Summary of
TTT4150 Navigation Systems

Morten Fyhn Amundsen

May 26, 2016

Contents

1 Introduction

1.1 A brief history of navigation

You want to know where you are. Latitude is easy to determine with a sextant, with e.g. the North star or the sun at its highest point. By keeping the home port time with a clock onboard, and measuring time where you are, you can estimate longitude. But clocks (or "marine chronometers" if you're a huge nerd) were too inaccurate until John Harrison made a really nice one in the 18th century. Some guys also computed time by observing the moons of Saturn.

1.2 Radionavigation

Ground waves are below HF and hug the terrain. Sky waves are in the same frequency range and reflect back down from the ionosphere. Can communicate past the horizon, but hard to determine propagation time, which is necessary for navigation. Space waves are VHF and up, and go straight.

1.2.1 Trilateration

Use propagation time to estimate distance from source. Compute position from several distances. You need three sources for 2D positioning (or two sometimes). For 3D positioning, you need at least one beacon with a large Δh, which you have with satellites but not ground beacons.

1.2.2 Hyperbolic positioning

You get signals from two transmitters, and measure the difference in propagation time. If the difference is zero, your position is equally far from both, and therefore somewhere on a straight line. If the difference is nonzero, your position is a bit closer to one transmitter, and will be along a hyperbolic line. If you have another transmitter pair, each pair gives a line of possible positions, and your location should be at their intersection. Figure 1 illustrates this, with time differences of 0.3 units between antenna 1 and 2, and -0.1 units between 1 and 3. Since you only worry about differences, your clock offset cancels, assuming all transmitter clocks are synced.

1.2.3 Doppler positioning

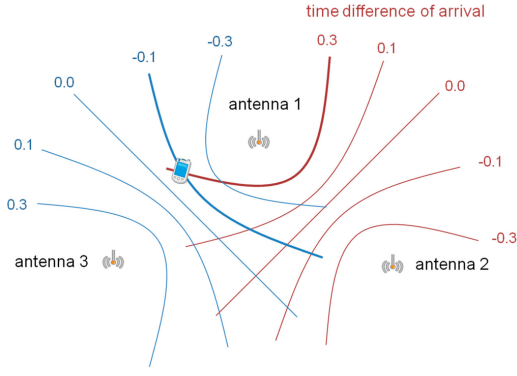Observe the doppler shift of a satellite. If you know the orbit, you can determine your position.

Figure 1: The principle of hyperbolic navigation

2 Coordinate frames

2.1 Reference frames

2.1.1 Notation

- $\boldsymbol{v}^n$—velocity given in frame n
- R_p^n—rotation from frame p to n
- $\boldsymbol{\omega}_{ip}^p$—angular rate of frame p w.r.t. i, represented in p.

2.1.2 Inertial frames and the ECI frame $\{i\}$

A frame in which Newton's laws of motion apply, i.e. it does not accelerate. All inertial sensors measure relative to an inertial frame. *Earth-centered intertial* (ECI) is an example of this, with origin at the Earth center, not rotating with the earth. (But as the Earth rotates around the sun, ECI is not exactly inertial.)

2.1.3 ECEF frame $\{e\}$

Earth-centered, earth-fixed. Rotates relative to ECI with $\omega_{ie} \approx 7.292 \times 10^{-5}$ rad s^{-1}.

2.1.4 Geographic frame $\{g\}$

Origin equal to platform origin projected onto the reference ellipsoid. z-axis down, normal to ellipsoid surface. x-axis north, y-axis east. Not inertial.

2.1.5 Geocentric frame

Similar to the geographic frame, but z-axis down toward Earth center. Also not inertial.

2.1.6 Tangent plane

North, east, down as used in daily life. Origin at a fixed point on Earth. Coincides with the geographic frame when the system origin coincides with the tangent frame origin.

2.1.7 Body frame $\{b\}$

Attached to the vehicle of interest, often origin at center of gravity. x-axis forward, z-axis down, and y-axis completes the right-handed orthogonal coordinate system (i.e. to the right).

2.1.8 Platform frame

Frame centered somewhere on the sensor platform. Usually non-moving wrt. the body frame, so R_p^b is constant.

2.1.9 Instrument frame

The frame along which an instrument measures. (An IMU measures acceleration along the instrument axes, and rotation rate around the instrument frame axes.)

