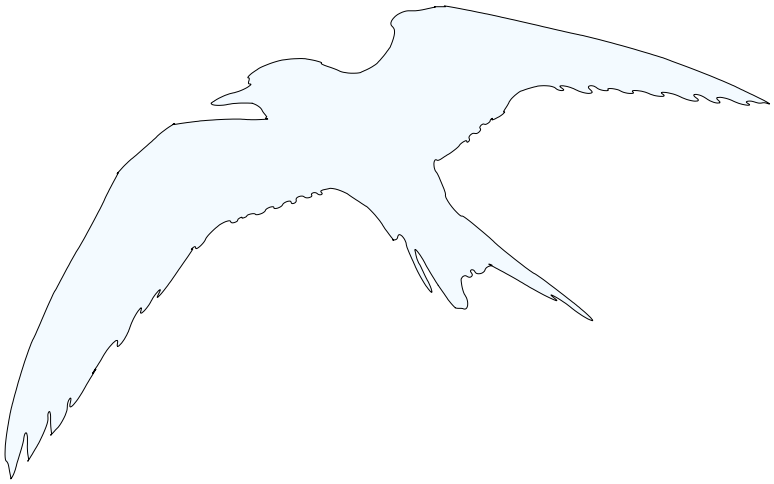
Methods of navigation Maa-6.3285

Martin Vermeer (martin.vermeer@tkk.fi)

November 19, 2015



Course Description

Workload 3 cr

Teaching Period I-II, Lectured in the autumns of odd years.

Learning Outcomes The student understands the basics of "technological navigation": stochastic processes, the Kalman filter, inertial navigation, the use of GPS in real time. He knows how to derive relevant formulas that describe the dynamic behaviour in time of various systems, and the observation equations for the observation types to be used; and how to build a working Kalman solution with these.

Content 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.

Foreknowledge Maa-6.203 or Maa-6.2203 is recommended.

Equivalences Replaces course Maa-6.285.

Target Group

Completion Completion in full consists of the exam and the calculation exercises.

Workload by Component

- Lectures 13 × 2 h = 26 h
- Independent study 24 h
- Calculation exercises 6 ×5 h = 30 h (independent work)
- Total 80 h

Grading The grade of the exam becomes the grade of the course, 1-5

Study Materials Lecture notes. Background material Strang and Borre: Linear Algebra, Geodesy, and GPS Strang and Borre [1997]

Teaching Language English

Course Staff and Contact Info Martin Vermeer, Gentti 4th floor, name@aalto.fi

Reception times By agreement

CEFR-taso

Lisätietoja

Cover picture:

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

(Original english translation: Susanna Nordsten)

Contents

1 Fundamentals of navigation

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 in 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 by two modern technologies: GPS (Global Positioning System) and inertial navigation. Also processing technologies have developed: specifically the recursive linear filter or Kalman filter should be mentioned here.

1.2 History

Old history

Humans have always been discovering the world around them and travelled often long distances[1]. 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.

[1] "Navigare necesse est".

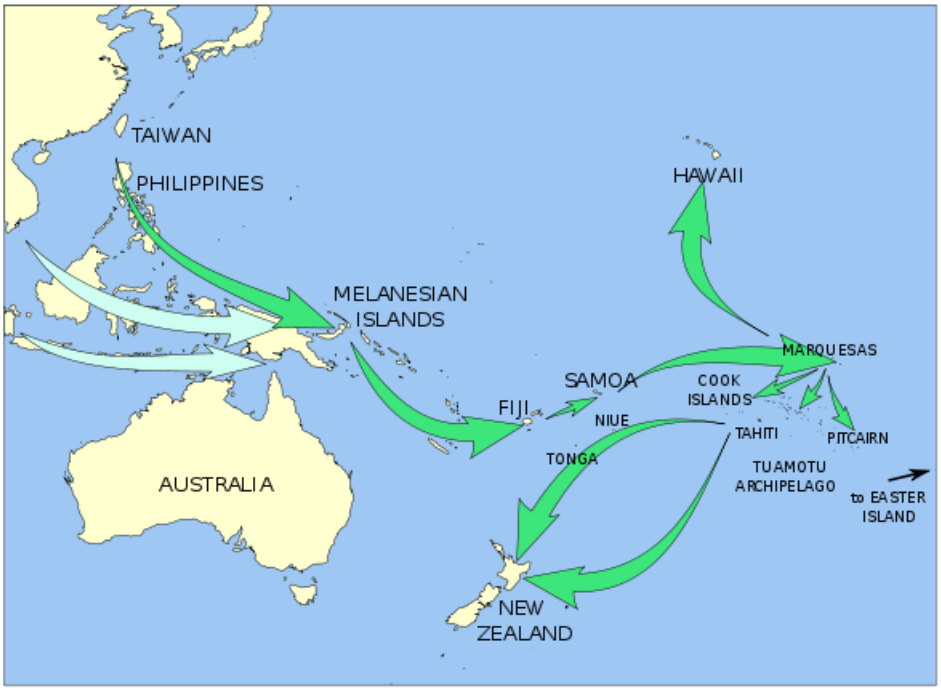


Figure 1.1: Polynesian migration routes, © 2008 Wikimedia Commons / David Hall

Using landmarks this way requires *mapping*, i.e., 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* (http://en.wikipedia.org/wiki/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 (http://www.bbc.co.uk/news/world-africa-11615613) and the archipelagos of the Pacific Ocean got their human settlements (http://www.paulwaters.com/migrate.htm, http://en.wikipedia.org/wiki/Polynesian_navigation, http://www.exploratorium.edu/neverlost/).

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* (http://www.scq.ubc.ca/the-compasses-of-birds/) *have always navigated.*

Figure 1.2: Barnacle geese in autumn migration. © 2006 Wikipedia

Navigation

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

▷ 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.

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.

Figure 1.3: John Harrison's chronometer H5. © 2007 Wikipedia

- ▷ Latitude is easy to get. The height of the celestial pole above the horizon. In the daytime from the Sun.

- ▷ Longitude is the problem: it presupposes the use of an accurate time standard (*chronometer*). Cf. Sobel [1995]. Alternatively, astronomical methods like using the moons of Jupiter as a "clock". Later, distribution of time signals by radio communication, which was not possible until the 20th century.

In the 20th century radio technological methods came into use. The most common is probably *DECCA*, which is 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 colored curve, a *hyperbole*. Every slave station forms with the master a bundle of hyperboles drawn in its own color. 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 also GLONASS) are based on a three-dimensional counterpart of the hyperbolic method.

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 V2, already included a gyroscope based control system. In this case navigation is *guidance*.

The guidance system of the V2 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 closed ("*Brennschluss*"). Physically the turning was done with the aid of small "air rudders" ("control vanes") connected to the tail, that changed the direction of the hot gases coming from the motor. Cf. http://en.wikipedia.org/wiki/V2_rocket[2].

Nowadays complete inertial navigation is used in airplanes and spacecraft. Many other computer based technologies such as satellite positioning (GPS/GNSS) are nowadays used.

1.3 A vehicle's movements

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*. In photogrammetry, we use the term Euler angles.

[2] In fact these were dual rudders: the part sticking into the exhaust stream consisted of graphite and burned up quickly. But by then the rocket was up to speed and the external rudders took over.

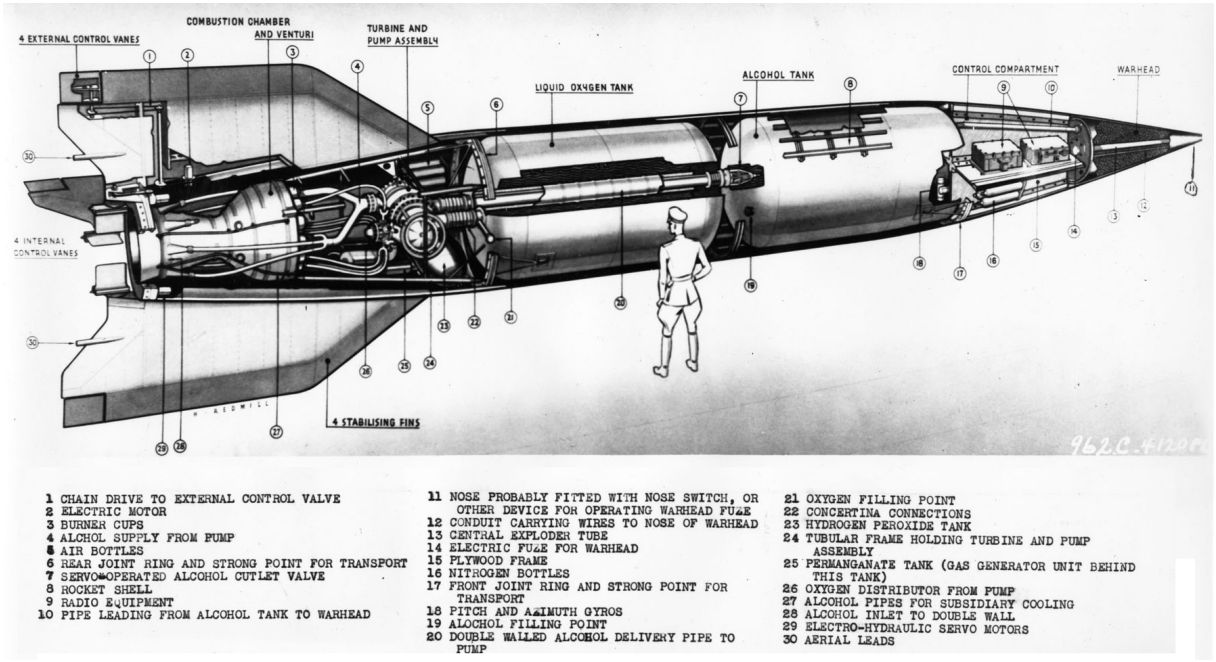


Figure 1.4: German rocket weapon V2. Photo U.S. Air Force

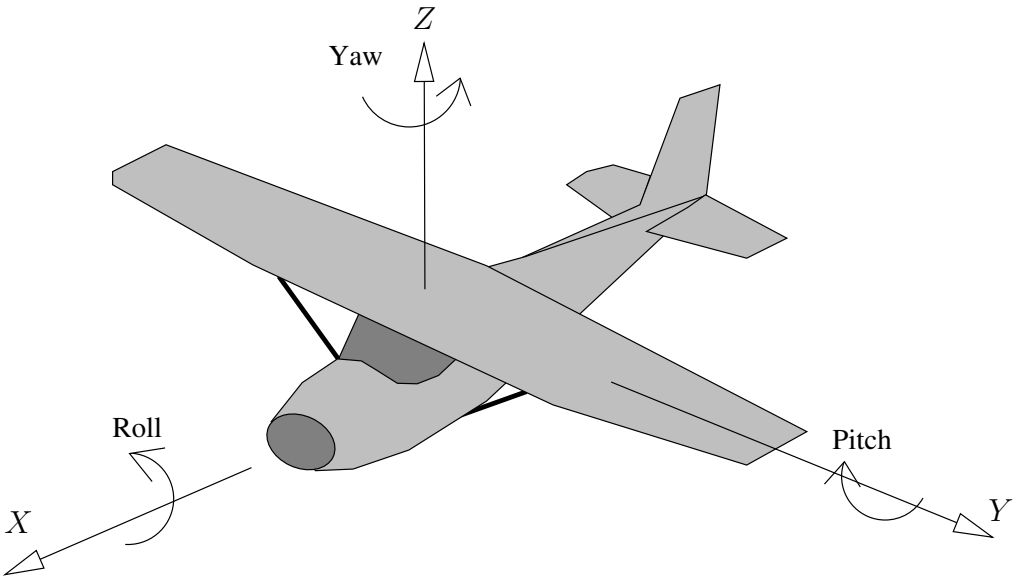


Figure 1.5: The attitude angles of a vehicle

1.4 Technologies

Technologies suitable for both navigation and geodetic position finding are:

1. GPS, Global Positioning System; today we use the term GNSS, Global Navigation Satellite Systems, to which also belong GLONASS (Russia), 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 The real time property

The definition of real time:

Guaranteed latency

Which means a process that has a latency of 1 month can be real time (if 1 month is guaranteed), but another process with a latency of 1 msec is not real time (if the latency is usually less than 1 msec, but it could sometimes be 2 msec, or 10 msec, or even more. . .)

1.6 Basic concepts

- ▷ Stochastic processes
- ▷ Linear estimation
- ▷ Kalman filtering, dynamic model, observation model, statistical model
- ▷ inertial navigation, mechanisation
- ▷ satellite orbit

In the following, these concepts will be discussed systematically.

2 Stochastic processes

2.1 Stochastic quantities

Cf. Strang and Borre [1997] pages 515-541.

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

First the *stochastic quantity* is defined as follows (the underscore is the traditional notation for this):

> A stochastic quantity $\underline{x}$ *is a series of realizations* $x_1, x_2, x_3, \ldots, x_i, \ldots,$ or x_i, $i = 1, \ldots, \infty$.

For example 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.

The value space of the stochastic quantity can be a discrete set (as above) or a continuous set.

A measurement is a stochastic, usually real-valued, quantity.

A measured distance is a real-valued stochastic quantity $\underline{s}$. Realizations $s_i \in \mathbb{R}$.

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

A vector measurement produced by GPS from a point A to a point B is a *stochastic vector quantity* $\underline{\mathbf{x}}$. The realizations belong to the three-dimensional vector space: $\mathbf{x}_i \in \mathbb{R}^3$.

2.2 Stochastic processes

A stochastic process is a stochastic quantity, the value space of which is a *function space*, so each realization of the stochastic quantity ("throwing dice") is a *function*. Most oftenly the function's argument is *time* t.

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

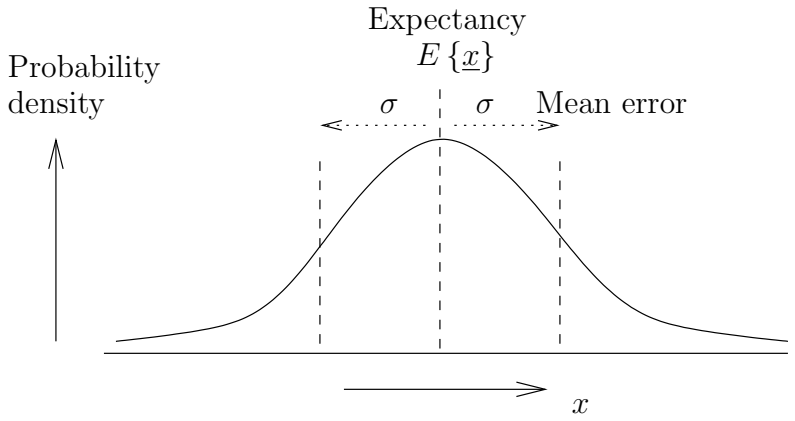


Figure 2.1: The Gaussian bell curve

In real life repeating the test can be difficult or impossible. As an example the temperature of Kaisaniemi in Helsinki $\underline{T}^{\mathrm{Kais}}(t)$. History can not be precisely repeated, so from this stochastic process we only have one realization $T_1^{\mathrm{Kais}}(t)$, the historical time series of Kaisaniemi. Other realizations $T_i^{\mathrm{Kais}}(t)$, $i = 2, \ldots, \infty$ exist only as theoretical constructs without any hope of observing them.

In such cases it is often assumed, that the result will be same if *the same process shifted in time* is used as realization. So for example

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

where Δt is an appropriately chosen time shift, which of course will have to be large enough. This hypothesis is called the *ergodicity hypothesis*.

2.3 On the sample mean

There is often a situation where some quantity x is measured several times and we have available realizations of the stochastic measurement quantity $\underline{x}$, which all of course differ in different ways from the "real" value x – which *we don't 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 known, 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, one realization of which is an estimate.

In the stochastic quantity's value space (domain) x *is defined a probability density function* $p(x)$, that describes the probability, that the value of one realization happens to be x. Often (but not always!) it can be assumed that $p(x)$ is so called *gaussian curve* or normal distribution, the "bell curve".

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

Because x must have some value, we know that the total probability is 1:

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

The definition of the expected value or expectancy E is:

$$E\left\{\underline{x}\right\} \equiv \int_{-\infty}^{+\infty} xp\left(x\right) dx.$$

The expected value is not the same as average; the connection is that the average of $\underline{x}$'s first n realizations,

$$\overline{x}^{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{2.1}$$

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

Above, the first group of n realizations is called *the sample*, *and* $\overline{x}^{(n)}$ *is* the sample average.

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

$$\mathrm{Var}\left(\underline{x}\right) \equiv E\left\{\left(\underline{x} - E\left\{\underline{x}\right\}\right)^2\right\}.$$

The square root of variance is precisely the standard deviation *or mean error* σ, look at the picture above:

$$\sigma^2 = \mathrm{Var}\left(\underline{x}\right).$$

Unfortunately the variance, like the expected value, can not be calculated straightforwardly. Instead it is *estimated* from the sample x_i, $i = 1, \ldots, n$. If the sample average $\overline{\underline{x}}$ already exists, and assuming that the realizations x_i are statistically independent from each other and all have the same mean error σ[1], follows the estimate of the variance σ^2 as follows:

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

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

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

where $\underline{x}_i$ is a stochastic quantity the successive realizations of which are simply $x_i, x_{i+n}, x_{i+2n}, \ldots$ (a "fork variate").

It is intuitively clear – and assumed without proof – that

$$\forall i:\ E\left\{\underline{x}_i\right\} = E\left\{\underline{x}\right\}.$$

[1]This is called the i.i.d. assumption, "independent and identically distributed".

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

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

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

Its variance is next estimated:

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

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 strict proofs, look at a statistics text book.

2.4 Optimality of the average value

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

$$\widehat{x} = \sum_{i=1}^{n} a_i \underline{x}_i,\ \sum_{i=1}^{n} a_i = 1,$$

the average

$$\widehat{x} \equiv \overline{\underline{x}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i \tag{2.2}$$

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

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

assuming, that $\underline{x}_i$ don't correlate with each other, and that $\mathrm{Var}(\underline{x}_i) = \sigma^2$.

Now, minimizing the expression

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

by using the additional constraint

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

yields

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

From which the claim follows.

2.5 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:

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

This is a very simple example of sequential linear filtering, *the Kalman-filter* (chapter 3). Note that, by using this procedure, it is possible to obtain a value for $\underline{\overline{x}}^{(n)}$ "on the fly", while observations are being collected, before all observations are in. This is precisely the advantage of using the Kalman filter.

2.6 Covariance, correlation

When there are two stochastic quantities $\underline{x}$ and $\underline{y}$, *the covarience between them can be calculated as*

$$\mathrm{Cov}\left(\underline{x},\underline{y}\right) \equiv E\left\{\left(\underline{x} - E\left\{\underline{x}\right\}\right)\left(\underline{y} - E\left\{\underline{y}\right\}\right)\right\}.$$

The covariance describes how the random variations of $\underline{x}$ and $\underline{y}$ behave similarly.

Besides covariance, *correlation* is defined as:

$$\mathrm{Corr}\left(\underline{x},\underline{y}\right) \equiv \frac{\mathrm{Cov}\left(\underline{x},\underline{y}\right)}{\sqrt{\mathrm{Var}\left(\underline{x}\right)\mathrm{Var}\left(\underline{y}\right)}}$$

Correlation can never be more than 1.0 (or less than -1.0)[2]. Often the correlation is expressed as a percentage, 100% is the same as 1.0.

[2] Eric Weisstein gives the following proof (http://mathworld.wolfram.com/StatisticalCorrelation.html).

Define *normalized* variates:

$$\underline{\xi} \equiv \frac{\underline{x}}{\sqrt{\mathrm{Var}\left(\underline{x}\right)}}, \underline{\eta} \equiv \frac{\underline{y}}{\sqrt{\mathrm{Var}\left(\underline{y}\right)}}.$$

Then, because of linearity:

$$\mathrm{Cov}\left(\underline{\xi},\underline{\eta}\right) = \frac{\mathrm{Cov}\left(\underline{x},\underline{y}\right)}{\sqrt{\mathrm{Var}\left(\underline{x}\right)\mathrm{Var}\left(\underline{y}\right)}} = \mathrm{Corr}\left(\underline{x},\underline{y}\right).$$

These variances are positive:

$$\begin{aligned} 0 \le \mathrm{Var}\left(\underline{\xi}+\underline{\eta}\right) &= \mathrm{Var}\left(\underline{\xi}\right) + \mathrm{Var}\left(\underline{\eta}\right) + 2\mathrm{Cov}\left(\underline{\xi},\underline{\eta}\right), \\ 0 \le \mathrm{Var}\left(\underline{\xi}-\underline{\eta}\right) &= \mathrm{Var}\left(\underline{\xi}\right) + \mathrm{Var}\left(\underline{\eta}\right) - 2\mathrm{Cov}\left(\underline{\xi},\underline{\eta}\right); \end{aligned}$$

when also

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

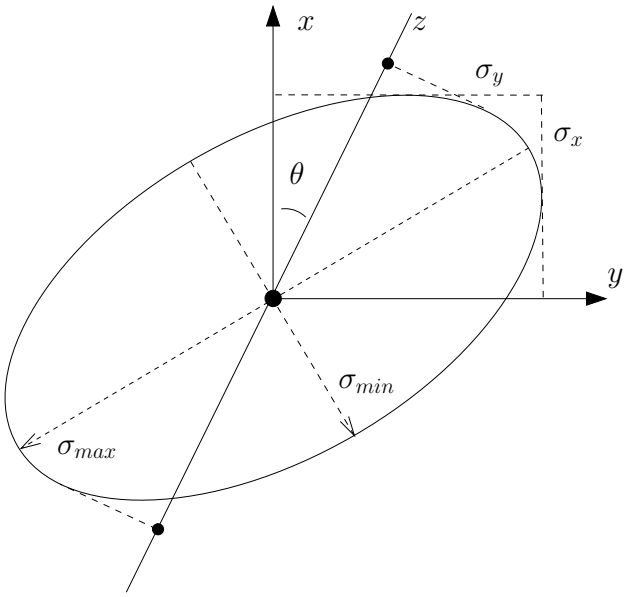


Figure 2.2: Error ellipse

When dealing with two stochastic processes, we often draw an error ellipse (figure 2.2). Compare this picture with the earlier picture of the bell curve. There the expected value is marked as $E\{\underline{x}\}$ (in the middle) and mean error $\pm\sigma$. In the error ellipse picture the central point represents the expected values of $\underline{x}$ and $\underline{y}$ the ellipse itself corresponds to the mean error values $\pm\sigma$. It can be said that the measurement value will probably fall inside the ellipse (that's why the name is error ellipse). If the ellipse is cut by the line z, the linear combination of $\underline{x}$ and $\underline{y}$ is obtained:

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

the point pair of which on the tangent to the ellipse represents precisely the mean error

---

and similarly $\mathrm{Var}\left(\underline{\eta}\right) = 1$, it follows that

$$-1 \leq \mathrm{Cov}\left(\underline{\xi},\underline{\eta}\right) = \mathrm{Corr}\left(\underline{x},\underline{y}\right) \leq 1.$$

of the quantity $\underline{z}$ [3]:

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

and from this $\sigma_z = \sqrt{\mathrm{Var}\left(\underline{z}\right)}$. The mean error σ_z has two extremal values, $\sigma_{\min}$ and $\sigma_{\max}$, look at the picture.

If $\sigma_{\min} = \sigma_{\max}$, or the ellipse is oriented along the axes of the extremal values $\sigma_{\min}$ and $\sigma_{\max}$ the correlation between $\underline{x}$ and $\underline{y}$ 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 of $\underline{x}$'s real value – or a good estimate – helps to estimate the $\underline{y}$ better. This is called *regression*.

2.7 Auto- and crosscovariance of a stochastic process

If instead of a stochastic quantity there is a stochastic *process* $\underline{x}\left(t\right)$, we can calculate the derived function called the *autocovariance*as follows:

$$A_x(t_1,t_2) \equiv \mathrm{Cov}\left(\underline{x}\left(t_1\right), \underline{x}\left(t_2\right)\right).$$

Often, in case of so called stationary processes (in other words, the properties of the process don't depend on absolute time but they are constant), one can write

$$A_x\left(t_1,t_2\right) = A_x\left(t_1, t_2 - t_1\right) \equiv A_x\left(t,\Delta t\right) = A_x(\Delta t) \equiv \mathrm{Cov}\left(\underline{x}\left(t\right), \underline{x}\left(t+\Delta t\right)\right)$$

independent of the value of t.

If there are two stochastic processes $\underline{x}\left(t\right)$ and $\underline{y}\left(t\right)$, one can obtain the derived function called *cross-covariance*.

$$C_{xy}(t_1,t_2) \equiv \mathrm{Cov}\left(\underline{x}\left(t_1\right), \underline{y}\left(t_2\right)\right),$$

[3] In matrix notation we can write

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

ja

$$\mathrm{Var}\begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(\underline{x}) & \mathrm{Cov}(\underline{x},\underline{y}) \\ \mathrm{Cov}(\underline{x},\underline{y}) & \mathrm{Var}(\underline{y}) \end{bmatrix};$$

this implies

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

i.e., the same result. This illustrates the *law of propagation of variances.*.

and again in the case of stationary processes

$$C_{xy}(\Delta t) \equiv \mathrm{Cov}\left(\underline{x}\,(t)\,, \underline{y}\,(t+\Delta t)\right).$$

Often the cross-covariance is called simply

$$C_{xy} \equiv C_{xy}\,(0)\,.$$

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

2.8 "White noise" and "random walk"

Noise is a stochastic process with an expected value of 0:

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

White noise is noise that consists of all possible frequencies. The mathematical way of describing this is saying that the autocovariance

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

In other words, the process values $\underline{n}\,(t_1)$ and $\underline{n}\,(t_2)$ do not correlate at all, no matter how close $t_2 - t_1$ is to zero.

Nevertheless we would have

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

And furthermore it holds that

$$\int_{-\infty}^{+\infty} A_n\,(\tau)\,d\tau = Q.$$

Here we assume all the time stationarity.

Perhaps you may want to stare at the above formulas for a while. Here we have a function $A_n\,(\tau)$ which is "almost everywhere" zero (namely if $\tau \neq 0$) but in the only point where it isn't zero (namely if $\tau = 0$) it is infinite! And furthermore, the integral function over the τ domain produces the finite value Q!

Such a function does not actually exist. It is a matemathical auxiliary device called *distribution*. It is the *delta-function,* named after the quantum physicist Paul Dirac:

$$A_n\,(\tau) = Q\delta\,(\tau)\,. \tag{2.3}$$

Intuitively we can have a mental picture of how such a "function" is built.

First the following block function is defined:

$$\delta_b(\tau) = \begin{cases} 0 & \text{if } \tau > \frac{b}{2} \text{ or } \tau < -\frac{b}{2} \\ \frac{1}{b} & \text{if } -\frac{b}{2} \leq \tau \leq \frac{b}{2} \end{cases}$$

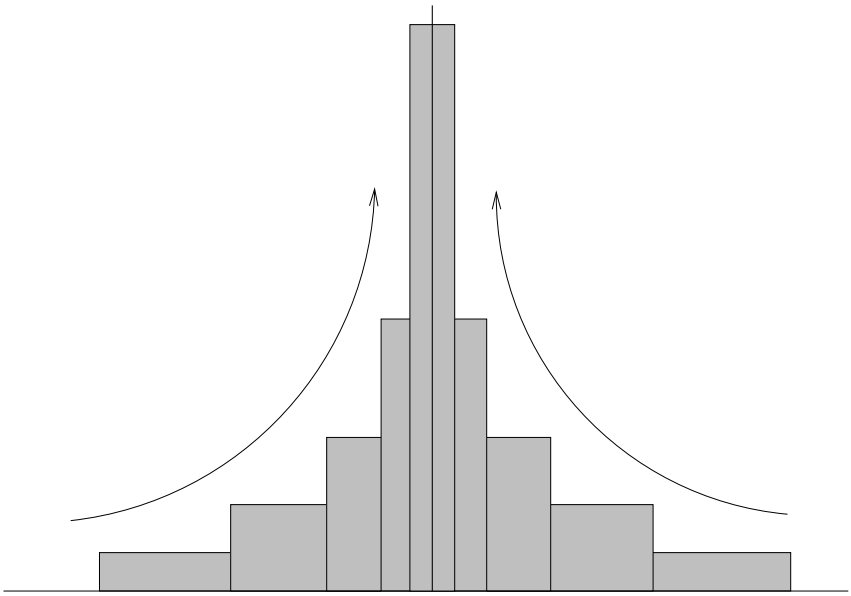


Figure 2.3: The Dirac delta function as the limit of block functions

Obviously the integral of this function

$$\int_{-\infty}^{+\infty} \delta_b(\tau)\, d\tau = 1 \tag{2.4}$$

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

Now let in the limit $b \to 0$. Then $\delta_b(0) \to \infty$, and to every τ value $\tau \neq 0$ there is always a corresponding bounding value for b under which $\delta_b(\tau) = 0$.

The handling rule of distributions is simply, that first we integrate, and then in the result obtained we let $b \to 0$.

