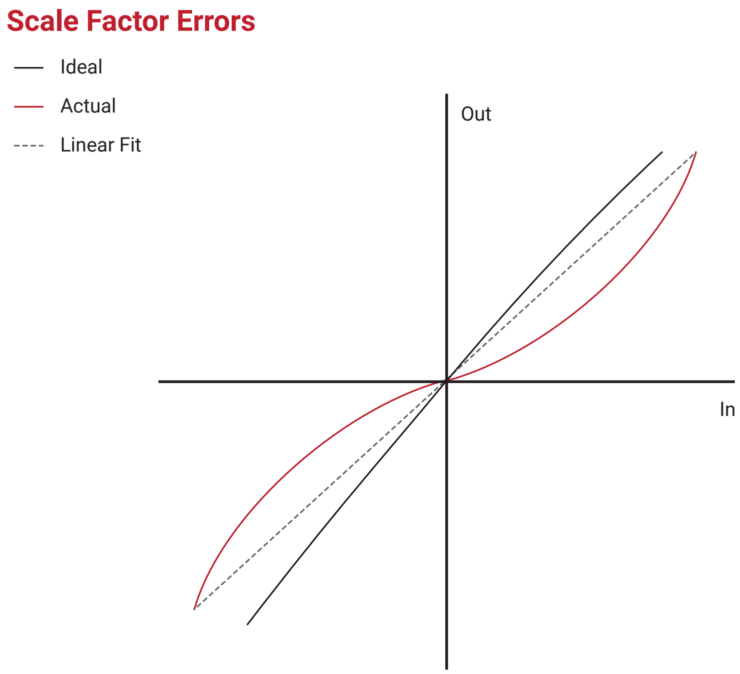


FIGURE 3.1

9.81 m/s^2. However, with a scale factor error of 0.1%, or 1.000 ppm, the output value from the sensor will instead be 9.82 m/s^2.

Scale Factor Nonlinearity (ppm or %FS)

The scale factor can also have errors associated with the ratio being non-linear, known as nonlinearity errors, as shown in Figure 3.1. The linearity error of the scale factor is also described as either a parts per million error (ppm), or as a percentage of the full scale range of the sensor.

3.1.5 Orthogonality Errors

When mounting and aligning sensors to an IMU, it is impossible to mount them perfectly orthogonal to each other. As a result, orthogonality errors are caused by a sensor axis responding to an input that should be orthogonal to the sensing direction. The two main types of orthogonality errors are cross-axis sensitivity and misalignment, both of which are often used interchangeably.

Cross-Axis Sensitivity

As defined here, cross-axis sensitivity is an orthogonality error caused by a sensor axis being sensitive to an input on a different axis. In other words, an input in the x-axis may also be partially sensed in the z-axis.

Misalignment

As defined here, misalignment is an orthogonality error resulting from a rigid-body rotation that offsets all axes relative to the expected input axes, while maintaining strict orthogonality between the sensing axes (x-axis remains orthogonal to y- and z-axes). In other words, the internal sensing axes do not align to the axes marked on the case of the IMU. Such misalignment errors exist throughout a full system (e.g. the IMU case is not mounted perfectly aligned to the vehicle, the camera lens is not perfectly aligned to the camera case, etc.). So most high-performance applications require the performance of a separate misalignment calibration across the full end-to-end system (e.g. internal IMU sensing axes to camera lens), and the factory-calibrated misalignment errors of an IMU are largely irrelevant.

3.1.6 Acceleration Sensitivity for Gyroscopes

In an ideal world, a gyroscope would only measure angular rate and would have no sensitivity to linear acceleration. However, in practice gyroscopes are susceptible to linear accelerations due to their asymmetrical design and manufacturing tolerances. The two types of linear acceleration sensitivities are referred to as g-sensitivity and g^2-sensitivity.

g-Sensitivity

An acceleration sensitivity in which a gyroscope experiences a bias shift when subjected to a constant linear acceleration is known as a g-sensitivity. Gyroscopes must be tested for sensitivity to linear accelerations both parallel and perpendicular to the sensing axis.

g^2-Sensitivity OR Vibration Rectification

The g^2-sensitivity specification is an acceleration sensitivity that causes a bias shift in the output of a gyroscope due to oscillatory linear accelerations. As with g-sensitivity, gyroscopes must be tested for sensitivity to linear accelerations both parallel and perpendicular to the sensing axis.

3.2 IMU CALIBRATION & CHARACTERIZATION

Sensor calibration is a method of improving a sensor's performance by removing structural errors from the measurements of the sensor. Structural errors are differences between a sensor's expected output and its measured output, which show up consistently every time a new measurement is taken. By removing the structural errors, calibration provides a means of improving the overall accuracy of the underlying sensors. These errors are removed through various processes utilizing different equipment and test setups, including rate tables, tumble tests, and thermal chambers.

Sensor characterization incorporates a wide range of tests, including Allan Variance and vibration, to assess the inherent characteristics of the inertial sensors—specifications like noise that cannot be improved through a calibration process. Characterization testing can also be used to verify the post-calibration performance of a sensor to determine parameters such as scale factor error, using the same general testing process used for calibration (though independent of the actual calibration data collection).

3.2.1 Sensor Model

Each sensor is calibrated using the linear sensor model shown in Equation 3.1. This equation applies for gyroscopes, accelerometers, and magnetometers. A least-squares fit is performed on the data collected during the calibration process to determine the values of each of the parameters in the model.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & S_z \end{bmatrix} \begin{bmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_3 & 1 & \alpha_4 \\ \alpha_5 & \alpha_6 & 1 \end{bmatrix} \left(\begin{bmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{bmatrix} - \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix} \right) \tag{3.1}$$

The left-hand side vector is the output from the sensor that has been calibrated for any scale factor errors, misalignment errors, and biases, and should match the truth data after calibration. The two matrices on the right-hand side are known as the scale factor matrix and misalignment matrix, respectively, which contain the sensor's scale factor and misalignment calibration parameters. The scale factor matrix and the misalignment matrix above can also be combined into a single matrix without loss of generality. The last term consists of the uncalibrated measured values and sensor biases. Typically, the biases are the only parameters in the sensor model equation that change significantly over the life of the part and that may require occasional calibration by the user.

Additional terms can be added to the model if they represent significant error sources, including (but not limited to):

- Non-linearity: Rather than a single value for the scale-factor, a polynomial or piece-wise continuous model of the scale factor can be used to address non-linearities across the measurement range.
- g-sensitivity: A gyroscope's bias sensitivity to linear acceleration can be calibrated and used to offset the bias in real time using the measured (and calibrated) accelerometer values.
- Temperature sensitivity: Most calibration parameters in these models have a sensitivity to temperature, so calibrating across the operational temperature range is required. This results in either a look-up table for each of these parameters versus temperature or a polynomial representation.
- Thermal ramp-rate sensitivity: Some advanced sensor calibration models also incorporate a sensitivity to thermal ramp rate, rather than simply assuming a constant temperature.

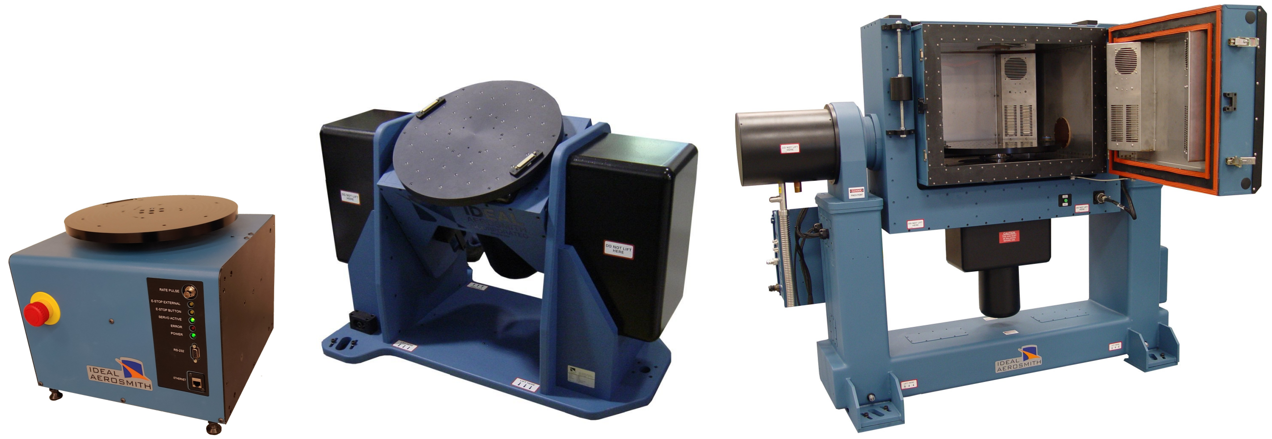
3.2.2 Calibration Process

Accelerometers are typically calibrated through a process known as a *tumble test*. During this test, static measurements are taken in different orientations. A tumble test is usually performed by mounting the accelerometer on a cube or a multi-axis turntable that allows each face to be rotated. This exercises the accelerometer by aligning each sensor axis (X, Y, and Z) to the direction of, the direction opposite of, and the directions perpendicular to known gravity as it is rotated, providing a ± 1 g measurement.

Gyroscopes are calibrated using a precise device known as a *rate table*. As shown in Figure 3.2a, rate tables have a circular platform connected to a brushless electric motor capable of providing very precise angular and angular rate outputs. The circular rotating platform allows the gyroscope to rotate and be calibrated across a range of angular rates. Rate tables can be purchased as either single-axis (Fig. 3.2a), 2-axis (Fig. 3.2b), or 3-axis systems. A 2-axis system allows for all three gyro axes to be calibrated in a single setup (the tumble test can also be performed), whereas a 3-axis system has the added benefit of being able to reconstruct arbitrary motion profiles.

Each of the calibration parameters from the sensor model vary with temperature. A thermal calibration over the sensor's operating temperature range can be performed using a thermal chamber to help mitigate errors due to temperature. As shown in Figure 3.2c, often times a rate table calibration and a tumble test calibration are performed in a temperature controlled thermal chamber to conduct both calibration processes simultaneously.

Calibration Rate Tables (Images Courtesy of Ideal Aerosmith)



(a) Single-Axis

(b) Two-Axis

(c) Integrated Thermal Chamber

FIGURE 3.2

3.2.3 Allan Variance

An Allan Variance test is a method used for identifying the noise properties of an inertial sensor. A sensor is mounted statically inside a thermal chamber set to a constant temperature, with data logged at high rate for an extended period (typically several hours). The data is downsampled at various time constants (averaged down to lower sampling rates), and the variances of each of these downsampled measurements can be calculated at each of the different time constants. Despite its name, most of the time the results are plotted as standard deviations (square root of variance) versus the downsampled time constants on a log-log scale, as seen in Figure 3.3.

The plot shown in Figure 3.3 illustrates two key specifications for an inertial sensor: (a) the Angle Random Walk (equivalent to noise density, see Section 3.1), and (b) the in-run bias stability. Noise density is roughly equal to the standard deviation at an averaging time constant of one second, while the minimum standard deviation value on the plot is the in-run bias stability.

3.2.4 Vibration Testing

While most calibration processes and characterization tests used to double-check the performance of those parameters utilize steady-state motion, vibration testing allows for the characterization of sensor response relative to time-varying inputs. Typically, a single-axis shaker table is used, requiring the test to be run multiple times to excite all axes. In addition to identifying any unexpected response a sensor may have to vibrations at particular levels or frequencies, vibration testing can be used to determine a number of sensor specifications including:

- Bandwidth: By performing a test known as a sine-sweep—inducing single-frequency vibrations across a range of frequencies—the sensor bandwidth can be determined by finding the frequency associated with the −3 dB point on the response.
- g-sensitivity & g^2-sensitivity: Gyroscopes' sensitivity to acceleration can be readily determined through vibration testing, both using sine-sweeps and more general random vibration profiles.

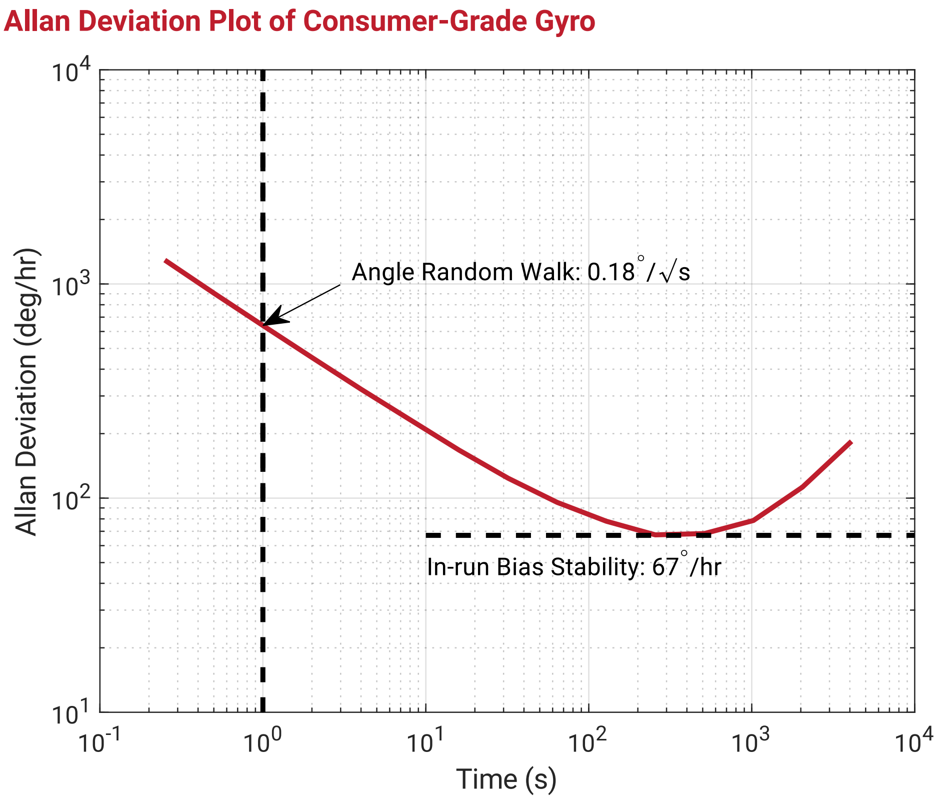


FIGURE 3.3

3.3 INS ERROR BUDGET

An inertial navigation system (INS) uses data collected from an IMU to form a self-contained navigation system by integrating the IMU measurements to track the position, velocity, and orientation of an object relative to an initial position, velocity, and orientation. When combined with data from a GNSS receiver through advanced Kalman filtering, the INS relative navigation solution can be tied to an absolute inertial reference frame, with no drift or long-term error accumulation. But absent such aiding, the relative INS navigation solution is subject to drift over time.

A variety of error sources within the inertial sensors' measurements themselves lead to unbounded error growth in the INS navigation solution, such as bias, noise, scale factor errors, misalignments, temperature dependencies, and gyro g-sensitivity. These sources of error and how they are specified are discussed more in-depth in Section 3.1. This error budget provides the means to calculate how those inertial sensor errors propagate to errors in the INS attitude, position, and velocity solutions over time.

3.3.1 Attitude

The attitude of a system is calculated by integrating angular rate (angular velocity) as measured by the gyros over a defined time period. For the purposes of this analysis we consider single-axis motion, as the non-linear coupling of attitude makes multi-axis analysis impossible in the general case. The equation for a measured angular rate for a single axis can be represented with error sources as follows:

$$\tilde{\omega} = (1 + k)\omega_t + b_g + \eta_g \tag{3.2}$$

where $\tilde{\omega}$ is the measured angular rate, ω_t is the true angular rate, k is the scale factor error, b_g is the time-varying bias, and η_g is the random Gaussian noise (defined by the angle random walk (ARW) specification).

The error in the attitude calculation is found by integrating the difference between the measured angular rate and the true angular rate:

$$\theta_{err} = \int_0^t (\tilde{\omega} - \omega_t)dt \tag{3.3}$$

Substituting the prior definition of $\tilde{\omega}$ and the relationship between noise integrals and ARW yields:

$$\theta_{err} = (\text{ARW})\sqrt{t} + \int_0^t (k\omega_t + b_g)dt \tag{3.4}$$

Static

We will first consider attitude errors under stationary—non-rotating and non-accelerating—conditions. If the bias can be considered constant over the time interval in question (depending on the time constant of the minimum on the Allan variance curve—see Section 3.1), then the angular error can be given by:

$$\theta_{err} = b_g t + (\text{ARW})\sqrt{t} \tag{3.5}$$

For longer durations, where bias is a non-constant factor, the integral of the bias must be handled as a variable rather than a constant. Let the variable bias, which will drift over time, be equal to some initial bias error, b_{g_0}, plus a time-varying bias, $b'_g(t)$, such that:

$$\int_0^t b_g dt = \int (b_{g_0} + b'_g(t))dt = b_{g_0} t + \int_0^t b'_g(t)dt \tag{3.6}$$

Substitute this equation for non-constant bias into the original equation for attitude error and it results in the following:

$$\theta_{err} = b_{g_0} t + (\text{ARW})\sqrt{t} + \int b'_g(t)dt \tag{3.7}$$

It is important to note that b_{g_0} will always be greater than or equal to the in-run bias stability—hence the criticality of that specification. In dynamic GNSS/INS applications where the gyro bias is tracked utilizing imperfect GNSS measurements, $b_{g_0}$is typically five to ten times greater than the in-run bias stability.

Dynamic

Building on this stationary case, now consider a situation in which the system is rotating about a single axis. Under these dynamic conditions, ω_t is non-zero, and we will define the integral of the true angular velocity over time to be $\Delta\theta$. This yields an updated angular error of:

$$\theta_{err} = k\Delta\theta + b_{g_0} t + (\text{ARW})\sqrt{t} + \int_0^t b'_g(t)dt \tag{3.8}$$

Note that the scale factor error k produces an error proportional to the total angle traversed over the time period, so oscillatory motions see a negligible contribution from scale factor errors.

Additional Error Terms

Along with the error sources included in the previous equations, there are a few additional errors that are present under general dynamic conditions, but are either typically negligible or only present under multi-axis motion. These include:

- g-sensitivity and g^2-sensitivity (bias shifts due to accelerations): typically small relative to other error sources except in the case of extreme dynamics.
- Earth's angular rotation (~15°/hour): plays a negligible role in calculating attitude error while heading error remains reasonably bounded.
- Cross-axis sensitivity: primarily of consideration for multi-axis motion, the contribution is similar to scale factor errors, producing angular errors proportional to the total change in angle.
- Initial attitude error (prior to start of integration): for multi-axis motion, initial alignment errors are similar to cross-axis sensitivities.
- Nonlinear behavior: attitude is inherently nonlinear and the coupling present in multi-axis motion can quickly dominate the error budget as attitude errors exceed 5°.

