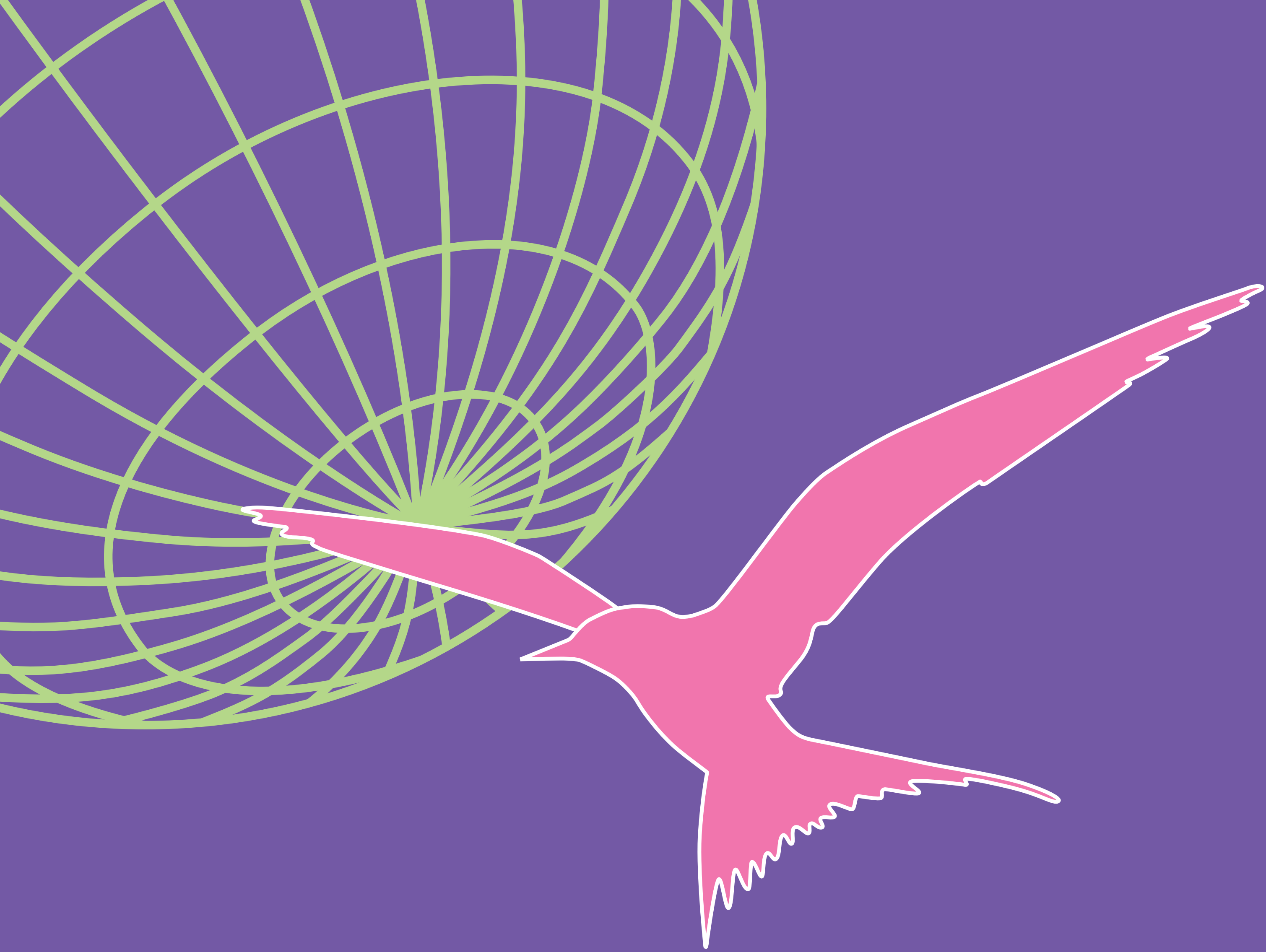


METHODS OF NAVIGATION

An introduction to technological navigation

Martin Vermeer

Aalto University

Aalto University publication series
SCIENCE + TECHNOLOGY 6/2020

Methods of navigation

An introduction to technological navigation

Martin Vermeer

Aalto University
School of Engineering
Department of Built Environment

Aalto University publication series
SCIENCE + TECHNOLOGY 6/2020

© 2020 Martin Vermeer

ISBN 978-952-64-0208-6 (pdf)
ISSN 1799-490X (pdf)
http://urn.fi/URN:ISBN:978-952-64-0208-6

Helsinki 2020

Finland

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Martin Vermeer

Name of the publication
Methods of navigation – An introduction to technological navigation

Publisher School of Engineering

Unit Department of Built Environment

Series Aalto University publication series SCIENCE + TECHNOLOGY 6/2020

Field of research Geodesy

Language English

Abstract

Historically, humankind has always navigated. Technological navigation originated in seafaring, because on the open ocean, measurements are needed in order to determine one's own location as a part of navigation.

Aircraft, rockets and spacecraft as well as vehicles moving on dry land, and even pedestrians, all "navigate" by means of modern technologies. This development is mainly due to two technologies: satellite positioning, such as GPS (the Global Positioning System) and inertial navigation. Also information and communication technologiy has evolved: especially recursive linear filtering or the Kalman filter. Furthermore, small and inexpensive digital sensors are revolutionising everyday navigation.

Subjects explained in this book are the fundamentals of navigation, stochastic processes, the Kalman filter, inertial navigation technology and methods, GNSS signal structure, carrier-phase measurement and ambiguities, real-time GNSS positioning and navigation, communication solutions and standards for differential corrections, GNSS base stations and networks, satellite-based augmentation systems, airborne gravimetry, sensor fusion and sensors of opportunity.

Keywords navigation, technological navigation, seafaring, aviation, rockets, missiles, aircraft, spacecraft, satellite positioning, inertial navigation, stochastic processes, Kalman filter, global navigation satellite systems, carrier phase, ambiguities, pseudo-random codes, real-time kinematic GNSS positioning, differential GNSS positioning, satellite-based augmentation systems, airborne gravimetry, sensor fusion, sensors of opportunity.

ISBN (pdf) 978-952-64-0208-6

ISSN (PDF) 1799-490X

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 370 **urn** http://urn.fi/URN:ISBN:978-952-64-0208-6

Preface

Cover picture: Navigation is not a human invention. The Arctic tern (*Sterna paradisaea*) flies every year from the Arctic area to the Southern Ocean and back.

Egevang et al. (2010)

This course was taught by the author during 2001–2010 at Helsinki University of Technology and during 2010–2015 at Aalto University as part of the degree programmes in positioning and navigation, geomatics, and geoinformatics. The idea behind the course was to give students of the geospatial sciences a basic understanding of the technologies and methods underlying real-time positioning and its uses for navigation, on land and sea, in the air and in space.

The main subjects taught are the fundamentals of navigation, stochastic processes, the Kalman filter, inertial navigation technology and methods, GNSS signal structure, carrier-phase measurement and ambiguities, real-time GNSS positioning and navigation, communication solutions and standards for differential corrections, GNSS base stations and networks, satellite-based augmentation systems, airborne gravimetry, sensor fusion and sensors of opportunity.

Helsinki, 4th December, 2020,

Martin Vermeer

Acknowledgements

Susanna Nordsten drafted an early English translation of the manuscript in 2002, using a prototype version of the "branches" facility of LyX, described here: multilingual.pdf. Keijo Inkilä pointed out the Woodbury identity.

The English language was competently checked by Finnish Translation Agency Aakkosto Oy. The cover was designed by Hanna Sario from Unigrafia Creative. Laura Mure and Henri Linnanketo helped with the practicalities of publishing.

Useful remarks from students Megersa Mekonen and Gedamu Amare from Addis Ababa University were helpful in improving the exposition.

Several map images were drawn using Generic Mapping Tools (Wessel et al., 2013).

This content is licensed under the *Creative Commons Attribution 4.0 International* (CC BY 4.0) licence, except as noted in the text or otherwise apparent.

Contents

Chapters

List of Tables

List of Figures

Acronyms

ABCDEFGHIJKLMNOPQRSTUVWXZ

A

AFSCN Air Force Satellite Control Network (GPS) 166

A-GNSS assisted GNSS 296

APPS Automatic Precise Positioning Service 257

AR(1) first-order autoregressive process 41

ARAIM advanced RAIM, uses more than one GNSS on frequencies L1 and L5 for even stronger integrity 244, 267

B

BDSBAS BeiDou Satellite-Based Augmentation System, China's SBAS under development 241

BeiDou BeiDou Navigation Satellite System (BDS), Chinese global navigation satellite system 14, 171, 260, 265, 268–270

BFSK binary frequency-shift keying 315

BKG *Bundesamt für Kartographie und Geodäsie*, Federal Agency for Cartography and Geodesy 217, 219

BOC binary offset carrier 170, 184–187, 196, 261, 265, 267, 269

BPSK binary phase-shift keying 168, 170, 184, 185, 261, 265, 267, 269

C

CDMA code-division multiple access 170, 187, 255, 260, 264, 265, 270

CHAMP Challenging Minisatellite Payload 279

CPU central processing unit 323

CRC cyclic redundancy check 315

CSM Command and Service Module (Apollo) 52

D

Decca maritime hyperbolic radionavigation system (obsolete) 5, 296

DGPS differential GPS 217, 218

DLL delay-locked loop 188

DNS Domain Name System 317

DOP dilution of precision (GNSS) 202, 243

DSL digital subscriber line 315

E

ECEF Earth centred, Earth fixed 10

EDAS EGNOS Data Access Service 258

EDGE Enhanced Data Rates for GSM Evolution 317

EGM2008 Earth Gravity Model 2008 120

EGNOS European Geostationary Navigation Overlay Service, an SBAS for the European area xiv, xvi, xvii, 241, 249, 250, 252, 258, 266, 267, 270

EOP Earth orientation parameters 208

ESA European Space Agency 250, 255, 265

ETRF European Terrestrial Reference Frame, usually with a year number. A realisation of ETRS 10

ETRS European Terrestrial Reference System, defined as co-moving with the Eurasian tectonic plate xiv, 10

EUREF IAG Regional Reference Frame Sub-Commission for Europe 10

F

FDMA frequency-division multiple access 262, 270

FFT fast Fourier transform 36, 313

FKP *Flächenkorrekturparameter*, Areal Correction Parameters, network RTK technique 236

FRS Fellow of the Royal Society (of London) 37, 104

G

GAGAN GPS-Aided Geo Augmented Navigation, an SBAS for the Indian area xv, 241, 256, 270

Galileo European global navigation satellite system xv, xvi, 14, 171, 184–186, 244, 260–262, 265–267, 269, 270

GBAS ground-based augmentation system 14, 256–258

GDGPS Global Differential GPS 216, 257

GIA glacial isostatic adjustment 208

GJU Galileo Joint Undertaking 266

GLONASS Russian, globally operating navigation satellite system xvi, 14, 171, 179, 218, 260–265, 270

GMS Ground Monitor Station (MSAS) 254

GNSS global navigation satellite systems, generic name xiii–xv, xvii, 6, 11–13, 16, 139, 191, 197, 202, 204–215, 217–220, 225, 229, 235, 237–239, 243, 244, 246, 255, 259, 261, 264, 266, 271, 272, 276, 278, 279, 282, 284, 292, 293, 296, 298–300, 316

GOCE Gravity field and steady-state Ocean Circulation Explorer 279, 280

GPRS General Packet Radio Services 317

GPS Global Positioning System xiii, xvi, 1, 6, 9, 10, 13, 14, 141, 165–172, 174–177, 179, 181–185, 187, 188, 191, 193, 195, 196, 203, 204, 217, 218, 238, 241, 244–249, 253, 256, 258–267, 269, 270, 296, 319

GRACE Gravity Recovery and Climate Experiment 279, 280

GRS80 Geodetic Reference System 1980 288

GSA European GNSS Agency, earlier European GNSS Supervisory Authority 266

GSM Global System for Mobile Communications xiv, 317

GUI graphical user interface 320

GUS Ground Uplink Station (WAAS) 249–251

H

HAS high-accuracy service (Galileo) 267

HTTP Hypertext Transfer Protocol 219, 316, 318

I

IAG International Association of Geodesy xiv, 9, 10

ICAO International Civil Aviation Organization 247, 248, 256

ICD Interface Control Document 262, 264, 268, 269

ICMP Internet Control Message Protocol 317

IERS International Earth Rotation and Reference Systems Service 9

IGS International GNSS Service 212

IGSO inclined geostationary orbit 260, 270

i.i.d. independent and identically distributed 28, 49, 201, 243

IMU inertial measurement unit 103, 191, 291, 292, 298

INLUS Indian Navigation Link Upload Station (GAGAN) 256, 257

INMCC Indian Mission Control Centre (GAGAN) 256, 257

INRES Indian Reference Station (GAGAN) 256, 257

IOD issue of data, broadcast ephemeris time stamp 246

IP Internet Protocol 315, 317

ITRF International Terrestrial Reference Frame, usually with a year number, a realisation of the ITRS 9, 10, 264

ITRS International Terrestrial Reference System xv, 10

ITU International Telecommunications Union 315

J

J_2 The dynamic flattening (“gravitational flattening”) of the Earth 161

JATO jet-assisted take-off 277, 280

JAXA Japanese Aerospace Exploration Agency 253

JPL Jet Propulsion Laboratory 257

K

KASS Korea Augmentation Satellite System, under development 241

KKJ National Map Grid Co-ordinate System (obsolete) 11

L

LAAS local-area augmentation system 256

LAMBDA Least-squares Ambiguity Decorrelation Adjustment 226, 228, 229, 239

LFSR linear feedback shift register 177

LIFO last in, first out data structure 320, 321

LOR lunar orbit rendezvous 51

M

MAC Master-Auxiliary Concept, network RTK technique 207, 236

MAX Master-Auxiliary Corrections, network RTK technique 207, 236, 237, 239

MCC Master Control Centre (EGNOS) 250, 252

MCS Master Control Station (GPS, MSAS) 166, 254

MEMS microelectronic motion sensor 111, 112, 296, 298, 300

MSAS Multi-functional Satellite Augmentation System xiv, xvi, 241, 250, 253, 254, 270

MSM Multiple Signal Message (RTCM-SC104 v. 3) 217

N

NASA National Aeronautics and Space Administration (USA) 115

NLES Navigation Land Earth Station (EGNOS) 250, 252

NTP Network Time Protocol 318

NTRIP Networked Transport of RTCM via Internet Protocol 219, 220, 258

NWP numerical weather prediction 239

O

OS open service (Galileo) 267, 319, 320

P

PFA probability of false alarm (RAIM) 243

PLL phase-locked loop 189

PMD probability of missed detection (RAIM) 243

PPP precise point positioning 177

PRN pseudo-random noise 170, 216, 245, 248–250, 253, 255, 299, 307

PRS President of the Royal Society 101

PRS public regulated service (Galileo) 267

PSD power spectral density 44–47, 184, 187

PZ-90 Parameters of the Earth 1990, GLONASS official reference frame 263, 264

Q

QZSS Quasi-Zenith Satellite System, a Japanese SBAS 171, 241, 253–255, 258, 270

R

RAIM receiver autonomous integrity monitoring xiii, xvi, 14, 241–244, 248, 258, 267

RDS Radio Data System 235

RIMS Ranging and Integrity Monitoring Station (EGNOS) 250, 252

RINEX Receiver Independent Exchange Format 258

RTCM RTCM-SC104: Radio Technical Commission for Maritime Services Special Committee 104, a set of standards for differential GNSS xvi, xvii, 216–220, 236, 237, 258

RTK real-time kinematic positioning xiv, xvi, xviii, 11, 13, 197, 207, 214, 217, 218, 235, 238, 239

S

SAR search and rescue 267

SBAS satellite-based augmentation system xiii, xiv, xvii, xviii, 14, 207, 217, 241, 244–247, 250, 253, 255–258, 261, 266, 270

SDCM System for Differential Corrections and Monitoring, Russian SBAS under development 241

SISNeT Signal in Space through the Internet, an application that makes the EGNOS signal available over the Internet 258

SoL safety of life 247, 261, 267

SSB single sideband modulation 312

T

TCP Transmission Control Protocol 257, 316, 318

TDM time-division multiplexing 265

TDMA time-division multiple access 296

TDOA time difference of arrival (positioning) 296

TEC total electron content, unit TECU 203, 256

TECU total electron content unit, 10^{16} electrons per m^2 xvii, 203

TOA time of arrival (positioning) 296, 300

U

UAV unmanned aerial vehicle, “drone” 297, 319

UDP User Datagram Protocol 317, 318

USB Universal Serial Bus 318

UTC Universal Time Co-ordinated 142, 263

V

V-2 (*Vergeltungswaffe 2*, "Retaliation Weapon 2"). German medium-range guided ballistic missile. Also A4 ("*Aggregat 4*") 6, 7, 112

VHF Very High Frequency, 30 – 300 MHz xviii, 241, 256

VOR VHF Omnidirectional Range, aviation navigation beacon 241, 256

VRS virtual reference station, network RTK technique 235, 239

W

WAAS Wide Area Augmentation System, an SBAS for the North American area xv, xviii, 241, 244, 248–251, 258, 270

WGS84 World Geodetic System 1984, a set of global reference frames created and maintained by the US Department of Defense 9, 10, 263

WLAN wireless local-area network 281, 295, 299

WMS Wide Area Master Station (WAAS) 249, 251

WRS Wide Area Reference Station (WAAS) 249, 251

X

XOR exclusive "or" operation 167, 170, 178–181, 304–306

Z

ZTD zenith total delay 237

ZUPT zero-velocity update 289

Fundamentals of navigation

1

1.1 Introduction

“Navigation” originates from the Latin word *navis*, ship. In other words, navigation is seafaring. A broader understanding of navigation is: finding and following a suitable route, recursively when appropriate. This includes determining one’s own location during the journey. And today’s navigation, at least outside our everyday personal sphere, is invariably technological.

Navigation is related to geodesy, because *location* is also a theme in geodesy. However in geodesy, the positions of objects are usually treated as constants or as very slowly changing.

So, the differences between navigation and traditional geodetic positioning are:

- In navigation, the location data is needed *immediately* or at most after a certain maximum delay. This is called the *real-time* requirement. tosiaikaisuus
- In navigation, the location data are *variable*, time-dependent.

Modern navigation is not limited to seafaring. Aeroplanes, missiles, and spacecraft as well as vehicles that move on dry land, and even pedestrians, often “navigate” with the aid of modern technology. This is mainly due to two technologies: satellite positioning, such as GPS (the Global Positioning System), and inertial navigation. In addition, information and communication technologies have developed: the recursive linear filter or Kalman filter should be mentioned in particular. Finally, sensor technologies have produced a host of small and inexpensive digital sensors that are revolutionising everyday navigation.

FIGURE 1.1. Life is navigation. Votive ship in the Admiralty Church of Karlskrona, Sweden. (Wikimedia Commons, Votive offering, cropped). The model is of the corvette Carlskrona launched in 1841. She capsized and sank in a squall off the coast of Cuba in 1846, taking with her 114 of her crew of 131.

1.2 History

1.2.1 Old history

Humans have always been discovering the world around them, often travelling long distances. Navigation has always been a necessity.[1]

Before the invention of technological methods of measurement and guidance, one was dependent on landmarks and distances estimated from travel time. This is why old maps drawn on the basis of travellers' tales and notes are often distorted in weird ways.

Using landmarks this way requires *mapping*: constructing a description of the world in the form of a map. The journey is then *planned* and executed by constantly comparing the actual place with the intended destination according to the travel plan.

Navigation with the help of landmarks and high technology is used by for example *cruise missiles*: they fly by the height contours of a digital terrain model they have stored in their memories.

If, for example in seafaring, landmarks are lacking, one can use

[1] "*Navigare necesse est*".

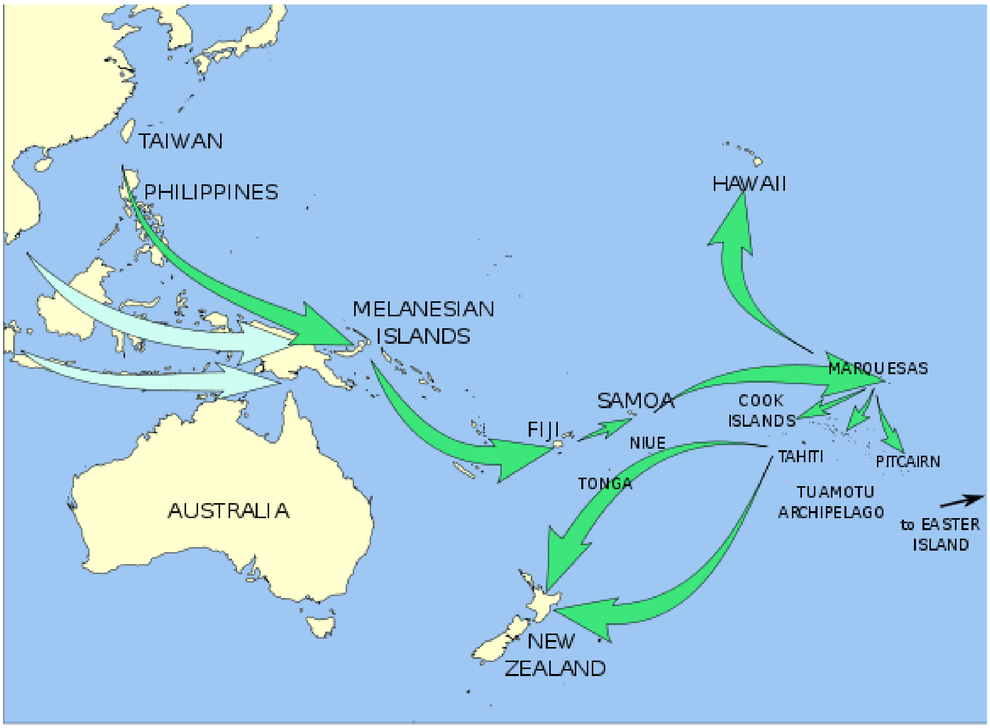


FIGURE 1.2. Polynesian migration routes, Wikimedia Commons, Polynesian migration.

a method called *dead reckoning*, Wikipedia, Dead reckoning. In this method one estimates where one *should* be based on travel direction and speed. Sources of error in dead reckoning are sea currents — in aviation, winds — and more generally the fact that the forecast weakens with time. merkintälasku

With these primitive methods, seafaring is somewhat safe only near the coast. However, this is the way in which the Phoenicians are believed to have already travelled around the continent of Africa, Sinjab (2010), and the archipelagos of the Pacific Ocean gained their human settlements. Wikipedia, Polynesian navigation; Kawaharada; Exploratorium, Never Lost.

See also Diamond (1999).

And, of course, *birds* have always navigated, Lindsay (2006).

1.2.2 Navigation at sea

Seafaring on the open ocean presupposes *measurement*, because there are no landmarks.[2]

2

FIGURE 1.3. Barnacle geese in autumn migration, Wikimedia Commons, Barnacle geese.

Pohjantähti

- Direction is the easiest. At night, the North Star (Polaris) shows the direction of north. In the daytime the Sun can be used, although in a more complicated way. On a cloudy day, the polarisation of

FIGURE 1.4. John Harrison's chronometer H5. Wikimedia Commons, Harrison's chronometer H5.

[2]At least no obvious ones. Some have wondered how Polynesian seafarers managed to find relatively tiny archipelagos like Hawaii, failing to grasp that the islands' area of influence on clouds, sea currents and birdlife is quite a bit larger than just the real estate sticking out of the water — for those with eyes to see.

the light of the sky can be used to help locate the Sun.

The magnetic compass made finding north easier under all conditions. However, the magnetic north is not the geographic north, and the difference between them, magnetic declination, depends on location and changes with time. eranto

- Latitude is also easy to obtain: it is the elevation angle of the celestial pole above the horizon. In the daytime the Sun may be used: at the upper culmination or solar noon, the elevation η of the Sun above the horizon may be observed. One also needs the solar declination or celestial latitude δ, the angular distance of the Sun from the celestial equator, as given by astronomical tables. The latitude φ may now be calculated as tähtitieteellinen keskipäivä eta ηH delta $\delta\Delta$ phi $\varphi\phi\Phi$

$$\varphi = \delta \overset{\text{north}}{\underset{\text{south}}{\pm}} (90^\circ - \eta).$$

  Here, the plus sign applies when the elevation of the Sun over the southern horizon is observed — usually in the northern hemisphere — whereas the minus sign applies when the Sun is due north, usually in the southern hemisphere.

- Longitude is a problem because of the rotation of the Earth. This means that the orientation of the Earth relative to the Sun and stars changes rapidly with the time of day. Using the Sun or stars for longitude determination demands knowledge of this orientation, which in turn requires knowing the absolute time using an accurate time standard, or *chronometer*. See Sobel (1995). Astronomical methods like using the moons of Jupiter as a “clock” have also been studied, starting with Galileo (Koberlein, 2016). In the 20th century the dissemination of time signals by radio became common.

In the 20th century, radio technological positioning methods came also into use. The most well-known is probably Decca, which was based on hyperbolic positioning.

One “master” station and two or more “slave” or auxiliary stations transmit radio waves whose phase angles, serving as time signals, are synchronised. The on-board receiver measures the travel-time difference between the waves received from master and auxiliary. On the nautical chart is marked the set of points having the same difference in travel time as a coloured curve, a *hyperbola*. Every auxiliary station forms with the master a bundle of hyperbolas drawn in its own colour.

The intersection point of two different-coloured hyperbolas gives the position of the ship. So, at least two auxiliaries are needed in addition to the master station.

Modern satellite positioning methods, like Transit/NNSS (no longer in use) and GPS and other global navigation satellite systems, are based on a three-dimensional version of the hyperbolic method.

1.2.3 The modern era

Aviation and space research have brought with them the need for automated three-dimensional navigation. Although the first aeroplanes could be flown by hand without any instruments, the first modern missile, the German V-2, already included a gyroscope-based control system. In this case, navigation is *guidance*.

The guidance system of the V-2 was primitive. The missile was launched vertically into the air, where it turned to the desired direction with the help of its gyroscope platform. The missile accelerated until it reached a pre-determined velocity, at which point the propellant supply was shut off ("*Brennschluss*"). Physically the steering was done with the aid of small "air and jet rudders" ("*Luft- und Strahlruder*") connected to the tail, that changed the direction of the hot gases coming from the
3 engine.[3]

Nowadays complete inertial navigation is used in aeroplanes and spacecraft, as are other computer-based technologies such as satellite positioning by GNSS, global navigation satellite systems.

1.3 Vehicle movements and co-ordinate frames

vertauskehys A moving vehicle has several co-ordinate reference frames relevant to it, see figure 1.6:

1. the body frame: x' pointing in the direction of motion, y' pointing sideways to port, and z' pointing roughly up.

---

[3]See Wikipedia, V-2 rocket. In fact these were dual rudders: the parts sticking into the hot gas stream from the engine, the "jet rudders", were made of graphite and burned up quickly. But by then, the rocket was up to speed and the external "air rudders" took control.
Today's rockets use gimballed, hydraulically actuated engines for precise thrust-vector control.

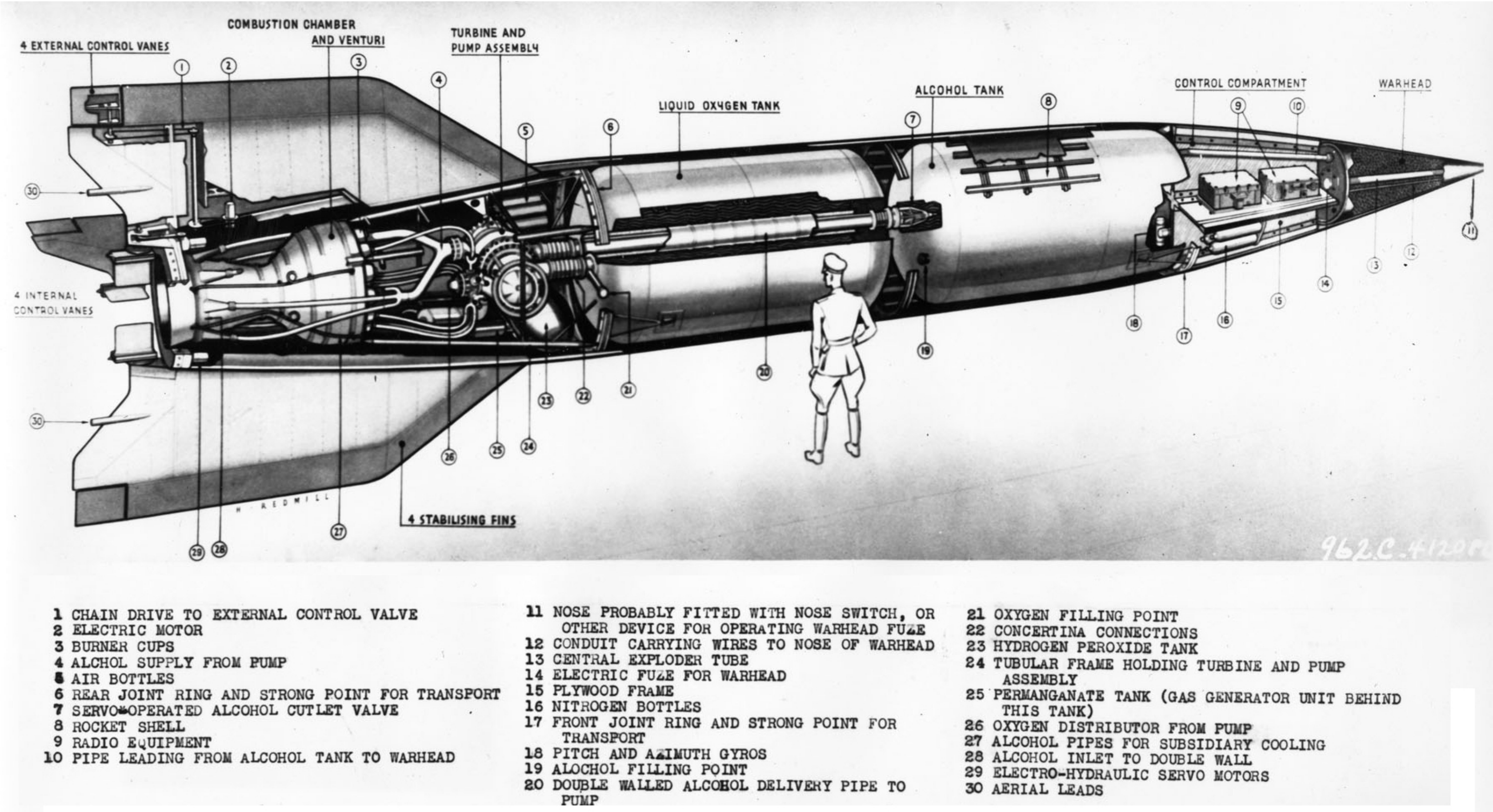


FIGURE 1.5. German V-2 rocket weapon. Image US Air Force.

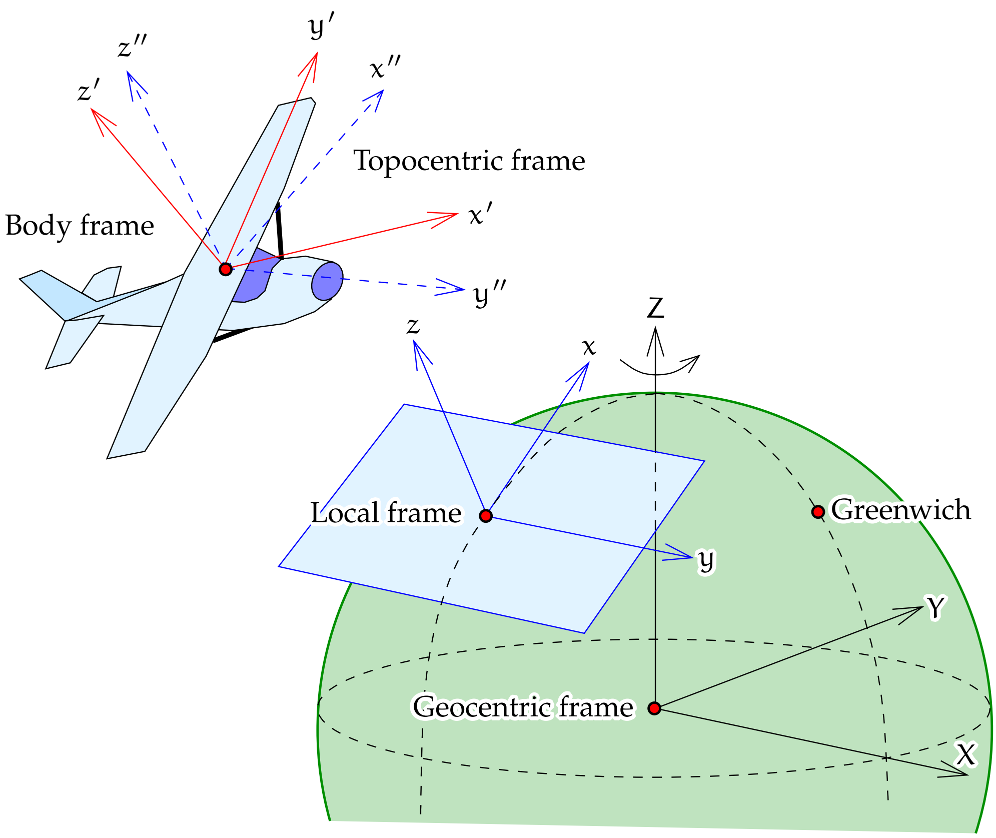


Figure 1.6. Various co-ordinate frames in navigation.

2. the topocentric frame, also north-east-up: the x'' axis pointing north (in geodesy) or east (in photogrammetry), the z'' axis pointing up along the local plumb line, and the y'' axis perpendicular to both, pointing either north or east.

luotiviiva

3. a local or regional terrestrial frame, with x and y being map-projection co-ordinates and z the height defined in a local height system from an agreed reference surface along the local plumb line.

   This is a quasi-Cartesian reference frame often used in aerial mapping.

vertauspinta

4. a geocentric reference frame (X, Y, Z), see section 1.4.

Between the body frame and each of these external frames exists a transformation characterised by three *shift* or *translation* parameters and three *rotation* parameters, the Euler angles. For a moving vehicle, all six are continuous functions of time, as are their first derivatives of time, known as *velocities* and *rotation rates*.

The *attitude* of a vehicle can be described relative to three axes. The

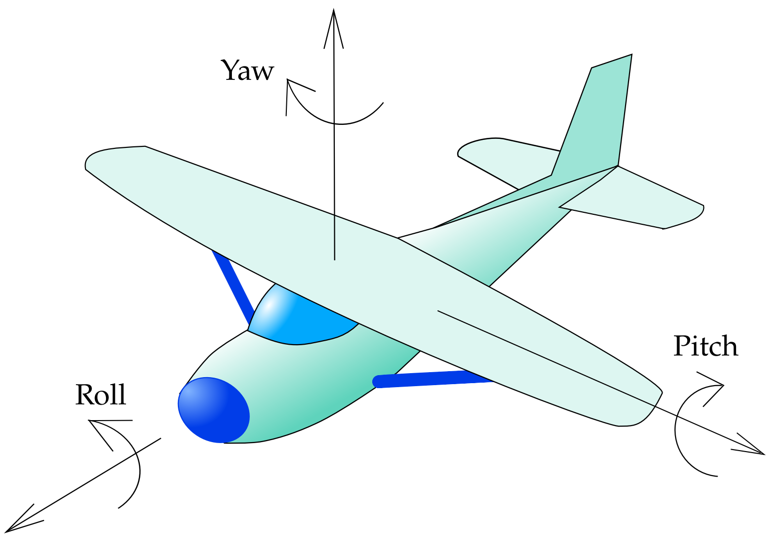


FIGURE 1.7. Attitude angles of a vehicle.

motion about the direction of travel is called *roll*, that about the vertical axis *yaw*, and that about the horizontal (left-right) axis *pitch*. We also use the term *Euler angles*, for example in photogrammetry, $(\omega, \varphi, \kappa)$, which are however slightly differently defined.

omega $\omega\Omega$
kappa κK

1.4 Geocentric reference frames

The *geocentricity* of a reference frame means that

- The origin is in the centre of mass of the Earth or very close to it.
- The Z axis points in the direction of the Earth's rotation axis.

Furthermore, for a co-rotating reference frame

- The X axis lies in the plane of the Greenwich meridian and points to the intersection of equatorial and Greenwich-meridian planes.
- The Y axis is perpendicular to the other two.

As such, GPS produces co-ordinates in the WGS84 reference frame, the geocentric frame originally used by the GPS. It is maintained by the US Department of Defense, and there have been a number of versions.

The international geodetic research community, through the IAG, the International Association of Geodesy, has provided its own geocentric frames through a service called IERS, the International Earth Rotation and Reference Systems Service. The frames have names of type ITRFyy: International Terrestrial Reference Frame where yy is the year of publication.

Nowadays these frames agree with WGS84 to the centimetre level.

In the European area, the IAG Regional Reference Frame Sub-Commission for Europe (EUREF) has similarly provided European geocentric reference frames called ETRFyy: European Terrestrial Reference Frame. The ETRF frames have been designed in such a way, that point co-ordinates on the Eurasian continental plate do not change, so the frame *moves with the plate*. The frames are realisations of the co-ordinate reference system ETRS-89.

In many European countries as well as in scientific circles, realisations of ETRS-89 are used, like in Finland EUREF-FIN. The moment of definition or *epoch* of ETRS-89 is the beginning of 1989, the moment when it coincided with ITRS, the International Terrestrial Reference System.

The WGS84, ITRF and ETRF co-ordinate reference frames are all geocentric. They are also terrestrial, attached to the solid Earth and co-rotating: ECEF, "Earth centred, Earth fixed". These are the kind of co-ordinates produced by satellite positioning equipment.

These reference frames are *not inertial*: use of inertial equipment will immediately show that they rotate at a rate of one full turn every sidereal day, $23^h56^m4^s$. Of course, looking at the sky on a starbright night will show that, too. . . .

Newton's laws of motion only apply in an inertial frame. Inertial devices have to be carefully tuned for use in a frame co-rotating with the Earth, as will be seen in section 5.9.

Conventionally, the inertial reference frame is also geocentric, with the origin placed in the Earth's centre of mass, but with the X axis pointing not along the Greenwich meridional plane but to the vernal equinox, the place of the Sun at the start of spring, when it moves from the southern to the northern hemisphere.

The time scale of geocentric reference frames is that of the geopotential of mean sea level. This becomes relevant when using precise atomic clocks both on the Earth and in space.

1.5 Non-geocentric reference frames

Ever since the early 1990s, GPS has been available for creating precisely geocentric reference frames. Most nations of the world, supported by the international geodetic community, have taken this opportunity and the official reference frames in most of them are currently ITRS-based.

This does not mean, however, that the old co-ordinate frames have vanished. Millions of point co-ordinates in old frames languish in old documents, like parcel boundaries and digital zoning and infrastructure maps. Municipalities have expended substantial effort in transforming these data sets to a geocentric reference. It is no longer recommended to use the old co-ordinates.

Real-time kinematic (RTK) positioning, essentially a navigation technique, is a widely used data collection method for digital mapping survey work. This is how the "navigation solution" can be used in mapping surveying. As a benefit, there is no more post-measurement office work. The collected data — which can be quite voluminous, millions of points — goes directly into a spatial database after a limited amount of manual work, such as type encoding according to a catalogue. tosiaikainen

If accuracy demands are at the metre level, even code-based differential GNSS is suitable.

However, if one wishes to work in a local or national non-geocentric system like KKJ, the Finnish Map Grid Co-ordinate System, things get difficult if one also wishes to retain the superior accuracy obtained from GNSS measurements.

Some RTK-GNSS systems enable the following way of measuring:

- Measure several points known in KKJ on the edges of the measurement area and feed in their KKJ co-ordinates.
- Measure the new points to be measured in the area.
- Return to a known point to check if there has been a jump in the integer unknown ("*cycle slip*") of the carrier-wave phase serving as the observable. vaihekatko
- The device itself calculates the parameters of a transformation formula using the known points and transforms on the fly all the newly measured points into KKJ. The transformation used is usually a Helmert transformation in space.

The drawback of this method is that the original accuracy of the measurement data drops irreversibly to the always weaker local accuracy of KKJ. This is why the method is in practice obsolete and is no longer used: the co-ordinates of known points used in RTK surveys must be geocentric.

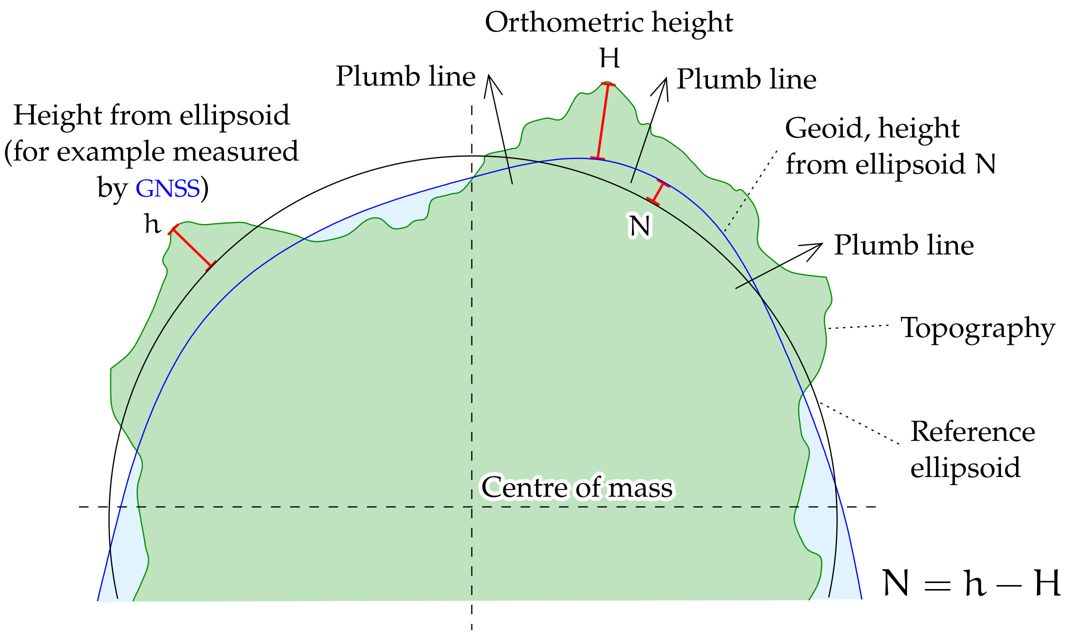


FIGURE 1.8. Height or elevation systems and reference surfaces.

1.6 The vertical reference

vertaus-ellipsoidi

GNSS positioning is often used in height determination. Then, there arises the problem that the heights are also geocentric; in other words, they are heights above the geocentric, mathematically defined reference ellipsoid. Traditional heights from geodesy on the other hand are above "mean sea level", more precisely the *geoid*. The geoid, or mathematical figure of the Earth, is an equipotential surface of the Earth's gravity field.

Figure 1.8 explains the different height reference surfaces and their connections.

1.7 Basic concepts and technologies

In the following chapters, these basic concepts and technologies will be discussed systematically. We shall see that there is a considerable overlap of technologies and approaches that are suitable for both navigation and geodetic location-finding.

Ideas, concepts and technologies to be discussed:

- Stochastic processes and their properties, starting from the basics of stochastic variables, estimation and averaging, covariance, and correlation. Then, time series and linear regression are described, touching upon serial correlation. Auto- and cross-covariance are presented, as are the well-known stochastic processes white noise, random walk, and the Gauss-Markov process. Finally, power spectral density and its link with the autocovariance function are discussed. tehon spektraalitiheys

- The Kalman filter is presented as an example of linear estimation and the least-squares method. The state vector, the dynamic model, the observation model, and the statistical models related to them are presented. It is shown how the dynamic model can be presented in either discrete or continuous form and how the state estimator and state variance propagate in time. Then, it is shown how observations are used to update the state optimally using the Kalman update equation. Several calculation and application examples follow.
- Inertial navigation is presented starting from physical principles, hardware components used, and technical solutions. The mathematics of navigation in the system of the solid Earth is developed. The stabilised platform and gyro compass are explained, followed by the Schuler pendulum applied to navigation in one dimension on a spherical Earth. Mechanisation is discussed and a simplified solution for two-dimensional navigation on the curved surface of a rotating Earth is presented as an example.
- Satellite orbits are discussed, first in terms of Kepler orbits and then in terms of rotating Hill co-ordinates, the basis for describing the relative motions of two orbiting bodies. The Clohessy-Wiltshire formalism is developed. This finds application in navigation in orbit and the *rendezvous* problem.
- GPS, the Global Positioning System, generically GNSS, global navigation satellite systems. The underlying technologies are presented, focusing on the physics of electromagnetic wave propagation, polarisation, and modulation, and the mathematics of pseudo-random codes. The power spectral densities of the signals from navigation satellites are derived. Pseudorange measurement techniques are described that use either the carrier phase or the codes modulated on the carrier. The behaviour of atomic clocks is discussed, and the technique of carrier-smoothed code measurement is presented. vaihetuettu koodimittaus
- Use of GNSS in navigation. The real-time kinematic (RTK) measurement technique is extensively discussed. Observation equations

kokonaisluku-tuntematon
tukiasema

are formulated and the measurement geometry, ambiguity resolution, use of networks of base stations, data standards for disseminating differential corrections, and modelling atmospheric propagation delays are studied.

- Satellite-based augmentation systems (SBAS). How they work, why they are valuable, how they are standardised, and which countries have deployed such systems for their respective air spaces. Related or complementary techniques, like RAIM, receiver autonomous integrity monitoring, and GBAS, ground-based augmentation systems are also presented.

eheyden valvonta

- The new, post-GPS satellite navigation systems GLONASS (Russia), Compass/BeiDou (China) and Galileo (Europe). These systems, their satellite constellations and orbits, frequencies and modulation techniques used, are described in detail.
- A short intermezzo on mapping gravity from the air or from space, techniques having a lot in common with navigation within the Earth's gravity field.
- Sensor fusion and sensors of opportunity are explained, seasoned with a few interesting examples from an active research field.

Self-test questions

tosiaikaisesti

1. What does "in real time" mean?
2. What is dead reckoning?
3. What are the limitations of landmark navigation?
4. How do you find your own latitude in the daytime? At night?
5. Why does longitude determination require a precise chronometer?
6. What are the names of the three axes around which a vehicle can turn?
7. How does a hyperbolic positioning system work?
8. What is a geocentric co-ordinate reference frame?
9. What is the difference between an inertial and a co-rotating reference frame?
10. What different kinds of height exist? What is the role of the geoid?

Stochastic processes

2

2.1 Stochastic variables and processes

A common way to describe an uncertain quantity that changes in a random fashion over time, is as a *stochastic process*. A stochastic process is a generalisation of the concept of a stochastic variable to functions.

For background reading, Strang and Borre (1997) pages 515–541 may serve.

2.1.1 *Stochastic variables*

A *stochastic variable* is defined as follows:

> *A* stochastic variable $\underline{x}$ *is a sequence of realisations* $x_1, x_2, x_3, \ldots$ *or* x_i, $i = 1, 2, 3, \ldots,$ *of a variable* x. *Every realisation value has a certain* probability $p(x)$ *of happening. If we repeat the realisations, or "throws", again and again, the percentage of that value happening tends towards this probability value.*

The traditional notation for stochasticity is an underscore.

The value set or co-domain of a stochastic variable can be a *discrete* or a *continuous* set. The above definition describes discrete stochastic variables.

Examples of discrete stochastic variables are

- Dice throwing. Each throw is one realisation. In this case $x_i \in \{1, 2, 3, 4, 5, 6\}$, a discrete value set. For a fair die, $p(k) = \frac{1}{6}$, $k = 1, \ldots, 6$. As the word "fair" suggests, a die can be used as an impartial decision-making instrument, like in a board game.
- Throwing coins. $x_i \in \{0, 1\}$, $0 =$ heads, $1 =$ tails. Coins are also used for impartial choosing, like by a soccer referee at the start of the game.

- A product quality test: accepted or rejected, $x_i \in \{0, 1\}$. The difference with coin throws is that for a fair coin, the probabilities are $p(0) = p(1) = 0.5$. For a quality test there is no such condition. The manufacturer just wants $p(1)$, the probability of rejection, to be small.
- The codes sent by a bank when doing business using a code calculator. $x_i \in \{1, 2, \ldots, N\}$. A special feature of this case is that the value set is discrete but its size N is large. The purpose of the randomness here is to prevent lucky guessing.

A measurement is usually a real-valued, continuous stochastic variable.

Examples of measurement:

- A *measured distance* is a real-valued, continuous stochastic variable $\underline{s}$. Realisations or measurement values s_i, $i = 1, 2, \ldots$ are in the
1 value set $\{\, s \in \mathbb{R} \mid s > 0 \,\}$, the positive real numbers.[1]
- A vector measurement produced by GNSS from a point A to a point B is a *stochastic vector variable* $\underline{\mathbf{x}}$. Every realisation consists of three components and belongs to a three-dimensional vector space: $\mathbf{x}_i \in \mathbb{R}^3$, $i = 1, 2, \ldots$.

alpha αA

- A measured horizontal angle $\underline{\alpha}$, realisations $\alpha_i \in [0, 2\pi)$, $i = 1, 2, \ldots$.

  For example, angle measurement with a theodolite: the value set
2 is $\{\, \alpha \in \mathbb{R} \mid 0 \leqslant \alpha < 2\pi \,\}$, a subset of the real numbers.[2]

With continuous stochastic variables we speak of probability *density* and not of the probability of a certain realisation value x — as the probability of its precise realisation will be zero. The probability of a realisation falling within a certain interval $I = (x_1, x_2)$ is computed as the integral

$$p(I) = \int_{x_1}^{x_2} p(x)\, dx.$$

Often, the *cumulative* probability density distribution is encountered, the integral

$$P(x) \stackrel{\text{def}}{=} \int_{-\infty}^{x} p(x')\, dx'. \tag{2.1}$$

With this definition,

$$p(I) = P(x_2) - P(x_1).$$

---

[1]More precisely: the positive *rational* numbers, $\{\, s \in \mathbb{Q} \mid s > 0 \,\}$. One cannot measure real values and write them up in a finite number of digits.

[2]More precisely: a subset of the rational numbers, $\{\, \alpha \in \mathbb{Q} \mid 0 \leqslant \alpha < 2\pi \,\}$.

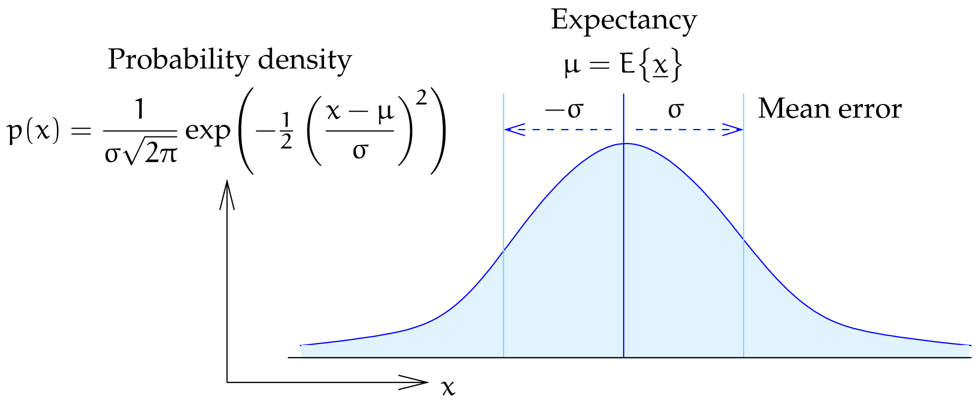


FIGURE 2.1. Gaussian bell curve or normal distribution.

If we assume that the probability distribution is *normal*, in other words following the Gaussian[3] bell curve, figure 2.1, then the equation for it is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{2.2}$$

where σ is the mean error or standard deviation of the distribution, and μ its expectancy, both to be defined later.

sigma $\sigma\Sigma$
mu μM
odotusarvo

2.1.2 *Stochastic processes and time series*

A stochastic *process* is a stochastic variable, the value set of which is a *function space*: each realisation of the stochastic variable (“die throw”) is an entire function.

The argument of the function is usually the time t, but can also be for example the location as latitude and longitude (φ, λ) on the surface of the Earth.

phi $\varphi\phi\Phi$
lambda $\lambda\Lambda$

A *time series* is a discrete series of values obtained from a stochastic process. The series is obtained by specialising, or *sampling*, the argument t to more or less regularly spaced, chosen values t_j, $j = 1, 2, \ldots$. In other words, a time series is a stochastic process that is being regularly measured.

A stochastic process — or a time series — is described as *stationary* if its statistical properties do not change when the argument t is replaced by the argument $t + \Delta t$.

Examples of stochastic processes:

[3]Johann Carl Friedrich Gauss (1777–1855) was a German mathematician and universal genius. *Princeps mathematicorum*.

- The temperature $\underline{T}(t)$ of an experimental device as a function of time t. Different realisations $T_i(t)$ are obtained by repeating the experiment: $i = 1, 2, \ldots$.
- The temperature from the Kaisaniemi weather station in downtown Helsinki $\underline{T}^{\text{Kais}}(t)$. History cannot be precisely repeated: there is only one realisation of this stochastic process, $T_1^{\text{Kais}}(t)$, the historical time series of Kaisaniemi. Other realisations $T_i^{\text{Kais}}(t)$, $i = 2, 3, \ldots$ exist only as theoretical constructs.

History does not repeat itself, but for *ergodic* processes it rhymes. It is often assumed that the result of studying the statistical properties of a process will be same if *the same process shifted in time* by varying amounts are used as realisations. For example

$$T_{i+1}(t) = T_i(t + \Delta t),$$

in which Δt is the time shift, the choice of which depends on the subject of study: for example, the Kaisaniemi time series for different years. This assumption is called the *ergodicity hypothesis*.

2.2 The sample average

2.2.1 General

One often encounters the situation where some quantity x was measured a number of times and we have several realisations of this stochastic measurement $\underline{x}$ available.

Of course all realisations differ in various ways from the "real" value x, *which we do not know*. If we did, we would not need to measure! We can however calculate, using the realisations or measurement values of the quantity that we have, an estimate for x that is "as good as possible". The computation techniques for doing this are called *estimation*.

The estimate is itself a *realisation* of the *estimator*: the estimator itself is a stochastic quantity, its realisations being estimates.

On the value set or co-domain of the stochastic quantity, the set of all possible values x, a *probability density function* $p(x)$ is defined. This function represents the probability that the value of one realisation happens to be inside a narrow interval around x, divided by the width of the interval. It is also the derivative of the cumulative probability density function $P(x)$, equation 2.1.

Often it is assumed that $p(x)$ has the form of the so-called Gaussian curve or *normal distribution*, the "bell curve", equation 2.2 and figure 2.1. The results presented below do not depend on the assumption of a normal distribution if not mentioned otherwise.

Because the variable x must assume *some* value, it follows that the total probability is 1, or as a percentage, 100 %:

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1.$$

The definition of the expected value or *expectancy* E is

$$E\{\underline{x}\} \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} x\,p(x)\,dx.$$

Expectancy is not the same as average: expectancy is a theoretical concept, while the average is calculated from measurements. There is an important connection though: the average of the first n realisations of variable $\underline{x}$,

$$\overline{x}^{(n)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.3}$$

is probably the closer to the expectancy $E\{\underline{x}\}$, the larger n is. This law based on experience is called the empirical *law of large numbers*.

In equation 2.3, the first set of n realisations is called the *sample*, and $\overline{x}^{(n)}$ is the *sample average*. otos otoskeskiarvo

Now that the expectancy has been defined, next define the *variance*:

$$\mathrm{Var}\{\underline{x}\} \stackrel{\text{def}}{=} E\left\{\left(\underline{x} - E\{\underline{x}\}\right)^2\right\}.$$

The square root of the variance is the standard deviation or mean error σ, see figure 2.1:

$$\sigma^2 = \mathrm{Var}\{\underline{x}\}.$$

The variance, like the expectancy, is a theoretical value that cannot be exactly known. It can however be *estimated* from the sample x_i, $i = 1, \ldots, n$. If the sample average $\overline{x}^{(n)}$ has already been calculated, an estimator of the variance σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}^{(n)}\right)^2.$$

Because the sampling can be repeated as often as one wishes, the sample average $\overline{x}^{(n)}$ itself also becomes a stochastic quantity,

$$\underline{\overline{x}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i,$$

in which $\underline{x}_i$ is a stochastic quantity the successive realisations of which are simply $(x_i)_j, j = 1, 2, \ldots$, in which now j is a new realisation counter.

When the x_i are given, we may for example construct

$$\overline{x}_j^{(n)} = \frac{1}{n}\sum_{i=1}^{n}(x_i)_j\,, \qquad (x_i)_j \stackrel{\text{def}}{=} x_{i+n(j-1)},$$

(a "comb variate"). To clarify, assume that the sample size is $n = 10$. Then, the sample average as a stochastic quantity is $\underline{\overline{x}}^{(10)}$, with successive realisations

$$\overline{x}_1^{(10)} = \tfrac{1}{10}\sum_{i=1}^{10} x_i, \quad \overline{x}_2^{(10)} = \tfrac{1}{10}\sum_{i=11}^{20} x_i, \quad \overline{x}_3^{(10)} = \tfrac{1}{10}\sum_{i=21}^{30} x_i, \quad \ldots.$$

It is intuitively clear that

$$E\{\underline{x}_i\} = E\{\underline{x}\}, \qquad i \in \{1, \ldots, n\}.$$

The expectancy of $\underline{\overline{x}}^{(n)}$ is

$$E\left\{\underline{\overline{x}}^{(n)}\right\} = \frac{1}{n}\sum_{i=1}^{n} E\{\underline{x}_i\} = E\{\underline{x}\},$$

harhaton estimaattori

the expectancy of $\underline{x}$. This kind of estimator is called *unbiased*.

Its variance is estimated with the equation

$$\widehat{\mathrm{Var}}\left\{\underline{\overline{x}}^{(n)}\right\} = \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left(\underline{x}_i - \underline{\overline{x}}^{(n)}\right)^2 = \frac{1}{n}\widehat{\sigma^2}.$$

In other words, the mean error of the sample average decreases proportionally to $^1/_{\sqrt{n}}$ when the size of the sample n increases.

2.2.2 Optimality of the average value

Among all unbiased estimators of x based on sample $\underline{x}_i, i = 1, \ldots, n$:

$$\widehat{x} = \left\{\sum_{i=1}^{n} w_i\,\underline{x}_i \;\middle|\; \sum_{i=1}^{n} w_i = 1\right\},$$

the *average*

$$\widehat{x} = \underline{\overline{x}}^{(n)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i \tag{2.4}$$

varianssien kasautuminen

minimises its variance. According to the propagation law of variances, the variance is

$$\mathrm{Var}\{\underline{\overline{x}}^{(n)}\} = \sum_{i=1}^{n} w_i^2 \,\mathrm{Var}\{\underline{x}_i\} = \sigma^2 \sum_{i=1}^{n} w_i^2,$$

assuming that the $\underline{x}_i$ do not correlate with each other and that $\mathrm{Var}\{\underline{x}_i\} = \sigma^2$.

Now, the constrained minimisation of the expression

$$\left\{ \sum_{i=1}^{n} w_i^2 \,\middle|\, \sum_{i=1}^{n} w_i = 1 \right\}$$

yields

$$w_i = \frac{1}{n},$$

from which the claim follows. The coefficients w_i are called *weights* and the situation in which they are all equal is called unweighted averaging.

2.2.3 *Computing the sample average one step at a time*

Instead of calculating the sample average directly, it can also be calculated *step by step* as follows:

$$\begin{aligned} \underline{\overline{x}}^{(n+1)} &= \frac{n}{n+1}\,\underline{\overline{x}}^{(n)} + \frac{1}{n+1}\,\underline{x}_{n+1}, \\ \mathrm{Var}\{\underline{\overline{x}}^{(n+1)}\} &= \left(\frac{n}{n+1}\right)^2 \mathrm{Var}\{\underline{\overline{x}}^{(n)}\} + \left(\frac{1}{n+1}\right)^2 \sigma^2. \end{aligned} \tag{2.5}$$

This is a simple example of sequential linear filtering, *the Kalman filter* (chapter 3). By using this procedure, it is possible to obtain a value for $\underline{\overline{x}}^{(n)}$ "on the fly", while observations are still being collected, after every new observation, without having to re-process the whole set of observations. This is precisely the advantage of using the Kalman filter.

Equations 2.5 can also be understood as a weighted average between the observations done so far and the new observation:

$$\begin{aligned} \underline{\overline{x}}^{(n+1)} &= W\,\underline{\overline{x}}^{(n)} + w\,\underline{x}_{n+1}, \\ \mathrm{Var}\{\underline{\overline{x}}^{(n+1)}\} &= W^2\,\mathrm{Var}\{\underline{\overline{x}}^{(n)}\} + w^2\sigma^2, \end{aligned}$$

with

$$w = \frac{1}{n+1}, \qquad W = \sum_{i=1}^{n} w = nw = \frac{n}{n+1}, \qquad W + w = 1.$$

2.3 Covariance, correlation

2.3.1 Definitions and properties

When two stochastic quantities $\underline{x}$ and $\underline{y}$ are given, the *covariance* between them is

$$\operatorname{Cov}\{\underline{x},\underline{y}\} \stackrel{\text{def}}{=} \operatorname{E}\left\{\left(\underline{x}-\operatorname{E}\{\underline{x}\}\right)\left(\underline{y}-\operatorname{E}\{\underline{y}\}\right)\right\}.$$

The covariance describes how strong the common random variations of $\underline{x}$ and $\underline{y}$ are.

Besides covariance, *correlation* is defined as

$$\operatorname{Corr}\{\underline{x},\underline{y}\} \stackrel{\text{def}}{=} \frac{\operatorname{Cov}\{\underline{x},\underline{y}\}}{\sqrt{\operatorname{Var}\{\underline{x}\}\operatorname{Var}\{\underline{y}\}}}. \tag{2.6}$$

Correlation can never be more than 1.0 or less than −1.0. Eric Weisstein gives the following proof (Wolfram MathWorld, Statistical Correlation).

xi ξ,Ξ eta ηH

Define the normalised quantities $\underline{\xi}$ and $\underline{\eta}$:

$$\underline{\xi} \stackrel{\text{def}}{=} \frac{\underline{x}}{\sqrt{\operatorname{Var}\{\underline{x}\}}}, \qquad \underline{\eta} \stackrel{\text{def}}{=} \frac{\underline{y}}{\sqrt{\operatorname{Var}\{\underline{y}\}}}.$$

By propagation of variances

$$\operatorname{Cov}\{\underline{\xi},\underline{\eta}\} = \frac{\operatorname{Cov}\{\underline{x},\underline{y}\}}{\sqrt{\operatorname{Var}\{\underline{x}\}\operatorname{Var}\{\underline{y}\}}} = \operatorname{Corr}\{\underline{x},\underline{y}\}.$$

The following variances are non-negative:

$$\begin{aligned}0 &\leqslant \operatorname{Var}\{\underline{\xi}+\underline{\eta}\} = \operatorname{Var}\{\underline{\xi}\}+\operatorname{Var}\{\underline{\eta}\}+2\operatorname{Cov}\{\underline{\xi},\underline{\eta}\},\\ 0 &\leqslant \operatorname{Var}\{\underline{\xi}-\underline{\eta}\} = \operatorname{Var}\{\underline{\xi}\}+\operatorname{Var}\{\underline{\eta}\}-2\operatorname{Cov}\{\underline{\xi},\underline{\eta}\}.\end{aligned}$$

When also

$$\operatorname{Var}\{\underline{\xi}\} = \frac{\operatorname{Var}\{\underline{x}\}}{\sqrt{\operatorname{Var}\{\underline{x}\}\operatorname{Var}\{\underline{x}\}}} = 1$$

and similarly $\operatorname{Var}\{\underline{\eta}\} = 1$, it follows that

$$-1 \leqslant \operatorname{Cov}\{\underline{\xi},\underline{\eta}\} = \operatorname{Corr}\{\underline{x},\underline{y}\} \leqslant 1.$$

Often, the correlation is expressed as a percentage, 100 % being the same as 1.0.

In the case of two stochastic processes one may draw an *error ellipse*, figure 2.2. Compare this picture with the earlier picture of the bell

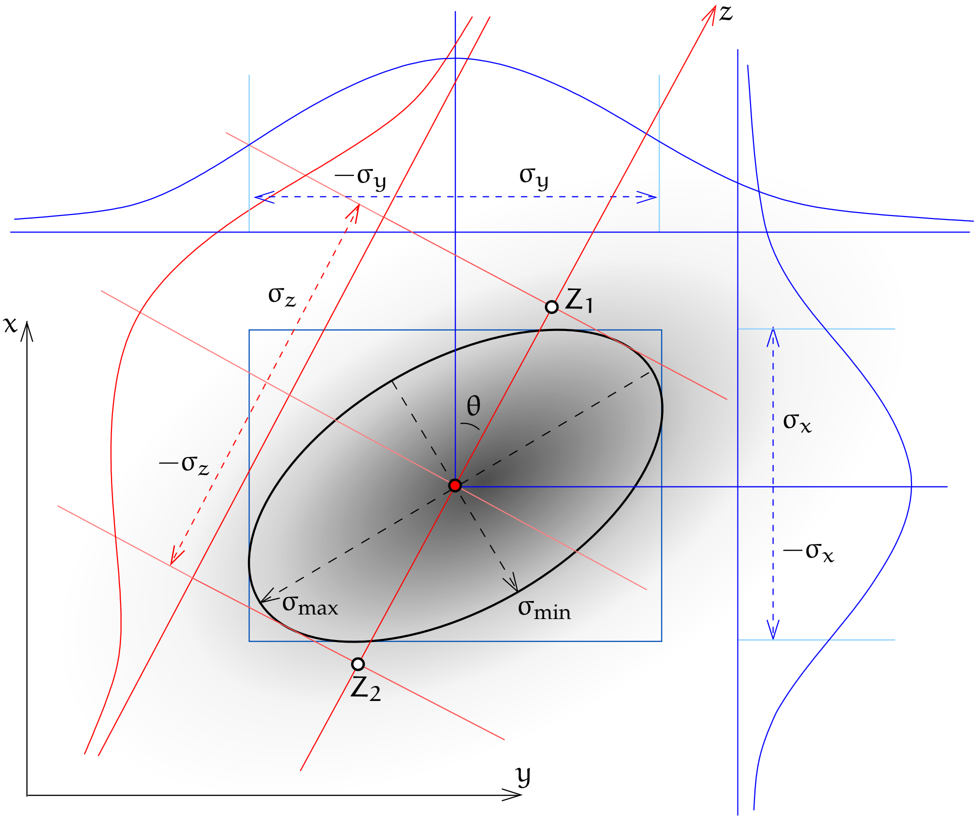


FIGURE 2.2. Error ellipse with probability density distributions. The shading indicates probability density.

curve, figure 2.1. There, the expectancy is marked as $\mu = E\{\underline{x}\}$ (in the middle) and the mean error as $\pm\sigma$. In the error ellipse of figure 2.2, the central point represents the expectancies of $\underline{x}$ and $\underline{y}$. The ellipse itself is the two-dimensional version of the mean error $\pm\sigma$ in figure 2.1.

The measurement values will with certain probabilities fall inside or outside the ellipse: this gives the error ellipse its name. However, these probabilities are not the same for a one-dimensional interval $[\mu - \sigma, \mu + \sigma]$ as for a two-dimensional ellipse or a three-dimensional

TABLE 2.1. Probabilities (%) of being outside the one-sigma, two-sigma, and three-sigma bounds, for one, two, and three dimensions.

	σ	2σ	3σ
One dimension (interval)	31.7	4.6	0.3
Two dimensions (ellipse)	60.6	13.5	1.1
Three dimensions (ellipsoid)	80.1	26.1	2.9

error ellipsoid. See table 2.1.

2.3.2 Correlation, error ellipse, and regression

If the ellipse is intersected by a line z, the linear combination of $\underline{x}$ and $\underline{y}$
theta $\vartheta\theta\Theta$ is obtained, with θ the direction angle of the line:

$$\underline{z} = \underline{x}\cos\theta + \underline{y}\sin\theta.$$

The points Z_1 and Z_2 in which the line intersects the projecting tangents
4 of the ellipse represent the mean error of the quantity $\underline{z}$:[4]

$$\begin{aligned}\operatorname{Var}\{\underline{z}\} &= \mathrm{E}\left\{\left(\underline{z}-\mathrm{E}\{\underline{z}\}\right)^2\right\} = \\ &= \mathrm{E}\left\{\left(\cos\theta\left(\underline{x}-\mathrm{E}\{\underline{x}\}\right)+\sin\theta\left(\underline{y}-\mathrm{E}\{\underline{y}\}\right)\right)^2\right\} = \\ &= \cos^2\theta\operatorname{Var}\{\underline{x}\} + 2\sin\theta\cos\theta\operatorname{Cov}\{\underline{x},\underline{y}\} + \sin^2\theta\operatorname{Var}\{\underline{y}\},\end{aligned}$$

and from this $\sigma_z = \sqrt{\operatorname{Var}\{\underline{z}\}}$. Considered as a function of the direction angle θ, the mean error $\sigma_z(\theta)$ has two extremal values, $\sigma_{\min}$ and $\sigma_{\max}$, figure 2.2.

If $\sigma_{\min} = \sigma_{\max}$, or the co-ordinate axes point in the directions of the extremal values $\sigma_{\min}$ and $\sigma_{\max}$, the correlation between $\underline{x}$ and $\underline{y}$ vanishes. In that case they really are independent from each other and knowing the real value of one does not help in estimating the other.

If the correlation does not vanish, knowledge of — or a good estimate of — the real value of $\underline{x}$ helps to estimate $\underline{y}$ better. This method is called *regression*.

---

[4] In matrix notation

$$\underline{z} = \begin{bmatrix}\cos\theta & \sin\theta\end{bmatrix}\begin{bmatrix}\underline{x}\\ \underline{y}\end{bmatrix}$$

and

$$\operatorname{Var}\left\{\begin{bmatrix}\underline{x}\\ \underline{y}\end{bmatrix}\right\} = \begin{bmatrix}\operatorname{Var}\{\underline{x}\} & \operatorname{Cov}\{\underline{x},\underline{y}\}\\ \operatorname{Cov}\{\underline{x},\underline{y}\} & \operatorname{Var}\{\underline{y}\}\end{bmatrix}.$$

This implies

$$\operatorname{Var}\{\underline{z}\} = \begin{bmatrix}\cos\theta & \sin\theta\end{bmatrix}\begin{bmatrix}\operatorname{Var}\{\underline{x}\} & \operatorname{Cov}\{\underline{x},\underline{y}\}\\ \operatorname{Cov}\{\underline{x},\underline{y}\} & \operatorname{Var}\{\underline{y}\}\end{bmatrix}\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix},$$

the same result. This illustrates the *law of propagation of variances*.

2.3.3 The effect of ignoring correlation

If two stochastic quantities $\underline{x}_1$ and $\underline{x}_2$ are given, a weighted average may be calculated:

$$\underline{\overline{x}} = w_1\,\underline{x}_1 + w_2\,\underline{x}_2,$$

in which w_1 and w_2 are the weights, and $w_1 + w_2 = 1$. The variance of this linear combination will be

$$\begin{aligned}\mathrm{Var}\{\underline{\overline{x}}\} &= w_1^2\sigma_1^2 + 2\,w_1w_2\sigma_{12} + w_2^2\sigma_2^2 = \\ &= w_1^2\sigma_1^2 + 2\,w_1\left(1-w_1\right)\sigma_{12} + \left(1-w_1\right)^2\sigma_2^2, \end{aligned}\quad (2.7)$$

in which $\sigma_1^2 \stackrel{\text{def}}{=} \mathrm{Var}\{\underline{x}_1\}$, $\sigma_2^2 \stackrel{\text{def}}{=} \mathrm{Var}\{\underline{x}_2\}$, and $\sigma_{12} \stackrel{\text{def}}{=} \mathrm{Cov}\{\underline{x}_1, \underline{x}_2\}$.

The optimum is the point where the derivative with respect to w_1 vanishes:

$$\frac{d}{dw_1}\,\mathrm{Var}\{\underline{\overline{x}}\} = 2\,w_1\sigma_1^2 + 2\left(1-2\,w_1\right)\sigma_{12} - 2\left(1-w_1\right)\sigma_2^2 = 0$$

$$\implies w_1\left(\sigma_1^2 - \sigma_{12}\right) - w_2\left(\sigma_2^2 - \sigma_{12}\right) = 0.$$

The optimum weights are now

$$w_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \qquad w_2 = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}.$$

In the special case[5] in which $\sigma_1^2 = \sigma_2^2 \stackrel{\text{def}}{=} \sigma^2$, the optimal weights are $w_1 = w_2 = \frac{1}{2}$. 5

Ignoring the statistical interdependence σ_{12} will give for the optimal weights

$$w_1' = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \qquad w_2' = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}.$$

Also here, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $w_1' = w_2' = \frac{1}{2}$.

When ignoring σ_{12}, we obtain for the variance of the average

$$\begin{aligned}\mathrm{Var}'\{\underline{\overline{x}}\} &= \left(w_1'\right)^2\sigma_1^2 + \left(w_2'\right)^2\sigma_2^2 = \\ &= w_1'\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\sigma_1^2 + w_2'\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\sigma_2^2 = \left(w_1' + w_2'\right)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}.\end{aligned}$$

In the case $\sigma_1^2 = \sigma_2^2 = \sigma^2$ this yields

$$\mathrm{Var}'\{\underline{\overline{x}}\} = \tfrac{1}{2}\sigma^2.$$

The *actual* variance will however be, equation 2.7:

$$\mathrm{Var}\{\underline{\overline{x}}\} = \left(\left(w_1'\right)^2\sigma_1^2 + \left(w_2'\right)^2\sigma_2^2\right) + 2\,w_1'w_2'\sigma_{12} =$$

[5]This special case is obtained by normalising $\underline{\xi}_1 \stackrel{\text{def}}{=} \underline{x}_1/\sigma_1$, $\underline{\xi}_2 \stackrel{\text{def}}{=} \underline{x}_2/\sigma_2$. Note that then, $\mathrm{Cov}\{\underline{\xi}_1, \underline{\xi}_2\} = \mathrm{Corr}\{\underline{\xi}_1, \underline{\xi}_2\} = \mathrm{Corr}\{\underline{x}_1, \underline{x}_2\}$.

$$
\begin{aligned}
&= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} + 2\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\sigma_{12} = \\
&= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\left(1 + 2\frac{\sigma_{12}}{\sigma_1^2+\sigma_2^2}\right) = \\
&= \mathrm{Var}'\{\underline{\overline{x}}\}\left(1 + 2\frac{\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\frac{\sigma_{12}}{\sigma_1\sigma_2}\right) = \\
&= \mathrm{Var}'\{\underline{\overline{x}}\}\left(1 + 2\frac{\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}\mathrm{Corr}\{\underline{x}_1, \underline{x}_2\}\right).
\end{aligned}
$$

In the case $\sigma_1^2 = \sigma_2^2 = \sigma^2$ this yields

$$\mathrm{Var}\{\underline{\overline{x}}\} = \mathrm{Var}'\{\underline{\overline{x}}\}\left(1 + \mathrm{Corr}\{\underline{x}_1, \underline{x}_2\}\right).$$

nu νN

tehollinen otoskoko

The factor $\nu \stackrel{\text{def}}{=} 1 + \mathrm{Corr}\{\underline{x}_1, \underline{x}_2\}$ is often expressed as an "effective sample size" in the form

$$n_{\text{eff}} = \frac{n}{\nu},$$

in which n is the actual sample size, in this case $n = 2$.

For a correlation of $\mathrm{Corr}\{\underline{x}_1, \underline{x}_2\} = 0.8$ we find $\nu = 1.8$, $\mathrm{Var}\{\underline{\overline{x}}\} = 1.8 \cdot \mathrm{Var}'\{\underline{\overline{x}}\}$, and $n_{\text{eff}} = {}^2/_{1.8} = 1.111\ldots$.

This is easily extended to the case of n values $\underline{x}_i$, $i = 1, \ldots, n$ with identical variances σ^2. The average is ($w_i = {}^1/_n$, $i = 1, \ldots, n$):

$$\underline{\overline{x}} = \sum_{i=1}^{n} w_i\,\underline{x}_i = \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i,$$

and the variance of the average is

$$
\begin{aligned}
\mathrm{Var}\{\underline{\overline{x}}\} &= \sum_{i=1}^{n} w_i^2\sigma^2 + \sum_{\substack{j,k=1\\ j\neq k}}^{n} w_j w_k \sigma_{jk} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n^2}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\sigma^2\frac{\sigma_{jk}}{\sigma^2} = \\
&= \frac{\sigma^2}{n}\left(1 + \frac{1}{n}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathrm{Corr}\{\underline{x}_j, \underline{x}_k\}\right) = \\
&= \mathrm{Var}'\{\underline{\overline{x}}\}\left(1 + \frac{1}{n}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathrm{Corr}\{\underline{x}_j, \underline{x}_k\}\right).
\end{aligned}
$$

An interesting special case arises when the correlations are powers of a constant $\rho \in (0, 1)$, falling off with index distance:

rho ρR

$$\mathrm{Corr}\{\underline{x}_j, \underline{x}_k\} = \rho^{|j-k|}.$$

6 With the definition[6] $m \stackrel{\text{def}}{=} |j - k|$:

$$\mathrm{Var}\{\underline{\bar{x}}\} = \mathrm{Var}'\{\underline{\bar{x}}\}\left(1 + \frac{2}{n}\sum_{m=1}^{n-1}\sum_{\ell=1}^{n-m}\rho^m\right) =$$
$$= \mathrm{Var}'\{\underline{\bar{x}}\}\left(1 + 2\sum_{m=1}^{n-1}\frac{n-m}{n}\rho^m\right) =$$
$$= \mathrm{Var}'\{\underline{\bar{x}}\}\left(1 + 2\sum_{m=1}^{n-1}\rho^m - \frac{2}{n}\sum_{m=1}^{n-1}m\rho^m\right).$$

Then, in the limit $n \to \infty$:

$$\mathrm{Var}\{\underline{\bar{x}}\} \approx \mathrm{Var}'\{\underline{\bar{x}}\}\left(1 + 2\frac{\rho}{1-\rho}\right) = \mathrm{Var}'\{\underline{\bar{x}}\}\frac{1+\rho}{1-\rho}.$$

The take-home message from this is that in the presence of correlation, stochastic quantities contain less *independent* information than it appears. Ignoring correlation leads one to believe that one knows more than one actually knows. It should be clear why this is dangerous.

This analysis may be generalised to other estimators besides the average. More on the effective sample size applied to serially correlated time series will be presented in subsection 2.4.2.

2.4 Linear regression of time series

In real life, as mentioned in subsection 2.1.2, stochastic processes are always provided as *time series*, sequences of values given for discrete points, or epochs, in time. Often, the behaviour of such time series is to first order linear in time and we may wish to study this linear behaviour by regression. There are some slings and arrows to remember here.

2.4.1 *Least-squares regression in the absence of autocorrelation*

Linear regression starts from the well-known equation

$$y = a + bx,$$

and a set of point pairs (x_i, y_i), $i = 1, \ldots, n$ are given. We wish to determine the intercept a and the trend b. This is actually an *observation equation*

$$\underline{y}_i = a + bx_i + \underline{n}_i,$$

[6]The variance matrix has $n-1$ side diagonals on each side, the lengths of which are $n-m$.

in which the stochastic process $\underline{n}_i$ models the noise or random error in the measurement process.

Another way to write this observation equation is

$$\underline{y}_i + \underline{v}_i = \widehat{a} + \widehat{b}x_i,$$

in which now $\underline{v}_i$ is the *residual* of observation $\underline{y}_i$, and $\widehat{a}$ and $\widehat{b}$ are the estimators of the unknowns a and b.

We assume the observation noise to have constant variance independently of i ("homoscedasticity"), and the covariances to vanish
7 identically[7] ("white noise"):

$$\begin{aligned} \mathrm{Var}\{\underline{n}_i\} &= \sigma^2, \\ \mathrm{Cov}\{\underline{n}_i, \underline{n}_j\} &= 0, \qquad i \neq j. \end{aligned}$$

This is called the *statistical model*.

We may write the observation equations into the form

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_n \end{bmatrix} + \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \vdots \\ \underline{v}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix},$$

in which $\underline{\boldsymbol{\ell}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{y}_1 & \underline{y}_2 & \cdots & \underline{y}_n \end{bmatrix}^{\mathsf{T}}$ is the vector of observations, $\underline{\mathbf{v}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_n \end{bmatrix}^{\mathsf{T}}$ that of residuals, in an n-dimensional abstract vector space $\mathbb{R}^n$, and $\widehat{\mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \widehat{a} & \widehat{b} \end{bmatrix}^{\mathsf{T}}$ is the vector of unknowns (parameters). Finally

$$A \stackrel{\text{def}}{=} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

rakennematriisi is the *design matrix*. This form of presentation is referred to as the
8 *function model*:[8]

$$\underline{\boldsymbol{\ell}} + \underline{\mathbf{v}} = A\widehat{\mathbf{x}}.$$

Based on the assumed statistical model, we may compute the least-
9 squares solution using the *normal equation*:[9]

[7]This set of assumptions about the observations is called *i.i.d.*, "independent and

$$\left(A^{\mathsf{T}}A\right)\widehat{\mathbf{x}} = A^{\mathsf{T}}\underline{\boldsymbol{\ell}}.$$

The normal matrix is

$$A^{\mathsf{T}}A = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix},$$

or according to Cramer's[10] rule

$$\left(A^{\mathsf{T}}A\right)^{-1} = \frac{1}{n\sum x^2 - \left(\sum x\right)^2}\begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix}.$$

From this follows

$$\widehat{\mathbf{x}} = \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \left(A^{\mathsf{T}}A\right)^{-1} A^{\mathsf{T}}\underline{\ell} = \left(A^{\mathsf{T}}A\right)^{-1}\begin{bmatrix} \sum \underline{y} \\ \sum x\underline{y} \end{bmatrix}$$

and

$$\widehat{a} = \frac{\sum x^2 \sum \underline{y} - \sum x \sum x\underline{y}}{n\sum x^2 - \left(\sum x\right)^2}, \qquad \widehat{b} = \frac{-\sum x \sum \underline{y} + n\sum x\underline{y}}{n\sum x^2 - \left(\sum x\right)^2},$$

which are the least-squares *estimators* of the unknowns. Their precisions — uncertainties, mean errors — are obtained by formal error propagation as the square roots of the diagonal elements of the inverted normal matrix $\left(A^{\mathsf{T}}A\right)^{-1}$, scaled by the factor σ, the mean error of unit weight: virheiden kasautuminen

$$\sigma_a = \sigma\sqrt{\frac{\sum x^2}{n\sum x^2 - \left(\sum x\right)^2}}, \qquad \sigma_b = \sigma\sqrt{\frac{n}{n\sum x^2 - \left(\sum x\right)^2}}. \tag{2.8}$$

Most often we are specifically interested in the trend b, meaning that we should compare the value $\widehat{b}$ obtained with its own mean error σ_b. If

identically distributed". This includes the further assumption of an identical shape of the distributions of all observations.

[8]An alternative, popular notation is $\underline{y} = \beta X + \underline{\varepsilon}$, with X the *explanatory* or *independent* variables, $\underline{y}$ the *explained* or *response* or *dependent* variables, and β the *model parameters*. The correspondence is $\underline{y} \sim \underline{\boldsymbol{\ell}}$, $X \sim A$, and $\beta \sim \mathbf{x}$, but $\underline{\varepsilon} \not\sim -\underline{\mathbf{v}}$. $\underline{\varepsilon} = \underline{\mathbf{n}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{n}_1 & \underline{n}_2 & \cdots & \underline{n}_n \end{bmatrix}^{\mathsf{T}}$ is the vector of *errors*, while $\underline{\mathbf{v}}$ is the vector of *residuals*.

[9]Equivalently

$$\left(X^{\mathsf{T}}X\right)\widehat{\beta} = X^{\mathsf{T}}\underline{y}.$$

[10]Gabriel Cramer (1704–1752) was a Swiss mathematician.

σ is not known *a priori*, it should be evaluated from the *residuals*: the sum of squared residuals

$$\sum \underline{v}^2 = \sum \left(\widehat{a} + \widehat{b}x - \underline{y}\right)^2$$

has an expectancy of $(n-2)\,\sigma^2$, in which $n-2$ is the number of *degrees of freedom* or overdetermination, 2 being the number of unknowns estimated. So

$$\widehat{\sigma^2} = \frac{\sum \underline{v}^2}{n-2}$$

is an unbiased estimator of σ^2.

2.4.2 Serial correlation

The assumption made above, that the observational errors $\underline{n}_i$ are uncorrelated among themselves, is often *wrong*. Nevertheless, least-squares regression is such a simple method — available for example in popular spreadsheets and scientific computation software — that it is often used even though the zero-correlation requirement is not met.

If the autocorrelation of the noise process $\underline{w}_i$ does not vanish, we can often model it as a Gauss-Markov process. Such a process is described in discrete form as a *Markov chain*:

$$\underline{w}_{i+1} = \rho \underline{w}_i + \underline{n}_i, \tag{2.9}$$

11 vaimennus in which[11] $\rho = \operatorname{Corr}\{\underline{w}_{i+1}, \underline{w}_i\}$ is a suitable damping parameter, $0 < \rho < 1$, and $\underline{n}_i$ is a truly non-correlating white-noise process:

$$\begin{aligned} \operatorname{Var}\{\underline{n}_i\} &= \sigma_n^2, \\ \operatorname{Cov}\{\underline{n}_i, \underline{n}_j\} &= 0, \qquad i \neq j. \end{aligned} \tag{2.10}$$

It follows from equation 2.9 and $\operatorname{Cov}\{\underline{w}_i, \underline{n}_i\} = 0$ that

$$\sigma_w^2 = \rho^2\sigma_w^2 + \sigma_n^2 \implies \sigma_w^2 = \frac{\sigma_n^2}{1-\rho^2}. \tag{2.11}$$

[11] Because

$$\begin{aligned} \operatorname{Cov}\{\underline{w}_{i+1}, \underline{w}_i\} &= \operatorname{Cov}\{(\rho\underline{w}_i + \underline{n}_i), \underline{w}_i\} = \rho \operatorname{Var}\{\underline{w}_i\} \\ \implies \rho = \frac{\operatorname{Cov}\{\underline{w}_{i+1}, \underline{w}_i\}}{\operatorname{Var}\{\underline{w}_i\}} &= \frac{\operatorname{Cov}\{\underline{w}_{i+1}, \underline{w}_i\}}{\sqrt{\operatorname{Var}\{\underline{w}_{i+1}\}\operatorname{Var}\{\underline{w}_i\}}} = \operatorname{Corr}\{\underline{w}_{i+1}, \underline{w}_i\} \end{aligned}$$

based on stationarity: $\operatorname{Var}\{\underline{w}_i\} = \operatorname{Var}\{\underline{w}_{i+1}\} = \sigma_w^2$.

Write the original observation equation

$$\underline{y}_i = a + bx_i + \underline{w}_i$$

two times, multiplied the second time around by $-\rho$:

$$\begin{aligned} \underline{y}_{i+1} &= a \quad\;\; + bx_{i+1} + \underline{w}_{i+1}, \\ -\rho\underline{y}_i &= -\rho a - \rho bx_i \;\; - \rho\underline{w}_i, \end{aligned}$$

and sum together:

$$\underline{y}_{i+1} - \rho\underline{y}_i = a\left(1-\rho\right) + b\left(x_{i+1} - \rho x_i\right) + \overbrace{\left(\underline{w}_{i+1} - \rho\underline{w}_i\right)}^{\underline{n}_i}.$$

This equation is of the form

$$\underline{Y}_i = A + bX_i + \underline{n}_i,$$

the equation for the non-correlated linear regression, in which $\underline{n}_i = \underline{w}_{i+1} - \rho\underline{w}_i$ is *white* noise, equation 2.10, and

$$A = a\left(1-\rho\right), \qquad X_i = x_{i+1} - \rho x_i, \qquad \underline{Y}_i = \underline{y}_{i+1} - \rho\underline{y}_i. \tag{2.12}$$

This approach is known as the Cochrane-Orcutt method. (Wikipedia, Cochrane-Orcutt estimation). The recipe now is:

1. Compute X_i and $\underline{Y}_i$ according to the above equations 2.12.
2. Solve $\widehat{A}$ and $\widehat{b}$ according to non-correlated linear regression. The method for this was described in subsection 2.4.1.
3. Compute $\widehat{a} = \widehat{A}\big/\left(1-\rho\right)$.
4. The non-correlated linear regression method will give mean errors for the estimators $\widehat{A}$ and $\widehat{b}$. Compute $\sigma_a = \sigma_A\big/\left(1-\rho\right)$.
5. The mean error σ_b computed considering serial correlation will be larger than that obtained without doing so. Of course the ratio $\widehat{b}\big/\sigma_b$, which is used to judge whether the trend b differs significantly from zero, will also be smaller.

   What is happening here is that the observations *contain less information* on the unknowns than we think they contain. Every new data point of the time series will not contain the amount of information we think it contains, because it will provide partly the same information as previous points.

As the time series is assumed to be equispaced, it follows, with $\Delta X = X_{i+1} - X_i$ and $\Delta x = x_{i+1} - x_i$, that $\Delta X = (1-\rho)\,\Delta x$. If furthermore we use "centre-of-mass co-ordinates", meaning

$$\sum_{i=1}^{n-1} X_i = \sum_{i=1}^{n} x_i = 0,$$

it will also follow that

$$\sum_{i=1}^{n-1} X_i^2 \approx (1-\rho)^2 \sum_{i=1}^{n} x_i^2. \tag{2.13}$$

Now we apply equation 2.8:

$$\mathrm{Var}\left\{\widehat{b}\right\} = \sigma_b^2 = \sigma^2 \frac{n}{n\sum x^2 - \cancel{\left(\sum x\right)^2}} = \frac{\sigma^2}{\sum x^2}.$$

Note that this equation is valid for *uncorrelated* observations only.

1. Consider the Cochrane-Orcutt transformed observations, making the substitutions $\sigma \to \sigma_n$, $x \to X$ and using equation 2.13, obtaining
$$\mathrm{Var}\left\{\widehat{b}\right\} = \frac{\sigma_n^2}{\sum X^2} = \frac{1}{(1-\rho)^2}\frac{\sigma_n^2}{\sum x^2}. \tag{2.14}$$
2. Consider the fiction that the original observations are uncorrelated with mean error σ_w, and write, with the substitution $\sigma \to \sigma_w$ and equation 2.11:
$$\mathrm{Var}'\left\{\widehat{b}\right\} = \frac{\sigma_w^2}{\sum x^2} = \frac{1}{1-\rho^2}\frac{\sigma_n^2}{\sum x^2}. \tag{2.15}$$

Together, equations 2.14 and 2.15 yield

$$\mathrm{Var}\left\{\widehat{b}\right\} = \mathrm{Var}'\left\{\widehat{b}\right\}\frac{1-\rho^2}{(1-\rho)^2} = \mathrm{Var}'\left\{\widehat{b}\right\}\frac{1+\rho}{1-\rho}.$$

The "effective sample size" n_{eff} is smaller than the true number of observations n, by a factor of ν (Lettenmaier, 1976):

$$\nu = \frac{1+\rho}{1-\rho}.$$

For example, if $\rho = 0.8$, it follows from this rule of thumb that $\nu = 9$! It is like you have only one-ninth of the number of data points you thought you had.

In this case all mean errors computed naively, using ordinary least-squares regression ignoring the autocorrelation, must be multiplied by the square root of this factor ν, that is, tripled. This applies especially to the mean error σ_b of the trend estimator $\widehat{b}$.

The take-away from this is

- if there is serial correlation (autocorrelation) in the data, a simple linear regression will give a *too-optimistic picture* of the trend parameter $\widehat{b}$'s mean error, and thus also of the statistical significance of its difference from zero. In this case, the above Cochrane-Orcutt method may be used. Judging the presence of serial correlation can be done by looking at the residuals. Finding the proper value of the autocorrelation ρ poses its own challenge.
- If the data is given as an equispaced time series, $x_i = x_0 + (i-1)\,\Delta t$, then the Markov-chain parameter ρ is related in a simple way to its *correlation length*. From equation 2.9 follows

  $$\underline{w}_{i+1} = \rho \underline{w}_i + \underline{n}_i \implies \underline{w}_{i+1} = \underline{w}_i e^{-\Delta t/\tau} + \underline{n}_i,$$

  in which τ is the correlation length expressed in units of time: tau τT

  $$\tau = -\,{}^{\Delta t}\!/_{\ln \rho}\,.$$

The general problem of regressing observations correlated in time is an example of *weighted* least-squares.

2.5 Auto- and cross-covariance of a stochastic process

2.5.1 *Autocovariance*

Given is a stochastic process $\underline{x}(t)$. A derived function called the *autocovariance*, with two time arguments t and t', may be calculated as follows:

$$A_x(t,t') \stackrel{\text{def}}{=} \operatorname{Cov}\left\{\underline{x}(t), \underline{x}(t')\right\}.$$

Stationary processes are processes with statistical properties that do not change with time: the properties do not depend on absolute time but stay the same over the course of time. For stationary processes it holds, with $t' = t + \Delta t$, that

$$A_x(t,t') = A_x(t, t+\Delta t) = \operatorname{Cov}\left\{\underline{x}(t), \underline{x}(t+\Delta t)\right\} \stackrel{\text{def}}{=} A_x(\Delta t), \quad (2.16)$$

independently of the value of time t.

The autocovariance function $A_x(t) \stackrel{\text{def}}{=} A_x(t,t)$ is simply called the variance function of process $\underline{x}(t)$. For a stationary process it is a constant $A_x = A_x(0)$.

The autocovariance function of a stationary process is symmetric:

$$A_x(\Delta t) = A_x(t, t') = \mathrm{Cov}\{\underline{x}(t), \underline{x}(t')\} = \mathrm{Cov}\{\underline{x}(t'), \underline{x}(t)\} = \\ = A_x(t', t) = A_x(-\Delta t). \quad (2.17)$$

2.5.2 Cross-covariance

If there are two different stochastic processes $\underline{x}(t)$ and $\underline{y}(t)$, one obtains the derived function called *cross-covariance*:

$$C_{xy}(t, t') \stackrel{\text{def}}{=} \mathrm{Cov}\{\underline{x}(t), \underline{y}(t')\}.$$

Again in the case of stationary processes

$$C_{xy}(\Delta t) \stackrel{\text{def}}{=} \mathrm{Cov}\{\underline{x}(t), \underline{y}(t + \Delta t)\}. \quad (2.18)$$

Note that $A_x(\Delta t) = A_x(-\Delta t)$, but (generally) $C_{xy}(\Delta t) = C_{yx}(-\Delta t) \neq C_{xy}(-\Delta t)$! Often, the term cross-covariance simply denotes

$$C_{xy}(t) \stackrel{\text{def}}{=} C_{xy}(t, t) = \mathrm{Cov}\{\underline{x}(t), \underline{y}(t)\}.$$

For stationary processes, this is a constant: $C_{xy}(t) = C_{xy}(0) = C_{xy}$.

2.5.3 Auto- and cross-correlation

With the covariances defined like this, one can also define the auto- and cross-correlation functions in the familiar way. *Autocorrelation* is

$$\mathrm{Corr}_x(t, t') \stackrel{\text{def}}{=} \frac{A_x(t, t')}{\sqrt{A_x(t, t)\, A_x(t', t')}}, \quad \mathrm{Corr}_x(\Delta t) \stackrel{\text{def}}{=} \frac{A_x(\Delta t)}{A_x(0)}, \quad (2.19)$$

and *cross-correlation*

$$\mathrm{Corr}_{xy}(t, t') \stackrel{\text{def}}{=} \frac{C_{xy}(t, t')}{\sqrt{A_x(t, t)\, A_y(t', t')}}, \quad \mathrm{Corr}_{xy}(\Delta t) \stackrel{\text{def}}{=} \frac{C_{xy}(\Delta t)}{\sqrt{A_x(0)\, A_y(0)}},$$

where the rightmost equations assume stationarity.

It is seen from equation 2.19 that, if $\Delta t = 0$, the autocorrelation is 1. Otherwise it is always in the interval $[-1, +1]$, as was already shown in connection with equation 2.6. Also the cross-correlation is always in the interval $[-1, +1]$, but generally $\mathrm{Corr}_{xy}(0) \neq 1$!

2.5.4 Stationary processes centred on zero

Consider stationary processes centred on zero:

$$E\{\underline{x}(t)\} = E\{\underline{y}(t)\} = 0.$$

Then, the following useful integral equation exists for the cross-covariance function:

$$C_{xy}(\Delta t) = E\{\underline{x}(t)\,\underline{y}(t+\Delta t)\} = \frac{1}{T}E\left\{\int_0^T \underline{x}(t)\,\underline{y}(t+\Delta t)\,dt\right\}. \quad (2.20)$$

What this says is that we can calculate the integral expression, the average over time

$$\frac{1}{T}\int_0^T x(t)\,y(t+\Delta t)\,dt$$

from observations of the processes $\underline{x}(t)$ and $\underline{y}(t)$ within the finite time interval $[0, T]$, and use it as an *unbiased estimator* of $C_{xy}(\Delta t)$. In practice, one would sample the process discretely as a time series.

Here, a special case of the law of large numbers applies:

$$T \to \infty \implies \frac{1}{T}\int_0^T x(t)\,y(t+\Delta t)\,dt \overset{\text{stoch}}{\to} C_{xy}(\Delta t).$$

Note that in this calculation, we are only using a single realisation of processes $\underline{x}(t)$ and $\underline{y}(t)$, not an ensemble of realisations. That this is allowed is called *ergodicity*.

2.5.5 Cross-covariance and convolution

The circular convolution integral is defined as

$$x \underset{T}{\otimes} y = \left(x \underset{T}{\otimes} y\right)(t) \overset{\text{def}}{=} \oint_0^T x(t-t')\,y(t')\,dt'. \quad (2.21)$$

In the definition it is assumed that the functions x and y are periodic with a period T. Therefore, an equivalent definition is

$$x \underset{T}{\otimes} y = \oint_{-T/2}^{T/2} x(t-t')\,y(t')\,dt'.$$

This circular convolution approaches the ordinary convolution in the limit $T \to \infty$:

$$x \otimes y = \int_{-\infty}^{\infty} x(t-t')\,y(t')\,dt'.$$

In practice, for large T, they may be taken as equivalent and the requirement of periodicity is relaxed.

The integral in 2.20 is visibly similar to the circular convolution integral 2.21. Defining $\tau \stackrel{\text{def}}{=} t' - t$ we obtain (circular is circular!):

$$\left(x \underset{T}{\otimes} y\right)(t) = \oint_0^T x(-\tau)\, y(t+\tau)\, dt' = \oint_0^T x(-\tau)\, y(t+\tau)\, d\tau.$$

Substitute $\tau \to t$ and $t \to \Delta t$:

$$\left(x \underset{T}{\otimes} y\right)(\Delta t) = \int_0^T x(-t)\, y(t+\Delta t)\, dt.$$

This equals the integral in equation 2.20, save for a minus sign. It follows that

$$C_{xy}(\Delta t) = \frac{1}{T} E\left\{x(-t) \underset{T}{\otimes} y(t)\right\}.$$

In the general theory with complex valued stochastic processes, the cross-covariance, equation 2.18, is defined as

$$C_{xy}(\Delta t) \stackrel{\text{def}}{=} \operatorname{Cov}\{\underline{x}^{\dagger}(t), \underline{y}(t+\Delta t)\},$$

liittoluku in which the symbol $\dagger$ stands for the complex conjugate. Then integral 2.20 becomes

$$C_{xy}(\Delta t) = \frac{1}{T} E\left\{\int_0^T x^{\dagger}(t)\, y(t+\Delta t)\, dt\right\} = \frac{1}{T} E\left\{x^{\dagger}(-t) \underset{T}{\otimes} y(t)\right\}.$$

Of course the autocovariance function is then similarly

$$A_x(\Delta t) = \operatorname{Cov}\{\underline{x}^{\dagger}(t), \underline{x}(t+\Delta t)\} = \frac{1}{T} E\left\{x^{\dagger}(-t) \underset{T}{\otimes} x(t)\right\}. \qquad (2.22)$$

The circular convolution has the useful property that it can be efficiently calculated using the fast Fourier transform (FFT).

2.6 White noise and random walk

2.6.1 White noise

Noise may be defined as a stochastic process with an expected value of zero:

$$E\{\underline{n}(t)\} = 0.$$

12 *White noise* is noise that consists uniformly of all possible frequencies.[12] The mathematical way of describing this is to say that the autocovariance

$$A_n(\Delta t) = 0, \qquad \Delta t \neq 0.$$

[12]The name is based on the analogy with white light, which thanks to Newton (Davidson and Tchourioukanov) we know to consist of all the frequencies of visible light.

In other words, the process values $\underline{n}(t)$ and $\underline{n}(t')$ do not correlate at all, no matter how close $\Delta t = t - t'$ is to zero. Nevertheless

$$A_n(0) = \infty.$$

Furthermore it holds that

$$\int_{-\infty}^{+\infty} A_n(\tau)\, d\tau \stackrel{\text{def}}{=} Q_n,$$

with Q_n a finite real value.

The above equations deserve some reflection. This is a function $A_n(\tau)$ which is "almost everywhere" zero — namely everywhere that $\tau \neq 0$ — but in the only point where it is not zero — the point $\tau = 0$ — it is infinite! And furthermore, the integral of the function over its domain $\mathbb{R}$ produces the finite value Q_n!

2.6.2 *The delta function*

Such a function does not actually exist. There is however a mathematical device, the *delta function* or distribution $\delta(t)$, named after the quantum physicist Paul Dirac:[13]

delta $\delta\Delta$
13

$$A_n(\Delta t) = Q_n\, \delta(\Delta t). \tag{2.23}$$

Intuitively we can appreciate how such a "function" may be built, figure 2.3.

First, the following block function is defined of width ϵ:

epsilon $\varepsilon\epsilon E$

$$\delta_\epsilon(\tau) = \begin{cases} 0 & \text{if } \tau > \frac{1}{2}\epsilon \ \text{ or } \ \tau < -\frac{1}{2}\epsilon, \\ {}^1/_\epsilon & \text{if } -\frac{1}{2}\epsilon \leqslant \tau \leqslant \frac{1}{2}\epsilon. \end{cases}$$

Clearly, the integral of this function

$$\int_{-\infty}^{+\infty} \delta_\epsilon(\tau)\, d\tau = 1$$

and $\delta_\epsilon(\tau) = 0$ if $|\tau|$ is large enough.

Let $\epsilon \to 0$. In this limit, $\delta_\epsilon(0) \to \infty$, and for every value $\tau \neq 0$ there is always a bounding value ϵ for which it holds that $|\tau| > \frac{1}{2}\epsilon \implies \delta_\epsilon(\tau) = 0$.

---

[13] Paul Adrien Marie Dirac FRS (1902–1984) was a leading British theoretical physicist and quantum theorist.

