Chapter 1

Early times

Today's global satellite navigation systems have predecessors[1]. The concepts evolved steadily from the initial tracking of Sputnik signals. This triggered the development of the first satellite navigation system TRANSIT, which played a crucial role in the development of satellite technologies in general and satellite navigation in particular. The initial systems used the time evolution of the Doppler shift for estimating the user's position. With the development of stable atomic clocks, range measurements became another option. Ranges were initially measured using side tones.

1.1 Sputnik - How it all started

The geophysical year 1957 had seen a number of discussions about artificial satellites. The US had announced the imminent launch of their first space vehicle - Vanguard, which was to weight 1.8 kg. On the Soviet side, Korolyov worked on the first Intercontinental Ballistic Missile (ICBM). Soviet missiles needed to carry a larger weight (5.7 tons) for a longer distance (7000 km) than their US counter parts. The US could potentially launch from Europe and had more compact nuclear war heads. Korolyov correspondingly developed the most powerful launcher of his time - the famous R7 rocket. The construction included a total of 20 rocket engines. The majority of these engines were arranged on four external attachments forming the first stage. They were discarded after burnout. The remaining inner rocket would then form the second stage. This arrangement allowed to fire all engines at once, which was important in the early days.

Korolyov negotiated the permission to launch a satellite into an orbit around the earth with his superiors. The agreement was that the launch would take place after two successful take-offs of the R7. The second take-off succeeded in August 1957. During the following weeks, he constructed a simple satellite - Sputnik 1. The launch was on October 4th, 1957 and became a tremendous success for the Soviet Union. The glory of Soviet space conquest was further supported by the launch of a dog in a 500 kg satellite a month later, by hitting the moon with Luna 2 on September 13th, 1959, and by bringing Yuri Gagarin into orbit on April 12th, 1961. On the US side, the first two Vanguard satellites never reached orbit. Wernher von Braun and his team would later restore US self-confidence with a series of successes starting with Explorer 1 on January 31st, 1958 and ending with Amstrong and Aldrin walking on the moon on July 20th, 1969. Explorer 1 led to the discovery of the radiation belt surrounding earth by van Allen.

The R7 rocket turned out to be rather unsuitable as an ICBM. Loading the fuel was taking 8-12 hours. However, the R7 became the most successful launcher ever built. It developed into the workhorse of Soviet and Russian science missions until today. It was named Sputnik, Vostok,

[1]**The present notes are part of a book that is under preparation. They are copyrighted and only provided for your exclusive personal use. Any reproduction or redistribution in any form is prohibited.**

Luna, Molniya, Voskhod, Poljod, Soyuz, and Soyuz-Fregat (see Figure 1.1). It also brought Western European scientific satellites into orbit, e.g. Venus and Mars Express, and is extensively used in the launch of Galileo satellites: GIOVE A and GIOVE B, the four in orbit validation satellites, as well as the 14 next satellites of the Galileo constellation.

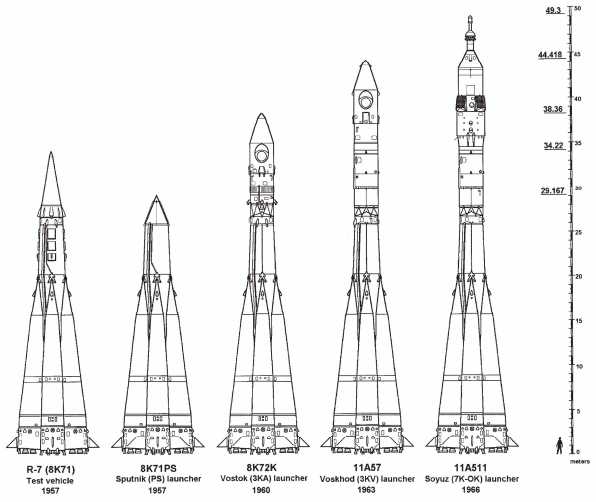

Figure 1.1: Sergey Pavlovich Korolyov (1907-1966) pioneer of the Soviet space program. Different variants of Korolyov's R7 rocket, which launched Sputnik 1 in 1957 and the first experimental Galileo satellite - GIOVE A - in 2005. [Courtesy: NASA]

The satellite Sputnik was a gas filled sphere with a diameter of 58.5 cm and a weight of 83.6 kg. It was around 50 times heavier than the first US satellite Vanguard. Sputnik carried a thermometer, a pressure gauge, and a battery powered radio frequency transmitter. It had 4 antennas with a length between 2.4 and 2.9 m pointing on one side of the satellite (see Figure 1.2). The pressure gauge was used to detect leakages in the 2mm thick aluminium-alloy hull that could have been caused by micro-meteorites. The sensor reading was encoded into the length of the beeps transmitted. The carrier frequencies of the transmission were 20.005 MHz and 40.002 MHz and the power was 1 Watt. The satellite's orbit had a perigee at 228 km, an apogee at 947 km, and thus an orbital period 96.2 minutes.

Sputnik's signals were observed around the globe, and it was soon discovered that the pattern of the Doppler shift could be used to determine the distance from the observer, as well as the time of nearest passage (see Figure 1.3). Three groups came to this conclusion: Brown, Green, Jr., Howland, Lerner, Manasse, and Pettengill at MIT [1]; Peterson at the Stanford Research Institute [2]; and Guier and Weiffenbach at John Hopkins University [3]. The latter team inspired McClure to consider the inverse problem, i.e., to determine the observers position based on the known orbit of the satellite. This idea was of strategic importance since it meant that Polaris nuclear submarines could potentially navigate with a compact receive antenna and thus stay undiscovered. Together McClure and his friend Kershner conceived the first satellite navigation system Transit[2] during one week-end in March 1958. Their concept included satellites on a polar orbit with two ultra-stable transmitters, whose carriers were modulated with information about the orbit parameters, as well as a tracking system which would use the same two signals for determining the orbits of the satellites. Two frequencies were used to compensate for the ionospheric dispersion (see [5] for more details). The Soviet Union designed a very similar system called Tsikada.

These origins justify considering Sputnik as well as the observations made on October 5th, 1957

[2]Transit is often capitalized into TRANSIT. Since the Kershner uses Transit we will adhere to this writing.

and the following days as the first steps towards satellite navigation. Definitively, March 1958 is the latest date to consider for the "invention" of satellite navigation. The next section introduces the Doppler effect in order to make the conceptions of the early times understandable.