3.3.2 Velocity & Position

In order to determine the velocity and position error that arises from an INS, another series of integrations must be performed. There are certain challenges that arise when multiple integrations are performed including attitude dependence and accounting for the presence of gravity in the calculations.

For the case of velocity and position, we will consider a non-rotating case with single-axis accelerations. With an understanding of fundamental kinematics, given some acceleration vector, velocity can be found by integrating the acceleration solution over a finite period of time. Likewise, position is found by integrating the velocity solution over a finite period of time.

When dealing with linear acceleration in this manner, it is important to differentiate between acceleration in a horizontal direction and a vertical direction. Similar to the measurement equation used to find the angular rate of the gyro, the linear acceleration of the system can be modeled by this equation for both directions:

$$\tilde{a} = (1+k)a_t + b_a + \eta_a + \begin{cases} g\sin\theta_{err} & \text{horizontal} \\ g(1-\cos\theta_{err}) & \text{vertical} \end{cases} \tag{3.9}$$

where $\tilde{a}$ is the measured linear acceleration, k is the scale factor error, a_t is the true linear acceleration, b_a is the bias, η_a is the Gaussian random noise defined by Velocity Random Walk (VRW), and g is gravity. The gravity terms exist because the accelerometers measure both linear acceleration and gravity, and the pitch and roll estimates are required to subtract out the gravity signal—leading to an attitude dependence for the position and velocity integrals.

Assuming the attitude error is small over the integration time, a small angle approximation reduces those terms as follows:

$$\begin{aligned} g\sin\theta_{err} &\approx g\theta_{err} \quad \text{(horizontal)} \\ g(1-\cos\theta_{err}) &\approx 0 \quad \text{(vertical)} \end{aligned} \tag{3.10}$$

Following that simplification, the equation for angular error under static conditions found in the previous section can be substituted. For the purposes of this analysis, we assume the accelerometer bias is constant, though it is actually time varying like the gyro.

Velocity

The velocity error is found by integrating the acceleration and adding that to an initial velocity error at the start of the integration:

$$V_{err} = V_{err_0} + \int_0^t (\tilde{a} - a_t)dt \tag{3.11}$$

By employing substitutions similar to those used in Section 3.3.1, Equation 3.11 can be rewritten:

$$V_{err} = V_{err_0} + \int_0^t (ka_t + b_a + \eta_a + g\theta_{err})dt \tag{3.12}$$

$$V_{err} = V_{err_0} + \int_0^t \left[ka_t + b_a + \eta_a + g(b_g t + (\text{ARW})\sqrt{t})\right] dt \tag{3.13}$$

$$V_{err} = V_{err_0} + k\Delta V + b_a t + (\text{VRW})\sqrt{t} + g\left[\frac{1}{2}b_g t^2 + \frac{2}{3}(\text{ARW})t^{\frac{3}{2}}\right] \tag{3.14}$$

The final term (multiplied by g) can be excluded when considering errors in the vertical channel. As with the attitude integration, scale factor errors only act on the change in velocity over the time period of interest and are negligible for oscillatory accelerations.

Position

Finding position error requires integrating Equation 3.14 and adding it to an initial position error:

$$P_{err} = P_{err_0} + \int_0^t V_{err}dt \tag{3.15}$$

$$P_{err} = P_{err_0} + \int_0^t \left[V_{err_0} + k\Delta V + b_a t + (\text{VRW})\sqrt{t} + g\left(\frac{1}{2}b_g t^2 + \frac{2}{3}(\text{ARW})t^{\frac{3}{2}}\right)\right] dt \tag{3.16}$$

$$P_{err} = P_{err_0} + k\Delta P + V_{err_0}t + \frac{1}{2}b_a t^2 + \frac{2}{3}(\text{VRW})t^{\frac{3}{2}} + g\left[\frac{1}{6}b_g t^3 + \frac{4}{15}(\text{ARW})t^{\frac{5}{2}}\right] \tag{3.17}$$

Finally, a solution for position error has been found that factors all significant sources of error found in inertial navigation. As a reminder, this solution takes into account a linear acceleration in the purely horizontal axis. Removing the last component in the equation would result in a solution based on vertical linear acceleration.

This position integration result shows why most INS systems are assessed initially on their gyro performance, particularly the in-run bias stability: the positioning errors proportional to the gyro bias grow as a function of time cubed!

3.3.3 Position Errors Over Time

Table 3.1 presents position errors over multiple time periods in a purely static condition, as calculated from the last equations derived above, to demonstrate how important and effective error budgeting is for an INS system to determine the best possible navigation solution. We have used representative values for each grade of inertial sensor as seen in Table 3.2 (these numbers do not reflect any particular sensor on the market), and assumed initial position and velocity errors are zero. Care must be taken during these calculations to ensure all parameters are in SI units.

Note that these calculations assume the gyro biases are known to the accuracy of the in-run bias stability (which is unrealistic in a dynamic scenario), and the errors contributed by the random walk terms represent 1-sigma standard deviations, and could therefore be ~3x higher. The other assumptions made (static, no initial velocity errors) also act to make these numbers lower than would be encountered in a real-world application. And once you reach the high-end performance of navigation-grade inertial sensors, additional error sources not discussed here become relevant.

Error Over Time by Sensor Grade

GRADE/TIME	1 s	10 s	60 s	10 min	1 hr
Consumer	6 cm	6.5 m	400 m	200 km	39.000 km
Industrial	6 mm	0.7 m	40 m	20 km	3.900 km
Tactical	1 mm	8 cm	5 m	2 km	400 km
Navigation	<1 mm	1 mm	50 cm	100 m	10 km

TABLE 3.1

Error Terms by Sensor Grade

GRADE	ACCELEROMETER BIAS (mg)	VELOCITY RANDOM WALK (m/s/$\sqrt{hr}$)	GYRO BIAS (deg/hr)	ANGLE RANDOM WALK (deg/$\sqrt{hr}$)
Consumer	10	1	100	2
Industrial	1	0.1	10	0.2
Tactical	0.1	0.03	1	0.05
Navigation	0.01	0.01	0.01	0.01

TABLE 3.2

3.4 GNSS ERROR BUDGET

The GNSS error budget describes the many factors involved in the GNSS system that determines how accurately a receiver may determine its position, velocity, and time (PVT). Knowledge of these error sources is useful in determining issues that may occur while using a GNSS system. Certain sources of error apply to the individual pseudorange measurements from each satellite, while others can be considered at the level of the PVT solution.

3.4.1 Pseudorange Errors

Table 3.3 provides a list of the various errors impacting individual pseudorange measurements and their impact on those measurements. The total error listed is the result of accumulating those individual errors in a sum-squared sense and represents the typical errors of a standalone GNSS receiver. As discussed in Section 1.5, most receivers track SBAS satellites which provide substantial corrections to each of these terms, bringing the error down to the more commonly seen 2 m specification for consumer GNSS receivers. And more advanced differential GNSS corrections can eliminate some of these errors entirely.

Orbit Errors

Any difference between a satellite's actual position and expected position will ripple through the entire PVT determination process. While the ground segment monitors and updates these positions, this still accounts for a large portion of the pseudorange error budget, typically contributing about 2.5 m of error.

GNSS Pseudorange Error Budget

ERROR SOURCE	ERROR CONTRIBUTION (m RMS)
Orbital	2.5
Satellite Clock	2
Receiver Noise	0.3
Ionospheric	5
Tropospheric	0.5
Multipath	1
Total	*11.3*

TABLE 3.3

Satellite Clock

GNSS trilateration uses the speed of light to measure distances, which means that clock errors are equivalent to range errors. A nanosecond-scale error in the satellite's atomic clock time from relative to GNSS system time results in 0.3 m of pseudorange error and such clock errors typically contribute about 2 m of error.

Receiver Noise

Noise received on the GNSS antenna or from within the receiver itself can contribute a small but not insignificant error to the solution, accounting for around 0.3-1 m of error. Receiver design and antenna quality can significantly impact this error term.

Ionospheric Delay

The ionosphere is the first layer of the atmosphere that a GNSS satellite signal must enter after the vacuum of space, as seen in Figure 3.4. Ranging from 50 km to 1000 km above Earth's surface, the ionosphere contains ionized gases that act as a dispersive medium and slow the propagation speed of radio waves. Since the refractive index is less than one, the signal speed through the medium slows slightly and the phase increases. The ionosphere is also constantly in flux, varying depending on solar activity, time of year, and time of day which therefore cannot be precisely modeled. Due to the many factors involved, ionospheric delay errors can be around 5 m, the largest source of GNSS error.

As mentioned previously, SBAS and other differential GNSS techniques can be used to largely eliminate ionospheric errors. Furthermore, the dispersive nature of the ionosphere causes a higher frequency signal to be slowed less than lower a frequency signal. Due to this, the L1 band experiences less effect than the L2 band, which experiences less than the L5 band. A multi-frequency receiver may use these differences to estimate ionospheric delay directly without external corrections.

Tropospheric Delay

When a signal enters the troposphere, additional effects come into play affecting signal transmission quality. The troposphere extends from the earth's surface to a height of about 10-16 km, as seen in Figure 3.4. This layer contains all weather on earth, and most of the atmosphere's water vapor. While accurate models exist to predict this delay, changes in pressure, temperature, and humidity still affect signal transmission, resulting in about 0.5 m error.

Multipath

Multipath error occurs when GNSS satellite signals bounce off solid objects such as buildings and terrain resulting in the same initial signal taking multiple paths to get to the receiver. This can result in error effects of 1 m or more for the receiver. Figure 3.5 shows some instances of multipath error.

3.4.2 Dilution of Precision

Beyond the errors in the individual pseudoranges, the geometry of the visible (and tracked) satellites contributes to the total PVT error budget. This is known as Dilution of Precision (DOP) and can be calculated in each dimension, including time (since clock errors result in distance errors) and can be combined in a total parameter known as Geometric Dilution of Precision (GDOP). Definitions for the different types of DOP are in Table 3.4.

The GDOP represents a sensitivity of the final PVT solution to errors in the pseudoranges. As such, GDOP values are considered ideal when they are low, with values over 5 considered poor. To get a more intuitive understanding of GDOP, the Figure 3.6 contains a few scenarios with varying GDOP. In Figure 3.6a, many well-distributed satellites

Atmospheric Effects on a GNSS Signal

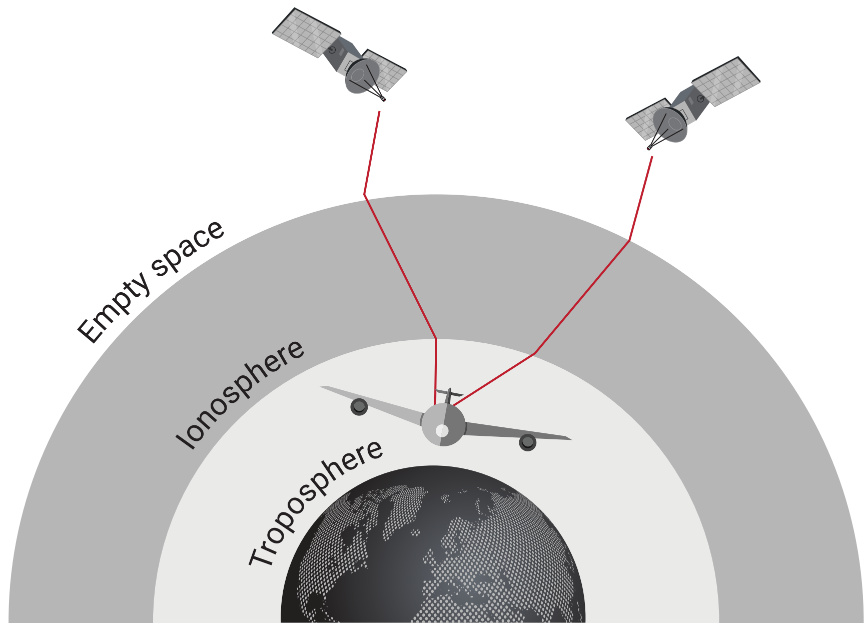


FIGURE 3.4

Multipath Errors

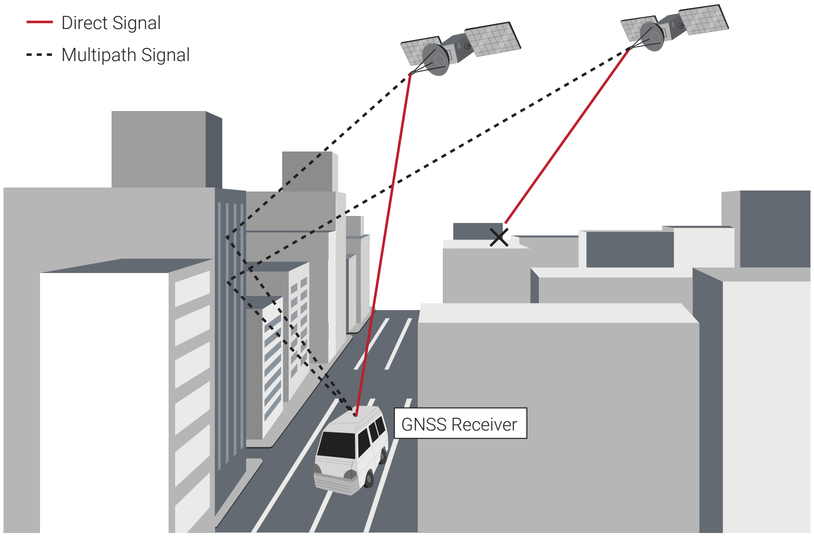


FIGURE 3.5

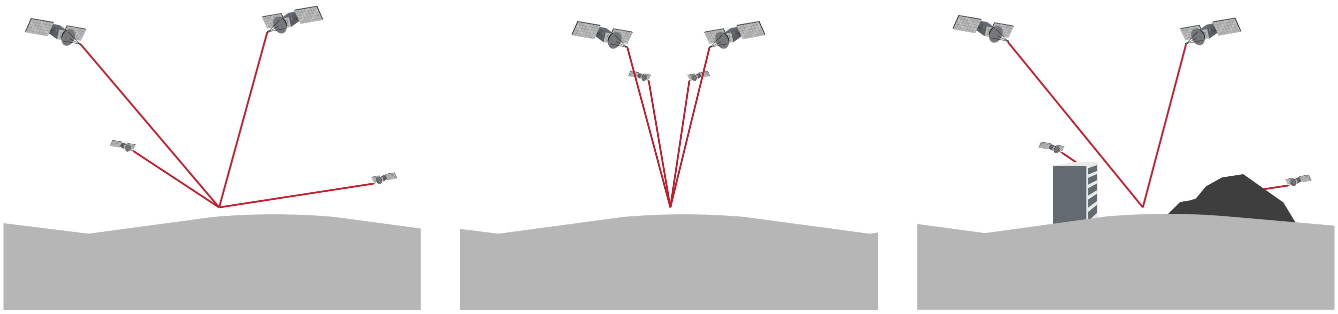
Dilution of Precision Definitions

DILUTION OF PRECISION	ACRONYM	DEFINITION
Horizontal	HDOP	$\sqrt{\sigma_N^2 + \sigma_E^2}$
Vertical	VDOP	σ_D
Positional	PDOP	$\sqrt{\text{HDOP}^2 + \text{VDOP}^2}$
Time	TDOP	σ_t
Geometric	GDOP	$\sqrt{\text{PDOP}^2 + \text{TDOP}^2}$

TABLE 3.4

are visible to the receiver, yielding low GDOP values. Meanwhile, Figure 3.6b shows the visible satellites grouped together overhead, which increases VDOP significantly and, subsequently, high GDOP values. Figure 3.6c contains obstructions that both reduce the number of visible satellites and leave the remaining satellites grouped together, both of which increase GDOP. Increasing the number of available satellites is especially important in this type of scenario, which is why a multi-constellation GNSS receiver is much better suited to urban canyon environments than a GPS-only receiver.

Various GDOP Conditions



(a) Good Geometry—Good GDOP

(b) Bad Geometry—Poor GDOP

(c) Bad Visibility—Poor GDOP

FIGURE 3.6

3.5 MAGNETIC ERROR SOURCES

Using a magnetometer is a common way to obtain a system's heading, particularly in applications that do not have access to GNSS. While seemingly straightforward, using a magnetometer to accurately estimate the heading can actually prove to be quite challenging. The Earth's magnetic field is quite weak and there often exists a number of different error sources that can impact the heading accuracy.

To mathematically model the various error sources that can be present in a set of measurements, a sensor model is often established. Magnetometers are commonly modeled using Equation 3.18, where the measurement vector on the left of the equation, $\tilde{m}$, contains the magnetic readings as measured by the magnetometer.

$$\tilde{\boldsymbol{m}} = S_I^{-1}(\boldsymbol{m}_E + \boldsymbol{m}_{e(t)}) + \boldsymbol{b}_{HI} + \boldsymbol{m}_{i(t)} \tag{3.18}$$

In a perfectly clean magnetic environment, this measurement vector would simply be Earth's magnetic field ($\boldsymbol{m}_E$). However, in a majority of real-world applications, the local magnetic field measured by the magnetometer often consists of a combination of Earth's magnetic field and magnetic fields created by nearby objects, known as magnetic disturbances. These magnetic disturbances include objects that are external to the system in the surrounding environment ($\boldsymbol{m}_{e(t)}$) as well as objects that are fixed with respect to the sensor ($\boldsymbol{m}_{i(t)}$). Hard iron distortions ($\boldsymbol{b}_{HI}$) and soft iron distortions (S_I) can also be present in the magnetic measurements, which bias and distort Earth's magnetic field. To account for error sources present in the magnetometer measurements, a compensation model can be constructed by rearranging Equation 3.18.

$$\boldsymbol{m}_c = S_I(\tilde{\boldsymbol{m}} - \boldsymbol{b}_{HI}) \tag{3.19}$$

This model is often utilized in what is known as a hard and soft iron (HSI) calibration to remove the impact of hard and soft iron distortions present in the magnetic readings. Hard and soft iron distortions as well as the calibration used to account for them are discussed more in depth in Section 3.6.

3.5.1 Earth's Magnetic Field

Earth's magnetic field is a self-sustaining magnetic field that resembles a magnetic dipole with one end near Earth's geographic North Pole and the other near Earth's geographic South Pole. This magnetic field is characterized by both a strength and direction. While the direction of Earth's magnetic field contains both a horizontal component and a vertical component, magnetic-based heading is calculated from only the horizontal component of the magnetic field as the horizontal component points to the magnetic North Pole of the earth.

The strength of this magnetic field varies across the earth with strengths as low as 0.3 Gauss in South America to over 0.6 Gauss in northern Canada. Although the Earth's magnetic field is relatively stable over time, electric currents in the ionosphere can cause daily alterations which can deflect surface magnetic fields by as much as 1°. Normally daily variations in field strength are on the order of 0.25 mG, which would equate to about a 0.03° variation in heading. This small of a change in heading is on the same order of magnitude as the resolution of most MEMS-based magnetometers, so in most cases the Earth's magnetic field can be considered constant with respect to time.