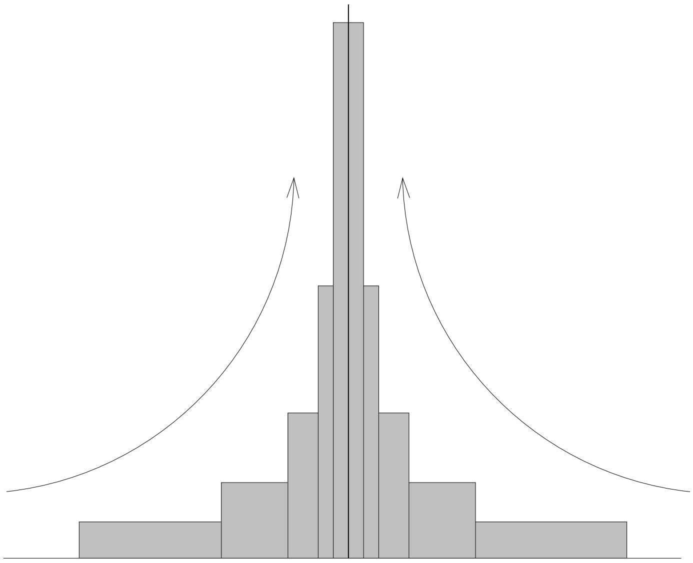


FIGURE 2.3. Dirac delta function as the limit of block functions.

The delta function has the important *reproducing property*, that for arbitrary functions f,

$$f(t) = \int_{-\infty}^{\infty} \delta(t-t')\, f(t')\, dt' \iff f = \delta \otimes f. \tag{2.24}$$

So, convolution with the delta function is the neutral operator on the function space, a little like multiplication with the unit matrix, that is the Kronecker delta tensor — the same symbol! — is on a vector space.

The handling rule for distributions is simply that first we integrate and only then we let in the result obtained $\epsilon \to 0$.

2.6.3 *Random walk*

A "random walk" is obtained when white noise is integrated over time:

$$\frac{d}{dt}\underline{w}(t) = \underline{n}(t), \tag{2.25}$$

in which $\underline{w}(t)$ is the random walk and $\underline{n}(t)$ is white noise.

Let the autocovariance function of the noise $\underline{n}$ be

$$A_n(\Delta t) = Q_n\, \delta(\Delta t), \tag{2.23}$$

in which Q_n is the variance of the noise $\underline{n}$ as defined above.

14 Then we integrate this function, obtaining a *random walk* or Wiener[14]
15 process:[15]

[14]Norbert Wiener (1894–1964) was a Jewish American mathematician and philosopher and the founder of cybernetics.

$$\underline{w}_0(t) = \int_{t_0}^{t} \underline{n}(\tau)\, d\tau.$$

We include a reference to the starting time of integration t_0 as a subscript 0, as $\underline{w}$ depends on it.

Note that

$$E\{\underline{w}_0(t)\} = \int_{t_0}^{t} E\{\underline{n}(\tau)\}\, d\tau = 0.$$

The autocovariance function is obtained as

$$A_{w,0}(t,t') = E\left\{\left(\underline{w}_0(t) - E\{\underline{w}_0(t)\}\right)\left(\underline{w}_0(t') - E\{\underline{w}_0(t')\}\right)\right\} =$$
$$= E\{\underline{w}_0(t)\,\underline{w}_0(t')\} =$$
$$= E\left\{\int_{t_0}^{t} \underline{n}(\tau)\, d\tau \int_{t_0}^{t'} \underline{n}(\tau')\, d\tau'\right\} = \int_{t_0}^{t} \overbrace{\int_{t_0}^{t'} E\{\underline{n}(\tau')\,\underline{n}(\tau)\}\, d\tau'}^{I}\, d\tau.$$

Here, the integral[16] 16

$$I = \int_{t_0}^{t'} E\{\underline{n}(\tau')\,\underline{n}(\tau)\}\, d\tau' = \int_{t_0}^{t'} A_n(\tau - \tau')\, d\tau' =$$
$$= Q_n \int_{t_0}^{t'} \delta(\tau - \tau')\, d\tau' = \begin{cases} Q_n & \text{if } t' > \tau, \\ \frac{1}{2} Q_n & \text{if } t' = \tau, \\ 0 & \text{if } t' < \tau. \end{cases}$$

It follows that

$$A_{w,0}(t,t') = Q_n \int_{t_0}^{t} \left(\int_{t_0}^{t'} \delta(\tau - \tau')\, d\tau'\right) d\tau =$$
$$= Q_n \cdot (t' - t_0) + 0 \cdot (t - t') = Q_n \cdot (t' - t_0). \qquad (2.26)$$

Here it has been assumed that the autocovariance of the noise function $\underline{n}$ is *stationary*, in other words, that Q_n is a constant. This can be generalised to the case in which $Q_n(t)$ is a function of time:

$$A_{w,0}(t,t') = \int_{t_0}^{t'} Q_n(\tau)\, d\tau. \qquad (2.27)$$

In both equations, 2.26 and 2.27, it is assumed that $t' \leqslant t$.

It is important to note that a random-walk process is *non-stationary*: the variance

$$\mathrm{Var}\{\underline{w}_0(t)\} = A_{w,0}(t,t) = Q_n \cdot (t - t_0)$$

will grow without bound with time t.

[15]Strictly speaking, a Wiener process is the special case which is normally distributed with $Q_n = 1$.

[16]We assume that the delta function has been defined symmetrically, so $\int_{-\infty}^{0} \delta(\tau)\, d\tau = \int_0^{\infty} \delta(\tau)\, d\tau = \frac{1}{2}$.

2.7 Coloured noise

Let us study next the simple differential equation in time

$$\frac{d}{dt}\underline{x}(t) = -k\,\underline{x}(t) + \underline{n}(t), \tag{2.28}$$

in which $\underline{n}(t)$ is white noise, of which the autocovariance function is $Q_n\,\delta(\Delta t)$. Q_n and k are constants.

17 The process thus described is called a *stationary Gauss-Markov*[17] *process*. Compared to the random-walk equation 2.25, this contains a term $-k\,\underline{x}(t)$, an influence driving $\underline{x}$ back towards zero against the dispersing influence of the noise $\underline{n}$.

The solution of this differential equation is

$$\underline{x}(t) = e^{-kt}\left(\underline{x}(t_0)\,e^{kt_0} + \int_{t_0}^{t} \underline{n}(\tau)\,e^{k\tau}\,d\tau\right) \tag{2.29}$$

as can be verified by substitution: differentiating equation 2.29 and applying the Leibniz product rule yields

$$\begin{aligned}\frac{d}{dt}\underline{x}(t) &= -k\,\overbrace{e^{-kt}\left\{\underline{x}(t_0)\,e^{kt_0} + \int_{t_0}^{t} \underline{n}(\tau)\,e^{k\tau}\,d\tau\right\}}^{\underline{x}(t)} + \\ &\qquad + e^{-kt}\frac{d}{dt}\left\{\int_{t_0}^{t} \underline{n}(\tau)\,e^{k\tau}\,d\tau\right\} = \\ &= -k\,\underline{x}(t) + e^{-kt}\,\underline{n}(t)\,e^{kt} = -k\,\underline{x}(t) + \underline{n}(t).\end{aligned}$$

The solution also satisfies the initial condition.

If we assume that the initial value $\underline{x}(t_0)$ is errorless and that the autocovariance function of $\underline{n}$ is

$$A_n(t, t') = Q_n\,\delta(t - t'),$$

we obtain from solution 2.29 the autocovariance function of $\underline{x}$:

$$\begin{aligned}A_x(t, t') &= e^{-k(t+t')}\,E\left\{\int_{t_0}^{t'} \underline{n}(\tau')\,e^{k\tau'}\,d\tau' \int_{t_0}^{t} \underline{n}(\tau)\,e^{k\tau}\,d\tau\right\} = \\ &= e^{-k(t+t')}\int_{t_0}^{t'} e^{k\tau'}\overbrace{\int_{t_0}^{t} E\left\{\underline{n}(\tau')\,\underline{n}(\tau)\right\} e^{k\tau}\,d\tau}^{\mathrm{I}}\,d\tau'.\end{aligned}$$

[17] Andrey Andreyevich Markov (1856–1922) was a Russian mathematician who did important work on stochastic processes.

The integral

$$I = \int_{t_0}^{t} E\{\underline{n}(\tau')\,\underline{n}(\tau)\}\, e^{k\tau}\, d\tau = \int_{t_0}^{t} A_n(\tau', \tau)\, e^{k\tau}\, d\tau =$$
$$= Q_n \int_{t_0}^{t} \delta(\tau - \tau')\, e^{k\tau}\, d\tau = \begin{cases} Q_n\, e^{k\tau'} & \text{if } t > \tau', \\ \frac{1}{2} Q_n\, e^{k\tau'} & \text{if } t = \tau', \\ 0 & \text{if } t < \tau'. \end{cases}$$

So, in the case $t' > t$:

$$A_x(t, t') = Q_n\, e^{-k(t+t')} \left(\int_{t_0}^{t} e^{2k\tau'}\, d\tau' + \int_{t}^{t'} 0\, d\tau' \right) =$$
$$= \frac{Q_n}{2k} e^{-k(t+t')} \left(e^{2kt} - e^{2kt_0} \right),$$

and in the case $t' < t$:

$$A_x(t, t') = Q_n\, e^{-k(t+t')} \int_{t_0}^{t'} e^{2k\tau'}\, d\tau' = \frac{Q_n}{2k} e^{-k(t+t')} \left(e^{2kt'} - e^{2kt_0} \right).$$

In both cases we obtain

$$A_x(t, t') = \frac{Q_n}{2k} \left(e^{-k|t-t'|} - e^{-k(t+t'-2t_0)} \right), \tag{2.30}$$

which is also valid for $t = t'$.

The situation in which $t, t' \gg t_0$, the stationary state into which the system will settle long after starting, when the initial state $\underline{x}(t_0)$ has been forgotten, yields

$$A_x(t, t') \stackrel{\text{def}}{=} A_x(t - t') \approx \frac{Q_n}{2k} e^{-k|t-t'|} = \frac{Q_n}{2k} e^{-k|\Delta t|} = A_x(0) e^{-k|\Delta t|}. \tag{2.31}$$

In this case one speaks of *coloured noise*,[18] and the process thus obtained is called a stationary Gauss-Markov process. 18

One also encounters the term first-order autoregressive or AR(1) process, although this usually refers to a discrete sequence or time series. Let a first-order autoregressive sequence be $\underline{x}_i$, in which each member only depends on the previous member of the sequence:

$$\underline{x}_{i+1} = \rho\, \underline{x}_i + \underline{n}_i, \tag{2.32}$$

[18] The name again uses the unequal brightness distribution with frequency of coloured light as a metaphor.

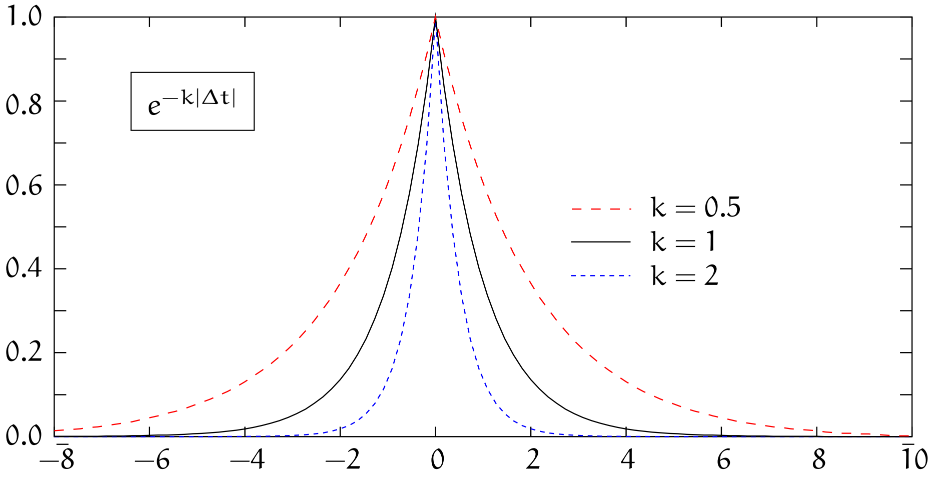


FIGURE 2.4. Autocovariance function of a stationary Gauss-Markov process as a function of the time difference $\Delta t = t - t'$. It is assumed that $A_x(0) = Q_n/2k = 1$.

in which $\underline{x}_i \stackrel{\text{def}}{=} \underline{x}(t_i)$ and $\underline{x}_{i+1} \stackrel{\text{def}}{=} \underline{x}(t_{i+1})$. When $\Delta t \stackrel{\text{def}}{=} t_{i+1} - t_i$ is small, comparing with equation 2.28 gives $\rho = e^{-k\,\Delta t} \approx 1 - k\,\Delta t$.

Equation 2.32 is an example of a *Markov chain*, in which each member $\underline{x}_{i+1}$ of the sequence, or chain, is derived only from, and can be optimally estimated using only, the previous member $\underline{x}_i$. The older members $\underline{x}_j$, $j < i$ are not involved. A Markov chain is thus *memoryless*. This is called the Markov property.

Define

$$\widetilde{Q}_n \stackrel{\text{def}}{=} \frac{1}{k^2} Q_n.$$

The surface area under the A_x curve is

$$\int_{-\infty}^{+\infty} A_x(\tau)\, d\tau = \frac{\widetilde{Q}_n k^2}{2k} \cdot \int_{-\infty}^{\infty} e^{-k|\tau|}\, d\tau = \frac{\widetilde{Q}_n k}{2} \cdot 2 \int_0^{\infty} e^{-k\tau}\, d\tau =$$
$$= -\widetilde{Q}_n \left[e^{-k\tau} \right]_0^{\infty} = \widetilde{Q}_n,$$

from which k has vanished.

- In the edge case $k \to \infty$ we make the autocovariance function A_x progressively narrower, but keep the surface area under the curve of the function the same. Then

$$\widetilde{A}_x(t - t') \to \widetilde{Q}_n\, \delta(t - t'). \tag{2.33}$$

Equation 2.33 corresponds to a degeneration of the equation 2.28, in which not only $k \to \infty$, but also the noise variance $Q_n \to \infty$.

Like this:

$$\frac{d}{dt}\underline{x}(t) = -k\underline{x}(t) + \overbrace{k\underline{\widetilde{n}}(t)}^{\underline{n}(t)} \implies \underline{x}(t) = \underline{\widetilde{n}}(t) - \frac{1}{k}\frac{d}{dt}\underline{x}(t).$$

In the limit $k \to \infty$ this goes to

$$\underline{x}(t) = \underline{\widetilde{n}}(t),$$

in which the "rescaled" noise

$$\underline{\widetilde{n}}(t) \stackrel{\text{def}}{=} \frac{1}{k}\underline{n}(t)$$

has a finite variance of

$$\widetilde{Q}_n = \frac{1}{k^2}Q_n,$$

consistent with the autocovariance equation 2.33.

- The other edge case, $k \to 0$, is *random walk*, already presented in subsection 2.6.3. It thus is a Gauss-Markov process with an infinitely long time constant. In that case one must use the whole equation 2.30:

$$A_{x,0}(t, t') = \frac{Q_n}{2k}\left(e^{-k|t-t'|} - e^{-k(t+t'-2t_0)}\right). \tag{2.30}$$

In this case, if $t > t'$, we get

$$A_{x,0}(t) = \frac{Q_n}{2k}\left(1 - e^{-2k(t-t_0)}\right) \approx Q_n \cdot (t - t_0),$$

the same as obtained in subsection 2.6.3.

The differential equation is obtained from equation 2.28 by substituting $k = 0$:

$$\frac{d}{dt}\underline{x}(t) = \underline{n}(t),$$

so $\underline{x}$ is the time integral of the white noise $\underline{n}$, equation 2.25.

Table 2.2 gives a summary of the main properties of the Gauss-Markov process and its edge cases.

Often, the model of equation 2.28 is used to produce a "coloured" noise process in cases where we know beforehand that the properties of the process are of that type. This is done by adding one unknown to the vector of unknowns and one equation to the system of equations. We shall see in section 4.4 how to do this with the Kalman filter.

TABLE 2.2. Summary of the properties of various stochastic processes.

Name	k	Equation	Autocovariance
Random walk	0	$\frac{d}{dt}\underline{x} = \underline{n}$	$Q_n \cdot (t - t_0)$
Gauss-Markov process	$\in (0, \infty)$	$\frac{d}{dt}\underline{x} = -k\underline{x} + \underline{n}$	$\frac{Q_n}{2k} e^{-k\lvert t-t'\rvert}$
White noise	∞	$\underline{x} = \underline{n}$	$Q_n\, \delta(t - t')$

2.8 Power spectral density (PSD)

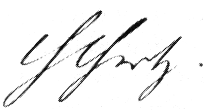
2.8.1 Definition

We often want to study stochastic processes in terms of their *spectrum*, in other words the presence of various frequency constituents in the process. This can be done by using the *Fourier*[19] *transform*.

19 tehon spektraalitiheys 20

For a stationary process, the Fourier transform of the autocovariance function is called the *power spectral density* function (PSD). We choose $t' = 0$, fixing the arbitrary origin on the time axis so that $\Delta t = t - t' = t$. This yields (Wiener–Khinchin[20]–Einstein theorem):

$$\mathcal{A}_x(f) \stackrel{\text{def}}{=} \mathcal{F}\{A_x(t)\} = \int_{-\infty}^{+\infty} A_x(t)\, e^{-2\pi i f t}\, dt, \tag{2.34}$$



21

in which $A_x(t) = A_x(\Delta t)$ (equation 2.16) and assuming that the integral exists. Here, f is the *frequency*, expressed in Hz (after Heinrich Hertz[21]), or cycles or oscillations or periods per second, or s^{-1}.

Analogically we may also define the cross-PSD of two functions:

$$\mathcal{C}_{xy}(f) \stackrel{\text{def}}{=} \mathcal{F}\{C_{xy}(t)\} = \int_{-\infty}^{+\infty} C_{xy}(t)\, e^{-2\pi i f t}\, dt.$$

The inverse operation using the inverse Fourier transform yields

$$A_x(\Delta t) = A_x(t) = \mathcal{F}^{-1}\{\mathcal{A}_x(f)\} = \int_{-\infty}^{+\infty} \mathcal{A}_x(f)\, e^{2\pi i f t}\, df.$$

Therefore, $t = 0$ gives

$$A_x(0) = \int_{-\infty}^{+\infty} \mathcal{A}_x(f)\, df,$$

[19]Joseph Fourier (1768–1830) was a French mathematician and physicist best known as the discoverer of Fourier series.

[20]Aleksandr Yakovlevich Khinchin (1894–1959) was a Soviet mathematician and contributor to probability theory.

[21]Heinrich Rudolf Hertz (1857–1894) was a German physicist, the first to generate and detect radio waves.

so the variance of process $\underline{x}$ is the same as the total surface area under its PSD curve, in other words the total power of the process.

The autocovariance function is symmetric, equation 2.17. It follows that the PSD, equation 2.34, is *real-valued* and also symmetric:

$$\mathcal{A}_x(f) = \int_{-\infty}^{+\infty} A_x(t) \left(\cos(-2\pi ft) + i \sin(-2\pi ft)\right) dt =$$
$$= 2\int_0^{\infty} A_x(t) \cos 2\pi ft \, dt.$$

In addition, it is non-negative everywhere, see appendix A.

For the cross-PSD this does not hold as it is not symmetric:

$$C_{xy}(t, t') = C_{yx}(t', t) \neq C_{xy}(t', t).$$

2.8.2 Gauss-Markov process

The autocovariance function of a Gauss-Markov process is given by equation 2.31:

$$A_x(\Delta t) = \frac{Q_n}{2k} e^{-k|\Delta t|}.$$

From this follows the PSD by integration according to equation 2.34, again choosing $t' = 0$ and hence $\Delta t = t$:

$$\begin{aligned}
\mathcal{A}_x(f) &= \int_{-\infty}^{+\infty} A_x(t)\, e^{-2\pi i f t}\, dt = \frac{Q_n}{2k}\int_{-\infty}^{+\infty} e^{-k|t|} e^{-2\pi i f t}\, dt = \\
&= \frac{Q_n}{2k}\int_0^{\infty}\left(e^{-kt-2\pi i f t} + e^{-kt+2\pi i f t}\right) dt = \\
&= \frac{Q_n}{2k}\left[\frac{1}{-k-2\pi i f} e^{(-k-2\pi i f)t} + \frac{1}{-k+2\pi i f} e^{(-k+2\pi i f)t}\right]_0^{\infty} = \\
&= \frac{Q_n}{2k}\frac{1}{k^2+4\pi^2 f^2}\begin{bmatrix} -k+2\pi i f & -k-2\pi i f \end{bmatrix}\begin{bmatrix} \left[e^{(-k-2\pi i f)t}\right]_0^{\infty} \\ \left[e^{(-k+2\pi i f)t}\right]_0^{\infty} \end{bmatrix} = \\
&= \frac{Q_n}{2k}\frac{1}{k^2+4\pi^2 f^2}\begin{bmatrix} -k & 2\pi i f \end{bmatrix}\begin{bmatrix} \left[e^{-kt}(2\cos 2\pi f t)\right]_0^{\infty} \\ \left[e^{-kt}(-2i\sin 2\pi f t)\right]_0^{\infty} \end{bmatrix} = \\
&= \frac{Q_n}{2k}\frac{1}{k^2+4\pi^2 f^2}\begin{bmatrix} -k & 2\pi i f \end{bmatrix}\begin{bmatrix} -2 \\ 0 \end{bmatrix} = \frac{Q_n}{4\pi^2 f^2 + k^2}.
\end{aligned}$$

We can also write

$$\mathcal{A}_x(f) = \frac{Q_n}{2k}\int_0^{\infty} e^{-kt}\left(e^{-2\pi i f t} + e^{2\pi f i t}\right) dt =$$
$$= \frac{Q_n}{k}\int_0^{\infty} e^{-kt} \cos 2\pi f t \, dt$$

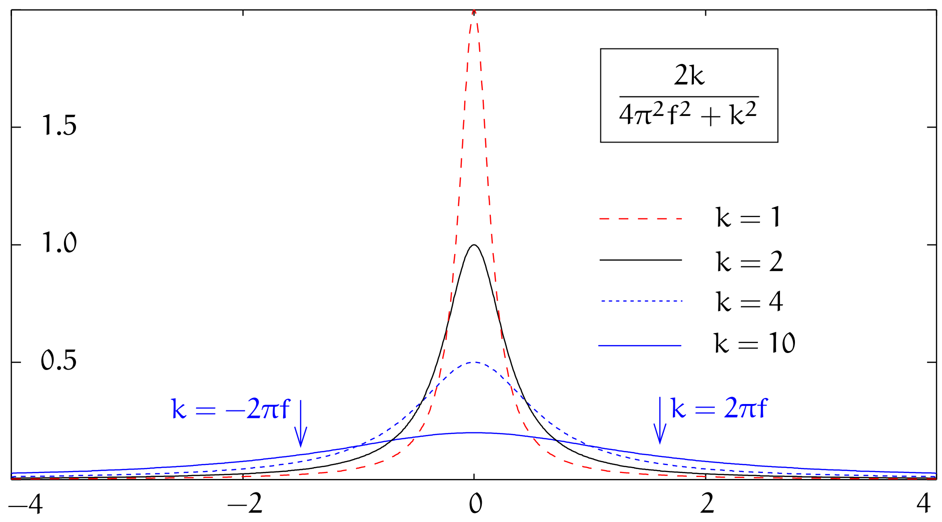


FIGURE 2.5. Power spectral density (PSD) of a Gauss-Markov process as a function of frequency. Assumed is $A_x(0) = 1$.

and try our luck with tabulations of integrals (Wolfram Functions, $\int e^{bx} \cos cx \, dx$), or with symbolic-algebra software, like the online integrator of Wolfram Research. The result is also[22]

$$\mathcal{A}_x(f) = \frac{Q_n}{4\pi^2 f^2 + k^2} = A_x(0) \frac{2k}{4\pi^2 f^2 + k^2}. \tag{2.35}$$

See Jekeli (2001) equation (6.75). Figure 2.5 plots the values of this function for $Q_n = 2k$, that is keeping the variance of $\underline{x}$, which equals $A_x(0) = Q_n/2k$, at unity regardless of k. This again keeps the surface area under the curve constant.

kantafunktio

We need to explain here the occurrence of negative values of the frequency f. In the Fourier integral used, the basis functions are complex, of type $e^{2\pi i f t}$. But we can write

$$e^{2\pi i f t} = \cos 2\pi f t + i \sin 2\pi f t,$$

and for a negative frequency $-f$:

$$e^{-2\pi i f t} = \cos 2\pi f t - i \sin 2\pi f t.$$

Addition and subtraction now yield

$$\cos 2\pi f t = \tfrac{1}{2}\left(e^{-2\pi i f t} + e^{2\pi i f t}\right), \quad \sin 2\pi f t = \tfrac{1}{2} i \left(e^{-2\pi i f t} - e^{2\pi i f t}\right).$$

Both of these are real-valued.

[22] An expression of this form is sometimes called a Cauchy-Lorentz distribution.

We thus see that for each (positive) frequency f there are *two independent Fourier basis functions*. The choice between either $\cos 2\pi f t$ and $\sin 2\pi f t$, or $e^{2\pi i f t}$ and $e^{-2\pi i f t}$, is free.

2.8.3 Squared-exponential autocovariance

The autocovariance function of a Gauss-Markov process is given by equation 2.31:

$$A_x(\Delta t) = A_x(0)\exp(-k\,|\Delta t|).$$

Often, a similar autocovariance function is encountered that takes the form

$$A_x(\Delta t) = A_x(0)\exp\left(-k^2\Delta t^2\right),$$

called a squared-exponential or Gaussian autocovariance function.

The power spectral density, again with $t' = 0$ and hence $\Delta t = t$, will be:

$$\mathcal{A}_x(f) = \int_{-\infty}^{+\infty} A_x(t)\exp(-2\pi i f t)\,dt = 2\int_0^{+\infty} A_x(t)\cos 2\pi f t\,dt =$$
$$= 2A_x(0)\int_0^{+\infty}\exp\left(-k^2t^2\right)\cos 2\pi f t\,dt = A_x(0)\frac{\sqrt{\pi}}{k}\exp\left(-\frac{\pi^2 f^2}{k^2}\right),$$

(Wolfram Functions, $\int \exp bx^2 \cos cx\,dx$). We see that the power spectral density function is of the same squared-exponential form as the autocovariance function, but with frequency f as the function argument.

Processes with the Gaussian autocovariance function have found some popularity in the theory of machine learning. Under the bonnet, however, the theoretical basis appears to be not very different from that of more traditional least-squares collocation (Heiskanen and Moritz, 1967, chapter 7) also known as kriging.

2.8.4 White noise

The PSD of white noise may be computed using expression 2.23:

$$A_n(\Delta t) = Q_n\,\delta(\Delta t) = Q_n\,\delta(t - t'),$$

which by choosing $t' = 0$ yields

$$\mathcal{A}_n(f) = \int_{-\infty}^{+\infty} Q_n\,\delta(t)\,e^{-2\pi i f t}\,dt = Q_n e^0 = Q_n,$$

using the reproducing property of the delta function, equation 2.24. Here we see why a process with a Dirac delta type autocovariance function is called *white* noise: the power spectral density is a constant Q_n all over the spectrum, for all frequencies f, just like is the case for white light — at least within the optical window of the electromagnetic spectrum.

However, there is a problem: the total power of such a process integrated over the whole spectrum is

$$\int_{-\infty}^{\infty} \mathcal{A}_n(f)\, df = \int_{-\infty}^{\infty} Q_n\, df = \infty = A_n(0).$$

This "ultraviolet catastrophe" illustrates the physical impossibility of a true white-noise process and the unrealness of the Dirac delta function. The delta function should really only be used safely packaged inside some integral.

It is better to study the limit for $k \to \infty$ of the Gauss-Markov process 2.35. Let us calculate its total power:

$$\begin{aligned}\int_{-\infty}^{\infty} \mathcal{A}_x(f)\, df &= 2\int_0^{\infty} \mathcal{A}_x(f)\, df = 2\int_0^{\infty} A_x(0)\frac{2k}{4\pi^2 f^2 + k^2} df = \\ &= 2\, A_x(0) \int_0^{\infty} \frac{2k}{k^2\left(\left(\frac{2\pi f}{k}\right)^2 + 1\right)} \frac{k}{2\pi}\, d\left(\frac{2\pi f}{k}\right) = \\ &= 2\, A_x(0)\, \frac{1}{\pi}\left[\arctan \frac{2\pi f}{k}\right]_0^{\infty} = 2A_x(0)\frac{1}{\pi}\frac{\pi}{2} = A_x(0).\end{aligned}$$

We see again that total power always equals variance irrespective of the value of k. By equation 2.35:

$$\mathcal{A}_x(f) = A_x(0)\frac{2k}{4\pi^2 f^2 + k^2},$$

so for $|2\pi f| \ll k$:

$$\mathcal{A}_x(f) \approx \frac{2}{k} A_x(0)$$

and for $|2\pi f| \gg k$:

$$\mathcal{A}_x(f) \approx A_x(0)\frac{2k}{4\pi^2 f^2} \approx 0.$$

We see, as we also see by inspecting figure 2.5, that this process has a roughly flat spectrum within $|2\pi f| < k$, and around $|2\pi f| \approx k$ goes down smoothly to zero for the areas $|2\pi f| > k$. The cut-off points $k = \pm 2\pi f$ are marked in the figure: they are the half-height points of the curve. Furthermore, the area between the cut-off points contains half of the total signal power:

$$\begin{aligned}\int_{-k/2\pi}^{k/2\pi} \mathcal{A}_x(f)\, df &= 2A_x(0)\frac{1}{\pi}\left[\arctan \frac{2\pi f}{k}\right]_{f=0}^{k/2\pi} = \\ &= A_x(0)\frac{2}{\pi}\arctan 1 = A_x(0)\frac{2}{\pi}\frac{\pi}{4} = \tfrac{1}{2}A_x(0).\end{aligned}$$

The larger is k, the lower and broader is the spectrum, but the total power remains at $A_x(0)$.

Self-test questions

1. What is the autocovariance of a stochastic process?
2. What is the cross-covariance of two stochastic processes?
3. When is a stochastic process stationary?
4. What is white noise?
5. What is a random walk?
6. Describe a Gauss-Markov process.
7. Show that random walk and white noise are limiting cases of Gauss-Markov.
8. What is a Markov chain?
9. What is power spectral density and how is it related to the autocovariance function?
10. What is homoscedasticity? What is *i.i.d.*?
11. How does linear regression of temporally correlated observations differ from that of uncorrelated observations?

Exercise 2–1: Normalisation of the normal distribution

Verify that the integral over the normal distribution, equation 2.2:

$$\int_{-\infty}^{+\infty} p(x)\, dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx,$$

produces 100 % total probability.

Hint Consider instead the two-dimensional stochastic variable

$$\underline{z} = \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix},$$

with joint probability density distribution $p(x, y) = p(x)\, p(y)$, and integrate in polar co-ordinates. Don't forget the determinant of Jacobi for polar co-ordinates. Now you understand where the 2π in the equation comes from.

Exercise 2–2: Effective sample size

Show that for the time series of subsection 2.4.2 it holds that

$$\mathrm{Var}\{\widehat{a}\} = \mathrm{Var}'\{\widehat{a}\}\,\frac{1+\rho}{1-\rho}.$$

3 The Kalman filter

The Kalman filter is a linear, predictive filter. Much like a coffee filter, which filters coffee from coffee-grounds, the Kalman filter filters the signal (the so-called *state vector*) from the noise of both the observation and the motion process. The Kalman filter is an optimal estimator in the least-squares sense.

The inventors of the Kalman filter were Rudolf Kalman[1] and Richard Bucy[2] in 1960–1961 (Kalman, 1960; Kalman and Bucy, 1961). The invention was extensively used in the space programme as well as in connection with missile guidance systems.

One important application of the Kalman filter was the *orbital rendezvous problem*: two spacecraft have to meet and exchange crew or materials while being either close together with only a small relative velocity or even mechanically docked. This problem, which nowadays is operational routine with the maintenance and servicing of the International Space Station, was considered a great challenge when the Apollo lunar programme was proposed. The technique of lunar orbit rendezvous (LOR), which won the day, played a key role in carrying through the Apollo Moon landings on schedule (Dickinson, 2014).

Nevertheless the Kalman filter is a generally applicable and broadly used estimation technique, not only in navigation but also, for example, in economics and meteorology.

The Kalman filter, with its origins in control theory and signal processing, has multiple links and commonalities with least-squares methods that originated in the geosciences, like adjustment, collocation

---

[1] Rudolf Emil Kálmán (1930–2016) was a Hungarian-born American electrical engineer, mathematician and inventor.

[2] Richard Snowden Bucy (1935–2019) was an American mathematician.

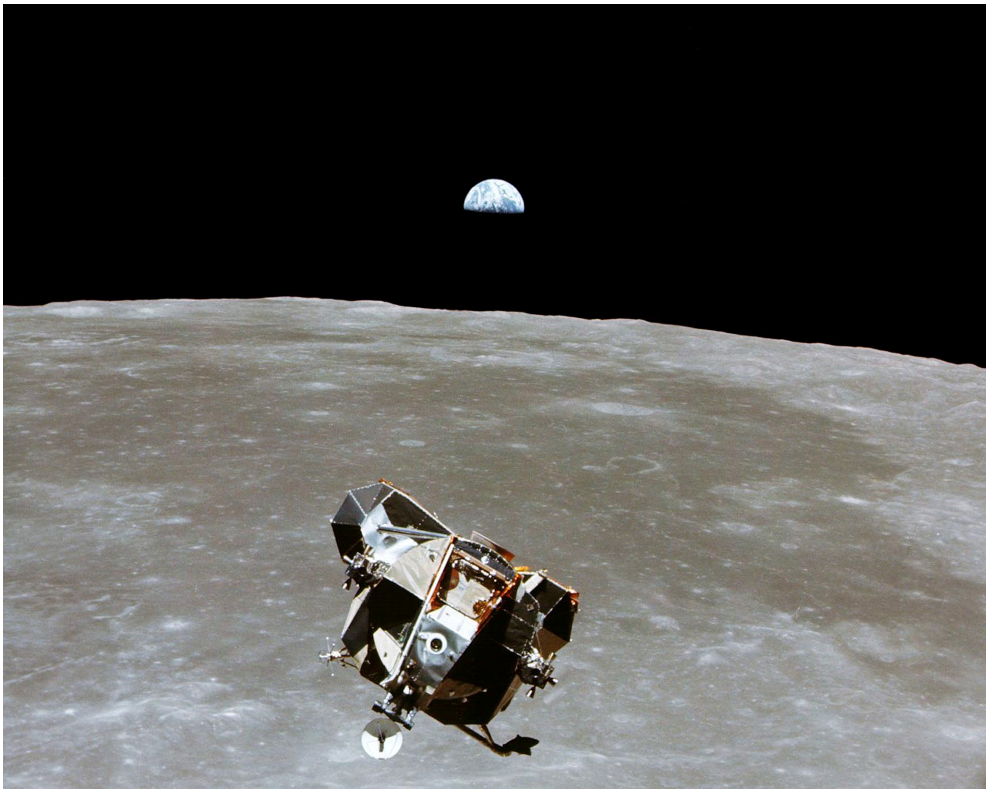


Figure 3.1. Lunar orbit rendezvous: two astronauts on their way home, bringing along a collection of rocks and film rolls. Note that both the mother ship — the Command and Service Module, CSM — and the lunar lander, seen here photographed from the mother ship, were equipped with inertial navigation systems. This saved the day during the Apollo 13 mission, when the Service Module was rendered inoperative by an on-board explosion.
Click for pronunciation (e-book).

and kriging. The differences in terminology should not obscure this.

For background reading, Strang and Borre (1997) pages 543–583 is recommended. A good link is Welch and Bishop, a good slide set Welch and Bishop (2001).

The Kalman filter consists of two parts:

1. The *dynamic model*. It describes how the state vector evolves over time.
2. The *observation model*. It describes how observations are made which contain information on the state vector at the time of observation.

Both of these models contain stochasticity: the dynamic model also describes random effects on the development of the system over time, like perturbations of a satellite orbit, while the observation model also describes the effect of observation uncertainty. ratahäiriö

The Kalman filter is special in the sense that the state vector propagates forwards in time step by step, and the observations are used to correct the state-vector estimate at times when observations have been made. Because of this, the Kalman filter does not demand high number-crunching power or the handling of large matrices. It can be implemented on-board a vehicle to be used in real time. tosiaikaisesti

3.1 The state vector

The state vector is a formal vector, an element of an abstract vector space, that describes completely the state of a dynamic system, either at a given point in time or *epoch* t or as a function of time; in other words a stochastic process.

We demonstrate this by a concrete example. A particle moving freely in space has three location co-ordinates and three velocity components. Let an orthonormal basis in this space be $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$. Then we write the location vector $\mathbf{x}$ and the velocity vector $\mathbf{v}$ of the particle as ortonormaali kanta

$$\mathbf{x} \stackrel{\text{def}}{=} x\mathbf{i} + y\mathbf{j} + z\mathbf{k}, \qquad \mathbf{v} \stackrel{\text{def}}{=} \dot{\mathbf{x}} = \dot{x}\mathbf{i} + \dot{y}\mathbf{j} + \dot{z}\mathbf{k}.$$

A vector in space is often identified with the formal vector of its three components on the agreed orthonormal basis $\beta \stackrel{\text{def}}{=} \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$: beta βB

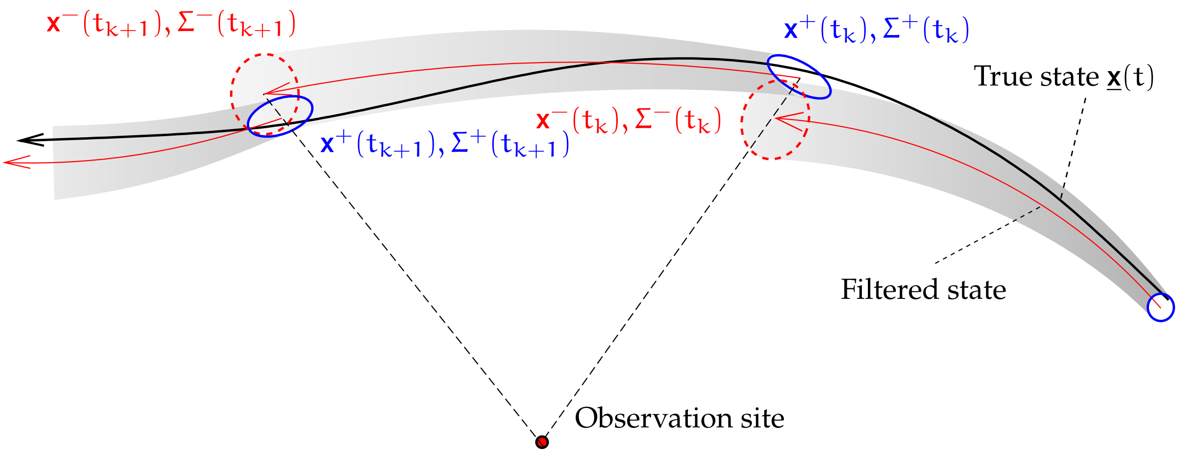
$$\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \iff \mathbf{x}_\beta \stackrel{\text{def}}{=} \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

although $\mathbf{x}$ and $\mathbf{x}_\beta$ are conceptually different things. We often semi-carelessly leave the β off when the meaning is clear, so

$$\mathbf{x}_\beta \sim \mathbf{x}.$$

Now the *state vector* of the particle becomes

$$\mathbf{x} = \begin{bmatrix} \mathbf{x} \\ \mathbf{v} \end{bmatrix} = \left[\begin{array}{c} x \\ y \\ z \\ \hline \dot{x} \\ \dot{y} \\ \dot{z} \end{array}\right], \tag{3.1}$$



(a)

Kalman filter

$$\frac{d}{dt}\underline{\mathbf{x}}(t) = F(\underline{\mathbf{x}}, t) + \underline{\mathbf{n}}(t) \overset{\text{lin}}{\approx} F(t)\,\underline{\mathbf{x}}(t) + \underline{\mathbf{n}}(t),$$

$$\frac{d}{dt}\mathbf{x}^-(t) = F(\mathbf{x}^-, t) \overset{\text{lin}}{\approx} F(t)\,\mathbf{x}^-(t),$$

$$\frac{d}{dt}\Sigma^-(t) = F(t)\,\Sigma^-(t) + \Sigma^-(t)\,F^T(t) + Q_n(t).$$

$$\underline{\mathbf{x}}(t_{k+1}) = \Phi_k^{k+1}\underline{\mathbf{x}}(t_k) + \underline{\mathbf{w}}_k^{k+1}, \qquad \underline{\mathbf{w}}_k^{k+1} = \int_k^{k+1}\Phi_t^{k+1}\underline{\mathbf{n}}(t)\,dt,$$

$$\mathbf{x}^-(t_{k+1}) = \Phi_k^{k+1}\,\mathbf{x}^+(t_k), \qquad \frac{d}{dt}\Phi_0^t = F(t)\,\Phi_0^t,$$

$$\Sigma^-(t_{k+1}) = \Phi_k^{k+1}\Sigma^+(t_k)\left(\Phi_k^{k+1}\right)^T + \Theta_k^{k+1}, \qquad \Theta_k^{k+1} = \int_k^{k+1}\Phi_t^{k+1}Q_n(t)\left(\Phi_t^{k+1}\right)^T dt.$$

(b)

Dynamic model

$$\underline{\ell}_k = H(\underline{\mathbf{x}}(t_k)) + \underline{\mathbf{m}}_k \overset{\text{lin}}{\approx} H_k\,\underline{\mathbf{x}}(t_k) + \underline{\mathbf{m}}_k, \qquad \operatorname{Var}\{\underline{\mathbf{m}}_k\} = R_k,$$

$$\mathbf{x}^+(t_k) = \mathbf{x}^-(t_k) - K_k\cdot\left(H(\mathbf{x}^-(t_k)) - \underline{\ell}_k\right) \overset{\text{lin}}{\approx} \mathbf{x}^-(t_k) - K_k\cdot\left(H_k\,\mathbf{x}^-(t_k) - \underline{\ell}_k\right),$$

$$\Sigma^+(t_k) = (I - K_k H_k)\,\Sigma^-(t_k), \qquad K_k = \Sigma^-(t_k)\,H_k^T\cdot\left(H_k\,\Sigma^-(t_k)\,H_k^T + R_k\right)^{-1}.$$

(c)

Observation model and update step

FIGURE 3.2. The Kalman filter, equations summary. Note the convention used that, in integral bounds or state transitions, we abbreviate t_k to k, t_{k+1} to $k+1$, and so on.

in which the location and velocity vectors are

$$\boldsymbol{x} = \boldsymbol{x}_\beta = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \boldsymbol{v} = \boldsymbol{v}_\beta = \dot{\boldsymbol{x}}_\beta = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix}.$$

In this case the state vector is seen to have six elements or *degrees of freedom*.

The whole state vector, each of its elements and the two subvectors are all functions of time:

$$\mathbf{x} = \mathbf{x}(t), \quad \boldsymbol{x} = \boldsymbol{x}(t), \quad \boldsymbol{v} = \boldsymbol{v}(t), \quad x = x(t), \quad y = y(t), \quad z = z(t).$$

If the particle is not a point but an extended object, its attitude angles or *Euler angles* also enter into the state vector. Then we already have nine elements. In a system of several particles every particle contributes its own elements, three location and three velocity components, to the state vector.

The state vector may also contain elements that model the behaviour of a mechanical device, like an inertial measurement unit.

3.2 The dynamic model

The dynamic model characterises the behaviour of the state vector in time. The state vector as defined in the previous section is a vector-valued stochastic process that is a function of time t.

The basic form of the dynamic model presented here is an *ordinary differential equation* in time of the state vector. There is an alternative form in which the change of the state vector over time is described in a discrete fashion, as a state transition from epoch t_k to epoch t_{k+1}. We will present this discrete form in section 3.4.

We present first the linear case. Next, we present the non-linear case applying linearisation, which yields the so-called *extended Kalman filter*.

3.2.1 *The linear case*

In the linear case the dynamic model looks like

$$\frac{d}{dt}\underline{\mathbf{x}}(t) = F(t)\,\underline{\mathbf{x}}(t) + \underline{\mathbf{n}}(t), \tag{3.2}$$

in which $\underline{\mathbf{x}}$ is the state vector and $\underline{\mathbf{n}}$ is the *dynamic noise* representing the random variability of the real motion. The underline denotes that a quantity is stochastic. Matrix F — which may be a function of time — is the coefficient matrix of the model.

The state vector has as many elements as are needed to fully describe the instantaneous state of the system. The dynamic noise has the same number of elements. The coefficient matrix is a square matrix which also has this same number of both rows and columns.

Realistic *statistical attributes* have to be defined for the dynamic noise $\underline{\mathbf{n}}$: often it is assumed that it is white noise, the autocovariance of which is

$$A_n(\Delta t) = Q_n\,\delta(\Delta t). \tag{2.23}$$

This will make the randomness in the state vector $\underline{\mathbf{x}}$ behave in a way resembling random walk, subsection 2.6.3. If the state vector has more than one element, both A_n and Q_n will be square matrices of the same size.

3.2.2 Linearisation

The more general non-linear case, called in the literature the *extended* Kalman filter, is

$$\frac{d}{dt}\underline{\mathbf{x}}(t) = \mathsf{F}(\underline{\mathbf{x}}, t) + \underline{\mathbf{n}}(t), \tag{3.3}$$

in which $\mathsf{F}(\underline{\mathbf{x}}, t)$ is a vectorial function. From this, the linear case easily

vertausarvo

follows: choose an *approximate* or *reference value* $\underline{\mathbf{x}}^{(0)}(t)$, a function of time, for the state vector. We demand from this approximate value consistency with the dynamic model:

$$\frac{d}{dt}\underline{\mathbf{x}}^{(0)}(t) = \mathsf{F}\left(\underline{\mathbf{x}}^{(0)}, t\right). \tag{3.4}$$

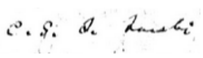
Now we linearise by subtracting equation 3.4 from equation 3.3 and doing a Taylor expansion:

$$\frac{d}{dt}\left(\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(0)}\right) = \mathsf{F}(\underline{\mathbf{x}}, t) + \underline{\mathbf{n}}(t) - \mathsf{F}\left(\underline{\mathbf{x}}^{(0)}, t\right) \approx F(t)\,\left(\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(0)}\right) + \underline{\mathbf{n}}(t),$$

which, with the definition $\Delta\underline{\mathbf{x}} \stackrel{\text{def}}{=} \underline{\mathbf{x}} - \underline{\mathbf{x}}^{(0)}$, is already of form 3.2:

$$\frac{d}{dt}\Delta\underline{\mathbf{x}}(t) = F(t)\,\Delta\underline{\mathbf{x}}(t) + \underline{\mathbf{n}}(t).$$

From this one may again drop the deltas if there is no risk of misunderstanding.



The elements of the Jacobi[3] matrix F of the function $\mathsf{F}(\cdot)$ used above

3

are

$$F_{ij}(t) = \frac{\partial}{\partial x_j} F_i(\overbrace{x_1, \ldots, x_j, \ldots, x_n}^{\mathbf{x}}, t)\Bigg|_{\mathbf{x}=\mathbf{x}^{(0)}}, \qquad i, j = 1, \ldots, n, \qquad (3.5)$$

in which the x_j are the n components of $\mathbf{x}$. In the example state vector 3.1 they are

$$\mathbf{x} = \left[\begin{array}{ccc|ccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array}\right]^{\mathsf{T}} = \left[\begin{array}{ccc|ccc} x & y & z & \dot{x} & \dot{y} & \dot{z} \end{array}\right]^{\mathsf{T}}.$$

The derivatives in equation 3.5 are evaluated at the approximate or reference values given in the vector

$$\mathbf{x}^{(0)} = \left[\begin{array}{ccc|ccc} x^{(0)} & y^{(0)} & z^{(0)} & \dot{x}^{(0)} & \dot{y}^{(0)} & \dot{z}^{(0)} \end{array}\right]^{\mathsf{T}}.$$

The function $F(\mathbf{x}, t)$ itself is a *vector* with n components, $F_i(\mathbf{x}, t)$, $i = 1, \ldots, n$.

3.2.3 State propagation

State propagation is done by integrating the differential equation that is the dynamic model. In the general case, the model is

$$\frac{d}{dt}\underline{\mathbf{x}}(t) = F(\underline{\mathbf{x}}, t) + \underline{\mathbf{n}}(t). \qquad (3.3)$$

In calculating state evolution, we are not integrating the actual state vector $\underline{\mathbf{x}}(t)$, as we do not know it. We can only integrate the *state estimator* $\mathbf{x}^-(t)$. The noise term $\underline{\mathbf{n}}$ is left out in the absence of actual knowledge: its best estimator is zero.

So, impose the same dynamic law on the estimator:

$$\frac{d}{dt}\mathbf{x}^-(t) = F(\mathbf{x}^-, t),$$

and the integration starts from some sensible initial state $\mathbf{x}^-(t_0)$. If the state estimator is close enough to the approximate state $\mathbf{x}^{(0)}(t)$, equation 3.4, we may use the same linearisation as for the true state and arrive at the linearised equation

$$\frac{d}{dt}\left(\mathbf{x}^- - \mathbf{x}^{(0)}\right) = F(t)\,\left(\mathbf{x}^- - \mathbf{x}^{(0)}\right) \implies \frac{d}{dt}\Delta\mathbf{x}^-(t) = F(t)\,\Delta\mathbf{x}^-(t).$$

[3]Carl Gustav Jacob Jacobi (1804–1851) was a Jewish German mathematician mainly remembered for his theory of elliptic functions. He died of smallpox at only 46 years of age.

From this, the deltas may again be dropped:

$$\frac{d}{dt}\mathbf{x}^-(t) = F(t)\,\mathbf{x}^-(t). \tag{3.6}$$

In the linear case, the equation for the true state is

$$\frac{d}{dt}\underline{\mathbf{x}}(t) = F(t)\,\underline{\mathbf{x}}(t) + \underline{\mathbf{n}}(t). \tag{3.2}$$

Equation 3.6 for the estimator is the same equation but without the noise.

Notation used:

$\mathbf{x}^-$ — *a priori* state estimator, the estimator before the update step, to be described later. Also used for the generic state estimator.

$\mathbf{x}^+$ — *a posteriori* state estimator, after the update step.

Notations like $\widehat{\mathbf{x}}^{k-1}$ and $\widehat{\mathbf{x}}^{k}$ can also be found in the literature, with k the sequence number of the update. The "hat" symbol denotes an estimator.

3.3 Example: satellite motion

An example is offered by the motion of a spacecraft in the Earth's gravitational field. The field is approximated by that of a point mass $GM_\oplus$:

$$\frac{d^2}{dt^2}\begin{bmatrix}\underline{x}\\ \underline{y}\\ \underline{z}\end{bmatrix} = -\frac{GM_\oplus}{\left(\underline{x}^2+\underline{y}^2+\underline{z}^2\right)^{3/2}}\begin{bmatrix}\underline{x}\\ \underline{y}\\ \underline{z}\end{bmatrix} + \begin{bmatrix}\underline{n}_x\\ \underline{n}_y\\ \underline{n}_z\end{bmatrix},$$

in which $\underline{n}_x, \underline{n}_y, \underline{n}_z$ describe perturbations like the randomly varying effect of atmospheric drag or the irregularities of the Earth's gravitational field, which make the spacecraft's motion slightly unpredictable.

This is a system of second-order differential equations. The state vector is extended by adding the *velocity components* to it, yielding a system of first-order equations:

$$\frac{d}{dt}\overbrace{\begin{bmatrix}\underline{x}\\ \underline{y}\\ \underline{z}\\ \hline \underline{\dot{x}}\\ \underline{\dot{y}}\\ \underline{\dot{z}}\end{bmatrix}}^{\underline{\mathbf{x}}(t)} = \overbrace{\begin{bmatrix}\begin{bmatrix}\underline{\dot{x}}\\ \underline{\dot{y}}\\ \underline{\dot{z}}\end{bmatrix}\\ -\dfrac{GM_\oplus}{\left(\underline{x}^2+\underline{y}^2+\underline{z}^2\right)^{3/2}}\begin{bmatrix}\underline{x}\\ \underline{y}\\ \underline{z}\end{bmatrix}\end{bmatrix}}^{F(\underline{\mathbf{x}},t)} + \overbrace{\begin{bmatrix}0\\ 0\\ 0\\ \hline \underline{n}_x\\ \underline{n}_y\\ \underline{n}_z\end{bmatrix}}^{\underline{\mathbf{n}}(t)}.$$

This system of equations is non-linear. Linearisation yields

$$\frac{d}{dt}\overbrace{\left[\begin{array}{c}\Delta\underline{x}\\ \Delta\underline{y}\\ \Delta\underline{z}\\ \hline \Delta\underline{\dot{x}}\\ \Delta\underline{\dot{y}}\\ \Delta\underline{\dot{z}}\end{array}\right]}^{\Delta\underline{\mathbf{x}}(t)} \approx \overbrace{\left[\begin{array}{c|c} \begin{matrix}0&0&0\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}1&0&0\\0&1&0\\0&0&1\end{matrix} \\ \hline \frac{GM_\oplus}{r^5}\begin{pmatrix}3x^2-r^2&3xy&3xz\\3yx&3y^2-r^2&3yz\\3zx&3zy&3z^2-r^2\end{pmatrix} & \begin{matrix}0&0&0\\0&0&0\\0&0&0\end{matrix}\end{array}\right]^{(0)}}^{F(t)} \overbrace{\left[\begin{array}{c}\Delta\underline{x}\\ \Delta\underline{y}\\ \Delta\underline{z}\\ \hline \Delta\underline{\dot{x}}\\ \Delta\underline{\dot{y}}\\ \Delta\underline{\dot{z}}\end{array}\right]}^{\Delta\underline{\mathbf{x}}(t)} + \overbrace{\left[\begin{array}{c}0\\0\\0\\ \hline \underline{n}_x\\ \underline{n}_y\\ \underline{n}_z\end{array}\right]}^{\underline{\mathbf{n}}(t)}, \tag{3.7}$$

in which $r \stackrel{\text{def}}{=} \sqrt{x^2+y^2+z^2}$ is the distance from the centre of the Earth.

Other assumptions are:

- There is a suitable *approximate* or *reference orbit*, a set of approximate or reference values forming an approximate state vector vertausrata

  $$\mathbf{x}^{(0)}(t) = \left[\begin{array}{ccc|ccc} x^{(0)}(t) & y^{(0)}(t) & z^{(0)}(t) & \dot{x}^{(0)}(t) & \dot{y}^{(0)}(t) & \dot{z}^{(0)}(t)\end{array}\right]^\mathsf{T},$$

  relative to which the delta quantities in equation 3.7 have been calculated:

  $$\Delta\underline{\mathbf{x}}(t) \stackrel{\text{def}}{=} \underline{\mathbf{x}}(t) - \mathbf{x}^{(0)}(t).$$

  These are all functions of time.
- The elements of the coefficient matrix $F(t)$ are evaluated using these approximate values, equation 3.5.

Each element in the approximate state vector is a function of time, as is the vector itself. The vector is an exact solution of the non-linear dynamic model, equation 3.4.

3.3.1 *The gravitation-gradient tensor*

The partitioned version of dynamic equation 3.7 above is

$$\frac{d}{dt}\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\end{bmatrix} = \begin{bmatrix}0 & I\\ \mathcal{M}^{(0)} & 0\end{bmatrix}\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\end{bmatrix} + \begin{bmatrix}0\\ \underline{\mathbf{n}}\end{bmatrix}, \tag{3.8}$$

in which I is the size 3×3 unit matrix or Kronecker[4] tensor, and $\mathcal{M}$ is a tensor called the gravitation-gradient tensor. The component matrix of the tensor, on the basis $\beta = \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ of the (x, y, z) co-ordinate frame, is kanta

[4]Leopold Kronecker (1823–1891) was a German mathematician who contributed to number theory and algebra.

$$\mathcal{M}_\beta = \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{bmatrix} \left(-\frac{GM_\oplus}{r^3} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) =$$

$$= \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{bmatrix} \frac{GM_\oplus}{r} =$$

$$= \begin{bmatrix} \frac{\partial^2}{\partial x^2} & \frac{\partial^2}{\partial x \partial y} & \frac{\partial^2}{\partial x \partial z} \\ \frac{\partial^2}{\partial y \partial x} & \frac{\partial^2}{\partial y^2} & \frac{\partial^2}{\partial y \partial z} \\ \frac{\partial^2}{\partial z \partial x} & \frac{\partial^2}{\partial z \partial y} & \frac{\partial^2}{\partial z^2} \end{bmatrix} \frac{GM_\oplus}{r} =$$

$$= \frac{GM_\oplus}{r^5} \begin{bmatrix} 3x^2 - r^2 & 3xy & 3xz \\ 3yx & 3y^2 - r^2 & 3yz \\ 3zx & 3zy & 3z^2 - r^2 \end{bmatrix}. \quad (3.9)$$

Here, a central, spherically symmetric gravitational field is assumed.

The component matrix of the tensor consists of partial derivatives with respect to the place of the components

$$-\frac{GM_\oplus}{r^3} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

of the gravitation or attraction vector. Remember that the gravitation vector is itself the *gradient of the geopotential*. The tensor contains all second partial derivatives with respect to the place of the geopotential

$$\frac{GM_\oplus}{r}.$$

häiriö The tensor $\mathcal{M}$ describes the way in which a small perturbation $\Delta\mathbf{x} = \begin{bmatrix} \Delta x & \Delta y & \Delta z \end{bmatrix}^T$ in the satellite location translates into an acceleration perturbance $\Delta\mathbf{a}$:

$$\Delta\mathbf{a} = \begin{bmatrix} \Delta\ddot{x} \\ \Delta\ddot{y} \\ \Delta\ddot{z} \end{bmatrix} = \frac{d}{dt}\Delta\mathbf{v} = \frac{d}{dt} \begin{bmatrix} \Delta\dot{x} \\ \Delta\dot{y} \\ \Delta\dot{z} \end{bmatrix} = \mathcal{M}\,\Delta\mathbf{x} = \mathcal{M} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}.$$

3.3.2 Choosing approximate values

The most important thing when choosing the approximate or reference values — in addition to them being suitably close to the real values —

is that they be *physically consistent*; in other words, that they together present a genuinely possible orbital motion within the assumed gravitational field. For this reason it is perhaps more appropriate to call them *reference values*.

For a central gravitational field, suitable approximate values are given by the *Kepler orbital motion* around the centre of the attractive force $GM_\oplus$, see section 6.1.

An even simpler source of approximate values is *constant circular motion*. This is a good choice if the orbital eccentricity is close to zero.

If a model of the gravitational field is available that is more complicated than a central field, one must integrate the approximate values over time using this more accurate field model. Nevertheless the above linearised dynamic model, equation 3.7, will still be good for integrating the difference quantities $\Delta\mathbf{x}$ and $\Delta\mathbf{v}$, as long as these remain *numerically small*. This is one of the benefits of linearisation.

3.4 The discrete dynamic model

Definition:

> *The* state variance *is the expected square difference of the* estimator *of the state vector from the state vector's* true value, *as follows:*

$$\Sigma^-(t) = \operatorname{Var}\left\{\mathbf{x}^-(t)\right\} \stackrel{\text{def}}{=} E\left\{\left(\mathbf{x}^-(t) - \underline{\mathbf{x}}(t)\right)\left(\mathbf{x}^-(t) - \underline{\mathbf{x}}(t)\right)^{\mathsf{T}}\right\}. \quad (3.10)$$

The symbol used is Σ. sigma $\sigma\Sigma$

Of course it should be noted that the state vector $\underline{\mathbf{x}}(t)$ is a stochastic process *to which we do not have access*. We can only know values of its estimator $\mathbf{x}^-(t)$.

3.4.1 The state transition matrix

Assume that matrix F is constant. For a small time step $\Delta t = t_{i+1} - t_i$ it is approximately true that

$$\mathbf{x}^-(t_{i+1}) \approx \mathbf{x}^-(t_i) + F\,\Delta t \cdot \mathbf{x}^-(t_i) = (I + F\,\Delta t)\,\mathbf{x}^-(t_i).$$

This shows that the elements of the state vector $\mathbf{x}^-(t_{i+1})$ for epoch t_{i+1} are linear combinations of the elements of the state vector $\mathbf{x}^-(t_i)$ for the previous epoch t_i. If for an even smaller time step δt it holds that $\Delta t = n\,\delta t$, one obtains by repeatedly applying the above equation:

$$\mathbf{x}^-(t_{i+1}) = (I + F\,\delta t)^n\,\mathbf{x}^-(t_i).$$

phi φϕΦ The matrix Φ:

$$\Phi_i^{i+1} \stackrel{\text{def}}{=} (I + F\,\delta t)^n$$

is called the *state transition matrix* between epochs t_i and t_{i+1}, and we write

$$\mathbf{x}^-(t_{i+1}) = \Phi_i^{i+1}\mathbf{x}^-(t_i). \tag{3.11}$$

By substituting $\delta t = {}^{\Delta t}/_{n}$ we obtain

$$\Phi_i^{i+1} = \left(I + \frac{F\,\Delta t}{n}\right)^n.$$

For ordinary real numbers, we have the classical equation

$$e^x = \exp(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = \\ = \lim_{\nu\to\infty}\left(1 + \frac{1}{\nu}\right)^{\nu x} = \lim_{\nu\to\infty}\left(\left(1 + \frac{1}{\nu}\right)^{\nu}\right)^x,$$

in which we see the definition of the number e:

$$e = \lim_{\nu\to\infty}\left(1 + \frac{1}{\nu}\right)^{\nu}.$$

This is why we write, generalising the exp and ln functions to square matrices:

$$\Phi_i^{i+1} = \lim_{n\to\infty}\exp\left(\ln\left(I + \frac{F\,\Delta t}{n}\right)^n\right) = \lim_{n\to\infty}\exp\left(n\ln\left(I + \frac{F\,\Delta t}{n}\right)\right) = \\ = \exp\left(n\frac{F\,\Delta t}{n}\right) = \exp(F\,\Delta t) = e^{F\,\Delta t}. \tag{3.12}$$

This is to be interpreted as a Taylor expansion:

$$\Phi_i^{i+1} = e^{F\,\Delta t} = I + F\,\Delta t + \frac{1}{2}F^2\Delta t^2 + \frac{1}{6}F^3\Delta t^3 + \cdots$$

Observe that for the state transition matrix the *transitive property* holds:

$$\Phi_i^{i+2} = \Phi_{i+1}^{i+2}\cdot\Phi_i^{i+1},$$

in other words, to transition the state from $\mathbf{x}(t_i)$ to $\mathbf{x}(t_{i+2})$, one may transition first from t_i to t_{i+1} and then from t_{i+1} to t_{i+2}.

3.4.2 *Example: satellite motion*

Differential equation 3.8 can be converted, for small values of Δt, to

$$\begin{bmatrix} \Delta\underline{\mathbf{x}}(t_{i+1}) \\ \Delta\underline{\mathbf{v}}(t_{i+1}) \end{bmatrix} - \begin{bmatrix} \Delta\underline{\mathbf{x}}(t_i) \\ \Delta\underline{\mathbf{v}}(t_i) \end{bmatrix} \approx \Delta t \begin{bmatrix} 0 & I \\ \mathcal{M}_i & 0 \end{bmatrix} \begin{bmatrix} \Delta\underline{\mathbf{x}}(t_i) \\ \Delta\underline{\mathbf{v}}(t_i) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{\mathbf{n}} \end{bmatrix}$$
$$\Longrightarrow \begin{bmatrix} \Delta\underline{\mathbf{x}}(t_{i+1}) \\ \Delta\underline{\mathbf{v}}(t_{i+1}) \end{bmatrix} \approx \begin{bmatrix} I & I\,\Delta t \\ \mathcal{M}_i\,\Delta t & I \end{bmatrix} \begin{bmatrix} \Delta\underline{\mathbf{x}}(t_i) \\ \Delta\underline{\mathbf{v}}(t_i) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{\mathbf{n}} \end{bmatrix},$$

with the definitions $\Delta t \stackrel{\text{def}}{=} t_{i+1} - t_i$ and $\mathcal{M}_i \stackrel{\text{def}}{=} \mathcal{M}^{(0)}(t_i)$.

We see that the state transition matrix is

$$\Phi_i^{i+1} = \begin{bmatrix} I & I\,\Delta t \\ \mathcal{M}_i\,\Delta t & I \end{bmatrix}.$$

Likewise, the matrix of the next state transition will be

$$\Phi_{i+1}^{i+2} = \begin{bmatrix} I & I\,\Delta t \\ \mathcal{M}_{i+1}\,\Delta t & I \end{bmatrix}.$$

Concatenating, or multiplying, them yields

$$\Phi_i^{i+2} = \Phi_{i+1}^{i+2} \cdot \Phi_i^{i+1} = \begin{bmatrix} I & I\,\Delta t \\ \mathcal{M}_{i+1}\,\Delta t & I \end{bmatrix} \begin{bmatrix} I & I\,\Delta t \\ \mathcal{M}_i\,\Delta t & I \end{bmatrix} =$$
$$= \begin{bmatrix} I + \mathcal{O}(\Delta t^2) & (I + I)\,\Delta t \\ (\mathcal{M}_{i+1} + \mathcal{M}_i)\,\Delta t & I + \mathcal{O}(\Delta t^2) \end{bmatrix}.$$

Repeating this for multiple small time steps Δt suggests that the following integral equation applies generally:

$$\Phi_0^t = \begin{bmatrix} I & I \cdot (t - t_0) \\ \int_0^t \mathcal{M}(\tau)\,d\tau & I \end{bmatrix}.$$

3.4.3 *Propagation of state and state variance*

We look at the propagation in time of both $\mathbf{x}^-(t)$ and $\underline{\mathbf{x}}(t)$ according to the discrete model.

The discrete propagation equation for the state estimator is

$$\mathbf{x}^-(t_{i+1}) = \Phi_i^{i+1}\,\mathbf{x}^-(t_i), \qquad (3.11)$$

with Φ_i^{i+1} the state transition matrix. In the absence of noise $\underline{\mathbf{n}}(t)$, it also holds that

$$\underline{\mathbf{x}}(t_{i+1}) = \Phi_i^{i+1}\,\underline{\mathbf{x}}(t_i)$$

and subtraction yields

$$\left(\mathbf{x}^-(t_{i+1}) - \underline{\mathbf{x}}(t_{i+1})\right) = \Phi_i^{i+1}\left(\mathbf{x}^-(t_i) - \underline{\mathbf{x}}(t_i)\right),$$

varianssien kasautuminen

after which propagation of variances yields (equation 3.10):

$$\Sigma^-(t_{i+1}) = \Phi_i^{i+1}\Sigma^-(t_i)\left(\Phi_i^{i+1}\right)^T.$$

3.4.4 State propagation in the presence of noise

In the presence of dynamic noise, the following discrete dynamic equations apply:

$$\begin{aligned} \mathbf{x}^-(t_{k+1}) &= \Phi_k^{k+1}\,\mathbf{x}^+(t_k), \\ \underline{\mathbf{x}}(t_{k+1}) &= \Phi_k^{k+1}\,\underline{\mathbf{x}}(t_k) + \underline{\mathbf{w}}_k^{k+1}, \\ \Sigma^-(t_{k+1}) &= \Phi_k^{k+1}\Sigma^+(t_k)\left(\Phi_k^{k+1}\right)^T + \Theta_k^{k+1}. \end{aligned} \tag{3.13}$$

These equations were already presented in figure 3.2b. We note that here, the symbols $\mathbf{x}^+(t_k)$ and $\Sigma^+(t_k)$ refer to state and state variance after (*a posteriori*) the observation update at time t_k, see section 3.7. Both t_k and t_{k+1} here refer to successive updates.

In subsections 3.5.2 and 3.5.3 we shall show how these equations are connected to their continuous counterparts, the differential equations also shown in figure 3.2b.

In equation 3.13 according to figure 3.2b:

$$\underline{\mathbf{w}}_k^{k+1} = \underline{\mathbf{w}}_{t_k}^{t_{k+1}} = \int_k^{k+1} \Phi_t^{k+1}\,\underline{\mathbf{n}}(t)\,dt, \tag{3.14}$$

and the autocovariance of $\underline{\mathbf{w}}$ is

$$\Theta_k^{k+1} = \Theta_{t_k}^{t_{k+1}} = A_{w,k}(t_{k+1}, t_{k+1}) = \int_k^{k+1} \Phi_t^{k+1} Q_n(t)\left(\Phi_t^{k+1}\right)^T dt. \tag{3.15}$$

The interpretation of equations 3.14 and 3.15 is the following. Each element of dynamic noise $\underline{\mathbf{n}}(t)\,dt$ in the interval (t_k, t_{k+1}) and the variance matrix $Q_n(t)\,dt$ of each such element are transitioned forwards in time from moment t to moment t_{k+1} by multiplying with the state transition matrix Φ_t^{k+1}. The variance matrix is multiplied both with the state transition matrix from the left and with its transpose from the right, as is always done based on variance propagation law.

After that, integration over the "carried forward" elements is carried out.

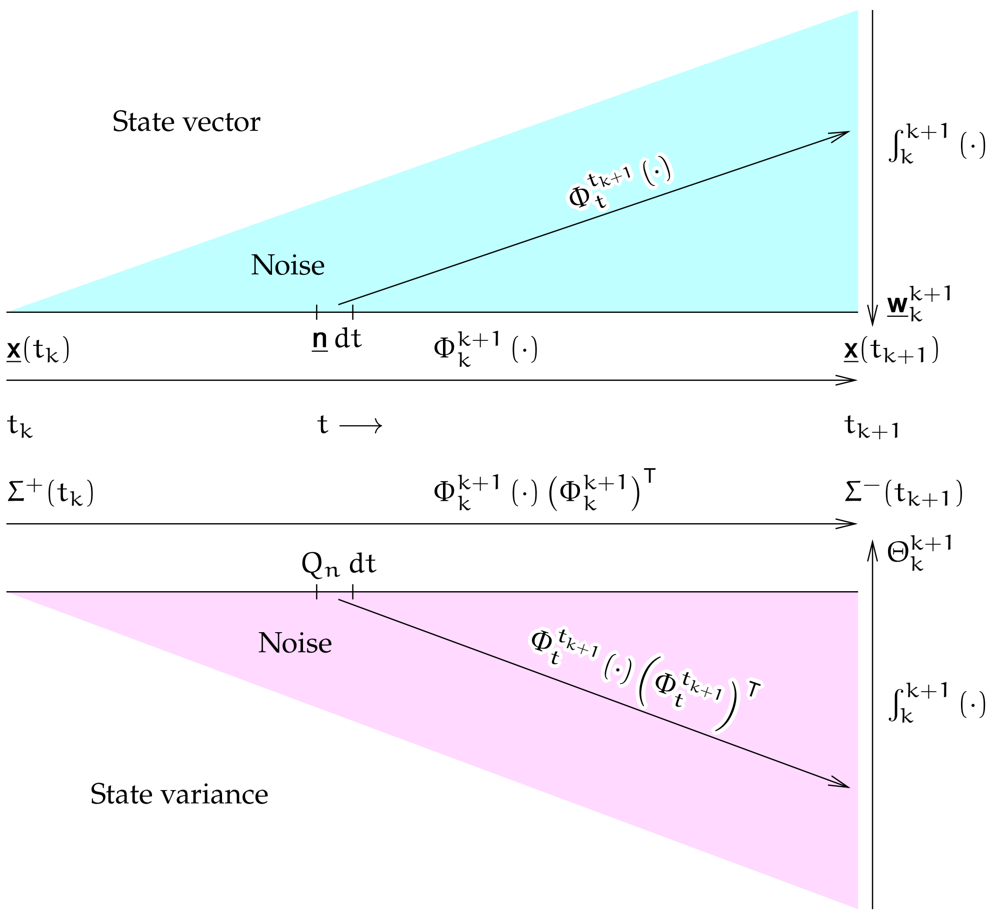


FIGURE 3.3. Propagation of state vector and state variance in the presence of noise.

The matrix Θ_k^{k+1} thus represents the accumulated variance of the dynamic noise over the time span from (t_k, t_{k+1}). Simply summing, or integrating, the variance contributions from elements dt is allowed because for white noise, the elements are statistically independent of each other. theta $\vartheta\theta\Theta$

A diagram describing the propagation process is given in figure 3.3.

If the time step $\Delta t = t_{k+1} - t_k$ is short and thus the Φ matrices are close to unity,

$$\underline{\mathbf{w}}_k^{k+1} \approx \int_k^{k+1} \underline{\mathbf{n}}(t)\,dt, \qquad \Theta_k^{k+1} \approx \int_k^{k+1} Q_n(t)\,dt.$$

The assumption is that the dynamic noise $\underline{\mathbf{n}}(t)$ is white. In this case, $\underline{\mathbf{w}}_k(t)$ is a *random walk*, an integral over white noise, which we encountered earlier.

3.5 The differential equation for the state variance

3.5.1 The differential equation for the state transition matrix

We may also derive differential equations that describe the propagation of the state variance matrix and state transition matrix in time. It is assumed that matrix $F(t)$ is a function of time.

Differentiate equation 3.11:

$$\mathbf{x}^-(t) = \Phi_0^t\,\mathbf{x}^-(t_0) \implies \frac{d}{dt}\mathbf{x}^-(t) = \left(\frac{d}{dt}\Phi_0^t\right)\mathbf{x}^-(t_0).$$

Substitute into equation 3.6:

$$\frac{d}{dt}\mathbf{x}^-(t) = F(t)\,\mathbf{x}^-(t) = F(t)\,\Phi_0^t\,\mathbf{x}^-(t_0),$$

yielding

$$\left(\frac{d}{dt}\Phi_0^t\right)\cancel{\mathbf{x}^-(t_0)} = F(t)\,\Phi_0^t\,\cancel{\mathbf{x}^-(t_0)} \implies \frac{d}{dt}\Phi_0^t = F(t)\,\Phi_0^t. \qquad (3.16)$$

With the initial condition $\Phi_0^0 = I$ we can by integration calculate matrix Φ_0^t.

There is no closed solution, except when F is a constant or if

$$F(t')\,F(t) = F(t)\,F(t') \qquad (3.17)$$

for arbitrary t and t' (Wikipedia, Magnus expansion). In that case, like equation 3.12,

$$\Phi_0^t = \exp\left(\int_0^t F(\tau)\,d\tau\right), \qquad (3.18)$$

to be interpreted as a Taylor expansion:

$$\Phi_0^t = I + \int_0^t F(\tau)\,d\tau + \frac{1}{2}\left(\int_0^t F(\tau)\,d\tau\right)^2 + \frac{1}{6}\left(\int_0^t F(\tau)\,d\tau\right)^3 + \cdots.$$

This generalises equation 3.12 for a time variable matrix $F(t)$.

A system for which the exponent equation 3.18 does *not* work is the satellite example of subsection 3.4.2:

$$F(t')\,F(t) = \begin{bmatrix} 0 & I \\ \mathcal{M}^{(0)}(t') & 0 \end{bmatrix}\begin{bmatrix} 0 & I \\ \mathcal{M}^{(0)}(t) & 0 \end{bmatrix} =$$
$$= \begin{bmatrix} \mathcal{M}^{(0)}(t) & 0 \\ 0 & \mathcal{M}^{(0)}(t') \end{bmatrix} \neq F(t)\,F(t').$$

3.5.2 *The differential equation for the state variance*

In order to derive a differential equation for the state variance matrix Σ, we start from equation 3.13:

$$\Sigma^-(t) = \Phi_0^t\,\Sigma^-(t_0)\,\left(\Phi_0^t\right)^{\mathsf{T}} + \Theta_0^t.$$

Differentiate and use equation 3.16:

$$\frac{d}{dt}\Sigma^-(t) = \left(\frac{d}{dt}\Phi_0^t\right)\Sigma^-(t_0)\,\left(\Phi_0^t\right)^{\mathsf{T}} + \Phi_0^t\,\Sigma^-(t_0)\left(\frac{d}{dt}\Phi_0^t\right)^{\mathsf{T}} + \frac{d}{dt}\Theta_0^t =$$

$$= F(t)\,\Phi_0^t\,\Sigma^-(t_0)\,\left(\Phi_0^t\right)^{\mathsf{T}} + \Phi_0^t\,\Sigma^-(t_0)\,\left(\Phi_0^t\right)^{\mathsf{T}} F^{\mathsf{T}}(t) + \frac{d}{dt}\Theta_0^t.$$

If we let $t_0 \to t$, then $\Phi_0^t \to I$, and

$$\frac{d}{dt}\Sigma^-(t) = F(t)\,\Sigma^-(t) + \Sigma^-(t)\,F^{\mathsf{T}}(t) + \frac{d}{dt}\Theta_0^t.$$

With the integral 3.15 we obtain in the same limit $t_0 \to t$, $\tau \to t$, and tau τT
thus $\Phi_\tau^t \to I$, the result

$$\frac{d}{dt}\Theta_0^t = \frac{d}{dt}\int_0^t \left(\Phi_\tau^t\right) Q_n(\tau)\,\left(\Phi_\tau^t\right)^{\mathsf{T}} d\tau \approx \frac{d}{dt}\int_0^t Q_n(\tau)\,d\tau = Q_n(t),$$

and thus

$$\frac{d}{dt}\Sigma^-(t) = F(t)\,\Sigma^-(t) + \Sigma^-(t)\,F^{\mathsf{T}}(t) + Q_n(t). \tag{3.19}$$

3.5.3 *Differential equations for the state noise*

We start from equation 3.13 with noise included,

$$\underline{\mathbf{x}}(t) = \Phi_0^t\,\underline{\mathbf{x}}(t_0) + \underline{\mathbf{w}}_0^t. \tag{3.20}$$

Divide the time interval $\left[t_0, t\right]$ into n subintervals of size δt, so

$$t_i = t_0 + j\,\delta t, \qquad j = 1, \ldots, n$$

and $t = t_n$. Expand

$$\underline{\mathbf{x}}(t) = \Phi^t_{n-1}\,\underbrace{\underline{\mathbf{x}}(t_{n-1})}_{\Phi^{n-1}_{n-2}\,\underbrace{\underline{\mathbf{x}}(t_{n-2})}_{\Phi^{n-2}_{n-3}\,\underbrace{\underline{\mathbf{x}}(t_{n-3})}_{\cdots\cdots\cdots} + \underline{\mathbf{w}}^{n-2}_{n-3}} + \underline{\mathbf{w}}^{n-1}_{n-2}} + \underline{\mathbf{w}}^t_{n-1} =$$

$$= \Phi^t_{n-3}\,\underline{\mathbf{x}}(t_{n-3}) + \underline{\mathbf{w}}^t_{n-1} + \Phi^t_{n-1}\,\underline{\mathbf{w}}^{n-1}_{n-2} + \Phi^t_{n-2}\,\underline{\mathbf{w}}^{n-2}_{n-3} + \cdots =$$

$$= \Phi_0^t\,\underline{\mathbf{x}}(t_0) + \sum_{i=0}^{n-1} \Phi^t_{n-i}\,\underline{\mathbf{w}}^{n-i}_{n-i-1} = \Phi_0^t\,\underline{\mathbf{x}}(t_0) + \sum_{j=1}^{n} \Phi^t_j\,\underline{\mathbf{w}}^j_{j-1}.$$

Use the approximation

$$\underline{\mathbf{w}}_{j-1}^{j} \approx \delta t \cdot \underline{\mathbf{n}}(t_j),$$

so

$$\underline{\mathbf{x}}(t) = \Phi_0^t\, \underline{\mathbf{x}}(t_0) + \delta t \sum_{j=1}^{n} \Phi_j^t\, \underline{\mathbf{n}}(t_j),$$

which is recognised as the rectangle-rule approximation to an integral. The limit $n \to \infty$ yields

$$\underline{\mathbf{x}}(t) = \Phi_0^t\, \underline{\mathbf{x}}(t_0) + \int_0^t \Phi_\tau^t\, \underline{\mathbf{n}}(\tau)\, d\tau,$$

so with equation 3.20 it follows that

$$\underline{\mathbf{w}}_0^t = \int_0^t \Phi_\tau^t\, \underline{\mathbf{n}}(\tau)\, d\tau. \tag{3.21}$$

Similarly, the autocovariance function of this variant of a random walk is

$$\Theta_0^t = A_{w,0}(t,t) = \int_0^t \Phi_\tau^t Q_n(\tau) \left(\Phi_\tau^t\right)^{\mathsf{T}} d\tau. \tag{3.22}$$

The twist is the state transition matrix inside the integral.

Finally, note that in the foregoing equations, the following transitive property can be spotted:

$$\underline{\mathbf{w}}_0^t = \overbrace{\Phi_{t'}^t\, \underline{\mathbf{w}}_0^{t'} + \underline{\mathbf{w}}_{t'}^t}^{\mathrm{I}}. \tag{3.23}$$

This can be proven by substituting the integral expression 3.21:

$$\begin{aligned}\overbrace{\Phi_{t'}^t\, \underline{\mathbf{w}}_0^{t'} + \underline{\mathbf{w}}_{t'}^t}^{\mathrm{I}} &= \Phi_{t'}^t \overbrace{\int_0^{t'} \Phi_\tau^{t'}\, \underline{\mathbf{n}}(\tau)\, d\tau}^{\underline{\mathbf{w}}_0^{t'}} + \overbrace{\int_{t'}^t \Phi_\tau^t\, \underline{\mathbf{n}}(\tau)\, d\tau}^{\underline{\mathbf{w}}_{t'}^t} = \\ &= \int_0^{t'} \overbrace{\Phi_{t'}^t \Phi_\tau^{t'}}^{\Phi_\tau^t}\, \underline{\mathbf{n}}(\tau)\, d\tau + \int_{t'}^t \Phi_\tau^t\, \underline{\mathbf{n}}(\tau)\, d\tau = \overbrace{\int_0^t \Phi_\tau^t\, \underline{\mathbf{n}}(\tau)\, d\tau}^{\underline{\mathbf{w}}_0^t}.\end{aligned}$$

3.5.4 *Integration summary*

We obtained the differential equation of the state variance

$$\frac{d}{dt}\Sigma^-(t) = F(t)\,\Sigma^-(t) + \Sigma^-(t)\,F^{\mathsf{T}}(t) + Q_n(t). \tag{3.19}$$

The equation is also suitable for integrating matrix $\Sigma(t)$ in the case where $F(t)$ is a function of time and not a constant. Then, one uses equation 3.6:

$$\frac{d}{dt}\mathbf{x}^-(t) = F(t)\,\mathbf{x}^-(t) \tag{3.6}$$

for integrating the state estimator itself.

In the discrete case, equation 3.11 together with equation 3.16 can be used to integrate the state $\mathbf{x}^-(t)$ over time between update events. If matrix F is constant, the state transition matrix Φ_k^{k+1} is directly obtained by equation 3.12.

All this however assumes that matrix F *exists*, in other words the function $F(\mathbf{x}, t)$ can be *linearised*.

3.6 The observation model

The evolution of the state vector in time would not be very interesting if it could not be *observed* in some way. The observation model in the linear case is

$$\underline{\boldsymbol{\ell}} = H\,\underline{\mathbf{x}} + \underline{\mathbf{m}}\,,$$

in which $\underline{\ell}$ is the observable, in the general case a vector, $\underline{\mathbf{x}}$ is the state vector — the "true value" — and $\underline{\mathbf{m}}$ is the "noise" or random error of the observation process. H is the *observation matrix*.[5] It is assumed that the expectancy $E\{\underline{\mathbf{m}}\} = 0$. The variance matrix of the observation vector is

5 odotusarvo

$$R \stackrel{\text{def}}{=} E\{\underline{\mathbf{m}}\,\underline{\mathbf{m}}^{\mathsf{T}}\}.$$

Let the moment of observation, or epoch, be t. The estimator of the state vector calculated forwards to this moment is $\mathbf{x}^- = \mathbf{x}^-(t)$. From this we calculate an estimator for the observable:

$$\widehat{\boldsymbol{\ell}} = H\,\mathbf{x}^-.$$

Now, a zero quantity or "*closing error*" — a quantity the expectancy $E\{\cdot\}$ of which is zero — can be formed as follows:

$$\underline{\mathbf{y}} = \widehat{\boldsymbol{\ell}} - \underline{\boldsymbol{\ell}} = H\,\mathbf{x}^- - \underline{\boldsymbol{\ell}} = H\cdot(\mathbf{x}^- - \underline{\mathbf{x}}) - \underline{\mathbf{m}},$$

and thus

$$E\{\underline{\mathbf{y}}\} = H\cdot\left(E\{\mathbf{x}^-\} - E\{\underline{\mathbf{x}}\}\right) - E\{\underline{\mathbf{m}}\} = H\cdot 0 - 0 = 0,$$

based on the assumption $E\{\mathbf{x}^-\} = E\{\underline{\mathbf{x}}\}$, that $\mathbf{x}^-$ is an *unbiased* estimator of the state $\underline{\mathbf{x}}$.

harhaton estimaattori

[5] This is the same as in least-squares adjustment the A matrix or *design matrix*.

In the non-linear case, H is not a matrix but a function $H(\underline{\mathbf{x}})$ of the state vector:

$$\underline{\boldsymbol{\ell}} = H(\underline{\mathbf{x}}) + \underline{\mathbf{m}}, \qquad \widehat{\boldsymbol{\ell}} = H(\mathbf{x}^-),$$

and the difference is

$$\underline{\mathbf{y}} = \widehat{\boldsymbol{\ell}} - \underline{\boldsymbol{\ell}} = H(\mathbf{x}^-) - \left(H(\underline{\mathbf{x}}) + \underline{\mathbf{m}}\right) \approx H \cdot \left(\mathbf{x}^- - \underline{\mathbf{x}}\right) - \underline{\mathbf{m}}\,.$$

The elements of matrix H are defined by

$$H_{ij} = \frac{\partial}{\partial x_j} H_i(\overbrace{x_1, \ldots, x_j, \ldots, x_n}^{\mathbf{x}})\bigg|_{\mathbf{x}=\mathbf{x}^{(0)}}.$$

H is the *Jacobi matrix*, or matrix of partial derivatives, of the function $H(\mathbf{x})$. The matrix is evaluated at the approximate values for the state vector contained in $\mathbf{x}^{(0)}$, for the observation epoch t.

To calculate the variances and covariances, it is assumed that both $\mathbf{x}^-$ and $\underline{\mathbf{x}}$ are statistically independent from $\underline{\mathbf{m}}$. These are sensible assumptions, as the observation process is usually physically completely independent from the system dynamic process, and the observation processes at different epochs are independent of each other.

Calculate

$$\begin{aligned}
\mathrm{Var}\left\{\underline{\mathbf{y}}\right\} &= E\left\{\underline{\mathbf{y}}\,\underline{\mathbf{y}}^{\mathsf{T}}\right\} = H\,E\left\{(\mathbf{x}^- - \underline{\mathbf{x}})\,(\mathbf{x}^- - \underline{\mathbf{x}})^{\mathsf{T}}\right\} H^{\mathsf{T}} + R = \\
&= H\Sigma^- H^{\mathsf{T}} + R, \\
\mathrm{Cov}\left\{\underline{\mathbf{y}}, \mathbf{x}^-\right\} &= E\left\{\underline{\mathbf{y}} \cdot (\mathbf{x}^- - \underline{\mathbf{x}})^{\mathsf{T}}\right\} = H\Sigma^-, \qquad \mathrm{Cov}\left\{\mathbf{x}^-, \underline{\mathbf{y}}\right\} = \Sigma^- H^{\mathsf{T}}.
\end{aligned}$$

3.7 Updating

The difference $\underline{\mathbf{y}}$ between the observable $\widehat{\boldsymbol{\ell}}$ calculated from the estimated state vector $\mathbf{x}^-$ and the real observation $\underline{\boldsymbol{\ell}}$ has an expectancy of zero. The update step makes optimal use of this fact.

6 So, an enhanced estimator is formed,[6]

$$\mathbf{x}^+ = \mathbf{x}^- - K\underline{\mathbf{y}} = \mathbf{x}^- - K \cdot (H\mathbf{x}^- - \underline{\boldsymbol{\ell}}) = \mathbf{x}^- - K \cdot \left(H \cdot (\mathbf{x}^- - \underline{\mathbf{x}}) - \underline{\mathbf{m}}\right),$$

so

$$(\mathbf{x}^+ - \underline{\mathbf{x}}) = (I - KH)\,(\mathbf{x}^- - \underline{\mathbf{x}}) + K\,\underline{\mathbf{m}}.$$

vahvistus-matriisi The matrix K is called the Kalman *gain matrix*.

[6]The plus sign used as a superscript here denotes the state vector *after* (*a posteriori*) the use of an observation in the update step. Other notations are found as well, for example the subscripts k and k + 1 referring to the states before and after.

According to definition 3.10 we may use this to derive the update equation for the state variance:

$$\Sigma^{+} = (I - KH)\,\Sigma^{-}\,(I - KH)^{\mathsf{T}} + KRK^{\mathsf{T}}. \tag{3.24}$$

The *optimal* solution is obtained by choosing

$$K = \Sigma^{-}H^{\mathsf{T}}\left(H\Sigma^{-}H^{\mathsf{T}} + R\right)^{-1}, \tag{3.25}$$

which gives as the solution

$$\begin{aligned} \mathbf{x}^{+} &= \mathbf{x}^{-} - K\cdot\left(H\mathbf{x}^{-} - \underline{\boldsymbol{\ell}}\right) = \\ &= \mathbf{x}^{-} - \Sigma^{-}H^{\mathsf{T}}\left(H\Sigma^{-}H^{\mathsf{T}} + R\right)^{-1}\left(H\mathbf{x}^{-} - \underline{\boldsymbol{\ell}}\right). \end{aligned} \tag{3.26}$$

if we call

$$P \stackrel{\text{def}}{=} \left(H\Sigma^{-}H^{\mathsf{T}} + R\right)^{-1} \implies K = \Sigma^{-}H^{\mathsf{T}}P,$$

we can open up equation 3.24 as

$$\Sigma^{+} = \overbrace{\left(I - \Sigma^{-}H^{\mathsf{T}}PH\right)\Sigma^{-}\left(I - \Sigma^{-}H^{\mathsf{T}}PH\right)^{\mathsf{T}}}^{\mathrm{I}} + \overbrace{\Sigma^{-}H^{\mathsf{T}}PRPH\Sigma^{-}}^{\mathrm{II}},$$

in which

$$\begin{aligned} \mathrm{I} &= \left(I - \Sigma^{-}H^{\mathsf{T}}PH\right)\Sigma^{-}\left(I - \Sigma^{-}H^{\mathsf{T}}PH\right)^{\mathsf{T}} = \\ &= \Sigma^{-} - \Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-} - \Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-} + \overbrace{\Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}}^{\mathrm{III}} \end{aligned}$$

and

$$\begin{aligned} \mathrm{III} + \mathrm{II} &= \Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-} + \Sigma^{-}H^{\mathsf{T}}PRPH\Sigma^{-} = \\ &= \Sigma^{-}H^{\mathsf{T}}P\overbrace{\left(H\Sigma^{-}H^{\mathsf{T}} + R\right)}^{P^{-1}}PH\Sigma^{-} = \Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}, \end{aligned}$$

so

$$\begin{aligned} \Sigma^{+} &= \Sigma^{-} - \Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-} \cancel{-\Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}} \cancel{+\Sigma^{-}H^{\mathsf{T}}PH\Sigma^{-}} = \\ &= \Sigma^{-} - \Sigma^{-}H^{\mathsf{T}}\overbrace{\left(H\Sigma^{-}H^{\mathsf{T}} + R\right)^{-1}}^{P}H\Sigma^{-}. \end{aligned}$$

This variance update equation may still be shortened as follows:

$$\Sigma^{+} = \Sigma^{-} - \Sigma^{-}H^{\mathsf{T}}\left(H\Sigma^{-}H^{\mathsf{T}} + R\right)^{-1}H\Sigma^{-} = (I - KH)\,\Sigma^{-}, \tag{3.27}$$

based on the definition of the Kalman gain matrix K.

Perhaps more intuitively summarised:

$$\begin{aligned} \mathbf{x}^{+} &= \mathbf{x}^{-} - \operatorname{Cov}\{\mathbf{x}^{-}, \underline{\mathbf{y}}\} \operatorname{Var}^{-1}\{\underline{\mathbf{y}}\} \underline{\mathbf{y}}, \\ \operatorname{Var}\{\mathbf{x}^{+}\} &= \operatorname{Var}\{\mathbf{x}^{-}\} - \operatorname{Cov}\{\mathbf{x}^{-}, \underline{\mathbf{y}}\} \operatorname{Var}^{-1}\{\underline{\mathbf{y}}\} \operatorname{Cov}\{\underline{\mathbf{y}}, \mathbf{x}^{-}\}, \end{aligned} \tag{3.28}$$

which looks like a regression of the state vector estimator $\mathbf{x}^{-}$ with respect to the "closing error" $\underline{\mathbf{y}}$.

So the *update equations* for the Kalman filter have been found for both the state vector and its variance matrix.

We can find many ways to calculate these equations effectively and precisely in the literature. The main issue, nevertheless, is that the variance matrix of the "closing error"

$$\operatorname{Var}\{\underline{\mathbf{y}}\} = H\Sigma H^{\mathsf{T}} + R$$

is the size of vector $\underline{\mathbf{y}}$. And the size of $\underline{\mathbf{y}}$ equals the number of simultaneous observations. This is why the Kalman filter is also called a *sequential filter*, because it handles the observations one epoch at a time and not, like for example traditional adjustment calculus, all at once.

3.8 The optimality of the update

Equations 3.28 are *optimal* in the least-squares sense. We can prove this as follows, with a little simplification.

We start by calculating, using the first equation of 3.28:

$$\operatorname{Cov}\{\mathbf{x}^{+}, \underline{\mathbf{y}}\} = \operatorname{Cov}\{\mathbf{x}^{-}, \underline{\mathbf{y}}\} - \operatorname{Cov}\{\mathbf{x}^{-}, \underline{\mathbf{y}}\} \operatorname{Var}^{-1}\{\underline{\mathbf{y}}\} \operatorname{Var}\{\underline{\mathbf{y}}\} = 0, \tag{3.29}$$

remembering that, by definition, $\operatorname{Cov}\{\underline{\mathbf{y}}, \underline{\mathbf{y}}\} = \operatorname{Var}\{\underline{\mathbf{y}}\}$.

Suppose now that there existed an alternative updated state $\mathbf{x}^{\times}$, an unbiased estimator that was even better than the standard update $\mathbf{x}^{+}$. Say

$$\mathbf{x}^{\times} = \mathbf{x}^{+} + C\underline{\mathbf{y}}$$

for some coefficient matrix C. Then, because of equation 3.29, we would have

$$\operatorname{Var}\{\mathbf{x}^{\times}\} = \operatorname{Var}\{\mathbf{x}^{+}\} + C \operatorname{Var}\{\underline{\mathbf{y}}\} C^{\mathsf{T}}.$$

Because $\operatorname{Var}\{\underline{\mathbf{y}}\}$ is positive definite, the expression

$$\operatorname{Var}\{\mathbf{x}^{\times}\} - \operatorname{Var}\{\mathbf{x}^{+}\} = C \operatorname{Var}\{\underline{\mathbf{y}}\} C^{\mathsf{T}}$$

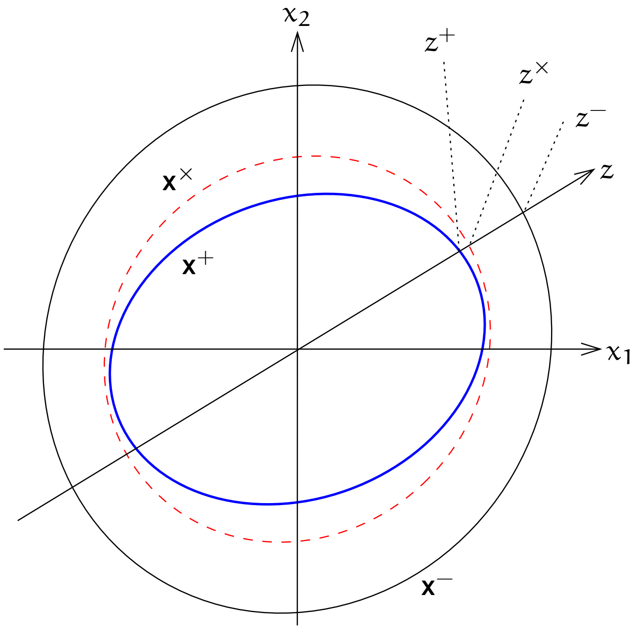


FIGURE 3.4. The error ellipse of the optimal estimator (blue) is completely surrounded by the error ellipses of other estimators — including that of the pre-update estimator.

is positive semidefinite, and

$$\mathrm{Var}\{\mathbf{x}^\times\} - \mathrm{Var}\{\mathbf{x}^+\} = 0$$

only if $C = 0$.

In other words, for an arbitrary linear combination

$$\underline{z} = \sum_i c_i \underline{x}_i, \qquad z^\times = \sum_i c_i x_i^\times, \qquad z^+ = \sum_i c_i x_i^+,$$

it holds that

$$\mathrm{Var}\{z^\times\} - \mathrm{Var}\{z^+\} = \mathbf{c}C\,\mathrm{Var}\{\underline{\mathbf{y}}\}\,C^\mathsf{T}\mathbf{c}^\mathsf{T}$$

in which $\mathbf{c} \stackrel{\text{def}}{=} \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}$. This only vanishes if $\mathbf{c}C = 0$, otherwise

$$\mathrm{Var}\{z^\times\} - \mathrm{Var}\{z^+\} > 0.$$

The two-dimensional case is presented graphically in figure 3.4.

So, the variance ellipse of the *optimal estimator* $\mathbf{x}^+$ — more generally a (hyper-)ellipsoid — always lies *entirely inside* the variance ellipse of any alternative estimator $\mathbf{x}^\times$, or at worst touches it from the inside. The same holds for the variances of an arbitrary linear combination $\underline{z}$ of the components.

Self-test questions

1. What is the state vector of the Kalman filter?
2. What is the dynamic model of the Kalman filter?
3. How is the dynamic model linearised?
4. What is the state transition matrix?
5. What is the relationship between the state transition matrix and the coefficient matrix F of the linear dynamic model?
6. What is the observation model of the Kalman filter?
7. What is the update step of the Kalman filter?

Exercise 3–1: A simple two-dimensional dynamic model

The following two-dimensional, Gauss-Markov like dynamic model is given:

$$\frac{d}{dt}\begin{bmatrix} \underline{x}(t) \\ \underline{y}(t) \end{bmatrix} = \begin{bmatrix} -k & 1 \\ 0 & -k \end{bmatrix}\begin{bmatrix} \underline{x}(t) \\ \underline{y}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n}(t) \end{bmatrix}. \tag{3.30}$$

Here, $k \in (0, 1)$ is a small attenuation constant.

1. Because the matrix
$$F = \begin{bmatrix} -k & 1 \\ 0 & -k \end{bmatrix}$$
is constant, it follows that condition 3.17 holds and we may express the state transition matrix as an exponential expansion. First, derive a general expression for the powers of F, F^n, $n \in \mathbb{N}$, in this exponential expansion.
2. What is F^0? What is F^{-1}?
3. Derive the exponential $e^{F\Delta t}$ as a Taylor expansion in Δt. Write out the first four terms.
4. Show that the state transition matrix Φ_i^{i+1} in the equation has the following form:
$$\begin{bmatrix} \underline{x}(t_{i+1}) \\ \underline{y}(t_{i+1}) \end{bmatrix} = \overbrace{e^{-k\Delta t}\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}}^{\Phi_i^{i+1}}\begin{bmatrix} \underline{x}(t_i) \\ \underline{y}(t_i) \end{bmatrix} + \cdots \tag{3.31}$$
Here we have left out the noise term for simplicity, and $\Delta t = t_{i+1} - t_i$.

Hint *Verify* that equation 3.16:

$$\frac{d}{dt}\Phi_0^t = F(t)\,\Phi_0^t$$

holds for the state transition matrix given in the above equation 3.31.

Simplify the notation by the substitutions $t_i \to 0$ and $t_{i+1} \to t$.

Exercise 3–2: A transitive property for the dynamic noise variance

Derive the transitive property, similar to that for the dynamic noise $\underline{\mathbf{w}}_0^t$ in subsection 3.5.3:

$$\Theta_0^t = \Phi_{t'}^t \Theta_0^{t'} \left(\Phi_{t'}^t\right)^\mathsf{T} + \Theta_{t'}^t.$$

Show, by substitution of equation 3.22, that this is an identity.

Examples and applications of the Kalman filter

4

4.1 Example 1: one-dimensional motion

In this example we shall not concern ourselves with physical units. It is assumed that metres are used for distances or co-ordinates and seconds are used for time.

Question Assume, in one spatial dimension, the dynamic model for the state vector

$$\underline{\mathbf{x}} = \begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}$$

to be

$$\frac{d^2}{dt^2}\underline{x} = \underline{n}.$$

The model is linear. Convert it to a matric equation with only first derivatives of time:

$$\frac{d}{dt}\overbrace{\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}}^{\underline{\mathbf{x}}} = \overbrace{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}}^{F}\overbrace{\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}}^{\underline{\mathbf{x}}} + \overbrace{\begin{bmatrix} 0 \\ \underline{n} \end{bmatrix}}^{\underline{\mathbf{n}}}.$$

Here, $\underline{n}$ is white noise with an autocovariance of $Q_n = 1$. Assume furthermore that an estimate of the initial state at $t = 0$ is given:

$$\widehat{\mathbf{x}}(0) = \begin{bmatrix} \widehat{x}(0) \\ \widehat{v}(0) \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix},$$

$$\Sigma(0) = \begin{bmatrix} \Sigma_{xx}(0) & 0 \\ 0 & \Sigma_{vv}(0) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix},$$

meaning that there is no useful velocity information.

1. Propagate this state information forwards to the moment $t = 5$, in other words calculate

$$\widehat{\mathbf{x}}(5) = \mathbf{x}^-(5),$$
$$\Sigma(5) = \Sigma^-(5).$$

2. At the moment $t = 5$, an observation yielding a value of 3 is made:

$$\underline{\ell}' = x^-(5) + \underline{m} = 3.$$

The observation random error or noise $\underline{m}$ has variance $R = 3$.

(a) What does the H matrix look like? And the K matrix?

(b) Calculate the *a posteriori* state $\mathbf{x}^+(5)$, $\Sigma^+(5)$.

3. Alternatively calculate the outcome using a standard least-squares adjustment. The dynamic model is

$$\widehat{x}(t) = \widehat{x}(0) + \widehat{v}(0) \cdot t,$$

the unknowns to be estimated are $\widehat{x}(0)$ and $\widehat{v}(0)$, forming the abstract vector $\widehat{\mathbf{x}} = \begin{bmatrix} \widehat{x}(0) & \widehat{v}(0) \end{bmatrix}^T$, and the observation equations are

$$\underline{\ell}_1 + \underline{v}_1 = \widehat{x}(0),$$
$$\underline{\ell}_2 + \underline{v}_2 = \widehat{x}(5).$$

The observation vector and its variance matrix are

$$\underline{\boldsymbol{\ell}} = \begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad S = \begin{bmatrix} \Sigma_{xx}(0) & 0 \\ 0 & R \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}. \tag{4.1}$$

Answer

1. $\widehat{x}(5) = \widehat{x}(0) + \widehat{v}(0) \cdot 5 = 4$. Because the coefficient matrix

$$F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

the state transition matrix is obtained as

$$\Phi_0^5 = e^{F\Delta t} = \exp\left(\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}\right) =$$
$$= I + \begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}^2 + \cdots =$$
$$= \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix},$$

because

$$\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}^n = 0, \quad n > 1.$$

Now, with equations 3.13 and 3.15:

$$\Sigma^-(5) = \Sigma(5) =$$

$$= \Phi_0^5 \, \Sigma(0) \, \left(\Phi_0^5\right)^T + \overbrace{\int_0^5 \Phi_t^5 \, Q_n(t) \, \left(\Phi_t^5\right)^T dt}^{\Theta_0^5} =$$

$$= \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 5 & 1 \end{bmatrix} +$$

$$+ \int_0^5 \begin{bmatrix} 1 & 5-t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 5-t & 1 \end{bmatrix} dt =$$

$$= \begin{bmatrix} 25\,002 & 5000 \\ 5000 & 1000 \end{bmatrix} + \int_0^5 \begin{bmatrix} (5-t)^2 & 5-t \\ 5-t & 1 \end{bmatrix} dt =^{1}$$

$$= \begin{bmatrix} 25\,002 & 5000 \\ 5000 & 1000 \end{bmatrix} + \begin{bmatrix} -\frac{1}{3}(5-t)^3 & -\frac{1}{2}(5-t)^2 \\ -\frac{1}{2}(5-t)^2 & t \end{bmatrix}_0^5 =$$

$$= \begin{bmatrix} 25\,002 & 5000 \\ 5000 & 1000 \end{bmatrix} + \begin{bmatrix} \frac{125}{3} & \frac{25}{2} \\ \frac{25}{2} & 5 \end{bmatrix} \approx$$

$$\approx \begin{bmatrix} 25\,043.7 & 5012.5 \\ 5012.5 & 1005.0 \end{bmatrix}.$$

(The numbers presented are rounded from the exact values used in the calculation.)

