Methods of navigation

An introduction to technological navigation

Martin Vermeer

13th April 2020

1martin.vermeer@aalto.fi
Cover picture:

Navigation is no human invention. The arctic tern (*Sternula paradisaea*) flies every year from the Arctic to the Antarctic Sea and back. Egevang et al. (2010)

(Original English translation: Susanna Nordsten)
Preface

Fundamentals of navigation, stochastic processes, Kalman filter, inertial navigation and mechanization, the real time concept, GPS navigation, on-the-fly ambiguity resolution, use of GPS base stations, its data communication solutions and standards; navigation and geographic information systems; topical subjects.

Helsinki, 13th April 2020,

Martin Vermeer
Contents

Chapters

1. Fundamentals of navigation .......................... 1
2. Stochastic processes ................................. 11
3. The Kalman filter .................................... 35
4. Examples and applications of the Kalman filter .... 57
5. Inertial navigation .................................... 79
6. Navigation and satellite orbits ........................ 109
7. Measuring gravity from the air and from space .......... 135
8. GPS navigation and base stations ...................... 143
9. RTK measurements .................................. 153
11. Satellite-based augmentation systems (SBAS) ......... 185
12. Sensor fusion, sensors of opportunity ................ 195
A. Real time systems and networks .................... 207
B. Useful matric equations ............................. 221

Preface

List of Figures xi
List of Tables xiii
Acronyms xv

1. Fundamentals of navigation .......................... 1
# Contents

1.1 Introduction ............................................. 1  
1.2 History .................................................. 2  
1.3 A vehicle’s movements and reference frames .......... 5  
1.4 The role of technologies ................................. 10  
1.5 Basic concepts ........................................... 10  

2. Stochastic processes ........................................ 11  
2.1 Stochastic variables and processes ..................... 11  
2.2 The sample average ...................................... 13  
2.3 Covariance, correlation .................................. 17  
2.4 Auto- and crosscovariance of a stochastic process ... 19  
2.5 White noise and random walk ......................... 20  
2.6 “Coloured noise,” Gauss–Markov process .......... 23  
2.7 Power spectral density (PSD) .......................... 26  
2.8 Linear regression of time series ....................... 29  
Self-test questions .......................................... 32  

3. The Kalman filter ........................................... 35  
3.1 The state vector ......................................... 37  
3.2 The dynamic model ...................................... 39  
3.3 Example: satellite motion ............................... 40  
3.4 State propagation ........................................ 43  
3.5 Observational model .................................... 51  
3.6 Updating .................................................. 52  
3.7 The optimality of the update ......................... 54  

4. Examples and applications of the Kalman filter ..... 57  
4.1 Example 1: one-dimensional motion ................... 57  
4.2 Example 2: spinning wheel ............................ 60  
4.3 Example 3: parachute jumper .......................... 62  
4.4 Modelling of realistic statistical behaviour ........ 64  
4.5 The Kalman filter as sequential adjustment ....... 66  
4.6 Using the Kalman filter “from both ends” .......... 67  
Self-test questions .......................................... 74  
Exercise 4–1: Simple Kalman-filter example ........... 75  
Exercise 4–2: A bit more complicated Kalman-filter example . 76
5. Inertial navigation
   5.1 Principle ................................................. 79
   5.2 Parts of a inertial device ................................ 81
   5.3 Implementation ............................................. 87
   5.4 Inertial navigation in the system of the solid Earth . . . 90
   5.5 The stabilised platform ..................................... 94
   5.6 The gyro compass ........................................... 96
   5.7 The Schuler pendulum ....................................... 98
   5.8 Mechanization .............................................. 102
   5.9 On the Earth’s surface in two dimensions ................. 106
   5.10 Initialization of an inertial device ......................... 106

   6.1 The Kepler orbit ........................................... 109
   6.2 Use of Hill co-ordinates ................................... 118
   6.3 Transformation from inertial system to Hill system ....... 120
   6.4 Series expansion for a central force field ................ 121
   6.5 Equations of motion in the Hill system .................... 122
   6.6 Solving the Hill equations .................................. 123
   6.7 Another solution ............................................ 125
   6.8 The state transition matrix .................................. 126
   Self-test questions .............................................. 130
   Exercise 6–1: Kepler orbit ..................................... 131
   Exercise 6–2: Rendez-vous ...................................... 132

7. Measuring gravity from the air and from space ......... 135
   7.1 Vectorial airborne gravimetry .............................. 136
   7.2 Scalar airborne gravimetry .................................. 136
   7.3 Studying gravitation from space ............................ 137
   7.4 Using the Kalman filter in airborne gravimetry .......... 138
   7.5 Present state of airborne gravimetry ...................... 140

8. GPS navigation and base stations ....................... 143
   8.1 The GPS system ............................................ 143
   8.2 GPS satellites and signal structure ....................... 144
8.3 Pseudorange smoothing ........................................ 145
8.4 Differential navigation ........................................ 147
8.5 The RTCM standard ........................................... 149
8.6 RTCM-over-Internet (NTRIP protocol) .................... 151

9. RTK measurements ................................................. 153
  9.1 Introduction .................................................... 153
  9.2 GPS observations and unknowns ......................... 153
  9.3 The carrier phase observable .............................. 159
  9.4 The RTK method analyzed .................................. 162
  9.5 Clock and atmospheric errors .............................. 164
  9.6 Using double differences ................................... 164
  9.7 Fast ambiguity resolution .................................. 165
  9.8 A geometric analysis of RTK measurement .............. 169
  9.9 Network RTK .................................................. 176

  10.1 GPS modernization .......................................... 177
  10.2 The Russian GLONASS system ............................ 178
  10.3 The European Galileo system ............................ 180
  10.4 The Chinese BeiDou system .............................. 182

11. Satellite-based augmentation systems (SBAS) ............. 185
  11.1 Integrity and Safety of Life ............................... 185
  11.2 WAAS .......................................................... 186
  11.3 Intermezzo: LAAS, or GBAS .............................. 188
  11.4 EGNOS ........................................................ 190
  11.5 MSAS .......................................................... 190
  11.6 QZSS .......................................................... 192
  11.7 Internet-based Global DGPS .............................. 193

12. Sensor fusion, sensors of opportunity ....................... 195
  12.1 Case: Sky Map .............................................. 196
  12.2 Zero-velocity update ....................................... 197
  12.3 Integration of GNSS and IMU ............................. 200
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Propagation of state vector and state variance in the presence of noise</td>
<td>47</td>
</tr>
<tr>
<td>3.4</td>
<td>Optimality and error ellipses</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>The Kalman filter used forward and backward in time</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Random walk, solution in both directions</td>
<td>72</td>
</tr>
<tr>
<td>5.1</td>
<td>A gyroscope and a ring laser</td>
<td>82</td>
</tr>
<tr>
<td>5.2</td>
<td>A gyro wheel and its moments of inertia</td>
<td>85</td>
</tr>
<tr>
<td>5.3</td>
<td>Accelerometer principle</td>
<td>86</td>
</tr>
<tr>
<td>5.4</td>
<td>Pendulous accelerometer</td>
<td>87</td>
</tr>
<tr>
<td>5.5</td>
<td>Gyroscopic pendulous accelerometer</td>
<td>88</td>
</tr>
<tr>
<td>5.6</td>
<td>Sagnac interferometer</td>
<td>89</td>
</tr>
<tr>
<td>5.7</td>
<td>Stabilized-platform inertial device ST-124</td>
<td>90</td>
</tr>
<tr>
<td>5.8</td>
<td>The principle of the stabilised platform</td>
<td>95</td>
</tr>
<tr>
<td>5.9</td>
<td>Principle of the gyro compass</td>
<td>96</td>
</tr>
<tr>
<td>5.10</td>
<td>One-dimensional navigation carriage</td>
<td>100</td>
</tr>
<tr>
<td>5.11</td>
<td>Schuler response loop</td>
<td>101</td>
</tr>
<tr>
<td>5.12</td>
<td>Error propagation of one-dimensional mechanization</td>
<td>105</td>
</tr>
<tr>
<td>6.1</td>
<td>Kepler’s orbital elements</td>
<td>111</td>
</tr>
<tr>
<td>6.2</td>
<td>Orbital eccentricities</td>
<td>113</td>
</tr>
<tr>
<td>6.3</td>
<td>Deriving the R matrix</td>
<td>117</td>
</tr>
<tr>
<td>6.4</td>
<td>Hill co-ordinate frame</td>
<td>118</td>
</tr>
<tr>
<td>6.5</td>
<td>Libration</td>
<td>126</td>
</tr>
<tr>
<td>6.6</td>
<td>Linear drift</td>
<td>127</td>
</tr>
<tr>
<td>7.1</td>
<td>Ethiopian airborne gravity survey</td>
<td>140</td>
</tr>
<tr>
<td>7.2</td>
<td>Lockheed Hercules taking off from the NorthGRIP site</td>
<td>141</td>
</tr>
<tr>
<td>8.1</td>
<td>GPS constellation</td>
<td>144</td>
</tr>
<tr>
<td>8.2</td>
<td>The principle of phase modulation</td>
<td>145</td>
</tr>
<tr>
<td>8.3</td>
<td>Differential positioning</td>
<td>148</td>
</tr>
<tr>
<td>8.4</td>
<td>The geometry of computing the precision of differential positioning</td>
<td>148</td>
</tr>
<tr>
<td>9.1</td>
<td>Real-time kinematic GPS measurement</td>
<td>154</td>
</tr>
</tbody>
</table>
List of Tables

9.2 GPS code tracking ........................................ 161
9.3 Phase tracking by a Costas discriminator or loop .... 162
9.4 Ambiguity resolution ...................................... 167
9.5 Ambiguity resolution ...................................... 168
9.6 Geometry in the case of one base station .............. 170
9.7 Barycentric co-ordinates ................................ 174
9.8 The geometry of differential GPS ...................... 175
10.1 Galileo’s frequencies .................................... 181
11.1 WAAS ground segment .................................. 189
11.2 EGNOS ground segment ................................ 191
11.3 The concept of operation of a QZSS orbit .......... 192
12.1 A simple Kalman filter with and without zero-velocity updates .............................. 199
12.2 Attitude determination by GNSS ....................... 201
12.3 The principle of a MEMS rotation sensor ............ 204
A.1 Amplitude modulation and bandwidth ................ 208
A.2 “Frequency shift keying” (FSK) modulation .......... 211
A.3 An example of a protocol stack ....................... 212
A.4 Procedure call and stack allocation .................. 217

List of Tables

2.1 Probabilities of being outside the one-sigma, two-sigma, and three-sigma bounds, for one, two and three dimensions. 18
2.2 A summary of the properties of various stochastoc processes 26
5.1 Mechanisation simulation in one dimension, octave code 104
8.1 Carrier waves of the original GPS signal .............. 145
8.2 Pseudo-random codes of the GPS signal .............. 146
8.3 Positioning precision as a function of vector length and orbit error .......................... 148
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4</td>
<td>The message types of the RTCM SC-104 format</td>
<td>150</td>
</tr>
<tr>
<td>10.1</td>
<td>New Global Satellite Navigation Systems</td>
<td>177</td>
</tr>
<tr>
<td>11.1</td>
<td>The various approach categories according to ICAO</td>
<td>186</td>
</tr>
<tr>
<td>11.2</td>
<td>The WAAS satellites</td>
<td>187</td>
</tr>
<tr>
<td>11.3</td>
<td>EGNOS satellites</td>
<td>188</td>
</tr>
<tr>
<td>11.4</td>
<td>MSAS satellites</td>
<td>190</td>
</tr>
<tr>
<td>Acronyms</td>
<td>Page Numbers</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>AFSCN Air Force Satellite Control Network (GPS)</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>APPS Automatic Precise Positioning Service</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>AR(1) first-order autoregressive process</td>
<td>24, 31, 154</td>
<td></td>
</tr>
<tr>
<td>BeiDou Chinese global satellite navigation system</td>
<td>9, 177, 182, 183</td>
<td></td>
</tr>
<tr>
<td>BPSK Binary Phase-Shift Keying</td>
<td>143</td>
<td></td>
</tr>
<tr>
<td>FRS Fellow of the Royal Society (of London)</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>CDMA Code Division Multiple Access</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>CHAMP Challenging Minisatellite Payload</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>CPU central processing unit</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>CRC cyclic redundancy check</td>
<td>211</td>
<td></td>
</tr>
<tr>
<td>CSM Command and Service Module (Apollo)</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Decca Maritime hyperbolic radionavigation system, now obsolete</td>
<td>4, 202, 205</td>
<td></td>
</tr>
<tr>
<td>DGPS differential GPS</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>DSL Digital Subscriber Line</td>
<td>210</td>
<td></td>
</tr>
<tr>
<td>EDGE Enhanced Data Rates for GSM Evolution</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>EGM2008 Earth Gravity Model 2008</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>EGNOS European Geostationary Navigation Overlay Service, an SBAS for the European area</td>
<td>181, 190, 191</td>
<td></td>
</tr>
<tr>
<td>ESA European Space Agency</td>
<td>180, 190</td>
<td></td>
</tr>
<tr>
<td>ETRF European Terrestrial Reference Frame, usually with a year number. A realization of ETRS 7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>EUREF-89 also ETRF89. A realization of ETRS with definition epoch 1989.0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>FDMA Frequency Division Multiple Access</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>FFT fast Fourier transform</td>
<td>209</td>
<td></td>
</tr>
<tr>
<td>FSK Frequency-Shift Keying</td>
<td>210, 211</td>
<td></td>
</tr>
<tr>
<td>Galileo European global satellite navigation system</td>
<td>9, 177–182, 190</td>
<td></td>
</tr>
<tr>
<td>GBAS ground-based augmentation system</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>Acronyms</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>GDGPS</td>
<td>Global Differential GPS 147, 193</td>
<td></td>
</tr>
<tr>
<td>GJU</td>
<td>Galileo Joint Undertaking 180</td>
<td></td>
</tr>
<tr>
<td>GLONASS</td>
<td>Russian, globally available satellite navigation system 9, 177–180</td>
<td></td>
</tr>
<tr>
<td>GNSS</td>
<td>Global Navigation Satellite Systems, generic name for global satellite</td>
<td></td>
</tr>
<tr>
<td></td>
<td>positioning systems 5, 7, 9, 11, 133, 134, 136, 138, 149, 177, 186, 190,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200, 203–205, 211</td>
<td></td>
</tr>
<tr>
<td>GOCE</td>
<td>Gravity field and steady-state Ocean Circulation Explorer 136</td>
<td></td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Services 212</td>
<td></td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System 1, 5, 7, 9, 10, 109, 110, 131, 141–144, 147, 151,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>152, 154, 155, 164, 165, 167, 173, 174, 177–183, 185–188, 192, 193</td>
<td></td>
</tr>
<tr>
<td>GRACE</td>
<td>Gravity Recovery and Climate Experiment 136</td>
<td></td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications 212</td>
<td></td>
</tr>
<tr>
<td>GUI</td>
<td>graphical user interface 215</td>
<td></td>
</tr>
<tr>
<td>GUS</td>
<td>Ground Uplink Station (WAAS) 187, 189</td>
<td></td>
</tr>
<tr>
<td>HTTP</td>
<td>Hypertext Transfer Protocol 149, 211</td>
<td></td>
</tr>
<tr>
<td>ICAO</td>
<td>International Civil Aviation Organization 186</td>
<td></td>
</tr>
<tr>
<td>ICD</td>
<td>Interface Control Document 179</td>
<td></td>
</tr>
<tr>
<td>ICMP</td>
<td>Internet Control Message Protocol 213</td>
<td></td>
</tr>
<tr>
<td>IERS</td>
<td>International Earth Rotation and Reference Systems Service 7</td>
<td></td>
</tr>
<tr>
<td>IGS</td>
<td>inclined geostationary orbit 177, 183</td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>independent and identically distributed 15, 29</td>
<td></td>
</tr>
<tr>
<td>IMU</td>
<td>Inertial measurement unit 81, 199, 200, 204</td>
<td></td>
</tr>
<tr>
<td>INS</td>
<td>inertial navigation system 9</td>
<td></td>
</tr>
<tr>
<td>IONEX</td>
<td>Ionosphere Map Exchange Format 154</td>
<td></td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol 211, 213</td>
<td></td>
</tr>
<tr>
<td>ITRF</td>
<td>International Terrestrial Reference Frame, usually with a year number, a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>realization of the ITRS 7</td>
<td></td>
</tr>
<tr>
<td>JATO</td>
<td>Jet-Assisted Take-Off 139</td>
<td></td>
</tr>
<tr>
<td>JPL</td>
<td>Jet Propulsion Laboratory 193</td>
<td></td>
</tr>
<tr>
<td>KKJ</td>
<td>National Map Grid Co-ordinate System (obsolete) 7, 8</td>
<td></td>
</tr>
<tr>
<td>LAAS</td>
<td>Local-Area Augmentation System 188</td>
<td></td>
</tr>
<tr>
<td>LAMBDA</td>
<td>Least-squares Ambiguity Decorrelation Adjustment 165–167</td>
<td></td>
</tr>
<tr>
<td>LOR</td>
<td>lunar orbital rendez-vous 35</td>
<td></td>
</tr>
<tr>
<td>MCC</td>
<td>Master Control Centre (EGNOS) 190, 191</td>
<td></td>
</tr>
<tr>
<td>MCS</td>
<td>Master Control Station (GPS) 142</td>
<td></td>
</tr>
<tr>
<td>MEMS</td>
<td>Micro-Electronic Motion Sensor 86, 87, 203, 204</td>
<td></td>
</tr>
<tr>
<td>MSAS</td>
<td>Multi-functional Satellite Augmentation System 190, 192</td>
<td></td>
</tr>
</tbody>
</table>
Acronyms

NLES Navigation Land Earth Station (EGNOS) 190, 191
NTP Network Time Protocol 213
NTRIP Networked Transport of RTCM via Internet Protocol 149

OM Order of Merit 21
OS operating system 215

PLL Phase-locked loop 159
PRC pseudo-random code 144, 187
PRN pseudo-random noise 187, 188, 190, 192, 205
PSD power spectral density 26–28

QZSS Quasi-Zenith Satellite System, a Japanese SBAS 192

RAIM Receiver Autonomous Integrity Monitoring 186
RDS Radio Data System 174
RIMS Ranging and Integrity Monitoring Station (EGNOS) 190, 191
RINEX Receiver Independent Exchange Format 193
RTCM RTCM-SC104: Radio Technical Commission for Maritime Services
Special Committee SC-104, a set of standards for differential GNSS 147–149

RTK real-time kinematic positioning 7, 148, 149, 151, 167, 174

SBAS Satellite-Based Augmentation Systems 178, 181, 185, 186, 188, 190, 192

TCP Transmission Control Protocol 193, 211, 213
TDOA time difference of arrival 202
TOA time of arrival 203

UDP User Datagram Protocol 213
USB Universal Serial Bus 214
UTC Universal Time Co-ordinated 111, 152, 179

V-2 (Vergeltungswaffe 2, Revenge Weapon 2). German medium range guided
ballistic missile. Also A4 (“Aggregat 4”) 5, 6, 87

VHF Very High Frequency, 30 – 300 MHz 188
VOR VHF Omnidirectional Range, aviation navigation beacon 188

WAAS Wide-Area Augmentation System, an SBAS for the North American
area 185–188, 190

WGS84 World Geodetic System 1984, A set of reference frames created and
maintained by the U.S. Department of Defense 7

WLAN Wireless Local-Area Network 195, 202, 205
WMS Wide-area Reference Station (WAAS) 187, 189
WRS Wide-area Reference Station (WAAS) 187, 189
ZUPT zero-velocity update 197
1.1 Introduction

“Navigation” originates from the Latin word navis, ship. In other words, navigation is seafaring. Nowadays the meaning of navigation is approximately: finding and following a suitable route. This includes determining one’s own location during the journey.

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

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

1. in navigation the location data is needed immediately or at least after certain maximum delay. This is called the real-time requirement.
2. in navigation the position data are variable, time dependent.

Nowadays navigation is not limited to seafaring. Airplanes, 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 caused mainly by two modern technologies: GPS (Global Positioning System) and inertial navigation. Also data processing technologies have developed: specifically the recursive linear filter or Kalman filter should be mentioned here. Finally, sensor technologies have produced a host of small and inexpensive digital sensors that are revolutionizing everyday navigation.
1.2 History

1.2.1 Old history

Humans have always been discovering the world around them and travelled often long distances. Navigation has always been a necessity. Before the existence of modern 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, i.e., constructing a pre-existing description of the world in the form of a map. The journey is then planned and executed by comparing all the time the actual place with the target place according to the travel plan.

In case that the landmarks are missing, for example in shipping, one can use a method called dead reckoning, Wikipedia, Dead reckoning. Here it is estimated where one should be based on travel direction and speed. The sources of error in this method apparently are sea currents (in aviation, winds) and more commonly that the forecast weakens with time.

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

See also Diamond (1999).

Navigation with the help of landmarks, but also using hi-tech, is used by, e.g., cruise missiles: they fly by the contour lines of a digital terrain model they have stored in their memories.

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.

---

1“Navigare necesse est.”

2At 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
Direction is the easiest. At night, the North Star (Polaris) shows the North direction. In the daytime, the Sun can be used, although in a more complicated way. On a cloudy day the polarization of sky light can be used to help locate the Sun.

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 magnetic compass made finding North easier under all conditions. Yet the magnetic North is not the geographical North, and the difference between them depends on position and changes with time.

- Latitude is easy to get: it is the height of the celestial pole above the horizon. In the daytime, the Sun may be used: at upper culmination — solar noon — on the Northern hemisphere, the elevation of the Sun over the Southern horizon equals the sum of its declination — from astronomical tables — and co-latitude, i.e., 90° minus latitude.

- Longitude is the problem: its determination requires the use of an accurate time standard (chronometer). See Sobel (1995). Also, astronomical methods like using the moons of Jupiter as a “clock” have been studied. In the 20th century the distribution of time signals by radio became common.

In the 20th century radio technological methods came into use. The most well known is probably **Decca**, which was based on hyperbolic positioning. One “master”-station and two or more “slave”-stations transmit synchronized time signals modulated onto the radio waves. The on-board receiver measures the travel time difference between the
waves received from master and slave. On the nautical chart is marked the set of points of the same difference in travel time, as a coloured curve, a hyperbola. Every slave station forms with the master a bundle of hyperboles drawn in its own colour. The intersection point of two hyperboles gives the position of the ship. So, at least two slaves are needed in addition to the master station.

Modern satellite positioning methods, like Transit (no longer in use) and GPS (and other global navigation satellite systems) are based on a three-dimensional counterpart 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 airplanes 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 very primitive. The missile was launched vertically into the air, where it turned to the right direction with the help of its gyroscope platform, and accelerated until reaching a pre-determined velocity, at which point the propellant supply was shut off (“Brennschluss”). Physically the turning 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 motor.\footnote{See Wikipedia, V-2 rocket. In fact these were dual rudders: the part sticking into the exhaust stream, the “jet rudder”, consisted of graphite and burned up quickly. But by then, the rocket was up to speed and the external “air rudders” took over.}

Nowadays complete inertial navigation is used in airplanes and spacecraft, as are other computer based technologies such as satellite positioning, i.e., GNSS (Global Navigation Satellite Systems).

1.3 A vehicle’s movements and reference frames

A moving vehicle has two co-ordinate reference frames relevant to it:

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

2. the external reference frame: this can be a local or regional terrestrial frame, with X, Y being map projection co-ordinates –
Figure 1.4. German rocket weapon V-2. Photo U.S. Air Force.
the X axis typically pointing North and the Y axis pointing east – and Z again height defined in some local height system from an agreed reference surface, taken as a co-ordinate along an axis pointing to the local zenith. This is a quasi-Cartesian reference frame often used in aerial mapping.

3. Alternatively, a geocentric external reference frame.

Geocentricity means above all, that

- the origin is in the Earth’s centre of mass (or very close to it);
- The Z axis points in the direction of the Earth’s rotation axis;
- The X axis lies in the plane of the Greenwich meridian and points to the intersection point of equator and Greenwich meridian, and the Y axis is perpendicular to both.

This is a terrestrial co-ordinate frame, attached to the solid Earth and “co-rotating”). These are co-ordinates most naturally produced by GNSS positioning equipment.

1.3.1 Geocentric reference frame

As such, GNSS systems produce coordinates in the WGS84 reference frame, the geocentric frame originally used by the GPS system. It is maintained by the US Department of Defense, and there have been a number of versions.

The international geodetic research community has provided its own geocentric frames, through a service called IERS, the (International Earth Rotation Service). The frames have names like ITRF(yy) and ETRF(yy): ITRF = International Terrestrial Reference Frame, ETRF = European Terrestrial Reference Frame. yy is the year of publication.

Nowadays these frames agree with WGS84 on the centimetre level.

The ETRF systems have been designed in such a way, that point coordinates on the Eurasian continental plate do not change, so the system moves with the plate. In many countries, as well as in scientific circles, the ETRF-89, also known as EUREF-89, is used. Its moment of definition (epoch) is the beginning of 1989.

1.3.2 Non-geocentric reference frames

When we want to work in a local or national, non-geocentric system, like KKJ (Map Grid Co-ordinate System) in Finland, things get more difficult if we want to retain the accuracy obtained from the GPS system.
Real-time kinematic (RTK) positioning, essentially a navigation technique, is a widely used data collection method for mapping surveying work and spatial data bases. If accuracy demands are on the level of 1 – 2 meters, even code based differential GPS is suitable.

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

- Measure several points known in KKJ on the edge of the measurement area, and feed in their KKJ coordinates;
- Measure the points to be measured in the area;
- Return to the known point to check if there has been some jump of the total value in the phase of the carrier wave ("cycle slip").
- The device calculates itself the transformation formula (Helmert transformation in space) with the help of the known points and transforms all regular measuring points to KKJ with it.

The disadvantage of this system is, that the original accuracy of the measurement data drops irreversibly in KKJ almost every time to the weakest local accuracy. If this is acceptable, it is a practicable solution in local surveying.

This is how the “navigation solution” can be used in mapping surveying. As benefit, there is no post-measurement office work remaining. The collected data — which can be quite voluminous, millions of points — goes directly into a spatial data base after a limited amount of manual work, such as type encoding according to a catalogue.

1.3.3 Vehicle attitude

The attitude of a vehicle can be described relative to three axes. The motion about the direction of travel is called roll, that about the vertical axis yaw, and that about the horizontal (left-right) axis pitch. \( \kappa, \varphi, \omega \).

1.3.4 Co-ordinate transformation

Between the body frame and the external frame 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.
A vehicle’s movements and reference frames

Figure 1.5. The attitude angles of a vehicle.

1.3.5 The vertical reference

When GPS — or any other positioning solution that doesn’t directly depend on the Earth’s gravity field, like also inertial navigation (INS) or GPS-INS integration — is used in height determination, there arises the problem that also the heights are geocentric, in other words, they are elevation above the geocentric, mathematically defined reference ellipsoid. Traditional elevations on the other hand are above “mean sea level,” more precisely, the geoid. See figure 1.6.

![Figure 1.6. Height or elevation systems.](image)
1.4 The role of technologies

Technologies suitable for both navigation and geodetic position finding are:

1. **GPS**, the Global Positioning System; today we use the term **GNSS**, Global Navigation Satellite Systems, to which also belong the Russian **GLONASS** system, Compass/BeiDou (China) and the upcoming **Galileo** (Europe).

2. inertial navigation

3. Kalman filtering

4. Automatic guidance, mostly for missiles and launch vehicles, but also for aircraft and experimentally for road vehicles.

1.5 Basic concepts

- Stochastic processes
- Linear estimation
- Kalman filtering, dynamic model, observation model, statistical model
- inertial navigation, mechanisation
- satellite orbit
- Use of **GPS** in navigation, including the role of networks of base stations
- The new, post-**GPS** satellite navigation systems
- Satellite-based augmentation systems
- Sensor fusion and sensors of opportunity.

In the following, these basic concepts will be discussed systematically.
### Stochastic processes

#### 2.1 Stochastic variables and processes


An often used way to describe quantities that change in time and are uncertain, is that of **stochastic processes**.

#### 2.1.1 Stochastic variables

A **stochastic variable** is defined as follows:

*A stochastic variable* \( x \) is a series of realizations \( x_1, x_2, x_3, \ldots, x_i, \ldots \), or \( x_i, i = 1, 2, \ldots \). Every realization value has a certain probability of happening. If we repeat the realizations, or “throws,” often enough, the long-term percentage of that value happening tends towards this probability value.

The traditional notation for stochasticity is an underscore.

The value set of the stochastic variable can be a **discrete** or a **continuous** set. Examples of discrete stochastic variables are

- Dice throwing. Each throw is one realization. In this case \( x_i \in \{1, 2, 3, 4, 5, 6\} \).
- Throwing coins. \( x_i \in \{0, 1\} \), 0 = heads, 1 = tails.

A **measurement** is a stochastic, usually real-valued, variable. Examples:

- A *measured distance* is a real-valued, i.e., continuous, stochastic variable \( s \). Realizations are \( s_i \in \mathbb{R}, i = 1, 2, \ldots \).
- A vector measurement produced by GNSS from a point A to a point B is a **stochastic vector variable** \( \mathbf{x} \). Every realization consists of three components and belongs to the three-dimensional vector space: \( x_i \in \mathbb{R}^3 \).
Stochastic processes

- Measured horizontal angle $\alpha$, realizations $\alpha_i \in [0, 2\pi)$.

For example angle measurement with a theodolite: the domain is 
\[ \{ \alpha \in \mathbb{R} \mid 0 \leq \alpha < 2\pi \} \], a subset of real values.\(^1\)

In these cases of continuous stochastic variables we speak of probability density and not the probability of a certain realization value $x$. The probability of a realization falling within a certain interval $(x_1, x_2)$ is computed as the integral

\[ p = \int_{x_1}^{x_2} p(x) \, dx. \]

if we assume that the probability distribution is normal, i.e., the Gaussian bell curve, figure 2.1, then the equation for it is

\[ p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right), \quad (2.1) \]

where $\sigma$ is the mean error (standard deviation) of the distribution, and $\mu$ its expectancy.

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 realization of the stochastic variable (“dice throw”) is a function.

The argument is usually the time $t$, but can also be for example location $(\varphi, \lambda)$ on the Earth’s surface.

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

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

Examples:

- The temperature of the experimental device $T(t)$ as the function of time $t$ Different realizations $T_i(t)$ are obtained by repeating the test: $i = 1, 2, \ldots$.

  In real life repeating the test can be difficult or impossible.

\(^1\)More precisely: a subset of rational values \( \{ \mathbb{Q} \mid 0 \leq \alpha < 2\pi \} \)
The temperature from Kaisaniemi weather station, Helsinki $T_{Kais}(t)$. History can not be precisely repeated: from this stochastic process we only have one realization $T_{Kais}^1(t)$, the historical time series of Kaisaniemi. Other realizations $T_{Kais}^i(t)$, $i = 2, 3, \ldots$ exist only as theoretical constructs without any hope of observing them.

It is often assumed, that the result will be same if the same process shifted in time is used as realization. For example

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

in which $\Delta t$ is the time shift, the choice of which depends on the subject of study. This hypothesis is called the ergodicity hypothesis.

### 2.2 The sample average

#### 2.2.1 General

One often encounters situation where some quantity $x$ is measured several times and we have available realizations of the stochastic measurement quantity $x$.

All realizations of course differ in different ways from the “real” value $x$ — which we don’t even know. Estimation is computing an “as good as possible” estimate for $x$ from the realizations of the stochastic measurement quantity. The “real value” $x$ is not known: if it were, we wouldn’t have to measure now would we?

The estimate is itself a realization of the estimator: the estimator itself is a stochastic quantity, its realizations being estimates.

On the stochastic quantity’s value set (codomain) $x$ is defined a probability density function $p(x)$, that describes the probability, that the value of one realization happens to be $x$. Often, it is assumed that $p(x)$ is so called Gaussian curve or normal distribution, the “bell curve,” equation 2.1.

The results presented below do not depend on the assumption of a gaussian distribution if not mentioned otherwise.

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

$$\int_{-\infty}^{+\infty} p(x) \, dx = 1.$$
The definition of the expected value or *expectancy* $E$ is

$$E(x) \overset{\text{def}}{=} \int_{-\infty}^{+\infty} xp(x) \, dx.$$  

The expectancy is not the same as average: expectancy is a theoretical concept, while the average can be calculated from measurements. Another connection is that the average of the first $n$ realisations of variable $x$,

$$\bar{x}^{(n)} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i,$$

(2.2)

is probably the closer to $E(x)$, the bigger $n$ is. This law based on experience is called the (empirical) *law of big numbers*.

In equation 2.2, the first group of $n$ realizations is called the *sample*, and $\bar{x}^{(n)}$ is the sample average.

Now that the expected value has been defined, we define next the *variance*:

$$\text{Var} \{ \bar{x} \} \overset{\text{def}}{=} E \left\{ (\bar{x} - E(\bar{x}))^2 \right\}.$$  

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

$$\sigma^2 = \text{Var} \{ \bar{x} \}.$$  

The variance, like the expected value, is a theoretical value that can not be calculated. Instead it is *estimated* from the sample $x_i$, $i = 1, \ldots, n$. If the sample average $\bar{x}$ already exists, and assuming that the realizations $x_i$ are statistically independent from each other and all have the same
The sample average mean error\footnote{This is called the \textit{i.i.d.} assumption, “independent and identically distributed.”} $\sigma$, follows the estimate of the variance $\sigma^2$ as follows:

\[ \hat{\sigma}^2 \overset{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}^n)^2. \]

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

\[ \bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i, \]

in which $x_i$ is a stochastic quantity the successive realizations of which are simply $x_i, x_{i+n}, x_{i+2n}, \ldots$ (a “fork variate”). For illustration, assume that the sample size is $n = 10$. Then, the sample mean as a stochastic quantity is $\bar{x}^{(10)}$, with successive realizations

\[ \bar{x}_1^{(10)} = \frac{1}{10} \sum_{i=1}^{10} x_i, \quad \bar{x}_2^{(10)} = \frac{1}{10} \sum_{i=11}^{20} x_i, \quad \bar{x}_3^{(10)} = \frac{1}{10} \sum_{i=21}^{30} x_i, \quad \ldots \]

It is intuitively clear — and assumed without proof — that

\[ \mathbb{E}\{x_i\} = \mathbb{E}\{\bar{x}\}, \quad i \in \{1, \ldots, n\}. \]

The expected value of the quantity $\bar{x}^{(n)}$ is

\[ \mathbb{E}\{\bar{x}^{(n)}\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{x_i\} = \mathbb{E}\{\bar{x}\}, \]

which is the same as the expectancy of $\bar{x}$; that kind of estimator is called \textit{unbiased}.

Its variance is estimated with the equation

\[ \hat{\text{Var}}\{\bar{x}^{(n)}\} = \frac{1}{n (n-1)} \sum_{i=1}^{n} (x_i - \bar{x}^{(n)})^2 = \frac{1}{n} \hat{\sigma}^2. \]

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

This all is presented here without proofs, which can be found in statistics textbooks.
2.2.2 Optimality of the average value

From all unbiased estimators of $x$ based on sample $x_i$, $i = 1, \ldots, n$:

$$\hat{x} = \sum_{i=1}^{n} a_i x_i, \quad \sum_{i=1}^{n} a_i = 1,$$

the average

$$\hat{x} \overset{\text{def}}{=} \bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{(2.3)}$$

minimizes the variance of $\hat{x}$. The variance is calculated as follows:

$$\operatorname{Var}\{\hat{x}\} = \sum_{i=1}^{n} a_i^2 \operatorname{Var}\{x_i\} = \sigma^2 \sum_{i=1}^{n} a_i^2,$$

assuming that $x_i$ don’t correlate with each other, and that $\operatorname{Var}\{x_i\} = \sigma^2$.

Now, minimizing the expression

$$\sum_{i=1}^{n} a_i^2$$

under the additional constraint

$$\sum_{i=1}^{n} a_i = 1$$

yields

$$a_i = \frac{1}{n}.$$

From which the claim follows.

2.2.3 Computing the sample average one step at a time

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

$$\bar{x}^{(n+1)} = \frac{n}{n+1} \bar{x}^{(n)} + \frac{1}{n+1} x_{n+1},$$

$$\operatorname{Var}\{\bar{x}^{(n+1)}\} = \left(\frac{n}{n+1}\right)^2 \operatorname{Var}\{\bar{x}^{(n)}\} + \left(\frac{1}{n+1}\right)^2 \sigma^2.$$

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 $\bar{x}^{(n)}$ “on the fly,” while observations are being collected, before all observations are in. This is precisely the advantage of using the Kalman filter.
2.3 Covariance, correlation

When given are two stochastic quantities $x$ and $y$, the covariance between them can be calculated as

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

The covariance describes how strongly the random variations of $x$ and $y$ behave similarly.

Besides covariance, correlation is defined as:

$$\text{Corr}\{x, y\} \overset{\text{def}}{=} \frac{\text{Cov}\{x, y\}}{\sqrt{\text{Var}\{x\}\text{Var}\{y\}}}.$$  \hspace{1cm} (2.4)

Correlation can never be more than 1.0 — or less than $-1.0$. Often the correlation is expressed as a percentage, 100% being the same as 1.0.

In case of two stochastic processes we often draw an error ellipse (figure 2.2). Compare this picture with the earlier picture of the bell curve, figure 2.1. There the expectancy is marked as $E\{x\}$ (in the middle) and mean error $\pm \sigma$. In the error ellipse figure 2.2 the central point represents the expected values of $\bar{x}$ and $\bar{y}$; the ellipse itself corresponds to the mean error values $\pm \sigma$.

\[\begin{align*}
\text{Define normalized variates:} \\
\xi &\overset{\text{def}}{=} \frac{x}{\sqrt{\text{Var}\{x\}}}, \\
\eta &\overset{\text{def}}{=} \frac{y}{\sqrt{\text{Var}\{y\}}}.
\end{align*}\]

Then, because of linearity

$$\text{Cov}\{\xi, \eta\} = \frac{\text{Cov}\{x, y\}}{\sqrt{\text{Var}\{x\}\text{Var}\{y\}}} = \text{Corr}\{x, y\}.$$  \hspace{1cm} (2.4)

These variances are positive:

\[\begin{align*}
0 \leq \text{Var}\{\xi + \eta\} &= \text{Var}\{\xi\} + \text{Var}\{\eta\} + 2\text{Cov}\{\xi, \eta\}, \\
0 \leq \text{Var}\{\xi - \eta\} &= \text{Var}\{\xi\} + \text{Var}\{\eta\} - 2\text{Cov}\{\xi, \eta\};
\end{align*}\]

when also

$$\text{Var}\{\xi\} = \frac{\text{Var}\{x\}}{\left(\sqrt{\text{Var}\{x\}}\right)^2} = 1$$

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

$$-1 \leq \text{Cov}\{\xi, \eta\} = \text{Corr}\{x, y\} \leq 1.$$
One can say that the measurement values will, with a certain probability, fall inside or outside the ellipse: this is what gave the error ellipse its name. However, this probability is not the same for a one-dimensional interval \([\mu - \sigma, \mu + \sigma]\) as for a two-dimensional ellipse, or for a three-dimensional error ellipsoid for that matter. See table 2.1.

If the ellipse is intersected by the line \(z\), the linear combination of \(x\) and \(y\) is obtained:

\[
z = x \cos \theta + y \sin \theta,
\]

the point pair of which on the tangents to the ellipse represents the mean error of the quantity \(z\):

\[
\text{Var}\{z\} = \mathbb{E}\{(z - \mathbb{E}\{z\})^2\}
\]

<table>
<thead>
<tr>
<th></th>
<th>(\sigma)</th>
<th>(2\sigma)</th>
<th>(3\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One dimension (interval)</td>
<td>31.7%</td>
<td>4.6%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Two dimensions (ellipse)</td>
<td>60.6%</td>
<td>13.5%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Three dimensions (ellipsoid)</td>
<td>80.1%</td>
<td>26.1%</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

4In matrix notation we can write

\[
z = \begin{bmatrix} \cos \theta & \sin \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}
\]
\[ = E \left\{ \left( \cos \theta (x - E(x)) + \sin \theta (y - E(y)) \right)^2 \right\} = \]
\[ = \cos^2 \theta \text{Var}\{x\} + 2 \sin \theta \cos \theta \text{Cov}\{x, y\} + \sin^2 \theta \text{Var}\{y\}, \]

and from this \( \sigma_z = \sqrt{\text{Var}\{z\}} \). As a function of the direction angle \( \theta \) The mean error \( \sigma_z \) has two extremal values, \( \sigma_{\text{min}} \) and \( \sigma_{\text{max}} \), see figure 2.2.

If \( \sigma_{\text{min}} = \sigma_{\text{max}} \), or the ellipse is oriented along the axes of the extremal values \( \sigma_{\text{min}} \) and \( \sigma_{\text{max}} \) the correlation between \( x \) and \( y \) disappears. In that case they really are independent from each other and knowing the real value of one doesn’t help in estimating the other.

If the correlation doesn’t vanish, the knowledge – or a good estimate – of \( x \)’s real value helps to estimate the \( y \) better. The method is called regression.

### 2.4 Auto- and crosscovariance of a stochastic process

If instead of a stochastic quantity there is a stochastic process \( x(t) \), one can calculate the derived function called the autocovariance as follows:

\[ A_x(t, t') \equiv \text{Cov}\{x(t), x(t')\}. \]

In case of so called stationary processes, processes the properties of which do not depend on absolute time but stay the same over the course of time, one may write

\[ A_x(t, t') = A_x(t, t + \Delta t) = \text{Cov}\{x(t), x(t + \Delta t)\} \equiv A_x(\Delta t), \]

with \( t' = t + \Delta t \), independent of the value of \( t \).

If there are two different stochastic processes \( x(t) \) and \( y(t) \), one can obtain the derived function called cross-covariance:

\[ C_{xy}(t, t') \equiv \text{Cov}\{x(t), y(t')\}. \]

ja

\[ \text{Var}\left\{ \begin{bmatrix} x \\ y \end{bmatrix} \right\} = \begin{bmatrix} \text{Var}\{x\} & \text{Cov}\{x, y\} \\ \text{Cov}\{x, y\} & \text{Var}\{y\} \end{bmatrix}. \]

This implies

\[ \text{Var}\{z\} = \begin{bmatrix} \cos \theta & \sin \theta \end{bmatrix} \begin{bmatrix} \text{Var}\{x\} & \text{Cov}\{x, y\} \\ \text{Cov}\{x, y\} & \text{Var}\{y\} \end{bmatrix} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}, \]

i.e., the same result. This illustrates the law of propagation of variances.
and again in the case of stationary processes

\[ C_{xy}(\Delta t) \overset{\text{def}}{=} \text{Cov}\{x(t), y(t + \Delta t)\}. \]

Often, the term cross-covariance is used simply to denote

\[ C_{xy} \overset{\text{def}}{=} C_{xy}(0) = \text{Cov}\{x(t), y(t)\}. \]

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

\[ \text{Corr}_x(\Delta t) \overset{\text{def}}{=} \frac{A_x(t, t')}{\sqrt{A_x(t, t)A_x(t', t')}} = \frac{A_x(\Delta t)}{A_x(0)}, \] \hspace{1cm} (2.5)

and cross correlation

\[ \text{Corr}_{xy}(\Delta t) \overset{\text{def}}{=} \frac{C_{xy}(\Delta t)}{\sqrt{A_x(0)A_y(0)}}, \]

all the time assuming stationarity.

It is seen from equation 2.5, that, if \( \Delta t = 0 \), the autocorrelation is 1. Otherwise it is always between \(-1\) and \(+1\), as was already show in connection with equation 2.4. Also the cross correlation is always in the interval \([-1, +1]\).

### 2.5 White noise and random walk

Noise may be considered as a stochastic process with an expected value of 0:

\[ E\{n(t)\} = 0. \]

White noise is noise that consists uniformly of all possible frequencies.\(^5\) The mathematical way of describing this is saying that the autocovariance

\[ A_n(\Delta t) = 0 \]

if \( \Delta t \neq 0 \). In other words, the process values \( n(t_1) \) and \( n(t_2) \) do not correlate at all, no matter how close \( t_2 - t_1 \) is to zero.

Nevertheless

\[ A_n(0) = \infty. \]

---

\(^5\)The name is based on an analogy with white light, which we know thanks to Newton (Davidson and Tchourioukanov) to consist of all the frequencies of visible light.
Furthermore it holds that
\[ \int_{-\infty}^{+\infty} A_n(\tau) \, d\tau = Q. \]

Perhaps you may want to reflect on the above equations for a while. Here 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 \( Q \! \)!

Such a function does not actually exist. There is however a mathematical device, the delta function or distribution, named after the quantum physicist Paul Dirac:

\[ A_n(\tau) = Q \delta(\tau). \quad (2.6) \]

Intuitively we can imagine how such a “function” may be built, figure 2.3.

First the following block function is defined:

\[ \delta_\epsilon(\tau) = \begin{cases} 0 & \text{if } \tau > \frac{1}{2} \epsilon \text{ or } \tau < -\frac{1}{2} \epsilon \\ 1/\epsilon & \text{if } -\frac{1}{2} \epsilon \leq \tau \leq \frac{1}{2} \epsilon \end{cases} \]

Clearly the integral of this function

\[ \int_{-\infty}^{+\infty} \delta_\epsilon(\tau) \, d\tau = 1 \quad (2.7) \]

Paul Adrien Marie Dirac OM FRS (1902-1984) was a leading British theoretical physicist and quantum theorist.
and \( \delta_\varepsilon(\tau) = 0 \) if |\( \tau \)| is large enough.

Let \( \varepsilon \to 0 \). In this limit \( \delta_\varepsilon(0) \to \infty \), and for every value \( \tau \neq 0 \) there is always a bounding value \( \varepsilon \) for which it holds that |\( \tau \)| \( > \varepsilon \) \( \implies \delta_\varepsilon(\tau) = 0 \).

The handling rule of distributions is simply, that first we integrate, and in the result obtained we let \( \varepsilon \to 0 \).