"Random walk" is obtained if white noise is integrated over time. Let the autocovariance of the noise $\underline{n}$ be

$$A_n(\Delta t) = Q\delta(\Delta t).$$

Then we integrate this function:

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

Note that

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

The autocovariance function is obtained as:

$$\begin{aligned} A_x(t_1,t_2) = &\quad E\{(\underline{x}(t_2) - E\{\underline{x}(t_2)\})(\underline{x}(t_1) - E\{\underline{x}(t_1)\})\} = \\ = &\quad E\{\underline{x}(t_2)\,\underline{x}(t_1)\} = \\ = &\quad E\left\{\int_{t_0}^{t_2} \underline{n}(\tau_2)\, d\tau_2 \int_{t_0}^{t_1} \underline{n}(\tau_1)\, d\tau_1\right\} = \\ = &\quad \int_{t_0}^{t_2} \left[\int_{t_0}^{t_1} E\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\}\, d\tau_1\right] d\tau_2. \end{aligned}$$

Here

$$\begin{aligned}
&\int_{t_0}^{t_1} E\left\{\underline{n}\left(\tau_1\right)\underline{n}\left(\tau_2\right)\right\} d\tau_1 = \\
&= \int_{t_0}^{t_1} A_n\left(\tau_2-\tau_1\right) d\tau_1 = \\
&= Q\int_{t_0}^{t_1} \delta\left(\tau_2-\tau_1\right) d\tau_1 = \begin{cases} Q & \text{if} \quad t_1 > \tau_2 \\ 0 & \text{if} \quad t_1 < \tau_2 \end{cases}
\end{aligned}$$

From this it follows that

$$\begin{aligned}
A_x\left(t_1,t_2\right) &= Q\int_{t_0}^{t_2}\left[\int_{t_0}^{t_1}\delta\left(\tau_2-\tau_1\right)d\tau_1\right]d\tau_2 = \\
&= Q\left(t_1-t_0\right)+0.\left(t_2-t_1\right) = \\
&= Q\left(t_1-t_0\right). \qquad (2.5)
\end{aligned}$$

In this derivation it has been assumed that the autocovariance of the noise function $\underline{n}$ is *stationary*, in other words, that Q is a constant. This can easily be generalized to the case where $Q\left(t\right)$ is a function of time:

$$A_x\left(t_1,t_2\right) = \int_{t_0}^{t_1} Q\left(t\right) dt. \qquad (2.6)$$

In both equations (2.5, 2.6) it is assumed that $t_1 \leq t_2$.

2.9 Power Spectral Density

Definition

We may also want to study these 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). As follows [4]:

$$\widetilde{A_x}\left(f\right) = \mathcal{F}\left\{A_x\left(t\right)\right\} = \int_{-\infty}^{+\infty} A_x\left(t\right)\exp\left(-2\pi i f t\right) dt, \qquad (2.7)$$

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

$$\widetilde{C_{xy}}\left(f\right) = \mathcal{F}\left\{C_{xy}\left(t\right)\right\} = \int_{-\infty}^{+\infty} C_{xy}\left(t\right)\exp\left(-2\pi i f t\right) dt.$$

[4] Note that we write here t for the time argument, which however represents a time *difference*. Earlier we used Δt.

The inverse operation using the inverse Fourier transform yields

$$A_x(t) = \mathcal{F}^{-1}\left\{\widetilde{A_x}(f)\right\} = \int_{-\infty}^{+\infty} \widetilde{A_x}(f) \exp(2\pi i f t)\, df.$$

Therefore, for $t = 0$ we obtain

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

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

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

$$A_x(\Delta t) = A_x(t_2 - t_1) = A_x(t_2, t_1) = A_x(t_1, t_2) = A_x(t_1 - t_2) = A_x(-\Delta t),$$

it follows that the PSD is always *real valued*; additionally it is always non-negative,

$$\widetilde{A_x}(f) \geq 0 \quad \forall f.$$

For cross-PSDs this does not hold: we have

$$C_{xy}(t_2, t_1) = C_{yx}(t_1, t_2) \neq C_{xy}(t_1, t_2),$$

as opposed to

$$\begin{aligned} A_x(t_2, t_1) &= E\left\{\left(x(t_2) - E\left\{x(t_2)\right\}\right)\left(x(t_1) - E\left\{x(t_1)\right\}\right)\right\} = \\ &= E\left\{\left(x(t_1) - E\left\{x(t_1)\right\}\right)\left(x(t_2) - E\left\{x(t_2)\right\}\right)\right\} = A_x(t_1, t_2). \end{aligned}$$

.

White noise

The PSD of white noise may be computed as follows using the expression (2.3):

$$A_n(t) = Q\delta(t),$$

from which

$$\widetilde{A_n}(f) = \int_{-\infty}^{+\infty} Q\delta(t) \exp(-2\pi i t f)\, dt = Q \exp(0) = Q \;\forall f,$$

using the δ function's integration property (2.4). Here we see why a process with a Dirac δ type autocovariance function is called *white* noise: the power spectral density is a constant all over the spectrum, for all frequencies f, just like is the case for white light.

3 The Kalman filter

Cf. Strang and Borre [1997] pages 543-583.

Link list: http://www.cs.unc.edu/~welch/kalman/.

A good slideshow: http://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_Slides_08.pdf.

The Kalman filter is a linear, predictive filter. Like a coffee filter filters coffee from coffee-grounds, the Kalman filter filters the signal (the *state vector*) from the noise of the observation 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. Nevertheless the Kalman filter is generally applicable and already used not only in navigation but also in economics, meteorology and so on.

The Kalman filter consists of two parts:

1. The *dynamic model*; it describes the process of motion, according to which the state vector evolves over time.

2. The *observation model*; it describes the observational quantities that tell something about the state vector at the time of observation.

Both of these models contain statistics: the dynamic model contains statistics describing the non-determinacy of the development of the system described, e.g., the random perturbations of a satellite orbit, while the observational model contains a description of observational uncertainty.

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

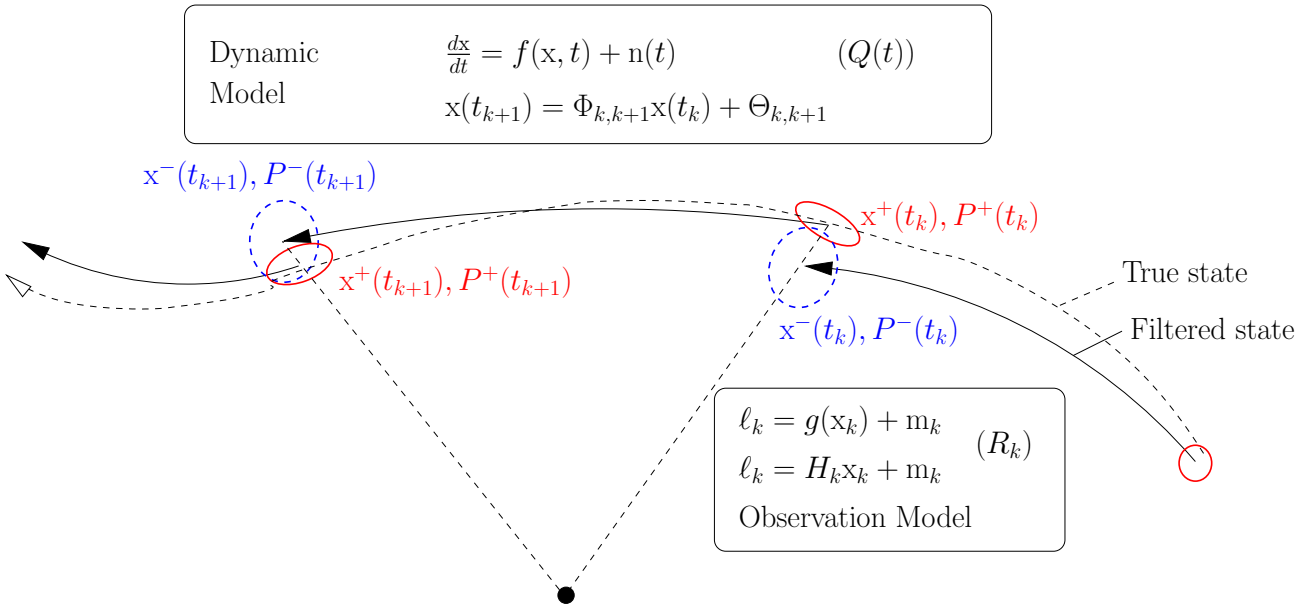


Figure 3.1: The Kalman filter

3.1 The state vector

The state vector is a formal vector (element of an abstract vector space) that describes completely the state of a dynamic system. E.g. a particle moving freely in space has three position co-ordinates and three velocity components; the state vector becomes

$$\underline{x} = \begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix}, \tag{3.1}$$

where the position vector is $\begin{bmatrix} x & y & z \end{bmatrix}^T$ and the velocity vector $\begin{bmatrix} \dot{x} & \dot{y} & \dot{z} \end{bmatrix}^T$[1]. In this case the state vector has six elements or *degrees of freedom*.

If the particle is not a point but an extended object, also its orientation angles (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 positions and three velocities, to the state vector.

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

[1] Alternative notation: place vector $x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$, velocity vector $\dot{x}\mathbf{e}_1 + \dot{y}\mathbf{e}_2 + \dot{z}\mathbf{e}_3$, where $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is an orthonormal base in $\mathbb{R}^3$.

3.2 The dynamic model

The dynamic model characterizes the state vector's behaviour in time. The state vector is a (vectorial, i.e. vector valued) stochastic process as a function of time t .

The dynamic model in the *linear case* looks like:

$$\frac{d}{dt}\underline{\mathrm{x}} = \Phi \cdot \underline{\mathrm{x}} + \underline{\mathrm{n}}, \tag{3.2}$$

where $\underline{\mathrm{x}} = \underline{x}(t)$ is the state vector, $\underline{\mathrm{n}} = \underline{\mathrm{n}}(t)$ is the dynamic noise (in other words, how inaccurately the equations of motion above actually apply) and Φ (also possibly dependent on time) is the coefficient matrix.

The more general *non-linear* case is:

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

where $F(\cdot)$ is a (vectorial) function. The linear case is easily obtained from this by choosing an *approximate value* $\mathrm{x}^{(0)}$ for the state vector. We demand from this approximate value (also a function of time!) *consistency* with the functional model:

$$\frac{d}{dt}\mathrm{x}^{(0)} = F\left(\mathrm{x}^{(0)}\right).$$

Now we *linearize* by subtraction and Taylor expansion:

$$\frac{d}{dt}\left(\underline{\mathrm{x}} - \mathrm{x}^{(0)}\right) = F(\underline{\mathrm{x}}) + \underline{\mathrm{n}} - F\left(\mathrm{x}^{(0)}\right) \approx \Phi \cdot \left(\underline{\mathrm{x}} - \mathrm{x}^{(0)}\right) + \underline{\mathrm{n}},$$

which already is of the form (3.2) if we write $\underline{\mathrm{x}} - \mathrm{x}^{(0)} \to \Delta\underline{\mathrm{x}}$:

$$\frac{d}{dt}\Delta\underline{\mathrm{x}} = \Phi \cdot \Delta\underline{\mathrm{x}} + \underline{\mathrm{n}},$$

from which one may drop the deltas.

The elements of the function $F(\cdot)$'s Jakobi matrix F used above are $\Phi_{ij} = \dfrac{\partial}{\partial x_j}F_i(\mathrm{x})$, where the x_j are the components of x: e.g., for the example state vector given in 3.1, $x_2 = y$, $x_6 = \dot{z}$, etc.

Realistic *statistical attributes* have to be given to the dynamic noise; often it is assumed that it is white noise (cf. above), the autocovariance of which is

$$A_{\mathrm{n}}(t_1, t_2) = Q(t_1)\,\delta(t_2 - t_1). \tag{3.3}$$

3.3 Example: a Kepler orbit

As an example the motion of a spacecraft in the Earth's gravitational field:

$$\frac{d^2}{dt^2}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = -\frac{GM}{(x^2+y^2+z^2)^{\frac{3}{2}}}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix},$$

where n_x, n_y, n_z are, e.g., the unknown effect of air drag or the irregularities of the Earth's gravitational field, etc.

Unfortunately this is a second order differential equation. The state vector is extended by adding *the velocities* to it:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \\ \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ -\frac{GM}{(x^2+y^2+z^2)^{\frac{3}{2}}}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ n_x \\ n_y \\ n_z \end{bmatrix}.$$

This system of equations is non-linear. Linearizing yields

$$\frac{d}{dt}\left[\begin{array}{c} \Delta x \\ \Delta y \\ \Delta z \\ \hline \Delta\dot{x} \\ \Delta\dot{y} \\ \Delta\dot{z} \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline GM\frac{3x^2-r^2}{r^5} & GM\frac{3xy}{r^5} & GM\frac{3xz}{r^5} & 0 & 0 & 0 \\ GM\frac{3yx}{r^5} & GM\frac{3y^2-r^2}{r^5} & GM\frac{3yz}{r^5} & 0 & 0 & 0 \\ GM\frac{3zx}{r^5} & GM\frac{3zy}{r^5} & GM\frac{3z^2-r^2}{r^5} & 0 & 0 & 0 \end{array}\right]\left[\begin{array}{c} \Delta x \\ \Delta y \\ \Delta z \\ \hline \Delta\dot{x} \\ \Delta\dot{y} \\ \Delta\dot{z} \end{array}\right] + \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline n_x \\ n_y \\ n_z \end{array}\right], \tag{3.4}$$

where $r = \sqrt{x^2+y^2+z^2}$ is the distance from the Earth's entre. It is also assumed that

1. there is a proper set of approximate values $\left[\begin{array}{ccc|ccc} x^{(0)} & y^{(0)} & z^{(0)} & \dot{x}^{(0)} & \dot{y}^{(0)} & \dot{z}^{(0)} \end{array}\right]^T$, relative to which the Δ quantities have been calculated, and that
2. *the elements of the coefficient matrix are evaluated using those approximate values.*

Each element in the state vector is a function of time: $x^{(0)}(t)$ etc.

The "partitioned" version of the formula above would be:

$$\frac{d}{dt}\begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{v} \end{bmatrix} = \begin{bmatrix} 0 & I \\ M & 0 \end{bmatrix}\begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{v} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{n} \end{bmatrix},$$

where

$$\begin{aligned} M &= \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{bmatrix}\begin{bmatrix} -\frac{GM}{r^3}x \\ -\frac{GM}{r^3}y \\ -\frac{GM}{r^3}z \end{bmatrix} = \\ &= \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{bmatrix}\begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{bmatrix}\frac{GM}{r} = \\ &= \begin{bmatrix} \frac{\partial^2}{\partial x^2} & \frac{\partial^2}{\partial x\partial y} & \frac{\partial^2}{\partial x\partial z} \\ \frac{\partial^2}{\partial y\partial x} & \frac{\partial^2}{\partial y^2} & \frac{\partial^2}{\partial y\partial z} \\ \frac{\partial^2}{\partial z\partial x} & \frac{\partial^2}{\partial z\partial y} & \frac{\partial^2}{\partial z^2} \end{bmatrix}\frac{GM}{r} = \\ &= \frac{GM}{r^5}\begin{bmatrix} 3x^2-r^2 & 3xy & 3xz \\ 3yx & 3y^2-r^2 & 3yz \\ 3zx & 3zy & 3z^2-r^2 \end{bmatrix} \end{aligned} \tag{3.5}$$

is called *the gravitational gradient tensor* also known as the *Marussi tensor*.

The Marussi tensor is the partial derivatives matrix of the gravitation vector $\frac{GM}{r^3}\mathbf{x}$ with respect to place. Remembering that the gravitation vector is the *gradient of the geopotential*, it follows that the tensor is also the second partial derivatives matrix of the geopotential $\frac{GM}{r}$ with respect to place. All these formulas assume a central gravitational field.

The gravitational gradient tensor describes how a small perturbation in the satellite's location $\begin{bmatrix} \Delta x & \Delta y & \Delta z \end{bmatrix}^T$ converts into an acceleration perturbance $\frac{d}{dt}\begin{bmatrix} \Delta \dot{x} & \Delta \dot{y} & \Delta \dot{z} \end{bmatrix}^T = \begin{bmatrix} \Delta \ddot{x} & \Delta \ddot{y} & \Delta \ddot{z} \end{bmatrix}^T$.

The most important thing when choosing the set of approximate values is, that it be *physically consistent*, in other words it describes a really possible orbital motion inside the assumed gravitational field.

In the case of a central gravitational field, a suitable set of approximate values is the Kepler orbit, or, more simply, a constant circular motion. In the formula above, the set of approximate values chosen can be precisely those of Kepler orbital motion around the centre of the attractive force GM.

Now, if the we have a gravitational field model that is more accurate than the central field approximation, one must integrate the approximate values using this more accurate field model. Nevertheless the above linearized dymamic model Eq. (3.4) will still be good for integrating the difference quantities $\Delta\mathbf{x}, \Delta\mathbf{v}$, as long as these are *numerically small*. This is one of the benefits of linearization.

3.4 State propagation

State propagation is done by integrating the formula (3.2). More precisely, the formula integrated (again in the linear case) is

$$\frac{d}{dt}\underline{\mathrm{x}}(t) = \Phi \cdot \underline{\mathrm{x}}(t)\,.$$

In the case of the state estimator $\underline{\mathrm{x}}^-$ [2] this is simple:

$$\begin{aligned} \underline{\mathrm{x}}^-(t_1) &\approx \underline{\mathrm{x}}^-(t_0) + \Phi\Delta t \cdot \underline{\mathrm{x}}^-(t_0) = \\ &= (I + \Phi\Delta t)\,\underline{\mathrm{x}}^-(t_0) \end{aligned}$$

if $\Delta t = t_1 - t_0$ is small. As we can immediately see, the elements of $\underline{\mathrm{x}}(t_1)$ are *linear combinations* of the elements of $\underline{\mathrm{x}}(t_0)$. If $t_1 - t_0 = n\delta t$, δt small, it follows, by repeated application of the above formula, that

$$\underline{\mathrm{x}}^-(t_1) = (I + \Phi\delta t)^n\,\underline{\mathrm{x}}^-(t_0)\,.$$

[2] Notation used: $\underline{\mathrm{x}}^-$ is the state estimator before the (later to be described) update step; $\underline{\mathrm{x}}^+$ is the state estimator after this step. In the literature, also the notations $\widehat{\mathrm{x}}^{i-1}$ and $\widehat{\mathrm{x}}^i$, where the "hat" is the mark of the estimator, can be found.

The matrix

$$\Phi_0^1 = (I + \Phi\delta t)^n$$

is called the state transition matrix between epochs t_0 and t_1 ; we can write

$$\underline{\mathrm{x}}^-(t_1) = \Phi_0^1 \underline{\mathrm{x}}^-(t_0).$$

If we write $\delta t = \Delta t/n$, we obtain

$$\Phi_0^1 \;=\; \left(I + \frac{\Phi\Delta t}{n}\right)^n.$$

For simple numbers, we have the classical formula

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

where we see the definition of the number e:

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

For this reason we write sometimes (generalizing the exp function to square matrices):

$$\Phi_0^1 = \exp\left\{\ln\left(I+\frac{\Phi\Delta t}{n}\right)^n\right\} = \exp\left\{n\,\ln\left(I+\frac{\Phi\Delta t}{n}\right)\right\} \approx \exp\left\{n\frac{\Phi\Delta t}{n}\right\} = e^{\Phi(t_1-t_0)}. \tag{3.6}$$

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

$$\Phi_{t_0}^{t_2} = \Phi_{t_1}^{t_2}\cdot\Phi_{t_0}^{t_1},$$

in other words, to transition the state from $\underline{x}(t_0)$ to $\underline{x}(t_2)$, you may transition first from t_0 to t_1 and then from t_1 to t_2.

Definition. We define as the *state variance* the square difference of its *estimator* from its *true value* – itself a stochastic process to which of course we do not have access[3]! – as follows:

$$P^-(t) = \mathrm{Var}\left(\underline{\mathrm{x}}^-(t)\right) \equiv E\left\{\left(\underline{\mathrm{x}}^-(t) - \underline{\mathrm{x}}(t)\right)\left(\underline{\mathrm{x}}^-(t) - \underline{\mathrm{x}}(t)\right)^T\right\}, \tag{3.7}$$

then

$$P^-(t_1) = \left(\Phi_0^1\right)P^-(t_0)\left(\Phi_0^1\right)^T + \int_{t_0}^{t_1} Q(t)\,dt, \tag{3.8}$$

where we have used the formula (2.6), and assumed that the dynamic noise $\underline{\mathrm{n}}$ is white.

[3] An interesting philosophical issue. Does the sound of the wind in the trees exist when no-one is listening?

We may also derive differential equations that describe the development of the state variance matrix and state transition matrix in time. If the times t_0 and t are close to each other, we may write

$$\Phi_{t_0}^{t} \approx I + F(t)(t - t_0),$$

where *now* the coefficient matrix $F(t)$ is allowed to be a function of time. Then

$$\frac{d}{dt}\left(\Phi_{t_0}^{t}\right) = F(t) + \frac{dF}{dt}(t - t_0) \approx F(t).$$

Let us now consider the situation where t and t_1 are close together, but t_1 and $t_0$far apart. Then we have

$$\frac{d}{dt}\left(\Phi_{t_0}^{t}\right) = \frac{d}{dt}\left(\Phi_{t_1}^{t}\Phi_{t_0}^{t_1}\right) = \frac{d}{dt}\left(\Phi_{t_1}^{t}\right)\Phi_{t_0}^{t_1} \approx F(t)\Phi_{t_0}^{t_1} \approx F(t)\Phi_{t_0}^{t}. \tag{3.9}$$

With the initial condition

$$\Phi_{t_0}^{t_0} = I$$

we can by numerical integration obtain the matrix $\Phi_{t_0}^{t_1}$. We can also write (without proof), in full analogy with eq. (3.6):

$$\Phi_{t_0}^{t_1} = \exp\left\{\int_{t_0}^{t_1} F(t)\,dt\right\},$$

which is also handy for calculation.

This is the more general case of (3.6) in the case where F depends on time. (The notation $\Phi_{t_0}^{t_1} = \Phi_0^1$ differs a little from that used earlier.)

In order to derive a differential equation for the state variance matrix P we start from equation (3.8):

$$P^{-}(t) = \left(\Phi_{t_0}^{t}\right)P^{-}(t_0)\left(\Phi_{t_0}^{t}\right)^{T} + \int_{t_0}^{t} Q(\tau)\,d\tau,$$

where we have substituted $t \to \tau$ ja $t_1 \to t$. In case $t - t_0$ is small, the result is using formula (3.9)

$$\begin{aligned}
\frac{d}{dt}P^{-}(t) &= \left(\frac{d}{dt}\Phi_{t_0}^{t}\right)P^{-}(t_0)\left(\Phi_{t_0}^{t}\right)^{T} + \left(\Phi_{t_0}^{t}\right)P^{-}(t_0)\left(\frac{d}{dt}\Phi_{t_0}^{t}\right)^{T} + Q(t) = \\
&= F(t)\Phi_{t_0}^{t}P^{-}(t_0)\left(\Phi_{t_0}^{t}\right)^{T} + \Phi_{t_0}^{t}P^{-}(t_0)\left(\Phi_{t_0}^{t}\right)^{T}F^{T}(t) + Q(t) = \\
&= F(t)P_0^{-}(t) + P_0^{-}(t)F^{T}(t) + Q(t),
\end{aligned} \tag{3.10}$$

in which $P_0^{-}(t)$:

$$P_0^{-}(t) = \left(\Phi_{t_0}^{t}\right)P^{-}(t_0)\left(\Phi_{t_0}^{t}\right)^{T},$$

is computed by integrating the differential equation

$$\frac{d}{dt}P_0^{-}(t) = F(t)P_0^{-}(t) + P_0^{-}(t)F^{T}(t). \tag{3.11}$$

The equation (3.10) is suitable for integrating the matrix P also in the case where F is time dependent.

All this however *assumes* that the matrix F *exists,* i.e., the function $F(\underline{x})$ *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 (*linear case*) is:

$$\underline{\ell} = H \cdot \underline{\mathrm{x}} + \underline{\mathrm{m}},$$

where $\underline{\ell}$ is the observation quantity (vector), $\underline{\mathrm{x}}$ is the state vector ("the real value") and $\underline{\mathrm{m}}$ is the "noise", i.e., the uncertainty, of the observation process. H is the *observation matrix*[4]. As the variance of the noise is given the variance matrix R; $E\{\underline{\mathrm{m}}\} = 0$ and $E\left\{\underline{\mathrm{m}}\ \underline{\mathrm{m}}^{\mathrm{T}}\right\} = R$ (as $E\{\underline{\mathrm{m}}\} = 0$, this is noise after all).

Let the observation moment be t; *the estimator* of the state vector propagated to this moment is[5] $\underline{\mathrm{x}}^{-}(t) = \underline{\mathrm{x}}^{-}$. From this value one can now calculate the observation quantity as:

$$\widehat{\ell} = H\underline{\mathrm{x}}^{-}.$$

Now a the zero quantity (a quantity the expected value $E\{\cdot\}$ is zero) is constructed as:

$$\begin{aligned} \underline{\mathrm{y}} &= \widehat{\ell} - \underline{\ell} = \\ &= H\left(\underline{\mathrm{x}}^{-} - \underline{\mathrm{x}}\right) - \underline{\mathrm{m}} \end{aligned}$$

and thus

$$\begin{aligned} E\left\{\underline{\mathrm{y}}\right\} &= H\left(E\left\{\underline{\mathrm{x}}^{-}\right\} - E\left\{\underline{\mathrm{x}}\right\}\right) - E\left\{\underline{\mathrm{m}}\right\} = \\ &= H \cdot 0 - 0, \end{aligned}$$

by using the assumption $E\{\underline{\mathrm{x}}^{-}\} = \underline{\mathrm{x}}$, i.e., $\underline{\mathrm{x}}^{-}$ is an *unbiased estimator* of $\underline{\mathrm{x}}$.

The nonlinear case: Then, H is not a matrix but a function $H(\underline{\mathrm{x}})$ of the state vector. We write

$$\underline{\ell} = H(\underline{\mathrm{x}}) + \underline{m}$$

and

$$\widehat{\ell} = H\left(\underline{\mathrm{x}}^{-}\right),$$

after which

$$\underline{y} = \widehat{\ell} - \underline{\ell} = H \cdot \left(\underline{\mathrm{x}}^{-} - \mathrm{x}\right) - \underline{m},$$

and the elements of the matrix H are defined by

$$H_{ij} = \frac{\partial}{\partial x_j} H_i(\mathrm{x}),$$

the *Jacobian matrix* (matrix of partial derivatives) of the function $H(\mathrm{x})$.

[4] This is the same as in the case of least squares adjustment the A matrix or "design matrix".

[5] The minus or plus sign used as a superscript is an often used notation to denote the state "before" and "after" (*a priori, a posteriori*) the use of an observation in the update step. Other notations are found as well, e.g., the subscripts i and $i+1$.

Let us also calculate

$$\begin{aligned} \mathrm{Var}\left(\underline{\mathrm{y}}\right) &= E\left\{\underline{\mathrm{y}}\underline{\mathrm{y}}^T\right\} = \\ &= HE\left\{\left(\underline{\mathrm{x}}^- - \underline{\mathrm{x}}\right)\left(\underline{\mathrm{x}}^- - \underline{\mathrm{x}}\right)^T\right\} H^T + R = \\ &= HP^-H^T + R, \end{aligned}$$

while assuming that $\underline{\mathrm{x}}^-$ and $\underline{\mathrm{m}}$ do not correlate with each other.

Also

$$\begin{aligned} \mathrm{Cov}\left(\underline{\mathrm{y}}, \underline{\mathrm{x}}^-\right) &\equiv E\left\{\underline{\mathrm{y}}\left(\underline{\mathrm{x}}^- - \underline{\mathrm{x}}\right)^T\right\} = \\ &= HP^-, \end{aligned}$$

by assuming that $\underline{\mathrm{m}}$ and $\underline{\mathrm{x}}^-$ – and $\underline{\mathrm{x}}$ – do not correlate (logical assumption; usually the observation process is physically completely independent from the orbital motion process, an the observation processes at different epoch are independent of each other)

Also with

$$\mathrm{Cov}\left(\underline{\mathrm{x}}^-, \underline{\mathrm{y}}\right) = P^-H^T.$$

3.6 Updating

The update step is now exploiting optimally the fact that the difference between the observation quantity's value $\widehat{\ell}$ calculated from the estimated state vector $\underline{\mathrm{x}}^-$ and the really observed observation quantity $\underline{\ell}$ has an expected value of zero.

So an enhanced estimator is constructed

$$\begin{aligned} \underline{\mathrm{x}}^+ &= \underline{\mathrm{x}}^- + K\underline{\mathrm{y}} = \\ &= \underline{\mathrm{x}}^- + K\left(H\left(\underline{\mathrm{x}}^- - \underline{\mathrm{x}}\right) + \underline{\mathrm{m}}\right), \end{aligned}$$

so

$$\left(\underline{\mathrm{x}}^+ - \underline{\mathrm{x}}\right) = \left(I + KH\right)\left(\underline{\mathrm{x}}^- - \underline{\mathrm{x}}\right) + K\underline{\mathrm{m}}.$$

Here the matrix K is called the Kalman "gain matrix".