2. Matrix $H = \begin{bmatrix} 1 & 0 \end{bmatrix}$. So[2] 2

$$H\Sigma^- H^T + R = 25\,043.7 + 3 = 25\,046.7.$$

[1] Alternatively, expansion into polynomials:

$$\int_0^5 \begin{bmatrix} (5-t)^2 & 5-t \\ 5-t & 1 \end{bmatrix} dt = \int_0^5 \begin{bmatrix} 25 - 10t + t^2 & 5-t \\ 5-t & 1 \end{bmatrix} dt =$$

$$= \begin{bmatrix} 25t - 5t^2 + \frac{1}{3}t^3 & 5t - \frac{1}{2}t^2 \\ 5t - \frac{1}{2}t^2 & t \end{bmatrix}_0^5 =$$

$$= \begin{bmatrix} 125 - 125 + \frac{1}{3} \cdot 125 & 25 - \frac{1}{2} \cdot 25 \\ 25 - \frac{1}{2} \cdot 25 & 5 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{125}{3} & \frac{25}{2} \\ \frac{25}{2} & 5 \end{bmatrix}.$$

[2] Notation: $\Sigma^- = \Sigma(5)$ is the *a priori* state variance matrix, just before making the observation at time $t = 5$.

The K matrix is

$$K = \Sigma^- H^T \left(H\Sigma^- H^T + R\right)^{-1} = \\ = \begin{bmatrix} 25\,043.7 \\ 5012.5 \end{bmatrix} \cdot \frac{1}{25\,046.7} = \begin{bmatrix} 0.999\,88 \\ 0.200\,13 \end{bmatrix}.$$

Next, calculate the zero quantity or "closing error"

$$\underline{\mathbf{y}} = H\mathbf{x}^-(5) - \underline{\ell}' = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 0 \end{bmatrix} - 3 = 1.$$

Then

$$\mathbf{x}^+(5) = \mathbf{x}^-(5) - K\underline{\mathbf{y}} = \\ = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.999\,88 \\ 0.200\,13 \end{bmatrix} \cdot 1 = \begin{bmatrix} 3.000\,12 \\ -0.200\,13 \end{bmatrix}.$$

Project this back to the moment $t = 0$:

$$\begin{aligned} \widehat{x}(0) &= x^+(5) - v^+(5) \cdot 5 = \\ &= 3.000\,12 - (-0.200\,13) \cdot 5 = 4.000\,77, \\ \widehat{v}(0) &= v^+(5) = -0.200\,13. \end{aligned} \tag{4.2}$$

The updated state variance matrix $\Sigma^+(5)$ is

$$\Sigma^+(5) = (I - KH)\,\Sigma^-(5) = \\ = \left(I - \begin{bmatrix} 0.999\,88 \\ 0.200\,13 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}\right) \begin{bmatrix} 25\,043.7 & 5012.5 \\ 5012.5 & 1005.0 \end{bmatrix} = \\ = \begin{bmatrix} 2.999\,64 & 0.600\,38 \\ 0.600\,38 & 1.866\,27 \end{bmatrix}. \tag{4.3}$$

3. Because

$$\begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \end{bmatrix} + \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \end{bmatrix} = \begin{bmatrix} \widehat{x}(0) \\ \widehat{x}(5) \end{bmatrix} = \\ = \begin{bmatrix} \widehat{x}(0) \\ \widehat{x}(0) + 5 \cdot \widehat{v}(0) \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & 0 \\ 1 & 5 \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} \widehat{x}(0) \\ \widehat{v}(0) \end{bmatrix}}^{\widehat{\mathbf{x}}},$$

rakennematriisi

the design matrix is

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 5 \end{bmatrix},$$

and the observation vector $\underline{\ell}$ and the observation variance matrix S are given, equations 4.1. We obtain the solution

$$A^T S^{-1} A = \frac{1}{6}\begin{bmatrix} 5 & 10 \\ 10 & 50 \end{bmatrix},$$

$$\widehat{\mathbf{x}} = \left(A^T S^{-1} A\right)^{-1} A^T S^{-1} \underline{\ell} = \frac{1}{5}\begin{bmatrix} 20 \\ -1 \end{bmatrix}.$$

This is the same result, practically, as that under point 2, equations 4.2. The variance of the solution is

$$\mathrm{Var}\{\widehat{\mathbf{x}}\} = \left(A^T S^{-1} A\right)^{-1} = \frac{1}{5}\begin{bmatrix} 10 & -2 \\ -2 & 1 \end{bmatrix},$$

which is not directly comparable to the earlier result 4.3 as this result refers to the moment in time or epoch $t = 0$. Furthermore, the Kalman solution contains the effect of the dynamic noise Q_n, which is not considered in the traditional least-squares solution.

4.2 Example 2: spinning wheel

Question An industrial machine has a wheel with radius r spinning at an angular velocity $\omega(t)$, where t is time. The instantaneous angular velocity varies randomly: the angular acceleration has the properties of "white noise". omega $\omega\Omega$

1. Write the state vector of this system. How many elements are needed?
2. Write the dynamic model of the system.
3. A reflective prism is attached to the edge of the wheel in order to make measurements. The rotation is monitored using laser ranging. The measuring device is at a great distance from the machine, within the plane of the wheel. Write the observation model.
4. Linearise the observation model.
5. In what way, as a consequence of linearisation, does the dynamic model change?

alpha αA

Answer

1. The state vector of this system contains the angular position $\alpha(t)$. However, it is given that the *angular acceleration* $\frac{d}{dt}\omega(t)$ has the properties of white noise. This makes it a good idea to include the angular velocity into the state vector as well. Thus we obtain for the state vector
$$\mathbf{x}(t) = \begin{bmatrix} \alpha(t) \\ \omega(t) \end{bmatrix}.$$

2. The dynamic model in the Kalman filter is a system of equations of the form
$$\frac{d}{dt}\underline{\mathbf{x}}(t) = F\big(\underline{\mathbf{x}}(t), t\big) + \underline{\mathbf{n}},$$
in which $\underline{\mathbf{x}}$ is the state vector of the system and $\underline{\mathbf{n}}$ is the dynamic noise vector.
The state vector is given above. Write
$$\frac{d}{dt}\begin{bmatrix} \underline{\alpha} \\ \underline{\omega} \end{bmatrix} = \begin{bmatrix} \underline{\omega} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n}_\omega \end{bmatrix}, \qquad (4.4)$$
in which the first equation, $\frac{d}{dt}\underline{\alpha} = \underline{\omega}$, expresses the definition of angular velocity ω, and the second equation, $\frac{d}{dt}\underline{\omega} = \underline{n}_\omega$, expresses the given fact that the angular acceleration has the properties of white noise.
We observe that the dynamic model found is *linear*:
$$\frac{d}{dt}\begin{bmatrix} \underline{\alpha} \\ \underline{\omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \underline{\alpha} \\ \underline{\omega} \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n}_\omega \end{bmatrix}.$$

3. If the distance to a prism on the edge of the wheel is observed from far away, the observation equation may be written as
$$\underline{s} = d + r\cos\underline{\alpha} + \underline{m}. \qquad (4.5)$$
The angle α is reckoned from the prism position furthest away from the observing instrument. Assume that d, the distance between the instrument and the wheel axis, is known. If it is not, it should be added to the state vector with the added dynamic equation $\frac{d}{dt}\underline{d} = 0$.

4. This model is non-linear: the dependence of the observable on the state-vector element is a cosine.

   We linearise as follows: define consistent approximate values for which

$$s^{(0)} = d + r\cos\alpha^{(0)}$$

   and subtract this from equation 4.5. The result is a Taylor expansion truncated after the first, linear term in $\Delta\alpha$:

$$\Delta\underline{s} = r\,\frac{\partial}{\partial\alpha}\cos\alpha\Big|_{\alpha=\alpha^{(0)}} \cdot \Delta\underline{\alpha} + \underline{m}.$$

   Here, the logical definitions $\Delta\underline{s} \overset{\text{def}}{=} \underline{s} - s^{(0)}$ and $\Delta\underline{\alpha} \overset{\text{def}}{=} \underline{\alpha} - \alpha^{(0)}$ have been applied.

   Partial differentiation yields

$$\underbrace{\Delta\underline{s}}_{\underline{\boldsymbol{\ell}}} = -r\sin\alpha^{(0)}\,\Delta\underline{\alpha} + \underline{m} =$$

$$= \overbrace{\begin{bmatrix} -r\sin\alpha^{(0)} & 0 \end{bmatrix}}^{H} \overbrace{\begin{bmatrix} \Delta\underline{\alpha} \\ \Delta\underline{\omega} \end{bmatrix}}^{\underline{\mathbf{x}}} + \overbrace{\underline{m}}^{\underline{\mathbf{m}}},$$

   a linear Kalman observation equation of type

$$\underline{\boldsymbol{\ell}} = H\underline{\mathbf{x}} + \underline{\mathbf{m}},$$

   if we write formally

$$\underline{\boldsymbol{\ell}} = \begin{bmatrix} \Delta\underline{s} \end{bmatrix}, \qquad H = \begin{bmatrix} -r\sin\alpha^{(0)} & 0 \end{bmatrix},$$

$$\underline{\mathbf{x}} = \begin{bmatrix} \Delta\underline{\alpha} \\ \Delta\underline{\omega} \end{bmatrix}, \qquad \underline{\mathbf{m}} = \begin{bmatrix} \underline{m} \end{bmatrix}.$$

5. A consistent dynamic model for an approximate state vector is

$$\frac{d}{dt}\begin{bmatrix} \underline{\alpha}^{(0)} \\ \underline{\omega}^{(0)} \end{bmatrix} = \begin{bmatrix} \underline{\omega}^{(0)} \\ 0 \end{bmatrix},$$

   subtraction of which from dynamic model 4.4 yields

$$\frac{d}{dt}\begin{bmatrix} \Delta\underline{\alpha} \\ \Delta\underline{\omega} \end{bmatrix} = \begin{bmatrix} \Delta\underline{\omega} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n}_\omega \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \Delta\underline{\alpha} \\ \Delta\underline{\omega} \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n}_\omega \end{bmatrix},$$

   with $\Delta\omega \overset{\text{def}}{=} \underline{\omega} - \omega^{(0)}$. This is the *linearised* dynamic model, with the state vector now consisting of delta quantities.

4.3 Example 3: a parachute-jumper

Question

1. Write the dynamic equations for a parachute-jumper in one dimension, with only the height z as a co-ordinate. The acceleration of gravity g is a constant, the braking acceleration caused by air drag is proportional to the velocity of falling and air density ρ, which may be described by the equation

rho ρR

$$\rho = \rho_0 \exp\left(-\frac{z}{H}\right).$$

The constant H is the scale height of the atmosphere and ρ_0 is air density at sea level.

2. A reflective tag is attached to the jumper in order to obtain measurements. A tacheometer on the ground measures the distance to this reflector. The horizontal distance between the tacheometer and touch-down point is given. The jumper comes down vertically, there is no wind.

Write the observation model.

Answer

1. The dynamic model is, with k the air-drag constant:

$$\frac{d^2}{dt^2}\underline{z} = -g - k\rho\,\dot{\underline{z}} + \underline{n} = -g - k\rho_0 \exp\left(-\frac{\underline{z}}{H}\right)\dot{\underline{z}} + \underline{n}.$$

Define the state vector as $\begin{bmatrix} \underline{z} & \dot{\underline{z}} \end{bmatrix}^T$ and obtain as the dynamic model the first-order differential equation

$$\frac{d}{dt}\begin{bmatrix} \underline{z} \\ \dot{\underline{z}} \end{bmatrix} = \begin{bmatrix} \dot{\underline{z}} \\ -g - k\rho_0\,\dot{\underline{z}} \exp\left(-\frac{\underline{z}}{H}\right) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n} \end{bmatrix}. \tag{4.6}$$

This equation is non-linear. Write

$$\begin{bmatrix} \underline{z} \\ \dot{\underline{z}} \end{bmatrix} = \begin{bmatrix} z^{(0)} \\ \dot{z}^{(0)} \end{bmatrix} + \begin{bmatrix} \Delta\underline{z} \\ \Delta\dot{\underline{z}} \end{bmatrix},$$

in which

$$\frac{d}{dt}\begin{bmatrix} z^{(0)} \\ \dot{z}^{(0)} \end{bmatrix} = \begin{bmatrix} \dot{z}^{(0)} \\ -g - k\rho_0\,\dot{z}^{(0)} \exp\left(-\frac{z^{(0)}}{H}\right) \end{bmatrix}. \tag{4.7}$$

This equation may be integrated if initial conditions are given, for example $z^{(0)}(t_0) = h$, $\dot{z}^{(0)}(t_0) = 0$, with h the nominal aircraft height.

Subtract equation 4.7 from equation 4.6:

$$\frac{d}{dt}\begin{bmatrix} \underline{z} - z^{(0)} \\ \underline{\dot z} - \dot z^{(0)} \end{bmatrix} =$$

$$= \begin{bmatrix} \underline{\dot z} - \dot z^{(0)} \\ -k\rho_0\left(\underline{\dot z}\exp\left(-\frac{\underline{z}}{H}\right) - \dot z^{(0)}\exp\left(-\frac{z^{(0)}}{H}\right)\right) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n} \end{bmatrix}$$

$$\Longrightarrow \frac{d}{dt}\begin{bmatrix} \Delta\underline{z} \\ \Delta\underline{\dot z} \end{bmatrix} = \begin{bmatrix} \Delta\underline{\dot z} \\ -k\rho_0\,\Delta\left(\underline{\dot z}\exp\left(-\frac{\underline{z}}{H}\right)\right) \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n} \end{bmatrix},$$

in which $\Delta\underline{z} = \underline{z} - z^{(0)}$, $\Delta\underline{\dot z} = \underline{\dot z} - \dot z^{(0)}$ and (linearisation)

$$\Delta\left(\underline{\dot z}\exp\left(-\frac{\underline{z}}{H}\right)\right) = \underline{\dot z}\exp\left(-\frac{\underline{z}}{H}\right) - \dot z^{(0)}\exp\left(-\frac{z^{(0)}}{H}\right) =$$

$$= \underline{\dot z}\exp\left(-\frac{\underline{z}}{H}\right) - \underline{\dot z}\exp\left(-\frac{z^{(0)}}{H}\right) +$$

$$+\underline{\dot z}\exp\left(-\frac{z^{(0)}}{H}\right) - \dot z^{(0)}\exp\left(-\frac{z^{(0)}}{H}\right) =$$

$$= \underline{\dot z}\cdot\Delta\left(\exp\left(-\frac{\underline{z}}{H}\right)\right) + \Delta\underline{\dot z}\cdot\exp\left(-\frac{z^{(0)}}{H}\right) \approx$$

$$\approx \dot z^{(0)}\left(-\frac{1}{H}\exp\left(-\frac{z^{(0)}}{H}\right)\Delta\underline{z}\right) + \exp\left(-\frac{z^{(0)}}{H}\right)\Delta\underline{\dot z} =$$

$$= \exp\left(-\frac{z^{(0)}}{H}\right)\left(\Delta\underline{\dot z} - \frac{\dot z^{(0)}}{H}\Delta\underline{z}\right).$$

Abbreviate

$$Z \stackrel{\text{def}}{=} \exp\left(-\frac{z^{(0)}}{H}\right).$$

Substitution then yields

$$\frac{d}{dt}\overbrace{\begin{bmatrix} \Delta\underline{z} \\ \Delta\underline{\dot z} \end{bmatrix}}^{\underline{x}} = \overbrace{\begin{bmatrix} \Delta\underline{\dot z} \\ -k\rho_0 Z\left(\Delta\underline{\dot z} - \frac{\dot z^{(0)}}{H}\Delta\underline{z}\right) \end{bmatrix}}^{F(\underline{x},t)} + \overbrace{\begin{bmatrix} 0 \\ \underline{n} \end{bmatrix}}^{\underline{n}} =$$

$$= \overbrace{\begin{bmatrix} 0 & 1 \\ k\rho_0\frac{\dot z^{(0)}}{H}Z & -k\rho_0 Z \end{bmatrix}}^{F(t)}\overbrace{\begin{bmatrix} \Delta\underline{z} \\ \Delta\underline{\dot z} \end{bmatrix}}^{\underline{x}} + \begin{bmatrix} 0 \\ \underline{n} \end{bmatrix}.$$

This is the linearised dynamic model.

2. Let the horizontal distance between the touch-down point of the parachutist and the tacheometer be r. Then the measured distance is

$$s = \sqrt{r^2 + z^2}$$

and the observation equation

$$\underline{s} = \sqrt{r^2 + \underline{z}^2} + \underline{m}.$$

Linearisation — $s = s_0 + \Delta s$ with $s_0 = \sqrt{r^2 + z_0^2}$ — yields, with

$$\frac{\partial s}{\partial z} = \frac{\partial \left(z^2\right)}{\partial z} \cdot \frac{\partial}{\partial \left(z^2\right)} \sqrt{r^2 + \left(z^2\right)} = 2z \cdot \frac{1}{2} \frac{1}{\sqrt{r^2 + \left(z^2\right)}} = \frac{z}{s},$$

the linearised observation equation

$$\overbrace{\Delta \underline{s}}^{\underline{\ell}} = \frac{z_0}{s_0} \Delta \underline{z} + \underline{m} = \overbrace{\left[\begin{array}{cc} \frac{z_0}{s_0} & 0 \end{array}\right]}^{H} \overbrace{\left[\begin{array}{c} \Delta \underline{z} \\ \Delta \underline{\dot{z}} \end{array}\right]}^{\mathbf{x}} + \overbrace{\underline{m}}^{\mathbf{m}}.$$

4.4 Modelling of realistic statistical behaviour

Coloured-noise processes, specifically Gauss-Markov processes, are very often used to model stochastic processes found in real life.

For example we may know that a measured stochastic process consists of two parts. One of these is the *signal* $\underline{s}$, a process that we are interested in, which varies rapidly on a time scale τ_s. We know this signal to have an expectancy of zero. The other is a systematic disturbance or *bias* $\underline{b}$, that we want to get rid of. We also know that this disturbance is slowly varying on a time scale characterised by the constant τ_b.

tau τT

odotusarvo

The state vector may be written as $\left[\begin{array}{cc} \underline{s} & \underline{b} \end{array}\right]^T$ and the dynamic model equation as

$$\frac{d}{dt} \left[\begin{array}{c} \underline{s} \\ \underline{b} \end{array}\right] = \left[\begin{array}{cc} -1/\tau_s & 0 \\ 0 & -1/\tau_b \end{array}\right] \left[\begin{array}{c} \underline{s} \\ \underline{b} \end{array}\right] + \left[\begin{array}{c} \underline{n}_s \\ \underline{n}_b \end{array}\right].$$

Here, τ_b is the long time constant of the disturbance process, which will thus be slowly varying. A much shorter value may be chosen for the time constant τ_s. It should, however, be chosen realistically. If observations are obtained at a time interval Δt, then we must have $\tau_s \gtrsim 2\,\Delta t$ in order for the process $\underline{s}$ to be realistically determinable from the observations.

The observation equation is

$$\underline{\ell} = \underline{s} + \underline{b} + \underline{m},$$

with $\underline{m}$ and its variance R representing the observation noise or uncertainty. If observations are obtained sufficiently densely in time, we may obtain separate estimates for the signal process $\underline{s}$ and the slowly changing bias $\underline{b}$. In order for this to work, we should attach realistic autocovariance values to the processes $\underline{n}_s$ and $\underline{n}_b$. It is also required that $E\{\underline{s}\} = 0$. If it is not, the systematic part of $\underline{s}$ will end up in the estimator $\widehat{b}$ produced by the filter.

This is a case of *spectral filtering* by Kalman filter. The low-frequency part, including zero frequency, goes to estimator $\widehat{b}$, the high-frequency part goes to $\widehat{s}$. The boundary between the two spectral areas is not sharp.

Somewhat the opposite situation arises if we have a measured stochastic process consisting of a rapidly varying noise part and a slowly varying signal. Assume that the noise is not white, but rather, "coloured". Let us call it $\underline{c}$. It has a correlation length τ_c. If we are interested only in the lower-frequency constituents of the signal $\underline{s}$, we may again apply a Kalman filter:

$$\frac{d}{dt}\begin{bmatrix} \underline{s} \\ \underline{c} \end{bmatrix} = \begin{bmatrix} -1/\tau_s & 0 \\ 0 & -1/\tau_c \end{bmatrix}\begin{bmatrix} \underline{s} \\ \underline{c} \end{bmatrix} + \begin{bmatrix} \underline{n}_s \\ \underline{n}_c \end{bmatrix}.$$

We choose τ_s according to the part of the spectrum of $\underline{s}$ that we are interested in — but always $\tau_s > \tau_c$. The time constant τ_c should be chosen realistically, to capture and remove to the maximum extent the real noise present in the process. The observation equation is again

$$\underline{\ell} = \underline{s} + \underline{c} + \underline{m}.$$

The earlier described technique for extracting a rapidly varying signal from a background of slowly varying bias was used for extracting a signal on underground mass concentrations or *mascons* on the Moon from Lunar Orbiter Doppler tracking data (Tapley and Schutz, 1975). The technique is called "dynamic model compensation".

4.5 The Kalman filter as sequential adjustment

The update step of the Kalman filter may also be written as a parametric adjustment problem.

The "observations" are the real observation vector for the update $\underline{\boldsymbol{\ell}}'$ and the *a priori* estimated state vector $\mathbf{x}^-(t)$ for the moment t of the update.

The parametric observation equations are in the standard form

$$\overbrace{\begin{bmatrix} \underline{\boldsymbol{\ell}}' \\ \mathbf{x}^-(t) \end{bmatrix}}^{\underline{\boldsymbol{\ell}}} + \overbrace{\begin{bmatrix} \underline{\mathbf{v}}' \\ \underline{\mathbf{v}}'' \end{bmatrix}}^{\underline{\mathbf{v}}} = \overbrace{\begin{bmatrix} H \\ I \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} \mathbf{x}^+(t) \end{bmatrix}}^{\widehat{\mathbf{x}}}.$$

The design matrix is seen to be

$$A = \begin{bmatrix} H \\ I \end{bmatrix}.$$

The variance matrix of the "observations" is

$$S = \operatorname{Var}\left\{ \begin{bmatrix} \underline{\boldsymbol{\ell}}' \\ \mathbf{x}^- \end{bmatrix} \right\} = \begin{bmatrix} R & 0 \\ 0 & \Sigma^- \end{bmatrix},$$

and the *a posteriori* state-vector solution is

$$\begin{aligned} \mathbf{x}^+ &= \left(A^T S^{-1} A\right)^{-1} A^T S^{-1} \underline{\boldsymbol{\ell}} = \\ &= \left(H^T R^{-1} H + (\Sigma^-)^{-1}\right)^{-1} \left(H^T R^{-1} \underline{\boldsymbol{\ell}}' + (\Sigma^-)^{-1} \mathbf{x}^-\right). \end{aligned} \tag{4.8}$$

The *a posteriori* state variance is

$$\Sigma^+ = \left(H^T R^{-1} H + (\Sigma^-)^{-1}\right)^{-1}. \tag{4.9}$$

Now we apply the Woodbury identity derived in appendix C,

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + V A^{-1} U\right)^{-1} V A^{-1}, \tag{C.5}$$

in this way:

$$\left(H^T R^{-1} H + (\Sigma^-)^{-1}\right)^{-1} = \Sigma^- - \Sigma^- H^T \left(R + H \Sigma^- H^T\right)^{-1} H \Sigma^-. \tag{4.10}$$

Substitution yields

$$\begin{aligned} \mathbf{x}^+ &= \left(\Sigma^- - \Sigma^- H^T \left(R + H\Sigma^- H^T\right)^{-1} H \Sigma^-\right) \left(H^T R^{-1} \underline{\boldsymbol{\ell}}' + (\Sigma^-)^{-1} \mathbf{x}^-\right) = \\ &= \Sigma^- H^T R^{-1} \underline{\boldsymbol{\ell}}' + \mathbf{x}^- - \Sigma^- H^T \left(R + H\Sigma^- H^T\right)^{-1} H \left(\Sigma^- H^T R^{-1} \underline{\boldsymbol{\ell}}' + \mathbf{x}^-\right) = \\ &= \mathbf{x}^- + \overbrace{\Sigma^- H^T R^{-1} \underline{\boldsymbol{\ell}}' - \Sigma^- H^T \left(R + H\Sigma^- H^T\right)^{-1} H \Sigma^- H^T R^{-1} \underline{\boldsymbol{\ell}}'}^{I} - \\ &\qquad - \Sigma^- H^T \left(R + H \Sigma^- H^T\right)^{-1} H \mathbf{x}^-, \end{aligned}$$

in which

$$
\begin{aligned}
I = &-\Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1} H\Sigma^- H^\mathsf{T} R^{-1}\underline{\boldsymbol{\ell}}' = \\
&= -\Sigma^- H^\mathsf{T}\cancel{\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1}\left(R + H\Sigma^- H^\mathsf{T}\right)} R^{-1}\underline{\boldsymbol{\ell}}' + \\
&\quad + \Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1}\cancel{R\cdot R^{-1}}\,\underline{\boldsymbol{\ell}}' = \\
&\qquad = -\Sigma^- H^\mathsf{T} R^{-1}\underline{\boldsymbol{\ell}}' + \Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1}\underline{\boldsymbol{\ell}}',
\end{aligned}
$$

yielding

$$
\begin{aligned}
\mathbf{x}^+ = \mathbf{x}^- &+ \cancel{\Sigma^- H^\mathsf{T} R^{-1}\underline{\boldsymbol{\ell}}' - \Sigma^- H^\mathsf{T} R^{-1}\underline{\boldsymbol{\ell}}'} + \\
&+ \Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1}\underline{\boldsymbol{\ell}}' - \Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1} H\mathbf{x}^- = \\
&= \mathbf{x}^- - \Sigma^- H^\mathsf{T}\left(H\Sigma^- H^\mathsf{T} + R\right)^{-1}\left(H\mathbf{x}^- - \underline{\boldsymbol{\ell}}'\right). \qquad (4.11)
\end{aligned}
$$

In addition, equations 4.9 and 4.10 yield

$$
\Sigma^+ = \Sigma^- - \Sigma^- H^\mathsf{T}\left(R + H\Sigma^- H^\mathsf{T}\right)^{-1} H\Sigma^-. \qquad (4.12)
$$

Equations 4.11 and 4.12 are precisely the update equations of the Kalman filter. Compared to equations 4.8 and 4.9, the matrix to be inverted has the size of the vector of observables $\underline{\boldsymbol{\ell}}'$ and not that of the state vector $\underline{\mathbf{x}}$. Often the matrix size is even 1×1, a simple number.[3] Being able to compute inverse matrices more quickly makes real-time applications easier.

3 tosiaikainen

From the preceding discussion we see that sequential adjustment is the same as Kalman filtering in the case where the state vector is constant in time. The calculation procedure in adjustment generally is parametric adjustment with observation equations, whereas in the Kalman-filter case, adjustment by condition equations is used.

4.6 Using the Kalman filter "from both ends"

In airborne gravimetry, chapter 12, the Kalman filter can be used to process the collected observations "on the fly", as they come in. However, generally we will want to use post-processing for extracting the highest-quality results from the raw observations. For this, batch-processing techniques such as least-squares collocation are available.

---

[3]... or may be reduced to such, if the observations made at one epoch are statistically independent of each other. Then they may be formally processed sequentially: that is, separately.

It is however also possible to harness the Kalman filter for this, by processing the observations both forwards and backwards in time and optimally combining the results of both. This may be a pre-processing step for batch processing. We will describe this approach below.

4.6.1 Observation equations and normal equations

If observations $\underline{\ell}_k$, $k = 1, \ldots, n$ are available and the dynamic model is the system of differential equations

$$\frac{d}{dt}\mathbf{x}(t) = F(t)\,\mathbf{x}(t) \tag{4.13}$$

— written here non-stochastically without dynamic noise $\underline{\mathbf{n}}$ — one may write

$$\mathbf{x}(t_k) = \Phi_0^k\,\mathbf{x}(t_0),$$

phi φϕΦ

in which Φ_0^k is the state transition matrix, which can be computed. Thus, the observation equations may be written into the standard form

$$\underline{\ell}_k + \underline{v}_k = H_k\,\widehat{\mathbf{x}}(t_k) = H_k\Phi_0^k\,\widehat{\mathbf{x}}(t_0), \qquad k = 1, \ldots, n,$$

a traditional system of observation equations

$$\underline{\boldsymbol{\ell}} + \underline{\mathbf{v}} = A\widehat{\mathbf{X}},$$

in which the design matrix, vectors of observations, residuals, and unknowns are

$$A = \begin{bmatrix} H_1\Phi_0^1 \\ \vdots \\ H_k\Phi_0^k \\ \vdots \\ H_n\Phi_0^n \end{bmatrix}, \qquad \underline{\boldsymbol{\ell}} = \begin{bmatrix} \underline{\ell}_1 \\ \vdots \\ \underline{\ell}_k \\ \vdots \\ \underline{\ell}_n \end{bmatrix}, \qquad \underline{\mathbf{v}} = \begin{bmatrix} \underline{v}_1 \\ \vdots \\ \underline{v}_k \\ \vdots \\ \underline{v}_n \end{bmatrix}, \qquad \widehat{\mathbf{X}} = \begin{bmatrix} \widehat{\mathbf{x}}(t_0) \end{bmatrix}.$$

From this we see that the least-squares solution can be obtained by solving an adjustment problem.

We may divide the observations into two parts, "before" ("<") and "after" (">") a certain point in time. Then the vector of observations, its variance-covariance matrix, and the design matrix of the observation equations are

$$\underline{\boldsymbol{\ell}} = \begin{bmatrix} \underset{<}{\underline{\boldsymbol{\ell}}} \\ \underset{>}{\underline{\boldsymbol{\ell}}} \end{bmatrix}, \qquad S = \begin{bmatrix} \underset{<}{S} & 0 \\ 0 & \underset{>}{S} \end{bmatrix}, \qquad A = \begin{bmatrix} \underset{<}{A} \\ \underset{>}{A} \end{bmatrix}.$$

In this way, separate normal equations are formed:

$$\left(\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A}\right)\underset{<}{\widehat{\mathbf{x}}}=\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}},\qquad \left(\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A}\right)\underset{>}{\widehat{\mathbf{x}}}=\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}},$$

with solutions

$$\underset{<}{\widehat{\mathbf{x}}}=\left(\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A}\right)^{-1}\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}},\qquad \underset{>}{\widehat{\mathbf{x}}}=\left(\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A}\right)^{-1}\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}},$$

and separate solution variances

$$\underset{<}{\Sigma}=\left(\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A}\right)^{-1},\qquad \underset{>}{\Sigma}=\left(\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A}\right)^{-1}.$$

On the other hand, the single normal equation of the *full adjustment* is

$$A^{\mathsf{T}}S^{-1}A\,\widehat{\mathbf{x}}=A^{\mathsf{T}}S^{-1}\,\underline{\boldsymbol{\ell}}.$$

We assume the observations $\underset{<}{\underline{\boldsymbol{\ell}}}$ and $\underset{>}{\underline{\boldsymbol{\ell}}}$ to be statistically independent. This is why matrix S is block diagonal, and we may decompose the normal equation as follows:

$$\begin{bmatrix}\underset{<}{A}^{\mathsf{T}} & \underset{>}{A}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\underset{<}{S} & 0\\ 0 & \underset{>}{S}\end{bmatrix}^{-1}\begin{bmatrix}\underset{<}{A}\\ \underset{>}{A}\end{bmatrix}\widehat{\mathbf{x}}=\begin{bmatrix}\underset{<}{A}^{\mathsf{T}} & \underset{>}{A}^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\underset{<}{S} & 0\\ 0 & \underset{>}{S}\end{bmatrix}^{-1}\begin{bmatrix}\underset{<}{\underline{\boldsymbol{\ell}}}\\ \underset{>}{\underline{\boldsymbol{\ell}}}\end{bmatrix}$$

or

$$\left(\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A}+\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A}\right)\widehat{\mathbf{x}}=\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}}+\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}}.$$

Define *weight matrices*

$$\underset{<}{P}\stackrel{\text{def}}{=}\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A},\qquad \underset{>}{P}\stackrel{\text{def}}{=}\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A},\tag{4.14}$$

and the normal equation becomes

$$\left(\underset{<}{P}+\underset{>}{P}\right)\widehat{\mathbf{x}}=\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}}+\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}}.$$

The solution is

$$\begin{aligned}\widehat{\mathbf{x}}&=\left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\left(\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}}+\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}}\right)=\\ &=\left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\left(\underset{<}{P}\underbrace{\left(\overbrace{\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{A}}^{\underset{<}{P}}\right)^{-1}\underset{<}{A}^{\mathsf{T}}\underset{<}{S}^{-1}\underset{<}{\underline{\boldsymbol{\ell}}}}_{\underset{<}{\widehat{\mathbf{x}}}}+\underset{>}{P}\underbrace{\left(\overbrace{\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{A}}^{\underset{>}{P}}\right)^{-1}\underset{>}{A}^{\mathsf{T}}\underset{>}{S}^{-1}\underset{>}{\underline{\boldsymbol{\ell}}}}_{\underset{>}{\widehat{\mathbf{x}}}}\right)=\\ &=\left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\left(\underset{<}{P}\,\underset{<}{\widehat{\mathbf{x}}}+\underset{>}{P}\,\underset{>}{\widehat{\mathbf{x}}}\right),\end{aligned}$$

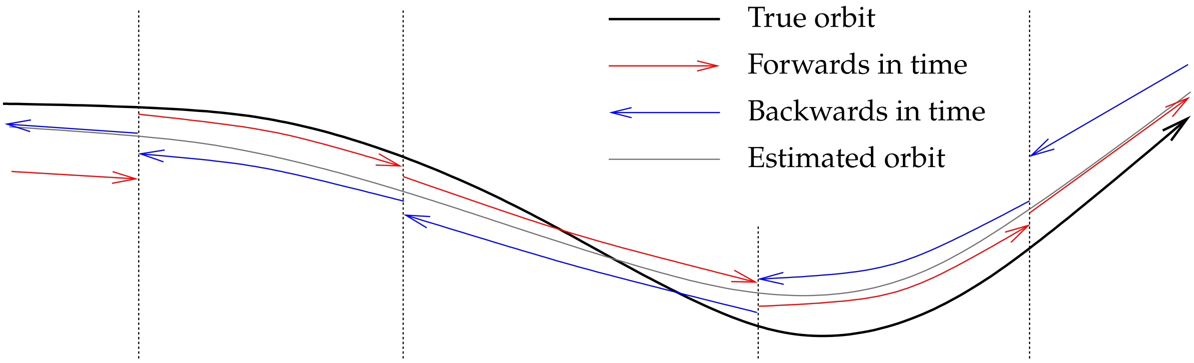


FIGURE 4.1. The Kalman filter used forwards and backwards in time.

and

$$\Sigma = \left(\underset{<}{A}^T \underset{<}{S}^{-1} \underset{<}{A} + \underset{>}{A}^T \underset{>}{S}^{-1} \underset{>}{A}\right)^{-1} = \left(\underset{<}{P} + \underset{>}{P}\right)^{-1} = \left(\underset{<}{\Sigma}^{-1} + \underset{>}{\Sigma}^{-1}\right)^{-1}$$

is the variance matrix of the solution from the full adjustment.

This shows that the separate solutions "<" and ">" can be "stacked" or combined into the full solution:

$$\widehat{\mathbf{x}} = \left(\underset{<}{P} + \underset{>}{P}\right)^{-1} \left(\underset{<}{P}\,\underset{<}{\widehat{\mathbf{x}}} + \underset{>}{P}\,\underset{>}{\widehat{\mathbf{x}}}\right), \qquad \Sigma = \left(\underset{<}{P} + \underset{>}{P}\right)^{-1}, \tag{4.15}$$

in other words the *weighted average* of the partial solutions.

4.6.2 The forwards-backwards Kalman filter

It is now important here that the partial tasks — "before", "<", and "after", ">" — can also be solved using the Kalman filter! In other words, we may for all times $t \in [t_0, t_n]$ calculate separately

1. The solution of the Kalman filter from the starting time t_0 forwards, by integrating the dynamic model and updating the state vector and its variance matrix using the observations $\underline{\ell}_1, \ldots, \underline{\ell}_k, \ldots, \underline{\ell}_n$.
2. The Kalman filter solution from the final moment t_n backwards in time, integrating the dynamic model and updating the state vector and the variance matrix using the observations $\underline{\ell}_n, \ldots, \underline{\ell}_k, \ldots, \underline{\ell}_1$ in reverse order. The Kalman filter equation to be used for that is, based on equation 4.13:
$$\frac{d}{dt'}\mathbf{x}'(t') = F'(t')\,\mathbf{x}'(t'),$$
in which $t' = -t$, $\mathbf{x}'(t') = \mathbf{x}(t)$ and $F'(t') = -F(t)$.
3. The total solution by combining the partial solutions is obtained using the above equations 4.15 and 4.14.

In this way, the advantages of the Kalman method may also be exploited in a post-processing situation. And note that equation 4.13, which we used in our derivation, contains no dynamic noise $\underline{\mathbf{n}}$. In reality, this limitation does not exist: one *can* include dynamic noise in the Kalman-filter model. Integrate forwards and backwards, take the weighted average of the two results for the whole time line, and obtain the full solution. This is a significant advantage over batch processing.

4.6.3 *Example 4: random walk in both directions*

A random walk is described by the equation

$$\frac{d}{dt}\underline{w}(t) = \underline{n}(t), \tag{2.25}$$

with autocovariance according to the equation

$$A_{w,0}(t,t) = Q_n \cdot (t - t_0)\,. \tag{2.26}$$

The random walk is observed at two points in time, t_1 and t_2, obtaining observation values

$$\underline{\ell}_1 = \underline{w}(t_1), \qquad\qquad \underline{\ell}_2 = \underline{w}(t_2),$$

both assumed to be errorless.

Then, in the interval (t_1, t_2) the forwards solution is

$$\underset{<}{\widehat{w}}(t) = \underline{\ell}_1, \qquad\qquad \underset{<}{\Sigma}(t) = Q_n \cdot (t - t_1)\,, \tag{4.16}$$

and the backwards solution

$$\underset{>}{\widehat{w}}(t) = \underline{\ell}_2, \qquad\qquad \underset{>}{\Sigma}(t) = Q_n \cdot (t_2 - t)\,. \tag{4.17}$$

Do the weighted averaging between forwards and backwards solutions. Start from equation 4.15:

$$\widehat{\mathbf{x}} = \left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\left(\underset{<}{P}\underset{<}{\widehat{\mathbf{x}}}+\underset{>}{P}\underset{>}{\widehat{\mathbf{x}}}\right) = \left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\underset{<}{P}\underset{<}{\widehat{\mathbf{x}}} + \left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\underset{>}{P}\underset{>}{\widehat{\mathbf{x}}}.$$

Rewrite the coefficients:

$$\left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\underset{<}{P} = \left(\underset{<}{\Sigma}^{-1}+\underset{>}{\Sigma}^{-1}\right)^{-1}\underset{<}{\Sigma}^{-1} =$$
$$= \left(\underset{<}{\Sigma}\left(\underset{<}{\Sigma}^{-1}+\underset{>}{\Sigma}^{-1}\right)\right)^{-1} = \left(I+\underset{<}{\Sigma}\,\underset{>}{\Sigma}^{-1}\right)^{-1} = \underset{>}{\Sigma}\left(\underset{>}{\Sigma}+\underset{<}{\Sigma}\right)^{-1}$$

and similarly

$$\left(\underset{<}{P}+\underset{>}{P}\right)^{-1}\underset{>}{P}=\underset{<}{\Sigma}\left(\underset{>}{\Sigma}+\underset{<}{\Sigma}\right)^{-1}.$$

So

$$\widehat{\mathbf{x}}=\underset{>}{\Sigma}\left(\underset{>}{\Sigma}+\underset{<}{\Sigma}\right)^{-1}\underset{<}{\widehat{\mathbf{x}}}+\underset{<}{\Sigma}\left(\underset{>}{\Sigma}+\underset{<}{\Sigma}\right)^{-1}\underset{>}{\widehat{\mathbf{x}}}.$$

Applied to our state vector of one element:

$$\widehat{w}(t)=\frac{\underset{>}{\Sigma}(t)}{\underset{<}{\Sigma}(t)+\underset{>}{\Sigma}(t)}\underset{<}{\widehat{w}}(t)+\frac{\underset{<}{\Sigma}(t)}{\underset{<}{\Sigma}(t)+\underset{>}{\Sigma}(t)}\underset{>}{\widehat{w}}(t).$$

Substitute variances 4.16 and 4.17:

$$\begin{aligned}\widehat{w}(t)=&\frac{Q_n\cdot(t_2-t)}{Q_n\cdot(t-t_1)+Q_n\cdot(t_2-t)}\underset{<}{\widehat{w}}(t)+\\&+\frac{Q_n\cdot(t-t_1)}{Q_n\cdot(t-t_1)+Q_n\cdot(t_2-t)}\underset{>}{\widehat{w}}(t)=\\&=\frac{t_2-t}{t_2-t_1}\underset{<}{\widehat{w}}(t)+\frac{t-t_1}{t_2-t_1}\underset{>}{\widehat{w}}(t)=\frac{t_2-t}{t_2-t_1}\ell_1+\frac{t-t_1}{t_2-t_1}\ell_2,\end{aligned}$$

amounting to linear interpolation between the observation epochs. And the variance is

$$\begin{aligned}\Sigma(t)=&\left(\underset{<}{\Sigma}^{-1}(t)+\underset{>}{\Sigma}^{-1}(t)\right)^{-1}=\\&=Q_n\cdot\left(\frac{1}{t-t_1}+\frac{1}{t_2-t}\right)^{-1}=Q_n\frac{(t-t_1)(t_2-t)}{t_2-t_1},\end{aligned}$$

a quadratic expression, being zero at both ends and maxing out in the middle at $\frac{1}{4}(t_2-t_1)Q_n$. The standard deviation is the square root of this, up to $\sigma=\frac{1}{2}\sqrt{(t_2-t_1)Q_n}$ in the middle. See figure 4.2.

4.6.4 *Example 5: estimating a constant*

Let x be an unknown constant to be estimated. It has been observed at epoch 1, observation value 7, mean error ± 2, and at epoch 2, observation value 5, mean error ± 1.

Question Formulate the observation equations of an ordinary adjustment problem and the variance matrix of the observation vector. Compute $\widehat{x}$.

Answer

$$\underline{\boldsymbol{\ell}}+\underline{\mathbf{v}}=A\widehat{x},$$

in which the observation vector, its variance matrix, and the design matrix are

$$\underline{\boldsymbol{\ell}}=\begin{bmatrix}\underline{\ell}_1\\\underline{\ell}_2\end{bmatrix}=\begin{bmatrix}7\\5\end{bmatrix},\ S=\begin{bmatrix}R_1&0\\0&R_2\end{bmatrix}=\begin{bmatrix}4&0\\0&1\end{bmatrix},\ A=\begin{bmatrix}1\\1\end{bmatrix}.$$

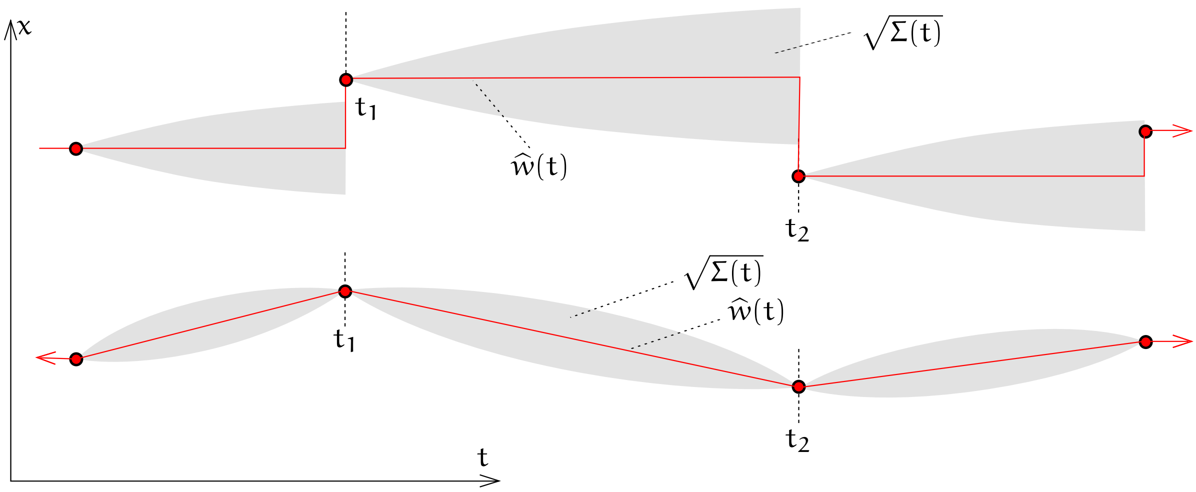


FIGURE 4.2. Random walk, solution and uncertainty. Above: solution forwards in time, below: combined solution in both directions. The line segments in the lower graph are called "Brownian bridges" (Wikipedia, Brownian bridge).

Solution:

$$\widehat{x} = \left(A^T S^{-1} A\right)^{-1} A^T S^{-1} \underline{\boldsymbol{\ell}} =$$
$$= \tfrac{4}{5} \cdot \begin{bmatrix} \tfrac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \end{bmatrix} = \frac{27}{5} = 5.4. \qquad (4.18)$$

Variance matrix of the solution:

$$\Sigma = \left(A^T S^{-1} A\right)^{-1} = \tfrac{4}{5} = 0.8. \qquad (4.19)$$

Question Write the dynamic equations for the Kalman filter. Remember that x *is a constant*.

Answer The general dynamic equation in the discrete case is

$$\underline{x}(t_{k+1}) = \Phi_k^{k+1} \underline{x}(t_k) + \underline{w}_k^{k+1}$$

in which $\Phi_k^{k+1} = \begin{bmatrix} 1 \end{bmatrix}$, a size 1×1 unit matrix, and $\underline{w}_k^{k+1} = 0$: deterministic motion, no dynamic noise. So

$$x(t_{k+1}) = x(t_k).$$

Alternatively, we write the differential equation:

$$\frac{d\underline{x}}{dt} = F\underline{x} + \underline{n}.$$

In this case, $F = 0$ and there is no dynamic noise, $\underline{n} = 0$:

$$\frac{dx}{dt} = 0.$$

Question Write the update equations for the Kalman filter — superscript "$-$" means *a priori*, "$+$" *a posteriori*:

$$x_k^+ = x_k^- - K_k \left(H_k x_k^- - \underline{\ell}_k \right), \qquad \Sigma_k^+ = \left(I - K_k H_k \right) \Sigma_k^-,$$

vahvistus-matriisi

in which the gain matrix is

$$K_k = \Sigma_k^- H_k^T \left(R_k + H_k^T \Sigma_k^- H_k \right)^{-1}.$$

What do the matrices H and K look like in this case?

Answer Because in this case the observation $\underline{\ell}_k = x_k + \underline{m}$ — we observe directly the state — we have $H_k = \left[\begin{array}{c} 1 \end{array} \right]$, a size 1×1 matrix, the only element of which is 1.

$$K_k = \frac{\Sigma_k^-}{R_k + \Sigma_k^-}.$$

If the original Σ_k^- is large, then $K \sim 1$.

$$\begin{aligned} x_k^+ &= x_k^- - \frac{\Sigma_k^-}{R_k + \Sigma_k^-} \left(x_k^- - \underline{\ell}_k \right) = \\ &= \overbrace{\frac{\Sigma_k^-}{R_k + \Sigma_k^-}}^{w_k} \underline{\ell}_k + \overbrace{\frac{R_k}{R_k + \Sigma_k^-}}^{W_k} x_k^- = \frac{\Sigma_k^- \underline{\ell}_k + R_k x_k^-}{R_k + \Sigma_k^-}, \\ \Sigma_k^+ &= \left(1 - K_k \right) \Sigma_k^- = \frac{R_k}{R_k + \Sigma_k^-} \Sigma_k^-. \end{aligned}$$

- The *a posteriori* state x_k^+ is the weighted average of the *a priori* state x_k^- and the observation $\underline{\ell}_k$ with weights W_k and w_k, $w_k + W_k = 1$.
- the poorer the *a priori* state variance Σ_k^- is compared to the observation variance R_k, the more the updated state variance Σ_k^+ will be an improvement.

Question Calculate manually through both observation events and give the *a posteriori* state estimate x_2^+ and its variance matrix Σ_2^+. The initial value of the state x_1^- is 0, and its initial variance matrix — meaning its variance — is set to "numerically infinite":

$$\Sigma_1^- = 100.$$

Answer First step:

$$K_1 = \frac{\Sigma_1^-}{R_1 + \Sigma_1^-} = \frac{100}{4 + 100} = \frac{100}{104}.$$

So

$$x_1^+ = x_1^- - K_1\left(x_1^- - \ell_1\right) = 0 - \frac{100}{104}\left(0 - 7\right) = 6.73 = x_2^-,$$
$$\Sigma_1^+ = \left(I - K_1\right)\Sigma_1^- = \left(1 - \frac{100}{104}\right)100 = \frac{400}{104} = 3.85 = \Sigma_2^-.$$

Second step:

$$K_2 = \frac{\Sigma_2^-}{R_2 + \Sigma_2^-} = \frac{3.85}{1 + 3.85} = 0.79.$$
$$x_2^+ = x_2^- - K_2\left(x_2^- - \ell_2\right) = 6.73 - 0.79\left(6.73 - 5\right) =$$
$$= 6.73 - 0.79 \cdot 1.73 = 5.36.$$
$$\Sigma_2^+ = \left(I - K_2\right)\Sigma_2^- = \left(1 - 0.79\right) \cdot 3.85 = 0.81.$$

These are close to the adjustment results 4.18 and 4.19.

Self-test questions

1. In the above example 4.1, provide the physical units for $\underline{x}$, $\underline{v}$, $\underline{n}$, Q, Σ, Φ_0^5, $\underline{\ell}'$, H, R, S, and K.
2. How many elements does the state vector contain when describing the motion in space of a point object?
3. And how many elements does it contain if the object is an extended, rigid object, like an aircraft?
4. What does the dynamic noise describe?
5. What does the observation noise describe?

   What is the dimension of the constant k in the above parachute-jumper problem 4.3?

   (a) time^{-1}

   (b) $\dfrac{\text{mass}}{\text{length}^3\ \text{time}}$

   (c) $\dfrac{\text{length}^3}{\text{mass time}}$

   (d) $\dfrac{\text{mass}}{\text{time}}$.

Exercise 4–1: A simple Kalman-filter example

1. Consider the following dynamic model for a one-element state vector x:

$$\frac{d}{dt}\underline{x}(t) = 0 + \underline{n}(t),$$

in which $\underline{n}(t)$ is white noise with an autocovariance of $Q_n = 1\,{}^{m^2}\!/\!{}_s$. The initial state estimate is $\widehat{x}(0) = 0$ m and the *a priori* state variance is $\Sigma(0) = 100\,\mathrm{m}^2$. Compute the state estimate for epoch $t = 3$ s, in other words, $x^-(3)$ and $\Sigma^-(3)$.

2. At epoch $t = 3$ s we observe $\underline{x}$: the observation equation is

$$\underline{\ell} = \underline{x}(3) + \underline{m},$$

in which $\underline{m}$ is the observation random error or noise, of which
4 variance R is given.[4] The observation value is $\ell = 10$ m. Compute

(a) the Kalman gain matrix K

(b) the *a posteriori* state estimate $x^+(3)$ and its variance $\Sigma^+(3)$.

Exercise 4–2: A somewhat more complicated Kalman-filter example

Assume now that the dynamic model is

$$\frac{d}{dt}\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{n} \end{bmatrix},$$

with the autocovariance of the white-noise process $\underline{n}$ being $50\,{}^{m^2}\!/\!{}_{s^3}$, meaning that

$$Q_n = \begin{bmatrix} 0 & 0 \\ 0 & 50\,{}^{m^2}\!/\!{}_{s^3} \end{bmatrix}.$$

The initial state estimate $\widehat{x}(0) = \widehat{v}(0) = 0$ and its variance matrix

$$\Sigma(0) = \begin{bmatrix} 100\,\mathrm{m}^2 & 0 \\ 0 & 100\,{}^{m^2}\!/\!{}_{s^2} \end{bmatrix}$$

are given.

1. Compute the state transition matrix Φ_0^3.
theta $\vartheta\theta\Theta$
2. Compute the state-noise variance θ_0^3.
3. Compute the *a priori* state estimate

$$x^-(3), v^-(3), \Sigma^-(3)$$

at epoch $t = 3$ s. Note that now, $x^-(3)$ and $v^-(3)$ are correlated.

[4]Take for the variance the day number of the month of your birthday!

4. The observation equation is again

$$\underline{\ell} = \underline{x}(3) + \underline{m}.$$

If $\underline{\ell} = 10\,\mathrm{m}$, compute

(a) the Kalman gain matrix K

(b) the *a posteriori* state variance $\Sigma^+(3)$.

Out of interest, compute the determinants of the variance matrices $\Sigma(0)$, $\Sigma^-(3)$ and $\Sigma^+(3)$. The square root of this determinant multiplied by π equals the surface area of the error ellipse in (x, v) space and is a sensible overall measure for the imprecision of the state vector.

Assume R to be the same as in the previous exercise or keep it as a symbol.

Hint: analyse carefully example 1 in section 4.1.

Exercise 4–3: The parachute-jumper revisited

In example 4.3, assume that the air drag is proportional to the *square* of the velocity of falling. Re-derive and re-linearise the dynamic model.

Inertial navigation

5

5.1 Principle

Inertial navigation is based on Isaac Newton's[1] first law of mechanics, also known as the law of continuity of motion, or the *law of inertia*:

> *"Every object persists in its state of rest or uniform motion in a straight line unless it is compelled to change that state by forces impressed on it."*

This suggests that it will be possible to reconstruct the motion of a vehicle from an initial state — location and velocity — by just measuring continuously all the forces acting upon the vehicle, without any reference to external objects or signals. This is what inertial navigation does.

A good text on inertial navigation and its use in geodesy is Jekeli (2001).

We shall first describe how inertial navigation is done in an inertial or free-falling frame in free space, an idealised situation far away from local sources of gravitation. Inertial navigation within the Earth's gravity field is discussed later.

In inertial navigation, the following quantities are measured continuously:

1. the three-dimensional *acceleration* of the vehicle,

$$\mathbf{a}(t) \stackrel{\text{def}}{=} \frac{d^2}{dt^2}\mathbf{x}(t) = \frac{d^2x(t)}{dt^2}\mathbf{i} + \frac{d^2y(t)}{dt^2}\mathbf{j} + \frac{d^2z(t)}{dt^2}\mathbf{k},$$

[1] Sir Isaac Newton PRS (1643–1727) was an English universal genius who derived the mathematical foundation of the physics of motion as applied to astronomy and geophysics.

or, as an abstract vector of components or co-ordinates,

$$\mathbf{a}_\beta(t) = \frac{d^2}{dt^2}\mathbf{x}_\beta(t) = \begin{bmatrix} \frac{d^2x(t)}{dt^2} & \frac{d^2y(t)}{dt^2} & \frac{d^2z(t)}{dt^2} \end{bmatrix}^T.$$

Here,

$$\mathbf{x}(t) \stackrel{\text{def}}{=} x(t)\,\mathbf{i} + y(t)\,\mathbf{j} + z(t)\,\mathbf{k}$$

is the three-dimensional location of the object, and

$$\mathbf{x}_\beta(t) \stackrel{\text{def}}{=} \begin{bmatrix} x(t) & y(t) & z(t) \end{bmatrix}^T$$

ortonormaali kanta

beta βB

is the abstract vector of its co-ordinates on the orthonormal basis $\beta \stackrel{\text{def}}{=} \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$.

2. The *attitude* of the vehicle:

$$\begin{aligned} R &= R_3(\alpha_3)\,R_2(\alpha_2)\,R_1(\alpha_1) = \\ &= \begin{bmatrix} c_3 & s_3 & 0 \\ -s_3 & c_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_2 & 0 & -s_2 \\ 0 & 1 & 0 \\ s_2 & 0 & c_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_1 & s_1 \\ 0 & -s_1 & c_1 \end{bmatrix} = \\ &= \begin{bmatrix} c_2c_3 & c_1s_3 + s_1s_2c_3 & s_1s_3 - c_1s_2c_3 \\ -c_2s_3 & c_1c_3 - s_1s_2s_3 & s_1c_3 + c_1s_2s_3 \\ s_2 & -s_1c_2 & c_1c_2 \end{bmatrix}, \end{aligned}$$

in which $c_i \stackrel{\text{def}}{=} \cos\alpha_i$, $s_i \stackrel{\text{def}}{=} \sin\alpha_i$, $i = 1, 2, 3$.

alpha αA

The attitude is described by three unknowns, $\alpha_i(t)$, $i = 1, 2, 3$, *Euler angles*, that are functions of time and change with the movements of the vehicle.

Before the journey begins, matrix $R(t_0)$, or equivalently, the attitude angles $\alpha_i(t_0)$, $i = 1, 2, 3$, have to be determined with sufficient accuracy. During the journey, the attitude changes or angular rates $\frac{d}{dt}\alpha_i$ are measured with the help of three gyroscopes, and are integrated in time in order to obtain the instantaneous attitude $\alpha_i(t)$, and thus matrix $R(t)$.

One measures continuously a total of six parameters: three angular rates (velocities) and three linear accelerations.

Now, the *acceleration* measured with three accelerometers is, expressed on vehicle axes $\beta' \stackrel{\text{def}}{=} \{\mathbf{i}', \mathbf{j}', \mathbf{k}'\}$:

$$\tilde{\mathbf{a}}_{\beta'}(t) = R(t)\,\mathbf{a}_\beta(t),$$

in which the tilde denotes a measured quantity.

Now, the data processing unit of the inertial device *integrates* the measured accelerations $\widetilde{\boldsymbol{a}}$ after the transformation

$$\boldsymbol{a}_{\beta}(t) = R^{-1}(t)\,\widetilde{\boldsymbol{a}}_{\beta'}(t) = R^{T}(t)\,\widetilde{\boldsymbol{a}}_{\beta'}(t)$$

in three dimensions, and twice in a row. The first integration yields the velocity vector of the vehicle, the second its location.

As follows, in inertial co-ordinates β:

$$\begin{aligned}\boldsymbol{v}_{\beta}(t) &= \boldsymbol{v}_{\beta}(t_0) + \int_{t_0}^{t} R^{T}(\tau)\,\widetilde{\boldsymbol{a}}_{\beta'}(\tau)\,d\tau,\\ \boldsymbol{x}_{\beta}(t) &= \boldsymbol{x}_{\beta}(t_0) + \int_{t_0}^{t} \boldsymbol{v}_{\beta}(\tau)\,d\tau,\end{aligned} \tag{5.1}$$

in which $\boldsymbol{x}_{\beta}(t_0)$ and $\boldsymbol{v}_{\beta}(t_0)$ are integration constants. They may represent, for example, the location of the launch site and the knowledge that the spacecraft is standing still on the launch pad.

As seen in equations 5.1, the accuracy of position $\boldsymbol{x}(t)$ will get progressively poorer with time, because both the measurement of acceleration $\widetilde{\boldsymbol{a}}$ and the determination of the rotation matrix R are imprecise: the error in these measurements will accumulate through integration. This accumulation happens even twice on top of each other, because there are two successive integrations.

A trick often used to preserve the precision of inertial navigation on the Earth's surface is to *halt* regularly and do a "zero-velocity update". Then, $\boldsymbol{v}(t_1) = 0$ for some moment in time $t_1 > t_0$, and the first integral, the velocity integral, will start again from a known starting value.

nollanopeus-päivitys

5.2 Parts of an inertial device

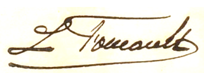
An inertial device or IMU, inertial measurement unit, contains the following measuring components:

1. gyroscopes
2. accelerometers.

5.2.1 *The gyroscope*

A gyroscope is a rapidly spinning flywheel, the inertia of which makes it difficult to change the direction of its axis of rotation or spin axis.

The name "gyroscope" was invented by Léon Foucault,[2] who success-



2

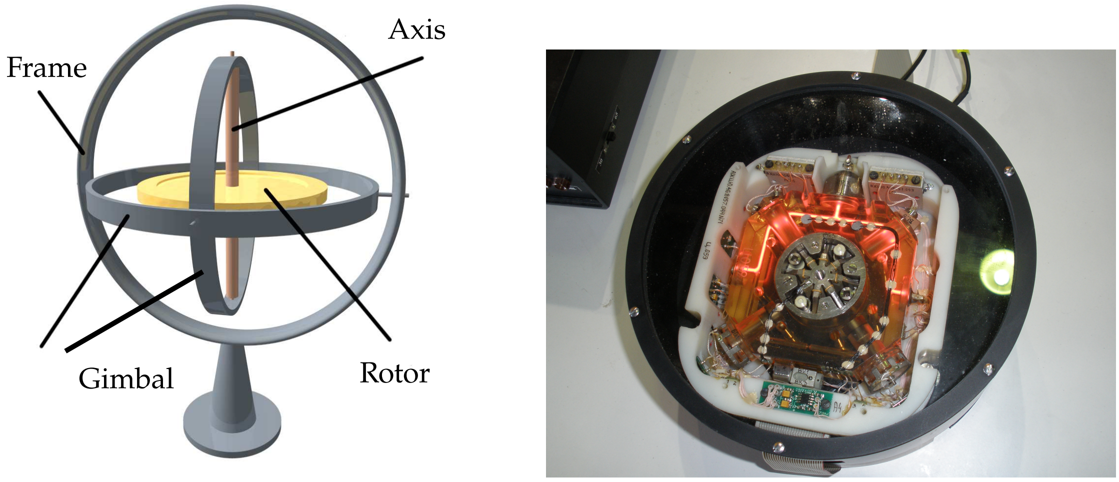


FIGURE 5.1. On the left, a gyroscope. Wikimedia Commons, Gyroscope. Click for animation. On the right, a ring-laser gyroscope used in aviation. Nockson (2011).

fully built one in order to demonstrate the rotation of the Earth, after having done so successfully with his eponymous pendulum. The name comes from the Greek and refers literally to seeing the rotation of the Earth (Sommeria, 2017).

The motion of a rotating body is described in an inertial frame by the
3 following equation:[3]

$$\mathbf{N} = \frac{d\mathbf{L}}{dt} = J\frac{d\boldsymbol{\omega}}{dt} + \left(\frac{d}{dt}J\right)\boldsymbol{\omega}, \tag{5.2}$$

in which

vääntö-momentti	$\mathbf{N}$	torque
pyörähdys-momentti	$\mathbf{L} = J\boldsymbol{\omega}$	angular momentum
omega $\omega\Omega$	$\boldsymbol{\omega}$	angular velocity
	J	*inertial tensor*, in terms of components on the basis β of the (x, y, z)

[2]Jean Bernard Léon Foucault FRS (1819–1868) was a French physicist best known for his empirical studies on the rotation of the Earth.

[3]This equation is similar to Newton's second law of motion:

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\boldsymbol{v}}{dt},$$

in which $\mathbf{F}$ is the (linear) force and $\boldsymbol{v}$ the (linear) velocity. $\mathbf{p} = m\boldsymbol{v}$ is the *momentum* or amount of (linear) motion. m, the mass, corresponds to the inertial tensor J, but is a scalar, and a constant in Newtonian mechanics.

co-ordinate frame:

$$J_\beta = \begin{bmatrix} J_{xx} & J_{xy} & J_{xz} \\ J_{xy} & J_{yy} & J_{yz} \\ J_{xz} & J_{yz} & J_{zz} \end{bmatrix},$$

a matrix of 3×3 elements.

The component matrix J_β of the inertial tensor J of an object is symmetric and positive definite. Its elements are the following integrals:

$$\begin{aligned} J_{xx} &= \iiint \rho(x,y,z)\left(y^2+z^2\right)\,dx\,dy\,dz, \\ J_{yy} &= \iiint \rho(x,y,z)\left(x^2+z^2\right)\,dx\,dy\,dz, \\ J_{zz} &= \iiint \rho(x,y,z)\left(x^2+y^2\right)\,dx\,dy\,dz, \\ J_{xy} &= -\iiint \rho(x,y,z)\,xy\,dx\,dy\,dz, \\ J_{xz} &= -\iiint \rho(x,y,z)\,xz\,dx\,dy\,dz, \\ J_{yz} &= -\iiint \rho(x,y,z)\,yz\,dx\,dy\,dz, \end{aligned}$$

in which ρ is the matter density. So rho ρR

$$J_\beta = \iiint \rho(x,y,z) \begin{bmatrix} y^2+z^2 & -xy & -xz \\ -xy & x^2+z^2 & -yz \\ -xz & -yz & x^2+y^2 \end{bmatrix} dx\,dy\,dz.$$

The result obviously depends on the choice of co-ordinate frame (x,y,z). The *origin* has a large influence: by choosing it to lie far outside the object, one can make the elements of J_β arbitrarily large! Therefore, when talking about the inertial tensor as a property of an object, we always choose its centre of mass as the origin:

$$\mathbf{x}_{\text{com}} = \frac{1}{M}\iiint \rho(\mathbf{x})\,\mathbf{x}\,d\mathcal{V},$$

or

$$\begin{aligned} x_{\text{com}} &= \frac{1}{M}\iiint \rho(x,y,z)\,x\,dx\,dy\,dz, \\ y_{\text{com}} &= \frac{1}{M}\iiint \rho(x,y,z)\,y\,dx\,dy\,dz, \\ z_{\text{com}} &= \frac{1}{M}\iiint \rho(x,y,z)\,z\,dx\,dy\,dz, \end{aligned}$$

with

$$M = \iiint \rho(x,y,z)\,dx\,dy\,dz$$

the total mass of the object. After this, in computing J we use

$$\widetilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_{\mathrm{com}}.$$

As for the axes orientation, it is well-known that a symmetric tensor can always be brought *on main axes* by a rotation of the co-ordinate frame. In this case the inertial tensor assumes the diagonal form

$$J_{\beta'} = \begin{bmatrix} J_x & 0 & 0 \\ 0 & J_y & 0 \\ 0 & 0 & J_z \end{bmatrix}. \tag{5.3}$$

The elements J_x, J_y, and J_z are called the *principal moments of inertia*. They are actually the eigenvalues of the J tensor.

The faster the gyroscope spins, the larger the vectorial rotation rate $\boldsymbol{\omega}$ and thus the angular momentum $\mathbf{L}$, and the more torque $\mathbf{N}$ is needed to turn the gyroscope's axis of rotation.

Equation 5.2 applies in an *inertial* frame. In a general non-inertial frame β, the equation for the time derivative of a vector yields

$$\begin{aligned} \mathbf{N}_\beta &= \frac{d}{dt}\mathbf{L}_\beta + \left\langle \boldsymbol{\omega}_\beta^{[\beta]} \times \mathbf{L}_\beta \right\rangle = \\ &= J_\beta \frac{d}{dt}\boldsymbol{\omega}_\beta + \left(\frac{d}{dt}J_\beta\right)\boldsymbol{\omega}_\beta + \left\langle \boldsymbol{\omega}_\beta^{[\beta]} \times J_\beta \boldsymbol{\omega}_\beta \right\rangle. \end{aligned} \tag{5.4}$$

The symbol $\times$ designates the exterior or vector product. The vector $\boldsymbol{\omega}^{[\beta]}$ is the angular velocity of the non-inertial frame as distinct from the angular velocity $\boldsymbol{\omega}$ of the body itself.

In a non-inertial frame in which the inertial tensor J_β is on main axes,
4 and furthermore is *constant* and does not depend on time,[4] we have, based on equation 5.4:

$$\begin{aligned} N_x &= J_x \frac{d}{dt}\omega_x + \omega_y^{[\beta]} J_z \omega_z - \omega_z^{[\beta]} J_y \omega_y, \\ N_y &= J_y \frac{d}{dt}\omega_y + \omega_z^{[\beta]} J_x \omega_x - \omega_x^{[\beta]} J_z \omega_z, \\ N_z &= J_z \frac{d}{dt}\omega_z + \omega_x^{[\beta]} J_y \omega_y - \omega_y^{[\beta]} J_x \omega_x. \end{aligned} \tag{5.5}$$

Vector $\boldsymbol{\omega}^{[\beta]}$ is the angular velocity of the frame. Note that for J to be constant, the frame needs to be connected only to the *figure* of the body, not the body itself. For example for a rotationally symmetric gyroscope

[4]The assumption that such a frame exists is natural for a flywheel, but not, for example, for the Earth, which changes shape continuously.

rotor, the x and y axes need not spin with the rotor as long as the z axis is aligned with the spin axis. Then, $\omega_z^{[\beta]} = 0$ though $\omega_z \neq 0$.

If $\omega_z^{[\beta]} = \omega_z$, then $\boldsymbol{\omega}^{[\beta]} = \boldsymbol{\omega}$, and the classical form of the Euler equations in the co-rotating body frame is obtained:

$$\begin{aligned} N_x &= J_x \frac{d}{dt}\omega_x + (J_z - J_y)\,\omega_y\omega_z, \\ N_y &= J_y \frac{d}{dt}\omega_y + (J_x - J_z)\,\omega_z\omega_x, \\ N_z &= J_z \frac{d}{dt}\omega_z + (J_y - J_x)\,\omega_x\omega_y. \end{aligned} \tag{5.6}$$

These equations apply for a freely spinning body, such as the Earth in space.[5] 5

Building a good gyroscope is a difficult engineering art. The measurement precisions of these devices are impressive: the stability of the rotation axis of a top-quality mechanical gyroscope can be of the order of $0.^\circ 0001$ per hour.

A gyroscope consists of a rotor and an axis that is mounted in bearings on both ends into a frame, also called a table, surrounding the rotor. The frame may consist of several rings and axes or *gimbals*, a so-called *Cardan suspension*.

Let the spin axis be the z axis. If the direction of the spin axis is constant in an external, inertial frame,

$$\omega_x^{[\beta]} = \omega_y^{[\beta]} = 0,$$

---

[5]If, in equations 5.6, we zero the torque $\mathbf{N} = 0$, the first two equations become, with $\Delta J = J_z - J_x = J_z - J_y$ for a rotationally symmetric body,

$$J_x \frac{d}{dt}\omega_x = -\Delta J\,\omega_y\omega_z, \quad J_y \frac{d}{dt}\omega_y = \Delta J\,\omega_z\omega_x,$$

and with $C = \Delta J\,\omega_z / J_x = \Delta J\,\omega_z / J_y \stackrel{\text{def}}{=} \Delta J\,\omega_z / J_0$:

$$\frac{d}{dt}\omega_x = -C\omega_y, \quad \frac{d}{dt}\omega_y = C\omega_x,$$

yielding by cross-substitution the second-order equations

$$\frac{d^2}{dt^2}\omega_x = -C^2\omega_x, \quad \frac{d^2}{dt^2}\omega_y = -C^2\omega_y,$$

both having periodic solutions with period

$$T = \frac{2\pi}{C} = 2\pi\frac{J_0}{\Delta J\,\omega_z} = \frac{J_0}{\Delta J}\frac{2\pi}{\omega_z} = \frac{J_0}{\Delta J}\text{ sidereal days.}$$

Here, it was assumed that $J_x = J_y \stackrel{\text{def}}{=} J_0$. Thus, we have found the Euler free nutation of the Earth, a circular motion which is one component of polar motion.

and we may choose $\omega_z^{[\beta]} = 0$. Also $\omega_x = \omega_y = 0$, so

$$\frac{d}{dt}\omega_x = \frac{d}{dt}\omega_y = 0.$$

If in addition the rotor spin velocity is constant, meaning

$$\frac{d}{dt}\omega_z = 0,$$

it follows from equations 5.5, that

$$N_x = \frac{d}{dt}L_x = 0, \quad N_y = \frac{d}{dt}L_y = 0, \quad N_z = \frac{d}{dt}L_z = 0.$$

So in this case, no torques are acting on the gyroscope.

And as in this co-ordinate frame, the inertial tensor J is on main axes,

6 equation 5.3, and[6]

$$\boldsymbol{\omega} = \omega_z \mathbf{k} \implies \omega \stackrel{\text{def}}{=} \|\boldsymbol{\omega}\| = \omega_z,$$

it follows that

$$\mathbf{L} = J\boldsymbol{\omega} = J_z\omega_z\mathbf{k} = J_z\omega\mathbf{k} \implies L \stackrel{\text{def}}{=} \|\mathbf{L}\| = L_z = J_z\omega_z = J_z\omega. \quad (5.7)$$

So both vectors, the angular velocity $\boldsymbol{\omega}$, magnitude ω, and the angular momentum $\mathbf{L}$, magnitude L, are always aligned with the gyroscope's physical spin axis and each other. This remains true to good approximation in an inertial frame, if the gyroscope's spin axis turns only

7 slowly.[7] It is also exactly true in the gyroscope body frame, because $\omega_x = \omega_y = 0$ as the gyroscope is mechanically constrained to spin around its axis of symmetry only.

We study the behaviour of the gyroscope using the principle of conservation of angular momentum $\mathbf{L}$, which is always aligned with the spin axis. The only way to get the spin axis of the gyroscope to turn is to apply a torque to it. Applying a torque N_x around the x axis will lead to a change in angular momentum around that axis of

$$\frac{d}{dt}L_x = N_x.$$

Because the total angular momentum is $L = J_z\omega$ and it is always pointing along the spin axis of the gyroscope, it follows that the change

theta $\vartheta\theta\Theta$ in direction θ of the spin axis must be[8]

8

---

[6]It is assumed that the vector $\boldsymbol{\omega}$ points along the positive z axis.

[7]For illustration, a typical gyroscope spin rate is 20 000 rotations per minute or over 300 Hz. This is a thousand times faster than typical vehicle rotations, and over a million times the angular rate of a low Earth orbit satellite.

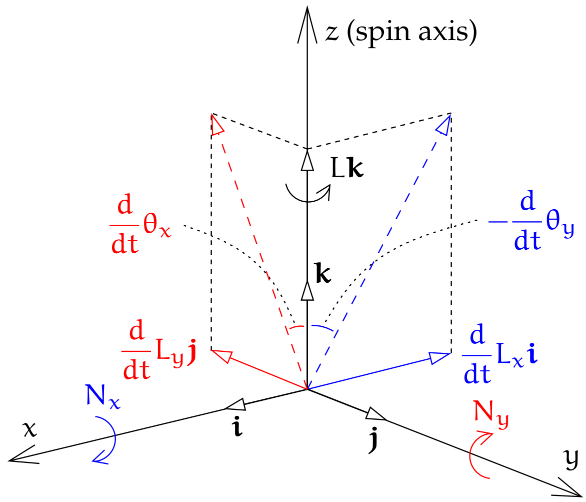


FIGURE 5.2. How torquing causes precession of the gyroscope spin axis. Note the corkscrew rule: $N_x, N_y, \frac{d}{dt}L_x, \frac{d}{dt}L_y$ and $\frac{d}{dt}\theta_y$ are negative, L and $\frac{d}{dt}\theta_x$ are positive.

$$\omega_y^{[\beta]} = \frac{d}{dt}\theta_y = \frac{\frac{d}{dt}L_x}{L} = \frac{N_x}{J_z\omega}. \tag{5.8}$$

Similarly

$$\omega_x^{[\beta]} = \frac{d}{dt}\theta_x = -\frac{\frac{d}{dt}L_y}{L} = -\frac{N_y}{J_z\omega}.$$

So, a torque N_x around the x axis causes a turning, or *precession*, of the spin axis around the y axis, and a torque N_y around the y axis causes the spin axis to turn around the x axis. See figure 5.2.

Conventionally, three mutually orthogonal axes are defined: the input axis, around which torques are applied, the gyroscope spin axis, and

---

[8]This can also be obtained from the first equation 5.5:

$$N_x = J_x\frac{d}{dt}\omega_x + \omega_y^{[\beta]}J_z\omega_z - \omega_z^{[\beta]}J_y\omega_y = \omega_y^{[\beta]}L,$$

because in the body frame, due to the mechanical constraint,

$$\omega_x = \omega_y = 0$$

and thus also

$$J_z\omega_z = J_z\omega = L.$$

The same applies for the second equation and $\omega_x^{[\beta]}$:

$$N_y = J_y\frac{d}{dt}\omega_y + \omega_z^{[\beta]}J_x\omega_x - \omega_x^{[\beta]}J_z\omega_z = -\omega_x^{[\beta]}L.$$

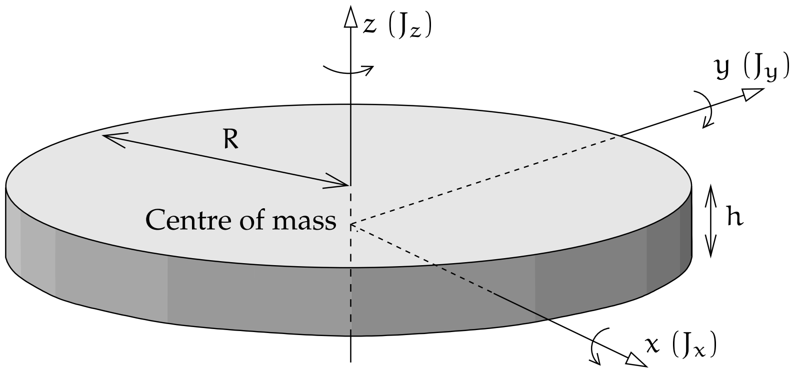


FIGURE 5.3. A gyroscope rotor and its principal moments of inertia.

the output axis, around which precessional motion takes place.

A cylinder of radius R, figure 5.3, has a moment of inertia about its z axis of

$$J_z = \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \iint_{\text{circular disc}} \rho \left(x^2 + y^2\right) dx\, dy\, dz =$$
$$= \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_0^{2\pi} \int_0^R r^2 \cdot r\, dr\, d\theta\, dz =$$
$$= \rho \overbrace{\int_{-\frac{1}{2}h}^{\frac{1}{2}h} dz}^{h} \overbrace{\int_0^{2\pi} d\theta}^{2\pi} \overbrace{\int_0^R r^2 \cdot r\, dr}^{\frac{1}{4}R^4} = \frac{1}{2}\left(\rho \cdot \pi R^2 \cdot h\right) R^2 = \frac{1}{2} M R^2,$$

in which $r^2 = x^2 + y^2$ and $M = \rho \cdot \pi R^2 \cdot h$ is the total mass. ρ is the matter density of the cylinder.

For a flat cylinder (h, and thus z, are small) we may also calculate, with the definition $Y(x) \stackrel{\text{def}}{=} \sqrt{R^2 - x^2}$:

$$J_x = \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_{-R}^{+R} \int_{-Y(x)}^{+Y(x)} \rho \left(y^2 + z^2\right) dy\, dx\, dz \approx$$
$$\approx \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_{-R}^{+R} \int_{-Y(x)}^{+Y(x)} y^2\, dy\, dx\, dz =$$
$$= \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_0^{2\pi} \int_0^R (r \sin\theta)^2\, r\, dr\, d\theta\, dz =$$
$$= \rho \overbrace{\int_{-\frac{1}{2}h}^{\frac{1}{2}h} dz}^{h} \overbrace{\int_0^{2\pi} \sin^2\theta\, d\theta}^{\pi} \overbrace{\int_0^R r^2 \cdot r\, dr}^{\frac{1}{4}R^4} = \frac{1}{4}\left(\rho \cdot \pi R^2 \cdot h\right) R^2 = \frac{1}{4} M R^2.$$

In both, we changed the integration co-ordinates (x, y) to polar co-ordinates (θ, r): $x = r\cos\theta$, $y = r\sin\theta$, $dx\, dy = r\, dr$.

Similarly, of course, $J_y = J_x = \frac{1}{4} M R^2 = \frac{1}{2} J_z$.

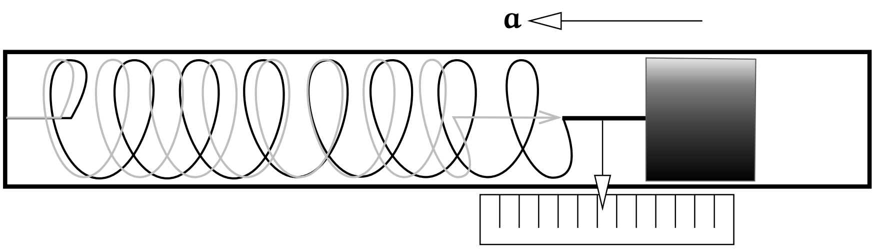


FIGURE 5.4. Principle of a spring accelerometer.

5.2.2 *The accelerometer*

A primitive accelerometer can be built by combining a spring, a test mass and a scale. The stretching of the spring is proportional to the test mass and to the acceleration and can be read from the scale.

Automatic readout is possible, for example capacitively or by using a piezoelectric sensor. In fact, a microelectronic acceleration sensor (a type of MEMS, microelectronic motion sensor) works in just this way.

The accelerometers are attached to the same frame onto which the gyroscopes are mounted. The measurement axes are made as parallel as possible.

Modern accelerometers may be very sensitive, for example 10 ppm $\approx$ 10 mGal.[9] Desirable traits besides sensitivity are *linearity* and good behaviour under circumstances of large variations of acceleration, or *vibration* (rocket launch!). 9

Sensors used for measuring ambient gravity to high precision are a special case. They are comparable to gravimeters: if they are based on the elasticity of matter, they similarly demand regular calibration. They change over time, so-called *drift*. käynti

An alternative type of accelerometer is the so-called *pendulous* type. Here, a mass is attached to the end of a beam. Acceleration makes the beam deflect, which is measured by a sensor. The signal goes to an actuator which restores the deflection to zero. It is thus a *nulling* sensor, toimilaite which in principle guarantees linear behaviour. Moreover, this type of accelerometer does not suffer from drift. However, the quality of measurement will depend on the quality with which a known restoring force can be generated, for example as an electric current through an electromagnetic actuator.

[9]The unit mGal is used for gravity or acceleration and equals $10^{-5}\,\mathrm{m}/\mathrm{s}^2$. Ambient gravity is $\approx 9.81\,\mathrm{m}/\mathrm{s}^2 = 981\,000\,\mathrm{mGal}$.

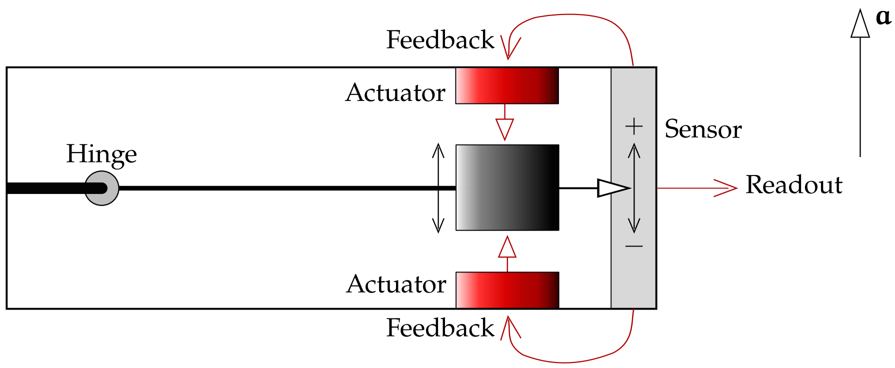


FIGURE 5.5. Pendulous accelerometer.

Pendulous accelerometers, but *gyroscope-based*, have been used in missiles for a long time, as they offer the highest precision. In these, the beam is also the axis of a spinning gyroscope, and the mass at the end of the beam exerts a torque that makes the gyroscope precess around the direction of acceleration, the z axis. This precessional motion is
10 measured and constitutes the output of the accelerometer.[10] The nulling
vääntötoimi-laite feedback from the y axis to a torquer on the z axis compensates the beam torque by imparting the correct amount of precession. This ensures that there is no torque around the z axis acting back on the gyroscope, and thus keeps the gyroscope spin axis perpendicular to the z axis. See figure 5.6.

Because of the strategic importance of inertial navigation — missiles — good accelerometers, like good gyroscopes, were long hard to obtain and expensive. Nowadays the situation is better. Modern accelerometers are often MEMS (microelectronic motion sensor) based and inexpensive. They are however not as precise as gyroscope-based ones.

5.3 Implementation

A popular introduction to inertial navigation and inertial measurement units is given in King (1998).

There are two, very different, general approaches for implementing an inertial measurement unit:

1. strapdown solution
2. stabilised-platform solution.

[10]For example the German V-2 used an integrating gyroscope-based accelerometer to turn off the propellant supply (“*Brennschluss*”) when the intended terminal velocity had been reached (Wikipedia, PIGA accelerometer).

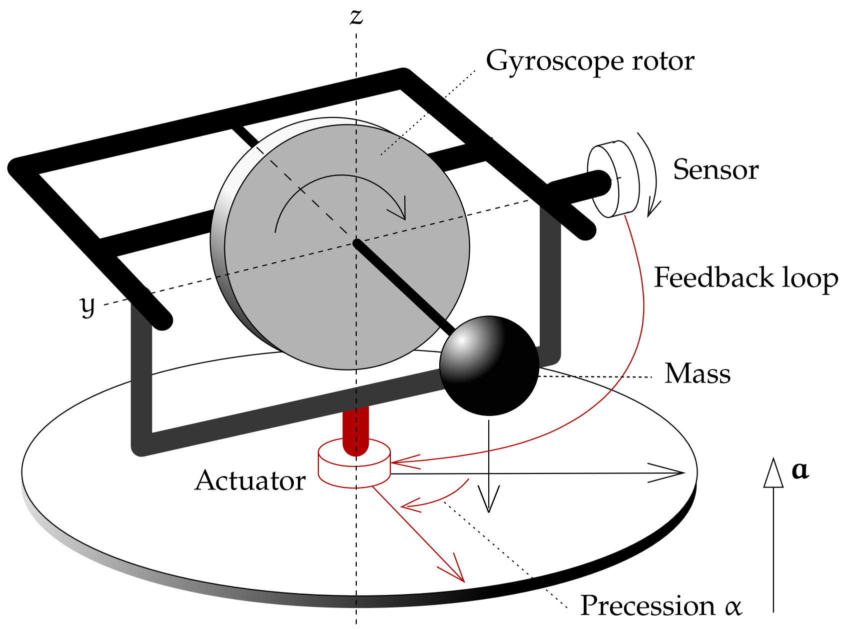


FIGURE 5.6. Gyroscopic pendulous accelerometer. The angle α is the total precession angle, which is proportional to the integral over time of the sensed acceleration.

5.3.1 Strapdown solution

In a strapdown solution, the gyroscope platform is rigidly connected to the body of the vehicle. When the attitude of the vehicle changes, the ends of the axis of the gyroscope push against its frame with a force that is accurately measured with force sensors. From the force $\mathbf{F}$, the torque $\mathbf{N}$ is obtained:

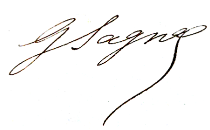
$$\mathbf{N} = \langle \boldsymbol{\ell} \times \mathbf{F} \rangle,$$

in which $\boldsymbol{\ell}$ is the length of the gyroscope's axis as a vector: "torque is arm times force". The symbol $\times$ designates again the exterior or vector product. The angular turning rates $\frac{d}{dt}\alpha_i$ of the vehicle are calculated from the measured torques using equation 5.2:

$$\frac{\mathbf{N}}{\omega} = J\left(\frac{1}{\omega}\frac{d\boldsymbol{\omega}}{dt}\right).$$

Here, the parenthesised expression represents the turning of the gyroscope's spin vector, which corresponds to the turning of the vehicle's body frame. Clearly, each gyroscope can only give the turning rates in the two directions perpendicular to its spin axis. When there are three gyroscopes, each is used to sense rotations around only one axis.

A different technological solution is offered by the so-called *ring-laser gyroscope*, which is based on the interference of light: the Sagnac[11]



11

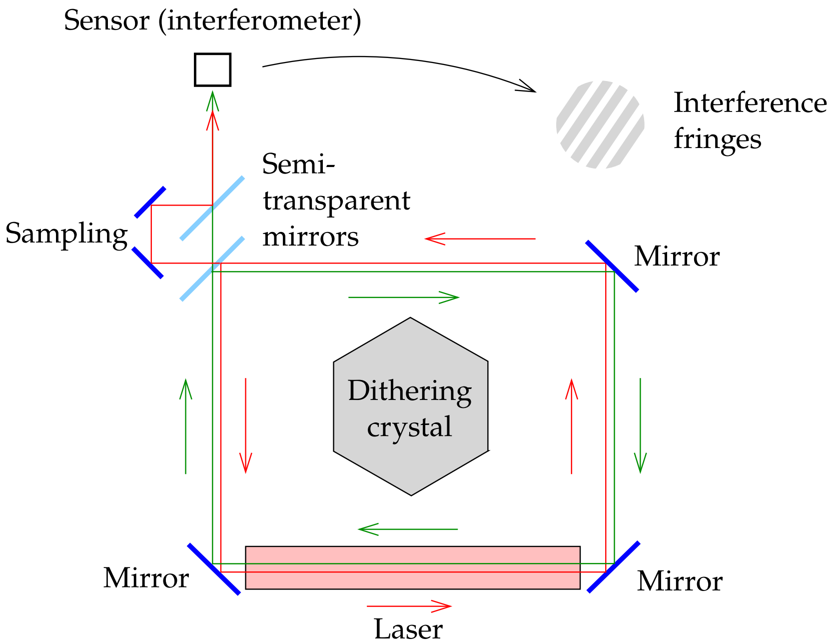


FIGURE 5.7. A ring-laser gyroscope based on the Sagnac interferometer.

phenomenon, 1913. In the device, monochromatic laser light travels in a ring in two opposite directions. Without rotation, the light forms a *standing wave*, the nodes of which do not move. However, even a small rotation of the ring will make the nodes move within the ring in the opposite direction, so they remain in the same place in a non-rotating frame.

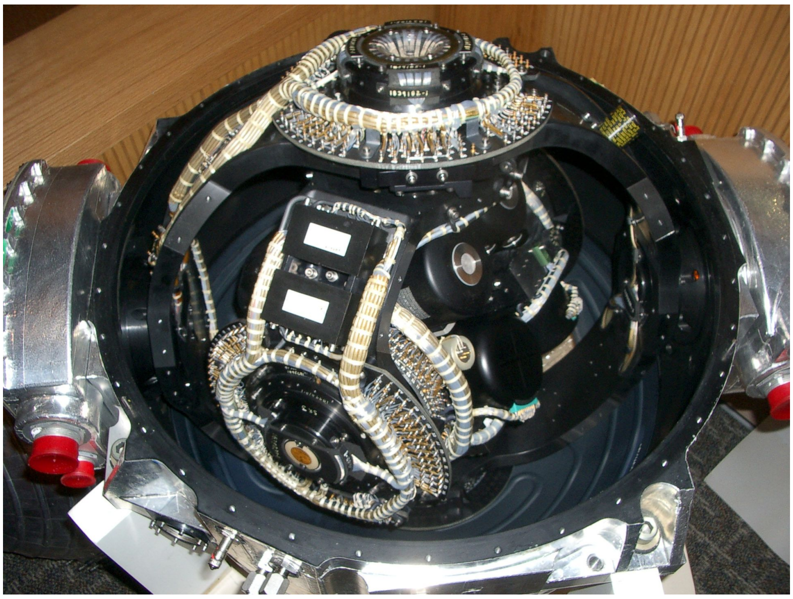
The simplest way to build a ring laser is to use fixed mirrors. Nowadays, a long optic fibre is often used that is wrapped around the circumference of a ring thousands of times. In this way, the effect is amplified thousands of times, gaining sensitivity. The achievable sensitivities are as high as 0.000 01 degrees per hour. MathPages, The Sagnac Effect.

5.3.2 Stabilised-platform solution

In the stabilised-platform solution, the whole gyroscope system is mounted inside a three-axis, freely turning Cardan ring suspension. Because of this, although the attitude of the vehicle changes, the gyroscopic frame retains its attitude in space.

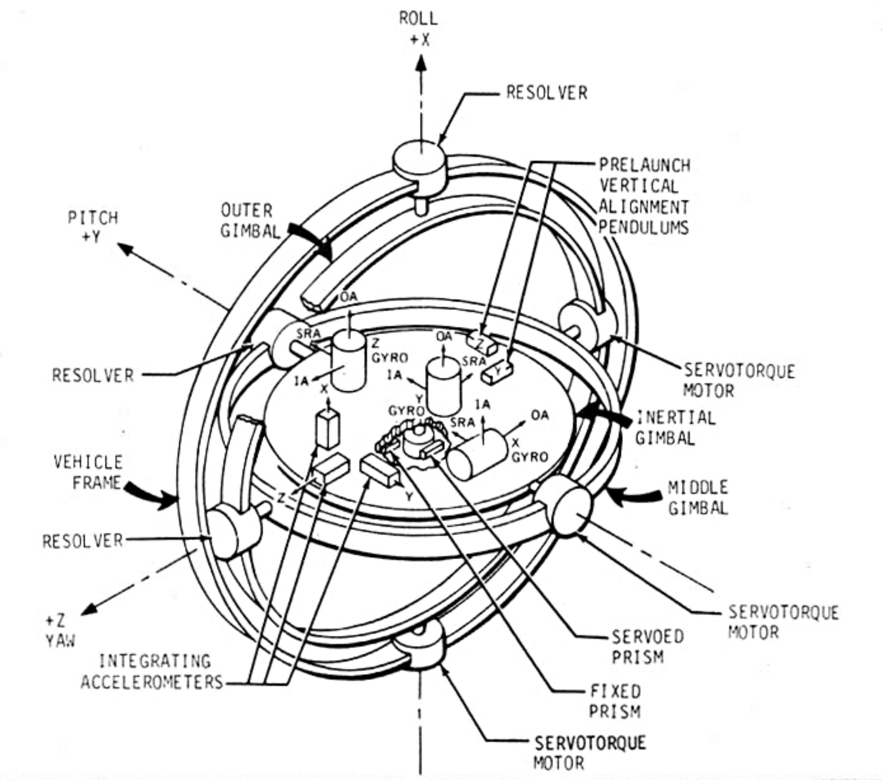
Figure 5.8 shows a three-axis device which has been used in deep-space missions, using three gyroscopes to preserve its attitude with

[11]Georges Sagnac (1869–1928) was a French physicist and student of physical optics.



(a)

Photograph, Durbin (2004)



(b)

Schematic, NASA. IA = input axis, OA = output axis, and SRA = spin reference axis

FIGURE 5.8. ST-124 inertial device. This stabilised-platform type device was used on the Saturn-V lunar rocket.

respect to an inertial frame. Each gyroscope is inside its own gyroscope assembly. The case of the assembly is drawn as a small cylinder.

Each gyroscope assembly is a self-contained unit designed to only respond to rotations around one axis, the input axis. It is also called a single-degree-of-freedom gyroscope, see Jekeli (2001), figure 3.1.

In this inertial device, the gyroscope assemblies are used as nulling sensors: any observed rotation is fed back to the corresponding "servo-torque motor" to cancel it, keeping the attitude of the inner platform, the "inertial gimbal", unchanged in inertial space. The output from the device as a whole is the angles from the three "resolvers", which are in fact the attitude angles α_i, $i = 1, 2, 3$ of the rocket's body in inertial space.

Note that the spin axis of each gyroscope is perpendicular to the axis of the gyroscope assembly's case. The latter axis is also the output axis (OA) of the gyroscope.

In *terrestrial* applications, instead of an inertial non-rotating frame, one uses a *local* frame connected to the solid Earth. The gyroscope axes are kept aligned with the axes of a topocentric frame:

- two directions x and y in the horizontal plane
- the up direction z.

The two horizontal gyroscope axes will not necessarily be aligned with the north and east directions, but pointing in two other, mutually perpendicular directions in the horizontal plane. These directions differ from the north and east directions by an azimuth angle α, which changes over time. Conceptually however, it may make sense to think
12 in terms of virtual "north" and "east" gyroscopes.[12]

Also, in terrestrial applications including aviation, only two horizontal accelerometers will be actually used. The height is taken care of by other means, as inertial navigation over longer time spans is not sufficiently accurate for this.

To keep the axes of the horizontal gyroscopes in the horizontal plane, appropriate torques are applied to the gyroscope frame with the help of *torquers*. The torques needed can be calculated analogically or digitally in connection with solving for the position of the device. The approach

[12]It is actually possible to make an inertial platform follow the north by applying an appropriate torque to the gyroscopes, see Jekeli (2001, subsection 4.2.2) and section 5.10.

is called *Schuler tuning*.

5.4 Inertial navigation in the system of the solid Earth

For an alternative source, see Cooper (1987) pages 104–107, offering a slightly different approach.

5.4.1 *Earth rotation*

Write the vector of place in inertial space $\mathbf{x}_{\beta'}$ as a function of the vector of place in a co-ordinate frame co-rotating with the Earth $\mathbf{x}_\beta$:

$$\mathbf{x}_{\beta'} = R(\theta)\,\mathbf{x}_\beta, \tag{5.9}$$

in which θ is the *Greenwich sidereal time*, an angle describing the orientation of the Earth in space. Its time derivative $\omega_\oplus = \frac{d}{dt}\theta$ is the angular velocity of the Earth's rotation.

In equation 5.9 both the inertial and the co-rotating frame have their z axes oriented along the Earth's rotation axis. We conventionally describe vectors in space as two different abstract vectors of their components in both these frames:

$$\mathbf{x}_\beta = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \mathbf{x}_{\beta'} = \begin{bmatrix} x' \\ y' \\ z \end{bmatrix}.$$

The rotation matrix between these two frames is

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the time derivative of this matrix is (chain rule):

$$\frac{d}{dt}R(\theta) = \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d\theta}{dt}.$$

Differentiation by the Leibniz product rule yields for the velocity vector

$$\mathbf{v}_{\beta'} = \frac{d}{dt}\mathbf{x}_{\beta'} = \frac{d}{dt}\left(R(\theta)\,\mathbf{x}_\beta\right) = R(\theta)\,\frac{d}{dt}\mathbf{x}_\beta + \left(\frac{d}{dt}R(\theta)\right)\mathbf{x}_\beta =$$
$$= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{v}_\beta + \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d\theta}{dt} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Here, the second term equals

$$\left(\frac{d}{dt}R(\theta)\right)\boldsymbol{x}_\beta = \frac{d\theta}{dt}\begin{bmatrix} -x\sin\theta - y\cos\theta \\ x\cos\theta - y\sin\theta \\ 0 \end{bmatrix}.$$

Defining the angular velocity or rotation vector of the Earth as

$$\boldsymbol{\omega}_\oplus = \boldsymbol{\omega}_{\oplus,\beta} = \boldsymbol{\omega}_{\oplus,\beta'} = \begin{bmatrix} 0 \\ 0 \\ \omega_\oplus \end{bmatrix} = \frac{d\theta}{dt}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{5.10}$$

yields

$$\begin{aligned} R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x}_\beta \right\rangle &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \left\langle \frac{d\theta}{dt}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right\rangle = \\ &= \frac{d\theta}{dt}\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} -y \\ x \\ 0 \end{bmatrix} = \frac{d\theta}{dt}\begin{bmatrix} -x\sin\theta - y\cos\theta \\ x\cos\theta - y\sin\theta \\ 0 \end{bmatrix}, \end{aligned}$$

the same result. So

$$\left(\frac{d}{dt}R(\theta)\right)\boldsymbol{x}_\beta = R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x}_\beta \right\rangle.$$

It follows that

$$\boldsymbol{v}_{\beta'} = R(\theta)\,\boldsymbol{v}_\beta + R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x}_\beta \right\rangle. \tag{5.11}$$

Take a snapshot when the inertial and co-rotating axes are parallel: $\theta = 0 \implies R(\theta) = I$, yielding

$$\boldsymbol{v}_{\beta'} = \boldsymbol{v}_\beta + \left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x}_\beta \right\rangle \iff \boldsymbol{v}^* = \boldsymbol{v} + \left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x} \right\rangle.$$

Thus we have demonstrated the well-known *result*:

> *The effect of rotational motion of the co-ordinate frame on the time derivative of a vector can be presented as the cross product of the rotation vector $\boldsymbol{\omega}_\oplus$ with this vector.*

This applies generally, so, for example

$$\begin{aligned} \dot{R}(\theta)\,\boldsymbol{v} &= R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{v} \right\rangle, \\ \dot{R}(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x} \right\rangle &= R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{x} \right\rangle \right\rangle, \end{aligned}$$

results that we shall use next. Differentiating equation 5.11 again yields the acceleration:

$$\begin{aligned}
\mathbf{a}_{\beta'} &= R(\theta)\,\mathbf{a}_\beta + \dot{R}(\theta)\,\mathbf{v}_\beta + \frac{d}{dt}\Big(R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x}_\beta \right\rangle\Big) = \\
&= R(\theta)\,\mathbf{a}_\beta + R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{v}_\beta \right\rangle + \\
&\quad + \Big(R(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{v}_\beta \right\rangle + \dot{R}(\theta)\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x}_\beta \right\rangle\Big) = \\
&= R(\theta)\Big(\mathbf{a}_\beta + 2\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{v}_\beta \right\rangle + \Big\langle \boldsymbol{\omega}_\oplus \times \left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x}_\beta \right\rangle\Big\rangle\Big).
\end{aligned}$$

By putting again $\theta = 0$, we find

$$\mathbf{a}_{\beta'} = \mathbf{a}_\beta + 2\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{v}_\beta \right\rangle + \Big\langle \boldsymbol{\omega}_\oplus \times \left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x}_\beta \right\rangle\Big\rangle. \tag{5.12}$$

5.4.2 *The acceleration*

The three-dimensional geocentric co-ordinate frame (x, y, z) defined on the rotating Earth is not inertial: for the acceleration, equation 5.12 applies. In symbolic notation:

$$\mathbf{a}^* = \mathbf{a} + 2\left\langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \right\rangle + \Big\langle \boldsymbol{\omega}_\oplus \times \left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x} \right\rangle\Big\rangle, \tag{5.13}$$

in which

$\mathbf{a}^*$	acceleration in the inertial frame
$\mathbf{a}$	acceleration relative to the Earth's surface, in other words, in an Earth-fixed, "co-rotating" frame
$\boldsymbol{\omega}_\oplus$	rotation vector of the Earth, a constant axial vector pointing along the rotation or z axis
$\mathbf{v}$	velocity in the same frame co-rotating with the Earth
$\mathbf{x}$	geocentric location of the vehicle.

In equation 5.13, the second term on the right-hand side is the so-called *Coriolis acceleration* and the third term is the *centrifugal acceleration*.