2.2 Ellipsoids
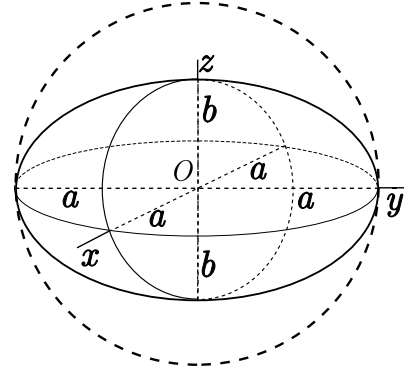
Figure 2: An oblate ellipsoid

An ellipsoid is defined by its semi-major axis a and semi-minor axis b (see Figure 2). Its *eccentricity* e is defined by

$$e^2 = \frac{a^2 - b^2}{a^2}. \tag{1}$$

and its flattening f is

$$f = \frac{a - b}{a}. \tag{2}$$

2.2.1 Radii of curvature

M is the radius of curvature of a meridian at a given geodetic latitude ϕ. N is the radius of curvature normal to the meridian at a given ϕ. Because the Earth is oblate, M is largest at the poles, and smallest at the equator. At the poles, $N = M$, because the radius is the same in all directions.

2.3 Geodetic coordinates

Defined by

- latitide ϕ,
- longitude λ, and
- height h above the ellipsoid.

2.3.1 Definitions of height

Three surface models are used: The ellipsoid surface, the geoid surface, and the true surface of the earth. The *geoid* is based on what the mean sea level would be based on local gravitation and the earth rotation.

- Orthometric height: Height above the geoid.

- Geoid height: Height of the geoid above/below the ellipsoid.
- Ellipsoidal height: Height above the ellipsoid, measured normal to the surface. This is what GPS uses.

2.4 Earth-Centered, Earth-Fixed (ECEF) coordinates

Defined by Cartesian coordinates X, Y, and Z, along the axes in Figure 2. The positive x-axis intersects the surface at $\phi = \lambda = 0°$. (The frame is fixed to Earth and rotates with it.)

2.5 Conversion from geodetic to ECEF

For a sphere of radius R we have

$$
\begin{aligned}
X &= (R + h) \cos \phi \cos \lambda \\
Y &= (R + h) \cos \phi \sin \lambda \\
Z &= (R + h) \sin \phi.
\end{aligned}
\tag{3}
$$

For an ellipsoid with semi-minor axis a and semi-major axis b, we have

$$
\begin{aligned}
X &= (N + h) \cos \phi \cos \lambda \\
Y &= (N + h) \cos \phi \sin \lambda \\
Z &= \left(N[1 - e^2] + h\right) \sin \phi
\end{aligned}
\tag{4}
$$

where the *normal radius of curvature* is

$$
N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}}.
\tag{5}
$$

2.6 Conversion from ECEF to geodetic

Easy for a sphere:

$$
\begin{aligned}
\phi &= \operatorname{atan} \frac{Z^2}{\sqrt{X^2 + Y^2}} \\
\lambda &= \operatorname{atan} \frac{Y}{X} \\
h &= \sqrt{X^2 + Y^2} \cos \phi + Z \sin \phi - R
\end{aligned}
\tag{6}
$$

For an ellipse, ECEF to geodetic conversion can be done iteratively or by one of several closed-form solutions.

2.6.1 Iterative method

The longitude λ is found as

$$
\lambda = \operatorname{atan} \frac{X}{Y}.
\tag{7}
$$

The latitude ϕ is found by iterating

$$
\phi_{k+1} = \operatorname{atan} \left(\frac{Z}{\rho} + \frac{N_k e^2 \sin \phi_k}{\rho} \right)
\tag{8}
$$

where

$$
\rho = \sqrt{X^2 + Y^2}.
\tag{9}
$$

When ϕ is sufficiently accurate, the height h is found by

$$
h = \rho \cos \phi + Z \sin \phi - N(1 - e^2 \sin^2 \phi)
\tag{10}
$$

2.6.2 Bowring's closed-form method

Only valid near the Earth surface.

$$
\begin{aligned}
\phi &= \operatorname{atan} \frac{Z + e^2 b \sin^3 \mu}{\rho - e^2 a \cos^3 \mu} \\
\mu &= \operatorname{atan} \frac{Za}{\rho b}
\end{aligned}
\tag{11}
$$

2.6.3 Vermeille's closed-form method

Closed-form, valid quite far from Earth (including satellite orbits). Really mathy.