Now according to the definition (3.7) we may use this to derive the propagation equation for the state variance:

$$P^+ = \left(I + KH\right)P^-\left(I + KH\right)^T + KRK^T. \tag{3.12}$$

"The optimal" solution is obtained by choosing

$$K = -P^-H^T\left(HP^-H^T + R\right)^{-1},$$

which gives as a solution the state propagation equation

$$\underline{\mathrm{x}}^{+} = \underline{\mathrm{x}}^{-} - P^{-}H^{T}\left(HP^{-}H^{T} + R\right)^{-1}\left(H\underline{\mathrm{x}}^{-} - \underline{\ell}\right).$$

if we call

$$\Pi \equiv \left(HP^{-}H^{T} + R\right)^{-1},$$

we can re-write eq. (3.12):

$$\begin{aligned} P^{+} &= \left(I - P^{-}H^{T}\Pi H\right)P^{-}\left(I - P^{-}H^{T}\Pi H\right)^{T} + P^{-}H^{T}\Pi R\Pi HP^{-} = \\ &= P^{-} - P^{-}H^{T}\Pi HP^{-} - P^{-}H^{T}\Pi HP^{-} + \\ &\quad + P^{-}H^{T}\Pi HP^{-}H^{T}\Pi HP^{-} + P^{-}H^{T}\Pi R\Pi HP^{-} = \\ &= P^{-} - 2P^{-}H^{T}\Pi HP^{-} + P^{-}H^{T}\Pi HP^{-} = \\ &= P^{-} - P^{-}H^{T}\Pi HP^{-} = P^{-} - P^{-}H^{T}\left(HP^{-}H^{T} + R\right)^{-1}HP^{-}. \end{aligned}$$

Perhaps more intuitively summarized:

$$\underline{\mathrm{x}}^{+} = \underline{\mathrm{x}}^{-} - \mathrm{Cov}\left(\underline{\mathrm{x}}^{-}, \underline{\mathrm{y}}\right)\mathrm{Var}^{-1}\left(\underline{\mathrm{y}}\right)\underline{\mathrm{y}}, \tag{3.13}$$

$$\mathrm{Var}\left(\underline{\mathrm{x}}^{+}\right) = \mathrm{Var}\left(\underline{\mathrm{x}}^{-}\right) - \mathrm{Cov}\left(\underline{\mathrm{x}}^{-}, \underline{\mathrm{y}}\right)\mathrm{Var}^{-1}\left(\underline{\mathrm{y}}\right)\mathrm{Cov}\left(\underline{\mathrm{y}}, \underline{\mathrm{x}}^{-}\right), \tag{3.14}$$

some kind of regression of the state vector $\underline{\mathrm{x}}$ with respect to the "closing error" $\underline{\mathrm{y}}$.

So the updating formulas for the Kalman-filter have been found for both the state vector and its variance matrix.

Remark. We may still shorten the variance update equation as follows:

$$P^{+} = P^{-} - P^{-}H^{T}\left(HP^{-}H^{T} + R\right)^{-1}HP^{-} = \left(I + KH\right)P^{-},$$

based on the definition of K.

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

$$\mathrm{Var}\left(\underline{\mathrm{y}}\right) = HPH^{T} + R$$

is the size of vector $\underline{\mathrm{y}}$. And $\underline{\mathrm{y}}$'s size is the amount 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 Kalman-filter

The formulas (3.13, 3.14) 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

$$\mathrm{Cov}\left(\underline{\mathrm{x}}^{+},\underline{\mathrm{y}}\right)=\mathrm{Cov}\left(\underline{\mathrm{x}}^{-},\underline{\mathrm{y}}\right)-\mathrm{Cov}\left(\underline{\mathrm{x}}^{-},\underline{\mathrm{y}}\right)\mathrm{Var}^{-1}\left(\underline{\mathrm{y}}\right)\mathrm{Var}\left(\underline{\mathrm{y}}\right)=0 \tag{3.15}$$

(remember that $\mathrm{Cov}\left(\underline{\mathrm{y}},\underline{\mathrm{y}}\right)=\mathrm{Var}\left(\underline{\mathrm{y}}\right)$). So the updated state vector $\underline{\mathrm{x}}^{+}$ *is orthogonal* to the "closing error vector" $\underline{\mathrm{y}}$.

Assume now that there was an alternative $\underline{\mathrm{x}}^{\times}$, that was even better than $\underline{\mathrm{x}}^{+}$Write

$$\underline{\mathrm{x}}^{\times}=\underline{\mathrm{x}}^{+}+A\underline{\mathrm{y}}.$$

Then, because of the formula (3.15) we would have

$$\mathrm{Var}\left(\underline{\mathrm{x}}^{\times}\right)=\mathrm{Var}\left(\underline{\mathrm{x}}^{+}\right)+A\mathrm{Var}\left(\underline{\mathrm{y}}\right)A^{T}.$$

So because $\mathrm{Var}\left(\underline{\mathrm{y}}\right)$ is positive-definite,

$$\mathrm{Var}\left(\underline{\mathrm{x}}^{\times}\right)-\mathrm{Var}\left(\underline{\mathrm{x}}^{+}\right)$$

is always positive-semidefinite, and

$$\mathrm{Var}\left(\underline{\mathrm{x}}^{\times}\right)-\mathrm{Var}\left(\underline{\mathrm{x}}^{+}\right)=0$$

happens if $A=0$. In other words, for an arbitrary linear combination $\underline{z}=\sum_i c_i\underline{x}_i$ (so $\underline{z}^{\times}=c_i\underline{x}_i^{\times}$, $\underline{z}^{+}=c_i\underline{x}_i^{+}$) it holds that

$$\mathrm{Var}\left(\underline{z}^{\times}\right)-\mathrm{Var}\left(\underline{z}^{+}\right)=0$$

if $A=0$, and otherwise we have

$$\mathrm{Var}\left(\underline{z}^{\times}\right)-\mathrm{Var}\left(\underline{z}^{+}\right)\geq 0.$$

The issue can be represented in the two-dimensional special case graphically like in figure 3.2. So, the variance ellipse of the optimal estimator $\underline{\mathrm{x}}^{+}$ (more generally a (hyper-) ellipsoid) is always entirely inside (or at worst, touching from the inside) the variance ellipse of the alternative estimator $\underline{\mathrm{x}}^{\times}$, and the same holds also for the variances of an arbitrary linear combination $\underline{\ell}$ of the components.

3.8 An example computation

Question:

Assume the dynamical model for the state vector

$$\underline{x}=\begin{bmatrix}\underline{x}\\ \underline{v}\end{bmatrix}$$

to be

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

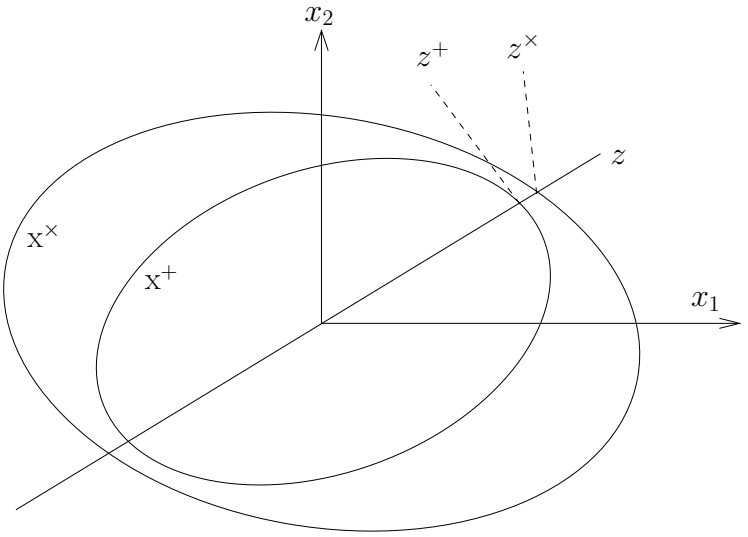


Figure 3.2: The error ellipse of the optimal estimator is completely surrounded by the error ellipses of other estimators

Here, $\underline{n}$ is white noise with an autocovariance of $Q = 1$. Furthermore, assume the initial state to be given as

$$\begin{bmatrix} x(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \; P(0) = \begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix}$$

(i.e., no real velocity information is actually given).

(Use *Matlab!*)

1. Propagate this state information forward to $t = 5$, i.e., calculate $\mathrm{x}(5)$, $P(5)$.
2. At $t = 5$, a further observation, value: 3, is made:

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

   where the variance of $\underline{m}$ is given as 3. Calculate the *a posteriori* state $\mathrm{x}^+(5), P^+(5)$.
3. Calculate alternatively the outcome using a standard *least-squares adjustment*. We have as our dynamic model

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

   unknowns to be estimated $x(0)$ and $v(0)$, and observation equations

   $$\begin{aligned} \ell_1 + v_1 &= x(0) \\ \ell_2 + v_2 &= x(5) \end{aligned}$$

   and the observation vector

   $$\ell = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \; Q_{\ell\ell} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}.$$

Answer:

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

$$\begin{aligned} \Phi_0^5 &= e^{F\Delta t} = e^{\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}} = I + \begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}^2 + \dots \\ &= \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix}, \end{aligned}$$

because

$$\begin{bmatrix} 0 & \Delta t \\ 0 & 0 \end{bmatrix}^n = 0,\ n > 1.$$

Then

$$\begin{aligned} P(5) &= \Phi_0^5 P(0) \left(\Phi_0^5\right)^T + Q\Delta t = \\ &= \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 1000 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 5 & 1 \end{bmatrix} + 5 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \\ &= \begin{bmatrix} 25002 & 5000 \\ 5000 & 1005 \end{bmatrix}. \end{aligned}$$

2. The matrix $H = \begin{bmatrix} 1 & 0 \end{bmatrix}$. So $HP^-H^T + R = 25002 + 3 = 25005$. The K matrix is

$$K = -P^-H^T\left(HP^-H^T + R\right)^{-1} = -\begin{bmatrix} 25002 \\ 5000 \end{bmatrix} \cdot \frac{1}{25005} = \begin{bmatrix} -0.999880023995201 \\ -0.199960007998400 \end{bmatrix}.$$

Next, we compute

$$\underline{y} = H\mathrm{x}^-(5) - \underline{\ell} = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} 4 \\ 0 \end{bmatrix} - 3 = 1.$$

Then,

$$\mathrm{x}^+(5) = \mathrm{x}^-(5) + K\underline{y} = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.99988 \\ 0.19996 \end{bmatrix} \cdot 1 = \begin{bmatrix} 3.00012 \\ -0.19996 \end{bmatrix}.$$

(We can project this back to $t = 0$: we then find $\widehat{x}(0) = 3.00012 - 5 \cdot (-0.19996) = 3.9999$, and $\widehat{v}(0) = v^+(5) = -0.19996$.)
For the $P^+(5)$ matrix we find

$$\begin{aligned} P^+(5) &= (I + KH)\,P^-(5) = \\ &= \begin{bmatrix} 1 - 0.999880023995201 & 0 \\ -0.199960007998400 & 1 \end{bmatrix}\begin{bmatrix} 25002 & 5000 \\ 5000 & 1005 \end{bmatrix} = \\ &= \begin{bmatrix} 2.99964 & 0.59988 \\ 0.59988 & 5.19996 \end{bmatrix}. \end{aligned}$$

3. The A matrix is

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

and the $Q_{\ell\ell}$ matrix and ℓ vector are given. We obtain:

$$A^T Q_{\ell\ell}^{-1} A = \begin{bmatrix} 0.83333 & 1.66667 \\ 1.66667 & 8.33333 \end{bmatrix},$$

$$\widehat{\mathrm{x}} = \left(A^T Q_{\ell\ell}^{-1} A\right)^{-1} A^T Q_{\ell\ell}^{-1} \underline{\ell} = \begin{bmatrix} 4.00000 \\ -0.20000 \end{bmatrix}.$$

The same, practically, as the result under point 2. For the solution variance we find

$$\mathrm{Var}\left(\widehat{\mathrm{x}}\right) = \left(A^T Q_{\ell\ell}^{-1} A\right)^{-1} = \begin{bmatrix} 2 & -0.4 \\ -0.4 & 0.2 \end{bmatrix},$$

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

4 The Kalman filter in practical use

4.1 "Coloured noise", Gauss-Markov process

Let us study the simple dynamic equation

$$\frac{d\underline{x}}{dt} = -k\underline{x} + \underline{n}, \tag{4.1}$$

where $\underline{n}$ is white noise, of which the autocovariance function is $Q\delta(t_2 - t_1)$, and k is a constant. The solution of this differential equation is

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

The solution satisfies also the initial condition.

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

$$A_n(t_1, t_2) = Q(t_1)\,\delta(t_1 - t_2),$$

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

$$\begin{aligned} A_x(t_1, t_2) &= \\ &= e^{-k(t_1+t_2)} E\left\{\int_{t_0}^{t_1} \underline{n}(\tau_1)\, e^{k\tau_1} d\tau_1 \int_{t_0}^{t_2} \underline{n}(\tau_2)\, e^{k\tau_2} d\tau_2\right\} = \\ &= e^{-k(t_1+t_2)} \int_{t_0}^{t_1} e^{k\tau_1} \left[\int_{t_0}^{t_2} E\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\}\, e^{k\tau_2} d\tau_2\right] d\tau_1. \end{aligned}$$

Here

$$\begin{aligned} &\int_{t_0}^{t_2} E\{\underline{n}(\tau_1)\,\underline{n}(\tau_2)\}\, e^{k\tau_2} d\tau_2 = \\ &= \int_{t_0}^{t_2} A_n(\tau_2 - \tau_1)\, e^{k\tau_2} d\tau_2 = \\ &\quad Q\int_{t_0}^{t_2} \delta(\tau_2 - \tau_1)\, e^{k\tau_2} d\tau_2 = \begin{cases} Qe^{k\tau_1} & \text{jos } t_2 > \tau_1 \\ 0 & \text{jos } t_2 < \tau_1 \end{cases} \end{aligned}$$

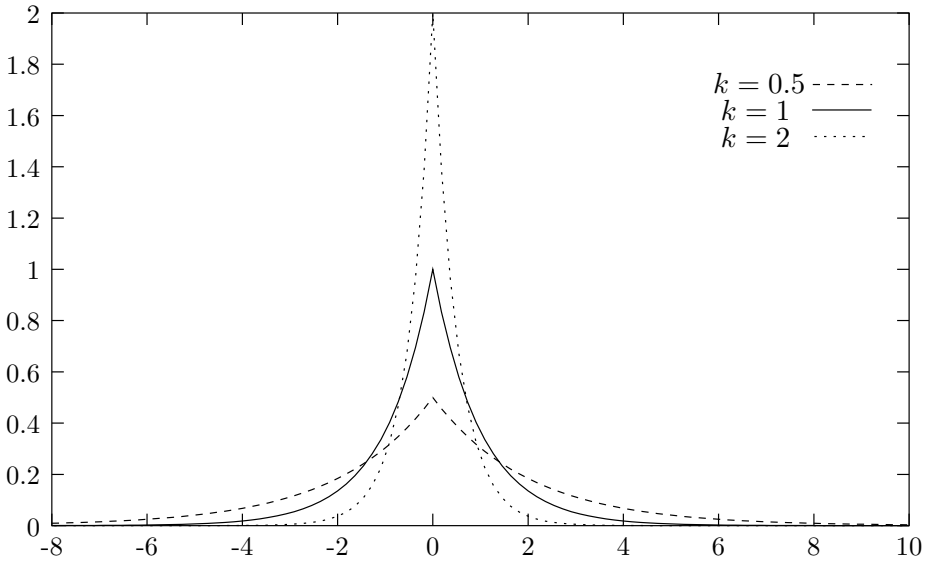


Figure 4.1: Gauss-Markov processes autocovariance function

So assuming that $t_2 < t_1$:

$$\begin{aligned} A_x(t_1, t_2) &= Qe^{-k(t_1+t_2)} \left[\int_{t_0}^{t_2} e^{2k\tau_1} d\tau_1 + \int_{t_2}^{t_1} 0 \, d\tau_1 \right] = \\ &= \frac{Q}{2k} e^{-k(t_1+t_2)} \left[e^{2kt_2} - e^{2kt_0} \right]. \end{aligned}$$

In this case where $t_2 > t_1$ this gives:

$$\begin{aligned} A_x(t_1, t_2) &= Qe^{-k(t_1+t_2)} \int_{t_0}^{t_1} e^{2k\tau_1} d\tau_1 = \\ &= \frac{Q}{2k} e^{-k(t_1+t_2)} \left[e^{2kt_1} - e^{2kt_0} \right]. \end{aligned}$$

In both cases we get

$$A_x(t_1, t_2) = \frac{Q}{2k} \left[e^{-k|t_1-t_2|} - e^{-k(t_1+t_2-2t_0)} \right]. \tag{4.2}$$

In the situation where $t_1, t_2 \gg t_0$ (stationary state long after starting) we obtain

$$A_x(t_2 - t_1) \equiv A_x(t_1, t_2) \approx \frac{Q}{2k} e^{-k|t_2-t_1|}. \tag{4.3}$$

In this (stationary) case *we talk about coloured noise* and the process above is called a (first order) *Gauss-Markov* process, also an autoregressive (AR(1)) process.

Let us also write

$$Q \equiv qk^2.$$

Then the surface area under the $A_x(t_2 - t_1)$ 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 $A_x(t_2 - t_1)$ becoming extremely narrow, but the surface area under the curve of the function does not change. In other words:

$$A_x(t_2 - t_1) = q\delta(t_2 - t_1).$$

This corresponds to the formula's (4.1) degeneration, where not only $k \to \infty$, but also variance of the noise $\underline{n}$, i.e., $Q \to \infty$. So:

$$\frac{d\underline{x}}{dt} = k\underline{x} - k\underline{\nu} \;\Rightarrow\; \underline{x} = \underline{\nu} - k^{-1}\frac{d\underline{x}}{dt} \approx \underline{\nu},$$

where the variance of the noise $\underline{\nu} \equiv -\frac{\underline{n}}{k}$ is $q = Qk^{-2}$.

The other borderline case case, where $k \to 0$, is the same as the case presented above (section 2.8). So "random walk" is a Gauss-Markov process the time constant of which is infinitely long. In that case we have to use the whole formula (4.2):

$$A_x(t_1, t_2) = \frac{Q}{2k}\left[e^{-k|t_1 - t_2|} - e^{-k(t_1 + t_2 - 2t_0)}\right].$$

In this case, if $t_2 \approx t_1 \equiv t$, we get

$$\begin{aligned} A_x(t) &= \frac{Q}{2k}\left[1 - e^{-2k(t - t_0)}\right] \approx \\ &\approx Q(t - t_0), \end{aligned}$$

which is in practice the same as in chapter 2.8.

The corresponding dynamic equation is obtained from the formula (4.1) by substituting $k = 0$:

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

so $\underline{x}$ is the time-integral of the white noise $\underline{n}$ as it should be.

Summary	k	dynamic model	autocovariance
Random walk	0	$\frac{d\underline{x}}{dt} = \underline{n}$	$Q(t - t_0)$
Gauss-Markov process	$\in (0, \infty)$	$\frac{d\underline{x}}{dt} = -k\underline{x} + \underline{n}$	$\frac{Q}{2k}e^{-k\lvert t_1 - t_2\rvert}$
White noise	∞	$\underline{x} = \frac{\underline{n}}{k}$	$Qk^{-2}\delta(t_1 - t_2)$

Often the model used to generate the "coloured" noise (4.1) or the process – in case where we know beforehand that the properties of the process are of that type.This is easily done by adding one unknown x to the state vector and one equation to the dynamic model of the Kalman filter.

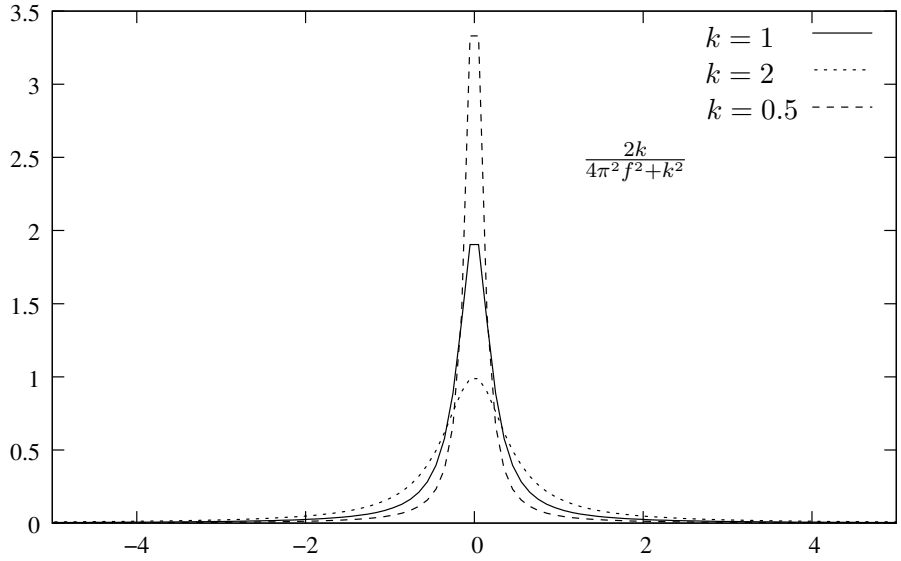


Figure 4.2: Power Spectral Density (PSD) of a Gauss-Markov process

Power spectral density of a Gauss-Markov process

We have the auto-covariance function as Eq. (4.3):

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

From this follows the PSD by integration (2.7):

$$\begin{aligned}\widetilde{A_x}(f) &= \int_{-\infty}^{+\infty} A_x(t) \exp(-2\pi i f t)\, dt = \\ &= \frac{Q}{2k} \int_{-\infty}^{+\infty} \exp(-k|t|) \exp(-2\pi i f t)\, dt.\end{aligned}$$

This integral isn't quite easy to evaluate; it is found in tabulations of integrals and can also be done using symbolic algebra software, like Wolfram's on-line integrator. The result is[1]

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

cf. Jekeli [2001] Eq. (6.75). In the figure are plotted values of this function for $Q = 2k$ – i.e., we keep the *variance* of $\underline{x}$, which is equal to $A_x(0) = Q/2k$, at unity – with $k = 0.5, 1, 2$.

4.2 Modelling of realistic statistical behaviour

Coloured noise, or 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

[1] A formula of this form is sometines called a Cauchy-Lorentz distribution.

$\underline{x}$ consists of the quantity we are interested in, $\underline{s}$ – which may be rapidly varying around zero – , and a systematic "disturbance" which we want to get rid of. We also know that this disturbance is slowly varying, with a time constant of τ_b. Let us call the disturbance $\underline{b}$. Then we may write the state vector as $\left[\begin{array}{cc} \underline{s} & \underline{b} \end{array} \right]^T$ and the dynamic equations as, e.g.,

$$\frac{d}{dt} \left[\begin{array}{c} \underline{s} \\ \underline{b} \end{array} \right] = \left[\begin{array}{cc} -1/\tau_s & 0 \\ 0 & -1/\tau_b \end{array} \right] \left[\begin{array}{c} \underline{s} \\ \underline{b} \end{array} \right] + \left[\begin{array}{c} \underline{n}_s \\ \underline{n}_b \end{array} \right].$$

Here, the τ_b is the (long) time constant of the bias process, which will thus be slowly varying; for for τ_s we may choose a much shorter time constant[2]. However, it should be chosen realistically. If measurements are obtained at a time interval Δt, $\tau_s \gg \Delta t$ in order for the process s to be realistically determinable from the observations.

The observation or Kalman update equation is

$$\underline{\ell} = \underline{s} + \underline{b} + \underline{m},$$

with $\underline{m}$ (variance R) representing the observational uncertainty. If observations are obtained at a sufficient density in time, we may obtain separate estimates for the signal process $\underline{s}$ and the slowly varying noise $\underline{b}$. In order for this to work, we should attach realistic auto-covariances to $\underline{n}_s$ and $\underline{n}_b$. Even then, it is a *requirement* in this case that $E\{\underline{s}\} = 0$. If it is not, the systematic part of $\underline{s}$ will end up in the $\widehat{b}$ estimate produced by the filter.

This is a case of spectral filtering by Kalman filter. The low frequency part, including zero frequency, goes to $\underline{b}$; the high frequency part goes to $\underline{s}$. 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's call it $\underline{c}$. It has a correlation length τ_c. Now if we are interested only in the signal's $\underline{s}$ lower frequency constituents, we may again apply a Kalman filter:

$$\frac{d}{dt} \left[\begin{array}{c} \underline{s} \\ \underline{c} \end{array} \right] = \left[\begin{array}{cc} -1/\tau_s & 0 \\ 0 & -1/\tau_c \end{array} \right] \left[\begin{array}{c} \underline{s} \\ \underline{c} \end{array} \right] + \left[\begin{array}{c} \underline{n}_s \\ \underline{n}_c \end{array} \right].$$

Here, we choose τ_s according to the part of the spectrum of $\underline{s}$ that we are interested in (but always $\tau_s > \tau_c$); τ_c should be chosen realistically, to capture and remove as much as possible the real noise in the process. Our observation or update equation is again

$$\underline{\ell} = \underline{s} + \underline{c} + \underline{m}.$$

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

[2] We may also choose an entirely different type of model, if we know that Gauss-Markov is not realistic for $\underline{s}$.

4.3 GPS observations and unknowns

GPS observations are described as *pseudoranges* and given by the equation

$$p = \rho + c\,(\Delta t - \Delta T) + d_{\text{ion}} + d_{\text{trop}}, \tag{4.4}$$

where

$\rho = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2}$ is the spatial distance between satellite $\begin{bmatrix} x & y & z \end{bmatrix}^T$ and ground station $\begin{bmatrix} X & Y & Z \end{bmatrix}^T$,

Δt is the satellite clock error,

ΔT is the receiver clock error, and

$d_{\text{ion}}, d_{\text{trop}}$ are the ionospheric and tropospheric effects.

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

$\mathbf{x} = \begin{bmatrix} x & y & z \end{bmatrix}^T$,

$\mathbf{X} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$,

$\Delta t, \Delta T$.

Satellite orbit determination

We can propose the following *observation equation* ($\underline{m}_p$ representing the observational uncertainty):

$$\underline{p} = \sqrt{(\underline{x} - X)^2 + \left(\underline{y} - Y\right)^2 + (\underline{z} - Z)^2} + c\,(\underline{\Delta t} - \Delta T) + d_{\underline{ion}} + d_{\underline{trop}} + \underline{m}_p.$$

This is the observation equation for *orbit determination*. In it, the ground station (tracking station) position is given and treated as non-stochastic: $\begin{bmatrix} X & Y & Z \end{bmatrix}^T$. 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 ΔT. The satellite clock, however, is being estimated.

For this situation we identify the *state vector* as

$$\underline{\mathrm{x}} = \begin{bmatrix} \underline{\mathbf{x}} \\ \underline{\mathbf{v}} \\ \underline{\Delta t} \\ \underline{d}_{\text{ion}} \\ \underline{d}_{\text{trop}} \end{bmatrix}.$$

As before, we introduced the velocity vector $\mathbf{v}$, so we can write the Kalman dynamical equations as a first-order differential equation.

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 formula 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

$$\begin{aligned}\Delta\mathbf{x} &= \mathbf{x} - \mathbf{x}_0,\\ \Delta\mathbf{v} &= \mathbf{v} - \mathbf{v}_0,\\ \Delta\left(\Delta t\right) &= \Delta t - \Delta t_0.\end{aligned}$$

Now, the linearized equations for $\mathbf{x}, \mathbf{v}$ are

$$\frac{d}{dt}\begin{bmatrix}\underline{\Delta\mathbf{x}}\\ \underline{\Delta\mathbf{v}}\end{bmatrix} = \begin{bmatrix}0 & I\\ M & 0\end{bmatrix}\begin{bmatrix}\underline{\Delta\mathbf{x}}\\ \underline{\Delta\mathbf{v}}\end{bmatrix} + \begin{bmatrix}0\\ \underline{n}_a\end{bmatrix},$$

where M is the earlier derived (for a central force field, Eq. (3.5)) *gravity gradient tensor*, I is the 3×3 unit matrix, and n_a is here introduced as the *dynamic noise* of satellite motion.

How do we model the behaviour of the satellite clock $\underline{\Delta t}$? Typically this is done as a *random walk* process. As follows:

$$\frac{d}{dt}\underline{\Delta t} = \underline{n}_t. \tag{4.5}$$

Modelling the tropo- and ionosphere 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 Gauß-Markov modelling is often used in this case, with a pragmatic choice of the time parameter τ. 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 $\underline{\Delta d}_{\text{ion}}, \underline{\Delta d}_{\text{trop}}$. For the ionosphere, this could be the model included with the satellite broadcast ephemeris (not very good), or the published IONEX models (not available in real time). For the troposphere, the standard Hopfield or Saastamoinen models may be considered.

