AAE 590

Space Traffic Management

Fall 2023

CAROLIN FRUEH

Version 6.0

Contents

Chapter 1

Some Notes on Probabilities

Very few things are certain in life... and even fewer quantities in engineering, in general, and Space Traffic Management (STM) and Space Situational Awareness (SSA), in particular. Every measurement, for example, is a carrier not only of information but also of uncertainty.

Hence some notes on probabilites are in order.

1.1 Random Variables

Random Variable - A Definition A random variable X is, in the simplest terms, a variable that takes on values at random and realizations of the random variable may be thought of as the outcomes of some random experiment.

1.1.1 Univariate

Probability Distribution Function also called Cumulative distribution function (CDF) The manner of specifying the probability with which different values are taken by the random variable is by the probability distribution function $F(x)$:

$$F_X(x) = F(x) := \Pr(X \leq x) \tag{1.1}$$

F represents the probability that the random variable X has values less than a particular value denoted by x.

Note that we sometimes write the probability distribution function as $F_X(x)$, but oftentimes the subscript is dropped for notational simplicity.

Probability Density Function (pdf) Often, one is interested in the probability in a local vicinity of a given value. In this case the probability density function $p(x)$ is used:

$$p_X(x) = p(x) := \frac{dF(x)}{dx} \tag{1.2}$$

By applying the definition of the derivative it follows that

$$p(x) = \frac{dF(x)}{dx} = \lim_{dx \to 0} \frac{F(x+dx) - F(x)}{dx} = \lim_{dx \to 0} \frac{\Pr(x \leq X \leq x+dx)}{dx} \tag{1.3}$$

That is, the interpretation of $p(x)$ is that it is the *density* of probability of the event that X takes on in the vicinity of x.

This function is finite if the probability that X takes a value in the infinitesimal interval between x and $x+dx$ is an infinitesimal of order dx. This is usually true for a continous random variable.

The inverse relationship between the distribution and density function is

$$F(x) = \int_{-\infty}^{x} p(u)du \tag{1.4}$$

The characteristic of any probability distribution is hence

$$F(\infty) = \int_{-\infty}^{\infty} p(u)du = 1 \tag{1.5}$$

since the total probability of the random variable occurring pver all possible values must be equal to one.

This is the same as saying that the probability density function must integrate to unity when the limits of integration are taken to be the support of the density. The **support** is often from minus infinity to infinity.

The discrete valued random variable pdf If X takes on any of a set of discrete values, x_i, with nonzero probabilities p_i, $p(x)$ is infinite at these values of x.

This is expressed as a series of Dirac delta "functions" weighted by the appropriate probabilities

$$p(x) = \sum_i p_i \delta(x - x_i) \tag{1.6}$$

An example of such a random variable is the outcome of the roll of a die.

The Dirac Delta, $\delta(x)$, describes a functional relationship that is zero everywhere, except at $x = 0$, where it is infinite in such a way that the integral of the function across the singularity is unity.

Sifting Property An important property of the Dirac delta, which follows from this definition, is that for a finite-valued function $g(x)$ that is continuous at $x = x_0$

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)dx = g(x_0) \tag{1.7}$$

Hybrid Random Variables A random variable may take on values over a continuous range and, additionally, take a discrete set of values with nonzero probability.

The resulting probability density function includes both a finite function of x and an additive set of probability-weighted delta functions; such a distribution is called *hybrid* or *mixed*.

1.1.2 Multivariate

The simultaneous consideration of more than one random variable is often necessary.

Joint Probability Distribution or Joint Cumultative Distribution Function In the case of two random variables, for instance, the probability of the occurrence of pairs of values in a given range is given by the joint probability distribution or joint cumultative distribution function

$$F(x,y) = \Pr(X \leq x \text{ and } Y \leq y) \tag{1.8}$$

where X and Y are the random variables of interest.