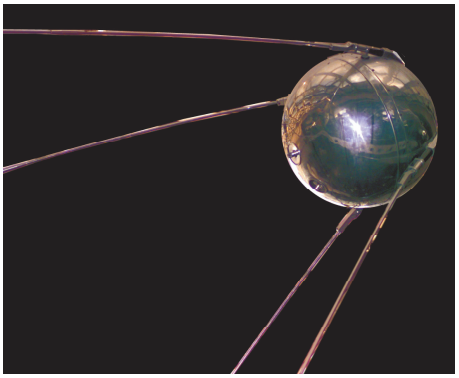

Figure 1.2: Sputnik - the first manmade satellite - was a gas filled sphere, with a transmitter and 4 antennas. [Courtesy: NASA]

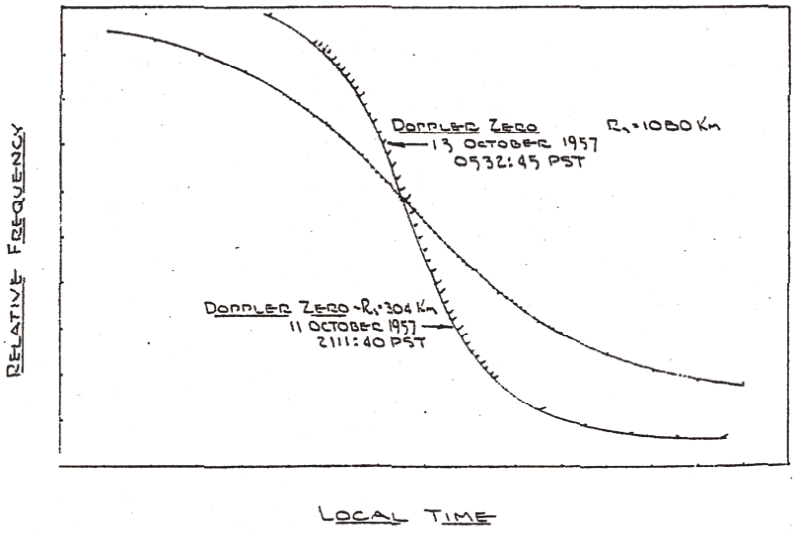

Figure 1.3: Doppler traces for two passages at different distances from the observer. Historical record made by Peterson. [Reproduced from [2]]

1.2 Doppler Positioning

1.2.1 Doppler Shift

The Doppler frequency shift was suggested by Christian Doppler in 1848 in order to explain the color of the stars. The difference in color was later found to be primarily due to the differing surface

temperature of the stars. The frequency of the wave measured by an observer moving relatively to the source was correctly predicted by Doppler and found a wide use in physics and engineering until today.

The Doppler shift is best explained by considering the lines of nodes of the wave shown in Figure 1.4. In a homogenous and isotropic medium, a wave propagates in a spherically symmetric manner, i.e., the maxima, minima, and nodes are circles, which expand with the phase velocity c. In the case of a stationary source all circles are centered around the source. However, if the source moves, the center of the circles are shifted according to that movement. This leads to the pattern shown on the right of Figure 1.4. The wave appears effectively compressed in the direction of the movement. The frequency depends both on the relative velocity and on the direction of observation. The latter is parameterized by the angle θ between the velocity vector and the direction of observation. The size of the Doppler shift is easily computed by considering Figure 1.5. At time t, the node

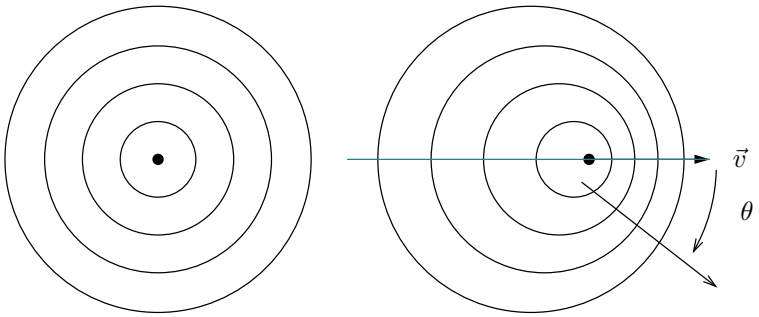

Figure 1.4: The nodes of the wave on the left are associated with a stationary source. They describe circles with a radius that increases with time. The nodes on the right are associated with a moving source. The circles are now centered around the position of the source at the time of transmission.

generated at time 0 will have a radius ct, with c being the velocity of light[3]. This is the large circle in the figure. Similarly, the node generated at time $t + \Delta t'$ will have a radius $c(t - \Delta t')$. This is the smaller circle centered around[4] $\vec{r} = (v\Delta t', 0, 0)^T$. The displacement is due to the movement of the source during the time interval $\Delta t'$. The distance between the two circles in the direction of θ is determined by simple geometry:

$$c\Delta t(\theta) = ct - \sqrt{(c(t - \Delta t')\sin\theta)^2 + (c(t - \Delta t')\cos\theta + v\Delta t')^2} \;\simeq\; c\Delta t' - v\Delta t'\cos\theta. \tag{1.1}$$

The last equation was obtained from the Taylor expansion $\sqrt{1 + x} \simeq 1 + x/2$, which holds for $v \ll c$, i.e. for non-relativistic velocities. The Doppler effect for relativistic velocities will be derived in Chapter 8. If the circles are associated with two consecutive nodes, $\Delta t'$ becomes the period T' of the wave, which is the inverse of its frequency f'. The primes indicate that the period and the frequency are measured in a system in which the source is at rest. The frequency f of the wave emitted by a moving source, observed in a "stationary system" under the angle θ is:

$$f = \frac{1}{\Delta t} \simeq (1 + \frac{v}{c}\cos\theta)\, f'. \tag{1.2}$$

The approximation $1/(1 - x) = 1 + x$ was used to obtain this equation. The latter approximation could have been combined with the previous expansion of the square root as well.

[3]Physical constants and other parameters are summarized in Chapter 17.

[4]Throughout this book vectors are written in column form. The row form, which is used in the running text, is obtained by transposition.

With $\vec{n}$ denoting the direction of the observer, the Doppler shift f_D can also be written in vector notations[5]:

$$f_D = f - f' = \frac{\vec{v} \cdot \vec{n}}{c} f'. \tag{1.3}$$

In this equation we have opted to write an equality, although this does not fully reflect the physical reality.

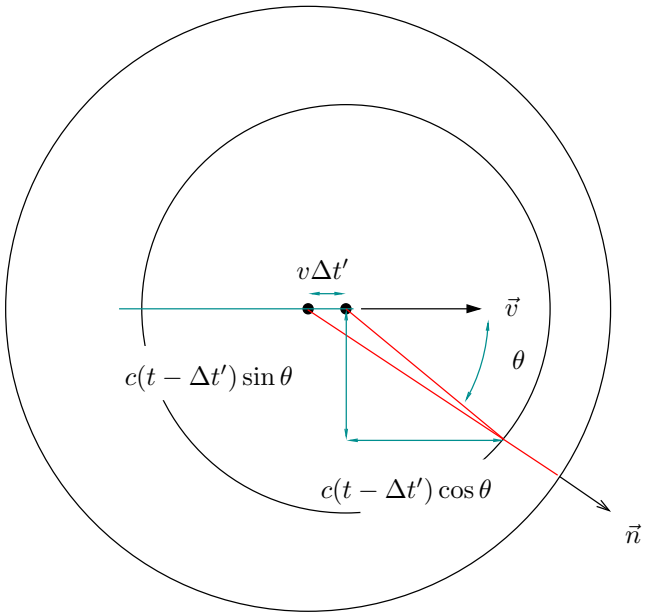

Figure 1.5: The distance between two lines of nodes is computed from geometry. The outer circle represents the line of nodes generated at time 0, while the inner circle represents the line of nodes generated at time $\Delta t'$.