Summarizing:

$$\frac{d}{dt}\begin{bmatrix}\underline{\Delta\mathbf{x}}\\ \underline{\Delta\mathbf{v}}\\ \Delta\left(\underline{\Delta t}\right)\\ \underline{\Delta d}_{\text{ion}}\\ \underline{\Delta d}_{\text{trop}}\end{bmatrix} = \begin{bmatrix}0 & I & & & \\ M & & & & \\ & 0 & & & \\ & & -\frac{1}{\tau_{\text{ion}}} & & \\ & & & & -\frac{1}{\tau_{\text{trop}}}\end{bmatrix}\begin{bmatrix}\underline{\Delta\mathbf{x}}\\ \underline{\Delta\mathbf{v}}\\ \Delta\left(\underline{\Delta t}\right)\\ \underline{\Delta d}_{\text{ion}}\\ \underline{\Delta d}_{\text{trop}}\end{bmatrix} + \begin{bmatrix}0\\ n_a\\ n_t\\ n_{\text{ion}}\\ n_{\text{trop}}\end{bmatrix}.$$

Station position determination

Starting from the same equation (4.4) we construct a different observation equation, as follows:

$$\underline{p} = \sqrt{(x-\underline{X})^2 + (y-\underline{Y})^2 + (z-\underline{Z})^2} + c\,(\Delta t - \underline{\Delta T}) + d_{\underline{ion}} + d_{\underline{trop}} + \underline{m}_p.$$

This is the observation equation for *geodetic positioning*. Here, the satellite orbital elements and clock are assumed known, i.e., $\begin{bmatrix} x & y & z \end{bmatrix}^T$ and Δt are known or precisely computable from available ephemeris. Now the *state vector* is

$$\underline{\mathrm{x}} = \begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \\ \underline{\Delta T} \\ \underline{d}_{\mathrm{ion}} \\ \underline{d}_{\mathrm{trop}} \end{bmatrix},$$

where $\underline{\mathbf{V}} = \frac{d}{dt}\underline{\mathbf{X}}$. Here, the new problem is to model the behaviour of the $\underline{\mathbf{X}}, \underline{\mathbf{V}}$ of the ground station.

In case the ground station is fixed, we may choose as the model

$$\underline{\mathbf{V}} = 0,$$

i.e., simply

$$\frac{d}{dt}\underline{\mathbf{X}} = 0.$$

In case we know that the stations are moving, but slowly and with constant velocity (e.g., plate tectonics, postglacial rebound), we may write

$$\frac{d}{dt}\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{V}} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The Kalman filter will gradually improve the estimates $\widehat{\mathbf{X}}, \widehat{\mathbf{V}}$ over time as more observations $\underline{p}$ are being processed. Some existing GPS processing software (GYPSY/OASIS) uses 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 Eq. (4.3), the variance of such a process is $Q/2k$, when the process equation is

$$\frac{d}{dt}\mathbf{A} = -k\mathbf{A} + \mathbf{n}_A.$$

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 manoeuver), and α 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 $\mathbf{n}_A$:

$$Q = \frac{2\alpha}{\tau_A}.$$

Thus we get as the complete dynamic equation:

$$\frac{d}{dt}\begin{bmatrix}\mathbf{X}\\ \mathbf{V}\\ \mathbf{A}\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & -1/\tau_A\end{bmatrix}\begin{bmatrix}\mathbf{X}\\ \mathbf{V}\\ \mathbf{A}\end{bmatrix} + \begin{bmatrix}0\\ 0\\ \frac{2\alpha}{\tau_A}\mathbf{n}_1\end{bmatrix},$$

where $\mathbf{n}_1$ stands for "unit variance white noise", 3-vectorial in this case.

Both α and τ_A will depend on the kind of vehicle we are considering. Large α and short τ_A is often referred to as a "high dynamic" environment, which is challenging for designing GPS receivers.

About clock modelling

Clocks are typically modelled as random walk processes, see Eq. (4.5):

$$\frac{d}{dt}\underline{c} = \underline{n}_c,$$

where now $\underline{c}$ 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 Eq. (2.5) 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 $\underline{n}_c$, and t_0 some starting time at which we have an exact value for Δt. We see that the variance grows linearly with time.

Let us compute the *difference* between two values $\underline{\delta c} \equiv \underline{c}(t_2) - \underline{c}(t_1)$. The *variance* of this difference is

$$\begin{aligned}\mathrm{Var}(\delta c) &= \mathrm{Var}\{c(t_2)\} + \mathrm{Var}\{c(t_1)\} - 2\mathrm{Cov}\{c(t_1), c(t_2)\} =\\ &= Q(t_2 - t_0) + Q(t_1 - t_0) - 2Q(t_1 - t_0) =\\ &= Q(t_2 - t_1),\end{aligned}$$

as was to be expected[3]. Obviously also, the *expected value* of $\underline{\delta c}$ vanishes:

$$E\{\underline{\delta c}\} = 0.$$

Now, suppose we have a time series of values

$$c(t_i),\ i = 1, \ldots, n,$$

with constant

$$\delta t = t_{i+1} - t_i.$$

[3] Why?

Then one can show that the expression

$$AV_{\delta t}\left(\underline{c}\right) = \frac{1}{n-1}\sum_{i=1}^{n-1}\left[\underline{c}\left(t_{i+1}\right) - \underline{c}\left(t_i\right)\right]^2 \tag{4.6}$$

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 (http://www.allanstime.com/AllanVariance/)[4]. For true random walk behaviour, $Q\delta t$, and thus $AV_{\delta t}\left(\underline{c}\right)$, should be strictly proportional to δt, and Q follows as the proportionality constant.

About ambiguity resolution

We may write the observation equation of carrier phase as (metric units):

$$P = \rho + c\left(\Delta t - \Delta T\right) + D_{\text{ion}} + D_{\text{trop}} + \lambda N. \tag{4.7}$$

Here, N identifies the *ambiguity*, an integer value identifying the number of whole wavelengths that cannot be determined from carrier phase measurements alone.

The way to handle the ambiguity in a Kalman filter may be to introduce an ambiguity unknown $\underline{N}$ to the state vector, but make it a real-valued state initially. As the filter progresses in time, the state variance attached to $\underline{N}$ 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 (4.7), 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 $\underline{N}_i$, $i = 1, \ldots, n$, 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 $\underline{N}_i$ lies close enough to a set of integer values, which forms 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 (http://www.lr.tudelft.nl/live/pagina.jsp?id=acd3da86-7b14-44e7-9de2-0d04c7c1a316&lang=en).

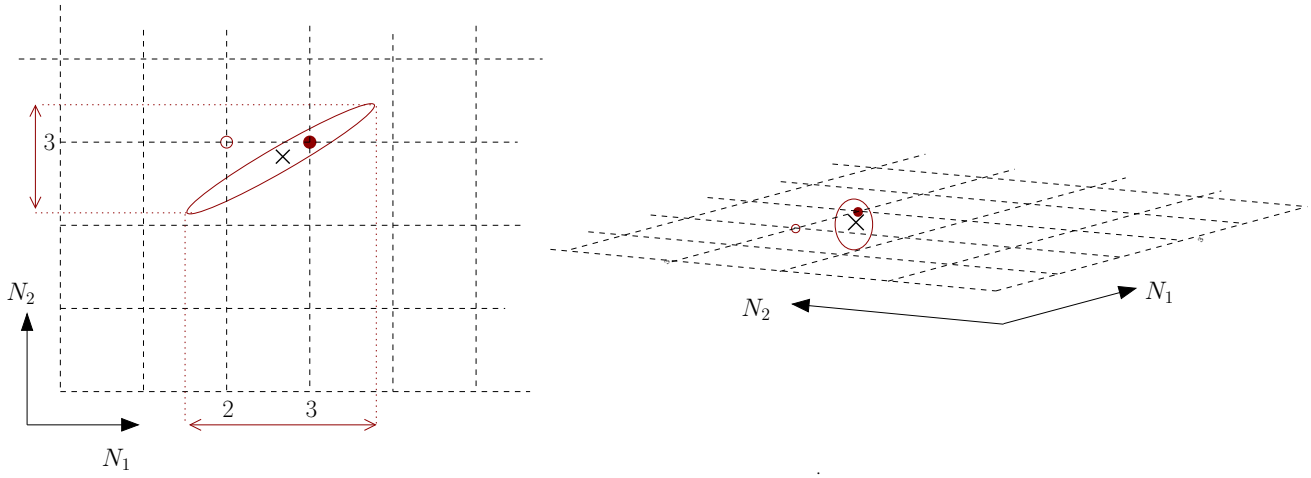
4.4 Examples

Kalman filter (2)

Question:

> In an industrial machine there is a wheel with radius r spinning at an angular velocity $\omega\left(t\right)$, where t is the time. The instantaneous angular velocity varies randomly: the angular acceleration has the properties of "white noise".

[4] Undoubtedly students of spatial information analysis will recognise this as very similar to the *semi-variogram* used in connection with the Kriging technique.



(Left) 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 $(3, 3)$.

(Right) multiple N_i variance ellipsoids are often very elongated "cigars" as depicted. In the LAMDA 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).

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:

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

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

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

where $\underline{\mathrm{x}}$ is the system's state vector and $\underline{\mathrm{n}}$ is the dynamical noise vector.

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

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

where the first equation $\frac{d}{dt}\underline{\alpha} = \underline{\omega}$ expresses the definition of angular velocity ω, and the second equation $\frac{d}{dt}\underline{\omega} = \underline{n}_\omega$ expresses the given fact that the angular acceleration has the properties of white noise.

We observe that the dynamical model found is *linear*.

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

$$\underline{\ell} = d + r\cos\underline{\alpha} + \underline{m}$$

(if we count α from the prism position furthest away from the observing instrument). Here, d is the distance between the instrument and the centre of the wheel. (We may assume for simplicity that it is known. If not, $\underline{d}$ should be added to the state vector with a dynamical equation of $\frac{d}{dt}\underline{d} = 0$ – aside remark.)

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, yielding (Taylor expansion into the first, linear term in $\Delta\alpha$):

$$\underline{\Delta\ell} = r\,\frac{\partial}{\partial\alpha}\cos\alpha\bigg|_{\alpha=\alpha_0} \cdot \underline{\Delta\alpha} + \underline{m},$$

where the logical definitions $\Delta\ell = \underline{\ell} - \ell_0$ and $\Delta\alpha = \underline{\alpha} - \alpha_0$ have been applied. Doing the partial differentiation yields

$$\Delta\ell = -r\sin\alpha_0\underline{\Delta\alpha} + \underline{m},$$

which is a linear equation of the standard Kalman observation equation type

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

if we write formally

$$\begin{aligned} \underline{\ell} &= [\underline{\Delta\ell}], \\ H &= \begin{bmatrix} -r\sin\alpha_0 & 0 \end{bmatrix}, \\ \underline{\mathrm{x}} &= \begin{bmatrix} \underline{\Delta\alpha} \\ \underline{\Delta\omega} \end{bmatrix}. \end{aligned}$$

Kalman filter (3)

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 can be described by the formula

   $$\rho = \rho_0 e^{-z/\sigma}$$

   (the constant σ is the scale height of the atmosphere, ρ_0 is air density at sea level).
2. A reflective tag is attached to the jumper in order to obtain measurements. A tacheometer on the ground measures the distance to this reflector. The horizontal distance between 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 a constant[5]):

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

   Define the state vector as $\left[\begin{array}{cc} \underline{z} & \underline{\dot{z}} \end{array}\right]^T$ and obtain as the dynamic model (first order differential equations):

   $$\frac{d}{dt}\left[\begin{array}{c} \underline{z} \\ \underline{\dot{z}} \end{array}\right] = \left[\begin{array}{c} \underline{\dot{z}} \\ -g + k\underline{\dot{z}}\rho_0 e^{-\underline{z}/\sigma} \end{array}\right] + \left[\begin{array}{c} 0 \\ \underline{n} \end{array}\right].$$

   This is non-linear; if we write

   $$\left[\begin{array}{c} \underline{z} \\ \underline{\dot{z}} \end{array}\right] = \left[\begin{array}{c} z_0 \\ \dot{z}_0 \end{array}\right] + \left[\begin{array}{c} \underline{\Delta z} \\ \underline{\Delta \dot{z}} \end{array}\right],$$

   where (completely computable if initial conditions are given)

   $$\frac{d}{dt}\left[\begin{array}{c} z_0 \\ \dot{z}_0 \end{array}\right] = \left[\begin{array}{c} \dot{z}_0 \\ -g + k\rho_0 \dot{z}_0 e^{-z_0/\sigma} \end{array}\right],$$

   we obtain (remember that (linearization) $\Delta\left(\dot{z}e^{-z/\sigma}\right) \approx -\frac{\dot{z}}{\sigma}e^{-z/\sigma}\Delta z + e^{-z/\sigma}\Delta\dot{z}$):

   $$\begin{aligned} \frac{d}{dt}\left[\begin{array}{c} \underline{\Delta z} \\ \underline{\Delta \dot{z}} \end{array}\right] &\approx \left[\begin{array}{c} \underline{\Delta \dot{z}} \\ -\frac{k\rho_0}{\sigma}\dot{z}_0 e^{-z_0/\sigma}\underline{\Delta z} + k\rho_0 e^{-z_0\sigma}\underline{\Delta \dot{z}} \end{array}\right] + \left[\begin{array}{c} 0 \\ \underline{n} \end{array}\right] = \\ &= \left[\begin{array}{cc} 0 & 1 \\ -\frac{k\rho_0}{\sigma}\dot{z}_0 e^{-z_0/\sigma} & k\rho_0 e^{-z_0\sigma} \end{array}\right]\left[\begin{array}{c} \underline{\Delta z} \\ \underline{\Delta \dot{z}} \end{array}\right] + \left[\begin{array}{c} 0 \\ \underline{n} \end{array}\right], \end{aligned}$$

   the linearized version of the dynamic model.

[5] A negative constant, because $\dot{z}$ is negative as well when z grows upward.

2. Let the horizontal distance between the touch-down point of the parachutist and the tacheometer be ℓ. Then the measured distance is

$$s = \sqrt{\ell^2 + z^2}$$

and the observation equation

$$\underline{s} = \sqrt{\ell^2 + \underline{z}^2} + \underline{m}.$$

Linearization ($s = s_0 + \Delta s$ where $s_0 = \sqrt{\ell^2 + z_0^2}$) yields

$$\underline{\Delta s} = \frac{z_0}{s_0}\underline{\Delta z} + \underline{m} = \begin{bmatrix} \frac{z_0}{s_0} & 0 \end{bmatrix} \begin{bmatrix} \underline{\Delta z} \\ \underline{\Delta \dot{z}} \end{bmatrix} + \underline{m}.$$

5

Inertial navigation

5.1 Principle

In inertial navigation, the following quantities are measured continuously:

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

$$\frac{d^2\mathbf{x}'(t)}{dt^2} = \begin{bmatrix} \frac{d^2x'(t)}{dt^2} \\ \frac{d^2y'(t)}{dt^2} \\ \frac{d^2z'(t)}{dt^2} \end{bmatrix};$$

   here $\mathbf{x}' \equiv \begin{bmatrix} x'(t) & y'(t) & z'(t) \end{bmatrix}^T$ is the object's three dimensional coordinates in the *object* coordinate system.

2. The *attitude* of the vehicle:

$$R = R_3(\alpha_3) R_2(\alpha_2) R_1(\alpha_1) =$$

$$\begin{bmatrix} \cos\alpha_3 & \sin\alpha_3 & 0 \\ -\sin\alpha_3 & \cos\alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\alpha_2 & 0 & -\sin\alpha_2 \\ 0 & 1 & 0 \\ \sin\alpha_2 & 0 & \cos\alpha_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_1 & \sin\alpha_1 \\ 0 & -\sin\alpha_1 & \cos\alpha_1 \end{bmatrix}$$

$$= \begin{bmatrix} \cos\alpha_2\cos\alpha_3 & \cos\alpha_1\sin\alpha_3 + \sin\alpha_1\sin\alpha_2\cos\alpha_3 & \sin\alpha_1\sin\alpha_3 - \cos\alpha_1\sin\alpha_2\cos\alpha_3 \\ -\cos\alpha_2\sin\alpha_3 & \cos\alpha_1\cos\alpha_3 - \sin\alpha_1\sin\alpha_2\sin\alpha_3 & \sin\alpha_1\cos\alpha_3 + \cos\alpha_1\sin\alpha_2\sin\alpha_3 \\ \sin\alpha_2 & -\sin\alpha_1\cos\alpha_2 & \cos\alpha_1\cos\alpha_2 \end{bmatrix},$$

so the transformation matrix between the global and object coordinates is:

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

at the moment of beginning of the journey t_0, where $\mathbf{x}$ and $\mathbf{x}'$ are global (often inertial) and object coordinates, respectively. The attitude is described by *three unknowns*, $\alpha_i(t)$, $i = 1, \ldots, 3$, 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 $\frac{d\alpha_i}{dt}$ are measured with the help of three gyroscopes as discussed later, and are integrated 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 $\mathbf{a} = \left[\begin{array}{ccc} \frac{d^2x}{dt^2} & \frac{d^2y}{dt^2} & \frac{d^2z}{dt^2} \end{array} \right]^T$ after the transformation

$$\mathbf{a} = R^{-1}\mathbf{a}'$$

in three dimensions, and *twice.* The first integration produces the object's (vehicle's) *velocity vector*, the second the *position* of the object.

As follows:

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \left[\mathbf{v}(t_0) + \int_{t_0}^{\theta} \mathbf{a}(\tau)\, d\tau \right] d\theta, \tag{5.1}$$

where $\mathbf{x}(t_0)$ and $\mathbf{v}(t_0)$ are *integration constants.*

As shown in the formula (5.1) the accuracy of position $\mathbf{x}(t)$ gets progressively poorer with time, because the acceleration measurements $\mathbf{a}(\tau)$ 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 $\mathbf{v}(t_1) = 0$, $t_1 > t_0$ and the inner (velocity) integral starts again from a known starting value.

5.2 Parts of a inertial device

An inertial device contains the following measuring parts:

1. Gyroscopes
2. Accelerometers

Gyroscope

A gyroscope is a rapidly spinning flywheel that tries not to change its axis of rotation. We can write the Euler equation as follows:

$$\mathbf{N} = \frac{d\mathbf{L}}{dt} = J\frac{d\overrightarrow{\omega}}{dt}, \tag{5.2}$$

where

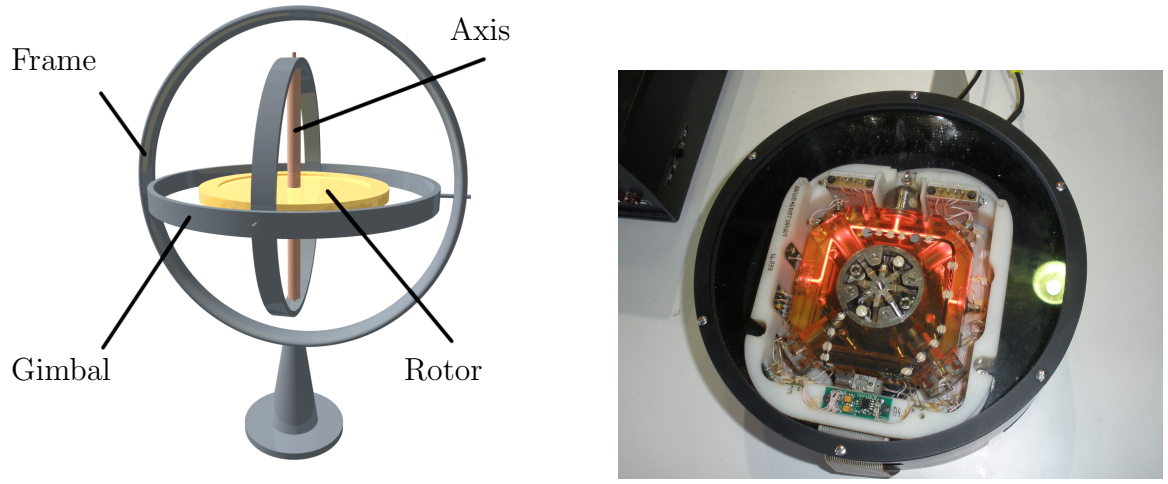


Figure 5.1: A gyroscope. On the right, a ring-laser gyro used in aviation. Wikipedia

$\mathbf{N}$	torque
$\mathbf{L}$	angular momentum
$\vec{\omega}$	angular velocity
J	*Inertial tensor:* $J \equiv \begin{bmatrix} J_{xx} & J_{xy} & J_{xz} \\ J_{xy} & J_{yy} & J_{yz} \\ J_{xz} & J_{yz} & J_{zz} \end{bmatrix}$ (a 3×3 sized matrix! This matrix is symmetric and positive definite.)

The faster the gyroscope rotates, the more torque is needed to turn its axis of revolution.

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 onto a frame, also called *table*, surrounding the wheel.

The above equation (5.2) can be remembered by analogy to the Newtonian Second Law of Motion:

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

where $\mathbf{F}$ 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 above, but is in this case a scalar. We assume all the time (which is natural for a flywheel, but not, e.g., for the whole Earth, which may change shape) that J (and m) is a constant.

The inertial tensor J of an object can be computed:

$$\begin{aligned}
J_{xx} &= \iiint \rho(x,y,z)\left(y^2+z^2\right)dxdydz,\\
J_{yy} &= \iiint \rho(x,y,z)\left(x^2+z^2\right)dxdydz,\\
J_{zz} &= \iiint \rho(x,y,z)\left(x^2+y^2\right)dxdydz,\\
J_{xy} &= -\iiint \rho(x,y,z)\,xydxdydz,\\
J_{xz} &= -\iiint \rho(x,y,z)\,xzdxdydz,\\
J_{yz} &= -\iiint \rho(x,y,z)\,yzdxdydz.
\end{aligned}$$

The result obviously depends on the choice of co-ordinate system (x, y, z). The *origin* has a large influence: by choosing it to lie far outside the object, we can make the elements of J arbitrarily large! Therefore, when talking about the inertial tensor of an object, we always choose the origin in the centre of mass:

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

or

$$\begin{aligned}
x_{\text{com}} &= \iiint \rho(x,y,z)\,xdxdydz,\\
y_{\text{com}} &= \iiint \rho(x,y,z)\,ydxdydz,\\
z_{\text{com}} &= \iiint \rho(x,y,z)\,zdxdydz,
\end{aligned}$$

after which we use in the computations

$$\mathbf{x}' = \mathbf{x} - \mathbf{x}_{\text{com}}.$$

As for the axes *orientation*, it is well known that a symmetric matrix can always be rotated – i.e., a co-ordinate system transformation – to *main axes*. In this case the inertial tensor assumes the diagonal form

$$J = \begin{bmatrix} J_1 & 0 & 0\\ 0 & J_2 & 0\\ 0 & 0 & J_3 \end{bmatrix}.$$

The J_i are called the *moments of inertia*.

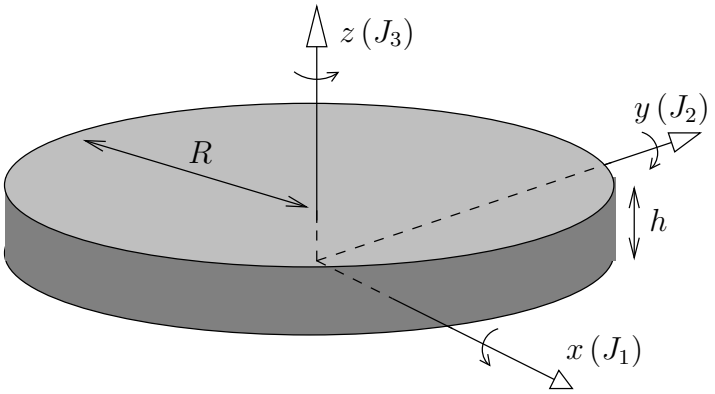


Figure 5.2: A gyro wheel and its moments of inertia

For a cylinder of radius R, one can show that the moment of inertia about the axis of the cylinder is

$$\begin{aligned} J_3 &= \int_0^h \iint_{\text{ympyrä}} \rho\left(x^2+y^2\right) dxdy\,dz \\ &= 2\pi h\rho \cdot \iint_{\text{ympyrä}} r^2\, r dr = \\ &= \frac{1}{2}\pi\rho h R^4 = \frac{1}{2}MR^2, \end{aligned} \tag{5.3}$$

where $M = \rho \cdot \pi R^2 \cdot h$ is the total mass. For a flat cilinder (h, and thus z, are small) we may also calculate

$$\begin{aligned} J_1 &= \int_{-h/2}^{h/2}\int_{-R}^{+R}\int_{-\sqrt{R^2-x^2}}^{+\sqrt{R^2-x^2}} \rho\left(y^2+z^2\right) dydxdz \approx \\ &\approx \int_{-h/2}^{h/2}\int_{-R}^{+R}\int_{-\sqrt{R^2-x^2}}^{+\sqrt{R^2-x^2}} \rho y^2 dydxdz = \\ &= \frac{h\rho}{3}\int_{-R}^{+R}\left[\left(R^2-x^2\right)^{\frac{3}{2}} + \left(R^2-x^2\right)^{\frac{3}{2}}\right] dx = \\ &= \frac{2h\rho}{3}\int_{-R}^{+R}\left(R^2-x^2\right)^{\frac{3}{2}} dx = \\ &= \frac{2h\rho}{3}\left[x\left(5R^2-2x^2\right)\sqrt{R^2-x^2} + \frac{3}{8}R^4 \arctan\frac{x}{\sqrt{R^2-x^2}}\right]_{-R}^{+R} = \\ &= \frac{1}{4}h\rho\cdot R^4\left[\frac{\pi}{2}+\frac{\pi}{2}\right] = \frac{1}{4}\left(\pi\rho h R^2\right)R^2 = \frac{1}{4}MR^2. \end{aligned} \tag{5.4}$$

Also of course $J_2 = J_1 = \frac{1}{4}MR^2 = \frac{1}{2}J_3$.

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.

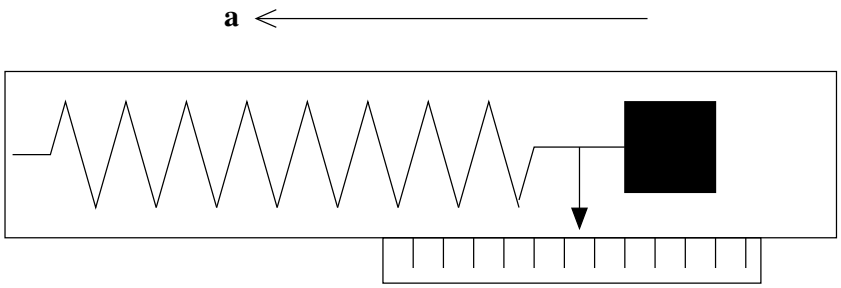


Figure 5.3: Accelerometer principle

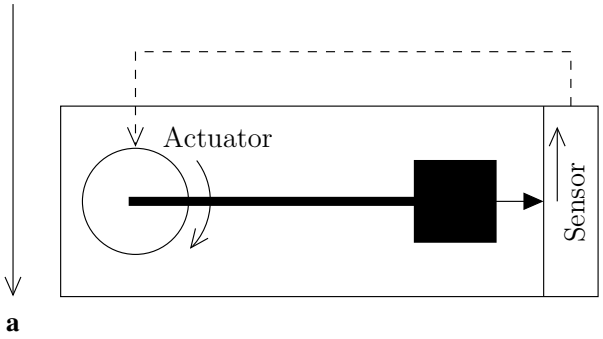


Figure 5.4: Pendulous accelerometer

Automatic read-out is possible, e.g., capacitively or with the aid of a piezo-sensor[1].

The accelerometers are attached to the same frame in which also the gyroscopes are suspended. The measurement axes are made as parallel as possible.

Modern accelerometers are very sensitive, e.g., 10 ppm. If they are based on the elasticity of matter, they demand careful, regular 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 excentrically to a beam. Acceleration makes the beam deflect, which is sensed by a sensor. The signal goes to an actuator on the pendulum's axis, which restores the deflection to zero. It is thus a *nulling* sensor, which is necessary to guarantee linear behaviour. Pendulous accelerometers are used in the highest precision devices. They do not suffer from drift.

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.

5.3 Implementation

There are two, very different, general approaches for implementing an inertial measurement unit:

[1]In fact, micromechanical acceleration sensors (MEMS) work in precisely this way.

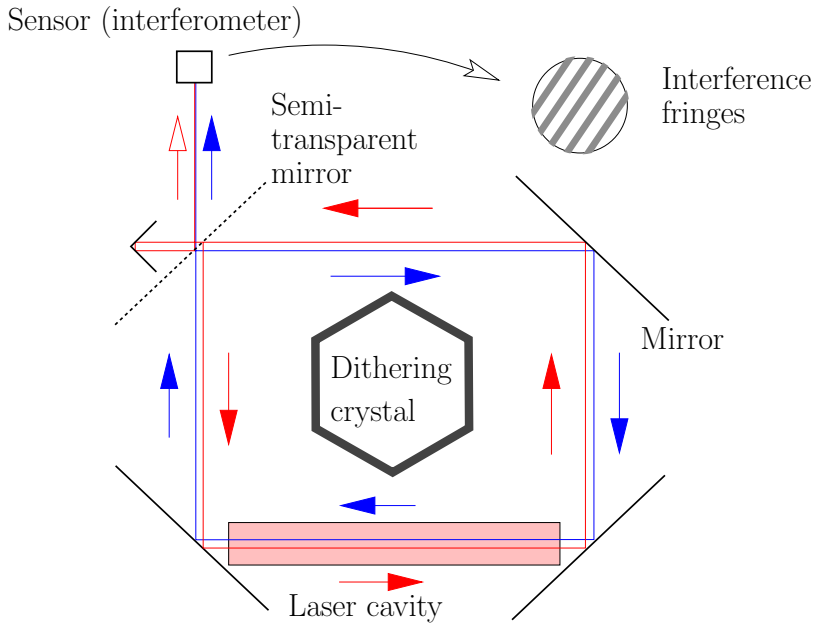


Figure 5.5: SAGNAC-interferometer

1. Strapdown solution

2. Stabilized platform solution

Strapdown solution

In a Strapdown solution the gyroscope platform is rigidly connected to the vehicle's body. When the vehicle turns, 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 formula:

$$\mathbf{N} = \mathbf{F} \wedge \overrightarrow{\ell},$$

where $\overrightarrow{\ell}$ is the length of the gyroscope's axis as a vector: "torque is force times arm". The symbol $\wedge$ is the exterior vector product.