### 2.5.1 Autocovariance of random walk

“Random walk” is obtained if white noise is integrated over time. Let the autocovariance of the noise \( n \) be

\[
A_n(\Delta t) = Q \delta(\Delta t).
\]

Then we integrate this function (a random walk or Wiener\(^7\) process):

\[
w(t) = \int_{t_0}^{t} n(\tau) \, d\tau.
\]  

(2.8)

Note that

\[
E\{w(t)\} = \int_{t_0}^{t} E\{n(\tau)\} \, d\tau = 0.
\]

The autocovariance function is obtained as

\[
A_w(t, t') = E\left\{ (w(t) - E\{w(t)\})(w(t') - E\{w(t')\}) \right\} =
E\{w(t)w(t')\} =
E\left\{ \int_{t_0}^{t} n(\tau) \, d\tau \int_{t_0}^{t'} n(\tau') \, d\tau' \right\} =
= \int_{t_0}^{t} \left( \int_{t_0}^{t'} E\{n(\tau)n(\tau')\} \, d\tau' \right) \, d\tau.
\]

Here

\[
\int_{t_0}^{t'} E\{n(\tau)n(\tau')\} \, d\tau' = \int_{t_0}^{t'} A_n(\tau - \tau') \, d\tau' =
Q \int_{t_0}^{t'} \delta(\tau - \tau') \, d\tau' = \begin{cases} Q & \text{if } t' > \tau, \\ 0 & \text{if } t' \leq \tau. \end{cases}
\]

It follows that

\[
A_w(t, t') = Q \int_{t_0}^{t} \left( \int_{t_0}^{t'} \delta(\tau - \tau') \, d\tau' \right) \, d\tau =
= Q(t' - t_0) + 0(t - t') = Q(t' - t_0). \quad (2.9)
\]

\(^7\)Norbert Wiener (1894–1964) was a Jewish American mathematician and philosopher and the founder of cybernetics.
Here it has been assumed that the autocovariance of the noise function \( n \) is stationary, in other words, that \( Q \) is a constant. This can be generalized to the case where \( Q(t) \) is a function of time:

\[
A_w(t, t') = \int_{t_0}^{t'} Q(\tau) \, d\tau. \tag{2.10}
\]

In both equations (2.9, 2.10) it is assumed that \( t' \leq t \).

### 2.6 “Coloured noise,” Gauss–Markov process

Let us study the simple differential equation in time

\[
\frac{d\chi}{dt} = -k\chi + n, \tag{2.11}
\]

where \( n \) is white noise, of which the autocovariance function is \( Q(\Delta t) \), and \( k \) is a constant. The solution of this differential equation is

\[
\chi(t) = e^{-kt} \left\{ \chi(t_0) e^{kt_0} + \int_{t_0}^{t} n(\tau) e^{k\tau} \, d\tau \right\}.
\]

The solution satisfies also the initial condition.

If we assume that the initial value \( \chi(t_0) \) is errorless and that the autocovariance function of \( n \) is

\[
A_n(t, t') = Q(t) \delta(t - t'),
\]

we obtain the autocovariance function of \( \chi \):

\[
A_\chi(t, t') = e^{-k(t+t')} E \left\{ \int_{t_0}^{t} n(\tau') e^{k\tau'} \, d\tau' \int_{t_0}^{t} n(\tau) e^{k\tau} \, d\tau \right\} = e^{-k(t+t')} \int_{t_0}^{t} e^{k\tau'} \int_{t_0}^{t} E \{ n(\tau') n(\tau) \} e^{k\tau} \, d\tau' \, d\tau.
\]

Here,

\[
\int_{t_0}^{t} E \{ n(\tau') n(\tau) \} e^{k\tau} \, d\tau = \int_{t_0}^{t} A_n(\tau - \tau') e^{k\tau} \, d\tau = \begin{cases} Q e^{k\tau'} & \text{if } t > \tau' \\ 0 & \text{if } t < \tau' \end{cases}
\]

So, assuming that \( t' > t \):

\[
A_\chi(t, t') = Q e^{-k(t+t')} \left( \int_{t_0}^{t} e^{2k\tau'} \, d\tau' + \int_{t}^{t'} 0 \, d\tau' \right) = \frac{Q}{2k} e^{-k(t+t')} \left( e^{2kt} - e^{2kt_0} \right).
\]
In the case in which \( t > t' \), this gives:

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

In both cases, we obtain

\[
A_x(t, t') = \frac{Q}{2k} e^{-k|t-t'|} - e^{-k(t+t'-2t_0)}.
\]

In this case we talk about *coloured noise*\(^8\) and the process thus obtained is called a *Gauss–Markov process*, also a (first-order) autoregressive (AR(1)) process\(^9\).

Let us also write

\[
Q \overset{\text{def}}{=} qk^2.
\]

Then the surface area under the \( A_x(t-t') \) curve is

\[
\int_{-\infty}^{+\infty} A_x(\tau) \, d\tau = \frac{qk}{2} \cdot 2 \int_{0}^{\infty} e^{k\tau} \, d\tau = q,
\]

a constant if \( q \) is constant.

The extreme case \( k \to \infty \) leads to the autocovariance function \( A_x(t-t') \) becoming extremely narrow, but the surface area under the curve of the function does not change. In other words:

\[
A_x(t-t') = q \delta(t-t').
\]

\(^8\)The name again uses the uneven brightness distribution with frequency of coloured light as a metaphor.

\(^9\)Autoregressive processes can also be discretely described as autoregressive sequences \( x_i \), which, for a first-order process, will only depend on the previous member of the sequence:

\[
x_{i+1} = ax_i + n_{i+1},
\]

in which \( x_i \overset{\text{def}}{=} x(t_i) \), and \( x_{i+1} \overset{\text{def}}{=} x(t_{i+1}) \). With \( \Delta t \overset{\text{def}}{=} t_{i+1} - t_i \) and comparing with equation 2.11 we have \( a = (1 - k\Delta t) \). Equation 2.14 is an example or special case of a *Markov chain*, in which each member \( x_{i+1} \) of the sequence, or chain, is derived only from, and can be optimally estimated using only, the previous member \( x_i \) and not from older members \( x_j \), \( j < i \). A Markov chain is thus *memoryless*, a property also called the Markov property.
“Coloured noise,” Gauss–Markov process

\[ k = 0.5 \]
\[ k = 1 \]
\[ k = 2 \]

\[ \text{Figure 2.4. Autocovariance function of a Gauss–Markov process as function of the time difference } \Delta t = t - t'. \]

This corresponds to a degeneration of the equation 2.11, in which not only \( k \to \infty \), but also variance of the noise \( \tilde{n} \): \( Q \to \infty \). So:

\[
\frac{dx}{dt} = kx - k\tilde{n} \implies x = \tilde{n} - \frac{1}{k} \frac{dx}{dt} \to \tilde{n},
\]
in which the variance of the noise \( \tilde{n} \) def = \( -\frac{n}{k} \) is \( q = Qk^{-2} \).

The other borderline case case, where \( k \to 0 \), is the same as the case presented above (section 2.5). So “random walk” is a Gauss–Markov process the time constant of which is infinitely long. In that case one must use the whole equation 2.12:

\[
A_x(t, t') = \frac{Q}{2k} \left( e^{-k|t-t'|} - e^{-k(t+t'-2t_0)} \right).
\]

In this case, if \( t \approx t' \) def \( \bar{t} \), we get

\[
A_x(\bar{t}) = \frac{Q}{2k} \left( 1 - e^{-2k(\bar{t}-t_0)} \right) \approx Q(\bar{t} - t_0),
\]
which is in practice the same as in chapter 2.5.

The corresponding differential equation is obtained from equation 2.11 by substituting \( k = 0 \):

\[
\frac{dx}{dt} = n_x \tag{2.15}
\]
so \( x \) is the time-integral of the white noise \( n \) as it should be.
Table 2.2. A summary of the properties of various stochastic processes.

<table>
<thead>
<tr>
<th></th>
<th>k</th>
<th>Equation</th>
<th>Autocovariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random walk</td>
<td>0</td>
<td>$\frac{d}{dt}x = n$</td>
<td>$Q(t - t_0)$</td>
</tr>
<tr>
<td>Gauss–Markov process</td>
<td>(0, $\infty$)</td>
<td>$\frac{d}{dt}x = -kx + n$</td>
<td>$(Q/2k) e^{-k</td>
</tr>
<tr>
<td>White noise</td>
<td>$\infty$</td>
<td>$kx = n$</td>
<td>$(Q/k^2) \delta(t - t')$</td>
</tr>
</tbody>
</table>

Often the model of equation 2.11 is used to generate a “coloured” noise process in cases where we know beforehand that the properties of the process are of that type. This is easily done by adding one unknown to the vector of unknowns and one equation to the system of equations. We shall see later how to do this with the Kalman filter, section 4.4.

### 2.7 Power spectral density (PSD)

#### 2.7.1 Definition

We often want to study stochastic processes in terms of their spectrum, i.e., the presence of various frequency constituents in the process. This can be done by using the Fourier transform.

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 $\Delta t = t - t' = t$. Which by choosing $t' = 0$ yields (Wiener–Khinchin$^{10}$–Einstein theorem):

$$\tilde{A}_x(f) = \mathcal{F}\{A_x(\Delta t)\} = \int_{-\infty}^{+\infty} A_x(t)e^{-2\pi ift} \, dt, \quad (2.16)$$

assuming it exists. Here, $f$ is the frequency, which is expressed, e.g., in Hz (after Heinrich Hertz$^{11}$), i.e., cycles per second, or s$^{-1}$. Analogically we may also define the cross-PSD of two functions:

$$\tilde{C}_{xy}(f) = \mathcal{F}\{C_{xy}(\Delta t)\} = \int_{-\infty}^{+\infty} C_{xy}(t)e^{-2\pi ift} \, dt.$$  

The inverse operation using the inverse Fourier transform yields

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

$^{10}$Alexander Yakovlevich Khinchin (1894–1959) was a Soviet mathematician and contributor to probability theory.

$^{11}$Heinrich Rudolf Hertz (1857–1894) was a German physicist, the first to generate and detect radio waves.
Therefore, for $t = 0$ we obtain

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

So the variance of process $x$ is the same as the total surface area under its PSD curve.

Because the auto-covariance function is symmetric, i.e.

$$A_x(\Delta t) = A_x(t, t') = A_x(t', t) = A_x(-\Delta t),$$

it follows that the PSD, equation 2.16, is real valued:

$$\tilde{A}_x(f) = \int_{-\infty}^{+\infty} A_x(t) \left( \cos(-2\pi ft) + i \sin(-2\pi ft) \right) \, dt = 2 \int_{0}^{\infty} A_x(t) \cos(2\pi ft) \, dt.$$

In addition, it is everywhere non-negative,

$$\tilde{A}_x(f) \geq 0 \ \forall f.$$

For cross-PSDs this does not hold: we have

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

as opposed to

$$A_x(t, t') = E\{ (x(t) - E\{x(t)\}) (x(t') - E\{x(t')\}) \} = E\{ (x(t') - E\{x(t')\}) (x(t) - E\{x(t)\}) \} = A_x(t', t).$$

## 2.7.2 White noise

The PSD of white noise may be computed as follows using the expression (2.6):

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

Which by choosing $t' = 0$ yields

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

using the $\delta$ function’s integration property, equation 2.7. Here we see why a process with a Dirac $\delta$ type autocovariance function is called white noise: the modulus $Q$ of the power spectral density is a constant all over the spectrum, for all frequencies $f$, just like is the case for white light.
2.7.3 Gauss–Markov process

The auto-covariance function of a Gauss–Markov process is given by equation 2.13:

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

From this follows the PSD by integration according to equation 2.16, again choosing $t' = 0$:

$$\widetilde{A}_x(f) = \int_{-\infty}^{+\infty} A_x(\Delta t) e^{-2\pi if t} \, dt = \frac{Q}{2k} \int_{-\infty}^{+\infty} e^{-k|t|} e^{-2\pi if t} \, dt = \frac{Q}{k} \int_0^{\infty} e^{-kt} \cos(2\pi ft) \, dt.$$  

This integral isn’t quite obvious to evaluate; it is found in tabulations of integrals (Wolfram Functions, $\int e^{bx} \cos cx \, dx$) and can also be done using symbolic algebra software, like the online integrator of Wolfram Research. The result is\(^{12}\)

$$\widetilde{A}_x(f) = \frac{Q}{4\pi^2f^2 + k^2} = \frac{2kA_x(0)}{4\pi^2f^2 + k^2}.$$  

See Jekeli (2001) equation (6.75). In the figure are plotted values of this function for $Q = 2k$ — so we keep the variance of $x$, which equals $A_x(0) = Q/2k$, at unity — with $k = 0.5, 1, 2$.

\(^{12}\) A formula of this form is sometimes called a Cauchy–Lorentz distribution.
2.8 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 in time. Often, the development 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.

2.8.1 Least-squares regression in absence of autocorrelation

Linear regression starts from the well known equation

\[ y = a + bx, \]

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

\[ y_i = a + bx_i + n_i, \quad (2.17) \]

in which the stochastic process \(n_i\) models the stochastic uncertainty of the measurement process, i.e., the noise.

Another way to write this observation equation is

\[ y_i + v_i = \hat{a} + \hat{b}x_i \]

in which now \(v_i\) is the residual of observation \(y_i\), and \(\hat{a}\) and \(\hat{b}\) are the estimators of the unknowns \(a\) and \(b\).

We assume the noise to behave so, that the variance is a constant independent of \(i\) ("homoskedasticity"), and that the covariance vanishes identically\(^{13}\) ("white noise"):

\[
\text{Var}\{n_i\} = \sigma^2, \\
\text{Cov}\{n_i, n_j\} = 0, \text{ if } i \neq j.
\]

This is called the statistical model.

We may write the observation equations into the form

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} + 
\begin{bmatrix}
v_1 \\
v_2 \\
\vdots \\
v_n
\end{bmatrix} = 
\begin{bmatrix}
1 & x_1 \\
1 & x_2 \\
\vdots & \vdots \\
1 & x_n
\end{bmatrix} 
\begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix},
\]

\(^{13}\)This set of assumptions is often called i.i.d., “independent and identically distributed”
in which now \( y \) \( = \) \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T \) is the vector of observations, \( \mathbf{v} \) \( = \) \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T \) that of residuals in an \( n \)-dimensional abstract vector space, and \( \mathbf{x} \) \( = \) \begin{bmatrix} a & b \end{bmatrix}^T \) is the vector of unknowns (parameters), and

\[
\mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots \\ 1 & x_n \end{bmatrix}
\]

is the design matrix. This way of presentation is referred to as the functional model.

Based on the assumed statistical model we may compute the least-squares solution with the help of the normal equations:

\[
(A^T A) \hat{x} = A^T y.
\]

More concretely

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

or (Cramèr’s rule):

\[
(A^T A)^{-1} = \frac{1}{n \sum x^2 - (\sum x)^2} \begin{bmatrix} \sum x^2 - \sum x \\ -\sum x & n \end{bmatrix},
\]

from which

\[
\hat{a} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}, \quad \hat{b} = \frac{-\sum x \sum y + n \sum xy}{n \sum x^2 - (\sum x)^2},
\]

which are the least-squares estimators of the unknowns. Their precision (uncertainty, mean error) is obtained by formal error propagation from the diagonal elements of the inverted normal matrix \( (A^T A)^{-1} \), scaled by the factor \( \sigma \), the mean error of unit weight:

\[
\sigma_a = \sigma \sqrt{\frac{\sum x^2}{n \sum x^2 - (\sum x)^2}}, \quad \sigma_b = \sigma \sqrt{\frac{n}{n \sum x^2 - (\sum x)^2}}.
\]

Most often we are particularly interested in the trend \( b \), meaning that we should compare the value \( \hat{b} \) obtained with its own mean error \( \sigma_b \). If \( \sigma \) is not known \( a \ priori \), it should be evaluated from the residuals: the square sum of residuals

\[
\sum v^2 = \sum (\hat{a} + \hat{b}x - y)^2
\]
Linear regression of time series

has the expectancy of \((n - 2) \sigma^2\), where \(n - 2\) is the number of degrees of freedom (overdetermination), \(2\) being the number of unknowns estimated. Or

\[
\hat{\sigma}^2 = \frac{\sum v^2}{n - 2}.
\]

### 2.8.2 Serial correlation

The assumption made above, that the observational errors \(n_i\) are uncorrelated between themselves, is often wrong. Nevertheless, least-squares regression is such a simple method — available for example in popular spreadsheets and pocket calculators — that it is often used even though the zero correlation requirement is not fulfilled.

If the autocorrelation of the noise process \(w_i\) does not vanish, we can often model it as a so-called AR(1) — auto-regressive first-order or Gauss–Markov process. Such a process is described as a Markov chain:

\[
w_{i+1} = \rho w_i + n_i, \tag{2.18}
\]

where \(\rho\) is a suitable damping parameter, \(0 < \rho < 1\), and \(n\) is a truly non-correlating white-noise process:

\[
\text{Var}\{n_i\} = \sigma^2, \quad \text{Cov}\{n_i, n_j\} = 0, \quad i \neq j. \tag{2.19}
\]

Write the original observation equation

\[
y_i = a + bx_i + w_i
\]

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

\[
y_{i+1} = a + bx_{i+1} + w_{i+1},
-\rho y_i = -\rho a - \rho bx_i - \rho w_i,
\]

and sum together:

\[
y_{i+1} - \rho y_i = a (1 - \rho) + b(x_{i+1} - \rho x_i) + (w_{i+1} - \rho w_i). \tag{2.20}
\]

This equation is of the form

\[
Y_i = A + bX_i + n_i, \tag{2.21}
\]

the equation for the non-correlated linear regression, in which \(n_i\) is white noise, equation 2.19, and

\[
A = a (1 - \rho),
X_i = x_{i+1} - \rho x_i, \tag{2.22}
Y_i = y_{i+1} - \rho y_i.
\]

The recipe now is:
1. Compute $X_i$ and $Y_i$ according to above equations 2.22.

2. Solve $\hat{A}$ and $\hat{b}$ according to non-correlated linear regression.

3. Compute $\hat{A} = (1 - \rho)^{-1} \hat{A}$.

4. The ratio between $\sigma_n^2$ and $\sigma_w^2$: from equation 2.20 it follows that
   \[ n_i = W_i + 1 - \rho W_i, \]
   so
   \[ \sigma_w^2 = \rho^2 \sigma_n^2 + \sigma_n^2 \implies (1 - \rho^2) \sigma_w^2 = \sigma_n^2. \]
   One estimates an empirical variance $\hat{\sigma}_w^2$ from the modified observation equation 2.21. This differs from the “naively” calculated variance estimate $\hat{\sigma}_w^2$ of the original observation equation 2.17:
   \[ E\left\{ \hat{\sigma}_w^2 \right\} = \frac{E\left\{ \hat{\sigma}_n^2 \right\}}{1 - \rho^2}. \]

5. From point 4 we may conclude that
   \[ \sigma_{b,AR(1)}^2 = \frac{\sigma_{b,nocorr}^2}{1 - \rho^2}, \]
   in which $\sigma_{b,nocorr}^2$ is again the “naively” calculated variance of the trend parameter.

Conclusion:
- if there is serial correlation (autocorrelation) in the data, a simple linear regression will give a too optimistic picture of the trend parameter $b$’s mean error, and thus also of the statistical significance of its difference from zero.
- If the data is given as an equi-spaced function of time, i.e., $x_i = x_0 + (i - 1) \Delta t$, then the parameter $\rho$ of the AR(1) process is related in a simple way to its correlation length. From equation 2.18:
   \[ W_{i+1} = \rho W_i + n_i \implies W_{i+1} = W_i e^{-\Delta t/\tau} + n_i, \]
   in which $\tau$ is the correlation length expressed in units of time.

Self-test questions

1. Verify that the integral over the normal distribution, equation 2.1:
   \[ \int_{-\infty}^{+\infty} p(x) \, dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma \sqrt{2\pi}} \exp\left( -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right) \, dx, \]
   produces 100% total probability.
**Hint:** consider instead the two-dimensional stochastic variable

\[ x = \begin{bmatrix} x \\ y \end{bmatrix} \]

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 2\( \pi \) in the equation comes from.
The Kalman filter


The Kalman filter is a linear, predictive filter. Like a coffee filter 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 movement process.

The inventors of the Kalman filter were Rudolf Kalman and Richard Bucy in the years 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 was the orbital rendez-vous 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 program was proposed. The technique of lunar orbital rendez-vous (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 generally applicable and broadly used not only in navigation but also for example in economics and meteorology.

The Kalman filter consists of two parts:

---

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

2Richard Snowden Bucy (born 1935) is an American mathematician
Figure 3.1. Lunar orbital rendez-vous: two astronauts and a bag of rocks returning home. 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 CSM was rendered inoperable by an on-board explosion.

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 statistics: the dynamic model describes random effects on the development of the system over time, like perturbations of a satellite orbit, while the observational model describes the effect of observational uncertainty.

The Kalman filter is special in the sense that the state vector propagates forward in time step by step, and also the observations are used to correct the state-vector estimate only at times when observations are made. Because of this, the Kalman filter does not demand high number-
The state vector crunching power or the handling of big matrices. It can be easily implemented onboard a vehicle and in real time.

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 at a given point in time or epoch $t$.

We demonstrate this be a concrete example. A particle moving freely in space has three position co-ordinates and three velocity components. Let an orthonormal basis in this space be

\[ \{i, j, k\} \]

Then we write as the position vector of the particle

\[ x \overset{\text{def}}{=} xi + yj + zk, \]

and as the velocity vector

\[ v \overset{\text{def}}{=} \dot{x}i + \dot{y}j + \dot{z}k. \]

A vector in space is often identified with the formal vector of its three components on the agreed orthonormal basis \( \{i, j, k\} \). So, we write carelessly

\[ x = x_i + y_j + z_k = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \]

though $x$ and $\bar{x}$ are conceptually different things.

Now the state vector of the particle becomes

\[
\bar{x} = \begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x,
\]

in which the position and velocity vectors are

\[
x = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad v = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix}.
\]
The Kalman filter

(a) Kalman filter

\[
\frac{dx(t)}{dt} = F(x, t) + \nu(t) \overset{\text{lin}}{\approx} F(t) \cdot x(t) + \nu(t), \quad A_n(t_1, t_2) = \int_1^{t_2} Q(t) \, dt,
\]

\[
x(t_{k+1}) = \Phi_{k+1}^k x(k) + \omega_{k+1}^k, \quad w_{k+1}^k = \int_k^{k+1} \Phi_{t}^{k+1} \nu(t) \, dt,
\]

\[
\Sigma^-(t_{k+1}) = \left( \Phi_{k+1}^k \right) \Sigma^-(t_k) \left( \Phi_{k+1}^k \right)^T + \Theta_{k+1}^k,
\]

\[
\Theta_{k+1}^k = A_w(t_{k+1}, t_{k+1}) = \int_k^{k+1} \Phi_{t}^{k+1} Q(t) \left( \Phi_{t}^{k+1} \right)^T \, dt.
\]

(b) Dynamic model

\[
\ell_k = H(x^-(t_k)) + \bar{m}_k \overset{\text{lin}}{\approx} H_k \cdot x^-(t_k) + \bar{m}_k, \quad \text{Var}\{m_k\} = R_k,
\]

\[
x^+(t_k) = x^-(t_k) + K_k \left( H(x^-(t_k)) - \ell_k \right) \overset{\text{lin}}{\approx} x^-(t_k) + K_k \left( H_k \cdot x^-(t_k) - \ell_k \right),
\]

\[
\Sigma^+(t_k) = (I + K_k H_k) \Sigma^-(t_k),
\]

\[
K_k = \Sigma^H_k (H_k \Sigma^-(t_k) H_k^T + R_k)^{-1}.
\]

(c) The observation model and update step

Figure 3.2. The Kalman filter. 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.
The dynamic model

The underline denotes that a quantity is stochastic.

In this case the state vector has six elements or degrees of freedom.

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

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

If the particle is not a point but an extended object, also its orientation angles or Euler angles enter into the state vector. Then we already have nine elements. In a system of several particles every particle contributes its own elements, three position 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 navigation device.

### 3.2 The dynamic model

The dynamic model characterizes the behaviour of the state vector in time. The state vector as defined in the precious section is a vector valued stochastic process as 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. We present the linear case first, and then the non-linear case, applying linearisation, producing the so-called extended Kalman filter.

In the linear case the dynamic model looks like

$$\frac{d}{dt} \bar{x}(t) = F(t) \cdot \bar{x}(t) + \bar{n}(t), \quad (3.2)$$

in which $\bar{x}$ is the state vector, $\bar{n}$ is the dynamic noise, in other words, how inaccurately these equations of motion actually describe the real motion. $F$ — also possibly dependent on time — is the coefficient matrix.

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

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

$$\frac{d}{dt} \bar{x}(t) = F(\bar{x}, t) + \bar{n}(t), \quad (3.3)$$

in which $F(\bar{x}, t)$ is a vectorial function. From this, the linear case easily follows: choose an approximate value $\bar{x}^{(0)}(t)$ for the state vector. We
demand from this approximate value — also a function of time! — consistency with the functional model:

\[
\frac{d}{dt} x^{(0)}(t) = F(x^{(0)}, t). \tag{3.4}
\]

Now we linearize by subtracting equations 3.4 and 3.3 from each other and doing a Taylor expansion:

\[
\frac{d}{dt}(x - x^{(0)}) = F(x, t) + n(t) - F(x^{(0)}, t) \approx F(t) \cdot (x - x^{(0)}) + n(t),
\]

which already is of the form 3.2 if we write \( \Delta x \) as:

\[
\frac{d}{dt} \Delta x(t) = F(t) \cdot \Delta x(t) + n(t),
\]

from which one may again drop the deltas.

The elements of the Jacobi\(^3\) matrix \( F \) of the function \( F(\cdot) \) used above are

\[
F_{ij}(t) = \frac{\partial}{\partial x_j} F_i(x, t) \bigg|_{x=x^{(0)}}, \quad i, j = 1, \ldots, n,
\]

where the \( x_j \) are the \( n \) components of \( x \). For the example state vector given in equation 3.1 they are

\[
\begin{bmatrix}
    x_1 & x_2 & x_3 & x_4 & x_5 & x_6
\end{bmatrix}^T = \begin{bmatrix}
    x & y & z & \dot{x} & \dot{y} & \dot{z}
\end{bmatrix}^T.
\]

These derivatives are evaluated at the approximate values in vector \( x^{(0)} \).

\( F \) itself is also a vector with \( n \) components, \( F_i, \ i = 1 \ldots, n. \)

Realistic statistical attributes have to be defined for the dynamic noise: often it is assumed that it is white noise, the autocovariance of which is (equation 2.6):

\[
A_n(t_1, t_2) = Q \delta(t_2 - t_1).
\]

### 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 \( GM \):

\[
\frac{d^2}{dt^2} \begin{bmatrix}
    \dot{x} \\
    \dot{y} \\
    \dot{z}
\end{bmatrix} = -\frac{GM_{\oplus}}{(x^2 + y^2 + z^2)^{3/2}} \begin{bmatrix}
    \dot{x} \\
    \dot{y} \\
    \dot{z}
\end{bmatrix} + \begin{bmatrix}
    n_x \\
    n_y \\
    n_z
\end{bmatrix},
\]

\(^3\)Carl Gustav Jacob Jacobi (1804–1851) was a brilliant Jewish German mathematician. He died of smallpox only 46 years of age.
Example: satellite motion

in which \( n_x, n_y, n_z \) describe, e.g., the randomly varying effects of air drag or the irregularities of the Earth's gravitational field, which make the spacecraft motion slightly unpredictable.

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

\[
\frac{d}{dt} \begin{bmatrix} x(t) \\ y(t) \\ z(t) \\ \dot{x}(t) \\ \dot{y}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} F(x(t)) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{n_x}{n_x} \\ \frac{n_y}{n_y} \\ \frac{n_z}{n_z} \end{bmatrix}.
\]

This system of equations is non-linear. Linearisation yields

\[
\frac{d}{dt} \begin{bmatrix} \Delta x(t) \\ \Delta y(t) \\ \Delta z(t) \\ \Delta \dot{x}(t) \\ \Delta \dot{y}(t) \\ \Delta \dot{z}(t) \end{bmatrix} = \begin{bmatrix} F(t) \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{GM_\oplus x}{r^3} (3x^2 - \frac{M_\oplus}{r^3} x^2) \\ 0 \\ \frac{GM_\oplus y}{r^3} (3y^2 - \frac{M_\oplus}{r^3} y^2) \end{bmatrix} \begin{bmatrix} \Delta x(t) \\ \Delta y(t) \\ \Delta z(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{n_x}{n_x} \\ \frac{n_y}{n_y} \\ \frac{n_z}{n_z} \end{bmatrix}.
\]

in which \( r = \sqrt{x^2 + y^2 + z^2} \) is the distance from the Earth's centre. It is also assumed that

1. there is an appropriate set of approximate values available to form an approximate state vector

\[
\begin{bmatrix} x^{(0)} \\ y^{(0)} \\ z^{(0)} \\ \dot{x}^{(0)} \\ \dot{y}^{(0)} \\ \dot{z}^{(0)} \end{bmatrix}^T,
\]

relative to which the \( \Delta \) quantities have been calculated, and that

2. the elements of the coefficient matrix \( F(t) \) are evaluated using those approximate values.

Each element in the state vector is a function of time, as is the vector itself.
3.3.1 The Eötvös tensor

The “partitioned” version of dynamic equation 3.5 above is

$$\frac{d}{dt} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ M & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} + \begin{bmatrix} 0 \\ n \end{bmatrix},$$

(3.6)
in which I is the size 3×3 unit matrix, and

\[
\mathcal{M} = \begin{bmatrix}
\frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
\frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\
\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) = \end{align}

\[
\begin{align}
= \frac{GM_\oplus}{r^3} \left[ \begin{array}{ccc}
\frac{\partial^2}{\partial x^2} & \frac{\partial^2}{\partial x \partial y} & \frac{\partial^2}{\partial x \partial z} \\
\frac{\partial^2}{\partial y \partial x} & \frac{\partial^2}{\partial y^2} & \frac{\partial^2}{\partial y \partial z} \\
\frac{\partial^2}{\partial z \partial x} & \frac{\partial^2}{\partial z \partial y} & \frac{\partial^2}{\partial z^2}
\end{array} \right] \frac{GM_\oplus}{r} =
\begin{align}
= \frac{GM_\oplus}{r^3} \begin{bmatrix}
3x^2 - r^2 & 3xy & 3xz \\
3yx & 3y^2 - r^2 & 3yz \\
3zx & 3zy & 3z^2 - r^2
\end{bmatrix}
\end{align}
\]

(3.7)
is called the gravitational gradient tensor, also known as the Eötvös tensor.

The Eötvös tensor is the matrix of partial derivatives with respect to place of the gravitation vector \((GM_\oplus/r^3)\mathbf{x}\). Remember that the gravitation vector is itself the gradient of the geopotential. The tensor contains all second partial derivatives with respect to place of the geopotential \(GM_\oplus/r\). All these equations assume a central gravitational field.

The tensor \(\mathcal{M}\) describes how a small perturbation \(\Delta x = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix}^T\) in the satellite location converts into an acceleration perturbation

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

3.3.2 Choosing approximate values

The most important thing when choosing the approximate values is, that they be physically consistent, in other words that they together describe a really possible orbital motion inside the assumed gravitational field.
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.

Now if the gravitational field model available is more complicated than a central field, one must integrate the approximate values over time using this more accurate field model. Nevertheless the above linearized dynamic model, equation 3.5, will still be good for integrating the difference quantities \( \Delta x, \Delta v \), as long as these remain *numerically small*. This is one of the benefits of linearization.

### 3.4 State propagation

*State propagation* is done by integrating the differential equation, or dynamic model, in the linear case equation 3.2:

\[
\frac{d}{dt} x(t) = F(t) \cdot x(t) + \eta.
\]

In the actual calculation we are not integrating the actual state vector \( x(t) \), as we do not know it. We can only integrate the *state estimator* \( x^-(t) \). Also the noise term \( \eta \) is left out in the absence of actual knowledge on it. So

\[
\frac{d}{dt} x^-(t) = F(t) \cdot x^-(t). \tag{3.8}
\]

**Notation used:**
- \( x^- \) is the *a priori* state estimator, the estimator before the (later to be described) update step.
- \( x^+ \) is the *a posteriori* state estimator, after this step.

In the literature, also the notations \( \hat{x}^{i-1} \) and \( \hat{x}^i \) can be found, with \( i \) the sequence number of the update. The “hat” symbol denotes an estimator.

**Definition:**

*The state variance is the expected square difference of its estimator from its true value — itself a stochastic process to which of course we do not have access! — as follows:*

\[
\Sigma^-(t) = \text{Var}\{x^-(t)\} \overset{\text{def}}{=} E\left\{ (x^-(t) - \hat{x}(t)) (x^-(t) - \hat{x}(t))^T \right\}. \tag{3.9}
\]
### 3.4.1 State transition matrix

For small time steps $\Delta t = t_{k+1} - t_k$ we may write approximately

$$x^-(t_{k+1}) \approx x^-(t_k) + F \Delta t \cdot x^-(t_k) = (I + F \Delta t) x^-(t_k).$$

We see that the elements of state vector $x^-(t_{k+1})$ are *linear combinations* of the elements of the earlier state vector $x^-(t_k)$. If for a small time interval $\delta t$ it holds that $\Delta t = t_{k+1} - t_k = n\delta t$, one obtains by repeatedly applying the above equation:

$$x^-(t_{k+1}) = (I + F \delta t)^n x^-(t_k).$$

The matrix

$$\Phi_k^{k+1} \overset{\text{def}}{=} (I + F \delta t)^n$$

is called the *state transition matrix* between epochs $t_k$ and $t_{k+1}$, and we can write

$$x^-(t_{k+1}) = \Phi_k^{k+1} x^-(t_k).$$

(3.10)

By substituting $\delta t = \Delta t / n$ we obtain

$$\Phi_k^{k+1} = \left( I + \frac{F \Delta t}{n} \right)^n.$$

For ordinary real numbers we have the classical equation

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

$$= \lim_{y \to \infty} \left( 1 + \frac{1}{y} \right)^{yx} = \lim_{y \to \infty} \left( \left( 1 + \frac{1}{y} \right)^y \right)^x,$$

in which we see the definition of the number $e$:

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

This is why we write, generalizing the $\exp$ and $\ln$ functions to square matrices:

$$\Phi_k^{k+1} = \exp\left( \ln \left( I + \frac{F \Delta t}{n} \right)^n \right) = \exp\left( n \ln \left( I + \frac{F \Delta t}{n} \right) \right) \approx$$

$$\approx \exp\left( n \frac{F \Delta t}{n} \right) = \exp\left( F \left( t_{k+1} - t_k \right) \right).$$

(3.11)

This is to be interpreted as a Taylor expansion:

$$\Phi_k^{k+1} = \exp(F \Delta t) = I + F \Delta t + \frac{1}{2} F^2 \Delta t^2 + \frac{1}{6} F^3 \Delta t^3 + \ldots$$

We observe that for the state transition matrix the *transitive property* holds:

$$\Phi_k^{k+2} = \Phi_k^{k+1} \cdot \Phi_k^{k+1},$$

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

≡ ≮ ≫
3.4.2 Example: satellite motion

The differential equation 3.6 can be converted, for small values of $\Delta t$ and leaving off the noise term, to

$$
\begin{bmatrix}
\Delta x \\
\Delta v
\end{bmatrix}_{i+1} -
\begin{bmatrix}
\Delta x \\
\Delta v
\end{bmatrix}_i 
\approx \Delta t
\begin{bmatrix}
0 & I \\
\text{M} & 0
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta v
\end{bmatrix}_i
$$

$$
\Rightarrow
\begin{bmatrix}
\Delta x \\
\Delta v
\end{bmatrix}_{i+1} =
\begin{bmatrix}
I & I \Delta t \\
\text{M} \Delta t & I
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta v
\end{bmatrix}_i,
$$

with the definitions

$$
\Delta x_i \overset{\text{def}}{=} \Delta x(t_i), \quad \Delta v_i \overset{\text{def}}{=} \Delta v(t_i), \quad \text{M} \overset{\text{def}}{=} M(t_i), \quad \Delta t \overset{\text{def}}{=} t_{i+1} - t_i.
$$

We see that the state transition matrix is

$$
\Phi_i^{i+1} \approx
\begin{bmatrix}
I & I \Delta t \\
\text{M}(t_i) \Delta t & I
\end{bmatrix}.
$$

Likewise, the matrix of the next state transition will be

$$
\Phi_i^{i+2} \approx
\begin{bmatrix}
I & I \Delta t \\
\text{M}(t_{i+1}) \Delta t & I
\end{bmatrix}.
$$

Concatenating, or multiplying, them yields

$$
\Phi_i^{i+2} = \Phi_i^{i+1} \cdot \Phi_i^{i+1} =
\begin{bmatrix}
I & I \Delta t \\
\text{M}(t_{i+1}) \Delta t & I
\end{bmatrix}
\begin{bmatrix}
I & I \Delta t \\
\text{M}(t_i) \Delta t & I
\end{bmatrix}
= 
\begin{bmatrix}
I + O(\Delta t^2) & (I + I) \Delta t \\
(\text{M}(t_{i+1}) + \text{M}(t_i)) \Delta t & I + O(\Delta t^2)
\end{bmatrix}.
$$

Repeating this for multiple small time steps $\Delta t$ suggests that the following integral equation applies:

$$
\Phi_i^{k+1} = 
\begin{bmatrix}
I & I (t_{k+1} - t_k) \\
\int_k^{k+1} \text{M}(t) \, dt & I
\end{bmatrix}.
$$

3.4.3 Propagation of state and state variance

We look at the discrete propagation in time of both $\bar{x}(t)$ and $\bar{x}(t)$. Write equation 3.2:

$$
\frac{d}{dt} \bar{x}(t) = F(t) \bar{x}(t) + n(t), \quad (3.12)
$$

and equation 3.8:

$$
\frac{d}{dt} \bar{x}^- (t) = F(t) \bar{x}^- (t), \quad (3.13)
$$
and subtract:
\[ \frac{d}{dt} (x^+(t) - x(t)) = F(t) (x^+(t) - x(t)) - n(t). \]

The discrete propagation equation 3.10 for the state estimator:
\[ x^-(t_{k+1}) = \Phi^{k+1}_k x^-(t_k) \] 
(3.14)
with \( \Phi^{k+1}_k = \Phi^{k+1}_k \) the state transition matrix. In the absence of noise \( n(t) \) it also holds that
\[ x(t_{k+1}) = \Phi^{k+1}_k x(t_k), \]
and subtraction yields
\[ (x^-(t_{k+1}) - x(t_{k+1})) = \Phi^{k+1}_k (x^-(t_k) - x(t_k)), \]
after which propagation of variances yields
\[ \Sigma^-(t_{k+1}) = (\Phi^{k+1}_k) \Sigma^-(t_k) (\Phi^{k+1}_k)^T. \]

3.4.4 State propagation in the presence of noise

In this case the following equations apply:

\[
\begin{align*}
    x^-(t_{k+1}) &= \Phi^{k+1}_k x^-(t_k) \\
    x(t_{k+1}) &= \Phi^{k+1}_k x(t_k) + w^k_{t_{k+1}} \\
    \Sigma^-(t_{k+1}) &= (\Phi^{k+1}_k) \Sigma^-(t_k) (\Phi^{k+1}_k)^T + \Theta^{k+1}_k
\end{align*}
\] 
(3.15)

in which, as we will show,
\[ w^k_{t_{k+1}} = w^k_{t_{k+1}} \overset{\text{def}}{=} \int_k^{k+1} \Phi^{k+1}_t n(t) \, dt, \] 
(3.16)
and the autocovariance of \( w \)
\[ \Theta^{k+1}_k = \Theta^{k+1}_k \overset{\text{def}}{=} A_w(t_{k+1}, t_{k+1}) = \int_k^{k+1} \Phi^{k+1}_t Q(t) (\Phi^{k+1}_t)^T \, dt. \] 
(3.17)

The interpretation of the equations 3.16 and 3.17 is, that each element of noise \( n(t) \, dt \) in the interval \( (t_k, t_{k+1}) \), and the variance matrix of each element, \( Q(t) \, dt \), are transitioned forward in time from the moment \( t \) to the moment \( t_{k+1} \) by multiplying with the state transition matrix.
State propagation

\[ \Phi_{t}^{k+1}(\cdot) \]

\[ \Sigma_{k}^{-} \]

\[ Q \ dt \]

\[ w_{k}^{k+1} \]

\[ x_{k} \]

\[ t_{k} \]

\[ t_{k+1} \]

\[ x_{k+1} \]

\[ \Theta_{k}^{k+1} \]

\[ \int_{k}^{k+1}(\cdot) \]

\[ \Sigma_{k+1}^{-} \]

\[ \Phi_{k}^{k+1}(\cdot) \]

\[ t \rightarrow \]

\[ \int_{k}^{k+1}(\cdot) \]

\[ \Phi_{t}^{k+1}(\cdot) \]

\[ \left( \Phi_{k}^{k+1} \right) \left( \Phi_{k}^{k+1} \right)^{T} \]

\[ \left( \Phi_{t}^{t_{k+1}} \right) \left( \Phi_{t}^{t_{k+1}} \right)^{T} \]

\[ n_{t} \ dt \]

\[ Q(t) \ dt \]

\[ w(t) \]

\[ \hat{w}(t) \]

\[ \hat{\Theta}(t) \]

\[ \int_{k}^{k+1} Q(t) \ dt \]

\[ \int_{k}^{k+1} n(t) \ dt \]

\[ W_{k}^{k+1} \]

\[ \Theta_{k}^{k+1} \]

\[ \Theta_{k}^{k+1} \approx \int_{k}^{k+1} Q(t) \ dt \]

\[ \Theta_{k}^{k+1} \approx \int_{k}^{k+1} Q(t) \ dt \]

Figure 3.3. The propagation of state vector and state variance in the presence of noise.

\[ \Phi_{t}^{t_{k+1}}, \text{ in case of the variance matrix both from the left and from the right. After that, the integration over the “brought forward” elements is carried out.} \]

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

\[ w_{k}^{k+1} \approx \int_{k}^{k+1} n(t) \ dt, \quad \Theta_{k}^{k+1} \approx \int_{k}^{k+1} Q(t) \ dt. \]

The assumption is that the dynamic noise \( n(t) \) is white. Note that now \( w(t) \) is a random walk, an integral over white noise, which we encountered earlier. A diagram describing the propagation process is given in figure 3.3.

3.4.5 Differential equation for state transition matrix

We may also derive differential equations that describe the development of the state variance matrix and state transition matrix in time.
Differentiate equation 3.14:
\[ x^-(t) = \Phi^i_0 x^-(t_0) \implies \frac{d}{dt} x^-(t) = \frac{d}{dt} (\Phi^i_0) x^-(t_0). \]

Substitute equation 3.13:
\[ \frac{d}{dt} x^-(t) = F(t) x^-(t) = F(t) \Phi^i_0 x^-(t_0), \]
yielding
\[ \frac{d}{dt} (\Phi^i_0) x^-(t_0) = F(t) \Phi^i_0 x^-(t_0) \implies \frac{d}{dt} (\Phi^i_0) = F(t) \Phi^i_0. \quad (3.18) \]

With the initial condition \( \Phi^i_0 = I \) we can by numerical integration obtain the matrix \( \Phi^i_0 \).

There is no closed solution, except when \( F \) is constant, or if \( F(t)F(t) = F(t)F(t) \) for arbitrary \( t, t' \) (Wikipedia, Magnus expansion). In that case, like equation 3.11,
\[ \Phi^i_t = \exp \left( \int_0^t F(\tau) d\tau \right), \quad (3.19) \]
to be interpreted as a Taylor expansion:
\[ \Phi^i_t = I + \int_0^t F(\tau) d\tau + \frac{1}{2} \left( \int_0^t F(\tau) d\tau \right)^2 + \frac{1}{6} \left( \int_0^t F(\tau) d\tau \right)^3 + \ldots \]
which generalises equation 3.11 for a time variable matrix \( F \).

A system for which the exponent equation 3.19 does not work, is the satellite example of subsection 3.4.2:
\[
F(t)F(t) = \begin{bmatrix} 0 & 1 \\ M(t') & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ M(t) & 0 \end{bmatrix} = \begin{bmatrix} M(t) & 0 \\ 0 & M(t') \end{bmatrix} \neq F(t)F(t').
\]

### 3.4.6 Differential equation for state variance

In order to derive a differential equation for the state variance matrix \( \Sigma \) we start from equation 3.15:
\[
\Sigma^-(t) = \left( \Phi^i_0 \right) \Sigma^-(t_0) \left( \Phi^i_0 \right)^T + \Theta^i_0.
\]

Differentiate and use equation 3.18:
\[
\frac{d}{dt} \Sigma^-(t) = \left( \frac{d}{dt} \Phi^i_0 \right) \Sigma^-(t_0) \left( \Phi^i_0 \right)^T + \left( \Phi^i_0 \right) \Sigma^-(t_0) \left( \frac{d}{dt} \Phi^i_0 \right)^T + \frac{d}{dt} \Theta^i_0 =
\]
\[
= F(t) \Phi^i_0 \Sigma^-(t_0) \left( \Phi^i_0 \right)^T + \Phi^i_0 \Sigma^-(t_0) \left( \Phi^i_0 \right)^T F(t) + \frac{d}{dt} \Theta^i_0 =
\]
\[
= F(t) \Sigma^-(t) + \Sigma^-(t) F(t) + \frac{d}{dt} \Theta^i_0. \quad (3.21)
\]
Here we defined
\[ \Sigma_0^-(t) \overset{\text{def}}{=} (\Phi_0^t) \Sigma^-(t_0) (\Phi_0^t)^\top, \]
which is computed by integrating the differential equation
\[ \frac{d}{dt} \Sigma_0^-(t) = F(t) \Sigma_0^-(t) + \Sigma_0^-(t) F^\top(t) \tag{3.22} \]
from the initial value \( \Sigma_0^-(t_0) = \Sigma^- (t_0) \), i.e., also the initial value for integrating equation 3.21.

We may combine the two equations 3.20 and 3.22 as follows:
\[ \Sigma^-(t) = \Sigma_0^-(t) + \Theta_0^t, \tag{3.23} \]
allowing the separate integration of \( \Sigma_0^-(t) \) and \( \Theta_0^t \).