5.4.3 *Fundamental equation of inertial navigation*

Accelerometers measure in practice the *combined effect* of local gravity and the geometric motions of the vehicle. In other words, the acceleration measured *in the vehicle* is, in an inertial frame,

$$\widetilde{\mathbf{a}} = \mathbf{a}^* - \mathbf{g}^*(\mathbf{x}), \tag{5.14}$$

or with equation 5.13,

$$\widetilde{\mathbf{a}} = \mathbf{a} + \overbrace{2\left\langle \boldsymbol{\omega}_{\oplus} \times \mathbf{v} \right\rangle}^{\text{Coriolis acceleration}} + \overbrace{\left\langle \boldsymbol{\omega}_{\oplus} \times \left\langle \boldsymbol{\omega}_{\oplus} \times \mathbf{x} \right\rangle \right\rangle}^{\text{centrifugal acceleration}} - \mathbf{g}^{*}\left(\mathbf{x}\right),$$

in which

- $\widetilde{\mathbf{a}}$ on-board measured acceleration vector. This is $-\widetilde{\mathbf{g}}$, the opposite of "gravity" as sensed inside the vehicle[13]
- $\mathbf{a}^{*}$ acceleration in the inertial frame
- $\mathbf{a}$ acceleration in a co-rotating frame
- $\mathbf{g}^{*}(\mathbf{x})$ *gravitational* acceleration as a function of place $\mathbf{x}$, being the acceleration of free fall in an inertial frame.

For a spherically symmetric Earth, $\mathbf{g}^{*}$ can be calculated directly from Newton's law of gravitation:

$$\mathbf{g}^{*}(\mathbf{x}) \approx -GM_{\oplus}\frac{\mathbf{x}}{\|\mathbf{x}\|^{3}},$$

but more complex models may also be used, such as the normal gravity field of an ellipsoid of revolution where the effect of the Earth's flattening is included, and even very detailed Earth gravitational field models, such as EGM2008 (Earth Gravity Model 2008).

Write

$$\mathbf{g} = \mathbf{g}^{*} - \left\langle \boldsymbol{\omega}_{\oplus} \times \left\langle \boldsymbol{\omega}_{\oplus} \times \mathbf{x} \right\rangle \right\rangle, \tag{5.15}$$

where $\mathbf{g}$ is the *gravity vector*, the resultant of gravitation and centrifugal acceleration, the acceleration of free fall in a co-rotating frame. Then

$$\widetilde{\mathbf{a}} = \mathbf{a} + 2\left\langle \boldsymbol{\omega}_{\oplus} \times \mathbf{v} \right\rangle - \mathbf{g}(\mathbf{x}). \tag{5.16}$$

Equations 5.14 and 5.16 are both referred to as the *fundamental equation of inertial navigation*. They both allow us to simultaneously and dynamically, "on the fly", integrate location $\mathbf{x}$, velocity $\mathbf{v}$, and acceleration $\mathbf{a}$, provided that the external field $\mathbf{g}$ or $\mathbf{g}^{*}$ is given and that on-board acceleration measurements $\widetilde{\mathbf{a}}$ are carried out.

[13] It is important to understand this. The acceleration $\widetilde{\mathbf{a}} = -\widetilde{\mathbf{g}}$ as sensed by acceleration sensors represents the acceleration *of the vehicle* with respect to a free-fall frame. Gravity $\widetilde{\mathbf{g}} = -\widetilde{\mathbf{a}}$ as sensed by gravity sensors represents the acceleration of free fall *inside the vehicle* with respect to the vehicle frame. In other words, the pseudo-force experienced inside the vehicle.

In a frame co-rotating with the Earth, the rotation of our planet causes a slow turning in the east-west direction of the vector of gravity sensed by the accelerometers, relative to the inertial directions defined by the gyroscopes, even though the vehicle were standing still on the ground.

This phenomenon is used to orient the gyroscope frame correctly relative to the local north direction — or equivalently, to solve the local north direction in the system of the gyroscope frame — before, for example, the take-off of an aeroplane or launch of a rocket. Furthermore, the accelerometers give the direction of local gravity, the plumb line, straight away. Together, the two directions are enough to orient the whole frame — except on the North or South Poles. luotiviiva

5.5 The stabilised platform

Let us first study the *stabilised platform*, which is implemented by a gyroscope that is attached to a frame which is kept aligned with the local horizon. A stabilised platform serves, for example, as the mounting platform for a sea or airborne gravimeter, because the measurement axis of the instrument has to be constantly aligned with the local plumb line to within a few minutes of arc.

In the stabilised-platform solution, one uses a feedback loop called a *Schuler loop* to control the direction of the gyroscope's spin axis so that it, and the inner ring it is mounted in, remain in the horizontal plane. This happens in such a way that trying to turn the gyroscope frame in the horizontal plane — around the vertical axis of the frame — causes the gyroscope to *precess*: the spin axis of the gyroscope itself turns up- or downwards.

The stabilised platform requires a *suitable sensor* that detects that the axis of the gyroscope is out of the horizontal plane by an angle θ. The sensor sends a signal $f(\theta)$ related to θ through the feedback loop to the motor controller or *actuator* of the vertical axis, figure 5.9. To construct such a loop that works well even in the presence of vehicle motion is challenging, and we defer to section 5.7 for a discussion of this.

When the instrument is standing on solid ground, it is easier. Assume that the torque about the vertical axis is simply made proportional to the sensed axis deviation θ from the horizontal plane. Then the change of θ with time is

$$\frac{d\theta}{dt} = -k_1\theta,$$

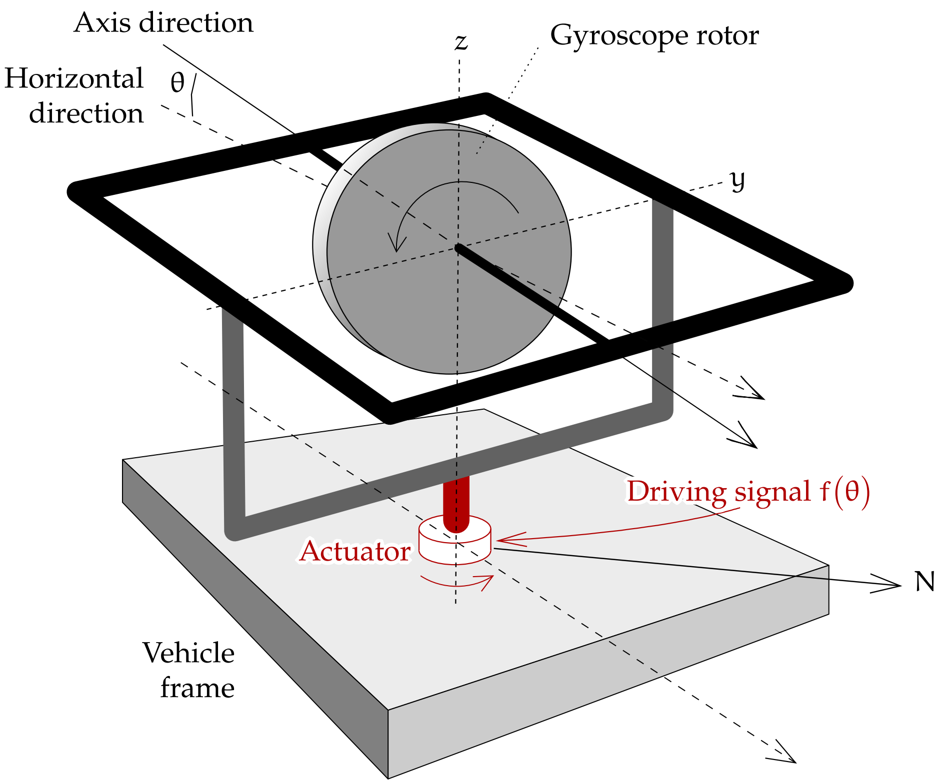


FIGURE 5.9. Principle of the stabilised platform. The driving signal produces a precessional motion that keeps the gyroscope axis within the plane of the horizon. Conventionally, the vertical or *z* axis, around which a torque is applied, is called the input axis, and the horizontal y axis, around which precession is produced, the output axis.

with the solution

$$\theta(t) = \theta(t_0)\, e^{-k_1(t-t_0)},$$

in other words, the deviation goes to zero exponentially. We can make this happen with suitable speed by tuning the constant k_1 of the feedback loop. This technique may be used to level a stabilised platform in preparation for flight.

5.6 The gyrocompass

The feedback loop visible in figure 5.10 of the gyrocompass again makes use of *the rotation of the Earth*. Because the Earth rotates around her axis, the plane of the horizon is tilting all the time. The eastern horizon sinks, the western rises. A freely suspended, spinning gyroscope, the spin axis of which was initially in the horizontal plane, will no longer be horizontal after an elapse of time.

If the rotational angular velocity of the Earth is $\omega_\oplus$, then the time

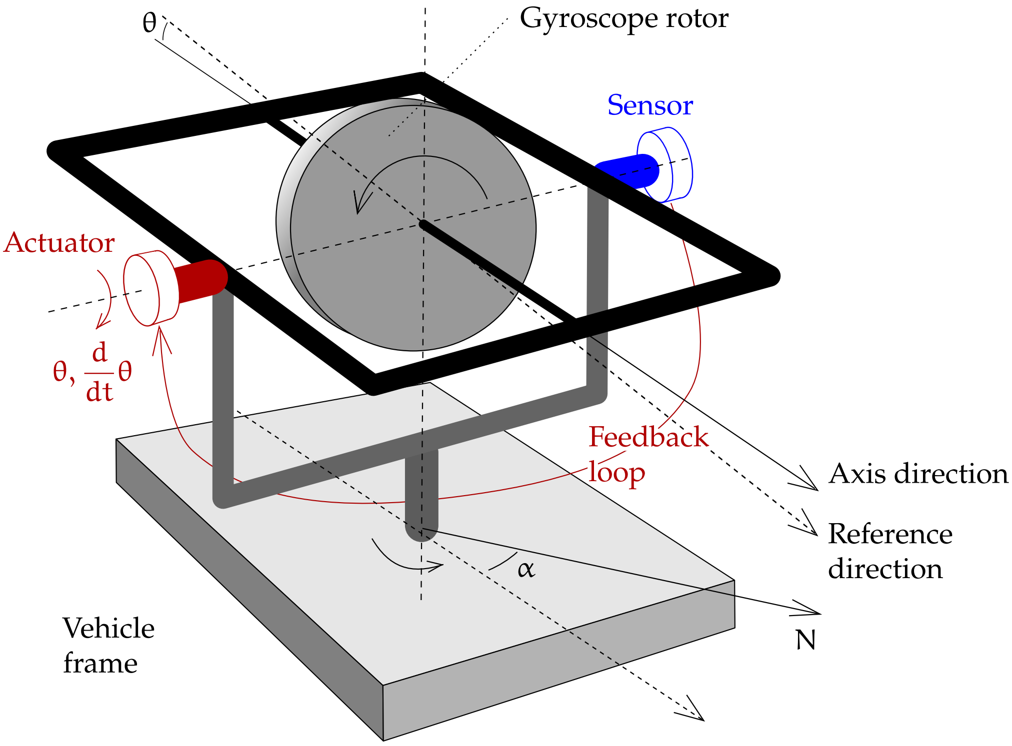


FIGURE 5.10. Principle of the gyrocompass. The feedback loop produces a precessional motion that makes the gyro's spin axis turn to the north.

derivative of the angle θ will be, because of this phenomenon,

$$\frac{d\theta}{dt} = \omega_\oplus \cos\varphi \sin\alpha,$$

in which φ is the latitude and α the azimuth of the gyroscope spin axis. phi φϕΦ

The feedback loop gets from the sensor the *time derivative* $\frac{d}{dt}\theta$ of the angle θ and feeds it after suitable amplification into the actuator. As the actuator tries to turn the gyroscope axis back into the horizontal plane, the outcome will be *precession* about the vertical axis: α changes. We write the equation, with positive k_2,

$$\frac{d\alpha}{dt} = -k_2 \frac{d\theta}{dt} = -k_2 \omega_\oplus \cos\varphi \sin\alpha. \tag{5.17}$$

If α is small enough, we have $\sin\alpha \approx \alpha$ and the solution is

$$\alpha(t) \approx \alpha(t_0)\, e^{-k_2 \omega_\oplus (t-t_0)\cos\varphi}.$$

This means that α goes exponentially and asymptotically to zero and so *the gyroscope axis turns to the north*. This is how a gyrocompass works. Of course, this assumes that the device stays at the same spot and remains level — or in practice that it moves only slowly, like on a ship.

A more general way to build a working gyrocompass uses θ itself in addition to its time derivative. If we write

$$\frac{d\alpha}{dt} = -k_3\theta,$$

we obtain by differentiation

$$\frac{d^2\alpha}{dt^2} = -k_3\frac{d\theta}{dt} = -k_3\omega_\oplus \cos\varphi \sin\alpha \approx -k_3\omega_\oplus \cos\varphi \cdot \alpha.$$

This is a *harmonic oscillator,* with solutions

$$\alpha(t) = \cos\left(t\sqrt{k_3\omega_\oplus \cos\varphi}\right), \qquad \alpha(t) = \sin\left(t\sqrt{k_3\omega_\oplus \cos\varphi}\right).$$

Unfortunately these solutions are periodic and do not converge to the north direction $\alpha = 0$. The best solution is obtained by combining θ and $\frac{d}{dt}\theta$ in the following way:

$$\frac{d^2\alpha}{dt^2} = -k_2\,\omega_\oplus \cos\varphi \frac{d\alpha}{dt} - k_3\,\omega_\oplus \cos\varphi \cdot \alpha$$

or

$$\frac{d^2\alpha}{dt^2} + \omega_\oplus \cos\varphi \left(k_2\frac{d\alpha}{dt} + k_3\,\alpha\right) = 0.$$

This is a general second-order ordinary differential equation. Depending on coefficients k_2 and k_3, it will have wave-like, exponentially damped, or *critically damped* solutions, see Wikipedia, Damping ratio. The last-mentioned alternative is best suited for a functioning compass.

vaimennus

If the circular frequency of the oscillation is $\omega \stackrel{\text{def}}{=} \sqrt{k_3\,\omega_\oplus \cos\varphi}$, and

$$k_2 = \frac{2\omega}{\omega_\oplus \cos\varphi},$$

we obtain the critically damped case

$$\frac{d^2\alpha}{dt^2} + 2\omega\frac{d\alpha}{dt} + \omega^2\alpha = 0,$$

of which the general solution is

$$\alpha(t) = (a + bt)\, e^{-\omega t},$$

in which a and b are arbitrary constants determined by the initial conditions.

Often k_3, the harmonic restoration coefficient, is implemented by attaching a heavy semi-ring rigidly to the inner ring of the gyroscope frame in such a way that the semi-ring extends downwards. This weight tries then to pull the spin axis of the gyroscope back to the horizontal plane. The damping coefficient k_2 is again implemented traditionally by using a viscous fluid in the bearings of the inner ring.

vaimennus-kerroin

5·7 The Schuler pendulum

5·7·1 Principle

A *Schuler*[14] *pendulum* is a pendulum the length of which is the same as the Earth's radius R. If that kind of pendulum were practically possible, for example as a mass at the end of a long massless rod, its period would be (in a one-g gravity field!) 14

$$P_{\text{Schuler}} = 2\pi\sqrt{\frac{R}{g}},$$

in which g is gravity on the Earth's surface.

This period, the *Schuler period*, $P_{\text{Schuler}} = 84.4$ minutes, is the same as the orbital period of an Earth satellite near the Earth's surface — if the atmosphere didn't exist.

Although it is impossible to build a pendulum this long, one could very well imagine a pendulum with a period of P_{Schuler}, for example an extended object suspended from a point very close to its centre of mass.[15] 15

The simplest pendulum of all is a mass on the end of a massless rod. Let its length be ℓ. If the pendulum swings out of the vertical by an angle[16] θ, the restoring force will be 16

$$F = mg\sin\theta,$$

and, as its mass is m, we may compute its acceleration using Newton's second law, as follows:

$$m\frac{d^2(-\theta\ell)}{dt^2} = mg\sin\theta \implies \frac{d^2\theta}{dt^2} \approx -\frac{g}{\ell}\theta, \qquad (5.18)$$

an oscillation equation, of which one solution is

$$\theta(t) = \sin\left(t\sqrt{\frac{g}{\ell}}\right),$$

[14]Maximilian Joseph Johannes Eduard Schuler (1882–1972) was a German pioneer of navigation technology, Wikipedia, Max Schuler.

[15]Sadly, not even this will work: as Schuler showed already, the distance between the suspension point and centre of mass of the pendulum would, for any reasonable pendulum size, be sub-micrometre.

[16]For consistency with other angles, the angle θ is reckoned positive for negative linear displacement of the mass.

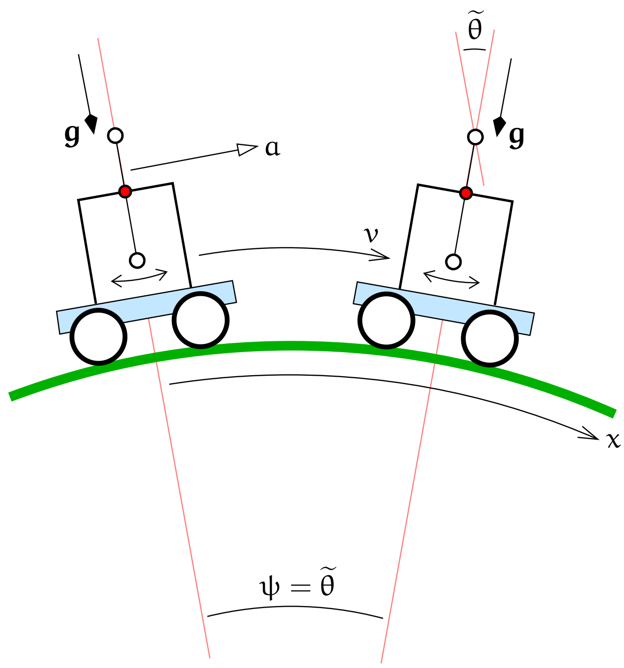


FIGURE 5.11. One-dimensional carriage with a Schuler pendulum moving on the curved surface of the Earth.

from which follows the period

$$P = 2\pi\sqrt{\frac{\ell}{g}}.$$

5.7.2 *Pendulum on a carriage*

Mount this pendulum on a carriage that moves in the horizontal direction with a linear acceleration $a(t)$. The mass of the pendulum will, in the frame of the carriage, experience an equally large, oppositely directed acceleration $-a(t)$. Because the length of the pendulum is ℓ, its angular acceleration is

$$\frac{d^2\widetilde{\theta}}{dt^2} = \frac{1}{\ell}\left(a - g\cdot\left(\widetilde{\theta} - \psi\right)\right).$$

Here, $\widetilde{\theta}$ stands for the deviation of the pendulum from the plumb line of the *starting point*.

The distance that the carriage has travelled as a function of time will be $x(t)$. This linear distance is related to the acceleration by

$$\frac{d^2x(t)}{dt^2} = a(t).$$

The same distance expressed as a *geocentric angular distance*, in other psi $\psi\Psi$ words an angle ψ viewed from the centre of the Earth, is $\psi(t) = {}^{x(t)}/_{R}$.

This quantity, which also represents the change in the local plumb line along the journey, obeys the differential equation

$$\frac{d^2\psi(t)}{dt^2} = \frac{1}{R}a(t). \tag{5.19}$$

Subtraction yields

$$\frac{d^2}{dt^2}\left(\widetilde{\theta} - \psi\right) = \frac{1}{\ell}\left(a - g(\widetilde{\theta} - \psi)\right) - \frac{a}{R}.$$

Now assume that the length of the pendulum $\ell = R$. Then, with $\theta \stackrel{\text{def}}{=} \widetilde{\theta} - \psi$, it follows that

$$\frac{d^2\theta(t)}{dt^2} = -\frac{g}{R}\theta(t). \tag{5.20}$$

Here, θ is the angle between the pendulum and the *local* plumb line. Equation 5.20 is identical to pendulum equation 5.18.

One solution of equation 5.20 is $\theta = 0$ identically. So

Even though the carriage moves and accelerates in a horizontal direction, the pendulum remains pointing to the centre of the Earth.

This is the defining property of the Schuler pendulum.

5.7.3 Implementation in an inertial device

A stabilised platform based inertial device implements a pair of response or feedback loops, so-called *Schuler loops*, that make the whole gyroscope frame act like a Schuler pendulum. Whenever the frame turns out of the horizontal level, the accelerometers of the horizontal directions x and y measure the projection of gravity $\mathbf{g}$ onto the tilting plane, and send correcting impulses to the corresponding gyroscope frame's actuators. This is how the frame tracks the local horizontal level.

According to equation 5.19:

$$\frac{d^2\psi(t)}{dt^2} = \frac{1}{R}a(t), \tag{5.19}$$

in which $a(t)$ is the linear acceleration in the x direction and R the radius of the Earth. The angle function $\psi(t)$ describes how the local direction of the plumb line changes along the journey.

The angular momentum in the gyroscope rotor is according to equation 5.7:

$$L = L_x = J_x\omega, \tag{5.21}$$

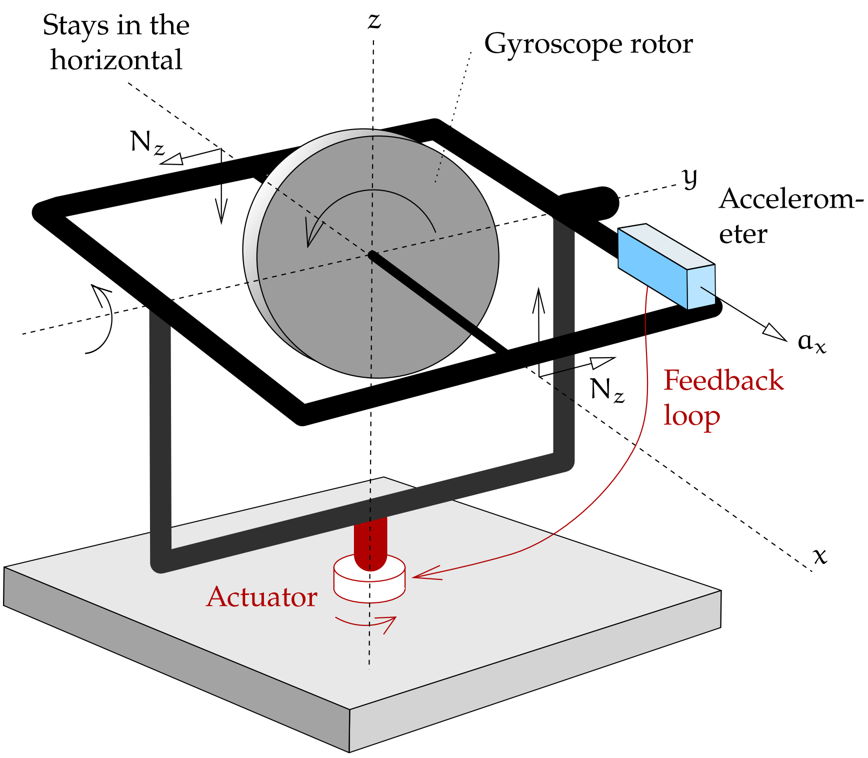


FIGURE 5.12. Schuler response loop in the x direction.

in which J_x is the moment of inertia of the rotor around its spin axis, and ω the spin rate. This assumes that the spin axis of the gyro lies in the plane of the horizon (x, y), specifically along the x axis direction.

In the geometry depicted in figure 5.12:

$$\mathbf{L} = L_x\mathbf{i} + L_y\mathbf{j} + L_z\mathbf{k} \approx J_x\omega\mathbf{i} + L_y\mathbf{j} + L_z\mathbf{k},$$

with $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ the orthonormal basis of the (x, y, z) frame. Assume that vectors $\mathbf{i}$ and $\mathbf{j}$ lie in the horizontal plane of the starting point, and $\mathbf{k}$ points up.

The absolute total amount of direction change that the gyroscope spin axis undergoes, reckoned from the starting point, is $\widetilde{\theta} = \theta + \psi$. Add together the equations 5.20 and 5.19:

$$\left.\begin{aligned}\frac{d^2}{dt^2}\theta(t) &= -\frac{g}{R}\theta(t)\\ \frac{d^2}{dt^2}\psi(t) &= \frac{1}{R}a(t)\end{aligned}\right\} \Longrightarrow \frac{d^2}{dt^2}\widetilde{\theta}(t) = \frac{1}{R}\overbrace{\left(a(t) - g\,\theta(t)\right)}^{\widetilde{a}(t)}. \qquad (5.22)$$

The angles θ, ψ, and $\widetilde{\theta}$ are assumed small. Now, affecting a gyroscope axis direction change of $\widetilde{\theta}$ means, with equation 5.21, a change in the
17 angular momentum about the z axis of[17]

$$L_z = -\widetilde{\theta} L = -\widetilde{\theta} J_x \omega.$$

The second derivative of $\widetilde{\theta}$ is

$$\frac{d^2}{dt^2}\widetilde{\theta} = -\frac{1}{J_x\omega}\frac{d^2}{dt^2}L_z.$$

With equation 5.22:

$$\frac{d^2}{dt^2}L_z = -J_x\omega\frac{d^2}{dt^2}\widetilde{\theta} = -\frac{J_x\omega}{R}\widetilde{a} \implies \frac{d}{dt}L_z = -\frac{J_x\omega}{R}\int \widetilde{a}(t)\,dt.$$

By equation 5.2:

$$N_z = \frac{d}{dt}L_z = -\frac{J_x\omega}{R}\int \widetilde{a}(t)\,dt. \qquad (5.23)$$

In equation 5.23, N_z is the required torque around the z axis, see figure 5.12. This is how a Schuler loop is implemented.

In equation 5.23, the measured acceleration is $\widetilde{a} = a - g\theta$, in which a is the geometric acceleration and $g\theta$ the component of gravity acting in the gyroscope spin axis direction.

According to equation 5.23, the Schuler loop is implemented either on the hardware level — in older equipment, the factor $-J_x\omega/R$ is a device constant and integration is done analogously in hardware[18] — or in the software of an inertial device. There are always *two* Schuler loops, one for the x direction and one for the y direction. 18

5.8 Mechanisation

Because a real-life inertial device is quite a lot more complicated than the simple principles, modelling the behaviour of all the parts must be done carefully. This model is called the *mechanisation* of the inertial device.

A one-dimensional carriage on the surface of a spherical Earth, figure 5.11, is presented as a simple example of mechanisation.

The velocity is by definition

$$\frac{dx(t)}{dt} = v(t).$$

Acceleration is *measured* continuously by an acceleration sensor, the measured value being $\widetilde{a}(t)$. This measured quantity, a function of time, consists of two parts:

---

[17]It is assumed that the vector $\boldsymbol{\omega}$ points in the positive x direction.

[18]For example by using the integrating gyroscopic accelerometer of section 5.2.2.

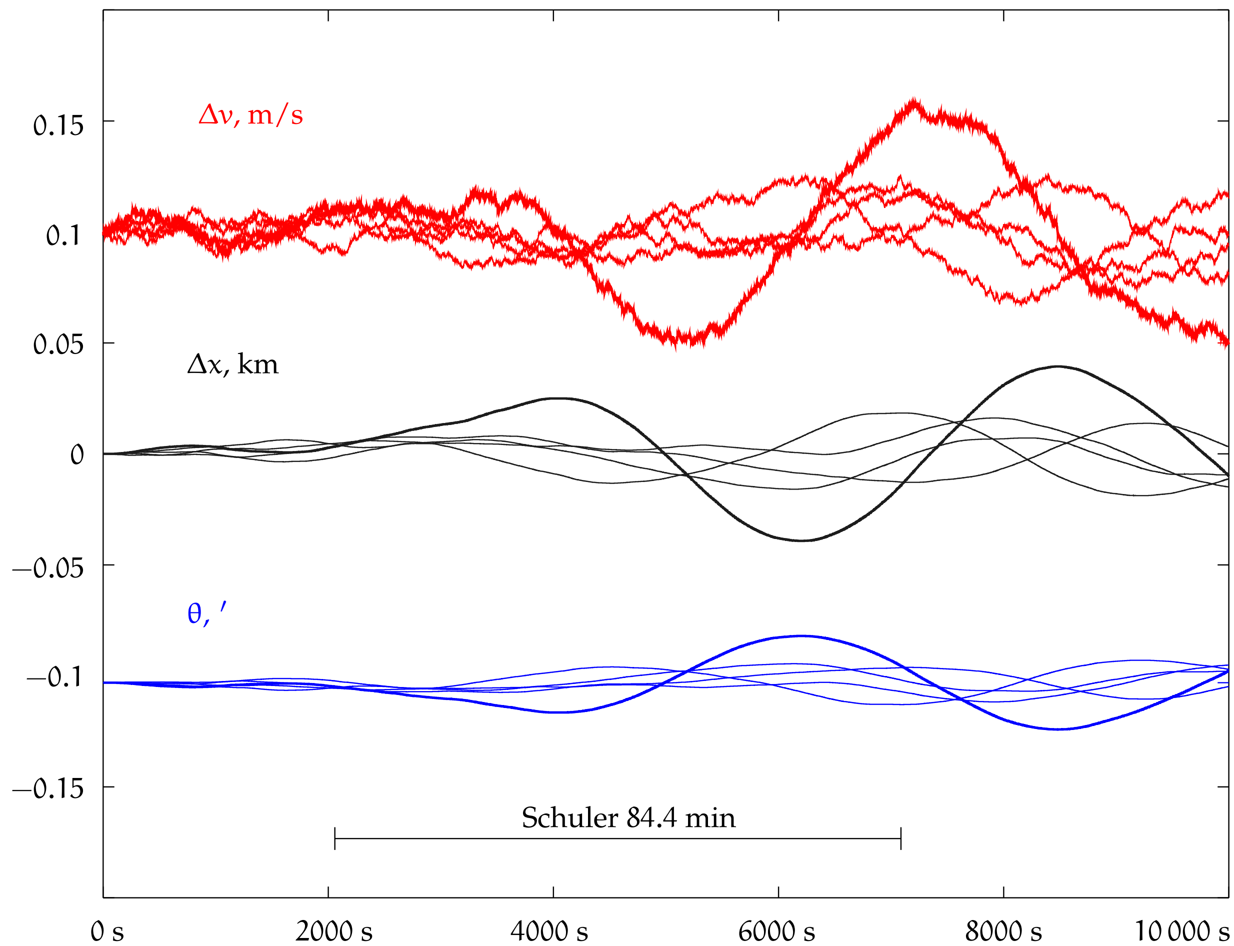


FIGURE 5.13. Error propagation of the one-dimensional mechanisation of equation 5.27. Values used: $\underline{n}_a$ is ± 100 mGal, $\underline{n}_g = 0$. The Δv and θ curves are offset for clarity.

- the geometric acceleration

$$a(t) = \frac{d^2x(t)}{dt^2} = \frac{dv(t)}{dt}$$

- the projection of the gravity vector onto the accelerometer's axis, $g\,\theta(t)$, in which $\theta(t)$ is the angle of tilt of the carriage from the local vertical or plumb-line direction.

The result is

$$\frac{dv(t)}{dt} = \tilde{a}(t) + g\,\theta(t), \qquad (5.24)$$

in which the function $\tilde{a}(t)$ is the result of a continuous measurement process.

Finally we discuss the Schuler loop. The angle of deflection θ behaves like a Schuler pendulum and tries to revert to zero according to equation

5.20:

$$\frac{d^2\theta(t)}{dt^2} = -\frac{g}{R}\theta(t). \tag{5.20}$$

Next, we *linearise*. Define *approximate values*, the functions of time $x^{(0)}(t)$ and $v^{(0)}(t)$, and the difference quantities $\Delta x(t) \stackrel{\text{def}}{=} x(t) - x^{(0)}(t)$ and $\Delta v(t) \stackrel{\text{def}}{=} v(t) - v^{(0)}(t)$, as follows:

$$\frac{dx^{(0)}(t)}{dt} = v^{(0)}(t), \quad \frac{dv^{(0)}(t)}{dt} = \tilde{a}(t), \tag{5.25}$$

in which $\tilde{a}(t)$ is assumed continuously measured,[19] and

19

$$\frac{d\,\Delta x(t)}{dt} = \Delta v(t), \quad \frac{d\,\Delta v(t)}{dt} = g\,\theta(t).$$

Now we can substitute into equation 5.20 the derivative

$$g\,\theta(t) = \frac{d\,\Delta v(t)}{dt},$$

with the result

$$\frac{d^2\theta(t)}{dt^2} = -\frac{1}{R}\frac{d\,\Delta v(t)}{dt}.$$

By integrating — meaning leaving out one $\frac{d}{dt}$ from each side — we obtain

$$\frac{d\theta(t)}{dt} = -\frac{1}{R}\Delta v(t), \tag{5.26}$$

and the complete Kalman-filter dynamic model is obtained:

$$\frac{d}{dt}\overbrace{\begin{bmatrix}\Delta\underline{x}\\ \Delta\underline{v}\\ \underline{\theta}\end{bmatrix}}^{\underline{x}} = \overbrace{\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & g\\ 0 & -{}^{1}\!/_{R} & 0\end{bmatrix}}^{F}\overbrace{\begin{bmatrix}\Delta\underline{x}\\ \Delta\underline{v}\\ \underline{\theta}\end{bmatrix}}^{\underline{x}} + \overbrace{\begin{bmatrix}0\\ \underline{n}_a\\ \underline{n}_g\end{bmatrix}}^{\underline{n}}. \tag{5.27}$$

We have added the possible (stochastic) noise terms $\underline{n}_a, \underline{n}_g$ of the acceleration sensor and the gyroscope.

This solution works in this way, that we continuously and precisely integrate the real-time approximate or reference values $v^{(0)}(t)$ and $x^{(0)}(t)$, equations 5.25, and with the help of the Kalman filter 5.27 we integrate $\Delta\underline{x}(t), \Delta\underline{v}(t)$, and $\underline{\theta}(t)$. tosiaikainen vertausarvo

In figure 5.13 we see how these quantities $\Delta x, \Delta y$ and θ behave over time. The oscillatory behaviour on the Schuler time scale is evident: in the solution, both the angle of deflection θ of the carriage and the velocity perturbation Δv — and also the location perturbation häiriö

TABLEAU 5.1. Mechanisation simulation in one dimension, `octave` code.

```
s = [1:10000];

x(s) = 0;
v(s) = 0;
th(s) = 0;

g = 9.8; R = 6378137;

for j = 1:5
  for i=1:9999
    v(i+1) = v(i) + g*th(i) + 0.001*(rand() - 0.5);
    x(i+1) = x(i) + v(i);
    th(i+1) = th(i) - v(i)/R + 0.00000000*(rand() - 0.5);
  endfor
  hold on
  plot(s, 0.001*x, 'b');
  plot(s, v + 0.1, 'c');
  plot(s, 57*60*th - 0.1, 'm');
endfor

print -dpdf "schuler.pdf"
```

20 Δx — swing harmonically like a Schuler pendulum,[20] with the period $P_{\text{Schuler}} = 84.4$ minutes. The height must be obtained in another way, for example in an aircraft by means of an atmospheric pressure sensor.

5.9 On the Earth's surface in two dimensions

This may be generalised to two dimensions. In this way, a "navigator" may be built that works on the surface of the Earth. Each of the elements in the state vector in equation 5.27 should be duplicated into a north and an east component.

---

[19]It would be appropriate to consider this measured $\tilde{a}(t)$, and all quantities derived from it, as stochastic. We don't do that here, but see $\underline{n}_a$ in equation 5.27.

[20]If the angle θ has for example an amplitude $A_\theta = 1' = 2.9 \cdot 10^{-4}$ rad, it follows from the equation

$$\frac{d\,\Delta v}{dt} = g\theta,$$

that

- Δv's amplitude is $A_{\Delta v} = g\sqrt{R/g}\,A_\theta = 2.3\,{}^{m}/_{s}$

As geographical co-ordinates φ and λ are in global use more practical than map co-ordinates x_{north} and x_{east}, we make the following substitutions. Replace Δx_{north} by latitude $\Delta\varphi = \Delta x_{\text{north}}/M$ and Δx_{east} by longitude $\Delta\lambda = \Delta x_{\text{east}}/N\cos\varphi$. This yields the linearised derivatives lambda $\lambda\Lambda$

$$\frac{d}{dt}\Delta\varphi = \frac{1}{M}\Delta v_{\text{north}}, \quad \frac{d}{dt}\Delta\lambda = \frac{\Delta v_{\text{east}}}{N\cos\varphi},$$

ignoring the dependence of $\Delta\lambda$ on φ through $\cos\varphi$. $N(\varphi)$ and $M(\varphi)$ are the transversal and meridional radii of curvature, respectively, of the reference ellipsoid. When not at sea level, N and M should be replaced by $N+h$ and $M+h$, respectively, h being the height above the reference ellipsoid. The slight dependence on latitude φ of both is ignored. vertaus-ellipsoidi

We gloss over the circumstance that, depending on implementation, the actual horizontal gyroscope and accelerometer axes may not be aligned with the north and east directions.

The *turning* of the north and east directions when travelling in longitude demands special treatment. Naively, one would write equation 5.24 for the velocity as

$$\frac{d}{dt}v_{\text{north}} = \widetilde{a}_{\text{north}} + g\,\theta_{\text{north}}, \quad \frac{d}{dt}v_{\text{east}} = \widetilde{a}_{\text{east}} + g\,\theta_{\text{east}},$$

but that would be too simple. There are extra terms:

$$\begin{aligned} \frac{d}{dt}v_{\text{north}} &= \widetilde{a}_{\text{north}} + g\,\theta_{\text{north}} - v_{\text{east}}\frac{d}{dt}\alpha, \\ \frac{d}{dt}v_{\text{east}} &= \widetilde{a}_{\text{east}} + g\,\theta_{\text{east}} + v_{\text{north}}\frac{d}{dt}\alpha, \end{aligned}$$

in which α is the heading or azimuth angle, reckoned clockwise from the north.

The angles θ_{north} and θ_{east} are the north and east tilt angles with respect to the local plumb line, equation 5.26. There are similar extra terms for these tilt angles, and one more term for Earth rotation:

$$\begin{aligned} \frac{d}{dt}\theta_{\text{north}} &= -\frac{1}{M}\Delta \underline{v}_{\text{north}} - \theta_{\text{east}}\frac{d}{dt}\alpha, \\ \frac{d}{dt}\theta_{\text{east}} &= -\frac{1}{N}\Delta \underline{v}_{\text{east}} + \theta_{\text{north}}\frac{d}{dt}\alpha - \omega_{\oplus}\cos\varphi. \end{aligned}$$

∘ Δx's amplitude is $A_{\Delta x} = \sqrt{R/g}\,A_{\Delta v} = 1855\,\text{m}$.

$$
\frac{d}{dt}\overbrace{\begin{bmatrix} \Delta\underline{\varphi} \\ \Delta\underline{\lambda} \\ \hline \Delta\underline{v}_{\text{north}} \\ \Delta\underline{v}_{\text{east}} \\ \hline \Delta\underline{\theta}_{\text{north}} \\ \Delta\underline{\theta}_{\text{east}} \\ \hline \Delta\underline{\alpha} \end{bmatrix}}^{\underline{x}} = \overbrace{\left[\begin{array}{cc|cc|cc|c} 0 & 0 & \frac{1}{M} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{N\cos\varphi} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -\dot{\alpha} & g & 0 & 0 \\ 0 & 0 & \dot{\alpha} & 0 & 0 & g & 0 \\ \hline 0 & 0 & -\frac{1}{M} & 0 & 0 & -\dot{\alpha} & 0 \\ \omega_{\oplus}\sin\varphi & 0 & 0 & -\frac{1}{N} & \dot{\alpha} & 0 & 0 \\ \hline \dot{\alpha}_{\varphi} & 0 & 0 & \frac{\tan\varphi}{N} & 0 & 0 & 0 \end{array}\right]}^{F} \overbrace{\begin{bmatrix} \Delta\underline{\varphi} \\ \Delta\underline{\lambda} \\ \hline \Delta\underline{v}_{\text{north}} \\ \Delta\underline{v}_{\text{east}} \\ \hline \Delta\underline{\theta}_{\text{north}} \\ \Delta\underline{\theta}_{\text{east}} \\ \hline \Delta\underline{\alpha} \end{bmatrix}}^{\underline{x}} + \overbrace{\begin{bmatrix} 0 \\ 0 \\ \hline \underline{n}_{a,\,\text{north}} \\ \underline{n}_{a,\,\text{east}} \\ \hline \underline{n}_{g,\,\text{north}} \\ \underline{n}_{g,\,\text{east}} \\ \hline \underline{n}_{g,\alpha} \end{bmatrix}}^{\underline{n}}. \tag{5.28}
$$

FIGURE 5.14. Mechanisation equation, inertial navigation on the Earth's surface.

The corresponding delta equations are

$$
\begin{aligned}
\frac{d}{dt}\Delta v_{\text{north}} &= g\,\Delta\theta_{\text{north}} - \Delta v_{\text{east}}\frac{d}{dt}\alpha, \\
\frac{d}{dt}\Delta v_{\text{east}} &= g\,\Delta\theta_{\text{east}} + \Delta v_{\text{north}}\frac{d}{dt}\alpha, \\
\frac{d}{dt}\Delta\theta_{\text{north}} &= -\frac{1}{M}\Delta\underline{v}_{\text{north}} - \Delta\theta_{\text{east}}\frac{d}{dt}\alpha, \\
\frac{d}{dt}\Delta\theta_{\text{east}} &= -\frac{1}{N}\Delta\underline{v}_{\text{east}} + \Delta\theta_{\text{north}}\frac{d}{dt}\alpha + \omega_{\oplus}\sin\varphi\,\Delta\varphi.
\end{aligned}
$$

The azimuth derivative $\frac{d}{dt}\alpha = \dot{\alpha}$ is taken from the integration of the reference value $\alpha^{(0)}$, equation 5.29.

In two dimensions, the mechanisation equation now looks like equation 5.28.

The coefficient matrix F is evaluated using approximate or reference val-
21 ues, which are to be *precisely* integrated over time using the equations[21]

$$
\frac{d}{dt}\begin{bmatrix} \varphi^{(0)} \\ \lambda^{(0)} \end{bmatrix}(t) = \begin{bmatrix} v_{\text{north}}^{(0)} \Big/ M(\varphi^{(0)}) \\ v_{\text{east}}^{(0)} \Big/ N(\varphi^{(0)})\cos\varphi^{(0)} \end{bmatrix}(t),
$$

$$
\begin{aligned}
\frac{d}{dt}\begin{bmatrix} v_{\text{north}}^{(0)} \\ v_{\text{east}}^{(0)} \end{bmatrix}(t) = \begin{bmatrix} \widetilde{a}_{\text{north}} \\ \widetilde{a}_{\text{east}} \end{bmatrix}(t) + g\begin{bmatrix} \theta_{\text{north}}^{(0)} \\ \theta_{\text{east}}^{(0)} \end{bmatrix}(t) &+ \\
+ \begin{bmatrix} -v_{\text{east}}^{(0)} \\ v_{\text{north}}^{(0)} \end{bmatrix}(t)\cdot\frac{d}{dt}\alpha^{(0)}(t)&,
\end{aligned}
$$

---

[21]In order to actually form $\widetilde{a}_{\text{north}}$ and $\widetilde{a}_{\text{east}}$ when measured are $\widetilde{a}_x$ and $\widetilde{a}_y$, of course the precise azimuth α is needed.

$$\frac{d}{dt}\begin{bmatrix}\theta^{(0)}_{\text{north}}\\ \theta^{(0)}_{\text{east}}\end{bmatrix}(t) = \begin{bmatrix}-\theta^{(0)}_{\text{east}}\\ \theta^{(0)}_{\text{north}}\end{bmatrix}(t)\cdot\frac{d}{dt}\alpha^{(0)}(t) - \begin{bmatrix}0\\ \omega_{\oplus}\cos\varphi^{(0)}(t)\end{bmatrix}.$$

The equation for the heading or azimuth is

$$\frac{d}{dt}\alpha = \frac{\tan\varphi}{N}\left(v_{\text{east}} + \omega_{\oplus} N\cos\varphi\right) = \frac{\tan\varphi}{N}v_{\text{east}} + \omega_{\oplus}\sin\varphi,$$

including an Earth rotation term: the stabilised platform is also a Foucault pendulum.

For the approximate or reference value it holds that

$$\frac{d}{dt}\alpha^{(0)} = \frac{\tan\varphi^{(0)}}{N(\varphi^{(0)})}v^{(0)}_{\text{east}} + \omega_{\oplus}\sin\varphi^{(0)}, \tag{5.29}$$

and the linearised equation is

$$\frac{d}{dt}\Delta\alpha = \frac{\tan\varphi}{N}\Delta v_{\text{east}} + \dot{\alpha}_{\varphi}\Delta\varphi, \qquad \dot{\alpha}_{\varphi} \approx \frac{v_{\text{east}}}{N\cos^2\varphi} + \omega_{\oplus}\cos\varphi.$$

This may suffice as an illustration of the complexity of mechanisation.

In practical application, only the approximate or reference state $\mathbf{x}^{(0)}(t)$ is integrated exactly. The linearised dynamic model 5.28 and the coefficient matrix F are used to propagate the state variance $\Sigma(t)$ using equation 3.19. And of course matrix F is used to bring the linearised state estimator $\Delta\mathbf{x}^{-}(t)$ forward. These deltas should always be kept small, similar in size to the uncertainties given by the state variance. Linearisation is approximation, and the above derived matrix F is not exact. sigma $\sigma\Sigma$

5.10 Initialisation of an inertial device

How one *levels* and *orients* an inertial measurement unit is also of interest. When not moving, the accelerometers of the inertial device act as clinometers and the feedback loops can be used to make the gyroscope axes turn to the horizontal plane.

The north direction can be obtained by using the device as a gyrocompass, by sensing how the local vector of gravity slowly turns around the south-north axis. This allows the initial azimuth $\alpha(t_0)$ to be determined.

Then, equation 5.29 is used to computationally track the heading $\alpha(t)$ during flight. This is called a *wandering-azimuth* solution.

Alternatively, the initial azimuth $\alpha(t_0)$ may be zeroed, and during flight the gyroscopes torqued to keep them, and the platform, pointing north and $\alpha(t)$ at zero. This is called *north-following*. The compensating torque needed is, analogously to equation 5.8, for example for the x gyroscope:

$$N_y = -L\frac{\frac{d}{dt}L_y}{L} = \\ = -J_x\omega\left(-\frac{d}{dt}\alpha\right) = J_x\omega\left(\frac{\tan\varphi}{N(\varphi)}v_{\mathrm{east}} + \omega_\oplus \sin\varphi\right). \quad (5.30)$$

Here is assumed that the gyroscope spin axis is the positive x axis and that the axis geometry is according to figure 5.2.

At airports, one often sees a tableau giving the precise ($\pm 0.1'$) geographical latitude and longitude of the gate. This is in fact used to initialise the co-ordinates in the inertial navigation platform used on an airliner. These gate co-ordinates are also found in *Jeppesen*. Levelling and orientation to the north of the inertial platform are performed while standing at the gate.

Self-test questions

1. What is a zero-velocity update?
2. Which are the two types of sensors commonly found in an inertial measurement unit? What do they measure?
3. In footnote 5 on page 107 the equations for the free nutation of the Earth are derived. If one substitutes values for J_0 and ΔJ from the literature, a period of around 306 sidereal days is obtained. We know however from observation that the period of the Earth's free nutation, or Chandler wobble, is about 433 days. What might cause the discrepancy?
4. Which two technical solutions exist for an inertial measurement unit?
5. How does a gyrocompass work?
6. How does a stabilised platform work? What is it used for?
7. What is a Schuler pendulum?
8. What is meant by the mechanisation of an inertial navigation device?
9. How, in an aircraft before take-off or a spacecraft before launch, is an inertial measurement unit initialised? Which two vectors are

used for this purpose? At which two points on the Earth's surface would this initialisation fail?

10. Why may it be problematic to use a "north-following" inertial platform on an aircraft flying over the pole?
11. What is the equation corresponding to equation 5.30 for the y gyroscope?

Exercise 5–1: Tennis-racket theorem

Prove the tennis-racket or intermediate-axis theorem using the Euler equations 5.6 in co-rotating co-ordinates for a freely spinning body.

Exercise 5–2: Gyrocompass equation

In the simple equation 5.17 for the gyrocompass, the right-hand side has a minus sign and is assumed to be negative. How would a gyrocompass behave if the right-hand side were positive? Do not use the approximation $\sin\alpha \approx \alpha$!

Exercise 5–3: Schuler period

Show that the Schuler period for a spherical celestial body of homogeneous density depends only on this density.

Navigation and orbital motion

6

Understanding satellite orbits and their geometry is fundamental to using the Kalman filter to navigate in space by means of observations made in real time. It is also essential in the context of terrestrial GNSS navigation, when calculating the locations of the satellites from the orbital elements, first in space and then in the local sky of the navigator. tosiaikaisesti

This is discussed more extensively in Hofmann-Wellenhof et al. (1997), chapter 4.

An understanding of the relative dynamics between two orbiting objects is also important. Here, the approach of Hill (1886) and the work of Clohessy and Wiltshire (1960) are relevant and will be presented.

6.1 The Kepler orbit

If it is assumed that the satellite moves in a central force field — for example the spherically symmetric gravitational field of a mass point or a spherical Earth — it follows that the orbit of the satellite is a *Kepler orbit*. Johannes Kepler (1571–1630) discovered the laws of planetary orbital motion bearing his name based on a masterly analysis of the extensive and uniquely precise naked-eye observations of the planet Mars by Tycho Brahe (1546–1601), see Physics Classroom, Kepler's Three Laws.

The motion of a satellite can be described with vector equations as follows:

$$\frac{d}{dt}\mathbf{x} = \mathbf{v}, \qquad \frac{d}{dt}\mathbf{v} = -\frac{GM_\oplus}{\|\mathbf{x}\|^3}\mathbf{x}.$$

Here, $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and $\mathbf{v} = \dot{x}\mathbf{i} + \dot{y}\mathbf{j} + \dot{z}\mathbf{k}$ are the location and velocity vectors of the satellite in three-dimensional space, in which an orthonormal basis $\beta \stackrel{\text{def}}{=} \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ is defined. The combined abstract ortonormaali kanta

beta βB

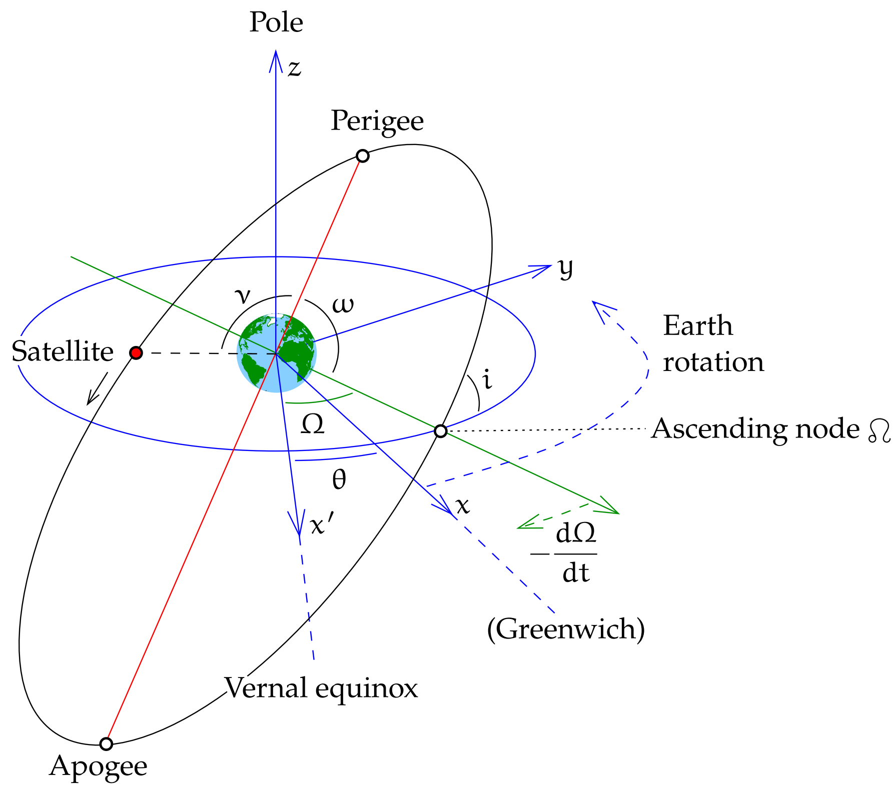


FIGURE 6.1. Kepler's orbital elements in space.

vector

$$\mathbf{x} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{x}_\beta & \mathbf{v}_\beta \end{bmatrix}^T = \begin{bmatrix} x & y & z \,|\, \dot{x} & \dot{y} & \dot{z} \end{bmatrix}^T$$

kanta on the agreed basis β is the *state vector* of the dynamic system formed by the orbiting satellite.

Kepler orbital elements are just a different way of writing the state vector, based on the knowledge that the orbit is a plane, eccentric ellipse

1 with the attracting body in one of the foci.[1] Wikipedia, Orbital elements gives a good description.

See figures 6.1 and 6.2 and tableau 6.1. The mean anomaly M is only a linear measure of elapsed time, scaled to the period P of the satellite and reckoned from the moment of its passage through the perigee t_P:

$$M(t) \stackrel{\text{def}}{=} 2\pi \frac{t - t_P}{P}. \tag{6.3}$$

nu νN E and ν again, the eccentric and the true anomaly, are purely geometric

[1] An ellipse has two focal points. It may be defined as the set of points for which the *sum* of the distances to these foci is constant. This also means that if one places a light source in one focus of an ellipse made of reflective material, an image of it will appear in the other focus. See Wikipedia, Dandelin spheres.

TABLEAU 6.1. Kepler's orbital elements.

Ω	right ascension or astronomical longitude of the ascending node. The ascending node is the point at which the satellite crosses the equator going from the southern to the northern hemisphere. The zero point of longitudes is the point on the celestial sphere where the ecliptic plane and the equatorial plane intersect, the vernal equinox: the place of the Sun at the start of spring, when it moves from the southern to the northern hemisphere.
i	inclination, the tilt angle of the orbital plane relative to the equatorial plane. The inclination of the orbital plane of GPS satellites is 55°.
ω	argument of perigee. The angular distance within the orbital plane between the ascending node and the perigee of the satellite orbital ellipse, the point at which the satellite in its orbit is closest to the Earth.
a	the semi-major axis of the satellite orbital ellipse.
e	the eccentricity of the satellite orbital ellipse. $1-e^2 = {}^{b^2}/_{a^2}$, in which b is the semi-minor axis.

ν, E, M describe the location of the satellite in its orbit as a function of time:

$\nu(t)$	true anomaly
$E(t)$	eccentric anomaly
$M(t)$	mean anomaly.

The connections between them are

$$E(t) = M(t) + e \sin E(t), \tag{6.1}$$

$$\frac{\tan \frac{1}{2}\nu(t)}{\tan \frac{1}{2}E(t)} = \sqrt{\frac{1+e}{1-e}}. \tag{6.2}$$

---

quantities.

In figure 6.1, the angle θ is the *Greenwich sidereal time*, which describes the globe's orientation relative to the stars. Greenwich sidereal time consists of a yearly and a daily component,[2] which are caused by the Earth's orbital and rotational motions, respectively. theta $\vartheta\theta\Theta$ 2

---

[2]Greenwich sidereal time is calculated to a precision of a few minutes as follows:

1. Take the month value from the following table:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
6 38	8 40	10 30	12 33	14 31	16 33	18 31	20 34	22 36	0 34	2 36	4 35

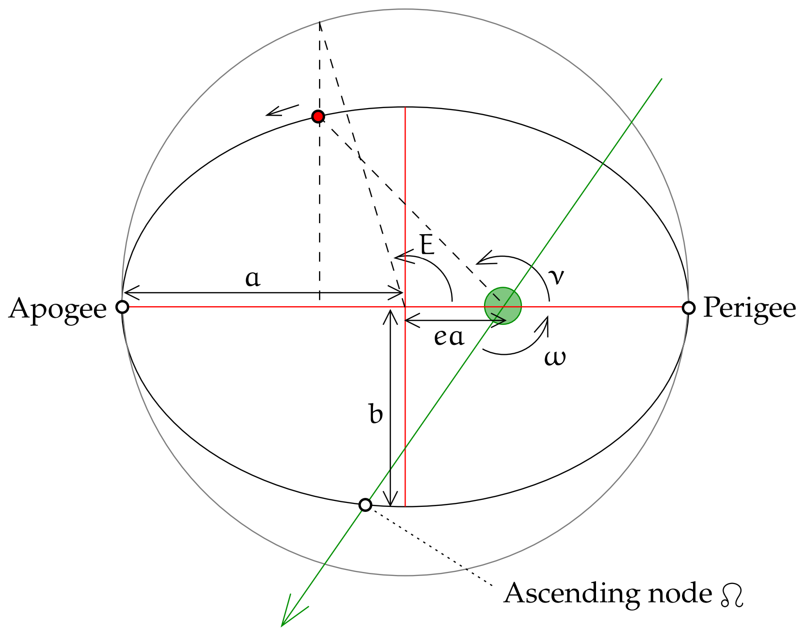


FIGURE 6.2. Kepler's orbital elements in the plane.

So we have obtained an alternative way of presenting the state vector:

$$\mathbf{a} = \begin{bmatrix} a & e & M & i & \omega & \Omega \end{bmatrix}^{\mathsf{T}}.$$

In a central force field the elements of this state vector are constants, except the mean anomaly $M(t)$, equation 6.3. If the force field is not precisely central, the other orbital elements may also change slowly with time. For example, the flattening of the Earth causes a slow turning of the ascending node, making $\Omega(t)$ time-dependent. Such time-dependent Kepler elements, like $\Omega(t)$ and $\omega(t)$ for a flattened Earth, are called *osculating elements*.[3]

omega $\omega\Omega$

3

6.1.1 The radius vector

See figure 6.3 of the orbital plane. To express the radius vector r in terms of the eccentric anomaly E, we first observe that the axes ratio of

2. Add to this four minutes for every day of the month.
3. Add to this the clock time (UTC, or Greenwich mean time).

For computing the *local* sidereal time, add to this the local east longitude converted to time units: $15° = 1^h$, $1° = 4^m$, $15' = 1^m$.
Precision will be $\pm 3^m$ because the table is not the same from year to year: it varies with the leap-year cycle.

[3]from Latin *osculatio*, kiss.

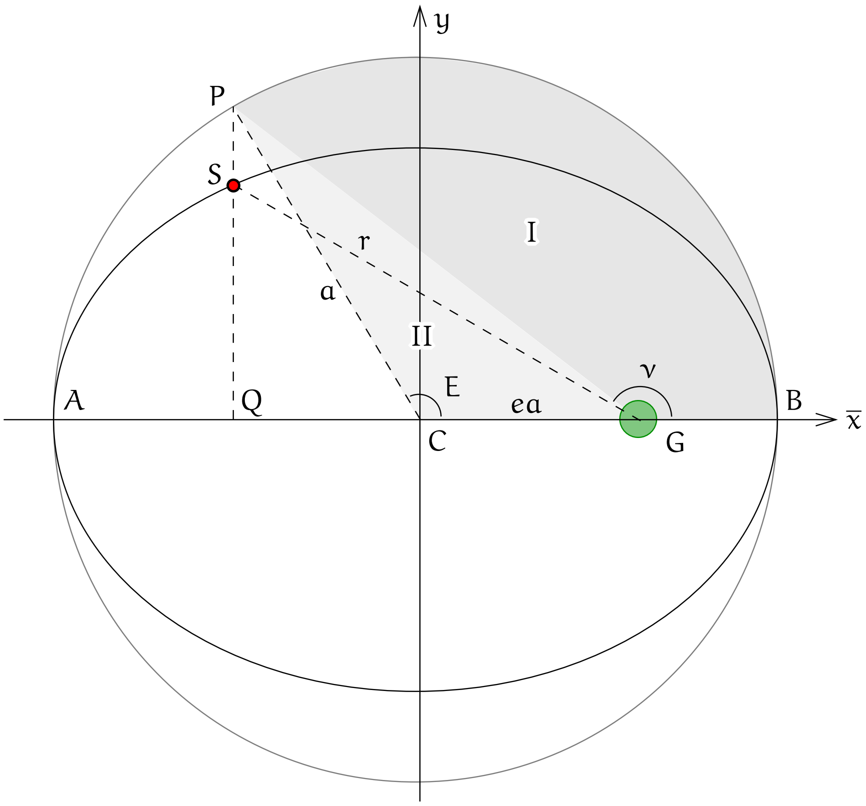


FIGURE 6.3. An eccentric orbit.

the orbital ellipse is

$$\frac{b}{a} = \sqrt{1 - e^2} = \frac{SQ}{PQ}.$$

Then

$$\begin{aligned} r = \sqrt{SQ^2 + QG^2} &= \sqrt{(1 - e^2)\,PQ^2 + (QC + CG)^2} = \\ &= \sqrt{(1 - e^2)\,a^2 \sin^2 E + (-a\cos E + ea)^2} = \\ &= \sqrt{(1 - e^2)\,a^2\,(1 - \cos^2 E) + a^2 \cos^2 E - 2ea^2 \cos E + e^2 a^2} = \\ &= a\sqrt{1 \cancel{- \cos^2 E} \cancel{- e^2} + e^2 \cos^2 E \cancel{+ \cos^2 E} - 2e\cos E \cancel{+ e^2}} = \\ &= a\sqrt{(1 - e\cos E)^2} = a\,(1 - e\cos E)\,. \qquad (6.4) \end{aligned}$$

To derive the radius vector r expressed in the true anomaly ν we start from the equation of the ellipse in rectangular co-ordinates centred on the centre C of the ellipse:

$$\frac{\overline{x}^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Expressing the co-ordinates $\overline{x}$ and y in ν gives

$$\frac{(r\cos\nu + ea)^2}{a^2} + \frac{r^2 \sin^2 \nu}{a^2\,(1 - e^2)} = 1$$

or

$$\left(1-e^2\right)\left(r\cos\nu + ea\right)^2 + r^2\sin^2\nu = a^2\left(1-e^2\right).$$

Reorganising terms yields

$$\left(\left(1-e^2\right)\cos^2\nu + \left(1-\cos^2\nu\right)\right)r^2 + \left(2\left(1-e^2\right)ea\cos\nu\right)r + \\ + \left(\left(1-e^2\right)e^2a^2 - a^2\left(1-e^2\right)\right) = 0$$

or

$$\left(1-e^2\cos^2\nu\right)r^2 + \left(2\left(1-e^2\right)ea\cos\nu\right)r + \left(-a^2\left(1-e^2\right)^2\right) = 0,$$

a quadratic equation, which has the standard solution. The discriminant Δ is

delta δΔ

$$\Delta = \left(2\left(1-e^2\right)ea\cos\nu\right)^2 - 4\left(1-e^2\cos^2\nu\right)\left(-a^2\left(1-e^2\right)^2\right) = \\ = 4a^2\left(1-e^2\right)^2\left(\left(e\cos\nu\right)^2 + \left(1-e^2\cos^2\nu\right)\right) = 4a^2\left(1-e^2\right)^2.$$

The solution is

$$r_{1,2} = \frac{-2\left(1-e^2\right)ea\cos\nu \pm \sqrt{\Delta}}{2\left(1-e^2\cos^2\nu\right)} = \\ = \frac{-2a\left(1-e^2\right)e\cos\nu \pm 2a\left(1-e^2\right)}{2\left(1-e^2\cos^2\nu\right)} = \frac{a\left(1-e^2\right)\left(\pm 1 - e\cos\nu\right)}{1-e^2\cos^2\nu}.$$

Of these, we take the positive solution:

$$r = \frac{a\left(1-e^2\right)\left(1-e\cos\nu\right)}{1-e^2\cos^2\nu} = \frac{a\left(1-e^2\right)}{1+e\cos\nu}.$$

We can thus calculate the satellite's instantaneous radius

$$r(t) = a\left(1 - e\cos E(t)\right) = \frac{a\left(1-e^2\right)}{1+e\cos\nu(t)}.$$

6.1.2 *Conversion between mean and eccentric anomalies*

For this, we will use Kepler's second law, the law of areas. It says that, per unit of time, the radius vector of the satellite will always sweep over the same surface area. It is a special case of the conservation of angular momentum.

pyörähdys-momentti

See figure 6.3 again. The law of areas applies not only to the motion of the satellite S itself, but also to the motion of its corresponding point P. All areas are just a/b times larger. This means that the area of the sector PGB, marked I in the figure, must be a fraction $M/2\pi$ of that of the full circle. The area of the full circle is πa^2. Thus

$$I = \frac{M}{2\pi}\pi a^2 = \tfrac{1}{2}a^2 M. \tag{6.5}$$

Similarly the sector PCB, comprising the areas I and II together, is, qua surface area, geometrically, a fraction $E/2\pi$ of the circle. Thus

$$I + II = \frac{E}{2\pi}\pi a^2 = \tfrac{1}{2}a^2 E. \tag{6.6}$$

Finally, the area of triangle PCG is

$$II = \tfrac{1}{2}CG \cdot PQ = \tfrac{1}{2}ae \cdot a \sin E = \tfrac{1}{2}a^2 e \sin E. \tag{6.7}$$

The three results 6.5, 6.6 and 6.7 together produce the *Kepler equation* 6.1:

$$E = M + e \sin E,$$

from which E can be solved iteratively if M is given.

6.1.3 Conversion between eccentric and true anomalies

For the x co-ordinate of the satellite it holds that

$$x = r\cos\nu = a\cos E - ea \implies r\cos\nu = a\left(\cos E - e\right).$$

The half-angle substitution

$$\cos\nu = 2\cos^2\tfrac{1}{2}\nu - 1$$

yields

$$2r\cos^2\tfrac{1}{2}\nu - r = a\left(\cos E - e\right)$$

or, with equation 6.4,

$$2r\cos^2\tfrac{1}{2}\nu - a\left(1 - e\cos E\right) = a\left(\cos E - e\right)$$

or

$$2r\cos^2\tfrac{1}{2}\nu = a\big(\left(\cos E - e\right) + \left(1 - e\cos E\right)\big) = a\left(1 - e\right)\left(\cos E + 1\right).$$

With another half-angle substitution

$$\cos E = 2\cos^2\tfrac{1}{2}E - 1$$

this yields

$$2r\cos^2\tfrac{1}{2}\nu = 2a\left(1 - e\right)\cos^2\tfrac{1}{2}E. \tag{6.8}$$

Similarly, using the relations

$$\cos\nu = 1 - 2\sin^2\tfrac{1}{2}\nu, \qquad \cos E = 1 - 2\sin^2\tfrac{1}{2}E$$

yields

$$\begin{aligned}
r - 2r\sin^2\tfrac{1}{2}\nu &= a\,(\cos E - e) \\
&\implies a\,(1 - e\cos E) - 2r\sin^2\tfrac{1}{2}\nu = a\,(\cos E - e) \\
&\implies -2r\sin^2\tfrac{1}{2}\nu = a\big((\cos E - e) - (1 - e\cos E)\big) = \\
&= a\,(1+e)\,(\cos E - 1) = -2a\,(1+e)\sin^2\tfrac{1}{2}E \\
&\implies 2r\sin^2\tfrac{1}{2}\nu = 2a\,(1+e)\sin^2\tfrac{1}{2}E. \qquad (6.9)
\end{aligned}$$

Division of equation 6.9 by equation 6.8 yields

$$\tan^2\tfrac{1}{2}\nu = \frac{1+e}{1-e}\tan^2\tfrac{1}{2}E \implies \frac{\tan\tfrac{1}{2}\nu}{\tan\tfrac{1}{2}E} = \sqrt{\frac{1+e}{1-e}},$$

result 6.2.

6.1.4 *Rectangular co-ordinates and time derivatives*

The time derivative of r is obtained by differentiating equation 6.4:

$$\frac{dr}{dt} = ae\sin E\frac{dE}{dt}.$$

The Kepler equation 6.1 yields

$$\frac{dE}{dt} = \frac{dM}{dt} + e\cos E\frac{dE}{dt} = \frac{2\pi}{P} + e\cos E\frac{dE}{dt},$$

so

$$\frac{dE}{dt} = \frac{2\pi}{P\,(1 - e\cos E)},$$

in which P is the orbital period. In the orbital plane

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r\cos\nu \\ r\sin\nu \end{bmatrix} = \begin{bmatrix} a\,(\cos E - e) \\ b\sin E \end{bmatrix}.$$

Differentiating this with respect to time yields

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -a\sin E \\ b\cos E \end{bmatrix}\frac{dE}{dt} = \frac{2\pi}{P\,(1 - e\cos E)}\begin{bmatrix} -a\sin E \\ b\cos E \end{bmatrix}. \qquad (6.10)$$

6.1.5 Transformation to the geocentric frame

We can transform the obtained two-dimensional vectors in the orbital plane into three-dimensional vectors in space by using the rotation angles ω, i, and Ω. If we write on the basis α of the orbital plane: alpha αA

$$\mathbf{x}_\alpha \stackrel{\text{def}}{=} \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} r\cos\nu \\ r\sin\nu \\ 0 \end{bmatrix} = \begin{bmatrix} a\,(\cos E - e) \\ b\sin E \\ 0 \end{bmatrix}, \tag{6.11}$$

we get on a geocentric basis β:

$$\mathbf{x}_\beta \stackrel{\text{def}}{=} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R(\Omega, i, \omega)\,\mathbf{x}_\alpha,$$

in which the rotation matrix

$$R(\Omega, i, \omega) = R(\Omega)\,R(i)\,R(\omega), \qquad R(i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{bmatrix},$$
$$R(\Omega) = \begin{bmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R(\omega) = \begin{bmatrix} \cos\omega & -\sin\omega & 0 \\ \sin\omega & \cos\omega & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{6.12}$$

To summarise:

$$R(\Omega, i, \omega) =$$
$$= \left[\begin{array}{c|c|c} \cos\Omega\cos\omega - \sin\Omega\sin\omega\cos i & -\cos\Omega\sin\omega - \sin\Omega\cos\omega\cos i & \sin\Omega\sin i \\ \sin\Omega\cos\omega + \cos\Omega\sin\omega\cos i & -\sin\Omega\sin\omega + \cos\Omega\cos\omega\cos i & -\cos\Omega\sin i \\ \sin\omega\sin i & \cos\omega\sin i & \cos i \end{array}\right]. \tag{6.13}$$

See figure 6.4.

The geocentric co-ordinates thus obtained are in an inertial, also called astronomical, frame. The origin of the longitudes is the direction to the vernal equinox. If we want to obtain the co-ordinates of the satellite in a frame co-rotating with the Earth, with Greenwich as the origin of longitudes, we must calculate the geographical longitude of the ascending node

$$\ell = \Omega - \theta, \tag{6.14}$$

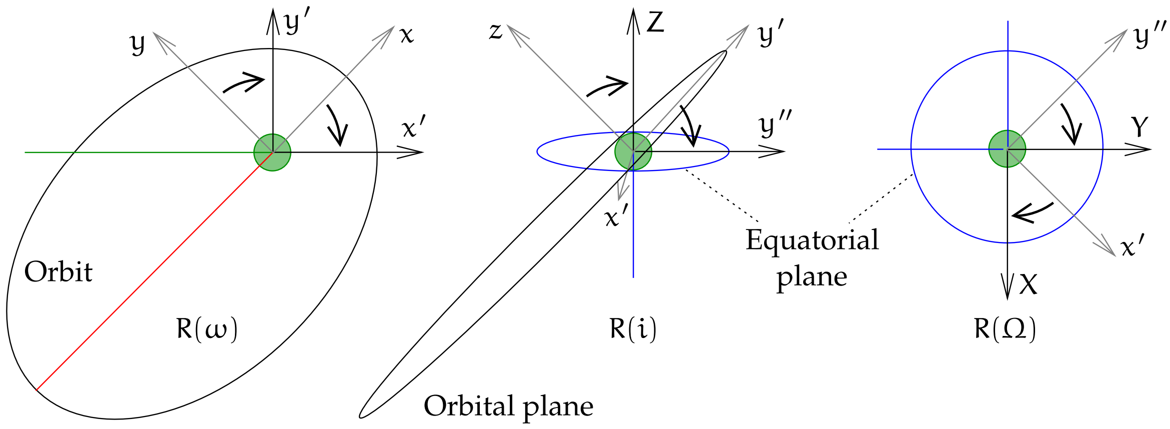


FIGURE 6.4. Deriving the R matrix.

in which θ is Greenwich sidereal time. Now, in matrix equation 6.13 replace the right ascension Ω of the ascending node with the longitude ℓ, and call the resulting matrix $R'(\ell, i, \omega)$.

The velocity vector is obtained by differentiating with respect to time, equation 6.10:

$$\frac{d}{dt}\mathbf{x}_\alpha = \frac{2\pi}{P\,(1 - e\cos E)} \begin{bmatrix} -a\sin E \\ b\cos E \\ 0 \end{bmatrix}, \tag{6.15}$$

from which the geocentric equivalent follows:

$$\frac{d}{dt}\mathbf{x}_\beta = \frac{d}{dt}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R\frac{d}{dt}\mathbf{x}_\alpha.$$

This is valid in inertial co-ordinates. In a frame co-rotating with the Earth, basis β', we find

$$\frac{d}{dt}\mathbf{x}_{\beta'} = R'\frac{d}{dt}\mathbf{x}_\alpha + \left(\frac{d}{dt}R'\right)\mathbf{x}_\alpha,$$

where now $R'(\ell, i, \omega) = R'(\Omega - \theta, i, \omega)$ contains the sidereal time θ, equation 6.14, and is thus time-dependent:

$$\frac{d}{dt}R' = \frac{d\ell}{dt}\frac{d}{d\ell}R' = -\frac{d\theta}{dt}\frac{d}{d\ell}R' = -\omega_\oplus\frac{d}{d\ell}R' = -\omega_\oplus \cdot$$
$$\cdot \left[\begin{array}{c|c|c} -\sin\ell\cos\omega - \cos\ell\sin\omega\cos i & \sin\ell\sin\omega - \cos\ell\cos\omega\cos i & \cos\ell\sin i \\ \cos\ell\cos\omega - \sin\ell\sin\omega\cos i & -\cos\ell\sin\omega - \sin\ell\cos\omega\cos i & \sin\ell\sin i \\ 0 & 0 & 0 \end{array}\right],$$

with $\omega_\oplus$ the rotation rate of the Earth.

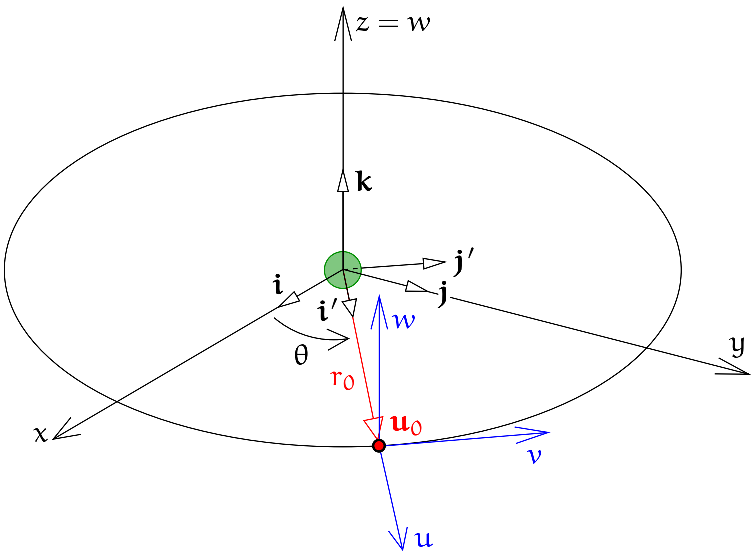
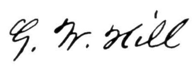


FIGURE 6.5. The Hill co-ordinate frame.

6.2 Use of Hill co-ordinates



The Hill co-ordinate frame was invented by George W. Hill[4] in connection with the study of the motion of the Moon. It replaces the standard way of describing orbital motion in an inertial co-ordinate frame (x, y, z) centred on the centre of motion, like the Sun. Instead, a non-inertial, co-moving and co-rotating frame (u, v, w) is used, the origin of which is centred on the Earth[5] and which rotates uniformly at the same mean rate as the Earth around the Sun, one rotation per year. As the distance of the Moon from the Earth is only $0.26\,\%$ of that between the Earth and the Sun, the mathematics of the solar influence can be effectively linearised.

A modification of the method models the motion of an Earth satellite relative to a fictitious point orbiting the Earth in a circular orbit with the same period as the satellite. Of course, an important simplification compared to the original application is that the fictitious point, unlike the Earth, has no attraction of its own. This approach has been fruitful for studying orbital perturbations and the rendezvous problem. ratahäiriö

[4]George William Hill (1838–1914) was an American astronomer and mathematician who studied the three-body problem. Lunar motion is a classical three-body problem in which the effects of the Earth and Sun are of similar magnitude.
Hill was a person who valued solitude. He preferred to work alone and did his ground-breaking work at his family farm in West Nyack, New York state.

[5]More precisely: on the common centre of mass of the Earth and Moon.

Write according to figure 6.5:

$$\mathbf{x} = \mathbf{u} + \mathbf{u}_0. \tag{6.16}$$

Expand the vectors into components:

$$\begin{aligned}
\mathbf{x} &= x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = (x\cos\theta + y\sin\theta)\,\mathbf{i}' + (y\cos\theta - x\sin\theta)\,\mathbf{j}' + z\mathbf{k},\\
\mathbf{u} &= u\mathbf{i}' + v\mathbf{j}' + w\mathbf{k},\\
\mathbf{u}_0 &= r_0\mathbf{i}'.
\end{aligned}$$

Form abstract component vectors on the orthonormal basis $\beta \stackrel{\text{def}}{=} \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$:

$$\mathbf{x}_\beta = \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

and on the orthonormal basis $\beta' \stackrel{\text{def}}{=} \{\mathbf{i}', \mathbf{j}', \mathbf{k}\}$:

$$\mathbf{x}_{\beta'} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \mathbf{u}_{\beta'} = \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \quad \mathbf{u}_{0,\beta'} = \begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix}.$$

Equation 6.16 becomes

$$R\mathbf{x}_\beta = \mathbf{x}_{\beta'} = \mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}, \tag{6.17}$$

in which matrix R represents a rotation of the co-ordinate frame around the z axis by an angular amount θ:

$$R = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Vector $\mathbf{x}_\beta$ consists of components in the inertial frame and vector $\mathbf{u}_{\beta'}$ of components in the frame co-rotating with the satellite. The $\mathbf{i}'$ or u axis points outwards ("upwards"), the $\mathbf{j}'$ or v axis forwards in the direction of flight, and the $\mathbf{k}$ or w or z axis points perpendicularly out of the orbital plane to "port". The origin of the (u, v, w) frame moves at constant velocity in a circular orbit: the angular velocity according to Kepler's third law is

$$n = \frac{d\theta}{dt} = \sqrt{\frac{GM_\oplus}{r_0^3}}.$$

r_0 is the orbital radius and also the distance of the origin of the (u, v, w) frame from that of the (x, y, z) frame.

Equation 6.17 can be inverted:

$$\mathbf{x}_\beta = R^{-1}\left(\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\right) = R^T\left(\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\right),$$

because for an orthogonal matrix $RR^T = I \iff R^{-1} = R^T$.

6.3 Transformation from inertial frame to Hill frame

We derive equations for the first and second derivatives with respect to time of vector $\mathbf{x}$ and matrix R and substitute. After that, we multiply both sides of the equation with matrix R.

We obtain by differentiation with the Leibniz product rule:

$$\begin{aligned}
\dot{\mathbf{x}}_\beta &= \dot{R}^T\left(\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\right) + R^T\dot{\mathbf{u}}_{\beta'},\\
\ddot{\mathbf{x}}_\beta &= \ddot{R}^T\left(\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\right) + 2\dot{R}^T\dot{\mathbf{u}}_{\beta'} + R^T\ddot{\mathbf{u}}_{\beta'}.
\end{aligned}$$

The derivatives of rotation matrix R are obtained using the chain rule:

$$\dot{R} = \frac{dR}{dt} = \frac{dR}{d\theta}\frac{d\theta}{dt} = \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n,$$

$$\ddot{R} = \frac{d^2R}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 = \begin{bmatrix} -\cos\theta & -\sin\theta & 0 \\ \sin\theta & -\cos\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n^2.$$

Substitution yields

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = n^2 \overbrace{\begin{bmatrix} -\cos\theta & \sin\theta & 0 \\ -\sin\theta & -\cos\theta & 0 \\ 0 & 0 & 0 \end{bmatrix}}^{\ddot{R}^T} \begin{bmatrix} u + r_0 \\ v \\ w \end{bmatrix} +$$

$$+ 2n \overbrace{\begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix}}^{\dot{R}^T} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} +$$

$$+ \overbrace{\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}}^{R^T} \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

By multiplying from the left with matrix R we obtain

$$\begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = n^2 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u + r_0 \\ v \\ w \end{bmatrix} + \\ + 2n \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}. \quad (6.18)$$

6.4 Series expansion for a central force field

The equation for a central force field in the (x, y, z) frame is

$$\ddot{\mathbf{x}} = -\frac{GM_\oplus}{\|\mathbf{x}\|^3}\mathbf{x} \implies R\ddot{\mathbf{x}}_\beta = -\frac{GM_\oplus}{\|\mathbf{x}\|^3} R\mathbf{x}_\beta = -\frac{GM_\oplus}{\|\mathbf{x}\|^3}\left(\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\right),$$

in components:

$$\begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -\frac{GM_\oplus}{\|\mathbf{x}\|^3} \begin{bmatrix} u + r_0 \\ v \\ w \end{bmatrix},$$

in which

$$\|\mathbf{x}\| = \|\mathbf{u} + \mathbf{u}_0\| = \|\mathbf{u}_{\beta'} + \mathbf{u}_{0,\beta'}\| = \sqrt{(u + r_0)^2 + v^2 + w^2}.$$

The Taylor expansion about the origin of the (u, v, w) frame yields

$$\begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -\frac{GM_\oplus}{r_0^3} \begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + \mathcal{M}^{(0)}_{\beta'} \cdot \begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

where the *gravitation-gradient* tensor matrix $\mathcal{M}^{(0)}_{\beta'}$ consists of the partial derivatives

$$\mathcal{M}^{(0)}_{\beta'} = \begin{bmatrix} \frac{\partial}{\partial u} & \frac{\partial}{\partial v} & \frac{\partial}{\partial w} \end{bmatrix} \left(-\frac{GM_\oplus}{\|\mathbf{u} + \mathbf{u}_0\|^3} \begin{bmatrix} u + r_0 \\ v \\ w \end{bmatrix} \right)\Bigg|_{u,v,w=0} = \\ = -\frac{GM_\oplus}{r_0^3} \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

equation 3.9 evaluated in the point $\mathbf{u}_0$.

According to Kepler's third law

$$\frac{GM_\oplus}{r_0^3} = n^2.$$

By combining we obtain

$$\begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \\ = -n^2 \left(\begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right). \quad (6.19)$$

6.5 Equations of motion in the Hill frame

By combining equations 6.18 and 6.19 we obtain

$$0 = n^2 \left(\begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right) + \\ + n^2 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u + r_0 \\ v \\ w \end{bmatrix} + \\ + 2n \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

Simplification gives

$$0 = n^2 \begin{bmatrix} -3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} + 2n \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

As the end result, by separately extracting the equations for the components u, v and w:[6]

$$\ddot{u} = 2n\dot{v} + 3n^2 u, \quad \ddot{v} = -2n\dot{u}, \quad \ddot{w} = -n^2 w,$$

in which the third is recognised as a classical harmonic oscillator.

[6]We can spot here the pseudo-forces occurring in a rotating co-ordinate frame, in particular the Coriolis terms $2n\dot{v}$ and $-2n\dot{u}$, which are velocity dependent. The centrifugal force is more hidden, n^2u. The remainder, $2n^2u$ and $-n^2w$, are gravitational gradients of the central field.

6.6 Solving the Hill equations

6.6.1 The w equation

We first solve the easiest equation, the third one,

$$\ddot{w} = -n^2 w.$$

Let us first try the general periodic solution,

$$w(t) = A\sin(Bt + C).$$

Twice differentiation and substitution yields

$$-AB^2 \sin(Bt + C) = -n^2 A\sin(Bt + C) \implies B = \pm n.$$

Thus the solution is

$$w(t) = A\sin(\pm nt + C),$$

in which A and C are arbitrary constants. The sine sum formula

$$\sin(\pm nt + C) = \sin(\pm nt)\cos C + \cos(\pm nt)\sin C$$

yields

$$w(t) = A_1 \sin nt + A_2 \cos nt,$$

in which $A_1 = \pm A\cos C$ and $A_2 = A\sin C$ are again arbitrary constants.

The velocity is obtained by differentiation:

$$\dot{w}(t) = nA_1 \cos nt - nA_2 \sin nt.$$

State transition from time t_0 to time $t_1 = t_0 + \Delta t$:

$$\begin{aligned}
w(t_1) = A_1 \sin n(t_0 + \Delta t) + A_2 \cos n(t_0 + \Delta t) &= \\
= A_1 (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) &+ \\
+ A_2 (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) &= \\
= (A_1 \sin nt_0 + A_2 \cos nt_0) \cos n\Delta t &+ \\
+ (A_1 \cos nt_0 - A_2 \sin nt_0) \sin n\Delta t &= \\
= w(t_0)\cos n\Delta t + \frac{1}{n}\dot{w}(t_0)\sin n\Delta t&, \\
\dot{w}(t_1) = nA_1 \cos n(t_0 + \Delta t) - nA_2 \sin n(t_0 + \Delta t) &= \\
= nA_1 (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) &- \\
- nA_2 (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) &= \\
= n(A_1 \cos nt_0 - A_2 \sin nt_0)\cos n\Delta t &- \\
- n(A_1 \sin nt_0 + A_2 \cos nt_0)\sin n\Delta t &= \\
= \dot{w}(t_0)\cos n\Delta t - nw(t_0)\sin n\Delta t&.
\end{aligned}$$

As a matric equation:

$$\overbrace{\begin{bmatrix} w \\ \dot{w} \end{bmatrix}(t_1)}^{\mathbf{x}(t_1)} = \overbrace{\begin{bmatrix} \cos n\Delta t & (\sin n\Delta t)/n \\ -n\sin n\Delta t & \cos n\Delta t \end{bmatrix}}^{\Phi_0^1} = \overbrace{\begin{bmatrix} w \\ \dot{w} \end{bmatrix}(t_0)}^{\mathbf{x}(t_0)}. \tag{6.20}$$

6.6.2 The u and v equations

$$\ddot{u} = 2n\dot{v} + 3n^2 u, \qquad \ddot{v} = -2n\dot{u}. \tag{6.21}$$

These are to be solved together. Let us try again a periodic solution:

$$u(t) = A\sin nt + B\cos nt, \qquad v(t) = C\sin nt + D\cos nt. \tag{6.22}$$

Substitution yields

$$\begin{aligned} &-n^2(A\sin nt + B\cos nt) = \\ &\quad = 2n\cdot n(C\cos nt - D\sin nt) + 3n^2(A\sin nt + B\cos nt), \\ &-n^2(C\sin nt + D\cos nt) = -2n\cdot n(A\cos nt - B\sin nt). \end{aligned}$$

Consider the sine and cosine terms separately and express C and D into A and B:

$$\begin{aligned} -n^2A &= -2n^2D + 3n^2A, & -n^2B &= 2n^2C + 3n^2B, \\ -n^2C &= 2n^2B, & -n^2D &= -2n^2A, \end{aligned}$$

or

$$\begin{aligned} -A &= -2D + 3A, & -B &= 2C + 3B, \\ -C &= 2B \implies C = -2B, & -D &= -2A \implies D = 2A. \end{aligned}$$

Substitution into equations 6.22 yields the general solution

$$u(t) = A\sin nt + B\cos nt, \qquad v(t) = -2B\sin nt + 2A\cos nt.$$

As a matric equation:

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} A & B \\ -2B & 2A \end{bmatrix} \begin{bmatrix} \sin nt \\ \cos nt \end{bmatrix}.$$

This solution is called a *libration movement*, figure 6.6. It is a periodic movement, the centre of which is the origin of the Hill frame $u = v = 0$.

In the inertial frame, the satellite describes an elliptical Kepler orbit around the origin $x = y = 0$. The period of the Hill solution is $2\pi/n$, the same as that of the Kepler orbit.

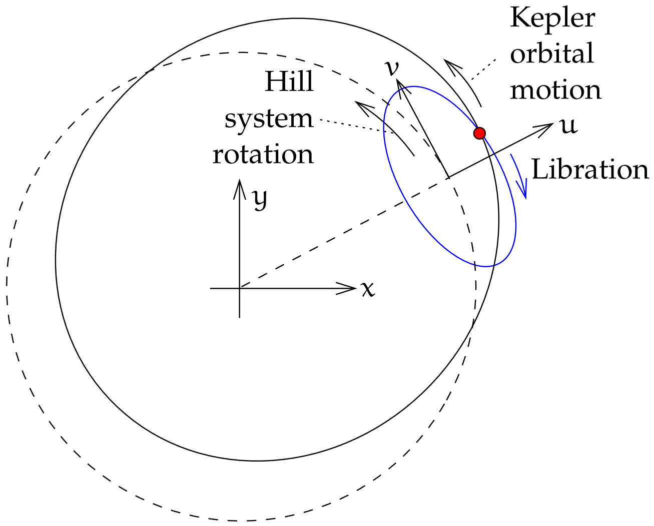


FIGURE 6.6. Libration.

6.7 Another solution

This is not however the end of the story. For a change, let us try a *linear* non-periodic solution:

$$u(t) = Et + F, \qquad v(t) = Gt + H.$$

Substitute this into the original differential equations 6.21 and express E and G into F and H:

$$0 = 2nG + 3n^2 (Et + F), \qquad 0 = -2nE,$$

from which

$$E = 0, \qquad G = -\tfrac{3}{2}nF.$$

The solution is

$$u(t) = F, \qquad v(t) = -\tfrac{3}{2}Fnt + H,$$

in which F and H are arbitrary constants. This represents an *orbital motion with a period different from* $^{2\pi}/_{n}$. The orbital radius is $r_0 + F$, the orbit's angular velocity $n - \frac{3}{2}Fn$ (Kepler's third law!) and the satellite is at the moment $t = 0$ in its orbit ahead of the origin of the (u, v, w) frame by a distance H. See figure 6.7.

Because the system of differential equations is linear, we may freely combine the above periodic and linear solutions.

6.8 The Clohessy-Wiltshire model

We derive the state transition matrix.

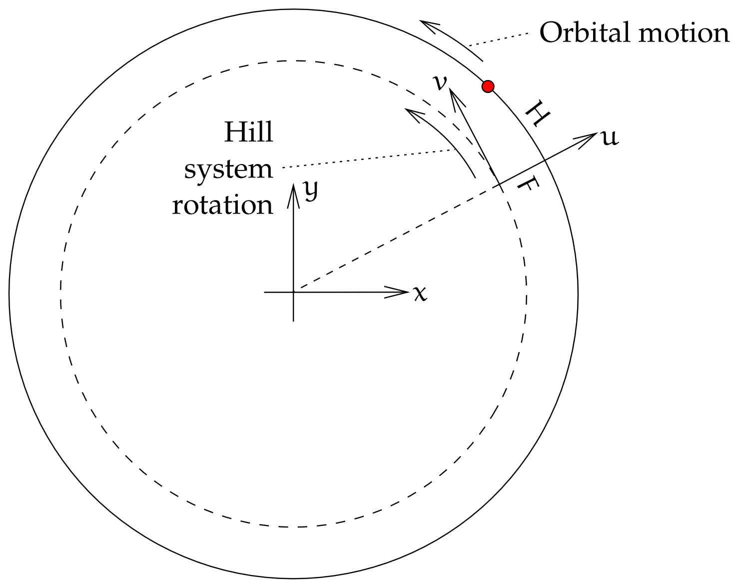


FIGURE 6.7. Linear drift for a satellite with a period different from that of the Hill frame.

6.8.1 The general case

Let us look only at the (u, v) plane. The general solution is

$$\begin{aligned} u(t) &= A \sin nt + B \cos nt + F, \\ v(t) &= -2B \sin nt + 2A \cos nt - \tfrac{3}{2} Fnt + H. \end{aligned}$$

The velocity components are obtained by differentiation:

$$\begin{aligned} \dot{u}(t) &= nA \cos nt - nB \sin nt, \\ \dot{v}(t) &= -2nA \sin nt - 2nB \cos nt - \tfrac{3}{2} Fn. \end{aligned}$$

Write for the initial epoch t_0:

$$u(t_0) = A \sin nt_0 + B \cos nt_0 + F = \mathcal{S}(t_0) + F, \tag{6.23}$$

$$\begin{aligned} v(t_0) &= -2B \sin nt_0 + 2A \cos nt_0 - \tfrac{3}{2} Fnt_0 + H = \\ &= 2\mathcal{C}(t_0) - \tfrac{3}{2} Fnt_0 + H, \end{aligned}$$

$$\dot{u}(t_0) = nA \cos nt_0 - nB \sin nt_0 = n\mathcal{C}(t_0), \tag{6.24}$$

$$\dot{v}(t_0) = -2nA \sin nt_0 - 2nB \cos nt_0 - \tfrac{3}{2} Fn = -2n\mathcal{S}(t_0) - \tfrac{3}{2} Fn. \tag{6.25}$$

The definitions used here are

$$\mathcal{S}(t) \stackrel{\text{def}}{=} A \sin nt + B \cos nt, \qquad \mathcal{C}(t) \stackrel{\text{def}}{=} A \cos nt - B \sin nt.$$

Write for the epoch t_1, applying the sum formulas for sine and cosine:

$$\begin{aligned} u(t_1) = u(t_0 + \Delta t) = & \\ &= A \sin n(t_0 + \Delta t) + B \cos n(t_0 + \Delta t) + F = \\ &= A \sin nt_0 \cos n\Delta t + A \cos nt_0 \sin n\Delta t + \\ &+ B \cos nt_0 \cos n\Delta t - B \sin nt_0 \sin n\Delta t + F = \\ &= \cos n\Delta t\, \mathcal{S}(t_0) + \sin n\Delta t\, \mathcal{C}(t_0) + F = \\ &= u(t_0) + (\cos n\Delta t - 1)\, \mathcal{S}(t_0) + \sin n\Delta t\, \mathcal{C}(t_0). \end{aligned}$$

Similarly

$$\begin{aligned} v(t_1) &= 2A \cos n(t_0 + \Delta t) - 2B \sin n(t_0 + \Delta t) - \tfrac{3}{2}Fnt_1 + H = \\ &= 2A \cos nt_0 \cos n\Delta t - 2A \sin nt_0 \sin n\Delta t - \\ &- 2B \sin nt_0 \cos n\Delta t - 2B \cos nt_0 \sin n\Delta t - \tfrac{3}{2}Fnt_1 + H = \\ &= \cos n\Delta t\, (2A \cos nt_0 - 2B \sin nt_0) - \\ &- \sin n\Delta t\, (2A \sin nt_0 + 2B \cos nt_0) - \tfrac{3}{2}Fnt_1 + H = \\ &= v(t_0) + 2\, (\cos n\Delta t - 1)\, \mathcal{C}(t_0) - 2 \sin n\Delta t\, \mathcal{S}(t_0) - \tfrac{3}{2}Fn\, \Delta t. \end{aligned}$$

Substitute equation 6.23, $F = u(t_0) - \mathcal{S}(t_0)$, into this, obtaining

$$\begin{aligned} v(t_1) = v(t_0) + 2\, (\cos n\Delta t - 1)\, \mathcal{C}(t_0) - 2 \sin n\Delta t\, \mathcal{S}(t_0) - & \\ - \tfrac{3}{2}u(t_0)\, n\, \Delta t + \tfrac{3}{2}\mathcal{S}(t_0)\, n\, \Delta t &= \\ = v(t_0) + 2\, (\cos n\Delta t - 1)\, \mathcal{C}(t_0) &+ \\ + \left(\tfrac{3}{2}n\, \Delta t - 2 \sin n\Delta t\right) \mathcal{S}(t_0) - \tfrac{3}{2}u(t_0)\, n\, \Delta t&. \end{aligned}$$

Next, the velocities:

$$\begin{aligned} \dot{u}(t_1) &= nA\, (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) - \\ &- nB\, (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) = \\ &= \cos n\Delta t\, (nA \cos nt_0 - nB \sin nt_0) - \\ &- \sin n\Delta t\, (nA \sin nt_0 + nB \cos nt_0) = \\ &= \cos n\Delta t\, n\mathcal{C}(t_0) - \sin n\Delta t\, n\mathcal{S}(t_0) \end{aligned}$$

and

$$\begin{aligned} \dot{v}(t_1) &= -2nA\, (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) - \\ &- 2nB\, (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) - \tfrac{3}{2}Fn = \\ &= \cos n\Delta t\, (-2nA \sin nt_0 - 2nB \cos nt_0) + \\ &+ \sin n\Delta t\, (-2nA \cos nt_0 + 2nB \sin nt_0) - \tfrac{3}{2}Fn = \\ &= \dot{v}(t_0) - (\cos n\Delta t - 1)\, 2n\mathcal{S}(t_0) - 2 \sin n\Delta t\, n\mathcal{C}(t_0). \end{aligned}$$

Introduce the notations

$$s \stackrel{\text{def}}{=} \sin n\Delta t, \qquad c \stackrel{\text{def}}{=} \cos n\Delta t.$$

Summary:

$$\begin{aligned}
u(t_1) &= u(t_0) + (c-1)\,\mathcal{S}(t_0) + s\mathcal{C}(t_0),\\
v(t_1) &= v(t_0) + 2\,(c-1)\,\mathcal{C}(t_0) + \left(\tfrac{3}{2}n\,\Delta t - 2s\right)\mathcal{S}(t_0) - \tfrac{3}{2}u(t_0)\,n\,\Delta t,\\
\dot u(t_1) &= nc\mathcal{C}(t_0) - ns\mathcal{S}(t_0),\\
\dot v(t_1) &= \dot v(t_0) - 2n\,(c-1)\,\mathcal{S}(t_0) - 2ns\mathcal{C}(t_0).
\end{aligned}$$

Combine equations 6.23 and 6.25:

$$\tfrac{3}{2}nu(t_0) + \dot v(t_0) = -\tfrac{1}{2}n\mathcal{S}(t_0).$$

From this and from equation 6.24:

$$\mathcal{S}(t_0) = -\left(3u(t_0) + \frac{2}{n}\dot v(t_0)\right), \qquad \mathcal{C}(t_0) = \frac{\dot u(t_0)}{n}.$$

Substitution yields

$$\begin{aligned}
u(t_1) &= u(t_0) - (c-1)\left(3u(t_0) + \frac{2}{n}\dot v(t_0)\right) + \frac{s}{n}\dot u(t_0),\\
v(t_1) &= v(t_0) + \frac{2\,(c-1)}{n}\dot u(t_0) -\\
&\qquad - \left(\tfrac{3}{2}n\,\Delta t - 2s\right)\left(3u(t_0) + \frac{2}{n}\dot v(t_0)\right) - \tfrac{3}{2}u(t_0)\,n\,\Delta t =\\
&= v(t_0) + \frac{2\,(c-1)}{n}\dot u(t_0) - (6n\,\Delta t - 6s)\,u(t_0) - \frac{3n\,\Delta t - 4s}{n}\dot v(t_0),\\
\dot u(t_1) &= c\dot u(t_0) + s\left(3nu(t_0) + 2\dot v(t_0)\right),\\
\dot v(t_1) &= \dot v(t_0) + (c-1)\,\left(6nu(t_0) + 4\dot v(t_0)\right) - 2s\dot u(t_0) =\\
&= 6n\,(c-1)\,u(t_0) - 2s\dot u(t_0) + (4c-3)\,\dot v(t_0).
\end{aligned}$$

As a matric equation, with state transition matrix Φ_0^1: phi φϕΦ

$$\overbrace{\left[\begin{array}{c} u \\ v \\ \hline \dot u \\ \dot v \end{array}\right]}^{\underline{x}(t_1)}(t_1) = \overbrace{\left[\begin{array}{cc|cc} 4-3c & 0 & s/n & -2\,(c-1)/n \\ 6s-6n\,\Delta t & 1 & 2\,(c-1)/n & 4s/n - 3\,\Delta t \\ \hline 3ns & 0 & c & 2s \\ 6n\,(c-1) & 0 & -2s & 4c-3 \end{array}\right]}^{\Phi_0^1}\overbrace{\left[\begin{array}{c} u \\ v \\ \hline \dot u \\ \dot v \end{array}\right]}^{\underline{x}(t_0)}(t_0). \tag{6.26}$$

This state transition matrix, together with equation 6.20 for the w co-ordinate, is known as the "Clohessy-Wiltshire model", Clohessy and Wiltshire (1960).

We can write this equation 6.26 in partitioned form on the basis β':

$$\begin{aligned}
\mathbf{u}(t_1) &= (\Phi_{11})_0^1\,\mathbf{u}(t_0) + (\Phi_{12})_0^1\,\dot{\mathbf{u}}(t_0),\\
\dot{\mathbf{u}}(t_1) &= (\Phi_{21})_0^1\,\mathbf{u}(t_0) + (\Phi_{22})_0^1\,\dot{\mathbf{u}}(t_0),
\end{aligned}$$

with, for example,

$$(\Phi_{11})_0^1 = \begin{bmatrix} 4 - 3\cos n\Delta t & 0 \\ 6\sin n\Delta t - 6n\,\Delta t & 1 \end{bmatrix}.$$

6.8.2 *The case of small Δt*

Write the system of second-order differential equations

$$\ddot{u} = 2n\dot{v} + 3n^2 u, \qquad \ddot{v} = -2n\dot{u}$$

as a first-order system:

$$\frac{d}{dt}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}.$$

7 For a small time step[7] Δt:

$$\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_1) \approx \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) + \Delta t \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) =$$

$$= \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 3n^2\,\Delta t & 0 & 1 & 2n\,\Delta t \\ 0 & 0 & -2n\,\Delta t & 1 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0).$$

One may verify for each matrix element that this is the same as equation 6.26, with the substitutions $c = \cos n\Delta t \approx 1$ and $s = \sin n\Delta t \approx n\,\Delta t$.

Self-test questions

1. What are Kepler's three laws of planetary motion?

[7]"Small" in relation to the orbital period, meaning $n\,\Delta t \ll 1$.

2. What are the six Kepler orbital elements?
3. What three alternative orbital elements are used to describe the position of a satellite in its orbit?
4. What are Hill co-ordinates?
5. Describe the libration movement found as one solution of the Hill equations and its connection to Kepler orbital motion.
6. Write out explicitly the other three state transition matrices $(\Phi_{12})_0^1$, $(\Phi_{21})_0^1$, and $(\Phi_{22})_0^1$ in the Clohessy–Wiltshire model 6.26.

Exercise 6–1: Kepler orbit

1. Derive the dynamic model of the Kepler state vector. Assuming that the force field is central, write the following linearised dynamic model equation explicitly:

$$\frac{d}{dt}\Delta\mathbf{a} = F\,\Delta\mathbf{a},$$

in which $\mathbf{a} \stackrel{\text{def}}{=} \begin{bmatrix} a & e & M & i & \omega & \Omega \end{bmatrix}^T$. The original non-linear model should be linearised by choosing suitable approximate or reference values $\mathbf{a}^{(0)}(t)$, obtained by precise integration over time from an initial $\mathbf{a}^{(0)}(t_0)$ using orbital mechanics. The delta quantities relate to these. Kepler's third law is also needed: vertausarvo

$$GM_\oplus P^2 = 4\pi^2 a^3,$$

in which P is the orbital period.

2. Due to flattening of the Earth, the right ascension Ω of the orbit's ascending node changes slowly (the so-called *nodal precession*) according to the following equation:

$$\frac{d\Omega}{dt} = -\frac{3}{2}\sqrt{\frac{GM_\oplus}{a^3}}\left(\frac{a_\oplus}{a}\right)^2 J_2 \cos i. \qquad (6.27)$$

This is a long-term average. A circular orbit is assumed, eccentricity $e \approx 0$. $a_\oplus$ is the equatorial radius of the Earth, and J_2 is the so-called dynamic flattening, a dimensionless number.
How does this affect the coefficient matrix F of the dynamic model derived above?