1.2.2 Determining Trajectories from the Doppler Shift

The distance h of the nearest point between a satellite transmitting a carrier and an observer can be determined from the pattern of the Doppler shift. In order to ease the computation, the satellite shall be assumed to be on a linear trajectory tangential to its circular orbit at the nearest point. In this case, the angle θ can easily be determined from the geometry shown in Figure 1.6. Let $\vec{V} = (V, 0)^T$ be the velocity of the satellite and $\vec{R} = \vec{V}t + (0, h)^T$ be its orbit in a coordinate system in which the observer is at rest, then the cosine can be computed as a function of t (see Figure 1.6):

$$\cos\theta = -\frac{\vec{R}}{\|\vec{R}\|} \cdot \frac{\vec{V}}{\|\vec{V}\|} = -\frac{\mathrm{sign}(t)}{\sqrt{1 + \frac{h^2}{V^2 t^2}}}.$$

In this expression, the velocity V is an orbital parameter only, while h depends on the position of the observer as well. The derivative of the Doppler shift at $t = 0$:

$$\frac{d}{dt} f_D \bigg|_{t=0} = -\frac{V^2}{hc} f',$$

[5]The dot between the vectors $\vec{x}$ and $\vec{y}$ describes the scalar product $\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y}$. This definition deviates from the usual one, which would require writing $\vec{x}^T \cdot \vec{y}^T$.

is a simple function of h. The corresponding expression for a circular orbit was determined by Giger [6]:

$$\frac{d}{dt} f_D \Big|_{t=0} = -\frac{R_e}{R_e + h} \frac{V^2}{hc} f',$$

with R_e being the radius of the earth. The distance h can be easily determined from the measurements and was used by Brown, Green, Howland, Lerner, Manasse, and Pettengill [1], as well as by Peterson [2] to determine the distance between an observer and Sputnik in 1957. Figure 1.7 shows the ground track of the orbit determined by Brown et al. on the days immediately after the launch. A ground track is the succession of locations on earth, where the satellite is in the zenith. Alternatively, it can be described as the sequence of pierce points of the vector joining the center of the earth and the satellite's location through the surface of the earth. The computation of ground tracks is described in the Appendix B to Chapter 4. The detailed determination of orbits will not

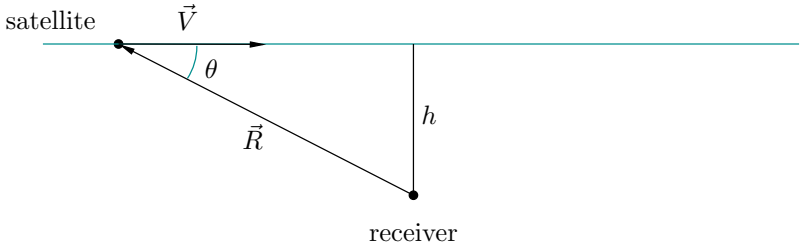

Figure 1.6: Simplified model of a satellite passing the observer at distance h on a linear orbit.

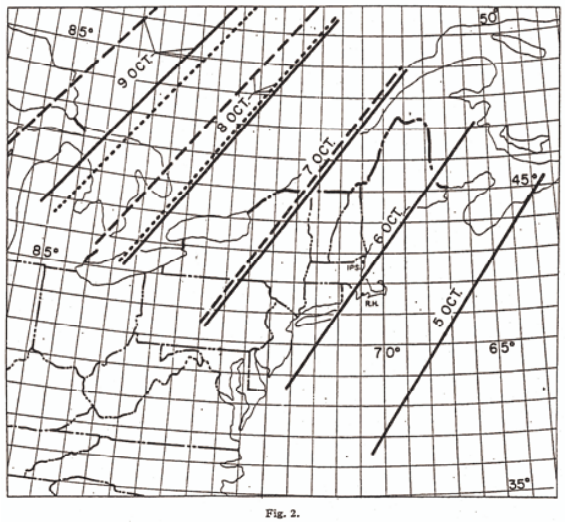

Figure 1.7: Sputnik - observed and predicted ground track over the US on October 5-7, 1957. Historical data from Brown et al. [Reproduced from [1]]

be developed further. Their predictability is an essential element in their computation.

1.3 Transit

1.3.1 System Overview

The development of Transit was initiated in 1958 at the dawn of the satellite area. Everything had to be invented from satellite technology, satellite control to signal generation and navigation algorithms. The stabilization of the satellite was achieved using a boom and the gradient of the gravitational field. Electronics were operated in vacuum by the US, while the Soviet Union placed its electronics in gas-filled tanks, which made electronic packaging and cooling easier but required tight seals around cables to external components. Power was experimentally provided by nuclear reactors, ultimately by panel mounted solar cells (see [7] for details on satellite development).

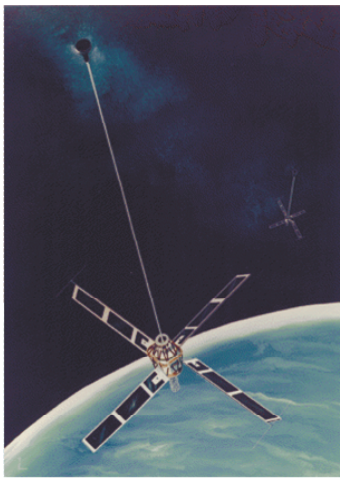

Figure 1.8: Oscar satellite first launched on October 6th 1964. The satellite was equipped with solar panels and a boom for stabilizing the attitude by taking advantage of the gradient of the gravitational field. [Public Domain, Wikipedia]

The nominal system configuration of Transit consisted of 4 satellites revolving on polar orbits at an altitude of 1075 km with a period of 107 minutes[6] (see Figure 1.9). These satellites transmitted carriers at 150 MHz and 400 MHz. They were modulated by a synchronization pattern for detecting two minutes intervals, and a navigation message containing a description of the satellite's orbit. Let $b(t)$ be the modulation, then the signal transmitted on each frequency f' is given by:

$$s(t) = b(t)\cos(2\pi f' t).$$

The parametrization of the Transit orbits is mainly of historical interest. It is very similar to the parametrization of the GPS orbits described in detail in Chapter 4. This description required the development of a global reference system and the accurate modeling of the earth gravitational field. Both were initiated, and later developed to a very high level of sophistication by Transit, see Chapter 9. Transit was operated by the Navy Astronautic Group at Point Mugu, California. Tracking stations were located at Prospect Harbor (Maine), Rosemount (Minnesota) and Wahiawa (Hawaii). They sent their Doppler measurements to the computing center at Point Mugu, which then determined the orbits and generated the predictions for several hours. Furthermore, the US Naval Observatory (USNO) was synchronizing Transit time to Universal Time Coordinated

[6]The description of the Transit System is based on [4] and [8].

Figure 1.9: The Transit constellation consisted of 4-7 satellites in a polar orbit. In May 1975 it consisted of 5 satellites [9].

(UTC). The navigation message was uplinked every 12 hours, and the clock on the satellite reset to approximate UTC to the best possible degree. The system became operational in December 1963 and was made available to civil users in late 1967. Transit was shutdown in 1996, once GPS, the next generation satellite navigation systems had achieved full operational capabilities.