3.5.2 Sources of Magnetic Disturbances

In an ideal scenario, a magnetometer would purely measure Earth's magnetic field and very accurately estimate the heading. Unfortunately, most applications consist of nearby objects that bias and distort Earth's background magnetic field, leading to errors in the calculated heading. These objects are commonly referred to as magnetic disturbances and can be characterized in a few different ways.

Hard and Soft Iron Distortions

Magnetic disturbances can be classified as either a hard iron distortion or a soft iron distortion. Hard iron distortions are created by objects that produce a magnetic field (i.e. speaker, magnet, etc.) and cause a permanent bias in the local magnetic field. On the other hand, soft iron distortions are commonly caused by metals such as nickel and iron (i.e. batteries) and distort and stretch the local magnetic field. Refer to Section 3.6 for more information on hard and soft iron distortions.

Internal vs. External Disturbances

Regardless of the type of distortion, magnetic disturbances are considered to be either internal or external disturbances. Internal magnetic disturbances are caused by objects that are fixed with respect to the magnetometer. This would include electronics, other sensors, and structures that are rigidly mounted with respect to the magnetometer. External magnetic disturbances include any objects producing a magnetic field in the surrounding environment that are not fixed with respect to the magnetometer. Things like keys, phones, computers, electronics, office furniture, and vehicles in the surrounding environment that are not rigidly attached to the sensor are common sources of external magnetic disturbances.

Time-Varying Disturbances

Magnetic disturbances can be further characterized based on whether or not they vary over time. Time-invariant disturbances are constant over time and include things like ferrous metals or electronics that are constantly running. On the other hand, time-varying disturbances change their magnetic field over time. Electric motors running at different speeds, such as the rotors on a quad-copter, or powering on and off of electronics are common examples of time varying disturbances.

Accounting for Magnetic Disturbances

Internal magnetic disturbances that are non-time varying can be accounted for using a calibration known as a hard & soft iron (HSI) calibration. For more information on an HSI calibration, see Section 3.6. Advanced filtering techniques can be used to mitigate the impact of external disturbances in the environment, but their effectiveness varies by manufacturer and application.

3.5.3 Tilt-Compensation

When using a magnetometer, the sensor is often oriented in such a way that the pitch and roll angles are non-zero (in contrast to the use of a handheld magnetic compass). Since the Earth's magnetic field contains both a horizontal and vertical component, a magnetometer will measure a portion of Earth's vertical magnetic field component in each of its axes when tilted. To remove the impact of the vertical component of Earth's magnetic field, a technique known as tilt-compensation must be utilized in which the estimated pitch and roll angles are used. The use of tilt-compensation is critical, as many places around the world (including most of the U.S. and Europe) contain a vertical component of the magnetic field that is stronger than the horizontal component.

The following equations apply tilt-compensation to the magnetometer data and calculate the corrected magnetic heading angle, where $\tilde{m}_x$, $\tilde{m}_y$, and $\tilde{m}_z$ are the measured magnetometer data in the body frame.

$$X = \tilde{m}_x \cos\theta + \tilde{m}_y \sin\theta \sin\phi + \tilde{m}_z \sin\theta \cos\phi \tag{3.20}$$

$$Y = \tilde{m}_y \cos\phi - \tilde{m}_z \sin\phi \tag{3.21}$$

$$\psi_c = \tan^{-1}\frac{-Y}{X} \tag{3.22}$$

Due to the dependence on pitch and roll, tilt-compensation couples errors in pitch and roll into errors in heading. If a magnetometer is to be used in an application experiencing non-zero pitch and roll angles, an accelerometer or tilt-sensor will be needed at the very least to obtain estimates of the pitch and roll angles. In such applications, a full Attitude and Heading Reference System (AHRS) or Inertial Navigation System (INS) is recommended for use, which calculates the pitch and roll angles and automatically applies tilt-compensation to the magnetic measurements.

3.5.4 Operation Near Poles

Near the equator, pitch and roll errors do not have as much of an impact on the magnetic heading since the inclination angle is small and the majority of the magnetic field is in the horizontal plane. However, for operation near the Earth's magnetic poles, the inclination angle is near 90° and the vertical magnetic field component is much stronger than the horizontal. In this case, the accuracy of the pitch and roll angles becomes much more important in correctly applying tilt-compensation to obtain an accurate magnetic heading. A pitch or roll error of 0.5° equates to about a 1° error in the magnetic heading in the southern part of the U.S. Comparatively, at the Earth's magnetic poles, this same 0.5° pitch or roll error causes a 2.5° error in the heading.

3.6 MAGNETOMETER HARD & SOFT IRON CALIBRATION

A magnetometer is a sensor used to measure the strength and direction of the local magnetic field surrounding a system. This magnetic field measurement can then be compared to models of Earth's magnetic field to determine the heading of a system with respect to magnetic North. However, in most real-world applications, the magnetic field measured will be a combination of both Earth's magnetic field and magnetic fields created by nearby objects, commonly referred to as magnetic disturbances. In order to obtain an accurate heading estimate, the impact of nearby magnetic disturbances must be mitigated. Internal magnetic disturbances that are non-time varying can be accounted for using a hard & soft iron (HSI) calibration. For more information on magnetic disturbances, refer to Section 3.5.

3.6.1 Magnetometer Sensor Model

In order to mathematically correct sensor measurements for various error sources, a sensor model must be established. As mentioned in Section 3.5, a magnetometer is commonly modeled using the following equation, in which the measurement vector, $\tilde{\boldsymbol{m}}$, consists of a combination of Earth's magnetic field and magnetic fields created by nearby magnetic disturbances.

$$\tilde{\boldsymbol{m}} = S_I^{-1}(\boldsymbol{m}_E + \boldsymbol{m}_{e(t)}) + \boldsymbol{b}_{HI} + \boldsymbol{m}_{i(t)} \tag{3.23}$$

From this measurement model, a compensation model can be constructed to account for the hard iron, b_{HI}, and soft iron, S_I, distortions present in the local magnetic field resulting from internal, non-time varying magnetic disturbances.

$$\boldsymbol{m}_c = S_I(\tilde{\boldsymbol{m}} - \boldsymbol{b}_{HI}) \tag{3.24}$$

$$\begin{bmatrix} m_{c_x} \\ m_{c_y} \\ m_{c_z} \end{bmatrix} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \tilde{m}_x - b_{H_0} \\ \tilde{m}_y - b_{H_1} \\ \tilde{m}_z - b_{H_2} \end{bmatrix}$$

3.6.2 Hard and Soft Iron Distortions

Magnetic measurements will be subjected to distortion. These distortions are considered to fall in one of two categories: hard iron or soft iron. Hard iron distortions are created by objects that produce a magnetic field. A speaker or piece of magnetized iron for example will cause a hard iron distortion. If the piece of magnetic material is physically attached to the same reference frame as the sensor, then this type of hard iron distortion will cause a permanent bias in the sensor output. Soft iron distortions are considered deflections or alterations in the existing magnetic field. These distortions will stretch or distort the magnetic field depending upon which direction the field acts relative to the sensor. This type of distortion is commonly caused by metals such as nickel and iron. In most cases hard iron distortions will have a much larger contribution to the total uncorrected error than soft iron.

3.6.3 Visualizing Hard and Soft Iron Distortions

A common way of visualizing and correcting hard and soft iron distortions is to plot the output of the magnetometer on a 2D graph. The following examples show measurements taken by the magnetometer as the device is slowly rotated about the z-axis.

No Distortions

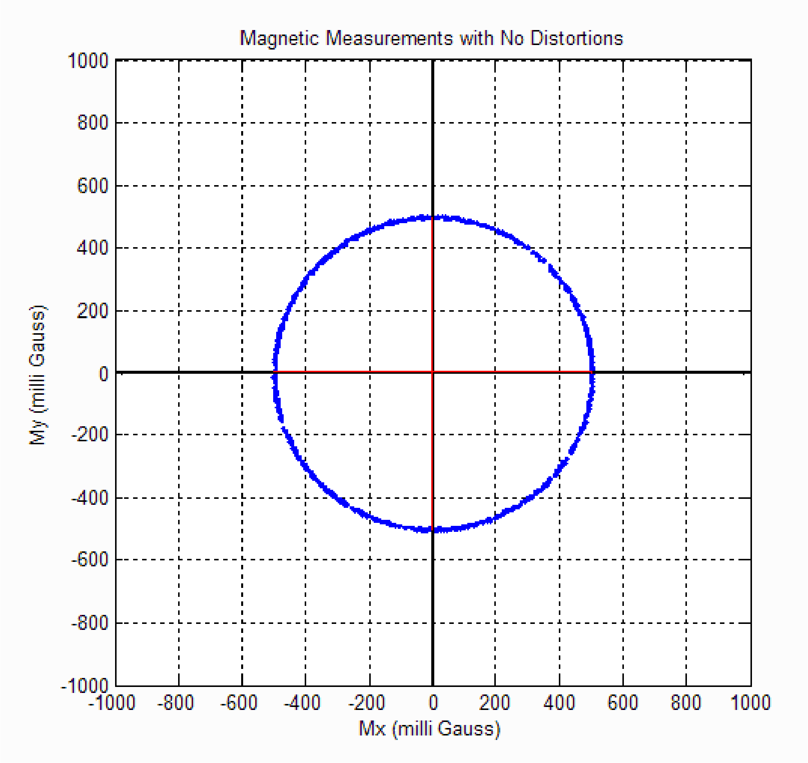


FIGURE 3.7

Case 1- No Distortions

In the event that there are no hard or soft iron distortions present, the measurements should form a circle centered at X=0, Y=0, illustrated in Figure 3.7. The radius of the circle equals the magnitude of the magnetic field.

Case 2 - Hard Iron Distortions

Hard iron distortions will cause a permanent bias to be present in the magnetic measurements, which leads to a shift in the center of the circle. As shown in Figure 3.8, the center of the circle is now at X=200, Y=100 with hard iron distortions present. From this it can be concluded that there is 200 mG hard iron bias in the X-axis and 100 mG hard iron bias in the Y-axis.

Case 3 - Hard and Soft Iron Distortions

Hard iron distortions will only shift the center of the circle away from the origin, they will not distort the shape of the circle in any way. Soft iron distortions, on the other hand, distort and warp the existing magnetic fields. When plotting the magnetic output, soft iron distortions are easy to recognize as they will distort the circular output into an elliptical shape.

Figure 3.9 illustrates the impact that both hard and soft iron distortions have on the magnetometer output—the circle has been distorted into an ellipse that is not centered at the origin. The center of the ellipse is still located at X=200 and Y=100 since the hard iron distortions are the same as before. Every ellipse has a major and minor axis which corresponds to the long and short dimensions, respectively. As shown in Figure 3.9, the ellipse has its major axis aligned 30° up from the body frame X direction, caused by the soft iron distortions.

Hard and Soft Iron Distortions in 3D Case

While plotting the magnetic measurements on a 2D graph provides a straightforward visualization of the impact of hard and soft iron distortions, in reality the magnetic measurements consist of a full 3D vector. When no distortions are present, the full magnetic measurement vector forms a sphere centered at the origin. Similar to the 2D case, hard iron distortions shift the center of the sphere away from the origin while soft iron distortions distort the sphere into an ellipsoid.

3.6.4 Eliminating Hard and Soft Iron Distortions

It is possible to eliminate the effects of both hard and soft iron distortions on the magnetometer outputs through the use of a hard and soft iron (HSI) calibration. An HSI calibration is a method used to map the biased and distorted ellipsoid back into a sphere centered at the origin. To do so, this calibration is typically implemented with one of two

Hard Iorn Distortions

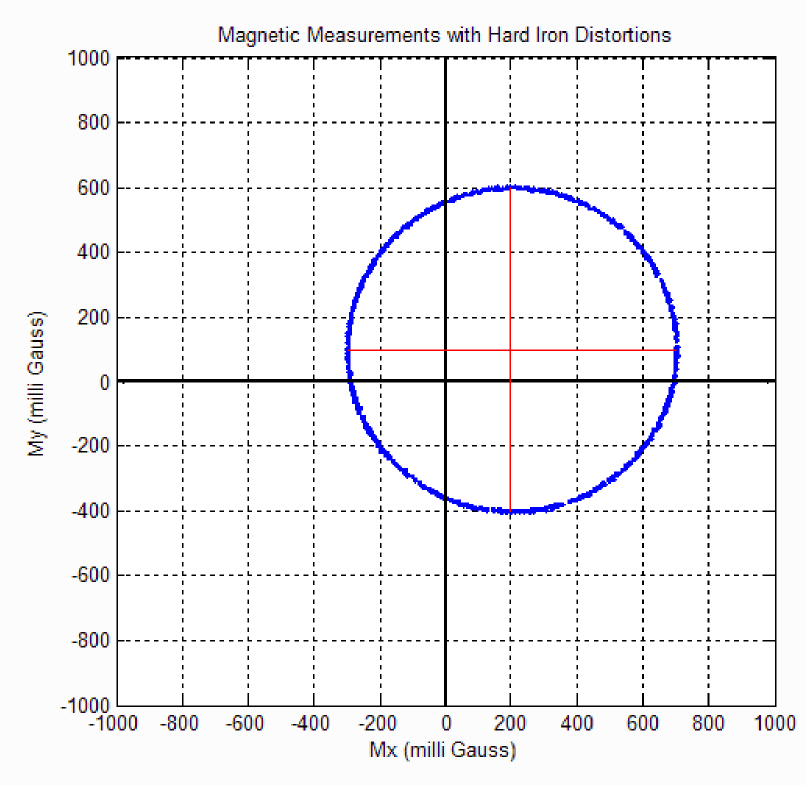


FIGURE 3.8

Hard and Soft Iron Distortions

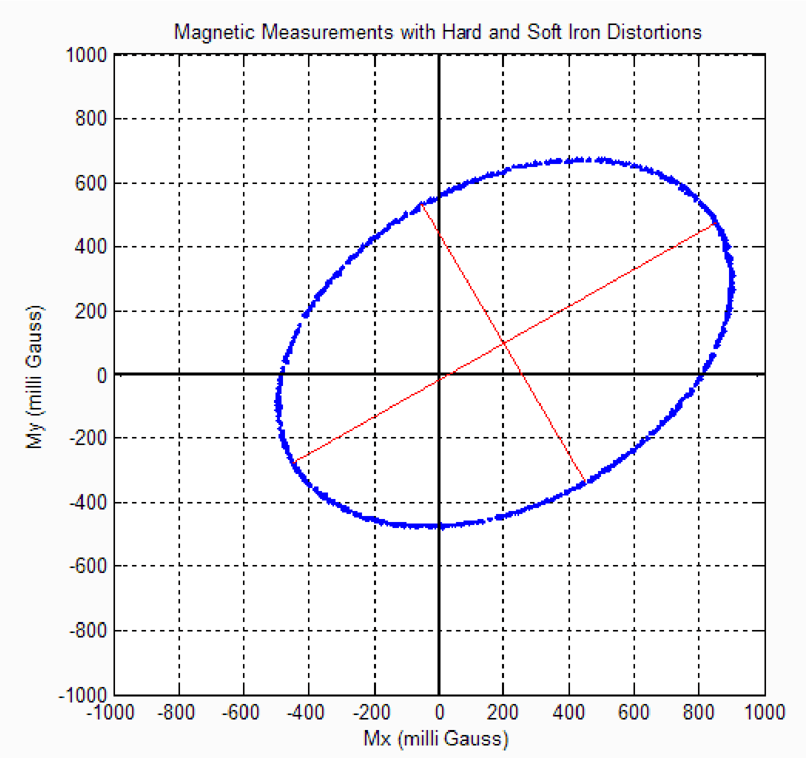


FIGURE 3.9

different types of motion, a 2D calibration or a 3D calibration, depending on the range of movement feasible during calibration.

A 2D calibration is generally performed by rotating the magnetometer about the gravity vector in a few 360° circles and is often sufficient if the sensor will stay between 5° to 10° of level in both pitch and roll. Outside of that range, a full 3D calibration is strongly recommended which entails rotating the sensor as much as possible about all axes to cover a sphere of measurement. In either case, hard iron, $\boldsymbol{b}_{HI}$, and soft iron, S_I, compensation parameters are estimated from the magnetometer measurements using Equation 3.24. For the case shown in Figure 3.9, the corresponding hard and soft iron calibration parameters are shown in Equation 3.25.

$$m_c = \begin{bmatrix} .75 & -.1443 & 0 \\ -.1443 & 0.9167 & 0 \\ 0 & 0 & 1.0 \end{bmatrix} \begin{bmatrix} \tilde{m}_x - 0.2 \\ \tilde{m}_y - 0.1 \\ \tilde{m}_z - 0.0 \end{bmatrix} \tag{3.25}$$

The hard iron parameters are set to offset the center of the ellipse by the 200 mG in the x-axis and 100 mG in the y-axis shown in Figure 3.9 back to the origin. The matrix of soft iron parameters shape the ellipse back into a sphere. Keep in mind, as mentioned previously, that the hard and soft iron distortions influence all axes of the sensor and the above example could be repeated for both rotations about the x- and y- axes.

4 HARDWARE

Electronic components can be damaged or important data corrupted by improper setup, unregulated interference, and assumptions of hardware specifications. These issues can be mitigated by following best practices in ensuring stable voltages, stable input/output signals, and comparing application requirements with physical hardware limitations. Understanding the various components of the hardware interface is crucial to successful implementation of an inertial navigation system.

Multiple methods for transmitting and receiving serial data may be used, each with different advantages and hardware requirements. Synchronous communication has more significant hardware requirements for clocking, but enables faster communication between more devices. Asynchronous communication has more economical hardware requirements, but is less efficient and restricted to communication between two devices. The requirements and capabilities of each should be carefully examined before selecting the appropriate method for a particular application.

4.1 ASYNCHRONOUS SERIAL COMMUNICATION

Asynchronous serial communication is a communication interface in which the signals used are not synchronized to each other using a common clock signal. Instead, start and stop bits are used to indicate the beginning and end of a data message. This type of communication utilizes a point-to-point type of interface, meaning that only two devices can be linked together to communicate. These two devices must also agree upon the rate at which bits will be transmitted and received, known as the baud rate, since there is no clock signal to indicate such transitions. Furthermore, asynchronous serial communication can be implemented in either a full-duplex (independent transmit and receive lines) or half-duplex (shared transmit/receive line) configuration, making it a versatile communication protocol that can be used in many different applications.

The asynchronous serial communication interface utilizes a receiving signal (RX) and a transmitting signal (TX). When connecting two devices to communicate in full-duplex mode, the RX pin of one device must connect to the TX pin of the other device, as shown in Figure 4.1. Asynchronous serial communication is most commonly implemented using a universal asynchronous receiver-transmitter (UART). UARTs are typically employed in microcontrollers, but can also exist as individual integrated circuits (ICs).