An alternative solution is to use a so called *ring laser gyroscope* that is based on the interference of light (the SAGNAC phenomenon, 1913). In the device monochromatic laser light travels in a ring in two opposite directions. Without rotation, the light forms a "standing wave" where the nodes don't move. However, even a small rotation will cause the nodes to move to the opposite direction relative to the ring. The simplest way to build a ring laser is to use stationary mirrors; nowadays often a long optical fibre is used that is wound around 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. (http://www.mathpages.com/rr/s2-07/2-07.htm).

Stabilized platform -solution

In this solution the whole gyroscope system is suspended inside a three-axis, freely turning cardanic ring system. Because of this, although the attitude of the vehicle changes, the gyroscopic frame (gyroscope table) retains its position in (inertial) space.

In practice one often uses instead of an inertial reference frame, a *local* frame connected to the solid Earth. One tries to keep the three axes of the gyroscope aligned with the *topocentric* axes triad:

1. North direction x
2. East direction y
3. Up direction z

To achieve this goal, appropriate torques are applied to the frame of the gyroscope with the help of *torquers*. The needed torques can be calculated anologically or digitally in connection with solving for the position of the device.

5.4 Inertial navigation in the system of the solid Earth

Cf. Cooper [1987] p. 104-107 (a slightly different approach)

Earth rotation

We can write the vector of place in inertial space as a function of the vector of place in a co-ordinate system co-rotating with the Earth, as follows:

$$\mathbf{x}_i = R(\theta)\,\mathbf{x},$$

where θ is the sidereal time. Its time derivative $\omega = \dot{\theta}$ is the angular velocity of the Earth's rotation. The matrix

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For the velocity we find by differentiation:

$$\begin{aligned} \mathbf{v}_i &= R(\theta)\,\mathbf{v} + \dot{R}(\theta)\,\mathbf{x} = \\ &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{v} + \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d\theta}{dt} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \\ &= R(\theta)\,\mathbf{v} + R(\theta)\,\langle \overrightarrow{\omega} \wedge \mathbf{x} \rangle\,, \end{aligned}$$

if we define[2]

$$\overrightarrow{\omega} \equiv \frac{d\theta}{dt} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

By suitably choosing $t = \theta = 0$, we get

$$\mathbf{v}_i = \mathbf{v} + \overrightarrow{\omega} \wedge \mathbf{x}.$$

By repeating the differentiation we obtain the accelerations:

$$\begin{aligned} \mathbf{a}_i &= R(\theta)\,\mathbf{a} + \dot{R}(\theta)\,\mathbf{v} + \frac{d}{dt}\left\{R(\theta)\left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle\right\} = \\ &= R(\theta)\,\mathbf{a} + R(\theta)\left\langle\overrightarrow{\omega} \wedge \mathbf{v}\right\rangle + \left(R(\theta)\left\langle\overrightarrow{\omega} \wedge \mathbf{v}\right\rangle + \dot{R}(\theta)\left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle\right) = \\ &= R(\theta)\left\{\mathbf{a} + 2\left\langle\overrightarrow{\omega} \wedge \mathbf{v}\right\rangle + \left\langle\overrightarrow{\omega} \wedge \left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle\right\rangle\right\}. \end{aligned}$$

By putting again $\theta = 0$ we find

$$\mathbf{a}_i = \mathbf{a} + 2\left\langle\overrightarrow{\omega} \wedge \mathbf{v}\right\rangle + \left\langle\overrightarrow{\omega} \wedge \left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle\right\rangle.$$

Acceleration

The problem is that on the rotating Earth the before mentioned three-dimensional coordinate system (x, y, z) *is not inertial*. We can write:

$$\mathbf{a}_i = \mathbf{a} + 2\left\langle\overrightarrow{\omega} \wedge \mathbf{v}\right\rangle + \left\langle\overrightarrow{\omega} \wedge \left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle\right\rangle,$$

where

[2] ... because for an arbitrary vector $\mathbf{x}$

$$\dot{R}(\theta)\,\mathbf{x} = \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d\theta}{dt} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{d\theta}{dt} \begin{bmatrix} -x\sin\theta - y\cos\theta \\ x\cos\theta - y\sin\theta \\ 0 \end{bmatrix}$$

and also

$$\begin{aligned} R(\theta)\left\langle\overrightarrow{\omega} \wedge \mathbf{x}\right\rangle &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \left\{ \begin{bmatrix} 0 \\ 0 \\ \omega \end{bmatrix} \wedge \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right\} = \\ &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\omega y \\ \omega x \\ 0 \end{bmatrix} = \\ &= \omega \begin{bmatrix} -x\sin\theta - y\cos\theta \\ x\cos\theta - y\sin\theta \\ 0 \end{bmatrix}, \end{aligned}$$

in other words, the same result. Thus we can *conclude*:

> The effect of rotational motion on the time derivative of a vector can be presented as the cross product of the rotation vector $\overrightarrow{\omega}$ with this vector.

$\mathbf{a}_i$	acceleration in inertial system
$\mathbf{a}$	acceleration relative to the Earth's surface, in other words, in an Earth-fixed, "co-rotating" system
$\overrightarrow{\omega}$	Earth's rotation vector (*constant*)
$\mathbf{v}$	velocity in the Earth-fixed system
$\mathbf{x}$	the geocentric location of the vehicle

In the above formula the second term on the right side is the so called *Coriolis-force* and the third term is the *centrifugal* force.

Fundamental formula of inertia navigation

Linear accelerometers measure in general the *combined effect* of the acceleration of the vehicle and the local gravitation. In other words, the measured acceleration is

$$\mathbf{t} = \mathbf{a} + 2\left\langle \overrightarrow{\omega} \wedge \mathbf{v} \right\rangle + \left\langle \overrightarrow{\omega} \wedge \left\langle \overrightarrow{\omega} \wedge \mathbf{x} \right\rangle \right\rangle - \mathbf{g}_i\left(\mathbf{x}\right), \tag{5.5}$$

where

$\mathbf{t}$	measured acceleration vector (three components)
$\mathbf{g}_i$	*gravitational* acceleration as the function of place $\mathbf{x}$.

It is often assumed that $\mathbf{g}$ can be calculated straight from Newton's gravitation formula:

$$\mathbf{g}_i \approx -GM\frac{\mathbf{x}}{\|\mathbf{x}\|^3},$$

but also more complex models are used, such as the normal field of an ellipsoid of revolution (where the Earth's oblateness and the influence of its rotation are included) and even very detailed Earth gravitational field models, such as EGM96 (Earth Gravity Model 1996).

Often we write still

$$\mathbf{g} \equiv \mathbf{g}_i - \left\langle \overrightarrow{\omega} \wedge \left\langle \overrightarrow{\omega} \wedge \mathbf{x} \right\rangle \right\rangle,$$

where $\mathbf{g}$ is the *gravity vector,* the resultant of gravitation and centrifugal force. Then

$$\mathbf{t} = \mathbf{a} + 2\left\langle \overrightarrow{\omega} \wedge \mathbf{v} \right\rangle - \mathbf{g}\left(\mathbf{x}\right). \tag{5.6}$$

With the help of the formula (5.6) 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 Earth-fixed system*, and after that integrating first $\mathbf{v}$, and then $\mathbf{x}$, both also in the Earth-fixed system. Formulas (5.5, 5.6) are both referred to as *the fundamental formula of inertial navigation.*

Note that in the Earth-fixed system, the Earth's rotation causes a slow turning in the East-West direction of the vector of gravity sensed by the accelerometers, relative to the

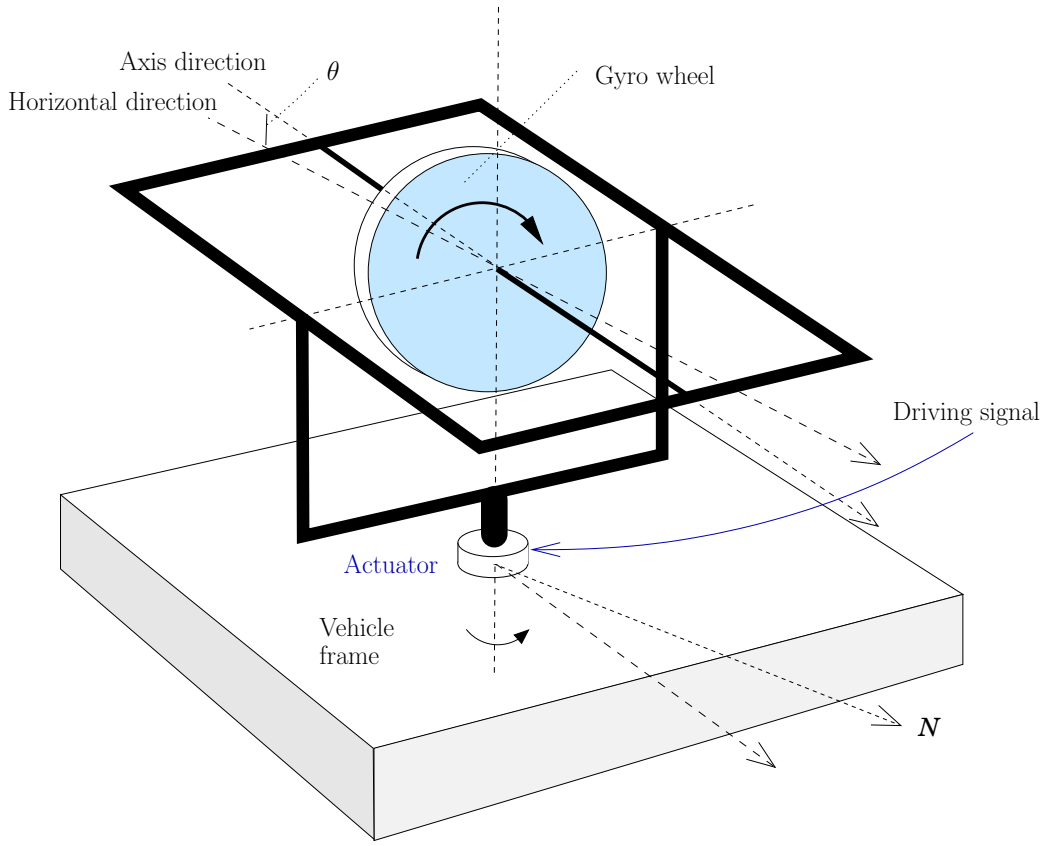


Figure 5.6: The principle of a stable table. The driving signal produces a precessional motion that keeps the gyro's axis within the horizon plane

inertial directions defined by the gyros, even though the vehicle is standing still on the ground. This phenomenon may be used to orient the gyroscope frame correctly relative to the local North direction (or equivalently, to solve the local North direction in the gyroscope frame's system!) before, e.g., the take-off of an aeroplane or launch of a missile. On the other hand, the accelerometers give right 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 pole.

5.5 Stable table with one axis

Let us first look at the *stable table*, i.e., a gyroscope that is attached to a frame, which is kept aligned with the local horizon. In the stable table solution one uses a *feedback loop* to control the gyroscope axis' direction 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 rotate the gyroscope frame in the horizontal plane (around the vertical axis) causes the gyroscope to *precess*. The rotational axis of the gyroscope turns up- or downwards. The stable table requires a suitable sensor that detects that the gyro's axis is out of the horizontal plane (angle θ), which sends a signal through the feedback loop to the motor, or *actuator*, of the vertical axis (cf. picture 5.6). More about this later.

We write as a formula, that the torque about the vertical axis is made proportional to the sensed *axis deviation* θ from the horizontal plane. Then the change of θ with time is

$$\frac{d\theta}{dt} = -k_1\theta$$

i.e.

$$\theta\,(t) = \theta\,(t_0)\,e^{-k_1(t-t_0)},$$

in other words, the deviation goes to zero exponentially. By tuning the constant of feedback (or equivalently the constant k_1 in the formula) we can make this happen with suitable speed.

5.6 The gyro compass

The feedback loop visible in the gyrocompass picture again makes use of *Earth rotation.* Because the Earth rotates around its axis, the horizontal plane is tilting all the time. The Eastern horizon sinks, the Western rises. A freely suspended, spinning gyroscope, which initially was in the horizontal plane, wouldn't be any more after an elapse of time.

If the rotational velocity of the Earth is ω, then the time derivative of the angle θ will be, because of this phenomenon,

$$\frac{d\theta}{dt} = \omega\cos\varphi\sin\alpha,$$

where φ is the latitude and α the azimuth of the gyroscope axis.

The feedback loop takes from the sensor the angle θ's *time derivative* and feeds it, after suitable amplification, into the actuator. As it tries to turn the gyroscope's axis toward the vertical direction, the end effect will be precession about the vertical axis: α changes. We write the formula

$$\frac{d\alpha}{dt} = -k_2\frac{d\theta}{dt} = -k_2\,\omega\cos\varphi\sin\alpha.$$

If α is small enough, we have $\sin\alpha \approx \alpha$ and the solution is

$$\alpha\,(t) \approx \alpha\,(t_0)\,e^{-k_2\omega\cos\varphi(t-t_0)}.$$

I.e., α goes exponentially to zero and *the gyroscope axis to the North.* Thus we have invented the *gyro compass.* Of course this assumes that the table remains horizontal and that the whole device stays in the same spot (or in practice moves only slowly, e.g., a ship.)

Another way to build a working gyrocompass uses θ itself rather than its time derivative; if we write

$$\frac{d\alpha}{dt} = -k_3\theta,$$

we obtain by differentiation

$$\frac{d^2\alpha}{dt^2} = -k_3\frac{d\theta}{dt} = -k_3\,\omega\cos\varphi\sin\alpha \approx -k_3\,\omega\cos\varphi\cdot\alpha.$$

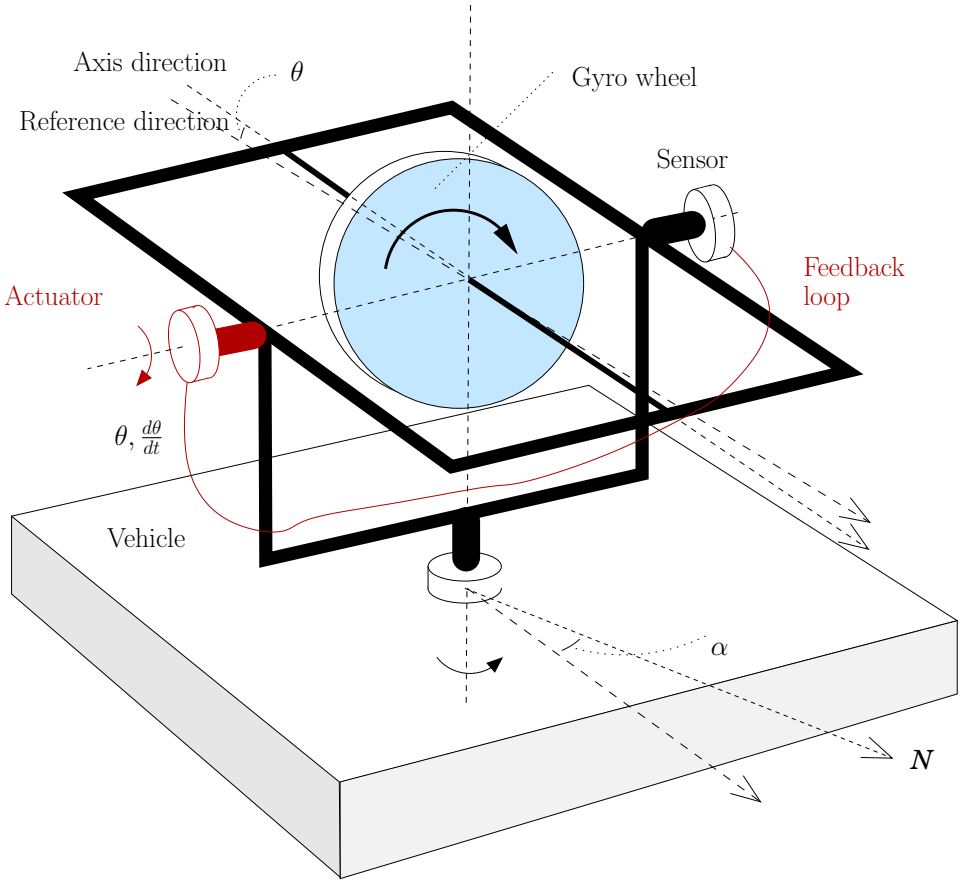


Figure 5.7: The principle of the gyro compass. The feedback loop produces a precessional motion that makes the gyro's axis turn to the North

This is a *harmonic oscillator,* some of the solutions of which are

$$\alpha(t) = \cos\left(t\sqrt{k_3\,\omega\cos\varphi}\right),$$
$$\alpha(t) = \sin\left(t\sqrt{k_3\,\omega\cos\varphi}\right).$$

Unfortunately these solutions are periodic and do not converge to the North direction ($\alpha = 0$). The best solution is obtained by combining θ and $\frac{d\theta}{dt}$ in the following way:

$$\frac{d^2\alpha}{dt^2} = -k_2\,\omega\cos\varphi\frac{d\alpha}{dt} - k_3\,\omega\cos\varphi\cdot\alpha,$$

leading to the following differential equation

$$\frac{d^2\alpha}{dt^2} + \omega\cos\varphi\left[k_2\frac{d\alpha}{dt} + k_3\alpha\right] = 0.$$

This is a general second order ordinary differential equation. Depending on the coefficients k_2 and k_3, it will have periodic, exponentially (over-)damped and *critically damped* solutions[3]. The last mentioned is the best for a functioning compass.

[3]Cf. https://en.wikipedia.org/wiki/Damping.

If we write the *inverse of the oscillation time* $\tau = \sqrt{k_3\,\omega\cos\varphi}$, and

$$k_2 = \frac{2\tau}{\omega\cos\varphi},$$

we obtain

$$\frac{d^2\alpha}{dt^2} + 2\tau\frac{d\alpha}{dt} + \tau^2 = 0$$

and the general solution in this case is

$$\alpha\,(t) = (a + bt)\,e^{-\tau t},$$

where a and b are arbitrary constants (given by the initial conditions).

Often k_3 (the harmonic restoration coefficient) is implemented by attaching rigidly a semi-ring to the inner ring of the gyroscope, which extends downward and to which a weight is attached. This tries then to pull the rotation axis of the gyroscope back to the horizontal plane. k_2 (the damping factor) again is implemented traditionally by using a viscous fluid in the bearings of the inner ring.

5.7 Schuler pendulum

Principle

A Schuler[4] *pendulum* is a pendulum, the length of which is the same as the Earth's radius $R = 6378$ km. If that kind of pendulum was physically possible, for example as a mass at the end of a long rod, its period would be (in a one-g gravity field!)

$$T_S = 2\pi\sqrt{\frac{R}{g}},$$

where g is gravity on the Earth's surface.

"By coincidence" this period, $T_S = 84.4$ min, is the same as the orbital period of an Earth satellite near the Earth surface.

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.

Let the length of a simple pendulum (i.e., a test mass on the end of a massless bar) be ℓ. If the pendulum swings out of the vertical by an angle θ then the back pulling force will be

$$F = -mg\sin\theta,$$

and as its mass is m, it follows that the acceleration is

$$\frac{d^2\ell\theta}{dt^2} = -\frac{mg\sin\theta}{m} \;\Rightarrow\; \frac{d^2\theta}{dt^2} \approx -\frac{g}{\ell}\theta,$$

[4]Max Schuler (1882–1972), saksalainen insinööri, https://en.wikipedia.org/wiki/Max_Schuler

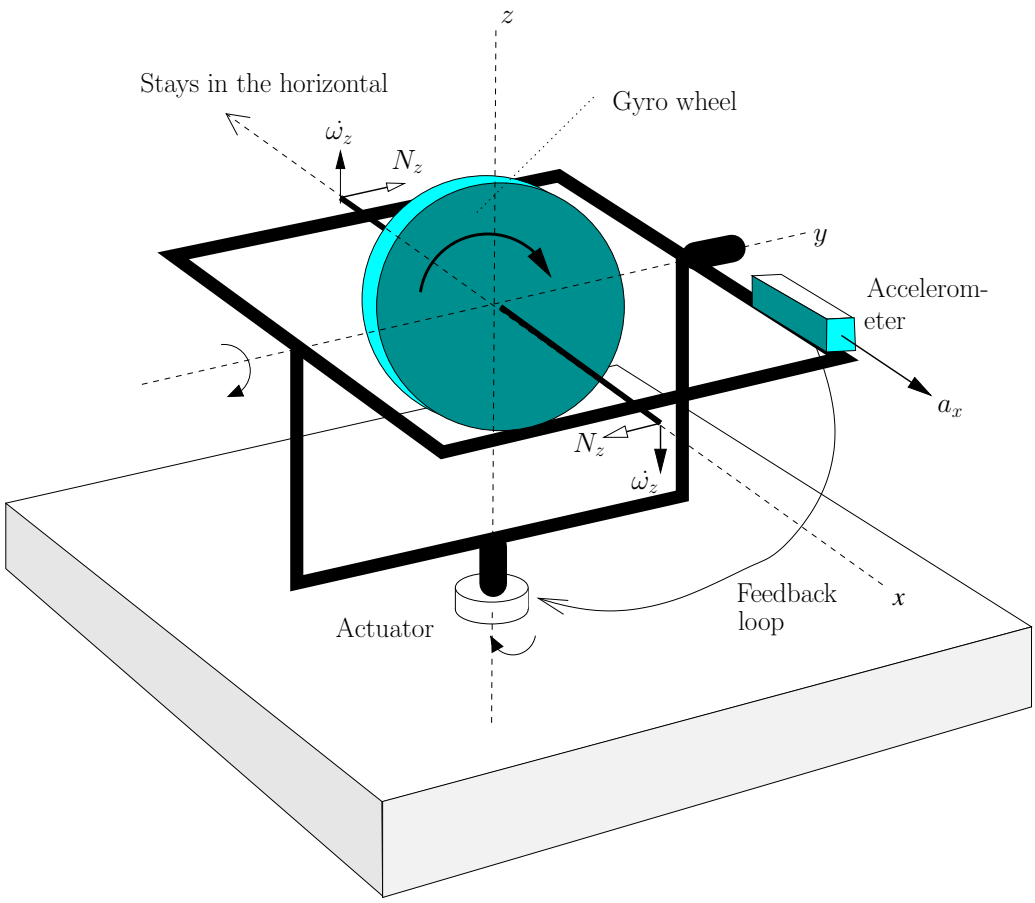


Figure 5.8: Schuler response loop

the oscillation equation, of which one solution is

$$\theta(t) = \sin\left(t\sqrt{\frac{g}{\ell}}\right),$$

from which it follows that the period is

$$T = 2\pi\sqrt{\frac{\ell}{g}}.$$

The pendulum on a carriage

If this pendulum is put on a carriage that accelerates linearly in the horizontal direction with an acceleration a, the test mass will, in the system of the carriage, experience a equally large but oppositely directed acceleration $-a$. Because the length of the pendulum is ℓ, it follows that the angular acceleration is

$$\frac{d^2\theta}{dt^2} = \frac{a}{\ell},$$

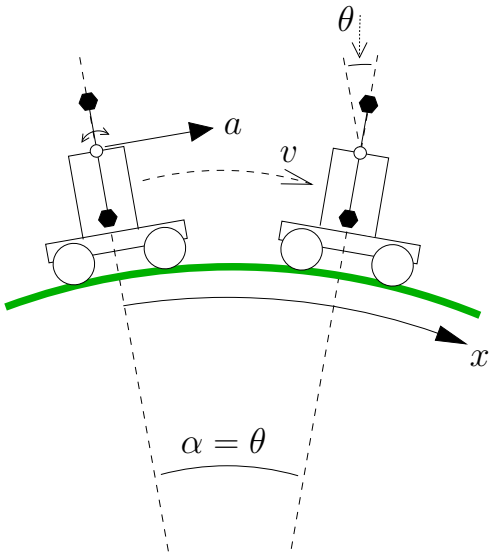


Figure 5.9: One-dimensional carriage with a Schuler pendulum on the curved Earth surface

and after a certain time Δt the reached angular deviation is

$$\theta = \frac{1}{2}\frac{a}{\ell}\Delta t^2. \tag{5.7}$$

The distance that the carriage has travelled after this same time is

$$s = \frac{1}{2}a\Delta t^2$$

and this distance expressed as an angle viewed from the centre of the Earth is

$$\alpha = \frac{1}{2}\frac{a}{R}\Delta t^2. \tag{5.8}$$

By comparing the formulas (5.7) and (5.8) we can see that if $\ell = R$, then $\alpha = \theta$. So,

> **Even though the carriage moves in a horizontal direction, the pendulum points all the time to the centre of the Earth.**

This is the so called *Schuler pendulum*'s essential property.

Implementation in an inertial device

In a stabilized-platform inertial device feedback loops (*Schuler loop*) 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 $\mathbf{g}$ 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 the pendulum formula

$$\frac{d^2}{dt^2}\theta = \frac{a}{\ell}, \tag{5.9}$$

where a is the accelerometer's measured acceleration in the x direction.

We may write geometrically for the deviation of the gyro spin axis out of the horizontal plane θ

$$\frac{d\theta}{dt} = \frac{d}{dt}\left(\frac{\omega_z}{\widetilde{\omega}}\right) = \frac{1}{\widetilde{\omega}}\frac{d}{dt}\omega_z.$$

The angular acceleration is now

$$\frac{d^2}{dt^2}\theta = \frac{1}{\widetilde{\omega}}\frac{d}{dt}\frac{d}{dt}\omega_z, \tag{5.10}$$

the acceleration of turning the rotational velocity vector *of the gyroscope*[5] $\widetilde{\omega}$ in the z direction.

Now substituting Eq. (5.9) into Eq. (5.10) and integration yields

$$\frac{d}{dt}\omega_z = \frac{\widetilde{\omega}}{\ell}\int a dt.$$

Using Euler's equation

According to Euler's formula (5.2)

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

where J is the inertial tensor; due to the symmetry of the gyro wheel, it is (in the standard orientation) a diagonal matrix:

$$J = \begin{bmatrix} J_{xx} & 0 & 0 \\ 0 & J_{yy} & 0 \\ 0 & 0 & J_{zz} \end{bmatrix},$$

where $J_{yy} = J_{zz} \approx \frac{1}{2}J_{xx}$ for a thin circular disk (cf. Eqs. 5.3, 5.4). Then the third Euler equation is

$$N_z = J_{zz}\frac{d\omega_z}{dt} = J_{zz}\frac{\widetilde{\omega}}{\ell}\int a dt,$$

where J_{zz} is the gyro wheel's moment of inertia around its z axis, and N_z the needed torque around the z axis (cf. figure 5.8). Now

$$N_z = J_{zz}\widetilde{\omega}\int \frac{a}{\ell}dt \approx \frac{\widetilde{L}}{R}\int a dt, \tag{5.11}$$

where R is the radius of the Earth, approximately 6378 km, and $\widetilde{L} = J_{zz}\widetilde{\omega}$ is a quantity of dimension "angular momentum", formula (5.2).

[5] Here we talk about the spinning of the gyrocope, not the Earth!

According to formula (5.11) the Schuler loop is implemented either on the hardware-level (older equipment; the factor $\frac{\widetilde{L}}{R}$ is a device constant, and integration is done by 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 Mechanisation

Cf., e.g., http://www.frc.ri.cmu.edu/~alonzo/pubs/reports/kalman_V2.pdf, http://www.frc.ri.cmu.edu/~alonzo/pubs/reports/nav.pdf.

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 mechanisation* of the inertial device.

As a simple example of mechanisation is treated a one dimensional carriage on the surface of a spherical Earth. Cf. figure 5.9.

First it can be pointed out that according to the definition, the velocity is

$$\frac{dx}{dt} = v.$$

Acceleration is *measured* continuously by an acceleration sensor; the measured value is $a(t)$. However this measured quantity (function of time) consists of two parts,

1. the geometric acceleration $\frac{d^2x}{dt^2} = \frac{dv}{dt}$, and
2. the component of gravity projected onto the accelerometer's axis, θg, where $\theta(t)$ is the angle of tilt of the carriage from the local vertical.

The final outcome is

$$\frac{dv}{dt} = a - \theta g,$$

or (remember that a differential equation is a statement on the properties of *functions*):

$$\frac{dv(t)}{dt} = a(t) - \theta(t) g,$$

where the quantity $a(t)$ is the result of a continuous measurement process.