3. Is it allowable to use equation 6.27 to calculate the development in time of the approximate or reference value $\Omega^{(0)}(t)$?

4. Relationships 6.11, 6.12, and 6.15 between co-ordinates and velocity components in the orbital plane and orbital elements were derived:

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} a(\cos E - e) \\ b\sin E \end{bmatrix},$$

$$\boldsymbol{v} = \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \frac{2\pi}{P(1 - e\cos E)} \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} -a\sin E \\ b\cos E \end{bmatrix}.$$

Linearise the relationship between the state vectors

$$\Delta\mathbf{x} = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\dot{x} \\ \Delta\dot{y} \end{bmatrix} \text{ and } \Delta\mathbf{a} = \begin{bmatrix} \Delta a \\ \Delta e \\ \Delta\omega \\ \Delta E \end{bmatrix}.$$

Remember $b = a\sqrt{(1-e^2)}$ and Kepler's third law.

5. *Observation station.* How does one model the station's three-dimensional trajectory

$$\begin{bmatrix} X(t) & Y(t) & Z(t) \end{bmatrix}^T$$

in inertial co-ordinates as a result of Earth rotation? Assuming
8 that Earth rotation is uniform and the station location fixed,[8] write the dynamic model for the station co-ordinates.

Exercise 6–2: Rendezvous

Consider the so-called *rendezvous* problem of two spacecraft. Both are in circular orbits, but the orbit of the quarry is higher. This exercise looks at using a rocket boost to first change the circular orbit into an elliptical
9 Hohmann[9] transfer orbit, lifting its apogee to the quarry's orbit. After half an orbital period, at the apogee, a second boost is issued to make the orbit circular again.

For simplicity, we forget about the third co-ordinate w.

1. Specialise the state transition equation 6.26 to the case where $n\,\Delta t = \pi$.

[8]So, do not consider polar motion or length-of-day variations, solid-Earth tides, plate tectonics, glacial isostatic adjustment, and so on. In short: ignore geodynamics.

[9]Walter Hohmann (1880–1945) was a German engineer and a pioneer of theoretical astronautics. Hohmann (1925).

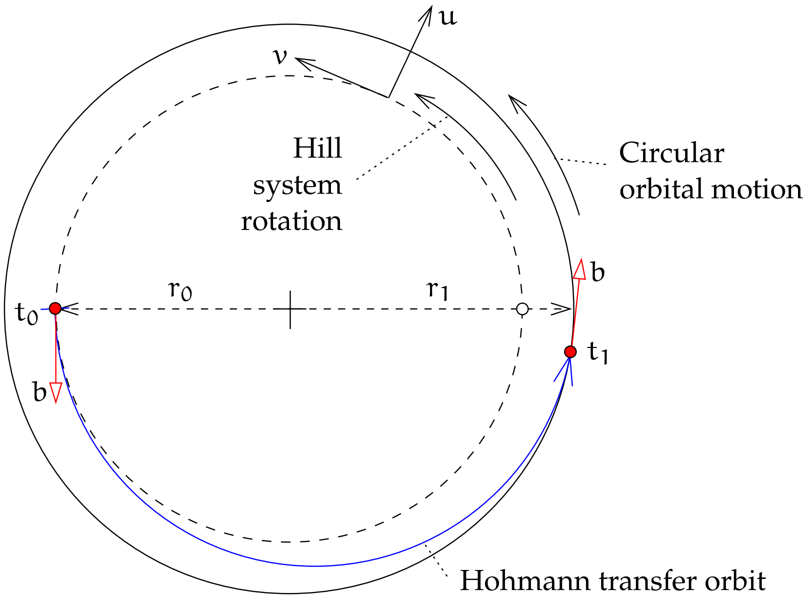


FIGURE 6.8. Rendezvous and Hohmann transfer orbit.

2. If the co-ordinates of place at starting epoch t_0 are given,

$$\mathbf{u}(t_0) = \begin{bmatrix} u \\ v \end{bmatrix}(t_0) = 0,$$

   simplify the equation. Which columns can be removed?

3. If furthermore is given ($b =$ "boost", velocity change):

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix}(t_0) = \begin{bmatrix} 0 \\ b \end{bmatrix},$$

   which value of $u(t_1)$ and of the distance $r_1 = r_0 + u(t_1)$ from the centre of attraction can be achieved with this boost? Simplify the equation even further.

4. What will be the velocity $\dot{v}(t_1)$ of the spacecraft at moment t_1 relative to the *departure* Hill frame?

5. What is the linear velocity of the departure Hill frame at distance r_1?

6. What is the velocity of the spacecraft at moment t_1 in a *non-rotating* frame?

7. What is the circular-orbit angular velocity n' and linear velocity $n'r_1$ at this distance r_1 according to Kepler's third law? Express in n, r_0, and r_1.

8. How large should the second boost at apogee be to get to circular velocity at the new distance r_1? Analyse carefully!

9. What will be the co-ordinate $v(t_1)$ in the departure Hill frame? In other words, by how much will the chaser be "behind"?

10. The craft carries on board a device for measuring the distance $\|\mathbf{u}\|$ between the spacecraft, as well as their velocity of approach $-\|\dot{\mathbf{u}}\|$. Form the *observation equations* if the state vector is

$$\mathbf{x} = \left[\begin{array}{ccc|ccc} u & v & w & \dot{u} & \dot{v} & \dot{w} \end{array}\right]^{\mathsf{T}}.$$

7 Technologies of satellite navigation

7.1 The Global Positioning System GPS

The Global Positioning System GPS was the first satellite system that could be used in real time for navigation. This was due to the choice of orbital geometry and the number of satellites, but especially to signal encoding suited for the simultaneous reception of ranging signals from multiple satellites. tosiaikaisesti

The GPS consists of *three segments*: the space, control, and user segments. The space segment consists of at least 24, in practice 27–31 satellites, including "active spares". There are six orbital planes, 60° apart, containing four satellites each. The orbital inclination is 55°. lohko

The geometry of the whole GPS constellation and of individual satellites repeats after $23^h 56^m$: one sidereal day or two GPS satellite orbital periods. After this time, observers on the rotating Earth as well as individual satellites are, in an inertial geocentric co-ordinate frame, in the same places in space again. The observation geometry repeats.

The GPS satellites belong to a number of technology generations or *blocks*. The old satellites of Block I, II, and IIA are no longer in operation. The newest type is Block III, the first of which was launched in 2018. The contract for the next block, Block IIIF ("III Follow-on"), has been awarded to Lockheed Martin.

The chosen geometry means that almost anywhere on Earth, at almost any time, one can see at least eight satellites above the horizon. Usually, more satellites can be seen, up to even twelve.

More on GPS and navigation is found in Strang and Borre (1997) pages 495–514. A good, though older, textbook on GPS positioning in general is Hofmann-Wellenhof et al. (1997).

The *control segment* comprises (Control Segment, status 2020):

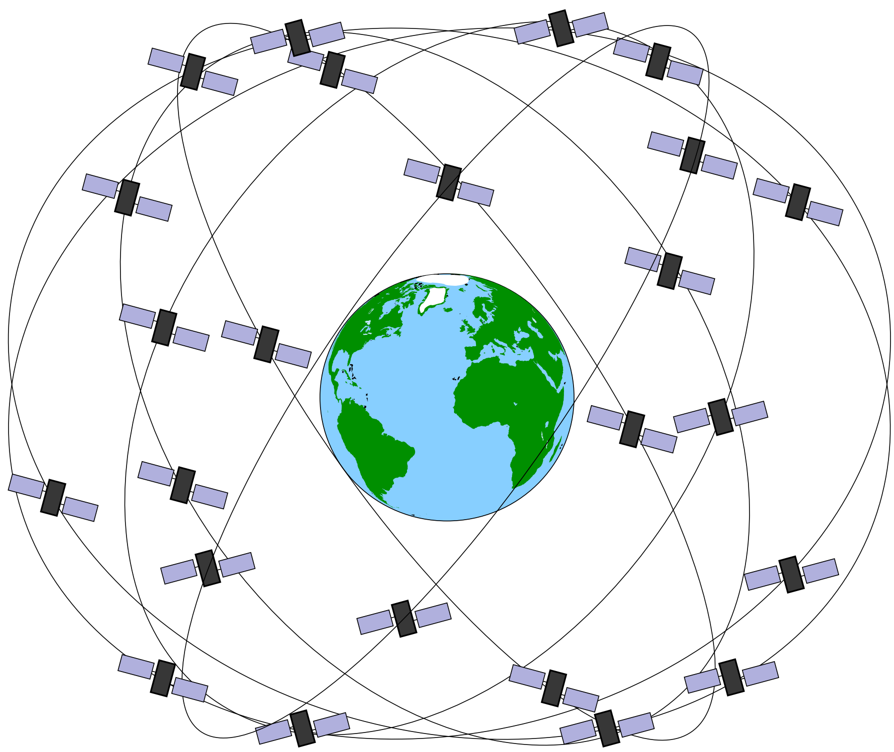


FIGURE 7.1. GPS constellation. Baseline constellation with 24 satellites.

- Sixteen monitoring stations around the globe. Unlike ordinary GPS observation stations, these have precise atomic clocks.
- Of the stations, four are "antenna stations" through which new orbital elements and other information are uploaded to the satellites' memories, usually twice every 24 hours. Seven more Air Force Satellite Control Network (AFSCN) stations can also be used for this.
- The data processing and control centre (MCS, Master Control Station) is located at Schriever Air Force Base in Colorado Springs, and its backup (Alternate MCS) is in Vandenberg, California.

7.2 GPS satellites and signal structure

The radio signal transmitted by a GPS satellite consists of a *carrier wave*, so-called *pseudo-random noise codes* modulated on it, and the *navigation message*. Both carrier wave and codes may be used for positioning. See the Interface Control Documents (GPS ICD) for details.

We mention separately the navigation message, which also contains the almanac. Each satellite broadcasts the almanac, which contains approximate orbital information for all satellites in the constellation.

The bit rate is only 50 bits per second and the total length of the message is 12.5 minutes.

Because of the low bit rate, it may take many minutes for a cold-started receiver to lock on to a sufficient number of satellites for a first fix. The same applies for a receiver transported between continents, as it not only will be initially looking for the wrong satellites in the local sky, but also assuming wrong Doppler shifts for them.

The navigation message is modulated on top of, that is, XORed with, both the C/A code on L1 and the P(Y) code on L1 and (optionally) L2. With the modernisation of GPS, a new navigation message called CNAV has been introduced, which is transmitted both as part of L2C on L2 and on L5. Furthermore, there is a new military navigation message called MNAV.

7.2.1 Carrier wave

The carrier wavelength is about 20 cm. The precision of positioning using carrier-phase measurements is 1 % of this or about 2 mm. Precise positioning and navigation is based on measuring the phase of the carrier wave. Dual-frequency receivers are always used for this purpose, so the ionospheric propagation delay may be eliminated. See table 7.1.

The orbital motion of the satellites causes a Doppler shift in the received signal, ranging from -5 cycles per millisecond when the satellite is moving away from the receiver to $+5$ cycles per millisecond when the satellite is approaching. If the receiver itself moves, that causes its own Doppler shift.

The receiver needs to be able to track the shifted frequency. The shift can be computed from the almanac.

The Doppler shift itself can be used as a GPS observable, in addition to pseudo-random code and carrier-phase measurements.

A well-known problem when using the phase of the carrier wave as an observable is that all waves look the same, so *ambiguity resolution* is needed. Multiple approaches have been developed for this. kokonaisluku-tuntematon

7.2.2 Codes

The "pseudo-wavelength" of the P-code is 29.3 m. This is the speed of light divided by the P-code "chip rate" or bit rate or bit frequency or

Table 7.1. Carrier waves of the GPS signal. The modernisation of the GPS system adds the frequency L5, intended mostly for use in aviation.

Carrier	Frequency (MHz)	Wavelength (cm)	Multiple of base frequency 10.23 MHz
L1	1575.42	19.0	154×
L2	1227.60	24.4	120×
L5	1176.45	25.5	115×

code frequency, yielding the spacing between bits on the radio wave in flight. Positioning precision is at best 1 % of this, or about 30 cm.

The chip rate or bit rate or bit frequency or code frequency of the C/A code corresponds to a "wavelength" of 293 m. See table 7.2.

When a receiver starts measurement, it first has to lock on to the C/A code. This is easy, as the code is short: it is a sequence of 1023 bits that repeats every millisecond. The receiver compares, for every satellite, the incoming signal with a receiver-generated replica of the code for that satellite, for the 1023 different possible delay values. When a match is found, the receiver locks on to the code and starts extracting the navigation message.

jäljitelmä

The C/A code, due to its short repeat period, only gives the pseudo-range to the satellite modulo 300 km, and with limited precision. The navigation message contains a "handover word" every six seconds telling the receiver precisely how far the much longer P(Y) code has advanced, allowing it to lock on to that code as well. The P(Y) code is a week-long segment of a much longer (37 weeks!) pseudo-random code, a different week for each satellite. This allows determination of the pseudorange to the satellite both unambiguously and precisely.

7.2.3 Modulations

The codes are modulated onto the carrier wave using so-called *phase modulation*, figure 7.2. The phase can be in one of two states separated by a phase-shift of $180° = \pi$. For code bits having the value 0, the carrier phase is kept as is, meaning a phase offset of zero. For code bits equalling 1, the carrier phase is advanced by 180°, meaning the carrier is "flipped" or multiplied with -1. The technique is called "binary phase-shift keying" (BPSK).

binaarinen vaiheavainnus

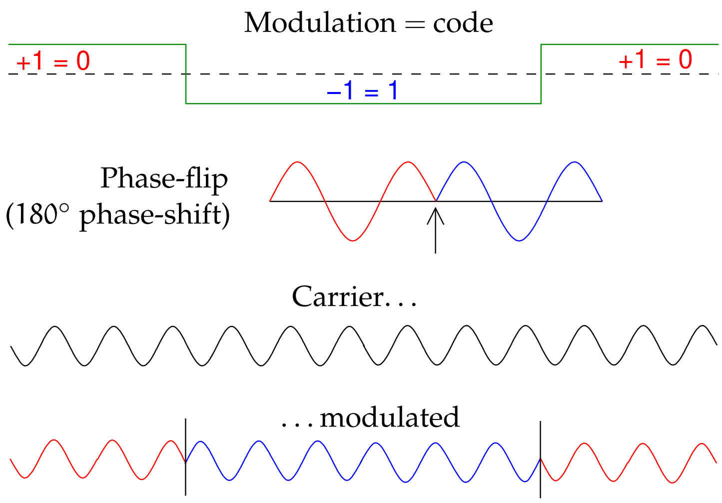
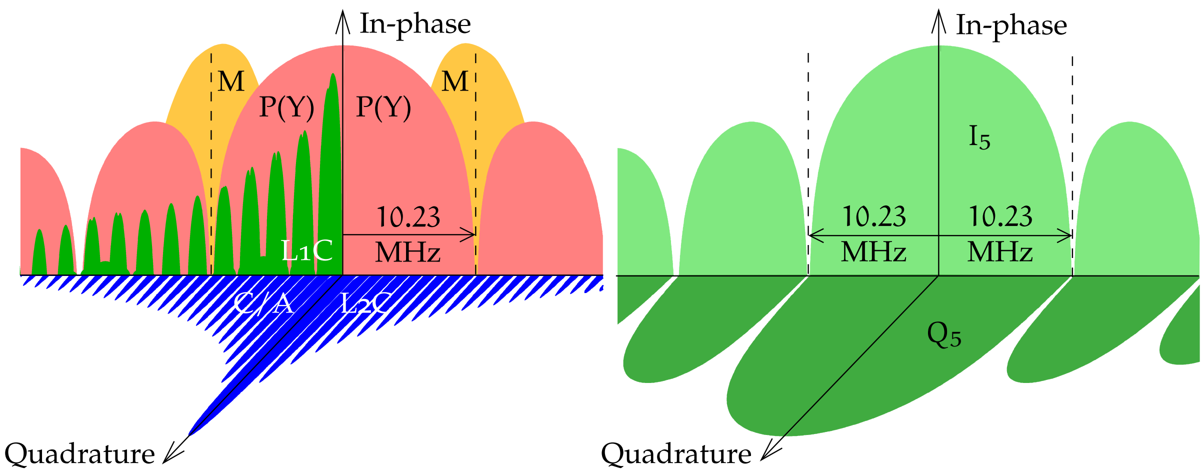


FIGURE 7.2. Principle of phase modulation.

Phase-shifts of 90° and 270° are also used in the GPS. These can also be used to encode a stream of bit values of 1 or 0. Thus, a single carrier can contain two bit streams side by side, with a phase-shift between them of $90° = \frac{1}{2}\pi$. These two bit streams are called the "in-phase" suora vaihe and "quadrature" components of the modulation. The total modulated carrier is the sum of these two components. For example on carrier L1, the P(Y) code is (conventionally) modulated in-phase and the C/A code in quadrature.



(a) **(b)**

L1, 1575.42 MHz L2, 1227.60 MHz L5, 1176.45 MHz

FIGURE 7.3. Power spectra of the GPS codes for in-phase and quadrature carrier-phase angles. L1 and L2 are given in a combined figure: C/A and L1C are modulated on L1, while L2C is similarly modulated on L2.

TABLE 7.2. The different pseudo-random noise (PRN) codes modulated on the GPS signal. The modernisation of the GPS has added to this a largish number of new codes, both on the old carrier frequencies L1 and L2 and on the new frequency L5. The markings "I" and "Q" denote in-phase and quadrature modulations.

Abbrev.	Name	Code frequency (Mb/s)	Modulation type	Repeat period	Carrier wave	I/Q
C/A	Coarse/Acquisition[b]	1.023	BPSK	1 ms	L1	Q
P, P(Y)[a]	Precise/Protected[b]	10.23	BPSK	1 week	L1, L2	I
L1C	L1 civil, data[b]	1.023	BOC(1, 1)	10 ms	L1	I
	pilot	1.023	TMBOC[c]	10 ms		
	overlay	100 b/s	BPSK	18 s		
L2C	L2 civil, moderate (CM)[b]	0.5115	BPSK	20 ms	L2	Q[d]
	long (CL)	0.5115	BPSK	1.5 s		
M	Military	5.115	BOC(10, 5)		L1, L2	I
I5	L5 data[b]	10.23	BPSK	1 ms	L5	I
Q5	L5 pilot	10.23	BPSK	1 ms		Q

[a] The Y, or P(Y), code is obtained by modulating — XOR-ing — the P-code with the classified W code, in order to prevent "spoofing", that is generation and transmission of a fake GPS signal.

[b] Carries navigation message.

[c] The pilot uses time multiplexed TMBOC (1, 1)/(6, 1).

[d] Default. I is optional.

7.2.4 Code-division multiple access

The pseudo-random noise codes modulated on the GPS carrier waves are individual to each satellite, and act like a "fingerprint" for the satellite. This is necessary, because all GPS satellites transmit on the same carrier frequencies, so-called *code-division multiple access* (CDMA). The codes are carefully constructed to be "orthogonal" to each other, so that separating out the signals from the different satellites arriving at the receiver antenna will be as easy as possible.

koodijako-kanavointi

For the C/A code, so-called Gold codes are used, to be discussed in section 7.4. They are 1023 bits long and thus repeat every millisecond. Much longer bit sequences are used for the P-code. These codes look random, but are generated by a known algorithm that receivers can

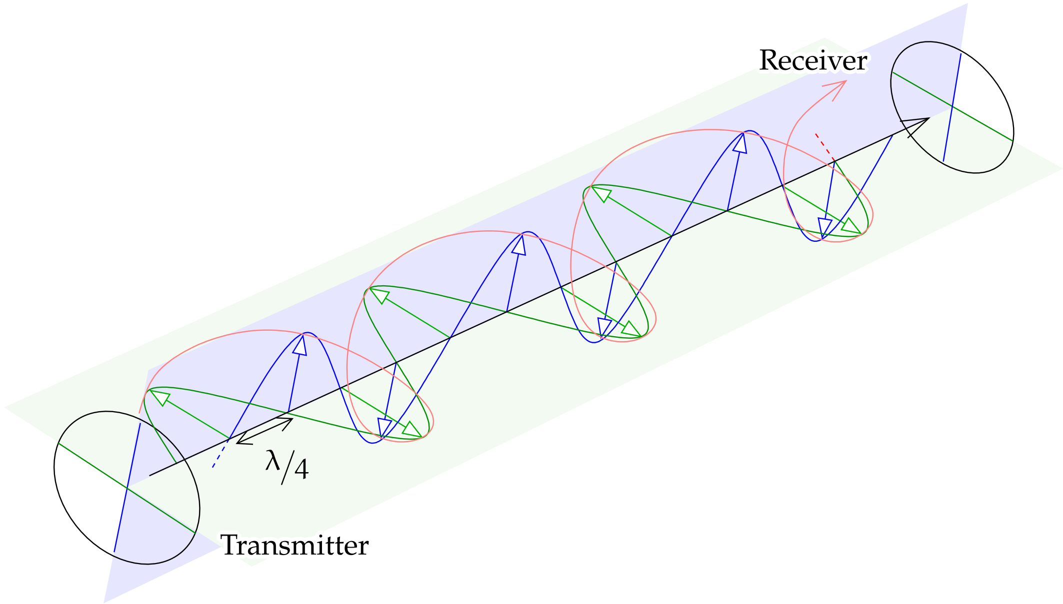


FIGURE 7.4. Circularly polarised radio waves propagating from transmitter to receiver. A circularly polarised wave is obtained by combining two waves that are linearly polarised in mutually perpendicular directions, with a phase offset of $\pi/2$ between them.

reproduce.

7.2.5 Polarisation

The radio transmissions of the GPS are right-hand or clockwise circularly polarised. This means that, looking along the propagation path, the field vector turns corkscrew fashion in the clockwise direction, one full round per wave period. See figure 7.4. The same applies for GLONASS, Galileo, BeiDou, and QZSS. The reason for this solution is to make it easy for a receiver antenna to reject signals that have undergone a single reflection from, for example, reflective surfaces near the antenna. These unwanted reflections are known as the *multipath* problem. monitie

This rejection is based upon the circumstance that reflection from a perfectly reflective surface, like a metallic surface, turns clockwise circularly polarised radiation into anti-clockwise polarised radiation.[1] [1]
Clockwise polarised radiation can be considered as consisting of two

[1] For non-metallic surfaces, like water or soil, it is more complicated. For perpendicular incidence, the reflected ray is also anti-clockwise circularly polarised like from a metallic surface, while for other incidence angles, the reflected radiation is elliptically polarised.

components that are linearly polarised in directions perpendicular to each other, and that differ by $\pi/2$ in their phase angle. Precise GPS receivers use a crossed-dipole type of antenna, where two perpendicular components are combined while applying a phase delay of $\pi/2$ to one of them. For clockwise polarised radiation, this means that the two components are in phase with each other and add up when combined. The components of the anti-clockwise polarised radiation on the other hand will be in anti-phase, and cancel fully or partly when combined.

7.2.6 Relativity theory

The frequency of the received GPS signal is relativistically shifted, due both to the receiver being much deeper inside the Earth's gravitational potential field than the satellites, and to the satellites moving at a considerable speed. A constant correction of $-4.4647 \cdot 10^{-10}$ f to frequency f is applied to all satellite clocks as they are being manufactured. This takes care of the bulk of the relativistic effect between the satellite orbit and mean sea level. A small, variable correction due to orbital eccentricity should still be applied by the receiver.[2]

There is also the issue of simultaneity, the so-called *Sagnac phenomenon* on the rotating Earth. In relativity theory, there is no absolute simultaneity. If one does the GPS calculations in a frame co-rotating with the Earth, a Sagnac correction to the measured pseudoranges will need to be made in the receiver that depends on the local speed of Earth rotation and the distance of the satellite from the meridian plane of the observation site (Poutanen, 2017, equation 4.19):

$$\delta_{\text{Sagnac}} p = \frac{1}{c} \langle \mathbf{V}' \cdot \mathbf{x} \rangle = \frac{1}{c} \left\langle \langle \boldsymbol{\omega}_\oplus \times \mathbf{X} \rangle \cdot \mathbf{x} \right\rangle,$$

omega $\omega\Omega$ in which $\mathbf{V}' = \langle \boldsymbol{\omega}_\oplus \times \mathbf{X} \rangle$ is the linear velocity vector of the observation site due to Earth rotation — pointing due east — and $\mathbf{x}$ is the geocentric location vector of the satellite, $\mathbf{X}$ that of the receiver. Note that the correction can be interpreted as the movement in the direction of the satellite which the observer undergoes in an inertial frame while the signal is flying from satellite to receiver at speed c.

This is an important consideration when comparing atomic clocks on different continents via the GPS. See Ashby (2003).

[2] In the early days of the GPS there was not enough on-board calculating capacity to make this correction on the satellite (Ashby, 2003)!

7.3 The carrier-wave corkscrew

7.3.1 *The effect of antenna orientation*

To analyse the effect of antenna orientation on the observed carrier phase, see figure 7.5, depicting the wave reaching the antenna from bottom left. The blue ellipse shows the foreshortening effect of projecting the field vector onto the horizontal plane of the antenna. In this plane, an incoming, circularly polarised wave of linear frequency f and thus of circular frequency $\omega = 2\pi f$, may be described, leaving out the arbitrary time origin and amplitude, by

$$\mathbf{E} = \cos(\omega t)\,\widetilde{\mathbf{i}} + \sin(\omega t)\cos\zeta\,\widetilde{\mathbf{j}},$$

in which $\widetilde{\mathbf{j}}$ is the unit vector in the direction of the line of sight projected onto the horizontal plane and $\widetilde{\mathbf{i}}$ the unit vector perpendicular to it within the horizontal plane. ζ is the zenith angle of the satellite. If $\mathbf{i}$ and $\mathbf{j}$ are the two unit vectors in the directions of the antenna's crossed dipoles, the signals in those dipoles will be zeta ζZ

$$\begin{aligned}
s_i = \left\langle \mathbf{E}\cdot\mathbf{i}\right\rangle = \cos(\omega t)\left\langle \widetilde{\mathbf{i}}\cdot\mathbf{i}\right\rangle + \sin(\omega t)\cos\zeta\left\langle \widetilde{\mathbf{j}}\cdot\mathbf{i}\right\rangle = \\
= \cos\omega t\cos\delta\phi + \sin\omega t\cos\zeta\sin\delta\phi, \\
s_j = \left\langle \mathbf{E}\cdot\mathbf{j}\right\rangle = \cos(\omega t)\left\langle \widetilde{\mathbf{i}}\cdot\mathbf{j}\right\rangle + \sin(\omega t)\cos\zeta\left\langle \widetilde{\mathbf{j}}\cdot\mathbf{j}\right\rangle = \\
= \sin\omega t\cos\zeta\cos\delta\phi - \cos\omega t\sin\delta\phi.
\end{aligned}$$

Here, $\delta\phi$ is the azimuth difference between line of sight and antenna. phi φϕΦ

Applying a phase advance relative to s_i of $\pi/2$ to s_j yields

$$s_j^* = \cos\omega t\cos\zeta\cos\delta\phi + \sin\omega t\sin\delta\phi,$$

and addition yields

$$\begin{aligned}
s_i + s_j^* = \cos(\omega t)\left(\cos\delta\phi + \cos\zeta\cos\delta\phi\right) + \\
+ \sin(\omega t)\left(\cos\zeta\sin\delta\phi + \sin\delta\phi\right) = \\
= (1+\cos\zeta)\left(\cos\omega t\cos\delta\phi + \sin\omega t\sin\delta\phi\right) = \\
= (1+\cos\zeta)\cos(\omega t - \delta\phi). \quad (7.1)
\end{aligned}$$

Note that a positive $\delta\phi$ makes the distance to the satellite appear longer: the same phase angle occurs for a larger value of t, the time of reception.

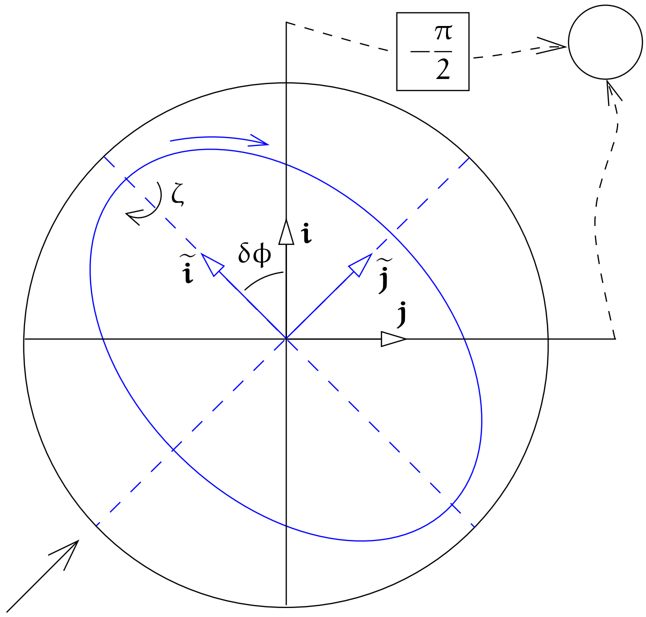


FIGURE 7.5. Reception of circularly polarised radiation.

Subtraction again yields

$$\begin{aligned}
s_i - s_j^* = \cos(\omega t)\,(\cos\delta\phi - \cos\zeta\cos\delta\phi) + \\
+ \sin(\omega t)\,(\cos\zeta\sin\delta\phi - \sin\delta\phi) = \\
= (1 - \cos\zeta)\,(\cos\omega t\cos\delta\phi - \sin\omega t\sin\delta\phi) = \\
= (1 - \cos\zeta)\cos(\omega t + \delta\phi).
\end{aligned}$$

This is conceptually equivalent to what addition does to anti-clockwise polarised radiation.

So, the rejection of anti-clockwise polarised radiation is complete for radio waves coming from the zenith, $\cos\zeta = 1$. For other satellite elevations, the rejection is incomplete, and for a satellite on the horizon, the antenna will only pick up the horizontally polarised component, regardless of whether it is part of a clockwise or anti-clockwise polarised wave. So in that case, no rejection of reflections takes place. This is unfortunate, as most multipath reflections come from nearby surfaces in directions close to the horizontal.

7.3.2 *Carrier-phase wind-up*

For carrier-phase measurements, the circular polarisation of the GPS radio waves introduces an unwelcome complication: rotating the antenna around its main axis will change the carrier-phase measurement. It is

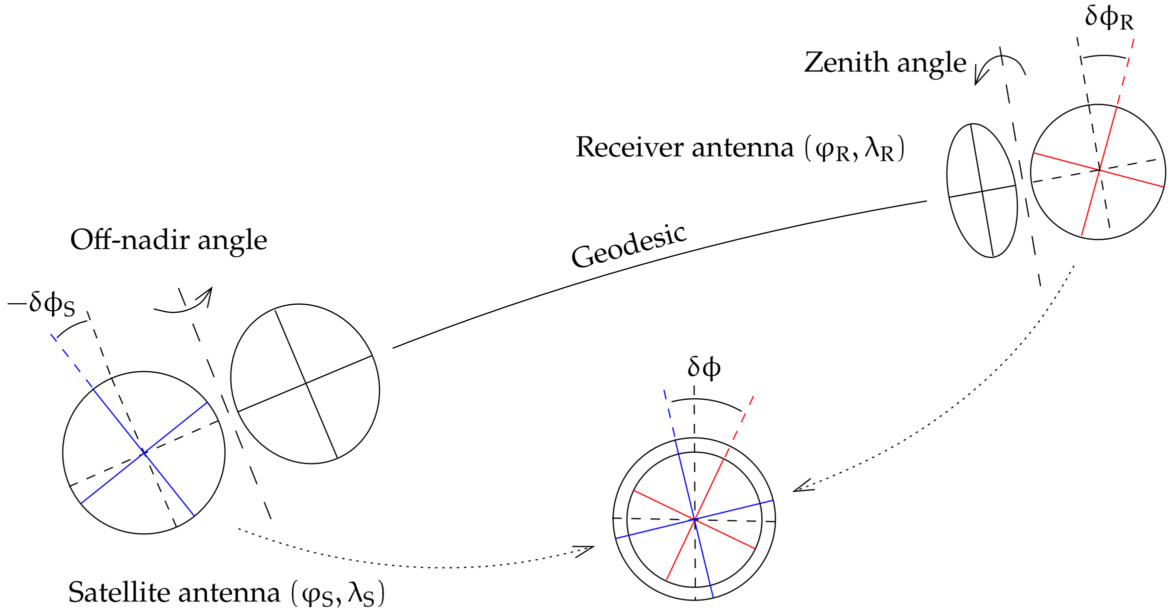


FIGURE 7.6. Phase wind-up. The total wind-up angle for a carrier-phase observation is $\delta\varphi = \delta\varphi_R - \delta\varphi_S$, the difference between the orientation offsets of receiver and satellite antennas from the line-of-sight azimuth. To make this calculation, each antenna has its main axis rotated around a horizontal line (long dashes) into the line of sight, bringing the antenna planes with the crossed dipoles into planes perpendicular to the line of sight. Now, the wind-up angle is the angle around the line of sight between the two crossed-dipole pairs.

The figure gives the situation projected onto the geocentric directional sphere, so the line of sight projects onto the geodesic.

like when you turn a bottle with a cork with a corkscrew in it and keep the corkscrew handle fixed: the corkscrew will move either in or out, depending on the direction of turning the bottle.

Of course the corkscrew handle does not stay fixed either: the GPS satellites turn in their orbits, tracking the Sun for their electric power. But this is something that can be modelled. For the receiver antenna, it is up to the user. The effect can be as large as half a wavelength. For an explanation of this phenomenon of *phase wind-up*, see figure 7.6. vaihekelaus

At both the receiver antenna and the satellite antenna, the phase wind-up effect amounts to the difference in azimuth of the line of sight and the antenna's crossed dipole, equation 7.1. It is assumed that the line of sight itself does not contribute, which is true if the plumb lines at receiver and satellite lie in the same plane; in other words, the Earth luotiviiva

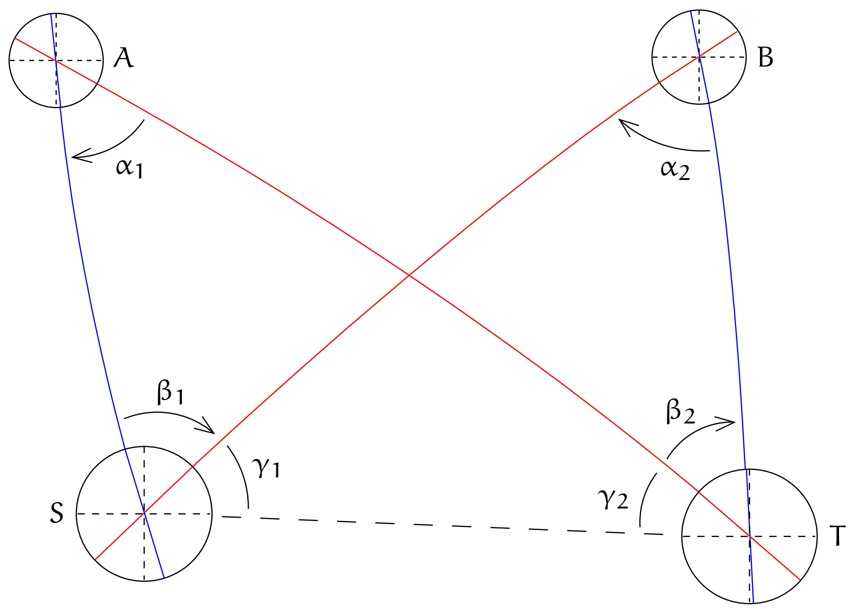


FIGURE 7.7. Phase wind-up effect for double-difference observations. The figure gives the situation projected onto the geocentric directional sphere, so lines of sight project onto geodesics.

is a sphere. See Wu et al. (1993).

For traditional geodetic measurements, it turns out that the difference observations used for calculating such networks are insensitive to changes in individual antenna orientation, both of a receiver and of a satellite. For local networks, the effect of the geometry also cancels out from the difference observations.

For larger networks, of a size of 1000 km and more, this is no longer the case and the effect of the geometry becomes significant. Figure 7.7 shows how for large networks there is a net non-zero wind-up effect for double differences. Every antenna, of receiver A or B or satellite S or T, contributes a part equal to the angle between the azimuths to a pair of other antennas: α_1 and α_2 between two satellite antennas as seen from a receiver, β_1 and β_2 between two receiver antennas as seen from a satellite. These parts should be carefully combined with the appropriate algebraic signs. See figure 7.8.

alpha αA
beta βB

In the figure, ϵ is the *spherical excess* of a triangle on the geocentric directional unit sphere. Symmetrically, also $\delta\phi_{AB}^{ST} = \epsilon_{TAB} - \epsilon_{SAB}$. The spherical excess equals the surface area of the triangle on the directional unit sphere. For small triangles, that is, local geodetic networks, this is negligible.

epsilon $\varepsilon\epsilon$E
palloylijäämä

Of course for GPS processing techniques that do not rely on construct-

$$
\begin{array}{rcl}
-\delta\phi_{AB}^{ST} = \alpha_1 - (-\beta_1) & - & \big(\alpha_2 - (-\beta_2)\big) \\
0 = \gamma_1 + \gamma_2 & - & (\gamma_1 + \gamma_2) \\
\hline
-\delta\phi_{AB}^{ST} = \underbrace{\alpha_1 + (\beta_1 + \gamma_1) + \gamma_2}_{180^\circ + \epsilon_{AST}} & - & \underbrace{\big(\alpha_2 + (\beta_2 + \gamma_2) + \gamma_1\big)}_{180^\circ + \epsilon_{BST}}
\end{array}
$$

$$
\Longrightarrow \quad \delta\phi_{AB}^{ST} = \epsilon_{BST} - \epsilon_{AST}.
$$

FIGURE 7.8. Phase wind-up and spherical excess for double-difference observations.

ing difference observations, such as precise point positioning (PPP), both the receiver and satellite antenna orientation must be explicitly taken into account.

When navigating, for example an antenna mounted on an aircraft will turn with the aircraft. Rotating the antenna around its own vertical axis will change the phase of the signal received from all satellites by the same angle, irrespective of the satellite zenith angle (García-Fernández et al., 2008). It follows that the phase-change will be absorbed into the receiver-clock unknown. Between-satellite difference observables will remain unaffected. Tight turns should be avoided, however, for example by using only measurements in straight flight. This may be significant when precise GPS measurements are used as part of a data product, like in airborne gravimetry, chapter 12.

7.4 The number theory of GPS

Nowadays Internet banking and commerce are critically dependent on number theory. This exotic branch of mathematics was once practised only for its own sake and not expected to ever have any practical use. Today, it allows strong encryption and authentication, so that people dare to trust the Internet with their money.

It should not be surprising that number theory has also insinuated itself into satellite positioning, for constructing the pseudo-random noise codes used to modulate the GPS carrier waves.

The starting point is the so-called linear feedback shift register (LFSR), figure 7.9. This example is of a shift register counting four cells or

x	y	$x \oplus y$
0	0	0
1	0	1
0	1	1
1	1	0

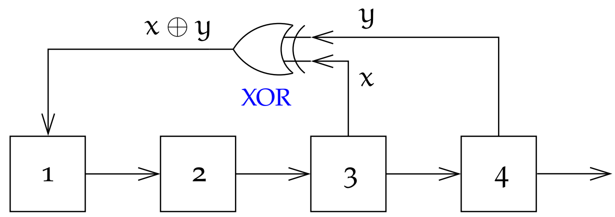


FIGURE 7.9. Linear feedback shift register. On the left, the definition of the XOR operation.

flipflops, each containing a bit value of 1 or 0. When the shift register is "triggered" by a clock signal, the content of every cell shifts to the cell on its right. Also, cells 3 and 4, through the XOR gate, create a bit value going into cell 1.

There are, with $n = 4$, $2^n = 16$ different possible states of the register content: all four-bit binary numbers. We exclude the state of all zeros from this, because the XOR of two zeros $0 \oplus 0 = 0$, so the register would never come out of the state of all zeros.

In this geometry, every trigger creates a new four-bit number and it can be shown that the full set of non-zero four-bit numbers is traversed in a random-looking sequence. The total count of these numbers is

$$2^n - 1 = 15.$$

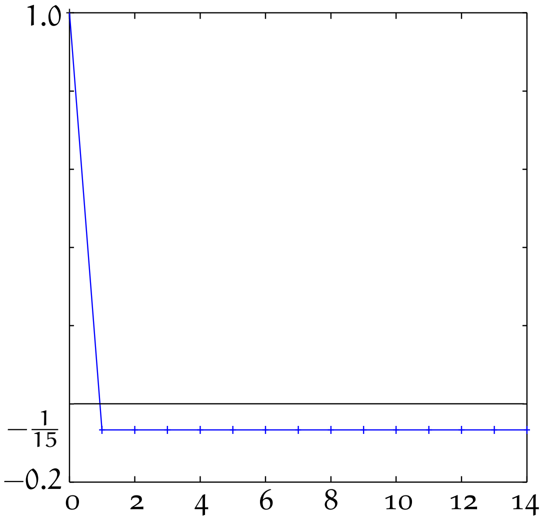
See table 7.3. After that, the sequence returns to the initial value and the cycle repeats.

The bit sequence in cell 4 is also the output bit sequence produced by the register.

- The number of ones in the sequence equals $2^{n-1} = 8$, the number of zeros is one less or $2^{n-1} - 1 = 7$. If all four-bit numbers were members of the set, one would expect as many last digits 1 as last digits 0 to occur. But the number 0000 is not a member of the set!
- Starting the sequence at a number different from 1000 will produce *the same sequence*, but cyclically shifted.
- It can be shown (appendix B) that XORing x with a cyclically shifted version x^α of itself produces yet another differently cyclically shifted version x^β.
- A sequence x and the same sequence x^α after any cyclical shift α are minimally correlated: the number of matching bits equals $2^{n-1} - 1$. Stated differently, $x \oplus x^\alpha = x^\beta$ contains 2^{n-1} ones and

TABLE 7.3. Sequence of values from the register of figure 7.9. On the right, the cyclical autocorrelation function.

	1	2	3	4	$3 \oplus 4$	Out
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	1	0
4	1	0	0	1	1	1
5	1	1	0	0	0	0
6	0	1	1	0	1	0
7	1	0	1	1	0	1
8	0	1	0	1	1	1
9	1	0	1	0	1	0
10	1	1	0	1	1	1
11	1	1	1	0	1	0
12	1	1	1	1	0	1
13	0	1	1	1	0	1
14	0	0	1	1	0	1
15	0	0	0	1	1	1
16	1	0	0	0	0	0



$2^{n-1} - 1$ zeros.

- As there are $2^n - 1$ different cyclically shifted versions of a single sequence x which are all minimally correlated, one might be tempted to use them to distinguish the individual satellites from each other. However, this will not work,[3] as the GPS observable is 3
precisely the shift in time of the signal. We need to find codes that are *essentially* different, it being impossible to convert one into the other by a mere cyclical shift.
- Robert Gold showed in 1967 that if two sequences x and y of the same lengths but generated by different shift-register feedback geometries are suitably chosen, then all sequences $x \oplus y^\alpha$ constructed as XORs of x and y for various cyclical shifts α
  - are essentially different
  - also have minimal cross-correlation with each other
  - their autocorrelation function resembles the delta function, with a peak at the origin and small values elsewhere.

[3]Actually GLONASS uses this type of sequences. This is the reason it must use different carrier frequencies for different satellites, see section 11.2.

$$\left.\begin{array}{cc|c} 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0\end{array}\right\} \quad \begin{tikzcd} x,y \arrow[r, leftrightarrow, "\overline{x}=1-2x", "\overline{y}=1-2y"'] \arrow[d, "\oplus"'] & \overline{x},\overline{y} \arrow[d, "\times"] \\ x\oplus y \arrow[r, leftrightarrow, "\begin{array}{c|c}0&+1\\ 1&-1\end{array}"] & \overline{x}\times\overline{y}=\overline{x\oplus y}\end{tikzcd} \quad \left\{\begin{array}{cc|c} -1&-1&+1\\ -1&+1&-1\\ +1&-1&-1\\ +1&+1&+1\end{array}\right.$$

FIGURE 7.10. Commutative diagram explaining the relationship between code bits and electromagnetic signal values, and how the XOR operator maps to multiplication.

See appendix B.

The concept of "minimal correlation" needs to be explained. If we have two bit sequences x and y with elements x_i, y_i, $i = 1, \ldots, 2^n - 1$, these bits will be represented in the GPS signal modulation as $0°$ and $180°$ phase shifts, or multiplication of the carrier by $+1$ and -1. So, the bit sequences x, y are represented as signal-value sequences $\overline{x}$, $\overline{y}$ containing a value -1 for a bit 1 and a value $+1$ for a bit 0. Expressed as linear relationships, $\overline{x} = 1 - 2x$, $\overline{y} = 1 - 2y$. See diagram 7.10.

Then, the correlation — "cyclical cross-correlation" — is

$$\begin{aligned} \mathrm{Corr}\{\overline{x}, \overline{y}\} &= \frac{1}{2^n - 1} \sum_{i=1}^{2^n-1} \overline{x}_i \overline{y}_i = \frac{1}{2^n - 1} \sum_{i=1}^{2^n-1} \overline{x_i \oplus y_i} = \\ &= \frac{1}{2^n - 1} \left(C_0(x \oplus y) - C_1(x \oplus y) \right), \end{aligned}$$

which will be close to zero if the numbers $C_0(x \oplus y)$ and $C_1(x \oplus y)$ of bits $0 = +1$ and bits $1 = -1$ in $x \oplus y$ are close to equal.

In table 7.3 the cyclical autocorrelation function

$$\begin{aligned} \mathrm{Corr}\{\overline{x}, \overline{x}^\alpha\} &= \frac{1}{2^n - 1} \sum_{i=1}^{2^n-1} \overline{x_i \oplus x_i^\alpha} = \\ &= \frac{1}{2^n - 1} \left(C_0(x \oplus x^\alpha) - C_1(x \oplus x^\alpha) \right) \end{aligned} \tag{7.2}$$

as a function of cyclical shift α is displayed. It is seen that the function value is unity for $x = x^\alpha$, that is, no shift. For all non-zero shifts it equals $-\frac{1}{15}$. This result holds generally, with the value being $-{}^1\!/_{(2^n - 1)}$. So, for large values of n, the function resembles the Dirac delta function, subsection 2.6.2. What this means is that a correlator can unambiguously identify the correct time shift between a received code and a locally generated replica of the same code.

TABLE 7.4. Sequences of register values for an alternative register geometry.

	1 2 3 4	2 ⊕ 4	Out		1 2 3 4	2 ⊕ 4	Out		1 2 3 4	2 ⊕ 4	Out
1	1 0 0 0	0	0	1	1 0 0 1	1	1	1	0 1 1 0	1	0
2	0 1 0 0	1	0	2	1 1 0 0	1	0	2	1 0 1 1	1	1
3	1 0 1 0	0	0	3	1 1 1 0	1	0	3	1 1 0 1	0	1
4	0 1 0 1	0	1	4	1 1 1 1	0	1	4	0 1 1 0	1	0
5	0 0 1 0	0	0	5	0 1 1 1	0	1				
6	0 0 0 1	1	1	6	0 0 1 1	1	1				
7	1 0 0 0	0	0	7	1 0 0 1	1	1				

The geometry of figure 7.9 was not chosen arbitrarily, as can be easily shown by building an alternative shift register that XORs cells 2 and 4 instead of 3 and 4, table 7.4. We need maximum-length sequences, also called *m-sequences*, that repeat after $2^n - 1$ register shifts. The branch of mathematics studying this field is called Galois theory, after Évariste Galois.[4]

4

The C/A code of the GPS is produced by two linear feedback shift registers of different geometry and a length of 10 bits each, producing 1023-bit long codes. These are XORed with each other for different values of the cyclical shift, producing the satellite-specific "Gold codes" modulated onto the outgoing carrier. See figure B.1.

7.5 The power spectrum of the GPS signal

The original pseudo-random noise code modulations of the GPS signal consist of simple, rectangular blocks, like the function

$$B(t) = \begin{cases} 1 & \text{if } t \in \left(-\frac{1}{2}, \frac{1}{2}\right), \\ 0 & \text{if } t \notin \left(-\frac{1}{2}, \frac{1}{2}\right). \end{cases} \tag{7.3}$$

It is not hard to show that the Fourier transform of this function is $(f \neq 0)$:

$$\mathcal{F}\{B\} = \int_{-\infty}^{+\infty} B(t) e^{-2\pi i f t} dt = \int_{-1/2}^{+1/2} e^{-2\pi i f t} dt =$$
$$= -\frac{1}{2\pi i f}\left[e^{-2\pi i f t}\right]_{-1/2}^{+1/2} = -\frac{1}{2\pi i f}\left(e^{-\pi i f} - e^{+\pi i f}\right) =$$

[4]Évariste Galois (1811–1832) was a French mathematician and number theorist who invented finite field theory or Galois theory. He died at age 20 of injuries suffered in a duel.

$$= \frac{1}{\pi f} \frac{e^{+\pi i f} - e^{-\pi i f}}{2i} = \frac{\sin \pi f}{\pi f}. \quad (7.4)$$

The case $f = 0$ yields $\mathcal{F}\{B\} = \int_{-1/2}^{+1/2} dt = 1$. This function is known as the sinc function, defined as

$$\operatorname{sinc} x \stackrel{\text{def}}{=} \begin{cases} \dfrac{\sin x}{x} & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

The modulation $C(t)$ of the GPS signal, whether the C/A code or the odotusarvo P(Y) code, may be considered a stationary process, the expectancy of which vanishes:

$$E\{C(t)\} = 0.$$

For a stationary process, the autocovariance function is (equation 2.16):

$$A_C(\Delta t) = \operatorname{Cov}\{\underline{C}(t), \underline{C}(t + \Delta t)\} = E\{\underline{C}(t)\, \underline{C}(t + \Delta t)\}. \quad (7.5)$$

It is also clear that the function $C(t)$, being pseudo-random noise, consists of bits that are uncorrelated with each other. This means that

$$A_C(\Delta t) = 0 \quad \text{if } |\Delta t| > 1,$$

because then, in equation 7.5, every bit value is multiplied only with values from other bits, with which it has zero correlation.

The only values of Δt for which the multiplication in equation 7.5 produces something of non-zero expectancy are those for which $-1 < \Delta t < 1$. For those values, the covariance equals the overlapping area of block $B(t)$ and the same block $B(t + \Delta t)$ shifted in time by Δt. We see that this is a triangle function $T(\Delta t)$ defined as

$$T(\Delta t) = \begin{cases} \Delta t + 1 & \text{if } \Delta t \in [-1, 0], \\ 1 - \Delta t & \text{if } \Delta t \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

See figure 7.11.

This triangle function is thus the autocovariance function $A_C(\Delta t)$ of the pseudo-random code $C(t)$. And looking at this figure, we also see that T equals the *convolution* of the block function B with itself,

$$T = B \otimes B \iff T(\Delta t) = \int_{-\infty}^{+\infty} B(t)\, B(\Delta t - t)\, dt.$$

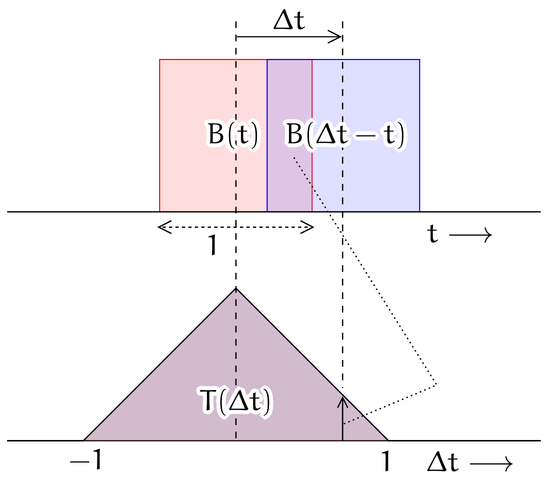


FIGURE 7.11. Autocovariance function of the GPS modulation.

According to the convolution theorem of Fourier theory, this means that the Fourier transform of T is, with equation 7.4,

$$\mathcal{F}\{T\} = \mathcal{F}\{B\}^2 = \left(\frac{\sin \pi f}{\pi f}\right)^2.$$

Thus we have found the Fourier transform of the autocovariance function of the GPS pseudo-random code. This is also known as the *power spectral density* function, equation 2.34: tehon spektraalitiheys

$$\mathcal{A}_C(f) = \mathcal{F}\{A_C\} = \mathcal{F}\{T\} = \left(\frac{\sin \pi f}{\pi f}\right)^2. \tag{7.6}$$

This function for the C/A and P(Y) codes, properly scaled for their respective bit frequencies and on a decibel scale, is plotted in figure 7.12.

In deriving this equation, it was assumed that the width of a code bit is unity. If it is something else, like $1/f_c$, in which f_c is the "code frequency" in bits per second, then f in equation 7.6 must be replaced by f/f_c. Also, the result should be divided by f_c.[5] 5

With this, equation 7.6 becomes

$$\mathcal{A}_C(f) = \frac{1}{f_c}\left(\frac{\sin(\pi\, f/f_c)}{\pi\, f/f_c}\right)^2 = f_c\left(\frac{\sin(\pi\, f/f_c)}{\pi f}\right)^2.$$

[5]Think of it in this way: if the code frequency f_c is one megahertz, then you can choose the microsecond as your new unit of time, and each bit will have a width of unity. The "new" frequency, expressed in megahertz, is now $f' = f/f_c$, and the power spectral density is found in watts per megahertz, to be divided by f_c to obtain it in the standard unit of watts per hertz.

Tableau 7.5. Power spectrum of the GPS signal, calculation code.

```
f = -20:0.1:20;
psdP = (sin(pi.*f/10.23)./(pi.*f/10.23)).^2;
psdCA = (sin(pi.*f/1.023)./(pi.*f/1.023)).^2;
hold on
axis([-20 20 -40 0])
plot(f, 10*log(psdP)/log(10), '-m')
plot(f, 10*log(psdCA)/log(10), '-b')
xlabel('Frequency offset (MHz)')
ylabel('Power spectral density (PSD) (dB)')
print 'GPS-PSD.pdf', '-dpdf'
```

7.6 BOC, binary offset carrier modulation

The original GPS adopted a modulation technique known as BPSK, binary phase-shift keying. The Galileo system uses a newer technique called BOC, binary offset carrier. The GPS's new M-code also uses the new technique.

In the original BPSK technique, every bit is represented by a single block of value either $+1$, for a 0 bit, or -1, for a 1 bit. The block function resulting from this code is then multiplied with the carrier, introducing $180°$ phase flips whenever a 1 bit is followed by a 0 bit or the reverse.

In BOC, what changes is that the single block representing every bit is

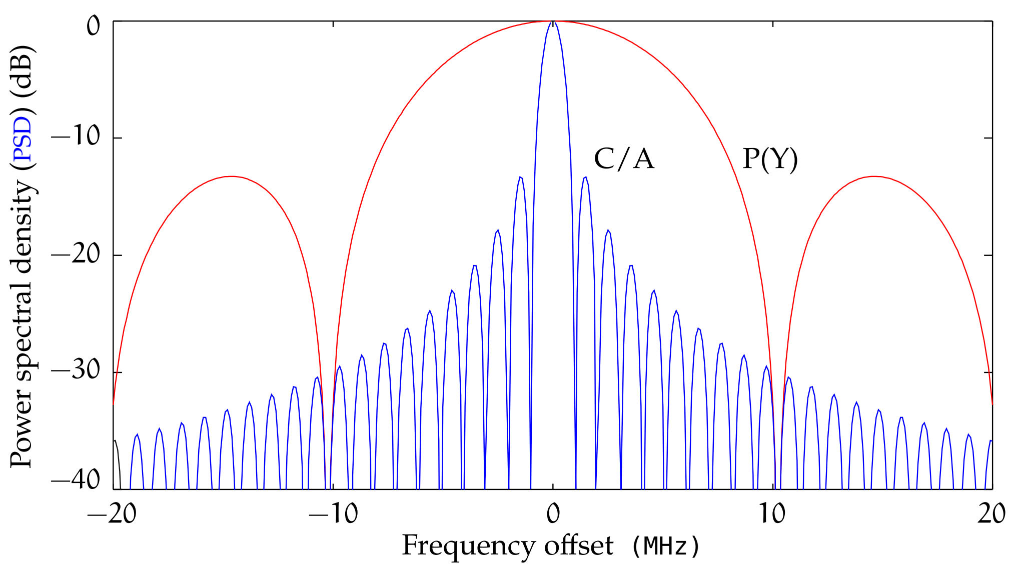


Figure 7.12. Power spectrum of the original GPS signal.

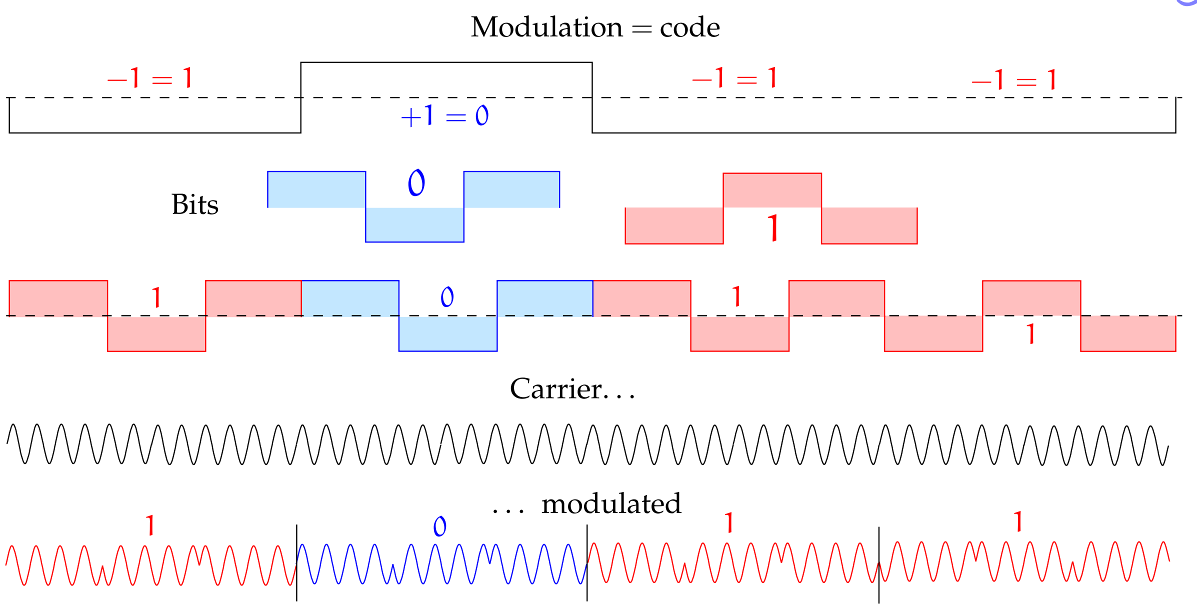


FIGURE 7.13. BOC, binary offset carrier modulation. An example.

replaced by a sequence of (for example) three blocks, see figure 7.13. Furthermore, in this example case (when there is an odd number of sub-blocks within a bit) there is an extra phase flip at every bit boundary. The net result is, that there will only be phase flips at bit boundaries $0 \rightarrow 0$ or $1 \rightarrow 1$, but not at boundaries $0 \rightarrow 1$ or $1 \rightarrow 0$.

It is possible to design BOC modulations for which most of the signal power goes to two side bands, one on each side of the carrier. Each side band is centred on a "subcarrier", offset from the carrier by a frequency difference of (in this example) $1.5\times$ the code frequency. Little power remains in the neighbourhood of the carrier. This makes it easier to separate the signal from signals that are already modulated on the same carrier using the BPSK technique, which puts most of the power near the carrier.

This is the situation for Galileo's E1 signal, which has the same carrier frequency as the GPS L1 carrier. The same applies for the GPS's new M-code, which on L1 overlays the pre-existing BPSK modulations of the C/A and P(Y) codes.[6] 6

[6]From the military viewpoint the advantage is, that they can selectively jam the C/A code, denying its and the P-code's use to the enemy, while continuing to use the M-code themselves.

Look for example at a two-block bit function:

$$B_2(t) = \begin{cases} 1 & \text{if } t \in \left(-\frac{1}{2}, 0\right), \\ -1 & \text{if } t \in \left(0, \frac{1}{2}\right), \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the Fourier transform is

$$\begin{aligned} \mathcal{F}\{B_2\} = \int_{-\infty}^{+\infty} B_2(t)\, e^{-2\pi i f t} dt = \\ = \int_{-1/2}^{0} e^{-2\pi i f t} dt - \int_{0}^{+1/2} e^{-2\pi i f t} dt = \\ = -\frac{1}{2\pi i f}\left[e^{-2\pi i f t}\right]_{-1/2}^{0} + \frac{1}{2\pi i f}\left[e^{-2\pi i f t}\right]_{0}^{+1/2} = \\ = \frac{1}{2\pi i f}\left(e^{\pi i f} - 2 + e^{-\pi i f}\right) = \frac{1}{2\pi i f} e^{-\pi i f}\left(e^{\pi i f} - 1\right)^2 = \\ = \frac{2i}{\pi f}\left(\frac{e^{\pi i f/2} - e^{-\pi i f/2}}{2i}\right)^2 = i\frac{\sin^2(\pi f/2)}{\pi f/2}. \end{aligned}$$

With the identity $\sin \pi f = 2 \sin(\pi f/2) \cos(\pi f/2)$ this becomes

$$\mathcal{F}\{B_2\} = i\frac{\sin(\pi f)\sin(\pi f/2)}{\pi f \cos(\pi f/2)},$$

the standard form found in the literature.

This transform is imaginary valued: this is because B_2 is antisymmetric. For symmetric functions, the autocovariance function and the self-convolution are the same; for antisymmetric functions, they are each other's opposites, equation 2.22. So the power spectral density becomes

$$\begin{aligned} \mathcal{A}_C(f) = \mathcal{F}\{A_C\} = \mathcal{F}\{B_2^\dagger(-t) \otimes B_2(t)\} = \\ -\mathcal{F}\{B_2(t) \otimes B_2(t)\} = \left(\frac{\sin(\pi f)\sin(\pi f/2)}{\pi f \cos(\pi f/2)}\right)^2. \end{aligned}$$

Figure 7.14 shows the effect of moving power away from around the carrier. Galileo uses an implementation of BOC called AltBOC, Shivaramaiah and Dempster (2009).

A popular symbolic notation for BOC modulations is $\mathrm{BOC}(f_s, f_c)$, with f_s the frequency of the subcarrier, which is *half* of the number of blocks per second, and f_c the number of code bits, units of information, per

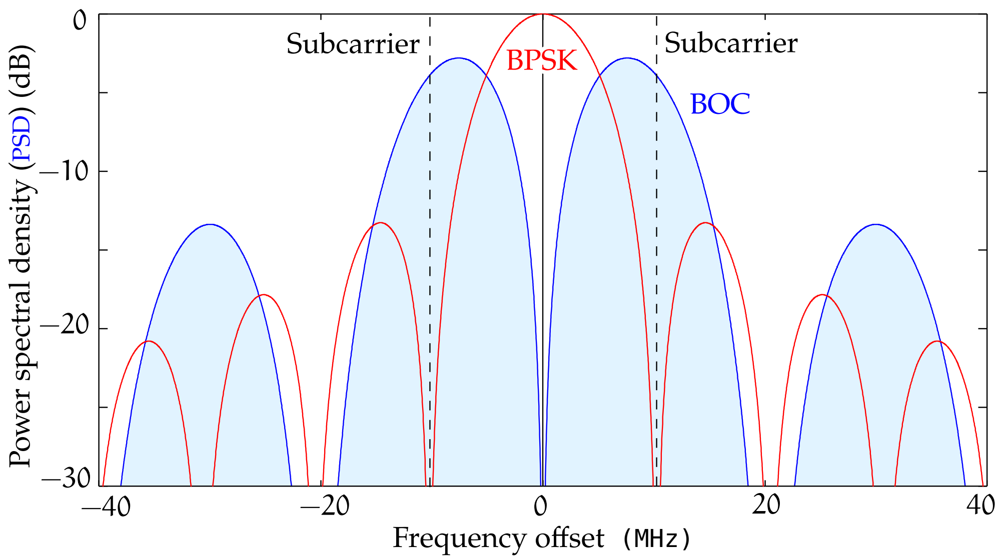


FIGURE 7.14. Example of how BOC moves power to the side bands.

second. With this notation, the generic equation for the power spectral density is

$$\mathcal{A}_C(f) = f_c \left(\frac{\sin(\pi\, f/f_c)\, \sin(\pi\, f/2f_s)}{\pi f \cos(\pi\, f/2f_s)} \right)^2,$$

see Ma et al. (2020). This is the most common BOC variant, also referred to as $\mathrm{BOC}_{\sin}(f_s, f_c)$, or "sine BOC".

7.7 Code and carrier-phase measurement

Due to the use of code-division multiple access (CDMA), the antenna of a GPS receiver on Earth will receive signals from all GPS satellites in the local sky, all transmitting on the same carrier frequencies. As the transmission power is limited and needs to be evenly spread out over one hemisphere of the Earth, the receivers need to be really sensitive. Therefore, the received signal is already pre-amplified in the antenna.

Before further amplification, the signal is multiplied, or *mixed*, with a reference frequency from the receiver clock, producing two derived signals at frequencies equal to the sum and difference of the signal and reference frequency. Of these, the difference frequency, also called the *intermediate frequency*, passes through a bandpass filter that also helps to suppress external noise. This intermediate-frequency signal is further amplified — without risk of the amplified signal leaking back into the antenna, so-called *crosstalk*. Only then is the amplified signal made available to the receiver's processing circuitry.

vertaustaajuus

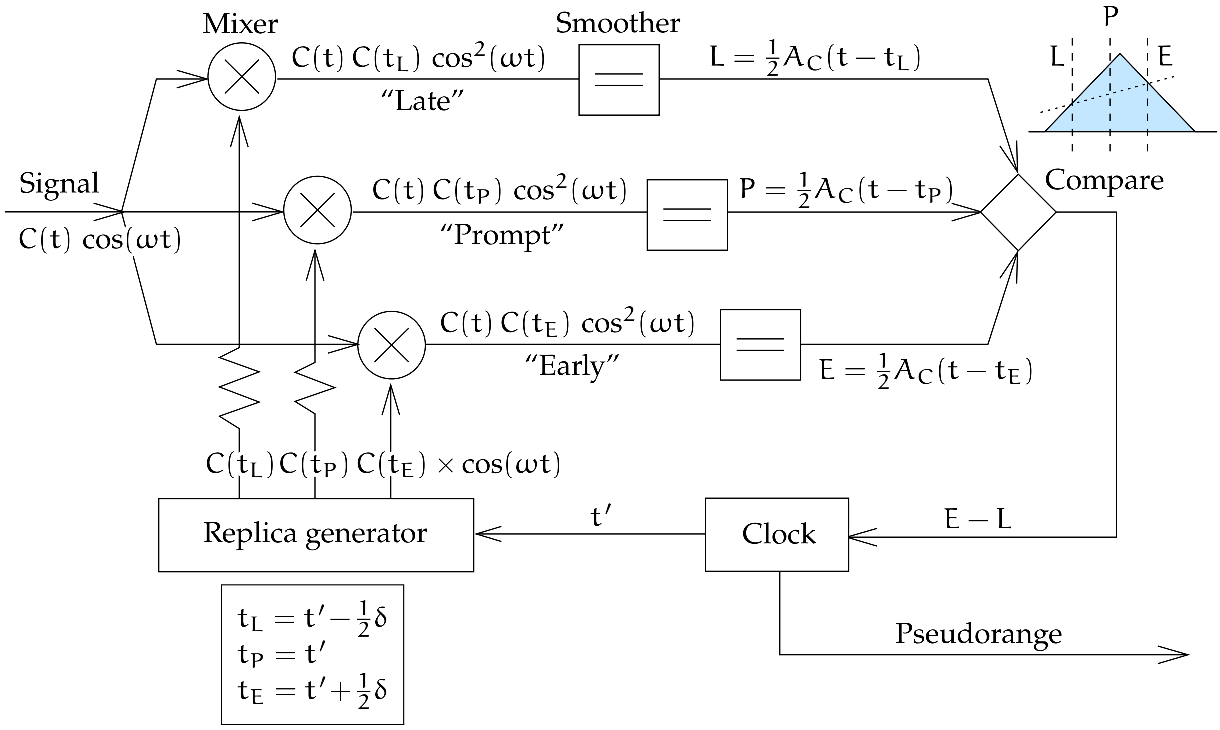


FIGURE 7.15. GPS code tracking.

Code measurement is relatively easy using a *correlator*, which correlates the received signal with signals self-generated by the receiver. Generally, a DLL, a *delay-locked loop*, is used as a correlator.

Each GPS satellite has its own pseudo-random noise code, which differs from that of every other satellite in a way that makes it as easy as possible to separate them. In a correlator, three synthetic signals or "replicas" are generated for each satellite: E (*early*), L (*late*) and P (*prompt*). See figure 7.15. The time difference δ between early and late equals one code chip, a microsecond for the GPS C/A code. Each generated signal is multiplied, or "mixed", with the incoming satellite signal. The result is the *correlation* of the two signals: a value which is positive only if the signals are nearly synchronous.[7]

Figure 7.11 shows the triangular shape of the autocovariance function of the pseudo-random code. The cross-covariance function between signal and replica has (assuming a coherent DLL) the same shape. So, in the case of exact synchronisation, it holds that $P = \frac{1}{2} \cdot 100\,\%$, $E = L = \frac{1}{2} \cdot 50\,\%$. The difference $E - L$, within the range $\left[-\frac{1}{2}, +\frac{1}{2}\right]$, will

[7] Here it is assumed for simplicity that the carrier phases of the incoming and replica signals are already synchronised. If this assumption for a *coherent* DLL does not apply, a more complicated, but also more robust, *non-coherent* DLL must be used.

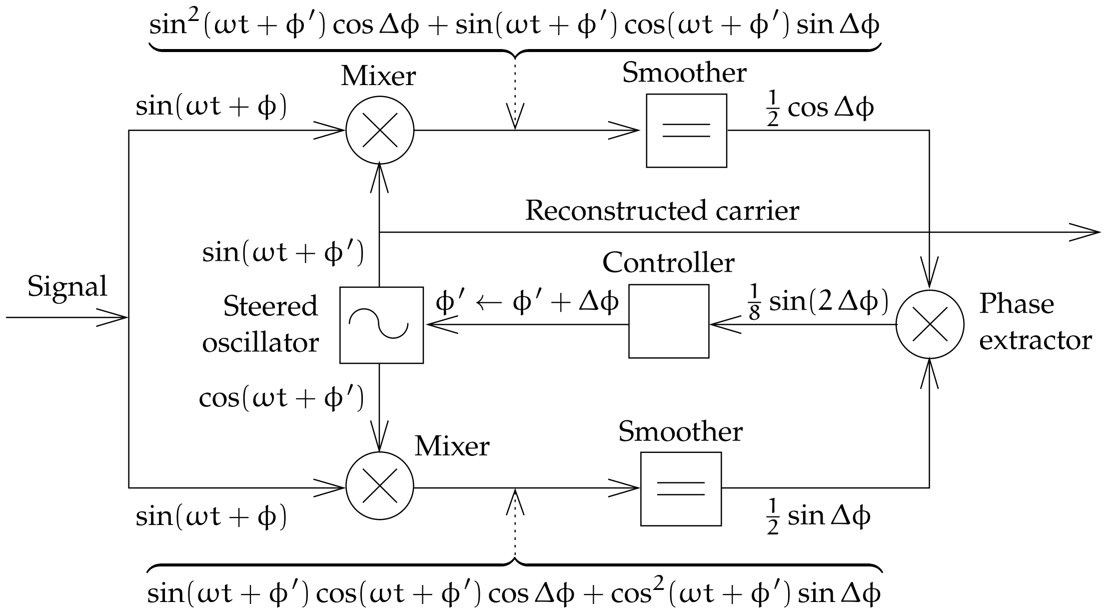


FIGURE 7.16. Carrier-phase tracking by a Costas discriminator, also known as a Costas loop.

indicate lack of synchronisation.

If the E mixer gives a high value, then the clock of the code generator is behind and it must be sped up. If, on the other hand, the L mixer gives the higher value, then the clock of the code generator must be slowed down.

Measurement of the carrier phase is unusual in consumer-grade devices. It is also more difficult. Generally, a so-called PLL or phase-locked loop is used for the purpose.

A commonly used phase-tracking device is the so-called Costas[8] 8
discriminator or loop. It observes the phase offset of the carrier wave against that of the receiver's reference oscillator and steers the receiver's reference frequency so that the difference vanishes. The reference frequency is observed and integrated into the carrier-phase observable.

A conceptual diagram of a Costas loop is given in figure 7.16. It is seen how the incoming signal is "mixed" or multiplied with generated reference carrier replicas, which are 90° or $\pi/2$ apart in phase. The outputs of these mixing operations are smoothed and then interpreted by a phase extractor to be the sine and cosine components of the

[8]John Peter Costas (1923–2008) was an American electrical engineer. His famous invention, the Costas loop, came out of his work on designing a better radio receiver (Costas, 1956).

difference $\Delta\phi = \phi - \phi'$ between the reference phase offset ϕ' and the signal phase offset ϕ being sought.

The phase extractor outputs the product $\frac{1}{8}\sin(2\,\Delta\phi)$ of its inputs $\frac{1}{2}\cos\Delta\phi$ and $\frac{1}{2}\sin\Delta\phi$. This is the steering signal for the reference carrier generator, which has its frequency steered to make the difference vanish.

The Costas discriminator is unable to distinguish the carrier from minus the carrier, that is the carrier phase shifted by $180° = \pi$:

$$\widetilde{\phi} = \pi + \phi \implies \sin(2\,\Delta\widetilde{\phi}) = \sin(2\,\Delta\phi).$$

This means that its unit of ambiguity resolution is half a wavelength,
lambda λΛ $\lambda/2$.

It also means that the Costas discriminator is a good way to "demodulate" the incoming signal, meaning removing the codes from the carrier: the output from the steered oscillator is the unmodulated carrier wave. Another simple way to achieve this would be by "squaring" the signal, multiplying it with itself. This makes phase flips of 180° just vanish, and all modulations, the pseudo-random codes as well as the navigation message, are stripped from the carrier. The same happens (except for the navigation message) if the incoming signal is multiplied with the synchronised replica signal.

In both the code-tracking and phase-tracking loops, there are com-
tasaaja ponents called "smoothers". They could also be called integrators, averagers, or low-pass filters. They can be conceptually described by the equation

$$\frac{d}{dt}\underline{y}(t) = -\frac{\underline{y}(t)}{T} + \underline{x}(t). \tag{7.7}$$

Here, T is the characteristic time scale of the smoother. The driving process is $\underline{x}(t)$, the smoothed resulting process $\underline{y}(t)$.

Equation 7.7 is somewhat similar to that of a Gauss-Markov process, equation 2.28, with time scale $T = {}^{1}\!/_{k}$. However, $\underline{x}(t)$ is a general process, unlike $\underline{n}(t)$ which is white noise. The general solution is also the same, equation 2.29:

$$\underline{y}(t) = \exp\left(-\frac{t}{T}\right)\left(\underline{y}(t_0)\,\exp\left(\frac{t_0}{T}\right) + \int_{t_0}^{t} \underline{x}(\tau)\,\exp\left(\frac{\tau}{T}\right)d\tau\right).$$

Here we see how the values of the driving process $\underline{x}(\tau)$ are attenuated over time by

$$\exp\left(-\frac{t-\tau}{T}\right),$$

in which $t - \tau$ is the age of the value $\underline{x}(\tau)$. This is also called a "fading-memory filter". The longer is T, the larger is the set of values $\underline{x}(\tau)$ over which the integral carries out averaging. tau τT

For the highest sensitivity in recovering the phase offsets and code delays between signal and replica, the smoothing time T should be long. This causes difficulty, however, in so-called high-dynamics applications such as manoeuvring jet fighters, where the Doppler shift of the received GPS signal will vary rapidly and unpredictably. With a too-long smoothing time, the tracking loops will lose their lock on the incoming signal. For this reason, receivers for high-dynamics use have short smoothing times at the cost of lower sensitivity and higher measurement uncertainty.

An alternative approach to handling high dynamics is the integration of GPS with a sensor capable of independently determining the changes in Doppler shift: an inertial measurement unit (IMU). Even a simple, inexpensive one will do.

7.8 Clock modelling

Clocks, such as the atomic clocks in GNSS satellites, are typically modelled as random-walk processes, equation 2.25. Let the clock-reading offset from the true time at time t be $c(t)$. Assume that the clock is driven by an oscillator of nominal frequency f_0. Then, the phase angle at time t will be in radians

$$\phi(t) = 2\pi f_0 t + \Delta\phi(t),$$

with $\Delta\phi(t)$ the phase offset, and the clock offset will be in seconds

$$c(t) = \frac{\Delta\phi(t)}{2\pi f_0}.$$

The frequency offset $\Delta f(t)$ is in hertz

$$f(t) = f_0 + \Delta f(t) = f_0 + \frac{1}{2\pi}\frac{d}{dt}\Delta\phi(t) = f_0 + f_0\frac{d}{dt}c(t)$$

and the *fractional frequency offset*, or clock drift, is a dimensionless number:[9] normalisoitu taajuus-poikkeama 9

$$d(t) \stackrel{\text{def}}{=} \frac{\Delta f(t)}{f_0} = \frac{d}{dt}c(t). \tag{7.8}$$

It is often assumed that the function $d(t)$ behaves like white noise, in which case equations 2.25 and 7.8 tell us that the clock offset $c(t)$ will kellon käynti behave like a random walk. Let the autocovariance function of the clock drift $\underline{d}(t)$ be, equation 2.23:

$$A_d(t, t') = Q_d\, \delta(t - t'),$$

delta δΔ in which δ is the Dirac delta function and Q_d the assumed variance. From equation 2.26 it follows that the autocovariance of the clock offset $\underline{c}$ is

$$A_{c,0}(t, t') = Q_d \cdot (t' - t_0)\,, \tag{7.9}$$

with t_0 some suitable starting time. We see that the variance grows linearly with time.

The average of the drift $\underline{d}(t)$ over a sampling interval $\tau = t_i - t_{i-1}$ is

$$\langle \underline{d} \rangle_i \stackrel{\text{def}}{=} \frac{1}{\tau} \int_{t_{i-1}}^{t_i} \underline{d}(t)\, dt = \frac{1}{\tau} \left(\underline{c}(t_i) - \underline{c}(t_{i-1}) \right). \tag{7.10}$$

10 Now the *Allan variance* is defined as this:[10] [11]

11

$$\sigma_A^2(\tau) \stackrel{\text{def}}{=} \tfrac{1}{2} E\left\{ \left(\langle \underline{d} \rangle_{i+1} - \langle \underline{d} \rangle_i \right)^2 \right\}. \tag{7.11}$$

12 The quantity is named after David W. Allan[12] (Allan's Time).

It may also be expressed in the clock offset $\underline{c}$:

$$\sigma_A^2(\tau) = \frac{1}{2\tau^2} E\left\{ \left(\underline{c}(t_{i+1}) - 2\underline{c}(t_i) + \underline{c}(t_{i-1}) \right)^2 \right\}.$$

An empirical estimator can be constructed as, for example,

$$\widehat{\sigma_A^2}(\tau) = \frac{1}{2n} \sum_{i=0}^{n-1} \left(\langle \underline{d} \rangle_{i+1} - \langle \underline{d} \rangle_i \right)^2.$$

sigma σΣ The square root $\sigma_A(\tau)$ of the Allan variance is called the *Allan deviation*.

Because $E\{\underline{d}\} = 0$, it follows that

$$E\left\{ \langle \underline{d} \rangle_{i+1} - \langle \underline{d} \rangle_i \right\} = 0$$

and equation 7.11 becomes

$$\sigma_A^2(\tau) = \tfrac{1}{2} \operatorname{Var}\left\{ \langle \underline{d} \rangle_{i+1} - \langle \underline{d} \rangle_i \right\}. \tag{7.12}$$

---

[9] The interpretation may be $^{\mathrm{Hz}}/_{\mathrm{Hz}}$ or $^{\mathrm{s}}/_{\mathrm{s}}$.

[10] In the literature, the symbol $\sigma_y^2(\tau)$ is used.

[11] Those familiar with spatial information analysis will recognise this as being similar to the *semivariogram* used in connection with the kriging technique.

[12] David Wayne Allan (born 1934) is an American atomic-clock physicist.

If the clock offset $\underline{c}(t)$ is a true random walk, the variances are, based on equations 7.10 and 7.9,

$$\begin{aligned}\mathrm{Var}\{\langle\underline{d}\rangle_i\} &= \frac{1}{\tau^2}\mathrm{Var}\{\underline{c}(t_i) - \underline{c}(t_{i-1})\} = \\ &= \frac{1}{\tau^2}\left(\mathrm{Var}\{\underline{c}(t_i)\} + \mathrm{Var}\{\underline{c}(t_{i-1})\} - 2\,\mathrm{Cov}\{\underline{c}(t_i), \underline{c}(t_{i-1})\}\right) = \\ &= \frac{1}{\tau^2}\left(Q_d \cdot (t_i - t_0) + Q_d \cdot (t_{i-1} - t_0) - 2Q_d \cdot (t_{i-1} - t_0)\right) = \\ &= \frac{1}{\tau^2}Q_d \cdot (t_i - t_{i-1}) = \frac{1}{\tau}Q_d,\end{aligned}$$

and thus

$$\mathrm{Var}\left\{\langle\underline{d}\rangle_{i+1}\right\} = \mathrm{Var}\left\{\langle\underline{d}\rangle_i\right\} = \frac{1}{\tau}Q_d, \tag{7.13}$$

and the covariance is obviously

$$\mathrm{Cov}\left\{\langle\underline{d}\rangle_i, \langle\underline{d}\rangle_{i+1}\right\} = 0, \tag{7.14}$$

as all $\{\, \underline{d}(t) \,|\, t \in (t_{i-1}, t_i)\,\}$ are uncorrelated with all $\{\, \underline{d}(t) \,|\, t \in (t_i, t_{i+1})\,\}$.

Expand equation 7.12 and substitute equations 7.13 and 7.14:

$$\begin{aligned}\sigma_A^2(\tau) &= \tfrac{1}{2}\mathrm{Var}\left\{\langle\underline{d}\rangle_{i+1} - \langle\underline{d}\rangle_i\right\} = \\ &= \tfrac{1}{2}\mathrm{Var}\left\{\langle\underline{d}\rangle_{i+1}\right\} + \tfrac{1}{2}\mathrm{Var}\left\{\langle\underline{d}\rangle_i\right\} - 2 \cdot \tfrac{1}{2}\mathrm{Cov}\left\{\langle\underline{d}\rangle_i, \langle\underline{d}\rangle_{i+1}\right\} = \\ &= \tfrac{1}{2}\frac{1}{\tau}Q_d + \tfrac{1}{2}\frac{1}{\tau}Q_d - 0 = \frac{1}{\tau}Q_d.\end{aligned}$$

This is the power law for frequency white noise.

In a double-logarithmic graph of $\sigma_A^2(\tau)$ against τ, this is a straight line sloping down by 45°. Plotting this graph is the accepted way of characterising clock behaviour (Allan, 1966).

Real clocks usually behave in this way for short sampling times τ. For longer times, frequency drift will bend the curve upwards. taajuuden ryömintä

Although the Allan variance was originally invented for characterising clocks, it is also used for gyroscopes and accelerometers.

7.9 Carrier-smoothed code measurement

In this method the absolute pseudorange comes from the code measurement, but its fractional-wavelength part from the phase measurement. vaihetuettu koodimittaus In kinematic applications of GPS it is often advantageous to *smooth* the raw pseudorange code measurements by using the much more smooth and geometrically precise, but in principle ambiguous, carrier-phase measurements.

Assume as given the code pseudorange measurements p_1 and p_2 (in metres) and the carrier-phase measurements ϕ_1 and ϕ_2 (in angular units, radians), at epochs t_k, $k = 1, 2, \ldots$.

Firstly, construct a *prediction equation* for the current *a priori* pseudorange from the previous *a posteriori* one:

$$p^-(t_k) = p^+(t_{k-1}) + \frac{\lambda}{2\pi}\left(\phi(t_k) - \phi(t_{k-1})\right). \tag{7.15}$$

Here, λ is the carrier wavelength. This equation is valid for both frequencies L1 (f_1) and L2 (f_2) as well as for the "widelane" observable defined as

leveäkuja

$$p_{WL} = \frac{f_1 p_1 - f_2 p_2}{f_1 - f_2}, \qquad \phi_{WL} = \phi_1 - \phi_2.$$

Equation 7.15 can be understood as a Kalman filter *dynamic model*: the state is the pseudorange $\underline{p}(t)$. The phase-change term $\phi(t_k) - \phi(t_{k-1})$ may be considered known, which is justified given that phase measurements have superior precision compared to code measurements.

Next, add to this Kalman filter an *observation equation*: the observation is simply the current pseudorange

$$\underline{\ell} = \underline{p}(t_k) + \underline{m},$$

the precision of which is given as observation variance $R = \mathrm{Var}\{\underline{m}\} = E\{\underline{m}^2\}$. The *update equation* is

$$p^+(t_k) = p^-(t_k) - KH\left(p^-(t_k) - \underline{\ell}\right),$$

in which

$$H = \left[\ 1\ \right], \qquad K = \Sigma^- H^T \left(H\Sigma^- H^T + R\right)^{-1} = \frac{\Sigma^-}{\Sigma^- + R},$$

and thus

$$p^+(t_k) = \overbrace{\frac{R}{\Sigma^-(t_k) + R}}^{W} p^-(t_k) + \overbrace{\frac{\Sigma^-(t_k)}{\Sigma^-(t_k) + R}}^{w} \underline{\ell}.$$

So the *a posteriori* pseudorange is a weighted average of, on the one hand, the predicted and carrier-smoothed *a priori* pseudorange, and on the other, the currently observed one. The weights W and w satisfy the condition

$$W + w = 1,$$

see subsection 2.2.3.

For the variance update we find

$$\Sigma^{+}(t_k) = (I - KH)\,\Sigma^{-}(t_k) = \frac{R}{\Sigma^{-}(t_k) + R}\Sigma^{-}(t_k).$$

For the variance propagation in the dynamic model, between epochs, we have simply $\Sigma^{-}(t_k) = \Sigma^{+}(t_{k-1})$.

It is possible to include *cycle slip detection* into the procedure: the testing variate is the difference or zero quantity or "closing error" vaihekatko

$$\underline{y} = \left(p^{-}(t_k) - \underline{\ell}\right),$$

of which we know the mean error to be

$$\sigma_{y} = \sqrt{H\Sigma^{-}(t_k)H^{\mathsf{T}} + R} = \sqrt{\Sigma^{-}(t_k) + R}.$$

This works best for the widelane linear combination thanks to its large effective wavelength, 86 cm.

This Kalman filter can run as a continuous process in the receiver — or in post-processing software, then without the real-time advantage. tosiaikaisuus The estimate

$$\left\{ p^{-}(t) \,\middle|\, t \in \left[t_{k-1}, t_k\right) \right\} = p^{+}(t_{k-1}) = p^{-}(t_k), \qquad k = 2, 3, \ldots$$

output by the filter will be significantly smoothed compared to the time series of raw observations $\underline{\ell}_k = \underline{p}(t_k) + \underline{m}$.

Self-test questions

1. What is the modulation technique used by the radio transmissions of the Global Positioning System?
2. What are the two parts of the relativistic frequency shift of the received GPS signal?
3. Describe the phenomenon of carrier-phase wind-up.
4. What are Gold codes? What does it mean that they are mutually nearly orthogonal?

5. How is it possible for a GPS receiver to separate out the signals from the satellites in the sky, in spite of them all broadcasting on the same carrier frequency?
6. What is a replica code and how does a code correlator work?
7. What is a high-dynamics environment and what does it demand of a GPS receiver?
8. Why is BOC, binary offset carrier modulation, useful?
9. How are receiver clocks usually modelled?
10. What is the Allan variance?
11. Explain carrier-smoothed code measurement.

Exercise 7–1: The three-block bit representation

The representation of every code bit by a sequence of three blocks as shown in figure 7.13 can be written as

$$B_3(t) = \begin{cases} -1 & \text{if } t \in \left(-\frac{1}{2}, -\frac{1}{6}\right) \cup \left(\frac{1}{6}, \frac{1}{2}\right), \\ 1 & \text{if } t \in \left(-\frac{1}{6}, \frac{1}{6}\right), \\ 0 & \text{otherwise.} \end{cases}$$

1. Derive the power spectral density

$$\mathcal{A}_C(f) = \mathcal{F}\left\{B_3^\dagger(-t) \otimes B_3(t)\right\} = \mathcal{F}\left\{B_3\right\}^2$$

of the modulation $C(t)$ produced by this representation.

Hint Write the block sequence as the sum of only two simple blocks.

2. How would you change the height of the central block of the sequence in order to make the power spectral density at the origin $f = 0$ vanish?

Real-time GNSS observations

8

8.1 Observation equations for GNSS

In this chapter, in order to study the use of real-time GNSS observations for navigation, the observation equations of GNSS measurement are developed first. Initially, the equations are presented in their general, non-stochastic, non-linear form, for both code and carrier-phase measurement. Then, the equations are linearised. Next, they are specialised by choosing which variables in the equation are known and which are unknown, the latter to be included in the vector of unknowns, also called the state vector. tosiaikainen

The atmospheric parameters in the equations demand special treatment. A separate analysis of this is presented.

At the end of the chapter, differential GNSS measurements are introduced, including the real-time kinematic (RTK) measurement technique. Also, methods and standards for establishing a data link between a reference GNSS receiver and a moving receiver are presented. vertaus-vastaanotin

8.1.1 Observation equation for code measurement

The most commonly used GNSS observation type is the *code pseudorange*, given by the equation

$$p = \rho + c\,(\Delta T - \Delta t) + d_{\text{ion}} + d_{\text{trop}}, \tag{8.1}$$

in which

$$\rho = \|\mathbf{X} - \mathbf{x}\| = \sqrt{(X-x)^2 + (X-y)^2 + (Z-z)^2}$$

is the spatial distance between satellite location $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and ground-station or receiver location $\mathbf{X} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$ computed using the Pythagoras theorem. rho ρR

Δt is the satellite clock offset.

ΔT is the receiver clock offset.

d_{ion} is the "ionospheric" delay in signal propagation caused by the ionised fraction of the atmosphere.

d_{trop} is the "tropospheric" delay caused by the neutral fraction of the atmosphere.

The prefix "pseudo" in the term pseudorange refers to the presence of the clock offsets Δt and ΔT, which makes the pseudorange behave differently in computations compared to the true ranges that are observed for example by laser ranging.

Equation 8.1 may be called a "proto observation equation". It contains more variables than can ever be determined in a single computation. From this equation, actual observation equations may be derived, the form of which will depend on what we choose to be the unknowns to be estimated by the Kalman filter. Unknowns that can be included in the state vector of a Kalman filter are $\mathbf{x}$, $\mathbf{X}$, Δt, and ΔT.

Inclusion of the atmospheric propagation parameters d_{trop} and d_{ion} is problematic. The form of equation 8.1 shows that every satellite will have its own parameters d_{trop} and d_{ion} for every epoch. These delays are dependent on the satellite elevation. Therefore they cannot be resolved from observations using this equation. These corrections may be available from some external source, like an atmospheric model.

Leaving the atmospheric parameters aside for now, we obtain as the proto observation equation for code measurements

$$p = \rho + c\,(\Delta T - \Delta t)\,. \tag{8.2}$$

8.1.2 Observation equation for carrier phase

A similar equation as equation 8.1 for code pseudoranges is also found for carrier-phase observables:

$$\begin{aligned}\overline{P} = P + \lambda N = \lambda\left(\frac{\phi}{2\pi} + N\right) = \lambda\left(\frac{\overline{\phi}}{2\pi}\right) = \\ = \rho + c\,(\Delta T - \Delta t) + D_{ion} + D_{trop},\end{aligned} \tag{8.3}$$

in which

$$\rho = \|\mathbf{X} - \mathbf{x}\| = \sqrt{(X-x)^2 + (Y-y)^2 + (Z-z)^2}$$

is the geometric distance between satellite, location $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, and receiver, location $\mathbf{X} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$.

ϕ is the measured or "raw" phase-difference angle in radians. At signal acquisition, $\phi(t_0) \in [0, 2\pi)$. Note that what is actually being measured is by how much the phase of a reference oscillator in the receiver is ahead of the phase of the incoming satellite signal. Therefore when the distance to the satellite increases, the measured phase angle also increases. phi $\varphi\phi\Phi$ vertaus-oskillaattori

$\overline{\phi}$ is the phase-difference angle including the number of full wavelengths N needed to make equation 8.3 valid. The number $N \in \mathbb{Z}$[1] is called the (integer) *ambiguity number*, and it holds that kokonaisluku-tuntematon

$$\overline{\phi} = \phi + 2\pi N.$$

$\overline{P} = \lambda \, \overline{\phi}/2\pi$ is close in value to the code pseudorange.

$P = \lambda \, \phi/2\pi$ is not.

$\Delta t, \Delta T$ are the satellite and receiver clock offsets.

D_{ion}, D_{trop} are the propagation delays caused by "ionosphere" and "troposphere" — in fact $D_{ion} = -d_{ion}$ and $D_{trop} = d_{trop}$.

λ is the wavelength of the carrier. lambda $\lambda\Lambda$

Also here we leave the atmospheric propagation parameters out of consideration in the first instance, leaving us with the proto observation equation

$$\overline{P} = \lambda \left(\frac{\phi}{2\pi} + N \right) = \rho + c \left(\Delta T - \Delta t \right). \tag{8.4}$$

8.2 Linearisation of the observation equations

The above equations 8.2 and 8.4 must be linearised in order to obtain the update equations of the Kalman filter 3.25, 3.26, and 3.27:

$$K = \Sigma^- H^\mathsf{T} \left(H \Sigma^- H^\mathsf{T} + R \right)^{-1}, \qquad \begin{aligned} \mathbf{x}^+ &= \mathbf{x}^- - K \left(H\mathbf{x}^- - \underline{\boldsymbol{\ell}} \right), \\ \Sigma^+ &= (I - KH)\, \Sigma^-. \end{aligned}$$

[1] Half-integer numbers may occur so that $2N \in \mathbb{Z}$, for example when using a Costas discriminator, see section 7.7.

8.2.1 Estimating receiver locations

Assume that the unknowns to be determined are the elements of the state vector

$$\mathbf{x} = \begin{bmatrix} \mathbf{X} \\ \Delta T \end{bmatrix}. \tag{8.5}$$

They are the location co-ordinates X, Y, and Z of the receiver as well as the receiver clock offset ΔT. We thus assume that the satellite location $\mathbf{x}$ and clock offset Δt are *not* intended to be determined, but are known.

vertausarvo Choose the following set of approximate or reference values:

$$\begin{aligned} p^{(0)} &= \rho^{(0)} + c\left(\Delta T^{(0)} - \Delta t\right), \\ \rho^{(0)} &= \sqrt{\left(X^{(0)} - x\right)^2 + \left(Y^{(0)} - y\right)^2 + \left(Z^{(0)} - z\right)^2}, \quad \Delta T^{(0)} = 0. \end{aligned} \tag{8.6}$$

This amounts to forming the approximate state vector

$$\mathbf{x}^{(0)} = \begin{bmatrix} \mathbf{X}^{(0)} \\ 0 \end{bmatrix}, \tag{8.7}$$

with $\mathbf{X}^{(0)} = \begin{bmatrix} X^{(0)} & Y^{(0)} & Z^{(0)} \end{bmatrix}^\mathsf{T}$.

Subtract equation 8.6 from equation 8.2, expand into a Taylor series, and identify the delta quantities defined in the usual way,

$$\Delta X \stackrel{\text{def}}{=} X - X^{(0)}, \quad \Delta Y \stackrel{\text{def}}{=} Y - Y^{(0)}, \quad \Delta Z \stackrel{\text{def}}{=} Z - Z^{(0)},$$

and

$$\begin{aligned} \Delta p &\stackrel{\text{def}}{=} p - p^{(0)} \approx \\ &\approx \left.\frac{\partial p}{\partial X}\right|_{\mathbf{X}=\mathbf{X}^{(0)}} \Delta X + \left.\frac{\partial p}{\partial Y}\right|_{\mathbf{X}=\mathbf{X}^{(0)}} \Delta Y + \left.\frac{\partial p}{\partial Z}\right|_{\mathbf{X}=\mathbf{X}^{(0)}} \Delta Z + c\,\Delta T. \end{aligned}$$

Identify the observation matrix and compute the partial derivatives:

$$\begin{aligned} H &= \begin{bmatrix} \frac{\partial p}{\partial X} & \frac{\partial p}{\partial Y} & \frac{\partial p}{\partial Z} & \frac{\partial p}{\partial\,\Delta T} \end{bmatrix}^{(0)} = \\ &= \begin{bmatrix} \frac{X - x}{\rho} & \frac{Y - y}{\rho} & \frac{Z - z}{\rho} & c \end{bmatrix}^{(0)} = \begin{bmatrix} \frac{\mathbf{X} - \mathbf{x}}{\rho} & c \end{bmatrix}^{(0)}, \end{aligned} \tag{8.8}$$

rakennematriisi the design or observation matrix of the Kalman filter.

The linearised state vector is obtained by subtracting equation 8.7 from equation 8.5, and we drop the delta as is customary:

$$\mathbf{x} = \Delta\mathbf{x} \stackrel{\text{def}}{=} \begin{bmatrix} \Delta X & \Delta Y & \Delta Z & \Delta T \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \Delta\mathbf{X} \\ \Delta T \end{bmatrix}. \tag{8.9}$$

All elements are functions of time.

Assume that at a certain point in time t, some function of the state vector is observed. The linearised observation equation, written as a stochastic equation, is

$$\underline{\ell} \stackrel{\text{def}}{=} \Delta\underline{p} = H\,\underline{\mathbf{x}}(t) + \underline{m}, \tag{8.10}$$

with H and $\underline{\mathbf{x}}(t)$ as defined above, and $\underline{m}$ the random observation error or noise.

We can study on an epoch-by-epoch basis how precisely the observations are able to fix the solution for the state vector 8.9. If we have pseudorange observations to n different satellites, the observation equation becomes

$$\overbrace{\begin{bmatrix} \Delta\underline{p}_1 \\ \Delta\underline{p}_2 \\ \vdots \\ \Delta\underline{p}_n \end{bmatrix}}^{\underline{\ell}} = \overbrace{\begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_n \end{bmatrix}}^{A} \underline{\mathbf{x}} + \overbrace{\begin{bmatrix} \underline{m}_1 \\ \underline{m}_2 \\ \vdots \\ \underline{m}_n \end{bmatrix}}^{\underline{m}}.$$

Assuming the observations to be statistically independent and of equal mean error[2] σ — what mathematicians call the *i.i.d.* assumption — allows us to use ordinary least squares to obtain the solution

[2] sigma $\sigma\Sigma$

$$\widehat{\mathbf{x}} = \left(A^{\mathsf{T}}A\right)^{-1} A^{\mathsf{T}}\underline{\boldsymbol{\ell}}$$

and its variance matrix

$$\mathrm{Var}\left\{\widehat{\mathbf{x}}\right\} = \sigma^2 \left(A^{\mathsf{T}}A\right)^{-1}.$$

It is worth noting that matrix $\left(A^{\mathsf{T}}A\right)^{-1}$ is completely determined by the satellite geometry; in other words, where every satellite is in the sky of the observer. All other factors affecting the uncertainty of the solution $\widehat{\mathbf{x}}$, such as the quality of the receiver and the circumstances of measurement, are (under the assumption made) contained in the single number σ, called the *mean error of unit weight*.

Matrix $\left(A^{\mathsf{T}}A\right)^{-1}$ informs us on the precision with which the unknowns can be determined: the variance of every unknown equals the corresponding diagonal element of $\left(A^{\mathsf{T}}A\right)^{-1}$ multiplied by σ^2. The

[2]This is physically unrealistic: the observational uncertainty is much larger for satellites close to the horizon, due to the longer signal path inside the atmosphere.

so-called DOP quantities often used for analysing the quality of the GNSS observation geometry are various sums of these diagonal elements.

In section 8.4 we will discuss the propagation in time of the state vector in the Kalman filter.

8.2.2 Estimating satellite locations

The momentary satellite position $\mathbf{x}$ and clock offset Δt are assumed known, meaning computable from the ephemeris. This assumption is not precisely true. We shall discuss in section 8.5 how to mitigate this assumption by using a differential technique with base stations.

tukiasema

A conceptual approach to this is to consider the base-station observations as observations to be used for determining the satellite unknowns x, y, z, and Δt, while the base-station co-ordinates X, Y, and Z are known. The H matrix for this problem is

$$H = \left[\begin{array}{ccc} -\dfrac{\mathbf{X}-\mathbf{x}}{\rho} & c & -c \end{array} \right]^{(0)}$$

with the corresponding state vector

$$\mathbf{x} = \left[\begin{array}{ccc} \Delta\mathbf{x} & \Delta T & \Delta t \end{array} \right]^{T}. \tag{8.11}$$

The estimates of the elements $\Delta\mathbf{x}$ and Δt of this state vector are then used to correct the broadcast ephemeris before using it in computing the location of the moving receiver or *rover*.

Equation 8.11 is also the general form of the state vector when estimating satellite orbits using pseudorange observations from known tracking stations. However, if the stations have synchronised atomic clocks, the station clock unknowns ΔT will drop out.

8.2.3 The carrier-phase observable

For the carrier-phase observable $P \stackrel{\text{def}}{=} \overline{P} - \lambda N$, equation 8.4, we also have to include the integer ambiguity N in the linearised observation equation, leading to

$$\begin{aligned} H &= \left[\begin{array}{ccccc} \dfrac{\partial P}{\partial X} & \dfrac{\partial P}{\partial Y} & \dfrac{\partial P}{\partial Z} & \dfrac{\partial P}{\partial \Delta T} & \dfrac{\partial P}{\partial N} \end{array} \right]^{(0)} = \\ &= \left[\begin{array}{ccccc} \dfrac{X-x}{\rho} & \dfrac{Y-y}{\rho} & \dfrac{Z-z}{\rho} & c & -\lambda \end{array} \right]^{(0)} = \left[\begin{array}{ccc} \dfrac{\mathbf{X}-\mathbf{x}}{\rho} & c & -\lambda \end{array} \right]^{(0)}. \end{aligned}$$

The corresponding state vector is

$$\mathbf{x} = \begin{bmatrix} \Delta X & \Delta Y & \Delta Z & \Delta T & N \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \Delta \mathbf{X} \\ \Delta T \\ N \end{bmatrix}.$$

In all cases, one linearised observation equation must be formed for every satellite and for every carrier frequency (for GPS, L1, L2 or L5) of observation.

Of course it is allowable to process these observations in the Kalman filter individually and sequentially, even if they have been made simultaneously — provided they are statistically independent of each other.

8.3 Atmospheric modelling

Atmospheric propagation delays are commonly divided into delays due to the "ionosphere" — more precisely, the ionised fraction of the atmosphere — and the "troposphere" — more precisely the neutral atmosphere, to which the stratosphere also belongs.

The mechanisms involved are very different.