Asynchronous serial communication using a UART interface is very commonly used due to the minimal amount of wires needed for communication and very simplistic protocol required for sending messages. It allows the ability to modify the data packet based on the needs of the application and does not require a separate clock signal to transmit data. However, a UART interface can only be used to communicate between two devices and requires that the baud rates and the bit packets on both devices be the same, or data will be misinterpreted.

4.1.1 Configuration

Data transmitted using asynchronous serial communication or via a UART is sent as packets of bits. These packets contain a start bit, a configurable number of data bits (5-9), an optional parity bit, and a configurable number of stop bits (1-2). The most common structure of a UART bit packet is known as 8-N-1, corresponding to eight data bits, no parity bit, and one stop bit. These bits combined with one start bit create a bit packet that is a total of ten bits long. Both devices communicating via the serial bus must be configured for the same bit packets and transmitting those bits at the same speed, known as the baud rate. The serial port configuration is often prepended with the baud rate: 115200-8-N-1.

Start & Stop Bits

The start and stop bits are known as synchronization bits as they indicate to the receiving device when the packet begins and ends. Asynchronous serial communication data lines are held in a high idle state when not transmitting data. The start bit transitions the data line from a high (1) to a low (0) state. Once the receiving device identifies this

Asynchronous Serial Communication Schematic

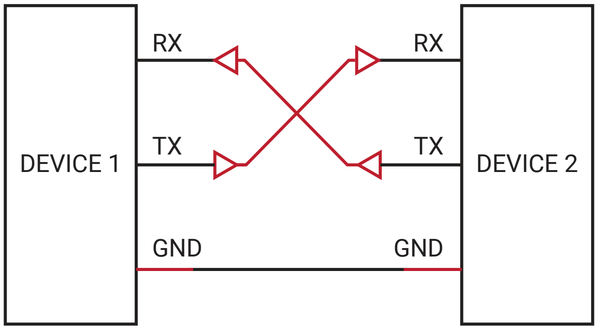


FIGURE 4.1

transition as the start bit, the 5-9 data bits are read at the specified baud rate. The stop bit indicates the end of the data packet by pulling the data line back to the high (1) idle state.

Parity Bit

The parity bit is an optional bit that provides a very low-level form of error detection as data bits can be altered from things like electromagnetic interference or lengthy data lines. If used, this bit can either be specified as an odd parity or an even parity. An odd parity determines whether the data bits in the bit packet contain an odd number of 1-bits. If there are an odd number of 1-bits, the parity bit is set to 0, if not the parity bit is set to 1. This ensures that the data bits combined with the parity bit contain an odd number of 1-bits. Similarly, an even parity will set the parity bit to 0 if the number of 1-bits in the data message is even, otherwise, the parity bit will be set to 1. If one of the data bits have flipped value during transmission, the parity bit will indicate that the number of 1-bits is incorrect. However, the parity bit is not often used as it is unlikely to detect that the message is incorrect if more than one-bit has flipped.

Baud Rate

An important parameter when using asynchronous serial communication or when interfacing with a UART is how fast data can be transmitted across a serial line. The number of bits per second sent over a UART is defined as the baud rate. Possible baud rates span a wide range and can be almost any value, but since both devices must support the same baud rate, certain values have become standard baud rates. As the baud rate increases, the amount of time required to send or receive data decreases. Table 4.1 provides a list of standard baud rates and the amount of time required to transmit 100 bytes of data using the standard 8-N-1 configuration (requiring 10-bits per byte of data).

Baud Rates

BAUD RATE	TIME FOR 100 BYTES
9600	104.2 ms
19200	52.1 ms
38400	26.0 ms
57600	17.4 ms
115200	8.7 ms
230400	4.3 ms
460800	2.2 ms
921600	1.1 ms

TABLE 4.1

4.1.2 Hardware Implementations

Asynchronous serial communication can be implemented in many different ways, depending upon the application it is being used in. Some of the more common standards include transistor-transistor logic (TTL), RS-232, and RS-422 or RS-485. Each of these implementations determine whether a signal is low (0-bit) or high (1-bit) based on different thresholds for the amount of voltage transmitted across a line. Due to the different voltage levels used, connecting two devices with different hardware standards can cause damage to one or both devices.

Transistor-Transistor Logic (TTL) Level

Transistor-transistor logic (TTL) is a physical implementation of asynchronous serial communication, ideal for board-level communication as the wiring must be less than a few feet long. This type of hardware implementation uses a single-ended type of signal, meaning that the voltages sent across the communication lines are referenced to the ground signal. TTL is most often utilized when communicating through a UART interface in microcontrollers or integrated circuits.

As shown in Figure 4.2, the low state (0-bit) is considered any voltage between 0 V and 0.8 V, while the high state (1-bit) is regarded as any voltage between 2 V and 5 V. The difference in acceptable voltage levels for input and output allows for loss and noise on the signal line. TTL idles in the high state.

Voltage Levels for TTL Serial

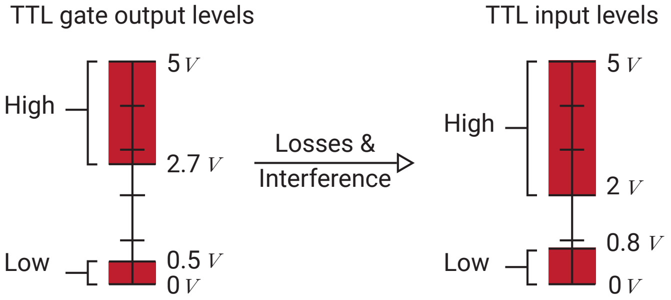


FIGURE 4.2

RS-232

The most common type of hardware implementation for asynchronous serial communication is known as RS-232. This type of interface often uses cables to connect devices and commonly supports cable lengths of up to 10 m, allowing its use in many applications. Similar to TTL, RS-232 also utilizes single-ended signals and idles in the high-state, but uses voltages ranging from −15 V to 15 V. The voltage threshold for the low-state (0-bit) ranges from 3 V to 15 V, while the threshold for the high-state (1-bit) ranges from −15 V to −3 V.

RS-422/RS-485

The RS-422 hardware implementation of asynchronous serial communication is a standard frequently used in industry. Differing from TTL and RS-232, RS-422 uses differential signals which measures the voltage difference between two wires rather than comparing a signal to a common ground. Using differential signals provides greater robustness to interference and cable losses, allowing RS-422 to operate over much greater distances than RS-232 or TTL.

Typically, the differential signals are referred to as RX^+ and RX^- lines for receive, and TX^+ and TX^- lines for transmit. However, these lines are sometimes referred to instead as RX^A, RX^B and TX^A, TX^B or as the inverted signal and the non-inverted signal. Unfortunately, the RS-422 standard does not specify the nomenclature of the particular signals, only the signal levels. This means that each manufacturer of a device can choose their own signal names and pinouts. Due to this, care must be taken when connecting two devices to ensure that the signals are matched correctly.

RS-422/RS-485 systems can be configured using either a full-duplex or half-duplex arrangement. A full-duplex configuration contains both RX and TX communication signals allowing the two devices to transmit data to each other independently and simultaneously. The half-duplex configuration utilizes a single shared communication signal, meaning that only one device can transmit data at a time, and half the wires are needed. Half-duplex configurations are advantageous when there are constraints on the number of wires to be used or if the communication is only in one direction.

4.2 SYNCHRONOUS SERIAL COMMUNICATION

Synchronous serial communication relies on synchronized clocks between the devices on the serial bus, allowing each to sample and transmit data at known intervals. Compared to asynchronous serial communication, this eliminates the need for start and stop bits, thereby increasing throughput. One of the most common synchronous serial communication protocols is the Serial Peripheral Interface (SPI), a board-level communication standard that uses a shared clock line for synchronization and typically operates at bit rates exceeding 10 MHz.

4.2.1 Serial Peripheral Interface (SPI)

The serial peripheral interface (SPI) is a communication interface used to send data between multiple devices. These devices are organized into a master and slave configuration, in which the master has control over the slaves and the slaves receive instruction from the master. The most common implementation of SPI consists of a configuration in which a single device is the master, and the remainder of the devices are slaves. SPI is a synchronous communication protocol that transmits and receives information simultaneously with high data transfer rates and is designed for board-level communication over short distances.

The SPI communication interface is advantageous when needing to communicate between multiple devices. It offers a higher data transfer rate than many other types of communication interfaces and allows for data to be sent and received at the same time. However, SPI also demands more signal lines or wires than other types of communication. There is also no standard message protocol for communicating over SPI, meaning that every device could have its own convention for data message formatting.

SPI Signals

There are four signals required to implement SPI communication, as listed in Table 4.2, with all but the MISO line controlled by the master. Chip Select (CS), sometimes referred to as Slave Select (SS), is also often denoted as $\overline{CS}$ or $\overline{SS}$ because a particular chip/slave is active when that line is pulled low by the master (the line over the top indicates an inverted signal).

Required SPI Signals

SIGNAL	DESCRIPTION
MOSI	Data: Master Out - Slave In
MISO	Data: Master In - Slave Out
SCLK	Serial Clock
CS	Chip Select

TABLE 4.2

As shown in Figure 4.3, four wires are required to connect each of these signal lines between a single master and a single slave. These wires connect to the same signal on both devices, namely SCLK connects to SCLK, MOSI to MOSI, MISO to MISO, and CS to CS. In a multi-slave configuration, all signal lines are shared among all slaves, with the exception of the CS line which is independently controlled for each slave.

SPI Wiring

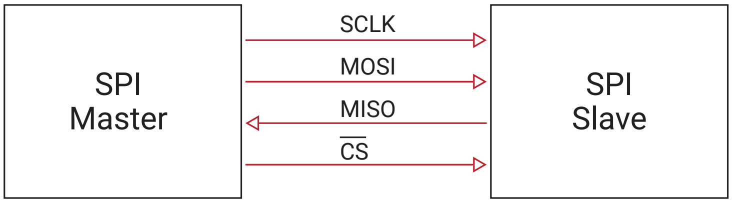


FIGURE 4.3

Clock Signal

The clock signal is generated by the master device to a specific frequency and is used to synchronize the data being transmitted and received between devices. This signal can be configured by the master by using two properties known as clock polarity (CPOL) and clock phase (CPHA). Clock polarity determines the polarity of the clock signal and can be configured to idle either low (0) or high (1). A clock signal that idles low has a high pulse and a rising leading edge, whereas a clock signal that idles high has a low pulse and a falling leading edge as illustrated in Figure 4.4.

As displayed in Figure 4.4, the clock phase determines the timing in which the data is to be modified and read. If the clock phase is set to zero, the data is modified on the trailing edge of the clock signal and the data is read on the leading edge. Conversely, if this property is set to one, data is changed on the leading edge of the clock signal and read on the trailing edge. As the clock cycles, data is sent bit by bit, simultaneously, over the MOSI and MISO lines.

Clock Polarity and Phase

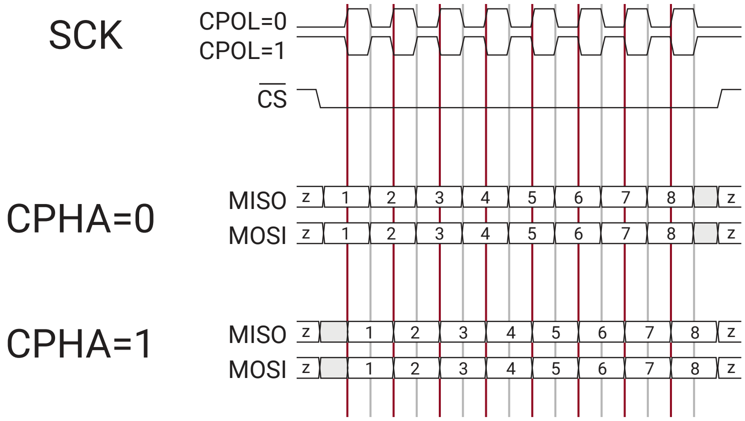


FIGURE 4.4

MOSI and MISO Signals

There are two data lines used in SPI communication known as MOSI and MISO. The MOSI signal sends data out from the master and is received by all slaves. Similarly, the MISO data line transmits data from one of the slave devices to the master device.

Chip Select Signal

The chip select pin is utilized by the master to select which slave to communicate with. This line for the specific slave should be pulled low when the master wants to communicate with the slave. If multiple slave devices are used on the same bus, then each slave will have its own dedicated chip select line, while sharing the clock and data lines. When the master is finished communicating with the slave, the chip select line is pulled back high.

4.3 OTHER COMMUNICATION STANDARDS

While SPI and UARTs are two of the more common communication interfaces used, there are numerous other protocols including Inter-Integrated Circuit, Controller Area Network, Industrial Ethernet, and Universal Serial Bus. Each of these interfaces have their own advantages and drawbacks that make them useful for different types of applications.

4.3.1 Inter-Integrated Circuit (I2C)

Inter-Integrated Circuit (I2C) is a communication interface that is often considered to be a mix between SPI and asynchronous serial communication. This interface is ideal for board-level communication and allows multiple master devices to communicate with multiple slave devices at a rate of either 100 kbps or 400 kbps. Only two signals are needed for communication between a master and a slave, a clock signal (SCL) and a data signal (SDA). Rather than using a slave select line to communicate with a specific slave, the master calls an address unique to each slave. I2C is a popular communication interface used in many board-level applications due to its easy implementation and ability to communicate with multiple devices.

4.3.2 Controller Area Network (CAN)

The Controller Area Network (CAN) communication interface was created by Bosch in the mid-1980s for use in automotive applications. CAN utilizes a network of devices that allows every message to be transmitted to each device on the network. A specific device can then determine what to do with the message, depending upon whether or not it is pertinent to the device's function. A few years after it was first developed, CAN was accepted as the international standard for communication in the automotive industry. Since then, CAN has become a popular communication interface in industrial, medical, transportation, and aerospace applications.

4.3.3 Industrial Ethernet

The industrial Ethernet communication protocol is becoming a popular interface, particularly in industrial automation applications. This communication interface offers higher data transfer speeds, longer cable length, and a more flexible network configuration with the capability to connect to numerous devices. There are many widely used industrial Ethernet protocols, such as EtherCAT, EtherNet/IP, PROFINET, POWERLINK, Sercos III, and Time-Sensitive Networking (TSN).

4.3.4 Universal Serial Bus (USB)

Universal Serial Bus (USB) is a communication interface that was developed as a standard for connecting devices to personal computers, taking the place of the larger ports used in these computers. This interface is constantly changing and utilizes different types of data transfer, higher data transfer rates, and allows for up to 127 devices to be connected at a time. One of the few drawbacks of USB is that the length of the cable used must be 5 meters or less. Nevertheless, USB is used in a large number of applications today.

4.4 ELECTRICAL INTERFACES

Outside of the data transmission pins described in the communication sections, there are a number of other electrical interfaces for powering and synchronizing to inertial systems that require care. Unregulated voltage, floating signal pins, and electrostatic discharge can have unintended effects on an electronic system. Voltage regulators, pull up/down resistors, and standard grounding practices are used to prevent these issues.

4.4.1 Voltage Regulation

A voltage regulator is a device that is designed to automatically regulate a voltage level when a stable and constant voltage is required. Voltage regulators are able to maintain a steady output voltage for significant changes in input voltage or load conditions. There are two types of voltage regulators: linear voltage regulators and switching voltage regulators.

Linear Regulator

Linear voltage regulators operate like a variable resistor, using closed feedback loops to maintain a constant output voltage as the input voltage or current load varies. All linear regulators require a higher input than the output meaning that the output voltage will always be less than the input voltage. Voltage regulation is accomplished by dissipating the excess voltage as heat. Like a variable resistor, the current draw at the input and the output are identical. A linear regulator's efficiency is equal to the ratio of the output voltage to the input voltage.

The minimum voltage difference between the input and the output voltage needed to supply a desired current is called the dropout voltage. Having a larger voltage difference between the input and the output or having a larger current means that there will be more power loss in the form of dissipated heat. For cases where the input voltage is very close in value to the output voltage, low-dropout (LDO) regulators are required.

Linear regulators are designed to have the advantage of having a very low ripple output voltage with very little noise coupled into the DC output. This makes them useful for sensitive analog electronics, like many inertial sensors or RF systems, but their low efficiency for large voltage drops limits their utility.

Switching Regulator

Switching regulators are a type of voltage regulator that rapidly switch a series device on and off, such that the average voltage at the output is correct. Capacitors and inductors are used to smooth this on-off voltage into a smooth, steady output voltage, though with considerably higher output voltage ripple than a linear regulator.

Unlike linear regulators, a switching regulator is able to produce output voltages that are higher than the input voltage or of opposite polarity. There are two main types of switching regulators: step-up, or boost regulators, and step-down, or buck converters. Step-up regulators provide a higher output voltage by increasing or stepping-up the input voltage. Conversely, step-down regulators perform the opposite action of the step-up regulator by lowering the input voltage to a desired regulated output.

Most switching regulators operate at 100s of kilohertz which is a common source of EMI in inertial systems. Furthermore, the requirements for capacitors and inductors to smooth and stabilize the output generally requires more components and board space than a linear regulator. However, switching regulators often operate at efficiencies of 80-90%, making them ideal for large voltage drops.

4.4.2 General Purpose Input/Output (GPIO)

A general purpose input/output is an uncommitted digital signal pin on an integrated circuit whose behavior (being in an input or output state) is controlled by the user. Hence, they can be configured as either an input or an output pin.

Input & Output Modes

When a GPIO is configured as an input pin, there are three configuration modes that can be assigned to the pin: high-impedance, pull-up, or pull-down. As stated in the tri-state logic section, high-impedance is the default state and is a floating pin. The pull-up mode uses an internal resistor to connect the input pin to Vcc. The pull-down mode uses an internal resistor to connect the input pin to ground.

When a GPIO in configured as an output pin, there are two configuration modes that can be assigned to the pin: push-pull and open-drain. The push-pull pin mode connects the output pin to either Vcc or ground and thus has the ability to source or sink the current. The open-drain mode has two the ability to achieve two output states: low-impedance and high-impedance, referring to the amount of resistance to current flow. Unlike the push-pull mode, open-drain only has the ability to sink current. In order to achieve a high state, an additional pull-up resistor is required to connect to the desired voltage level.

Tri-State Logic

In digital electronics, tri-state or three-state logic states that a pin can assume three independent states: logical 0, logical 1, and high-impedance. Logical 0 means that the connection is to the ground reference, logical 1 means that the connection is to Vcc and lastly the high-impedance state, also known as Hi-Z, effectively removes the devices influence from the rest of the circuit. Alternatively, the term “floating” can be used for a pin which is in Hi-Z mode due to the fact that the pin is connected to neither Vcc nor ground.