Joint Probability Density Function The corresponding joint probability density function is

$$p(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{1.9}$$

The individual probability distribution and density functions for X and Y can be found from the joint distribution and density functions. For example

$$F_X(x) = F(x,\infty) \tag{1.10}$$

$$p_X(x) = \int_{-\infty}^{\infty} p(x,y)dy \tag{1.11}$$

and similar relationships hold for $F_Y(y)$ and $p_Y(y)$.

Independence If X and Y are independent, the event $X \leq x$ is independent of the event $Y \leq y$; thus, the probability of the joint occurrence of these events is the product of the probabilities of the individual events:

$$F(x,y) = \Pr(X \leq x \text{ and } Y \leq y) \tag{1.12}$$

$$= \Pr(X \leq x)\Pr(Y \leq y) \tag{1.13}$$

$$= F_X(x)F_Y(y) \tag{1.14}$$

Similarly, for the joint probability density function of two independent random variables:

$$p(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{1.15}$$

$$= \frac{\partial^2 F_X(x)F_Y(y)}{\partial x \partial y} \tag{1.16}$$

$$= \frac{\partial F_X(x)}{\partial x}\frac{\partial F_Y(y)}{\partial y} \tag{1.17}$$

$$= p_X(x)p_Y(y) \tag{1.18}$$

1.2 Statistics of Random Variables

The pdf can provide comprehensive probability distribution information. However, it might be not always available. Also, one is interested in statistical information that allows to characterize a given probability distribution. Therefore, so-called moments are used.

First moment The expectation, which is the same as the mean and the same as the first moment of a random variable is defined as the integration (continuous) or sum (discrete) of all values that the random variable may take, each weighted by the probability with which the value is taken. The probability, in the limit as $dx \to 0$, that X takes a value in the infinitesimal interval of width dx near x is $p(x)dx$.

Therefore, the expectation of X, which we denote by $\mathrm{E}\{X\}$ or μ_X is (continuous case)

$$\mu_X = \mathrm{E}\{X\} = \int_{-\infty}^{\infty} xp(x)dx \tag{1.19}$$

The values of the random variable might not be of direct interest, but rather the expectation of a function of the random variable is sought. Assume Y is a function of the random variable X via

$$Y = f(X) \tag{1.20}$$

Then Y is itself a random variable with a distribution derivable from the distribution of X.

Thus the expectation of *any* function of X can be calculated directly from the distribution of X by the integral

$$\mathrm{E}\{Y\} = \mathrm{E}\{f(X)\} = \int_{-\infty}^{\infty} f(x)p(x)dx \tag{1.21}$$

Second Raw Moment and Root-Mean-Square Raw moments are moments around zero. The second raw moment, which is the same as the mean squared value. From the definition of the expected value, the expectation or mean of the square of X is

$$\mathrm{E}\{X^2\} = \int_{-\infty}^{\infty} x^2 p(x)dx \tag{1.22}$$

The root-mean-squared (rms) value of X is the square root of $\mathrm{E}\{X^2\}$.

Second Central Moment and Standard Deviation Central moment are moments around the mean. Central and raw moments are the same, if the mean is zero. The variance, which is the same as the second central moment of a random variable is the mean squared deviation of the random variable from its mean; it is often denoted by σ^2, where

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mathrm{E}\{X\})^2 p(x)dx \tag{1.23}$$

$$= \mathrm{E}\{(X - \mathrm{E}\{X\})^2\} \tag{1.24}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mathrm{E}\{X\})^2 p(x)dx \tag{1.25}$$

$$= \int_{-\infty}^{\infty} (x^2 - 2x\mathrm{E}\{X\} + \mathrm{E}\{X\}^2) p(x)dx \tag{1.26}$$

$$= \int_{-\infty}^{\infty} x^2 p(x)dx - 2\mathrm{E}\{X\} \int_{-\infty}^{\infty} xp(x)dx + \mathrm{E}\{X\}^2 \int_{-\infty}^{\infty} p(x)dx \tag{1.27}$$

$$= \mathrm{E}\{X^2\} - 2\mathrm{E}\{X\}\mathrm{E}\{X\} + \mathrm{E}\{X\}^2 \tag{1.28}$$

$$\sigma^2 = \mathrm{E}\{X^2\} - \mathrm{E}\{X\}^2 \tag{1.29}$$

The square root of the variance, or σ, is the standard deviation of the random variable.