2.7 Direction cosine matrix

You have a vector $\boldsymbol{v}_1$ given in frame ϕ_1. The unit vectors along the axes of ϕ_1 are I_1, J_1, and K_1. Similarly, the unit vectors along the axes of ϕ_2 are I_2, J_2, and K_2.

Then, the direction cosine matrix between the frames is

$$
\boldsymbol{R}_2^1 = \begin{pmatrix} \cos(I_2, I_1) & \cos(I_2, J_1) & \cos(I_2, K_1) \\ \cos(J_2, I_1) & \cos(J_2, J_1) & \cos(J_2, K_1) \\ \cos(K_2, I_1) & \cos(K_2, J_1) & \cos(K_2, K_1) \end{pmatrix}
\tag{12}
$$

3 Spherical geometry
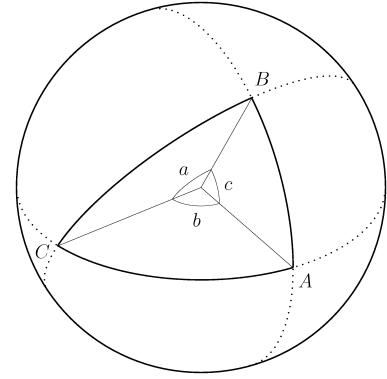
Figure 3: A triangle on a sphere

A triangle on a sphere is defined by three points on the sphere, as in Figure 3. Each corner A, B, and C has an angle, and each line/arc segment has an angle a, b, and c.

Modified cosine rules

$$
\begin{aligned}
\cos a &= \cos b \cdot \cos c + \sin b \cdot \sin c \cdot \cos A \\
\cos b &= \cos c \cdot \cos a + \sin c \cdot \sin a \cdot \cos B \\
\cos c &= \cos a \cdot \cos b + \sin a \cdot \sin b \cdot \cos C
\end{aligned}
\tag{13}
$$

and sine rules

$$
\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}
\tag{14}
$$

apply.

3.1 Intersection between great circles

Each great circle is defined by two points on the surface of a sphere.

1. Convert the points to cartesian coordinates.
2. blabla

4 Signal processing

Figure 4 shows the main components of GNSS receiver signal processing.

Figure 4: Signal processing in a GNSS receiver

4.1 Acquisition
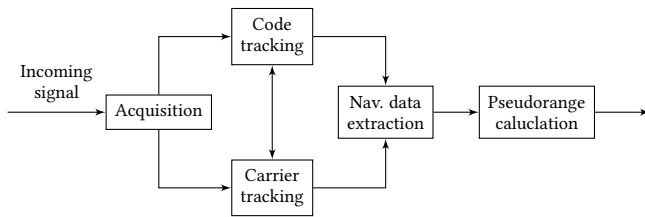
Each satellite transmits a PRN signal independent from transmitted data. The signal uses more bandwith than the navigation data needs. The receiver correlates the incoming signal with a replica of the PRN code.

All GNSS signals share the same medium, and most use CDMA to share it. With CDMA, each satellite has its own PRN code. These have low crosscorrelation with other PRN codes.

4.2 Code and carrier tracking

The code delay and carrier phase is tracked for each satellite in view, in separate "channels" in the software. The code delay is the misalignment between the incoming signal code and the carrier-generated replica. The carrier phase reflects the relative motion between the satellite and the user. Both are tracked by *tracking loops* (phase-lock and delay-lock loops).

4.3 Navigational data extraction

4.4 Pseudorange extraction

5 GPS coordinate frames, time reference, and orbits

5.1 Global coordinate systems

5.1.1 Conventional terrestrial reference system (CTRS)

The *pole of rotation* moves relative to the true pole. The *Conventional terrestrial pole* (CTP) is a defined position corresponding to the average position of the pole of rotation between 1900 and 1905. With this, CTRS is defined with

- origin at Earth mass center,
- z-axis through the CTP,
- x-axis through the intersection of CTP's equatorial plane and a reference meridian, and
- y-axis to complete a right-handed system.