Pull-Up & Pull-Down Resistors

A schematic showing the implementation of pull-up and pull-down resistors is shown in Figure 4.5. A pull-up resistor holds an input pin in a high state unless the input is shorted to ground, whereas a pull-down resistor keeps an input pin in a low state unless it is shorted to Vcc. This is useful to guarantee that in an input pin is in a known state when nothing is otherwise connected to it without impeding normal operation. Connecting the pins directly to Vcc or ground would render the pin inoperable because a device connected to that pin would create a short circuit when attempting to signal low or high, respectively.

Pull-Up & Pull-Down Resistors

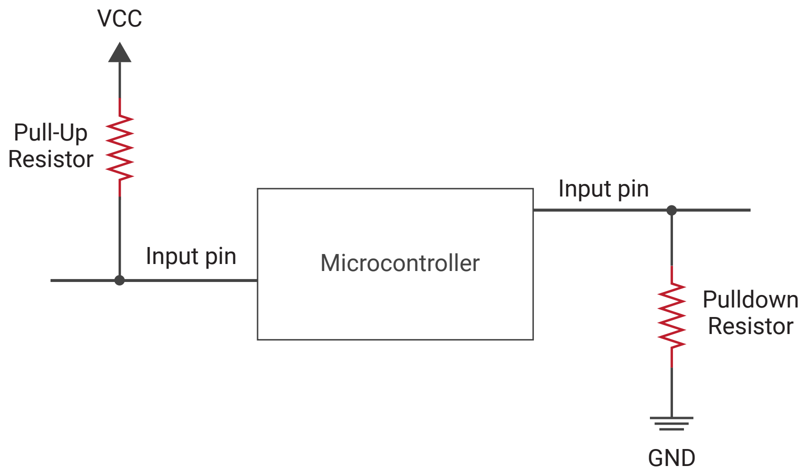


FIGURE 4.5

A typical pull-up/down resistor value ranges from 1-10 kΩ, but can go as high as 100 kΩ depending on the application. A lower valued resistor is considered a ”stronger” pull-up/down, while the higher valued resistors are considered ”weaker”, based on the current draw that is required to toggle the state of the pin. A stronger pull-up/down requires a greater current draw, and therefor reduces the likelihood that transients on a signal line (noise, EMI, etc) will cause a false trigger on the input pin.

It is important to note that a weaker pull-up/pull-down will result in a slower voltage change on an input pin due to the coupling of the resistor and the wire capacitance which forms an RC circuit. The larger the product between these two values, the more time is required for the capacitance to charge and discharge due to the resultant time constant value.

IO Speed (Slew Rate)

GPIO speed is what controls the rate at which a signal can change between high and low states. Common terms for IO speed include “frequency” and “slew rate” as they are generally used synonymously. The speed of the signal is dependent on signal path capacitance as the capacitance will slow voltage rise/fall times of the signal. Higher IO speeds increase the rate of change of the output voltage meaning that the rise time of the signal is reduced.

However, this requires high power consumption and causes an increase in radiated noise such as Electromagnetic Interference (EMI).

4.4.3 Electrostatic Discharge (ESD)

Electrostatic discharge is the sudden flow of electricity between two objects that have different charges and number of electrons. The flow of electricity from one object to another creates a dielectric breakdown, sometimes creating a visible spark. ESD is an important phenomenon to be aware of due to its ability to cause damage to electrical components and potentially start fires or create explosions. Damage caused by ESD can be difficult to detect and may result in reduced system life, leading to an early failure long after the actual ESD event.

Therefore, it is important to protect sensitive components, especially microchips, during the manufacturing, assembly, and shipping processes from potential ESD occurrences. Additionally, users must be mindful when working with a device or component by hand as they may have accumulated an electric charge that can be dissipated via ESD through to the device. Due to this, it is important to follow proper safety and grounding procedure to combat any possible instances of ESD, such as wearing grounded wrist-straps at workbenches.

4.5 RF HARDWARE

The Global Navigation Satellite System (GNSS) relies on the use of radio-frequency (RF) waves to transmit the navigation message. Understanding how devices receive the waves is an important part of using the system. Once transmitted, a GNSS satellite's signal must be received by a device using an antenna. There are an assortment of antenna types available with different characteristics, cabling, and connectors as well as a variety of different antenna accessories.

4.5.1 GNSS Antennas

The three major antenna types used in GNSS systems are patch, helical, and choke-ring antennas. The shape of the sensing element that captures the RF waves is the major difference between them. Antenna design can also affect the gain, a measure of how well an antenna receives the intended RF energy. Receiving antennas are characterized by how efficiently they can capture the incoming signal from the direction of interest. This is known as the gain and is specified in units of decibels (dB). In general, larger antennas contain larger sensing elements which leads to higher gains.

Each antenna also has a location where the signal is tracked known as the phase center, which can vary over temperature and may be important in some applications.

Active vs. Passive Antennas

Passive antennas consist of only an RF receiving element and draw no power from the receiver while active antennas contain a low-noise amplifier and must receive power from the receiver through the RF connector. Active antennas typically add 3-50 mA of current draw to a system. The low-noise amplifier allows an active antenna to compensate for cable loss and increases the gain closer to unity.

Patch Antennas

Patch antennas, also called microstrip antennas, consist of a sheet of metal that acts as the sensing element separated by an insulator from a larger sheet called a ground plane. This allows for a low-profile shape that is good for mounting on flat surfaces. The ground plane helps reduce multipath, but also makes patch antennas directional. Patch antennas can be cheaply and easily fabricated on printed circuit boards and are commonly used in mobile electronics. Patch antenna construction and an example product form can be seen in Figure 4.6.

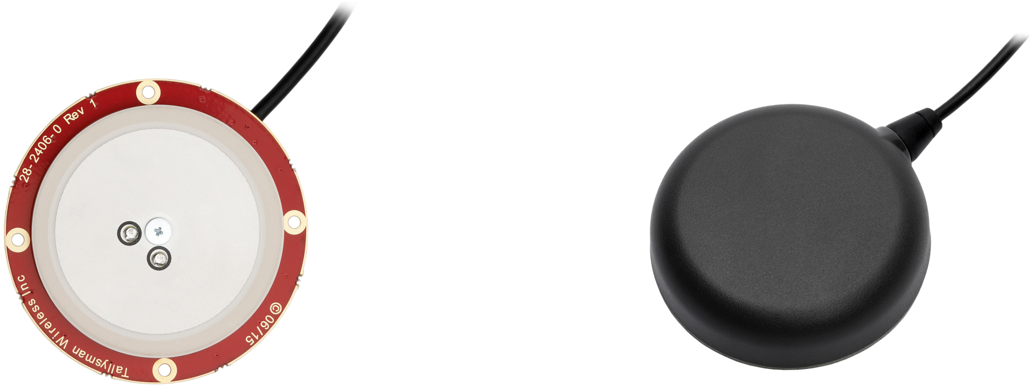
Helical Antennas

Helical or helix antennas are made of one or more wires wound into the form of a helix, and take a cylindrical shape, as seen in Figure 4.7. While most often mounted over a ground plane, omnidirectional designs can be achieved by omitting the ground plane. Helical antennas are versatile since they can operate in either normal mode or axial mode depending on the helix circumference relative to the intended wavelength. Normal mode is used when transmitting or receiving waves that are perpendicular to the helical axis, while the axial mode is for transmitting or receiving waves in the direction of the helical axis. These antennas can be made small enough for mobile applications or much larger. Due to their design, helical antennas are more susceptible to multipath.

Choke-Ring Antennas

Choke-ring antennas are a directional design that consists of a central receiving element and a series of hollow concentric rings which act to greatly remove multipath signals. The multipath rejection and high phase-center stability

Patch Antennas (Images Courtesy of Tallysman)

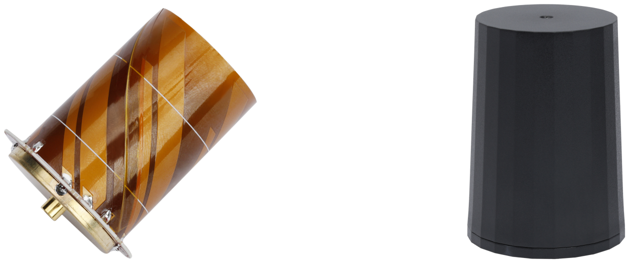


(a) Embedded

(b) Housed

FIGURE 4.6

Helical Antennas (Images Courtesy of Tallysman)



(a) Embedded

(b) Housed

FIGURE 4.7

of these antennas allow for the millimeter-level accuracy required for surveying applications. However, their large size makes them less ideal when mobility is required. A choke-ring antenna design is shown in Figure 4.8.

Choke-Ring Antenna (Image Courtesy of Tallysman)

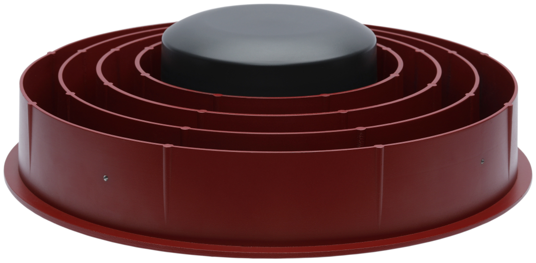


FIGURE 4.8

Ground Planes

The RF waves transmitted from the GNSS satellites can be susceptible to multipath interference, which occurs when the signals reflect off solid objects such as buildings and terrain prior to reaching the GNSS antenna. This causes the satellite signal to make multiple paths before reaching the GNSS antenna and can cause errors in the navigation solution. To prevent multipath interference from beneath the antenna, a ground plane is often mounted under the GNSS antenna. Ground planes can be any thin piece of metal, including foil, that block any multipath interference from reflecting up to the base of the antenna. Ground planes do not need to be electrically grounded.

4.5.2 RF Connections

GNSS signals occur below the noise floor and require special processing algorithms to recover the signal. For this reason, it is important to reduce any potential additional losses. Antenna characteristics, cable loss and connector loss can each contribute to how well a signal is received. The choice of cable, connector, and optionally a splitter, will have an effect on the GNSS signal strength.

Cables

Due to the high frequencies used in GNSS signal reception, coaxial cables are used. Coaxial cables consist of a central conductor surrounded by a dielectric, an outer conductor, and an outer insulator. Keeping cables short reduces GNSS signal strength loss, as does larger diameter cables.

Connectors

Figure 4.9 shows three common RF connectors: U.FL, SMA, and MMCX. Each varies in size and latching mechanism and force, so are useful for different applications.

U.FL connectors are the smallest, commonly used to attach an antenna directly on an exposed PCB near a GNSS chip. These connectors are not meant to be attached or removed much as this can wear them out quickly. They can operate up to 6 GHz and are typically used for short distances. Shown in Figure4.9a, a U.FL connector is typically only rated for ten connects and disconnects. Due to this, these connectors are designed for use as board to board connectors, rather than as panel mount connectors.

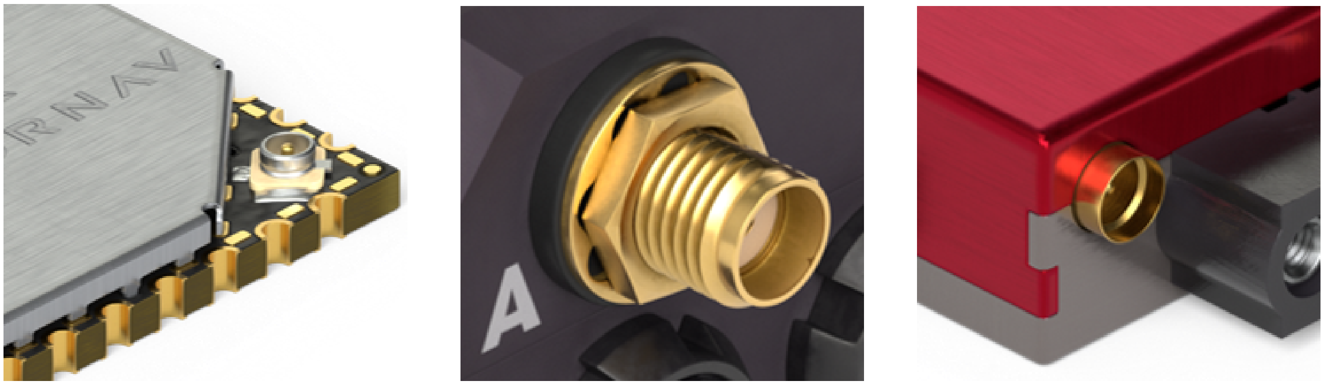
Sub-miniature version-A (SMA) connectors are available in male/female, but also in reverse polarity (typically denoted by RP) form that keeps the same electrical connections but puts the center pin on the threaded female connector. SMA connectors typically have performance up to 18 GHz and insertion loss as low as 0.17 dB. Figure 4.9b shows these connectors.

Micro-miniature coaxial connectors (MMCX) are smaller than SMA connectors and around a third of the weight. They have a 360-degree swivel mechanism and are popular in consumer electronics. MMCX connectors can operate up to 6 GHz with insertion loss of 0.3 dB and are designed for use as board to board connectors rather than as panel mount connectors. A MMCX connector is shown next to an SMA Connector in Figure 4.9c.

RF Splitters

RF splitters allow for multiple devices to receive the same GNSS signal. RF splitters divide a signal into 2 or more outputs, each having a fraction of the strength of the original. A 1-to-2 splitter will have a 3 dB decrease (50% power) on each output. Larger splitters are usually made from combinations of 1-to-2 splitters, so each additional division will lower the strength by 3 dB. It is important not to split a signal too many times or it may become difficult to recover for the GNSS receiver without adding another amplifier (which increases noise). In addition, splitters often

Common RF Connectors



(a) VN-200-SMD U.FL

(b) VN-210 SMA

(c) VN-200-CR MMCX

FIGURE 4.9

contain DC blocks on all but one antenna to prevent multiple powered antenna sources from being in parallel and damaging each other. The low-noise amplifier on an active antenna only needs one source of power.

4.6 ELECTROMAGNETIC INTERFERENCE (EMI)

When designing or integrating a product it is important to take electromagnetic interference (EMI) into consideration. Failing to account for the various types of EMI for both emissions and sensitivity can result in unexpected system behavior or even failure.

4.6.1 Transmission Modes

The two main modes of action for EMI are through conducted and radiated means. Conducted types involve noise that passes through a conductor into or out of a device, while radiated noise is received or transmitted wirelessly. It is also possible for a combination of noise types to be present within a system, and that multiple mitigation methods will be required. The various types can be seen in Figure 4.10, and testing for each of them is described in detail by MIL-STD-461, which is described in more detail in Appendix A.4.

Electromagnetic Compatibility

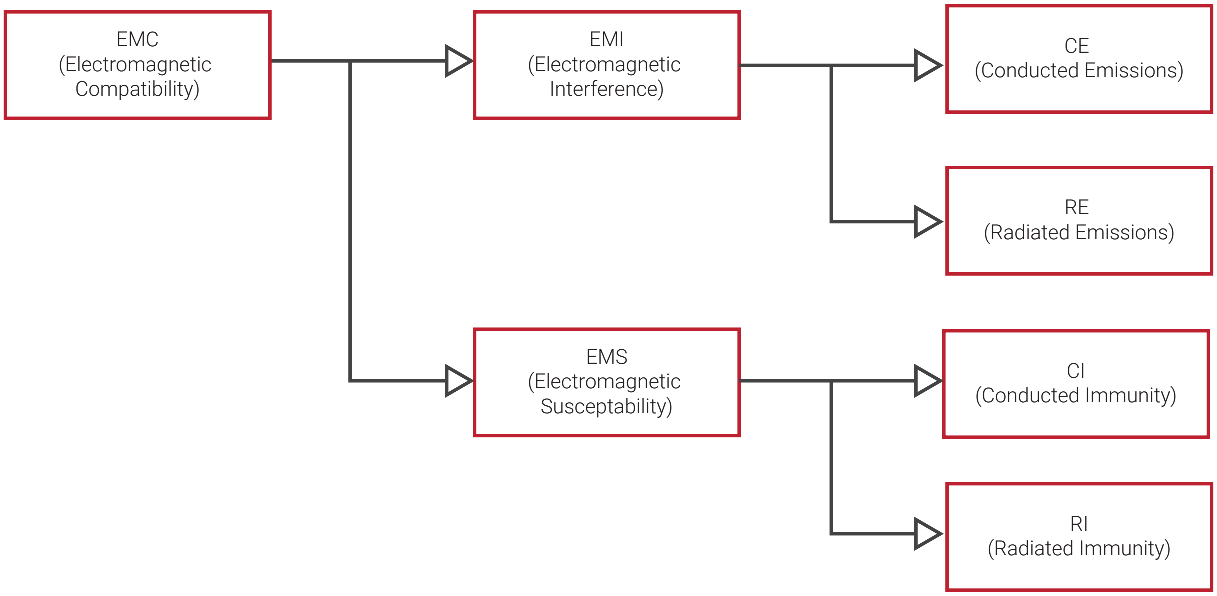

EMC (Electromagnetic Compatibility)
EMI (Electromagnetic Interference)
CE (Conducted Emissions)
RE (Radiated Emissions)
EMS (Electromagnetic Susceptability)
CI (Conducted Immunity)
RI (Radiated Immunity)

FIGURE 4.10

At lower frequencies, noise usually travels through conducted means, while at higher frequencies, it is more often radiated. It is also often the case that a device will be vulnerable to the same frequencies that it is emitting. It is also possible that what appears to be conducted noise may be the result of a radiation issue, such as cable cross-talk in cable bundles or PCB layout issues.

Conducted Susceptibility/Immunity (CS)

Conducted susceptibility is the vulnerability of a system to noise that enters through a conductive path, such as power inputs/returns. A device may share a power rail with multiple other devices and each one could contribute noise to the rail, so it is important to ensure that all devices can continue to function properly. At the printed circuit board level, most integrated circuits recommend decoupling capacitors located near the power pins to prevent high frequency noise from entering the IC. This can be applied to the overall system as well with some success, depending on the frequencies of noise involved. It may also be beneficial to include a low-pass filter designed to attenuate the frequency of noise present.

Conducted Emissions (CE)

Conducted Emissions is the noise that a system puts out onto its conductive connections that could interfere with another device. Locating and altering the noise source to no longer output noise that reaches the power rails is the ideal option. Where this is not possible, adding filtering can usually reduce this noise. This type of noise is commonly associated with switching power supplies/regulators. Cable cross-talk could also give the appearance of conducted noise. The use of individual twisted, shielded pairs in a cable can help reduce the effects.