1.3.2 Hyperbolic Positioning

The positioning method for Transit described below is due to Kershner and Newton [4]. It builds on a modified version of Equation (1.3). With $\vec{r}$, and $\vec{R}$ denoting the user's and satellite's positions respectively, and $\vec{s} = \vec{r} - \vec{R}$, as well as $s = \|\vec{s}\|$, the Doppler shift described by Equation (1.3) becomes:

$$f_D = -\frac{\dot{\vec{s}}}{c} \cdot \frac{\vec{s}}{\|\vec{s}\|} f' = -\frac{\dot{s}}{c} f'$$

with f' being the carrier frequency of the satellite. The signal at the receiver thus becomes:

$$u(t) = \alpha(t)\, b(t - \frac{s}{c}) \cos\Big(2\pi(f' + f_D)(t - \frac{s}{c}) + \varphi_0\Big) + n(t),$$

with $\alpha(t)$ and $n(t)$ being the attenuation, and noise, respectively, and $t - s/c$ being the retardation due to finiteness of the velocity of light. It would also occur if the receiver was moving with the satellite. The approximation $b(t - s/c) \simeq b(t)$ is an excellent one, since the data rates are very low.

A typical user receiver does not carry an ultra stable clock. Its clock may thus be offset[7] with respect to f':

$$f = f' + \delta.$$

It is with this clock that the receiver demodulates the signal by forming:

$$u(t)\cos(2\pi f t) = \tilde{b}(t)\cos(2\pi\Delta f t + \Delta\varphi) + \tilde{n}(t),$$

and thus measures Δf :

$$\Delta f = f_D - \delta = f_D + f' - f. \tag{1.4}$$

The quantities $\tilde{b}(t)$, and $\tilde{n}(t)$ are easily derived. Amongst others, they define the signal-to-noise ratio, and thus the accuracy of the frequency estimate. The details of frequency measurements are described in Chapter 6 in the section on frequency locked loop.

[7] The frequency offset of the receivers's local oscillator δ is assumed to be constant during the measurement process.

The user's receiver has the task of determining its position $\vec{r}$ and the frequency offset δ from the modified Doppler measurements Δf. In a first step the receiver demodulates the navigation message and determines the satellite's orbit. It then uses the two minute synchronization pattern to define the boundaries of the integration intervals $[t_{i-1}, t_i]$ with:

$$t_i = t_0 + n_i(f' + f_D), \qquad \text{and} \qquad n_i = \frac{120}{f'} i \text{ [s/Hz]},$$

and integrates the modified Doppler using Equation (1.4):

$$N_i = \int_{t_{i-1}}^{t_i} dt\, \Delta f(t) = -(s_i - s_{i-1})\frac{f'}{c} - \Delta,$$

with $s_i = s(t_i) = \|\vec{r}(t_i) - \vec{R}(t_i)\|$, and $\Delta = (t_i - t_{i-1})\delta$. The difference $t_{i+1} - t_i = 120(1 + f_d/f') = 120(1+v/c)$ also depends on the Doppler shift, which shall be neglect however, since $v/c \sim 2.5{\cdot}10^{-5}$ in the present setup.

The receiver integrates the Doppler over a number of consecutive intervals to obtain the differences

$$s_i - s_{i-1} = \|\vec{r}_i - \vec{R}_i\| - \|\vec{r}_{i-1} - \vec{R}_{i-1}\| = -(N_i + \Delta)\frac{c}{f'}. \tag{1.5}$$

These differences $s_i - s_{i-1}$ are negative as long as the satellite approaches the user. They become positive when it flies away. Each of the differences include three unknown position vectors $\vec{r}_i$ and one term Δ proportional to the unknown frequency offset Δ. Additionally each equation with the index $i, i-1, \ldots$ adds three components of the unknown vector $\vec{r}_j$ with $j = i-1, i-2, \ldots$. This is more unknowns than equations. If the user was able to relate his positions at the boundaries of the integration intervalls, the situation would change, and the total number of unknown would be limited to four. Three measurements or more and the knowledge that the user is on the surface of the earth, would then allow to solve the equations. A first class of users, who can relate their positions at different time instants are the surveyors. If they don't move their equipment, the latter is only displaced due to earth rotation. Some other users might have means to determine their position changes by inertial measurements. Submarines provide provide and example, in which this is the case. The need for the determination of the receiver's movement during the measurements is a clear weakness of Transit, which was accepted due to the initial lack of alternatives. Now, assume that the receiver can determine the sequence of difference vectors

$$\Delta\vec{r}_i = \vec{r}_i - \vec{r}_{i-1}$$

for $i, i-1, i-2$, then Equation (1.5) can be rewritten in the form:

$$\|\vec{r}_i - \vec{R}_i\| - \|\vec{r}_i - (\vec{R}_{i-1} + \Delta\vec{r}_i)\| = -(N_i + \Delta)\frac{c}{f'}. \tag{1.6}$$

The coordinate transformation:

$$\vec{y}_i = \vec{r}_i - \frac{1}{2}(\vec{R}_i + \vec{R}_{i-1} + \Delta\vec{r}_i),$$

and

$$\vec{e}_i = -\frac{1}{2}(\vec{R}_i - \vec{R}_{i-1} - \Delta\vec{r}_i),$$

leads to the following equivalent form of Equation (1.6):

$$\|\vec{y}_i + \vec{e}_i\| - \|\vec{y}_i - \vec{e}_i\| = -(N_i + \Delta)\frac{c}{f'}.$$

This equation describes a hyperboloid of revolution in the variable $\vec{y}_i$, with foci $\pm\vec{e}_i$, and semi-major axis $a_i = -(N_i + \Delta)\frac{c}{2f'}$ (see Appendix A for details). The vector

$$S_i = \frac{1}{2}(\vec{R}_i + \vec{R}_{i-1} + \Delta\vec{r}_i)$$

defines the coordinate transformation. It is the mid-point between the location of the satellite at time t_i and the its location at time t_{i-1} corrected by the movement of the user during that time. The transformation can be rewritten in the form

$$\vec{y}_i = \vec{r}_i - \vec{S}_i, \qquad \text{and} \qquad \vec{e}_i = -(\vec{R}_i - \vec{S}_i).$$

Thus the user is located on a hyperboloid easily described in an adequate coordinate system. The focus of that hyperboloid actually is at the location of the satellite at time t_i. Positive values of a_i imply that $\vec{y}_i \cdot \vec{e}_i > 0$, while negative values imply the reverse. This means that the hyperboloid is always open towards the receiver.

Considering a second interval provides another hyperboloid of revolution, which intersects the first one along a closed curve. Under the assumption that the frequency offset Δ is known and that the receiver is located on earth, the intersection of this closed curve with the surface of the earth typically provides two possible locations, from which the user has to choose the correct one (see Figure 1.10). This is unambiguous, whenever the locations are far apart. A degenerate situation occurs when the satellite flies over the receiver's location. In its proximity, the determination of the longitude becomes uncertain, since the hyperboloids intersect the ellipsoid in a tangential manner. In reality, the frequency offset δ is unknown. This is the reason for considering the intersection of three hyperboloids with the surface of the earth. The measurements N_i, N_{i-1}, N_{i-2} and the intersection with the earth, e.g. modeled as an ellipsoid of revolution with semi-major axis a_e, and linear eccentricity[8] e determine the parameters $\vec{r}_i = (x_i, y_i, z_i)^T$ and Δ through the non-linear equations

$$\begin{aligned} f_0(\vec{r}_i, \Delta) &= \frac{x_i^2 + y_i^2}{a_e^2} + \frac{z_i^2}{a_e^2(1-e^2)} - 1 \\ f_j(\vec{r}_i, \Delta) &= \|\vec{r}_i - (\vec{R}_{i-j} + \sum_{k=i-j+1}^{i} \Delta\vec{r}_k)\| - \|\vec{r}_i - (\vec{R}_{i-j-1} + \sum_{k=i-j}^{i} \Delta\vec{r}_k)\| + (N_{i-j} + \Delta)\frac{c}{f'}, \end{aligned}$$