Note: The rms value and the standard deviation are equal only for a zero-mean random variable.

Multivariate Variance: Covariance The second direct moment or covariance of two random variables X and Y is given by the expectation of the product of the deviations of the random variables from their respective means, such that

$$\mathrm{E}\{(X-\mathrm{E}\{X\})(Y-\mathrm{E}\{Y\})\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}\{X\})(y-\mathrm{E}\{Y\})p(x,y)dxdy \tag{1.30}$$

As before, we can expand out the product leads to:

$$\mathrm{E}\{(X-\mathrm{E}\{X\})(Y-\mathrm{E}\{Y\})\} = \mathrm{E}\{XY\}-\mathrm{E}\{X\}\mathrm{E}\{Y\} \tag{1.31}$$

Correlation The covariance, normalized by the standard deviations of X and Y, is called the correlation coefficient

$$\rho = \frac{\mathrm{E}\{XY\}-\mathrm{E}\{X\}\mathrm{E}\{Y\}}{\sigma_X\sigma_Y} \tag{1.32}$$

The correlation coefficient is a measure of the degree of linear dependence between X and Y:

- if X and Y are independent, $\rho = 0$
- if Y is a linear function of X, $\rho = \pm 1$

Note: if $\rho = 0$, it is not necessarily true that X and Y are independent. It can only be said that X and Y are uncorrelated. Independence of the random variables implies that they are uncorrelated, but the reverse is not true (e.g. $X \sim U(-1,1)$; $Y = X^2$).

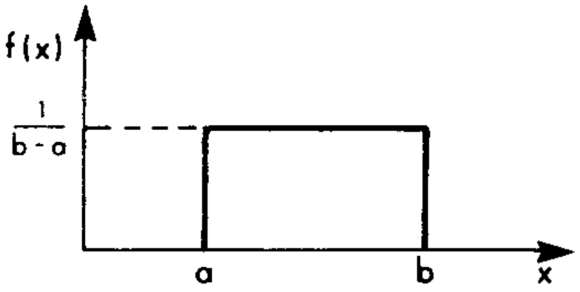
1.3 Some Probability Distributions

Three probability distributions that cannot be avoided in this class.

Uniform Distribution The uniform distribution is characterized by a uniform (constant) probability density over some finite interval.

The magnitude of the density function in this internal is the reciprocal of the interval width, as required to make the integral of the probability density function unity.

$$p(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{1.33}$$



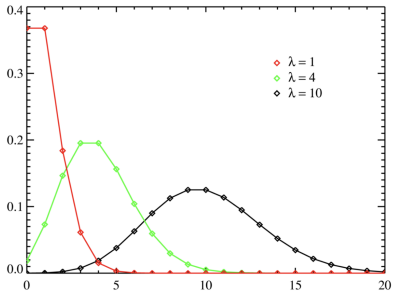
Mean and variance of the uniform distribution are:

$$\mathrm{E}\{X\} = \mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12} \tag{1.34}$$

The uniform distribution is used, e.g., in the characterization of the admissible regions in first orbit determination from only one measurement.