### 3.4.7 Differential equations for state noise

Start from equation 3.15, with noise included,
\[ x(t) = \Phi_0^t x(t_0) + w_0^t. \tag{3.24} \]
Differentiate and substitute equation 3.18:
\[
\begin{align*}
\frac{d}{dt} x(t) &= \frac{d}{dt} (\Phi_0^t x(t_0)) + \frac{d}{dt} w_0^t \\
&= F(t) \Phi_0^t x(t_0) + \frac{d}{dt} w_0^t = F(t) x(t) + \frac{d}{dt} w_0^t. \tag{3.25}
\end{align*}
\]
Compare with equation 3.12:
\[ \frac{d}{dt} x(t) = F(t) x(t) + n(t), \]
yielding — note that \( w \) accumulates all noise inputs from the time interval \( (t_0, t) \):^4
\[ \frac{d}{dt} w_0^t = n(t) = \Phi_0^t n(\tau). \]
Integration yields
\[ w_0^t = \int_0^t \Phi_0^\tau n(\tau) d\tau. \]
This describes a random walk, equation 2.8, with autocovariance function 2.10:
\[ \Theta_0^t = A_w(t, t) = \int_0^t \Phi_0^\tau Q(\tau) (\Phi_0^\tau)^\top d\tau. \]

^4 The reason that this is necessary is, that the two parts of equation 3.25 cannot actually be separately and consistently integrated in the presence of noise. The postulated form 3.24 of the discrete propagation equation should not be taken too literally.
An alternative and more brutal approach to showing this is the following. Write again equation 3.15:

\[
\dot{x}(t) = \Phi_t^t x(t_0) + w_t^t = \Phi_t^t (\Phi_0^t x(t_0) + w_0^t) + w_t^t,
\]

and assume \( \delta t \overset{\text{def}}{=} t' - t_0 \) small. Then

\[
w_0^t \approx \delta t \cdot \eta \left( \frac{1}{2} (t_0 + t') \right) = \delta t \cdot \eta \left( t_0 + \frac{1}{2} \delta t \right),
\]

and

\[
\dot{x}(t) = \Phi_t^t \left( \Phi_0^t x(t_0) + w_0^t \right) + w_t^t = \Phi_t^t \Phi_0^t x(t_0) + \Phi_t^t w_t^t + w_t^t = \Phi_t^t x(t_0) + \delta t \Phi_t^t \eta \left( t_0 + \frac{1}{2} \delta t \right) + w_t^t + \frac{\delta t}{2} \delta t.
\]

Repeating the procedure to completion with \( t - t_0 = n \delta t \) produces, with \( w_{t_0 + n \delta t} = 0 \),

\[
x(t) = \Phi_t^t x(t_0) + \delta t \cdot \sum_{i=0}^{n-1} \Phi_t^{t_0 + (i+\frac{1}{2}) \delta t} \cdot \eta \left( t_0 + \left( i + \frac{1}{2} \right) \delta t \right),
\]

or as an integral (rectangle rule)

\[
x(t) = \Phi_0^t x(t_0) + \int_0^t \Phi_t^\tau \eta(\tau) \, d\tau,
\]

so

\[
w_t^t = \int_0^t \Phi_t^\tau \eta(\tau) \, d\tau.
\]

Similarly the autocovariance function of this random walk is

\[
\Theta_t^t = \Phi_t^t \Phi_t^T = \int_0^t \Phi_t^\tau Q(\tau) (\Phi_t^\tau)^T \, d\tau.
\]

The twist is the state transition matrix inside the integral. For small \( t - t_0 \), the matrix may be ignored, i.e., replaced by the unit matrix.

### 3.4.8 Final integration

Equation 3.13, or equation 3.14 together with equation 3.18, can be used to integrate the state \( x^{-1}(t) \) over time between update events. If the matrix \( F \) is not time dependent, the state transition matrix \( \Phi_k^{k+1} \) is directly obtained by equation 3.11.

Equations 3.21–3.23 are suitable for integrating the matrix \( \Sigma(t) \) also in the case where \( F \) is a function of time and not a constant.

All this however assumes that the matrix \( F \) exists, i.e., the function \( F(x, t) \) can be linearised.
3.5 Observational model

The evolution of the state vector in time would not be very interesting, unless it could be observed in some way. The observational model in the linear case is:

\[ \ell = Hx + m, \]

in which \( \ell \) is the observable, in the general case a vector, \( x \) is the state vector ("the true value"), and \( m \) is the "noise," i.e., the uncertainty, of the observation process. \( H \) is the observation matrix\(^5\). The variance matrix of the observation vector is

\[ R \triangleq \mathbb{E}\{mm^T\}. \]

We assume that \( \mathbb{E}\{m\} = 0 \) for the noise \( m \).

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

\[ \hat{\ell} = Hx^-. \]

Now, a zero quantity — a quantity the expectancy \( \mathbb{E}\{\cdot\} \) of which is zero — can be constructed as

\[ y = \hat{\ell} - \ell = Hx^- - \ell = H(x^- - x) - m, \]

and thus

\[ \mathbb{E}\{y\} = H\left(\mathbb{E}\{x^-\} - \mathbb{E}\{x\}\right) - \mathbb{E}\{m\} = H \cdot 0 - 0 = 0, \]

based the assumption \( \mathbb{E}\{x^-\} = \mathbb{E}\{x\} \), that \( x^- \) is an unbiased estimator of the state \( x \).

In the nonlinear case, \( H \) is not a matrix but a function \( H(x) \) of the state vector:

\[ \ell = H(x) + m \]

and

\[ \hat{\ell} = H(x^-), \]

and the difference is

\[ y = \hat{\ell} - \ell = H(x^- - x) - m. \]

\(^5\)This is the same as in the least-squares adjustment A matrix or “design matrix.”
The elements of the matrix $H$ are defined by

$$H_{ij} = \left. \frac{\partial}{\partial x_j} H_i(x) \right|_{x=x(0)},$$

the Jacobi matrix, or matrix of partial derivatives, of the function $H(x)$. Evaluation of which is done at the approximate values for the state vector contained in $x^{(0)}$.

We calculate

$$\text{Var}\{y\} = E\{yy^T\} = H E\{(x^– - x)(x^– - x)^T\} H^T + R = H \Sigma^{-} H^T + R,$$

while assuming that $x^–$ and $m$ are statistically independent from each other, and

$$\text{Cov}\{y, x^–\} \overset{\text{def}}{=} E\{y (x^– - x)^T\} = H \Sigma^{-},$$

by furthermore assuming that $m$ and $x$ are statistically independent. These are sensible assumptions, as usually the observation process is physically completely independent from the system development process, and the observation processes at different epochs are independent of each other.

We also obtain

$$\text{Cov}\{x^–, y\} = \Sigma^{-} H^T.$$

### 3.6 Updating

The update step is exploiting optimally the fact that the difference between the value of the observable calculated from the estimated state vector $x^–$ and the real observation has an expectancy of zero.

So, an enhanced estimator is constructed\(^6\),

$$x^+ = x^– + Ky = x^– + K (Hx^– - \xi) = x^– + K (H(x^– - x) + m),$$

so

$$(x^+ - x) = (I + KH) (x^– - x) + Km.$$

Here, the matrix $K$ is called the Kalman gain matrix.

Now according to definition 3.9 we may use this to derive the update equation for the state variance:

$$\Sigma^+ = (I + KH) \Sigma^{-} (I + KH)^T + KRK^T. \tag{3.26}$$

\(^6\)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 $i$ and $i + 1$ referring to the states before and after.
The “optimal” solution is obtained by choosing

\[ K = -\Sigma^{-1}H^T (H\Sigma^{-1}H^T + R)^{-1}, \]

which gives as the solution

\[ x^+ = x^- - K (Hx^- - \ell) = x^- - \Sigma^{-1}H^T (H\Sigma^{-1}H^T + R)^{-1}(Hx^- - \ell). \]  

(3.27)

if we call

\[ P \overset{\text{def}}{=} (H\Sigma^{-1}H^T + R)^{-1} \implies K = -\Sigma^{-1}H^TP, \]

we can open up equation 3.26 as

\[ \Sigma^+ = (I - \Sigma^{-1}H^TP) \Sigma^- (I - \Sigma^{-1}H^TP)^T + \Sigma^{-1}H^TP \Sigma, \]

in which

\[ I = (I - \Sigma^{-1}H^TP) \Sigma^- (I - \Sigma^{-1}H^TP)^T = \Sigma^- - \Sigma^{-1}H^TP\Sigma^- - \Sigma^{-1}H^TP\Sigma^- + \Sigma^{-1}H^TP\Sigma^- H^TP\Sigma^- \]

and

\[ III + II = \Sigma^{-1}H^TPH\Sigma^{-1}H^TPH\Sigma^- + \Sigma^{-1}H^TP \Sigma =
\]

\[ = \Sigma^{-1}H^TP (H\Sigma^{-1}H^T + R) \Sigma = \Sigma^{-1}H^TP \Sigma^- , \]

so

\[ \Sigma^+ = \Sigma^- - \Sigma^{-1}H^TPH\Sigma^- \Sigma^{-1}H^TPH\Sigma^- + \Sigma^{-1}H^TP \Sigma^- =
\]

\[ = \Sigma^- - \Sigma^{-1}H^T (H\Sigma^{-1}H^T + R)^{-1}H\Sigma^- . \]

Shorten this variance update equation still as follows:

\[ \Sigma^+ = \Sigma^- - \Sigma^{-1}H^T (H\Sigma^{-1}H^T + R)^{-1}H\Sigma^- = (I + KH) \Sigma^- , \]

(3.28)

based on the definition of the Kalman gain matrix K.

Perhaps more intuitively summarized:

\[ x^+ = x^- - \text{Cov}\{x^-, y\} \text{Var}^{-1}\{y\} y, \]

\[ \text{Var}\{x^+\} = \text{Var}\{x^-\} - \text{Cov}\{x^-, y\} \text{Var}^{-1}\{y\} \text{Cov}\{y, x^-\} , \]

(3.29)
which looks like a regression of the state vector $\mathbf{x}$ with respect to the “closing error” $\mathbf{y}$.

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

In the literature we can find many ways to calculate these equations effectively and precisely. The main issue nevertheless is, that the variance matrix of the “closing error”

$$\text{Var}\{\mathbf{y}\} = \mathbf{H}\Sigma\mathbf{H}^T + \mathbf{R}$$

is the size of vector $\mathbf{y}$. And the size of $\mathbf{y}$ is 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, not (like for example in traditional adjustment calculus) all of them at once.

### 3.7 The optimality of the update

The equations 3.29 are optimal in the sense of the least-squares adjustment method. Proving it can be done as follows, with a little simplification.

We start by calculating

$$\text{Cov}\{\mathbf{x}^+, \mathbf{y}\} = \text{Cov}\{\mathbf{x}^-, \mathbf{y}\} - \text{Cov}\{\mathbf{x}^-, \mathbf{y}\} \text{Var}^{-1}\{\mathbf{y}\} \text{Var}\{\mathbf{y}\} = 0, \quad (3.30)$$
The optimality of the update

remembering that, by definition, \( \text{Cov}\{y, y\} = \text{Var}\{y\} \). So the updated state vector \( x^+ \) is orthogonal to the “closing error vector” \( y \).

Suppose now that there existed an alternative updated state \( x^\times \), that was even still better than the standard update \( x^+ \). Write

\[
x^\times = x^+ + Cy
\]

for some coefficient matrix \( C \). Then, because of equation 3.30, we would have

\[
\text{Var}\{x^\times\} = \text{Var}\{x^+\} + C \text{Var}\{y\} C^T.
\]

Because \( \text{Var}\{y\} \) is positive definite, the expression

\[
\text{Var}\{x^\times\} - \text{Var}\{x^+\} = C \text{Var}\{y\} C^T
\]

is positive semidefinite, and

\[
\text{Var}\{x^\times\} - \text{Var}\{x^+\} = 0
\]

happens only if \( C = 0 \).

In other words, for an arbitrary linear combination

\[
z = \sum_i c_i x_i, \quad z^\times = \sum_i c_i x^\times_i, \quad z^+ = \sum_i c_i x^+_i,
\]

it holds that

\[
\text{Var}\{z^\times\} - \text{Var}\{z^+\} = cC \text{Var}\{y\} C^T
\]

in which \( c \triangleq \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \). This only vanishes if \( cC = 0 \), otherwise

\[
\text{Var}\{z^\times\} - \text{Var}\{z^+\} > 0.
\]

The two-dimensional case is presented graphically in figure 3.4.

So, the variance ellipse of the optimal estimator \( x^+ \) — more generally a (hyper-)ellipsoid — lies always entirely inside the variance ellipse of the alternative estimator \( x^\times \), or at worst touches it from the inside. The same holds for the variances of an arbitrary linear combination \( z \) of the components.
Examples and applications of the Kalman filter

4.1 Example 1: one-dimensional motion

In this example, we shall not bother with physical units. Feel free to assume metres for distances or co-ordinates and seconds for time.

**Question:**

Assume, in one spatial dimension, the dynamical model for the state vector

\[
x = \begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix}
\]

to be

\[
\frac{d^2}{dt^2} x = n,
\]

or as a matrix equation with velocity \( v \),

\[
\frac{d}{dt} \begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ n \end{bmatrix}.
\]

Here, \( n \) is white noise with an autocovariance of \( Q = 1 \). Furthermore, assume that the initial state is given as the estimate

\[
\begin{bmatrix} \hat{x}(0) \\ \hat{v}(0) \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \quad \Sigma(0) = \begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix},
\]

meaning that we do not actually have any real velocity information.

1. Propagate this state information forward to \( t = 5 \), i.e., calculate

\[
\hat{x}(5) = x^-(5), \quad \Sigma(5) = \Sigma^-(5).
\]
2. At $t = 5$, a further observation yielding a value of 3 is made:

$$\ell = x^-(5) + m,$$

in which $m$ is the observation noise, with variance $R = 3$.

(a) What does the $H$ matrix look like? And the $K$ Matrix?

(b) Calculate the $a$ posteriori state $x^+(5), \Sigma^+(5)$.

3. Calculate alternatively the outcome using a standard least-squares adjustment. The dynamic model is

$$\hat{x}(t) = \hat{x}(0) + \hat{v}(0) \cdot t,$$

unknowns to be estimated are $\hat{x}(0)$ and $\hat{v}(0)$, and observation equations

$$\ell_1 + v_1 = \hat{x}(0),$$

$$\ell_2 + v_2 = \hat{x}(5),$$

and the observation vector and its variance matrix

$$\ell = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad R = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}.$$  (4.1)

Answer:

1. $\hat{x}(5) = \hat{x}(0) + \hat{v}(0) \cdot 5 = 4$. Because the matrix $F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, we obtain the state transition matrix as

$$\Phi_5^0 = e^{F \Delta t} = \exp \left( \begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix} \right) =$$

$$= \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix},$$

because

$$\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}^n = 0, \ n > 1.$$
Example 1: one-dimensional motion

Then
\[ \Sigma(5) = \Phi_0^5 \Sigma(0) (\Phi_0^5)^T + Q \Delta t = \]
\[ = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 5 & 1 \end{bmatrix} + 5 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \]
\[ = \begin{bmatrix} 25002 & 5000 \\ 5000 & 1005 \end{bmatrix}. \]

2. The matrix \( H = \begin{bmatrix} 1 & 0 \end{bmatrix} \). So
\[ H \Sigma^{-1}H^T + R = 25002 + 3 = 25005. \]

The \( K \) matrix is
\[ K = -\Sigma^{-1}H^T (H \Sigma^{-1}H^T + R)^{-1} = - \begin{bmatrix} 25002 \\ 5000 \end{bmatrix} \cdot \frac{1}{25005} = \]
\[ = \begin{bmatrix} -0.99988 \\ -0.19996 \end{bmatrix}. \]

Next, we compute the "closing error"
\[ y = Hx^-(5) - \ell = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 0 \end{bmatrix} - 3 = 1. \]

Then,
\[ x^+(5) = x^-(5) + Ky = \]
\[ = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.99988 \\ 0.19996 \end{bmatrix} \cdot 1 = \begin{bmatrix} 3.00012 \\ -0.19996 \end{bmatrix}. \]

We can project this back to \( t = 0 \):
\[ \hat{x}(0) = x^+(5) - v^+(5) \cdot 5 = \]
\[ = 3.00012 - (-0.19996) \cdot 5 = 3.9999, \quad (4.2) \]
\[ \hat{v}(0) = v^+(5) = -0.19996. \]

\(^1\)Notation: \( \Sigma^- = \Sigma(5) \) is now the \( a \) priori state variance matrix, i.e., just before making the observation at epoch \( t = 5 \).
The updated state variance matrix $\Sigma^+ (5)$ is

$$\Sigma^+ (5) = (I + KH) \Sigma^- (5) =$$

$$= \left( I + \begin{bmatrix} -0.99988 & 0.19996 \\ 0.19996 & -0.99988 \end{bmatrix} \right) \begin{bmatrix} 25002 & 5000 \\ 5000 & 1005 \end{bmatrix} =$$

$$= \begin{bmatrix} 2.99964 & 0.59988 \\ 0.59988 & 5.19996 \end{bmatrix}.$$  

3. The design matrix $A$ is

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 5 \end{bmatrix},$$

and the observation vector $\ell$ and the observation variance matrix $R$ are given, equations 4.1. We obtain

$$A^T R^{-1} A = \frac{1}{6} \begin{bmatrix} 5 & 10 \\ 10 & 50 \end{bmatrix},$$

$$\hat{x} = (A^T R^{-1} A)^{-1} A^T R^{-1} \ell = \frac{1}{5} \begin{bmatrix} 20 & -1 \end{bmatrix}.$$

This is the same result, practically, as that under point 2, equations 4.2. For the solution variance we find

$$\text{Var} \{ \hat{x} \} = (A^T R^{-1} A)^{-1} = \frac{1}{5} \begin{bmatrix} 10 & -2 \\ -2 & 1 \end{bmatrix},$$

which is not directly comparable to the earlier result as it refers to time $t = 0$. Furthermore, the Kalman solution contains the effect of the dynamic noise $Q$, which is not along in the standard least-squares solution.

### 4.2 Example 2: spinning wheel

**Question:**

In an industrial machine there is 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.”

1. Write the state vector of this system. How many elements are needed?
2. Write the dynamical model of the system.

3. A reflective prism is attached to the edge of the wheel in order to do measurements. The rotation is monitored by using laser distance measurement. The measuring device is at a great distance from the machine, within the plane of the wheel.

Write the observational model.

4. Linearize the observational model.

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. We shall see in question 2 that this makes it a good idea to include also the angular velocity into the state vector.

Thus we obtain for the state vector:

$$x(t) = \begin{bmatrix} \alpha(t) \\ \omega(t) \end{bmatrix}.$$

2. The dynamical model in the Kalman filter is a system of equations of the form

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

where $x$ is the system’s state vector and $n$ is the dynamical noise vector.

In our case we have the state vector above. We can write

$$\frac{d}{dt} \begin{bmatrix} \alpha \\ \omega \end{bmatrix} = \begin{bmatrix} \omega \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ n_\omega \end{bmatrix},$$

in which the first equation $\frac{d}{dt} \alpha = \omega$ expresses the definition of angular velocity $\omega$, and the second equation $\frac{d}{dt} \omega = n_\omega$, expresses the given fact that the angular acceleration has the properties of white noise.

We observe that the dynamical model found is linear.

3. If we observe the distance to a prism on the edge of the wheel from far away, we may write the observation equation

$$\ell = d + r \cos \alpha + m.$$
We reckon \(\alpha\) from the prism position farthest away from the observing instrument. We assume \(d\), the distance between the instrument and the centre of the wheel, known. If it is not, \(d\) should be added to the state vector with the added dynamical equation \(\frac{d}{dt}d = 0\).

4. This model is non-linear, i.e., the dependence of the observation quantity on the state vector element is a cosine. We linearize as follows: define consistent approximate values for which

\[ \ell_0 = d + r \cos \alpha_0 \]

and subtract this from the above. The result is a Taylor expansion truncated after the first, linear term in \(\Delta \alpha\):

\[ \Delta \ell = r \frac{\partial}{\partial \alpha} \cos \alpha \bigg|_{\alpha = \alpha_0} \cdot \Delta \alpha + m. \]

Here, the logical definitions \(\Delta \ell = \ell - \ell_0\) and \(\Delta \alpha = \alpha - \alpha_0\) have been applied.

Partial differentiation yields

\[ \Delta \ell = -r \sin \alpha_0 \Delta \alpha + m, \]

a standard linear Kalman observation equation, type

\[ \ell = Hx + m, \]

if we write formally

\[ \ell = \begin{bmatrix} \Delta \ell \end{bmatrix}, \quad H = \begin{bmatrix} -r \sin \alpha_0 & 0 \end{bmatrix}, \]

\[ x = \begin{bmatrix} \Delta \alpha \\ \Delta \omega \end{bmatrix}, \quad m = \begin{bmatrix} m \end{bmatrix}. \]

4.3 Example 3: parachute jumper

Question:

1. Write the dynamic equations for a parachute jumper in one dimension (only the height co-ordinate \(z\)). The gravity acceleration \(g\) is a constant, the braking acceleration caused by air drag is proportional to the velocity of falling and the air density, which may be described by the equation

\[ \rho = \rho_0 e^{-z/\sigma}. \]
Example 3: parachute jumper

The constant $\sigma$ is the scale height of the atmosphere, $\rho_0$ 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 tacheometer and touch-down point is given. The jumper comes down vertically, there is no wind.

Write the observational model.

**Answer:**

1. The dynamic model is ($k$ is a constant$^2$):

$$\frac{d^2}{dt^2} z = -g + k\dot{z}\rho + \mathbf{n} = -g + k\dot{z}\rho_0 e^{-z/\sigma} + \mathbf{n}.$$

Define the state vector as $\begin{bmatrix} z & \dot{z} \end{bmatrix}^T$ and obtain as the dynamic model the first-order differential equations

$$\frac{d}{dt} \begin{bmatrix} \dot{z} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} \dot{z} \\ -g + k\dot{z}\rho_0 e^{-z/\sigma} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{n} \end{bmatrix}.$$  

This is non-linear; if we write

$$\begin{bmatrix} \dot{z} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} z^{(0)} \\ \dot{z}^{(0)} \end{bmatrix} + \begin{bmatrix} \Delta z \\ \Delta \dot{z} \end{bmatrix},$$

in which

$$\frac{d}{dt} \begin{bmatrix} z^{(0)} \\ \dot{z}^{(0)} \end{bmatrix} = \begin{bmatrix} \dot{z}^{(0)} \\ -g + k\dot{z}^{(0)}\rho_0 e^{-z^{(0)}/\sigma} \end{bmatrix}.$$ 

This system of equations may be integrated if initial conditions are given.

Linearisation:

$$\Delta(\dot{z} e^{-z/\sigma}) \approx -(\dot{z}/\sigma) e^{-z/\sigma} \Delta z + e^{-z/\sigma} \Delta \dot{z}$$

$^2$ A negative constant, because $\dot{z}$ is negative as well when $z$ grows upward.
and

\[
\frac{d}{dt} \begin{bmatrix} \Delta \hat{z} \\ \Delta \dot{\hat{z}} \end{bmatrix} \approx \begin{bmatrix} \Delta \hat{z} \\ \Delta \dot{\hat{z}} \end{bmatrix} \approx \begin{bmatrix} -k\rho_0 \frac{\dot{\hat{z}}(0)}{\sigma} e^{-\frac{z(0)}{\sigma}} \Delta \hat{z} + k\rho_0 e^{-\frac{z(0)}{\sigma}} \Delta \dot{\hat{z}} \\ -k\rho_0 \frac{\dot{\hat{z}}(0)}{\sigma} e^{-\frac{z(0)}{\sigma}} + k\rho_0 e^{-\frac{z(0)}{\sigma}} \end{bmatrix} \begin{bmatrix} \Delta \hat{z} \\ \Delta \dot{\hat{z}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} n \end{bmatrix},
\]

the linearized version of the dynamic model.

2. Let the horizontal distance between the touch-down point of the parachutist and the tacheometer be \( \rho \). Then the measured distance is

\[
s = \sqrt{\rho^2 + z^2}
\]

and the observation equation

\[
s = \sqrt{\rho^2 + z^2} + m.
\]

Linearization — \( s = s_0 + \Delta s \) with \( s_0 = \sqrt{\rho^2 + z_0^2} \) — yields

\[
\Delta s = \frac{z_0}{s_0} \Delta z + m = \begin{bmatrix} \frac{z_0}{s_0} & 0 \end{bmatrix} \begin{bmatrix} \Delta \hat{z} \\ \Delta \dot{\hat{z}} \end{bmatrix} + m.
\]

### 4.4 Modelling of realistic statistical behaviour

Coloured noise, specifically Gauss-Markov processes, are very often used to model stochastic processes found in real life. Say, for example, that we know that a measured stochastic process \( x \) consists of the quantity \( z \) which we are interested in — which may vary rapidly on both sides of zero — and a systematic “disturbance” or bias, which we want to get rid of. We also know that this disturbance is slowly varying, on a time scale described by the constant \( \tau_b \). Call the disturbance \( b \). We may write the state vector as \( \begin{bmatrix} s & b \end{bmatrix}^T \) and the dynamic equations as, say,

\[
\frac{d}{dt} \begin{bmatrix} s \\ b \end{bmatrix} = \begin{bmatrix} -1/\tau_s & 0 \\ 0 & -1/\tau_b \end{bmatrix} \begin{bmatrix} s \\ b \end{bmatrix} + \begin{bmatrix} n_s \\ n_b \end{bmatrix}.
\]

Here, \( \tau_b \) is the long time constant of the bias process, which will thus be slowly varying. For the time constant \( \tau_s \) we may choose a much shorter
value. However, it should be chosen realistically. If measurements are obtained at a time interval \( \Delta t \), then we must have \( \tau_s \gg \Delta t \) in order for the process \( \hat{s} \) to be realistically determinable from the observations.

The observation equation is

\[
\ell = \hat{s} + b + m,
\]

with \( m \) (variance \( R \)) representing the observational noise of uncertainty. The variance of \( m \) is given as \( R \). If observations are obtained at a sufficient density in time, we may obtain separate estimates for the signal process \( \hat{s} \) and the slowly varying bias \( \hat{b} \). In order for this to work, we should attach realistic autocovariances to the processes \( n_s \) and \( n_b \). Even then, it is a requirement in this case that \( E(\hat{s}) = 0 \). If it is not, the systematic part of \( \hat{s} \) will end up in the estimate \( \hat{b} \) produced by the filter.

This is a case of spectral filtering by Kalman filter. The low-frequency part, including zero frequency, goes to \( \hat{b} \); the high-frequency part goes to \( \hat{s} \). However, 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 \( \hat{c} \). It has a correlation length \( \tau_c \). If we are interested in the lower-frequency constituents of the signal \( \hat{s} \), we may again apply a Kalman filter:

\[
\frac{d}{dt} \begin{bmatrix} \hat{s} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} -1/\tau_s & 0 \\ 0 & -1/\tau_c \end{bmatrix} \begin{bmatrix} \hat{s} \\ \hat{c} \end{bmatrix} + \begin{bmatrix} n_s \\ n_c \end{bmatrix}.
\]

We choose \( \tau_s \) according to the part of the spectrum of \( \hat{s} \) that we are interested in — but always \( \tau_s > \tau_c \). The time constant \( \tau_c \) should be chosen realistically, to capture and remove to the maximum extent the real noise present in the process. The observation equation is again

\[
\ell = \hat{s} + \hat{c} + m.
\]

The earlier described technique for extracting a rapidly varying signal from a background of slowly varying bias was used (Tapley and Schutz, 1975) already in 1975 for extracting signal on underground mass concentrations or mascons on the Moon from Lunar Orbiter tracking data. It is called “Dynamic Model Compensation”.

---

The image contains a page from a text discussing the modeling of realistic statistical behavior. It explains the importance of choosing parameters realistically, such as time intervals and system constants, to accurately determine signals from observations. The text delves into the details of observing equations and the application of Kalman filters for spectral filtering, highlighting the distinction between high-frequency and low-frequency components. It concludes with a reference to a specific technique used in 1975 for extracting signals from underground mass concentrations on the Moon using Lunar Orbiter tracking data.
4.5 The Kalman filter as sequential adjustment

We may write the update step of the Kalman filter also as a parametric adjustment problem.

The “observations” are the real observation vector for this epoch \( \ell' \) and the a priori estimated state vector \( x^- \). The parametric observation equations in the standard form are

\[
\begin{bmatrix}
\ell' \\
x^-
\end{bmatrix}
+ \begin{bmatrix}
\psi' \\
\psi''
\end{bmatrix} = \begin{bmatrix}
A \\
H \\
I
\end{bmatrix}
\begin{bmatrix}
\hat{x} \\
\chi
\end{bmatrix}.
\]

The design matrix is seen to be

\[
A = \begin{bmatrix}
H \\
I
\end{bmatrix}.
\]

The variance matrix of the “observations” is

\[
\Sigma = \text{Var}\left\{ \begin{bmatrix}
\ell' \\
x^-
\end{bmatrix} \right\} = \begin{bmatrix}
R & 0 \\
0 & \Sigma^-
\end{bmatrix},
\]

and the posterior state-vector solution is

\[
x^+ = (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} \ell' = \left( H^T R^{-1} \Sigma^{-1} H + (\Sigma^-)^{-1} \right)^{-1} \left( H^T R^{-1} \ell' + (\Sigma^-)^{-1} x^- \right).
\] (4.3)

As the posterior state variance we obtain

\[
\Sigma^+ = \left( H^T R^{-1} \Sigma^{-1} H + (\Sigma^-)^{-1} \right)^{-1}.
\] (4.4)

Now we exploit the second formula derived in appendix B:

\[
(A + UCV)^{-1} = A^{-1} - A^{-1} U \left( C^{-1} + VA^{-1} U \right)^{-1} VA^{-1},
\]

in this way:

\[
\left( H^T R^{-1} \Sigma^{-1} H + (\Sigma^-)^{-1} \right)^{-1} = \Sigma^- - \Sigma^- H^T \left( R + H \Sigma^- H^T \right)^{-1} H \Sigma^-.
\]

Substitution yields

\[
x^+ = \left( \Sigma^- - \Sigma^- H^T \left( R + H \Sigma^- H^T \right)^{-1} H \Sigma^- \right) \left( H^T R^{-1} \ell' + (\Sigma^-)^{-1} x^- \right) = \Sigma^- H^T R^{-1} \ell' + x^- - \Sigma^- H^T \left( R + H \Sigma^- H^T \right)^{-1} H (\Sigma^- H^T R^{-1} \ell' + x^-) = x^- + \Sigma^- H^T R^{-1} \ell' - \Sigma^- H^T \left( R + H \Sigma^- H^T \right)^{-1} H \Sigma^- H^T R^{-1} \ell' - \Sigma^- H^T \left( R + H \Sigma^- H^T \right)^{-1} H x^-,
\]
Using the Kalman filter “from both ends”

in which

\[
I = -\Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} H\Sigma^{-1}H^T R^{-1}\ell' = \\
= -\Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} \left( R + H\Sigma^{-1}H^T \right) R^{-1}\ell' + \Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} R^{-1}R = \\
= -\Sigma^{-1}H^T R^{-1}\ell' + \Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} \ell',
\]
yielding

\[
x^+ = x^- + \Sigma^{-1}H^T R^{-1}\ell' = -\Sigma^{-1}H^T R^{-1}\ell' + \\
+ \Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} \ell' - \Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} Hx^- = \\
= x^- + \Sigma^{-1}H^T \left( H\Sigma^{-1}H^T + R \right)^{-1} (\ell' - Hx^-). \quad (4.5)
\]

In addition

\[
\Sigma^+ = \Sigma^- - \Sigma^{-1}H^T \left( R + H\Sigma^{-1}H^T \right)^{-1} H\Sigma^-.
\quad (4.6)
\]

The equations 4.5 and 4.6 are precisely the update equations of the Kalman filter. Compared to the equations 4.3 and 4.4, the matrix to be inverted has the size of the vector of observables \(\ell'\) and not that of the state vector \(x\). Often the matrix size is even \(1 \times 1\), i.e., a simple number\(^3\). Being able to compute inverse matrices more quickly makes real-time applications easier.

From the preceding we see, that sequential adjustment is the same as Kalman filtering in the case that the state vector is constant. Although the computation procedure in adjustment generally is parametric adjustment (observation equations), whereas in the Kalman case, condition-equations adjustment is used.

**4.6 Using the Kalman filter “from both ends”**

In airborne gravimetry, chapter 7, 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 a least-squares collocation are available.

However, it is also possible to harness the Kalman filter for this, by processing the data both forward and backward in time and optimally

\(^3\) 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, i.e., separately.
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 and normal equations

If we have available the observations \(\ell_i, i = 1, \ldots, n\) and the functional model is the system of differential equations

\[
\frac{d}{dt}x(t) = F \cdot x(t)
\]

— written here non-stochastically and without dynamic noise \(n\) — we may write

\[x(t_i) = \Phi_0^i x(t_0),\]

in which \(\Phi_0^i\) is the state transition matrix to be computed. Thus, the observation equations may be written into the standard form

\[
\ell_i + v_i = H_i x(t_i) = H_i \Phi_0^i x(t_0), \quad i = 1, \ldots, n,
\]

a traditional system of observation equations

\[\ell + v = A \hat{x},\]

in which the design matrix, the vectors of observations, residuals, and unknowns are

\[
A = \begin{bmatrix}
H_1 \Phi_0^1 & \cdots & \cdots & \cdots \\
\vdots & & & \\
H_i \Phi_0^i & \cdots & \cdots & \cdots \\
\vdots & & & \\
H_n \Phi_0^n & \cdots & \cdots & \cdots \\
\end{bmatrix}, \quad \ell = \begin{bmatrix}
\ell_1 \\
\vdots \\
\ell_i \\
\vdots \\
\ell_n \\
\end{bmatrix}, \quad v = \begin{bmatrix}
v_1 \\
\vdots \\
v_i \\
\vdots \\
v_n \\
\end{bmatrix}, \quad \hat{x} = \begin{bmatrix}
x(t_0) \\
\end{bmatrix}.
\]

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

\[
\ell = \begin{bmatrix}
\ell < \\
\ell > \\
\end{bmatrix}, \quad R = \begin{bmatrix}
R & 0 \\
0 & R \\
\end{bmatrix}, \quad A = \begin{bmatrix}
A < \\
A > \\
\end{bmatrix}.
\]
Using the Kalman filter “from both ends”

This way, separate normal equations are formed:

\[
\begin{align*}
\left( A^T R^{-1} A \right) \hat{x} &= A^T R^{-1} \xi, \\
\left( A^T R^{-1} A \right) \hat{\xi} &= A^T R^{-1} \xi,
\end{align*}
\]

with solutions

\[
\begin{align*}
\hat{x} &= \left( A^T R^{-1} A \right)^{-1} A^T R^{-1} \xi, \\
\hat{\xi} &= \left( A^T R^{-1} A \right)^{-1} A^T R^{-1} \xi,
\end{align*}
\]

and separate solution variances

\[
\Sigma = \left( A^T R^{-1} A \right)^{-1}, \quad \Sigma = \left( A^T R^{-1} A \right)^{-1}.
\]

On the other hand, the single normal equation of the full adjustment is

\[
A^T R^{-1} A \hat{x} = A^T R^{-1} \xi.
\]

We assume the observations \( \xi \) and \( \ell \) to be statistically independent. This is why the matrix \( R \) is block diagonal, and we may decompose the normal equation as follows:

\[
\begin{align*}
\left[ \begin{array}{c} A \\ A \end{array} \right] \left[ \begin{array}{c} R \\ 0 \\ 0 \\ R \end{array} \right]^{-1} \left[ \begin{array}{c} A \\ A \end{array} \right] &= \left[ \begin{array}{c} A \\ A \end{array} \right] \left[ \begin{array}{c} R \\ 0 \\ 0 \\ R \end{array} \right]^{-1} \left[ \begin{array}{c} \xi \\ \xi \end{array} \right],
\end{align*}
\]

or

\[
\begin{align*}
\left( A^T R^{-1} A + A^T R^{-1} A \right) \hat{x} &= A^T R^{-1} \xi + A^T R^{-1} \xi
\end{align*}
\]

with solution

\[
\begin{align*}
\hat{x} &= \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} \left( A^T R^{-1} \xi + A^T R^{-1} \xi \right) = \\
&= \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} \left( A^T R^{-1} A \right) \left( A^T R^{-1} A \right)^{-1} A^T R^{-1} \xi + \\
&+ \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} \left( A^T R^{-1} A \right) \left( A^T R^{-1} A \right)^{-1} A^T R^{-1} \xi = \\
&= \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} \left( A^T R^{-1} A \right) \hat{x} + \\
&+ \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} \left( A^T R^{-1} A \right) \hat{x},
\end{align*}
\]

and

\[
\Sigma = \left( A^T R^{-1} A + A^T R^{-1} A \right)^{-1} = \left( \Sigma^{-1} + \Sigma^{-1} \right)^{-1}
\]

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:

\[
\hat{x} = \left( P + P \right)^{-1} \left( P \hat{x} + P \hat{x} \right), \quad \Sigma = \left( P + P \right)^{-1}, (4.8)
\]

in other words the weighted average of the partial solutions, with weight matrices \( P \overset{\text{def}}{=} A^T R^{-1} A \) and \( P \overset{\text{def}}{=} A^T R^{-1} A. \)
4.6.2 The forward-backward Kalman filter

Important here is now, that the partial tasks — “before”, <, and “after”, > — can be solved also with the help of the Kalman filter! In other words, we may for an arbitrary observation epoch \( t_i \) calculate separately

1. The solution of the Kalman filter from the starting epoch \( t_0 \) forward, by integrating the dynamical model and updating the state vector and its variance matrix using the observations \( \ell_1, \ldots, \ell_i, \ell_{i+1}, \ldots, \ell_n \), and

2. The Kalman filter solution from the final moment \( t_n \) backward in time integrating the dynamic modal, updating the state vector and the variance matrix using the observations \( \ell_n, \ldots, \ell_{i+1}, \ell_i, \ldots, \ell_1 \) (in reverse order). The Kalman filter equation to be used for that is, based on equation 4.7:

\[
\frac{d}{dt'} x(t') = -F \cdot x(t'),
\]

in which \( t' = -t \).

3. Combining the partial solutions obtained into a total solution using the above equations.

In this way, the advantages of the Kalman method may be exploited also in a post-processing situation. And note that equation 4.7, which we used in our derivation, contains no dynamic noise \( n \). In reality this limitation does not exist: one can include dynamic noise in the Kalman-filter model. Integrate forward and backward, 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: random walk in both directions

A random walk is described by equation 2.15:

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

with the autocovariance according to equation 2.9

$$A_x(t, t) = Q(t - t_0).$$

The random walk is observed at two points in time, \(t_1\) and \(t_2\), observation values

\[ \ell_1 = x(t_1), \quad \ell_2 = x(t_2), \]

both assumed exact.

Then, the forward solution is

\[ \hat{x}(t) = \ell_1, \quad \Sigma(t) = Q(t - t_1), \]

and the backward solution

\[ \hat{x}(t) = \ell_2, \quad \Sigma(t) = Q(t - t_2). \]

Do the weighted averaging between forward and backward. Equation 4.8:

\[
\hat{x}(t) = \frac{Q(t_2 - t)}{Q(t - t_1) + Q(t_2 - t)} \hat{x}(t) + \frac{Q(t - t_1)}{Q(t - t_1) + Q(t_2 - t)} \hat{x}(t) = \frac{t_2 - t}{t_2 - t_1} \hat{x}(t) + \frac{t - t_1}{t_2 - t_1} \hat{x}(t),
\]

amounting to linear interpolation between the observation epochs. Also

\[
\Sigma(t) = \left( \left( \Sigma(t) \right)^{-1} + \left( \Sigma(t) \right)^{-1} \right)^{-1}
= Q \left( \frac{1}{t - t_1} + \frac{1}{t_2 - t} \right)^{-1} = Q \frac{(t - t_1) (t_2 - t)}{t_2 - t_1},
\]
a parabola, being zero at both ends and maxing out in the middle at \(\frac{1}{4}Q\). See figure.

4.6.4 Example: estimating a constant

Let \(x\) be an unknown constant to be estimated. It has been observed at epoch 1, observation value 7, mean error \(\pm 2\), and at epoch 2, observation value 5, mean error \(\pm 1\).
Examples and applications of the Kalman filter

Figure 4.2. Random walk, solution in both directions. Solution and uncertainty.

Kysymys Formulate the observation equations of an ordinary adjustment problem and the variance matrix of the observation vector. Compute \( \hat{x} \).

Vastaus

\[
\ell + v = A \hat{x},
\]

in which

\[
\ell = \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \quad R = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.
\]

Solution:

\[
\hat{x} = (A^T R^{-1} A)^{-1} A^T R^{-1} \ell =
\]

\[
= \frac{4}{5} \cdot \begin{bmatrix} \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \end{bmatrix} = \frac{27}{5} = 5.4.
\]

Variance matrix:

\[
\Sigma = (A^T R^{-1} A)^{-1} = \frac{4}{5} = 0.8.
\]

Kysymys Write the dynamical equations for the Kalman filter. Remember that \( x \) is a constant.

Answer The general dynamical equation in the discrete case is

\[
x(t_{k+1}) = \Phi_k^{k+1} x(t_k) + \omega_k^{k+1}
\]

in which \( \Phi_k^{k+1} = \begin{bmatrix} 1 \end{bmatrix} \), a size 1 \times 1 unit matrix, and \( \omega_k^{k+1} = 0 \): deterministic motion, no dynamic noise. So

\[
x(t_{k+1}) = x(t_k).
\]

Alternatively we write the differential equation:

\[
\frac{dx}{dt} = F \cdot x + n
\]
Using the Kalman filter “from both ends”

In this case there is no dynamic noise, \( n = 0 \):

\[
\frac{dx}{dt} = 0.
\]

**Kysymys** Write the update equations for the Kalman filter — superscript “−” is *a priori*, “+” *a posteriori*:

\[
x_k^+ = x_k^- + K_k (\ell_k - H_k x_k^-), \quad \Sigma_k^+ = (I - K_k H_k) \Sigma_k^-,
\]

in which the gain matrix

\[
K_k = \Sigma_k^- H_k^T (R_k + H_k^T \Sigma_k^- H_k)^{-1}.
\]

How do the matrices \( H \) and \( K \) look in this case?

**Answer** Because in his case the observation \( \ell_k = x_k \) — we observe directly the state — we have \( H_k = \begin{bmatrix} 1 \end{bmatrix} \), a size \( 1 \times 1 \) matrix, the only element of which is 1.

\[
K_k = \frac{\Sigma_k^-}{R_k + \Sigma_k^-}.
\]

If the original \( \Sigma_k^- \) is large, then \( K \approx 1 \).

\[
x_k^+ = x_k^- + \frac{\Sigma_k^-}{R_k + \Sigma_k^-} (\ell_k - x_k^-) =
\]

\[
= \frac{\Sigma_k^-}{R_k + \Sigma_k^-} \ell_k + \frac{R_k}{R_k + \Sigma_k^-} x_k^- = \frac{\Sigma_k^- \ell_k + R_k x_k^-}{R_k + \Sigma_k^-},
\]

\[
\Sigma_k^+ = (1 - K_k) \Sigma_k^- = \frac{R_k}{R_k + \Sigma_k^-} \Sigma_k^-.
\]

- The *a posteriori* state \( x_k^+ \) is the weighted average of the *a priori* state \( x_k^- \) and the observation \( \ell_k \).
- the poorer the *a priori* state variance \( \Sigma_k^- \) compared to the observation precision \( R_k \), the more the updated state variance \( \Sigma_k^+ \) will improve.

**Kysymys** Calculate manually through both observation events and give the *a posteriori* state estimate \( \hat{x}_1 \) and its variance matrix. The initial value of the state \( x \) is 0, and for its initial variance matrix — i.e., variance — “numerically infinite”:

\[
\Sigma_0^+ = 100 = \Sigma_1^-.
\]

**Answer** First step:

\[
K_1 = 100 (4 + 100)^{-1} = \frac{100}{104}.
\]
so
\[ x_1^+ = x_1^- + K_1 (\ell_1 - x_1^-) = 0 + \frac{100}{104} (7 - 0) = 6.73 = x_2^-, \]
\[ \Sigma_1^+ = (I - K_1) \Sigma_1^- = \left(1 - \frac{100}{104}\right) 100 = \frac{400}{104} = 3.85 = \Sigma_2^- . \]

Second step:
\[ K_2 = 3.85 \left(1 + 3.85\right)^{-1} = 0.79, \]
\[ x_2^+ = x_2^- + K_2 (\ell_2 - x_2^-) = 6.73 + 0.79 (5 - 6.73) = \]
\[ = 6.73 - 0.79 \cdot 1.73 = 5.36. \]
\[ \Sigma_2^+ = (I - K_2) \Sigma_2^- = (1 - 0.79) \cdot 3.85 = 0.81. \]

### Self-test questions

1. In the above example 4.1, add the physical units for \( x, v, n, Q, \Sigma, \Phi, v_t, H, R \) and \( K \).

2. What is in the Kalman filter the state vector? ⭐
   
   (a) a vector of on/off bits describing the components of the system that are enabled or disabled
   
   (b) an abstract vector containing all parameters needed to completely describe the state of the system
   
   (c) the three-dimensional vector describing the current location of the vehicle in space
   
   (d) a vector of observation values used to estimate the state of the system.

3. How many elements does the state vector contain when describing the motion in space of a point object? ⭐

4. And how many elements does it contain if the object is an extended, rigid object, e.g., an aircraft? ⭐

5. Describe the dynamic model of the Kalman filter. ⭐

6. Describe the observation model of the Kalman filter. ⭐

7. What does the dynamic noise describe? ⭐
   
   (a) the imprecision or uncertainty by which the development over time of the system may be derived from the observations made
Exercise 4–1: Simple Kalman-filter example

(b) the acoustic or electronic noise level of the engine of the vehicle being tracked
(c) the imprecision with which the development over time of the system may be modelled by the dynamic model
(d) that part of the noise contribution that is dependent on time, not constant.

8. What does the observation noise describe? ★
(a) the imprecision or uncertainty of the observations made of the system
(b) the acoustic or electronic noise level of the observation instrument used
(c) the imprecision or uncertainty with which the development over time of the system may be derived from the observations made
(d) the extent to which the observations made fail to constrain the uncertainty of the state vector.


10. Show that random walk and white noise are limiting cases of Gauss–Markov.

11. How many elements does the state vector have in the above satellite orbit determination problem?

12. What is the dimension of the constant $k$ in the above parachute jumper problem? ★
(a) $\text{time}^{-1}$
(b) $\frac{\text{mass}}{\text{length}^3 \text{time}}$
(c) $\frac{\text{length}^3}{\text{mass time}}$
(d) $\frac{\text{mass}}{\text{time}}$

Exercise 4–1: Simple Kalman-filter example

1. Consider the following dynamical model, for a one-element state vector $x$:
\[
\frac{d}{dt} x = 0 + n.
\]
where \( \mathbf{n} \) is white noise with an autocovariance of \( Q = 1 \text{ m}^2/\text{s} \). The initial state estimate is \( \hat{x}(0) = 0 \text{ m} \) and the \textit{a priori} state variance as \( \Sigma(0) = 100 \text{ m}^2 \). Compute the state estimate for the epoch \( t = 3 \text{ s} \), i.e., \( \hat{x}^{-}(3) \) and \( \Sigma^{-}(3) \).

2. At epoch \( t = 3 \text{ s} \) we observe \( x \): the observation equation is

\[
\ell = x + m,
\]

where \( m \) is the observational uncertainty or “noise”, of which the variance \( R \) is given\(^4\). Our observation \textit{value} is \( \ell = 10 \text{ m} \). Compute

(a) the Kalman gain matrix \( K \), and

(b) the \textit{a posteriori} state estimate \( x^{+}(3) \) and its variance \( \Sigma^{+}(3) \).

\[\text{Exercise 4–2: A bit more complicated Kalman-filter example}\]

Assume now, that the dynamical model is

\[
\frac{\text{d}}{\text{d}t} \begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{n} \end{bmatrix},
\]

with the autocovariance of \( \mathbf{n} \) being \( Q = 50 \text{ m}^2/\text{s} \). Given is the initial state estimate \( \hat{x}(0) = \hat{v}(0) = 0 \) and its variance matrix

\[
\Sigma(0) = \begin{bmatrix} 100 \text{ m}^2 & 0 \\ 0 & 100 \text{ m}^2/\text{s}^2 \end{bmatrix}.
\]

1. Compute the \textit{state transition matrix} \( \Phi_{0}^{3} \),

2. Compute the \textit{a priori} state estimate

\[
x^{-}(3) = \hat{x}(3), \quad v^{-}(3) = \hat{v}(3), \quad \Sigma^{-}(3)
\]

at epoch \( t = 3 \text{ s} \). Note that now, \( \hat{x} \) and \( \hat{v} \) are \textit{correlated}.