Radiated Susceptibility/Immunity (RS)

Radiated Susceptibility is the vulnerability of a system to wireless noise from the surrounding environment. A device must be able to encounter a reasonable level of outside interference and still function properly. Proper shielding such as a Faraday cage equivalent chassis can help block out radiated noise, as well as twisted shielded pairs in cables. Filters and internal shielding around sensitive components may be necessary if it is not possible to prevent radiated noise from entering a device.

Radiated Emissions (RE)

Radiated Emissions is the noise a system puts out into its environment. It is important that this noise be at an acceptable level to not cause any issues withing surrounding systems or equipment. Design factors such as case material and cable selection can play a huge role in radiated emissions. A well shielded setup will radiate very little noise. If shielding the entire setup is not an option, it may be helpful to locate the radiating source within the device and find a way to attenuate it there, such as localized shielding or using a different part operating at a different frequency.

4.6.2 Mitigation Techniques

While mitigating the effects of EMI, whether emitted or received can influence design choices throughout the entire system, careful use of grounding, shielding, and filtering of problematic lines can significantly reduce EMI effects with minimal impact.

Shielding

The best way to shield a device is to emulate a Faraday cage. A Faraday cage surrounds an object with a conductor which helps block or greatly reduce most electromagnetic signals from reaching the inside. While a solid conductor works best, a mesh can be applied as long as the holes in the mesh are smaller than the wavelength of the signals being blocked. By selecting a metal chassis as a device enclosure, a Faraday cage-like effect is achieved.

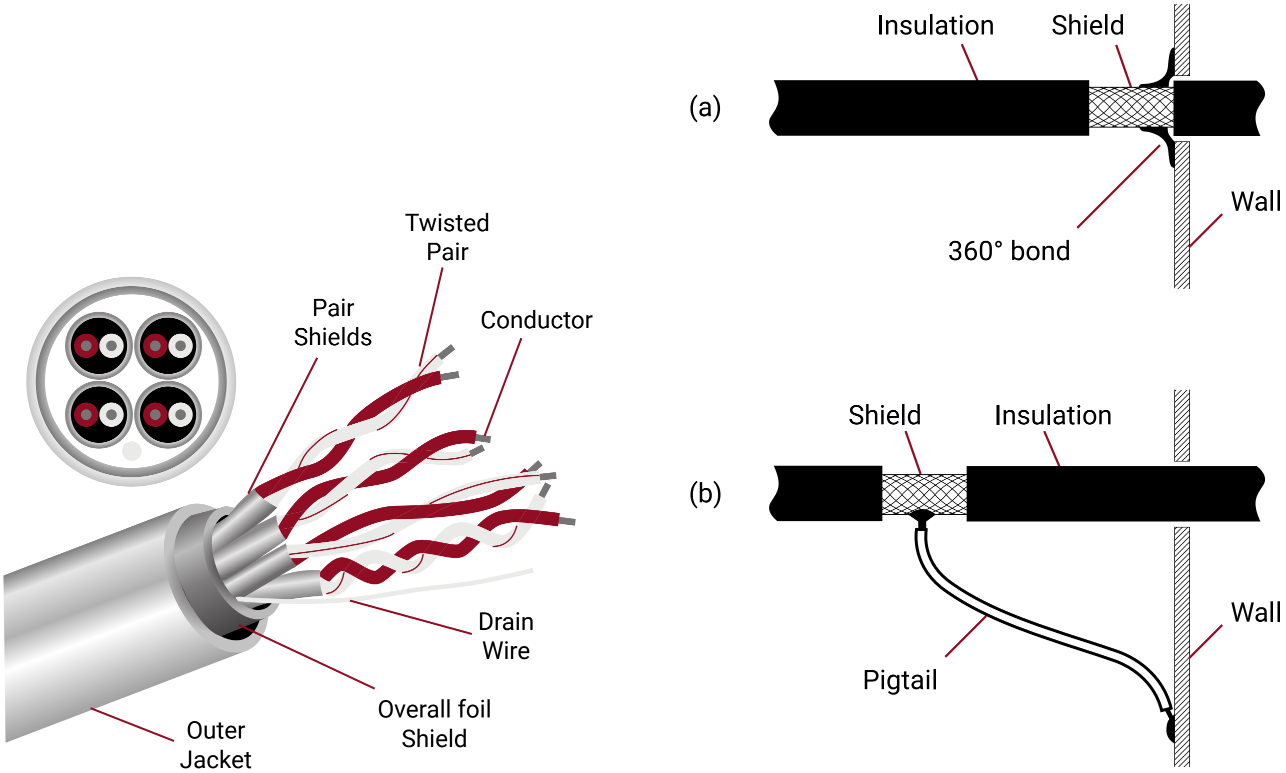
Cables and connectors will require openings in the cage that noise could still get through. Using metal connectors can help when paired with cable shielding. There are a few different types of cable shield and ways to attach that shield to the connector on each end. While an overall cable shield helps with emissions, individual pair shields as shown in Figure 4.11a can help in keeping noise from coupling into adjacent wires or becoming ’conducted emissions’ on power leads. Foil shield does a better job than braided shield at high frequency, but is not as mechanically strong. Some cables use a combination of braided and foil shielding to get the advantages of both. Using twisted pairs also helps to reduce common-mode noise, which is noise that occurs on both conductors.

Cables also need their shield tied to ground at both ends, where possible, to help with the Faraday effect. The way that this is done can also have an effect on EMI. Creating a 360° bond is more effective than just connecting the shield to the chassis with a wire ’pigtail’, as shown in Figure 4.11b.

Filtering

When shielding techniques are not enough, adding a filter can assist in conducted EMI reduction. A low-pass filter consisting of an inductor and capacitor can attenuate high frequencies while leaving DC voltages alone. A band-stop filter may be a better tool in some cases, as it can attenuate noise in a range while leaving the remaining frequencies

Cable Shielding & Grounding



(a) Cable Shielding

(b) Cable Grounding

FIGURE 4.11

unchanged. Many different filter configurations exist, with most having online calculators to help in designing to filter a certain problem frequency.

5 ALTERNATIVE NAVIGATION

While GNSS measurements are the most common aiding source for a navigation system, many applications do not have access to GNSS or operate in environments that are susceptible to the disruption of GNSS signals. In such cases, a variety of technologies can be used to provide a robust and reliable navigation solution, often in a multi-sensor approach.

5.1 GNSS-CHALLENGED ENVIRONMENTS

A major drawback of a GNSS-aided navigation system is how weak the GNSS signal is once it reaches a receiver on Earth. Consequently, it can be quite easy to disrupt GNSS, whether unintentional or deliberate. Such disruption could degrade the signal or even deny its reception altogether.

5.1.1 GNSS-Degraded Environments

GNSS-degraded environments occur when a GNSS signal is present but degraded in some way. This is most often caused by signal interference which causes a disturbance on the signal transmitted by the GNSS satellites. Such interference can be deliberate, as in the case of jamming or spoofing, or unintentional, such as disturbances from radio broadcasting signals or multipath interference. Even something as simple as a hardware malfunction along the RF path, such as a bad GNSS antenna or loose connection, could lead to degraded GNSS conditions. In these environments, specialized hardware and algorithms exist, as discussed in Section 5.2, that can help mitigate these challenges and allow GNSS to be used for navigation.

Jamming

Because GNSS signals are quite weak when received by an antenna on Earth, they can easily be overwhelmed by higher-power signals. Jamming is the intentional use of such higher-powered signals to drown out the signals transmitted by the GNSS satellites, as seen in Figure 5.1a. This prevents a GNSS receiver from tracking the GNSS signals and providing a navigation solution. Though a jamming system is typically operated from the ground, it can render GNSS unusable over a wide area—sometimes up to hundreds of kilometers.

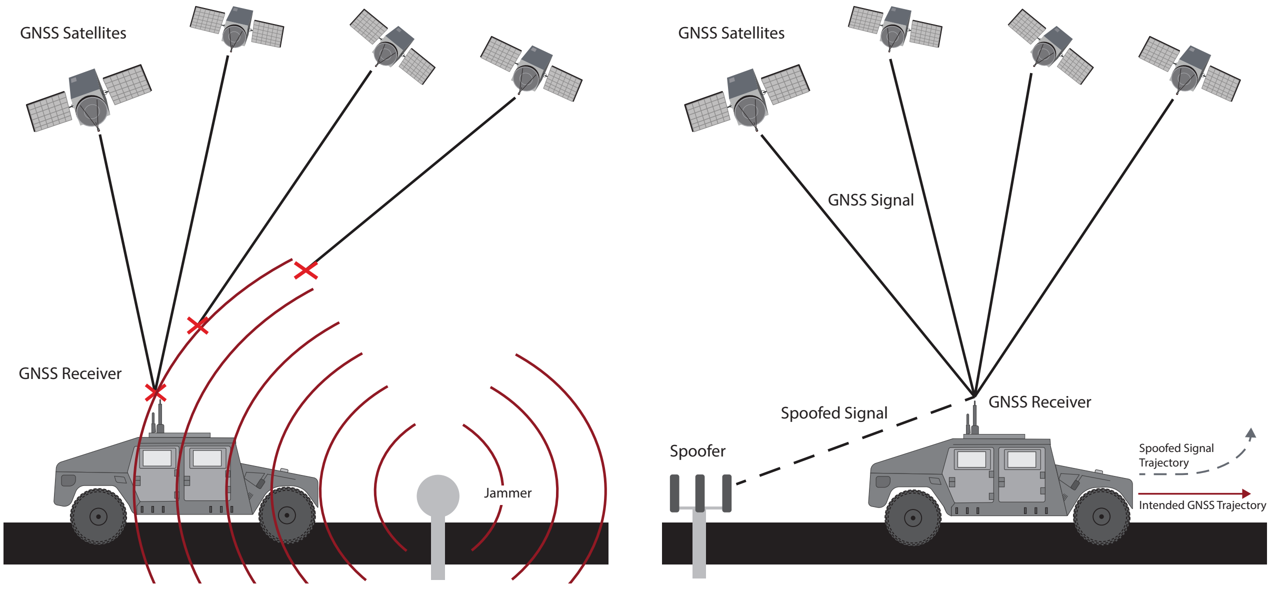
Spoofing

Spoofing is a deliberate attack in which false GNSS signals are generated with the intent to lead a GNSS receiver off track. Because the structure of many of the GNSS signals is publicly known, an adversary can construct signals closely resembling those from the GNSS satellites but with incorrect information. Often, a spoofer will first jam the GNSS signals they intend to spoof, preventing a user's GNSS receiver from tracking the GNSS signals. The spoofed signals are then broadcasted by the spoofer at a higher power level than the GNSS signals, causing the GNSS receiver to lock primarily onto the false signals rather than those from the GNSS satellites. This leads the GNSS receiver to calculate an incorrect navigation solution as shown in Figure 5.1b—most commonly an incorrect position, but potentially an incorrect time as well. All GNSS constellations are vulnerable to a spoofing attack, however, the complexity of generating consistent counterfeit signals increases as additional constellations and frequencies are added.

Unintentional Interference

GNSS signals are broadcast at specific frequencies in the L-band of the electromagnetic spectrum. This portion of the electromagnetic spectrum is highly regulated to prevent unintentional interference from other radio transmissions. However, signals in other bands can sometimes bleed over into the GNSS frequencies. Often this is caused by equipment operating too close to a GNSS antenna or higher-powered signals that have harmonics in the GNSS bands, such as television broadcasts. Similar to jamming, this unintentional interference overwhelms the GNSS signals causing a receiver to lose track of the signals on that GNSS frequency. Typically, only a single GNSS frequency is affected by unintended radio interference.

GNSS-Degraded Conditions



(a) Jamming

(b) Spoofing

FIGURE 5.1

Multipath Interference

Multipath interference occurs when GNSS signals reflect off solid objects or surfaces, such as buildings and terrain, resulting in the signal taking multiple paths to reach the antenna, as shown in Figure 5.2. These reflected signals are delayed compared to direct line-of-sight signals, causing errors in the pseudorange and carrier phase measurements. Multipath interference is most commonly encountered in areas heavily populated with tall buildings, known as urban canyon environments, though any system lacking a clear view of the sky is susceptible to such interference.

GNSS Multipath

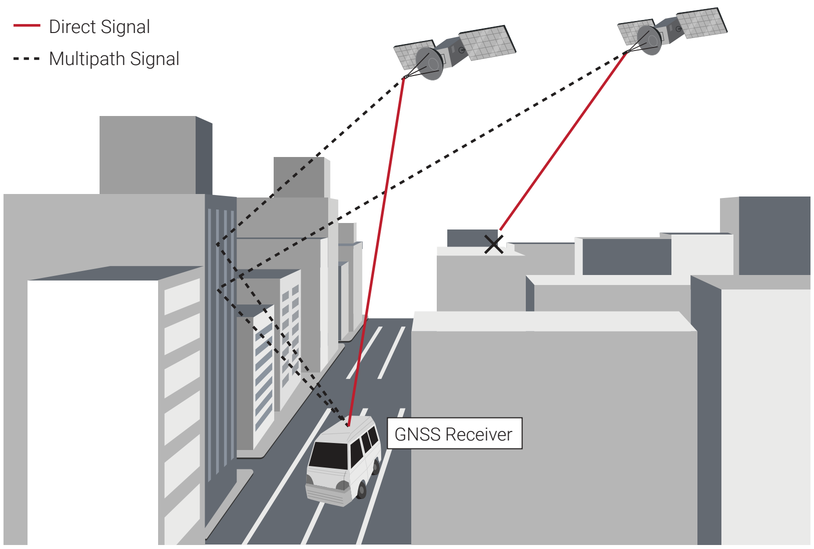


FIGURE 5.2

5.1.2 GNSS-Denied Environments

Environments that are completely devoid of a GNSS signal are often referred to as GNSS-denied environments. Such environments include indoor settings, underground or underwater applications, and GNSS jamming scenarios when access to anti-jamming hardware is not available. In such scenarios, other navigation sensors and techniques must be relied upon to provide localization information, often through a multi-sensor approach. Some of the more common localization approaches for GNSS-denied navigation are detailed in Sections 5.3 and 5.4.

5.2 ENHANCED GNSS TECHNOLOGIES

Increasingly, systems must navigate in GNSS-challenged environments in which GNSS may be susceptible to jamming, spoofing, or unintentional interference. The use of enhanced GNSS technologies, including specialized hardware, sophisticated algorithms, and additional GNSS signals, can provide robustness during operation in GNSS-degraded conditions.

5.2.1 Military GPS Receivers

As discussed in Section 1.5, a GPS satellite signal has three components: the carrier, the code, and the navigation message. The code portion of the signal allows all satellites on the band to transmit on the same frequency without interfering with one another. While standard GPS positioning utilizes publicly available code signals, these signals can be spoofed with false information, leaving users vulnerable to such attacks. To combat spoofing, the U.S. military developed encrypted code signals that are broadcast by the GPS satellites. These signals can only be decrypted by specialized GPS receivers: SAASM receivers and M-code receivers.

SAASM Receivers

At the creation of GPS, two different positioning services were offered: the Standard Positioning Service for public use and the Precise Positioning Service for authorized military users. During the 1990s, the publicly available service was intentionally degraded to limit the position accuracy that civil users could obtain through a feature known as Selective Availability (SA). The Precise Positioning Service utilized Precise code (P-code) which was encrypted with W-code to become P(Y)-code to protect authorized users from false GPS signals generated during a spoofing attack.

While Selective Availability was deactivated in the early 2000s, the P(Y)-code remains encrypted to provide anti-spoofing (AS) capabilities. To decrypt the P(Y)-code, authorized users must utilize a specialized receiver called a Selective Availability Anti-Spoofing Module (SAASM). SAASM receivers are commonly employed in military applications and are often a requirement for military systems needing GPS. While a SAASM receiver provides robustness to spoofing, the P(Y)-code signals are still relatively weak when received on Earth, leaving SAASM receivers vulnerable to jamming.

M-Code Receivers

To enhance the security of GPS systems for the U.S. military and its allies, an encrypted signal called M-code was developed. Representing the latest development of the GPS constellation, the M-code signal is an encrypted signal added onto the L1 and L2 GPS bands that allows for the GPS signal to be transmitted at a higher power without interference to the civilian C/A code or previous military P(Y)-code. Since M-code is an encrypted signal, authorized users must utilize a specialized receiver called an M-code receiver to decrypt the signal.

Similar to a SAASM receiver, an M-code receiver can reject false signals and provide robustness to spoofing. However, compared to a SAASM receiver, an M-code receiver provides the advantage of being more resistant to jamming due to the higher transmission power of the M-code signal. M-code is expected to replace P(Y)-code, with many military applications already upgrading from legacy SAASM receivers to M-code receivers to take advantage of the anti-jamming and anti-spoofing capabilities.

5.2.2 Anti-Jam Antennas

GNSS jammers drown out the relatively weak GNSS signals by generating a far stronger signal in the GNSS frequency band(s), preventing a GNSS receiver from tracking the real signal (see Section 5.1). A specialized antenna—the Controlled Reception Pattern Antenna (CRPA) (also referred to as an anti-jam antenna)—is often utilized to combat GNSS jamming as seen in Figure 5.3. The main advantage of a CRPA is that it can provide resistance to jamming by simply swapping out the GNSS antenna without impacting other components in the navigation system.

A CRPA consists of multiple antenna elements that are spatially distributed, making the CRPA larger than typical single-element patch antennas. Signals from different directions strike these different antenna elements at different times. Electronics inside the CRPA can amplify or attenuate different signals by varying the phase shift applied to each element when combining them into a single RF output used by the GNSS receiver. Amplifying a particular signal is often referred to as beamforming or beam-steering while attenuating a particular signal is referred to as nulling or null-steering.

While the algorithms behind beamforming (amplifying the GNSS satellite signals) and nulling (canceling the jamming signal(s)) are complex and proprietary to each manufacturer, they typically take advantage of two methods for distinguishing the jamming signal(s) from the satellite signals. The first is spatial distribution: GNSS satellites are typically above while jammers are typically at the horizon or below. The second is a difference in power: jammers

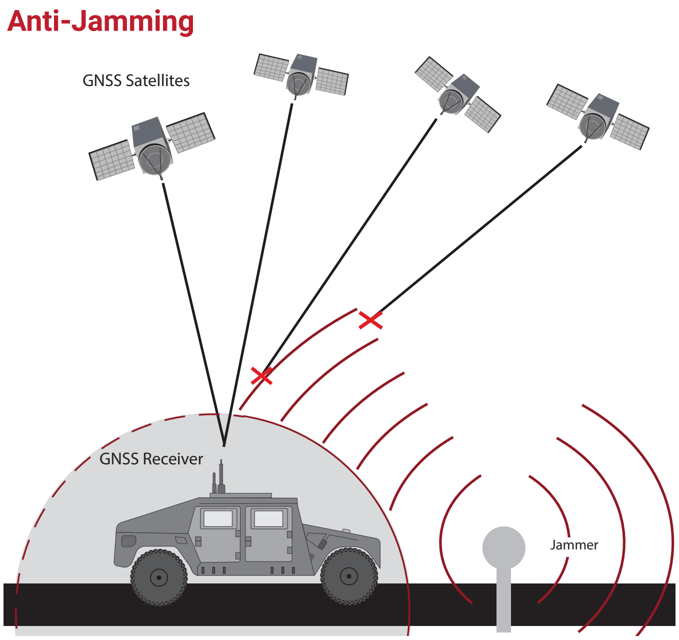


FIGURE 5.3

generate a signal many times stronger than the satellites. Because of these two distinctions, CRPA antennas can perform beamforming and nulling without feedback from the GNSS receiver.