Finally we treat the Schuler loop. The angle of deflection θ behaves like a Schuler pendulum and tries to revert to zero according to the following formula:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{R}\theta. \tag{5.12}$$

Let's determine the approximate values (functions of time) $x_0(t)$, $v_0(t)$, $\theta_0(t) \equiv 0$, and $\Delta x = x - x_0$, $\Delta v = v - v_0$ (linearization). Then

$$\begin{aligned} \frac{dx_0}{dt} &= v_0 \\ \frac{dv_0}{dt} &= a \end{aligned}$$

(continuously measured!) and

$$\begin{aligned} \frac{d\Delta x}{dt} &= \Delta v \\ \frac{d\Delta v}{dt} &= -\theta g. \end{aligned}$$

Now into the formula (5.12) can be substituted

$$g\theta = -\frac{d\Delta v}{dt},$$

with the result

$$\frac{d^2\theta}{dt^2} = \frac{1}{R}\frac{d\Delta v}{dt}.$$

By integrating (leaving out one $\frac{d}{dt}$ from each side) we obtain

$$\frac{d\theta}{dt} = \frac{1}{R}\Delta v,$$

and as the complete Kalman formula we obtain:

$$\frac{d}{dt}\begin{bmatrix} \Delta x \\ \Delta v \\ \theta \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -g \\ 0 & \frac{1}{R} & 0 \end{bmatrix}\begin{bmatrix} \Delta x \\ \Delta v \\ \theta \end{bmatrix} + \begin{bmatrix} 0 \\ n_a \\ n_g \end{bmatrix},$$

where we have added the possible 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 $x_0(t)$ and $v_0(t)$, and with the help of the Kalman filter $\Delta x, \Delta v$ and θ.

This is easily generalizised to two dimensions. In this way, a "navigator" may be built on the surface of the Earth. Note that in the solution both the angle of deflection θ of the carriage and the speed disturbance Δv (and also the position disturbance Δx) "oscillate" harmonically like the Schuler pendulum[6], 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.

[6] If the angle θ has, e.g., an amplitude $A_\theta = 1'' = 4.8 \cdot 10^{-6}\,\text{rad}$, it follows from formula

$$\frac{d\Delta v}{dt} = -g\theta,$$

that

- ▷ Δv's amplitude is $A_{\Delta v} = -g\sqrt{\frac{R}{g}}A_\theta = 4\,\text{cm}\,\text{s}^{-1}$, and
- ▷ Δx's amplitude is $A_{\Delta x} = \sqrt{\frac{R}{g}}A_{\Delta v} = 3\,\text{km}$.

5.9 Initialization of an inertial device

In all the previous theory we assume, that the Earth is a sphere and does not rotate. A physically more realistic theory is very complicated.

Interesting is also, how one *levels* and *orients* an inertial platform. In a state of no motion, the inertial device behaves approximately like a stable table. In this case the accelerometers act as *inclinometers* and through feedback loops we make the gyroscope axes turn into the horizontal plane.

The North orientation is obtained by using the device as a gyro compass, i.e., observing how the local vector of gravity slowly turns about the South-North axis.

On airports one often sees a tableau giving the precise ($\pm 0'.1$) 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.

6 Navigation and satellite orbits

The subjects of this chapter are more extensively presented in the books Poutanen [1998], chapter 3, 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 Kepler orbit

If it is assumed that the satellite moves in a central force field (i.e. a point-like or the sphere-like Earth gravitational field), it follows that the satellite's orbit is a *Kepler orbit.* Johannes Kepler (1571-1630) discovered it based on the observation material on the orbit of Mars by Tycho Brahe (1546-1601) (http://www.cvc.org/science/kepler.htm; http://www.glenbrook.k12.il.us/gbssci/phys/Class/circles/u6l4a.html).

As we have seen, we can describe the satellite's motion in rectangular coordinates like this:

$$\begin{aligned} \frac{d}{dt}\mathbf{x} &= \mathbf{v}; \\ \frac{d}{dt}\mathbf{v} &= -\frac{GM}{\|\mathbf{x}\|^3}\mathbf{x}. \end{aligned}$$

Here $\mathbf{x}$ and $\mathbf{v}$ are the position and velocity vectors in three-dimensional space. The combined vector $\underline{\mathrm{x}} \equiv \begin{bmatrix} \mathbf{x} & \mathbf{v} \end{bmatrix}^T = \begin{bmatrix} x & y & z & \dot{x} & \dot{y} & \dot{z} \end{bmatrix}^T$ is the *state vector* of the system.

Elements of the Kepler-orbit are only an alternative way of writing the state vector. See http://www.orbitessera.com/html/body_orbital_description.html, where is found a good description of all the Kepler elements, as well as useful links.

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 equatoral 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.$1 - e^2 = \frac{b^2}{a^2}$, where b is the semi-minor axis.

ν, E, M The position of the satellite in its orbit as the function of time:

$\nu\,(t)$ true anomaly

$E\,(t)$ eccentric anomaly

$M\,(t)$ mean anomaly

The connections between them:

$$E\,(t) = M\,(t) + e \sin E(t) \tag{6.1}$$

$$\frac{\tan \frac{1}{2}\nu(t)}{\tan \frac{1}{2}E(t)} = \sqrt{\frac{1+e}{1-e}}$$

Cf. 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 τ:

$$M\,(t) \equiv 2\pi \frac{t-\tau}{P}.$$

E and ν are purely geometrical quantities.

In the figure the angle θ 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:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
6 37	8 40	10 30	12 32	14 31	16 33	18 31	20 33	22 36	0 34	2 36	4 34

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: $15° = 1^{\mathrm{h}}$, $1° = 4^{\mathrm{m}}$, $15' = 1^{\mathrm{m}}$.

The precision of your result will be $\pm 4^{\mathrm{m}}$, because this table is not really constant from year to year: it varies with the leap year cycle.

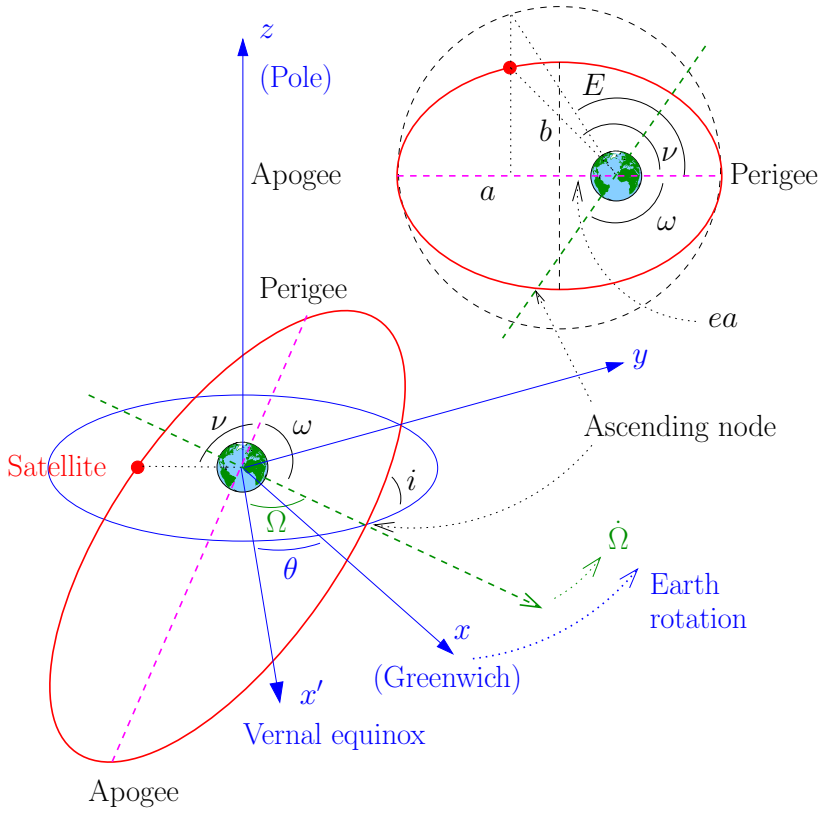


Figure 6.1: Kepler's orbital elements

So we have obtained an alternative way of presenting the state vector:

$$\underline{\mathrm{a}} = \begin{bmatrix} a \\ e \\ M \\ i \\ \omega \\ \Omega \end{bmatrix}$$

In a central force field the elements of this state vector are constants except $M(t)$, cf. above. In case the force field is not 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 Ω . This kind of time dependent Kepler elements (like for example $\Omega(t)$) are called *osculating elements*[2].

6.2 Computing rectangular coordinates from the orbital elements

We can calculate the satellite's instantaneous radius

$$r = a(1 - e\cos E) = \frac{a(1-e^2)}{1 + e\cos\nu},$$

[2]from Latin *ōsculārī*, to kiss

where E can be calculated from M by iterating the formula 6.1.

The time derivative of r is

$$\frac{dr}{dt} = ae \sin E \frac{dE}{dt};$$

from Eq. (6.1), the definintion of E, we get

$$\begin{aligned} \frac{dE}{dt} &= \frac{dM}{dt} + e \cos E \frac{dE}{dt} = \frac{2\pi}{P} + e \cos E \frac{dE}{dt} \quad \Rightarrow \\ \Rightarrow \quad \frac{dE}{dt} &= \frac{2\pi}{P\,(1 - e \cos E)}, \end{aligned}$$

yielding upon substitution

$$\frac{dr}{dt} = \frac{2\pi ae \sin E}{P\,(1 - e \cos E)} = \frac{2\pi a^2 e \sin E}{Pr}.$$

After that in the orbital plane

$$\begin{bmatrix} x \\ y \end{bmatrix} = r \begin{bmatrix} \cos \nu \\ \sin \nu \end{bmatrix}.$$

After this we can transform this two-dimensional vector into a three-dimensional space vector by using the rotation angles ω, i, Ω. If we write

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} r \cos \nu \\ r \sin \nu \\ 0 \end{bmatrix},$$

we get geocentrically

$$\mathbf{X} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R\mathbf{x},$$

where

$$R = \begin{bmatrix} \cos\Omega \cos\omega & -\cos\Omega \sin\omega & \sin\Omega \sin i \\ -\sin\Omega \sin\omega \cos i & -\sin\Omega \cos\omega \cos i & \\ \sin\Omega \cos\omega & -\sin\Omega \sin\omega & -\cos\Omega \sin i \\ +\cos\Omega \sin\omega \cos i & +\cos\Omega \cos\omega \cos i & \\ \sin\omega \sin i & \cos\omega \sin i & \cos i \end{bmatrix}.$$

The geocentric coordinates thus obtained are in an inertial (i.e., 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 θ_0 is Greenwich sidereal time, and put in the matrix formula above ℓ instead of Ω.

The velocity vector is obtained by differentiating with respect to time:

$$\begin{aligned}\frac{d}{dt}\mathbf{x} &= \begin{bmatrix} -r\sin\nu \\ r\cos\nu \\ 0 \end{bmatrix} \frac{d\nu}{dM}\frac{dM}{dt} + \frac{dr}{dt}\begin{bmatrix} \cos\nu \\ \sin\nu \\ 0 \end{bmatrix} \\ &= \frac{2\pi}{P}\begin{bmatrix} -r\sin\nu \\ r\cos\nu \\ 0 \end{bmatrix}\frac{d\nu}{dM} + \frac{2\pi ae\sin E}{P\left(1-e\cos E\right)}\begin{bmatrix} \cos\nu \\ \sin\nu \\ 0 \end{bmatrix};\end{aligned}$$

finding the derivative $\frac{d\nu}{dM}$ is left as a (hard) exercise, and note, that in case of a circular orbit $M = \nu$ i.e., $\frac{d\nu}{dM} = 1$.

6.3 Exercises

Kepler orbit

1. The Kepler state vector's dynamic model. Assuming that the force field is central, write explicitly the following dynamic model equation:

$$\frac{d}{dt}\Delta\underline{\mathrm{a}} = F\cdot\Delta\underline{\mathrm{a}},$$

   where $\underline{\mathrm{a}} = \begin{bmatrix} a & e & M & i & \omega & \Omega \end{bmatrix}^T$. For this you need to *linearize*: the delta quantities are referred to suitable approximate values. You also need *Kepler's third law*:

$$GM\cdot P^2 = 4\pi^2 a^3.$$

2. Due to flattening of the Earth, the ascending node's right ascension Ω changes slowly according to the following formula (circular orbit assumed, $e \approx 0$):

$$\dot{\Omega} = -\frac{3}{2}\sqrt{\frac{GM}{a^3}}\left(\frac{a_e}{a}\right)^2 J_2\cos i.$$

   a_e is the equatorial radius of the Earth J_2 the so called dynamic flattening (a dimensionless number).

   How does this affect the above matrix F?

3. [Difficult.] How does one transform a rectangular state vector $\underline{\mathrm{x}}$ into a Kepler vector $\underline{\mathrm{a}}$ and the reverse? In other words, we want in the following equation

$$\underline{\mathrm{x}} = A\underline{\mathrm{a}}$$

   the matrix A written out in components (*Linearization*!). For simplicity assume that e is small.

   Hint: write first $\underline{\mathrm{x}}$ as a function of $\underline{\mathrm{a}}$ and calculate the partial derivatives.

4. In a central force field, if we write

$$\underline{\mathrm{x}}\,(t_1) = \Phi_0^1 \underline{\mathrm{x}}\,(t_0)\,,$$

find the matrix Φ_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

$$\left[\begin{array}{ccc} X\,(t) & Y\,(t) & Z\,(t) \end{array}\right]^T$$

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 observational quantity as a function of the elements of the state vector $\underline{\mathrm{x}}$, and linearize.

7. Write the observation equations for the case of GPS, where the observation quantity is the *pseudo-range* (pseudorandom code measurement) to the satellite. What new problem comes up?

8. What new problem comes up in the case, that the observational quantity is the *carrier phase*?

7 Use of Hill co-ordinates

The Hill co-ordinate frame was invented by George W. Hill[1] in connection with the study of the motion of the Moon. The idea is to describe the motion, instead of in an inertial co-ordinate system (x, y, z) centred on the centre of motion (i.e., the Sun), in a co-rotating, non-inertial frame (u, v, w), 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 relative to a fictitious

[1] 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 where the effects of Earth and Sun are of similar magnitude.

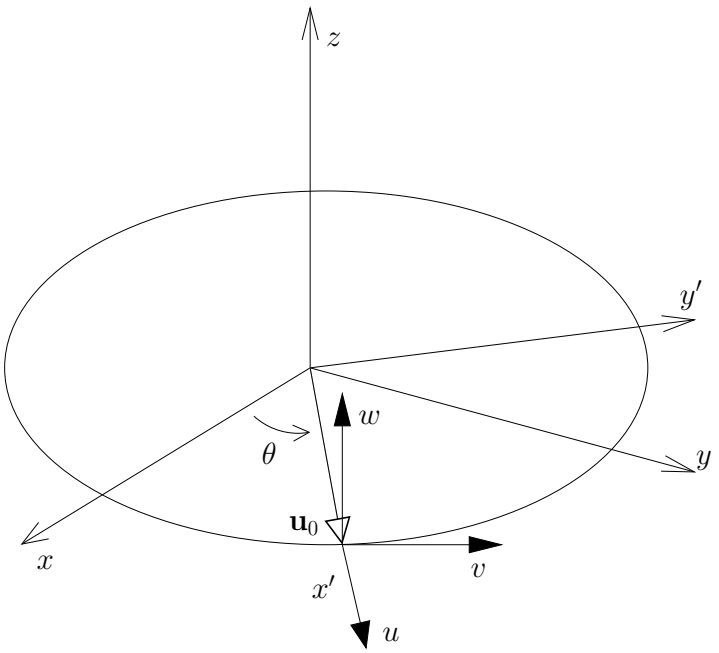


Figure 7.1: Hill co-ordinate frame

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 rendez-vous problem.

Write

$$\mathbf{u} = R\,\mathbf{x} - \mathbf{u}_0$$

where $\mathbf{u} = \begin{bmatrix} u & v & w \end{bmatrix}^T$, $\mathbf{x} = \begin{bmatrix} x & y & z \end{bmatrix}^T$, $\mathbf{u}_0 = \begin{bmatrix} r_0 & 0 & 0 \end{bmatrix}^T$ and the rotation matrix

$$R = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$\mathbf{x}$ is in the inertial system, $\mathbf{u}$ is in the system co-rotating with the satellite; the u axis points outward ("upward"), the v axis forward in the direction of flight, and the w axis (i.e., the z axis) perpendicularly out of the orbital plane to "port".

The satellite moves at constant velocity in a circular orbit: the angular velocity in according to Kepler's third law

$$n = \frac{d\theta}{dt} = \sqrt{\frac{GM}{r_0^3}}.$$

r_0 is the orbital radius and also the distance of the (u, v, w) system's origin from that of the (x, y, z) system.

We can invert the above formula as

$$\mathbf{x} = R^{-1}\left(\mathbf{u} + \mathbf{u}_0\right) = R^T\left(\mathbf{u} + \mathbf{u}_0\right),$$

because for an orthogonal matrix $RR^T = I \;\Leftrightarrow\; R^{-1} = R^T$.

7.1 Transformation from inertial system to Hill system

Derive formulas for the vector $\mathbf{x}$ and the matrix R's first and second derivatives and substitute. After that, multiply both sides of the equation with the matrix R.

We obtain by differentiation (product rule):

$$\begin{aligned} \dot{\mathbf{x}} &= \dot{R}^T\left(\mathbf{u} + \mathbf{u}_0\right) + R^T\dot{\mathbf{u}}, \\ \ddot{\mathbf{x}} &= \ddot{R}^T\left(\mathbf{u} + \mathbf{u}_0\right) + 2\dot{R}^T\dot{\mathbf{u}} + R^T\ddot{\mathbf{u}}. \end{aligned}$$

Here the derivatives of matrix R are (chain rule):

$$\dot{R} = \frac{dR}{dt} = \frac{dR}{d\theta}\frac{d\theta}{dt} = \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n$$

and

$$\ddot{R} = \frac{d^2R}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 = \begin{bmatrix} -\cos\theta & \sin\theta & 0 \\ -\sin\theta & -\cos\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n^2.$$

Substitution yields:

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} -\cos\theta & -\sin\theta & 0 \\ \sin\theta & -\cos\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n^2 \begin{bmatrix} u+r_0 \\ v \\ w \end{bmatrix} + \\ + 2\begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} n \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \\ + \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

By multiplying from the left with the R matrix we obtain[2]:

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} n^2 \begin{bmatrix} u+r_0 \\ v \\ w \end{bmatrix} + \\ + 2\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} n \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}. \qquad (7.1)$$

7.2 Series expansion for a central force field

The formula for a central force field in the (x, y, z) system is

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

i.e., (multiplying from the left by the R matrix):

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -\frac{GM}{\|\mathbf{x}\|^3} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \\ = -\frac{GM}{\|\mathbf{x}\|^3} \begin{bmatrix} u+r_0 \\ v \\ w \end{bmatrix} = -\frac{GM}{\|\mathbf{x}\|^3} (\mathbf{u} + \mathbf{u}_0),$$

[2]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 (α, β) axes turn with the satellite and remain in the same direction as the axes (u, v).

where

$$\|\mathbf{x}\| = \sqrt{x^2+y^2+z^2} = \|\mathbf{u}+\mathbf{u}_0\| = \sqrt{(u+r_0)^2+v^2+w^2}.$$

The Taylor expansion about the origin of the (u,v,w) system now yields

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -\frac{GM}{r_0^3}\mathbf{u}_0 + M\cdot\mathbf{u} = \\ = -\frac{GM}{r_0^3}\begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + M\cdot\begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

where the *gravity gradient matrix* M consists of the partial derivatives:

$$M = \begin{bmatrix} \frac{\partial}{\partial u} & \frac{\partial}{\partial v} & \frac{\partial}{\partial w} \end{bmatrix} \left[-\frac{GM}{\|\mathbf{x}\|^3}\begin{bmatrix} u+r_0 \\ v \\ w \end{bmatrix} \right] \Bigg|_{u,v,w=0} = \\ = -\frac{GM}{r_0^3}\begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(see eq. (3.5) applied to the situation $x = r_0, y = 0, z = 0$), and

$$\frac{GM}{r_0^3} = n^2$$

according to Kepler III.

By combining we obtain

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} = -n^2 \left\{ \begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right\}. \tag{7.2}$$

7.3 Equations of motion in the Hill system

By combining the equations (7.1) and (7.2) we obtain the result (in the absence of external forces)

$$0 = n^2 \left\{ \begin{bmatrix} r_0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right\} + \\ + \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} n^2 \begin{bmatrix} u+r_0 \\ v \\ w \end{bmatrix} + \\ + 2\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} n \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

Simplifying

$$0 = n^2 \begin{bmatrix} -3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} + 2 \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} n \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \begin{bmatrix} \ddot{u} \\ \ddot{v} \\ \ddot{w} \end{bmatrix}.$$

As the end result, by extracting separately the equations for the u, v and w components[3]:

$$\begin{aligned} \ddot{u} &= 2n\dot{v} + 3n^2 u \\ \ddot{v} &= -2n\dot{u} \\ \ddot{w} &= -n^2 w \end{aligned}$$

where the last is a classical harmonic oscillator.

7.4 Solving the Hill equations

Skip over this at first reading; complicated math.

w equation

We attempt first the easiest equation, the third one:

$$\ddot{w} = -n^2 w.$$

Let us first try the general periodic solution,

$$w(t) = A \sin(Bt + C).$$

Substitution yields

$$A \cdot B^2 \cdot -\sin(Bt + C) = -n^2 \cdot A \sin(Bt + C),$$

from which we conclude that

$$B = \pm n.$$

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 (*show*)

$$w(t) = A_1 \sin nt + A_2 \cos nt,$$

where $A_1 = \pm A \cos C$ and $A_2 = A \sin C$, again arbitrary constants.

[3] We can spot here the pseudo-forces occurring in a rotating co-ordinate frame, the centrifugal contributions (slightly hidden) $n^2 u$ and $n^2 v$, dependent upon place only, and the Coriolis terms $2n\dot{v}$ and $-2n\dot{u}$ which are velocity dependent.

u, v equations

$$\begin{aligned} \ddot{u} &= 2n\dot{v} + 3n^2 u \\ \ddot{v} &= -2n\dot{u} \end{aligned}$$

These are to be solved together. Let's try again a periodic solution:

$$\begin{aligned} u(t) &= A\sin nt + B\cos nt, \\ v(t) &= C\sin nt + D\cos nt. \end{aligned}$$

Substitution yields

$$\begin{aligned} -n^2(A\sin nt + B\cos nt) &= 2n^2(C\cos nt - D\sin nt) + 3n^2(A\sin nt + B\cos nt) \\ -n^2(C\sin nt + D\cos nt) &= -2n\cdot n(A\cos nt - B\sin nt) \end{aligned}$$

Consider now the sine and cosine terms separately and express C and D into A and B.

We find the general solution

$$\begin{aligned} u(t) &= A\sin nt + B\cos nt, \\ v(t) &= -2B\sin nt + 2A\cos nt. \end{aligned}$$

In matrix form:

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} A & B \\ -2B & 2A \end{bmatrix} \begin{bmatrix} \sin nt \\ \cos nt \end{bmatrix}.$$

This solution we call the *libration movement*, a periodic movement, the centre of which is the origin $u = v = 0$. In fact, the satellite describes a Kepler orbit that is elliptical, although the period is the same as that of the Hill system, $2\pi/n$.

7.5 Another solution

This isn't however end of story. Let's try for a change a *linear* non-periodic solution:

$$\begin{aligned} u(t) &= Et + F, \\ v(t) &= Gt + H. \end{aligned}$$

Substitute this into the original differential equation set and express E and G into F and H.

$$\begin{aligned} 0 &= 2nG + 3n^2(Et + F) \\ 0 &= -2nE \end{aligned}$$

from which

$$\begin{aligned} E &= 0, \\ G &= -\frac{3}{2}nF. \end{aligned}$$

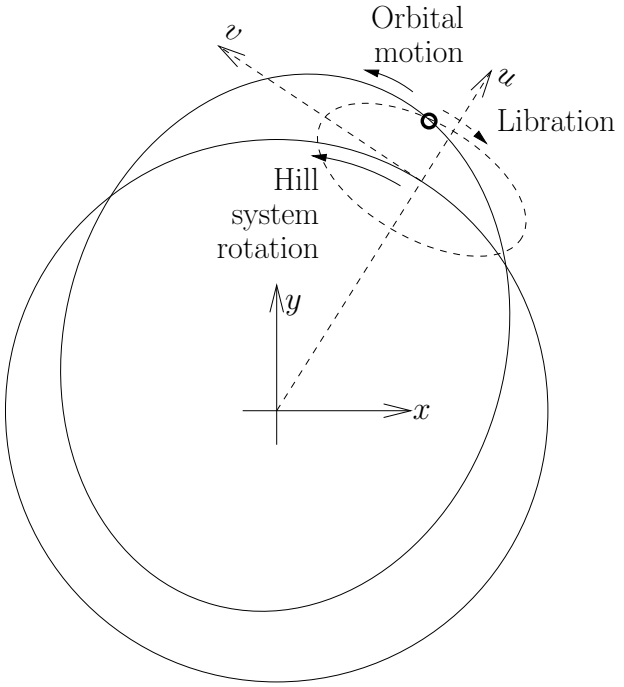


Figure 7.2: Libration

We obtain as the solution

$$\begin{aligned} u(t) &= F, \\ v(t) &= -\frac{3}{2}Fnt + H, \end{aligned}$$

F and H 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 III!) and the satellite is at the moment $t = 0$ in its orbit ahead of the origin of the (u, v, w) system by an amount H.

Combining solutions

Because the system of differential equations is linear, we may freely combine the above periodic and linear solutions.