with $j \in \{1, 2, 3\}$. Let $\xi^T = (\vec{r}_i^{\,T}, \Delta)$, then the system of non-linear equations $f(\xi) = 0$ can be solved numerically using the Newton-Raphson method, which is described in Appendix B. This requires considering the equations $f(\xi) + \operatorname{grad}_\xi(f)\,\Delta\xi = 0$ linearized around a first approximation $\vec{r}_0$ of the position and δ_0 of the frequency offset:

$$\begin{pmatrix} f_0(\xi_0) \\ f_1(\xi_0) \\ f_2(\xi_0) \\ f_3(\xi_0) \end{pmatrix} + \begin{pmatrix} \frac{2x_i}{a_e^2} & \frac{2y_i}{a_e^2} & \frac{2z_i}{a_e^2(1-e^2)} & 0 \\ (\vec{e}_1 - \vec{e}_0)_x & (\vec{e}_1 - \vec{e}_0)_y & (\vec{e}_1 - \vec{e}_0)_z & \frac{c}{f'} \\ (\vec{e}_2 - \vec{e}_1)_x & (\vec{e}_2 - \vec{e}_1)_y & (\vec{e}_2 - \vec{e}_1)_z & \frac{c}{f'} \\ (\vec{e}_3 - \vec{e}_2)_x & (\vec{e}_3 - \vec{e}_2)_y & (\vec{e}_3 - \vec{e}_2)_z & \frac{c}{f'} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta\Delta \end{pmatrix} =$$

$$f(\xi_0) + H(\xi_0)\,\Delta\xi = 0 \qquad (1.7)$$

with

$$\vec{e}_j = \frac{\vec{r}_i - (\vec{R}_{i-j} + \sum_{k=i-j+1}^{i} \Delta\vec{r}_k)}{\|\vec{r}_i - (\vec{R}_{i-j} + \sum_{k=i-j+1}^{i} \Delta\vec{r}_k)\|} = \frac{\vec{r}_{i-j} - \vec{R}_{i-j}}{\|\vec{r}_{i-j} - \vec{R}_{i-j}\|}$$

being the vector pointing from the satellite to the receiver at time t_{i-j}. Since a Transit satellite moves significantly during a time span of two minutes, the vectors $\vec{e}_j - \vec{e}_{j-1}$ are rather different.

[8] The linear eccentricity of an ellipse with semi-major axis a_e and semi-major axis b_e is $e = \sqrt{a_e^2 - b_e^2}$

Since the satellite essentially moves in a plane, one vector is dependent on the other two. The last column, however, typically prevents the matrix from becoming singular. Thus the equation can be resolved for $\Delta\xi$, providing a new approximation $\xi_1 = \xi_0 + \Delta\xi$. This is iterated until the desired level of convergence is achieved. Since the measurements are noisy, additional measurements can be taken. In this case the matrix is extended with additional rows. The solution is then derived using the least squares method described in Appendix A to Chapter 3.

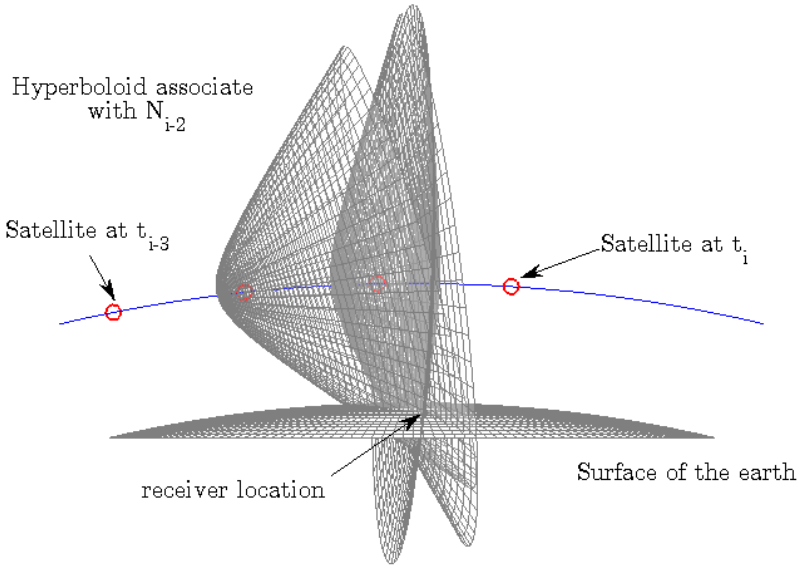

Figure 1.10: The satellite moves on its orbit (blue). It is in discrete positions at time $t_{i-3} \ldots t_i$. The receiver integrates the Doppler shift between these times, and for each such interval determines the associated hyperboloid of possible receiver locations. The intersection of three such hyperboloids with the surface of the earth determines the receiver location up to a two fold ambiguity.

Although Transit was a fantastic technical achievement in all respects, tremendously contributing to the development of geodesy, as well as navigation and satellite technology, it had clear limitations. The limited number of satellites in low altitudes resulted in long waiting times to obtain a fix. This could take up to 100 minutes, and was acceptable to stationary users or users with stable inertial platforms such as submarines but not to most others. Furthermore, the position determination method of Transit required that the receiver's movement be known accurately. An error of 1 nmi/h caused a positioning error of 0.2 nmi. These deficiencies triggered the next step, which aimed at a direct, continuous and simultaneous measurement of the distance between the user and an adequate number of satellites. The next section discusses corresponding work performed at the Naval Research Laboratory (NRL). The next chapter describes the developments carried out by the Air Force, as well as by the Joint Program Office. The latter effort ultimately led to the conception of the Global Positioning System (GPS).

1.4 Timation - Time Synchronous Positioning

The timation project (Time Navigation) was initiated in 1964 at the Naval Research Laboratory for providing position and time to a wider variety of users. It included three elements: the direct estimation of the distance between the receiver and satellites, the determination of position and time by the receiver, and the synchronization of stable clocks on the ground and on the satellites.

The technical approach to range estimation was based on Side Tone Ranging. This method had been introduced by Earp [10] in 1964, and proposed for satellite navigation by Easton in 1970. It is further explained below. The associated measurements are so-called pseudoranges. Pseudoranges differ from ranges in being the product of the velocity of light and a propagation time measured using satellite clocks, which are globally synchronized, on the transmit side and a user clock, which is not synchronized, on the receive side. The knowledge of at least four pseudoranges is sufficient for determining the position and the time offset of the local clock, as shall be developed in Chapter 3.

The most stable clocks available in the late 1960's were temperature stabilized quartz oscillators. They were considered in the initial phase of the Timation project. Such an oscillator flew on Timation I. In an attempt to improve the long term drift of the oscillator, Timation II was equipped with a compensation for quartz aging. It was a surprise that the performance was worse than for Timation I. The explanation was that the exposure to radiation compensated most of the quartz aging on Timation I, and that the correction thus overcompensated the phenomenon. Quartz oscillators were replaced by atomic clocks in a later phase of the project.