3. The observation equation is again

\[
\ell = x + m.
\]

If \( \ell = 10 \text{ m} \), compute

(a) the Kalman gain matrix \( K \), and

\[\text{\textcopyright Take for the variance, the day number of the month of your birthday!}\]
Exercise 4–2: A bit more complicated Kalman-filter example

(b) the *a posteriori* state variance $\Sigma^+(3)$.

(c) Just for fun, compute the determinants of the variance matrices $\Sigma(0)$, $\Sigma^-(3)$, and $\Sigma^+(3)$. This determinant, or surface area of the error ellipse, is a good measure for the imprecision of the state vector.

You may assume $R$ to be the same as in the previous exercise, or keep it as a symbol.
5

5.1 Principle

Inertial navigation is based on Isaac Newton’s first law of mechanics, 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.

In inertial navigation, the following quantities are measured continuously:

1. the three-dimensional acceleration of the object (vehicle),

\[
\frac{d^2x(t)}{dt^2} = \frac{d^2x'(t)}{dt^2} i' + \frac{d^2y'(t)}{dt^2} j' + \frac{d^2z'(t)}{dt^2} k',
\]

in object co-ordinates:

\[
\left[ \frac{d^2x'(t)}{dt^2} \quad \frac{d^2y'(t)}{dt^2} \quad \frac{d^2z'(t)}{dt^2} \right]^T.
\]

Here,

\[ x(t) \overset{\text{def}}{=} x'(t) i' + y'(t) j' + z'(t) k' \]

is the object’s three dimensional location,

\[ \bar{x}'(t) \overset{\text{def}}{=} \left[ \begin{array}{c} x'(t) \\ y'(t) \\ z'(t) \end{array} \right]^T \]
Inertial navigation is the vector of its co-ordinates in the object coordinate frame, an orthonormal basis of which is \( \{i', j', k'\} \).

2. The attitude of the vehicle:

\[
R = R_3(\alpha_3)R_2(\alpha_2)R_1(\alpha_1) = \\
\begin{bmatrix}
\cos \alpha_3 & \sin \alpha_3 & 0 \\
-\sin \alpha_3 & \cos \alpha_3 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\cos \alpha_2 & 0 & -\sin \alpha_2 \\
0 & 1 & 0 \\
\sin \alpha_2 & 0 & \cos \alpha_2
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \alpha_1 & \sin \alpha_1 \\
0 & -\sin \alpha_1 & \cos \alpha_1
\end{bmatrix}
\]

\[
= \\
\begin{bmatrix}
\cos \alpha_2 \cos \alpha_3 & \cos \alpha_1 \sin \alpha_3 + \sin \alpha_1 \sin \alpha_2 \cos \alpha_3 & \sin \alpha_1 \sin \alpha_3 - \cos \alpha_1 \sin \alpha_2 \cos \alpha_3 \\
-\cos \alpha_2 \sin \alpha_3 & \cos \alpha_1 \cos \alpha_3 - \sin \alpha_1 \sin \alpha_2 \sin \alpha_3 & \sin \alpha_1 \cos \alpha_3 + \cos \alpha_1 \sin \alpha_2 \sin \alpha_3 \\
\sin \alpha_2 & -\sin \alpha_1 \cos \alpha_2 & \cos \alpha_1 \cos \alpha_2
\end{bmatrix}
\]

We may also write in global co-ordinates, that

\[ x(t) \overset{\text{def}}{=} x(t) \, i + y(t) \, j + z(t) \, k, \]

in which

\[ \mathbf{x}(t) \overset{\text{def}}{=} \begin{bmatrix} x(t) & y(t) & z(t) \end{bmatrix}^T \]

is the vector of global location co-ordinates, and \( \{i, j, k\} \) the orthonormal base of the global co-ordinate frame.

\( R \) is now the transformation matrix between the global and object coordinates:

\[ \mathbf{x}'(t_0) = R(t_0) \mathbf{x}(t_0), \]

at the moment of beginning of the journey \( t_0 \). The vectors \( \mathbf{x} \) and \( \mathbf{x}' \) are global (often inertial) and object coordinates, respectively.

The attitude is described by three unknowns, \( \alpha_i(t) \), \( i = 1, \ldots, 3 \), Euler angles, that are functions of time and vary with the movements of the vehicle.

Before the journey begins, the matrix \( R(t_0) \), or equivalently, the attitude angles \( \alpha_i(t_0) \), \( i = 1, \ldots, 3 \), have to be determined with sufficient accuracy. During the journey the attitude changes \( \frac{d}{dt} \alpha_i \) are measured with the help of three gyroscopes as discussed later, and are integrated in time in order to obtain the instantaneous position \( \alpha(t) \), and thus \( R(t) \).

Generally one measures continuously six parameters, three linear accelerations and three angular velocities.
Now the data processing unit of the inertial device integrates the accelerations
\[
a = \frac{d^2x}{dt^2} i + \frac{d^2y}{dt^2} j + \frac{d^2z}{dt^2} k = \frac{d^2x'}{dt^2} i' + \frac{d^2y'}{dt^2} j' + \frac{d^2z'}{dt^2} k',
\]
after the transformation
\[
\bar{a} = R^{-1} \bar{a'}
\]
in three dimensions, and twice. The vectors
\[
\bar{a} = \left[ \frac{d^2x}{dt^2} \frac{d^2y}{dt^2} \frac{d^2z}{dt^2} \right]^T, \quad \bar{a'} = \left[ \frac{d^2x'}{dt^2} \frac{d^2y'}{dt^2} \frac{d^2z'}{dt^2} \right]^T
\]
are again the acceleration component vectors in the two co-ordinate frames.

The first integration yields the velocity vector of the object or vehicle, the second its position.

As follows:
\[
x(t) = x(t_0) + \int_{t_0}^{t} \left( v(t_0) + \int_{t_0}^{\tau} a(\tau) d\tau \right) d\tau', \quad (5.1)
\]
in which \( x(t_0) \) and \( v(t_0) \) are integration constants. They may represent, e.g., the location of the launch site, and the knowledge that the spacecraft is standing still on the launch platform.

As shown in the equation 5.1 the accuracy of position \( x(t) \) will get progressively poorer with time, because the measurements of acceleration \( a(\tau) \) are imprecise and the error in them accumulates through integration. This accumulation happens even twice, because there are two integrals inside each other.

An often used trick to preserve the precision of inertial navigation is to halt regularly ("zero-velocity update"). Then we obtain \( v(t_1) = 0 \) for some time \( t_1 > t_0 \), and the inner (velocity) integral will start again from a known starting value.

### 5.2 Parts of a inertial device

An inertial device, or IMU (inertial measurement unit) contains the following measuring parts:

1. gyroscopes
2. accelerometers.

---

\[1\]This description is of the way inertial navigation is done in an inertial frame in free space, far from planetary surfaces, e.g., in lunar and interplanetary missions.
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.

The motion of a rotating body is described by the following equation:

$$\mathbf{N} = \frac{d\mathbf{L}}{dt} = J \frac{d\mathbf{\omega}}{dt},$$

in which

- \(\mathbf{N}\) torque,
- \(\mathbf{L} = J \mathbf{\omega}\) angular momentum,
- \(\mathbf{\omega}\) angular velocity,
- \(J\) Inertial tensor:

$$J = \begin{bmatrix} J_{xx} & J_{xy} & J_{xz} \\ J_{xy} & J_{yy} & J_{yz} \\ J_{xz} & J_{yz} & J_{zz} \end{bmatrix}.$$

This equation is similar to Newton’s second law of motion:

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = m \frac{\mathbf{v}}{dt},$$

in which \(\mathbf{F}\) is the (linear) force and \(\mathbf{v}\) is the (linear) velocity. \(\mathbf{p} = m\mathbf{v}\) is the momentum or amount of (linear) motion. \(m\), the mass, corresponds to the inertial tensor \(J\), but is in this case a scalar. We assume all the time that \(J\) (and \(m\)) is constant. This is natural for a flywheel, but not, e.g., for the whole Earth, which changes shape continuously.
a matrix of $3 \times 3$ elements. This matrix is symmetric and positive definite.

The faster the gyroscope rotates — the vectorial rotation rate $\vec{\omega}$ —, the more torque $\mathbf{L}$ is needed to turn its axis of rotation.

Equation 5.2 applies in an inertial frame. In a frame connected to the body, we have, assuming that $J$ does not depend on time,

$$N = \frac{d\mathbf{L}}{dt} + \langle \vec{\omega} \times \mathbf{L} \rangle = J \frac{d\vec{\omega}}{dt} + \langle \vec{\omega} \times J \vec{\omega} \rangle. \quad (5.3)$$

The elements of the inertial tensor $J$ of an object can be computed:

$$J_{xx} = \iiint \rho(x, y, z) (y^2 + z^2) \, dx \, dy \, dz,$$

$$J_{yy} = \iiint \rho(x, y, z) (x^2 + z^2) \, dx \, dy \, dz,$$

$$J_{zz} = \iiint \rho(x, y, z) (x^2 + y^2) \, dx \, dy \, dz,$$

$$J_{xy} = -\iiint \rho(x, y, z) xy \, dx \, dy \, dz,$$

$$J_{xz} = -\iiint \rho(x, y, z) xz \, dx \, dy \, dz,$$

$$J_{yz} = -\iiint \rho(x, y, z) yz \, dx \, dy \, dz,$$

so

$$J = \iiint \rho(x, y, z) \begin{bmatrix} y^2 + z^2 & xy & xz \\ xy & x^2 + z^2 & yz \\ xz & yz & x^2 + y^2 \end{bmatrix} \, dx \, dy \, dz.$$

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 $J$ arbitrarily large! Therefore, when talking about the inertial tensor of an object, we always choose its the centre of mass as the origin:

$$x_{\text{com}} = \iiint \rho(x) \, x \, dV,$$

or

$$x_{\text{com}} = \iiint \rho(x, y, z) \, x \, dx \, dy \, dz,$$

$$y_{\text{com}} = \iiint \rho(x, y, z) \, y \, dx \, dy \, dz,$$

$$z_{\text{com}} = \iiint \rho(x, y, z) \, z \, dx \, dy \, dz,$$
after which we use in the computations

\[ x' = x - x_{\text{com}}. \]

As for the axes \textit{orientation}, it is well known that a symmetric matrix can always be brought on \textit{main axes} by a rotation of the Cartesian co-ordinate frame. In this case the inertial tensor assumes the diagonal form

\[ J = \begin{bmatrix} J_x & 0 & 0 \\ 0 & J_y & 0 \\ 0 & 0 & J_z \end{bmatrix}. \]

The elements \( J_x, J_y, \) and \( J_z \) are called the \textit{moments of inertia}. They are actually the eigenvalues of the \( J \) tensor. In this case we obtain from equation 5.3:

\[ N_x = \frac{dL_x}{dt} = J_x \frac{d\omega_x}{dt} + (J_z - J_y) \omega_y \omega_z, \]
\[ N_y = \frac{dL_y}{dt} = J_y \frac{d\omega_y}{dt} + (J_x - J_z) \omega_z \omega_x, \]
\[ N_z = \frac{dL_z}{dt} = J_z \frac{d\omega_z}{dt} + (J_y - J_x) \omega_x \omega_y. \]

These equations apply for a freely spinning body such as the Earth in space.\(^3\)

Building a good gyroscope is a difficult engineering art. A gyroscope consists of a wheel and an axis that is mounted in bearings on both ends

\[^3\text{If we zero the torque } N = 0, \text{ the first two equations become, with } \Delta J = J_z - J_x = J_z - J_y \text{ for a rotationally symmetric body,}

\[ J_x \frac{d\omega_x}{dt} = \Delta J \omega_y \omega_z, \quad J_y \frac{d\omega_y}{dt} = -\Delta J \omega_z \omega_x, \]

and with \( C_x = \Delta J \omega_z / J_x \) and \( C_y = \Delta J \omega_z / J_y \):

\[ \frac{d\omega_x}{dt} = C_x \omega_y, \quad \frac{d\omega_y}{dt} = -C_y \omega_x, \]

yielding by cross-substitution the second-order equations

\[ \frac{d^2\omega_x}{dt^2} = -C_x C_y \omega_x, \quad \frac{d^2\omega_y}{dt^2} = -C_x C_y \omega_y, \]

both having periodic solutions with period

\[ T = 2\pi \sqrt{1 / C_x C_y} = 2\pi \sqrt{J_x J_y / \Delta J \omega_z} = \frac{J'_x}{\Delta J} \text{ sidereal days}. \]

Here, we assumed that \( J_x = J_y \equiv J'. \) Thus, we have found the Euler free nutation of the Earth, one component of the polar motion.
onto a frame, also called table, surrounding the wheel. The frame may consist of several rings and axes or gimbals, a so-called Cardan suspension.

It is important to note that the gyroscope’s rotation axis itself is cannot move sideways in its own body co-ordinate frame. Therefore, in that frame, assuming that the gyroscope axis is the $z$ axis,

$$\omega_x = \omega_y = 0$$

and

$$N_x = \frac{dL_x}{dt} = 0, \quad N_y = \frac{dL_y}{dt} = 0, \quad N_z = \frac{dL_z}{dt} = J_z \frac{d\omega_z}{dt}. \quad (5.4)$$

So, in the body frame, both vectors, the rotation-rate vector $\vec{\omega}$ and the angular momentum $L = J \vec{\omega}$, are always aligned with the gyroscope’s physical spin axis.

For a cylinder of radius $R$, one can show that the moment of inertia about the axis of the cylinder is

$$J_z = \int_0^h \iiint_{\text{circular disc}} \rho (x^2 + y^2) \, dx \, dy \, dz =$$

$$= 2\pi h \rho \cdot \iiint_{\text{circular disc}} r^2 \, r \, dr = \frac{1}{2} \pi \rho h R^4 = \frac{1}{2} MR^2, \quad (5.5)$$

where $M = \rho \cdot \pi R^2 \cdot h$ is the total mass and $r^2 = x^2 + y^2$.  

---

**Figure 5.2.** A gyro wheel and its moments of inertia
For a flat cylinder \((h, \text{ and thus } z, \text{ are small})\) we may also calculate

\[
J_x = \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_{-R}^{R} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \rho \left( y^2 + z^2 \right) \, dy \, dx \, dz
\]

\[
\approx \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_{-R}^{R} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} y^2 \, dy \, dx \, dz =
\]

\[
= \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \int_{0}^{2\pi} \int_{0}^{R} (r \sin \theta) \, r \, dr \, d\theta \, dz =
\]

\[
= \rho \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \frac{1}{2} R^2 \, dz \int_{0}^{2\pi} \sin^2 \theta \, d\theta \int_{0}^{R} r^2 \, r \, dr =
\]

\[
= \frac{1}{4} \left( \rho \cdot h \pi R^2 \cdot h \right) R^2 = \frac{1}{4} MR^2, \quad (5.6)
\]

in which we changed to cylindrical co-ordinates \((\theta, r)\): \(x = r \cos \theta, y = r \sin \theta\). Similarly of course \(J_y = J_x = \frac{1}{4} MR^2 = \frac{1}{2} J_z\).

### 5.2.2 The accelerometer

A primitive accelerometer can easily be built by combining a spring, a scale and test mass. The stretching of the spring is proportional to the test mass, and the acceleration can be read from the scale.

Automatic read-out is possible, e.g., capacitively or with the aid of a piezo-sensor.

The accelerometers are attached to the same frame into which also the gyroscopes are mounted. The measurement axes are made as parallel as possible.

Modern accelerometers are very sensitive, e.g., 10 ppm \(\approx 10\) mGal. If they are based on the elasticity of matter, they demand careful, regular

---

4 In fact, micromechanical acceleration sensors (MEMS) work in precisely this way.

5 The unit mGal is used for measuring gravity or acceleration and equals \(10^{-5}\) m/s\(^2\). Ambient gravity is \(\approx 9.8\) m/s\(^2\) = 980 000 mGal.
calibration. They age (so called drift). Desirable traits, besides sensitivity, are linearity and good behaviour under circumstances of large variations of acceleration, or vibration (missile launch!).

An alternative type of accelerometer is the so-called pendulous type. Here, a mass is attached to the end a beam. Acceleration makes the beam deflect, which is sensed by a sensor. The signal goes to an actuator which restores the deflection to zero. It is thus a nulling sensor, which is necessary to guarantee linear behaviour. Also, this type of accelerometer does not suffer from drift.

Pendulous accelerometers, but gyroscope based, have been used in missiles for a long time, as they offer the highest precision. In these, instead of a mass, a spinning gyroscope is used at the end of the beam, and the nulling feedback applies a sideways torque that causes the gyro to precess as it returns to the null position. This precessional motion is measured and constitutes the output of the accelerometer.\(^6\)

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).

---

\(^6\)E.g., 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.
There are two, very different, general approaches for implementing an inertial measurement unit:

1. strapdown solution
2. stabilized-platform solution.

### 5.3.1 Strapdown solution

In a strapdown solution the gyroscope platform is rigidly connected to the vehicle’s body. When the attitude of the vehicle changes, the ends of the axes of the gyroscope push against its frame with a force that is accurately measured with a force sensor. From the force $F$ we obtain the torque $N$ with the following equation:

$$N = \left\langle F \times \vec{\ell} \right\rangle,$$

in which $\vec{\ell}$ is the length of the gyroscope’s axis as a vector: “torque is force times arm.” The symbol $\times$ designates the exterior or vectorial product.

An alternative solution is to use a so called ring-laser gyroscope that is based on the interference of light (the Sagnac\(^7\) phenomenon, 1913). In

---

\(^7\)Georges Sagnac (1869—1928) was a French physicist.
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 system.

The simplest way to build a ring laser is to use stationary mirrors, nowadays often a long optic fibre is used that is wrapped around the circumference of the ring thousands of times. So the effect multiplies many thousands of times and the sensitivity improves. Nowadays the sensitivity can be as high as $0.00001$ degrees per hour. MathPages, The Sagnac Effect.

### 5.3.2 Stabilized-platform solution

In this solution the whole gyroscope system is mounted inside a three-axis, freely turning cardanic ring system. Because of this, although the attitude of the vehicle changes, the gyroscopic frame retains its position in (inertial) space.

In *terrestrial* application one uses instead of an inertial reference frame, a *local* frame connected to the solid Earth. The three axes of the gyroscope are kept aligned with the *topocentric* frame’s axes triad:

- North direction $x$
To achieve this goal, appropriate torques are applied to the frame of the gyroscopes with the help of torquers. The needed torques can be calculated analogically or digitally in connection with solving for the position of the device. The approach is called Schuler tuning.

5.4 Inertial navigation in the system of the solid Earth

See Cooper (1987) pages 104–107, for a slightly different approach.

5.4.1 Earth rotation

Write the vector of place in inertial space as a function of the vector of place in a co-ordinate frame co-rotating with the Earth:

\[ \mathbf{x}' = R(\theta)\mathbf{x}, \]

in which \( \theta \) is the sidereal time, an angle describing the orientation of the Earth. Its time derivative \( \omega = \frac{d}{dt} \theta \) is the angular velocity of the Earth’s rotation.
Inertial navigation in the system of the solid Earth

In equation 5.7 it is assumed that both the inertial and the corotating co-ordinate 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:

\[
\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \mathbf{x'} = \begin{bmatrix} x' \\ y' \\ z \end{bmatrix},
\]

although actually \( x = xi + yj + zk = x'i' + y'j' + zk \) is one and the same vector in space.

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 its time derivative (chain rule)

\[
\dot{R}(\theta) = \begin{bmatrix}
-\sin \theta & -\cos \theta & 0 \\
\cos \theta & -\sin \theta & 0 \\
0 & 0 & 0
\end{bmatrix} \frac{d\theta}{dt}.
\]

For the velocity we find by differentiation (Leibniz product rule)

\[
\mathbf{v}' = R(\theta) \mathbf{v} + \dot{R}(\theta) \mathbf{x} =
\]

\[
= \begin{bmatrix}
\cos \theta & -\sin \theta & 0 \\
\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix} \mathbf{v} + \begin{bmatrix}
-\sin \theta & -\cos \theta & 0 \\
\cos \theta & -\sin \theta & 0 \\
0 & 0 & 0
\end{bmatrix} \frac{d\theta}{dt} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.
\]

If we define

\[
\vec{\omega} \overset{\text{def}}{=} \begin{bmatrix} 0 \\ 0 \\ \omega \end{bmatrix} = \frac{d\theta}{dt} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},
\]

we have

\[
\dot{R}(\theta) \mathbf{x} = \begin{bmatrix}
-\sin \theta & -\cos \theta & 0 \\
\cos \theta & -\sin \theta & 0 \\
0 & 0 & 0
\end{bmatrix} \frac{d\theta}{dt} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{d\theta}{dt} \begin{bmatrix}
-x \sin \theta - y \cos \theta \\
x \cos \theta - y \sin \theta \\
0
\end{bmatrix}.
\]
but also

$$R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \left( \begin{bmatrix} 0 \\ 0 \\ \omega \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) =$$

$$= \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\omega y \\ \omega x \\ 0 \end{bmatrix} = \omega \begin{bmatrix} -x \sin \theta - y \cos \theta \\ x \cos \theta - y \sin \theta \\ 0 \end{bmatrix},$$

the same result, so

$$\dot{R}(\theta) \overrightarrow{x} = R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right).$$

It follows that

$$\overrightarrow{v}' = R(\theta) \overrightarrow{v} + R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right). \tag{5.8}$$

Suitably choosing $\theta = 0$ yields

$$\overrightarrow{v}' = \overrightarrow{v} + \left( \overrightarrow{\omega} \times \overrightarrow{x} \right).$$

Thus we can conclude:

The effect of rotational motion of the reference frame on the time derivative of a vector can be presented as the cross product of the rotation vector $\overrightarrow{\omega}$ with this vector.

This applies generally, so, e.g.,

$$\dot{R}(\theta) \overrightarrow{v} = R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{v} \right),$$

$$\dot{R}(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) = R(\theta) \left( \overrightarrow{\omega} \times \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) \right),$$

results that we shall use next. Differentiating equation 5.8 again yields the acceleration:

$$\overrightarrow{a}' = R(\theta) \overrightarrow{a} + \dot{R}(\theta) \overrightarrow{v} + \frac{d}{dt} \left( R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) \right) =$$

$$= R(\theta) \overrightarrow{a} + R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{v} \right) + \left( R(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{v} \right) + \dot{R}(\theta) \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) \right) =$$

$$= R(\theta) \left( \overrightarrow{a} + 2 \left( \overrightarrow{\omega} \times \overrightarrow{v} \right) + \left( \overrightarrow{\omega} \times \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) \right) \right).$$

By putting again $\theta = 0$ we find

$$\overrightarrow{a}' = \overrightarrow{a} + 2 \left( \overrightarrow{\omega} \times \overrightarrow{v} \right) + \left( \overrightarrow{\omega} \times \left( \overrightarrow{\omega} \times \overrightarrow{x} \right) \right).$$
5.4.2 The acceleration

The three-dimensional coordinate frame \((x, y, z)\) defined on the rotating Earth is not inertial: for the acceleration, equation 5.9. applies. We may also write for vectors in space:

\[
a' = a + 2 (\vec{\omega} \times v) + (\vec{\omega} \times (\vec{\omega} \times x)),
\]

in which

- \(a'\) acceleration in the inertial system
- \(a\) acceleration relative to the Earth’s surface, in other words, in an Earth-fixed, “co-rotating” frame
- \(\vec{\omega}\) Earth’s rotation vector (constant)
- \(v\) velocity in the Earth-fixed system
- \(x\) the geocentric location of the vehicle in the same system.

In the above equation, the second term on the right side is the so called Coriolis force and the third term is the centrifugal force.

5.4.3 Fundamental equation of inertial navigation

Linear accelerometers measure in general the combined effect of the geometric acceleration of the vehicle and the local gravitation. In other words, the acceleration measured in the vehicle is

\[
t = a' - g'(x) = a + 2 (\vec{\omega} \times v) + (\vec{\omega} \times (\vec{\omega} \times x)) - g'(x),
\]

in which

- \(t\) on-board measured acceleration vector
- \(g'(x)\) gravitational acceleration as the function of place \(x\).

It is often assumed that \(g\), can be calculated straight from Newton’s law of gravitation:

\[
g'(x) \approx -GM_\oplus \frac{x}{\|x\|^3},
\]

but also more complex models are used, such as the normal gravity field of an ellipsoid of revolution where the effects of the Earth’s oblateness and her rotational motion are included, and even very detailed Earth gravitational field models, such as EGM2008 (Earth Gravity Model 2008).
Often we write still

\[ \mathbf{g} = \mathbf{g}' - \langle \mathbf{\omega} \times \langle \mathbf{\omega} \times \mathbf{x} \rangle \rangle, \]

where \( \mathbf{g} \) is the gravity vector, the resultant of gravitation and centrifugal force. Then

\[ \mathbf{t} = \mathbf{a} + 2 \langle \mathbf{\omega} \times \mathbf{v} \rangle - \mathbf{g}(\mathbf{x}). \]  \( \text{(5.10)} \)

With the help of equation 5.10 we can compute from the acceleration measurements \( \mathbf{t} \) and place \( \mathbf{x} \) and velocity \( \mathbf{v} \) (dynamically, “on the fly”) the acceleration \( \mathbf{a} \) in the co-ordinate frame co-rotating with the Earth.

After that, we integrate first \( \mathbf{v} \), and then \( \mathbf{x} \), both also in the co-rotating frame. The equations 5.9, 5.10 are both referred to as the fundamental equation of inertial navigation.

In a frame co-rotating with the Earth, the rotation of the 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 gyros, even though the vehicle is standing still on the ground. This phenomenon is used to orient the gyroscope frame correctly relative to the local north direction (equivalently, to solve the local North direction in the gyroscope frame’s system!) before for example the take-off of an aeroplane or launch of a missile. Furthermore, the accelerometers give straight away the direction of local gravity, the vertical. Together, the two directions are enough to orient the whole frame — except on the North or South Poles.

## 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 stabilized platform serves for example as the mounting platform for a sea or airborne gravimeter, because the measurement axis of the instrument has to be all the time aligned with the local vertical to within a few minutes of arc.

In the stabilised-platform solution one uses a feedback loop called 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 stabilized platform

**Figure 5.8.** The principle of the stabilised platform. The driving signal produces a precessional motion that keeps the gyro’s axis within the plane of the horizon.

the gyroscope to *precess:* the spin axis of the gyroscope itself turns up or downwards.

The stabilized platform requires a *suitable sensor* that detects that the gyro’s axis is out of the horizontal plane (angle $\theta$), which sends a signal through the feedback loop to the motor controller or *actuator* of the vertical axis, figure 5.8. To construct such a sensor that works well in spite of vehicle motions is challenging. See section 5.7 for more.

Assume that the torque about the vertical axis is made proportional to the sensed *axis deviation* $\theta$ from the horizontal plane. Then the change of $\theta$ with time is

$$\frac{d\theta}{dt} = -k_1 \theta$$

with the solution

$$\theta(t) = \theta(t_0) \exp(-k_1 (t - t_0)),$$

in other words, the deviation goes to zero exponentially. By tuning the constant $k_1$ of the feedback loop we can make this happen with suitable speed.
5.6 The gyro compass

The feedback loop visible in the gyrocompass figure 5.9 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 that initially was in the horizontal plane will no longer be horizontal after an elapse of time.

If the rotational angular velocity of the Earth is \( \omega \), then the time derivative of the angle \( \theta \) will be, because of this phenomenon,

\[
\frac{d\theta}{dt} = \omega \cos \varphi \sin \alpha,
\]

in which \( \varphi \) is the latitude and \( \alpha \) the azimuth of the gyroscope spin axis.

The feedback loop gets from the sensor the time derivative \( \frac{d\theta}{dt} \) of the angle \( \theta \) and feeds it, after suitable amplification, into the actuator. As the actuator tries to turn the gyroscope axis back toward the horizontal direction, the outcome will be precession about the vertical axis: \( \alpha \)
The gyro compass

changes. We write the equation

$$\frac{d\alpha}{dt} = -k_2 \frac{d\theta}{dt} = -k_2 \omega \cos \phi \sin \alpha.$$ 

If \( \alpha \) is small enough, we have \( \sin \alpha \approx \alpha \) and the solution is

$$\alpha(t) \approx \alpha(t_0) \exp(-k_2 \omega \cos \phi (t - t_0)).$$

I.e., \( \alpha \) goes exponentially and asymptotically to zero and thus the gyroscope axis turns to the North. Thus we have invented the gyro compass.

Of course this assumes that the device stays in 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 \( \theta \) itself in addition to its time derivative. If we write

$$\frac{d\alpha}{dt} = -k_3 \theta,$$

we obtain by differentiation

$$\frac{d^2\alpha}{dt^2} = -k_3 \frac{d\theta}{dt} = -k_3 \omega \cos \phi \sin \alpha \approx -k_3 \omega \cos \phi \cdot \alpha.$$ 

This is a harmonic oscillator, with solutions

$$\alpha(t) = \cos(t \sqrt{k_3 \omega \cos \phi}), \quad \alpha(t) = \sin(t \sqrt{k_3 \omega \cos \phi}).$$

Unfortunately these solutions are periodic and do not converge to the North direction \( \alpha = 0 \). The best solution is obtained by combining \( \theta \) and \( \frac{d\theta}{dt} \) in the following way:

$$\frac{d^2\alpha}{dt^2} = -k_2 \omega \cos \phi \frac{d\alpha}{dt} - k_3 \omega \cos \phi \cdot \alpha,$$

leading to the following differential equation

$$\frac{d^2\alpha}{dt^2} + \omega \cos \phi \left[k_2 \frac{d\alpha}{dt} + k_3 \alpha\right] = 0.$$

This is a general second-order ordinary differential equation. Depending on the 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.

If we write the inverse of the oscillation time \( \tau = \sqrt{k_3 \omega \cos \phi} \), and

$$k_2 = \frac{2\tau}{\omega \cos \phi},$$

then
we obtain the critically damped case
\[ \frac{d^2 \alpha}{dt^2} + 2\tau \frac{d\alpha}{dt} + \tau^2 \alpha = 0, \]
of which the general solution is
\[ \alpha(t) = (a + bt) e^{-\tau t}, \]
in which \( a \) and \( b \) are arbitrary constants determined by the initial conditions.

Often \( k_3 \), the harmonic restoration coefficient, is implemented by attaching rigidly a heavy semi-ring to the inner ring of the gyroscope frame, which extends downwards. This weight tries then to pull the rotation axis of the gyroscope back to the horizontal plane. The damping factor \( k_2 \) again is implemented traditionally by using a viscous fluid in the bearings of the inner ring.

### 5.7 The Schuler pendulum

#### 5.7.1 Principle

A Schuler\(^8\) pendulum is a pendulum, the length of which is the same as the Earth’s radius \( R = 6378 \text{ km} \). If that kind of pendulum were physically possible, for example as a mass at the end of a long massless rod, its period would be (in a one-g gravity field!)

\[ T_S = 2\pi \sqrt{\frac{R}{g}}, \]
in which \( g \) is gravity on the Earth’s surface.

This period, \( T_S = 84.4 \text{ min} \), 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, it is very well possible to build a pendulum with a period of \( T_S \), for example an extended object suspended from a point very close to its centre of mass.

The simplest pendulum of all is a test mass on the end of a massless bar. Let its length be \( \ell \). If the pendulum swings out of the vertical by an angle\(^9\) \( \theta \), the restoring force will be

\[ F = mg \sin \theta, \]

---

8 Max Schuler (1882–1972), was a German pioneer of navigation technology, Wikipedia, Max Schuler.

9 For consistency with other angles, the angle \( \theta \) is reckoned positive for negative linear displacement of the test mass.
and, as its mass is \( m \), we may compute its acceleration using Newton’s second law, as follows:

\[
md\frac{d^2(-\theta \ell)}{dt^2} = mg \sin \theta \implies \frac{d^2\theta}{dt^2} \approx -\frac{g}{\ell} \theta, \tag{5.11}
\]

the oscillation equation, of which one solution is

\[
\theta(t) = \sin \left(t \sqrt{\frac{g}{\ell}}\right),
\]

from which follows the period

\[
T = 2\pi \sqrt{\frac{\ell}{g}}.
\]

### 5.7.2 The pendulum on a carriage

Mount this pendulum on a carriage that moves in the horizontal direction with an linear acceleration \( a(t) \). The test mass of the pendulum will, in the frame of the carriage, experience a equally large, oppositely directed acceleration \(-a(t)\). Because the length of the pendulum is \( \ell \), it follows that its angular acceleration is

\[
\frac{d^2\bar{\theta}}{dt^2} = \frac{1}{\ell} \left(a - g \left(\bar{\theta} - \psi\right)\right).
\]

Here, \( \bar{\theta} \) stands for the deviation of the pendulum from the vertical of the starting point.

The distance that the carriage has travelled as a function of time will be \( x(t) \) and this distance expressed as a geocentric angular distance, i.e., an angle viewed from the centre of the Earth is \( \psi(t) = \frac{x(t)}{R} \). This quantity, which also represents the change in local vertical along the journey, obeys the differential equation

\[
\frac{d^2\psi(t)}{dt^2} = \frac{1}{R} a(t). \tag{5.12}
\]

Then we obtain by subtraction

\[
\frac{d^2}{dt^2} \left(\bar{\theta} - \psi\right) = \frac{1}{\ell} \left(a - g \left(\bar{\theta} - \psi\right)\right) - \frac{a}{R}.
\]

Now assume that the length of the pendulum \( \ell = R \). Then, with \( \theta \overset{\text{def}}{=} \bar{\theta} - \psi \), it follows that

\[
\frac{d^2\theta(t)}{dt^2} = -\frac{g}{R} \theta(t). \tag{5.13}
\]

Here, \( \theta \) is the angle between the pendulum and the local vertical. Equation 5.13 is identical to pendulum equation 5.11.

One solution of equation 5.13 is \( \theta = 0 \) identically. So
Even though the carriage moves and accelerates in a horizontal direction, the pendulum points all the time to the centre of the Earth.

This is the defining property of the Schuler pendulum.

### 5.7.3 Implementation in an inertial device

In a stabilized platform-based inertial device, so-called Schuler loops, are implemented that make the whole gyroscope frame act like a Schuler pendulum. Every time the frame turns out of the horizontal level, the accelerometers of the horizontal directions \((x, y)\) measure the projection of gravity \(g\) onto the tilting plane, and send correcting impulses to the corresponding gyroscope frame’s actuators. This is how the frame always tracks the local horizontal level.

According to equation 5.12:

\[
\frac{d^2\psi(t)}{dt^2} = \frac{1}{R} a(t),
\]

in which \(a(t)\) is the linear acceleration measured by the accelerometer in the \(x\) direction, and \(R\) the radius of the Earth. The angle function \(\psi(t)\) describes how the local direction of the vertical changes along the journey.
The angular momentum in the gyro wheel is according to Euler’s equation in the body frame 5.4:

\[ \mathbf{L} = \mathbf{J} \mathbf{\omega} = \mathbf{L}_x \mathbf{i} = \mathbf{J}_x \mathbf{\omega} \mathbf{i}, \]

in which \( \mathbf{J}_x \) is the moment of inertia around the gyro’s spin axis, \( \mathbf{\omega} \) the gyro’s spin rate, \( \mathbf{L}_x = \mathbf{J}_x \mathbf{\omega} \) the angular momentum around the \( \mathbf{x} \) axis, and \( \mathbf{i} \) the unit vector in the direction of the spin axis of the gyro. This assumes that the spin axis of the gyro lies in the plane of the horizon \( \mathbf{xy} \), specifically along the \( \mathbf{x} \) axis direction.

More generally in the geometry depicted in figure 5.11:

\[ \mathbf{L} = \mathbf{L}_x \mathbf{i} + \mathbf{L}_y \mathbf{j} + \mathbf{L}_z \mathbf{k} \approx \mathbf{J}_x \mathbf{\omega} \mathbf{i} + \mathbf{L}_y \mathbf{j} + \mathbf{L}_z \mathbf{k} \]

with \( \{ \mathbf{i}, \mathbf{j}, \mathbf{k} \} \) is the orthonormal basis of the \( \mathbf{(x, y, z)} \) frame.

In this, \( \mathbf{J}_x \) is the moment of inertia of the gyroscope wheel around its spin axis, the \( \mathbf{x} \) axis. Due to the symmetry of the gyro wheel, the inertial tensor \( \mathbf{J} \) is, in the geometry of figure 5.11, a diagonal matrix

\[
\mathbf{J} = \begin{bmatrix}
J_x & 0 & 0 \\
0 & J_y & 0 \\
0 & 0 & J_z \\
\end{bmatrix} \approx \begin{bmatrix}
\frac{1}{2} \mathbf{M} \mathbf{r}^2 & 0 & 0 \\
0 & \frac{1}{4} \mathbf{M} \mathbf{r}^2 & 0 \\
0 & 0 & \frac{1}{4} \mathbf{M} \mathbf{r}^2 \\
\end{bmatrix}
\]
by equations 5.5 and 5.6 for a thin circular disk of mass \( M \) and radius \( r \).

When the vehicle moves and the horizontal plane tilts away from the gyroscope spin axis direction, \( L_z \neq 0 \). Assume the deviation angle \( \psi \) to be small. Then we have geometrically

\[
\psi \approx \frac{L_z}{L_x} \approx \frac{1}{J_x \omega} L_z \implies \frac{d\psi}{dt} = \frac{d}{dt} \left( \frac{L_z}{L_x} \right) = \frac{1}{J_x \omega} \frac{d}{dt} L_z.
\]

The angular acceleration of the turning of the horizontal plane is

\[
\frac{d^2 \psi}{dt^2} = \frac{1}{J_x \omega} \frac{d^2}{dt^2} L_z. \tag{5.15}
\]

Substitution of equation 5.14 into equation 5.15 and integration yields

\[
\frac{d^2 L_z}{dt^2} = J_x \omega \frac{d^2 \psi}{dt^2} = J_x \tilde{\omega} \frac{a}{R} \implies \frac{dL_z}{dt} = \frac{J_x \tilde{\omega}}{R} \int a \, dt.
\]

By Euler’s equation 5.2 in an inertial frame

\[
N_z = \frac{dL_z}{dt} = \frac{J_x \tilde{\omega}}{R} \int a \, dt. \tag{5.16}
\]

In equation 5.16 \( N_z \) is the required torque around the \( z \) axis, see figure 5.11.

This is how a Schuler pendulum is obtained. \( R \) is the radius of the Earth, approximately 6378 km.

According to equation 5.16 the Schuler loop is implemented either on the hardware level — in older equipment, the factor \( J_x \tilde{\omega}/R \) is a device constant and integration is done analogously in hardware — 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.

### 5.8 Mechanization

Because a real-life inertial device is quite a lot more complicated than simple principles, the modelling of the behaviour of all the parts is to be done carefully. This model is called the mechanization of the inertial device.

As a simple example of mechanization is treated a one-dimensional carriage on the surface of a spherical Earth, figure 5.10.

Firstly, the velocity is by definition

\[
\frac{dx(t)}{dt} = v(t).
\]
Acceleration is measured continuously by an acceleration sensor, the measured value being \( a(t) \). This measured quantity, a function of time, consists of two parts:

- the geometric acceleration

\[
\frac{d^2x(t)}{dt^2} = \frac{dv(t)}{dt}
\]

- the projection of the gravity vector onto the accelerometer’s axis, \( \theta(t) \, g \), in which \( \theta(t) \) is the angle of tilt of the carriage from the local vertical.

Remember that a differential equation is a statement on the properties of functions! The result is

\[
\frac{dv(t)}{dt} = a(t) - g \, \theta(t),
\]

in which the function \( a(t) \) is the result of a continuous measurement process.

Finally we discuss the Schuler loop. The angle of deflection \( \theta \) behaves like a Schuler pendulum and tries to revert to zero according to equation 5.13:

\[
\frac{d^2\theta(t)}{dt^2} = -\frac{g}{R} \theta(t). \tag{5.17}
\]

Determine the approximate values, the functions of time \( x^{(0)}(t) \) and \( v^{(0)}(t) \), and the difference quantities \( \Delta x(t) \overset{\text{def}}{=} x(t) - x^{(0)}(t) \) and \( \Delta v(t) \overset{\text{def}}{=} v(t) - v^{(0)}(t) \): linearization. Then

\[
\frac{dx^{(0)}(t)}{dt} = v^{(0)}(t), \quad \frac{dv^{(0)}(t)}{dt} = a(t),
\]

in which \( a(t) \) is assumed continuously measured\(^{10}\) and

\[
\frac{d\Delta x(t)}{dt} = \Delta v(t), \quad \frac{d\Delta v(t)}{dt} = -g \, \theta(t).
\]

Now into the equation 5.17 can be substituted

\[
g \, \theta(t) = -\frac{d\Delta v(t)}{dt},
\]

with the result

\[
\frac{d^2\theta(t)}{dt^2} = \frac{1}{R} \frac{d\Delta v(t)}{dt}.
\]

\(^{10}\)It would be proper to consider this measured \( a(t) \), and all quantities derived from it, as stochastic. We don’t do that here, but see \( n_a \) below.
Inertial navigation

Tableau 5.1. Mechanisation simulation in one dimension, octave code.

```octave
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"
```

By integrating — i.e., leaving out one $\frac{d}{dt}$ from each side — we obtain

$$\frac{d\theta(t)}{dt} = \frac{1}{R} \Delta v(t),$$

and as the complete Kalman-filter equation we obtain

$$\frac{d}{dt} \begin{bmatrix} \Delta x \\ \Delta v \\ \theta \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -g \\ 0 & \frac{1}{R} & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \\ \theta \end{bmatrix} + \begin{bmatrix} 0 \\ n_a \\ n_g \end{bmatrix}, \quad (5.18)$$

where we have added the possible (stochastic) noise terms $n_a, n_g$ of the acceleration sensor and the gyro stabilization mechanism.

This solution works in this way, that we continuously integrate the real-time approximate values $v(0)(t)$ and $x(0)(t)$, and with the help of the Kalman filter $\Delta x(t), \Delta v(t)$ and $\theta(t)$.

In figure 5.12 we see how these quantities $\Delta x, \Delta y$ and $\theta$ behave over time. The oscillatory behaviour on the Schuler time scale is evident.

Note that in the solution both the angle of deflection $\theta$ of the carriage and the speed perturbation $\Delta v$ — and also the position perturbation
\[ \Delta x \text{ — “oscillate” harmonically like the Schuler pendulum}^{11}, \text{ with the period } T_S = 84.4 \text{ min. The height has to be obtained in another way, for example in an airplane by means of an atmospheric pressure sensor.} \]

---

\(^{11}\)If the angle \( \theta \) has, e.g., an amplitude \( A_\theta = 1' = 2.9 \cdot 10^{-4} \text{ rad} \), it follows from the equation

\[ \frac{d\Delta v}{dt} = -g\theta \]

that

- \( \Delta v \)'s amplitude is \( A_{\Delta v} = g\sqrt{\frac{g}{2}} A_\theta = 2.3 \text{ m/s} \) and
- \( \Delta x \)'s amplitude is \( A_{\Delta x} = \sqrt{\frac{g}{2}} A_{\Delta v} = 1855 \text{ m.} \)
5.9 On the Earth’s surface in two dimensions

This is easily generalized to two dimensions. In this way, a “navigator” may be built on the surface of the Earth.

In two dimensions, the mechanization equations look like

\[
\frac{d}{dt} \begin{bmatrix}
\Delta x_1 \\
\Delta x_2 \\
\Delta v_1 \\
\Delta v_2 \\
\theta_1 \\
\theta_2 \\
\Delta A
\end{bmatrix} = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -g & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & g & 0 \\
0 & 0 & 1/N & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1/M & 0 & 0 & 0 \\
0 & 0 & 0 & tan \varphi/N & 0 & 0 & 0 \\
\end{bmatrix} \begin{bmatrix}
\Delta x_1 \\
\Delta x_2 \\
\Delta v_1 \\
\Delta v_2 \\
\theta_1 \\
\theta_2 \\
\Delta A
\end{bmatrix} + \begin{bmatrix}
0 \\
0 \\
n_{a,1} \\
n_{a,2} \\
n_{g,1} \\
n_{g,2} \\
n_{g,A}
\end{bmatrix},
\]

in which \( \theta_1 \) and \( \theta_2 \) are the north and east tilt angles, and \( A \) is the heading or azimuth angle, clockwise from the north. The co-ordinates \( \Delta x_1 \) and \( \Delta x_2 \) are the north and east map co-ordinates.

The approximate values satisfy

\[
\frac{d}{dt} \begin{bmatrix}
x_1^{(0)} \\
x_2^{(0)} \\
v_1^{(0)} \\
v_2^{(0)}
\end{bmatrix} = \begin{bmatrix}
\frac{d}{dt} x_1^{(0)} \\
\frac{d}{dt} x_2^{(0)} \\
\frac{d}{dt} v_1^{(0)} \\
\frac{d}{dt} v_2^{(0)}
\end{bmatrix} = \begin{bmatrix}
a_1(t) \\
a_2(t)
\end{bmatrix}.
\]

The equation for the heading is

\[
\frac{dA}{dt} = \frac{tan \varphi}{N} (v_2 + \omega N \cos \varphi) = \frac{tan \varphi}{N} v_2 + \omega \sin \varphi,
\]

including an Earth rotation term: the stabilized platform is also a Foucault pendulum!

\( N(\varphi) \) and \( M(\varphi) \) are the transversal and meridional radii of curvature, respectively, so this mechanization is good to ellipsoidal precision.

For the approximate values it holds that

\[
\frac{dA^{(0)}}{dt} = \frac{tan \varphi}{N} v_2^{(0)} + \omega \sin \varphi,
\]

and it follows that

\[
\frac{d}{dt} \Delta A = \frac{tan \varphi}{N} \Delta v_2.
\]

5.10 Initialization of an inertial device

It is of interest how one levels and orients an inertial platform. When not moving, the accelerometers of the inertial device act as inclinometers.
Initialization of an inertial device

and the feedback loops make the gyroscope axes turn to the horizontal plane.

The north direction is obtained by using the device as a gyro compass, by observing how the local vector of gravity slowly turns about the south-north axis.

On airports, one often sees a tableau giving the precise (±0.1) geographic latitude and longitude of the gate. This is in fact used to initialize the co-ordinates in the inertial navigation platform used on a jetliner. Also levelling and orientation is performed while standing at the gate.
The subjects of this chapter are more extensively presented in the books Poutanen (2017), chapter 5, and Hofmann-Wellenhof et al. (1997), chapter 4. A good understanding of satellite orbits and their geometry is needed, if the Kalman-filter is used to improve the satellite orbit with the help of observations made in real time.

Also in the context of terrestrial GPS navigation this helps to understand how the locations of the GPS-satellites can be calculated from the orbital elements, first in space and then in the observer’s orb of heaven.