8.3.1 The ionosphere

In the ionosphere, it is the free electrons that impact propagation most. To good approximation, the delay is inversely proportional to the square of the frequency f:

$$d_{\text{ion}} = 40.3\,\text{m}^3\,\text{Hz}^2 \cdot \frac{1}{f^2} \int N_e \, ds, \tag{8.12}$$

in which N_e is the density of free electrons in electrons per m^3, which is integrated along the ray path. The expression

$$\text{TEC} \stackrel{\text{def}}{=} 10^{-16}\,\text{TECU}\,\text{m}^2 \cdot \int N_e \, ds \tag{8.13}$$

is called the *total electron content*, in units of TECU, 10^{16} electrons per square metre. Then, equation 8.12 becomes

$$d_{\text{ion}} = 40.3 \cdot 10^{16}\,{}^{\text{m}\,\text{Hz}^2}/_{\text{TECU}}\,\frac{1}{f^2}\text{TEC},$$

in which the frequency f is to be given in hertz. So, in metres

$$\begin{aligned} d_{\text{ion},1} &= 0.1624\,{}^{\text{m}}/_{\text{TECU}} \cdot \text{TEC}, \\ d_{\text{ion},2} &= 0.2674\,{}^{\text{m}}/_{\text{TECU}} \cdot \text{TEC}, \\ d_{\text{ion},5} &= 0.2912\,{}^{\text{m}}/_{\text{TECU}} \cdot \text{TEC} \end{aligned}$$

for the frequencies of the GPS.

The ionosphere is *dispersive*: the propagation delay depends on the frequency f and is therefore different for modulations — the so-called group delay — and for the carrier phase — the phase delay. In fact, the 3 carrier phase has a negative delay:[3]

$$D_{\text{ion}} = -d_{\text{ion}}.$$

The dispersive nature of the ionospheric propagation delay is what makes it possible to eliminate — or determine! — it using dual- or multiple-frequency GNSS receivers. It should be taken into account in formulating the observation equations: for example equation 8.1 becomes for a dual-frequency receiver

$$\begin{aligned} p_1 &= \rho + c\,(\Delta T - \Delta t) + d_{\text{ion},1} + d_{\text{trop}}, \\ p_2 &= \rho + c\,(\Delta T - \Delta t) + \left(\frac{f_1}{f_2}\right)^2 d_{\text{ion},1} + d_{\text{trop}}. \end{aligned}$$

From this, the shared unknown $d_{\text{ion},1}$ is either resolvable or can be eliminated. Define the ionosphere-free observable

$$p_{\cancel{\text{ion}}} \stackrel{\text{def}}{=} \frac{f_1^2 p_1 - f_2^2 p_2}{f_1^2 - f_2^2} = \rho + c\,(\Delta T - \Delta t) + d_{\text{trop}},$$

from which $d_{\text{ion},1}$ has been eliminated.

8.3.2 The troposphere

The troposphere is not dispersive: the propagation delay is independent of frequency. It can be expressed into the refractive index of air:

$$d_{\text{trop}} = D_{\text{trop}} = \int (n-1)\,ds = 10^{-6} \int N\,ds. \tag{8.14}$$

The more manageable $N \stackrel{\text{def}}{=} 10^6\,(n-1)$ is often used instead of the refractive index n. The refractive index depends on the temperature T, pressure p, and absolute humidity or water-vapour partial pressure e (Rüeger, 2002):

$$N = \frac{77.624\,\mathrm{K/hPa}}{T}(p - e) + \frac{64.70\,\mathrm{K/hPa}}{T}\left(1 + \frac{5748\,\mathrm{K}}{T}\right) e.$$

[3] So, the wave crests of the carrier move faster than light in vacuum! This is not a problem, as they are no material objects and do not carry information. They are more like the "cutting point" of a pair of scissors, which may also move faster than the blades (Wikipedia, Superluminal motion).

If values for these three are available along the path, the propagation delay is obtained from the above path integral 8.14. Standard models, like Hopfield and Saastamoinen, are simplified approximations to this, accepting ground measurements as input.

A special feature of the troposphere is the disproportionate impact of *water vapour* on the propagation delay: a given partial pressure of water vapour has approximately 18 times the impact of the other, "dry" gases in the atmosphere. This is related to the chemical polarity of the water molecule, making it interact more strongly with microwaves — like in a microwave oven. poolisuus

8.3.3 *The mapping function*

The closer to the horizon the GNSS satellite is in the local sky, the longer the signal path through the atmosphere will be. This dependence is irrelevant for modelling the propagation delay as a property of the *atmosphere*. Therefore the parameter d, which represents the propagation delay along a slant path through the atmosphere, is replaced as an unknown or state-vector element by the corresponding delay at the same location along a vertical path:

$$d = \mu(\zeta)\, d^{\perp}.$$

Here, $d^{\perp}$ is the *total zenith delay* and $\mu(\zeta)$ the so-called *mapping function*. mu μM

For a flat Earth, the mapping function will be $\mu(\zeta) = {}^{1}\!/\!_{\cos\zeta}$, with ζ the zenith angle of the satellite in the local sky. For the real, curved Earth, more precise mapping functions exist which take into account the interaction of the Earth's curvature with the vertical distribution of refracting matter. zeta ζZ

These new atmospheric delay parameters $d^{\perp}$ may be added to equation 8.2, changing it to

$$p = \rho + c\,(\Delta T - \Delta t) + \mu_{\text{ion}}\, d^{\perp}_{\text{ion}} + \mu_{\text{trop}}\, d^{\perp}_{\text{trop}}. \tag{8.15}$$

Estimating these new parameters will also give a new state vector $\mathbf{x}$ and observation matrix H, for the case of estimating receiver locations, equations 8.9 and 8.8:

$$\mathbf{x} \stackrel{\text{def}}{=} \begin{bmatrix} \Delta\mathbf{X} & \Delta T & d^{\perp}_{\text{ion}} & d^{\perp}_{\text{trop}} \end{bmatrix}^{\mathsf{T}}, \quad H = \begin{bmatrix} \dfrac{\mathbf{X}-\mathbf{x}}{\rho} & c & \mu_{\text{ion}} & \mu_{\text{trop}} \end{bmatrix}^{(0)}.$$

8.3.4 Multiple satellites

If there are multiple satellites within view at different elevation angles, this becomes

$$H = \begin{bmatrix} \dfrac{\mathbf{X}-\mathbf{x}^1}{\rho^1} & c & \mu^1_{\mathrm{ion}} & \mu^1_{\mathrm{trop}} \\ \dfrac{\mathbf{X}-\mathbf{x}^2}{\rho^2} & c & \mu^2_{\mathrm{ion}} & \mu^2_{\mathrm{trop}} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\mathbf{X}-\mathbf{x}^n}{\rho^n} & c & \mu^n_{\mathrm{ion}} & \mu^n_{\mathrm{trop}} \end{bmatrix}^{(0)},$$

with $\mu^S \stackrel{\mathrm{def}}{=} \mu(\zeta^S)$ for satellite S, different values for different satellites. The unknown zenith delays $d^\perp$ are common to all satellites and in principle resolvable in the Kalman filter.

If there are multiple sensors, of course observation equations must be formulated for all of them. If, for example, a GNSS receiver is able to measure on multiple frequencies, it will be possible to determine and eliminate precisely the propagation delay of the ionosphere. For a dual-frequency receiver and multiple satellites, the result is

$$\begin{aligned} p^1_{\cancel{\mathrm{ion}}} &= \rho^1 + c\left(\Delta T - \Delta t^1\right) + \mu^1_{\mathrm{trop}}\, d^\perp_{\mathrm{trop}}, \\ p^2_{\cancel{\mathrm{ion}}} &= \rho^2 + c\left(\Delta T - \Delta t^2\right) + \mu^2_{\mathrm{trop}}\, d^\perp_{\mathrm{trop}}, \\ &\cdots \\ p^n_{\cancel{\mathrm{ion}}} &= \rho^n + c\left(\Delta T - \Delta t^n\right) + \mu^n_{\mathrm{trop}}\, d^\perp_{\mathrm{trop}}, \end{aligned}$$

with the state vector $\mathbf{x}$ and the H matrix

$$\mathbf{x} = \begin{bmatrix} \Delta\mathbf{X} \\ \Delta T \\ d^\perp_{\mathrm{trop}} \end{bmatrix}, \qquad H = \begin{bmatrix} \dfrac{\mathbf{X}-\mathbf{x}^1}{\rho^1} & c & \mu^1_{\mathrm{trop}} \\ \dfrac{\mathbf{X}-\mathbf{x}^2}{\rho^2} & c & \mu^2_{\mathrm{trop}} \\ \vdots & \vdots & \vdots \\ \dfrac{\mathbf{X}-\mathbf{x}^n}{\rho^n} & c & \mu^n_{\mathrm{trop}} \end{bmatrix}^{(0)}.$$

There is only one atmospheric unknown left here, $d^\perp_{\mathrm{trop}}$. Because $\zeta^i \neq \zeta^j \implies \mu^i_{\mathrm{trop}} \neq \mu^j_{\mathrm{trop}}$, it should be possible to estimate it with enough satellites at different elevations.

Here it is assumed that the unknown $d^\perp_{\mathrm{trop}}$ is the same for all satellites and equal to the zenith path integral straight up from the receiver. This holds for a perfectly stratified atmosphere but is of course not precisely true for the real atmosphere, where the amount of refractive matter is not constant in the horizontal direction.

To address this, $d^\perp$ may be developed into a Taylor series:

$$d^\perp = d_0^\perp + \tan\zeta\,\cos\alpha\, d_{\text{north}}^\perp + \tan\zeta\,\sin\alpha\, d_{\text{east}}^\perp + \cdots,$$

in which ζ is the zenith angle and α the azimuth of the satellite. Now, instead of a single unknown to be estimated, there are three of them: $d_0^\perp$, $d_{\text{north}}^\perp$, and $d_{\text{east}}^\perp$, which can still be estimated with a single receiver at a single epoch if there are enough visible satellites. The state vector then becomes alpha αA

$$\mathbf{x} = \begin{bmatrix} \Delta\mathbf{X} & \Delta T & d_{\text{trop},0}^\perp & d_{\text{trop, north}}^\perp & d_{\text{trop, east}}^\perp \end{bmatrix}^T$$

and the observation matrix

$$H = \begin{bmatrix} \dfrac{\mathbf{X}-\mathbf{x}^1}{\rho^1} & c & \mu_{\text{trop}}^1 & \tan\zeta^1\cos\alpha^1\mu_{\text{trop}}^1 & \tan\zeta^1\sin\alpha^1\mu_{\text{trop}}^1 \\ \dfrac{\mathbf{X}-\mathbf{x}^2}{\rho^2} & c & \mu_{\text{trop}}^2 & \tan\zeta^2\cos\alpha^2\mu_{\text{trop}}^2 & \tan\zeta^2\sin\alpha^2\mu_{\text{trop}}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dfrac{\mathbf{X}-\mathbf{x}^n}{\rho^n} & c & \mu_{\text{trop}}^n & \tan\zeta^n\cos\alpha^n\mu_{\text{trop}}^n & \tan\zeta^n\sin\alpha^n\mu_{\text{trop}}^n \end{bmatrix}^{(0)}.$$

This approach is being used in GNSS processing software in modelling the tropospheric delay.

8.3.5 *Use of reference models*

A significant improvement is obtained by using *residual* ionosphere or troposphere corrections: differences relative to some suitable *a priori* model. The notation then becomes

$$\mu_{\text{ion}}\,\Delta d_{\text{ion}}^\perp = d_{\text{ion}} - d_{\text{ion}}^{(0)}, \qquad \mu_{\text{trop}}\,\Delta d_{\text{trop}}^\perp = d_{\text{trop}} - d_{\text{trop}}^{(0)},$$

in which $d_{\text{ion}}^{(0)}$ and $d_{\text{trop}}^{(0)}$ are values produced by the models.

For the ionosphere, the *a priori* model could be the Klobuchar model included with the satellite navigation message, which is very simple. Satellite-based augmentation systems (SBAS) broadcast ionosphere models for their areas of coverage, chapter 10, as do, implicitly, base-station networks for RTK using the MAC/MAX protocols as well, section 9.4.

For the troposphere, the known Hopfield or Saastamoinen models may be used. These models expect values for meteorological parameters like temperature, pressure and humidity at the measurement location. Measured values at the measurement location may be used for this, if they exist. Alternatively, interpolated values from nearby weather stations may be used — considering appropriately the vertical gradients of many parameters. And as a last resort, climatic averages for the measurement location may be used.

Then, equation 8.15 becomes

$$p = \rho + c\,(\Delta T - \Delta t) + d_{\text{ion}}^{(0)} + d_{\text{trop}}^{(0)} + \mu_{\text{ion}}\,\Delta d_{\text{ion}}^{\perp} + \mu_{\text{trop}}\,\Delta d_{\text{trop}}^{\perp}. \tag{8.16}$$

It is often advisable to split the tropospheric path delay into a "dry" and a "wet" part $\Delta d_{\text{dry}}^{\perp}$ and $\Delta d_{\text{wet}}^{\perp}$, as these will have different mapping functions μ_{dry} and μ_{wet}. This is especially the case if an air-pressure sensor is present, as is the case on aircraft. Air pressure is informative on the total amount of refractive matter in a column above the observation location, which helps to determine the dry and wet parts separately.

8.4 Dynamic models for various estimation problems

8.4.1 Satellite orbit determination

Specialise equation 8.16 to the following observation equation:

$$\begin{aligned} \underline{p} &= \underline{\rho} + c\,(\Delta T - \underline{\Delta t}) + d_{\text{ion}}^{(0)} + d_{\text{trop}}^{(0)} + \mu_{\text{ion}}\,\Delta \underline{d}_{\text{ion}}^{\perp} + \mu_{\text{trop}}\,\Delta \underline{d}_{\text{trop}}^{\perp} + \underline{m}, \\ \underline{\rho} &= \sqrt{(X - \underline{x})^2 + \left(Y - \underline{y}\right)^2 + (Z - \underline{z})^2}. \end{aligned}$$

Here $\underline{m}$ represents the observation noise or random error.

This is the observation equation for *orbit determination*. In it, the ground or tracking-station position is given and treated as non-stochastic: $\mathbf{X} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$.

In precise orbit determination, these co-ordinates may not be considered constant. It will be clear that in an inertial reference frame, the rotation of the Earth needs to be modelled. But Earth orientation parameters (EOP), like polar motion and length-of-day variations, must also be included. These time series are available on the Internet. Individual station motions due for example to plate tectonics or glacial isostatic adjustment (GIA) also need to be taken into account.

vertauskehys

The satellite position $\underline{\mathbf{x}} = \underline{x}\mathbf{i} + \underline{y}\mathbf{j} + \underline{z}\mathbf{k}$ is stochastic and is estimated by the filter.

The tracking-station clock is assumed to be known relative to GNSS system time, the offset being ΔT. The satellite clock offset $\underline{\Delta t}$ is estimated.

For this situation, we identify the *state vector* as

$$\underline{\mathbf{x}} = \begin{bmatrix} \underline{\mathbf{x}} & \underline{\mathbf{v}} & \underline{\Delta t} & \Delta \underline{d}_{\text{ion}}^{\perp} & \Delta \underline{d}_{\text{trop}}^{\perp} \end{bmatrix}^{\mathsf{T}}.$$

As usual, we brought the velocity vector $\underline{\mathbf{v}}$ into the state vector, so we can write the Kalman dynamic equations as first-order differential equations.

Next, we have to decide how to model the time behaviour of each state-vector element. For the location $\underline{\mathbf{x}}$ this is simple: by definition

$$\frac{d}{dt}\underline{\mathbf{x}} = \underline{\mathbf{v}}$$

exactly. For the approximate or reference satellite velocity we need to use the exact equations of motion of Newtonian theory. Available orbit predictions may be used as approximate values. In navigation, only the broadcast ephemeris transmitted by the satellites in real time will do. Subtraction and *linearisation* will produce delta quantities. If these are small, the central force field approximation may be a good enough linear model for their propagation. tosiaikaisesti

Let the approximate values be $\mathbf{x}^{(0)}(t)$, $\mathbf{v}^{(0)}(t)$, and $\Delta t^{(0)}(t)$ — the GNSS ephemeris always also contains satellite clock offsets. Then, the linearised or differential state-vector elements ("delta quantities") are

$$\Delta\underline{\mathbf{x}} = \underline{\mathbf{x}} - \mathbf{x}^{(0)}, \quad \Delta\underline{\mathbf{v}} = \underline{\mathbf{v}} - \mathbf{v}^{(0)}, \quad \Delta(\underline{\Delta t}) = \underline{\Delta t} - \Delta t^{(0)}.$$

Now, the linearised dynamic model for $\underline{\mathbf{x}}$ and $\underline{\mathbf{v}}$ is

$$\frac{d}{dt}\overbrace{\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\end{bmatrix}}^{\mathsf{x}} = \overbrace{\begin{bmatrix}0 & I\\ \mathcal{M}^{(0)} & 0\end{bmatrix}}^{F}\overbrace{\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\end{bmatrix}}^{\mathsf{x}} + \overbrace{\begin{bmatrix}0\\ \underline{\mathbf{n}}_a\end{bmatrix}}^{\mathsf{n}},$$

in which $\mathcal{M}^{(0)}$ is the *gravitation-gradient tensor*, earlier derived for a central force field, equation 3.9. I is the unit matrix of size 3×3, and $\underline{\mathbf{n}}_a$ is the dynamic noise of satellite motion.

How does one model the behaviour of the satellite clock $\underline{\Delta t}$? Often, the model chosen is a *random walk* process, equation 2.25, section 7.8:

$$\frac{d}{dt}\Delta(\underline{\Delta t}) = \underline{n}_t.$$

The tropospheric and ionospheric propagation delays are represented by zenith delays $\Delta\underline{d}^{\perp}_{\text{ion}}$ and $\Delta\underline{d}^{\perp}_{\text{trop}}$ from which the effect of the slanting path direction within the atmosphere, which depends on the zenith angle of the GNSS satellite in the local sky, has been eliminated, and from which values given by a reference model have been subtracted. vertausmalli

In this case, a Gauss-Markov model is often used with a sensible choice of the time parameter τ. The parameter should be chosen to tau τT represent the time scale of atmospheric change, perhaps a few hours. It is wise to split the tropospheric path delay into a "dry" and a "wet" part and form separate dynamic models for them, as they behave differently over time.

A summary in the form of a Kalman filter dynamic model is

$$\frac{d}{dt}\overbrace{\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\\ \Delta(\underline{\Delta t})\\ \Delta\underline{d}^{\perp}_{\text{ion}}\\ \Delta\underline{d}^{\perp}_{\text{trop}}\end{bmatrix}}^{\mathbf{x}} = \overbrace{\begin{bmatrix}0 & I & & & \\ \mathcal{M}^{(0)} & 0 & & & \\ & & 0 & & \\ & & & -\frac{1}{\tau_{\text{ion}}} & \\ & & & & -\frac{1}{\tau_{\text{trop}}}\end{bmatrix}}^{F}\overbrace{\begin{bmatrix}\Delta\underline{\mathbf{x}}\\ \Delta\underline{\mathbf{v}}\\ \Delta(\underline{\Delta t})\\ \Delta\underline{d}^{\perp}_{\text{ion}}\\ \Delta\underline{d}^{\perp}_{\text{trop}}\end{bmatrix}}^{\mathbf{x}} + \overbrace{\begin{bmatrix}0\\ \underline{\mathbf{n}}_a\\ \underline{n}_t\\ \underline{n}_{\text{ion}}\\ \underline{n}_{\text{trop}}\end{bmatrix}}^{\mathbf{n}}.$$

Note that inclusion of elements like $\Delta\underline{d}^{\perp}_{\text{ion}}$ or $\Delta\underline{d}^{\perp}_{\text{trop}}$, or even $\Delta d^{\perp}_{\text{dry}}$ and $\Delta d^{\perp}_{\text{wet}}$, into the state vector does not guarantee that good estimates of these processes will be produced! It all depends on the types of observations available. For example using a dual-frequency GNSS receiver will allow elimination of $\Delta\underline{d}^{\perp}_{\text{ion}}$. Also, if independent estimates of $\Delta\underline{d}^{\perp}_{\text{trop}}$ are available in real time, like from a GNSS tracking network in the area, then these may be used to eliminate the element from the state vector.

If none of this applies, then the estimate for the state-vector element will continue to have a large uncertainty, with strong correlation with other, similarly uncertain state-vector elements, which is numerically unpleasant.

8.4.2 Station location determination

Starting from the same equation 8.16, we form a different observation equation:

$$\underline{p} = \underline{\rho} + c\,(\underline{\Delta T} - \Delta t) + d^{(0)}_{\text{ion}} + d^{(0)}_{\text{trop}} + \mu_{\text{ion}}\,\Delta\underline{d}^{\perp}_{\text{ion}} + \mu_{\text{trop}}\,\Delta\underline{d}^{\perp}_{\text{trop}} + \underline{m},$$
$$\underline{\rho} = \sqrt{(\underline{X}-x)^2 + (\underline{Y}-y)^2 + (\underline{Z}-z)^2}.$$

This is the observation equation for *geodetic positioning*. Here, the satellite orbital elements and clock offset are assumed known: $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and Δt are precisely computable from the available ephemeris. Now the *state vector* is

$$\underline{\mathbf{x}} = \begin{bmatrix}\underline{\mathbf{X}} & \underline{\mathbf{V}} & \underline{\Delta T} & \Delta\underline{d}^{\perp}_{\text{ion}} & \Delta\underline{d}^{\perp}_{\text{trop}}\end{bmatrix}^T,$$

in which $\underline{\mathbf{V}}$ is the velocity vector of station motion. Here, the new problem is to model the behaviour of the vectors $\underline{\mathbf{X}}(t)$ and $\underline{\mathbf{V}}(t)$ of the ground station.

In the case that the ground station is fixed, we may choose as the model, in a co-ordinate frame co-rotating with the Earth,

$$\underline{\mathbf{V}} = \frac{d}{dt}\underline{\mathbf{X}} = 0.$$

If we know that the stations are moving, but slowly and with constant velocity — for example due to plate tectonics or glacial isostatic adjustment — we may write

$$\frac{d}{dt}\overbrace{\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \end{bmatrix}}^{\underline{x}} = \overbrace{\begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}}^{F}\overbrace{\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \end{bmatrix}}^{\underline{x}} + \overbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}^{\underline{n}}.$$

The Kalman filter will gradually improve the estimators $\widehat{\mathbf{X}}, \widehat{\mathbf{V}}$ over time as more observations $\underline{p}$ are processed.

Some existing GNSS processing software (GYPSY/OASIS) uses the Kalman filter in this way.

The same equations may also be used for *moving vehicles* — for example aircraft — but with complications. One could use the knowledge that the *acceleration* $\mathbf{A}$ of the vehicle is bounded and model it as coloured noise, for example as a Gauss-Markov process. According to equation 2.31, the variance of such a process is $Q_n/2k$, when the process equation is

$$\frac{d}{dt}\underline{\mathbf{A}} = -k\,\underline{\mathbf{A}} + \underline{\mathbf{n}}_A.$$

Now let $\tau_A = {}^1/k$ be the time constant of the motion. Typically it will be seconds, the time scale of vehicle motions. Also, let α be the typical order of magnitude of the accelerations occurring. By putting

$$\frac{Q_n}{2k} = \tfrac{1}{2}Q_n\tau_A = \alpha^2$$

we obtain for the variance of the driving noise $\underline{\mathbf{n}}_A$:

$$\mathrm{Var}\{\|\underline{\mathbf{n}}_A\|\} = E\{\|\underline{\mathbf{n}}_A\|^2\} = Q_n = \frac{2\alpha^2}{\tau_A}.$$

Thus we get as the complete dynamic model, also adding $\underline{\Delta T}$, $\Delta\underline{d}^{\perp}_{\mathrm{ion}}$, and $\Delta \underline{d}^{\perp}_{\text{trop}}$:

$$\frac{d}{dt}\overbrace{\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \\ \underline{\mathbf{A}} \\ \Delta \underline{T} \\ \Delta \underline{d}^{\perp}_{\text{ion}} \\ \Delta \underline{d}^{\perp}_{\text{trop}} \end{bmatrix}}^{\underline{x}} = \overbrace{\begin{bmatrix} 0 & I & & & & \\ & 0 & I & & & \\ & & -\frac{1}{\tau_A} I & & & \\ & & & 0 & & \\ & & & & -\frac{1}{\tau_{\text{ion}}} & \\ & & & & & -\frac{1}{\tau_{\text{trop}}} \end{bmatrix}}^{F} \overbrace{\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \\ \underline{\mathbf{A}} \\ \Delta \underline{T} \\ \Delta \underline{d}^{\perp}_{\text{ion}} \\ \Delta \underline{d}^{\perp}_{\text{trop}} \end{bmatrix}}^{\underline{x}} + \overbrace{\begin{bmatrix} 0 \\ 0 \\ \frac{2\alpha^2}{\tau_A} \underline{\mathbf{n}}_1 \\ \underline{n}_t \\ \underline{n}_{\text{ion}} \\ \underline{n}_{\text{trop}} \end{bmatrix}}^{\underline{n}}, \tag{8.17}$$

in which $\mathbf{X}$ and $\mathbf{V}$ are not linearised. If we wish to use pseudoranges as observations to update the state vector, the non-linear observation equation will have to be linearised, producing linearised $\Delta\mathbf{X}$ and $\Delta\mathbf{V}$, also to be used in dynamic model 8.17. We leave this as an exercise for the reader.

The noise vector $\underline{\mathbf{n}}_1$ stands for "unit-variance white noise", a three-dimensional vector in this case.

Both α and τ_A will depend on the type of vehicle. Large α and short τ_A is often referred to as a "high-dynamics" environment. Such an environment poses challenges for designing GNSS receivers due to rapid and unpredictable variations in Doppler shifts in signal reception frequency.

8.5 Differential positioning

A common problem with both code measurement and carrier-phase measurement is the imprecision of the satellite orbit predictions. In real time using the broadcast ephemeris transmitted by the satellites one can achieve a precision of ± 1 m in the three-dimensional satellite position.[4]

A solution to this is offered by the differential technique, meaning the use of a base station or a network of base stations. If the distance from the base station is sufficiently small, most of the orbit error will cancel out from the final result, see figure 8.1.

[4] However, the International GNSS Service IGS publishes "ultra-rapid" orbit predictions four times daily, which have an estimated precision of ± 5 cm and are available in real time (International GNSS Service). The precision estimate represents the mean error of each co-ordinate, σ_x, σ_y, and σ_z.

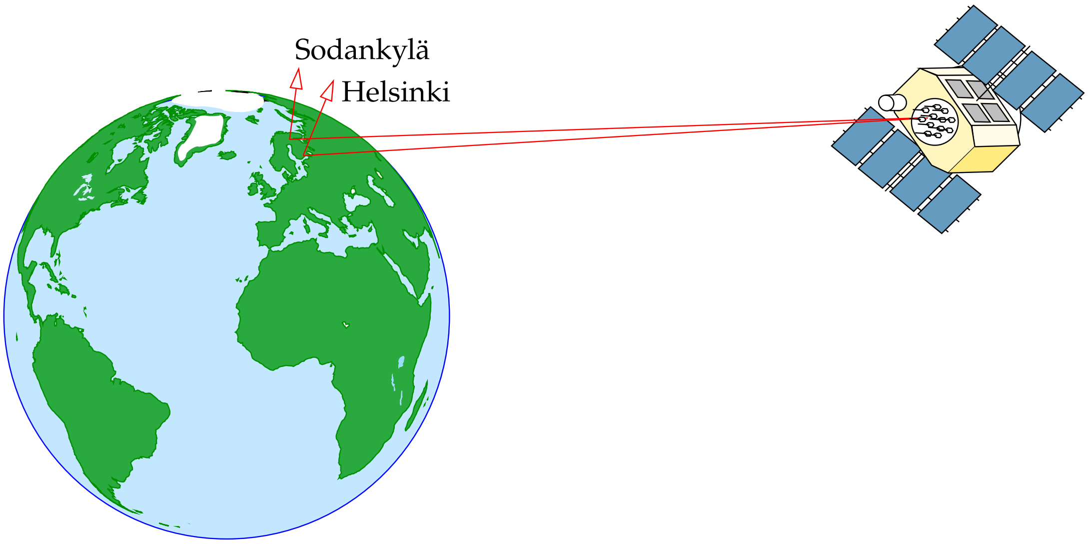


FIGURE 8.1. Differential positioning. The distance between two ground stations, in this case Helsinki and Sodankylä, is always *small* compared to the satellite orbital height, 20 000 km. Therefore the orbit error cancels out for the most part in differential measurement. The satellite clock error cancels out entirely.

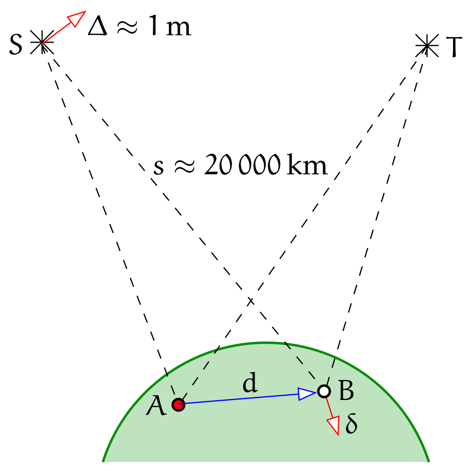
It is easy to make a back-of-the-envelope estimate of the precision of differential positioning using a geometric argument. See figure 8.2. If the geometric precision of the satellite orbit is called Δ and the distance of the satellite from the observer s — always longer than 20 000 km — we obtain for the positioning precision

$$\delta \approx \frac{d}{s}\Delta,$$

in which d is the length of the vector to be measured. Using this formula, the table in figure 8.2 is compiled.

Differential GNSS is widely used in geodetic GNSS processing. Every time software is used that builds so-called double-difference observables, the differential method is being used. Double differences are calculated by subtracting from each other not only the observations using two satellites but also the observations by two ground stations. This is how many of the sources of error in the inter-station vector solution are eliminated. The sources of error are in principle substantial, but change only slowly with place, such as

- orbit errors, satellite clock errors
- atmospheric signal propagation delays, divided into delays caused by the ionosphere and those caused by the troposphere



d (km)	δ (mm)
1	0.05
10	0.5
100	5
1000	50

FIGURE 8.2. Geometry of estimating the precision of differential positioning. The assumed orbit error $\Delta = 1$ m corresponds to the precision of today's broadcast ephemeris.

- errors caused by the antenna's phase-delay pattern depending on the direction (azimuth, elevation angle) to the satellite, and thus on the direction of the local plumb line. luotiviiva

8.6 Real-time kinematic positioning

We use the abbreviation RTK for real-time kinematic GNSS positioning.

The kinematic measuring method was invented by the American Benjamin Remondi. The method uses as an observable the carrier phase angle of the radio signal transmitted by a GNSS satellite. This observable suffers from the ambiguity problem: the phase is periodic, with a period of 2π. The measured phase angle is at the start of measurement assumed to be within the range of one cycle, $[0, 2\pi)$. Thereafter, as the range changes, the measured phase will track these changes including full cycles. The real range, however, differs from the range inferred from the phase by an integer number of whole cycles or wavelengths, and that integer cannot be determined by phase measurement alone.

During RTK measurement, the receiver remains "locked" to the phase of the GNSS carrier wave. As long as the lock holds and no "cycle slip" happens, the unknown integer associated with the phase measurement of the carrier wave does not change. vaihekatko

Now assume that somehow we can determine these integers, for example by visiting a known location, so that the ranges from the satellites can be calculated. After this resolution, the phase angles

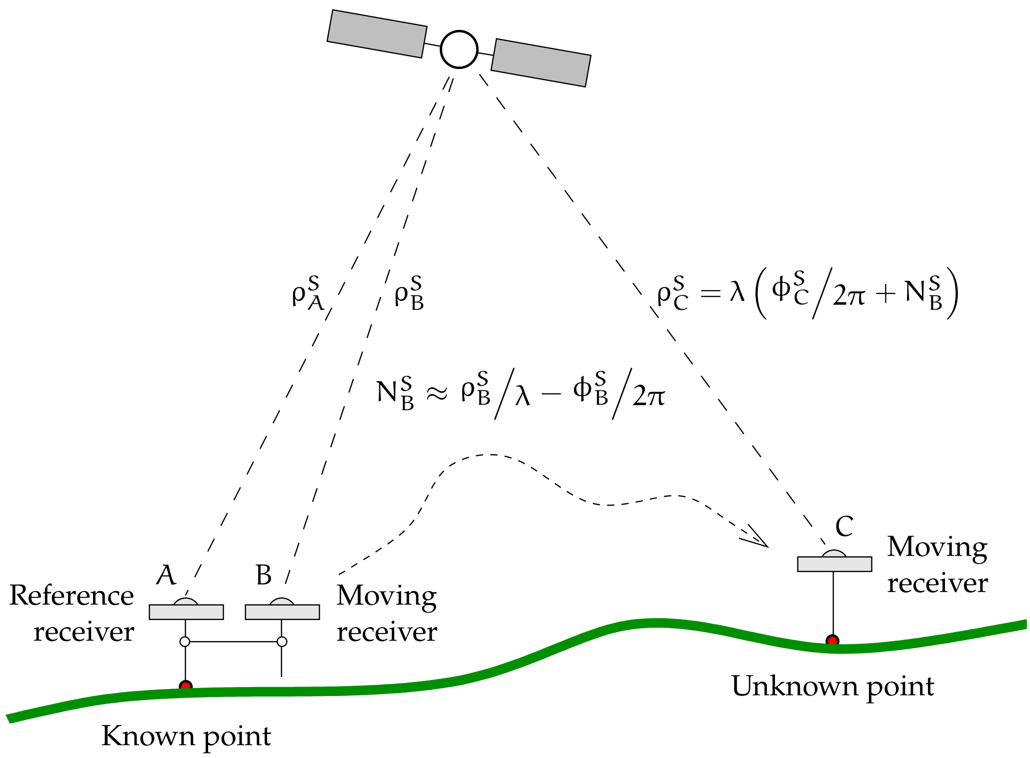


FIGURE 8.3. The idea of real-time kinematic GNSS positioning. Co-location allows the determination of ambiguity count N, allowing the determination of the true range to the satellite at unknown location C. It is assumed here, unrealistically, that all changes in the measured carrier phase ϕ are caused by changes in the geometric distance ρ between the satellite and receiver.

continuously tracked by the receiver now represent true ranges from the satellites.

After that, the receiver can be moved — this is where the "kinematic" comes in! — to other points, the unknown locations of which are then resolved in real time, if only in relation to the known initial point.

See figure 8.3.

The procedure described here does not yet involve the stationary reference receiver shown in figure 8.3. A reference receiver would not be needed if the error sources affecting the measurement of the range from the known location to the satellite could be assumed not to change over time. Of course in reality, they do.

The foremost error source is the offset, and especially the *drift*, of the kellon käynti
receiver clock — this is why the observables are called "pseudoranges" and not ranges. We shall see later in subsection 9.1.2 how to eliminate this error source by constructing difference observables.

Furthermore, there are satellite clock offsets and propagation delays along the changing signal path though the atmosphere. It is the changes over time of these errors caused by satellites and atmosphere that the reference receiver attempts to monitor, in order to allow the moving receiver to correct for them. Of course, this will only work in a limited neighbourhood of the reference receiver, where it "sees" the same satellite and atmospheric geometry as the moving receiver.

For real-time measurements, a *data link* is needed from the reference receiver to the moving receiver.

8.7 The data link

A *radio link* is used in real-time differential methods to transfer either the original observations or differential corrections from a "base station" at a known location to another, often moving, ground station at an unknown location. The various methods

- transfer either the phase of the carrier wave or the time-of-arrival offset of the pseudo-random noise (PRN) code modulated on the carrier wave
- vertausasema can use one reference station for an entire area or several stations by means of interpolation
- can interpolate separately for each user, who thus has to state their position: *singlecasting*, or let the users themselves interpolate: *broadcasting*. An in-beween method is *multicasting*, with separate transmissions to groups of users.

A radio broadcast network, a pair of radio modems, the mobile-phone network, or even a geostationary satellite can provide the data link. Coverage can be local (radio modems), national (the commercial ser-
5 vices Trimnet VRS™[5] and HxGN SmartNet™[6] in Finland), continental
6 (satellite-based augmentation systems), or global (GDGPS, Jet Propulsion Laboratory).

8.8 The RTCM standard

The Radio Technical Commission for Maritime Services (RTCM, http://www.rtcm.org/) is an independent organisation created in 1947. The

[5] VRS is a common-law trademark of Trimble Inc.

[6] HxGN SmartNet is a common-law trademark of Leica Geosystems AG.

member organisations number over a hundred, including manufacturers of radionavigation equipment, state agencies responsible for radiopositioning, shipbuilders, positioning service providers and academic institutions.

RTCM Special Commission 104 has designed the standard for a GPS differential data service bearing the name RTCM SC-104, RTCM-104, or "RTCM". A version in common use is RTCM 10402.3 or 2.3.

Message types are listed in tableau 8.1. Each message is a sequence of 30 bits.

The newest version is RTCM 10403.2 or 3.2, which however is not backwards-compatible with versions 2.x. It uses a more efficient data transfer mechanism than the 2.x protocol. Like versions 2.x, it is suitable for real-time kinematic measurement.

An important new development of the standard is MSM, Multiple Signal Message, which allows information related to all GNSS and SBAS systems to be presented and disseminated in real time in generic form, taking best advantage of the existence of multiple systems.

A nice summary of version 3 messages including MSM is given by the German research institute BKG: BKG, Ntrip and RTCM version 3.

There are many devices on the market that send and can use the message types listed in differential navigation, either by using the phases of the carrier waves (the RTK technique) or the pseudo-random codes modulated onto the carrier waves (the DGPS technique). In both cases, the navigation is *real-time*: the "age" of the position solution stays always below a specified limit.

The RTCM messages are sent on behalf of a *base station*. The location of the base station has been determined with geodetic precision using static GNSS positioning. Because the position is known, it is possible to calculate, using the orbit information transmitted by the satellites, what the pseudorange to each satellite *ought* to be.[7] Subtracting this value 7 from the the measured pseudorange yields the *correction*, to be coded into the message (message types 1, 20 and 21).[8] 8

[7]This logic does not consider the receiver clock offset. But when using double differences for positioning, as one does in real-time positioning, this does not matter as the offset is eliminated.

TABLEAU 8.1. Message types of the RTCM SC-104 format. Protocol version 2.3, lightly edited. An asterisk means: retired or not widely used.

Message type	Message title	Message type	Message title
1	DGPS corrections	20	RTK carrier-phase corrections
2	Delta DGPS corrections*	21	RTK pseudorange corrections
3	GNSS Reference station parameters	22	Extended reference station parameters*
4	Carrier surveying information	23	Antenna type definition record
5	GPS constellation health	24	Antenna reference point (ARP)
6	GPS null frame	25–26	Undefined
7	DGPS radiobeacon almanac*	27	Extended radiobeacon almanac
8	Pseudolite almanac*	31	Differential GLONASS corrections
9	High-rate DGPS corrections	32	GLONASS reference station parameters
10	P-code DGPS corrections*	33	GLONASS constellation health
11	C/A code L1/L2 delta corrections*	34	GLONASS differential corrections
12	Pseudolite station parameters*	35	GLONASS radiobeacon almanac
13	Ground transmitter parameters*	36	GLONASS special message
14	Surveying auxiliary message	37	GNSS system time offset
15	Ionospheric/tropospheric message	38–40	Undefined
16	GPS special message	41–43	Experimental (generic GNSS)
17	GPS ephemeris almanac	44–58	Undefined
18	Uncorrected carrier phases	59	Proprietary message
19	Uncorrected pseudoranges	60–63	Differential Loran C messages

The transmitted corrections are valid at the base station and a small area around it. The size of the area depends on the desired accuracy. Metre-level accuracy is obtained even hundreds of kilometres from the base station, but centimetre-level accuracy succeeds only out to about ten kilometres.

[8] In the case of RTK, often one rather transmits the original phase observations, types 18 and 19, but conceptually the matter is the same.

8.9 The NTRIP protocol

NTRIP stands for "Networked Transport of RTCM via Internet Protocol". It resulted from the realisation that, although differential corrections using the RTCM protocol can be transmitted using a pair of radio modems, the Internet offers a more attractive alternative. The mobile base-station network may be used for the transmission of corrections over the Internet to GNSS instruments operating in the field. The 3G, 4G, and 5G technology generations have ample data transfer capacity for this.

The protocol used is "streaming" based on the Hypertext Transfer Protocol (HTTP), which is already in massive use for transferring World-wide Web pages over the Internet. It has been adapted for the purpose of transferring GNSS data streams. It scales fairly well to providing multiple data streams to many users. Different data streams or sources are distinguished as named "mount points". NTRIP is in active development at the German federal research institute BKG, who maintain an open-source reference implementation. See BKG, NTRIP v. 1.0. suoratoisto

Self-test questions

1. What is the observation equation for code pseudorange measurements?
2. What is the observation equation for carrier-phase measurements?
3. What is a zenith propagation delay? What is a mapping function?
4. When determining the co-ordinates of ground stations using GNSS, what movements of these ground stations must be modelled?
5. How would one model the motions of a land vehicle or aircraft in a Kalman filter?
6. What does it mean when we say that the ionosphere is *dispersive*? How can the ionospheric propagation delay d_{ion} be eliminated from GNSS pseudorange observations?
7. Why is it a good idea to replace the atmospheric propagation delay unknown d by the zenith delay unknown $d^{\perp}$?
8. What reference models exist for calculating the ionospheric propagation delay of a GNSS signal?
9. What reference models exist for calculating the tropospheric

propagation delay? How are the models applied?

10. In what way does differential positioning improve positioning precision?
11. Why does real-time differential positioning require a data link?
12. What is a cycle slip?
13. Describe the RTCM standard. What types of observations does it include?
14. What is the added value offered by the NTRIP protocol?
15. What are the advantages and disadvantages of using the GNSS carrier phase as the observable, rather than the code pseudorange?

Exercise 8–1: Linearising the dynamic model 8.17

What would the equivalent linearised model look like? And what are the dynamic equations for the approximate values $\mathbf{X}^{(0)}$ and $\mathbf{V}^{(0)}$?

RTK navigation

9

9.1 RTK and ambiguities

9.1.1 *The simplest case*

Let us look at real-time kinematic measurement using two receivers, at first leaving atmospheric effects out of consideration, as allowed when the distance between receivers is short. Let A be the reference receiver on the base station and B the moving receiver or *rover*. We measure the phase of the carrier wave with both receivers. The measurement by the reference receiver is transferred over a data link to the rover, which calculates the difference of the phase measurements.

tosiaikainen

vertaus-vastaanotin

First, we measure the phase of the carrier wave *with both receivers on the known point*, the base station, see figure 8.3. From equation 8.4:

tukiasema

$$\frac{1}{2\pi}\phi^S_{A,1} + N^S_A = \frac{f}{c}\rho^S_A + f\tau^S_{A,1}, \qquad \frac{1}{2\pi}\phi^S_B + N^S_B = \frac{f}{c}\rho^S_B + f\tau^S_{B,1},$$

in which

$$\tau^S_{A,1} = \Delta T_A - \Delta t^S_1, \qquad \tau^S_{B,1} = \Delta T_B - \Delta t^S_1$$

are the differences between the receiver clock offsets ΔT_A and ΔT_B and the satellite clock offset Δt^S_1 for the same point in time. The subscript 1 refers to the initial situation with both receivers on the known point. $f = {}^{\omega}\!/_{2\pi} = {}^{c}\!/_{\lambda}$ is the linear frequency and c the speed of light.

omega $\omega\Omega$

The quantities $N^S_A, N^S_B \in \mathbb{Z}$ are unknown integer values, *ambiguities*. These are needed because the values ϕ can only be measured modulo 2π.

lambda $\lambda\Lambda$

kokonaisluku-tuntematon

phi $\varphi\phi\Phi$

Subtraction yields the difference between the phase measurements of the receivers at time 1, with $\rho^S_B = \rho^S_A$:

rho ρR

$$\frac{1}{2\pi}\phi^S_{AB} \stackrel{\text{def}}{=} \frac{1}{2\pi}\left(\phi^S_B - \phi^S_{A,1}\right) = f\,\Delta T_{AB} - \left(N^S_B - N^S_A\right), \qquad (9.1)$$

in which $\Delta T_{AB} = \Delta T_B - \Delta T_A$.

After that, the moving receiver is taken to the unknown point C, yielding

$$\frac{1}{2\pi}\phi^S_C + N^S_C = \frac{f}{c}\rho^S_C + f\tau^S_{C,2}, \qquad \frac{1}{2\pi}\phi^S_{A,2} + N^S_A = \frac{f}{c}\rho^S_A + f\tau^S_{A,2},$$

in which 2 refers to the new situation on the unknown point, and

$$\tau^S_{C,2} = \Delta T_C - \Delta t^S_2, \qquad \tau^S_{A,2} = \Delta T_A - \Delta t^S_2.$$

Subtraction yields

$$\begin{aligned}\frac{1}{2\pi}\phi^S_{AC} \overset{\text{def}}{=} \frac{1}{2\pi}\left(\phi^S_C - \phi^S_{A,2}\right) = \\ = \frac{f}{c}\left(\rho^S_C - \rho^S_A\right) + f\,\Delta T_{AC} - \left(N^S_C - N^S_A\right) = \\ = \frac{f}{c}\left(\rho^S_C - \rho^S_A\right) + f\,\Delta T_{AB} - \left(N^S_B - N^S_A\right). \quad (9.2)\end{aligned}$$

Here

$$\Delta T_{AC} = \tau^S_{C,2} - \tau^S_{A,2} = \Delta T_C - \Delta T_A \approx \Delta T_B - \Delta T_A = \Delta T_{AB}.$$

The following assumptions were made:

vaihekatko

1. No "cycle slip" has happened, so $N^S_C = N^S_B$.
2. The time elapsed is so short that $\Delta T_C \approx \Delta T_B = \Delta T_A + \Delta T_{AB}$, in which ΔT_{AB} is a *constant difference*, the clock-offset difference of the clocks of the two receivers. So

$$\tau^S_{B,1} = \tau^S_{A,1} + \Delta T_{AB}, \quad \tau^S_{C,2} = \tau^S_{A,2} + \Delta T_{AB}.$$

3. At the known point, the reference and moving receivers are in the same place,[1] so that $\rho^S_A = \rho^S_B$.

1

In equations 9.1 and 9.2, the left-hand sides are *measured*. Their subtraction yields as an observation equation

$$\frac{1}{2\pi}\phi^S_{AC} - \frac{1}{2\pi}\phi^S_{AB} = \frac{f}{c}\left(\rho^S_C - \rho^S_A\right),$$

in which the left-hand side is an "observed" quantity, a difference of differences of phase measurements, and on the right-hand side, ρ^S_C is a function of the co-ordinates $\mathbf{X}_C$ of the rover, which are the unknowns

[1]More practically, the (small) difference between their locations is precisely known.

of this adjustment problem. Linearisation will yield an observation equation to be used, for example, in a Kalman filter.

Note that both the quantity ΔT_{AB}, the clock difference between the receivers, assumed constant, and the ambiguity difference $N_B^S - N_A^S$ have been eliminated. This holds for all satellites $S = 1, \ldots, n$.

In this approach, assumption 2 is especially problematic. In reality, receiver clock offsets ΔT are strongly time-dependent, unless the receivers are equipped with atomic clocks, which field-grade receivers are not.

9.1.2 Using double differences

In this geometry, it is tempting to use *double differences*; in other words, observables constructed by taking the difference between observations using two different satellites. Then, at the base station or reference receiver we obtain

$$\begin{aligned}\frac{1}{2\pi}\phi_{AB}^{ST} \stackrel{\text{def}}{=} \frac{1}{2\pi}\left(\phi_B^T - \phi_A^T\right) - \frac{1}{2\pi}\left(\phi_B^S - \phi_A^S\right) = \\ = -\left(\left(N_B^T - N_A^T\right) - \left(N_B^S - N_A^S\right)\right) + f\tau_{AB}^{ST}, \quad (9.3)\end{aligned}$$

in which

$$\begin{aligned}\tau_{AB}^{ST} = \tau_{AB}^{T} - \tau_{AB}^{S} = \\ = \left(\left(\Delta T_B - \Delta t_1^T\right) - \left(\Delta T_{A,1} - \Delta t_1^T\right)\right) - \\ - \left(\left(\Delta T_B - \Delta t_1^S\right) - \left(\Delta T_{A,1} - \Delta t_1^S\right)\right) = 0,\end{aligned}$$

and similarly at the rover or moving receiver

$$\begin{aligned}\frac{1}{2\pi}\phi_{AC}^{ST} \stackrel{\text{def}}{=} \frac{1}{2\pi}\left(\phi_C^T - \phi_A^T\right) - \frac{1}{2\pi}\left(\phi_C^S - \phi_A^S\right) = \\ = \frac{f}{c}\left(\left(\rho_C^T - \rho_A^T\right) - \left(\rho_C^S - \rho_A^S\right)\right) - \\ - \left(\left(N_B^T - N_A^T\right) - \left(N_B^S - N_A^S\right)\right) + f\tau_{AC}^{ST}, \quad (9.4)\end{aligned}$$

in which again

$$\begin{aligned}\tau_{AC}^{ST} = \tau_{AC}^{T} - \tau_{AC}^{S} = \\ = \left(\left(\Delta T_C - \Delta t_2^T\right) - \left(\Delta T_{A,2} - \Delta t_2^T\right)\right) - \\ - \left(\left(\Delta T_C - \Delta t_2^S\right) - \left(\Delta T_{A,2} - \Delta t_2^S\right)\right) = 0.\end{aligned}$$

So, we have gotten rid of all the clock unknowns, without any assumptions about the receiver clocks, such as $\Delta T_C \stackrel{?}{=} \Delta T_B$ or $\Delta T_{A,2} \stackrel{?}{=} \Delta T_{A,1}$. This is a major advantage of using double differences.

In this case, the quantity that is solved for by putting the reference receiver and the moving receiver side by side is for two satellites S and T:

$$N_{AB}^{ST} \stackrel{\text{def}}{=} \left(N_B^T - N_A^T\right) - \left(N_B^S - N_A^S\right).$$

This is an integer. "Observe" the double-difference quantities, equation 9.3:

$$\frac{1}{2\pi}\phi_{AB}^{ST} = -N_{AB}^{ST}, \tag{9.5}$$

to all satellite pairs $S = 1, \ldots, n$, $T = S + 1, \ldots, n$, with n the number of satellites, and we round the values obtained to the nearest integer.

These values can be used to calculate the quantities

$$\rho_{AC}^{ST} \stackrel{\text{def}}{=} \left(\rho_C^T - \rho_A^T\right) - \left(\rho_C^S - \rho_A^S\right)$$

from the observations ϕ_{AC}^{ST}. Equation 9.4 yields

$$\frac{1}{2\pi}\phi_{AC}^{ST} + N_{AB}^{ST} = \frac{f}{c}\rho_{AC}^{ST}. \tag{9.6}$$

Here, ρ_C^S and ρ_C^T are again functions of the co-ordinates $\mathbf{x}_C$ of the rover. These co-ordinates are to be solved after linearisation using a least-squares technique, for example the Kalman filter.

The easiest solution is the *float solution*. It is obtained by just calculating the ambiguities as real numbers using equation 9.5, and then calculating the solution using equation 9.6 as an observation equation.

However, a precision advantage can be obtained by first fixing the ambiguities N_{AB}^{ST} to their proper integer values using equation 9.5 — provided this can be done unambiguously with confidence — and only after that, solving for $\mathbf{x}_C$, the location of the rover, using equation 9.6. This is the more precise *fix solution*.

9.2 Fast ambiguity resolution

The measurement method described above requires that both before the field measurement — in which measurements are made on a number of unknown points — and after it as a double check, the moving receiver is placed next to the reference receiver, so-called *co-location*.

Often co-location is difficult: the reference receiver may be outside the measurement area and be run by an external service provider. This is one reason why *fast ambiguity resolution* was invented. No co-location with a reference receiver, or even a visit to a known point, is necessary — although it is always useful as a check.

The method works best if the distance between the reference and moving receivers is so short, some 10–20 km, that the between-receivers differential atmosphere and orbit errors can be ignored. In this case, equation 9.6 is

$$\frac{1}{2\pi}\phi_{AC}^{ST} + N_{AC}^{ST} = \frac{f}{c}\rho_{AC}^{ST},$$

in which

$$N_{AC}^{ST} = \left(N_C^T - N_A^T\right) - \left(N_C^S - N_A^S\right), \quad \rho_{AC}^{ST} = \left(\rho_C^T - \rho_A^T\right) - \left(\rho_C^S - \rho_A^S\right).$$

Here, the range double differences ρ_{AC}^{ST} are purely geometric. If we write out for the satellites $S = 1, \ldots, n$, $T = S+1, \ldots, n$:

$$\rho_C^S = \sqrt{\left(X_C - x^S\right)^2 + \left(Y_C - y^S\right)^2 + \left(Z_C - z^S\right)^2},$$
$$\rho_C^T = \sqrt{\left(X_C - x^T\right)^2 + \left(Y_C - y^T\right)^2 + \left(Z_C - z^T\right)^2},$$

we can see that the only unknowns are the location co-ordinates of the moving receiver

$$\mathbf{X}_C = \begin{bmatrix} X_C & Y_C & Z_C \end{bmatrix}^T.$$

The location of the moving receiver is always known with an accuracy of a couple of metres with the help of GNSS code measurement, which has no ambiguity problem. Then it suffices to find from the set of all possible receiver locations — a search space belonging to $\mathbb{R}^3$ — only the places $\mathbf{X}_C$ for which *all* values, for all satellite pairs S, T,

$$N_{AC}^{ST} = \frac{f}{c}\rho_{AC}^{ST} - \frac{1}{2\pi}\phi_{AC}^{ST}$$

are integers, see figure 9.1.

Conversely, if there are n satellites, there are $n-1$ independent ambiguity values N_{AC}^{ST}. These ambiguity combinations or double differences are thus the elements of an $n-1$ -dimensional space. If each ambiguity has, for example, ten different possible values that are compatible with the approximate position obtained from the code measurement, this already gives 10^{n-1} different ambiguity double

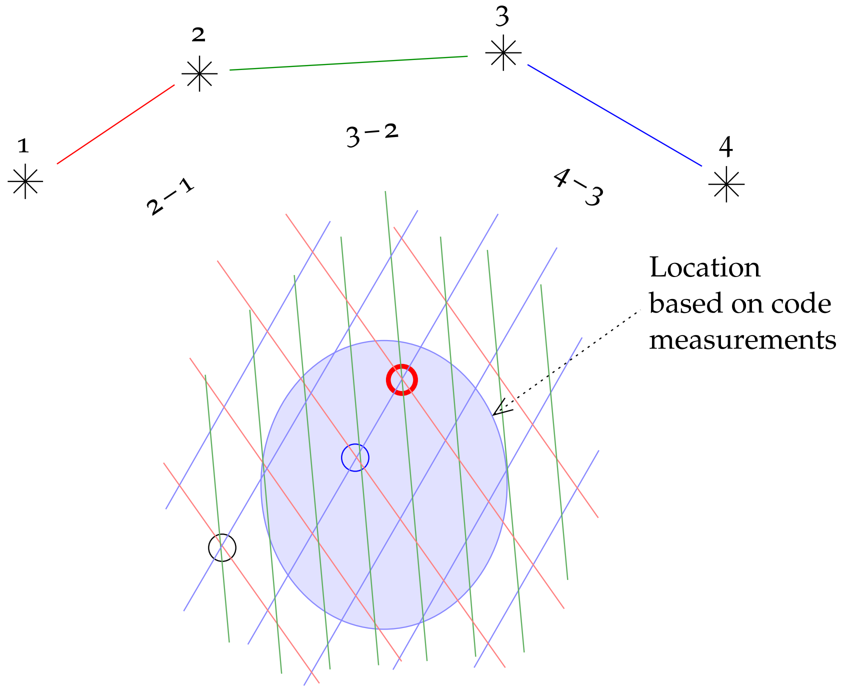


FIGURE 9.1. Ambiguity resolution. In this figure, the line bundles drawn (red, green and blue) represent the possible locations of the moving receiver that are compatible with the observed phase double differences. The ambiguity solution must be compatible with *all* double-difference observations (red circle), but also the contrast with the second-best fit (blue circle) should be clear.

tosiaikaisesti

differences. If there are eight satellites, this number is already ten million, probably too many possibilities to search in real time in a device with limited calculating power.

However, it is seen that of all the ambiguity alternatives, only a very small fraction are consistent with any *particular* position of the moving receiver: the consistent ambiguity combinations belong to *a three-dimensional subspace* of ambiguity space, one parametrisation of which is given by the rover co-ordinates $\begin{bmatrix} X_C & Y_C & Z_C \end{bmatrix}^T$, as already seen earlier.

Over the years, smart and efficient methods have been developed to resolve ambiguities in this internally consistent subspace, like the LAMBDA method (Least-squares AMBiguity Decorrelation Adjustment, Teunissen et al. 1997).

The ambiguity-resolution method described will succeed only if the distance between the reference and moving receivers is short enough, in practice under 10–20 km. In that case, we can take advantage of the fact that the satellites send their signals on two or even three different carrier frequencies, by constructing *widelane* observables. Ambiguity resolution is obtained immediately or after only a couple of epochs. leveäkuja

Ambiguity resolution is also possible for longer vectors, but this is more difficult, as the effect of the atmosphere has to be taken into account.

In the Kalman filter, the ambiguity is introduced to the state vector as an unknown N_{AC}^{ST}, but initially as a real-valued state. It is assumed that, in addition to the moving receiver C, there is a stationary reference receiver or network of receivers A in the area. The filter takes as input double-difference carrier-phase observations, type ϕ_{AC}^{ST}.

As the filter progresses in time, the state variance attached to N_{AC}^{ST} will become smaller and smaller, until it becomes possible to identify the real-valued ambiguity estimate with confidence with a single integer value.

In a practical situation, one will have not just one observation equation for each epoch — the double-difference equation 9.6 — but many. If the number of useable satellites in the sky is n, usually 5–12, the number of independent ambiguities N_{AC}^{ST} will be $n-1$. This set, or vector, of ambiguities will have a *state variance matrix* of size $(n-1)\times(n-1)$. Typically, the error ellipsoid of this matrix will be a very elongated "hyper-cigar", with three of its main axes long and the others short.

It needs to be analysed whether the whole *set*, or vector, of values N_{AC}^{ST} for all satellite pairs S, T lies near enough to a set of integer values, representing one point in a grid of points in the abstract vector space $\mathbb{R}^{n-1}$.

"Near enough" should be understood in terms of this state variance matrix. Let

$$\mathbf{v} \stackrel{\text{def}}{=} \begin{bmatrix} v_{AC}^{1,2} & v_{AC}^{1,3} & \cdots & v_{AC}^{1,n} \end{bmatrix}^{\mathsf{T}}, \quad v_{AC}^{ST} = \widehat{N}_{AC}^{ST} - N_{AC}^{ST},$$

be the vector of independent differences between the real-valued estimators

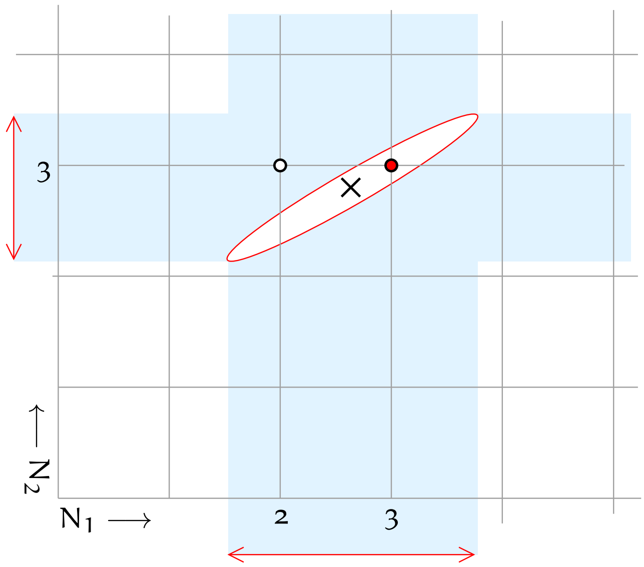
$$\widehat{\mathbf{N}} \stackrel{\text{def}}{=} \begin{bmatrix} \widehat{N}_{AC}^{1,2} & \widehat{N}_{AC}^{1,3} & \cdots & \widehat{N}_{AC}^{1,n} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{n-1}$$

and the integer-valued candidates

$$\mathbf{N} \stackrel{\text{def}}{=} \begin{bmatrix} N_{AC}^{1,2} & N_{AC}^{1,3} & \cdots & N_{AC}^{1,n} \end{bmatrix}^{\mathsf{T}} \in \mathbb{Z}^{n-1}.$$

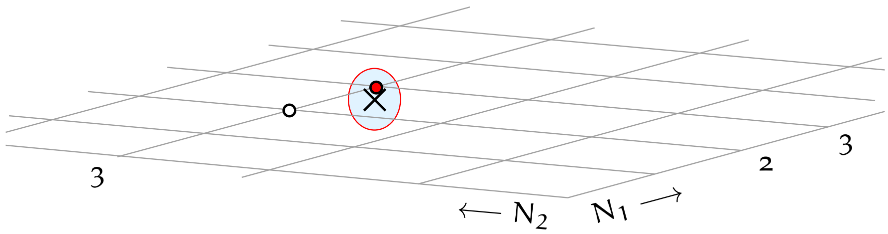
Then, the quantity serving as "distance" squared is

$$\underset{\sim}{\mathcal{E}} = \underline{\mathbf{v}}^{\mathsf{T}} \left(\mathrm{Var}\left\{\underline{\mathbf{v}}\right\}\right)^{-1} \underline{\mathbf{v}}.$$



(a)

The one-by-one ambiguity resolution method fails to resolve N_1 to 3



(b)

Multiple ambiguity state variance ellipsoids are often very elongated "cigars". The LAMBDA method resolves (N_1, N_2) to $(3, 3)$

FIGURE 9.2. Smart ambiguity resolution. In the LAMBDA method, the ambiguities are transformed into integer linear combinations that change the error ellipse to (almost) a circle. In the lower picture, the correct solution is easily seen to be the grid point nearest to the real-valued solution (cross).

This quantity is distributed according to the chi-squared distribution with $n - 1$ degrees of freedom, or χ^2_{n-1}. chi χX

So, the state variance matrix serves as a *metric tensor* for measuring the goodness of the ambiguity resolution.

Instead of using satellite 1 as the reference, other sets of $n - 1$ independent double differences, or their linear combinations, may also be used. Then, both $\underline{\mathbf{v}}$ and $\mathrm{Var}\{\underline{\mathbf{v}}\}$ will change, but $\underset{\sim}{\widetilde{\mathcal{E}}}$ is left invariant.

Not only should the solution be good enough, it should be sufficiently unique as well: there should be a clear *contrast* with the second-best solution. A technique for testing this is the Fisher F test.

Generally, this resolution of all ambiguities together will succeed well before they are successfully resolved one by one. Sophisticated algorithms have been developed for this, of which the LAMBDA method (TU Delft, LAMBDA) is an example.

9.3 A geometric analysis of RTK measurement

The real-time kinematic method may be used in two different ways or geometries:

1. by using one base station
2. by using a network of base stations.

In the following, the notation $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ denotes a triad of orthogonal unit vectors which form an orthonormal basis in three-dimensional space. Every vector in space can be written as a linear combination of these base vectors. ortonormaali kanta

First we look at how the orbit errors and clock offsets of GNSS satellites propagate into the position solution in the case of one base station; next we address the case of three base stations.

9.3.1 *One base station*

In the case of one base station we write the observable as follows, ignoring for the moment the ambiguities, the atmosphere, and other complicating factors:

$$\overline{P} = \sqrt{(X-x)^2 + (Y-y)^2 + (Z-z)^2} + c\,(\Delta T - \Delta t)\,.$$

Here, $\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ is the location vector of the satellite, $\mathbf{X} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$ that of the receiver, while $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ is an orthonormal basis, assumed to be geocentric. The quantity

$$\rho = \sqrt{(X-x)^2 + (Y-y)^2 + (Z-z)^2}$$

is the geometric distance between satellite and receiver.

Now we choose an *alternative* "line-of-sight" orthonormal basis in which $\widetilde{\mathbf{i}}$ points from the satellite to the base station, and $\widetilde{\mathbf{j}}$ and $\widetilde{\mathbf{k}}$ are orthogonal with respect to each other and $\widetilde{\mathbf{i}}$. These basis vectors may be kantavektori

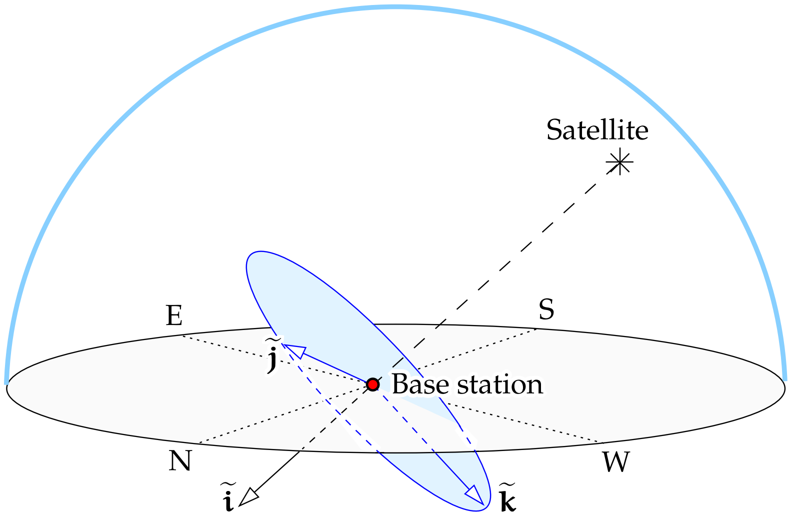


FIGURE 9.3. Orthonormal basis $\left\{\widetilde{\mathbf{i}}, \widetilde{\mathbf{j}}, \widetilde{\mathbf{k}}\right\}$ tied to the satellite line of sight.

constructed as follows:

$$\widetilde{\mathbf{i}} = \frac{\mathbf{X}_0 - \mathbf{x}}{\|\mathbf{X}_0 - \mathbf{x}\|}, \quad \widetilde{\mathbf{j}} = \frac{\left\langle \widetilde{\mathbf{i}} \times \mathbf{X}_0 \right\rangle}{\left\| \left\langle \widetilde{\mathbf{i}} \times \mathbf{X}_0 \right\rangle \right\|}, \quad \widetilde{\mathbf{k}} = \left\langle \widetilde{\mathbf{i}} \times \widetilde{\mathbf{j}} \right\rangle,$$

in which $\mathbf{X}_0$ is the geocentric location of the base station, a vector pointing in the direction of the zenith of the station. Note that the co-ordinates $\widetilde{X}$, $\widetilde{Y}$, and $\widetilde{Z}$ still have the geocentre as their origin:

$$\mathbf{X}_0 = X_0\mathbf{i} + Y_0\mathbf{j} + Z_0\mathbf{k} = \widetilde{X}_0\widetilde{\mathbf{i}} + \widetilde{Y}_0\widetilde{\mathbf{j}} + \widetilde{Z}_0\widetilde{\mathbf{k}},$$

with

$$\|\mathbf{X}_0\| = \sqrt{X_0^2 + Y_0^2 + Z_0^2} = \sqrt{\widetilde{X}_0^2 + \widetilde{Y}_0^2 + \widetilde{Z}_0^2} = R,$$

the radius of the Earth.

Let the effect of the orbital errors on the momentary location of the satellite in space be $\Delta\mathbf{x} = \Delta\widetilde{x}\,\widetilde{\mathbf{i}} + \Delta\widetilde{y}\,\widetilde{\mathbf{j}} + \Delta\widetilde{z}\,\widetilde{\mathbf{k}}$, and the satellite's clock offset Δt, already assumed to be small. Their effect on the pseudorange is

$$\Delta\overline{P} = \frac{\partial\overline{P}}{\partial\widetilde{x}}\Delta\widetilde{x} + \frac{\partial\overline{P}}{\partial\widetilde{y}}\Delta\widetilde{y} + \frac{\partial\overline{P}}{\partial\widetilde{z}}\Delta\widetilde{z} - c\Delta t. \tag{9.7}$$

Let

$$\rho_0 = \|\mathbf{X}_0 - \mathbf{x}\| = \sqrt{(X_0 - x)^2 + (Y_0 - y)^2 + (Z_0 - z)^2} =$$
$$= \sqrt{\left(\widetilde{X}_0 - \widetilde{x}\right)^2 + \left(\widetilde{Y}_0 - \widetilde{y}\right)^2 + \left(\widetilde{Z}_0 - \widetilde{z}\right)^2}$$

be the distance between satellite

$$\mathbf{x} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = \widetilde{x}\,\widetilde{\mathbf{i}} + \widetilde{y}\,\widetilde{\mathbf{j}} + \widetilde{z}\,\widetilde{\mathbf{k}}$$

and base station $\mathbf{X}_0$. Let furthermore the location of the moving receiver or "rover" be

$$\mathbf{X} = \widetilde{X}\widetilde{\mathbf{i}} + \widetilde{Y}\widetilde{\mathbf{j}} + \widetilde{Z}\widetilde{\mathbf{k}} = \left(\widetilde{X}_0 + \xi\right)\widetilde{\mathbf{i}} + \left(\widetilde{Y}_0 + \eta\right)\widetilde{\mathbf{j}} + \left(\widetilde{Z}_0 + \chi\right)\widetilde{\mathbf{k}}.$$

Here, ξ, η, and χ are now the co-ordinates of the rover relative to the base station in the frame agreed on above. The distance between the base station and rover is

xi $\xi\Xi$
eta ηH

$$s = \|\mathbf{s}\| = \sqrt{\xi^2 + \eta^2 + \chi^2}, \quad \mathbf{s} = \xi\widetilde{\mathbf{i}} + \eta\widetilde{\mathbf{j}} + \chi\widetilde{\mathbf{k}}.$$

Write out equation 9.7 separately for base station and rover:

$$\begin{aligned}\Delta\overline{P}_0 &= \frac{\widetilde{x} - \widetilde{X}_0}{\rho_0}\Delta\widetilde{x} + \frac{\widetilde{y} - \widetilde{Y}_0}{\rho_0}\Delta\widetilde{y} + \frac{\widetilde{z} - \widetilde{Z}_0}{\rho_0}\Delta\widetilde{z} - c\,\Delta t,\\ \Delta\overline{P} &= \frac{\widetilde{x} - \widetilde{X}}{\rho}\Delta\widetilde{x} + \frac{\widetilde{y} - \widetilde{Y}}{\rho}\Delta\widetilde{y} + \frac{\widetilde{z} - \widetilde{Z}}{\rho}\Delta\widetilde{z} - c\,\Delta t,\end{aligned}$$

in which

$$\begin{aligned}\rho = \|\mathbf{X} - \mathbf{x}\| &= \sqrt{(X - x)^2 + (Y - y)^2 + (Z - z)^2} =\\ &= \sqrt{\left(\widetilde{X} - \widetilde{x}\right)^2 + \left(\widetilde{Y} - \widetilde{y}\right)^2 + \left(\widetilde{Z} - \widetilde{z}\right)^2}.\end{aligned}$$

The difference between these is the error in the rover's determined location due to the distance from the base station:

$$\begin{aligned}\Delta\overline{P} - \Delta\overline{P}_0 = &\left(\frac{\widetilde{x} - \widetilde{X}}{\rho} - \frac{\widetilde{x} - \widetilde{X}_0}{\rho_0}\right)\Delta\widetilde{x} +\\ &+ \left(\frac{\widetilde{y} - \widetilde{Y}}{\rho} - \frac{\widetilde{y} - \widetilde{Y}_0}{\rho_0}\right)\Delta\widetilde{y} + \left(\frac{\widetilde{z} - \widetilde{Z}}{\rho} - \frac{\widetilde{z} - \widetilde{Z}_0}{\rho_0}\right)\Delta\widetilde{z}.\end{aligned}$$

The clock offset of the satellite has vanished from this.

Let us look closer at the coefficient

$$\left(\frac{\widetilde{x} - \widetilde{X}}{\rho} - \frac{\widetilde{x} - \widetilde{X}_0}{\rho_0}\right).$$

If we define the function

$$f\left(\widetilde{X}\right) \stackrel{\text{def}}{=} \frac{\widetilde{x} - \widetilde{X}}{\rho\left(\widetilde{x}, \widetilde{X}, \widetilde{y}, \widetilde{Y}, \widetilde{z}, \widetilde{Z}\right)},$$

this coefficient is

$$\begin{aligned}f\left(\widetilde{X}\right) - f\left(\widetilde{X}_0\right) &\approx \left.\frac{\partial f}{\partial \widetilde{X}}\right|_{\widetilde{X}=\widetilde{X}_0}\left(\widetilde{X} - \widetilde{X}_0\right) + \frac{1}{2}\left.\frac{\partial^2 f}{\partial \widetilde{X}^2}\right|_{\widetilde{X}=\widetilde{X}_0}\left(\widetilde{X} - \widetilde{X}_0\right)^2 + \cdots =\\ &= \left.\frac{\partial f}{\partial \widetilde{X}}\right|_{\widetilde{X}=\widetilde{X}_0}\xi + \frac{1}{2}\left.\frac{\partial^2 f}{\partial \widetilde{X}^2}\right|_{\widetilde{X}=\widetilde{X}_0}\xi^2 + \cdots,\end{aligned}$$

a Taylor expansion. Retaining only the first term yields with the Leibniz product rule

$$\left.\frac{\partial f}{\partial \widetilde{X}}\right|_{\widetilde{X}=\widetilde{X}_0} = -\frac{1}{\rho_0} + \frac{\left(\widetilde{x}-\widetilde{X}_0\right)^2}{\rho_0^3}, \tag{9.8}$$

and similarly for $f(\widetilde{Y})$ and $f(\widetilde{Z})$. We obtain

$$\Delta\overline{P} - \Delta\overline{P}_0 = \left(\frac{\left(\widetilde{x}-\widetilde{X}_0\right)^2}{\rho_0^3} - \frac{1}{\rho_0}\right)\xi\,\Delta\widetilde{x} + \\ + \left(\frac{\left(\widetilde{y}-\widetilde{Y}_0\right)^2}{\rho_0^3} - \frac{1}{\rho_0}\right)\eta\,\Delta\widetilde{y} + \left(\frac{\left(\widetilde{z}-\widetilde{Z}_0\right)^2}{\rho_0^3} - \frac{1}{\rho_0}\right)\chi\,\Delta\widetilde{z}.$$

As the co-ordinate axes are defined in the way described above, we have

$$\widetilde{x}-\widetilde{X}_0 = -\rho_0, \quad \widetilde{y}-\widetilde{Y}_0 = 0, \quad \widetilde{z}-\widetilde{Z}_0 = 0, \tag{9.9}$$

and we obtain

$$\Delta\overline{P} - \Delta\overline{P}_0 = -\frac{1}{\rho_0}\left(\eta\,\Delta\widetilde{y} + \chi\,\Delta\widetilde{z}\right). \tag{9.10}$$

We see that the error is linearly proportional to the distance from the base station, and that *only the distance sideways from the line of sight to the satellite has an effect.*

Note that the derivation of this equation is *symmetric* with respect to interchanging $\mathbf{X}$ and $\mathbf{X}_0$ in the definition of the basis $\left\{\widetilde{\mathbf{i}}, \widetilde{\mathbf{j}}, \widetilde{\mathbf{k}}\right\}$: we could have aligned the vector $\widetilde{\mathbf{i}}$ along the satellite line of sight to the rover location instead.

kanta

In reality, however, the pseudorange correction is not precisely linear: study the quadratic equation

$$f(\widetilde{X}) - f(\widetilde{X}_0) = \left.\frac{\partial f}{\partial \widetilde{X}}\right|_{\widetilde{X}=\widetilde{X}_0}\xi + \tfrac{1}{2}\left.\frac{\partial^2 f}{\partial \widetilde{X}^2}\right|_{\widetilde{X}=\widetilde{X}_0}\xi^2 + \cdots$$

We already derived in connection with equation 9.8:

$$\frac{\partial f}{\partial \widetilde{X}} = -\frac{1}{\rho} + \frac{\left(\widetilde{x}-\widetilde{X}\right)^2}{\rho^3}.$$

Differentiate once more:

$$\left.\frac{\partial^2 f}{\partial \widetilde{X}^2}\right|_{\widetilde{X}=\widetilde{X}_0} = \left.\left(-\frac{\widetilde{x}-\widetilde{X}}{\rho^3} - 2\frac{\widetilde{x}-\widetilde{X}}{\rho^3} + 3\frac{\left(\widetilde{x}-\widetilde{X}\right)^3}{\rho^5}\right)\right|_{\widetilde{X}=\widetilde{X}_0} = \\ = 3\left(\frac{\left(\widetilde{x}-\widetilde{X}_0\right)^3}{\rho_0^5} - \frac{\widetilde{x}-\widetilde{X}_0}{\rho_0^3}\right),$$

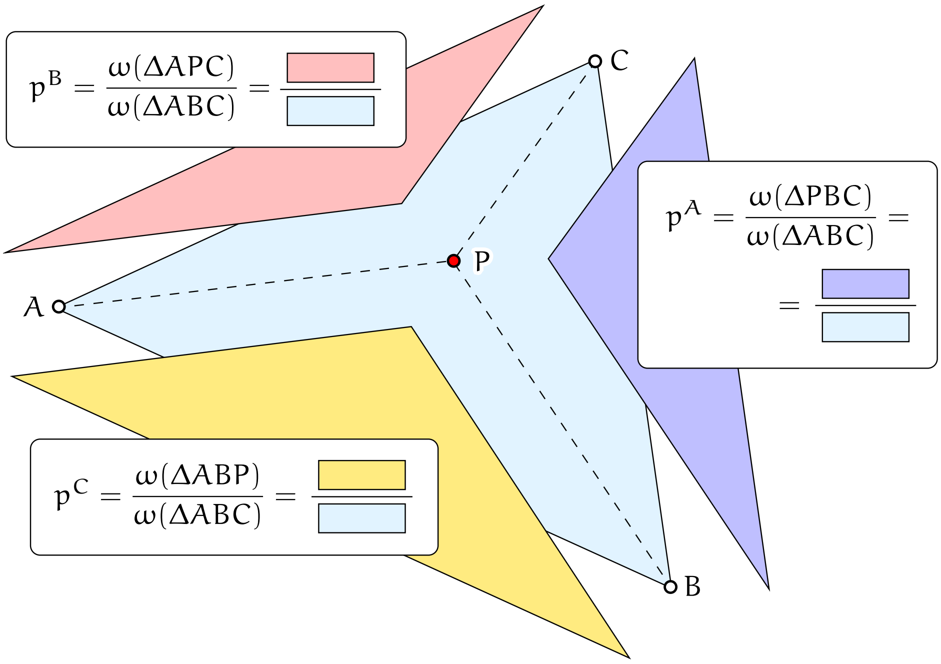


FIGURE 9.4. Barycentric co-ordinates. The ω quantities are the surface areas of triangles, computable as determinants, equation 9.12.

and similarly for $f(\widetilde{Y})$ and $f(\widetilde{Z})$. Again

$$\widetilde{x} - \widetilde{X}_0 = -\rho_0, \quad \widetilde{y} - \widetilde{Y}_0 = 0, \quad \widetilde{z} - \widetilde{Z}_0 = 0, \tag{9.9}$$

and

$$\left.\frac{\partial^2 f}{\partial \widetilde{X}^2}\right|_{\widetilde{X}=\widetilde{X}_0} = 0, \quad \left.\frac{\partial^2 f}{\partial \widetilde{Y}^2}\right|_{\widetilde{Y}=\widetilde{Y}_0} = 0, \quad \left.\frac{\partial^2 f}{\partial \widetilde{Z}^2}\right|_{\widetilde{Z}=\widetilde{Z}_0} = 0.$$

This is a somewhat surprising but pleasant result. Behind this is the linearisation of observation equation 9.7, the assumption that the satellite orbit error $\Delta\mathbf{x}$ is small. It actually is nowadays only of the order of metres, whereas the distance s of the rover from the base station may be many kilometres.

9.3.2 The case of three base stations

Because generally the case of a network of base stations can be reduced to the case of three base stations, we shall only study the latter.

Let us start from the equations derived above. Equation 9.10 is *linear* in the parameters η and χ, which we may interpret as the plane co-ordinates in two mutually orthogonal directions. $\Delta\widetilde{y}$ is the location error of the satellite in the "left-right" direction, $\Delta\widetilde{z}$ "upwards" on the celestial dome, towards the zenith, also perpendicular to the line of sight to the satellite.

Because equation 9.10 is bilinear in the co-ordinates η and χ, we may interpolate the correction linearly, when it has been determined at the three base stations. If the base stations A, B, and C and the measured corrections $\Delta\overline{P}_A$, $\Delta\overline{P}_B$, and $\Delta\overline{P}_C$ are given, we can compute the correction for an arbitrary point as follows:

$$\Delta\overline{P} = p^A\Delta\overline{P}_A + p^B\Delta\overline{P}_B + p^C\Delta\overline{P}_C. \tag{9.11}$$

In this equation, p^A, p^B, and p^C are the computation point's *barycentric co-ordinates* within the triangle, see figure 9.4:

$$p^A = \frac{\begin{vmatrix} 1 & 1 & 1 \\ \eta_B & \eta_C & \eta \\ \chi_B & \chi_C & \chi \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ \eta_A & \eta_B & \eta_C \\ \chi_A & \chi_B & \chi_C \end{vmatrix}},\ p^B = \frac{\begin{vmatrix} 1 & 1 & 1 \\ \eta_C & \eta_A & \eta \\ \chi_C & \chi_A & \chi \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ \eta_A & \eta_B & \eta_C \\ \chi_A & \chi_B & \chi_C \end{vmatrix}},\ p^C = \frac{\begin{vmatrix} 1 & 1 & 1 \\ \eta_A & \eta_B & \eta \\ \chi_A & \chi_B & \chi \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ \eta_A & \eta_B & \eta_C \\ \chi_A & \chi_B & \chi_C \end{vmatrix}}. \tag{9.12}$$

Here, the co-ordinates η and χ were used. For barycentric co-ordinates it holds that[2] $p^A + p^B + p^C = 1$, and they are all three *linear* in both the η and the χ co-ordinate. By simple substitution, one may ascertain that in the vertices of the triangle, for example in point A, $p^A = 1$ and $p^B = p^C = 0$ — the *reproducing* property.

Equation 9.10 contains only η and χ, and only linearly. The equation applies in three-dimensional space, where the location difference vector between the base station and rover is given by $\mathbf{s} = \xi\widetilde{\mathbf{i}} + \eta\widetilde{\mathbf{j}} + \chi\widetilde{\mathbf{k}}$.

The interpolation between the three base stations should take into account the curvature of the Earth: all four points must be in the same plane. Plane equations, like 9.11, are wrong in principle when used in the map plane. The error made is, however, negligible, even for triangles tens of kilometres in size. The theoretically correct way is to project the location of the rover onto the plane through the three base stations as shown in figure 9.5.

[2] Barycentric co-ordinates are *weights*, and equation 9.11 gives a weighted average of the vertex values.

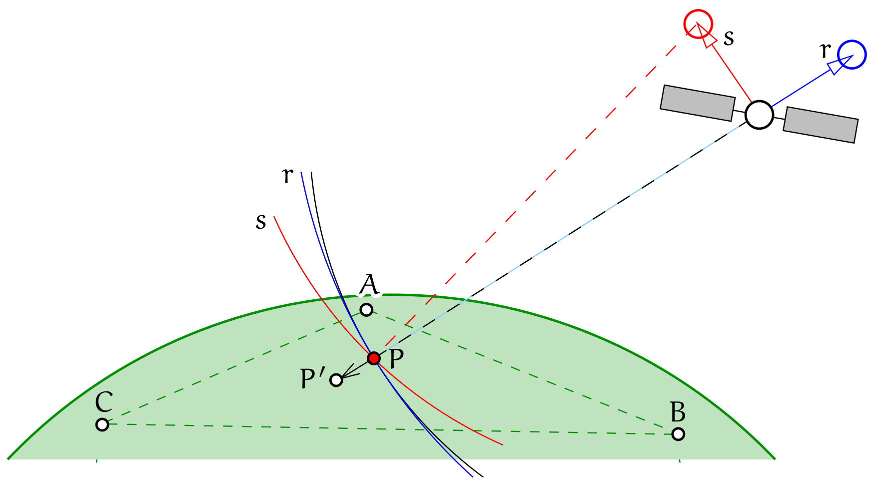


FIGURE 9.5. Geometry of differential GNSS. The radial satellite orbit error r does not strongly affect the difference measurements between different ground stations; the sideways orbit error s on the other hand affects in proportion to the distance between stations.
Also depicted is the way the differential correction is computed: the location P of the rover must be projected along the line of sight of the satellite onto the plane passing through the base stations ABC, yielding projection point P′. After that, the correction must be bilinearly interpolated.

9.4 Base-station networks

To implement a network RTK solution, several base stations are used and in some way the corrections given by these are *interpolated* between them to the location of the user.

Two different methods of data communication may be used:

1. The broadcast method: corrections are sent to many users at the same time. This may use, for example, the FM sideband of a radio broadcast (RDS, Radio Data System). The method scales well for a large number of users. The challenge with broadcasting is that the message cannot be tailored to the location of the user. Perhaps for this reason it has been popular for use with the less precise code-based differential GNSS. ULA
2. The singlecast method: the corrections are computed for one user and sent to them, for example by mobile Internet. The content of the correction message can be different for each user.

   One of the variants of the singlecast method is the *virtual reference station* (VRS) method, where base-station corrections are interpo- virtuaalitukiasema

> lated to the location of a "virtual base station" in the immediate vicinity of the observer. An important advantage of this approach is that the receiver does not know that the corrections are not coming from a real, physical RTCM base station, which all popular receiver brands know how to handle.

The most obvious interpolation technique is by brute force. It is assumed that the correction is continuous and slowly changing as a function of location on Earth. For a bilinear function, three base stations around the measurement area suffice.

One technique based on this approach is called FKP, *Flächenkorrekturparameter* or Areal Correction Parameters (Janssen, 2009). It is based on the estimation of separate bilinear models around base stations for the tropospheric and ionospheric propagation delays. The correction parameters are transmitted to all receivers in an area: this is a *broadcast* method. The rover performs the interpolation calculation for its location. RTCM 2.x message type 59, "proprietary message", has been used to transmit the corrections, tableau 8.1.

Other, similar interpolation methods are MAC (Master-Auxiliary Concept) and MAX (Master-Auxiliary Corrections), Leica Geosystems (2005). In these techniques, "as-raw-as-possible" carrier-phase range corrections from multiple base stations are broadcast to users in the field (Janssen, 2009). One station in the network is made the "master", and only its differential corrections are broadcast as such, the ionospheric ("dispersive") part and the remainder (tropospheric, orbits) separately. Also the raw carrier-phase observations of the master are broadcast. Only the *differences* of the differential corrections with those of the master are transmitted from the other stations in the network. These differences are numerically smaller and change more slowly with time.

In this approach, the rover does the complete interpolation job between all base stations. In order to limit the rover's calculation load, the network server has to pre-resolve the ambiguities of the base-station network, so that the corrections received by the rover for different base stations are at least "ambiguity compatible" with each other. This ambiguity resolution is made easier by the circumstance that the stations are in known, fixed locations.

Then, the rover need only resolve its own ambiguities with the master — and they will be the same with any other station in the network. Moreover, when only using the three nearest base stations, a rover moving into a neighbouring triangle can seamlessly switch base stations.

The RTCM standard 3.x allows for the compact format of the MAX multiple base-station messages to be transmitted.

Interpolation is a working solution at least for correcting the impact of satellite orbit errors and clock offsets, as their effect is by nature *deterministic*: it may be calculated accurately for the roving receiver's location, if the observations, or pseudorange corrections, at three base stations are given, and the base stations form a nice triangle around the measurement location.

For the atmospheric propagation delays however, unless inter-station distances are shorter than some 80 km, brute-force interpolation can be improved upon by atmospheric modelling. More elaborate atmospheric modelling is best done on the server side.

9.5 Modelling the atmosphere

The effect of the atmosphere is not deterministic but *stochastic*. This is why precise interpolation to the rover location will not work: the uncertainty grows with growing distance from the base station. How it grows depends on the statistical properties of the atmosphere. This brings us to the subject of atmospheric modelling.

The GNSS signal propagation delay caused by the atmosphere — both the ionosphere and the troposphere — reduced to a standard total delay in the direction of the zenith, the zenith total delay ZTD, forms a *stochastic process* $\underline{d}^{\perp}(\varphi, \lambda)$. The process is defined on the domain of geographic locations (φ, λ).

The process $\underline{d}^{\perp}(\varphi, \lambda)$ has a *signal covariance function* between two locations, which may be described by, for example, a Gauss-Markov type covariance formula:

$$\mathrm{Cov}\left\{\underline{d}^{\perp}(\varphi_1, \lambda_1), \underline{d}^{\perp}(\varphi_2, \lambda_2)\right\} = C_0 \exp\left(-\psi/\psi_0\right), \tag{9.13}$$

in which ψ is the angular distance between the points (φ_1, λ_1) and (φ_2, λ_2) as seen from the centre of the Earth. The constant C_0 is called the *signal variance*, the constant ψ_0 the *correlation length*. These constants may be chosen suitably, meaning realistically, for the troposphere and ionosphere of a certain time and place. psi $\psi\Psi$

Another approach is to model the atmosphere deterministically with a set of unknown parameters.[3] For example the global ionosphere may

3

be well-described by a spherical-harmonic expansion. A polynomial or Fourier expansion may also be considered. The unknown coefficients are estimated from the observations at the base stations.

Defining the *location* (φ, λ) of a propagation path is an interesting issue: it could be taken as the average location of the points on the path weighted by the density of refracting matter. For the ionosphere, this is often expressed as the location of a representative "pierce point" at average ionospheric height.

The quantity to be estimated may thus be represented as a general function of geographic location $d^{\perp}(\varphi, \lambda)$, the resolution of which requires measurements by a network of receivers. From here, it is a small step to introduce the third dimension and estimate $d(\varphi, \lambda, h)$ — or equivalently, N_e, the electron density, equation 8.13, or N, the index of refraction, equation 8.14 — throughout the atmospheric space. This technique is referred to as *atmospheric GNSS tomography*.

If a good numerical weather model is available, an operator of a GNSS base-station network may use it to improve, for example, the interpolation of tropospheric propagation delays.

Conversely, tropospheric wet propagation delay values obtained from observations of a base-station network may be *assimilated* into a numerical weather model to improve its real-time performance. Numerical weather models describe the behaviour of the atmosphere using physical model equations on a three-dimensional grid as a function of time. This is the subject of active research.

The current state of the art is that RTK positioning using a network of base stations is as precise as RTK positioning using one nearby base station, on condition that the distances between the base stations are at most about 80 km. The positioning quality is also preserved for some distance outside the coverage area of the network. The atmosphere is the limiting factor.

Self-test questions

1. What is the ambiguity problem? Describe approaches to resolving it.

---

[3]The GPS navigation message contains such a model, the Klobuchar ionosphere model, with eight parameters.

2. What is the advantage of using double differences for ambiguity resolution?
3. What is the advantage of using networks of RTK base stations, so-called network RTK?
4. What methods exist for interpolating corrections from a network of base stations to the location of the user?
5. Why does linear interpolation of corrections between base stations work well for satellite orbit errors, but not so well for atmospheric propagation delays?
6. What is a virtual reference station?
7. What is the advantage *for the user* of using the VRS method?
8. How, in the MAX interpolation method, is the size of the data to be transferred to the rover minimised?
9. Describe the use of the LAMBDA method for ambiguity resolution.
10. Why would the use of numerical weather prediction (NWP) models be advantageous for monitoring atmospheric water vapour content using GNSS?

    Hint The models describe atmospheric quantities as functions of three-dimensional place *and time*.

Exercise 9–1: Variance function of a difference of zenith propagation delays

If the covariance function of the zenith propagation delay $\underline{d}^{\perp}$ between two locations 1 and 2 is given by equation 9.13:

$$\operatorname{Cov}\left\{\underline{d}^{\perp}(\varphi_1,\lambda_1),\underline{d}^{\perp}(\varphi_2,\lambda_2)\right\} = C_0 \exp\left(-\psi/\psi_0\right),$$

derive the *variance function* of the zenith propagation delay inter-location difference quantity

$$\Delta_{12}\underline{d}^{\perp} \stackrel{\text{def}}{=} \underline{d}_2^{\perp} - \underline{d}_1^{\perp} = \underline{d}^{\perp}(\varphi_2,\lambda_2) - \underline{d}^{\perp}(\varphi_1,\lambda_1)$$

for two locations 1 and 2.

The function will be of the form

$$f(\psi) = \operatorname{Var}\left\{\Delta_{12}\underline{d}^{\perp}(\psi)\right\}.$$

Draw its graph assuming $C_0 = \psi_0 = 1$.

Satellite-based augmentation systems

10

Especially in aviation, it is not acceptable to use the GPS as the only means of navigation, because it does not give any availability or correctness guarantees. This circumstance is known as the *integrity problem*. eheysongelma

Satellite-based augmentation systems (SBAS) are a widely deployed way to address this problem. There are working SBAS systems in North America (WAAS), Europe (EGNOS), Japan (MSAS and QZSS, the latter being a hybrid augmentation/positioning system) and India (GAGAN). Russia (SDCM), China (BDSBAS) and the Republic of Korea (KASS) are also considering or already deploying systems compatible with the international aviation standard.

SBAS does not just offer integrity safeguards to those using the GPS, it also saves money by replacing traditional radionavigation and instrument-landing support equipment in airport areas, which is massive in size.[1] Thanks to SBAS, smaller airports also become reachable [1] by regular air traffic. Furthermore, fuel is saved as GPS and SBAS allow straight flights along the geodesic from airport to airport instead of along a polygon of radio beacons.

10.1 Receiver autonomous integrity monitoring

One more modest approach to safeguarding integrity in aviation is *receiver autonomous integrity monitoring* (RAIM), in which the redundancy eheyden valvonta of the GPS is exploited to detect a possible untrustworthiness of one or more GPS satellite signals (Wikipedia, Receiver autonomous integrity monitoring). The technique is similar to evaluating the robustness of geodetic networks against gross errors in terms of the ease with which

[1] For example a VOR beacon (VHF Omnidirectional Range) has a diameter of 44 ft or 13 m!

such gross errors can be detected, a concept known as *reliability* (Baarda, 1968).

There are two questions that need to be answered at any given moment:

1. Is any of the satellites that I am using transmitting an untrustworthy signal?
2. If this were the case, would I even be able to detect it?

The answer to the first question is obtained by statistical testing, the answer to the second is found by reliability assessment.

The observation equation to use is that for station-location determination from pseudorange observations, equation 8.10:

$$\Delta\underline{p} = H\underline{\mathbf{x}} + \underline{m},$$

but written as an estimation equation (subsection 2.4.1):

$$\Delta\underline{p} + \underline{v} = H\widehat{\mathbf{x}},$$

$\underline{v}$ being the residual.

For all visible satellites together, this becomes

$$\underline{\boldsymbol{\ell}} + \underline{\mathbf{v}} = A\widehat{\mathbf{x}},$$

with

$$\underline{\boldsymbol{\ell}} \stackrel{\text{def}}{=} \begin{bmatrix} \Delta\underline{p}^1 \\ \Delta\underline{p}^2 \\ \vdots \\ \Delta\underline{p}^n \end{bmatrix}, \quad \underline{\mathbf{v}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{v}^1 \\ \underline{v}^2 \\ \vdots \\ \underline{v}^n \end{bmatrix}, \quad A \stackrel{\text{def}}{=} \begin{bmatrix} H^1 \\ H^2 \\ \vdots \\ H^n \end{bmatrix}, \quad \widehat{\mathbf{x}} = \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \\ \Delta T \end{bmatrix}.$$

The ordinary least-squares solution to this is

$$\widehat{\mathbf{x}} = \left(A^T A\right)^{-1} A^T \underline{\boldsymbol{\ell}},$$

and the residuals are

$$\underline{\mathbf{v}} = A\widehat{\mathbf{x}} - \underline{\boldsymbol{\ell}}.$$

For RAIM in aviation, the test statistic to use is the sum of squares of the residuals of the least-squares solution:

$$\underline{\mathcal{E}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left(\underline{v}^i\right)^2 = \underline{\mathbf{v}}^T \underline{\mathbf{v}}, \tag{10.1}$$

in which $\underline{v}^i$ are the residuals and n is the number of satellites.

In ordinary least-squares, it is assumed that the observed pseudoranges are normally distributed and uncorrelated and all have the same uncertainty or mean error σ. This is called the *i.i.d.* property. Then, the quantity $\underline{\widetilde{\mathcal{E}}} \stackrel{\text{def}}{=} \underline{\mathcal{E}}/\sigma^2$ is distributed according to the chi-squared distribution χ^2_{n-4}. Here, $n - 4$ is the number of degrees of freedom — as there are four unknowns: the three receiver co-ordinates X, Y, and Z and the receiver clock offset ΔT. sigma $\sigma\Sigma$ chi χX

The first rule is that the probability of a "false alarm" (PFA) when there is no satellite malfunction, that is, an error of the first kind, should be no more than $1/15\,000$,[2] meaning a test significance level of $\alpha = 1 - 1/15\,000 = 99.9933\ldots$ %. The χ^2 distribution under the null hypothesis gives a threshold value for this, which will depend on the number of degrees of freedom. See table 10.1. 2 alpha αA

A second parameter for testing that needs to be set is the probability that a gross error in one of the pseudoranges will be detected. This value is called the *power* of the test, often called β. It is set to $\beta = 99.9$ %. This means that the probability of an error of the second kind, that is, missing a really occurring malfunction or "probability of missed detection" (PMD), is $1 - \beta = 0.1$ %. beta βB

If the satellite geometry is such that for the set value of α, a gross error detectable with probability β in one of the pseudoranges would produce an error in, for example, the horizontal location exceeding a specified "alarm limit", then RAIM must not be used. This requirement is known as *exterior reliability*, in this case of horizontal location.

Note that reliability is a property of the *satellite geometry* unrelated to any actual measurements! Reliability can be calculated pre-flight. It shares this feature with DOP, dilution of precision, a well-known quantity characterising the quality of GNSS positioning. DOP, however, is related to the *precision* of a GNSS positioning solution. It, too, is a planning tool.

There are different alarm limits for different phases of the flight. So, in a specific situation, RAIM may be useable for cruise flight but not for approach.

It is clear that the level of redundancy has to be sufficient for this

[2] The reason why the number is this low is psychological: too many false alarms may lead to appropriate alarms not being taken seriously.

TABLE 10.1. χ^2 test limits for $\alpha = 1 - {}^1\!/_{15\,000}$ and various numbers of degrees of freedom. These are the limits that the quantity $\underline{\widetilde{\mathcal{E}}} \stackrel{\text{def}}{=} \underline{\mathcal{E}}/\sigma^2$, computed by equation 10.1 from the residuals of the least-squares solution *during flight*, is tested against.

Number of satellites	Degrees of freedom	χ^2 limit for $\underline{\widetilde{\mathcal{E}}}$
5	1	15.90
6	2	19.23
7	3	21.95
8	4	24.39
9	5	26.65
10	6	28.79
11	7	30.83
12	8	32.81

to work. Four satellites is the minimum to allow positioning using code-based pseudoranges, but offers no redundancy. A rule of thumb in geodesy says that redundancy should be 50 % at least, which would mean eight satellites in the sky of the aircraft. This is achieved most of the time already with a minimum 24-satellite GPS constellation.

But it is not just about the number of satellites: they have to be in well-distributed positions in the sky that actually produce the level of reliability aimed for. Interactive planning tools exist on the Internet that help predict where and when RAIM is actually useable.

The current larger GPS constellation and the multitude of other working GNSS are helping to make this an attractive option. It is the subject of active research, to mention ARAIM, advanced RAIM using for example Galileo and GPS together in a dual-frequency (L1 and L5) receiver (GLAD).

10.2 Description of SBAS technology

Each SBAS is built for a target area — for example WAAS for North America — where a network of GPS monitoring stations is operating, keeping an eye on the correct functioning of each GPS satellite visible from the area in real time. Users receive information from the service through a transponder on a geostationary satellite.

tosiaikaisesti

eheys

In addition to integrity information, the service also provides differential corrections to pseudorange measurements to GPS satellites. These

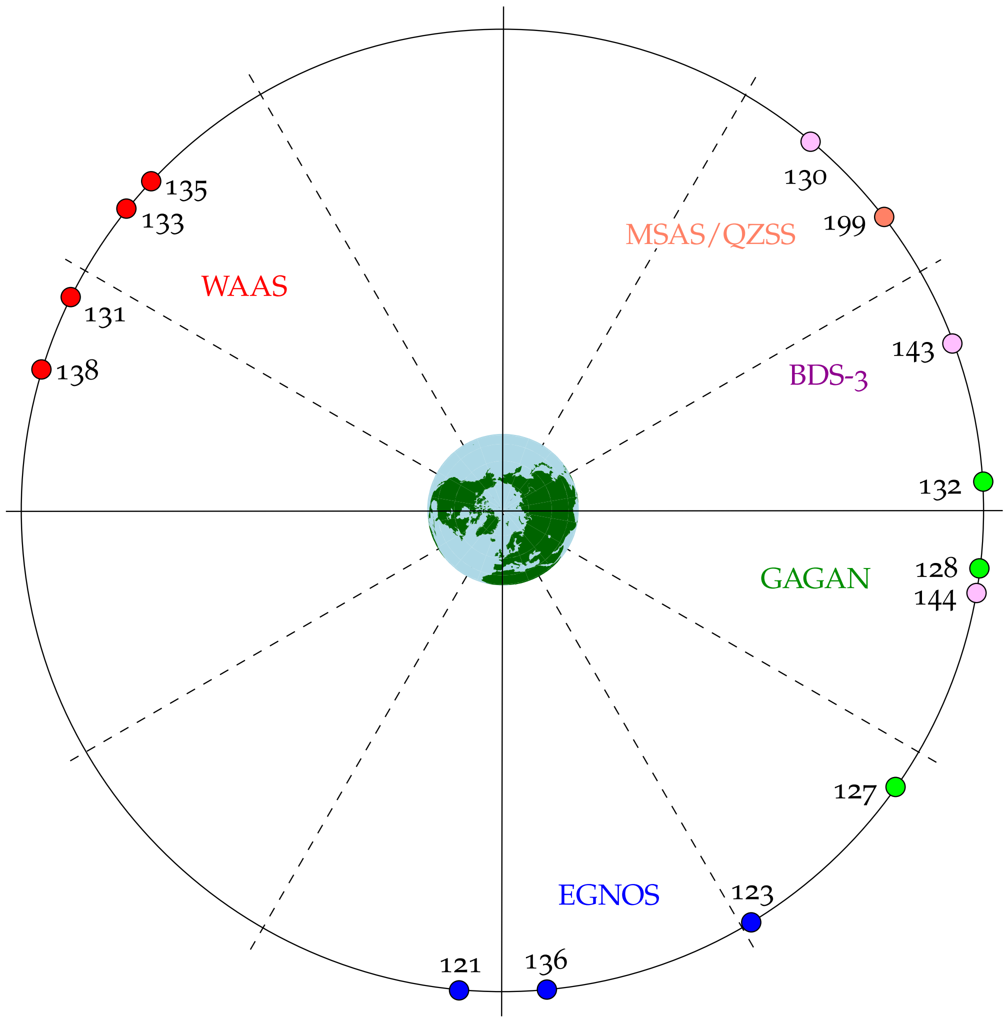


FIGURE 10.1. SBAS transponder geometry. Status 2020. The numbers refer to the pseudo-random noise (PRN) codes of the L1 frequency. On the L5 frequency, corresponding codes are used.

corrections are valid within the target area.

In principle, in order to provide differential corrections in an area, three stations in the corners of the area would suffice if the *only* causes of imprecise pseudoranges were the satellite clocks and the orbits given in the broadcast ephemeris. In practice, the need to include the effect of the atmosphere into the corrections necessitates a denser coverage of the area. And of course redundancy is always welcome.