7.6 The state transition matrix

The general case

Let us look only at the (u, v) plane. Then the general solution is

$$\begin{aligned} u(t) &= A\sin nt + B\cos nt + F, \\ v(t) &= -2B\sin nt + 2A\cos nt - \frac{3}{2}Fnt + H. \end{aligned} \tag{7.3}$$

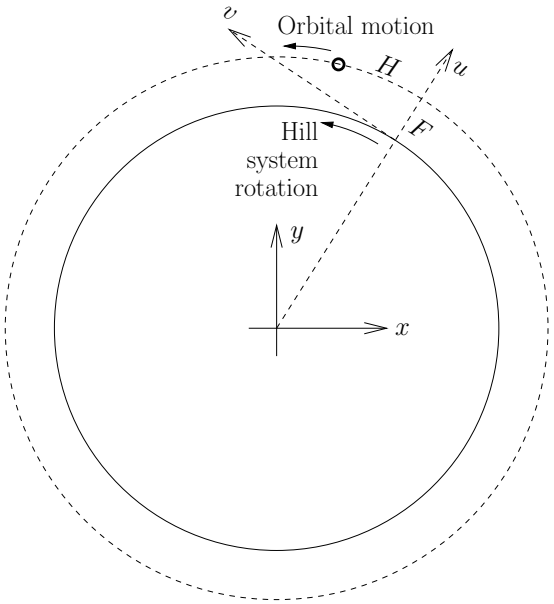


Figure 7.3: Linear drift

We obtain the velocity components too by differentiating:

$$\begin{aligned} \dot{u}(t) &= nA\cos nt - nB\sin nt, \\ \dot{v}(t) &= -2nA\sin nt - 2nB\cos nt - \frac{3}{2}Fn. \end{aligned} \tag{7.4}$$

We write for the initial epoch t_0:

$$\begin{aligned} u(t_0) &= A\sin nt_0 + B\cos nt_0 + F, \\ v(t_0) &= -2B\sin nt_0 + 2A\cos nt_0 - \frac{3}{2}Fnt_0 + H, \\ \dot{u}(t_0) &= nA\cos nt_0 - nB\sin nt_0, \\ \dot{v}(t_0) &= -2nA\sin nt_0 - 2nB\cos nt_0 - \frac{3}{2}Fn. \end{aligned}$$

We write for the epoch t_1, using the sum formulas for sine and cosine:

$$\begin{aligned} u(t_1) &= u(t_0+\Delta t) = \\ &= A\sin n(t_0+\Delta t) + B\cos n(t_0+\Delta t) + F = \\ &= A\sin nt_0\cos n\Delta t + A\cos nt_0\sin n\Delta t + B\cos nt_0\cos n\Delta t - B\sin nt_0\sin n\Delta t + F = \\ &= \cos n\Delta t\cdot(A\sin nt_0 + B\cos nt_0) + F + \sin n\Delta t\cdot(A\cos nt_0 - B\sin nt_0) = \\ &= u(t_0) + (\cos n\Delta t - 1)(A\sin nt_0 + B\cos nt_0) + \sin n\Delta t\cdot(A\cos nt_0 - B\sin nt_0) = \\ &= u(t_0) + (\cos n\Delta t - 1)(A\sin nt_0 + B\cos nt_0) + \sin n\Delta t\cdot\frac{1}{n}\dot{u}(t_0). \end{aligned}$$

Similarly

$$\begin{aligned}
v(t_1) &= 2A\cos n(t_0+\Delta t) - 2B\sin n(t_0+\Delta t) - \frac{3}{2}Fnt_1 + H = \\
&= 2A\cos nt_0\cos n\Delta t - 2A\sin nt_0\sin n\Delta t - 2B\sin nt_0\cos n\Delta t - 2B\cos nt_0\sin n\Delta t - \\
&\quad -\frac{3}{2}Fnt_1 + H = \\
&= \cos n\Delta t\cdot(2A\cos nt_0 - 2B\sin nt_0) - \sin n\Delta t\cdot(2A\sin nt_0 + 2B\cos nt_0) - \frac{3}{2}Fnt_1 + H = \\
&= v(t_0) + (\cos n\Delta t - 1)(2A\cos nt_0 - 2B\sin nt_0) - \sin n\Delta t\cdot(2A\sin nt_0 + 2B\cos nt_0) - \\
&\quad -\frac{3}{2}Fn\Delta t = \\
&= v(t_0) + (\cos n\Delta t - 1)\frac{2}{n}\dot{u}(t_0) - \sin n\Delta t\cdot(2A\sin nt_0 + 2B\cos nt_0) - \frac{3}{2}Fn\Delta t.
\end{aligned}$$

Substituting into this from the $u(t)$ formula

$$F = u(t_0) - A\sin nt_0 - B\cos nt_0$$

we obtain

$$\begin{aligned}
v(t_1) &= v(t_0) + (\cos n\Delta t - 1)\frac{2}{n}\dot{u}(t_0) - \sin n\Delta t\cdot(2A\sin nt_0 + 2B\cos nt_0) - \frac{3}{2}u(t_0)n\Delta t \\
&\quad +\frac{3}{2}n\Delta t\cdot(A\sin nt_0 + B\cos nt_0) = \\
&= v(t_0) + (\cos n\Delta t - 1)\frac{2}{n}\dot{u}(t_0) + \left(\frac{3}{2}n\Delta t - 2\sin n\Delta t\right)(A\sin nt_0 + B\cos nt_0) - \\
&\quad -\frac{3}{2}u(t_0)n\Delta t.
\end{aligned}$$

Next:

$$\begin{aligned}
\dot{u}(t_1) &= nA(\cos nt_0\cos n\Delta t - \sin nt_0\sin n\Delta t) - nB(\sin nt_0\cos n\Delta t + \cos nt_0\sin n\Delta t) = \\
&= \cos n\Delta t\cdot(nA\cos nt_0 - nB\sin nt_0) - \sin n\Delta t\cdot(nA\sin nt_0 + nB\cos nt_0) = \\
&= \cos n\Delta t\cdot\dot{u}(t_0) - \sin n\Delta t\cdot(nA\sin nt_0 + nB\cos nt_0)
\end{aligned}$$

and

$$\begin{aligned}
\dot{v}(t_1) &= -2nA\cdot(\sin nt_0\cos n\Delta t + \cos nt_0\sin n\Delta t) - 2nB\cdot(\cos nt_0\cos n\Delta t - \sin nt_0\sin n\Delta t) - \\
&\quad -\frac{3}{2}Fn = \\
&= \cos n\Delta t\cdot(-2nA\sin nt_0 - 2nB\cos nt_0) + \sin n\Delta t\cdot(-2nA\cos nt_0 + 2nB\sin nt_0) - \\
&\quad -\frac{3}{2}Fn = \\
&= \dot{v}(t_0) - (\cos n\Delta t - 1)(2nA\sin nt_0 + 2nB\cos nt_0) - 2\sin n\Delta t\cdot\dot{u}(t_0).
\end{aligned}$$

Calculate now by combining the $u(t_0)$- and $\dot{v}(t_0)$- formulas:

$$\begin{aligned}
\frac{3}{2}nu(t_0) + \dot{v}(t_0) &= -\frac{1}{2}n(A\sin nt_0 + B\cos nt_0) \Rightarrow \\
\Rightarrow\ A\sin nt_0 + B\cos nt_0 &= -\left(3u(t_0) + \frac{2}{n}\dot{v}(t_0)\right).
\end{aligned}$$

We obtain by substitution

$$\begin{aligned}
u(t_1) &= u(t_0) - (\cos n\Delta t - 1)\left(3u(t_0) + \frac{2}{n}\dot{v}(t_0)\right) + \sin n\Delta t \cdot \frac{1}{n}\dot{u}(t_0), \\
v(t_1) &= v(t_0) - (\cos n\Delta t - 1)\left(\frac{2}{n}\dot{u}(t_0)\right) - \left(\frac{3}{2}n\Delta t - 2\sin n\Delta t\right)\cdot\left(3u(t_0) + \frac{2}{n}\dot{v}(t_0)\right) - \\
&\quad -\frac{3}{2}u(t_0)\, n\Delta t, \\
\dot{u}(t_1) &= \cos n\Delta t \cdot \dot{u}(t_0) + \sin n\Delta t \cdot (3nu(t_0) + 2\dot{v}(t_0)), \\
\dot{v}(t_1) &= \dot{v}(t_0) + (\cos n\Delta t - 1)(6n\dot{u}(t_0) + 4\dot{v}(t_0)) - 2\sin n\Delta t \cdot \dot{u}(t_0).
\end{aligned}$$

As a matrix formula:

$$\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_1) = \begin{bmatrix} 4 - 3\cos n\Delta t & 0 & \frac{\sin n\Delta t}{n} & (\cos n\Delta t - 1)\frac{2}{n} \\ 6\sin n\Delta t - 6n\Delta t & 1 & -(\cos n\Delta t - 1)\frac{2}{n} & (4\sin n\Delta t - 3n\Delta t)\frac{1}{n} \\ 3n\sin n\Delta t & 0 & \cos n\Delta t & 2\sin n\Delta t \\ 0 & 0 & 6(\cos n\Delta t - 1) - 2\sin n\Delta t & 4\cos n\Delta t - 3 \end{bmatrix} \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0).$$

The case of small Δt

Write the system of differential equations

$$\begin{aligned}
\ddot{u} &= 2n\dot{v} + 3n^2 u \\
\ddot{v} &= -2n\dot{u}
\end{aligned}$$

as follows:

$$\frac{d}{dt}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}$$

i.e., for a small time step[4] Δt:

$$\begin{aligned}
\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_1) &= \begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) + \Delta t \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3n^2 & 0 & 0 & 2n \\ 0 & 0 & -2n & 0 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0) = \\
&= \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 3n^2\Delta t & 0 & 1 & 2n\Delta t \\ 0 & 0 & -2n\Delta t & 1 \end{bmatrix}\begin{bmatrix} u \\ v \\ \dot{u} \\ \dot{v} \end{bmatrix}(t_0).
\end{aligned}$$

You may verify, for each matrix element, that this is the same as the above in the limit $\Delta t \to 0$.

[4] "Small" in relation to the orbital period, i.e., $n\Delta t \ll 1$.

8 Airborne gravimetry and gradiometry

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 $\mathbf{x}(t_0)$ and the starting velocity $\mathbf{v}(t_0)$ we can calculate forward to get the instantaneous position and speed $\mathbf{x}(t), \mathbf{v}(t)$. However, if there is a *independent* source of information that gives the current place and velocity precisely enough – such as GPS – then we can harness inertial technology to survey the Earth's gravity field.

With the help of a well working GPS 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 the GPS system, continuously tracking the satellite's accurate three dimensional position.

Let the airplane's or satellite's position as a function of time be $\mathbf{x}(t)$, and its discrete measurement time series $\mathbf{x}_i \equiv \mathbf{x}(t_i)$. Then the geometrical acceleration can be approximated as follows:

$$\left.\frac{d^2}{dt^2}\mathbf{x}\right|_{t_i} \approx \frac{\mathbf{x}_{i+1} + \mathbf{x}_{i-1} - 2\mathbf{x}_i}{\Delta t^2},$$

where Δt is the interval between successive epochs $t_{i+1} - t_i$.

Let us assume that at the same time the airplane's *sensed* acceleration $\mathbf{a}$ is measured ("gravity") for example with acceleration sensors. At this point, for simplicity, we also assume that $\mathbf{x}$ and $\mathbf{a}$ are given in the same coordinate system, i.e., the directions of the acceleration measurement axes are the same as those of the location co-ordinate axes.

Then in an inertial reference system it holds that:

$$\mathbf{g} = \frac{d^2}{dt^2}\mathbf{x} + \mathbf{a}, \tag{8.1}$$

so:

> *gravitation* $\mathbf{g}$ is the sum of the "gravity" $\mathbf{a}$ felt inside of a vehicle, and geometrical acceleration.

8.1 Vectorial airborne gravimetry

If an airplane carries both an inertial device and a GPS receiver, we can measure both $\frac{d^2}{dt^2}\mathbf{x}\Big|_{t_i}$ and $\mathbf{a}(t_i)$, and we can calculate $\mathbf{g}(t_i)$. This is a method to survey the gravity field from the air. In practice the data streams generated from both the GPS device and the inertial device are fed into a Kalman filter, which outputs the plane's precise route and gravity profile. The gravity comes as a three-dimensional vector; the data rate is typically high, many epochs per second. Because of the airplane's motions, the differences in time of both $\frac{d^2}{dt^2}\mathbf{x}$ and $\mathbf{a}$ are large (thousands of milligals), but the final determination precision of $\mathbf{g}$ can be as good as a couple of mGals.

However, it must be said that this technique, vectorial airborne gravimetry, is not as good in precision as the next technique to be introduced, scalar airborne gravimetry. The reason is that the accelerometers in the inertial device, as precise as they are, suffer more from systematic problems, such as *drift*, than the best gravimeters.

8.2 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 gravitational 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 *stabilized table*; the stabilization is done with the aid of gyroscopes.

The gravimeter measures the gravity acceleration "felt" inside the vehicle, but only in the direction of the local vertical (plumbline). If the direction of the local vertical is $\mathbf{n}$ (*downwards*), the measured quantity is $\langle \mathbf{n} \cdot \mathbf{a} \rangle$.

We can write

$$\langle \mathbf{n} \cdot \mathbf{g} \rangle = \frac{d^2}{dt^2} \langle \mathbf{n} \cdot \mathbf{x} \rangle + \langle \mathbf{n} \cdot \mathbf{a} \rangle = \|\mathbf{g}\| \equiv g, \tag{8.2}$$

because the plumbline is in the direction of gravity.

In practice the equation (8.2) is written in a system rotating with the solid Earth, so we obtain:

$$\begin{aligned} g &= \langle \mathbf{n} \cdot \mathbf{a} \rangle + \frac{d^2}{dt^2} \langle \mathbf{n} \cdot \mathbf{x} \rangle + \left(\frac{v_\mathrm{e}}{R_\mathrm{e} + h} + 2\omega \cos\varphi \right) v_\mathrm{e} + \frac{v_\mathrm{n}^2}{R_\mathrm{n} + h} = \\ &= a_\mathrm{d} - \frac{d}{dt} v_\mathrm{u} + \left(\frac{v_\mathrm{e}}{R_\mathrm{e} + h} + 2\omega \cos\varphi \right) v_\mathrm{e} + \frac{v_\mathrm{n}^2}{R_\mathrm{n} + h}, \end{aligned}$$

where $v_\mathrm{u}, v_\mathrm{e}, v_\mathrm{n}$ are the velocity's "up", "east", "north" components, a_d is the *measured* acceleration inside the vehicle in the "down" direction, ω is the angular velocity of the Earth's rotational motion, and R_n and R_e are the Earth's radii of curvature in the meridional (North-South) and East-West directions. h and φ are the height and latitude. In the formula above, the two last terms are called the Eötvös correction. Cf. Wei and Schwarz [1997].

8.3 The research of gravitation in space

In the formula (8.1) the quantity $\mathbf{a}$ is about the magnitude of the Earth's surface gravity (about $10\,\mathrm{m\,s^{-2}}$), while the geometrical acceleration $\frac{d^2}{dt^2}\mathbf{x}$ is much smaller. In the ideal case this acceleration would be zero, which corresponds to measurements on the surface of the Earth. In both shipborne and airborne gravimetry this geometrical acceleration differs from zero and makes accurate measurement of gravity difficult. The movements of the vehicle are disturbances from the viewpoint of measuring.

In the measurement of the gravity field from space, the situation is the opposite. The local gravity $\mathbf{a}$ acting inside the satellite is zero (*weightlessness*) or very close to zero. The geometrical acceleration $\frac{d^2}{dt^2}\mathbf{x}$ is almost the magnitude of gravity at the Earth 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 GPS system – so-called "high-low satellite-to-satellite tracking" – and also the satellite's own, non-inertial motion $\mathbf{a}$ is measured with the aid of acceleration measurement devices (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.

At this moment there are three different gravity missions in flight or completed: CHAMP, GRACE and GOCE.

- ▷ CHAMP (http://op.gfz-potsdam.de/champ/index_CHAMP.html), a small German satellite, operated from 2000 to 2010 and produced a large amount of data.

- ▷ GRACE (http://www.csr.utexas.edu/grace/), a small American-German satellite pair, measures 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. It has been a great success already. An animation of its results can be found here:http://en.wikipedia.org/wiki/File:Global_Gravity_Anomaly_Animation_over_LAND.gif.

- ▷ GOCE (Gravity Field and Ocean Circulation Explorer) surveyed the Earth's gravity field 2009-2013 in great detail with the help of a so called *gravity gradiometer*, cf. http://www.esa.int/esaLP/LPgoce.html. The GOCE satellite contained a so-called *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 GPS navigation system and accelerometers included, in the case of GOCE even an array – a gradiometer – counting six extremely sensitive accelerometers.

8.4 Using the Kalman filter in airborne gravimetry

Let's start with formula (8.1). We can write (including the "dynamic noise" $\mathbf{n}$):

$$\frac{d^2}{dt^2}\mathbf{x} = \mathbf{a} - \mathbf{g} + \mathbf{n},$$

i.e.,

$$\frac{d}{dt}\begin{bmatrix}\mathbf{v}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{a}-\mathbf{g}\\ \mathbf{v}\end{bmatrix} + \begin{bmatrix}\mathbf{n}_a\\ \mathbf{0}\end{bmatrix}.$$

Here $\mathbf{a} = \mathbf{a}(t)$ is a *measured quantity*, but $\mathbf{g}$ is not.

Write

$$\mathbf{g} = \overrightarrow{\gamma} + \delta\mathbf{g},$$

where $\overrightarrow{\gamma}$ is a suitable reference value (e.g., *normal gravity*) and $\delta\mathbf{g}$ the *gravity disturbance*. We can model $\delta\mathbf{g}$ empirically as a Gauss-Markov process, eq. (4.1), so we can write

$$\frac{d}{dt}\delta\mathbf{g} = -\frac{\delta\mathbf{g}}{\tau} + \mathbf{n}_g,$$

where τ 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:

$$\frac{d}{dt}\begin{bmatrix}\mathbf{v}\\ \mathbf{x}\\ \delta\mathbf{g}\end{bmatrix} = \begin{bmatrix}0_{3,3} & 0_{3,3} & -I_3\\ I_3 & 0_{3,3} & 0_{3,3}\\ 0_{3,3} & 0_{3,3} & -\frac{1}{\tau}I_3\end{bmatrix}\begin{bmatrix}\mathbf{v}\\ \mathbf{x}\\ \delta\mathbf{g}\end{bmatrix} + \begin{bmatrix}\mathbf{a}-\overrightarrow{\gamma}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{n}_a\\ \mathbf{0}\\ \mathbf{n}_g\end{bmatrix}.$$

So the length of the state vector is 9. Note that the matrix is a 3×3 matrix consisting of 3×3 sized elements, i.e., a 9×9 matrix in total.

A more sophisticated way of handling takes into consideration that the gravity $\mathbf{g}$ is a function of place $\mathbf{x}$, which we don't actually know:

$$\mathbf{g}(\mathbf{x}) = \overrightarrow{\gamma}(\mathbf{x}_0) + M\Delta\mathbf{x} + \delta\mathbf{g}$$

or

$$\overrightarrow{\gamma}(\mathbf{x}) = \overrightarrow{\gamma}(\mathbf{x}_0) + M\Delta\mathbf{x},$$

Where $\mathbf{x}_0$ is the *approximate* position, given us by the linearization, see below. Here appears the *gradient matrix* M, Equation 3.5. Then also $\mathbf{x}$ and $\mathbf{v}$ must be linearized, and the difference states $\Delta\mathbf{x} \equiv \mathbf{x} - \mathbf{x}_0$, $\Delta\mathbf{v} = \mathbf{v} - \mathbf{v}_0$ must be used in the state vector, the equations being

$$\frac{d}{dt}\begin{bmatrix}\mathbf{v}_0\\ \mathbf{x}_0\end{bmatrix} = \begin{bmatrix}0_{3,3} & 0_{3,3}\\ I_3 & 0_{3,3}\end{bmatrix}\begin{bmatrix}\mathbf{v}_0\\ \mathbf{x}_0\end{bmatrix} + \begin{bmatrix}\mathbf{a}(t) - \overrightarrow{\gamma}(\mathbf{x}_0)\\ \mathbf{0}\end{bmatrix}.$$

Final result:

$$\frac{d}{dt}\begin{bmatrix}\Delta\mathbf{v}\\ \Delta\mathbf{x}\\ \delta\mathbf{g}\end{bmatrix} = \begin{bmatrix}0_{3,3} & -M & -I_3\\ I_3 & 0_{3,3} & 0_{3,3}\\ 0_{3,3} & 0_{3,3} & -\frac{1}{\tau}I_3\end{bmatrix}\begin{bmatrix}\Delta\mathbf{v}\\ \Delta\mathbf{x}\\ \delta\mathbf{g}\end{bmatrix} + \begin{bmatrix}\mathbf{n}_a\\ \mathbf{0}\\ \mathbf{n}_g\end{bmatrix}.$$

Figure 8.1: Lockheed Hercules C-120 taking off from Camp Summit, Greenland, 3,216 m above sea level. Note the JATO (Jet-Assisted Take-Off) bottles helping out. © 109th Airlift Wing, Air National Guard

The observation equations (updating equations) are again

$$\begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \\ \underline{\ell}_3 \end{bmatrix}_i = \underline{\mathbf{x}}(t_i) + \underline{\mathbf{m}},$$

where the "noise vector" $\underline{\mathbf{m}}$ describes the statistical uncertainty of GPS navigation. For both $\underline{\mathbf{n}}_a$ and $\underline{\mathbf{m}}$ we have to find suitable statistical models (variance matrices Q and R) based on the properties of the measurement devices.

8.5 Present state of airborne gravimetry

One of the first successful airborne gravimetric projects was Brozena [1991] , Greenland's gravity survey.

Many later measurements, often in Arctic or Antarctic locations, can be mentioned [Forsberg et al., 1996, 2011]. The logistics requirements of working there are typically "challenging", see figure 8.1.

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 expecially if the gravimetric data is meant for the determination of a geoid.

Recent examples of airborne gravity surveys include Ethiopia (Bedada 2010), Mongolia (Munkhtsetseg 2009), Indonesia (2010), and many more.

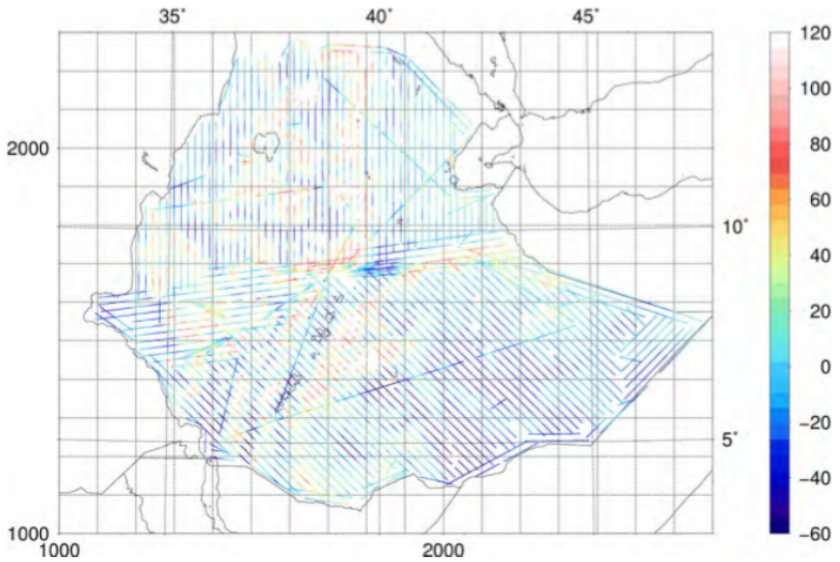


Figure 8.2: Ethiopian airborne gravity survey; measurement points. Gravity anomaly values in mGal

9 GPS-navigation and base stations

About the subject *GPS and Navigation*, cf. e.g. Strang and Borre [1997] pages 495-514.

9.1 Differential navigation

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 (IDGS, Jet Propulsion Lab).

- ▷ A radio broadcast network, a radio-modem pair, or a cell phone can provide the data link.

9.2 RTCM-standard

Radio Technical Commission for Maritime Services (RTCM, http://www.rtcm.org/) SC-104 has defined a standard group for GPS differential corrections. Message types are listed below.

Message Type	Message Title
1	DGPS corrections
2	Delta DGPS corrections
3	Reference station parameters
4	Carrier surveying information
5	Constellation health
6	Null frame
7	Marine radiobeacon almanacs
8	Pseudolite almanacs
9	High rate DGPS corrections
10	P code DGPS corrections
11	C/A code L1/L2 delta corrections
12	Pseudolite station parameters
13	Ground transmitter parameters
14	Surveying auxiliary message
15	Ionospheric/tropospheric message
16	Special message
17	Ephemeris almanac
18	Uncorrected carrier phase measurements
19	Uncorrected pseudorange measurements
20	RTK Carrier phase corrections
21	RTK pseudorange corrections
22-59	Undefined
60-63	Differential Loran C messages

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 pseudo random 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.

9.3 Pseudorange smoothing

In many kinematic applications of GPS, it is advantageous to *smooth* the raw pseudo-range code observables by using the much more smooth and noise-free, but ambiguous,

carrier phase measurements.

Let us assume we have as observations the code measurements p_1 and p_2 (metric units) and the carrier phases ϕ_1 and ϕ_2 (angular units i.e., radians), at a time t.

Firstly, we can construct a *prediction equation* for the current (*a priori*) pseudo-range from the previous one, by

$$p^{-}(t_i) = p(t_{i-1}) + \frac{\lambda}{2\pi}\left(\phi(t_i) - \phi(t_{i-1})\right). \tag{9.1}$$

This equation is valid for both frequencies 1 and 2, and also for the widelane observables defined as:

$$p_{WL} = \frac{f_1 p_1 - f_2 p_2}{f_1 - f_2},\ \phi_{WL} = \phi_1 - \phi_2.$$

Note that eq. (9.1) 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 the current $p(t_i)$ observation, the precision of which can be given as R_i. Now the *correction equation* is

$$p^{+}(t_i) = p^{-}(t_i) + KH\left(p^{-}(t_i) - p(t_i)\right),$$

where $H = [1]$, $K = -P^{-}H^T\left(HP^{-}H^T + R\right)^{-1} = -P^{-}/\left(P^{-} + R\right)$, and thus

$$p^{+}(t_i) = \frac{R_i}{P^{-}(t_i) + R_i}p^{-}(t_i) + \frac{P^{-}(t_i)}{P^{-}(t_i) + R_i}p(t_i).$$

So: the *a posteriori* pseudo-range is a weighted linear combination of the predicted and carrier-smoothed one and the currently observed one.

For the variance propagation we find

$$P^{+}(t_i) = (I + KH)P^{-}(t_i) = \frac{R_i}{P^{-}(t_i) + R_i}P^{-}(t_i).$$

(For the variance propagation in the dynamic model, between epochs, we have simply: $P^{-}(t_i) = P^{+}(t_{i-1})$.)

It is possible to include *cycle slip detection* into the procedure: the testing variate is the difference

$$\left(p^{-}(t_i) - p(t_i)\right),$$

of which we know the mean error to be:

$$\sigma = \sqrt{HP^{-}H^T + R} = \sqrt{P^{-} + R}.$$

This will work 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 post-processing software, but then without the real time advantage). The output $p^{+}(t)$ is significantly smoothed compared to the input one $p(t)$.

9.4 Base station and corrections

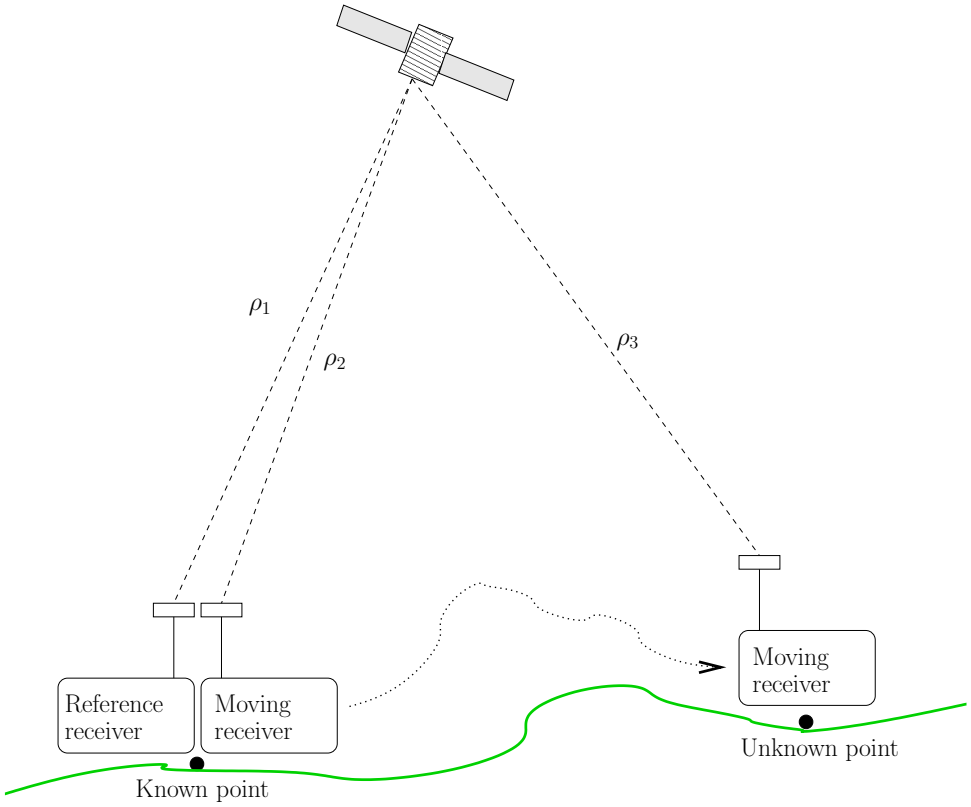
The *base station*, the position of which is measured precisely using static geodetic positioning, sends the RTCM-messages. Because the position is known, its possible to calculate what the pseudo distance to each satellite should be with the help of the satellite orbits. By subtracting this from the measured values we get the *correction* to be coded into the message (message types 1, 2, 20 and 21).[1] 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 accuracy is obtained even hundreds of kilometres from the base station, but cm-accuracy (only RTK-method) succeeds only out to about twenty kilometers.

The transmission of the correction messages can be done using many different techniques: radio, cellular phone, Internet (NTRIP,Networked Transport of RTCM via Internet Protocol). Ala kehittyy nopeasti.

9.5 RTK-measurements

RTK = 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. Cf. figure.



[1] In the case of RTK, often one rather transmits the original phase observations, types 18 and 19, but conceptually the matter is the same.

First, we measure the phase of the carrier wave *with both receivers on the known point:*

$$\begin{aligned}\varphi^S_{R_1} &= -f\frac{\rho_1}{c} - f\Delta\delta_1 + N_1,\\ \varphi^S_{R_2} &= -f\frac{\rho_2}{c} - f\Delta\delta_2 + N_2,\end{aligned}$$

where

$$\begin{aligned}\Delta\delta_1 &= \delta_{R_1} - \delta^{S_{(1)}},\\ \Delta\delta_2 &= \delta_{R_2} - \delta^{S_{(1)}}\end{aligned}$$

are the differences between the receiver's (R_1 reference receiver, R_2 moving receiver) clock offset and the simultaneous satellite clock offset $\delta^{S_{(1)}}$. The index (1) refers to the initial situation with both receivers on the known point.

the quantity N_i is an unknown integer value, the *ambiguity*, chosen so that the φ values are always in the interval $[0, 1)$.

After that, the moving receiver is moved to the unknown point R_3 and we obtain

$$\varphi^S_{R_3} = -f\frac{\rho_3}{c} - f\Delta\delta_3 + N_3,$$

where (now (2) refers to the new situation, on the unknown point):

$$\Delta\delta_3 = \delta_{R_2} - \delta^{S_{(2)}}.$$

The following assumptions:

1. There hasn't happened a "cycle slip", so $N_3 = N_2$.
2. the time elapsed is so short that both $\delta^{S_{(1)}} = \delta^{S_{(2)}}$ and $\delta_{R_2} = \delta_{R_1} + \Delta\delta_{12}$, where $\Delta\delta_{12}$ is a *constant difference* (clock error difference of the clocks of the two receivers); so $\Delta\delta_2 = \Delta\delta_1 + \Delta\delta_{12}$ and $\Delta\delta_3 = \Delta\delta_2 = \Delta\delta_1 + \Delta\delta_{12}$;
3. The reference and moving receivers are in the same place on the known point[2], so that $\rho_1 = \rho_2$.