### 6.1 The Kepler orbit

If it is assumed that the satellite moves in a central force field, for example the gravitational field of a mass point or a spherical Earth, it follows that the satellite’s orbit is a Kepler orbit. Johannes Kepler (1571–1630) discovered his model of planetary orbits based on the observation material on the orbit of the planet Mars by Tycho Brahe (1546–1601), see Physics Classroom, Kepler’s Three Laws.

As we have seen, we can describe the satellite’s motion vectorially like this:

\[
\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x} = \mathbf{v}, \quad \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{v} = -\frac{G M \oplus}{\|\mathbf{x}\|^3} \mathbf{x}.
\]

Here \( \mathbf{x} \) and \( \mathbf{v} \) are the position and velocity vectors in three-dimensional space. The combined vector

\[
\mathbf{x} \overset{\text{def}}{=} \begin{bmatrix} \mathbf{x} & \mathbf{v} \end{bmatrix}^T = \begin{bmatrix} x & y & z & \dot{x} & \dot{y} & \dot{z} \end{bmatrix}^T
\]

written in an agreed, rectangular (Cartesian) co-ordinate frame is the state vector of the system.
Elements of the Kepler orbit are only an alternative way of writing the state vector. Wikipedia, Orbital elements gives a good description.  

Ω right ascension of the ascending node, i.e., astronomical longitude. The zero point of this longitude is the place on celestial sphere where the ecliptic plane and the equatorial plane intersect, the “vernal equinox point”: the place of the Sun at the start of spring, when it goes from the Southern hemisphere to the Northern hemisphere.

i inclination, the orbital plane’s tilt angle relative to the equator. The inclination of the orbital plane for the GPS satellites is 55°.

ω argument of perigee. The angular distance between the ascending node and the perigee of the satellite orbit.

a the semi-major axis of the satellite orbit.

e the eccentricity of the satellite orbit. \[1 - e^2 = \frac{b^2}{a^2},\] in which b is the semi-minor axis.

v, E, M describe the position of the satellite in its orbit as the function of time:

\[v(t) \quad \text{true anomaly}\]
\[E(t) \quad \text{eccentric anomaly}\]
\[M(t) \quad \text{mean anomaly}.\]

The connections between them:

\[E(t) = M(t) + e \sin E(t) \quad (6.1)\]

\[\tan \frac{1}{2} v(t) = \sqrt{\frac{1 + e}{1 - e}} \quad (6.2)\]

See figure 6.1. The mean anomaly M is only a linear measure of elapsed time, scaled to the period P of the satellite and referred to the moment of its passage through the perigee \(\tau\):

\[M(t) \overset{\text{def}}{=} 2\pi \frac{t - \tau}{P}.\]

E and v are purely geometrical quantities.

In the figure the angle \(\theta_0\) is the sidereal time of Greenwich, which describes the globe’s attitude relative to the starry sky. Greenwich sidereal time consists of annual and daily components,\(^1\) that are caused by the Earth’s rotation and orbit movements, respectively.

\(^1\)Greenwich sidereal time is calculated as follows:
1. Take the month value from the following table:

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>37</td>
<td>40</td>
<td>30</td>
<td>32</td>
<td>31</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Jul</td>
<td>Aug</td>
<td>Sep</td>
<td>Oct</td>
<td>Nov</td>
<td>Dec</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>31</td>
<td>20</td>
<td>33</td>
<td>22</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>36</td>
<td>2</td>
<td>34</td>
<td>2</td>
<td>36</td>
<td>4</td>
<td>34</td>
</tr>
</tbody>
</table>

2. Add to this 4 (four) minutes for every day of the month;

3. Add to this the clock time (UTC or Greenwich mean time);

If you want to compute the local time, you have to add to this the longitude East of your location converted to time units: $15^\circ = 1^h$, $1^\circ = 4^m$, $15' = 1^n$.

The precision of your result will be $\pm 4^m$, because this table is not really constant from year to year: it varies with the leap year cycle.

Figure 6.1. Kepler’s orbital elements.
So we have obtained an alternative way of presenting the state vector:

\[
\mathbf{a} = \begin{bmatrix} a & e & M & i & \omega & \Omega \end{bmatrix}^T.
\]

In a central force field the elements of this state vector are constants except \( M(t) \), see above. If the force field is not precisely central, also the other orbital elements can change slowly with time. For example the Earth’s flattening causes the slow turning of the ascending node \( \Omega \). This kind of time dependent Kepler elements — like for example \( \Omega(t) \) — are called osculating elements\(^2\).

### 6.1.1 The radius vector

See figure 6.2 of the orbital plane. For expressing the radius vector \( r \) in terms of the eccentric anomaly \( E \) we first observe that the axes ratio of the orbital ellipse is

\[
\frac{b}{a} = \sqrt{1-e^2} = \frac{SQ}{PQ}.
\]

Then

\[
\begin{align*}
\sqrt{SQ^2 + QG^2} &= \sqrt{(1-e^2)PQ^2 + (QC + CG)^2} = \\
&= \sqrt{(1-e^2)PC^2 \sin^2 E + (-a \cos E + ea)^2} = \\
&= \sqrt{(1-e^2)a^2 (1-\cos^2 E) + a^2 \cos^2 E - 2ea^2 \cos E + e^2a^2} = \\
&= a\sqrt{1-e^2 \cos^2 E - e^2 + e^2 \cos^2 E - 2e \cos E + e^2} = \\
&= a\sqrt{(1-e \cos E)^2} = a (1 - e \cos E). \quad (6.3)
\end{align*}
\]

For deriving the radius vector \( r \) expressed in the true anomaly \( \nu \) we start from the equation for the ellipse in rectangular co-ordinates centred on the ellipse’s centre \( C \):

\[
\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.
\]

Expressing the co-ordinates \( x \) and \( y \) in \( \nu \) gives

\[
\frac{(r \cos \nu + ea)^2}{a^2} + \frac{r^2 \sin^2 \nu}{a^2 (1-e^2)} = 1
\]

or

\[
(1-e^2) (r \cos \nu + ea)^2 + r^2 \sin^2 \nu = a^2 (1 - e^2).
\]

\(^2\)from Latin osculārī, to kiss.
Reorganizing terms yields

\[
\left( (1 - e^2) \cos^2 \nu + (1 - \cos^2 \nu) \right) r^2 + \left( 2 (1 - e^2) e a \cos \nu \right) r + \\
+ \left( (1 - e^2) e^2 a^2 - a^2 (1 - e^2) \right) = 0
\]

or

\[
(1 - e^2 \cos^2 \nu) r^2 + \left( 2 (1 - e^2) e a \cos \nu \right) r + \left( -a^2 (1 - e^2)^2 \right) = 0,
\]

a quadratic equation with the standard solution. The discriminant is

\[
\text{Disc} = \left( 2 (1 - e^2) e a \cos \nu \right)^2 - 4 (1 - e^2 \cos^2 \nu) \left( -a^2 (1 - e^2)^2 \right) = \\
= 4a^2 (1 - e^2)^2 \left( e \cos \nu)^2 + (1 - e^2 \cos^2 \nu) \right) = 4a^2 (1 - e^2)^2.
\]

The solution is

\[
r_{1,2} = \frac{-2 (1 - e^2) e a \cos \nu \pm \sqrt{\text{Disc}}}{2 (1 - e^2 \cos^2 \nu)} = \\
= \frac{-2a \left( 1 - e^2 \right) e \cos \nu \pm 2a (1 - e^2)}{2 (1 - e^2 \cos^2 \nu)} = a \left(1 - e^2\right) \left( \pm 1 - e \cos \nu \right) \frac{1}{1 - e^2 \cos^2 \nu}.
\]
Of this, 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

See again figure 6.2. Kepler’s second law, the law of areas, applies not only to the motion of the satellite itself, but also to the motion of its corresponding point \( P \). All areas are just \( \frac{a}{b} \) times larger. This means that the area of the sector \( PGB \), marked “\( I \)” in the figure, must be a fraction \( \frac{M}{2\pi} \) of that of the full circle. The area of the full circle is \( \pi a^2 \).

Thus

\[ I = \frac{M}{2\pi} \pi a^2 = \frac{1}{2} a^2 M. \] (6.4)

Similarly the sector \( PCB \), comprising the areas “\( I \)” and “\( II \)” together, is qua surface area, geometrically, a fraction \( \frac{E}{2\pi} \) of the circle. Thus

\[ I + II = \frac{E}{2\pi} \pi a^2 = \frac{1}{2} a^2 E. \] (6.5)

Finally, the area of triangle \( PCG \) is

\[ II = \frac{1}{2} CG \cdot PQ = \frac{1}{2} ae \cdot a \sin E = \frac{1}{2} a^2 e \sin E. \] (6.6)

The three results 6.4, 6.5 and 6.6 together produce Kepler’s 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

We can write the satellite’s \( x \) co-ordinate as

\[ x = r \cos \nu = a \cos E - ea \implies r \cos \nu = a \left( \cos E - e \right). \]

Now, the half-angle substitution

\[ \cos \nu = 2 \cos^2 \frac{1}{2} \nu - 1 \]

yields

\[ 2r \cos^2 \frac{1}{2} \nu - r = a \left( \cos E - e \right) \]
or, with equation 6.3

\[ 2r \cos^2 \frac{1}{2} \nu - a(1 - e \cos E) = a(\cos E - e) \]

or

\[ 2r \cos^2 \frac{1}{2} \nu = a \left( (\cos E - e) + (1 - e \cos E) \right) = a(1 - e)(\cos E + 1). \]

With the other half-angle substitution

\[ \cos E = 2 \cos^2 \frac{1}{2} E - 1 \]

this yields

\[ 2r \cos^2 \frac{1}{2} \nu = 2a(1 - e) \cos^2 \frac{1}{2} E. \] (6.7)

Similarly using the relations

\[ \cos \nu = 1 - 2 \sin^2 \frac{1}{2} \nu, \quad \cos E = 1 - 2 \sin^2 \frac{1}{2} E \]

yields

\[ r - 2r \sin^2 \frac{1}{2} \nu = a(\cos E - e) \]

\[ \implies a(1 - e \cos E) - 2r \sin^2 \frac{1}{2} \nu = a(\cos E - e) \]

\[ \implies -2r \sin^2 \frac{1}{2} \nu = a((\cos E - e) - (1 - e \cos E)) = \]

\[ = a(1 + e)(\cos E - 1) \implies -2a(1 + e) \sin^2 \frac{1}{2} E \]

\[ \implies 2r \sin^2 \frac{1}{2} \nu = 2a(1 + e) \sin^2 \frac{1}{2} E. \] (6.8)

Division of equation 6.8 by equation 6.7, yields

\[ \tan^2 \frac{1}{2} \nu = \frac{1 + e}{1 - e} \tan^2 \frac{1}{2} E \implies \frac{\tan \frac{1}{2} \nu}{\tan \frac{1}{2} E} = \sqrt{\frac{1 + e}{1 - e}}, \]

result 6.2.

### 6.1.4 Rectangular co-ordinates and time derivatives

The time derivative of \( r \) is, equation 6.3:

\[ \frac{dr}{dt} = ae \sin E \frac{dE}{dt}. \]

From Kepler equation 6.1 we obtain

\[ \frac{dE}{dt} = \frac{dM}{dt} + e \cos E \frac{dE}{dt} = \frac{2\pi}{P} + e \cos E \frac{dE}{dt}. \]
so

\[ \frac{dE}{dt} = \frac{2\pi}{P(1 - e \cos E)}, \]

in which \( P \) is the period. In the orbital plane

\[ \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} r \cos \nu \\ r \sin \nu \\ 0 \end{bmatrix} = \begin{bmatrix} a \cos E - e \\ b \sin E \end{bmatrix}, \]

Differentiation with respect to time yields

\[ \begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix} = \begin{bmatrix} \frac{d}{dt} \left[ \begin{bmatrix} a \cos E - e \\ b \sin E \end{bmatrix} \right] \end{bmatrix} = \frac{2\pi}{P(1 - e \cos E)} \begin{bmatrix} -a \sin E \\ b \cos E \end{bmatrix}. \]

6.1.5 Transformation to the geocentric frame

We can transform the obtained two-dimensional vectors in the orbital plane into three-dimensional space vectors by using the rotation angles \( \omega, i, \Omega \). If we write in the co-ordinate frame of the orbital plane \(^3\)

\[ \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} r \cos \nu \\ r \sin \nu \\ 0 \end{bmatrix} = \begin{bmatrix} a \cos E - e \\ b \sin E \end{bmatrix}, \]

we get geocentrically

\[ \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R\mathbf{x}, \]

in which

\[ R(\Omega, i, \omega) = R(\Omega)R(i)R(\omega), \quad R(i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{bmatrix}, \]

\[ R(\Omega) = \begin{bmatrix} \cos \Omega & -\sin \Omega & 0 \\ \sin \Omega & \cos \Omega & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R(\omega) = \begin{bmatrix} \cos \omega & -\sin \omega & 0 \\ \sin \omega & \cos \omega & 0 \\ 0 & 0 & 1 \end{bmatrix}, \]

summarisingly

\[^3\text{Here, the overbar indicates that this is the abstract vector of the components of vector } \mathbf{x} \text{ in space.}\]
The Kepler orbit

\[ R(\Omega, i, \omega) = \\
\begin{bmatrix}
\cos \Omega \cos \omega - \sin \Omega \sin \omega \cos i & -\cos \Omega \sin \omega - \sin \Omega \cos \omega \cos i & \sin \Omega \sin i \\
\sin \Omega \cos \omega + \cos \Omega \sin \omega \cos i & \cos \Omega \cos \omega \cos i - \sin \Omega \sin \omega & -\cos \Omega \sin i \\
\sin \omega \sin i & \cos \omega \sin i & \cos i
\end{bmatrix}. \]

See figure 6.3.

The geocentric coordinates thus obtained are in an inertial or astronomical system. The origin of the longitudes is the direction to the vernal equinox. In case the satellite's coordinates are sought in a system co-rotating with the Earth (the origin of longitudes being Greenwich) we calculate

\[ \ell = \Omega - \theta_0, \]

where \( \theta_0 \) is Greenwich sidereal time, and put in the matrix equation above \( \ell \) instead of \( \Omega \).

The velocity vector is obtained by differentiating with respect to time:

\[ \frac{d}{dt} \vec{x} = \frac{2\pi}{P (1 - e \cos E)} \begin{bmatrix}
-a \sin E \\
b \cos E \\
0
\end{bmatrix}, \]

from which the geocentric equivalent follows:

\[ \frac{d}{dt} \vec{X} = \frac{d}{dt} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R \frac{d}{dt} \vec{x}. \]

This only holds in an inertial frame. In a frame co-rotating with the Earth we find

\[ \frac{d}{dt} \vec{X} = R \frac{d}{dt} \vec{x} + \left( \frac{d}{dt} R \right) \vec{x} \]
where now \( R \) contains the sidereal time and is thus time dependent:

\[
\frac{dR}{dt} = \begin{vmatrix}
-\sin \ell \cos \omega - \cos \ell \sin \omega \cos i & \sin \ell \sin \omega - \cos \ell \cos \omega \cos i & \cos \ell \sin i \\
\cos \ell \cos \omega - \sin \ell \sin \omega \cos i & -\cos \ell \sin \omega - \sin \ell \cos \omega \cos i & \sin \ell \sin i \\
\sin \omega \sin i & \cos \omega \sin i & \cos i
\end{vmatrix} \cdot \frac{d\theta_0}{dt}.
\]

### 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 co-rotating, non-inertial frame \((u, v, w)\) is used, the origin of which is centred on the Earth and which rotates at the same mean rate as the Earth, i.e., one rotation per year. As the distance of the Moon from the Earth is only 0.3\% of that between Earth and Sun, the mathematics of at least the solar influence can be effectively linearized.

A modification of the method models the motion of an Earth satellite

---

\(^4\)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 Earth and Sun are of similar magnitude.
Use of Hill co-ordinates

relative to a fictitious point orbiting the Earth in a circular orbit with the same period as the satellite. This approach has been fruitful for studying orbital perturbations and the \textit{rendez-vous} problem.

Looking at figure 6.4 we may write

$$x = u + u_0.$$  \hfill (6.9)

Expand the vectors into components:

$$u = ui' + vj' + wk,$$

$$x = xi + yj + zk =$$

$$(x \cos \theta + y \sin \theta + r_0) i' + (-x \sin \theta + y \cos \theta) j' + zk,$$

$$u_0 = r_0 i'.$$

Form abstract component vectors on the basis \{i, j, k\}:

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

and on the basis \{i', j', k\}:

$$\bar{x}' = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \bar{u}' = \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \quad \bar{u}_0 = \begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix}.$$  

Equation 6.9 becomes

$$R \bar{x} = \bar{x}' = \bar{u}' + \bar{u}_0'$$ \hfill (6.10)

in which the matrix $R$ is a rotation of the co-ordinate frame around the $z$ axis by an angular amount $\theta$:

$$R = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$  

$\bar{x}$ contains the components in the inertial frame, $\bar{u}'$ in the frame co-rotating with the satellite. The $i'$ or $u$ axis points outward ("upward"), the $j'$ or $v$ axis forward in the direction of flight, and the $k$ or $w$ or $z$ axis points perpendicularly out of the orbital plane to "port". The satellite moves at constant velocity in a circular orbit: the angular velocity according to Kepler’s third law is

$$n = \frac{d\theta}{dt} = \sqrt{\frac{GM_{\oplus}}{r_0^3}}.$$  

---

119
\( r_0 \) is the orbital radius and also the distance of the origin of the \((u, v, w)\) frame from that of the \((x, y, z)\) system.

Equation 6.10 can be inverted:

\[
\mathbf{x} = R^{-1} (\mathbf{u}' + \mathbf{u}_0') = R^T (\mathbf{u}' + \mathbf{u}_0'),
\]

because for an orthogonal matrix \(RR^T = I \iff R^{-1} = R^T\).

### 6.3 Transformation from inertial system to Hill system

We derive equations for the first and second derivatives of the vector \(\mathbf{x}\) and the matrix \(R\) and substitute. After that, we multiply both sides of the equation with the matrix \(R\).

We obtain by differentiation with the Leibniz product rule:

\[
\dot{\mathbf{x}} = \dot{R}^T (\mathbf{u}' + \mathbf{u}_0') + R^T \dot{\mathbf{u}}',
\]

\[
\ddot{\mathbf{x}} = \ddot{R}^T (\mathbf{u}' + \mathbf{u}_0') + 2\dot{R}^T \dot{\mathbf{u}}' + R^T \ddot{\mathbf{u}}'.
\]

Here 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

\[
\begin{bmatrix}
\dot{x} \\
\dot{y} \\
\dot{z}
\end{bmatrix} = n^2 \begin{bmatrix}
-cos \theta & sin \theta & 0 \\
-sin \theta & -cos \theta & 0 \\
0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
u + r_0 \\
v \\
w
\end{bmatrix} +
\]

\[
2n \begin{bmatrix}
-sin \theta & -cos \theta & 0 \\
cos \theta & -sin \theta & 0 \\
0 & 0 & 0
\end{bmatrix} \begin{bmatrix}
\dot{u} \\
\dot{v} \\
\dot{w}
\end{bmatrix} +
\]

\[
\begin{bmatrix}
-cos \theta & -sin \theta & 0 \\
sin \theta & cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\ddot{u} \\
\ddot{v} \\
\ddot{w}
\end{bmatrix}.
\]
By multiplying from the left with the matrix $R$ we obtain\(^5\)

$$
\begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\dot{x} \\
\dot{y} \\
\dot{z} \\
\end{bmatrix}
= n^2
\begin{bmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
u + r_0 \\
v \\
w \\
\end{bmatrix}
+
2n
\begin{bmatrix}
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
\dot{u} \\
\dot{v} \\
\dot{w} \\
\end{bmatrix}
+
\begin{bmatrix}
\ddot{u} \\
\ddot{v} \\
\ddot{w} \\
\end{bmatrix}.
\tag{6.11}
$$

### 6.4 Series expansion for a central force field

The equation for a central force field in the $(x, y, z)$ system is

$$
\ddot{x} = -\frac{GM}{\|x\|^3}x,
$$
in components on the basis $\{i', j', k\}$:

$$
\ddot{x}' = -\frac{GM}{\|x\|^3}x' = -\frac{GM}{\|x\|^3}(\bar{u}' + \bar{u}'_0),
$$
in which

$$
\|x\| = \|u + u_0\| = \|\bar{u}' + \bar{u}'_0\| = \sqrt{(u + r_0)^2 + v^2 + w^2}.
$$

\(^5\)Sometimes we use the notation

$$
\begin{bmatrix}
\alpha \\
\beta \\
\gamma \\
\end{bmatrix}
= \begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
z \\
\end{bmatrix}
= \begin{bmatrix}
u + r_0 \\
v \\
w \\
\end{bmatrix},
$$

This is a co-ordinate system with the same origin as $(x, y, z)$, but whose $(\alpha, \beta)$ axes turn with the satellite and remain in the same direction as the axes $(u, v)$. 

\[\]
The Taylor expansion about the origin of the \((u, v, w)\) frame now yields

\[
\begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\ddot{x} \\
\ddot{y} \\
\ddot{z}
\end{bmatrix}
= -\frac{GM}{r_0^3}
\begin{bmatrix}
r_0 \\
0 \\
0
\end{bmatrix}
+ \mathcal{M} \cdot \begin{bmatrix}
u \\
v \\
w
\end{bmatrix},
\]

where the gravitational-gradient tensor, the matrix \(\mathcal{M}\) consists of the partial derivatives:

\[
\mathcal{M} = \begin{bmatrix}
\frac{\partial}{\partial u} & \frac{\partial}{\partial v} & \frac{\partial}{\partial w}
\end{bmatrix}
\begin{bmatrix}
-\frac{GM\oplus}{\|x\|^3}
\begin{bmatrix}
u + r_0 \\
v \\
w
\end{bmatrix}
\end{bmatrix}
_{u,v,w=0}
= -\frac{GM\oplus}{r_0^3}
\begin{bmatrix}
-2 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix},
\]

equation 3.7 evaluated in the point \(u_0\).

According to Kepler’s third law

\[
\frac{GM\oplus}{r_0^3} = n^2.
\]

By combining we obtain

\[
\begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\ddot{x} \\
\ddot{y} \\
\ddot{z}
\end{bmatrix}
= -n^2
\begin{bmatrix}
r_0 \\
0 \\
0
\end{bmatrix}
+ \begin{bmatrix}
-2 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w
\end{bmatrix}
+ \begin{bmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
u + r_0 \\
v \\
w
\end{bmatrix}
\]

(6.12)

\section*{6.5 Equations of motion in the Hill system}

By combining the equations 6.11 and 6.12 we obtain in the absence of external forces:

\[
0 = n^2
\begin{bmatrix}
r_0 \\
0 \\
0
\end{bmatrix}
+ \begin{bmatrix}
-2 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w
\end{bmatrix}
+ n^2
\begin{bmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
u + r_0 \\
v \\
w
\end{bmatrix}
\]
Solving the Hill equations

\[
+ 2n \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.
\]

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 \(w\):\(^6\)

\[
\ddot{u} = 2n\dot{v} + 3n^2u, \quad \ddot{v} = -2n\dot{u}, \quad \ddot{w} = -n^2w,
\]

in which the third is recognized as a classical harmonic oscillator. These equations were found by Hill (1886) when studying the motion of the Moon.

6.6 Solving the Hill equations

This part may be skipped over at first-time reading. It contains complicated derivations.

6.6.1 The \(w\) equation

We first solve the easiest equation, the third one:

\[
\ddot{w} = -n^2w.
\]

Let us first try the general periodic solution,

\[
w(t) = A \sin(Bt + C).
\]

Substitution and twice differentiation yields

\[-AB^2 \sin(Bt + C) = -n^2A \sin(Bt + C),
\]

from which we conclude that

\[
B = \pm n.
\]

---

\(^6\)We can spot here the pseudo-forces occurring in a rotating co-ordinate frame, the centrifugal contributions (somewhat hidden) \(n^2u\) and \(n^2w\), dependent upon place, and the Coriolis terms \(2n\dot{v}\) and \(-2n\dot{u}\), which are velocity dependent.
Thus the solution is

\[ w(t) = A \sin(\pm nt + C), \]

where \( A, C \) are arbitrary constants. The sine decomposition 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:

\[ \dot{w}(t) = nA_1 \cos nt - nA_2 \sin nt. \]

The state transition:

\[
\begin{align*}
    w(t_1) &= A_1 \sin n(t_0 + \Delta t) + A_2 \cos n(t_0 + \Delta t) = \\
    &= A_1 (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) + \\
    & \quad + A_2 (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) = \\
    &= (A_1 \sin nt_0 + A_2 \cos nt_0) \cos n\Delta t + \\
    & \quad + (A_1 \cos nt_0 - A_2 \sin nt_0) \sin n\Delta t = \\
    &= w(t_0) \cos n\Delta t + \frac{1}{n}\dot{w}(t_0) \sin n\Delta t,
\end{align*}
\]

\[
\begin{align*}
    \dot{w}(t_1) &= nA_1 \cos n(t_0 + \Delta t) - nA_2 \sin n(t_0 + \Delta t) = \\
    &= nA_1 (\cos nt_0 \cos n\Delta t - \sin nt_0 \sin n\Delta t) - \\
    & \quad - nA_2 (\sin nt_0 \cos n\Delta t + \cos nt_0 \sin n\Delta t) = \\
    &= n (A_1 \cos nt_0 - A_2 \sin nt_0) \cos n\Delta t - \\
    & \quad - n (A_1 \sin nt_0 + A_2 \cos nt_0) \sin n\Delta t = \\
    &= \dot{w}(t_0) \cos n\Delta t - n\dot{w}(t_0) \sin n\Delta t.
\end{align*}
\]

In matric form:

\[
\begin{bmatrix}
    w \\
    \dot{w}
\end{bmatrix}(t_1) = 
\begin{bmatrix}
    \cos n\Delta t & \sin n\Delta t/n \\
    -n \sin n\Delta t & \cos n\Delta t
\end{bmatrix} 
\begin{bmatrix}
    w \\
    \dot{w}
\end{bmatrix}(t_0). \quad (6.13)
\]

\textbf{6.6.2 The } u \text{ and } v \text{ equations}

\[ \ddot{u} = 2n\dot{v} + 3n^2 u, \quad \ddot{v} = -2n\dot{u}. \]
Another solution

These are to be solved together. Let us try again a periodic solution:

\[ u(t) = A \sin nt + B \cos nt, \quad v(t) = C \sin nt + D \cos nt. \quad (6.14) \]

Substitution yields

\[
-n^2 (A \sin nt + B \cos nt) = \\
= 2n \cdot n (C \cos nt - D \sin nt) + 3n^2 (A \sin nt + B \cos nt), \\
- n^2 (C \sin nt + D \cos nt) = -2n \cdot n (A \cos nt - B \sin nt).
\]

Consider the sine and cosine terms separately and express C and D into A and B:

\[
-n^2 A = -2n^2 D + 3n^2 A, \quad -n^2 B = 2n^2 C + 3n^2 B, \\
- n^2 C = 2n^2 B, \quad -n^2 D = -2n^2 A,
\]

or

\[
-A = -2D + 3A, \quad -B = 2C + 3B, \\
-C = 2B \quad \Rightarrow \quad C = -2B, \quad -D = -2A \quad \Rightarrow \quad D = 2A.
\]

Substitution into equation 6.14 yields the general solution

\[ u(t) = A \sin nt + B \cos nt, \quad v(t) = -2B \sin nt + 2A \cos nt. \]

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 \textit{libration movement}. 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.

\section*{6.7 Another solution}

This is not however end of story. Let us try for a change a \textit{linear} non-periodic solution:

\[ u(t) = Et + F, \quad v(t) = Gt + H. \]
Substitute this into the original differential equations and express $E$ and $G$ into $F$ and $H$:

$$0 = 2nG + 3n^2 (Et + F), \quad 0 = -2nE,$$

from which

$$E = 0, \quad G = -\frac{3}{2}nF.$$

We obtain as the solution

$$u(t) = F, \quad v(t) = -\frac{3}{2}Fnt + H,$$

in which $F$ and $H$ are arbitrary constants. This represents an *orbital motion with a period different from* $\frac{2\pi}{n}$. 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 $H$.

Because the system of differential equations is linear, we may freely combine the above periodic and linear solutions.

### 6.8 The state transition matrix

#### 6.8.1 The general case

Let us look only at the $(u, v)$ plane. Then the general solution is

$$u(t) = A \sin nt + B \cos nt + F,$$

$$v(t) = -2B \sin nt + 2A \cos nt - \frac{3}{2}Fnt + H.$$
We obtain the velocity components too by differentiating:

\[
\dot{u}(t) = nA \cos nt - nB \sin nt,
\]

\[
\dot{v}(t) = -2nA \sin nt - 2nB \cos nt - \frac{3}{2}Fnt.
\]

We write for the initial epoch \(t_0\):

\[
u(t_0) = A \sin nt_0 + B \cos nt_0 + F = S(t_0) + F,
\]

\[
v(t_0) = -2B \sin nt_0 + 2A \cos nt_0 - \frac{3}{2}Fnt_0 + H =
\]

\[
= 2C(t_0) - \frac{3}{2}Fnt_0 + H,
\]

\[
\dot{u}(t_0) = nA \cos nt_0 - nB \sin nt_0 = nC(t_0),
\]

\[
\dot{v}(t_0) = -2nA \sin nt_0 - 2nB \cos nt_0 - \frac{3}{2}Fnt = -2nS(t_0) - \frac{3}{2}Fnt.
\]

Here we have defined

\[
S(t) \overset{\text{def}}{=} A \sin nt + B \cos nt,
\]

\[
C(t) \overset{\text{def}}{=} A \cos nt - B \sin nt.
\]

We write for the epoch \(t_1\) using the summation rules for sine and cosine:

\[
u(t_1) = u(t_0 + \Delta t) =
\]

\[
= A \sin nt_0 (t_0 + \Delta t) + B \cos nt_0 (t_0 + \Delta t) + F =
\]

\[
= A \sin nt_0 \cos n\Delta t + A \cos nt_0 \sin n\Delta t +
\]

\[
+ B \cos nt_0 \cos n\Delta t - B \sin nt_0 \sin n\Delta t + F =
\]

\[
= \cos (n\Delta t) \cdot S(t_0) + \sin (n\Delta t) \cdot C(t_0) + F =
\]

\[
= u(t_0) + (\cos n\Delta t - 1) S(t_0) + \sin (n\Delta t) C(t_0).
\]
Similarly

\[ v(t_1) = 2A \cos n(t_0 + \Delta t) - 2B \sin n(t_0 + \Delta t) - \frac{3}{2} F n t_1 + H = \]

\[ = 2A \cos n t_0 \cos n \Delta t - 2A \sin n t_0 \sin n \Delta t - \]

\[ - 2B \sin n t_0 \cos n \Delta t - 2B \cos n t_0 \sin n \Delta t - \frac{3}{2} F n t_1 + H = \]

\[ = \cos (n \Delta t) \cdot (2A \cos n t_0 - 2B \sin n t_0) - \]

\[ - \sin (n \Delta t) \cdot (2A \sin n t_0 + 2B \cos n t_0) - \frac{3}{2} F n t_1 + H = \]

Substitute into this equation 6.15: \( F = u(t_0) - S(t_0) \), obtaining

\[ v(t_1) = v(t_0) + 2 (\cos n \Delta t - 1) C(t_0) - 2 \sin (n \Delta t) S(t_0) - \]

\[ - \frac{3}{2} u(t_0)n \Delta t + \frac{3}{2} S(t_0)n \Delta t = \]

\[ = v(t_0) + 2 (\cos n \Delta t - 1) C(t_0) + \]

\[ + (\frac{3}{2} n \Delta t - 2 \sin n \Delta t) S(t_0) - \frac{3}{2} u(t_0)n \Delta t. \]

Next, the velocities:

\[ \dot{u}(t_1) = nA \left( \cos n t_0 \cos n \Delta t - \sin n t_0 \sin n \Delta t \right) - \]

\[ - nB \left( \sin n t_0 \cos n \Delta t + \cos n t_0 \sin n \Delta t \right) = \]

\[ = \cos (n \Delta t) \cdot (nA \cos n t_0 - nB \sin n t_0) - \]

\[ - \sin (n \Delta t) \cdot (nA \sin n t_0 + nB \cos n t_0) = \]

\[ = \cos (n \Delta t) \cdot n C(t_0) - \sin (n \Delta t) \cdot n \mathcal{S}(t_0) \]

and

\[ \dot{v}(t_1) = -2nA \cdot (\sin n t_0 \cos n \Delta t + \cos n t_0 \sin n \Delta t) - \]

\[ - 2nB \cdot (\cos n t_0 \cos n \Delta t - \sin n t_0 \sin n \Delta t) - \frac{3}{2} F n = \]

\[ = \cos (n \Delta t) \cdot (-2nA \sin n t_0 - 2nB \cos n t_0) + \]

\[ + \sin (n \Delta t) \cdot (-2nA \cos n t_0 + 2nB \sin n t_0) - \frac{3}{2} F n = \]

\[ = \dot{v}(t_0) - (\cos n \Delta t - 1) \cdot 2n \mathcal{S}(t_0) - 2 \sin (n \Delta t) \cdot n \mathcal{C}(t_0). \]

Here we introduce the notations

\[ s(\Delta t) \overset{\text{def}}{=} \sin n \Delta t, \quad c(\Delta t) \overset{\text{def}}{=} \cos n \Delta t. \]

Summary:

\[ u(t_1) = u(t_0) + (c - 1) S(t_0) + s \mathcal{C}(t_0), \]

\[ v(t_1) = v(t_0) + 2 (c - 1) \mathcal{C}(t_0) + (\frac{3}{2} n \Delta t - 2s) S(t_0) - \frac{3}{2} u(t_0)n \Delta t, \]

\[ \dot{u}(t_1) = nc \mathcal{C}(t_0) - ns \mathcal{S}(t_0), \]

\[ \dot{v}(t_1) = \dot{v}(t_0) - 2n (c - 1) S(t_0) - 2ns \mathcal{C}(t_0). \]
The state transition matrix

Calculate by combining equations 6.15 and 6.17:

\[
\frac{3}{2}nu(t_0) + v(t_0) = -\frac{1}{2}nS(t_0).
\]

From this and from equation 6.16:

\[
S(t_0) = -\left(3u(t_0) + \frac{2}{n}v(t_0)\right), \quad \bar{c}(t_0) = \frac{u(t_0)}{n}.
\]

Substitution yields

\[
\begin{align*}
    u(t_1) &= u(t_0) - (c - 1)\left(3u(t_0) + \frac{2}{n}v(t_0)\right) + \frac{s}{n} \cdot \dot{u}(t_0), \\
    v(t_1) &= v(t_0) + \frac{2}{n} (c - 1) \dot{u}(t_0) - \\
    &\quad - \left(\frac{3}{2}n\Delta t - 2s\right) \cdot \left(3u(t_0) + \frac{2}{n}v(t_0)\right) - \frac{3}{2}u(t_0)n\Delta t = \\
    &= v(t_0) + \frac{2}{n} (c - 1) \dot{u}(t_0) - (6n\Delta t - 6s) u(t_0) - \frac{3n\Delta t - 4s}{n} v(t_0), \\
    \dot{u}(t_1) &= cu(t_0) + s \left(3nu(t_0) + 2\dot{v}(t_0)\right), \\
    \dot{v}(t_1) &= \dot{v}(t_0) + (c - 1) \left(6nu(t_0) + 4\dot{v}(t_0)\right) - 2s\dot{u}(t_0) = \\
    &= 6n (c - 1) u(t_0) - 2s\dot{u}(t_0) + (4c - 3) \dot{v}(t_0).
\end{align*}
\]

As a matrix equation:

\[
\begin{bmatrix}
    \dot{u} \\
    \dot{v} \\
    \dot{\bar{c}}
\end{bmatrix}
(t_1) =
\begin{bmatrix}
    4 - 3c & 0 & \frac{s}{n} - \frac{2(c - 1)}{n} \\
    6s - 6n\Delta t & 1 & \frac{2(c - 1)}{n} & \frac{4s}{n} - 3\Delta t \\
    3ns & 0 & c & 2s \\
    6n(c - 1) & 0 & -2s & 4c - 3
\end{bmatrix}
\begin{bmatrix}
    u \\
    v \\
    \dot{u} \\
    \dot{v}
\end{bmatrix}
(t_0).
\]

This state transition matrix, together with equation 6.13 for the \(w\) co-ordinate, is known under the name “Clohessy–Wiltshire model”, Clohessy and Wiltshire (1960).

We can write this equation 6.18 in partitioned form:

\[
\begin{align*}
    \begin{bmatrix}
        \dot{u}' \\
        \dot{v}' \\
        \dot{\bar{c}}'
    \end{bmatrix}(t_1) &= (\Phi_{11})^1_0 \begin{bmatrix}
        \dot{u}' \\
        \dot{v}' \\
    \end{bmatrix}(t_0) + (\Phi_{12})^1_0 \begin{bmatrix}
        \dot{\bar{c}}'
    \end{bmatrix}(t_0), \\
    \begin{bmatrix}
        \dot{\bar{c}}'
    \end{bmatrix}(t_1) &= (\Phi_{21})^1_0 \begin{bmatrix}
        \dot{u}' \\
        \dot{v}' \\
    \end{bmatrix}(t_0) + (\Phi_{22})^1_0 \begin{bmatrix}
        \dot{\bar{c}}'
    \end{bmatrix}(t_0),
\end{align*}
\]

with, for example,

\[
(\Phi_{11})^1_0 =
\begin{bmatrix}
    4 - 3\cos(n\Delta t) & 0 \\
    6\sin(n\Delta t) - 6n\Delta t & 1
\end{bmatrix}.
\]
6.8.2 The case of small $\Delta t$

Write the system of second-order differential equations

$$\ddot{u} = 2n\dot{v} + 3n^2u, \quad \dot{v} = -2nu,$$

as a first-order system:

$$\frac{d}{dt}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}.$$

For a small time step $^7\Delta t$:

$$\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_1) \approx \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) + \Delta t \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) =$$

$$= \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 3n^2\Delta t & 0 & 1 & 2n\Delta t \\ 0 & 0 & -2n\Delta t & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0).$$

You may verify for each matrix element that this is the same as equation 6.18, with the substitutions $c = \cos n\Delta t \approx 1$ and $s = \sin n\Delta t \approx n\Delta t$.

We can go one step further still with the slightly better approximation $c = \cos n\Delta t \approx 1 - \frac{1}{2}n^2\Delta t^2$:

$$\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_1) \approx \begin{bmatrix} 1 + \frac{3}{2}n^2\Delta t^2 & 0 & \Delta t & n\Delta t^2 \\ 0 & 1 & -n\Delta t^2 & \Delta t \\ 3n^2\Delta t & 0 & 1 - \frac{1}{2}n^2\Delta t^2 & 2n\Delta t \\ -3n^3\Delta t^2 & 0 & -2n\Delta t & 1 - 2n^2\Delta t^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0).$$

Self-test questions

1. Write out explicitly the other three state transition matrices $(\Phi_{12})_0^1, (\Phi_{21})_0^1, (\Phi_{22})_0^1$ in the Clohessy–Wiltshire model 6.18.

---

^7“Small” in relation to the orbital period, i.e., $n\Delta t \ll 1.$
Exercise 6–1: Kepler orbit

1. Derive the dynamic model of the Kepler state vector. Assuming that the force field is central, write explicitly the following linearised dynamic model equation:

\[ \frac{d}{dt} \Delta a = F \cdot \Delta a, \]

in which \( a \) \( \text{def} \) \( \begin{bmatrix} a & e & M & i & \omega & \Omega \end{bmatrix}^T \). You need to linearise the original non-linear model by choosing suitable approximate values. The delta quantities are referred to these. You also need Kepler’s third law, with \( P \) the orbital period:

\[ GM_\oplus P^2 = 4\pi^2 a^3. \]

2. Due to flattening of the Earth, the right ascension \( \Omega \) of the orbit’s ascending node changes slowly according to the following equation:

\[ \dot{\Omega} = -\frac{3}{2} \sqrt{\frac{GM_\oplus}{a^3}} \left( \frac{a_\oplus}{a} \right)^2 J_2 \cos i. \]

A circular orbit is assumed, \( e \approx 0 \). \( a_\oplus \) is the equatorial radius of the Earth, \( J_2 \) is the so-called dynamic flattening (a dimensionless number).

How does this affect the matrix \( F \) of the dynamic model derived above?

3. [Complicated] How does one transform a rectangular state vector \( x \) into a Kepler vector \( a \) and the reverse? In other words, we want in the following equation

\[ x = Aa \]

the matrix \( A \) written out in components, in other words linearization. For simplicity assume that \( e \) is small.

Hint: write first \( x \) as a function of \( a \) and calculate the partial derivatives.

4. In a central force field, if we write

\[ x(t_1) = \Phi_0^1 x(t_0), \]

find the matrix \( \Phi_0^1 \) approximately (series expansion), if \( \Delta t = t_1 - t_0 \) is small. (Consult the literature.)
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 space as a result of the Earth’s rotation? Assuming that the Earth’s rotation is uniform and the place of the station fixed, write a dynamic model for the station co-ordinates.

6. Write the **observation equations** for the case, where we measure from the ground station *the distance* using a laser range finder. In other words, write the observable as a function of the elements of the state vector \(x\), and linearize.

7. Write the observation equations for the case of **GPS**, where the observation quantity is the *pseudorange* (pseudorandom code measurement) to the satellite. What new problem comes up?

8. What new problem comes up in the case, that the observable is the *carrier phase*?

---

**Exercise 6–2: Rendez-vous**

1. Specialize the state transition matrix 6.18 to the case where \(n\Delta t = \pi\). If given are, at starting epoch \(t_0\), the co-ordinates of place (for simplicity, we forget about the third co-ordinate \(w\)):

\[
\begin{bmatrix}
u' \ n(t_0)
\end{bmatrix}
\]

calculate the necessary initial velocity vector

\[
\begin{bmatrix}
u' \ n(t_0)
\end{bmatrix}
\]

to reach the origin or quarry at epoch \(t_1\), in other words,

\[
\begin{bmatrix}
u' \ n(t_1)
\end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

If the true velocity vector on epoch \(t_0\) is different, then rocket power will be needed to change it.

This is the so-called *rendez-vous* problem of two spacecraft. Analyse the situation.
2. Note that at the moment of *rendez-vous* $t_1$ the craft’s velocity

$$\dot{u}'(t_1) = \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix}(t_1)$$

within the system of the quarry is *not* zero. Calculate it. More use of rocket power is necessary if a docking is intended.

3. The craft carries on board a device for measuring the distance $\|u\|$ between them, as well as their velocity of approach—$\|\dot{u}\|$ Form the *observation equations*. 
Measuring gravity from the air and from space

The saying is well known:

“one guy’s noise is the other guy’s signal.”

Inertial navigation is based on assuming the Earth’s gravity field as known. Then, from the starting position \( x(t_0) \) and the starting velocity \( v(t_0) \), we can calculate forward in time to obtain the position and speed \( x(t), v(t) \) at a later time \( t \).

Conversely however, if there is an independent source of information that gives the current place and velocity precisely enough — such as any GNSS system — then we can harness inertial technology to survey the Earth’s gravity field.

With the help of a well working GNSS satellite navigation system it is nowadays possible to perform gravimetric measurements from the air. Also the study of the gravity field with the aid of satellites is based on the use of a GNSS system, continuously tracking the satellite’s accurate three dimensional position.

Let the position of the aircraft or satellite as a function of time be \( x(t) \), and its discrete measurement time series \( x_t \). Then the geometrical acceleration can be approximated as follows:

\[
\frac{d^2}{dt^2} x_{t_i} \approx \frac{x_{t_{i+1}} + x_{t_{i-1}} - 2x_t}{\Delta t^2},
\]

in which \( \Delta t \) is the time interval between successive epochs \( t_{i+1} - t_i \).

Let us assume that, simultaneously, the aircraft’s sensed acceleration \( a \) is measured (“gravity”) for example with acceleration sensors. At this point, for simplicity, we also assume that \( x \) and \( a \) are given in the same co-ordinate frame, and that especially the directions of the acceleration
measurement axes are the same as those of the location co-ordinate axes.

Then, in an inertial reference frame it holds that

\[ g = \frac{d^2}{dt^2}x + a, \]  

(7.1)
in words:

gravitation \( g \) is the sum of the “gravity” \( a \) sensed inside a vehicle, and geometrical acceleration.

### 7.1 Vectorial airborne gravimetry

If an aircraft carries both an inertial device and a GNSS receiver, we can measure both \( \frac{d^2}{dt^2}x \) \( t_i \) and \( a(t_i) \), and we can calculate \( g(t_i) \). This is a method to survey the gravity field from the air. In practice, the data streams generated from 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 vectorial airborne gravimetry, gravity is measured as a three-dimensional vector, The data rate is typically high, many epochs per second. Because of the motions of the aircraft, the variations over time of both geometric acceleration \( \frac{d^2}{dt^2}x \) and sensed acceleration \( a \) are large, thousands of milligals, but the final determination precision of \( g \) can be as good as a couple of milligals.

Vectorial airborne gravimetry is not on the same level, precision wise, as the next technique to be introduced, scalar airborne gravimetry. This is because the accelerometers in the inertial device, as precise as they are, suffer more from systematic problems, such as drift, than the best gravimeters.

### 7.2 Scalar airborne gravimetry

In this technique, a traditional gravimeter — a device for measuring gravity — is used. The gravimeter is modified in a way that makes it possible to make measurements in strongly varying gravity-acceleration environments. The modification, damping, is the same as the one that is made to make measurements at sea possible. The gravimeter is mounted on a stabilised platform, the stabilization being done with the aid of gyroscopes, see section 5.5.
The gravimeter measures the gravity acceleration “sensed” inside the vehicle, but only in the direction of the local vertical or plumbline. If the direction of the local vertical is the unit vector \( \mathbf{n} \) (downward), the measured quantity is \( \langle \mathbf{n} \cdot \mathbf{a} \rangle \).

Multiply \( \mathbf{n} \) with equation 7.1:

\[
\langle \mathbf{n} \cdot \mathbf{g} \rangle = \frac{d^2}{dt^2} \langle \mathbf{n} \cdot \mathbf{x} \rangle + \langle \mathbf{n} \cdot \mathbf{a} \rangle = \| \mathbf{g} \| \overset{\text{def}}{=} g, 
\]

because the plumbline is in the direction of gravity.