In practice, this is implemented as the coordinate system that best fits a number of reference points with assigned coordinates.

5.1.2 Conventional inertial reference system (CIRS)

An inertial frame is nonaccelerating. CTRS spins with the Earth, and is therefore not inertial. Defined with

- origin at Earth mass center,
- z-axis along the axis of rotation,
- x-axis in the equatorial plane, toward the vernal equinox,
- y-axis to complete a right-handed system.

5.1.3 International terrestrial reference frame (ITRF)

More accurate realization of CTRS than WGS 84 ECEF, used for research. Defined from a large number of reference stations around the world.

5.1.4 WGS 84

A realization of CTRS. It defines

- an ECEF Cartesian coordinate frame,
- an ellipsoid model of the Earth,
- a characterization of Earth's gravity field and geoid, and
- some constants.

and its parameters are:

- Semi-major axis a
- Reciprocal flattening $\frac{1}{f}$
- Earth's angular velocity ω_E
- Earth's gravitational constant GM
- Speed of light in vacuum c

5.2 GPS orbits
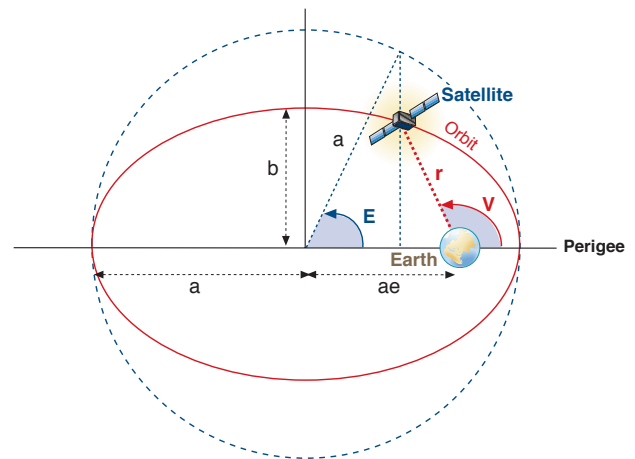
Figure 5: True anomaly V and eccentric anomaly E

5.2.1 Kepler's laws

The laws are modified for satellites orbiting Earth.

1. The satellite orbit is an ellipse with Earth at one focus.
2. The line between the satellite and Earth sweeps over equal areas during equal intervals of time. (I.e. faster close to Earth.)
3. The square of the orbital period is proportional to the cube of the semi-major axis of the orbit: $T^2 \propto a^3$
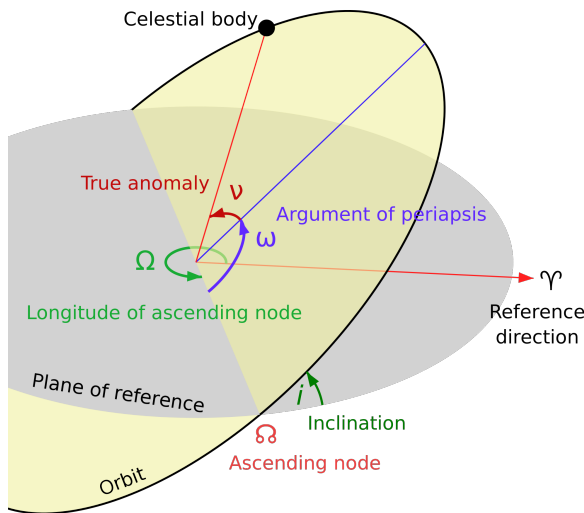
5.2.2 Keplerian elements

Figure 6: Some Kepler elements, expressed generally, not terrestrially

The Keplerian elements are given at a specific epoch, and are as follows:

a **Semi-major axis** *a*: Length of the semi-major axis of the elliptical orbit.

e **Eccentricity** *e*: The eccentricity of the elliptical orbit.

i **Orbital inclination** *i*: The orbital plane intersects the Earth center, but usually at an angle relative to the equator. This angle is the inclinaton.

Ω **Right ascension of the ascending node** Ω: First: The intersection between the equator and the orbital plane is called the line of nodes. The nodes are the two points where the line intersects the surface. The ascending node is the node above which the satellite ascends (passes in a south to north direction).
The RAAN is then the angle from the direction of the vernal equinox to the ascending node. The vernal equinox is a fixed point in the sky above equator, not rotating with Earth. (The vernal equinox is actually the ascending node of the Sun, meaning the sun has Ω = 0.)
In Figure 6, it is called "longitude of ascending node".