Side Tone Ranging (STR) uses a family of subcarriers to measure distances. A spherical wave transmitted by a satellite is described by

$$A' \frac{\cos[k'r - \omega'(t - \delta_S)]}{r^2},$$

with r being the distance between the observer and the satellite, t being the time of observation, δ_S being the phase offset with respect to an appropriate reference, $\omega' = 2\pi f'$ being the angular frequency of the transmitted signal, and $k' = \omega'/c$ being the associated wave vector. Note that the latter expression assumes that there is no dispersion, i.e., that the phase velocity is equal to the velocity of light in vacuum c.

The receiver demodulates the signal using a copy of the carrier generated by a local oscillator:

$$\cos[\omega(t - \delta)],$$

with ω being the receiver's estimate of the carrier frequency (including e.g. Doppler shifts), and δ being its clock offset. This leads to the following result:

$$\frac{A}{2} \frac{\cos[k'r - (\omega' - \omega)(t - \delta) + \omega'(\delta_S - \delta)]}{r^2} + \epsilon_A,$$

with a channel attenuation A/A' and white Gaussian additive amplitude noise ϵ_A. The above expression is independent of t if $\omega = \omega'$, i.e., the phase of the demodulated signal comes to rest if this condition is fulfilled. Defining ρ as the noisy estimate of the argument divided by k' leads to

$$\rho = r + c(\delta_S - \delta) + \epsilon. \tag{1.8}$$

The variable ρ is the desired pseudorange, i.e., the desired range r up to an offset. The demodulation process translates the amplitude noise ϵ_A into phase noise ϵ. This is discussed in detail in Chapter 6. The phasing δ_S of the satellite clocks with respect to a common reference is measured by the ground segment in a complex estimation process and broadcast in the navigation message. The offset of the receiver's clock δ is jointly determined with the position by using at least four phase measurements associated with four different satellites. The number of measurements is the same as for Transit. Their nature is different, however. In Transit three measurement intervals and a representation of the surface of the earth are used to determine the position and the frequency-offset of the clock. In pseudorange-based position-determination, the phases of the signals from four satellites are used to determine the position and time-offset of the clock. The details are described in Chapter 3.

The phase of the cosine wave can be separated into an integer $N \in \{0, 1, \ldots\}$, and a fractional part ϕ :

$$\rho = \lambda(\phi + N) + \epsilon.$$

The fractional part can always be measured. The integer part, i.e., the number of wavelengths λ laying between the satellite and the receiver, must be determined differently. In the absence of noise $\epsilon = 0$, the wavelength could be chosen large enough ($>$ maximum distance between satellite and receiver) to eliminate the ambiguity, i.e. to achieve $N = 0$. Since the noise limits the position accuracy to a value that is typically in the range of 1/100-th to 1/1000-th of a wavelength, this would lead to a rather inaccurate pseudorange as well as position and time determination. For this reason the process is iterated with a family of subcarriers, the lowest having the largest wavelength. The subcarriers have frequencies:

$$f_i = \nu_{i-1} f_{i-1} \qquad i = 1, \ldots, n,$$

which implies that the wavelengths are related by

$$\lambda_{i-1} = \nu_{i-1} \lambda_i, \qquad i = 1, \ldots, n.$$

The lowest subcarrier with the index $i = 0$ must either fulfill $N_0 = 0$ or its ambiguity must be inferred from physical restrictions, e.g., knowing that the user is on earth implies a given value of N_0. Each phase ϕ_i is directly observable, and provides the ambiguity for the next subcarrier:

$$N_i = \lfloor \nu_{i-1} \phi_{i-1} \rfloor, \qquad i = 1, 2, \ldots, n.$$

The phase of the last subcarrier ϕ_n provides the analog estimate of the position. This iterative determination of the pseudorange is shown in Figure 1.11 for three iterations and frequencies that double at each iteration. The specific family of subcarrier frequencies considered by Easton

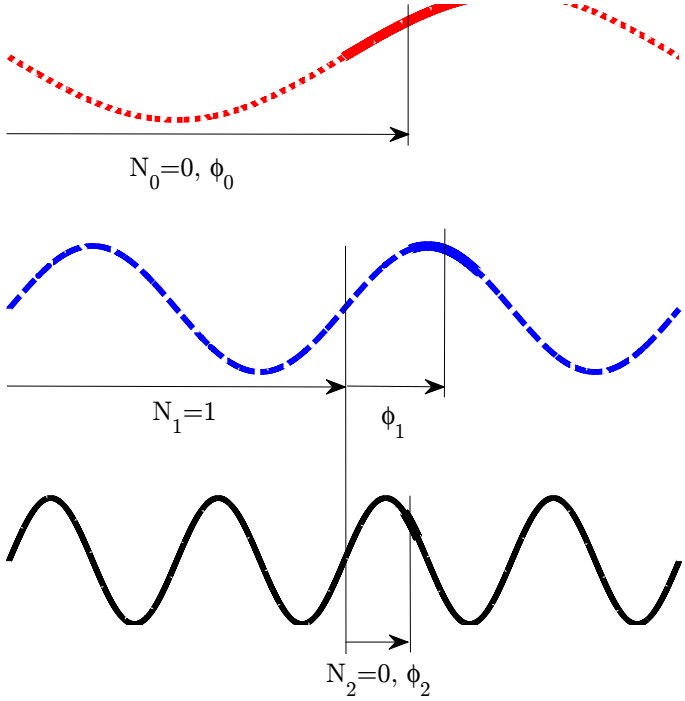

Figure 1.11: The accuracy of a measurement is proportional to the wavelength of the signal considered. The uncertainty is amplified by a factor of more than 100 in the picture for visualization purposes. The phase can be directly measured with the stated accuracy modulo an integer number of wavelengths. The latter integer is determined by evaluating the phases at the previous level.

included: 100 and 300 Hz, as well as 1, 3, 10, 30, and 100 kHz [11], i.e. $\nu = (\nu_1 \ldots \nu_6) = (3, 10/3, 3, 10/3, 3, 10/3)$. The pseudorange is then obtained from:

$$\begin{aligned} \hat{\rho} &= \lambda_1 N_1 + \lambda_2 N_2 + \ldots + \lambda_n (N_n + \phi_n) \\ &= \sum_{i=1}^{n} \lambda_i \lfloor \nu_{i-1} \phi_{i-1} \rfloor + \lambda_n \phi_n. \end{aligned}$$

This is very similar to the hierarchy meter, decimeter, centimeter, millimeter,... of the metric system.

The first satellite to implement the new approach described above was Timation I. It transmitted STR signals modulated on a carrier of 400 MHz. The satellite was successfully launched in May 67 and led to the first time transfer experiments with the US Naval Observatory (USNO) in October 67. Positioning in d space-dimensions would have required at least $d+1$ satellites and was thus not possible. Its successor Timation II included a number of improvements, in particular a second signal at 150 MHz for compensating the extra delay due to the propagation through the ionospheric plasma (see Chapter 7), as well as side tones up to 1 MHz, and the mentioned "improved" oscillators.

Timation transitioned into the Navigation Technology Satellite program. Timation IIIA also called NTS-1 was to become the first satellite to fly an atomic clock. It was launched into an orbit with an altitude of 13'800km and a small eccentricity[9] $\epsilon = 0.007$ in July 74. The clock was a Rb-normal based on a clock conceived by Efratom in Munich. The latter clock was the most compact design of its time, nearly 10 times lighter than any other clock with a comparable stability. It had been conceived by Jechart and Huebner. The main ranging method on NTS-1 was still STR. However, this satellite also carried the first spread spectrum payload provided by the USAF's programm 621B. This leads us to the origins of GPS in the next chapter.