Then

$$\Delta\varphi^S_{R_1R_2} \equiv \varphi^S_{R_2} - \varphi^S_{R_1} = (N_2 - N_1) + f\Delta\delta_{12} \tag{9.2}$$

and

$$\begin{aligned}\Delta\varphi^S_{R_1R_3} &\equiv \varphi^S_{R_3} - \varphi^S_{R_1} =\\ &= -f\frac{(\rho_3 - \rho_1)}{c} + (N_2 - N_1) + f\Delta\delta_{12}\,.\end{aligned} \tag{9.3}$$

In formula (9.2) the left hand side is *measured*. We get immediately $(N_2 - N_1) + f\Delta\delta_{12}$ to be substituted into the formula (9.3), and as the observation equation we get:

$$\Delta\varphi^S_{R_1R_3} - (N_2 - N_1) - f\Delta\delta_{12} = -f\frac{\rho_3 - \rho_1}{c},$$

[2]more generally, their difference in location is precisely known

where the left hand side is an "observed" quantity, and on the right hand side ρ_3 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 $N_2 - N_1 + f\Delta\delta_{12}$ is a real number, but $N_2 - N_1$ is an integer number. If there are many satellites to be used instead of just one, the satellite being S_k, , several quantities can be calculated on the known point

$$\nu^{S_k} \equiv N_2^{S_k} - N_1^{S_k} + f\Delta\delta_{12} \tag{9.4}$$

where however there is one and the same $\Delta\delta_{12}$. Let's *choose* the integer $N_2^{S_1} - N_1^{S_1}$ so that $f\Delta\delta_{12}$ is minimized (for example!). After that we can *calculate*

$$N_2^{S_k} - N_1^{S_k},\ k = 2, 3, \ldots$$

and they too have to be integers. If not, we have an *adjustment condition* that can be used to slightly improve the value $\Delta\delta_{12}$, for example we can minimize the $\left(N_2^k - N_1^k\right)$:n's sum of squared differences ($k = 1, 2, \ldots$) from among the nearest integers. After this the values $N_2^k - N_1^k$ can be rounded to the nearest integers.

As the final solution 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 $N_2^k - N_1^k$ are affected by the uncertainties of the atmosphere and satellite orbits, and will not be close enough to integers.

Other sources of error

In the most general case the quantities ν^k include not only the clock errors but also delays caused by the ionosphere and neutral atmosphere ("troposphere"). In that case we can write

$$\nu^{S_k} = N_2^k - N_1^k + f\left(\delta_{R_2} - \delta_{R_1}\right) + \frac{d_{12}^{\text{ion}}}{\lambda} + \frac{d_{12}^{\text{trop}}}{\lambda}.$$

In *real time* application both the clock error δ_{R_i} 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 ready to be used immediately.

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

$$\begin{aligned} \nabla\Delta\varphi_{R_1R_2}^{S_1S_2} &\equiv \left(\varphi_{R_2}^{S_2} - \varphi_{R_1}^{S_2}\right) - \left(\varphi_{R_2}^{S_1} - \varphi_{R_1}^{S_1}\right) = \\ &= \left(N_2^2 - N_1^2\right) - \left(N_2^1 - N_1^1\right) + f\Delta\delta^{12}, \end{aligned}$$

where

$$\begin{aligned}\Delta\delta^{12} &= \Delta\delta^2_{12} - \Delta\delta^1_{12} = \\ &= \left\{\left(\delta_{R_2} - \delta^{S_2}\right) - \left(\delta_{R_1} - \delta^{S_2}\right)\right\} - \\ &- \left\{\left(\delta_{R_2} - \delta^{S_1}\right) - \left(\delta_{R_1} - \delta^{S_1}\right)\right\} = \\ &= 0,\end{aligned}$$

and similarly

$$\begin{aligned}\nabla\Delta\varphi^{S_1S_2}_{R_1R_3} &\equiv \left(\varphi^{S_2}_{R_3} - \varphi^{S_2}_{R_1}\right) - \left(\varphi^{S_1}_{R_3} - \varphi^{S_1}_{R_1}\right) = \\ &= -f\frac{(\rho^2_3 - \rho^2_1) - (\rho^1_3 - \rho^1_1)}{c} + \\ &+ \left(N^2_2 - N^2_1\right) - \left(N^1_2 - N^1_1\right) + f\Delta\delta^{12}, \qquad (9.5)\end{aligned}$$

where again $\Delta\delta^{12} = 0$.

In this case the "ν quantity", that is solved by putting the reference receiver and the moving receiver side by side is

$$\nu^{S_1S_2} = \left(N^2_2 - N^2_1\right) - \left(N^1_2 - N^1_1\right)$$

for two satellites S_1 and S_2. *This is an integer.* We observe the quantity $\nabla\Delta\varphi^{S_kS_m}_{R_1R_2}$ to all satellite pairs $(k = 1, \ldots n,\ m = k+1, \ldots, n)$, where n is the number of satellites, and we round to the nearest integer. The values found after that can be used to compute the quantities $\left(\rho^{S_k}_3 - \rho^{S_k}_1\right) - \left(\rho^{S_m}_3 - \rho^{S_m}_1\right)$ from the observations $\nabla\Delta\varphi^{S_kS_m}_{R_1R_3}$.

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". This 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 formula (9.5) is

$$\begin{aligned}\nabla\Delta\varphi^{k_1k_2}_{R_1R_3} &\equiv \left(\varphi^{k_2}_{R_3} - \varphi^{k_2}_{R_1}\right) - \left(\varphi^{k_1}_{R_3} - \varphi^{k_1}_{R_1}\right) = \\ &= -f\frac{\left(\rho^{k_2}_3 - \rho^{k_2}_1\right) - \left(\rho^{k_1}_3 - \rho^{k_1}_1\right)}{c} + \left(N^{k_2}_2 - N^{k_2}_1\right) - \left(N^{k_1}_2 - N^{k_1}_1\right).\end{aligned}$$

Here the quantities

$$\nabla\Delta\rho^{k_1k_3}_{R_1R_3} \equiv \left(\rho^{k_2}_3 - \rho^{k_2}_1\right) - \left(\rho^{k_1}_3 - \rho^{k_1}_1\right)$$

are purely geometric. If we write

$$\rho^{k_i}_3 = \sqrt{\left(X^{k_i} - X_{R_3}\right)^2 + \left(Y^{k_i} - Y_{R_3}\right)^2 + \left(Z^{k_i} - Z_{R_3}\right)^2},\ i = 1, 2,$$

we can see, that the only unknowns here are the position of the moving receiver

$$\begin{bmatrix} X_{R_3} & Y_{R_3} & Z_{R_3} \end{bmatrix}^T .$$

The position of the moving receiver is always known with the accuracy of couple of metres with the help of the GPS-code measurement, when there is 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

$$\nabla\Delta N_{12}^{k_1 k_2} \equiv \left(N_2^{k_2} - N_1^{k_2}\right) - \left(N_2^{k_1} - N_1^{k_1}\right)$$

are integers.

Cf. figure 9.1. Conversely, if there are n satellites, there are $n-1$ different ambiguity values $\nabla\Delta N$. 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 possibillities 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 receiver: the consistent ambiguity combinations belong to the *a three-dimensional subspace* of ambiguity space, one parametrization of which is the co-ordinates $\begin{bmatrix} X_{R_3} & Y_{R_3} & Z_{R_3} \end{bmatrix}^T$, 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 (LAMBDA = Least-squares Ambiguity Decorrelation Adjustment, Teunissen et al. [1997]).

The introduced ambiguity resolution method succeeds only if the distance between the comparison 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, L_1 (1575.42 MHz) and L_2 (1227.60 MHz). The ambiguity resolution is obtained immediately or after only a couple of epochs.

Ambiguity resolution is also possible for longer vectors, but a lot more difficult, more laborious time consuming, because the errors caused by the atmosphere etc. have to be taken into account.

9.6 Network RTK

Tests are ongoing and already done, to implement a network RTK solution: here 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).

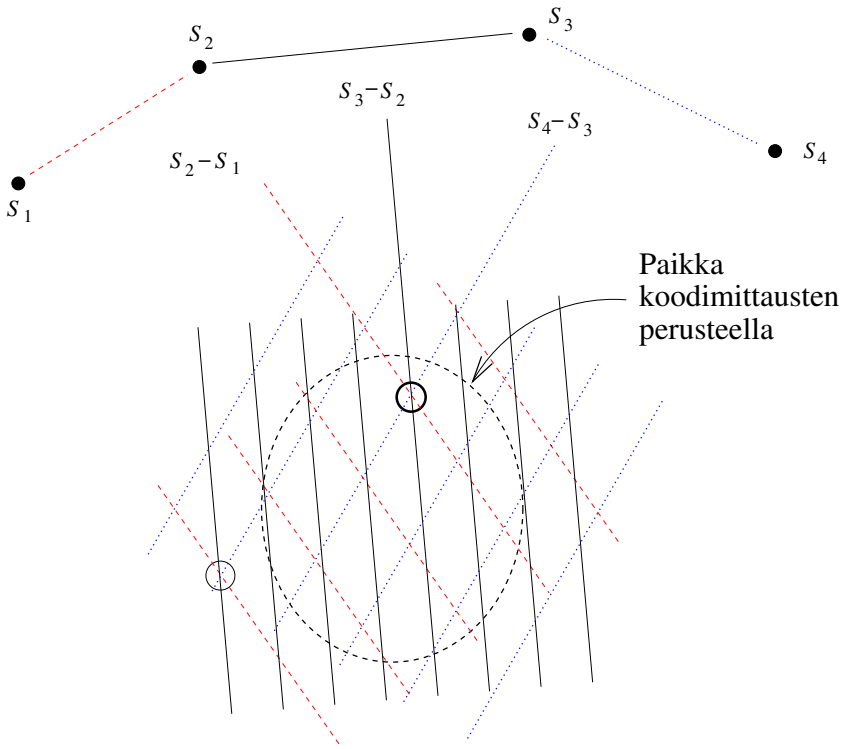


Figure 9.1: Ambiguity resolution

2. *One-to-one* ("singlecast") method: the corrections are computed for one user and sent to him, e.g., 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 etc. In principle this could improve the interpolation results, if the model is good.

In many places, like in Germany, is used Spectra Precision Terrasat *GPS-Network* software (http://www.terrasat.de/applications/refvirtual.htm), that is based on the virtual base station concept. Also in Finland this system is used in Geotrim's GNSSnet network.

9.7 Global DGPS

This system was invented and implemented by the Jet Propulsion Laboratory. The corrections sent via the Internet are globally valid.

IGDG, Internet-based Global Differential GPS. http://gipsy.jpl.nasa.gov/igdg/.

The system works as follows:

Each second a 560 bit message is send to the user. The message includes the three dimensional satellite position corrections (XYZ) and meter level satellite clock corrections to four (4) satellites, and a cm-level residual corrections to 32 satellites.

Thanks to this it is possible after 8 seconds, at the most, to reconstruct all the starting values of orbit and clock corrections to 32 satellites.

The resolution of the clock corrections is 1.5625 cm, the resolution of the orbit corrections is 6.25 cm.

The corrections are sent via Internet to the user using the TCP-protocol.

9.8 RTCM-over-Internet (NTRIP protocol)

"Networked Transport of RTCM via Internet Protocol".

Cf. http://igs.bkg.bund.de/pdf/NtripPaper.pdf. This is a promising method which has also been tested in Finland. From 2012 on, Indagon Oy offers the @Focus service based on NTRIP.

10

Real time systems and networks

Technological navigation will often depend on obtaining external data in real time from a communication network, as well as on on-board processing by equipment and software suitable for real time use. We shall consider those requirements next.

10.1 Communication networks

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.

Example: amplitude modulation

In Figure 10.1 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

$$\begin{aligned} A(t) &= \cos(2\pi F) \cdot \cos(2\pi f) = \\ &= \frac{1}{2}\left[\cos(2\pi[F+f]) + \cos(2\pi[F-f])\right], \end{aligned}$$

so we see that the new wave can be represented as the sum of two frequencies, $F+f$ and $F-f$.

Now, if the modulating wave contains a large number of different frequencies, $0 < f < f_{max}$, the resulting spectrum will contain signal in the full range $(F - f_{max}, F + f_{max})$. We say that the *band width consumption* is $2f_{max}$.

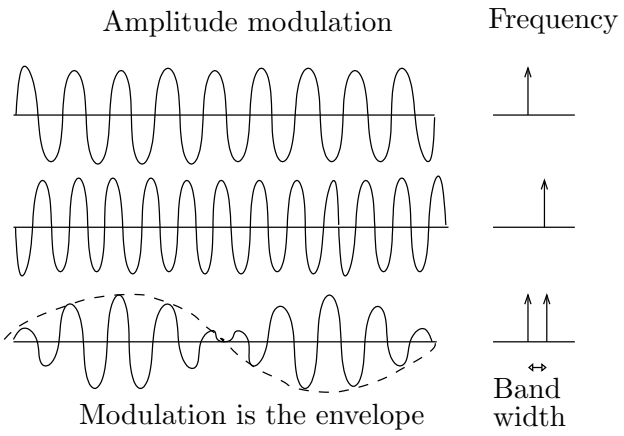


Figure 10.1: Amplitude modulation and bandwidth.

For broadcasting networks, bandwidth is a scarce and valuable resource, to be carefully allocated.

The Nyqvist theorem

One can show that in order to represent a function of time by sample points, the distance Δt between the sample points should never be more than 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 $\widetilde{A}(f)$ representation using the discrete Fourier transform. Numerically, typically the Fast Fourier Transform (FFT) is used.

Now, if we have a modulating function $a(t)$, that has as its highest contained frequency f_{max}, then its shortest contained period is $1/f_{max}$.The number of samples transmitted using amplitude modulation will then be max $2f_{max}$, i.e., precisely the effective bandwidth occupied by the modulated signal.

Switched connection networks

History

The first, still existing and wildly successful switched, or many-to-many, connection network is the telephone network.

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 phone cable, launching the first communications satellite (Telstar), and inventing Unix...

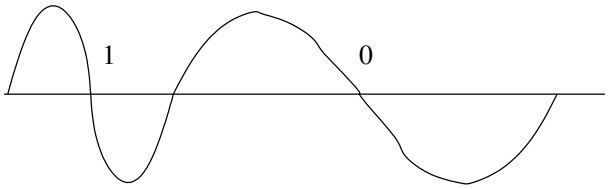


Figure 10.2: FSK-modulation.

Telephone is based on transmitting sound in electric form over a copper cable. This is still the way it happens for the few metres nearest to the consumer, although all in-between equipment is nowadays fully digital. Making a connections between two telephone customers was originally done by hand; already before 1900, the first, mechanical automatic switches were built. A number was dialled using a round disc, sending as many pulses as the number being encoded. This is called "pulse dialling". Today, faster tone dialling has 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 is 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, may soon change this.

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 on the same line (Digital Subscriber Line, DSL).

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 the form of (analogue) sound waves. This is done with a device called a *modem* (modulator-demodulator).

The picture 10.2 shows one technique (Frequency Shift Keying) 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 verify that the data received equals the data sent (Parity check, Cyclic Redundancy Check) even over noisy lines. 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/second is the best achievable.

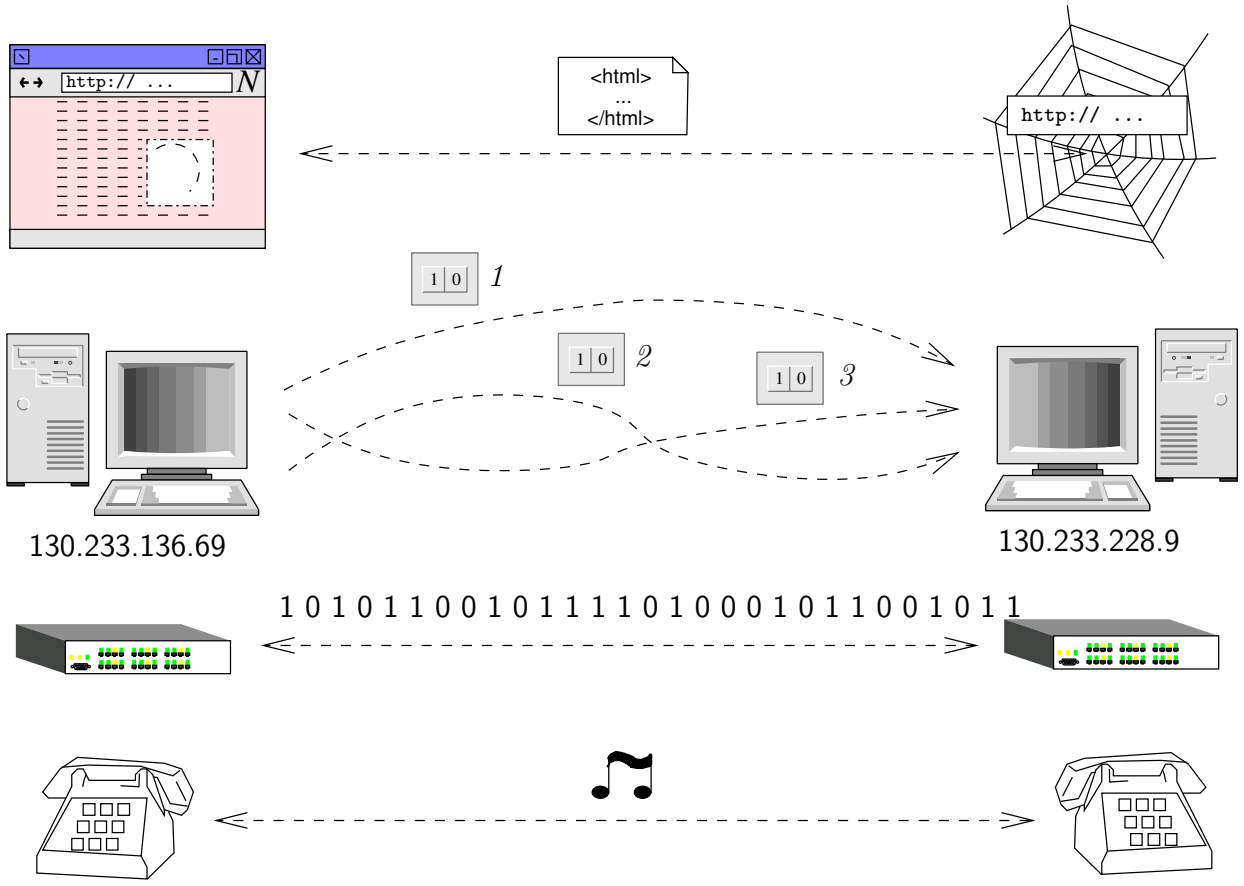


Figure 10.3: An example of a *protocol stack*

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 and 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 creeping down 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 are becoming mostly just general purpose computers, differently programmed. This phenomenon is known as *convergence*.

Mobile phones

Mobile phones based on GSM (Global System for Mobile Communications) can also be used for data transfer; data rates achievable are 9600-14400 bits/second. As GSM is a

natively digital telephony system, it wouldn't be correct to talk about "GSM modems", as is often done.

However, there is a development towards more advanced protocols such as GPRS (General Packet Radio Services) which allow an always-on digital connection with much higher data rates. This brings us to the following subject: packet switching networks.

Packet forwarding networks

With this we mean 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 functioning of the Internet, IP addresses, and domain name services is explained in many places (e.g., http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/ip.htm) 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" 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 if 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 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 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[1]. It is important to understand that these ports are purely software things; it is the networking software layer in the operating system

[1] There is a list of all services in the file `/etc/services`.

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 one thing that *none* of these Internet protocols is, 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.

10.2 Real time systems

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 and touch screen – of limited size – is indicated. Also ruggedness may be required depending on the navigation environment.

Operating systems

The hardware limitations mentioned obviously also limit what operating system software can be used. Typically found are "embedded" operating systems, like in mobile phones Symbian, in PDAs (Personal Digital Assistants) PalmOS, and more and more Windows CE, e.g., in the iPaq and friends, which however consume significantly more power.

In high-reliability operations, e.g., on spacecraft, also systems like the QNX and Wind River Systems[2] real time embedded operating systems are being used. In "hard" real time applications, the operating system should preferably not crash[3].

Linux/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 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.

Interrupts, masking, latency

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

[2] The Mars rovers Spirit and Opportunity use the Wind River Systems software.

[3] . . . which however the Spirit's system did, due to running out of *file handles.* But it came beautifully back up again.

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, making possible operation by lightly trained personnel. This is however quite demanding in 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., CPU resources are managed through the *scheduler*.

If we look at a single process[4], we can say that the path of execution is *linear*. This means that execution either proceeds to the next statement, or to a statement pointed to by a branching (`if`, `switch`, ...) statement. This makes it easy to keep track of the current *state* of the process: it can only be changed by statements that we have executed.

Looking at a procedure or subroutine 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 Program Counter in the calling procedure – on the *stack*. Next, any locally defined variables are also located on the top of the stack, which thus grows. When the flow of control meets the end of the procedure, first the local variables are deallocated, and then the top of the stack is moved back into the Program Counter of the CPU again, and we have returned to the calling procedure.

Interrupts change this whole picture. Computer hardware provides for a number of different interrupts, and they can happen at any time. When they happen, it is their responsibility not to change anything that could interfere with the processes that are currently executing. Interrupts are used, e.g., to service input/output devices that cannot wait. Every interrupt is connected to an interrupt service routine, which is executed when it is triggered.

Take the clock interrupt routine, for example. It is triggered 50 times a second, and its main function is to increment the software time register kept by the operating system software. But it is typically also responsible for *task 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, that it will need to continue running – is saved, and another process, with 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 *priority*.

Every process – but especially kernel or system level processes – have pieces in their code where 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

[4] ... and ignoring threading!

them, and if we are interrupted, we may 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.

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.

11

Navigation and GIS

Nowadays RTK-navigation is a widely used data collection method for mapping surveying work and GIS. If the accuracy demands are on the level of 1-2 meters, even code based DGPS is suitable, expecially now that the encryption of the GPS system is switched off.

This is how the "navigation solution" can be used in mapping surveying. The accuracy is not at the same level as in static measurement, but often this is completely acceptable. As benefit, there is no post-measurement work (office work), so it's work extensive. The collected data – which can be fairly voluminous, millions of points – goes directly into a GIS after a minimal amount of manual work (type coding for example).

11.1 Geocentric co-ordinate systems

In accurate navigation one has to be a accurate with the co-ordinate system. As itself the GPS gives the coordinates in the WGS84system, which is the system that the GPS system itself uses. More accurate *geocentric* systems, like ITRF-xx and ETRF-xx, are provided by IERS (International Earth Rotation Service) (IRTF = International Terrestrial Reference Frame; ETRF = European Terrestrial Reference Frame). At the accuracy level of about one decimetre, these systems are identical to WGS84.

The ETRF systems have as a useful practical property, that the coordinates of the continental platform of Eurasia, so the system *moves with the platform.* In many countries, as well as in scientific circles, the ETRF-89, also known as EUREF-89is used. Its moment of definition (epoch) is the beginning of 1989.

Geocentric system is a system for which:

- ▷ 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 points either to the vernal equinox point in the sky (astronomical co-ordinate system, inertial) or lies in the plane of the Greenwich meridian (terrestrial co-ordinate system, attached to the solid Earth and "co-rotating")

11.2 Non-geocentric systems

When we want to work in a local or national, non-geocentric system, like *kkj* (Map Grid Co-ordinate System) in Finland, things get a lot more difficult if we want to retain the accuracy obtained from the GPS system. 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 good general solution in local surveying.

11.3 Elevation systems

When GPS – or any other system 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 "the mean sea level", more precisely, the *geoid*. Cf. figure 11.1.

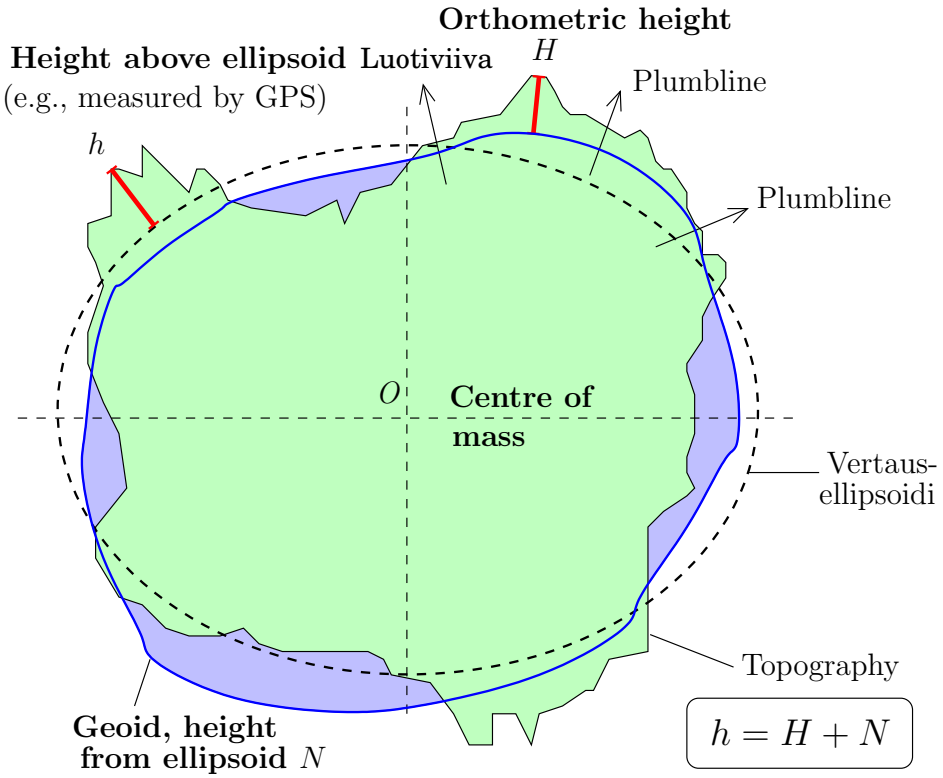


Figure 11.1: Height systems

Bibliography

Tullu Besha Bedada. *Absolute geopotential height system for Ethiopia*. PhD thesis, University of Edinburgh, 2010. URL: http://www.era.lib.ed.ac.uk/handle/1842/4726.

J.M. Brozena. The Greenland Aerogeophysics Project: Airborne gravity, topographic and magnetic mapping of an entire continent. In *From Mars to Greenland: Charting Gravity With Space and Airborne Instruments*, volume 110 of *International Association of Geodesy Symposia*, pages 203–214. IAG, Springer Verlag, 1991.

M. A. R. Cooper. *Control surveys in civil engineering*. Collins, Department of Civil Engineering, The City University, London, 1987.

Jared Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. Norton, 1999.

Carsten Egevang, Iain J. Stenhouse, Richard A. Phillips, Aevar Petersen, James W. Fox, and Janet R. D. Silk. Tracking of arctic terns *Sterna paradisaea* reveals longest animal migration. *Proc. Nat. Acad. of Sci.*, 2010. DOI: 10.1073/pnas.0909493107.

R. Forsberg, K. Hehl, U. Meyer, A. Gidskehaug, and L. Bastos. Development of an airborne geoid mapping system for coastal oceanography (AGMASCO). In *Proceedings International Symposium on Gravity, Geoid and Marine Geodesy (GraGeoMar96)*, volume 117 of *International Association of Geodesy Symposia*, pages 163–170, Tokyo, 1996. IAG, Springer Verlag.

Rene Forsberg, Arne V Olesen, Hasan Yildiz, and CC Tscherning. Polar gravity fields from goce and airborne gravity. In *4th International GOCE User Workshop*, TU Munich, Germany, 2011. URL: https://earth.esa.int/download/goce/4th_Int_GOCE_User_Wkshp_2011/Polar_Gravity_Fields_GOCE_Airborne%20Gravity_R.Forsberg.pdf.

B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *GPS Theory and Practice*. Springer-Verlag, fourth, revised edition, 1997. ISBN 3-211-82839-7.

Christopher Jekeli. *Inertial Navigation Systems with Geodetic Applications*. Walter de Gruyter, Berlin – New York, 2001.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng., Series 82D*, pages 35–45, 1960.

R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME, J. Basic Eng., Series 83D*, pages 95–108, 1961.

D. Munkhtsetseg. Geodetic network and geoid model of mongolia. In *Proceedings, GSEM (Geospatial Solutions for Emergency Management) 2009*, Beijing, China, 2009. ISPRM. URL: http://www.isprs.org/proceedings/XXXVIII/7-C4/121_GSEM2009.pdf.

Markku Poutanen. *GPS-paikanmääritys*. Ursan julkaisuja 64. Tähtitieteellinen yhdistys Ursa, 1998. ISBN 951-9269-89-4.

Dava Sobel. *Longitude. The true story of a lone genius who solved the greatest scientific problem of his time*. Penguin Books, New York, 1995.

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

B. D. Tapley and B. E. Schutz. Estimation of unmodeled forces on a lunar satellite. *Celestial Mechanics*, 12:409–424, December 1975.

P. J. G. Teunissen, P. J. de Jonge, and C. C. J. M. Tiberius. Performance of the LAMBDA method for Fast GPS Ambiguity Resolution. *Navigation*, 44(3):373–383, 1997.

M. Wei and K.P. Schwarz. Comparison of different approaches to airborne gravimetry by strapdown ins/gps. In J. Segawa, H. Fujimoto, and S. Okubo, editors, *Gravity, Geoid and Marine Geodesy*, volume 117 of *International Association of Geodesy Symposia*, pages 155–162, Tokyo, Japan, 1997. International Association of Geodesy, Springer.

Index