1. **Argument of perigee** *ω*: The perigee is the point of the orbit closest to Earth. The argument of perigee is the angle between the ascending node and the perigee, as seen from Earth.
In Figure 6, this is called "argument of periapsis".

2. **True anomaly** *ν*: The angle between the perigee and the satellite, as seen from Earth.

5.2.3 Satellite position and velocity

The orbital plane is specified fully by *a*, *e*, *i*, Ω, and *ω* defined above. If we know *ν* at an epoch, we can find the position at any given time. In addition to *true* anomaly *ν*, we have *eccentric* anomaly and *mean* anomaly are defined:

- **Eccentric anomaly** *E*: The angle between the perigee and the projection of the satellite onto a circle of radius *a*, as seen from the center of the orbit. See angle "E" in Figure 5.

- **Mean anomaly** *M*: The angle between the perigee and the place the satellite would have been had it been on a circular orbit with the same focus and period. See angle "M" in Figure 5.

Kepler's equation *E* and *M* are related by *Kepler's equation*:

$$M = E - e \sin E \tag{15}$$

5.2.4 Perturbations

Precession (period 25 800 years) and *nutation* (18.6 years) affect satellite orbits.

6 GPS measurements and error sources

Two measurements: Code tracking and carrier phase.

6.1 Measurement models

6.1.1 Code phase measurements

The satellite sends a code with a time stamp. The receiver gets the code, and calculates the propagation time by comparing the time stamp to the receiver time when the signal arrives. This propagation time is off by δt, because the receiver clock is less accurate and out of sync with the satellite clocks. It still calculates a receiver-satellite range with the imprecise propagation time. This range is also imprecise, and is called a *pseudorange*. It consists of the correct range plus a term $\delta t \cdot c$. We must estimate this δt to find the correct distance between user and satellite.

The measurement of propagation time is done like this: The satellite sends a 'pseudo-random' code, and the receiver generates an identical code. Receiver shifts its code until it aligns with the satellite code. Shift is equal to signal travel time. However, the cycle width is almost a microsecond, and *perfect* alignment is impossible. Alignment within 1–2 percent can be done, but gives meter-range errors.

6.1.2 Carrier phase measurements

The carrier signal frequency for the code is over a GHz (1.57 GHz?). The wavelength is rougly 20 cm, and if we align with 1 % accuracy, we're at millimeter level. As opposed to the code, carrier cycles are similar to each other. After narrowing down to meter level with code measurements, we must find out which carrier cycle we're on. We can measure the phase difference Δ_0, but we must determine the integer cycle ambiguity N.

6.2 Signal propagation modeling errors

GPS signals mostly travel at the speed of light, but at roughly 1000 km they are refracted by hitting the atmosphere. The direction change doesn't matter much, but the speed change is bad. For L-band radio signals, the ionosphere is dispersive (refractive index depends on signal frequency) and the troposphere is not.

6.2.1 Ionospheric delay

The ionosphere is made of ionized gases and its properties (electron density) change from day to day and year to year and so on. Ionized gas is refractive and dispersive to radio waves. The phase advance and group delay are proportional to electron count along the path, and the path depends on the satellite elevation.

If you have a dual-frequency GPS receiver, you can estimate the errors and remove them.

6.2.2 Tropospheric delay

The troposphere consists of dry gases (mostly N_2 and O_2) and water vapor. Not dispersive like the ionosphere. Propagation speed is lower than in free space, so apparent range is around 2.5 m to 25 m too long.

6.3 Measurement errors

First: High precision means low variability. High accuracy means average is near true value.

6.3.1 Receiver noise

All non-signal RF radiation sensed by the antenna, in the same band.

6.3.2 Multipath

Reflections of the signal so that the same signal (from the satellite) reaches the antenna more than once.

6.4 Differential GPS (DGPS)

Many GPS error sources are almost the same for all receivers in an area. With a receiver of known position, you can estimate the errors and broadcast them to nearby users, who then can use them to remove the errors.