A CRPA is capable of nulling up to $n-1$ jammers, where n is the number of antenna elements, but it is most effective when there are significantly more elements than jammers. Those extra antenna elements allow the optimization algorithms more degrees of freedom to effectively null the jammer(s) while maximizing the satellite signals. The effectiveness of this nulling, measured by the attenuation in decibels (dB), also varies between manufacturers due to the algorithms they employ.

5.2.3 Tightly-Coupled GNSS/INS

Most commonly, a GNSS receiver is combined with an INS system using a loosely-coupled approach, which incorporates a GNSS receiver's calculated position, velocity, and time (PVT) with the inertial sensor measurements to provide a fused position, velocity, time, and attitude solution. However, this approach necessitates that four direct line-of-sight satellites are in view of the GNSS antenna. Some environments are prone to structures that block and reflect a GNSS signal—the classical example of this is an urban canyon as shown in Figure 5.2.

In restricted visibility environments, it may not be possible to acquire a signal lock on four direct line-of-sight satellites. Rather than using the GNSS receiver's calculated navigation solution, a tightly-coupled GNSS/INS filter combines the raw GNSS pseudorange and Doppler measurements directly with the inertial sensor measurements in an extended Kalman filter, as displayed in Figure 5.4. Such an approach still estimates the fused position, velocity, time, and attitude of the system; however, this integration method allows even a single satellite to provide useful navigation information.

In clear sky conditions, the tightly-coupled GNSS/INS solution typically provides the same level of accuracy as the loosely-coupled GNSS/INS solution. As such, it is not as widely used due to its additional complexity. However, as long as care is taken to reject measurements susceptible to multipath, a tightly-coupled GNSS/INS solution can provide more accuracy than a free-inertial solution in GNSS-degraded environments. A more detailed comparison of tightly-coupled and loosely-coupled GNSS/INS integration architectures can be found in Section 1.7.

5.2.4 Additional Constellations and Frequencies

The last decade has seen a significant expansion in GNSS constellations and frequencies, well beyond the original GPS constellation and its L1 and L2 frequencies (see Section 1.2). Those additional constellations and frequencies help multi-constellation/multi-frequency receivers maintain GNSS tracking in various GNSS-challenged environments.

Multi-constellation receivers can double or triple the number of satellites available for tracking relative to a GPS-only receiver. This is particularly important when only a narrow view of the sky is available or in a significant multipath environment because it dramatically increases the likelihood of having direct line-of-sight visibility to four or more

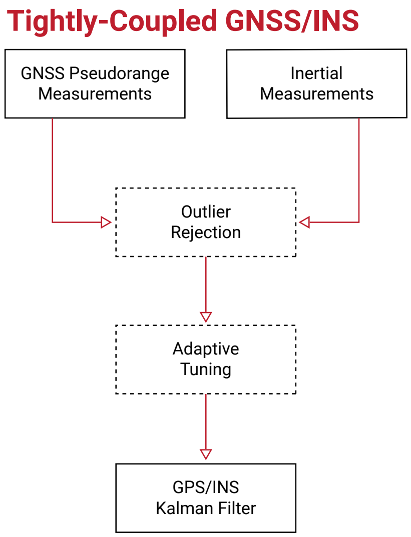


FIGURE 5.4

satellites as required for computing a full navigation solution.

Figure 5.5 and Table 5.1 show the wide range of GNSS frequencies now available across each of the different constellations. Unintentional interference often impacts only a small portion of the entire spectrum, so multi-frequency receivers provide robustness in those situations. Even intentional jamming or spoofing attacks often target only a single frequency or some small subset of frequencies.

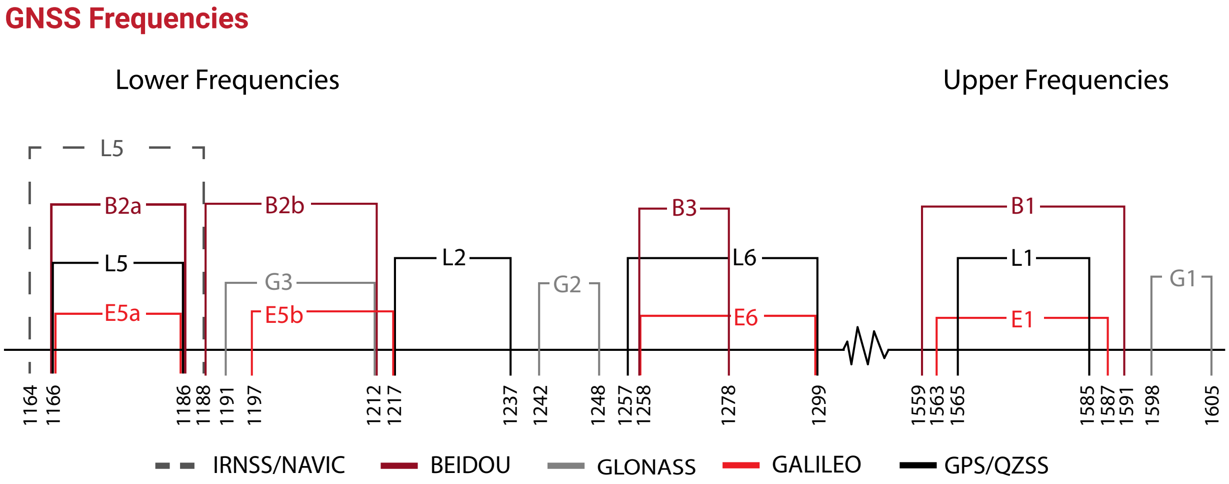


FIGURE 5.5

5.3 ABSOLUTE LOCALIZATION

In many systems, having access to an accurate navigation solution that is tied to an absolute frame of reference is critical to the success of its operation. The use of GNSS has become the most common approach for obtaining such global information; however, systems must increasingly operate in GNSS-challenged environments. As the need for suitable alternatives to GNSS continues to rise, the development of other absolute localization techniques is an expanding area of focus. While not an exhaustive list, this section provides a high-level overview of a few alternative techniques that can be used for absolute localization in GNSS-denied scenarios.

5.3.1 Environmental Maps

One common theme in many absolute localization techniques is the use of distinct features in the surrounding environment. These features can be geolocated and compiled into a corresponding map. When a system passes over the mapped area, the detected features are compared with the map and used to refine the system's navigation solution.

GNSS Frequencies

		FREQUENCY (MHz)		
CONSTELLATION	SIGNAL	LOWER	CENTER	UPPER
GPS	L1	1565.19	1575.42	1585.65
	L2	1217.37	1227.60	1237.83
	L5	1166.22	1176.45	1186.68
Galileo	E1	1563.14	1575.42	1587.70
	E5a	1166.22	1176.45	1186.68
	E5(altBOC)	1166.22	1191.80	1217.37
	E5b	1196.91	1207.14	1217.37
	E6	1258.29	1278.75	1299.21
GLONASS	G1	1598.06	1600.99	1605.37
	G2	1242.94	1248.06	1248.63
	G3	1191.80	1202.03	1212.26
BeiDou	B1	1559.05	1575.42	1591.79
	B2a	1166.22	1176.45	1186.68
	B2/B2b	1197	1207.14	1217
	B3	1258.29	1268.52	1278.75
IRNSS/NAVIC	L5	1164.45	1176.45	1188.45
QZSS	L1	1573.42	1575.42	1577.42
	L2	1226.58	1227.60	1228.62
	L5	1166.22	1176.45	1186.68
	L6	1257.75	1278.75	1299.75

TABLE 5.1

Vision-Based Navigation

Many satellites are equipped with cameras and possess the capabilities to photograph the surface of Earth. A system equipped with a camera, and flying at sufficient altitude, can also photograph the surface of the Earth and potentially match images with those taken by satellites as seen in Figure 5.6. If the satellite image is geolocated, then the position of the vehicle can also be determined using image processing techniques. A database of satellite images can be compiled so that wide geographic areas can be mapped using satellite imagery.

Successful image matching between satellite imagery and a real-time photographed image is sensitive to environmental conditions, such as lighting and the relative orientation of the cameras taking the image. Best results often require very similar lighting conditions and for the orientation of the cameras to be within a few degrees of each other. There must also be clear features between the two images for successful comparison, meaning vision-based navigation will not be possible when traveling over areas without unique markers, such as large bodies of water. Feature detection and tracking is also degraded in poor lighting environments and other conditions that decrease visibility, such as adverse weather. Because image matching is computationally expensive, such navigation systems require larger, more powerful processors.

LiDAR Digital Elevation Matching

A digital elevation model (DEM) is a model of Earth's surface that excludes features such as trees and buildings. These models can be produced using sensors such as a LiDAR scanner. An aerial vehicle equipped with a LiDAR scanner can scan the ground over which it flies and compare the scan with a DEM to determine if the scan matches the model of Earth's topography. Compiling a database of geolocated DEMs enables an aircraft to determine its global position in scenarios when a GNSS signal is not available.

This technique requires compiling a database of geolocated DEMs before matching can take place. Additionally, the topography must vary sufficiently in a spatial manner so that a scan can be matched to a unique DEM. Environments that are flat or vary with time, such as a beach, are not suitable for DEM matching since incorrect matches provide incorrect global position information.

Celestial Navigation

Before the creation of GNSS, navigation using the sun, moon, and stars was a common practice known as celestial navigation. The threat posed by GNSS jammers and spoofers has revived interest in celestial navigation and motivated the resurgance of terrestrial star trackers. A star tracker consists of a camera pointed at the sky and tuned

Vision Based Navigation

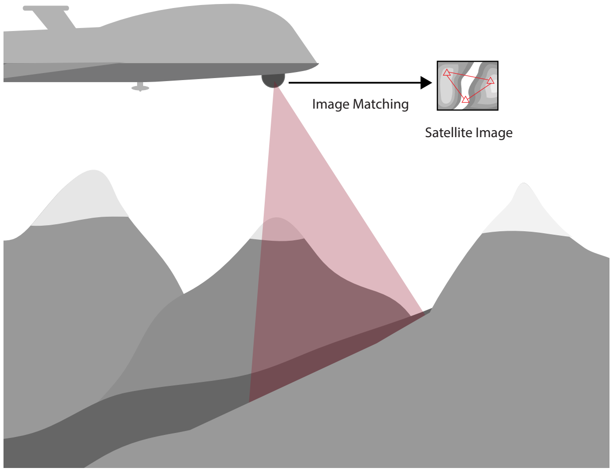


FIGURE 5.6

to observe light from stars. The star tracker maintains a catalog of different constellations or groups of stars that, when seen, can be used to provide global feedback as shown in Figure 5.7.

Celestial Navigation

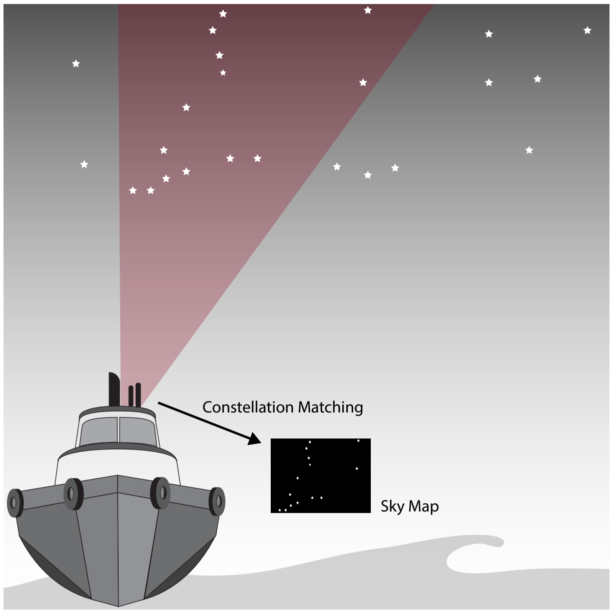


FIGURE 5.7

If the time, pitch, and roll of the star tracker are known, then the latitude and longitude of the star tracker can be estimated. However, the resulting position estimates are extremely sensitive to errors in the time, pitch, and roll estimates. Celestial navigation requires a clear view of the sky above the star tracker in order for constellations to be identified. Unfortunately, this technique cannot be relied upon during precipitating weather or even cloudy conditions.

Magnetic Anomaly Navigation

Researchers believe that many animals, such as salmon and sea turtles, use Earth's magnetic field to return to places they have previously traveled. It is these navigation capabilities of different animals that inspired magnetic anomaly navigation. Magnetic anomaly navigation uses the spatial deviation of Earth's magnetic crustal field, which is one part of Earth's total magnetic field. Other magnetic field components include the core magnetic field, induced fields, and magnetic disturbances due to ferrous materials and electrical currents.

Magnetic anomaly mapping seeks to isolate the crustal field using magnetic models and calibrations to remove the effects of other magnetic sources. When used with a magnetic anomaly map, which provides the expected value of Earth's crustal magnetic field, the location of the magnetometer can be estimated. Magnetic anomaly navigation is typically performed using a scalar magnetometer (rather than vector magnetic field measurements), and is sensitive to any magnetic disturbances around the magnetometer. These disturbances often bias and distort Earth's magnetic field and can be difficult to compensate for. This sensitivity to magnetic disturbances also presents challenges in generating the magnetic anomaly map of an area itself. Additional information on magnetic error sources can be found in Section 3.5.

5.3.2 Radio Frequency Positioning

While GNSS is the leading method of using radio frequency signals for positioning, it is not the only technology that is possible. Similar navigation approaches can be employed with other radio frequency signals, including alternative satellite-based signals as well as terrestrial-based signals.

Low-Earth Orbit (LEO) Satellite Navigation

Recently, low-earth orbit (LEO) satellite navigation has been introduced as an alternative satellite-based navigation technique to GNSS. These LEO satellites differ from GNSS satellites in that they operate much closer to the surface of Earth. Operating closer to Earth's surface means that the signals from LEO satellites have more power than a GNSS signal, allowing them to penetrate occlusions such as foliage and even be accessible indoors. This higher power also makes these signals more resistant to jamming.

Because LEO satellites were not designed explicitly for navigation, their ephemeris data and clock error information are often not publicly available. This means that a LEO receiver must also estimate the position, velocity, and clock errors of each LEO satellite in addition to its own position, velocity, and clock errors. Although a receiver capable of navigating using LEO satellites is not currently available on the commercial market, this navigation technique is an expanding area of development.

Signals of Opportunity

Aside from GNSS, many other signals exist that are widely available and can be used for navigation purposes, such as Wi-Fi, Bluetooth, and cellular signals. The principles of localization using these signals are similar to GNSS and rely on trilateration of distance measurements from the origin of the signal (e.g. the Wi-Fi router or cellular tower) and the receiver. These signals are often available in spaces where GNSS signals are not, such as indoors or in urban canyons, but each requires special hardware and software to track the signals and generate a position solution.

5.4 RELATIVE LOCALIZATION

Relative localization does not provide information about the global position of the system, but instead provides information relative to the surrounding environment. Consequently, the absolute position error of the system is no longer bounded and, given a sufficient amount of time, may increase beyond acceptable limits for a given application. This section highlights a few techniques that can be used to slow the accumulation of errors in the navigation solution. This list is not meant to be comprehensive or an implementation guide but is meant to provide insight into different possible approaches.

5.4.1 Free-Inertial Navigation

Free-inertial navigation, sometimes referred to as dead-reckoning, is a classical technique used to estimate a system's position, velocity, and orientation. In such an approach, an inertial navigation system (INS) integrates acceleration and angular rate measurements from an inertial measurement unit (IMU) to track the change in the position, velocity, and orientation of a system.

Often an INS incorporates absolute localization measurements, such as from GNSS or alternative techniques discussed in Section 5.3, into its advanced Kalman filtering algorithms, allowing for a drift-free navigation solution that is tied to an absolute inertial reference frame. Absent such aiding, an INS can only provide a relative navigation solution of how a system changes over time with respect to some initial starting point. Unfortunately, due to a variety of error sources within the inertial sensors themselves (e.g. bias, random walk, etc.), a free-inertial navigation solution will be subject to drift that accumulates exponentially over time. A more in-depth discussion on inertial sensor errors can be found in Section 3.1 and a detailed error budget for free-inertial navigation is presented in Section 3.3.

5.4.2 Odometry

A variety of technologies can produce an odometry measurement—a measured change in position and/or orientation from one time instant to another. Integrating these odometry measurements produces a navigation solution that drifts as a function of distance traveled. Combining odometry with inertial dead-reckoning can significantly reduce the drift rates of the two individual technologies.

Wheels

Wheeled systems can use wheel encoders to estimate how a system has moved. A wheel encoder has discrete markings at evenly spaced angular distances that the encoder can detect and determine the angle the wheel has turned. If the radii of the wheels are known, the angular distance can be converted into a change in position and heading assuming there is no wheel slip. Wheel odometry can be significantly improved by combining it with a gyro measuring heading rate, which can compensate for errors in parameters like wheel radius and help detect wheel slip.

Unlike free-inertial navigation, the accuracy of wheel odometry is dependent on the environment, not on time. In conditions where wheel slip is prevalent, wheel odometry does not produce accurate estimates of the vehicle's motion because the wheel is spinning but not translating. Additionally, wheel odometry is often limited to planar motion and cannot detect changes in elevation, roll, or pitch.

Visual

Cameras have become a popular sensor to pair with an IMU for GNSS-denied environments due to their small size, weight, and power consumption (SWaP). Computer vision algorithms enable a computer to track features in an image over time across several image frames and estimate how the camera has moved, as shown in Figure 5.8. When used in conjunction with an IMU, this technique is known as visual-inertial odometry. In the absence of a feature map, no information is available about the global position of the camera—only the motion of the camera relative to the features is measured.