Figure 1.12: NTS-1 Satellite with L-band and UHF-band antennas and the laser reflector in the fourth position. The solar panel are not yet mounted. [Courtesy NRL]

1.5 Summary

Sputnik was the first satellite orbiting the earth. The low altitude of its orbit was associated with a large Doppler shift of its radio signals. It was soon recognized that this Doppler shift could be used for estimating Sputnik's orbit from measurements taken by a small number of ground

[9]The eccentricity is the ratio of the linear eccentricity e and the semi-major axis a: $\epsilon = e/a = \sqrt{1 - b^2/a^2}$.

receivers. The knowledge of a satellite's orbit allowed to inverse the problem and to solve for the user position by integrating the Doppler shift over a sequence of periods. This approach was at the root of the Transit program. Transit was the first satellite navigation system. It was developed for positioning submarines but also played an important role in geodesy, and led to the first World Geodetic System (WGS). Transit's main weaknesses were that its users needed to be capable of determining their own movement during the measurement phase, as well as the latency caused by the positioning method, as well as due to the sparseness of the constellation. Finally, Transit was a two dimensional positioning system, which limited its use in aeronautics. Thus, work on developing a system that would provide three dimensional positioning, instantaneously, at any time and everywhere was initiated.

Appendix A Hyperboloid of Revolution

A hyperboloid is described by the equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

It can equivalently be parameterized by two variables $t \in (-\infty, \infty)$ and $\phi \in [0, 2\pi)$:

$$\begin{aligned} x &= a\cosh t \\ y &= b\sinh t\cos\phi \\ z &= c\sinh t\sin\phi \end{aligned}$$

The sections $y = \alpha$ or $z = \alpha$ are hyperbolas, this applies in particular for $\alpha = 0$. The sections $x = \alpha$ is an ellipse with semi- major axis $b\sqrt{\alpha^2/a^2 - 1}$ and $c\sqrt{\alpha^2/a^2 - 1}$ if $\alpha > a$. In the case that $c = b$, the ellipse becomes a circle. In this case, the hyperboloid is called a hyperboloid of revolution. It can be obtained by rotating the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

around the x-axis. The hyperbola and thus the hyperboloid have two foci in $\vec{e} = (e, 0, 0)^T$ and $-\vec{e}$ with

$$e = \sqrt{a^2 + b^2}.$$

The latter quantity is also called linear eccentricity of the hyperbola. The foci have the crucial property that

$$\|\vec{r} + \vec{e}\| - \|\vec{r} - \vec{e}\| = 2a. \tag{1.9}$$

The proof is left as an exercise. The comparison of the Equations (1.6) and (1.9) implies that the following quantities have to the be identified:

$$2\vec{e} = -(\vec{R}_i - \vec{R}_{i-1} - \Delta\vec{r}_i)$$

and

$$2a = -(\Delta N_i + \Delta)\frac{c}{f'}.$$

The former specifies the foci. The latter completes the characterization of the hyperboloid.

Appendix B The Newton-Raphson Method

Many problems lead to the determination of zeros ξ of non-linear functions $f(\xi)$. Such zeros can rarely be determined in closed form. The solution must be determined numerically. Newton and

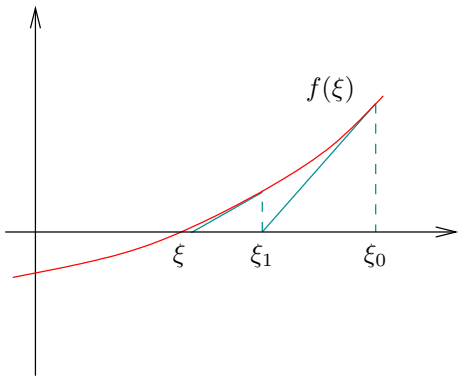

Figure 1.13: the Newton-Raphson method approximates the function by its tangent at the n-th approximation of the zero in order to compute the $n+1$-th approximation.

Raphson invented independently a very effective method in the second half of the 17th. century. Today, his method is best introduced through a Taylor expansion

$$f(\xi) = f(\xi_0) + f'(\xi_0)(\xi - \xi_0) + \sum_{m=2}^{\infty} \frac{f^{(m)}(\xi_0)}{m!}(\xi - \xi_0)^m.$$

Such an expansion converges if the function is analytical in the neighborhood of ξ_0. If the desired zero ξ is in that neighborhood, one can apply the following iterative procedure to determine increasingly accurate approximations of ξ: compute ξ_{n+1} from

$$f(\xi_n) + f'(\xi_n)(\xi_{n+1} - \xi_n) = 0,$$

i.e.:

$$\xi_{n+1} = N(\xi_n) = \xi_n - \frac{f(\xi_n)}{f'(\xi_n)}.$$

The iteration is started with an approximation ξ_0 of the zero ξ such the the above assumptions are met. A few iterations are shown in Figure 1.13. The residual error $\Delta_n = \xi_n - \xi$ can also be determined iteratively:

$$\begin{aligned}
\Delta_{n+1} &= \xi_{n+1} - \xi \\
&= N(\xi_n) - \xi \\
&= N(\xi + \Delta_n) - \xi \\
&= \Delta_n - \frac{f(\xi + \Delta_n)}{f'(\xi + \Delta_n)} \\
&= \Delta_n - \frac{\Delta_n f'(\xi) + \frac{1}{2}\Delta_n^2 f''(\xi) + \mathrm{O}(\Delta_n^3)}{f'(\xi) + \Delta_n f''(\xi) + \mathrm{O}(\Delta_n^2)} \\
&= \Delta_n^2 \frac{f''(\xi)}{2f'(\xi)} + \mathrm{O}(\Delta_n^3)
\end{aligned}$$

If the function is monotonic in the interval between ξ_0 and ξ, the convergence is quadratic, i.e., the number of correct decimals doubles at each step.

The Newton-Raphson Method for scalar functions can be generalized to k non-linear functions

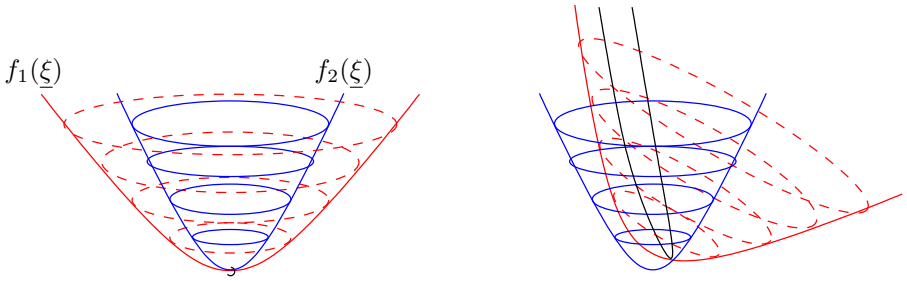

Figure 1.14: The Jakobian is non-singular if the intersection has maximal dimension. This is the case in the right figure but not in the left one.

in k unknown. In this case, the linearized equation reads

$$\begin{pmatrix} f_1(\underline{\xi}_n) \\ \vdots \\ f_k(\underline{\xi}_n) \end{pmatrix} + \begin{pmatrix} \text{grad}_{\underline{\xi}} f_1(\underline{\xi}_n) \\ \vdots \\ \text{grad}_{\underline{\xi}} f_k(\underline{\xi}_n) \end{pmatrix} \cdot (\underline{\xi}_{n+1} - \underline{\xi}_n) = 0.$$

The matrix is called the Jakobian, it must be non-singular, which means that the spaces must intersect each other in the maximum dimension. Figure 1.14 shows an example for $k = 2$.