An SBAS satellite, or rather an SBAS transponder on a general-purpose geostationary satellite, transmits a signal similar to the L1 signal of the GPS. The carrier frequency is also 1575.42 MHz and it is binary phase-shift modulated by a pseudo-random code, a 1023-bit Gold code at a bit frequency of 1.023 MHz, just like the GPS C/A code. This means that a minimally modified GPS receiver can make use of the signal.

However, instead of the navigation message of the GPS, which is broadcast at a bit rate of 50 Hz, the SBAS message is transmitted at a rate of 250 Hz. The length of a single message is also 250 bits, which takes one second to transmit. Transmission of all the necessary information is spread out over many messages. The information is classified as "fast" or "slow", depending on the allowed latency. For example part of the clock correction is "fast", while the ionosphere correction again is voluminous but only slowly changing.

The SBAS signal provides the following information:

- The pseudo-random code identifying the transponder using a
3 Gold-code number earmarked for SBAS use.[3] These could be used for pseudorange measurement, but offer little added value compared to the GPS.
- Differential corrections. An SBAS-capable GNSS receiver receives corrections for the orbit errors and the clock offsets of the satellites as well as for the effect of the ionosphere. As many SBAS receivers are single-frequency (L1 only), the ionospheric propagation delay is very significant — from metres to tens of metres — and needs to be corrected precisely and at sufficient spatial resolution.
- Integrity information.

The orbit corrections are long-term and slowly changing and are broadcast as corrections to the rectangular geocentric satellite locations and velocities calculated from the broadcast ephemeris. The correction formula is

$$\mathbf{x}_{\mathrm{corr}}(t) = \mathbf{x}_{\mathrm{be}}(t) + \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix} + \begin{bmatrix} \delta \dot{x} \\ \delta \dot{y} \\ \delta \dot{z} \end{bmatrix} (t - t_0),$$

with t_0 being the issue of data (IOD) time of both the broadcast ephemeris and the correction message.

The clock corrections for the GPS satellites consist of two parts:

kellon käynti
- a slow part broadcast as offset and drift corrections at a low rate
- a part that changes rapidly with time.

---

[3]The US Air Force is the global co-ordinator for assigning these code numbers. The code numbers for SBAS are for the frequencies L1 (C/A) and L5.

Table 10.2. The various approach categories according to ICAO. The category describes how well aircraft and airfields are equipped for instrument landings. An SBAS approach is only possible for some of the categories.

Category	Decision height (DH)	Visibility	Runway visual range (RVR)
Cat I	200 ft = 60 m	$\geqslant$ 2600 ft = 800 m, or	$\geqslant$ 1800 ft = 550 m
Cat II	100–200 ft = 30–60 m	-	$\geqslant$ 1200 ft = 350 m
Cat III A	< 100 ft = 30 m / none	-	$\geqslant$ 700 ft = 200 m
Cat III B	< 50 ft = 15 m / none	-	150–700 ft = 50–200 m
Cat III C	none	-	none

The ionospheric propagation delay, which changes slowly in time, is at the same ionosphere location the same for all GPS satellites. The delay values are broadcast in the form of a grid covering its area of validity, the area of the base stations used in its calculation. The broadcast values give the magnitude of the effect for a satellite in the zenith. The correction is dependent on user location: the user interpolates the effect from the grid to the ionospheric pierce point of each satellite signal, taking the angle of penetration into account. tukiasema

The SBAS message content is extensively presented in Sánchez and Berges (2006).

10.3 Integrity and safety of life

Integrity monitoring means that users are warned within a set time limit — for example six seconds — if the positioning signal of a GPS satellite exceeds set tolerance values. This is mandatory in safety-of-life applications. Safety of life (SoL) means that if the system does not function correctly, people may die. eheyden valvonta

A good example of this is approach and landing at an airport. If one descends in fog using GPS navigation and the height is many metres off without warning, an accident will happen. If the GPS height cannot be relied upon in fog or cloud and the pilot is warned of this, no landing is attempted. If, on the other hand, the precision guaranteed by the integrity system is within a few metres and there is no fog but low cloud, the pilot may go for a visual landing from, for example, 200 feet downwards.

Integrity makes operations possible in aviation that would be too

TABLE 10.3. Satellites carrying a WAAS transponder, status 2020. PRN is the pseudo-random noise code used, similar to those for GPS satellites.

Satellite	PRN	Longitude	Launch	Expiration
Inmarsat-3 F3	134	178° E	17 Dec 1996	Sept 2007
Inmarsat-3 F4	122	142° W	3 June 1997	Sept 2008
Telesat Anik F1R	138	107.°3 W	21 Nov 2000	-
Intelsat Galaxy 15	135	133° W	13 Oct 2005	-
Inmarsat-4 F3	133	117° W	18 Aug 2008	Nov 2017
Eutelsat 117 WB	131	117° W	14 June 2016[a]	-
SES-15	133	129° W	18 May 2017	-

[a]Operating since March 2018.

risky using only GPS. As such, GPS has already been used in RAIM mode, receiver autonomous integrity monitoring, but only on the straight parts of route flights and not during descent. It is required that when a GPS signal becomes unreliable, the pilot is alerted within the set time limit.

10.4 WAAS

WAAS (Wide Area Augmentation System) was declared operational in 2003. The system monitors the transmissions of the GPS and computes differential corrections for users while providing a certified *integrity level*. Precision after differential correction is of the order of ±2 m. Other sources give 1–2 m horizontally and 2–3 m vertically within the service area. Without WAAS, the precision of GPS positioning would only be ±15 m on average using a single-frequency GPS receiver, mostly due to the ionosphere.

eheystaso

From 2007 on, WAAS has been approved for guiding an aircraft safely to 200 feet height above the runway: ICAO Category I, table 10.2.

In 2014, 73 000 aircraft in North America were equipped with a WAAS receiver (NASA Spinoff).

WAAS is also being used outside aviation, for example in maritime navigation.

Table 10.4. Satellites carrying an EGNOS transponder, status 2020. PRN is the pseudo-random code number of the L1 transmission. The currently active EGNOS satellites also transmit on L5.

Satellite	PRN	Longitude	Launch	Expiration
Inmarsat-3 F1	131	64°.5 E	3 Apr 1996	Retired
Inmarsat-3 F2	120	15°.5 W	6 Sept 1996	1 Jan 2019
Artemis	124	21°.5 E	12 July 2001	2015
Inmarsat 4-F2	126	64°.0 E	8 Nov 2005	Retired
Astra SES-5	136	5°.0 E	9 July 2012	-
Astra 5B	123	31°.5 E	23 Mar 2014	-
Eutelsat 5WB	121	5°.0 W	9 Oct 2019	-

10.4.1 *The WAAS space segment*

The satellites that have carried WAAS transponders and their history are described in table 10.3, based on public sources.

The signal from a geostationary satellite is coded in the same way using pseudo-random noise (PRN) codes as the signals from the GPS satellites. The WAAS satellites have their own codes, which differ from each other and from the codes of the GPS satellites, and the correlator in the receiver distinguishes them by these codes.

10.4.2 *The WAAS ground segment*

The WAAS uses the following base stations in the United States and neighbouring countries:

1. WRS (Wide Area Reference Stations): 38 in North America, including seven in Alaska, one in Hawaii and one in Puerto Rico. This is also the area within which the disseminated corrections are precise.
2. Three WMS (Wide Area Master Stations). The corrections and integrity information are computed at these. The differential corrections are computed for the nodes of a grid covering the area.
3. Four GUS (Ground Uplink Stations). At these, the correction message is assembled and transmitted to a transponder on a geostationary satellite, which sends the message on to the users on the same frequencies used also by the GPS satellites: L1 and on the newer satellites L5, using similar identifying pseudo-random codes as the GPS satellites.

The GUS stations have large parabolic antennas for the uplink. Two different stations are used for each transponder available over the target area.

The data communication links between stations have been custom built. tosiaikainen The open Internet is not real time with acceptable reliability.

10.5 EGNOS

EGNOS, the European Geostationary Navigation Overlay Service, is a joint project by the European Commission, ESA and Eurocontrol, the joint European aviation safety organisation.

EGNOS is compatible and interoperable with WAAS. EGNOS started operations in July 2005.

lohko EGNOS consists of four functional parts or segments: ground segment, support segment, space segment and user segment.

The ground segment of EGNOS consists of the following components:

1. RIMS (Ranging and Integrity Monitoring Stations): 40. These receive the signal and forward the data to the Master Control Centres.
2. Two MCC (Master Control Centres), Ciampino (Italy) and Torrejon (Spain). These receive data from the RIMS and compute from it corrections and integrity data, which are forwarded.
3. Six NLES (Navigation Land Earth Stations). These forward the data to the different geostationary transponders.

As with WAAS, dedicated data connections have also been built between the EGNOS ground stations.

10.6 Japanese SBAS systems

10.6.1 MSAS

MSAS (MTSAT Satellite-based Augmentation System) was the first Japanese SBAS. It was designed to be compatible with WAAS and EGNOS. The system became operational in September 2007. Table 10.5 describes the space segment.

After MTSAT-1R was decommissioned in 2015, MTSAT-2, which had transmitted only signal for PRN 137, proceeded to transmit signal for

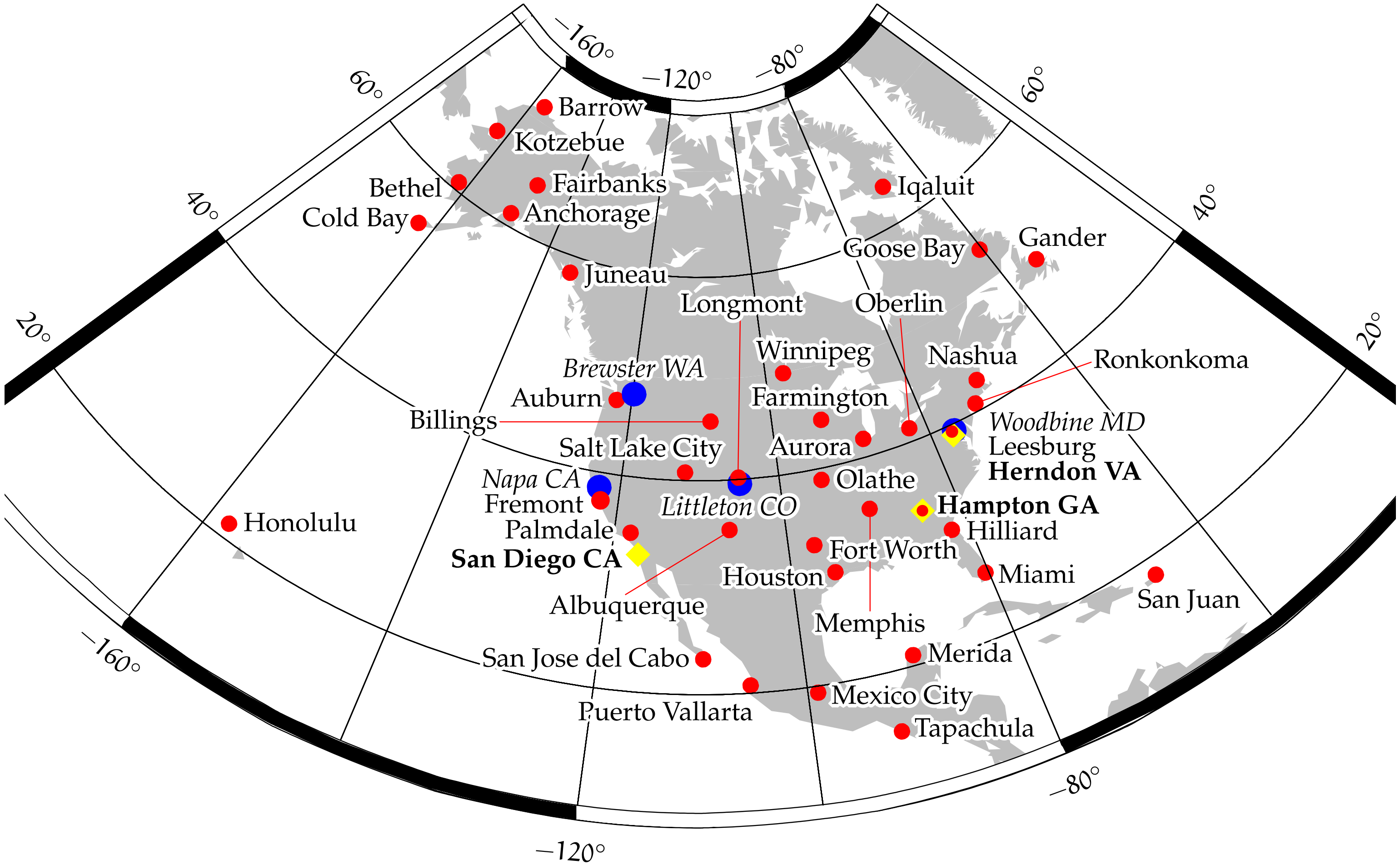


FIGURE 10.2. WAAS ground segment, status 2017. Red: WRS, **yellow: WMS**, *blue*: *GUS*. Note the stations in Canada and Mexico.

Red: RIMS, **yellow: MCC**, *blue*: *NLES*.

Not shown here are the far-field ground stations in South Africa (Hartebeesthoek), Canada (Montreal), Singapore, and French Guyana (Kourou).

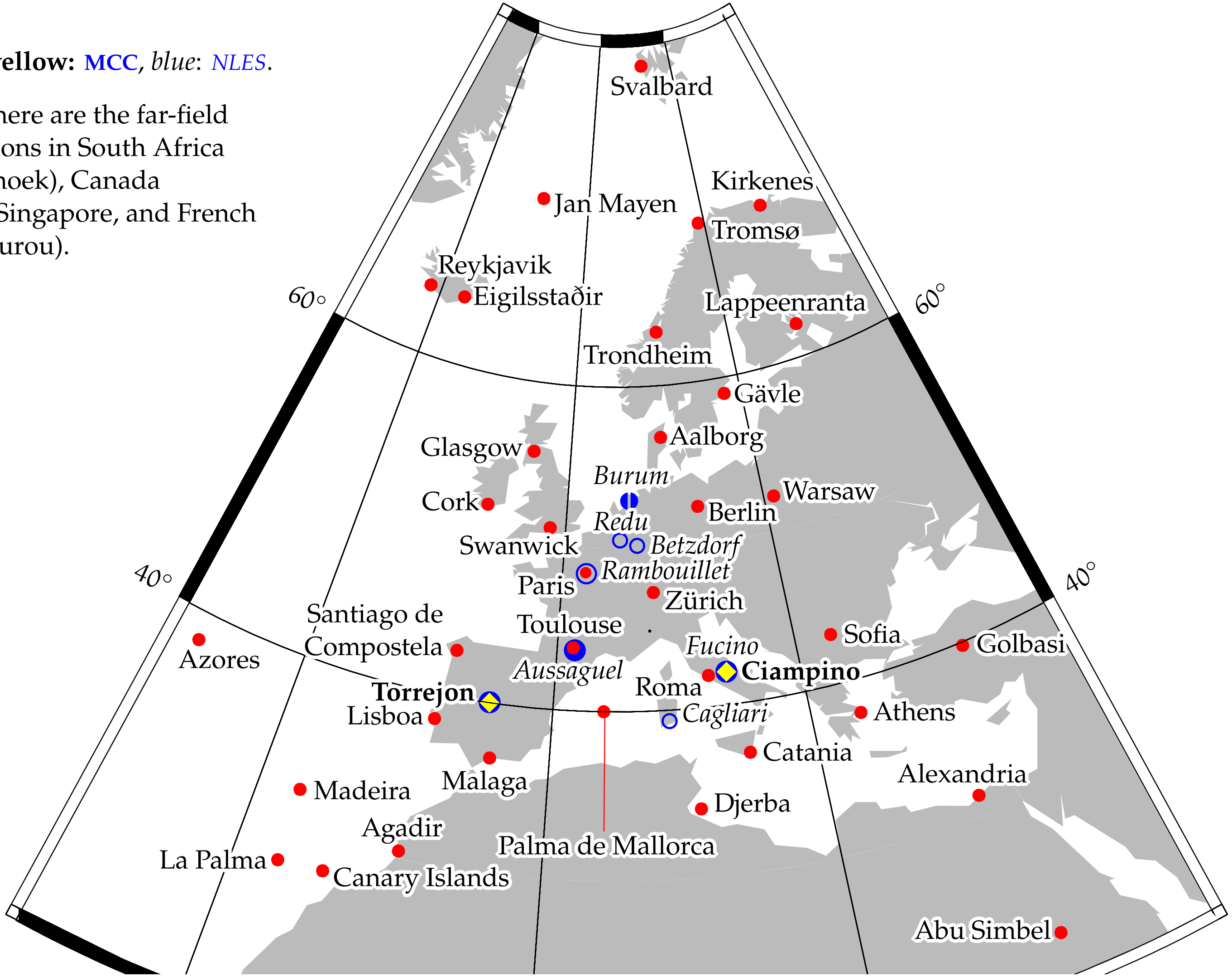


FIGURE 10.3. EGNOS ground segment, status 2019.

Table 10.5. Japanese satellites carrying an MSAS transponder. The MTSAT satellites were weather satellites and the SBAS transponder was only a small part of the payload. PRN is the pseudo-random code number of the transmission.

Satellite	PRN	Longitude	Launch	Expiration
MTSAT-1R (Himawari 6)	129	140°E	26 Feb 2005	Dec 2015
MTSAT-2 (Himawari 7)	137, 129	145°E	18 Feb 2006	May 2020
QSZZ-3	199	127°E	19 Aug 2017[a]	-

[a] Operating since April 2020.

both PRN 129 and PRN 137. In April 2020, the geostationary satellite QZSS-3 started transmitting signal for PRN 199.

MSAS has six monitoring stations on the ground. See Office of Aeronautical Satellite Systems, ATS Engineering Division, Japan Civil Aviation Bureau (2008) and figure 10.4.

The MSAS system, also called MSAS V1, is being replaced by QZSS, also called MSAS V2, in 2020.

10.6.2 QZSS

The Japanese Aerospace Exploration Agency JAXA has built a system named QZSS, Quasi-Zenith Satellite System, nickname *Michibiki* — "guidance". It works as an augmentation system for GPS using a signal structure that is compatible and interoperable with GPS. Unlike the other SBAS systems, it not only provides augmentation of GPS positioning, but also its own positioning signals meant to be used together with those of the GPS.

The satellite orbits have a period of one sidereal day and are called geosynchronous orbits. The inclination of the orbital planes is high, some 45°–53°, and the orbits are fairly round: the orbital eccentricity is around 0.07. The intention is that, at any time, there is one satellite hanging over Japan at a reasonably high elevation angle. The satellite is visible even from the "urban canyons" of Japan's big cities. The label "quasi-zenith" originates from this.

Always having one satellite close to the zenith in Japan requires three satellites in orbits with ascending nodes at the same geographic

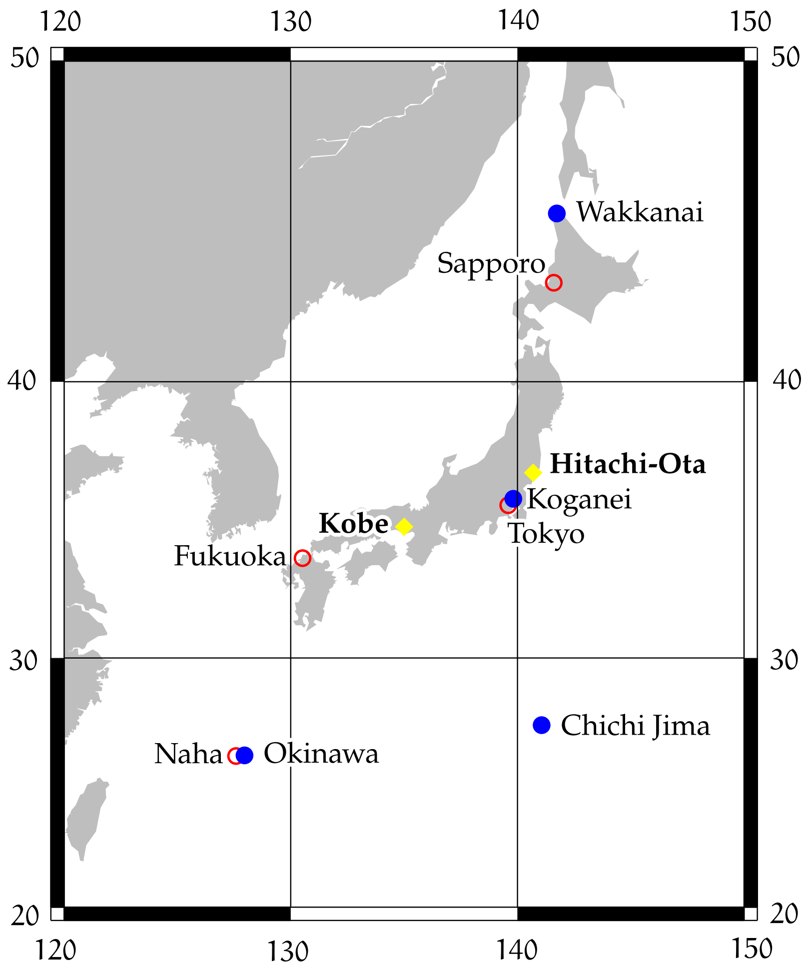


FIGURE 10.4. MSAS and QZSS ground segment, status around 2016. **Yellow**: Master Control Station (MCS). Ground Monitor Station (GMS), red: MSAS, blue: QZSS.
QZSS: Multiple tracking control stations and monitoring stations outside Japan not indicated.

longitude — and their right ascensions 120° apart. The satellites traverse the nodes at time intervals of one-third of their period, like the clubs of a juggler. The satellites are also suited, and especially so, for serving as communication satellites.

This type of orbit is often referred to as a "Tundra orbit". It is somewhat similar to a "Molnya orbit" as used by Soviet communication satellites, that had a similar operational concept of "apogee dwell" in order to serve high northern latitudes. However, Molnya orbits are highly eccentric and have an orbital period of only one-half of a sidereal day. Their inclination is $i = 63\overset{\circ}{.}4$, which satisfies the condition $5 \sin^2 i - 4 = 0$ for zero "apsidal precession" of the orbital ellipse within

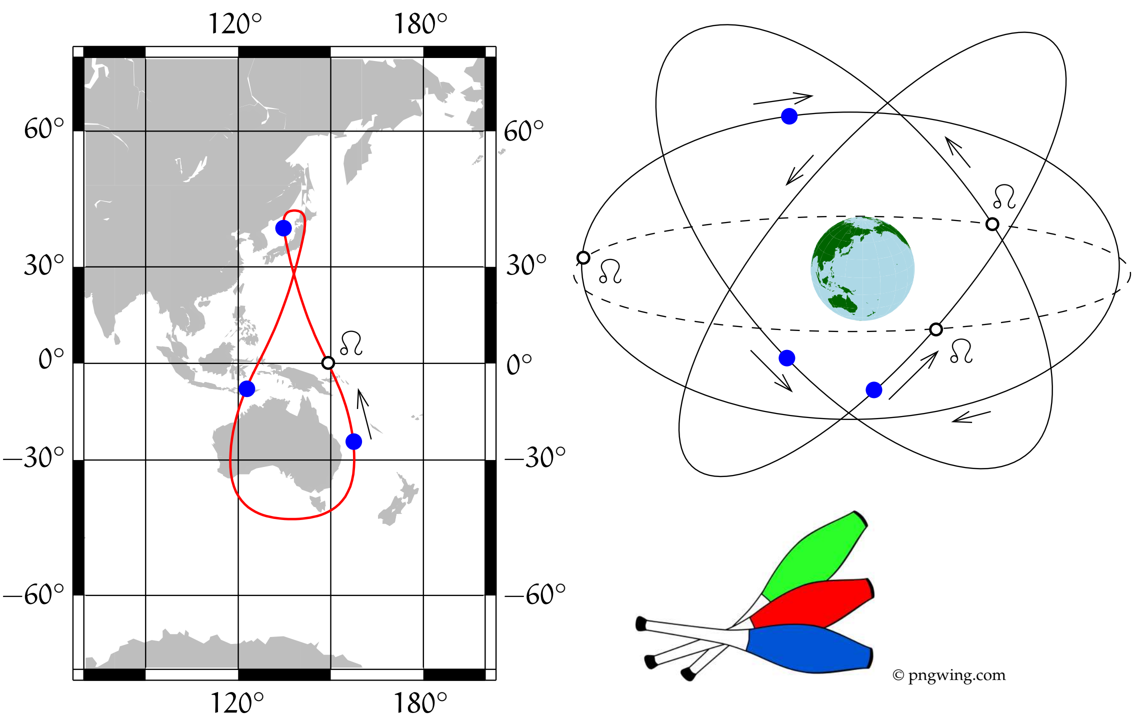


FIGURE 10.5. Concept of operation of the QZSS orbit. On the left, in a frame co-rotating with the Earth, on the right, in an inertial frame.

the orbital plane. This means that once the apogee has been placed due north — argument of perigee $\omega = 270°$ — it will stay there. omega $\omega\Omega$

Unlike other — geostationary — SBAS systems, this system offers *useable* observables of types pseudorange and carrier phase. It transmits in the same frequency bands as the GNSS systems using code-division multiple access (CDMA) with its own PRN codes, minimising interference. koodijako-kanavointi

The QZSS's first satellite was launched on 11 September 2010. Three more followed in 2017, and on 1 November 2018, the system was taken officially into service. An expansion to seven satellites is planned.

The ambitious QZSS is extensively described on the European Space Agency ESA's Navipedia web site (GMV, 2011). The system's performance standard and interface specification are found here: QZSS PS/IS.

10.7 The Indian GAGAN system

GAGAN, GPS-Aided Geo Augmented Navigation, is India's SBAS, primarily for aviation use in Indian airspace. It currently has three transponders in operation, see table 10.6.

TABLE 10.6. GAGAN satellites.

Satellite	PRN	Longitude	Launch
GSAT-8	127	55° E	21 May 2011
GSAT-10	128	83° E	29 Sept 2012
GSAT-15	132	93°.5 E	10 Nov 2015

There are 15 Indian Reference Stations (INRES), all within India, two Indian Mission Control Centres (INMCC) in Bangalore and Delhi, and three Indian Navigation Link Upload Stations (INLUS), two in Bangalore and one in New Delhi.

A particular problem in Indian airspace is the variable ionospheric delay, due to the geomagnetic equator passing through the country, and the equatorial electrojet within the ionosphere passing over it. The problem is addressed by an Indian research effort involving a network of 18 total electron content (TEC) monitoring stations all over India. It is also one reason why the GAGAN satellites transmit on both L1 and L5.

ekvatoriaalinen virtatihentymä

10.8 Ground-based augmentation systems

GBAS, ground-based augmentation systems, use a GPS base station within three miles of an airport to provide differential corrections and integrity data, just like an SBAS does. It supports navigation, approach, and landing in an air space of 20 miles surrounding the airport.

The original American implementation is called LAAS, local-area augmentation systems.

GBAS are intended to be used for instrument landings at busy airports. Its use is approved for ICAO Cat I approaches. Approval for Categories II and III has been under preparation for a long time.

The airport's GPS base station monitors the GPS satellites and computes differential corrections locally. GPS positioning using the corrections has an accuracy of ± 1 m. The corrections are transmitted to the aircraft by radio (VHF). Unlike SBAS, it does require special equipment at the airport. However, one such equipment set would replace many traditional installations, like VOR beacons, for every runway.

GBAS can also be used to guide ground vehicles within the airport area. The safety relevance of this is clear as it helps prevent aircraft and ground vehicles from being in the same place at the same time.

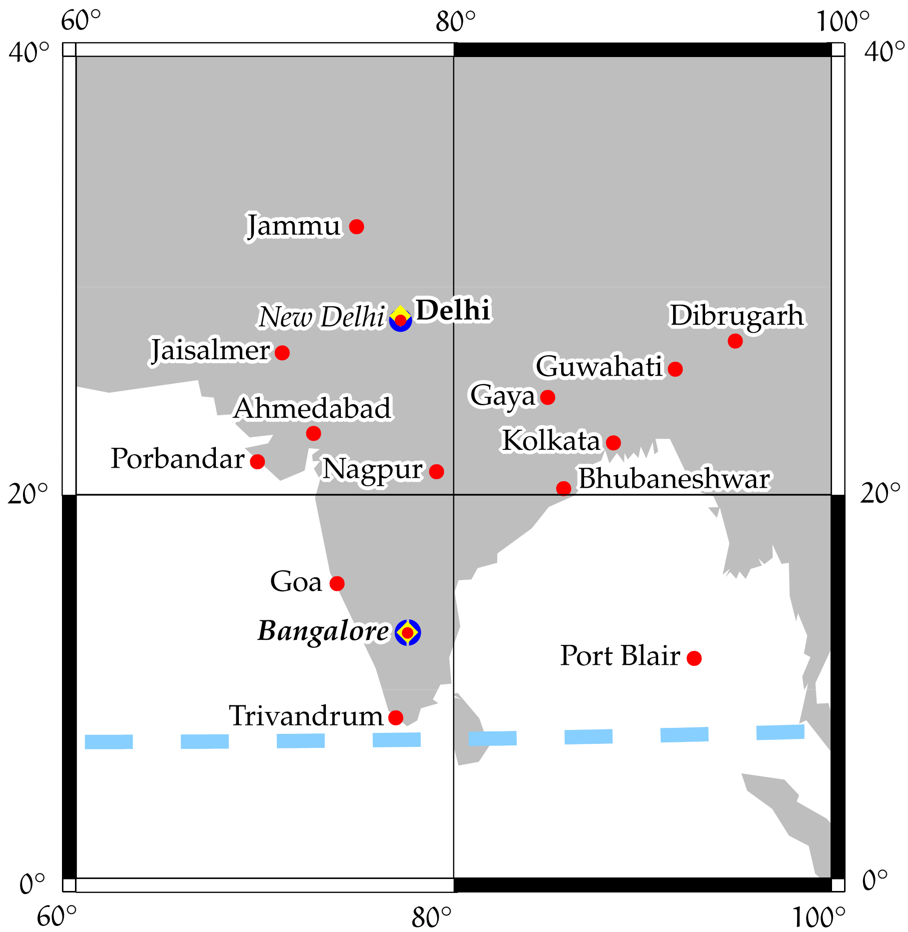


FIGURE 10.6. GAGAN ground segment, status 2017. Red: INRES, **yellow: INMCC**, *blue: INLUS*. Equatorial electrojet in light blue dashed.

GBAS is in operational use at many international airports. Like SBAS, it can replace traditional ground installations for instrument landings — two for each runway — and allows for flexible approach geometries, saving fuel.

10.9 Internet-based augmentation systems

GDGPS, the Global Differential GPS System (Jet Propulsion Laboratory, The Global Differential GPS System), was invented and implemented by the Jet Propulsion Laboratory, JPL. The system uses a global network of JPL-owned stations to produce a globally valid set of differential corrections. Part of the stations use atomic clocks.

The corrections are sent to the user via the Internet using the TCP protocol. The end-to-end latency can be under five seconds.

A simple user interface to GDGPS is offered called APPS, Automatic

Precise Positioning Service (GDGPS APPS). The user uploads RINEX files to a web page, and the results appear as a web page too.

We present the system here because of its similarity to a satellite-based augmentation system: it could be called an Internet-based augmentation system. Moreover, it served as a kind of prototype for WAAS (NASA Spinoff).

For EGNOS, a somewhat similar solution, SISNeT, has been developed, disseminating the same signal in space as broadcast by the geostationary transponders to users over the Internet (Chen et al., 2003). This is especially useful in high-latitude urban or forested areas, where reception of an uninterrupted signal from a geostationary satellite by a mobile user on the ground may be challenging.

SISNeT is today part of the broader service EDAS, EGNOS Data Access Service, which disseminates EGNOS ground-segment observations to the user by means of the RTCM and NTRIP protocols (EDAS; EDAS Service Definition Document).

maalohko

Self-test questions

1. How does a satellite-based augmentation system (SBAS) work? Which systems are currently in operation in the world?
2. What is GPS signal *integrity*?
3. What is RAIM, receiver autonomous integrity monitoring, and how does it work?
4. How is the data transfer from a geostationary SBAS satellite to users optimised, both for data volume and for latency?
5. In what ways does using an SBAS with the GPS save money and fuel?
6. How does the QZSS work? Describe in particular the choice of orbits.
7. What special problem does the Earth's magnetic field cause for SBAS in Indian airspace?
8. How do ground-based augmentation systems (GBAS) differ from SBAS?

The new era of satellite navigation

11

New global navigation satellite systems (GNSS) are coming, or have already come, online in addition to the well-established Global Positioning System. As these systems will be available and suitable for technological navigation, we present them in this chapter. A summary of the current status is presented in tableau 11.1.

11.1 GPS modernisation

Rapid developments in the field of electronics have made possible, and even easy, things that were unthinkable in the early days of GPS. An example of this are multi-frequency receivers, which back then were expensive specialised instruments. In addition, the monopoly of GPS has been broken. The needs of user communities now counting millions are better understood today and are being taken seriously. For these reasons, modernisation of the system was considered necessary.

The Global Positioning Systems Directorate is in charge of the modernisation effort. The objective of the modernisation is to upgrade the GPS by providing new features such as new civil and military signals and to enhance its performance.

11.1.1 Satellite generations

The successive technology generations of GPS satellites are called *blocks*.

Block I This series of eleven test satellites was used for testing the basic GPS concept. Ten reached orbit, the first in 1978.

Block II This was the first operational GPS satellite model. Nine satellites were launched in 1989–1990. None of these satellites are in operation any more, the last one being decommissioned in 2007.

TABLEAU 11.1. New global navigation satellite systems.

	GPS	GLONASS	Galileo	BeiDou
Orbital planes	6	3	3	3
Separation angle	60°	120°	120°	120°
Satellites per plane	4	8	8	8
Separation angle	90° [a]	45°	45°	45°
Satellites total (official)	$24+6$	$24+3$	$24+6$	$24+3$[b]
Orbit height (km)	20 200	19 140	23 222	21 528
Period	11^h58^m	11^h16^m	14^h05^m	12^h53^m
"Resonance" (orbits/sidereal days)	2/1	17/8	17/10	13/7
Orbital inclination	55°	64°.8	56°	55°

Carrier frequency (MHz)	GPS	GLONASS[c]	Galileo	BeiDou[d]
1600.995		L1		
1575.420	L1		E1	B1C
1561.098				B1I
1278.750			E6	
1268.520				B3I
1248.060		L2		
1227.600	L2			
1207.140			E5b	B2b
1202.025		L3		
1176.450	L5		E5a	B2a

[a] From 2011 onwards the satellites have been unevenly spaced, forming a so-called "expandable" constellation (SPS performance standard, 2020).

[b] There are three more satellites in inclined geosynchronous orbits (IGSO) as well as three in true geostationary orbits.

[c] The new CDMA transmissions.

[d] BeiDou-3.

Block IIA The letter A stands for "advanced". Nineteen satellites of this model were launched during 1990–1997. In 2019, the last satellite was no longer functioning properly and was switched off.

Block IIR The letter R stands for "replenishment". A total of 12 satellites were launched in 1997–2004.

Block IIR-M The letter M stands for "military". Eight satellites were launched in 2005–2009.

Block IIF The letter F stands for "follow-on". A total of 12 satellites were launched in 2010–2016.

Block III Ten satellites of this model will be built. The first launch took place in 2018.

Block IIIF Planned for 2026–2034.

11.1.2 New codes and frequencies

The first Block IIR-M satellite was launched on 25 September 2005. It was the first to transmit the new military M-code. It was also the first to transmit the new civil code L2C, modulated on the L2 carrier. L2C contains the new pseudo-random codes CM (Civil Moderate) and CL (Civil Long), the repeat periods of which are 10 230 and 767 250 bits. Both codes are generated at a modulation frequency ("chip rate") of 511 500 bits per second. They are combined by a technique called *time-division multiplexing*, meaning that bits are taken alternately from CM and CL to form a new sequence, like streams of cars merging at a motorway junction. In this new sequence, each bit is only half as long and the modulation frequency is 1 023 000 bits per second or 1.023 MHz. See figure 11.1. L2C also contains an improved navigation message, containing among other things the time-scale offsets between the GPS time scale and those of other GNSS. This is important in combined use with GLONASS and Galileo.

The new frequency L5, 1176.45 MHz, is meant for safety-of-life (SoL) type use, especially in aviation, see section 10.3. This frequency is in a band internationally reserved for aviation. Block IIF satellites carry this frequency for the first time. In addition, SBAS systems use the L5 frequency. Together with L1, it can be used to eliminate the ionospheric propagation delay from the measurements.

There is also a new civilian L1C code, an attractive alternative to the old C/A code on L1. The code is being broadcast since January 2020 by the first Block III satellite launched in December 2018. The code has a complicated structure, with a data and a pilot component, see figure 7.3. Both are BOC modulated. The pilot uses so-called TMBOC, a time multiplexed binary offset carrier. The BOC modulation provides a better separation from the BPSK-modulated C/A code. The pilot is

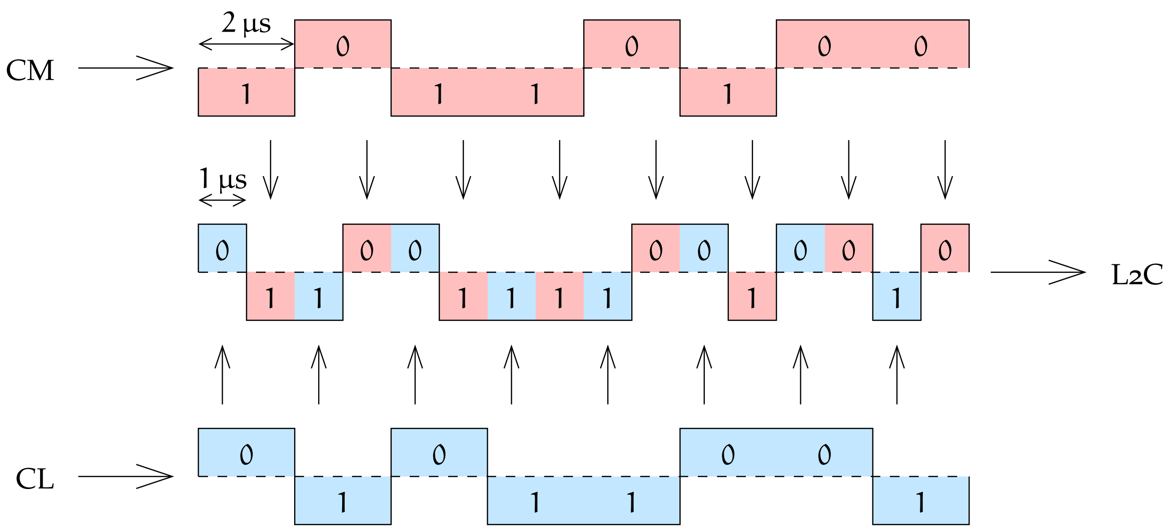


FIGURE 11.1. L2C and time-division multiplexing. L2C also contains a modernised navigation message CNAV at 50 bits per second.

also modulated by an overlay code that is unique for each satellite. According to the ICD (GPS ICD), both data and pilot are in the same phase as the P(Y) code carrier on L1.

So, there are no less than *four* civil GPS signals: C/A, L2C, L5 and L1C.

11.2 The Russian GLONASS system

The GLONASS system dates back to the Soviet era, with the decision to develop it taken in 1976. An official source is GLONASS news, which also shows the constellation status.

In April 2020 there were 27 satellites in orbit, of which 24 were operational. The number of satellites has grown slowly. After a long period of neglect in the 1990s, the system is back to full operationality again.

11.2.1 Frequencies and signals of the original system

The original GLONASS system is described by the Interface Control Document (ICD) from 2008, 65 pages in English, GLONASS ICD.

The original GLONASS differs from both GPS and Galileo in that *every satellite has its own carrier frequency*, a solution known as FDMA, "frequency-division multiple access". However, "antipodes" — satellites in the same orbital plane on opposite sides of the Earth — can use the

taajuusjako-kanavointi

same frequencies. The original frequencies L1 and L2 are

$$f_{L1} = f_{01} + K\Delta f_1,$$
$$f_{L2} = f_{02} + K\Delta f_2,$$

in which $K \in \{-7, -6, \ldots, +5, +6\}$ is the satellite's channel number found from the almanac, and

$$f_{01} = 9 \cdot 178\,\text{MHz} = 1602\,\text{MHz}, \quad \Delta f_1 = \tfrac{9}{16}\,\text{MHz} = 562.5\,\text{kHz},$$
$$f_{02} = 7 \cdot 178\,\text{MHz} = 1246\,\text{MHz}, \quad \Delta f_2 = \tfrac{7}{16}\,\text{MHz} = 437.5\,\text{kHz}.$$

The civil code — corresponding to C/A — has a bit frequency of 0.511 MHz,[1] the encrypted military code 5.11 MHz. The modulation technique is the same as for the GPS: phase modulation with a phase-shift of $\pi = 180°$ — "binary phase-shift keying". The civil code was originally only broadcast on L1, but all satellites currently operating broadcast it on both L1 and L2. The navigation message's bit frequency is 50 Hz.

1

binaarinen vaiheavainnus

The radiation transmitted by the satellites is clockwise circularly polarised, like that of the GPS.

11.2.2 *Time, reference system and constellation*

The GLONASS time scale is the same as UTC(SU), the realisation of UTC for the Russian Federation. This means that the leap seconds of UTC, for which times are reserved at the ends of December and June, enter into the GLONASS time scale. This is different from GPS practice: GPS time does not have leaps.

The navigation message contains the differences between UTC and UTC(SU) as well as announcements of upcoming leap seconds.

The co-ordinate reference frame used by GLONASS is PZ-90 ("Parameters of the Earth 1990"), which is geocentric and co-rotating but slightly different from WGS84. The broadcast ephemeris or orbital prediction gives the location and velocity of the satellite in space in rectangular co-ordinates in this frame, at time intervals of 30 minutes.[2] The latest

vertauskehys

2

[1] The length of the civil code is 511 bits, generated by a single linear feedback shift register of nine bits ($2^9 - 1 = 511$). This makes the repeat period of the code precisely one millisecond.

realisation of PZ-90, PZ-90.11, is within millimetres of ITRF2008, Subirana et al. (2011).

The nominal number of GLONASS satellites in a full constellation is 24 in three orbital planes, plus three spares. The orbital inclination angle is $64^{\circ}\!.8$, a high value serving better the area of the former Soviet Union. As for the GPS, the satellite geometry of GLONASS repeats after a sidereal day. Unlike the GPS however, there will then be a *different* satellite in the *same* place. The orbital height is 19 140 km and the orbital period $11^{\mathrm{h}}16^{\mathrm{m}}$, shorter than half a sidereal day. Only after eight days and 17 orbits will the same geometry repeat itself with the same individual satellites also in the same places. This solution reduces the resonant orbital perturbations, the correction of which consumes rather much thruster propellant for the GPS.

ratahäiriö

11.2.3 Modernisation

koodijako-kanavointi

GLONASS is being switched to CDMA. This technique, "code-division multiple access", is what the other GNSS systems use. This makes the system more compatible with the others and facilitates the design of multi-system receivers. The switch is taking place gradually as new satellites are launched.

The Interface Control Documents (ICD) describing the CDMA solution are GLONASS ICD CDMA (2016a,b,c,d). A third frequency L3 is also being taken into use,

$$f_{L3} = 235 \cdot 5.115\,\mathrm{MHz} = 1202.025\,\mathrm{MHz},$$

GLONASS ICD CDMA (2016d). The old L1 and L2 frequencies are being replaced by new ones:

$$f_{L1} = 313 \cdot 5.115\,\mathrm{MHz} = 1600.995\,\mathrm{MHz},$$
$$f_{L2} = 244 \cdot 5.115\,\mathrm{MHz} = 1248.06\,\mathrm{MHz},$$

close to the old central frequencies. All three frequencies carry both open and encrypted signals.

---

[2] The approach limits the calculation work to be done in the receiver. The Earth's gravity field is modelled as that of a simple ellipsoid of revolution. The acceleration caused by Sun and Moon is given as a separate, constant vector. The navigation message also contains an almanac with conventional extended Kelper orbital elements of all satellites.

TABLE 11.2. GLONASS pseudo-random noise. Current CDMA implementation.

Frequency (MHz)		Type	Modulation	Repeat period (ms)	Phase
L1	1600.995	Data[a]	BPSK(1)	2	Q, TDM[b]
		Pilot	BOC$(1, 1)$	8	
		Encrypted	BOC$(5, \frac{5}{2})$		I
L2	1248.06	CSI[c]	BPSK(1)	-	Q, TDM[b]
		Pilot	BOC$(1, 1)$	20	
		Encrypted	BOC$(5, \frac{5}{2})$		I
L3	1202.025	Data[a]	BPSK(10)	1	I
		Pilot	BPSK(10)	1	Q

[a] Carries a navigation message.

[b] Time-division multiplexing.

[c] Channel for service information.

For example, the bit frequency of the C/A-code equivalent L1 civil code remains *almost* unchanged at 0.5115 MHz, now *exactly* half that of the GPS C/A code bit frequency. The length of the code is now 1023 bits (data) and 4092 bits (pilot), taking precisely 2 ms and 8 ms to transmit. Also the bit frequency of the L2 civil code (channel for service information and pilot) is 0.5115 MHz. The bit frequency of the L3 codes is 10.23 MHz.

The navigation message is broadcast on L1 and L3. The bit frequency on L1 is 125 Hz and on L3 100 Hz.

Expanding the space segment to thirty satellites in six planes is planned for the future. In the longer term, harmonisation of the carrier frequencies to the GPS frequencies L1 and L5, also used by Galileo and BeiDou, and the Galileo frequency E5b, also used by BeiDou, is foreseen with compatible modulations, which would make the design of multi-system receivers and software easier still.

avaruuslohko

11.3 The European Galileo system

The Galileo system is a joint project of the European Commission and the European Space Agency (ESA). Its administrative history has been long and winding. In June 2003, the European Commission and ESA formed an organ called the "Galileo Joint Undertaking" (GJU), which chose a "*concessionaire*", a private "Galileo Operating Company" responsible for Galileo's daily operations and especially its commercialisation. In 2006, the GJU was disbanded and the "European GNSS Supervisory Authority" (GSA) created. In 2010 it was reorganised and given the new name "European GNSS Agency (GSA)".

An important motivation for Galileo was and is achieving technological independence from the United States, the importance of which was underscored by the latter's unilateral invasion of Iraq in March 2003. Galileo is officially a civilian system and is administered by civilian authorities, but is an integral part of the defence planning of many nations.

11.3.1 Satellites and orbits

There will be thirty Galileo satellites, in three orbital planes, in each of which are eight satellites and two spares. The distance between satellites within the orbital plane is 45°. The inclination of the orbital planes with respect to the equator is 56°.

The height of the orbits of the Galileo satellites is 23 222 km, clearly higher than the GPS orbits. The orbital period is $14^{\mathrm{h}}05^{\mathrm{m}}$.

After ten days and 17 revolutions, the same satellites will again be in the same places in the sky of the observer.

The first experimental Galileo satellite, GIOVE-A, was launched on 28 December 2005. There are (in 2020) 22 working satellites in orbit and more coming. Real-time information on the constellation is provided by the European GNSS Service Centre.

tosiaikainen

11.3.2 System description, components

The Galileo system is designed to be compatible with the GPS. It is also intended to work seamlessly with SBAS systems, like EGNOS in Europe.

Galileo's signal and frequency structure is complex, see ESA, Galileo Navigation Signals and Frequencies. The carriers are

E1 1575.42 MHz (the same as GPS L1)

E5a 1176.45 MHz (the same as GPS L5)

E5b 1207.14 MHz

E6 1278.75 MHz.

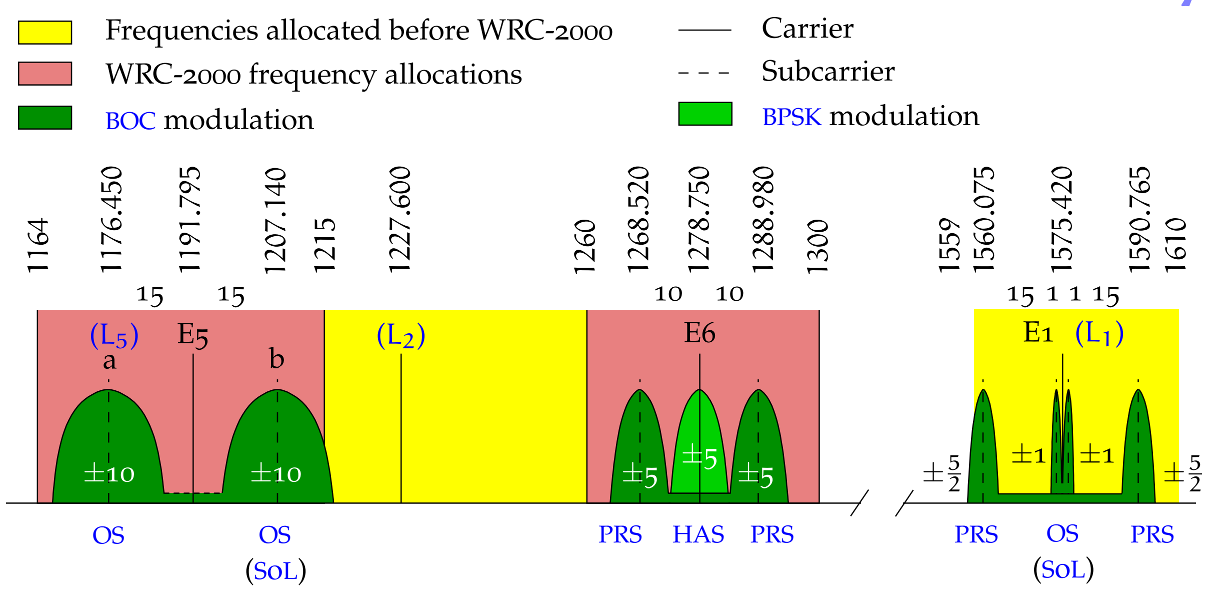


FIGURE 11.2. Galileo frequencies and modulations, unit MHz. WRC-2000 was the World Radiocommunications Conference in Istanbul 2000. This is a stylised diagram, in which the half-widths of the green areas (main lobes) represent code bit rates. Frequency differences and spreads are in "GPS megahertzes" of 1.023 MHz each.

See tableau 11.1 and Galileo OS SiS ICD.

11.3.3 Services

The services offered by Galileo are:

- Open service (OS). Useable by anyone. [E1, E5]
- Safety-of-life service (SoL). Galileo itself does not as such offer safety-of-life services, but EGNOS (section 10.5) does. Use of ARAIM, advanced RAIM, for integrity monitoring is also foreseen. [E1, E5b] eheyden valvonta
- High-accuracy service (HAS). An earlier name for this was "commercial service". [E6] tarkkuus-palvelu
- Public regulated service (PRS). This service is better protected against interference. Users include the police, border guards, defence forces, and peacekeeping forces. The fact that Galileo is called, "unlike the GPS, a civilian system", is perhaps not quite accurate. . . . [E1, E6] viranomais-palvelu

The Galileo satellites also carry search-and-rescue (SAR) transponders belonging to the Cospas-Sarsat system.

See Ávila Rodríguez (2011) and figure 11.2.

11.4 The Chinese BeiDou system

The name of the system, BeiDou, is the name of the constellation Big Dipper (*Ursa Major*), which is used to find the North Star (Polaris, α UMi). Thus, the name is symbolic for navigation. Internationally, the name "Compass" has also been used. The official English name is "BeiDou Navigation Satellite System". The official web site is BeiDou Navigation Satellite System.

11.4.1 BeiDou-1

The older BeiDou-1 system, which served the territory of China, consisted of three satellites in geostationary orbits. It was decommissioned in 2012.

It has been discovered that the system functioned autonomously: the ground station sent a signal through two satellites to a receiver at an unknown location, which also answered through two satellites — it was thus an *active* system. An accurate terrain model of China was used as an additional constraint. The computed position was sent back to the receiver. Reported positioning precisions ranged from ± 20 m to ± 100 m.

11.4.2 BeiDou-2 and -3

The BeiDou-2 and -3 systems are also called "Compass" in English.

The BeiDou-2 series comprises (2020) ten satellites, of which five are in geostationary orbits, offering positioning services in the Asia-Pacific region. Launch activity was intense in 2010–1012 (BeiDou Constellation Status).

Then, in November 2017, China started launching BeiDou-3 satellites at a high rate. On 27$^{\text{th}}$ December 2018, it was announced (Xinhua, 2018) that the service was globally available. Reaching a full constellation of 35 satellites, of which 5 are geostationary, appears on schedule for 2020.

Both BeiDou-2 and BeiDou-3 transmit in three frequency bands, B1, B2 and B3. The documented public-service signals of BeiDou-3 are B1C, B1I, B3I and B2a, see the Interface Control Documents BeiDou ICD
3 (2017, 2018a,b, 2019). There is currently no ICD for B2b.[3]

Table 11.3. BeiDou-3 pseudo-random noise.

Signal	Frequency (MHz)	Type	Modulation	Repeat period (ms)	Phase
B2a[a]	1176.45	Data[b]	BPSK(10)	1	I
		Pilot	BPSK(10)	1	Q
B2b	1207.14	Data[b]	BPSK(10)	1	
B3I	1268.52	Data[b]	BPSK(10)	1	
B1I[c]	1561.098	Data[b]	BPSK(2)	1	
B1C[d]	1575.42	Data[b]	BOC(1, 1)	10	I
		Pilot	BOC(1, 1) } QMBOC[e]	10	Q
			BOC(6, 1) } QMBOC[e]	10	I

[a]Identical to GPS L5.

[b]Carries a navigation message.

[c]Will be replaced by B2a.

[d]Resembles L1C.

[e]Quadrature multiplexed BOC.

- B1C operates on a carrier frequency of 1575.42 MHz, coinciding with the GPS L1 frequency. It resembles the GPS signal L1C and the Galileo E1 OS (open service) signal, except for the use of quadrature.
- B2a operates on a frequency of 1176.45 MHz, the same as the GPS frequency L5 and the Galileo frequency E5a. The modulation also appears identical to that of the GPS and Galileo.

See tableau 11.1 and table 11.3.

The BeiDou public-service signal is unencrypted and freely available. The precision is ±10 m for BeiDou-2, and the planned precision for BeiDou-3 is ±0.5 m. The BeiDou-3 signal also includes integrity information, see section 10.3. eheys

The orbits are a little higher than those of the GPS, the height being 21 528 km and the inclination 55°.

The BeiDou-3 system includes three satellites in so-called *inclined*

[3]Such an ICD has since appeared, with a publication date of July 2020: BeiDou ICD (2020).

geosynchronous orbits (IGSO) at 35 786 km, as well as three satellites in true geostationary orbits.

Relevant documents and current information on constellation status and satellite health are found at BeiDou Introduction.

Self-test questions

1. How does time-division multiplexing work?
2. Explain the difference between code-division multiple access (CDMA) and frequency-division multiple access (FDMA).
3. Why is the inclination of the orbits of the GLONASS satellites so much larger than the orbital inclinations of the other systems?
4. What are the various services offered by Galileo?
5. SBAS receivers work on two frequencies: GPS L1 and L5. This applies for all SBAS systems, WAAS, EGNOS, MSAS, QZSS, and GAGAN. Explain why this standard was inevitable in a field like aviation.
6. SBAS receivers work on two frequencies: GPS L1 and L5. These are identical to the Galileo frequencies E1 and E5a, and to the BeiDou frequencies B1C and B2A. Why do you think this choice has been made by the designers of Galileo and BeiDou?

Measuring gravity in flight

12

The saying is well-known:

> "One guy's noise is the other guy's signal."

Inertial navigation is based on assuming that the Earth's gravity field is known. Then, from the starting location $\mathbf{x}(t_0)$ and the starting velocity $\mathbf{v}(t_0)$, we can calculate location and velocity forwards in time, step by small step, to obtain the location and velocity $\mathbf{x}(t), \mathbf{v}(t)$ for any later time t.

Conversely however, if there is an *independent* source of information, such as a GNSS receiver, that gives the change in location and velocity over time with sufficient precision, then inertial technology can be harnessed to survey the Earth's gravity field.

Because well-working navigation satellite systems exist today, it is possible to perform gravimetric measurements from the air. Moreover, continuously tracking the precise three-dimensional motion of a satellite using GNSS is one technique for mapping the gravity field of the Earth.

Let the position of the aircraft or satellite in an inertial frame as a function of time be $\mathbf{x}^*(t)$ and its discrete measurement time series $\mathbf{x}_i^* \stackrel{\text{def}}{=} \mathbf{x}^*(t_i)$. Then, the geometric acceleration can be approximated as follows:

$$\mathbf{a}^*(t_i) = \left.\frac{d^2}{dt^2}\mathbf{x}^*\right|_{t_i} \approx \frac{\mathbf{x}_{i+1}^* + \mathbf{x}_{i-1}^* - 2\mathbf{x}_i^*}{\Delta t^2},$$

in which $\Delta t \stackrel{\text{def}}{=} t_{i+1} - t_i$ is the time interval between successive epochs.

Assume that at the same time the "gravity" $\widetilde{\mathbf{g}}$ *sensed* on board the aircraft is measured with a vector accelerometer. At this point, we also assume for simplicity that $\mathbf{x}^*$ and $\widetilde{\mathbf{g}}$ are given on the same axes frame, meaning that the directions of the acceleration measurement axes are the same as those of the location co-ordinate axes.

Then, in an inertial frame it holds according to equation 5.14 that

$$\mathbf{g}^* = \widetilde{\mathbf{g}} + \mathbf{a}^*, \tag{12.1}$$

with $\widetilde{\mathbf{g}} = -\widetilde{\mathbf{a}}$. Now by equation 5.15, gravity in a frame co-rotating with the Earth is

$$\mathbf{g} = \mathbf{g}^* - \Big\langle \boldsymbol{\omega}_\oplus \times \langle \boldsymbol{\omega}_\oplus \times \mathbf{x} \rangle \Big\rangle. \tag{12.2}$$

Substitute equation 5.13 into equation 12.1:

$$\mathbf{g}^* = \widetilde{\mathbf{g}} + \mathbf{a} + 2 \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle + \Big\langle \boldsymbol{\omega}_\oplus \times \langle \boldsymbol{\omega}_\oplus \times \mathbf{x} \rangle \Big\rangle.$$

Substituting this into equation 12.2 yields

$$\mathbf{g} = \widetilde{\mathbf{g}} + \mathbf{a} + 2 \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle. \tag{12.3}$$

12.1 Airborne vector gravimetry

If an aircraft carries both an inertial measurement unit and a GNSS receiver, we can determine $\mathbf{x}(t_i)$, $\mathbf{v}(t_i)$, $\mathbf{a}(t_i)$, and $\widetilde{\mathbf{g}}(t_i)$, and by equation 12.3 we can calculate $\mathbf{g}(t_i)$. This is a method for surveying the gravity field from the air.

In practice, the data streams generated by both the GNSS device and the inertial device are fed into a Kalman filter, which outputs the plane's precise route and gravity profile. In airborne vector gravimetry, gravity is measured as a three-dimensional vector. The data rate is typically high, many observation epochs per second. Because of the motions of the aircraft, the variations over time of both geometric acceleration $\mathbf{a}$ and sensed gravity $\widetilde{\mathbf{g}}$ on board are large, thousands of milligals, but the final determination precision of $\mathbf{g}$ can be as good as several milligals.

Airborne vector gravimetry does not reach quite the same level of precision as airborne scalar gravimetry, to be introduced next. This is because the accelerometers in the inertial device, precise as they are, are not as good as the best gravimeters.

12.2 Airborne scalar gravimetry

In this technique, a traditional *gravimeter*, an instrument for measuring gravity, is used. The gravimeter has been modified in a way that makes it possible to take measurements in environments where, in addition to gravity, strong, varying geometric accelerations have an impact. The vaimennus modification, *damping*, is the same as that which is made to make measurements at sea possible. The gravimeter is mounted on a *stabilised platform*, the stabilisation being done with the aid of gyroscopes, section 5.5.

The gravimeter measures the acceleration of free fall "sensed" inside the vehicle, but only in the direction of the local vertical or plumb line. If the direction of the local plumb line is the unit vector $\mathbf{n}$ (*downwards*), the measured quantity is $\widetilde{g} \stackrel{\text{def}}{=} \langle \mathbf{n} \cdot \widetilde{\mathbf{g}} \rangle$. luotiviiva

We start from equation 12.3:

$$\mathbf{g} = \widetilde{\mathbf{g}} + \mathbf{a} + 2 \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle. \tag{12.3}$$

Compute the scalar product with $\mathbf{n}$:

$$\langle \mathbf{n} \cdot \mathbf{g} \rangle = \langle \mathbf{n} \cdot \widetilde{\mathbf{g}} \rangle + \langle \mathbf{n} \cdot \mathbf{a} \rangle + 2 \Big\langle \mathbf{n} \cdot \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle \Big\rangle = \|\mathbf{g}\| \stackrel{\text{def}}{=} g,$$

because the direction of the plumb line is that of gravity. So

$$g = \widetilde{g} + \langle \mathbf{n} \cdot \mathbf{a} \rangle + 2 \Big\langle \mathbf{n} \cdot \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle \Big\rangle.$$

For an aircraft standing on the ground, $\mathbf{a} = \mathbf{v} = 0$, so $g = \widetilde{g}$. If the aircraft is flying, the term

$$\langle \mathbf{n} \cdot \mathbf{a} \rangle = -\frac{d}{dt} v_{\text{up}} + \frac{v_{\text{east}}^2}{N + h} + \frac{v_{\text{north}}^2}{M + h}$$

represents the vertical acceleration. Here, $v_{\text{up}} \stackrel{\text{def}}{=} \frac{d}{dt} h$, with h the height from the reference ellipsoid. Using the reference ellipsoid in this way brings in two terms from the downward curvature of the surface of the Earth: the vertical acceleration only equals the second derivative of height in the absence of horizontal motion. $M(\varphi)$ and $N(\varphi)$ are the radii of curvature in the meridional (south-north) and transversal (west-east) directions of the Earth's reference ellipsoid; in other words, the meridional and transversal radii of curvature. vertaus-ellipsoidi phi $\varphi\phi\Phi$

The term

$$2 \Big\langle \mathbf{n} \cdot \langle \boldsymbol{\omega}_\oplus \times \mathbf{v} \rangle \Big\rangle = 2\omega_\oplus \cos\varphi \cdot v_{\text{east}}$$

represents the Coriolis acceleration due to aircraft velocity in the west-east direction, which has a vertical component: the aircraft is a little lighter flying along with the rotation of the Earth, a little heavier flying against it.

Expressed in geodetic co-ordinates (φ, λ, h) we obtain lambda $\lambda\Lambda$

$$g = \widetilde{g} - \frac{d}{dt} v_{\text{up}} +$$

$$
+ (N+h)\overbrace{\left(\left(\omega_\oplus \cos\varphi + \frac{v_{\text{east}}}{N+h}\right)^2 - (\omega_\oplus \cos\varphi)^2\right)}^{\text{west-east, difference}} + \overbrace{\frac{v_{\text{north}}^2}{M+h}}^{\text{south-north}} =
$$

$$
= \tilde{g} - \frac{d}{dt}v_{\text{up}} + \left(\frac{v_{\text{east}}}{N+h} + 2\omega_\oplus \cos\varphi\right) v_{\text{east}} + \frac{v_{\text{north}}^2}{M+h}.
$$

Here, v_{east} and v_{north} are the east and north components of the velocity, and $\omega_\oplus$ is the angular velocity of the Earth's rotation. The symbols h and φ denote the height above the reference ellipsoid and geographic latitude.

omega $\omega\Omega$

The gravity g already contains the vertical part of the centrifugal acceleration caused by the Earth's rotation, $-(N+h)(\omega_\oplus \cos\varphi)^2$, just as the gravity vector $\mathbf{g}$ contains the centrifugal acceleration of Earth rotation, $-\left\langle \boldsymbol{\omega}_\oplus \times \left\langle \boldsymbol{\omega}_\oplus \times \mathbf{x}\right\rangle\right\rangle$, as a vector, equation 12.2.

In the above equation, the two last terms are called the Eötvös[1] correction, see Wei and Schwarz (1996).

12.3 Using the Kalman filter in airborne gravimetry

Re-write equation 12.3,

$$
\mathbf{g} = \tilde{\mathbf{g}} + \mathbf{a} + 2\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{v}\right\rangle, \tag{12.3}
$$

stochastically with the "dynamic noise" $\underline{\mathbf{n}}_a$ included:

$$
\underline{\mathbf{a}} = \frac{d^2}{dt^2}\underline{\mathbf{x}} = \underline{\mathbf{g}} - \underline{\tilde{\mathbf{g}}} - 2\left\langle \boldsymbol{\omega}_\oplus \times \underline{\boldsymbol{v}}\right\rangle + \underline{\mathbf{n}}_a.
$$

In geocentric co-ordinates we may write, with definition 5.10:

$$
\left\langle \boldsymbol{\omega}_\oplus \times \boldsymbol{v}\right\rangle = \left\langle \begin{bmatrix} 0 \\ 0 \\ \omega_\oplus \end{bmatrix} \times \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \right\rangle = \omega_\oplus \begin{bmatrix} -v_y \\ v_x \\ 0 \end{bmatrix} =
$$

$$
= \omega_\oplus \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \stackrel{\text{def}}{=} \omega_\oplus \Omega \boldsymbol{v}.
$$

Form a system of first-order differential equations:

$$
\frac{d}{dt}\begin{bmatrix} \underline{\mathbf{x}} \\ \underline{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \underline{\boldsymbol{v}} \\ \underline{\mathbf{g}} - \underline{\tilde{\mathbf{g}}} - 2\omega_\oplus \Omega \underline{\boldsymbol{v}} \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{\mathbf{n}}_a \end{bmatrix}.
$$

[1] Loránd baron Eötvös de Vásárosnamény (1848–1919) was a Hungarian physicist and student of gravitation.

Here $\widetilde{\mathbf{g}} = \widetilde{\mathbf{g}}(t)$ is a measured quantity, but $\mathbf{g} = \mathbf{g}(t)$ is not.

Write

$$\mathbf{g} = \boldsymbol{\gamma} + \delta\mathbf{g},$$

in which $\boldsymbol{\gamma}$ is a suitable reference value, for example calculated from the normal gravity field, and $\delta\mathbf{g}$ is the *gravity disturbance*. Model $\delta\mathbf{g}$ empirically as a Gauss-Markov process, equation 2.28:

gamma $\gamma\Gamma$
vertausarvo

$$\frac{d}{dt}\underline{\delta\mathbf{g}} = -\frac{\underline{\delta\mathbf{g}}}{\tau} + \underline{\mathbf{n}}_g,$$

in which τ is a suitably chosen empirical time constant. Its choice depends on the behaviour of the local gravity field (correlation length) and the flying speed and height.

tau τT

Now the dynamic model of the Kalman filter is

$$\frac{d}{dt}\begin{bmatrix} \underline{\mathbf{x}} \\ \underline{\mathbf{v}} \\ \underline{\delta\mathbf{g}} \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ 0 & -2\omega_\oplus\Omega & I \\ 0 & 0 & -\frac{1}{\tau}I \end{bmatrix}\begin{bmatrix} \underline{\mathbf{x}} \\ \underline{\mathbf{v}} \\ \underline{\delta\mathbf{g}} \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{\gamma} - \widetilde{\mathbf{g}} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \underline{\mathbf{n}}_a \\ \underline{\mathbf{n}}_g \end{bmatrix}.$$

So, the length of the state vector is 9. The coefficient matrix is of size 3×3 and consists of 3×3 sized submatrices, for a size of 9×9 elements in total.

A more sophisticated approach considers that gravity is a function of the unknown location $\mathbf{x}$:

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}(\mathbf{x}) \approx \boldsymbol{\gamma}\left(\mathbf{x}^{(0)}\right) + \mathcal{M}\left(\mathbf{x}^{(0)}\right)\Delta\mathbf{x},$$

in which $\mathbf{x}^{(0)} = \mathbf{x}^{(0)}(t)$ is the *approximate* location used in linearisation, see below. Here the *gravitation-gradient tensor* $\mathcal{M}$ pops up, equation 3.9.

Then, $\mathbf{x}$ and $\mathbf{v}$ must also be linearised, and linearised state elements $\Delta\mathbf{x} \stackrel{\text{def}}{=} \mathbf{x} - \mathbf{x}^{(0)}$ and $\Delta\mathbf{v} \stackrel{\text{def}}{=} \mathbf{v} - \mathbf{v}^{(0)}$ must be used in the state vector. The equation defining the approximate values is

$$\frac{d}{dt}\begin{bmatrix} \mathbf{x}^{(0)} \\ \mathbf{v}^{(0)} \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{x}^{(0)} \\ \mathbf{v}^{(0)} \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{\gamma}(\mathbf{x}^{(0)}) - \widetilde{\mathbf{g}} \end{bmatrix}.$$

The final result is obtained by subtraction:

$$\frac{d}{dt}\overbrace{\begin{bmatrix} \Delta\underline{\mathbf{x}} \\ \Delta\underline{\mathbf{v}} \\ \underline{\delta\mathbf{g}} \end{bmatrix}}^{\mathbf{x}} = \overbrace{\begin{bmatrix} 0 & I & 0 \\ \mathcal{M} & -2\omega_\oplus\Omega & I \\ 0 & 0 & -\frac{1}{\tau}I \end{bmatrix}^{(0)}}^{F}\overbrace{\begin{bmatrix} \Delta\underline{\mathbf{x}} \\ \Delta\underline{\mathbf{v}} \\ \underline{\delta\mathbf{g}} \end{bmatrix}}^{\mathbf{x}} + \overbrace{\begin{bmatrix} 0 \\ \underline{\mathbf{n}}_a \\ \underline{\mathbf{n}}_g \end{bmatrix}}^{\mathbf{n}}.$$

The GNSS observation equation for the update step is

$$\boldsymbol{\ell}_k = \underline{\mathbf{x}}(t_k) + \underline{\mathbf{m}} = \underline{\mathbf{x}}^{(0)}(t_k) + \Delta\underline{\mathbf{x}}(t_k) + \underline{\mathbf{m}},$$

in which the "noise vector" $\underline{\mathbf{m}}$ describes the statistical uncertainty of GNSS navigation.

For both the dynamic noise terms $\underline{\mathbf{n}}_a$ and $\underline{\mathbf{n}}_g$, representing the uncertainty of the on-board gravity measurement $\widetilde{\mathbf{g}}$ and the variability of the gravity field $\delta\underline{\mathbf{g}}$, and the GNSS observation noise $\underline{\mathbf{m}}$ we need to build suitable statistical models (variances $Q_n(t)$ and R) based on the properties of the measuring devices and the gravity field.

It is good to keep in mind the integrating nature of the Kalman filter. The measurements $\widetilde{\mathbf{g}}$ contain strong high-frequency, sub-second variations due to aircraft motions, which are readily absorbed into the estimates of the speed $\mathbf{v}$ and location $\mathbf{x}$ of the aircraft as they should, having no effect on $\delta\mathbf{g}$ with its long correlation time τ. For good spatial along-track resolution on the ground, the flight speed should be as low as possible. Propeller aircraft like the popular Canadian DHC-6 Twin
2 Otter®[2] are suitable for this.

12.4 Inter-sensor calibration

In the above theory, it was assumed that the GNSS positioning measurements and the gravimetric measurements refer to the same location inside the aircraft. Of course they do not: the GNSS antennas must be
monitie mounted on the roof — by the way a nightmarish cause of multipath reflections — whereas the gravimeter is located lower down inside. Integrating these instruments requires determining the vector offsets between them by calibration.

A good practice is *in-flight calibration*: the airframe is flexible, so these vectors will be different on the ground and in the air and will depend on fuel load. The vector offsets have to be included as unknowns in processing the measurements, usually with a Kalman filter, section 12.3.

12.5 Present state of airborne gravimetry

One of the first successful airborne gravimetric projects was Brozena (1992), the gravity survey of Greenland in the frame of the Greenland Aerogeophysics Project.

[2]Twin Otter is a registered trademark of Viking Air Ltd.

FIGURE 12.1. A Lockheed Hercules C-130[a] taking off — on skis! — from the NorthGRIP ice-coring site, Greenland, three kilometres above sea level. Note the JATO (jet-assisted take-off) bottles helping out on a runway that is also three kilometres long. Nielsen (2005).

[a]C-130 is a registered trademark of The Lockheed Martin Corporation.

Many later measurements, often in Arctic or Antarctic locations, may be mentioned (Forsberg et al., 1996, 2011). The logistic requirements of working there are commonly "challenging", see figure 12.1.

Airborne gravimetry is a suitable technique if the area to be surveyed is large and there are no earlier terrestrial gravity surveys available. *Homogeneity* is one the advantages of airborne and space gravimetry: the quality of the measurement is the same over large areas and long-range systematic errors are small. This is especially important if the gravimetric data is meant for the determination of a geoid model.

Examples of airborne gravity surveys include Ethiopia (Bedada, 2010), Mongolia (Munkhtsetseg, 2009), and many more countries.

12.6 Studying the gravity field of the Earth from space

In equation 12.3, the quantity $\widetilde{\mathbf{g}}$ is roughly the magnitude of the Earth's surface gravity (about $10\,{}^{\mathrm{m}}/_{\mathrm{s}^2}$), while the geometric acceleration $\mathbf{a}$ is much smaller. In the ideal case, the geometric acceleration would be

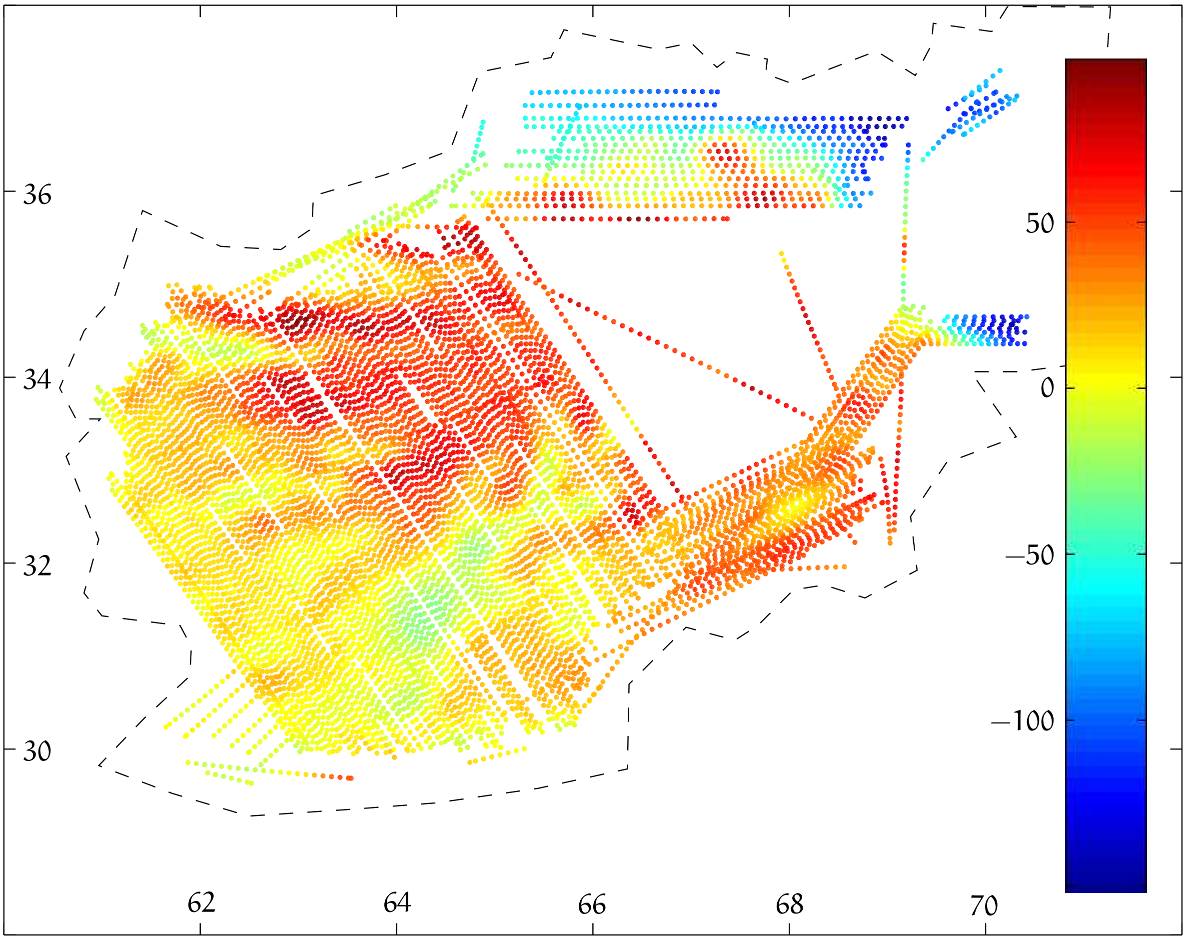


FIGURE 12.2. Airborne gravimetric survey of Afghanistan. Colours represent free-air anomalies at flight height in milligals (USGS Open-File Report 2008-1089). Original data decimated 50×.

zero, corresponding to measurements on the surface of the solid Earth. In both sea and airborne gravimetry, the geometric acceleration differs however clearly from zero, which complicates the accurate measurement of gravity. From the viewpoint of measuring, the movements of the vehicle are disturbances.

In the measurement of the gravity field from space, the situation is the opposite. In equation 12.1, the local gravity $\widetilde{\mathbf{g}}$ sensed inside the satellite is zero or very close to zero: *weightlessness*. The geometric acceleration $\mathbf{a}^*$ is almost the magnitude of gravity at the Earth's surface, because the satellite is falling freely the whole time while flying in orbit. The geometric acceleration is all the time being measured with the help of GNSS — so-called "high-low satellite-to-satellite tracking" — and the satellite's own non-inertial motion, or acceleration, $\widetilde{\mathbf{a}} = -\widetilde{\mathbf{g}}$ is also measured with the aid of accelerometers. The largest cause of non-inertial acceleration is atmospheric friction or drag, because the orbit of a satellite for studying the gravity field is chosen to be as low as possible, the typical height being 250–400 km.

During recent decades, three different gravity missions have flown: CHAMP, GRACE and GOCE.

- CHAMP (GFZ, CHAMP — Challenging Minisatellite Payload), a German satellite which operated from 2000 to 2010, producing a substantial amount of data.
- GRACE (University of Texas, GRACE — Gravity Recovery and Climate Experiment), an American-German satellite pair which measured with its special equipment the accurate distance between two satellites ("Tom" and "Jerry") flying in tandem. The technique is called low-low satellite-to-satellite tracking. The measurement objective was to monitor temporal changes in the Earth's gravity field. The mission, from 2003 to 2018, has been a great success. An animation of its results can be found in figure 12.3. The successor mission, GRACE-FO ("GRACE Follow-On"), is currently operating.
- GOCE (Gravity Field and Ocean Circulation Explorer) surveyed the Earth's gravity field in great detail during 2009–2013, using a *gravitation gradiometer*, see ESA, Introducing GOCE. The GOCE satellite contained an *ionic engine* in order to compensate for the air drag and make a low orbit possible. It was a design challenge to separate the gravitation-gradient measurements from the effects of air drag and the satellite's own slow rotation as it circled the Earth.

A GNSS receiver and accelerometers were included in all the satellites; in the case of GOCE even an array, a gradiometer, counting six extremely sensitive accelerometers.

Self-test questions

1. What technology development made gravimetric measurements from the air suddenly possible?
2. What are the differences between vector and scalar airborne gravimetry?
3. What is the difference between an on-board gravity measurement $\widetilde{\mathbf{g}}$ and an on-board acceleration measurement $\widetilde{\mathbf{a}}$?

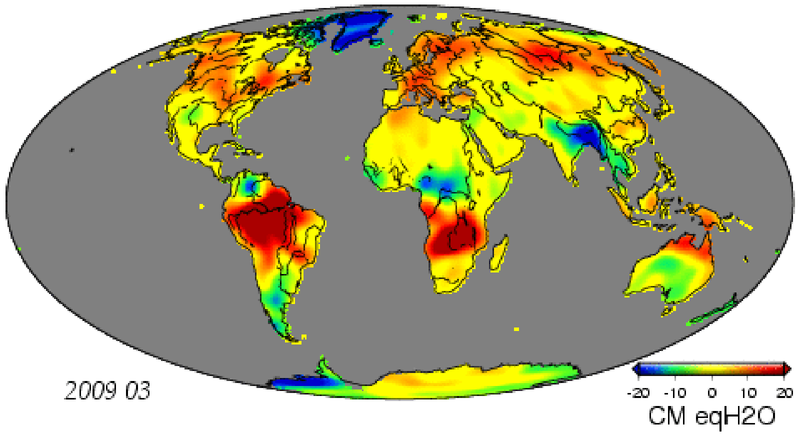


FIGURE 12.3. Results from the GRACE mission presented as a surface mass layer in centimetres of water equivalent. Click for animation (e-book).

4. What is the Eötvös correction?
5. Look at figure 12.1 and read the caption. Why would the Hercules need JATO rockets to help it take off (two reasons)? How does one build a runway on top of a continental ice sheet?
6. Why is the calibration method of *in-flight* calibration preferable for determining the offset vectors between different instruments ?
7. Why did the GOCE satellite have an ionic engine?
8. What is high-low and low-low satellite-to-satellite tracking?

Sensor fusion, sensors of opportunity

13

The art of technological navigation originated in the maritime sphere a long time ago. It was also adopted in aviation and spacefaring as those fields developed. Today, we see technological navigation proliferate into the lives of millions of ordinary citizens in the form of car and pedestrian navigation.

There is a multitude of sensors available, both intended for navigation, like satellite positioning and inertial sensors, and intended for other uses but adaptable for navigation, like mobile base-station networks and indoor wireless local-area (WLAN) networks. This begs the question of how to best use these sensors together in an integrated way.

Sensor fusion or *sensor integration* is the use of multiple sensors towards a reduction of uncertainty, for example in location determination during navigation (Wikipedia, Sensor fusion). The sensors may be similar, in which case we speak of homogeneous sensor fusion, or different, which is called heterogeneous sensor fusion.

Sensor fusion is often realised by using the output of the various sensors as input observations to a Kalman or similar filter. If the raw observations are used and the sensors are modelled inside the filter software, we speak of "tightly integrated" sensor fusion. If the sensors do their own processing before sending output to the filter software, we speak of "loosely integrated" sensor fusion.

Sensors of opportunity are sensors that serve some other purpose but can be harnessed for the purpose of positioning or navigation. Examples.

- The accelerometer in a mobile phone. Its primary purpose is to orient the screen display upright (portrait or landscape) depending on how the user is holding the phone, but it can be used to inform a navigator of the direction of local gravity.

- The speed meter in a car. Its primary purpose is to inform the driver of the speed of the car for safety and law abidance.
- The air-pressure sensor in an aircraft. Barometric pressure is used as a measure of altitude to keep aircraft safely apart vertically. However, air pressure is also a measure of the amount of atmospheric matter above the aircraft. This amount affects the propagation of the GNSS satellite signal.

Sensor fusion may also incorporate *background knowledge* that is not an actual sensor measurement. For example:

- A traditional technique is using known landmarks.
- A vehicle may stop and inform the filter software that now the velocity is zero — a *zero-velocity update*.

nollanopeus-päivitys

- A mobile phone used by a pedestrian or car navigator may assume that the upper edge of the screen is pointing in the direction of motion.
- A car is assumed to always have zero sideways velocity.
- A car is always assumed to follow the road, like in a tunnel outside GNSS reach.
- Upper bounds for accelerations in various directions may be assumed. For vehicles carrying human beings, such values may be imposed by what the human body will endure.

In the following, we shall discuss some examples of sensor-fusion and sensors-of-opportunity technologies. This is a broad and expanding field based on rapidly developing technologies. We will only scratch the surface here.

13.1 Case: Sky Map

An interesting example of the creative use of mobile-phone sensors is the open-source application Sky Map for the Android™[1] operating system (Google Play, Sky Map). It shows, after calibration, the stars in their proper locations when holding up the telephone to the sky. It even shows stars and other objects, including planets, that are below the horizon!

[1] Android is a common-law trademark of Google Inc.

Calibration involves rotating the device manually around three different axes: the "figure-of-eight" motion recommended by the manufacturer. Apparently, the application uses the three-axis accelerometer commonly found in mobile phones, as well as the magnetometer for finding the compass north.

We can analyse this situation as follows. Let the rotation matrix between the body frame of the telephone and an Earth-fixed frame be R. Let the acceleration vector of gravity be $\mathbf{g}$ and the magnetic field-strength vector be $\mathbf{m}$. It is assumed that both vectors are computable in the Earth-fixed frame and are measured in the telephone body frame.

Based on these two vectors, we may construct an *orthonormal basis* $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ by applying Gram-Schmidt orthonormalisation as follows: ortonormaali kanta

$$\mathbf{u} = \frac{\mathbf{g}}{\|\mathbf{g}\|}, \quad \mathbf{v} = \frac{\mathbf{m} - \langle \mathbf{u} \cdot \mathbf{m} \rangle}{\|\mathbf{m} - \langle \mathbf{u} \cdot \mathbf{m} \rangle\|}, \quad \mathbf{w} = \langle \mathbf{u} \times \mathbf{v} \rangle. \tag{13.1}$$

These vectors have in either frame the following components:

$$\begin{aligned}
\mathbf{u} &= a_{11}\mathbf{i} + a_{12}\mathbf{j} + a_{13}\mathbf{k} = b_{11}\mathbf{i}' + b_{12}\mathbf{j}' + b_{13}\mathbf{k}', \\
\mathbf{v} &= a_{21}\mathbf{i} + a_{22}\mathbf{j} + a_{23}\mathbf{k} = b_{21}\mathbf{i}' + b_{22}\mathbf{j}' + b_{23}\mathbf{k}', \\
\mathbf{w} &= a_{31}\mathbf{i} + a_{32}\mathbf{j} + a_{33}\mathbf{k} = b_{31}\mathbf{i}' + b_{32}\mathbf{j}' + b_{33}\mathbf{k}',
\end{aligned} \tag{13.2}$$

in which $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ and $\{\mathbf{i}', \mathbf{j}', \mathbf{k}'\}$ are the orthonormal bases of the Earth-fixed and telephone body frames, respectively.

With these definitions, we see that the matrices

$$A \stackrel{\text{def}}{=} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \qquad B \stackrel{\text{def}}{=} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

are both orthogonal. Write equation 13.2 as

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \end{bmatrix} = A \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix} = B \begin{bmatrix} \mathbf{i}' \\ \mathbf{j}' \\ \mathbf{k}' \end{bmatrix} \Longrightarrow \begin{bmatrix} \mathbf{i}' \\ \mathbf{j}' \\ \mathbf{k}' \end{bmatrix} = B^{-1}A \begin{bmatrix} \mathbf{i} \\ \mathbf{j} \\ \mathbf{k} \end{bmatrix}.$$

Here, the rotation matrix $R \stackrel{\text{def}}{=} B^{-1}A = B^{\mathsf{T}}A$ is uniquely defined and computable.

A condition for equations 13.1 to work is, of course, that the magnetic-field and gravity vectors are not parallel: $\mathbf{m} \nparallel \mathbf{g}$, which is fulfilled everywhere except on the magnetic poles. This is logical, as at those two locations, a magnetic compass is unusable as well!

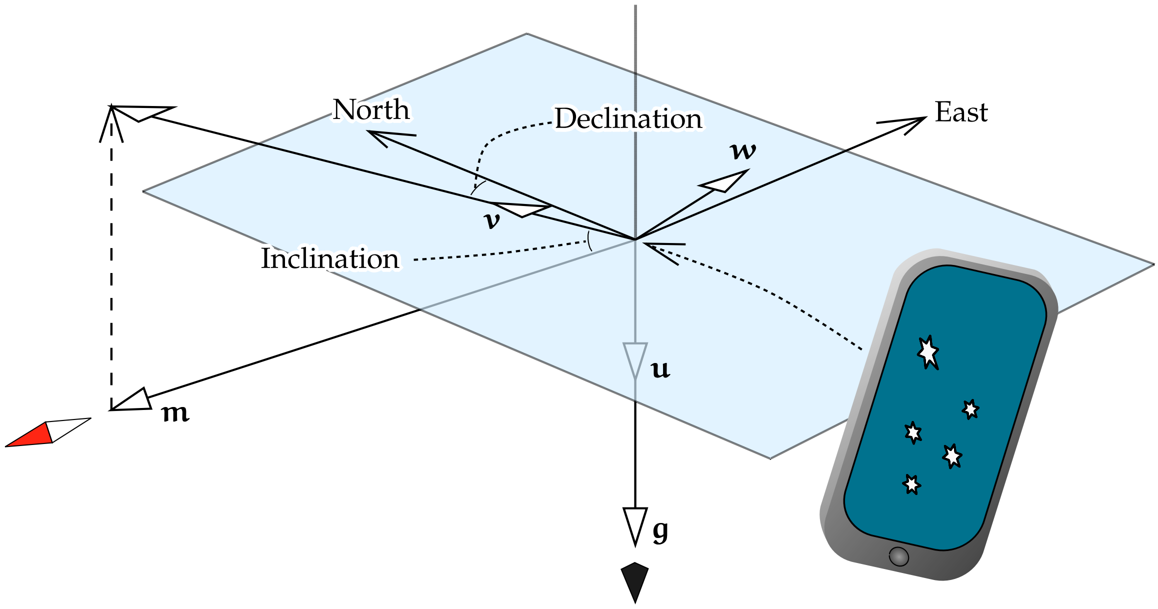


FIGURE 13.1. Orientation of a mobile phone using an accelerometer and a magnetometer.

More generally, determining two independent vectors in two different co-ordinate frames allows the determination of the rotation matrix between the two frames. We shall see that again in section 13.4 for the two vectors determined by three GNSS antennas not on a straight line.

The calibration of three-axis vector sensors is a subject in itself. We shall discuss two cases: calibrating the scales, directions and orthogonality of the sensor axes, and calibrating out a constant offset in the vector being measured.

13.1.1 Axes scale and direction calibration

The scales and directions of the axes of a three-axis accelerometer can be calibrated in the laboratory. The known strength and direction of local gravity $\mathbf{g}$ is used. The phone is mounted on a platform and rotated, with each of the three body axes in turn aligned with $\mathbf{g}$. The calibration equation is

$$\begin{bmatrix} g_x \\ g_y \\ g_z \end{bmatrix}_i = G\,\mathbf{g}_i, \quad i = 1, 2, 3,$$

with g_x, g_y, and g_z the raw readings from the sensor axes and $\mathbf{g}_i$ the gravity vector in the telephone body frame in each of the three

TABLEAU 13.1. Sky Map code which carries out the Gram-Schmidt method described in the text for determining the attitude of a mobile phone. From the GitHub code repository, Sky Map Devs, Stardroid.

```
/**
 * Calculates local North and Up vectors in terms of the phone's coordinate
 * frame from the magnetic field and accelerometer sensors.
 */
private void calculateLocalNorthAndUpInPhoneCoordsFromSensors() {
  Vector3 magneticNorthPhone;
  Vector3 magneticEastPhone;
  if (useRotationVector) {
    float[] rotationMatrix = new float[9];
    SensorManager.getRotationMatrixFromVector(rotationMatrix, rotationVector);
    // The up and north vectors are the 2nd and 3rd rows of this matrix.
    magneticNorthPhone = new Vector3(rotationMatrix[3], rotationMatrix[4],
          rotationMatrix[5]);
    upPhone = new Vector3(rotationMatrix[6], rotationMatrix[7],
          rotationMatrix[8]);
    magneticEastPhone = new Vector3(rotationMatrix[0], rotationMatrix[1],
          rotationMatrix[2]);
  } else {
    // TODO(johntaylor): we can reduce the number of vector copies done in here.
g   Vector3 down = acceleration.copy();
u   down.normalize();
    // Magnetic field goes *from* North to South, so reverse it.
    Vector3 magneticFieldToNorth = magneticField.copy();
m   magneticFieldToNorth.scale(-1);
    magneticFieldToNorth.normalize();
    // This is the vector to magnetic North *along the ground*.
    magneticNorthPhone = addVectors(magneticFieldToNorth,
    scaleVector(down, -scalarProduct(magneticFieldToNorth, down)));
v   magneticNorthPhone.normalize();
    upPhone = scaleVector(down, -1);
w   magneticEastPhone = vectorProduct(magneticNorthPhone, upPhone);
  }
  // The matrix is orthogonal, so transpose it to find its inverse.
  // Easiest way to do that is to construct it from row vectors instead
  // of column vectors.
  axesPhoneInverseMatrix = new Matrix33(magneticNorthPhone, upPhone,
        magneticEastPhone, false);
}
```

measurement positions.

Three equations are obtained,

$$\begin{bmatrix} g_{x,1} \\ g_{y,1} \\ g_{z,1} \end{bmatrix} = G \begin{bmatrix} g \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} g_{x,2} \\ g_{y,2} \\ g_{z,2} \end{bmatrix} = G \begin{bmatrix} 0 \\ g \\ 0 \end{bmatrix}, \quad \begin{bmatrix} g_{x,3} \\ g_{y,3} \\ g_{z,3} \end{bmatrix} = G \begin{bmatrix} 0 \\ 0 \\ g \end{bmatrix},$$

vertausarvo in which g is the ground-truth or reference value for the acceleration of gravity at the telephone location.

The solution for the *calibration matrix* G is

$$G = \frac{1}{g} \begin{bmatrix} g_{x,1} & g_{x,2} & g_{x,3} \\ g_{y,1} & g_{y,2} & g_{y,3} \\ g_{z,1} & g_{z,2} & g_{z,3} \end{bmatrix}.$$

With the calibration done, the calibrated measurement obtained is

$$\mathbf{g}_{\text{cal}} = G^{-1} \begin{bmatrix} g_x \\ g_y \\ g_z \end{bmatrix}.$$

For upmarket devices, this calibration is done after manufacture and the calibration matrix inscribed in firmware (read-only memory) on the telephone (Zhang et al., 2019).

13.1.2 Constant offset vector calibration

For the magnetic-field strength, not only will the local field often deviate from the global field due to local disturbances, but more importantly, due to the *field contribution from the telephone itself*, which contains not only magnetic materials but electric currents also generating magnetic fields. And these are at very close proximity![2]

The self-generated field rotates along with the phone while the external field does not. This allows separation of the two. As the software cannot know the configuration of the device it has been installed on, this needs to be done at least once, and whenever there is a suspicion that the self-generated field has changed.

Let m be the magnetic-field strength not considering this self-generated part and let the raw sensor measurements along the three axes be m_x, m_y and m_z. Let the self-generated field be described by

mu μM components μ_x, μ_y and μ_z, to be estimated. Then

[2]Hint: check if the cover of the protective case contains a magnet holding it closed!

$$m^2 = (m_x - \mu_x)^2 + (m_y - \mu_y)^2 + (m_z - \mu_z)^2 + \nu,$$

meaning that the sensor readings $\begin{bmatrix} m_x & m_y & m_z \end{bmatrix}^T$ are lying on a sphere of radius m centred on $\begin{bmatrix} \mu_x & \mu_y & \mu_z \end{bmatrix}^T$.

Write this as an observation equation:

$$\overbrace{\underline{m}_x^2 + \underline{m}_y^2 + \underline{m}_z^2}^{\underline{\ell}} + \underline{\nu} =$$
$$= \overbrace{m^2 - \mu_x^2 - \mu_y^2 - \mu_z^2}^{\widehat{M}} + 2m_x\widehat{\mu}_x + 2m_y\widehat{\mu}_y + 2m_z\widehat{\mu}_z.$$

The residual $\underline{\nu}$ for every sensor measurement represents how much the distance squared from the centre of the sphere differs from the radius squared of the sphere, m^2.

For n observations, the vector of observations and the design matrix are

$$\underline{\boldsymbol{\ell}} = \begin{bmatrix} \underline{m}_{x,1}^2 + \underline{m}_{y,1}^2 + \underline{m}_{z,1}^2 \\ \underline{m}_{x,2}^2 + \underline{m}_{y,2}^2 + \underline{m}_{z,2}^2 \\ \vdots \\ \underline{m}_{x,n}^2 + \underline{m}_{y,n}^2 + \underline{m}_{z,n}^2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 2m_{x,1} & 2m_{y,1} & 2m_{z,1} \\ 1 & 2m_{x,2} & 2m_{y,2} & 2m_{z,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2m_{x,n} & 2m_{y,n} & 2m_{z,n} \end{bmatrix}.$$

The vector of unknowns is

$$\widehat{\mathbf{x}} = \begin{bmatrix} \widehat{M} & \widehat{\mu}_x & \widehat{\mu}_y & \widehat{\mu}_z \end{bmatrix}^T.$$

The solution is

$$\widehat{\mathbf{x}} = \left(A^T A\right)^{-1} A^T \underline{\boldsymbol{\ell}},$$

after which

$$\widehat{m^2} = \widehat{M} + \widehat{\mu}_x^2 + \widehat{\mu}_y^2 + \widehat{\mu}_z^2.$$

Observations are needed in as many as possible different three-dimensional attitudes of the telephone as provided by the recommended figure-of-eight motion. The calibrated magnetic-field measurement is now

$$\mathbf{m}_{\text{cal}} = \begin{bmatrix} m_x - \widehat{\mu}_x \\ m_y - \widehat{\mu}_y \\ m_z - \widehat{\mu}_z \end{bmatrix}.$$

13.1.3 Combining accelerometer and magnetometer measurements

vertausellipsoidi

The gravity vector $\mathbf{g}$ is computable from latitude and longitude using, for example, the gravity formula of the GRS80 reference ellipsoid to a precision of better than four decimals.

For the magnetic-field vector $\mathbf{m}$, the situation is not so good. The uncertainty in the local direction of the magnetic field vector is a particular problem. If we accept the direction of the measured and calibrated gravity vector $\mathbf{g}_{\text{cal}}$ as correct, then the only interesting remaining information that the measured and calibrated magnetic-field vector $\mathbf{m}_{\text{cal}}$ can provide is the azimuth angle around the vertical axis.

See equation 13.1: uncertainty in $\mathbf{m}$ will cause only uncertainty in $\boldsymbol{v}$ and $\boldsymbol{w}$. As these are both in the plane perpendicular to $\mathbf{g}$, it follows that the uncertainty is also restricted to that plane. And as $\boldsymbol{v} \perp \boldsymbol{w}$ and $\|\boldsymbol{v}\| = \|\boldsymbol{w}\| = 1$, the error can be summarised as just one rotation angle in this plane.

The angle between the two vectors $\mathbf{g}_{\text{cal}}$ and $\mathbf{m}_{\text{cal}}$, equivalent to the magnetic inclination, is not of interest for orienting the phone. It only indicates the local deviations of the magnetic field from the model field due to, for example, local magnetic materials.

Due to the same effect, the azimuth given by the magnetometer is likely off even after the internal calibration described above. One way to correct for this, adopted by the application SkyView®,[3] is to use the stars. SkyView lets you switch the camera on so you can see on the same screen both the sky as presented by SkyView and the real sky as seen by the camera. A simple left-right slider then allows you to make the two coincide.

vertaussuunta

Instead of a magnetometer, a three-axis gyroscope can also be used to orient the azimuth. As a gyroscope does not use an external reference direction, the setting of the correct azimuth using the real sky will have to be done again each time. It will remain valid for some length of time depending on the quality of the gyroscope.

[3]SkyView is a registered trademark of Terminal Eleven LLC.

13.2 Zero-velocity update

An inertial measurement unit loses both location and attitude precision as time progresses due to the double integration over time of the measurements. One way to keep location precision within bounds is to make regular stops and inform the unit's software that now the velocity is zero. This will start the velocity integration from scratch again, interrupting the error propagation. This technique is called a zero-velocity update or ZUPT.

nollanopeus-päivitys

13.2.1 *Propagation of acceleration uncertainty*

The accelerometers have both random and systematic errors. The *systematic errors* can be eliminated by system calibration, determining scale and alignment errors and possibly their nonlinearities. The gyroscopes also have both random and systematic errors, for example attitude drift.

käynti

An alternative approach to eliminating systematic effects — like sensor scale and alignment errors and gyroscope drifts — is *modelling* them by including their model equations as part of the Kalman filter.

Random errors in acceleration will produce through integration random but correlated errors in velocity that grow with the square root of time:

$$\sigma_v \sim \sigma_a \sqrt{t} \tag{13.3}$$

in which σ_a is the random uncertainty of the acceleration, assumed to be white noise. Upon a second integration, the random uncertainty of location will grow with the power $\frac{3}{2}$ of time:

sigma $\sigma\Sigma$

$$\sigma_x \sim \sigma_v t \sim \sigma_a t\sqrt{t}. \tag{13.4}$$

13.2.2 *Propagation of gyroscope uncertainty*

The propagation of the directional uncertainty of a gyroscope is trickier. Assume that the direction error behaves like a random walk:

$$\sigma_\theta \sim \sigma_\omega \sqrt{t},$$

in which σ_ω is the random uncertainty of the angular rate $\omega(t) = \frac{d}{dt}\theta(t)$ at which the direction of the spin axis turns. The angular rate is assumed to be white noise, meaning that the axis direction angle $\theta(t)$ will be a random walk.

omega $\omega\Omega$

theta $\vartheta\theta\Theta$

The impact of gyroscope directional uncertainty on velocity and position uncertainty is not straightforward: it depends on the scenario of the journey. A simple scenario is a uniform acceleration from standstill up to cruise velocity v_c, which is reached after a time t_c.

In the acceleration phase, the impact on the acceleration vector is

$$\sigma_{a,\theta} \sim a\,\sigma_\theta \sim a\,\sigma_\omega \sqrt{t},$$

and on the velocity vector, through integration,

$$\sigma_{v,\theta} \sim \sigma_{a,\theta} t \sim a\,\sigma_\omega t\sqrt{t} \sim v\,\sigma_\omega \sqrt{t}.$$

This impact will be in the sideways direction for both acceleration and velocity vectors. The acceleration will be in the direction of the journey, as will be the cruise velocity during the journey. So we find also for the location

$$\sigma_{x,\theta} \sim \sigma_{v,\theta} t \sim v\,\sigma_\omega t\sqrt{t} \sim a\,\sigma_\omega t^2\sqrt{t}.$$

We see that the dependence on time is of power no less than $\frac{5}{2}$.

However, the acceleration phase will be of limited duration. If the acceleration phase is followed by a cruise phase, the uncertainty in the cruise velocity will propagate linearly in time:

$$\sigma_{v,\theta} \sim v_c\sigma_\omega \sqrt{t_c} \sim a\,\sigma_\omega t_c\sqrt{t_c} \implies \sigma_{x,\theta} \sim v_c\sigma_\omega t\sqrt{t_c} \sim a\,\sigma_\omega t\, t_c\sqrt{t_c}.$$

Note that it is the absence of acceleration that causes the growth of location uncertainty to be linear in time. In a general scenario in which the vehicle undergoes accelerations, for example changes in travel direction, throughout the journey, the uncertainty will grow faster than this.

13.2.3 Carrying out zero-velocity updates

Carrying out zero-velocity updates at intervals of Δt, assuming that the velocity errors in different update intervals are statistically independent (equations 13.3 and 13.4), will yield

$$\sigma_v \sim \sigma_a\sqrt{\Delta t} \implies \sigma_x \sim \left(\sigma_a\,\Delta t\sqrt{\Delta t}\right)\sqrt{n} = \sigma_a\,\Delta t\sqrt{t},$$

with n the number of zero-velocity updates: $t = n\,\Delta t$.

We see that the growth in positional uncertainty, which was earlier of the power of $\frac{3}{2}$ in elapsed time, has now been brought down to a dependence on the power of $\frac{1}{2}$ of time!