7 Position, velocity, and time (PVT) estimation

7.1 Position estimation with pseudoranges

The pseudorange measurement from satellite k after correction is

$$\rho_c^{(k)} = r^{(k)} + c \cdot \delta t_u + \tilde{\epsilon}_\rho^{(k)} \tag{16}$$

where δt_u is the user clock bias, and $\tilde{\epsilon}_\rho^{(k)}$ is the combined effect of residual errors. The position of the user at the time of measurement is

$$\boldsymbol{x} = (x, y, z) \tag{17}$$

and the position of satellite k at signal transmission time is

$$\boldsymbol{x}^{(k)} = \left(x^{(k)}, y^{(k)}, z^{(k)}\right). \tag{18}$$

The user-to-satellite range is

$$r^{(k)} = \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\right\| \tag{19}$$

and (16) can be written as

$$\rho_c^{(k)} = r^{(k)} + b + \tilde{\epsilon}_\rho^{(k)}. \tag{20}$$

The *geometry matrix* is

$$G = \begin{bmatrix} (-\boldsymbol{1}^{(1)})^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ (-\boldsymbol{1}^{(K)})^{\mathrm{T}} & 1 \end{bmatrix} \tag{21}$$

where K is the number of satellites in view, and $\boldsymbol{1}^{(k)}$ is the line of sight unit vector from the initial user position estimate to satellite k.

7.2 Dilution of precision (DOP)

DOP quantifies the effect of satellite geometry on the accuracy of positioning. A good geometry gives a low DOP value, and high accuracy. There are several variants, using $\boldsymbol{H} = (\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}$:

- Position dilution of precision

$$\text{PDOP} = \sqrt{H_{11} + H_{22} + H_{33}} \tag{22}$$

- Time dilution of precision

$$\text{TDOP} = \sqrt{H_{44}} \tag{23}$$

- Geometric dilution of precision

$$\text{GDOP} = \sqrt{H_{11} + H_{22} + H_{33} + H_{44}} \tag{24}$$

- Horizontal dilution of precision

$$\text{HDOP} = \sqrt{H_{11} + H_{22}} \tag{25}$$

- Vertical dilution of precision

$$\text{VDOP} = \sqrt{H_{33}} \tag{26}$$

8 Precise positioning with carrier phase

Code measurement accuracy tops out at the meter level. Precise positioning with carrier phase measurements, can achieve centimeter-level accuracy.

8.1 Carrier phase and integer ambiguity resolution: A simple model

Two antennas receive the same carrier signal. It's off by N full cycles, and a partial cycle. The partial cycle (phase shift) can be measured, but we don't know N. This gives integer ambiguity.

If we wait for the satellite to move a bit and measure again, we have two equations in two unknowns, and can solve for N. Satellite geometry change is essential for the quality of the estimates.

8.2 Precise point positioning (PPP)

PPP is centimeter-level positioning without a reference station, based on precise estimation of all significant error sources. The necessary orbit data is available in various latencies and accuracies. Near real-time is possible, but you get better accuracy if you wait a while.

The errors we must estimate are:

- Satellite ephemeris and clock error: You can get these from IGS.
- Ionospheric delay: Eliminate with dual-frequency measurements.
- Tropospheric delay: Must estimate.
- Multipath and receiver noise: At least no reference station multipath error now.
- Phase wind-up correction: Antenna rotation (rover or satellite) will change the carrier phase.
- Satellite antenna offsets: Orbit models give position of satellite mass center, but pseudorange measurements refer to the antenna phase center.

8.3 Carrier phase integer ambiguity

As mentioned, carrier phase has an integer ambiguity. This integer can be fixed.

8.3.1 Using code measurements to estimate integers

We can use the code measurement to estimate the carrier integer. But it's hard to use the fairly coarse code pseudorange to estimate this, and you need to average codes over a long time horizon to get a good enough estimate. Therefore, it doesn't really work.

9 Loxodromes

A *loxodrome* curve is an Earth course of constant bearing (also called a rhumb). Generally not a great circle, and therefore not the shortest route, but easy to use, as great circles require continuous change of bearing.