Exercises

1. Doppler-shift
   - Consider the situation shown in Figure 1.6, and assume that the satellite transmits a carrier of frequency f'. Derive the Doppler shift f_D observed at the receiver.
   - Define $\rho = \|\vec{r}\|$, compute $\dot{\rho}$ in terms of $\vec{r}$, and its derivatives.
   - The earth has a GM $= 398'600.4415$ [km^3/s^3]. Sputnik had a semi-major axis of 6'955.2 [km]. How large is the maximum Doppler shift at the surface of the earth for the 20 [MHz] carrier?
     Hint: Let the earth be a sphere and neglect the earth rotation.
   - Assume that the satellite also transmits a navigation message. Does the rate at which the information is received change over one pass? If yes by how much?
2. Hyperbolic and Elliptic Geometry
   - A hyperbola can be defined by the equation
     $$\|\vec{r} + \vec{e}\| - \|\vec{r} - \vec{e}\| = 2a$$
     i.e. Equation (1.9), with $\pm\vec{e}$ being the two foci. Derive this equation from the alternative definition provided in Appendix A.
     Hint: Use the parametrization in terms of hyperbolic functions.
   - Draw the hyperboloid of revolution defined by Equation (1.9).
   - An ellipse is defined by the equation
     $$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$
     Indicate coordinates, which lead to a simple representation of the ellipse.

- Find a similar representation to Equation (1.9) in the elliptical case.

3. The US navy navigation satellite system Transit employs a constellation of polar satellites. The orbits, the earth radius, and the transmit frequency are as defined in the text. The receiver are assumed to be at sea level.

   - How many satellites does a receiver need to see in order to determine its position?
   - How long is the arc on the equator from which the signal of a given satellite can be received at the time it crosses the equator?
     Hint: A satellite is visible in locations where its elevation is positive. Make a drawing and then compute the value.
   - Determine the angular frequency Ω of the orbit of Transit satellites.

4. Positioning with Transit: The receiver measures the Doppler shift as a function of time plotted in Figure 1.15. Let the satellites orbit in "Earth Fixed Earth Centered" coordinates, i.e. in the coordinate system that rotates with the earth, be:

$$\vec{r}_{\mathrm{SV}} = (R_e + h)(\sqrt{3}/2 \sin(\Omega t + \varphi), -0.5 \sin(\Omega t + \varphi), \cos(\Omega t + \varphi))^T,$$

   with Ω denoting the above angular frequency of the satellite orbit, and $\varphi = 2.0813$ rad.

   - Determine the receiver's latitude.
   - Compute the closest distance h between the satellite and the receiver.
     Hint: Do not apply hyperbolic positioning.
   - Compute the receiver's longitude. Is it unique?
   - Draw the receiver's position relative to the satellite orbit (all options if the position is not unique).

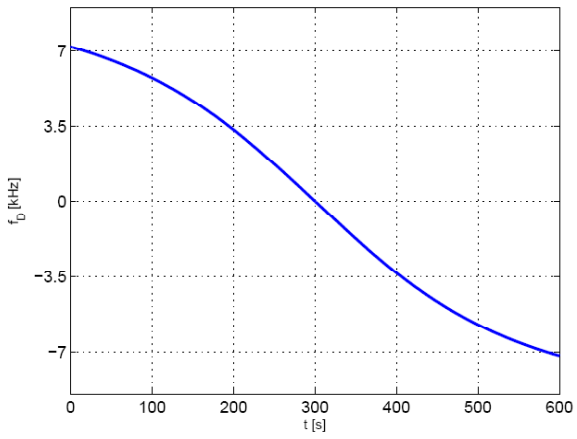

Figure 1.15: Measured Doppler shift over time.

5. Side Tone Ranging: Consider a satellite, which transmits a carrier modulated by a sinusoidal signal of frequency $f_n = 8$, 32, 160, 800, 4000, 20'000, and 100'000 [Hz]. Let $c = 299'792'458$ m/s be the velocity of light.

   - Compute the wavelength of the modulation of frequency f_n.

- Assume that the phase ϕ_n is corrupted by zero mean additive white Gaussian noise (AWGN) with a standard deviation $\sigma_n = \lambda_n/100$. Determine the probability that the error of the first phase measurement fulfills $|\hat{\phi}_1 - \phi_1| > 374$ [km].
- Compute a good approximation to the probability $p(|\hat{\phi} - \phi| < 30[\text{cm}])$.
  Hint: As an approximation you might consider the case that all estimates of the ambiguities are correct and that the final estimate is in the desired interval.
- Describe the next level of approximation.

Bibliography

[1] R.R. Brown, P.E. Green, Jr., B. Howland, R.M. Lerner, R. Manasse, G. Pettengill, "Radio Observation of the Russian Earth Satellite," *Proc. IRE*, pp. 1552-1553, Nov. 1957.

[2] A.M. Peterson, "Radio and Radar Tracking of the Russian Earth Satellite," *Proc. IRE*, pp. 1553-1555, Nov. 1957.

[3] W.H. Guier, G.C. Weiffenbach, "Theoretical Analysis of Doppler Radio Signals from Earth Satellites," *APL, The John Hopkins Univ., Bumblebee Series Report,* No. 276, Apr. 1958.

[4] R.B. Kershner, R.R. Newton, "The Transit System," *J. Inst. of Nav.,* vol. 15, pp. 129-144, Apr. 1962.

[5] W.H. Guier, G.C. Weiffenbach, "Genesis of Satellite Navigation," *John Hopkins APL Techn. Digest,* vol. 19, pp. 14-17, 1998.

[6] K. Giger, *Private Communication,* 2009.

[7] R.J. Danchik, "An Overview of Transit Development," *John Hopkins APL Techn. Digest,* vol. 19, pp. 18-26, 1998.

[8] B.W. Parkinson, T. Stansell, R. Beard, K. Gromov, "A History of Satellite Navigation," *J. of The Inst. of Nav.,* vol. 42, pp. 109-164, 1995.

[9] H.D. Black, R.E. Jenkins, L.L. Pryor, *Technical Memo. - The Transit System, 1975,* John Hopkins Univ., Appl. Phys. Lab., 1975.

[10] W.W. Earp, "Radiolocation System Transimitting Sideband Signals," *US Patent* 3'339'202, Aug. 1967 (filed Oct. 5th, 1964).

[11] R.L. Easton, "Navigation System Using Satellites and Passive Ranging Techniques," *US Patent* 3'789'409, Jan. 1974 (filed Oct. 8th, 1970).

[12] B. Parkinson, *Private Communication,* 2008.

[13] J.A. Buisson, R.L. Easton, Th.B. McCaskill, "Initial Results of the NAVSTAR GPS NTS-2 Satellite," *Internal Report Nav. Res. Lab.,* May 1978.