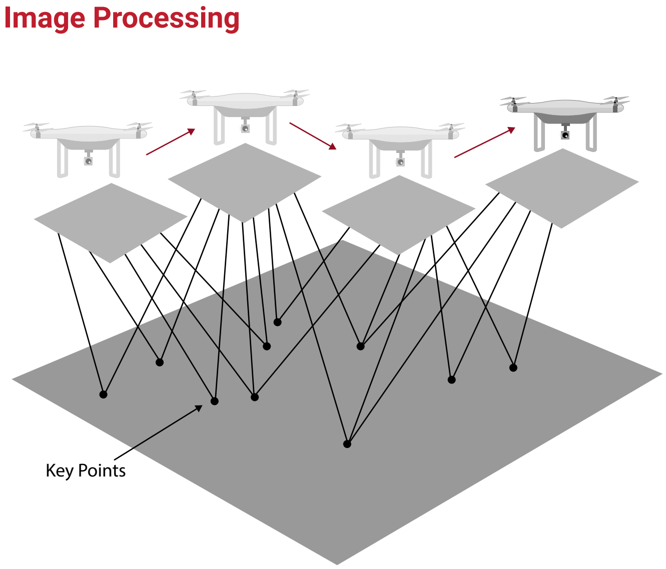


FIGURE 5.8

While visual-inertial odometry is attractive for many systems, not all applications are well-suited to navigate with this technique. To identify and track unique features within the camera images, there must be sufficient textural variation in the surrounding environment. Applications navigating in areas without such variation, including flying over large bodies of water, often lack unique features needed for visual odometry. The lighting conditions can also affect the quality of a visual odometry solution. A dark environment, such as at night, can prove difficult for an RGB camera and might require an IR camera for successful navigation.

LiDAR

Light Detection and Ranging (LiDAR) has become a popular measurement technique for aiding an IMU in GNSS-denied environments, an approach known as LiDAR-inertial odometry. A LiDAR sensor emits light pulses and times the round trip to measure the distance to reflective objects. Typically, tens or hundreds of thousands of measurements are taken per second, providing the capability to digitize an entire scene almost instantly. This produces a point cloud that can be compared with point clouds from previous scans to track features over time, similar to the visual-inertial odometry technique shown in Figure 5.8. Algorithms such as iterative closest point (ICP) are often used to align the point clouds and determine how the sensor has moved between scans.

While LiDAR sensors provide a dense representation of the surrounding environment, they are not well-suited for all applications. Geometrically uniform environments, such as a long, straight hallway, prove difficult for navigation using LiDAR-inertial odometry as often there is not a single unique way to align the point clouds measured at two different points. LiDAR sensors are also often heavier and have higher power consumption than other sensors in addition to having a higher cost, making them impractical for some SWaP-constrained systems.

5.4.3 Velocity Measurements

Certain technologies can be used to measure linear velocity directly, helping reduce integrated position drift. Unlike odometry, these systems typically do not provide any angular rate feedback to stabilize heading, a major error source in long-term position drift.

Airspeed

Aircraft are often equipped with a pitot tube that is capable of measuring both the static and total air pressure. Differencing these two quantities produces the dynamic air pressure or the pressure due to the motion of the vehicle. The dynamic pressure can be converted into an airspeed measurement that describes the speed of the vehicle with respect to the surrounding body of air. This is different from the ground speed, which is the speed of the vehicle with respect to the ground, as the ground is stationary while the surrounding air can be moving. When the surrounding air is still (i.e. no wind) the two quantities will be equal.

Incorporating airspeed measurements with inertial dead-reckoning, often referred to as airspeed aiding, can be useful in GNSS-denied scenarios because of its ability to provide corrections to the velocity estimate and thus slow the growth of error in the position estimates. For these measurements to be effective, an estimate of the velocity of the surrounding air (i.e. wind speed) must be available to calculate the correct ground speed or the resulting estimates will become biased. Since real-time data about the wind speed in any given location is not readily available it must be estimated in real time.

Doppler Velocity Log

Marine vessels are often equipped with an acoustic sensor called a Doppler Velocity Log (DVL) that sends acoustic pulses toward the sea floor and can measure the velocity of the vehicle by observing the frequency shift of the pulse. A DVL is often used with an IMU or INS to maintain an accurate velocity estimate of an underwater vehicle. Unfortunately, having an accurate velocity estimate does not mean that the position estimates stays accurate. Small errors in the velocity combined with heading errors translates into large position errors when integrated over a long period of time.

DVL requires operating within a certain range of the seafloor. This range is dependent on the frequency of the DVL pulse—the lower the frequency the longer the range. However, the size and weight of the DVL generally increases as the frequency decreases, requiring a trade-off between operating depth and the size of the sensor.

5.4.4 Simultaneous Localization and Mapping

Simultaneous localization and mapping (SLAM) is a GNSS-denied navigation technique where a user attempts to create a map of an environment while localizing themselves within that map. Solving SLAM is traditionally quite challenging as creating a map requires position information but obtaining position information often requires a map in the absence of a GNSS signal resulting in a circular dependency.

SLAM-based navigation utilizes an imager that can capture information about the surrounding environment. Such imagers often include cameras, LiDAR, and radar. Similar to the odometry navigation techniques described previously, features are extracted from the imager data and tracked over time to estimate how the system has moved relative to the features. As additional images are collected, a map of the surrounding environment can be constructed. The biggest difference between SLAM and odometry-based navigation techniques occurs when the navigation system revisits a location that it has already traveled. SLAM systems can identify when a location has been revisited and use this information in a technique called closing the loop. Such loop closures enable SLAM systems to remove large amounts of error from their navigation solution but comes at the cost of more computational complexity.

A EXAMPLES

When learning new concepts, some topics are best understood through in-depth examples. This appendix consists of supplementary information to expand upon various concepts covered in previous sections and includes a number of examples to help explain topics presented in this text.

A.1 ATTITUDE TRANSFORMATIONS

In order to better understand the attitude representation and transformation equations defined in Sections 2.3 and 2.4, the example below walks through a simple attitude representation transformation. An orientation of ($\psi = 102°$, $\theta = 20°$, $\phi = 14°$) defined by a (3-2-1) set of Euler angles will be transformed into an attitude represented by the quaternion.

First, the set of Euler angles must be converted into the associated DCM, shown in Equation A.1 (cos ϕ and sin ϕ have been abbreviated cϕ and sϕ, respectively).

$$
\begin{aligned}
C_{(3,2,1)} &= R_1(14°)R_2(20°)R_3(102°) \\
&= \begin{bmatrix} c(20°)c(102°) & c(20°)s(102°) & -s(20°) \\ -c(14°)s(102°)+s(14°)s(20°)c(102°) & c(14°)c(102°)+s(14°)s(20°)s(102°) & s(14°)c(20°) \\ s(14°)s(102°)+c(14°)s(20°)c(102°) & -s(14°)c(102°)+c(14°)s(20°)s(102°) & c(14°)c(20°) \end{bmatrix}
\end{aligned} \tag{A.1}
$$

This results in a DCM of

$$
C_{(3,2,1)} = \begin{bmatrix} -0.1954 & 0.9192 & -0.3420 \\ -0.9663 & -0.1208 & 0.2273 \\ 0.1676 & 0.3749 & 0.9118 \end{bmatrix} \tag{A.2}
$$

Each of the quaternion elements must then be calculated to determine the maximum value of the four quaternion elements.

$$
\begin{aligned}
q_1^2 &= \frac{1}{4}(1 - 0.1954 + 0.1208 - 0.9118) = 0.0034 \\
q_2^2 &= \frac{1}{4}(1 + 0.1954 - 0.1208 - 0.9118) = 0.0407 \\
q_3^2 &= \frac{1}{4}(1 + 0.1954 + 0.1208 + 0.9118) = 0.5570 \\
q_4^2 &= \frac{1}{4}(1 - 0.1954 - 0.1208 + 0.9118) = 0.3989
\end{aligned} \tag{A.3}
$$

As shown in Equation A.3, the q_3 element provides the maximum value of the four quaternion elements. Therefore, the formula corresponding to q_3 will be used to compute the remaining quaternion terms.

$$
\boldsymbol{q} = \frac{1}{4q_3}\begin{bmatrix} C_{31} + C_{13} \\ C_{23} + C_{32} \\ 4q_3^2 \\ C_{12} - C_{21} \end{bmatrix} = \frac{1}{4(0.7463)}\begin{bmatrix} 0.1676 + (-0.3420) \\ 0.2273 + 0.3749 \\ 4(0.5570) \\ 0.9192 - (-0.9663) \end{bmatrix} = \begin{bmatrix} -0.0584 \\ 0.2017 \\ 0.7463 \\ 0.6316 \end{bmatrix} \tag{A.4}
$$

A.2 NOISE CALCULATIONS

Since noise can be specified in a variety of different ways, it is important to understand how to convert each of these different specifications into a common form that can be used to accurately compare different sensors. The following examples work through a couple of these conversions.

Noise Density to Standard Deviation

To determine the noise standard deviation (σ) at a specified sample rate (SR) from a noise density (ND), the square root of the sampling rate should be multiplied by the noise density, as shown in Equation A.5

$$\sigma = \mathrm{ND}\sqrt{\mathrm{SR}} \tag{A.5}$$

For example, consider a noise density of 0.01 $^\circ/\mathrm{s}/\sqrt{\mathrm{Hz}}$ which needs to be converted into a standard deviation at a sampling rate of 100 Hz:

$$\begin{aligned} \sigma &= \left(0.01\,^\circ/\mathrm{s}/\sqrt{\mathrm{Hz}}\right)\sqrt{100\,\mathrm{Hz}} \\ &= 0.1\,^\circ/\mathrm{s} \end{aligned} \tag{A.6}$$

Random Walk to Standard Deviation

As shown in Equation A.7, the standard deviation (σ) of the drift due to noise can be found from the random walk (RW) by multiplying it by the square root of time (t).

$$\sigma = \mathrm{RW}\sqrt{t} \tag{A.7}$$

Suppose an angle random walk of 0.03 $^\circ/\sqrt{\mathrm{hr}}$ for a gyroscope, which needs to be converted into a standard deviation for a time of 100 s:

$$\begin{aligned} \sigma &= \left(0.03\,^\circ/\sqrt{\mathrm{hr}}\right)\left(\frac{1\sqrt{\mathrm{hr}}}{60\sqrt{\mathrm{s}}}\right)\sqrt{100\,\mathrm{s}} \\ &= 0.005^\circ \end{aligned} \tag{A.8}$$

A.3 LEAST SQUARES

To better understand the linear least squares estimation process described in Section 2.7, consider the collection of data points plotted in Figure A.1. Each of these points can be written as a linear function using Equation A.9.

$$z = mt + b \tag{A.9}$$

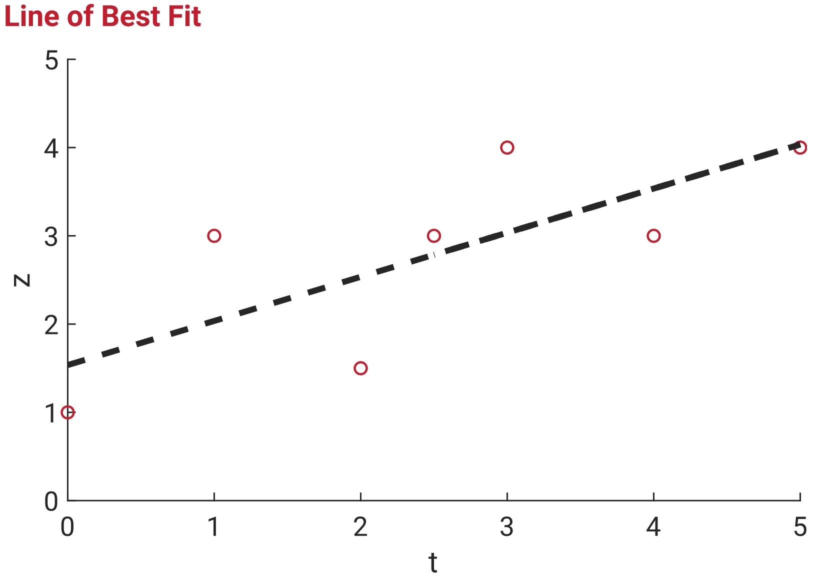


FIGURE A.1

This will form the system of equations shown in Equation A.10.

$$\begin{cases} 1 & = 0 + b \\ 1.5 & = 2m + b \\ 4 & = 3m + b \\ 3 & = 4m + b \\ 3 & = 2.5m + b \\ 4 & = 5m + b \\ 3 & = m + b \end{cases} \tag{A.10}$$

As seen in Equation A.11, these equations can also be written into an equivalent form using vectors and matrices.

$$\begin{bmatrix} 1 \\ 1.5 \\ 4 \\ 3 \\ 3 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 2.5 & 1 \\ 5 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} \tag{A.11}$$

Though a solution cannot be found to solve this system of equations, the linear least squares estimation technique can be used to estimate a line of best fit for this data. Equation A.11 follows the same form as Equation A.12,

$$\tilde{\boldsymbol{y}} = H\hat{\boldsymbol{x}} \tag{A.12}$$

where

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} 1 \\ 1.5 \\ 4 \\ 3 \\ 3 \\ 4 \\ 3 \end{bmatrix}, \; H = \begin{bmatrix} 0 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 2.5 & 1 \\ 5 & 1 \\ 1 & 1 \end{bmatrix}, \; \hat{\boldsymbol{x}} = \begin{bmatrix} m \\ b \end{bmatrix}$$

These matrices can then be used in the linear least squares solution to solve for the optimal slope and z-intercept of the line of best fit.

$$\begin{aligned} \hat{\boldsymbol{x}} &= (H^{\intercal}H)^{-1}H^{\intercal}\tilde{\boldsymbol{y}} \\ &= \left(\begin{bmatrix} 0 & 2 & 3 & 4 & 2.5 & 5 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 2.5 & 1 \\ 5 & 1 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 & 2 & 3 & 4 & 2.5 & 5 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1.5 \\ 4 \\ 3 \\ 3 \\ 4 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix} \end{aligned}$$

As shown in Figure A.1, the line of best fit that minimizes the residual errors for this collection of data is given by: $z = 0.5t + 1.5$

A.4 MIL-STD-461G

In an effort to define safe operating levels for military use, MIL-STD-461 was created to give a standard that devices be tested to so that even commercial-off-the-shelf parts could be designed and used in military applications. The standard lays out testing setups and requirements for a variety of EMI issues that may occur in military applications. In the event that a device does not pass the standard, the companies involved may agree to “agreed upon” exemptions when a signal does not seem as if it will be problematic in a given situation.

A.4.1 Test List

CE101 Conducted Emissions, Audio Frequency Currents, Power Leads

This test is to ensure that the equipment under test (EUT) outputs acceptable noise levels on the power leads in the 30 Hz to 10 kHz range. This is important because certain devices such as acoustic receivers or magnetic anomaly detectors may be sensitive to noise in this range.

CE102 Conducted Emissions, Radio Frequency Potentials, Power Leads

Similar to CE101, this test looks for power lead noise in the 10 kHz to 10 MHz range.

CE106 Conducted Emissions, Antenna Port

This test occurs on antenna ports of devices that will be transmitting, receiving, or amplifying. The frequency range involved depends on the highest frequency generated or received by the system, but could include anywhere from 10 kHz to 40 GHz.

CS101 Conducted Susceptibility, Power Leads

This test verifies the EUT's ability to operate while subject to signals coupled directly into the input power leads. The device must not exhibit any malfunction or deviation beyond the tolerances listed in the device specification.

CS103 Conducted Susceptibility, Antenna Port, Intermodulation

This test checks for intermodulation products from injecting noise in the 15 kHz to 10 GHz range on antenna ports.

CS104 Conducted Susceptibility, Antenna Port, Rejection of Undesired Signals

Similar to CS103, this test looks for spurious responses while injecting 30 Hz to 20 GHz noise onto the antenna port.

CS105 Conducted Susceptibility, Antenna Port, Cross-Modulation

CS105 looks for cross modulation on antenna ports of receivers that normally process amplitude-modulated RF signals.

CS109 Conducted Susceptibility, Structure Current

This test looks at the effect of structure currents on the EUT. It applies to EUTs with an operating frequency of 100 kHz or less, and an operating sensitivity of 1 uV or better.

CS114 Conducted Susceptibility, Bulk Cable Injection

This test is to verify EUT performance with 10 kHz- 200 MHz RF signals injected into each cable going to the device. The signal strength is adjusted depending on the actual current induced in the cable.

CS115 Conducted Susceptibility, Bulk Cable Injection, Impulse Excitation

Similar to CS114 but dealing with impulse signals (5 A for 30 ns) occurring at 30 Hz for a duration of 1 minute.

CS116 Conducted Susceptibility, Damped Sinusoidal Transients, Cables and Power Leads

This test looks at the results of applying damped sinusoidal transients at frequencies of 0.01, 0.1, 1, 10, 30, and 100 MHz coupled into all cables and power leads. The pulse frequency is between 0.5-1Hz and repeats for 5 minutes. Peak current during the test is 10 A, with less current required at frequencies over 30 MHz and under 1 MHz.

CS117 Conducted Susceptibility, Lightning Induced Transients, Cables and Power Leads

This test verifies EUT susceptibility to transients typically associated with lightning. Depending on which waveform is tested, the voltage used may be as high as 1500 V for an externally located device.

CS118 Conducted Susceptibility, Personnel Borne Electrostatic Discharge

This test applies to devices with a human-machine interface and involves applying up to +/-15 kV ESD discharges to areas of the device that would likely come into contact during normal use. The device must also be powered during testing.

RE101 Radiated Emissions, Magnetic Field RE102 Radiated Emissions, Electric Field

This test looks at the magnetic field emissions in the 30 Hz to 100 kHz range and includes cables but not antenna radiation. A loop sensor is used to measure the magnetic field.

RE102 Radiated Emissions, Electric Field

RE102 testing depends on the application environment the device will be used in. The emission range could be from 10 kHz to 18 GHz depending on if the environment will be stationary on the ground, internally or externally on a ship or aircraft, or in space. Different antennas are required to detect the emissions depending on frequency range.

RE103 Radiated Emissions, Antenna Spurious and Harmonic Outputs

RE103 may be used as an alternative to CE106, and is the preferred method when an active antenna is used. The test looks for harmonics of the device fundamental frequency in the frequency range of 10 kHz to 40 GHz, dependent on the highest frequency present.

RS101 Radiated Susceptibility, Magnetic Field

RS101 testing involves subjecting the device to external magnetic fields using a radiating loop in the 10 Hz to 100 kHz range.

RS103 Radiated Susceptibility, Electric Field

RS103 tests how the EUT performs while subjected to radiation having frequencies ranging from 2MHz to 18 GHz, with up to 40 GHz as an optional test. Some frequency ranges require vertical and horizontal antenna rotations to account for field polarization.

RS105 Radiated Susceptibility, Transient Electromagnetic Field

RS105 applies to EUTs that are installed in an external electromagnetic environment. The test uses a transient pulse generator to subject the device to a transient EM field. The transient involves a 50,000 V/m pulse lasting 100 nanoseconds.

This page intentionally left blank.

VectorNav Technologies, LLC
10501 Markison Rd
Dallas, TX 75238, USA

Tel: +1.512.772.3615
Email: sales@vectornav.com
Web: vectornav.com

Version 19-1000-R3