In practice the equation 7.2 is written in a system moving with the aircraft and rotating with the solid Earth, so we obtain

\[
g = \langle \mathbf{n} \cdot \mathbf{a} \rangle + \frac{d^2}{dt^2} \langle \mathbf{n} \cdot \mathbf{x} \rangle + \text{Centrifugal W–E} + \text{Centrifugal S–N}
\]

\[
= a_{\text{down}} - \frac{d}{dt} v_{\text{up}} + \left( \frac{v_{\text{east}}}{N + h} + 2\omega \cos \varphi \right) v_{\text{east}} + \frac{v_{\text{north}}^2}{M + h},
\]

in which \( v_{\text{up}}, v_{\text{east}}, v_{\text{north}} \) are the “up”, “east”, “north” components of the velocity, \( a_{\text{down}} \overset{\text{def}}{=} \langle \mathbf{n} \cdot \mathbf{a} \rangle \) is the measured acceleration inside the vehicle in the “down” direction, \( \omega \) is the angular velocity of the Earth’s rotational motion, and \( M \) and \( N \) are the Earth’s radii of curvature in the meridional (South-North) and West-East directions, i.e., the meridional and transversal radii of curvature. \( h \) and \( \varphi \) are height and latitude. In the above equation, the two last terms are called the Eötvös correction, see Wei and Schwarz (1996).

### 7.3 Studying gravitation from space

In equation 7.1 the quantity \( \mathbf{a} \) is about the magnitude of the Earth’s surface gravity (about \( 10 \text{ m/s}^2 \)), while the geometrical acceleration \( \frac{d^2}{dt^2} \mathbf{x} \) is much smaller. In the ideal case, the geometric acceleration would be zero, corresponding to measurements on the surface of the Earth. In both sea and airborne gravimetry it differs however from zero and complicates the accurate measurement of gravity. The movements of the vehicle are, from the viewpoint of measuring, disturbances.

In the measurement of the gravity field from space, the situation is the opposite. The local gravity \( \mathbf{a} \) sensed inside the satellite is zero (weightlessness) or very close to zero. The geometrical acceleration...
\( \frac{d^2}{dt^2} x \) is almost the magnitude of gravity at the Earth’s surface, because the satellite “falls” freely the whole time while flying in orbit. The geometrical acceleration is all the time being measured with the help of the GNSS system — so-called “high-low satellite-to-satellite tracking” — and also the satellite’s own, non-inertial motion \( a \) is measured with the aid of accelerometers. Its largest cause is atmospheric friction (drag), because the orbit of a satellite for measuring the gravity field is chosen to be as low as possible, the typical height of the orbit being 250 – 400 km.

During recent decades, three different gravity missions have flown: CHAMP, GRACE and GOCE.

- **CHAMP** (GFZ, CHAMP — Challenging Minisatellite Payload), a small German satellite, operated from 2000 to 2010 and produced a large amount of data.

- **GRACE** (University of Texas, GRACE — Gravity Recovery and Climate Experiment), a small American–German satellite pair which measured with its special equipment the accurate distance between two satellites ("Tom" and "Jerry") flying in tandem, in order to survey the Earth’s gravity field’s temporal changes. The mission, from 2003 to 2018, has been a great success. An animation of its results can be found here: Wikimedia Commons, GRACE animation. Currently operating is the successor, GRACE-FO ("GRACE Follow-on").

- **GOCE** (Gravity Field and Ocean Circulation Explorer) surveyed the Earth’s gravity field 2009–2013 in great detail using a gravity gradiometer, see ESA, Introducing GOCE. The GOCE satellite contained an ionic engine in order to compensate the air drag and make a low orbit possible. It was quite a challenge to separate the gravity-gradient measurements from the effects of air drag and the satellite’s own rotation as it circled the Earth.

In all the satellites there are a GNSS positioning system and accelerometers included, in the case of GOCE even an array — a gradiometer — counting six extremely sensitive accelerometers.

### 7.4 Using the Kalman filter in airborne gravimetry

Starting from equation 7.1 we can write, including the “dynamic noise” \( n \):

\[
\frac{d^2}{dt^2} x = a - g + n,
\]
Using the Kalman filter in airborne gravimetry

\[ \frac{d}{dt} \begin{bmatrix} v \\ x \\ \delta g \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{\tau} I \end{bmatrix} \begin{bmatrix} v \\ x \\ \delta g \end{bmatrix} + \begin{bmatrix} a - \gamma \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} n_a \\ 0 \\ n_g \end{bmatrix}. \]

Here \( a = a(t) \) is a measured quantity, but \( g = g(t) \) is not.

Write

\[ g = \gamma + \delta g, \]

in which \( \gamma \) is a suitable reference value (e.g., normal gravity) and \( \delta g \) is the gravity disturbance. We can model \( \delta g \) empirically as a Gauss–Markov process, equation 2.11, so we can write

\[ \frac{d}{dt} \delta g = -\frac{\delta g}{\tau} + n_g, \]

in which \( \tau \) is a suitable empirical time constant, the choice of which depends on the behaviour of the local gravity field (correlation length) and the flying speed and height.

Now the Kalman filter’s dynamic equations are

\[ \frac{d}{dt} \begin{bmatrix} v_0 \\ x_0 \\ \Delta x \\ \Delta v \\ \delta g \end{bmatrix} = \begin{bmatrix} 0 & 0 & -M & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\tau} I \end{bmatrix} \begin{bmatrix} v_0 \\ x_0 \\ \Delta x \\ \Delta v \\ \delta g \end{bmatrix} + \begin{bmatrix} a(t) - \gamma \end{bmatrix} + \begin{bmatrix} n_a \\ 0 \\ 0 \end{bmatrix}. \]

So, the length of the state vector is 9. The matrix is of size \( 3 \times 3 \) and consists of \( 3 \times 3 \) sized submatrices, so we have a \( 9 \times 9 \) matrix in total.

A more sophisticated way of handling takes into consideration that the gravity \( g \) is a function of place \( x \), which we don’t actually know:

\[ g(x) = \gamma(x_0) + M(x) \Delta x + \delta g \iff \gamma(x) = \gamma(x_0) + M(x) \Delta x, \]

in which \( x_0 = x_0(t) \) is the approximate position given by the linearization, see below. Here appears the gravitational-gradient tensor \( M \) pops up, equation 3.7.

Then, also \( x \) and \( v \) must be linearized, and linearized state elements \( \Delta x \equiv x - x_0, \Delta v \equiv v - v_0 \) must be used in the state vector. The equations defining the approximate values are

\[ \frac{d}{dt} \begin{bmatrix} v_0 \\ x_0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_0 \\ x_0 \end{bmatrix} + \begin{bmatrix} a(t) - \gamma(x_0) \end{bmatrix}. \]

The final result is

\[ \frac{d}{dt} \begin{bmatrix} \Delta v \\ \Delta x \\ \delta g \end{bmatrix} = \begin{bmatrix} 0 & -M & -I \\ 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{\tau} I \end{bmatrix} \begin{bmatrix} \Delta v \\ \Delta x \\ \delta g \end{bmatrix} + \begin{bmatrix} n_a \\ 0 \\ n_g \end{bmatrix}. \]
The observation equations for the update are

\[ \vec{\ell}_i = \vec{x}(t_i) + m, \]

in which the “noise vector” \( m \) describes the statistical uncertainty of GNSS navigation.

For both the dynamic noise \( n_a \) and the observation noise \( m \) we need to build suitable statistical models (variances \( Q(t) \) and \( R \)) based on the properties of the measuring devices.

### 7.5 Present state of airborne gravimetry

One of the first successful airborne gravimetric projects was Brozena (1992), Greenland’s gravity survey in the frame of the Greenland Aerogeophysics Project.

Many later measurements, often in Arctic or Antarctic locations, can be mentioned (Forsberg et al., 1996, 2011). The logistic requirements of working there are typically “challenging,” see figure 7.2.

Airborne gravimetry is a suitable technique, if the area to be surveyed is large and there are no earlier gravity surveys available. Homogeneity is one of airborne and space gravimetry’s advantages: the quality of the
measurement is the same over large areas and systematic errors over long distances are small. This is important especially if the gravimetric data is meant for the determination of a geoid model.

Recent examples of airborne gravity surveys include Ethiopia (Bedada, 2010), Mongolia (Munkhtsetseg, 2009), Indonesia (2010), and many more.
8 GPS navigation and base stations

About the subject GPS and Navigation, see e.g. Strang and Borre (1997) pages 495–514.

8.1 The GPS system

The GPS system consists of three segments: the space, control and user segment. The space segment consists of at least 24, in practice 26–30 satellites, including “active spares.” Four orbital planes, in each six satellites. The geometry, including satellite identities, repeats after 23h56m (one sidereal day, two GPS satellite orbital periods). The orbital inclination is 55°.

The existing GPS satellites belong to a number of technology generations or blocks. The old satellites of Block I and II are no longer in operation. The newest type is Block IIIA, the first of which 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, almost anytime, one can see at least four satellites above the horizon, usually more.


The control segment consists (2017) of

- sixteen tracking stations. 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 computing and control centre (MCS, Master Control Station) is in Colorado Springs, and its backup (Alternate MCS) is in Vandenberg, California.

8.2 GPS satellites and signal structure

The radio signal transmitted by a GPS satellite consists of a carrier wave and so-called pseudo random codes modulated on top of it. Both may be used for positioning.

In the following we present only the original signal structure, before the modernization programme.

Carrier wave: wavelength \( \sim 20 \text{ cm} \), precision of positioning \( 1\% \) of this, \( \sim 2\text{ mm} \). Precise navigation is based on measuring the phase of the carrier wave. For this purpose, always dual frequency receivers are used. See table 8.1.

Problem: all waves look the same, so we need ambiguity resolution.

Code: The “pseudo-wavelength” is 30 m (The P code “chip rate”, or distance between bits on the radio wave in flight): positioning precision is \( \sim 1\% \) of this, \( \sim 30 \text{ cm} \). The codes are modulated using so-called phase modulation, see figure 8.2.

More precisely: the chip rate or bit rate of the C/A code corresponds to a “wavelength” of 300 m, the corresponding number
Pseudorange smoothing

In this method, the absolute pseudorange comes from the code measurement, but its fractional-wavelength part from the phase measurement.

In many kinematic applications of GPS, it is advantageous to smooth the raw pseudorange code observables by using the much more smooth and geometrically precise, but ambiguous, carrier phase measurements.

Let us assume we have as observations the code measurements \( p_1 \) and \( p_2 \) (in metres) and the carrier phases \( \phi_1 \) and \( \phi_2 \) (in angular units,

<table>
<thead>
<tr>
<th>Carrier</th>
<th>Frequency (MHz)</th>
<th>Wavelength (cm)</th>
<th>Multiple of base frequency (10.23 MHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_1</td>
<td>1575.42</td>
<td>19.0</td>
<td>154\times</td>
</tr>
<tr>
<td>L_2</td>
<td>1227.60</td>
<td>24.4</td>
<td>120\times</td>
</tr>
</tbody>
</table>
Table 8.2. The different pseudo-random codes (PRC) modulated on the original GPS signal. The modernization of the GPS system adds to this a largeish number of new codes, both on the old carrier frequencies $L_1$ and $L_2$ and on the new frequency $L_5$.

<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Name</th>
<th>Modulation frequency</th>
<th>Repeat period</th>
<th>Carrier wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/A</td>
<td>Coarse / Acquisition</td>
<td>1.023 Mb/s</td>
<td>1 ms</td>
<td>L1</td>
</tr>
<tr>
<td>P, P(Y)</td>
<td>Precise / Protected</td>
<td>10.23 Mb/s</td>
<td>1 week</td>
<td>L1, L2</td>
</tr>
<tr>
<td></td>
<td>Navigation message</td>
<td>50 bit/s</td>
<td>continuous</td>
<td>L1, L2</td>
</tr>
</tbody>
</table>

† The Y, or P(Y), code is obtained by modulating — XOR-ing — the P code with the classified W code, to prevent “spoofing”, i.e., generation and transmission of a fake GPS signal.

i.e., radians), at a time $t$.

Firstly, we can construct a prediction equation for the current (a priori) pseudorange from the previous one:

$$p^-(t_i) = p(t_{i-1}) + \frac{\lambda}{2\pi} (\phi(t_i) - \phi(t_{i-1})). \quad (8.1)$$

This equation is valid for both frequencies 1 and 2, and also for the widelane observables defined as:

$$p_{WL} = \frac{f_1 p_1 - f_2 p_2}{f_1 - f_2}, \quad \phi_{WL} = \phi_1 - \phi_2.$$  

Note that eq. (8.1) can be interpreted as a Kalman filter dynamic equation: the state is $p(t)$ and its variance matrix can be modelled as $P^-(t)$. The phase correction term $\phi(t_i) - \phi(t_{i-1})$ may be considered known, which is justified given its superior precision compared to code measurements.

Next, we add to this Kalman filter an observation equation: it is simply for the current observation $p(t_i)$, the precision of which can be given as $R_i$. Now the correction equation is

$$p^+(t_i) = p^-(t_i) + KH (p^-(t_i) - p(t_i)),$$

in which

$$H = \begin{bmatrix} 1 \end{bmatrix},$$

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

and thus

$$p^+(t_i) = \frac{R_i}{\Sigma^-(t_i) + R_i} p^-(t_i) + \frac{\Sigma^-(t_i)}{\Sigma^-(t_i) + R_i} p(t_i).$$
Differential navigation

So: the *a posteriori* pseudorange is a weighted linear combination of the predicted and carrier-smoothed one and the currently observed one.

For the variance propagation we find

$$\Sigma^+(t_i) = (I + KH) \Sigma^-(t_i) = \frac{R_i}{\Sigma^-(t_i) + R_i} \Sigma^-(t_i).$$

(For the variance propagation in the dynamic model, between epochs, we have simply $\Sigma^-(t_i) = \Sigma^+(t_{i-1})$.)

It is possible to include *cycle slip detection* into the procedure: the testing variate is the difference

$$p^-(t_i) - p(t_i)$$

of which we know the mean error to be:

$$\sigma = \sqrt{H \Sigma^- H^T + R} = \sqrt{\Sigma^- + R}.$$

This works best for the wide lane linear combination because of its large effective wavelength, 86 cm.

This Kalman filter can run as a continuous process in the receiver—or in post-processing software, but then without the real-time advantage. The output $p^+(t)$ is significantly smoothed compared to the input $p(t)$.

\section{8.4 Differential navigation}

In practice the problem with both methods is the imprecision of the satellite orbits. In real time, by using *broadcast ephemeris*, one can achieve at best ±1 m positioning precision.

**Solution:** differential navigation, i.e., the use of a base station or base stations. If the distance from the base station is sufficiently small, the greatest part of the orbit error will cancel out from the final result, see figure 8.3.

The precision of differential positioning may be easily estimated using a geometric argument. See figure 8.4. If the geometric precision of the satellite orbit (*broadcast ephemeris*) is called $\Delta$ and the distance of the satellite from the observer $s$ (in practice over 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 we form table 8.3.
Figure 8.3. 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 (like the satellite’s clock error) cancels for the most part out in differential measurement.

Figure 8.4. The geometry of computing the precision of differential positioning.

Table 8.3. The precision of positioning as a function of the length of the vector and the orbit error. 1 m corresponds to the precision of today’s broadcast ephemeris.

<table>
<thead>
<tr>
<th>Vector length (km)</th>
<th>Orbit error (m)</th>
<th>Positioning error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>50</td>
</tr>
</tbody>
</table>
Differential GPS is widely used also in traditional geodetic GPS processing. Every time when 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 of two satellites but also the observations of 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 clocks
- Atmosphere (ionosphere, troposphere) errors
- Errors caused by the antenna’s phase delay pattern, depending on the direction (azimuth, elevation) and thus on the local vertical.

A radio link is used in real time differential methods to transfer the original observations or corrections from one ground station (the position of which is assumed to be known) to another (unknown, often moving) ground station. The various methods

- use either the phase of the carrier wave, or the delay of the PRN code modulated on the carrier wave, and
- can use one reference station for a whole area, or more stations to make interpolation possible; and those
- can interpolate a ready result for the user (on a known position; 1-to-1 method) or let the user interpolate himself (1-to-many method).
- Coverage can be local (the commercial services TrimNet VRS and Leica SmartNet in Finland) or global (GDGPS, Jet Propulsion Lab).
- A radio broadcast network, a radio-modem pair, or a cell phone can provide the data link.

### 8.5 The RTCM standard

Radio Technical Commission for Maritime Services (RTCM, [http://www.rtcm.org/](http://www.rtcm.org/)) SC-104 is an independent organization created in 1947. Member organizations are over 100, e.g., manufacturers of radio navigation equipment, state agencies responsible for radio positioning, shipbuilders, positioning service providers and academic institutions.

RTCM Special Commission 104 designed the standard for a GPS differential data service, carrying the name RTCM SC-104 (or RTCM-
104, or “RTCM”). A version in common use is 2.3.

Additionally there exists a version 3.0, which however is not downward compatible with the versions 2.x. It is suitable for real-time kinematic measurement and uses a more efficient data transfer mechanism than the 2.x protocol.

Message types are listed in table 8.4.

There are many devices on the market that send and can use the message types above in differential navigation either by using the phases of the carrier waves (RTK technique) or the pseudorandom codes modulated to the carrier waves (DGPS-technique). In both cases the navigation is real-time, the “age” of the position solution stays always below the specified limiting value.

The base station, the position of which is measured precisely using static geodetic positioning, sends the RTCM-messages. Because the position is known, its possible to calculate, using the satellite orbits,

### Table 8.4. The message types of the RTCM SC-104 format.

<table>
<thead>
<tr>
<th>Message Type</th>
<th>Message Title</th>
<th>Message Type</th>
<th>Message Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DGPS corrections</td>
<td>13</td>
<td>Ground transmitter parameters</td>
</tr>
<tr>
<td>2</td>
<td>Delta DGPS corrections</td>
<td>14</td>
<td>Surveying auxiliary message</td>
</tr>
<tr>
<td>3</td>
<td>Reference station parameters</td>
<td>15</td>
<td>Ionospheric / tropospheric message</td>
</tr>
<tr>
<td>4</td>
<td>Carrier surveying information</td>
<td>16</td>
<td>Special message</td>
</tr>
<tr>
<td>5</td>
<td>Constellation health</td>
<td>17</td>
<td>Ephemeris almanac</td>
</tr>
<tr>
<td>6</td>
<td>Null frame</td>
<td>18</td>
<td>Uncorrected carrier phase measurements</td>
</tr>
<tr>
<td>7</td>
<td>Marine radiobeacon almanacs</td>
<td>19</td>
<td>Uncorrected pseudorange measurements</td>
</tr>
<tr>
<td>8</td>
<td>Pseudolite almanacs</td>
<td>20</td>
<td>RTK carrier phase corrections</td>
</tr>
<tr>
<td>9</td>
<td>High rate DGPS corrections</td>
<td>21</td>
<td>RTK pseudorange corrections</td>
</tr>
<tr>
<td>10</td>
<td>P code DGPS corrections</td>
<td>22-59</td>
<td>Undefined</td>
</tr>
<tr>
<td>11</td>
<td>C/A code L1/L2 delta corrections</td>
<td>60-63</td>
<td>Differential Loran C messages</td>
</tr>
<tr>
<td>12</td>
<td>Pseudolite station parameters</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
what the pseudo distance to each satellite should be\(^1\). By subtracting this from the measured values, we get the \textit{correction} to be coded into the message (message types 1, 2, 20 and 21).\(^2\) 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 twenty kilometers.

### 8.6 RTCM-over-Internet (NTRIP protocol)

NTRIP stands for “Networked Transport of RTCM via Internet Protocol.” It resulted from the realization 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 existing mobile telephony base-station network enables the transmission of corrections over the Internet to GNSS instruments operating in the field. The 3G and 4G technology generations already have a well sufficient data transfer capacity for this.

The protocol used is a “streaming” one based on the Hypertext Transfer Protocol (HTTP) in massive use for transferring Worldwide Web pages over the Internet. It has been extended for the purpose of transferring GNSS data streams. It scales well to providing multiple data streams to even a large number of users. Different data streams or “sources” are distinguished as named “mount points.” See BKG, NTRIP v. 1.0.

---

\(^1\)This logic does not consider the receiver clock error. But, when using double differences for positioning, as one does in real-time positioning, this does not matter.

\(^2\)In the case of RTK, often one rather transmits the original phase observations, types 18 and 19, but conceptually the matter is the same.
9.1 Introduction

The abbreviation RTK stands for Real-Time Kinematic.

The kinematic measuring method was invented by the American Benjamin Remondi. It is based on the idea, that the receiver is “locked” to the phase of the GPS carrier wave and as long as the lock holds (no “cycle slip” happens), the integer value of the phase of the carrier wave is known. See figure.

9.2 GPS observations and unknowns

GPS observations are described as pseudoranges and given by the equation

\[ p = \rho + c (\Delta T - \Delta t) + d_{\text{ion}} + d_{\text{trop}}, \]  \hspace{1cm} (9.1)

in which

\[ \rho = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2} \]

is the spatial distance between satellite \([x\ y\ z]^T\) and ground station \([X\ Y\ Z]^T\),

\[ \Delta t \]

is the satellite clock error,

\[ \Delta T \]

is the receiver clock error, and

\[ d_{\text{ion}}, d_{\text{trop}} \]

are the ionospheric and tropospheric effects on signal propagation.

This equation can be written in different ways, depending on what we consider to be the unknowns to be estimated by the Kalman filter. Available unknowns that can be included in the Kalman filter are

\[ \bar{x} = \begin{bmatrix} x & y & z \end{bmatrix}^T \] satellite location,

\[ \bar{X} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T \] receiver location,
Figure 9.1. Real-time kinematic GPS measurement.

\[ \Delta t, \Delta T \] satellite and receiver clock errors.

9.2.1 Satellite orbit determination

We can propose the following observation equation (\( m_p \) representing the observational uncertainty):

\[
p = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2 + c(\Delta T - \Delta t) + d_{\text{ion}} + d_{\text{trop}} + m_p}.
\]

This is the observation equation for orbit determination. In it, the ground station (tracking station) position is given and treated as non-stochastic: \[ [X \ Y \ Z]^T \]. The satellite position is stochastic and to be estimated by the filter. The same applies for the clocks: the tracking station clock is assumed known relative to UTC, the deviation being \( \Delta T \). The satellite clock error \( \Delta t \), however, is being estimated.
GPS observations and unknowns

For this situation we identify the state vector as

\[
\mathbf{x} = \begin{bmatrix}
\mathbf{x} \\
\mathbf{v} \\
\Delta t \\
\mathbf{d}_{\text{ion}} \\
\mathbf{d}_{\text{ trop}}
\end{bmatrix}.
\]

As before, we introduced the velocity vector \( \mathbf{v} \), so we can write the Kalman dynamical equations as first-order differential equations.

Next, we have to decide how to model the time behaviour of these various state vector elements. For the location \( \mathbf{x} \) this is simple: we have

\[
\frac{d}{dt}\mathbf{x} = \mathbf{v},
\]

exactly. For the velocity, we use the equation for a central force field, and we linearize. As approximate values we can use available orbital predictions, e.g., broadcast or precise ephemeris: call these \( \mathbf{x}_0, \mathbf{v}_0, \Delta t_0 \) (these always also contain satellite clock corrections!). Then we may define linearized (differential) state vector elements

\[
\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_0, \quad \Delta \mathbf{v} = \mathbf{v} - \mathbf{v}_0, \quad \Delta (\Delta t) = \Delta t - \Delta t_0.
\]

Now, the linearized equations for \( \mathbf{x}, \mathbf{v} \) are

\[
\frac{d}{dt} \begin{bmatrix}
\Delta \mathbf{x} \\
\Delta \mathbf{v}
\end{bmatrix} = \begin{bmatrix}
0 & 1 \\
\mathcal{M} & 0
\end{bmatrix} \begin{bmatrix}
\Delta \mathbf{x} \\
\Delta \mathbf{v}
\end{bmatrix} + \begin{bmatrix}
0 \\
\mathbf{n}_a
\end{bmatrix},
\]

where \( \mathcal{M} \) is the earlier derived (for a central force field, equation 3.7) gravitational-gradient tensor, \( \mathbf{I} \) is the \( 3 \times 3 \) unit matrix, and \( \mathbf{n}_a \) is here introduced as the dynamic noise of satellite motion.

How do we model the behaviour of the satellite clock \( \Delta t \)? Often, this is done as a random walk process. As follows:

\[
\frac{d}{dt} \Delta t = \mathbf{n}_t.
\]

Modelling the tropospheric and ionospheric propagation effects is trickier. Note that we are here talking about the slant delay due to these atmospheric components along the satellite-receiver path, and most of the change in this delay will be due not to physical atmospheric changes, but rather, satellite motion causing the path to move to a different place in the atmosphere.
First order autoregressive (AR(1), also Gauss-Markov) modelling is often used in this case, with a pragmatic choice of the time parameter \( \tau \). This could be a few hours, i.e., a fraction of the time during which the GPS satellite is above the horizon. A significant improvement is obtained by using residual ionosphere or troposphere corrections, i.e., differences relative to some suitable a priori model. The notation becomes then \( \Delta d_{ion}, \Delta d_{trop} \). For the ionosphere, this could be the model included with the satellite broadcast ephemeris (The Klobuchar model, very simple), or the published IONEX models (not available in real time). For the troposphere, the standard Hopfield or Saastamoinen models may be considered.

Summarizing:

\[
\begin{bmatrix}
\Delta x \\
\Delta y \\
\Delta (\Delta t) \\
\Delta d_{ion} \\
\Delta d_{trop}
\end{bmatrix}
= \begin{bmatrix}
0 & 1 & M & 0 \\
0 & 0 & -1/\tau_{ion} & -1/\tau_{trop}
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta y \\
\Delta (\Delta t) \\
\Delta d_{ion} \\
\Delta d_{trop}
\end{bmatrix}
+ \begin{bmatrix}
0 \\
n_a \\
n_t \\
n_{ion} \\
n_{trop}
\end{bmatrix}
\]

9.2.2 Station position determination

Starting from the same equation 9.1 we construct a different observation equation, as follows:

\[
p = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2 + c (\Delta T - \Delta t)^2 + d_{ion} + d_{trop} + m_p}.
\]

This is the observation equation for geodetic positioning. Here, the satellite orbital elements and clock are assumed known, i.e., \([x \ y \ z]^T\) and \(\Delta t\) are known or precisely computable from available ephemeris. Now the state vector is

\[
x = \begin{bmatrix}
X \\
V \\
\Delta T \\
\Delta d_{ion} \\
\Delta d_{trop}
\end{bmatrix},
\]

in which \(V\) is the velocity vector of station motion. Here, the new problem is to model the behaviour of the \(X, V\) of the ground station.

In case the ground station is fixed, we may choose as the model

\[
V \overset{\text{def}}{=} \frac{\text{d}}{\text{d}t} x = 0.
\]
If we know that the stations are moving, but slowly and with constant velocity (e.g., plate tectonics, postglacial rebound), we may write

\[
\frac{d}{dt} \begin{bmatrix} X \\ V \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ V \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}.
\]

The Kalman filter will gradually improve the estimates \( \hat{X}, \hat{V} \) over time as more observations are being processed. Some existing GPS processing software (GYPSY/OASIS) use the Kalman filter in this way.

For moving vehicles (aircraft, e.g.) it gets more complicated. One could use the knowledge that the acceleration of the vehicle is bounded, and model it as a coloured noise (Gauss–Markov) process. According to equation 2.13, the variance of such a process is \( Q/2k \), when the process equation is

\[
\frac{d}{dt} A = -k A + n_A.
\]

Now let \( \tau_A = 1/k \) be the time constant of the motion (typically something like a second, the time in which the vehicle can manoeuvre), and \( \alpha \) the typical scale of the accelerations occurring. By putting

\[
\frac{Q}{2k} = \frac{1}{2} Q \tau_A = \alpha
\]

we obtain for the variance of the driving noise \( n_A \):

\[
Q = \frac{2\alpha}{\tau_A}.
\]

Thus we get as the complete dynamic equation:

\[
\frac{d}{dt} \begin{bmatrix} X \\ V \\ A \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \\ 0 & 0 & -1/\tau_A \end{bmatrix} \begin{bmatrix} X \\ V \\ A \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 2\alpha n_1/\tau_A \end{bmatrix},
\]

in which \( n_1 \) stands for “unit variance white noise,” three-dimensional vectorial in this case.

Both \( \alpha \) and \( \tau_A \) will depend on the kind of vehicle we are considering. Large \( \alpha \) and short \( \tau_A \) is often referred to as a “high dynamic” environment, which is challenging for designing GPS receivers.

9.2.3 About clock modelling

Clocks are typically modelled as random-walk processes, see equation 9.2:

\[
\frac{d}{dt} c = n_c,
\]
where now $c$ is the time error, i.e., the difference between clock reading and “true” time. (We changed the notation in this section in order to prevent later mix-ups.)

From equation 2.9 we know that the autocovariance of random walk is

$$A_c(t_1, t_2) = Q(t_1 - t_0),$$

with $Q$ the variance of the white noise process $n_c$, and $t_0$ some suitable starting time. We see that the variance grows linearly with time.

Let us compute the difference between two values $\delta c \overset{\text{def}}{=} c(t_2) - c(t_1)$. The variance of this difference is

$$\text{Var}\{\delta c\} = \text{Var}\{c(t_2)\} + \text{Var}\{c(t_1)\} - 2\text{Cov}\{c(t_1), c(t_2)\} = Q(t_2 - t_0) + Q(t_1 - t_0) - 2Q(t_1 - t_0) = Q(t_2 - t_1),$$

as was to be expected\(^1\). Obviously also, the expectancy of $\delta c$ vanishes:

$$\mathbb{E}\{\delta c\} = 0.$$

Now, suppose we have a time series of values

$$c(t_i), \ i = 1, \ldots, n,$$

with constant

$$\delta t = t_{i+1} - t_i.$$

Then one can show that the expression

$$AV_{\delta t}(c) = \frac{1}{n-1} \sum_{i=1}^{n-1} (c(t_{i+1}) - c(t_i))^2$$

has the expected value of, and is thus an unbiased estimator of, the variance $Q\delta t$. This empirically computable quantity is called the Allan variance, after David W. Allan\(^2\) (Allan’s Time, The Allan Variance). For true random walk behaviour, $Q\delta t$, and thus $AV_{\delta t}(c)$, should be strictly proportional to $\delta t$, and $Q$ follows as the proportionality constant.

---

\(^1\)Why?

\(^2\)Undoubtedly students of spatial information analysis will recognise this as very similar to the semivariogram used in connection with the Kriging technique.
9.3 The carrier phase observable

9.3.1 Observation equations

When we can write code pseudoranges as follows:

\[ p = \rho + c (\Delta T - \Delta t) + d_{\text{ion}} + d_{\text{trop}}, \]  

we can write carrier phase observables as follows:

\[ P = \lambda \left( \frac{\phi}{2\pi} \right) = \lambda \left( \frac{\phi}{2\pi} - N \right) = \rho + c (\Delta T - \Delta t) + D_{\text{ion}} + D_{\text{trop}} - \lambda N, \]  

in which

- \( \phi \) is the measured phase-difference angle in radians, \( \phi \in [0, 2\pi) \).
- \( \phi \) phase-difference angle including the number of full wavelengths \( N \) needed to make equation 9.4 valid. The number \( N \in \mathbb{N} \) is called the (integer) ambiguity number, and we have \( \phi = \phi + 2\pi N \).
- \( \Delta t \) is the satellite’s,
- \( \Delta T \) the receiver’s clock error,
- \( \rho = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2} \) the geometric distance between satellite, location \([x \ y \ z]^T\), \([X \ Y \ Z]^T\) and receiver, location, and
- \( D_{\text{ion}}, D_{\text{trop}} \) the propagation delays caused by iono- and troposphere.

5. In fact \( D_{\text{ion}} = -d_{\text{ion}} \) and \( D_{\text{trop}} = d_{\text{trop}} \).

\( \lambda \) is the wavelength (or semi-wavelength).

9.3.2 Linearized observation equations

The above observation equations 9.3, 9.4 must be linearized in order to obtain the update equations of the Kalman filter 3.27, 3.28:

\[ x^+ = x^- - K (Hx^- - \ell), \]
\[ \Sigma^+ = (I + KH) \Sigma^-, \]
\[ K = -\Sigma^- H^T (H \Sigma^- H^T + R)^{-1}. \]

We find from 9.3 for the code pseudorange observable \( \ell = p \):

\[ H = \begin{bmatrix} \frac{\partial p}{\partial x} & \frac{\partial p}{\partial y} & \frac{\partial p}{\partial z} & \frac{\partial p}{\partial \Delta T} \end{bmatrix}^T = \begin{bmatrix} \frac{X - x}{\rho} & \frac{Y - y}{\rho} & \frac{Z - z}{\rho} & c \end{bmatrix}^T, \]
the design or observation matrix of the Kalman filter. Note that the unknowns here are the positions of the vehicle, \(X, Y, Z\), and the receiver clock offset \(\Delta T\). All four are functions of time, but are assumed here determined at a certain point in time.

In section 9.2 we discuss the state vector and its propagation in time in the Kalman filter.

The momentaneous satellite position \(x, y, z\) and clock offset \(\Delta t\) on the other hand are assumed known, i.e., computable from the ephemeris. This assumption is not strictly valid and may cause errors. We already discussed how to mitigate it by using a differential technique with base stations, above in section 8.4. A conceptual approach to this is to consider the base-station observations as observations of the satellite unknowns \(x, y, z, \Delta t\), while the base-station co-ordinates \(X, Y, Z\) are known.

For the carrier phase observable 9.4 \(\ell = P\) we have to also include the integer ambiguity \(N\) in the linearized observation equation, leading to

\[
H = \begin{bmatrix}
\frac{\partial P}{\partial X} & \frac{\partial P}{\partial Y} & \frac{\partial P}{\partial Z} & \frac{\partial P}{\partial \Delta T} & \frac{\partial P}{\partial N}
\end{bmatrix} = \begin{bmatrix}
X - x & Y - y & Z - z & \frac{c}{\rho} & \lambda
\end{bmatrix}.
\]

Including atmospheric parameters to be estimated similarly will extend the state vector (vector of unknowns) \(x\), the propagation in time we already touched upon in section 9.2. In all cases, one linearized observation equation must be formed for every satellite and for every frequency (\(L_1\) or \(L_2\)) of observation. Of course in the Kalman filter, it is allowed to process these observations individually and sequentially, even if they happen simultaneously.

### 9.3.3 Carrier-phase measurement

Measurement of the carrier phase is unusual in consumer grade devices. More common is code measurement. Code measurement is relatively easy to implement by using a correlator, which correlates the signal in the receiver with self-generated signals.

In a correlator, three synthetic signals — “replicas” are generated for every satellite, i.e., every pseudo-random code: E (early), L (Late) and P (prompt). See figure 9.2 Each signal is multiplied, or ”mixed”, with the incoming satellite signal. The result is the correlation of the two signals: it is positive only, if the signals are synchronous.

If the E mixer gives a high value, then the clock of the code generator is fast and it must be slowed down. If, on the other hand, the L mixer
The carrier phase observable gives the higher value, then the clock of the code generator must be speeded up.

Observing the carrier phase is more difficult: for that, generally a so-called **PLL** (phase-locked loop) is used. The most commonly used type is a so-called Costas 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.

The Costas discriminator assumes that the code has already been removed from the carrier. Additionally it is unable to distinguish between the carrier and minus the carrier, which means that its unit of ambiguity resolution is half a wavelength, $\lambda/2$.

A conceptual diagram of a Costas loop is given in figure 9.3. It is seen how the incoming signal is “mixed,” i.e., multiplied, with generated reference carrier replicas $90^\circ$ or $\pi/2$ apart in phase. The outputs of these are smoothed, and then interpreted by a phase extractor to be the sine and cosine components of the “beat signal” having the phase offset $\phi$ being sought. The reference carrier generator then has its frequency “steered” to make this phase offset vanish.
162

RTK measurements

\[
a(t) \left( \sin^2(\omega t) \cos \phi + \sin(\omega t) \cos(\omega t) \sin \phi \right)
\]

**Figure 9.3.** Phase tracking by a Costas discriminator or loop

### 9.4 The RTK method analyzed

First, we measure the phase of the carrier wave *with both receivers on the known point, see figure 9.1:*

\[
\frac{1}{2\pi} \phi_A^S = \frac{f}{c} \rho_A^S + f \tau_A^S - N_A^S, \quad \frac{1}{2\pi} \phi_B^S = \frac{f}{c} \rho_B^S + f \tau_B^S - N_B^S,
\]

in which

\[
\tau_A^S = \Delta T_A - \Delta t_{(1)}^S, \quad \tau_B^S = \Delta T_B - \Delta t_{(1)}^S
\]

are the differences between the receiver’s (A reference receiver, B moving receiver) clock offset and the satellite clock offset \(\Delta t_{(1)}^S\) for the same point in time. The index \(1\) refers to the initial situation with both receivers on the known point.

The quantities \(N_A^S, N_B^S \in \mathbb{N}\) are unknown integer values, *ambiguities,* chosen so that the \(\phi\) values are always in the interval \([0, 2\pi)\). \(f = \omega/2\pi\) is the (circular) frequency, \(c\) the speed of light.

After that, the moving receiver is moved to the unknown point \(R_3\) and we obtain

\[
\frac{1}{2\pi} \phi_C^S = \frac{f}{c} \rho_C^S + f \tau_C^S - N_C^S,
\]
The RTK method analyzed

in which (now (2) refers to the new situation, on the unknown point):

\[ \tau^S_C = \Delta T_C - \Delta t^S_{(2)}. \]

The following assumptions:

1. There hasn’t happened a “cycle slip”, so \( N^S_S = N^S_B \).
2. The time elapsed is so short that both \( \Delta t^S_{(1)} = \Delta t^S_{(2)} \) and \( \Delta T_C = \Delta T_B = \Delta T_A + \Delta T_{AB} \), in which \( \Delta T_{AB} \) is a constant difference (clock error difference of the clocks of the two receivers); so \( \tau^S_B = \tau^S_A + \Delta T_{AB} \) and \( \tau^S_C = \tau^S_B = \tau^S_A + \Delta T_{AB} \).
3. The reference and moving receivers are in the same place on the known point\(^3\), so that \( \rho^s_A = \rho^s_B \).

Then

\[ \frac{1}{2\pi} \phi^S_{AB} \overset{\text{def}}{=} \frac{1}{2\pi} (\phi^S_B - \phi^S_A) = f \Delta T_{AB} - (N^S_B - N^S_A) \quad (9.5) \]

and

\[ \frac{1}{2\pi} \phi^S_{AC} \overset{\text{def}}{=} \frac{1}{2\pi} (\phi^S_C - \phi^S_A) = \frac{f}{c} (\rho^S_C - \rho^S_A) + f \Delta T_{AB} - (N^S_B - N^S_A). \quad (9.6) \]

In equation 9.5 the left-hand side is measured. We get immediately \( f \Delta T_{AB} - (N^S_B - N^S_A) \) to be substituted into equation 9.6, and as the observation equation we get

\[ \frac{1}{2\pi} \phi^S_{AC} + (N^S_B - N^S_A) - f \Delta T_{AB} = \frac{f}{c} (\rho^S_C - \rho^S_A), \]

where the left hand side is an “observed” quantity, and on the right hand side, \( \rho^S_C \) is a function of the unknown point’s coordinates (i.e., the unknowns in this adjustment problem). The linearization gives an observation equation, to be used either by a least squares adjustment routine or by a Kalman-filter.

Note that the quantity

\[ \gamma^S_{AB} \overset{\text{def}}{=} f \Delta T_{AB} - (N^S_B - N^S_A) \]

is a real number, even though \( N^S_B - N^S_A \) is an integer number. If there are more than one satellite to be used simultaneously, say, satellite \( S_k, k = 1, \ldots, n \), several quantities can be calculated on the known point

\[ \gamma^k_{AB} = f \Delta T_{AB} - (N^k_B - N^k_A), \quad k = 1, \ldots, n \]

\(^3\)more generally, their difference in location is precisely known
where however there is one and the same $\Delta T_{AB}$. Let us start by choosing the integer for satellite 1, $N^1_B - N^1_A$, such that, in equation 9.5, $\Delta T_{AB}$ is minimized. After that, we can calculate, also using equation 9.5 but for satellite $S_k$,

$$N^k_B - N^k_A = f\Delta T_{AB} - \frac{1}{2\pi}\phi^k_{AB}, \quad k = 2, 3, \ldots$$

and they too have to be integers. If not, we have an adjustment condition that can be used to improve the value $\Delta T_{AB}$, for example, we can minimize the sum of squared differences of $(N^k_B - N^k_A)$ (k = 1, 2, ... ) from among the nearest integers. After this, the values $(N^k_B - N^k_A)$ can be rounded to the nearest integers.

As the final result of this whole operation, we get more accurate observation quantities, so also more accurate estimators of the unknowns. But unfortunately it works only if the distance is relatively short, 10–20 km at the most. Otherwise, the values $N^k_B - N^k_A$ are affected by the uncertainties of the atmosphere and satellite orbits, and will not be close enough to integers.

### 9.5 Clock and atmospheric errors

In the most general case the quantities $\nu^k_{AB}$ include not only the clock errors but also delays caused by the ionosphere and neutral atmosphere ("troposphere"). In that case we can write

$$\nu^k_{AB} = f (\Delta T_B - \Delta T_A) - (N^k_B - N^k_A) + \frac{d^{\text{ion}}_{AB}}{\lambda} + \frac{d^{\text{trop}}_{AB}}{\lambda}.$$ 

In real-time application, both the receiver clock errors $\Delta T_A, \Delta T_B$ and the delays of the ionosphere and troposphere are modelled with suitable parameters as Gauss–Markov or random walk processes, suitably parametrized. Then, all the parameters, also the co-ordinates of the moving receiver, are estimated in real time with the help of the Kalman filter, and they are immediately available for use.

### 9.6 Using double differences

In the geometry above it is tempting to use double differences, in other words, observation quantities obtained by taking the difference between two satellites. Then at the base station we get
Fast ambiguity resolution

\[
\frac{1}{2\pi} \phi_{AB}^{ST} \overset{\text{def}}{=} \frac{1}{2\pi} (\phi_B^T - \phi_A^T) - \frac{1}{2\pi} (\phi_B^S - \phi_A^S) =
- \left( (N_B^T - N_A^T) - (N_B^S - N_A^S) \right) + f\tau_{AB}^{ST},
\]

in which

\[
\tau_{AB}^{ST} = \tau_{AB}^T - \tau_{AB}^S =
= \left( (\Delta T_B - \Delta t^T) - (\Delta T_A - \Delta t^T) \right) -
- \left( (\Delta T_B - \Delta t^S) - (\Delta T_A - \Delta t^S) \right) = 0,
\]

and similarly

\[
\frac{1}{2\pi} \phi_{AC}^{ST} \overset{\text{def}}{=} \frac{1}{2\pi} (\phi_C^T - \phi_A^T) - \frac{1}{2\pi} (\phi_C^S - \phi_A^S) =
= \frac{f}{c} \left( (\rho_C^T - \rho_A^T) - (\rho_C^S - \rho_A^S) \right) -
- \left( (N_B^T - N_A^T) - (N_B^S - N_A^S) \right) + f\tau_{AB}^{ST}, \quad (9.7)
\]

in which again \( \tau_{AB}^{ST} = 0 \).

In this case the “\( \nu \) quantity”, that is solved by putting the reference receiver and the moving receiver side by side, is

\[
\nu_{AB}^{ST} = - \left( (N_B^T - N_A^T) - (N_B^S - N_A^S) \right)
\]

for two satellites \( S \) and \( T \). This is an integer. We “observe” the double-difference quantities \( \phi_{AB}^{km} \) to all satellite pairs \( S_k, S_m \), \( k = 1, \ldots n \), \( m = k + 1, \ldots, n \) (\( n \) number of satellites), and we round the corresponding values \( \nu \) to the nearest integer. The values found after that can be used to compute the quantities \( (\rho_C^T - \rho_A^k) - (\rho_C^m - \rho_A^m) \) from the observations \( \phi_{AC}^{km} \).

\[9.7\] Fast ambiguity resolution

The measurement method described above before requires, that before field measurement (i.e., the movement of the moving receiver in the field and its occupation of the points to be measured) and in order to check also after measurement, the moving receiver can be placed next to the reference receiver (so called co-location).

Often this is somewhat difficult: the reference receiver may be outside the measurement area and be run by the a “service provider.”
is why fast ambiguity resolution was invented. It works best if the distance between the reference and moving receivers is so small that the
differential atmosphere and orbit errors between them can be ignored.
In this case the equation 9.7 is

\[
\frac{1}{2\pi} \phi_{AC}^{ST} \equiv \frac{1}{2\pi} (\phi_C^T - \phi_A^T) - \frac{1}{2\pi} (\phi_C^S - \phi_A^S) = \frac{1}{c} \left( (\rho_C^T - \rho_A^T) - (\rho_C^S - \rho_A^S) \right) - \left( (N_B^T - N_A^T) - (N_B^S - N_A^S) \right).
\]

Here the double-difference quantities, the distances

\[\rho_{AC}^{ST} \equiv (\rho_C^T - \rho_A^T) - (\rho_C^S - \rho_A^S)\]

are purely geometric. If we write, for the satellites \(S_k, k = 1, \ldots, n\),

\[\rho_k^C = \sqrt{(x^k - x_C)^2 + (y^k - y_C)^2 + (z^k - z_C)^2},\]

we can see, that the only unknowns here are the position of the moving receiver

\[
\begin{bmatrix}
X_C \\
Y_C \\
Z_C
\end{bmatrix}^T.
\]

The position of the moving receiver is always known with an accuracy of a couple of metres with the help of GPS code measurement, which
has no ambiguity problem. Then it’s sufficient, if we find from all the possible positions of the receivers (Searching space, belonging to the
set \(\mathbb{R}^3\)) only the places for which all the values, for satellite pairs \(S_k, S_m\),

\[N_{AB}^{km} \equiv (N_B^m - N_A^m) - (N_B^k - N_A^k)\]

are integers.

See figure 9.4. Conversely, if there are \(n\) satellites, there are \(n - 1\)
different ambiguity values \(N_{AB}^{km}\). The ambiguity combinations are thus
the elements of a \(n - 1\) dimensional space. In case each ambiguity has,
say, 10 different possible values that are compatible with the approximate
position obtained from the code measurement, this already gives \(10^{n-1}\)
different ambiguity combinations. If there are 8 satellites, this number is 10 million. Too many possibilities to search in real time in a device
that has limited calculating capacity.

However we can remark that of all the ambiguity alternatives only a
very small fraction is consistent with a particular position of the moving
Fast ambiguity resolution

Location based on code measurements

Figure 9.4. Ambiguity resolution

receiver: the consistent ambiguity combinations belong to the *a three-dimensional subspace* of ambiguity space, one parametrization of which is given by the co-ordinates $[X_C \ Y_C \ Z_C]^T$, as already remarked earlier.

In recent years there have been developed smart and efficient methods to resolve ambiguities in this consistent subspace, like the LAMBDA method (Least-squares Ambiguity Decorrelation Adjustment, Teunissen et al. 1997).

The ambiguity resolution method introduced will succeed only if the distance between the reference and moving receivers is short enough, in general under 10–20 km. In that case we can take advantage of the fact that the GPS satellites send their signal in two different frequencies, L1 and L2. The ambiguity resolution is obtained immediately or after only a couple of epochs.