9.1 Conformal maps

A *conformal map* distorts east-west and north-south by the same amount, which means that it preserves angles. Distortion depends only on position (actually only on latitude). The *Mercator map* is a special conformal map, on which a rhumb/loxodrome is a straight line.
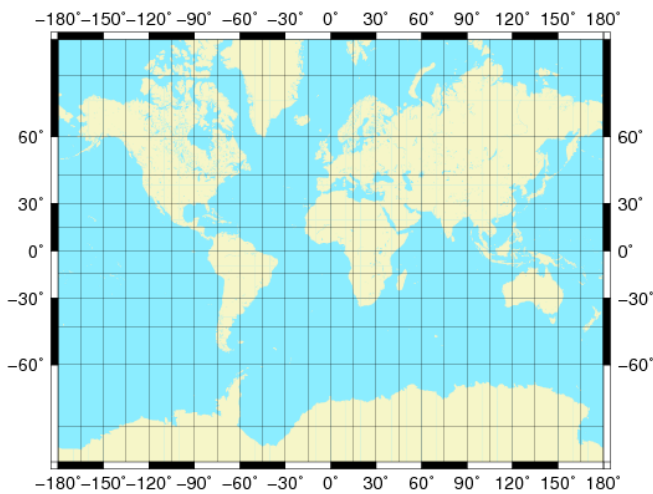

Figure 7: A Mercator map

10 Kalman filtering

A *Kalman filter* recursively estimates the state of a process, such that it minimizes the mean squared error. Or, maximize separation of signal and noise.

10.1 Process model

The *discretized* process is

$$x_k = \Phi x_{x-1} + w_{k-1}$$
$$z_k = H x_k + v_k \tag{27}$$

where

- Φ is the (state) *transition matrix*,
- H is the *design matrix*,
- w_k is random white process noise $p(w) \sim N(0, Q)$, and
- v_k is random white measurement noise $p(v) \sim N(0, R)$.

The transition matrix is calculated from the continuous-time representation

$$\dot{x} = Fx + Gu \tag{28}$$

as

$$\Phi = e^{F\Delta t} = I + F\Delta t + \frac{(F\Delta t)^2}{2!} + \dots \tag{29}$$

or, for time-variant F

$$\Phi_k = \mathcal{L}\left\{(sI - F)^{-1}\right\}_{t=\Delta t}. \tag{30}$$

10.2 Method

The goal is to compute an *a posteriori* estimate[1] $\hat{x}_k$ based on an *a priori* estimate[2] $\hat{x}_k^-$, a measurement z_k, and a measurement prediction $H\hat{x}_k^-$. This is how we do it:

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \tag{31}$$

The gain matrix K is chosen to minimize the a posteriori error covariance $P_k = E\left[e_k e_k^\mathsf{T}\right]$, where $e_k \equiv x_k - \hat{x}_k$. Some math leads to

$$K_k = P_k^- H^\mathsf{T} \left(HP_k^- H^\mathsf{T} + R\right)^{-1}. \tag{32}$$

We have $\lim_{R\to 0} K_k = H^{-1}$, which gives $\hat{x}_k = H^{-1}z_k$. Makes sense, because zero measurement error covariance R should let us trust the measurement fully. We also have $\lim_{P_k^-\to 0} K_k = 0$, giving $\hat{x}_k = \hat{x}_k^-$. Also makes sense, because zero a priori estimate error covariance P_k^- should make us not trust the measurement at all.

These equations are split into the time update equations

$$\hat{x}_k^- = \Phi_{k-1}\hat{x}_{k-1}$$
$$P_k^- = \Phi_{k-1}\hat{P}_{k-1}\Phi_{k-1}^\mathsf{T} + Q \tag{33}$$

and the measurement update equations

$$K_k = P_k^- H^\mathsf{T}\left(HP_k^- H^\mathsf{T} + R\right)^{-1}$$
$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$$
$$\hat{P}_k = (I - K_k H)P_k^-. \tag{34}$$

---

[1] *After* measuring.
[2] *Before* measuring.

Often, Q and R are simply considered tuning parameters. Large R means we expect a large measurement error variance, and the filter converges slowly because it cannot trust the measurements. A small R does the opposite: Converges fast, but possibly so fast that the estimate follows the measurement noise.

10.3 IMU + GNSS integration

GNSS has good long time accuracy, and gives measurements without drift. INS has good short time accuracy, but drifts a lot. Combining both in a *complementary filter* gives "the best of both worlds".