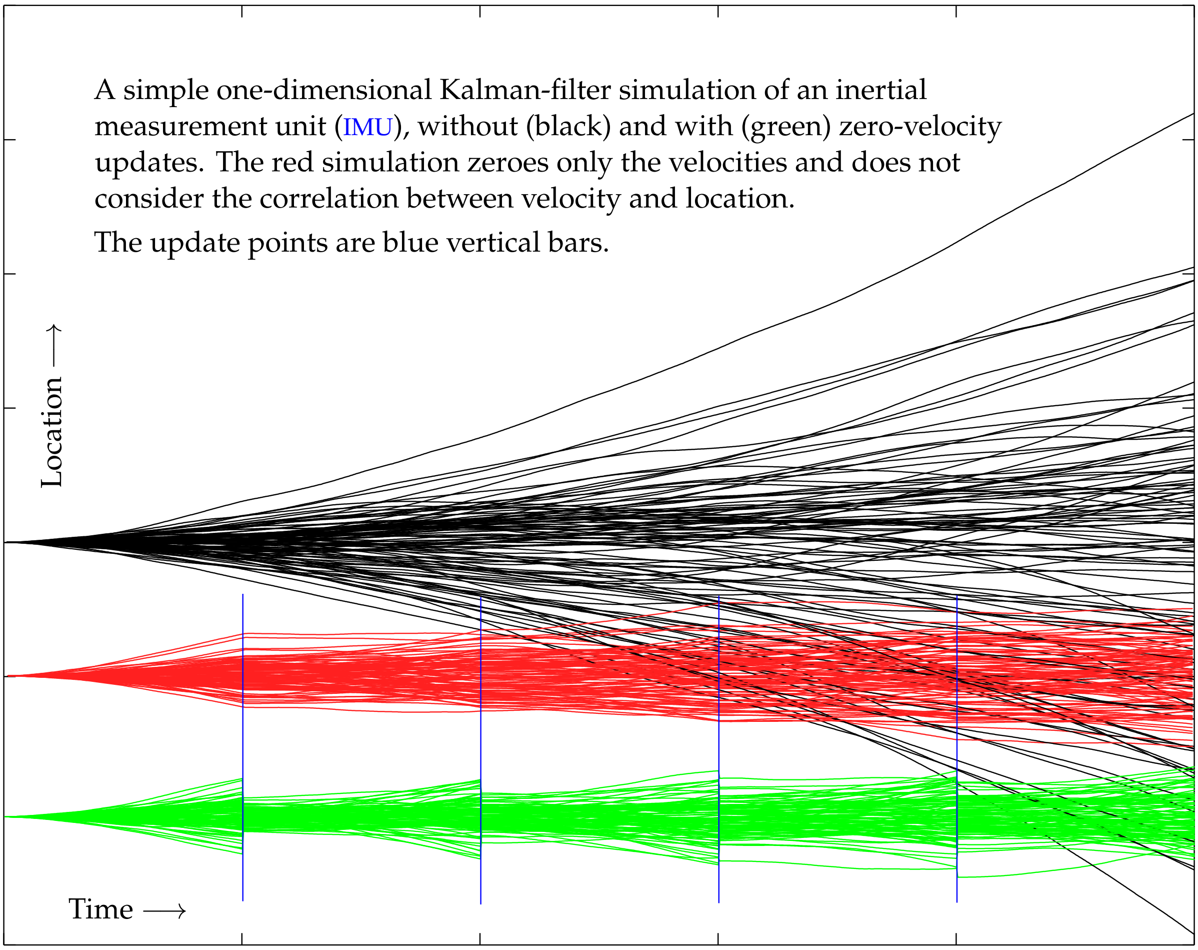


FIGURE 13.2. A simple Kalman filter without and with zero-velocity updates.

Figure 13.2 shows a simulation of a simple one-dimensional IMU model displaying this basic behaviour. The model is[4] 4

$$\overbrace{\begin{bmatrix} \underline{x}_{k+1} \\ \underline{v}_{k+1} \end{bmatrix}}^{\mathbf{x}_{k+1}} = \overbrace{\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}}^{\Phi_k^{k+1}} \overbrace{\begin{bmatrix} \underline{x}_k \\ \underline{v}_k \end{bmatrix}}^{\mathbf{x}_k} + \overbrace{\begin{bmatrix} 0 \\ \underline{n}_a \Delta t \end{bmatrix}}^{\underline{\mathbf{w}}_k^{k+1}},$$

with $\underline{n}_a$ being random uncorrelated error (white noise) representing the accelerometer measurement uncertainty. The quantity plotted is $\underline{x}_k = \underline{x}(t_k)$, the one-dimensional location. $\Delta t = t_{k+1} - t_k$ is the time step.

[4]The differential or continuous-time model corresponding to this is

$$\frac{d}{dt} \overbrace{\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}}^{\mathbf{x}} = \overbrace{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}}^{F} \overbrace{\begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}}^{\mathbf{x}} + \overbrace{\begin{bmatrix} 0 \\ \underline{n}_a \end{bmatrix}}^{\mathbf{n}}.$$

A zero-velocity update will reset the uncertainty of the velocity of the vehicle to zero. It does nothing to improve the uncertainties of the other state-vector elements, like the location or the directional uncertainty of the gyroscope — at least not directly. Consider, however, that all elements of the state vector are connected and their estimators intercorrelated in the Kalman filter. A complete zero-velocity update is really a Kalman update step.

If the gyroscope axis has drifted, this will show up as a misclosure between the velocity vector estimate $\boldsymbol{v}^{-}(t)$ maintained by the filter and the zero "observation" of the same vector provided by the zero-velocity update. So at the update step, the velocity is zeroed and, in addition, not only the location estimate $\mathbf{x}^{-}(t)$, but also the gyroscope spin axis direction estimate $\theta^{-}(t)$ are updated and their uncertainties reduced. In figure 13.2 the effect of this on location can be seen by comparing the green curves with the red ones.

13.3 Integration of GNSS and IMU

This is an example of *heterogeneous* sensor fusion.

An inertial measurement unit (IMU) as presented in section 5.2 contains three accelerometers and three gyroscopes. Using these, it can track its own linear movements and rotations. A positioning solution is obtained by integrating the measured acceleration twice in succession, first into velocity and then into location. Over time, the solution will progressively deteriorate.

The deterioration may be controlled if one does a "real" positioning at regular intervals, for example using GNSS. In this way, one may build a system that preserves its positioning precision even though the GNSS signal is intermittent, like in tunnels, under bridges, near high-rise buildings or in indoor space.

At the same time the precise attitude of the equipment is also available. GNSS alone cannot do that. On the other hand GNSS can, in the presence of accelerations, observe and eliminate the long-term attitude drift of an inertial device. A great example of this is the use of integrated IMU and GNSS in aerial photography or airborne laser scanning: every 180-degree turn between flight lines provides an acceleration vector that can be "seen" by both the inertial measurement unit and the GNSS

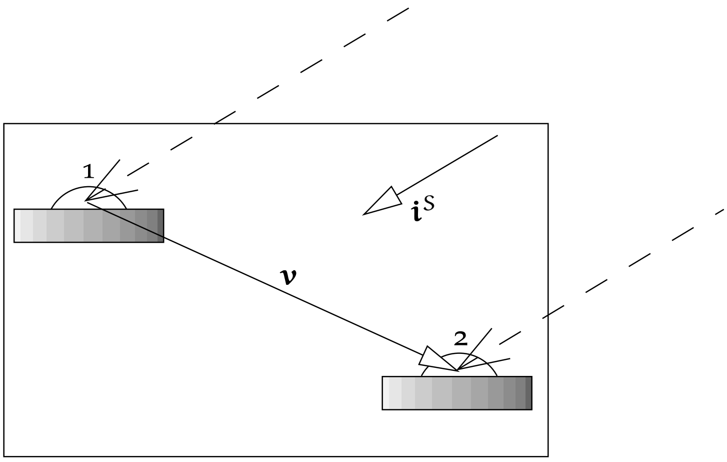


FIGURE 13.3. Attitude determination by GNSS.

receiver, connecting the instrumental axis directions of both. This does for attitude what zero-velocity updates do for location, section 13.2.

An example of such integrated equipment is the NovAtel SPAN®,[5] Synchronized Position Attitude Navigation, (Hexagon, SPAN GNSS Inertial Navigation Systems). 5

13.4 Attitude determination with GNSS

The use of multi-antenna GNSS receivers for attitude determination is an example of *homogeneous* sensor fusion.

A GNSS receiver may determine the attitude of a vehicle using several — at least three — antennas. The method used is real-time kinematic positioning over very short vectors. With two antennas mounted horizontally, vehicle heading can be determined. Three-dimensional attitude determination requires a third antenna not on the same line. tosiaikainen

As can be seen from figure 13.3, the same satellite is being observed from two different antennas. The observable is the difference between two measurements of the carrier phase. Based on equation 8.3, the raw measurement in metres is

$$P \stackrel{\text{def}}{=} \lambda \frac{\phi}{2\pi},$$

and the difference measurement between antennas 1 and 2 to satellite S is

$$\Delta P^S = P_2^S - P_1^S = \langle \boldsymbol{v} \cdot \mathbf{i}^S \rangle - N^S \lambda,$$

---

[5]SPAN is a registered trademark of NovAtel Inc.

in which $\boldsymbol{v}$ is the inter-antenna vector and $\mathbf{i}^S$ is the direction vector from the satellite, a unit vector, $\|\mathbf{i}^S\| = 1$. The number N^S is an integer describing the ambiguity, the integer number representing multiple alternative values of the observable.

kokonaisluku-tuntematon

lambda λΛ

We re-write this by reducing the observable to the interval $[0, \lambda)$:

$$\Delta P^S \bmod \lambda = \left\langle \boldsymbol{v} \cdot \mathbf{i}^S \right\rangle - \overline{N}^S \lambda.$$

Here we have to solve simultaneously for the vector $\boldsymbol{v}$ and the integer unknown (ambiguity) $\overline{N}^S$. Solving for the vector requires observations from at least three different satellites, and then the values $\overline{N}$ still remain undetermined:

$$\begin{aligned}
\Delta P^1 \bmod \lambda &= \left\langle \boldsymbol{v} \cdot \mathbf{i}^1 \right\rangle - \overline{N}^1 \lambda, \\
\Delta P^2 \bmod \lambda &= \left\langle \boldsymbol{v} \cdot \mathbf{i}^2 \right\rangle - \overline{N}^2 \lambda, \\
\Delta P^3 \bmod \lambda &= \left\langle \boldsymbol{v} \cdot \mathbf{i}^3 \right\rangle - \overline{N}^3 \lambda.
\end{aligned}$$

We know at least that the values $\overline{N}^1, \overline{N}^2, \overline{N}^3$ cannot be very large if the vector $\boldsymbol{v}$ is short: as $\Delta P^S \bmod \lambda$ lies in the interval $[0, \lambda)$ and $\left\langle \boldsymbol{v} \cdot \mathbf{i}^S \right\rangle$ in the interval $[-\|\boldsymbol{v}\|, \|\boldsymbol{v}\|]$, then $\overline{N}^S$ can only lie in the interval $\left(-\|\boldsymbol{v}\|/\lambda - 1, \|\boldsymbol{v}\|/\lambda\right]$. If the vector is, for example, 2 m long and the wavelength is 24 cm, then the only possible values for $\overline{N}^S$ are $-9, -8, \ldots, +7, +8$.

The solution is obtained as follows:

1. If we can see more than three satellites, we choose the three of them that together produce the best possible geometry. This is easy: traverse all triplets and compute their determinants $\left\langle \mathbf{i}^1 \cdot \left\langle \mathbf{i}^2 \times \mathbf{i}^3 \right\rangle \right\rangle$. The largest absolute value wins. If we can see for example ten satellites, we have to compute

$$\binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$$

   determinants.

2. Try out all possible values $\overline{N}^S$ for the three satellites 1, 2 and 3, and compute for every combination a vector solution $\boldsymbol{v}$. The total number of solutions to be computed is in the example case $18^3 = 5832$.

3. If the vector is, for example, 20 m long, the total number of solutions to be computed is already $168^3 = 4.7$ million. This

would require significant processing capacity. On the other hand, if we have the use of a dual-frequency instrument, *widelaning* will be available with an effective wavelength of 86 cm. Then we need only $48^3 = 110\,592$ different solutions. leveäkuja

4. Next, compute for each provisional solution $\widehat{\mathbf{v}}\left(\overline{N}^1, \overline{N}^2, \overline{N}^3\right)$ thus found, whether $\|\widehat{\mathbf{v}}\|$ is close enough to the known distance between the antennas. Solutions failing this condition can be rejected immediately.
5. After this we compute the observables of the other satellites:

$$\begin{aligned}
\Delta P^4 \bmod \lambda &= \left\langle \widehat{\mathbf{v}} \cdot \mathbf{i}^4 \right\rangle - \overline{N}^4 \lambda, \\
\Delta P^5 \bmod \lambda &= \left\langle \widehat{\mathbf{v}} \cdot \mathbf{i}^5 \right\rangle - \overline{N}^5 \lambda, \\
&\cdots \\
\Delta P^n \bmod \lambda &= \left\langle \widehat{\mathbf{v}} \cdot \mathbf{i}^n \right\rangle - \overline{N}^n \lambda.
\end{aligned}$$

   Each of the observation values on the left-hand side should agree within observation uncertainty with the corresponding right-hand side for one value $\overline{N}$. Generally this works for *all* satellites $S = 4, \ldots, n$ only for one provisional solution.
6. Using the set of values thus found, $\overline{N}^S$, $S = 1, \ldots, n$, the *final adjustment* is carried out in order to compute an optimal vector $\widehat{\mathbf{v}}$ from *all* observations.
7. When the vehicle with the instrument moves, as long as no cycle slips occur, the values $\overline{N}^S$ stay the same. Then one can continuously solve for $\widehat{\mathbf{v}}$ in real time from the observations collected. vaihekatko tosiaikaisesti

13.5 Modern radionavigation

The proliferation of base-station networks, like mobile telephony and wireless local-area networks (WLAN), have made radionavigation in two dimensions on the Earth's surface practical again. Techniques used are

Cell identity Here, the location is constrained by which base stations are within the range of the navigator. It is based on the fact that each base station, and sometimes each directional sector around the base station, transmits identifying information. The method is imprecise outside cities, more precise within them. solutunnus tukiasema

Signal strength Two or more base stations are needed, precision is typically low.

Range difference This resembles the Decca system of yore and similar ones used in maritime radionavigation. This is also called TDOA, time difference of arrival, pseudorange multilateration, or hyperbolic navigation. A position fix using this method requires the use of three base stations.

Range This is called a time-of-arrival (TOA) method. It requires back-and-forth signalling between the navigator and each base station, of which there must be at least two. As mobile telephony uses TDMA, time-division multiple access, a built-in method for mobile networks is "timing advance", needed to synchronise the time slots of mobile clients at varying distances from the base station.

aikajako-kanavointi

ajastusennakko

aikaikkuna

Furthermore, mobile-phone positioning by GNSS may be assisted by the base-station network by providing the satellite ephemeris faster than the satellites themselves are broadcasting them: the navigation message broadcast by the satellites has for GPS a bit rate of only 50 bits per second. Using the mobile network for this is called assisted GNSS or A-GNSS.

13.6 Microelectronic motion sensors (MEMS)

Microelectronic motion sensors (MEMS) are small and inexpensive accelerometers and gyroscopes. The manufacturing process consists of the same photolithography technique used in manufacturing computer integrated circuits together with micromachining.

Readout of MEMS sensors is often done capacitively. In both accelerometers and gyroscopes, small solid structures deform elastically. Two conductors separated by an air — or vacuum — gap form a capacitor, the capacitance of which will vary with the width of the gap. By connecting them in circuit together with an inductance (solenoid), an oscillator is formed, the frequency of which can be measured precisely. See Wikipedia, LC circuit.

käämi

6

A MEMS magnetometer is usually based on the Hall[6] effect: a conductor in which an electric current runs, placed in a magnetic field, will produce a sideways potential difference. If the direction of the magnetic field is x and that of the current y, the direction of the potential difference will be z. See Wikipedia, Hall effect.

[6]Edwin Herbert Hall (1855–1938) was an American physicist.

13.6.1 Accelerometers

These circuits measure accelerations by measuring, for example capacitively, the movement of a small test mass under the influence of the pseudo-force caused by acceleration. One model, the ADXL103 from Analog Devices, is capable of measuring accelerations under 1.7 g ($17\,{}^{\mathrm{m}}/_{\mathrm{s}^2}$) with a resolution of 1 mg or 1000 mGal.

To get an idea of what this means, consider that an acceleration of 1000 milligals during a minute produces a displacement of 18 metres. When used as a tilt meter, this corresponds to a tilt of 3.4 minutes of arc.

The *sensitivity* of the device can even be better than that after calibration.

The device also survives, for instance, dropping onto a concrete floor — momentary acceleration 3500 g! The size of the chip is $5 \times 5 \times 2$ mm (Analog Devices). Similar devices have also been fired from cannon and are used in "smart ammunition". älyammus

Prices are nowadays (2020) around a euro a piece even for three-axis sensors. They are a few millimetres in size. Applications include triggering sensors for automotive airbags and drop-protection triggering sensors for laptop hard drives. These mass markets have driven down prices.

13.6.2 Rotation sensors

Microelectronic gyroscopes or rotation sensors are commonly based on the oscillation of a tuning fork like structure, see Wikipedia, Vibrating structure gyroscope. Like a rotating object, a vibrating object also tries to stay within the same plane. Therefore, rotation of the whole device causes torque, the *Coriolis force*, which is measured, for example, capacitively. The measurement value is proportional to the rotation rate or angular velocity. See figure 13.4. A three-axis or three-degrees-of-freedom device contains three one-axis components. äänirauta vääntö

These devices are capable of precisions of $\pm 0.1\,{}^{\circ}/_{\mathrm{s}}$, which is four to six orders of magnitude poorer than a spinning-rotor or laser gyroscope, see subsections 5.2.1 and 5.3.1.

Fields of application include stabilising the image of video cameras, improving the steering response of cars, robotics, drones (UAVs, unmanned aerial vehicles), and many others.

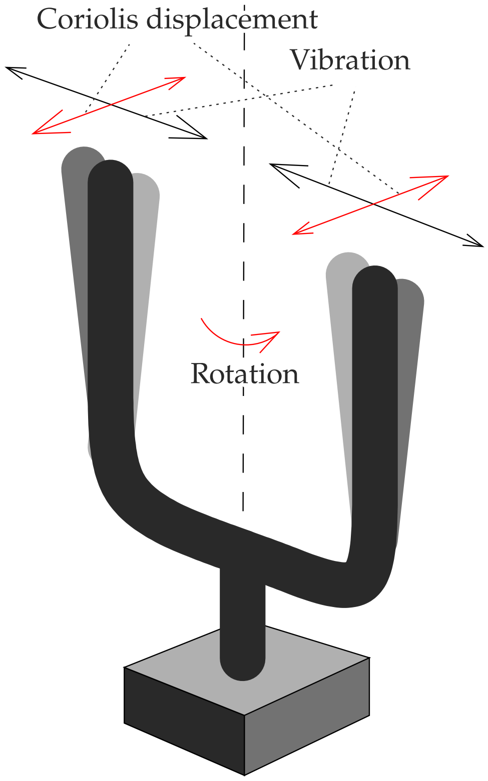


FIGURE 13.4. Principle of a MEMS rotation sensor. When the platform rotates, the vibration of the tuning fork (in the picture left-right) will cause a periodic displacement due to the Coriolis force, which is measured capacitively.

13.7 Pedestrian navigation

When a pedestrian uses a mobile phone for navigation, the application inside the phone assumes that the top edge of the screen is pointing in the direction of walking. One quickly gets used to this intuitive requirement: by holding the phone in this way, the map displayed on the screen of the phone will align with the surrounding landscape.

However, this requires the user to move by at least a few metres: the motion vector in the terrain as observed by GNSS will then be placed in
7 the map and used to turn the map image correctly.[7]

When applying inertial navigation technology in personal navigation, the principle of the *zero-velocity update* has been used in personal navigation solutions with an IMU to be mounted in boots. Every time

[7]The same applies with car navigation: the car has to move before the map display orients itself correctly. This may well lead to the first instruction to the driver being a 180-degree turn. This is assuming the application does not use the built-in compass or magnetometer for orientation if there is one.

the boot hits the ground, the velocity in the Kalman navigation filter is reset to zero and the other state-vector elements are updated.

An interesting application of this is guiding firefighters, specifically smoke divers, in a smoke-filled building. The self-contained nature and independence from external signals of inertial navigation is a major feature for this. For example, Godha et al. (2006).

13.8 Indoor navigation

This is a broad and rapidly developing field (Wikipedia, Indoor positioning system). We only scratch the surface here.

Indoors, the availability of GNSS may not be assumed and is usually lacking. However, other base-station types are either available (WLAN, wireless local-area network) or may be installed. The various positioning approaches available are similar to those using base stations for mobile telephony as presented in section 13.5.

Pseudolites are devices that transmit radio frequencies and signals like GNSS satellites do, but are installed on Earth. They transmit a carrier modulated with a pseudo-random code (PRN) similar to that of a GNSS satellite. For example, Zhao et al. (2018).

Acoustic positioning has been explored in conference settings, where speakers can be located and tracked, for example to steer cameras and directional microphones. This becomes especially valuable if conferences are streamed over the Internet, or there are remote participants using Internet meeting platforms like Adobe®[8] Connect™[9] 8
or Microsoft Teams®.[10] For a description and further references, see 9
Parviainen (2016). 10

Magnetic-field or WLAN or Bluetooth®[11] signal strength can be used 11
for positioning. The magnetic field inside a building varies irregularly due to the magnetic properties of building materials used. The same is true for the power of WLAN signals. These could be used for positioning, provided that they are first mapped, or "fingerprinted", throughout the building. See for example Chen et al. (2013); Liu et al. (2017); Mazlan

[8] Adobe is a registered trademark of Adobe Systems Incorporated.

[9] Connect is a common-law trademark of Adobe Systems Incorporated.

[10] Microsoft Teams is a registered trademark of Microsoft Corporation.

[11] Bluetooth is a registered trademark of The Bluetooth Special Interest Group.

et al. (2017). A downside is that the mapping needs to be repeated, especially if modifications are made to the building's interior.

Inertial sensors are useful in indoor navigation either on their own or in a sensor-fusion context. Their major advantage is independence from external signals. As indoor navigation is most often also pedestrian navigation, the zero-velocity update technique can be usefully applied.

Self-test questions

1. How does a MEMS rotation sensor work?
2. How does a MEMS magnetometer work? Why, if it is being used in a mobile phone, does it need to be calibrated? How is calibration done?
3. How, in the application Sky Map, does a mobile phone determine its attitude with respect to the stars?
4. Why would a (theoretical) gyrocompass in a mobile phone not work well when the phone is hand-held?
5. How does attitude determination with GNSS work?
6. What are the four different principles of operation of modern radionavigation using mobile base stations?
7. Why is integration of inertial navigation and GNSS navigation attractive?
8. Why are zero-velocity updates useful?
9. In the time-of-arrival (TOA) method, at least two base stations are needed. Is *unambiguous* positioning possible with only two base stations?
10. What indoor navigation techniques exist?
11. How can an *odometer* be integrated into a land vehicle navigation system? What benefits does it offer?
12. How can the *daylight recorder* described in Fox (2018) be used to determine latitude and longitude over time? When will this break down?

Power spectral density is non-negative

A

Showing this claim to be true is based on the observation that the autocovariance function can be seen as representing a variance matrix, which is always symmetric and positive definite.

Let $A_x(t, t') = A_x(\Delta t)$, $\Delta t = t' - t$, be the autocovariance function of a periodic signal $\underline{x}(t)$ of period T. Sample the signal at an interval δt, with $T = N\,\delta t$ and N the number of sample points. Now, the function $A_x(t_j, t_k)$, $j, k = 0, \ldots, N-1$ will fill a variance matrix of all these points together:

otospiste

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} A_x(t_0,t_0) & A_x(t_0,t_1) & A_x(t_0,t_2) & & \\ A_x(t_1,t_0) & A_x(t_1,t_1) & A_x(t_1,t_2) & & \\ A_x(t_2,t_0) & A_x(t_2,t_1) & A_x(t_2,t_2) & & \\ & & & \ddots & \\ & & & & A_x(t_{N-1},t_{N-1}) \end{bmatrix} = \\
&= \begin{bmatrix} A_x(0) & A_x(\delta t) & A_x(2\,\delta t) & & & \\ A_x(-\delta t) & A_x(0) & A_x(\delta t) & & & \\ A_x(-2\,\delta t) & A_x(-\delta t) & A_x(0) & & & \\ & & & \ddots & & \\ & & & & A_x(0) & A_x(\delta t) \\ & & & & A_x(-\delta t) & A_x(0) \end{bmatrix} = \\
&= \Big[A_x\big((k-j)\,\delta t\big)\Big]_{jk}, \qquad j,k = 0,\ldots,N-1,
\end{aligned}
$$

where stationarity is assumed. It is seen that in this case, the matrix Σ is Toeplitz circulant,[1] meaning that all rows, and all columns, are circularly shifted copies of each other. Also,

sigma $\sigma\Sigma$
1

[1] Otto Toeplitz (1881–1940) was a Jewish German mathematician and a historian and populariser of mathematics.

$$\begin{aligned} A_x(\Delta t) &= A_x(\Delta t + nT), \\ A_x\big((k-j)\,\delta t\big) &= A_x\big((k-j+nN)\,\delta t\big), \end{aligned} \qquad n \in \mathbb{Z},$$

meaning that the autocovariance function $A_x(\Delta t)$ is circular with period $T = N\,\delta t$.

Next, we build a vector $\mathbf{c}$ looking like this:

$$\mathbf{c}(f) \stackrel{\text{def}}{=} \begin{bmatrix} 1 & e^{2\pi i f \cdot \delta t} & e^{2\pi i f \cdot 2\,\delta t} & e^{2\pi i f \cdot 3\,\delta t} & \dots & e^{2\pi i f \cdot (N-1)\,\delta t} \end{bmatrix},$$

with f an integer multiple of $^1/_T$, so that $fT \in \mathbb{Z}$.

Calculate the quadratic form

$$\frac{\delta t}{N}\,\mathbf{c}(f)\,\Sigma\,\mathbf{c}^\dagger(f) = \frac{\delta t}{N} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} e^{2\pi i f \cdot j\,\delta t} A_x\big((k-j)\,\delta t\big)\, e^{-2\pi i f \cdot k\,\delta t} =$$
$$= \frac{\delta t}{N} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} e^{-2\pi i f\,(k-j)\,\delta t} A_x\big((k-j)\,\delta t\big).$$

Remember that $fN\,\delta t = fT \in \mathbb{Z}$, so also the exponent expression $e^{-2\pi i f\,(k-j)\,\delta t}$ is circular with period $T = N\,\delta t$: $e^{-2\pi i f N\,\delta t} = e^{-2\pi i f T} = 1$. Now we substitute $m = k - j$ and rearrange the sums:

$$\frac{\delta t}{N}\,\mathbf{c}(f)\,\Sigma\,\mathbf{c}^\dagger(f) = \frac{\delta t}{N} \sum_{j=0}^{N-1} \sum_{m=0}^{N-1} e^{-2\pi i f \cdot m\,\delta t} A_x\big(m\,\delta t\big) =$$
$$= \delta t \cdot \sum_{m=0}^{N-1} e^{-2\pi i f \cdot m\,\delta t} A_x(m\,\delta t) = \sum_{m=-\frac{1}{2}N}^{\frac{1}{2}N-1} A_x(\Delta t)\, e^{-2\pi i f\,\Delta t} \cdot \delta t, \qquad \text{(A.1)}$$

with $\Delta t = m\,\delta t$. N is assumed even.

The periodicity of both the exponent and A_x functions means that $m = k - j$ may be understood as the modulo expression $m \bmod N = (k - j) \bmod N$.

Because the quadratic form on the left-hand side of equation A.1 is Hermitian and contains a positive-definite Hermitian matrix Σ (meaning $\Sigma = \Sigma^\dagger$), it follows that it must be non-negative for any argument f.

The right-hand side of equation A.1 contains the discrete, circular Fourier transform of $A_x(\Delta t) = A_x(t)$. In the limit $T \to \infty$, $N \to \infty$, $\delta t \to 0$ the right-hand side approximates the integral 2.34,

$$\mathcal{A}_x(f) = \int_{-\infty}^{+\infty} A_x(t)\, e^{-2\pi i f t}\, dt,$$

for which thus the same must hold: it is non-negative for all values f, for which now all real values are allowed.

M-sequences and Gold codes

B

A maximum-length bit sequence or *m-sequence* may be defined by its generating mechanism, given in figure 7.9:

$$\begin{aligned} b_1(i) &= b_3(i-1) \oplus b_4(i-1), \\ b_2(i) &= b_1(i-1), \\ b_3(i) &= b_2(i-1), \\ b_4(i) &= b_3(i-1). \end{aligned}$$

These bits in the shift register can be identified with bits in the output sequence as follows:

$$\begin{aligned} x(i) &= b_4(i), \\ x(i+1) &= b_3(i), \\ x(i+2) &= b_2(i), \\ x(i+3) &= b_1(i). \end{aligned}$$

Substitution yields

$$\begin{aligned} x(i+3) = b_1(i) = b_3(i-1) \oplus b_4(i-1) = x(i) \oplus x(i-1) \\ \implies x(i+4) = x(i+1) \oplus x(i), \end{aligned} \tag{B.1}$$

which directly generates the sequence x.

This is for the register geometry in figure 7.9. In the generic case, we write

$$x(i+n) = \bigoplus_{\substack{k=0 \\ a(k)=1}}^{n-1} x(i+k), \tag{B.2}$$

in which $a(k) \in \{0, 1\}$, $k = 0, \ldots, n-1$ are bit values indicating which feedback connections or "taps" are present in the register.

Here, the XOR operator $\oplus$ may be usefully understood as *modulo-2 addition*: $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$. It has all the nice properties of addition, such as commutativity $x \oplus y = y \oplus x$ and associativity $(x \oplus y) \oplus z = x \oplus (y \oplus z)$.

Equation B.1 is *linear*. This means that if the equation holds for sequence x and sequence y, it will also hold for their sum $z = x \oplus y$:

$$\overbrace{x(i+4) \oplus y(i+4)}^{z(i+4)} = \big(x(i+1) \oplus x(i)\big) \oplus \big(y(i+1) \oplus y(i)\big) = \\ = \overbrace{\big(x(i+1) \oplus y(i+1)\big)}^{z(i+1)} \oplus \overbrace{\big(x(i) \oplus y(i)\big)}^{z(i)},$$

thanks to the associativity and commutativity of the $\oplus$ operator. To appreciate this on an intuitive level, just replace the symbol $\oplus$ — "cyclical plus" — by an ordinary plus symbol $+$.

This readily generalises to equation B.2, as it is clear that

$$\overbrace{x(i+n) \oplus y(i+n)}^{z(i+n)} = \left(\bigoplus_{\substack{k=0 \\ a(k)=1}}^{n-1} x(i+k) \right) \oplus \left(\bigoplus_{\substack{k=0 \\ a(k)=1}}^{n-1} y(i+k) \right) = \\ = \bigoplus_{\substack{k=0 \\ a(k)=1}}^{n-1} \overbrace{\big(x(i+k) \oplus y(i+k)\big)}^{z(i+k)}.$$

alpha αA This is a useful result: if $x^\alpha \neq x$ is an arbitrarily cyclically shifted version of x, and thus a legitimate m-sequence for this shift-register geometry,
beta βB then $x^\beta \stackrel{\text{def}}{=} x \oplus x^\alpha$ will itself be an m-sequence for the same geometry, yet another cyclically shifted version of x. All m-sequences for this geometry are cyclically shifted versions of each other.

A well-known property of an m-sequence of length $2^n - 1$ is that it contains 2^{n-1} ones and $2^{n-1} - 1$ zeros. This is because its every bit is the rightmost bit of the shift-register content, which cycles through all n-bit binary numbers except the number zero. If zero were included, the number of bits one and bits zero would be equal.

Thanks to this property, $x^\beta = x \oplus x^\alpha$ has one more bit value 1 than it has bit values 0. The cyclical autocorrelation follows with equation 7.2:

$$\mathrm{Corr}\{\overline{x}, \overline{x}^\alpha\} = \frac{1}{2^n - 1} \big(C_0(x \oplus x^\alpha) - C_1(x \oplus x^\alpha)\big) = \\ = \frac{1}{2^n - 1} \big(C_0(x^\beta) - C_1(x^\beta)\big) = -\frac{1}{2^n - 1},$$

for all cyclical shifts α except zero. Zero of course yields identically $x \oplus x = 0$, so

$$\mathrm{Corr}\{\overline{x}, \overline{x}\} = \frac{1}{2^n - 1} C_0(x \oplus x) = \frac{2^n - 1}{2^n - 1} = 1.$$

See the figure in table 7.3.

Robert Gold showed in 1967, that if one uses two different, equally long shift registers x and y with *different* geometries, then all combinations, or Gold codes,

$$g_0 \stackrel{\text{def}}{=} x \oplus y, \quad g_1 \stackrel{\text{def}}{=} x \oplus y^{\alpha}, \quad g_2 \stackrel{\text{def}}{=} x \oplus y^{\alpha'}, \quad g_3 \stackrel{\text{def}}{=} x \oplus y^{\alpha''}, \quad \ldots$$

with different *relative* cyclical shifts $0, \alpha, \alpha', \alpha'', \ldots$ between x and y are useful for separating signals from each other when their absolute time offsets are unknown.

Let us first look at the autocorrelation. Note that g_1 is *not* a cyclically shifted g_0. The sequence g_0 cyclically shifted by α equals[1] 1

$$g_0^{\alpha} \stackrel{\text{def}}{=} (x \oplus y)^{\alpha} = x^{\alpha} \oplus y^{\alpha}.$$

The cyclical autocorrelation function of g_0 is[2] $(\alpha \neq 0)$: 2

$$\begin{aligned} g_0 \oplus g_0^{\alpha} &= (x \oplus y) \oplus (x^{\alpha} \oplus y^{\alpha}) = (x \oplus x^{\alpha}) \oplus (y \oplus y^{\alpha}) = x^{\beta} \oplus y^{\gamma} \\ \Longrightarrow \mathrm{Corr}\{\overline{g}_0, \overline{g}_0^{\alpha}\} &= \frac{1}{2^n - 1}\Big(C_0\big(g_0 \oplus g_0^{\alpha}\big) - C_1\big(g_0 \oplus g_0^{\alpha}\big)\Big) = \\ &= \frac{1}{2^n - 1}\Big(C_0\big(x^{\beta} \oplus y^{\gamma}\big) - C_1\big(x^{\beta} \oplus y^{\gamma}\big)\Big). \end{aligned}$$

This makes it clear that x and y must be chosen such that all sequences $x^{\beta} \oplus y^{\gamma}$ — in practice, the sequences obtained when x and y are cyclically shifted with respect to *each other* in all $N = 2^n - 1$ possible ways — are *balanced*, meaning they contain approximately as many zeros as ones. It holds that gamma $\gamma\Gamma$

$$\big(x^{\beta} \oplus y^{\gamma}\big)^{N-\beta} = x^{\beta+(N-\beta)} \oplus y^{\gamma+(N-\beta)} = x^N \oplus y^{\delta} = x \oplus y^{\delta},$$

so

$$C_0\big(x^{\beta} \oplus y^{\gamma}\big) = C_0\big(x \oplus y^{\delta}\big), \qquad C_1\big(x^{\beta} \oplus y^{\gamma}\big) = C_1\big(x \oplus y^{\delta}\big),$$

and these two should be close to each other for all $\delta = 0, 1, \ldots, N-1$. delta $\delta\Delta$

[1] This means physically that both registers x and y are simultaneously triggered α times, and the output combined through an XOR gate. See figure B.1.

[2] You may not assume $\beta = \gamma$!

Fortunately such choices of the shift registers x and y can be found. Then, the autocorrelation function of g_0 — and all g_i — will have a unique peak at the origin: $\mathrm{Corr}\{\overline{g}_0, \overline{g}_0\} = \mathrm{Corr}\{\overline{g}_0^\alpha, \overline{g}_0^\alpha\} = 1$, and smallish values $\mathrm{Corr}\{\overline{g}_0, \overline{g}_0^\alpha\}$ elsewhere. This is a requirement for successful temporal correlation between satellite signal and receiver replica.

jäljitelmä

The off-peak values will however never all be quite as small as the value $-{}^1\!/_{(2^n-1)}$ for m-sequences.

Next, look at the cross correlation. For $i \neq j \iff \alpha \neq \alpha'$:

$$g_i \oplus g_j^\delta = (x \oplus y^\alpha) \oplus \left(x \oplus y^{\alpha'}\right)^\delta =$$
$$= \left(x \oplus x^\delta\right) \oplus \left(y^\alpha \oplus y^{\alpha'+\delta}\right) = \begin{cases} y^\gamma & \delta = 0, \\ x^\beta & \alpha = (\alpha' + \delta) \bmod N, \\ x^\beta \oplus y^\gamma & \text{otherwise.} \end{cases}$$

This is similar to what we found for the autocorrelation, and also the conclusion is the same. This is good news for using the sequences to distinguish and separate the signals from different satellites.

It is also not difficult to prove that the sequences $g_0 = x \oplus y$ and $g_1 = x \oplus y^\alpha$ are *essentially* different — meaning that they cannot be converted into each other by a cyclical shift. Proof by contradiction: if it were true that g_1 could be produced by a cyclical shift of g_0, we would have

$$g_0^\beta = g_1,$$

where β denotes this hypothetical, non-zero shift. Writing this out yields

$$(x \oplus y)^\beta = x \oplus y^\alpha \implies x^\beta \oplus y^\beta = x \oplus y^\alpha.$$

XORing from the right with y^α yields

$$x^\beta \oplus y^\beta \oplus y^\alpha = x \oplus \overbrace{y^\alpha \oplus y^\alpha}^{0} \implies x^\beta \oplus y^\beta \oplus y^\alpha = x,$$

and now XORing this from the left with x^β yields

$$\overbrace{x^\beta \oplus x^\beta}^{0} \oplus \overbrace{y^\beta \oplus y^\alpha}^{Y} = \overbrace{x^\beta \oplus x}^{x^\gamma} \implies \begin{cases} \beta = \alpha: & Y = 0 = x^\gamma, \\ \beta \neq \alpha: & Y = y^\delta = x^\gamma, \end{cases}$$

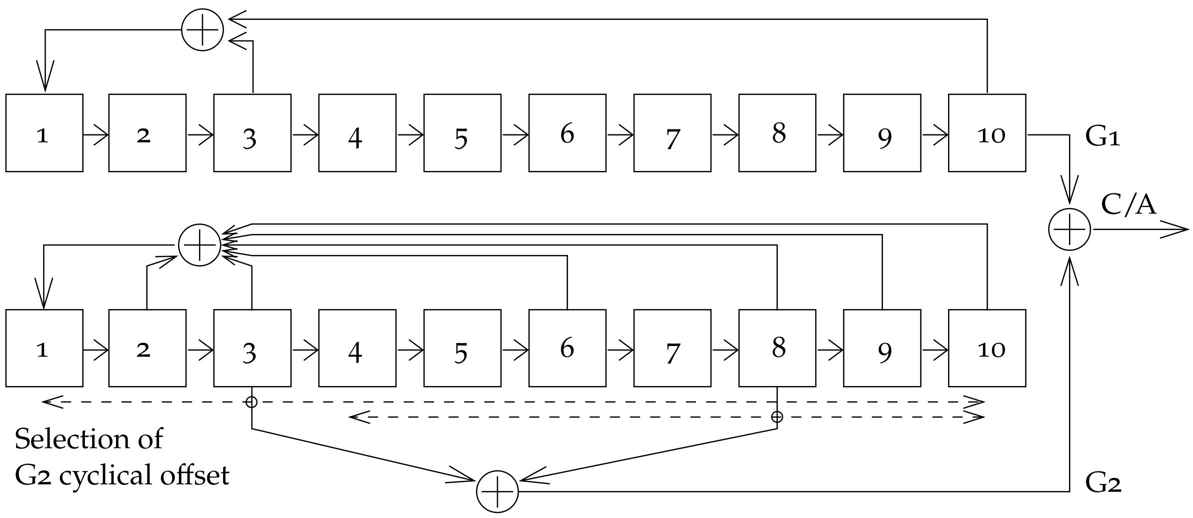


FIGURE B.1. Construction of the C/A code using two linear feedback shift registers, G1 and G2, of ten cells each and different geometries of the feedback taps. The cyclical offset of G2 relative to G1, which is different for each satellite, is set before launch by selecting and mixing two cell outputs: $y^\alpha \oplus y^{\alpha'} \to y^\beta$. Tapping from positions 3 and 8 as done in the figure produces the PRN code 31.

showing that a cyclically shifted x equals either zero or a cyclically shifted y, the latter meaning that x and y come from the same shift-register geometry: a contradiction.

The above proof was for the pair g_0 and g_1. Because of cyclicity, it holds for any g_i instead of g_0: picking the "seed" y out of its set of $N = 2^n - 1$ cyclically shifted versions and thus, which g_i is cast in the role of g_0, is arbitrary. If we call $\widetilde{y} \stackrel{\text{def}}{=} y^\alpha, \widetilde{x} \stackrel{\text{def}}{=} x^\alpha$, then

$$\widetilde{g}_0 \stackrel{\text{def}}{=} g_0^\alpha = (x \oplus y)^\alpha = x^\alpha \oplus y^\alpha = \widetilde{x} \oplus \widetilde{y}$$

and

$$y^{\alpha'} = (y^\alpha)^{\alpha' + (N-\alpha)} \stackrel{\text{def}}{=} \widetilde{y}^\beta, \qquad y^{\alpha''} = (y^\alpha)^{\alpha'' + (N-\alpha)} \stackrel{\text{def}}{=} \widetilde{y}^{\beta'}, \qquad \ldots$$

and the other g are

$$\widetilde{g}_1 = \widetilde{x} \oplus \widetilde{y}^\beta, \qquad \widetilde{g}_2 = \widetilde{x} \oplus \widetilde{y}^{\beta'}, \qquad \widetilde{g}_3 = \widetilde{x} \oplus \widetilde{y}^{\beta''}, \qquad \ldots \tag{B.3}$$

Also the choice which member of set B.3 to call g_1 is free.

See figure B.1 for a hardware implementation.

C The Woodbury matric identity

Start from these two matric equations:

$$AX + UY = I, \tag{C.1}$$
$$VX - C^{-1}Y = 0. \tag{C.2}$$

Add equation C.2 multiplied from the left by UC to equation C.1:

$$(A + UCV)\,X = I \implies X = (A + UCV)^{-1}. \tag{C.3}$$

Subtract equation C.1 multiplied from the left by VA^{-1} from equation C.2:

$$\left(-C^{-1} - VA^{-1}U\right)Y = -VA^{-1} \implies Y = \left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}.$$

Substitute back into equation C.1:

$$\begin{aligned} AX + U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1} &= I \\ \implies X &= A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}. \end{aligned} \tag{C.4}$$

Thus we found two different expressions C.3 and C.4 for matrix X, which must be *identical*. We obtain

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}, \tag{C.5}$$

the *Woodbury*[1] *matrix identity* (Keijo Inkilä, personal communication); [1] Wikipedia, Woodbury matrix identity.

[1] Max Atkin Woodbury (1917–2010) was an American mathematician.

Real-time systems and networks

D

General requirements for technological navigation are the ability to receive external data in real time over the data communications network as well as a processing capability with equipment and software appropriate and sufficient for real-time use. We shall consider those requirements next.

The definition of *real time* is tosiaikaisuus

Guaranteed latency.

So, a process that has a latency of a month can be real-time if a month is *guaranteed*, but another process with a latency of a millisecond is not real-time if the latency is usually less than a millisecond, but it could sometimes be two milliseconds, or ten, or even more.

D.1 Communication networks

D.1.1 Broadcasting networks

Broadcasting networks, one-to-many communication networks, are almost as old as the discovery of radio waves. Radio waves — carrier waves — can be used to carry signals in digital or analogue form. An example of the former is using the Morse code in radio telegraphy. Examples of analogue signals are sound (radio, radio telephony) and images (television). Measurement data or telemetry is almost always digital.

Information is carried on radio waves by *modulation*. Often used modulation techniques are amplitude modulation, frequency modulation and phase modulation.

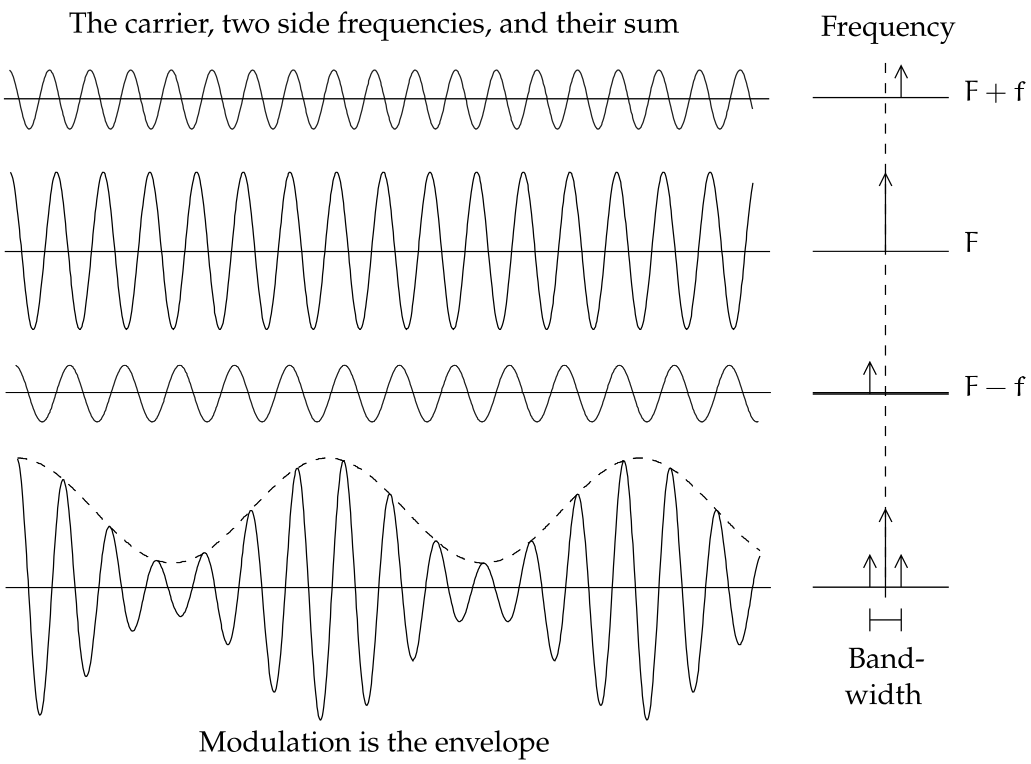


FIGURE D.1. Amplitude modulation and bandwidth.

D.1.1.1 Example: amplitude modulation

Figure D.1 shows how amplitude modulation places a signal — the dashed curve, for example a sound wave — on top of the carrier wave.[1] To the right we see what the spectrum of the modulated wave looks like.

If we call the carrier frequency F and the modulating signal (sound) frequency f, we can write the modulated signal as

$$A(t) = \cos(2\pi F t) \cdot \left(\tfrac{3}{2} + \cos 2\pi f t\right) =$$
$$= \tfrac{3}{2}\cos 2\pi F t + \tfrac{1}{2}\Big(\cos\big(2\pi\,(F+f)\,t\big) + \cos\big(2\pi\,(F-f)\,t\big)\Big).$$

We see that the contribution of the modulation can be represented as the semi-sum of two side frequencies, $F + f$ and $F - f$.

Now, if the modulating wave contains many different frequencies f, $0 < f < f_{max}$, the resulting spectrum will contain signal, or power, over

[1] In the example, it is assumed that the carrier wave is stronger than the modulations, as is the case in traditional amplitude modulation. However, this carrier contains a great deal of power and is therefore sometimes suppressed. One of the side bands may also be suppressed for single side band (SSB) modulation, saving half the bandwidth.

the full range $(F - f_{max}, F + f_{max})$. We say that the *bandwidth consumption* is $2f_{max}$.

For broadcasting networks, bandwidth is a scarce and valuable resource to be judiciously allocated.

D.1.1.2 The Nyquist Theorem

A common problem is representing a function of time by discrete sample points, or equivalently, representing a stochastic process by a time series. One can show that the time interval Δt between sample points should not be longer than *one-half the shortest period* present in the function. This is called the Nyquist[2] Theorem. The discrete representation of a function may be transformed, using the discrete Fourier transform, from the time-domain form $A(t)$ to the frequency-domain form $\mathcal{A}(f)$, and back again. If the Nyquist condition is satisfied, both forms contain the full information content of the original function. Numerically, the fast Fourier transform (FFT) is used in these computations.

otospiste

If as a modulating signal is given a function of time that has as its highest contained frequency f_{max}, then the shortest period contained in it is $^1/_{f_{max}}$. The number of samples transmitted per second will then have to be $2f_{max}$, precisely the bandwidth occupied by the amplitude modulated signal.

D.1.2 Circuit-switched networks

D.1.2.1 History

The first, wildly successful and still existing, switched or many-to-many connection network was the telephone network. It is actually circuit-switched: it establishes a temporary but persistent connection between speakers.

The invention of the telephone is usually credited to Alexander Graham Bell. In reality, like with the steam engine, the telescope, and many other very practical inventions, the time was ripe for it and many people, like Elisha Gray (who filed his patent a mere two hours after Bell!), Antonio Meucci (Carroll, 2002), and Thomas Edison, contributed valuable ideas before a working implementation became the basis of the first telephone network.

[2]Harry Nyquist (1889–1976) was a Swedish-born American electronic engineer.

For many years, American Telephone & Telegraph held a monopoly on telephone technology. From time to time there were anti-trust proceedings against the company, which is also credited with laying the first trans-Atlantic telephone cable, building the first active, or amplifier-based, communications satellite Telstar, and inventing the
3 UNIX®[3] operating system. . . .

Telephony is based on transmitting sound in electric form over a copper cable. This is still the way it happens with landline phones for the few metres nearest to the subscriber. All other equipment in the network is nowadays fully digital. Making a connection between two telephone customers was at the very beginning done by hand. As early as before 1900 the first mechanically automated exchanges were built.
4 A number was dialled using a rotating disc[4] with finger holes, sending as many pulses as the number printed under the chosen finger hole.
pulssivalinta This is called "pulse dialling". Since its introduction in 1963, faster tone
äänivalinta dialling has mostly, though not completely, replaced this.

The number system for telephones is a three-layer hierarchical system that is not controlled from a single point: a remarkable invention. It has aged well, in spite of being uniquely user-hostile: looking up telephone numbers was traditionally done manually using thick paper books. The world is divided into national domains with country codes. The United States has code 1, most larger countries have two-digit codes (for example Germany 49), while smaller, poorer countries like Finland have settled for three-digit codes (358). Under the national-domain level are trunk codes, typically (although not necessarily) for cities, within which individual subscribers have their numbers.

Attempts to make phone numbers "mnemonic" so that they can be more easily remembered have pretty much failed. New telephone concepts such as Internet telephony might in the longer run change
5 this. Remarkably, the applications Signal®[5] and WhatsApp®[6] use the
6 number of the mobile phone they are installed on as the user ID! Old habits die hard.

The digitalisation of the telephone network has made it possible to

---

[3] UNIX is a registered trademark of The Open Group.

[4] Like a clock dial, which gave the rotary telephone dial its name.

[5] Signal is a registered trademark of Signal Messenger LLC.

[6] WhatsApp is a registered trademark of WhatsApp Inc.

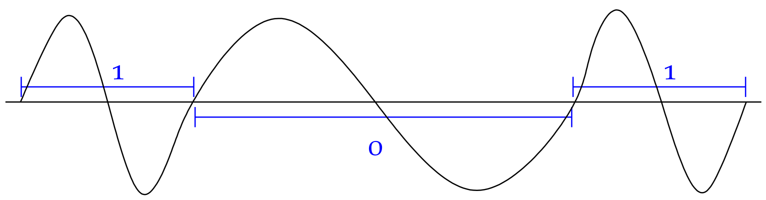


FIGURE D.2. "Binary frequency-shift keying" (BFSK) modulation.

offer customers "always-on" data connections, even over the last few metres of copper wire, which use frequencies above those used for audible sound. With a low-pass filter in between, one may even use voice and data simultaneously on the same line (digital subscriber line, DSL).

D.1.2.2 Modems

As the telephone network is designed for the transport of sound, for data transport to work on it, the data must be converted to and from sound in the form of analogue sound waves. This is done with a device called a *modem* (modulator-demodulator).

Figure D.2 shows one technique — binary frequency-shift keying, BFSK — often used for modulation: a logical 1 is encoded as a short (high-frequency) wave, a logical 0 as a long (low-frequency) wave. This is a simple, somewhat wasteful but effective and robust modulation technique.[7] In addition, checksums are transmitted in order to verify that the data received equals the data sent (parity check, cyclic redundancy check[8] CRC) even over noisy lines. Data compression is used if possible and speeds up the transfer of textual material in particular.

binaarinen taajuus-avainnus

7

8

There are a number of standards for modems, mostly created by the International Telecommunications Union ITU. Over a quality analogue line, 56k bits per second is the best achievable on audio frequencies using a traditional dial-up modem.

Using a modem to transfer data over a network designed for sound only is an example of a *protocol stack*: the lowest layer is sound transfer, upon which digital data transfer in the form of a bitstream is layered. Other layers can be placed on top of this: the Internet Protocol (IP)

[7]The technique was also used for cassette-tape storage peripherals for home computers like the Commodore 64.

[8]Both parity check and CRC are somewhat similar to the "casting out nines" check on manual calculations involving decimal numbers.

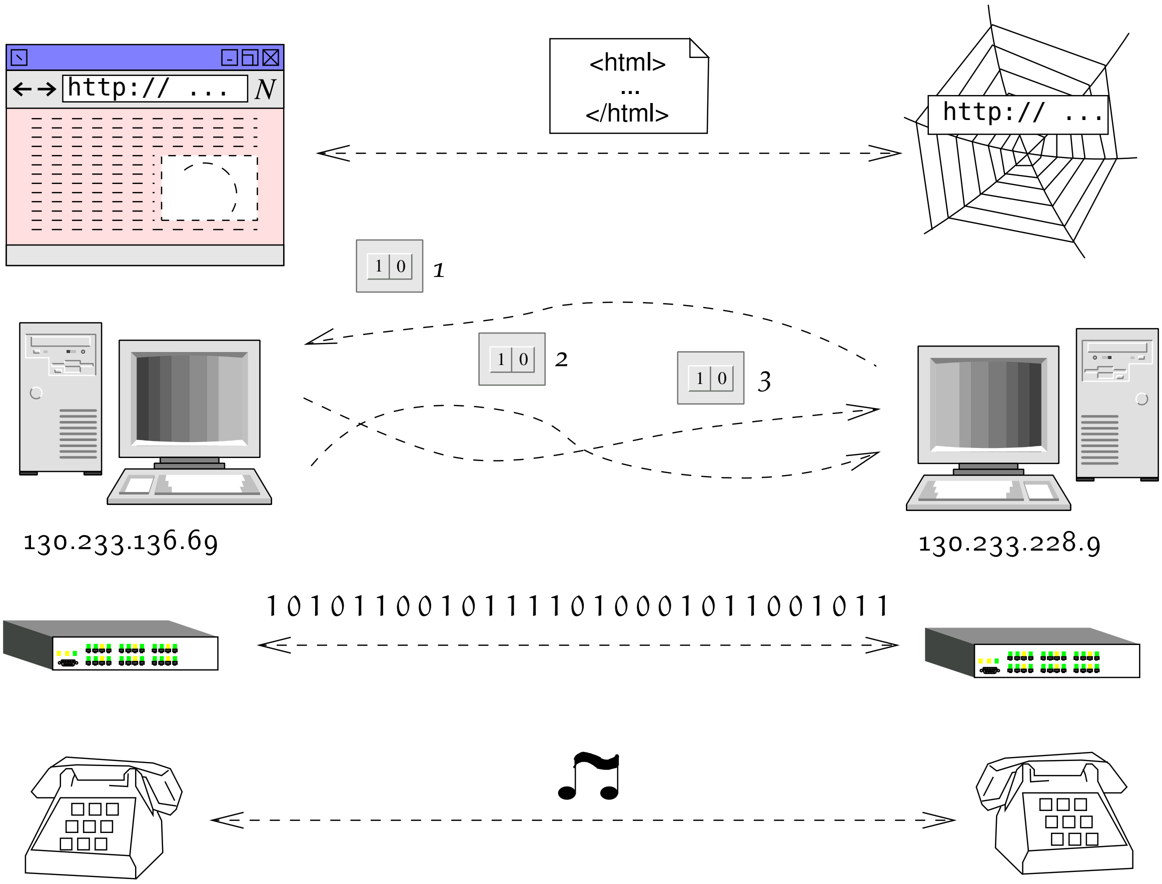


FIGURE D.3. Example of a protocol stack.

and the Transmission Control Protocol (TCP), to be discussed later, advanced protocols such as the World-Wide Web service HTTP, and so on. Establishing such a connection requires bringing up every layer of the stack in succession, from the ground up.

In a protocol stack, the higher layers are usually implemented in software, whereas the lowest layers are hardwired. For example telephone sound is transmitted traditionally as voltage fluctuations in a copper wire. As digital technology develops, however, the software spreads down the stack: for all but the last few metres, telephonic sound nowadays moves as digital bit patterns, often in fibre-optic cables.

This downward migration of software is leading to devices that previously were very different becoming almost the same at the hardware level: for example, a telephone and a television set — and a GNSS receiver — are becoming mostly just general-purpose computers, but differently programmed and with different peripherals. This phenomenon is known as *technological convergence*.

D.1.2.3 Mobile devices

Mobile phones based on GSM (Global System for Mobile Communications) can also be used for data transfer. Data rates achievable for the original GSM protocol were 9600–14 400 bits per second.

However, currently more advanced protocols such as GPRS (General Packet Radio Services) and EDGE (Enhanced Data Rates for GSM Evolution) are in use, allowing always-on digital connections with much higher data rates. Often these technologies are metaphorically referred to in terms of generations: 2G, 3G, and 4G.

This brings us to the following subject: packet-switching networks.

D.1.3 Packet-switching networks

The classic packet-switching network is the Internet. This is also a many-to-many communication network — but there the similarity with the telephone network ends. The Internet is based on the transfer of *packets* made up of data bytes and accompanying information. There is no way of telling how a particular packet will reach its destination — or indeed whether it will at all, and if so, how quickly.

The idea of packet-switching networks as an alternative to circuit-switching ones originated in military research: packet-switched network transfer is less vulnerable to localised network damage and thus harder to interrupt. The same applies for attempts at blocking the network: "the Internet interprets censorship as damage, and routes around it."

The functioning of the Internet, IP addresses, and the Domain Name System (DNS) is explained in many places[9] and we will not repeat those explanations here. There are a number of protocols built upon the Internet Protocol, the most important of which are 9

- ICMP (Internet Control Message Protocol), for example the well-known command `ping`[10] for checking network connectivity. 10
- UDP (User Datagram Protocol) is a connectionless protocol: a transmitter sends out packets and a receiver receives them — most of the time. There is no check on successful reception, reception in the correct order, or duplicate reception. But UDP's overhead is low, which is usually why it is used. For example the Network

[9] For example Wikipedia, Domain Name System.

[10] The name is onomatopoeic and mimics the sound of sonar, Muuss (undated).

Time Protocol NTP uses UDP. A time server just sends out packets for the clients to pick up and synchronise their clock to.

- TCP (Transmission Control Protocol) is a *connection-based* protocol. It establishes a connection between two hosts on the Internet, and then exchanges packets in both directions, until the connection is closed. It is thus a *bidirectional* protocol, but is always initiated from one side, normally the client side.

  The packets may travel from one host to the other over many different paths. The receiver places them in the proper order based on a *sequence number* contained in every packet. If a packet is missing and has timed out, a request to re-send is issued. Thus, TCP is *reliable*.

  The security of the connection is improved by each host randomly — or at least in a way that is not easily predictable — choosing the starting value of its packet counter for this connection. Such a connection could be hijacked in principle — a so-called "man-in-the-middle attack" — but this makes it a little harder.

For UDP and TCP, every packet contains two data fields called *source* and *destination port*. These are sixteen-bit numbers between 0 and 65 535 which are used to distinguish various service types from each other. [11] For example HTTP uses port 80 — usually.[11] One should understand that these "ports" are purely software matters: it is the networking software layer in the operating system that distinguishes these port numbers from each other and directs packets to appropriate server or client processes. Nothing like a *hardware* serial or parallel or USB port!

None of these Internet protocols are truly real-time — because the Internet is not. They are sometimes used in a real-time fashion, assuming that the latency on a transmission will never become very large, but that is a *gamble*: a fairly harmless one, for example, for music streaming. But already moderate congestion — locally or upstream — will make transmission times unpredictable.

D.2 Real-time systems

D.2.1 Hardware

tosiaikainen The digital hardware included in real-time systems used for navigation

[11]There is a list of all services in the file `/etc/services`.

will typically have modest processing power. Think for example of mobile phones: the dictate of low power consumption and small form factor limits the kinds of circuitry and how much of it one can use.

Another limitation may be that a full-size keyboard cannot be used and instead of a mouse, a touch screen of limited size is indicated. Antenna size may be limited, too. Physical ruggedness may also be required, depending on the navigation environment.

D.2.2 Operating systems

The hardware limitations mentioned clearly also limit what system software can be used. "Embedded" operating systems, such as embedded versions of Android™,[12] Windows®,[13] and Linux®,[14] are commonly 12
found, for example in UAVs. 13
14

In high-reliability operations like on spacecraft, systems like the QNX®[15] and Wind River Systems real-time embedded operating systems 15
are also being used. The Mars rovers Spirit, Opportunity, and Curiosity as well as the Mars lander InSight used and use the Wind River Systems VxWorks®[16] software. 16

In "hard" real-time applications, meaning applications for which the real-time requirement is unconditional, the operating system should preferably not crash. Spirit's system did once crash due to the memory filling up with handles for files in flash storage. After clean-up, it came beautifully back up again (Weiss, 2004).

It will be clear that when interfacing with various devices such as GPS and other sensors, the availability or ease of development of device drivers is critical. laiteajuri

As with technology development, hardware capability grows while size and power consumption diminish, more and more general consumer-grade operating systems, slightly adapted, are also finding their way into these constrained mobile platforms.

A basic operating system (OS) functions in the following way: upon

[12] Android is a common-law trademark of Google Inc.

[13] Windows is a registered trademark of Microsoft Corporation.

[14] Linux is a registered trademark of The Linux Foundation.

[15] QNX is a registered trademark of QNX Software Systems Ltd.

[16] VxWorks is a registered trademark of Wind River Systems Inc.

start-up, after operating-system, file-system and device-driver functions have been enabled, the initial process spawns all the background service processes (daemons) that need to run on this system and goes into multi-user mode. It loads a login process, presenting it to the user on one or more consoles connected to the system. When users log in, they are presented with a *shell* or command interpreter, allowing them to start their own user processes.

A consumer-grade OS will at this stage start up a windowing GUI (graphical user interface) as well, allowing operation by lightly trained personnel. This, however, demands extra resources. User processes can also be started from the GUI. These processes will then interact with the user through the GUI.

The defining property of an operating system is that it manages the system's various resources in a way that is transparent and abstracts from their technical details. Device drivers are one example of this. And,
vuorontaja for example, processor resources are managed through the *scheduler*.

D.2.3 Process flow

[17] For a single process,[17] the path of execution is *linear*. This means that after the execution of a statement, the process either proceeds to the next statement of the program, or to the statement pointed to by a branching (`if`, `switch`, ...) statement. This makes it easy to keep track of the current state of the process: it can only be changed by statements that are executed.

aliohjelma A *procedure*, also called a subroutine, function, or method, is only executed because another procedure, and ultimately the main program, called it in the course of *its* linear execution. A procedure is executed as follows: when it is called, it places a *return address*, the current value held by the program counter, on the *execution stack*. The stack is a LIFO — last in, first out — data structure used in connection with procedure calls.

Next, any parameter values or pointers to global variables contained in the procedure call are also placed on the top of the stack, making the stack grow. There are two ways of calling parameters or arguments with a procedure call: *call by name* and *call by value*.

nimi-välitteisyys

- *Call by name*: the name of a variable is used as a parameter or

[17] ... and ignoring threading!

argument to a procedure call. A pointer to the storage location allocated to the variable by the calling program is pushed onto the stack.

- *Call by value*: as a parameter of a procedure call is used either an expression or a constant. The expression is evaluated at execution time when the procedure is called, and the value obtained is placed on the stack. arvovälitteisyys

If it is desired that a result is returned from the computations within the procedure to the calling procedure, one must use call by name. Then, inside the procedure, an assignment is made to the variable, or rather to the storage location it points to, where the calling procedure which reserved this location will find it.

Local variables declared within the procedure are also allocated on the stack. The *scope* of these variables — their area of validity or visibility — is limited to within the procedure being executed. näkyvyysalue

When the flow of control meets the end of the procedure, first these local variables are deallocated. After that, the procedure's stacked parameters (constants, expressions or pointers) are deallocated, and finally the top of the stack is moved back into the program counter of the processor. Thus the flow of control has returned to the calling procedure, which continues from the statement following the procedure call.

Advantages of using an execution stack are:

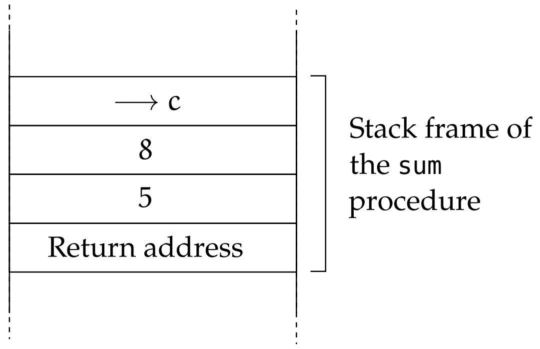
- Local variables declared inside procedures are released upon return from the procedure.
- Procedures can be *called recursively*: every instance of the procedure has its own stack frame with its own version, or instance, of the call parameters and local variables. This, of course, does require that the recursion must be ended at some point by a suitable condition: if not, a stack overflow error is inevitable.

D.2.4 Memory allocation

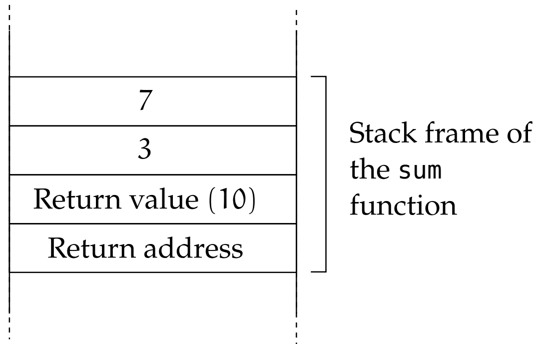
With a *stack*, memory is allocated linearly by means of a stack pointer. Memory is released in reverse order. The stack pointer always points to the topmost item on the stack, the last one that was allocated and the first one that will be released. LIFO — last in, first out.

In the case of a *heap*, the memory allocation also occurs linearly, but

```
procedure sum(a, b, c)
c = a + b
return
...
declare c
# Call by name c
call sum(5, 8, c)
print c
>> 13
```




```
function sum(a, b)
sum = a + b
return
...
declare c
c = sum(3, 7)
print c
>> 10
```




```
procedure sum(a, b, c)
c = a + b
return
...
declare a, b, c
a = 6
b = 5
# Call by name a, b, c
call sum(a, b, c)
print c
>> 11
```

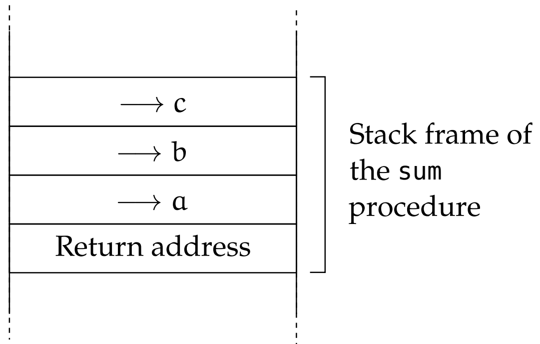


FIGURE D.4. Procedure call and stack allocation.

release occurs asynchronously. A heap is a suitable memory solution when storing variables of varying sizes, such as text strings. A text string is stored as a pointer on the stack that points to the content of the string on the heap. For example, when two strings are concatenated into a new string, the new string is put onto the heap and the pointers to the old strings are moved from the stack to a freed-memory list. The same happens with a string that was declared locally within a procedure, when control returns from the procedure.

When the memory of the heap has been used up, a compression operation called *garbage collection* is started, in which the memory allocation holes, the deleted old strings or "corpses", are also physically reclaimed.

D.2.5 Interrupts and masking

Interrupts are both similar to and different from procedure calls. The essential difference is that interrupts can happen at any time due to an external hardware event. Usually the event is a peripheral either having data available to be read from it or being ready for data to be written to it. The main reason for using interrupts is that, without them, the peripheral would have to be *polled* periodically, at a rate sufficient not to miss any transferable data. This polling consumes processor time. kiertokysely

Computer hardware provides for a number of different interrupts. When they are triggered, it is their responsibility not to change anything that could interfere with the processes scheduled for execution. Interrupts are used, for example, to service input-output devices that cannot wait. Every interrupt is associated with an *interrupt vector* — a memory address to be loaded into the program counter — to an *interrupt service routine* or interrupt handler, which is executed when the interrupt is triggered.

Take the clock interrupt routine, for example. It is triggered typically 50 times a second and gets its name from what it does: increment the time register kept by the operating-system software. It is usually also responsible for *task-switching* or *context-switching*, allowing multiple tasks to run seemingly simultaneously. vuoronvaihto

At every task switch, the *context* of the currently running process — the set of data including the CPU registers and in particular the program counter, that the process needs in order to continue running from the point where it was interrupted — is saved and another process, after loading *its* context data, is allowed to run during the next "time slice" of 0.02 s.

The decision on which process to *schedule* next is a subject on which thick books have been written. It should be a process that is "runnable" and not, for example, waiting for user input, and should have a high enough *priority*. Interactive processes are given high priority, while batch processes run under low priority. In a multi-user environment, there should be a mechanism for voluntarily lowering the priority of unattended batch runs: the `nice` command. vuoronnus

All processes — especially kernel or system-level processes — have snippets in their code during which it would be wrong or disastrous to be interrupted. An example of this is transferring data to memory from a hard-disc sector as it spins past the reading head. We humans know this all too well: there are certain tasks that we simply cannot do if we are not left in peace to do them, and if we are interrupted, we just have to start from the beginning again, if not worse. Computers are no different.

This is why it is possible for interrupts to be *masked* or disabled. Critical kernel routines will mask the clock and other interrupts and *unmask* them again when they finish.

D.2.6 Requirements for real-time use

tosiaikaisuus The requirements for real-time use are the following.

1. One must know in advance which processes will be running on the system. For example, an environment like a multi-user server that people can log in to and start user processes at will is not acceptable.
2. It must be known in advance what are the *longest* snippets of code, execution-time wise, that the various runnable processes contain *during which they may not be interrupted*. These durations should *all* be acceptably short.
3. The real-time critical processes should receive the highest priority, all others a lower priority.
4. The time interval for task-switching should be suitably short, 0.02 s may be too long. This depends on the task at hand.
5. The total processing capacity of the system should be sufficient
   (a) on average for all processes
   (b) at *all points in time* for *all* the real-time processes together.

Meeting "hard" real-time requirements is demanding and requires extensive load-testing as well as substantial overallocation of capacity. Often it is wise to ask if hard real-time is really what is needed and settle for something more modest. For example, audio or video streaming software for the Internet take the non-real-time nature of the Internet into account by *buffering*, collecting incoming data packets into a buffer from which an uninterrupted sound and video experience can be created, even if some packets arrive rather late. The price paid is a delay of a few seconds. This is very noticeable if the same, live source is also available in analogue form, but irrelevant when playing back stored content.

Bibliography

ABCDEFGHIJKLMNOPQRSTUWXZ

A

David W. Allan. Statistics of atomic frequency standards. *Proceedings of the IEEE*, 54(2):221–230, 1966. URL https://doi.org/10.1109/PROC.1966.4634. 193

Allan's Time. The Allan Variance. URL http://www.allanstime.com/AllanVariance/. Accessed 6th May, 2019. 192

Analog Devices. Data Sheet ADXL103/ADXL203 Rev. F. URL https://www.analog.com/media/en/technical-documentation/data-sheets/adxl103_203.pdf. Accessed 11th July, 2020. 297

Neil Ashby. Relativity in the Global Positioning System. *Living Reviews in Relativity*, 55, January 2003. URL https://doi.org/10.12942/lrr-2003-1. 172

Jose Ángel Ávila Rodríguez. Galileo Signal Plan. *ESA Navipedia*, 2011. URL https://gssc.esa.int/navipedia/index.php/Galileo_Signal_Plan. Accessed 16th June, 2020. 268

B

Willem Baarda. *A testing procedure for use in geodetic networks*, volume 2 number 5 of *Publications on geodesy, new series*. Netherlands Geodetic Commission, Delft, 1968. URL https://www.ncgeo.nl/downloads/09Baarda.pdf. Accessed 14th May, 2019. 242

Tulu Besha Bedada. *Absolute geopotential height system for Ethiopia*. PhD thesis, University of Edinburgh, 2010. URL https://era.ed.ac.uk/handle/1842/4726. Accessed 22nd September, 2021. 277

BeiDou Constellation Status. Test and Assessment Research Center of China Satellite Navigation Office. URL http://www.csno-tarc.cn/en/system/constellation. Accessed 17th September, 2021. 268

BeiDou ICD. BeiDou Navigation Satellite System Signal In Space Interface Control Document, Open Service Signal B2a (Version 1.0). PDF, China Satellite Navigation Office, 2017. URL http://en.beidou.gov.cn/SYSTEMS/ICD/201806/P020180608518432765621.pdf. Accessed 16th June, 2020. 268

BeiDou ICD. BeiDou Navigation Satellite System Signal In Space Interface Control Document, Open Service Signal B1C (Version 1.0). PDF, China Satellite Navigation Office, 2018a. URL http://en.beidou.gov.cn/SYSTEMS/ICD/201806/P020180608519640359959.pdf. Accessed 16th June, 2020. 268

BeiDou ICD. BeiDou Navigation Satellite System Signal In Space Interface Control Document, Open Service Signal B3I (Version 1.0). PDF, China Satellite Navigation Office, 2018b. URL http://en.beidou.gov.cn/SYSTEMS/ICD/201806/P020180608516798097666.pdf. Accessed 16th June, 2020. 268

BeiDou ICD. BeiDou Navigation Satellite System Signal In Space Interface Control Document, Open Service Signal B1I (Version 3.0). PDF, China Satellite Navigation Office, 2019. URL http://en.beidou.gov.cn/SYSTEMS/ICD/201902/P020190227702348791891.pdf. Accessed 16th June, 2020. 268

BeiDou ICD. BeiDou Navigation Satellite System Signal In Space Interface Control Document, Open Service Signal B2b (Version 1.0). PDF, China Satellite Navigation Office, 2020. URL http://en.beidou.gov.cn/SYSTEMS/ICD/202008/P020200803539206360377.pdf. Accessed 21st February, 2022. 269

BeiDou Introduction. CSNO-TARC, Test and Assessment Research Center of China Satellite Navigation Office. URL http://www.csno-tarc.cn/en/system/introduction. Accessed 17th April, 2020. 270

BeiDou Navigation Satellite System. Serve the World and Benefit Mankind. URL http://en.beidou.gov.cn/. Accessed 29th May, 2020. 268

BKG, Ntrip and RTCM version 3. Ntrip - Networked Transport of RTCM via Internet Protocol. URL https://igs.bkg.bund.de/ntrip. Accessed 17th September, 2021. 217

BKG, NTRIP v. 1.0. Networked Transport of RTCM via Internet Protocol (NTRIP) — Version 1.0. URL https://igs.bkg.bund.de/root_ftp/NTRIP/documentation/NtripDocumentation.pdf. Accessed 5th May, 2019. 219

John M. Brozena. The Greenland Aerogeophysics Project: Airborne gravity, topographic and magnetic mapping of an entire continent. International Association of Geodesy Symposia 110, pages 203–214, Vienna, Austria, 20th August, 1992. Springer, New York, New York. URL https://doi.org/10.1007/978-1-4613-9255-2. 276

C

Rory Carroll. Bell did not invent telephone, US rules. *Guardian*, 17th June 2002. URL https://www.theguardian.com/world/2002/jun/17/humanities.internationaleducationnews. Accessed 30th April, 2020. 313

Liang Chen, Ling Pei, Heidi Kuusniemi, Yuwei Chen, Tuomo Kröger, and Ruizhi Chen. Bayesian fusion for indoor positioning using Bluetooth fingerprints. *Wireless Personal Communications*, 70(4):1735–1745, 2013. URL https://doi.org/10.1007/s11277-012-0777-1. Accessed 18th May, 2019. 299

Ruizhi Chen, Felix Toran-Marti, and Javier Ventura-Traveset. Access to the EGNOS signal in space over mobile-IP. *GPS Solutions*, 7:16–22, 2003. URL https://doi.org/10.1007/s10291-003-0050-x. 258

William H. Clohessy and R. S. Wiltshire. Terminal guidance system for satellite rendezvous. *Journal of the Aerospace Sciences*, 27(9), 1960. URL https://doi.org/10.2514/8.8704. 139, 159

Control Segment. GPS Interface Control Documents, 2018. URL https://www.gps.gov/systems/gps/control/. Accessed 5th August, 2020. 165

Michael Alan Ralph Cooper. *Control Surveys in Civil Engineering*. Collins, Department of Civil Engineering, The City University, London, 1987. 117

John Peter Costas. Synchronous communications. *Proceedings of the IRE*, 44(12): 1713–1718, 1956. URL https://doi.org/10.1109/JRPROC.1956.275063. 189

D

Michael W. Davidson and Kirill I. Tchourioukanov. Newton's prism experiments. URL https://micro.magnet.fsu.edu/primer/java/scienceopticsu/newton/. Accessed 1st April, 2020. 36

TU Delft, LAMBDA. Download LAMBDA and Ps-LAMBDA source code. URL https://www.tudelft.nl/citg/over-faculteit/afdelingen/geoscience-remote-sensing/research/lambda/lambda/. Accessed 6th May, 2019. 229

Jared Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. Norton, New York, New York, 1999. 3

David Dickinson. Remembering John Houbolt: the man who gave us lunar orbit rendezvous. *Universe Today*, 2014. URL https://www.universetoday.com/111424/remembering-john-houbolt-the-man-who-gave-us-lunar-orbit-rendezvous/. Accessed 4th May, 2019. 51

Edgar Durbin. Wikimedia Commons, Teterboro, ST-124 uncovered, 2004. URL https://commons.wikimedia.org/wiki/File:ST-124_uncovered_(IMGP3445).JPG. © 2004 Edgar Durbin (GFDL). Accessed 4th May, 2019. 115

E

EDAS. EGNOS User Support. URL https://egnos-user-support.essp-sas.eu/new_egnos_ops/services/about-edas. Accessed 17th September, 2021. 258

EDAS Service Definition Document. EGNOS User Support, 2019. URL https://egnos-user-support.essp-sas.eu/new_egnos_ops/sites/default/files/documents/egnos_edas_sdd_in_force.pdf. Accessed 17th September, 2021. 258

Carsten Egevang, Iain J. Stenhouse, Richard A. Phillips, Ævar Petersen, James W. Fox, and Janet R. D. Silk. Tracking of Arctic terns *Sterna paradisaea* reveals longest animal migration. *Proceedings of the National Academy of Sciences*, 2010. URL https://doi.org/10.1073/pnas.0909493107. i

ESA, Galileo Navigation Signals and Frequencies. URL https://www.esa.int/Applications/Navigation/Galileo/Galileo_navigation_signals_and_frequencies. Accessed 22nd September, 2021. 266

ESA, Introducing GOCE. URL https://www.esa.int/Applications/Observing_the_Earth/GOCE/Introducing_GOCE. Accessed 5th May, 2019. 279

European GNSS Service Centre. Constellation Information. URL https://www.gsc-europa.eu/system-service-status/constellation-information. Accessed 16th April, 2020. 266

Exploratorium, Never Lost. URL https://www.exploratorium.edu/video/collections/never-lost-polynesian-navigation. Accessed 17th September, 2021. 3

F

René Forsberg, Klaus Hehl, Luísa Bastos, Arne Gidskehaug, and Uwe Mayer. Development of an airborne geoid mapping system for coastal oceanography (AGMASCO). In Segawa et al. (1996), pages 163–170. URL https://rd.springer.com/chapter/10.1007/978-3-662-03482-8_24. Accessed 17th September, 2021. 277

René Forsberg, Arne V. Olesen, Hasan Yildiz, and Carl Christian Tscherning. Polar gravity fields from GOCE and airborne gravity. In *4th International GOCE User Workshop*, TU Munich, Germany, 2011. URL https://earth.esa.int/eogateway/documents/20142/37627/Polar_Gravity_Fields_GOCE_Airborne_Gravity_R.Forsberg.pdf. Accessed 17th September, 2021. 277

James W. Fox. Intigeo® series geolocator, 2018. URL http://www.migratetech.co.uk/IntigeoSummary.pdf. Accessed 19th June, 2020. 300

G

Galileo OS SiS ICD. Galileo Open Service Signal-in-Space Interface Control Document. PDF, European Union, 2016. URL https://www.gsc-europa.eu/sites/default/files/sites/all/files/Galileo-OS-SIS-ICD.pdf. Accessed 9th August, 2020. 267

Miguel García-Fernández, Markus Markgraf, and Oliver Montenbruck. Spin rate estimation of sounding rockets using GPS wind-up. *GPS Solutions*, 12: 155–161, 2008. URL https://doi.org/10.1007/s10291-007-0074-8. 177

GDGPS APPS. Automatic Precise Positioning Service of the Global Differential GPS System. URL https://apps.gdgps.net/apps_file_upload.php. Accessed 10th May, 2019. 258

GFZ, CHAMP — Challenging Minisatellite Payload. URL https://www.gfz-potsdam.de/champ/. Accessed 5th May, 2019. 279

GLAD. Global ARAIM for Dual-Constellation webinar. European Global Navigation Satellite Systems Agency (GSA), 2020. URL https://www.euspa.europa.eu/newsroom/european-space-expo/global-araim-dual-constellation-webinar. Accessed 27th July, 2020. 244

Global Positioning Systems Directorate. URL https://en.wikipedia.org/wiki/Global_Positioning_Systems_Directorate. Accessed 17th September, 2021. 259

GLONASS ICD. Global Navigation Satellite System GLONASS Interface Control Document, Navigational Radiosignal in Bands L1, L2. PDF, Russian Institute of Space Device Engineering, Moscow, 2008. URL https://kb.unavco.org/kb/file.php?id=607. Edition 5.1. Accessed 6th March, 2022. 262

GLONASS ICD CDMA. Global Navigation Satellite System GLONASS Interface Control Document, General Description of Code Division Multiple Access Signal System. PDF, Russian Space Systems, JSC, Moscow, 2016a. URL http://web.archive.org/web/20220121130328/https://russianspacesystems.ru/wp-content/uploads/2016/08/ICD-GLONASS-CDMA-General.-Edition-1.0-2016.pdf. Edition 1.0. Accessed 21st May, 2022. 264

GLONASS ICD CDMA. Global Navigation Satellite System GLONASS Interface Control Document, Code Division Multiple Access Open Service Navigation Signal in L1 Frequency Band. PDF, Russian Space Systems, JSC, Moscow, 2016b. URL http://web.archive.org/web/20220126001255/http://russianspacesystems.ru:80/wp-content/uploads/2016/08/ICD-GLONASS-CDMA-L1.-Edition-1.0-2016.pdf. Edition 1.0. Accessed 21st May, 2022. 264

GLONASS ICD CDMA. Global Navigation Satellite System GLONASS Interface Control Document, Code Division Multiple Access Open Service Navigation Signal in L2 Frequency Band. PDF, Russian Space Systems, JSC, Moscow, 2016c. URL http://web.archive.org/web/20220126001258/https://russianspacesystems.ru/wp-content/uploads/2016/08/ICD-GLONASS-CDMA-L2.-Edition-1.0-2016.pdf. Edition 1.0. Accessed 21st May, 2022. 264

GLONASS ICD CDMA. Global Navigation Satellite System GLONASS Interface Control Document, Code Division Multiple Access Open Service Navigation Signal in L3 Frequency Band. PDF, Russian Space Systems, JSC, Moscow, 2016d. URL http://web.archive.org/web/20220126040633/https://russianspacesystems.ru/wp-content/uploads/2016/08/ICD-GLONASS-CDMA-L3.-Edition-1.0-2016.pdf. Edition 1.0. Accessed 21st May, 2022. 264

GLONASS news. GLONASS Information and Analysis Center for Positioning, Navigation and Timing. URL https://www.glonass-iac.ru/en/about_glonass/. Accessed 7th May, 2019. 262

GMV. QZSS. *ESA Navipedia*, 2011. URL https://gssc.esa.int/navipedia/index.php/QZSS. Accessed 16th June, 2020. 255

Saurabh Godha, Gérard Lachapelle, and Elizabeth Cannon. Integrated GPS/INS system for pedestrian navigation in a signal degraded environment. In *ION GNSS 2006, 26-29 September*, Fort Worth, Texas, 2006. URL https://schulich.ucalgary.ca/labs/position-location-and-navigation/files/position-location-and-navigation/godha2006_conference.pdf. Accessed 12th August, 2020. 299

Robert Gold. Optimal binary sequences for spread spectrum multiplexing. *IEEE Transactions on Information Theory*, 13(4):619–621, 1967. URL https://doi.org/10.1109/TIT.1967.1054048. 179, 305

Google Play, Sky Map. URL https://play.google.com/store/apps/details?id=com.google.android.stardroid&hl=en. Accessed 7th May, 2019. 282

GPS ICD. GPS Interface Control Documents, 2019. URL https://www.gps.gov/technical/icwg/. Accessed 5th June, 2020. 166, 262

H

Weikko A. Heiskanen and Helmut Moritz. *Physical Geodesy*. W. H. Freeman and Company, San Francisco, London, 1967. 47

Hexagon, SPAN GNSS Inertial Navigation Systems. URL https://novatel.com/products/gnss-inertial-navigation-systems. Accessed 22nd September, 2021. 293

George William Hill. On the part of the motion of the lunar perigee which is a function of the mean motions of the sun and moon. *Acta Mathematica*, 8: 1–36, 1886. URL https://doi.org/10.1007/BF02417081. 139

Bernhard Hofmann-Wellenhof, Herbert Lichtenegger, and James Collins. *GPS Theory and Practice*. Springer-Verlag, fourth, revised edition, 1997. 139, 165

Walter Hohmann. The attainability of heavenly bodies. Technical translation F44, NASA, 1925. URL https://archive.org/details/nasa_techdoc_19980230631. Accessed 12th August, 2020. 162

I

International GNSS Service. IGS Products. URL https://igs.org/products. Accessed 17th September, 2021. 212

J

Volker Janssen. A comparison of the VRS and MAC principles for network RTK. In *International Global Navigation Satellite Systems Society Symposium*, 2009. URL https://www.spatial.nsw.gov.au/__data/assets/pdf_file/0003/129414/2009_Janssen_IGNSS2009_VRS_vs_MAC.pdf. Accessed 8th August, 2020. 236

Christopher Jekeli. *Inertial Navigation Systems with Geodetic Applications*. Walter de Gruyter, Berlin – New York, 2001. 46, 101, 116

Jet Propulsion Laboratory, The Global Differential GPS System. URL https://www.gdgps.net. Accessed 22nd September, 2021. 257

K

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1st March 1960. URL https://doi.org/10.1115/1.3662552. 51

Rudolf E. Kalman and Richard S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108, 1st March 1961. URL https://doi.org/10.1115/1.3658902. 51

Dennis Kawaharada. The Settlement of Polynesia, Part 1. URL https://paulwaters.com/learning-hawaiian-culture/the-settlement-of-polynesia-part-1/. Accessed 3rd May, 2019. 3

A. D. King. Inertial navigation – forty years of evolution. *GEC Review*, 13(3): 140–149, 1998. URL http://www.imar-navigation.de/downloads/papers/inertial_navigation_introduction.pdf. Accessed 18th May, 2019. 112

Brian Koberlein. Galileo's discovery of Jupiter's moons, and how it changed the world. *Forbes*, 2016. URL https://www.forbes.com/sites/briankoberlein/2016/01/07/galileos-discovery-of-jupiters-moons-and-how-it-changed-the-world/. Accessed 2nd May, 2020. 5

L

Leica Geosystems. Take it to the MAX! White paper, 2005. URL https://www.smartnetna.com/documents/Leica_GPS_SpiderNET-Take_it_to_the_MAX_June2005_en.pdf. Accessed 16th August, 2020. 236

Dennis P. Lettenmaier. Detection of trends in water quality data from records with dependent observations. *Water Resources Research*, 12(5):1037–1046, 1976. URL https://doi.org/10.1029/WR012i005p01037. 32

Bethany Lindsay. The compasses of birds. *The Science Creative Quarterly*, 2006. URL http://www.scq.ubc.ca/the-compasses-of-birds/. Accessed 3rd May, 2019. 3

Tao Liu, Xing Zhang, Qingquan Li, and Zhixiang Fang. A visual-based approach for indoor radio map construction using smartphones. *Sensors*, 17 (8), 2017. URL https://doi.org/10.3390/s17081790. 299

M

Jiangang Ma, Yikang Yang, Hengnian Li, and Jisheng Li. Expressions for the autocorrelation function and power spectral density of BOC modulation based on convolution operation. *Mathematical Problems in Engineering*, 2020 (2063563), 2020. URL https://doi.org/10.1155/2020/2063563. 187

MathPages, The Sagnac Effect. URL https://www.mathpages.com/rr/s2-07/2-07.htm. Accessed 3rd August, 2020. 114

Muhammad Al Amin Amali Mazlan, M. H. Md Khir, Naufal M. Saad, and S. C. Dass. WiFi fingerprinting indoor positioning with multiple access points in a single base station using probabilistic method. *International Journal of Applied Engineering Research*, 12(6):1102–1113, 2017. URL https://www.ripublication.com/ijaer17/ijaerv12n6_45.pdf. Accessed 31st July, 2020. 299

Dalkhaa Munkhtsetseg. Geodetic network and geoid model of Mongolia. In *Proceedings, GSEM (Geospatial Solutions for Emergency Management) 2009*, Beijing, China, 2009. ISPRM. URL https://www.isprs.org/PROCEEDINGS/XXXVIII/7-C4/121_GSEM2009.pdf. Accessed 22nd September, 2021. 277

Michael John Muuss. The Story of the PING Program, undated. URL https://web.archive.org/web/20010107114600/https://ftp.arl.army.mil/~mike/ping.html. Accessed 17th September, 2021. 317

N

NASA Spinoff. NASA Brings Accuracy to World's Global Positioning Systems. URL https://spinoff.nasa.gov/Spinoff2019/ps_1.html. Accessed 12th June, 2020. 248, 258

Søren Vedel Nielsen. Wikimedia Commons, LC-130 take-off, Greenland, 2005. URL https://commons.wikimedia.org/wiki/File:LC130-Takeoff-Greenland.swn.jpg. © 2005 Søren Vedel Nielsen (GFDL, CC BY SA 2.5). Accessed 26th July, 2019. 277

James Robert Nockson. Wikimedia Commons, Ring laser gyroscope produced by Ukrainian "Arsenal" factory on display at MAKS-2011 airshow, 2011. URL https://commons.wikimedia.org/wiki/File:Ring_laser_gyroscope_at_MAKS-2011_airshow.jpg. © 2011 James Robert Nockson (CC BY-SA 3.0). Accessed 4th May, 2019. 104

O

Office of Aeronautical Satellite Systems, ATS Engineering Division, Japan Civil Aviation Bureau. Overview of MSAS, 2008. URL https://web.archive.org/web/20151106015542/http://www.unoosa.org/pdf/icg/2008/icg3/08-1.pdf. Accessed 28th May, 2020. 253

P

Mikko Parviainen. *Self-localization in Ad Hoc Indoor Acoustic Networks*. PhD thesis, Tampere University of Technology, 7th October 2016. URL https://researchportal.tuni.fi/en/publications/self-localization-in-ad-hoc-indoor-acoustic-networks. Accessed 9th September, 2021. 299

Physics Classroom, Kepler's Three Laws. URL https://www.physicsclassroom.com/class/circles/Lesson-4/Kepler-s-Three-Laws. Accessed 3rd August, 2020. 139

Markku Poutanen. *Satelliittipaikannus*. Ursa, Helsinki, 2017. ISBN 978-9-5259-8541-2. 172

Q

QZSS PS/IS. Performance Standard (PS-QZSS) and Interface Specification (IS-QZSS). Web site, Cabinet Office, National Space Policy Secretariat, 2020. URL https://qzss.go.jp/en/technical/ps-is-qzss/ps-is-qzss.html. Accessed 25th July, 2020. 255

R

Jean M. Rüeger. Refractive indices of light, infrared and radio waves in the atmosphere. *UNISURV report* S 68, School of Surveying and Spatial Information Systems, University of New South Wales, 2002. URL https://www.sage.unsw.edu.au/sites/sage/files/SAGE_collection/SpecialSeries/s68.pdf. 204

S

Daniel Porras Sánchez and César Pisonero Berges. The EGNOS SBAS message format explained. *ESA Navipedia*, 2006. URL https://gssc.esa.int/navipedia/index.php/The_EGNOS_SBAS_Message_Format_Explained. Accessed 18th June, 2020. 247

Jiri Segawa, Hiromi Fujimoto, and Shuhei Okubo, editors. *Proceedings, IAG International Symposium on Gravity, Geoid and Marine Geodesy (GraGeoMar96)*, International Association of Geodesy Symposia 117, Tokyo, Japan, 30th September – 5th October 1996. Springer-Verlag. 328, 335

Nagaraj C. Shivaramaiah and Andrew G. Dempster. The Galileo E5 AltBOC: understanding the signal structure. In *International Global Navigation Satellite Systems Society Symposium*, 2009. URL https://www.researchgate.net/publication/242169981_The_Galileo_E5_AltBOC_Understanding_the_Signal_Structure. Accessed 13th June, 2020. 186

Lina Sinjab. Replica Phoenician ship ends round-Africa journey. *BBC.com*, 2010. URL https://www.bbc.com/news/av/world-africa-11615613. Accessed 3rd May, 2019. 3

Sky Map Devs, Stardroid. URL https://github.com/sky-map-team/stardroid/blob/55b83c2aa46b071f62f3eb72996a5e7b6640ef7a/app/src/main/java/com/google/android/stardroid/control/AstronomerModelImpl.java. Accessed 22nd May, 2020. 285

Dava Sobel. *Longitude. The True Story of a Lone Genius who Solved the Greatest Scientific Problem of his Time*. Penguin, New York, 1995. 5

Joël Sommeria. Foucault and the rotation of the Earth. *Comptes Rendus Physique*, 18:520–525, 2017. URL https://doi.org/10.1016/j.crhy.2017.11.003. 104

SPS performance standard. Global Positioning System Standard Positioning Service Performance Standard, 2020. URL https://www.gps.gov/technical/ps/2020-SPS-performance-standard.pdf. Accessed 26th June, 2020. 260

Gilbert Strang and Kai Borre. *Linear Algebra, Geodesy, and GPS*. Wellesley — Cambridge Press, 1997. 15, 52, 165

Jaume Sanz Subirana, José Miguel Juan Zornoza, and Manuel Hernández-Pajares. Reference frames in GNSS. *ESA Navipedia*, 2011. URL https://gssc.esa.int/navipedia/index.php/Reference_Frames_in_GNSS. Accessed 16th June, 2020. 264

T

Byron D. Tapley and Bob E. Schutz. Estimation of unmodeled forces on a lunar satellite. *Celestial Mechanics*, 12:409–424, December 1975. URL https://doi.org/10.1007/BF01595388. 87

Peter J. G. Teunissen, Paul J. de Jonge, and Christiaan C. J. M. Tiberius. Performance of the LAMBDA method for fast GPS ambiguity resolution. *Navigation*, 44(3):373–383, 1997. URL https://doi.org/10.1002/j.2161-4296.1997.tb02355.x. 226

University of Texas, GRACE — Gravity Recovery and Climate Experiment. URL http://www2.csr.utexas.edu/grace/. Accessed 5th May, 2019. 279

U

USGS Open-File Report 2008-1089. Airborne Gravity Survey and Ground Gravity in Afghanistan: A Website for Distribution of Data. URL https://pubs.usgs.gov/of/2008/1089/Afghan_dataproc.html. Accessed 25th June, 2020. 278

W

Ming Wei and Klaus-Peter Schwarz. Comparison of different approaches to airborne gravimetry by strapdown INS/GPS. In Segawa et al. (1996), pages 155–162. URL https://rd.springer.com/chapter/10.1007/978-3-662-03482-8_23. Accessed 18th May, 2019. 274

Todd R. Weiss. Out-of-memory problem caused Mars rover's glitch. *Computerworld*, 2004. URL https://www.computerworld.com/article/2574759/out-of-memory-problem-caused-mars-rover-s-glitch.html. Accessed 30th April, 2020. 319

Greg Welch and Gary Bishop. The Kalman Filter. URL https://www.cs.unc.edu/~welch/kalman/. Accessed 4th May, 2019. 52

Greg Welch and Gary Bishop. Course 8: An introduction to the Kalman Filter. In *SIGGRAPH 2001, Los Angeles, August 12-17*. 2001. URL https://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_Slides_08.pdf. Accessed 4th May, 2019. 52

Paul Wessel, W. H. F. Smith, Remco Scharroo, J. Luis, and F. Wobbe. Generic Mapping Tools: improved version released. *EOS Trans. AGU*, 94(45): 409–410, 2013. URL http://dx.doi.org/10.1002/2013EO450001. ii

Wikimedia Commons, Barnacle geese. URL https://commons.wikimedia.org/wiki/File:MigrationFlock.jpg. © 2006 User:Thermos (CC BY-SA 2.5). Accessed 3rd May, 2019. 4

Wikimedia Commons, Gyroscope. URL https://commons.wikimedia.org/wiki/File:3D_Gyroscope-no_text.png. © 2006 User:LucasVB (PD). Accessed 3rd May, 2019. 104

Wikimedia Commons, Harrison's chronometer H5. URL https://commons.wikimedia.org/wiki/File:Harrison%27s_Chronometer_H5.JPG. © 2007 User:Racklever (GFDL). Accessed 3rd May, 2019. 4

Wikimedia Commons, Polynesian migration. URL https://commons.wikimedia.org/wiki/File:Polynesian_Migration.svg. © 2008 David Eccles (CC BY 4.0). Accessed 14th July, 2019. 3

Wikimedia Commons, Votive offering. URL https://commons.wikimedia.org/wiki/File:Votivskepp.JPG. © 2009 User:Svenboatbuilder (CC BY-SA 3.0 Unported). Accessed 21st August, 2020. 2

Wikipedia, Brownian bridge. URL https://en.wikipedia.org/wiki/Brownian_bridge. Accessed 10th May, 2020. 95

Wikipedia, Cochrane-Orcutt estimation. URL https://en.wikipedia.org/wiki/Cochrane%E2%80%93Orcutt_estimation. Accessed 30th June, 2020. 31

Wikipedia, Damping ratio. URL https://en.wikipedia.org/wiki/Damping_ratio. Accessed 4th May, 2019. 124

Wikipedia, Dandelin spheres. URL https://en.wikipedia.org/wiki/Dandelin_spheres. Accessed 27th June, 2020. 140

Wikipedia, Dead reckoning. URL https://en.wikipedia.org/wiki/Dead_reckoning. Accessed 3rd May, 2019. 3

Wikipedia, Domain Name System. URL https://en.wikipedia.org/wiki/Domain_Name_System. Accessed 7th May, 2019. 317

Wikipedia, Hall effect. URL https://en.wikipedia.org/wiki/Hall_effect. Accessed 29th May, 2020. 296

Wikipedia, Indoor positioning system. URL https://en.wikipedia.org/wiki/Indoor_positioning_system. Accessed 31st July, 2020. 299

Wikipedia, LC circuit. URL https://en.wikipedia.org/wiki/LC_circuit. Accessed 29th May, 2020. 296

Wikipedia, Magnus expansion. URL https://en.wikipedia.org/wiki/Magnus_expansion. Accessed 31st July, 2020. 66

Wikipedia, Orbital elements. URL https://en.wikipedia.org/wiki/Orbital_elements. Accessed 31st July, 2020. 140

Wikipedia, PIGA accelerometer. URL https://en.wikipedia.org/wiki/PIGA_accelerometer. Accessed 4th July, 2020. 112

Wikipedia, Polynesian navigation. URL https://en.wikipedia.org/wiki/Polynesian_navigation. Accessed 31st July, 2020. 3

Wikipedia, Receiver autonomous integrity monitoring. URL https://en.wikipedia.org/wiki/Receiver_autonomous_integrity_monitoring. Accessed 8th May, 2020. 241

Wikipedia, Max Schuler. URL https://en.wikipedia.org/wiki/Max_Schuler. Accessed 31st July, 2020. 125

Wikipedia, Sensor fusion. URL https://en.wikipedia.org/wiki/Sensor_fusion. Accessed 7th May, 2019. 281

Wikipedia, Superluminal motion. URL https://en.wikipedia.org/wiki/Superluminal_motion. Accessed 16th August, 2020. 204

Wikipedia, V-2 rocket. URL https://en.wikipedia.org/wiki/V2_rocket. Accessed 31st July, 2020. 6

Wikipedia, Vibrating structure gyroscope. URL https://en.wikipedia.org/wiki/Vibrating_structure_gyroscope. Accessed 2nd May, 2019. 297

Wikipedia, Woodbury matrix identity. URL https://en.wikipedia.org/wiki/Woodbury_matrix_identity. Accessed 31st July, 2020. 309

Wolfram Functions, $\int e^{bx} \cos cx \, dx$. URL https://functions.wolfram.com/ElementaryFunctions/Cos/21/01/02/04/01/01/0002/. Accessed 3rd April, 2020. 46

Wolfram Functions, $\int \exp bx^2 \cos cx \, dx$. URL https://functions.wolfram.com/ElementaryFunctions/Cos/21/01/02/04/01/05/0002/. Accessed 14th May, 2020. 47

Wolfram MathWorld, Statistical Correlation. URL https://mathworld.wolfram.com/StatisticalCorrelation.html. Accessed 4th May, 2019. 22

Jiuntsong Wu, Sienchong Wu, George A. Hajj, Willy I. Bertiger, and Stephen M. Lichten. Effects of antenna orientation on GPS carrier phase. *manuscripta geodaetica*, 18:91–98, 1993. 176

X

Xinhua. China's BeiDou officially goes global. *Silk Road News*, 2018. URL https://en.imsilkroad.com/p/125586.html. Accessed 22nd September, 2021. 268

Z

Jiexin Zhang, Alastair R. Beresford, and Ian Sheret. SensorID: Sensor calibration fingerprinting for smartphones. *IEEE Symposium on Security and Privacy (SP)*, 2019. URL https://doi.org/10.1109/SP.2019.00072. 286

Yinzhi Zhao, Peng Zhang, Jiming Guo, Xin Li, Jinling Wang, Fei Yang, and Xinzhe Wang. A new method of high-precision positioning for an indoor pseudolite without using the known point initialization. *Sensors*, 18(6), 2018. URL https://doi.org/10.3390/s18061977. 299

Index

ABCDEFGHIJKLMNOPQRSTUVWXZ