Ambiguity resolution is also possible for longer vectors, but much more difficult, laborious and time consuming, because the errors caused by the atmosphere etc. have to be taken into account.

In the Kalman filter, the ambiguity is introduced as an unknown $N$ to the state vector, but make it a real-valued state initially. As the
Resolution of multiple ambiguities works better than doing it one-by-one. The one-by-one method fails to resolve $N_1$ to 3, while the combined method resolves $(N_1, N_2)$ to $(3, 3)$

Multiple $N_i$ variance ellipsoids are often very elongated “cigars” as depicted. In the LAMBDA method, the ambiguities are transformed to integer linear combinations that change the error ellipse to (almost) a circle. In this picture, the correct solution is easily seen to be the one nearest to the point $(N_1, N_2)$

Figure 9.5. Ambiguity resolution.

filter progresses in time, the state variance attached to $N$ will become smaller and smaller, until it become possible to identify the real-valued ambiguity with confidence with a single integer value.

*Note*, however, that in a practical situation you will not have just one equation 9.4, but as many as there are useable GPS satellites in the sky, i.e., $4 - 12$. This means that we will have not one, but several $N_i$, $i = 1, \ldots, n$, with $n$ the number of satellites. This set of ambiguities will have a *variance-covariance matrix* of size $n \times n$. Now one should analyse if the whole set of $N_i$ lies close enough to a set of integer values, one point of a grid of points in the abstract vector space $\mathbb{R}^n$. “Close enough” should be understood in terms of this variance-covariance
matrix. Generally, this resolution of all ambiguities together will succeed well before any single one will be resolved successfully. Sophisticated algorithms have been developed for this – e.g., the LAMBDA technique (TU Delft, LAMBDA).

9.8 A geometric analysis of RTK measurement

The RTK 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 we use the notation \{i, j, k\} for a triad of orthogonal unit vectors. Every vector in space can be written as a linear combination of these base vectors. In a local horizon frame, the vector i points north, j east and k up to the zenith. This is a left-handed reference frame.

First we look at how the orbit and clock errors of GPS satellites propagate into the position solution in the case of one base station, next we address the case of three base stations.

9.8.1 One base station

In the case of one base station we may write the observable as follows, ignoring for a moment the atmosphere and other disturbances:

\[ P = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2 + c(\Delta T - \Delta t) + \ldots} \]

Here \( x = x_i + y_j + z_k \) is the position vector of the satellite, \( X = X_i + Y_j + Z_k \) that of the receiver, while \( \{i, j, k\} \) is an orthogonal triad of unit vectors. The expression

\[ \rho = \sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2} \]

is the geometric distance between satellite and receiver.

We choose here an alternative three-dimensional unit vector triad, in which \( i' \) points from the satellite to the base station, and \( j' \) and \( k' \) are orthogonal with respect to each other and \( i' \). These base vectors are constructed as follows:

\[ i' = -\frac{x - X}{\|x - X\|}, \quad j' = \frac{\langle i' \times X \rangle}{\|\langle i' \times X \rangle\|}, \quad k' = \langle i' \times j' \rangle. \quad (9.8) \]
Let the orbital errors (or their effect on the satellite’s position in space) be \( \Delta x = \Delta x' + \Delta y' + \Delta z' \), and the satellite’s clock error \( \Delta t \) (which we already assume to be small). Their effect on the pseudorange is

\[
\Delta P = \frac{\partial P}{\partial x} \Delta x + \frac{\partial P}{\partial y} \Delta y + \frac{\partial P}{\partial z} \Delta z + c \Delta t. \tag{9.9}
\]

Let 

\[
\rho_0 = \|x - X_0\| = \sqrt{(x - X_0)^2 + (y - Y_0)^2 + (z - Z_0)^2}
\]

be the distance between satellite \( x = x' + y' + z' \) and base station \( X_0 = X_0i' + Y_0j' + Z_0k' \). Let furthermore the location of the moving receiver ("rover") be \( X = Xi' + Yj' + Zk' = (X_0 + \xi)i' + (Y_0 + \eta)j' + (Z_0 + \chi)k' \). Here \( \xi, \eta \) and \( \chi \) are now the co-ordinates of the rover relative to the base station in the co-ordinate system agreed on above. The distance between base station and rover is

\[
s = \|s\| = \sqrt{\xi^2 + \eta^2 + \chi^2}, \quad s = \xi i' + \eta j' + \chi k'.
\]

Let us write out equation 9.9. We obtain

\[
\Delta P = \frac{x - X}{\rho} \Delta x + \frac{y - Y}{\rho} \Delta y + \frac{z - Z}{\rho} \Delta z + c \Delta t,
\]
A geometric analysis of RTK measurement

in other words, we get separately for base station and rover

\[ \Delta P_0 = \frac{x - X_0}{\rho_0} \Delta x + \frac{y - Y_0}{\rho_0} \Delta y + \frac{z - Z_0}{\rho_0} \Delta z + c \Delta t \]

\[ \Delta P = \frac{x - X}{\rho} \Delta x + \frac{y - Y}{\rho} \Delta y + \frac{z - Z}{\rho} \Delta z + c \Delta t. \]

The difference, i.e., the error in the rover’s position due to the distance from the base station, is

\[ \Delta P - \Delta P_0 = \left( \frac{x - X}{\rho} - \frac{x - X_0}{\rho_0} \right) \Delta x + \left( \frac{y - Y}{\rho} - \frac{y - Y_0}{\rho_0} \right) \Delta y + \left( \frac{z - Z}{\rho} - \frac{z - Z_0}{\rho_0} \right) \Delta z \]

From here, the satellite’s clock error has vanished.

Let us look closer at the coefficient

\[ \left( \frac{x - X}{\rho} - \frac{x - X_0}{\rho_0} \right). \]

If we define the function

\[ f(X) \overset{\text{def}}{=} \frac{x - X}{\rho(x, X, Y, Y, Z)}, \]

this coefficient is

\[ f(X) - f(X_0) \approx \frac{\partial f}{\partial X} \bigg|_{X=X_0} (X - X_0) + \frac{1}{2} \frac{\partial^2 f}{\partial X^2} \bigg|_{X=X_0} (X - X_0)^2 + \ldots = \]

\[ = \frac{\partial f}{\partial X} \bigg|_{X=X_0} \xi + \frac{1}{2} \frac{\partial^2 f}{\partial X^2} \bigg|_{X=X_0} \xi^2 + \ldots, \]

a Taylor expansion. If we retain from this only the first term, we obtain

\[ \frac{\partial f}{\partial X} \bigg|_{X=X_0} = -\frac{1}{\rho_0} + \frac{(x - X_0)^2}{\rho_0^3}, \]

and

\[ \Delta P - \Delta P_0 = \left( \frac{(x - X_0)^2}{\rho_0^3} - \frac{1}{\rho_0} \right) \xi \Delta x + \left( \frac{(y - Y_0)^2}{\rho_0^3} - \frac{1}{\rho_0} \right) \eta \Delta y + \left( \frac{(z - Z_0)^2}{\rho_0^3} - \frac{1}{\rho_0} \right) \chi \Delta z. \]

Because the co-ordinate axes are defined in the way described above, we have

\[ \bar{x} - \bar{X}_0 = \begin{bmatrix} x - X_0 \\ y - Y_0 \\ z - Z_0 \end{bmatrix} = \begin{bmatrix} -\rho_0 \\ 0 \\ 0 \end{bmatrix}, \]
and we obtain
\[ \Delta P - \Delta P_0 = -\frac{1}{\rho_0} (\eta \Delta y + \chi \Delta z). \] (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 direction vector to the satellite has an effect.

9.8.2 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 \( \eta \) and \( \chi \), which we may interpret as the plane co-ordinates in two mutually orthogonal directions. \( \Delta y \) is the position error of the satellite in the “left-right” direction, \( \Delta z \) “upwards” on the celestial dome, towards zenith, also perpendicular to the line of sight to the satellite.

Because equation (9.10) is bilinear in the co-ordinates \( (\eta, \chi) \), we may interpolate the correction linearly, when it has been determined at the three base stations. If we assume the base stations \( A \), \( B \) and \( C \), and the measured corrections \( \Delta P_A \), \( \Delta P_B \) ja \( \Delta P_C \), we can compute the correction for an arbitrary point as follows:
\[ \Delta P = p_A \Delta P_A + p_B \Delta P_B + p_C \Delta P_C, \] (9.11)
in which \( p_A \), \( p_B \) ja \( p_C \) are the computation point’s barycentric co-ordinates within the triangle:

\[ p_A = \begin{vmatrix} \eta_B & \eta_C & \eta \\ \chi_B & \chi_C & \chi \\ 1 & 1 & 1 \end{vmatrix}, \quad p_B = \begin{vmatrix} \eta_C & \eta_A & \eta \\ \chi_C & \chi_A & \chi \\ 1 & 1 & 1 \end{vmatrix}, \quad p_C = \begin{vmatrix} \eta_A & \eta_B & \eta \\ \chi_A & \chi_B & \chi \\ 1 & 1 & 1 \end{vmatrix}. \] (9.12)

Here we have the co-ordinates \( (\eta, \chi) \). For barycentric co-ordinates it holds that \( p_A + p_B + p_C = 1 \), and they are all three linear in both the \( \eta \) and the \( \chi \) co-ordinate. By simple substitution one may ascertain, that in the corner points, e.g., point \( A \), \( p_A = 1 \) and \( p_B = p_C = 0 \) — the reproducing property.
In reality, however, the pseudorange correction is not precisely linear: one should use the quadratic equation

\[
f(X) - f(X_0) = \frac{\partial f}{\partial X} \bigg|_{X=X_0} \xi + \frac{1}{2} \frac{\partial^2 f}{\partial X^2} \bigg|_{X=X_0} \xi^2 + \ldots
\]

We already derived

\[
\frac{\partial f}{\partial X} \bigg|_{X=X_0} = -\frac{1}{\rho_0} + \frac{(x-X_0)^2}{\rho_0^3}.
\]

Furthermore we may derive, that

\[
\frac{\partial^2 f}{\partial X^2} \bigg|_{X=X_0} = -\frac{x-X_0}{\rho_0^3} - 2\frac{x-X_0}{\rho_0^3} + 3\frac{(x-X_0)^3}{\rho_0^5} = 3 \left( \frac{(x-X_0)^3}{\rho_0^5} - \frac{x-X_0}{\rho_0^3} \right).
\]

Again

\[
\mathbf{x} - \mathbf{x}_0 = \begin{bmatrix} x - X_0 \\ y - Y_0 \\ z - Z_0 \end{bmatrix} = \begin{bmatrix} -\rho_0 \\ 0 \\ 0 \end{bmatrix},
\]

and

\[
\frac{\partial^2 f}{\partial X^2} \bigg|_{X=X_0} = 0, \quad \frac{\partial^2 f}{\partial Y^2} \bigg|_{Y=Y_0} = 0, \quad \frac{\partial^2 f}{\partial Z^2} \bigg|_{Z=Z_0} = 0.
\]

This is a surprising result, but not entirely surprising. Equation 9.10 applies in three-dimensional space, i.e., when the location difference vector between base station and rover is \(\mathbf{s} = \mathbf{x} = \xi \mathbf{i} + \eta \mathbf{j} + \chi \mathbf{k}\). In the equation we only see \(\eta\) and \(\chi\), but don’t blame the equation for that.

However, 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 when used in the map plane. However, plane equations are good when the rover’s position is projected onto the plane through the three base stations as shown in figure 9.8.

### 9.8.3 Summary

The procedure thus is the following:

1. For every satellite separately, compute the special co-ordinates \((\eta, \chi)\) of base stations A, B, C and rover, in a plane that goes through the
rover, which is perpendicular to the direction vector to the satellite. This requires the use of almanac data. As follows: project the co-ordinate difference vector with the rover onto the plane spanned by the vectors \( j' \) and \( k' \). In other words, compute

\[
\eta_A = \langle (X_A - X) \cdot j'(X) \rangle, \quad \chi_A = \langle (X_A - X) \cdot k'(X) \rangle,
\]

in which \( j'(X) \) and \( k'(X) \) are computed using equations 9.8, and do the same computation for base stations B and C.

2. Compute from those the barycentric weights, equation 9.12, by substituting \((\eta, \chi)\).

3. Do the interpolation of the \( \Delta P \) values.

See figure 9.8.

### 9.8.4 Modelling the atmosphere

The effect of the satellite orbit and clock errors is by its nature functional, i.e., it may in principle be calculated precisely for the roving receiver’s location, if we have been given the observations, or pseudorange corrections, in three base stations, and the base stations are in a pretty triangle around the measurement location. The effect of the atmosphere again is not functional but stochastic.

This is why precise prediction to the rover location will not work; the uncertainty grows with growing distance from the base station.
A geometric analysis of RTK measurement

Figure 9.8. The geometry of differential GPS. 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 linearly in the distance between stations.

We also depict the way the differential correction is computed: the location \(P\) of the rover must be projected on the plane going through the base stations ABC (projection point \(P'\)) and after that, the correction must be bilinearly interpolated.

How it grows, depends on the statistical properties of the atmosphere. The GPS signal delay caused by the atmosphere – both the ionosphere and the troposphere – (in case the satellite is in the zenith) forms a stochastic process, let’s say \(d(\varphi, \lambda)\), on the domain of geographical locations \((\varphi, \lambda)\). One can define a signal covariance function for it. E.g., the covariance between two locations can be described by a Gauss–Markov type covariance formula:

\[
\text{Cov}\{d(\varphi_1, \lambda_1), d(\varphi_2, \lambda_2)\} = C_0 e^{-\psi/\psi_0},
\]

in which \(\psi\) is the angular distance between the points \((\varphi_1, \lambda_1)\) and \((\varphi_2, \lambda_2)\) on the Earth surface. The quantity \(C_0\) is called the signal variance, the quantity \(\psi_0\) the correlation length. These quantities may be chosen suitably, i.e., realistically for the tropo- and ionosphere of a certain time and place.

Another approach is to model the atmosphere functionally, but with unknown parameters. E.g., the global ionosphere may be well described by a spherical harmonic expansion\(^4\). Also a polynomial or

\(^4\)In fact the GPS navigation message contains such a model with twelve parameters.
Fourier expansion may be considered. The unknown coefficients are estimated from the observations of the base stations.

The state of the art of this moment is, that the RTK positioning using a network of base stations is as precise as RTK positioning using one base station in the immediate vicinity, on condition that the distances between the base stations are at most approx. 80 km. The quality of positioning is also preserved for some distance outside the coverage area of the network of base stations. The atmosphere is the limiting factor.

9.9 Network RTK

To implement a network RTK solution, several base stations are used, and in some way the corrections are interpolated to the location of the user.

Two basic methods:

1. Broadcast method: corrections are sent to many users at the same time. Can use for example a radio transmission’s FM sideband (RDS, Radio Data System).

2. One-to-one (“singlecast”) method: the corrections are computed for one user and sent to them, for example by mobile phone or Internet. The content of the correction message can be different for each user.

One of the variants of the one-to-one method is the virtual base station method, where the calculation is done by interpolating base station corrections into a “virtual base station” in the vicinity of the observer.

Various interpolation techniques:

1. Brute force: here is assumed that the correction is continuous as a function of position on Earth. If assumed that this function is linear, three base stations around the measurement area are adequate.

2. Modelling of the atmosphere. In principle this could improve the interpolation results, if the model is good.
New Global Navigation Satellite Systems (GNSS) are coming, or have already come, on-line in addition to the well established Global Positioning System. As these will be useable and relevant to technological navigation, we discuss them here. See table 10.1.

10.1 GPS modernization

In charge of the modernization effort is the Global Positioning Systems Directorate.

10.1.1 Satellite types

The different types of GPS satellites are called blocks.

Block I consisted of test satellites used for testing the GPS concept.

Block II was the first operational GPS satellite type. None are any longer in operation.

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Orbital planes; separation angle</td>
</tr>
<tr>
<td>Satellites / plane; separation angle</td>
</tr>
<tr>
<td>Satellites total (official)</td>
</tr>
<tr>
<td>Orbit height (km)</td>
</tr>
<tr>
<td>Period</td>
</tr>
<tr>
<td>“Resonance” (orbits / sidereal day)</td>
</tr>
<tr>
<td>Inclination</td>
</tr>
</tbody>
</table>

† There are 3 – 5 more satellites in inclined geosynchronous orbits (IGSO) as well as five in true geostationary orbits.
Block IIA: A = “Advanced.” 19 satellites were launched in the years 1990–1997. The last one is (2018) dying.

Block IIR: the first one was launched in 1997.

Block IIR-M: “M” stands for “Military”. Eight satellites were launched in the years 2005–2009.

Block IIF: “F” stands for “Follow-on”. Twelve satellites were launched in the years 2010–2016.

Block IIIA: the first launch is planned for December 2018.

Block IIIF: planned for 2025–2034.

10.1.2 New codes and frequencies

The first Block IIR-M satellite was launched September 25, 2005. It transmits the new civil code L2C, modulated on the L2 carrier. It contains new pseudo-random codes, so-called CM (Civilian Moderate) and CL (Civilian Long) codes, the lengths of which are 10 230 and 767 250 bits. Together, the modulation frequency (“chip rate”) is 1 023 000 bits per second. L2C contains also an improved navigation message, to which belongs, among other things, the time scale offset of the GPS system (important in combined use with GLONASS and Galileo!) as well as an Alert-warning, which announces within six seconds, if one cannot rely on the satellite’s data (integrity).

The new frequency L5, 1176.45 MHz, is meant for Safety-of-Life type use especially in aviation, see section 11.1. This frequency is internationally reserved for aviation. Block IIF satellites carry this frequency as the first.

Also SBAS systems use the L5 frequency.

10.2 The Russian GLONASS system

See GLONASS Information and Analysis Center, which also shows the constellation status.

In December 2018 there were 26 satellites in orbit, of which 23 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.
10.2.1 Frequencies and signals

The GLONASS system is described by the Interface Control Document ICD from 2008, 65 pages in English, Anon. (2008).

The GLONASS system differs from GPS and Galileo in that every satellite has its own carrier frequency, a solution known as FDMA, “frequency division multiple access.” However, like GPS, it uses two frequency bands \( L_1 \), approximately 1.6 GHz, and \( L_2 \), approximately 1.2 GHz. Nowadays however the “antipodes” — satellites in the same orbital plane on opposite sides — can use the same frequencies. The frequencies are

\[
\begin{align*}
    f_{K1} &= f_{01} + K\Delta f_1, \\
    f_{K2} &= f_{02} + K\Delta f_2,
\end{align*}
\]

in which \( K \) is the satellite’s channel number — which can be found from the almanac — and

\[
\begin{align*}
    f_{01} &= 1602 \text{ MHz}, \quad \Delta f_1 = 562.5 \text{ kHz}, \\
    f_{02} &= 1246 \text{ MHz}, \quad \Delta f_2 = 437.5 \text{ kHz}.
\end{align*}
\]

The civil code — corresponding to C/A — has a bit frequency of 0.511 MHz, the encrypted military code 5.11 MHz. The modulation technique is the same as for GPS, phase modulation with a phase shift of \( \pi = 180^\circ \) — “Binary Phase-Shift keying.” The navigation message is only on \( L_1 \); its bit frequency is 50 Hz.

The satellites’ radiation is clockwise (“right-hand”) polarized, like GPS.

10.2.2 Time, reference system and constellation

The GLONASS time scale is the same as UTC (SU), i.e., the realization of UTC for the Russian Federation. This means that the leap seconds of UTC, which happen either at the end of December or of June, enter into the GLONASS time scale. This is different from the GPS practice: GPS time does not have leaps.

The differences between UTC and UTC(SU) as well as an announcement of upcoming leap seconds are contained in the navigation message.

The reference system used by GLONASS is PZ90 (“Parameters of the Earth 1990”), which is geocentric but slightly different from GRS90.
The ephemeris give the location of the satellite in space in this system, however in rectangular co-ordinates.

The nominal number of GLONASS satellites (in a full constellation) is 24 in three orbital planes, plus three spares. The inclination angle with the equator is 64°8, a high value serving the area of the old Soviet Union better. Like also for GPS, the satellite geometry of GLONASS repeats after a sidereal day. Differently from GPS, however, then there will be a different satellite in the same place. The orbital height is 19 140 km and the period 11 h 16 m, shorter than a sidereal day. Only after eight days (17 orbits), the same geometry will repeat itself also with the same individual satellites in the same places. This solution reduces the resonant orbital perturbations, which for the GPS system consume a lot of rocket propellant.

10.2.3 The future

There are plans to switch to CDMA, “code division multiple access,” in the near future, with a signal structure similar to that of the other systems. This would make the system more compatible with other GNSS and facilitate the design of multi-system receivers. Also planned is expanding the space segment to thirty satellites in six planes.

10.3 The European Galileo system

The Galileo system is a joint undertaking of the European Commission and ESA. They formed the “Galileo Joint Undertaking” (GJU), which again chooses a “concessionaire,” a private “Galileo Operating Company” responsible for Galileo’s daily operations and especially its commercialization.

10.3.1 Satellites and orbits

There will be thirty Galileo satellites, in three orbital planes, in each of which are nine satellites and one spare. The distance between satellites within the orbital plane is 40°. The inclination of the orbits relative to the equator is 56°.

The height of the Galileo satellites’ orbit is 23 222 km, higher than the GPS orbits.

The geometry is repeated every sidereal day, like the GPS geometry; the difference is though, like also for GLONASS, that after a day, the
The European Galileo system

Yellow: frequencies already in use
Red: new frequency reservations

Figure 10.1. Galileo’s frequencies.

same places are taken by different satellites.

Only after 10 days, or 17 revolutions, will the same satellites be again in the same places in the observer’s sky.

The first, experimental Galileo satellite, GIOVE-A, was launched on December 28, 2005. At the time of writing (2018) there are 18 working satellites in orbit, and more coming.

10.3.2 System description, components

The Galileo system is partially compatible with the GPS system. It is also intended to work seamlessly with SBAS systems, like (in Europe) EGNOS.

Galileo’s signal and frequency structure is complex, see ESA, Galileo Navigation Signals and Frequencies. The carriers are L1, 1575.42 MHz like also for GPS, E5a (1176.45 MHz), E5b (1207.14 MHz), and E6. E5a is also called L5.

10.3.3 Various services

The services offered by Galileo may be divided as follows:

- Open service (OS). Useable by anyone. [L1, E5]
- Safety-of-Life (SoL). [L1, E5b]
- Commercial (CS). [L1, E5, E6]
- Public regulated services (PRS). This includes the police, border guards, defence forces and peacekeeping forces as well. The fact
that Galileo is called, “as opposed to the GPS, a civilian system,” is perhaps not quite accurate . . . [L1, E6]

New with Galileo is the integrity service. It is part of the Safety-of-Life service.

### 10.4 The Chinese BeiDou system

About the name: BeiDou is the constellation Big Dipper (Ursa Major), which is used to find the North Star (Polaris, α UMi). Thus the name is symbolic for navigation.

#### 10.4.1 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 found out that the system functioned independently: the ground station sent a signal through two satellites to the roving receiver, which also answered through two satellites (it was thus an active system). Thus, the receiver location could be computed, at least on the territory of China. By using a terrain model, the precision could be improved. The computed position was sent to the receiver encrypted for military use.

#### 10.4.2 BeiDou 2 and 3

The BeiDou-2 and 3 systems are also called “Compass” in English.

The BeiDou-2 system comprises 14 satellites, of which 5 in geostationary orbits, offering positioning services in the Asia-Pacific region. It was completed in 2012.

Then, in November 2017, China started launching BeiDou-3 satellites at a high rate. On December 27, 2018 it was announced (Xinhua, 2018) that the service was globally available. Reaching a full constellation of 35 satellites, of which 5 geosatationary, is planned for 2020.

BeiDou-2 transmits three public-service signals, B1I, B2I and B3I, on three frequencies: 1561.1, 1207.14 and 1268.520 MHz. BeiDou-3 transmits four public-service signals, B1I, B1C, B2a and B3I, on three partly different frequencies: 1575.41, 1176.45, and 1268.520 MHz. See BeiDou Introduction.
The signal is modulated with pseudo-random codes in a similar way as the signals of GPS or Galileo. The BeiDou public-service signal is unencrypted and freely available. Precision is $\pm 10$ m for BeiDou-2, and the planned precision for BeiDou-3 is $\pm 0.5$ m. The BeiDou-3 signal also includes integrity information, see section 11.1.

The orbits are similar to those of GPS height 21 528 km, inclination 55°. There are however planned to be three satellites in so-called inclined geosynchronous orbits (IGSO) at 35 786 km.

Relevant documents and current information on constellation status and satellite health are found from the above source.
SBAS stands for Satellite Based Augmentation Systems. These systems address the circumstance that, especially in aviation, it is not possible to use the GPS system on its own as it does not provide any availability or correctness guarantees.

Each SBAS system is built for a target area — e.g., for WAAS, North America — where there is operating a network of GPS monitoring stations keeping an eye on the correct functioning of every GPS satellite in real time. Users receive information from the service through transponder on a geostationary satellite.

A side benefit is that the system, in addition to integrity information, also provides differential corrections to the GPS satellites’ pseudorange measurements, valid for the target area.

### 11.1 Integrity and Safety of Life

Integrity is, that the user is warned, in practice within six seconds, of the positioning signal of a certain GPS satellite appears to exceed certain tolerance values. In Safety-of-Life applications this is mandatory. Safety of Life means that, if the system doesn’t function correctly, people may die.

An example is approach and landing on an airport: if we descend in fog using GPS navigation, and the height is wrong by many metres without any warning, an accident will happen. If in fog or cloud the GPS height cannot be relied upon and the pilot is warned, he doesn’t land. If, on the other hand, the precision guaranteed by the integrity system is within several metres and there is no fog but low cloud, the pilot may land on visual from, e.g., 200 feet downward.
Table 11.1. The various approach categories according to ICAO. The category describes how well aircraft and airfield are equipped for instrumental landings. SBAS approach is possible only for part of the categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>“Decision height”</th>
<th>Visibility</th>
<th>Visibility on runway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat I</td>
<td>$\geq 200$ ft $= 60$ m, and</td>
<td>$\geq 800$ m, or</td>
<td>$\geq 550$ m $= 1800$ ft</td>
</tr>
<tr>
<td>Cat II</td>
<td>$100 - 200$ ft $= 30 - 60$ m, and</td>
<td>-</td>
<td>$\geq 350$ m $= 1200$ ft</td>
</tr>
<tr>
<td>Cat IIIA</td>
<td>$&lt; 100$ ft $= 30$ m / none; and</td>
<td>-</td>
<td>$\geq 200$ m $= 700$ ft</td>
</tr>
<tr>
<td>Cat III B</td>
<td>$&lt; 50$ ft $= 15$ m / none; and</td>
<td>-</td>
<td>$50 - 200$ m $= 150 - 700$ ft</td>
</tr>
<tr>
<td>Cat III C</td>
<td>none</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

11.2 WAAS

WAAS (Wide Area Augmentation System) was declared operational in 2003. The system monitors the GPS system and computes differential corrections for the users, as well as providing a certified integrity level. Precision after differential correction is of order $\pm 2$ 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 $\pm 15$ m using a single frequency GPS receiver, mostly due to the ionosphere.

Integrity again enables operations in aviation that would be too risky only using GPS. GPS as such has already been used in RAIM mode (Receiver Autonomous Integrity Monitoring), but only in the straight parts of route flights and not during descent. The requirement is, that at most six seconds after a GPS signal goes unreliable, the user is informed of this. From 2007 on, WAAS has been approved for guiding an aircraft safely to 200 feet height above an airport: ICAO Category I.

WAAS is also cheaper, because traditional radio navigation support equipment in airport areas is massive in size. Also fuel is saved, as GPS/WAAS allows straight flights from field to field instead of through a polygon over radio beacons.

A WAAS capable GNSS receiver receives corrections separately from the orbit and clock errors of the satellites (the latter the same everywhere but different for each satellite, and at least the clock correction rapidly changing in time) and the ionospheric effect (same for all satellites but location dependent, which is why it is computed in the form of a grid). The ionospheric correction is valid only in the area where there are base stations.
WAAS

Table 11.2. The WAAS satellites. These are communication satellites on which the WAAS transponder is only a small part of the payload. PRN is the pseudo-random code used, similar to the one for GPS satellites.

<table>
<thead>
<tr>
<th>Satellite</th>
<th>PRN</th>
<th>Longitude</th>
<th>Launch</th>
<th>Expiration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inmarsat-3 F3</td>
<td>134</td>
<td>178°E</td>
<td>Dec 17, 1996</td>
<td>Sep 2007</td>
</tr>
<tr>
<td>Inmarsat-3 F4</td>
<td>122</td>
<td>142°W</td>
<td>Jun 3, 1997</td>
<td>Sep 2008</td>
</tr>
<tr>
<td>Telesat Anik F1R</td>
<td>138</td>
<td>107°30’W</td>
<td>Nov 21, 2000</td>
<td>Operating</td>
</tr>
<tr>
<td>Intelsat Galaxy 15</td>
<td>135</td>
<td>133°W</td>
<td>Oct 13, 2005</td>
<td>Operating</td>
</tr>
<tr>
<td>Inmarsat-4 F3</td>
<td>133</td>
<td>117°W</td>
<td>Aug 18, 2008</td>
<td>Nov 2017</td>
</tr>
<tr>
<td>Eutelsat 117 West B</td>
<td>131</td>
<td>98°W</td>
<td>June 15, 2016</td>
<td>Operating since 2018</td>
</tr>
</tbody>
</table>

WAAS is also being used outside aviation, like in maritime navigation.

11.2.1 The WAAS space segment

The satellites that have carried WAAS transponders and their history are described in table 11.2.

The signal from a geostationary satellite is coded in the same way using pseudo-random codes like the signals from the GPS satellites. The WAAS satellites have their own PRC codes separate from each other and the codes of the GPS satellites, and the correlator in the receiver distinguishes them with the aid of these.

11.2.2 The WAAS ground segment

The American WAAS system uses the following base stations in the United States and neighbouring countries:

1. **WRS** (Wide-area Reference Stations): 38 in North America, including eight 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). At these, the corrections and integrity information are computed. The differential corrections are computed for the nodes of a grid.

3. Six **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 user on the same frequencies used also by the GPS satellites, i.e., $L_1$ and (in
Satellite-based augmentation systems (SBAS)

Table 11.3. EGNOS satellites. These are general communications satellites and the SBAS transponder is only a small part of the payload. PRN is the pseudo-random code number of the transmission.

<table>
<thead>
<tr>
<th>Satellite</th>
<th>PRN</th>
<th>Longitude</th>
<th>Launch</th>
<th>Expiration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inmarsat-3 F1</td>
<td>131</td>
<td>64°5 E</td>
<td>Apr 3, 1996</td>
<td>?</td>
</tr>
<tr>
<td>Inmarsat-3 F2</td>
<td>120</td>
<td>15°5 W</td>
<td>Sep 6, 1996</td>
<td>Jan 1, 2019</td>
</tr>
<tr>
<td>Artemis</td>
<td>124</td>
<td>21°5 E</td>
<td>Jul 12, 2001</td>
<td>2015</td>
</tr>
<tr>
<td>Inmarsat 4-F2</td>
<td>126</td>
<td>64°0 E</td>
<td>Nov 8, 2005</td>
<td>Operating</td>
</tr>
<tr>
<td>Astra SES-5</td>
<td>136</td>
<td>5°0 E</td>
<td>Jul 9, 2012</td>
<td>Operating</td>
</tr>
<tr>
<td>Astra 5B</td>
<td>123</td>
<td>31°5 E</td>
<td>Mar 2014</td>
<td>Operating</td>
</tr>
</tbody>
</table>

the newer satellites) L₅, using similar identifying pseudo-random codes at the GPS satellites.

These GUS stations have large parabolic antennas for the up-link, and there are two different ones used for each transponder available over the target area.

The data communication links between stations have been custom built. The open Internet is not real time with acceptable reliability.

11.3 Intermezzo: LAAS, or GBAS

LAAS (Local Area Augmentation System) is intended to enable the automatic landing of aircraft on busy airports: ICAO Category III. In the LAAS system, GPS receivers are used in the airport area, from the observations of which corrections are computed locally and transmitted to the aircraft by radio (VHF). Unlike WAAS, it does require special equipment on the airport. However, one such set of equipment would replace many traditional installations, like VOR beacons, for every runway.

LAAS can also be used to guide ground vehicles within the airport area.

LAAS is a so-called ground-based augmentation system (GBAS). Its use is approved for CAT I approaches. Like WAAS, it can replace multiple traditional ground installations for instrument landings at each airport — two for each runway — and allows for flexible approach geometries, saving fuel.
Figure 11.1. WAAS ground segment, status 2017. Red: WRS, green: WMS, blue: GUS. Note the stations in Canada and Mexico.
11.4 EGNOS

EGNOS is a joint project by the European Commission, ESA and Eurocontrol in (the joint European aviation safety organization). It is sometimes called GNSS-1, to distinguish it from Galileo, which is thus GNSS-2.

EGNOS is compatible, or “interoperable,” with WAAS. EGNOS started operations in July 2005.

EGNOS consists of four functional parts: the ground segment, support segment, space segment and user segment.

11.4.1 EGNOS ground segment

The ground segment of EGNOS consists of the following components:

1. RIMS (Ranging and Integrity Monitoring Stations): 45. Receive the signal and send it forward to the Master Control centres.
2. MCC (Master Control Centres): 4. Receive data from the RIMS and compute from it corrections and integrity data, which are sent forward.
3. NLES (Navigation Land Earth Stations) send the data forward to the different geostationary transponders: 6.

Like in the case of WAAS, also the EGNOS ground stations have dedicated data connections between them.

In the framework of the EGNOS Transafrica project, after 2002 were 10 EGNOS-RIMS established in Africa, in addition to Hartebeesthoek.

11.5 MSAS

MSAS (MTSAT Satellite-based Augmentation System) is a Japanese SBAS. It is compatible with WAAS and EGNOS. It has been operational since September 2007.
Not shown here are the far-field ground stations in South Africa (Hartebeesthoek), Canada (Montreal), Singapore, French Guyana (Kourou).

Red: RIMS, yellow: MCC, blue: NLES.

Figure 11.2. EGNOS ground segment, status 2017.
Satellite-based augmentation systems (SBAS)

MSAS has six monitoring stations on the ground. Currently (2018) only the MTSAT-2 satellite is transmitting signals for both PRN 129 and PRN 137. The system only offers horizontal positions, not height. Plans are that the MSAS system, also called MSAS V1, will be replaced by QZSS (also called MSAS V2) in 2020.

11.6 QZSS

Japan has built a system named QZSS, Quasi-Zenith Satellite System. It works as an augmentation system for GPS using the GPS signal structure.

The satellite orbits are 24-hour orbits, the inclination of which is high, approx. 45° – 53°, and very elongated, i.e., the orbital eccentricity is high. The intention is that at any time there would be one satellite hanging over Japan at a reasonably high elevation angle. From this, the label “quasi-zenith” originates. It would require three satellites in orbit. The satellites are also suited, and especially so, as communication satellites, like in its day the Soviet Molnya satellites, that had a similar operational concept.

Differently from other (geostationary) SBAS systems, this system offers useable measurables pseudorange and carrier phase.

The QZSS system’s first satellite was launched on September 11, 2010.
Three more followed in 2017, and on November 1, 2018, the system was taken officially into service.

11.7 Internet-based Global DGPS

GDGPS, the Global Differential GPS system (JPL, 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 via Internet to the user using the TCP-protocol. 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). It allows the user to upload RINEX files, and the results will be presented as a web page.
The art of technological navigation originated in the maritime sphere, and was adopted also into aviation and astronautics as those fields developed. Today, however, we see technological navigation proliferate into the life of millions of ordinary citizens, in the form of car or pedestrian navigation. There is a multitude of sensors available, both intended for navigation, like satellite positioning and augmentation systems, and intended for other uses but adaptable for navigation, like mobile-phone base station networks and indoor wireless local-area (WLAN) networks. This begs the question of how to use these sensors together in an integrated way.

Sensor fusion or sensor integration is the use of multiple sensors toward a reduction of uncertainty, e.g., in location determination during navigation (Wikipedia, Sensor fusion). The sensors may be similar — homogeneous sensor fusion — or different — heterogeneous sensor fusion.

Sensor fusion is often realized 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 are

- 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.
Sensor fusion may also incorporate background knowledge that is not actual sensor output. For example

- A vehicle may stop and inform the filter software that now the velocity is zero — a zero-velocity update.
- A mobile phone used by a pedestrian or car navigator may assume that the upper edge of the screen points in the direction of motion.
- A car is assumed to always have a zero velocity in the sideways direction.
- Maximum values 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 below, we shall discuss some examples of sensor-fusion or sensors-of-opportunity technologies. This is already a broad and expanding field based on rapidly developing technologies, and we will only scratch the surface here.

### 12.1 Case: Sky Map

An interesting example of the creative use of mobile-phone sensors is the application “Sky Map” for Android (Google Play, Sky Map). It shows, after calibration, the stars in their proper locations when you hold up the phone to the sky. It even shows stars and other objects, including planets, that are under the horizon!

Calibration involves rotating the device manually around three different axes. Apparently it exploits the three-axes accelerometer nowadays 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 phone’s body frame and an Earth-fixed frame be $R$. Let the acceleration of gravity and the magnetic field strength be $\mathbf{g}$ and $\mathbf{m}$, respectively. Based on these two vectors, we may construct an orthonormal triad by applying Gram–Schmidt orthonormalization. As follows:

$$
\mathbf{u}_1 = \frac{\mathbf{g}}{\|\mathbf{g}\|}, \quad \mathbf{u}_2 = \frac{\mathbf{m} - (\mathbf{u}_1 \cdot \mathbf{m})}{\|\mathbf{m} - (\mathbf{u}_1 \cdot \mathbf{m})\|}, \quad \mathbf{u}_3 = (\mathbf{u}_1 \times \mathbf{u}_2).
$$
These vectors have, on the basis of each frame, the following components:

\[ \mathbf{u}_i = u_{i1} \mathbf{e}_1 + u_{i2} \mathbf{e}_2 + u_{i3} \mathbf{e}_3 = u'_{i1} \mathbf{e}'_1 + u'_{i2} \mathbf{e}'_2 + u'_{i3} \mathbf{e}'_3, \quad i = 1, 2, 3, \]

in which \( \{ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \} \) and \( \{ \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3 \} \) are the orthonormal bases of the Earth-fixed and phone body frames, respectively.

With these definitions, we see that

\[
\begin{bmatrix}
  u'_{i1} \\
  u'_{i2} \\
  u'_{i3}
\end{bmatrix}
= R
\begin{bmatrix}
  u_{i1} \\
  u_{i2} \\
  u_{i3}
\end{bmatrix}, \quad i = 1, 2, 3,
\]

and the matrices

\[
\begin{bmatrix}
  u_{11} & u_{21} & u_{31} \\
  u_{12} & u_{22} & u_{32} \\
  u_{13} & u_{23} & u_{33}
\end{bmatrix}, \quad \begin{bmatrix}
  u'_{11} & u'_{21} & u'_{31} \\
  u'_{12} & u'_{22} & u'_{32} \\
  u'_{13} & u'_{23} & u'_{33}
\end{bmatrix}
\]

are both orthogonal. It follows that

\[
\begin{bmatrix}
  u'_{11} & u'_{21} & u'_{31} \\
  u'_{12} & u'_{22} & u'_{32} \\
  u'_{13} & u'_{23} & u'_{33}
\end{bmatrix}
= R
\begin{bmatrix}
  u_{11} & u_{21} & u_{31} \\
  u_{12} & u_{22} & u_{32} \\
  u_{13} & u_{23} & u_{33}
\end{bmatrix}
\]

making the rotation matrix \( R \) uniquely computable. A requirement for this is, of course, that the magnetic-field vector \( \mathbf{m} \parallel \mathbf{g} \), which is fulfilled everywhere except on the magnetic poles. This is logical, as at those two locations, a magnetic compass is useless as well!

### 12.2 Zero-velocity update

As we saw above, an inertial measurement units loses both location and attitude position progressively over time, due to the double integration 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 is called a zero-velocity update or ZUPT.

We can analyze the error propagation in this case. The accelerometers may have both random and systematic errors. The systematic errors can be eliminated by system calibration, determining both the scale errors and the alignment errors of the accelerometers, and possibly their nonlinearities. Also the gyros have both random and systematic (orientation drift) errors. An alternative approach to eliminate systematic
effects like sensor scale and alignment errors and gyro drifts is modelling them by including their model equations as part of the Kalman filter.

Random errors in acceleration will produce, by integration, random but correlated errors in velocity that grow with the square root of time:

\[ \sigma_v \sim \sigma_a \sqrt{t} \]

in which \( \sigma_a \) is the random uncertainty of the acceleration. Upon second integration, the random uncertainty of position will grow with the power \( \frac{3}{2} \) of time:

\[ \sigma_x \sim \sigma_a t^{3/2}. \]

The propagation of the directional uncertainty of the gyros is trickier. If we assume that the direction error behaves like a random walk:

\[ \sigma_\theta \sim \sigma_0 \sqrt{t}, \]

it follows that the impact on the acceleration vector is

\[ \sigma_{a,\theta} \sim a \sigma_0 \sqrt{t}, \]

and on the velocity vector, through integration,

\[ \sigma_{v,\theta} \sim a t \sigma_0 \sqrt{t}. \]

This impact will be in a sideways direction for both vectors. For location we find

\[ \sigma_{x,\theta} \sim (n \Delta t)^2 \sigma_0 \sqrt{t}. \]

Introducing zero-velocity updates at intervals \( \Delta t \) will give us (assuming the velocity errors in different update intervals are statistically independent)

\[ \sigma_v \sim \sigma_a \sqrt{\Delta t} \implies \sigma_x \sim \sigma_a \sqrt{n \Delta t} \sqrt{\Delta t} = \sigma_a \Delta t \sqrt{t} \]

with \( n \) the number of zero-velocity updates, \( t = n \Delta t \). Also

\[ \sigma_{a,\theta} \sim a \sigma_0 \sqrt{\Delta t} \implies \sigma_{v,\theta} \sim a \Delta t \sigma_0 \sqrt{\Delta t} \]

\[ \implies \sigma_{x,\theta} \sim a \sqrt{n \Delta t}^2 \sigma_0 \sqrt{\Delta t} = a \Delta t^2 \sigma_0 \sqrt{t}. \]

We see that the growth in positional uncertainty, which was earlier of powers \( \frac{3}{2} \) and \( \frac{5}{2} \) in elapsed time, has now been brought down to a dependency of power \( \frac{1}{2} \) on time!
A simple one-dimensional IMU Kalman-filter simulation, without (black) and with (red) zero-velocity updates. The update points are in blue.

\[ \begin{bmatrix} x_{i+1} \\ v_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ v_i \end{bmatrix} + \begin{bmatrix} 0 \\ n_a \Delta t \end{bmatrix}, \]

with \( n_a \) random, uncorrelated error (white noise) representing the accelerometer noise. The quantity plotted is \( x_i = x(t_i) \), the one-dimensional location. \( \Delta t = t_{i+1} - t_i \) is the time step.

\[ \frac{d}{dt} x(t) = v, \]
\[ \frac{d}{dt} v(t) = n_a. \]

**Figure 12.1.** A simple Kalman filter with and without zero-velocity updates.
12.3 Integration of GNSS and IMU

This is an example of heterogeneous sensor fusion.

An inertial measurement unit (IMU) as already presented in section 5.2 contains three accelerometers and three (laser) gyroscopes. With their aid it can follow its own (i.e., the equipment into which it is integrated) rotational movements as well as linear movements. As time elapses, the positioning solution, obtained by integrating the measured accelerations twice in succession, first into velocities, and then once more into locations, will deteriorate.

The deterioration may be controlled, if one at certain intervals performs a “real” positioning, e.g., using GNSS. Thus one may build a system that preserves its positioning precision even though the GNSS signal is patchy (tunnels, bridges, indoor space). At the same time we also obtain a precise orientation for the equipment.

An example of such an integrated equipment is the Novatel SPAN, Synchronized Position Attitude Navigation, (Hexagon, SPAN GNSS Inertial Navigation Systems).

The measurement precisions of these devices are impressive: the stability of rotation is (for a high quality mechanical gyroscope) 0.0001°/h.

12.4 GNSS multi-antenna system

This is an example of homogeneous sensor fusion.

Also with a GNSS system we may measure attitudes, by using several (at least three) different antennas. The method used is real-time kinematic positioning over very short vectors.

As can be seen from the figure, the same satellite is being observed from two different antennas. The observable is the difference between two measurements of carrier phase:

\[ \Delta P^s = P^s_2 - P^s_1 = \langle \mathbf{v} \cdot \mathbf{e}^s \rangle + k \lambda, \]

in which \( \mathbf{v} \) is the inter-antenna vector, \( \mathbf{e}^s \) is the direction vector to used satellite \( s \) (unit vector, \( \| \mathbf{e} \| = 1 \)) and \( k \) is an integer describing the ambiguities (potential for many alternative values).

We re-write this by reducing the observable to the interval \([0, \lambda)\):

\[ \Delta P^s \mod \lambda = \langle \mathbf{v} \cdot \mathbf{e}^s \rangle + k' \lambda. \]
Here, we have to solve simultaneously for the vector \( v \) and the integer unknown (ambiguity) \( k' \). Solving for the vector requires observations from at least three different satellites, and then the values \( k' \) remain still undetermined:

\[
\begin{align*}
\Delta P^1 \mod \lambda &= \langle v \cdot e^1 \rangle + k'^1 \lambda, \\
\Delta P^2 \mod \lambda &= \langle v \cdot e^2 \rangle + k'^2 \lambda, \\
\Delta P^3 \mod \lambda &= \langle v \cdot e^3 \rangle + k'^3 \lambda.
\end{align*}
\]

About the values \( k'^1, k'^2, k'^3 \) we know at least that they cannot be very large if the vector \( v \) is short: as \( \Delta P \mod \lambda \) lies in the interval \([0, \lambda)\) and \( \langle v \cdot e^i \rangle \) in the interval \([-\|v\|, \|v\|]\), then \( k'^i \) can only lie in the interval \([-\|v\|/\lambda, \|v\|/\lambda + 1]\). If the vector is, e.g., 2 m long, and the wavelength is 24 cm, then the only possible values for \( k'^i \) are: \(-8, -7, ..., +8, +9\).

The solution is obtained as follows:

1. Try out all possible values \( k'^i \) for three satellites 1, 2 and 3, and compute for every combination a vector solution \( v \). The total number of solutions to be computed is in the example case \( 18^3 = 5832 \).

2. If the vector is, e.g., 20 m long, then the total number of solutions to be computed is already \( 180^3 = 5.8 \) million. This requires already processing capacity. On the other hand, if we have the use of a dual-frequency device, we may use widelaning, the effective wavelength of which is 86 cm. Then we need only \( 48^3 = 110592 \) different solutions.

3. If we can see more than three satellites, we choose three of them,
which together produce the best possible geometry. This is easy: traverse all triplets and compute their determinant \( e_1 \cdot (e_2 \times e_3) \). The maximum value wins. If we can see, e.g., 10 satellites, we have to compute \( 10 \cdot 9 \cdot 8 = 720 \) determinants.

4. After this, we compute for every provisional solution thus found, \( \hat{v}_{k^a,k^b,k^c} \), whether \( \|\hat{v}\| \) is close enough to the known distance between the antennas. Solutions that are not, can be discarded immediately.

5. After this we compute again the observables of the other satellites:

\[
\begin{align*}
\hat{\Delta}P_4 \mod \lambda &= \langle \hat{v} \cdot e_4 \rangle + k^4 \lambda, \\
\hat{\Delta}P_5 \mod \lambda &= \langle \hat{v} \cdot e_5 \rangle + k^5 \lambda, \\
& \quad \ldots \\
\hat{\Delta}P_n \mod \lambda &= \langle \hat{v} \cdot e_n \rangle + k^n \lambda.
\end{align*}
\]

All the values on the left hand side should agree, within their measurement uncertainties, with the measured carrier phases for one value \( k^4, k^5, \ldots \). Generally, this only happens for all values \( k^s, s = 4, \ldots, n \) only in the case of one solution.

6. Using the set of values thus found, \( k^s, s = 4, \ldots, n \), the final adjustment is executed in order to compute \( v \) from all observations.

7. When the device (vehicle) moves, and no cycle slips occur, the values \( k^s \) stay the same. Then one can continuously solve for \( v \) in real time from the observations collected.

12.5 Modern radionavigation

The proliferation of base-station networks, e.g., for supporting mobile telephony, but also wireless LAN (WLAN), have made radio navigation in two dimensions, on the Earth’s surface, actual again. Techniques used may be

- cell identity. Here the location is constrained by which base stations are within the range of the navigator. The method is imprecise outside cities, more precise within them.
- signal strength. Two base stations needed, low precision.
- range difference — like the old Decca system and similar ones used in maritime radio navigation. This is also called TDOA, time
difference of arrival, also hyperbolic navigation. A position fix here requires the use of three base stations.

- range. This is called a time-of-arrival (TOA) method. This requires back-and-forth signaling between the navigator and each base station, of which there must be at least two.

Furthermore, the use of GNSS for mobile-phone positioning may be assisted by the base-station network, e.g., by providing satellite ephemeris faster than the satellites themselves are broadcasting them. This is called Assisted GNSS, or AGNSS.

### 12.6 Microelectronic Motion Sensors (MEMS)

MEMS devices are small and inexpensive acceleration and rotation sensors. The manufacturing process is often similar to that used in manufacturing computer micro circuits: photolithography.

#### 12.6.1 Accelerometers

These circuits measure accelerations by measuring, e.g., capacitively, the movement of a small test mass under the influence of acceleration (a pseudo-force). One model is capable of measuring down to $1.7\,g$ ($17\,m/s^2$) with stability $0.2\%$, i.e., $340\,mGal$. In order to get an idea what this means, let us say that an acceleration of $340\,mGal$ during a minute moves over a distance of $6\,m$. The sensitivity of the device is even better than that, several $mGal$. It also survives dropping, e.g., onto a concrete floor (acceleration $3500\,g$!). They have also been fired from cannon.

Prices are nowadays (2018) around a euro even for three-axis sensors. They are a few millimetres in size. Applications: e.g., triggering sensors for automotive airbags, and drop protection triggering sensors for laptop hard drives. These mass markets have driven down prices.

#### 12.6.2 Rotation sensors

Rotation sensors commonly are based on measuring the frequency of oscillation of a tuning fork, see Wikipedia, Vibrating structure gyroscope. Like a rotating object, also a vibrating object tries to stay within the same plane. The torque required for this, the Coriolis force, is measured, e.g., capacitively; the measurement value is proportional to the rotation rate. See figure 12.3. A three-axis or degrees-of-freedom device is manufactured from three one-axis components.
Sensor fusion, sensors of opportunity

**Figure 12.3.** The principle of a MEMS rotation sensor. When the substrate rotates, the vibration of the test mass (in the picture left and right) will cause a displacement due to the Coriolis force (up-down, also periodic), which is measured capacitively.

These devices are capable of a precision of $0.1^\circ$/s which is six orders of magnitude poorer than a "real" inertial device, see above.

Fields of application: stabilizing the image of video cameras, robotics, unmanned aerial devices (UAD), ... 

### 12.7 Pedestrian navigation

When a pedestrian uses a mobile phone for navigation, the app inside the phone will assume that the top edge of the screen will point in the direction of motion. One gets quickly used to this intuitive requirement: by holding the phone in this way, the map displayed on the phone’s screen will align with the surrounding landscape.

This does however require that the user executes some movement, a few metres at least: the motion vector in the terrain as observed by GNSS will then be identified in the map, and used to orient the map image correctly.\(^2\)

When applying inertial navigation technology in personal navigation, the principle of zero-velocity update has been used in personal navigation devices with an IMU to be mounted in boots. Every time a boot hits the ground, the velocity in the Kalman navigation filter is reset to zero.

\(^2\)The same applies with car navigation: the car has to move before the map display can be oriented correctly. This may well lead to the first instruction to the driver being a $180^\circ$ turn. This is assuming the app doesn’t use the built-in compass (magnetometer) for orientation if there is one.
Indoor navigation

An interesting application of this is guiding firefighters, i.e., smoke divers, in a smoke-filled building. For this, the self-contained nature and independence from external signals of inertial navigation is a major feature. For example, Godha et al. (2006).

### 12.8 Indoor navigation

This is a broad and rapidly developing field ([https://en.wikipedia.org/wiki/Indoor_positioning_system](https://en.wikipedia.org/wiki/Indoor_positioning_system)). We only scratch the surface here.

Indoors, the availability of GNSS may not be assumed and is typically lacking. However, other base station types are either available (WLAN, Wireless Local-Area Network) or may be installed. Techniques used for positioning are very similar to traditional radio positioning techniques like Decca. Various approaches available are similar to those using mobile telephony base stations as discussed in section 12.5.

*Pseudolites* are devices that transmit radio frequencies and signals like GNSS satellites, but are installed on Earth. They transmit a carrier modulated with a pseudo-random code (PRN) similar to that of a GNSS satellite, in one of the pseudo-random code slots especially reserved for this. For example, Zhao et al. (2018).

Acoustic positioning has been explored in conference settings, where speakers can be located and tracked, e.g., 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 Connect™ or Skype for Business™. For a description and further references, see Parviainen (2016).

Magnetic-field or WLAN or Bluetooth signal strength can be used for positioning. Inside a building, the magnetic field varies irregularly due to the magnetic properties of building materials used. The same applies to the power of WLAN signals. These could be used for positioning, provided that they are first mapped, or “fingerprinted,” throughout the building. See, e.g., Chen et al. (2013); Liu et al. (2017); Mazlan 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 remain useful in indoor navigation, on their own or in a sensor-fusion context. The major advantage is independence from external signals. As indoor navigation is very often pedestrian navigation, the zero-velocity update technique can be applied advantageously.
Real time systems and networks

Requirements for technological navigation often are the possibility to obtain external data in real time over the data communications network, as well as a processing capability with equipment and software sufficient and appropriate for real-time use. We shall consider those requirements next.

The definition of *real time*:

```
guaranteed latency.
```

Which means a process that has a latency of one month can be real time (if one month is *guaranteed*), but another process with a latency of one millisecond is not real time (if the latency is usually less than one millisecond, but it could sometimes be two, or ten, or even more. . . )

A.1 Communication networks

A.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 form, e.g., by using the Morse code (radio telegraphy), or in analogue form, like sound (radio telephony), images (television), or analogue or digital measurement data (telemetry).

Information is carried on radio waves by *modulation*. Modulation techniques used include amplitude modulation, frequency modulation and phase modulation.
The carrier, two side frequencies, and their sum

\begin{center}
\begin{tabular}{c}
\includegraphics[width=0.8\textwidth]{figure.png}
\end{tabular}
\end{center}

Modulation is the envelope

Figure A.1. Amplitude modulation and bandwidth.

\subsection{A.1.1.1 Example: amplitude modulation}

In Figure A.1 we see how amplitude modulation places a signal (the dashed curve, e.g., a sound wave) on top of the carrier wave. 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

$$A(t) = \cos(2\pi Ft) \cdot (\cos 2\pi ft + 1.5) =$$

$$= 1.5 \cos 2\pi ft + \frac{1}{2} \left( \cos(2\pi(F+f)t) + \cos(2\pi(F-f)t) \right),$$

so we see that the contribution of the modulation can be represented as the semi-sum of two frequencies, $F + f$ and $F - f$.

Now, if the modulating wave contains a large number of different frequencies, $0 < f < f_{\text{max}}$, the resulting spectrum will contain signal, or "power", over the full range $(F - f_{\text{max}}, F + f_{\text{max}})$. We say that the \textit{bandwidth consumption} is $2f_{\text{max}}$.

For broadcasting networks, bandwidth is a scarce and valuable resource, to be carefully allocated.
A.1.1.2 The Nyqvist theorem

One can show that in order to represent a function of time by sample points, the distance $\Delta t$ between the sample points should never be more than one-half the shortest period present in the function. This is called the Nyqvist Theorem. For a function satisfying Nyqvist’s condition, it is possible to transform it back and forth from the time domain $A(t)$ representation to the frequency domain $\tilde{A}(f)$ representation using the discrete Fourier transform. Numerically, typically the Fast Fourier Transform (FFT) is used for computing this.

Now, if we have a modulating function $a(t)$ that has as its highest contained frequency $f_{\text{max}}$, then its shortest contained period is $\frac{1}{f_{\text{max}}}$. The number of samples transmitted per second using amplitude modulation will then be at most $2f_{\text{max}}$, i.e., precisely the effective bandwidth occupied by the modulated signal.

A.1.2 Circuit-switched networks

A.1.2.1 History

The first, still existing and wildly successful switched, or many-to-many, connection network is the telephone network. It is actually circuit switched, i.e., 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 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 and Thomas Edison, contributed valuable ideas before a working implementation became the basis of the first telephone network.

For many years, American Telephone and Telegraph held a monopoly on telephone technology. Off and on, there were anti-trust proceedings against the company, which is also credited with laying the first trans-Atlantic telephone cable, launching the first communications satellite Telstar, and inventing Unix...
before 1900, the first mechanical, automatic switches were built. A number was dialled using a round disc, sending as many pulses as the number printed on the disc edge. This is called “pulse dialling.” Since its introduction in 1963, faster tone dialling has mostly, though not completely, replaced it.

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 extraordinarily user-hostile: looking up telephone numbers was traditionally done manually using thick paper books. The world is divided into national domains having country codes. The United States has code 1, most larger countries have two-digit codes (e.g., Germany 49), while smaller, poorer countries like Finland have settled for three-digit codes (358). Under the national domains are trunk codes, typically (but not necessarily) for cities, within which individual subscribers have their numbers.

Attempts to make phone numbers “mnemonic,” so they can be easier remembered, have pretty much failed. New telephone concepts, such as Internet telephony, might in the longer run change this\(^1\).

The digitization of the telephone network has also made possible to offer customers “always-on” data connections, even over last-few-metres copper, which use frequencies above those used for audible sound. Using a low-pass filter in-between, it is even possible to use voice and data simultaneously on the same line (Digital Subscriber Line, DSL).

\[\text{A.1.2.2 Modems}\]

Given that the phone network is designed for the transport of sound, it is necessary, in order to transport data on it, to convert this to and from sound in the form of (analogue) sound waves. This is done with a device called a \textit{modem} (modulator-demodulator).

The picture A.2 shows one technique — frequency shift keying, FSK — 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. Additionally, checksums are transmitted as well, in order to

\(^{1}\)Remarkably, the application WhatsApp uses the number of the mobile phone it is installed on as user ID! Old habits die hard.
verify that the data received equals the data sent (parity check, cyclic redundancy check\(^2\) CRC) even over noisy lines. Data compression is used if possible and speeds up especially the transfer of textual material.

There are a number of standards for modems, mostly created by the International Telecommunications Union. Over a good quality analogue line, 56k bits per second is the best achievable.

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 bit stream, is layered. Other layers can still be placed on top of this: the Internet Protocol (IP) and the Transmission Control Protocol (TCP) to be discussed later, advanced protocols such as the 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, typically the higher layers are implemented in software, whereas the lowest layers are hardwired. E.g., telephone sound is transmitted traditionally as voltage fluctuations in a copper wire. As digital technology develops, however, the software comes down in the stack: for all but the last few metres, nowadays telephone sound moves as digital bit patterns, often in optic-fibre cables.

This downward migration of software is leading to devices that previously were very different, to become almost the same on the hardware level. E.g., a telephone and a television set — and a GNSS receiver — are becoming mostly just general purpose computers, differently programmed and with different peripherals. This phenomenon is known as (technological) convergence.

\(^2\)Both parity check and CRC are somewhat similar to the “casting out nines” check on manual calculations.
A.1.2.3 Mobile phones

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 are 9600 – 14400 bits per second. As GSM is a natively digital telephony system, it wouldn’t be correct to talk about “GSM modems,” as has sometimes been heard.

However, currently more advanced protocols such as GPRS (General Packet Radio Services) and EDGE (Enhanced Data Rates for GSM Evolution) are in use, which allow always-on digital connections with much higher data rates. Often these technologies are described in terms of generations: 2G, 3G, 4G.

This brings us to the following subject: packet-switching networks.

A.1.3 Packet-switching networks

The classical packet-switching network is the Internet. Also this is 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 from military research: packet-switched network transfer is less vulnerable to localized network damage, and thus harder to interrupt.

The functioning of the Internet, IP addresses, and domain name services (DNS) is explained in many places\(^3\) and we will not repeat it here. There are a number of protocols built upon the Internet Protocol, the most important of which are

- **ICMP** (Internet Control Message Protocol), e.g., the well-known “ping”\(^4\) command for checking network connectivity.

- **UDP** (User Datagram Protocol) is a connectionless protocol. Essentially, a transmitter sends out packets, and a receiver receives them — most of the time. There is no check on successful reception, and not even of whether packets purported to come from the same source actually do. But UDP’s overhead is low, which is why it is sometimes used. E.g., the Network Time Protocol NTP uses UDP. A time server just sprays packets around for clients to pick up and synchronize their clocks 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, typically 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 safeguarded by each host randomly choosing the starting value of its packet counter for this

\(^3\)E.g., Wikipedia, Domain Name System.

\(^4\)The name is onomatopoeic and mimicks the sound of sonar, Muuss (undated).
connection. Such a connection could be hijacked in principle — a so-called “man-in-the-middle attack” — but it is not easy.

Every packet contains two data fields called *source port* and *destination port*. These are numbers between 1 and 65535 which are used to distinguish various service types from each other. E.g., HTTP uses port 80 — usually\(^5\). It is important to understand that these “ports” are purely software things: it is the networking software layer in the operating system that distinguishes these port numbers from each other and directs packets to appropriate server/client processes. Nothing like a (hardware) serial or parallel or USB port!

Note that *none* of these Internet protocols is *real time*. 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, e.g., for music streaming. But already modest congestion — locally or upstream — will make transmission times totally unpredictable.

### A.2 Real-time systems

#### A.2.1 Hardware

In real-time systems used for navigation, digital hardware included will typically have a rather low processing capacity. Think, e.g., of mobile phones: the dictate of low power consumption and small form factor limits what kinds of circuitry one can use, and how much of it.

Another limitation may be, that no full-blown keyboard may be used, and instead of a mouse, a stylus or a finger touch screen — of limited size — is indicated. Also physical ruggedness may be required depending on the navigation environment.

#### A.2.2 Operating systems

The hardware limitations mentioned obviously also limit what operating system software can be used. Typically found are “embedded” operating systems, like Windows Embedded.

In high-reliability operations, e.g., on spacecraft, also systems like the QNX and Wind River Systems\(^6\) real-time embedded operating systems

---

\(^5\)There is a list of all services in the file `/etc/services`.

\(^6\)The Mars rovers Spirit and Opportunity used and use the Wind River Systems
are being used. In “hard” real-time applications, the operating system should preferably not crash.7

Linux and Unix variants are also being used and have become recently quite popular, e.g., Android and the iPhone’s OS X.

It will be clear that, for interfacing with various devices such as GPS and other sensors, the availability — or easy development — of device drivers is critical.

As, with technology development, hardware capability grows while size and power consumption drops, more and more “general” consumer grade operating systems, slightly adapted, are finding their way also into these constrained mobile platforms.

A typical operating system functions in the following way: upon start-up, after operating system, file system and device driver functions have been enabled, the initial process goes into multi-user mode and spawns all the background service processes (deamons) that are supposed to run on this system. Then it loads a login process, presenting it to the user on one or more consoles connected to the system. When a user logs in, he is presented with a shell or command interpreter, allowing him to start his own user processes.

On consumer grade OSes, a windowing GUI or Graphical User Interface is started up as well at this stage, enabling operation by lightly trained personnel. This however demands extra resources. Also from the GUI, user processes can be started in addition to the system processes underlying OS and GUI operation. The defining property of an operating system is, that it manages the system’s various resources in a way that is transparent to the user. Device drivers are one example of this. And, e.g., processor resources are managed through the scheduler.

A.2.3 Process flow

Looking at a single process8, we can say that the path of execution is linear. This means that execution either proceeds to the next statement of the program, or to the statement pointed to by a branching (if, switch,

---

7... which however the Spirit’s system did, due to running out of file handles. But it came beautifully back up again.

8... and ignoring threading!
...) 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.

Looking at a procedure (also called subroutine or function or method), it is only executed because another procedure, and ultimately the main program, called it in the course of its linear execution. The way a procedure is executed is as follows: when it is called, it places a return address — the current value held by the program counter in the calling procedure — 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 references to global variables, in the procedure call are also placed on the top of the stack, which thus grows. There are two ways of calling parameters or arguments with a procedure: call by name and call by value.

- Call by name: this happens when the name of a variable is used as a parameter or 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: this happens when, 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 placed on the stack.

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. Inside the procedure then an assignment is made to the variable, i.e., to the storage location pointed to by it, where the calling procedure can find it.

Also local variables declared within the procedure are allocated on the stack. The scope of these variables — their validity or visibility — is limited to within the procedure being executed. When the flow of control meets the end of the procedure, first these local variables are deallocated, after that the procedure’s stacked parameters, and finally the top of the stack is moved back into the program counter of the processor again, and we have returned to the calling procedure, which continues from the statement following the procedure call.

Major advantages of using an execution stack are, that

1. Local variables declared inside procedures are released upon return from the procedure, and

2. Procedures can be called recursively: every instance of the procedure
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 c
c = sum(3, 7)
print c
>> 13

procedure sum(a, b, c)
c = a + b
return ...
declare a, b, c
a = 6
b = 5
# Call by name a, b
call sum(a, b, c)
print c
>> 11

Figure A.4. Procedure call and stack allocation

has its own stack frame with its own version of the call parameters and local variables. This does require of course, that the recursion must be ended at some point by a suitable condition: if not, a stack overflow error is inevitable.
A.2.4 Interrupts, masking, latency

Interrupts are both similar to procedure calls, and different. The essential difference is that interrupts can happen at any time, due to an external (hardware) event, typically from a peripheral that either has data available that it wants read, or expects data to be written. The main reason for using interrupts is that, without them, the peripheral would have to be polled periodically, at a sufficient rate in order not to miss any transferable data. This polling consumes processor time.

Computer hardware provides for a number of different interrupts. When they happen, it is their responsibility not to change anything that could interfere with the processes scheduled for execution. Interrupts are used, e.g., 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 it is triggered.

Take the clock interrupt routine, for example. It is triggered typically 50 times a second, and its gets its name from what it does, increment the time register kept by the operating system software. But it is usually also responsible for task switching or context switching, allowing the running of multiple tasks apparently simultaneously.

At every task switch, the context of the currently running process — the set of data, including CPU registers and especially the program counter, that it will need 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 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, e.g., waiting for user input —, and should have a high enough priority.

Every process — especially kernel or system level processes — have pieces in their code during which it would be wrong or disastrous to be interrupted. 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. Critical kernel routines will mask the clock interrupt, and unmask it again when finished.
A.2.5 Requirements for real-time use

Now, the requirements for real-time use are

1. We should know in advance which processes will be running on our system. An environment like a multi-user server into which people can log in and start user processes at will, is not acceptable.

2. We should know in advance what are the longest pieces of code, execution-time wise, that the various runnable processes contain during which they can 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.

5. The total processing capacity of the system should be sufficient
   (a) on average for all processes, and
   (b) at every point in time for all the real-time processes taken together.

Meeting hard real-time requirements is quite demanding and requires extensive testing and substantial over-allocation 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, i.e., 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.
Useful matric equations

B.1 The first equation

\[(A + B)^{-1} = \left(A (I + A^{-1}B)\right)^{-1} = \left(A (B^{-1} + A^{-1}) B\right)^{-1} = B^{-1} (A^{-1} + B^{-1})^{-1} A^{-1}.\]

Substitute

\[B^{-1} = (A^{-1} + B^{-1}) - A^{-1}\]

and obtain

\[(A + B)^{-1} = \left((A^{-1} + B^{-1}) - A^{-1}\right) (A^{-1} + B^{-1})^{-1} A^{-1} = A^{-1} - A^{-1} (A^{-1} + B^{-1})^{-1} A^{-1}.\]

B.2 The second equation

We write

\[B = UCV.\]

Study the following partitioned equation:

\[
\begin{bmatrix}
A & U \\
V & -C^{-1}
\end{bmatrix}
\begin{bmatrix}
D_{11} & D_{12} \\
D_{21} & D_{22}
\end{bmatrix}
= \begin{bmatrix}
I & 0 \\
0 & I
\end{bmatrix}.
\]

This may be written out as four matric equations:

\[
\begin{align*}
AD_{11} + UD_{21} &= I, \\
AD_{12} + UD_{22} &= 0, \\
VD_{11} - C^{-1}D_{21} &= 0, \\
VD_{12} - C^{-1}D_{22} &= I.
\end{align*}
\]
Of these four equations, we only need the first and the third in the sequel.

Add equation B.2 multiplied by UC to equation B.1:

\[(A + UC)D_{11} = I \implies D_{11} = (A + UC)^{-1}.\]

Subtract equation B.1 multiplied by VA\(^{-1}\) from equation B.2:

\[(C^{-1} - VA^{-1}U)D_{21} = -VA^{-1} \implies D_{21} = -(C^{-1} - VA^{-1}U)^{-1}VA^{-1}.\]

Substitute back into equation B.1:

\[AD_{11} - U(C^{-1} - VA^{-1}U)^{-1}VA^{-1} = I \implies D_{11} = A^{-1} + A^{-1}U(C^{-1} - VA^{-1}U)^{-1}VA^{-1}.\]

Now we have two different expressions for the sub-matrix \(D_{11}\), which have to be identical. Thus we obtain

\[(A + UC)^{-1} = A^{-1} + A^{-1}U(C^{-1} - VA^{-1}U)^{-1}VA^{-1},\]

the Woodbury matrix identity (Keijo Inkilä, personal communication), Wikipedia, Woodbury matrix identity.
Allan’s Time, The Allan Variance. The Allan Variance. URL


BeiDou Introduction. CSNO-TARC, Test and Assessment Research Center of China Satellite Navigation Office. URL


ESa, Introducing GOCE. URL https://www.esa.int/Our_Activities/Observing_the_Earth/GOCE/Introducing_GOCE. Accessed 5. toukokuuta 2019. 138


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. 123


Bibliography


Bibliography


Wolfram Functions, \( \int e^{bx} \cos cx \, dx \). URL http://functions.wolfram.com/ElementaryFunctions/Cos/21/01/02/04/01/01/0002/. Accessed 3. huhtikuuta 2020. 28


Index

A
acceleration
  of a vehicle, 79
accelerometer, 134, 136
    linear, 93
    pendulous, 87
      gyroscope based, 87
    three-axes, 196
Acoustic positioning, 205
actuator, 87, 95, 96, 100
adjustment calculus, 54
air drag, 41
airborne gravimeter, 94
airborne gravimetry, 135
Allan variance, 156
ambiguity number, 157
ambiguity resolution
  multiple ambiguities, 166
ambiguity unknown, 165
American Telephone and Telegraph, 209
angular momentum, 82
angular velocity, 82
  Earth rotation, 90
anomaly
  eccentric, 110
  mean, 110
  true, 110
antipodes, 179
anti-trust law, 209
approximate value, 39, 41
AR(1) process, 24
assignment, to variable, 216
Assisted GNSS (AGNSS), 203
attitude
  of a vehicle, 80
attitude determination, 200
autocorrelation, 20
autocovariance, 19, 40
autocovariance function, 23
autokovariansifunktio, 22
Automatic Precise Positioning Service, 193
autoregressive model, first order, 154
average, 16
azimuth
  of gyroscope axis, 96

B
bandwidth, 208
batch processing, 67
BeiDou-2, 182
BeiDou-3, 182
bell curve, 13
Bell, Alexander Graham, 209
Big Dipper (constellation), 182
Binary Phase-Shift keying, 179
binary phase-shift keying, see BPSK
Bluetooth, 205
body co-ordinate frame, 83, 85
BPSK, 143
Brennschluss, 87
broadcast ephemeris, 153
Bucy, Richard, 35

C
call by name, 216
call by value, 216
cardanic ring system, 89
cardanic suspension, 85
carrier phase, 131
casting out nines, 211
Cauchy-Lorentz distribution, 28
cell identity, 202
centrifugal force, 93
chain rule, 91
CHAMP, 136
chronometer, 4
circuit-switched network, 209
clock error, 157
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPS receiver</td>
<td>151</td>
</tr>
<tr>
<td>GPS satellite</td>
<td>151</td>
</tr>
<tr>
<td>clock interrupt</td>
<td>218</td>
</tr>
<tr>
<td>clock modelling</td>
<td>156</td>
</tr>
<tr>
<td>closing error</td>
<td>54</td>
</tr>
<tr>
<td>code division multiple access</td>
<td>180</td>
</tr>
<tr>
<td>coloured noise</td>
<td>23–25, 64, 65</td>
</tr>
<tr>
<td>concessionaire</td>
<td>180</td>
</tr>
<tr>
<td>context switching</td>
<td>218</td>
</tr>
<tr>
<td>context, of a process</td>
<td>218</td>
</tr>
<tr>
<td>convergence, of technologies</td>
<td>211</td>
</tr>
<tr>
<td>co-ordinate frame</td>
<td></td>
</tr>
<tr>
<td>co-ordinate, co-rotating</td>
<td>90, 93, 117</td>
</tr>
<tr>
<td>co-ordinate reference system</td>
<td></td>
</tr>
<tr>
<td>astronomical</td>
<td>116</td>
</tr>
<tr>
<td>inertial</td>
<td>116</td>
</tr>
<tr>
<td>Coriolis force</td>
<td>93, 203</td>
</tr>
<tr>
<td>correlation length</td>
<td>32, 137</td>
</tr>
<tr>
<td>correlator</td>
<td>158</td>
</tr>
<tr>
<td>Costas discriminator</td>
<td>159</td>
</tr>
<tr>
<td>country code</td>
<td>210</td>
</tr>
<tr>
<td>CRC</td>
<td>211</td>
</tr>
<tr>
<td>critical damping</td>
<td>97</td>
</tr>
<tr>
<td>cross correlation</td>
<td>20</td>
</tr>
<tr>
<td>cross-covariance</td>
<td>19</td>
</tr>
<tr>
<td>Cruise missiles</td>
<td>2</td>
</tr>
<tr>
<td>cycle slip</td>
<td>8, 161</td>
</tr>
<tr>
<td>cyclic redundancy check</td>
<td>211</td>
</tr>
<tr>
<td>cylinder, moments of inertia</td>
<td>85</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td></td>
</tr>
<tr>
<td>damping factor</td>
<td>98</td>
</tr>
<tr>
<td>damping, of a gravimeter</td>
<td>134</td>
</tr>
<tr>
<td>deamon</td>
<td>215</td>
</tr>
<tr>
<td>Decca</td>
<td>4</td>
</tr>
<tr>
<td>declaration, of variable</td>
<td>216</td>
</tr>
<tr>
<td>degree of freedom</td>
<td>31, 39</td>
</tr>
<tr>
<td>delta function</td>
<td>21</td>
</tr>
<tr>
<td>design matrix</td>
<td>51</td>
</tr>
<tr>
<td>destination port</td>
<td>214</td>
</tr>
<tr>
<td>device driver</td>
<td>215</td>
</tr>
<tr>
<td>differential correction</td>
<td>185</td>
</tr>
<tr>
<td>Digital Subscriber Line</td>
<td>210</td>
</tr>
<tr>
<td>Dirac, Paul</td>
<td>21</td>
</tr>
<tr>
<td>distribution</td>
<td>21</td>
</tr>
<tr>
<td>DNS</td>
<td>213</td>
</tr>
<tr>
<td>domain name service</td>
<td>213</td>
</tr>
<tr>
<td>double difference (observable)</td>
<td>147</td>
</tr>
<tr>
<td>drag</td>
<td></td>
</tr>
<tr>
<td>atmospheric</td>
<td>136</td>
</tr>
<tr>
<td>drift</td>
<td></td>
</tr>
<tr>
<td>of a gravimeter</td>
<td>134</td>
</tr>
<tr>
<td>of an accelerometer</td>
<td>87</td>
</tr>
<tr>
<td>DSL</td>
<td>210</td>
</tr>
<tr>
<td>dynamic model</td>
<td>10, 35, 39</td>
</tr>
<tr>
<td>dynamic noise</td>
<td>39, 68, 70</td>
</tr>
<tr>
<td>Dynamic Model Compensation</td>
<td>65</td>
</tr>
<tr>
<td><strong>E</strong></td>
<td></td>
</tr>
<tr>
<td>number $e$</td>
<td>44</td>
</tr>
<tr>
<td>eccentricity</td>
<td>110</td>
</tr>
<tr>
<td>economics</td>
<td>35</td>
</tr>
<tr>
<td>EDGE</td>
<td>212</td>
</tr>
<tr>
<td>Edison, Thomas Alva</td>
<td>209</td>
</tr>
<tr>
<td>EGM2008</td>
<td>93</td>
</tr>
<tr>
<td>Enhanced Data Rates for GSM Evolution</td>
<td>212</td>
</tr>
<tr>
<td>Eötvös correction</td>
<td>135</td>
</tr>
<tr>
<td>Eötvös tensor</td>
<td>42</td>
</tr>
<tr>
<td>epoch</td>
<td>51, 54</td>
</tr>
<tr>
<td>ergodicity</td>
<td>13</td>
</tr>
<tr>
<td>error ellipse</td>
<td>54</td>
</tr>
<tr>
<td>estimator</td>
<td></td>
</tr>
<tr>
<td>enhanced</td>
<td>52</td>
</tr>
<tr>
<td>of state vector</td>
<td>51</td>
</tr>
<tr>
<td>unbiased</td>
<td>51</td>
</tr>
<tr>
<td>Ethiopian airborne gravity survey</td>
<td>138</td>
</tr>
<tr>
<td>ETRF</td>
<td>7</td>
</tr>
<tr>
<td>ETRF-89</td>
<td>7</td>
</tr>
<tr>
<td>Euler angles</td>
<td>39</td>
</tr>
<tr>
<td>EUREF-89</td>
<td>7</td>
</tr>
<tr>
<td>Eurocontrol</td>
<td>190</td>
</tr>
<tr>
<td>European Commission</td>
<td>180, 190</td>
</tr>
<tr>
<td>execution stack</td>
<td>216</td>
</tr>
<tr>
<td>expectancy</td>
<td>14</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td></td>
</tr>
<tr>
<td>FDMA</td>
<td>179</td>
</tr>
<tr>
<td>file handle</td>
<td>215</td>
</tr>
<tr>
<td>file system</td>
<td>215</td>
</tr>
<tr>
<td>filter</td>
<td></td>
</tr>
<tr>
<td>sequential</td>
<td>54</td>
</tr>
<tr>
<td>flattening, of the Earth</td>
<td>130</td>
</tr>
<tr>
<td>Fourier transform</td>
<td>26, 209</td>
</tr>
<tr>
<td>Fourier transform, inverse</td>
<td>26</td>
</tr>
<tr>
<td>freely spinning body</td>
<td>84</td>
</tr>
<tr>
<td>frequency</td>
<td>26</td>
</tr>
<tr>
<td>frequency division multiple access</td>
<td>179</td>
</tr>
<tr>
<td>frequency shift keying</td>
<td>210</td>
</tr>
<tr>
<td>FSK</td>
<td>210</td>
</tr>
<tr>
<td>function space</td>
<td>12</td>
</tr>
<tr>
<td>fundamental equation of inertal navigation</td>
<td>94</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td></td>
</tr>
<tr>
<td>gain matrix, Kalman</td>
<td>52</td>
</tr>
<tr>
<td>Galileo Joint Undertaking</td>
<td>180</td>
</tr>
<tr>
<td>Galileo Operating Company</td>
<td>180</td>
</tr>
<tr>
<td>Galileo system</td>
<td>180</td>
</tr>
<tr>
<td>Gauss–Markov process</td>
<td>23, 24, 26, 28, 137</td>
</tr>
<tr>
<td>autocovariance function</td>
<td>25</td>
</tr>
</tbody>
</table>
power spectral density, 28
GDGPS, 193
General Packet Radio Services, 212
gzeocentric angular distance, 99
gimbar, 85
GIUVE-A, 181
Global Positioning System, see GPS
Global System for Mobile Communications, 212
GNSS, 5, 133
GNSS navigation system, 136
GNSS system, 136
GOCE, 136
GPRS, 212
GPS, 1
GRACE, 136
Gram–Schmidt orthonormalization, 196
Graphical User Interface (GUI), 215
gravimeter, 134
gravitational field, 40
gravitational-gradient tensor, 122, 153
gravity disturbance, 137
gravity gradiometer, 136
gravity vector, 94
Gray, Elisha, 209
Greenland Aerogeophysics Project, 138
Greenland's gravity survey, 138
Greenwich, 116
ground segment (EGNOS), 190, 191
GSM, 212
GSM modem, 212
guidance system, 35
gyro compass, 97, 106
gyroscope, 82
gyroscopic frame, 89

H
harmonic restoration coefficient, 98
Hartebeesthoek (South Africa), 191
Hertz, Heinrich, 26
high dynamic–environment, 155
Hill co–ordinate frame, 118
Hill, George W., 118
homoskedasticity, 29
Hopfield model, 154
hyper-ellipsoid, 55

I
ICMP, 213
inclination, 110, 180
inclinometer, 106
independent and identically distributed, 29
inertial device, 39, 134
inertial measurement unit, 81, 204
inertial navigation, 1, 10, 79
inertial sensor, 205
inertial technology, 133
inertial tensor, 82
inertianavigointi, 79
initialization, 107
instance, of a procedure, 216
instrument landing, 188
integrity, 185, 186
International Space Station, 35
International Telecommunications Union, 211
Internet Control Message Protocol, 213
Internet Protocol, 211, 213
interrupt, 218
interrupt service routine, 218
interrupt vector, 218
IONEX (ionosphere model), 154
ionic engine, 136
ionospheric effect, 151
IP, see Internet Protocol
IRTF, 7

J
Jacobi matrix, 40, 52
Jacobi, Carl, 40
Jacobi, determinant of, 33

K
Kaisaniemi, Helsinki, 13
Kalman filter, 1, 10, 26, 35, 134, 204
extended, 39
for inertial navigation, 104
Kalman, Rudolf, 35
Kepler orbit, 109
Khinchin, Alexander, 26
KKJ (Map Grid Co–ordinate System), 7
Klobuchar model, 154
Kourou (Ranskan Guyana), 191

L
LAMBDA method, 166
LAMBDA technique, 167
law of inertia, 79
least-squares collocation, 67
linear combination, 55
linear regression, 29
non-correlated, 32
linearity, of a sensor, 87
linearization, 40, 41, 103
Linux, 215
Local Area Augmentation System, 188
location, 1
Lunar Orbiter, 65

M
magnetic field, 205
magnetometer, 196, 204
main axes, of a symmetric matrix, 84
maritime navigation, 187
Markov chain, 24, 31
mascons, 65
Master Control Centre, 190
MCC, 190
mechanisation, 10
mechanization, 102
milligal, 134
Mobile telephony technology generations, 212
modem, 210
modulation, 207
amplitude, 207
frequency, 207
phase, 207
Molnya (satellite), 192
moment of inertia, 84
Morse code, 207
moving vehicle, 155
MSAS, 190
MTSAT Satellite-based Augmentation System, 190

N
navigation
  definition, 1
Navigation Land Earth Station, 190
Network Time Protocol, 213
Newton’s law of gravitation, 93
Newton’s second law of mechanics, 99
Newton’s second law of motion, 82
NLES, 190
node
  ascending, 110, 130
noise, 20
noise vector, 138
normal equation, 30
normal gravity, 137
normal gravity field, 93
North Star, 182
NTP, 213
Nyqvist Theorem, 209

O
observable, 51
observation equation, 65
  for geodetic positioning, 154
  for satellite orbit determination, 152
observation matrix, 51
observation model, 10, 36
observational model, 51
operating system, 215
Opportunity (Mars rover), 214
optic fibre, 89
optimality
  of average, 16
  of Kalman filter, 53, 54
orbital elements
  Kepler orbit, 109
orbital motion, Kepler, 43
orbital period
  of a satellite, 98
orthonormal basis, 38, 80, 101, 197
orthonormal triad, 196

P
packet (Internet), 213
packet-switching network, 212
parity check, 211
perigee
  argument of, 110
period
  of a pendulum, 98
phase-locked loop, 159
photolithography, 203
piezo-sensor, 86
ping, 213
pitch, 8
plate tectonics, 155
polarization, 179
polled I/O, 218
positioning
  geodetic, 1
postglacial rebound, 155
Power spectral density, 28
power spectral density (PSD), 26
precession
  of a gyroscope, 95
precise ephemeris, 153
priority, 218
procedure (program component), 216
process
  stationary, 19
product differentiation rule, 91
program counter, 216
propagation delay, 157
propagation of variances, 46
protocol stack, 211
pseudolite, 205
pseudo-random code, 187
pseudorange, 131, 151
pulse dialling, 210
PZ90, 179

Q
Quasi-Zenith Satellite System, 192
QZSS, 192

R
RAIM, 186
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>random walk</td>
<td>22, 25, 26, 47, 153</td>
</tr>
<tr>
<td>Ranging and Integrity Monitoring Station</td>
<td>190</td>
</tr>
<tr>
<td>real time</td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td>207</td>
</tr>
<tr>
<td>hard</td>
<td>219</td>
</tr>
<tr>
<td>use requirements</td>
<td>219</td>
</tr>
<tr>
<td>Receiver Autonomous Integrity Monitoring</td>
<td>186</td>
</tr>
<tr>
<td>rendez-vous</td>
<td>35, 118</td>
</tr>
<tr>
<td>lunar orbital</td>
<td>35, 36</td>
</tr>
<tr>
<td>return address, of procedure call</td>
<td>216</td>
</tr>
<tr>
<td>RIMS</td>
<td>190</td>
</tr>
<tr>
<td>ring-laser gyroscope</td>
<td>88</td>
</tr>
<tr>
<td>roll</td>
<td>8</td>
</tr>
<tr>
<td>RTCM</td>
<td>147</td>
</tr>
<tr>
<td>RTK</td>
<td>151</td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Saastamoinen model</td>
<td>154</td>
</tr>
<tr>
<td>Safety of Life</td>
<td>185</td>
</tr>
<tr>
<td>Safety-of-Life service</td>
<td>182</td>
</tr>
<tr>
<td>Sagnac phenomenon</td>
<td>88</td>
</tr>
<tr>
<td>Satellite Based Augmentation System</td>
<td>185</td>
</tr>
<tr>
<td>satellite clock correction</td>
<td>153</td>
</tr>
<tr>
<td>satellite clock modelling</td>
<td>153</td>
</tr>
<tr>
<td>satellite-to-satellite tracking</td>
<td></td>
</tr>
<tr>
<td>high-low</td>
<td>136</td>
</tr>
<tr>
<td>SBAS</td>
<td>185</td>
</tr>
<tr>
<td>scalar airborne gravimetry</td>
<td>134</td>
</tr>
<tr>
<td>scheduler</td>
<td>215</td>
</tr>
<tr>
<td>scheduling</td>
<td>218</td>
</tr>
<tr>
<td>Schuler loop</td>
<td>94, 100, 103</td>
</tr>
<tr>
<td>Schuler pendulum</td>
<td>98, 100, 102, 103</td>
</tr>
<tr>
<td>Schuler tuning</td>
<td>90</td>
</tr>
<tr>
<td>scope</td>
<td></td>
</tr>
<tr>
<td>of variable</td>
<td>216</td>
</tr>
<tr>
<td>sea gravimeter</td>
<td>94</td>
</tr>
<tr>
<td>sea gravimetry</td>
<td>135</td>
</tr>
<tr>
<td>seafaring</td>
<td>1</td>
</tr>
<tr>
<td>semi-major axis</td>
<td>110</td>
</tr>
<tr>
<td>semi-minor axis</td>
<td>110</td>
</tr>
<tr>
<td>Sensor fusion</td>
<td>195</td>
</tr>
<tr>
<td>sensor fusion</td>
<td></td>
</tr>
<tr>
<td>heterogeneous, 195, 200</td>
<td></td>
</tr>
<tr>
<td>homogeneous, 195, 200</td>
<td></td>
</tr>
<tr>
<td>loosely integrated, 195</td>
<td></td>
</tr>
<tr>
<td>tightly integrated, 195</td>
<td></td>
</tr>
<tr>
<td>sensor integration, 195</td>
<td></td>
</tr>
<tr>
<td>sensors of opportunity, 195</td>
<td></td>
</tr>
<tr>
<td>sequential adjustment, 67</td>
<td></td>
</tr>
<tr>
<td>serial correlation, 32</td>
<td></td>
</tr>
<tr>
<td>shell, 215</td>
<td></td>
</tr>
<tr>
<td>sidereal time</td>
<td>90</td>
</tr>
<tr>
<td>Greenwich</td>
<td>116</td>
</tr>
<tr>
<td>signal strength</td>
<td>202</td>
</tr>
<tr>
<td>Sky Map (Android app)</td>
<td>196</td>
</tr>
<tr>
<td>slant delay</td>
<td>153</td>
</tr>
<tr>
<td>source port</td>
<td>214</td>
</tr>
<tr>
<td>space segment (EGNOS)</td>
<td>190</td>
</tr>
<tr>
<td>spectral filtering</td>
<td>65</td>
</tr>
<tr>
<td>spectrum</td>
<td>26</td>
</tr>
<tr>
<td>Spirit (Mars rover)</td>
<td>214</td>
</tr>
<tr>
<td>spoofing</td>
<td>144</td>
</tr>
<tr>
<td>stabilised platform</td>
<td>94</td>
</tr>
<tr>
<td>stabilized platform</td>
<td>89</td>
</tr>
<tr>
<td>stabilised platform</td>
<td>134</td>
</tr>
<tr>
<td>stack overflow error</td>
<td>217</td>
</tr>
<tr>
<td>standing wave</td>
<td>89</td>
</tr>
<tr>
<td>state estimator</td>
<td>43</td>
</tr>
<tr>
<td>a posteriori</td>
<td>43</td>
</tr>
<tr>
<td>a priori</td>
<td>43</td>
</tr>
<tr>
<td>state propagation</td>
<td>43</td>
</tr>
<tr>
<td>state transition matrix</td>
<td>44</td>
</tr>
<tr>
<td>state variance</td>
<td>165</td>
</tr>
<tr>
<td>definition</td>
<td>43</td>
</tr>
<tr>
<td>state vector</td>
<td>35, 109</td>
</tr>
<tr>
<td>definition</td>
<td>38</td>
</tr>
<tr>
<td>stationarity</td>
<td>12, 20</td>
</tr>
<tr>
<td>stochastic model</td>
<td>10</td>
</tr>
<tr>
<td>statistical model</td>
<td>12, 39</td>
</tr>
<tr>
<td>strapdown inertial measurement unit</td>
<td>88</td>
</tr>
<tr>
<td>support segment (EGNOS)</td>
<td>190</td>
</tr>
<tr>
<td>symbolic algebra</td>
<td>28</td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>task switching</td>
<td>218, 219</td>
</tr>
<tr>
<td>telemetry</td>
<td>207</td>
</tr>
<tr>
<td>telephone directory</td>
<td>210</td>
</tr>
<tr>
<td>telephone network</td>
<td>209</td>
</tr>
<tr>
<td>Telstar (satellite)</td>
<td>209</td>
</tr>
<tr>
<td>time difference of arrival (TDOA)</td>
<td>202</td>
</tr>
<tr>
<td>time of arrival (TOA)</td>
<td>203</td>
</tr>
<tr>
<td>time series</td>
<td>12, 29</td>
</tr>
<tr>
<td>time slice</td>
<td>218</td>
</tr>
<tr>
<td>tone dialling</td>
<td>210</td>
</tr>
<tr>
<td>torque, 82, 88, 102</td>
<td></td>
</tr>
<tr>
<td>torquer, 90</td>
<td></td>
</tr>
<tr>
<td>touch screen</td>
<td>214</td>
</tr>
<tr>
<td>transformation matrix</td>
<td>80</td>
</tr>
<tr>
<td>transitivity</td>
<td>44</td>
</tr>
<tr>
<td>Transmission Control Protocol, 193, 211</td>
<td></td>
</tr>
<tr>
<td>transponder</td>
<td>185</td>
</tr>
<tr>
<td>WAAS, 187</td>
<td></td>
</tr>
<tr>
<td>tropospheric effect</td>
<td>151</td>
</tr>
<tr>
<td>trunk code</td>
<td>210</td>
</tr>
<tr>
<td>U</td>
<td></td>
</tr>
<tr>
<td>UDP, 213</td>
<td></td>
</tr>
<tr>
<td>unbiased estimator</td>
<td>15</td>
</tr>
<tr>
<td>Unix, 209, 215</td>
<td></td>
</tr>
<tr>
<td>update</td>
<td></td>
</tr>
</tbody>
</table>
of state variance, 52
update equations, Kalman filter, 54, 138
update step, Kalman filter, 52
*Ursa Major*, 182
User Datagram Protocol, 213
user segment (EGNOS), 190
UTC (SU), 179

**V**
variable, stochastic
definition, 11
variance ellipse, 55
vectorial airborne gravimetry, 134
velocity vector, 117
vernal equinox, 116
vibration, 87
visual landing, 185

**W**
WAAS, 185, 186
weighted average, 73
weightlessness, 135
WGS84, 7
WhatsApp, 210
white light, 20
white noise, 20, 26, 27, 29, 40, 47, 57, 60
PSD, 27
Wide Area Augmentation System, 186
Wiener process, 22
Wiener, Norbert, 22
Wiener–Khinchin–Einstein theorem, 26
wireless local-area network (WLAN), 202
Wireless Local-Area Network (WLAN), 205
wireless local-area network (WLAN), 195
Wolfram Research, 28

**Y**
yaw, 8

**Z**
zero quantity, 51
zero-velocity update, 81, 196, 197, 204, 205
ZUPT, 197