Poisson Distribution The Poisson distribution is discrete and characterized by the parameter λ:

$$p(n) = \frac{\lambda^n}{n!} e^{-\lambda} \qquad \text{for } n \in \mathbb{N}^+ \tag{1.35}$$

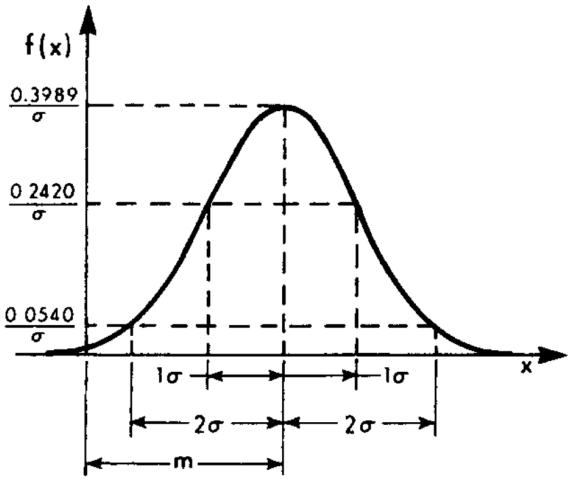


$$\mathrm{E}\{X\} = \mu = \lambda \qquad \sigma^2 = \lambda \tag{1.36}$$

Normal Distribution - Univariate The normal Gaussian probability density function is characterized by its mean μ and its variance σ^2:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.37}$$

The integral of the normal function is unity, which is required for this to be a valid probability density function. This is not the case for a general Gaussian function, without the normalizing amplitude factor $\frac{1}{\sigma\sqrt{2\pi}}$.



The area within the ± 1 σ bounds (centered about the mean) is approximately 0.68.
The area within the ± 2 σ bounds (centered about the mean) is approximately 0.95.
As an interpretation, the probability that a normally distributed random variable resides outside of the ± 2 σ bounds is approximately 0.05. It is important to note that these specific values hold only for the univariate case.

The distribution of a sum of independent normally distributed variables is also normally distributed. This is even true even if the random variables within the sum are not independent.

Normal Distribution - Multivariate For the case of n random variables that are jointly Gaussian, the probability density function takes the form:

$$p(\boldsymbol{x}) = |2\pi\boldsymbol{P}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.38}$$

with

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1.39}$$

The quantities $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are, respectively, the mean and covariance of the vector $\boldsymbol{x}$:

$$\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{x}\} \qquad \text{and} \qquad \boldsymbol{P} = \mathrm{E}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\} \tag{1.40}$$

That is, the definitions of the mean and covariance is taken on an element-wise basis.

For the mean, this implies that

$$\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{x}\} = \begin{bmatrix} \mathrm{E}\{x_1\} \\ \mathrm{E}\{x_2\} \\ \vdots \\ \mathrm{E}\{x_n\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{1.41}$$

Similarly, for the covariance:

$$\boldsymbol{P} = \begin{bmatrix} \mathrm{E}\{(x_1-\mu_1)(x_1-\mu_1)\} & \mathrm{E}\{(x_1-\mu_1)(x_2-\mu_2)\} & \cdots & \mathrm{E}\{(x_1-\mu_1)(x_n-\mu_n)\} \\ \mathrm{E}\{(x_2-\mu_2)(x_1-\mu_1)\} & \mathrm{E}\{(x_2-\mu_2)(x_2-\mu_2)\} & \cdots & \mathrm{E}\{(x_2-\mu_2)(x_n-\mu_n)\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\{(x_n-\mu_n)(x_1-\mu_1)\} & \mathrm{E}\{(x_n-\mu_n)(x_2-\mu_2)\} & \cdots & \mathrm{E}\{(x_n-\mu_n)(x_n-\mu_n)\} \end{bmatrix} \tag{1.42}$$

This means $\boldsymbol{P} = \boldsymbol{P}^T$; the covariance matrix is symmetric.

For the case $n = 1$, it follows that the vector-valued $\boldsymbol{x}$ becomes the scalar-valued x, with mean and covariance

$$\boldsymbol{\mu} = \mu \qquad \text{and} \qquad \boldsymbol{P} = \sigma^2 \tag{1.43}$$

This leads exactly to the univariate case:

$$p(x) = |2\pi\sigma^2|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)\sigma^{-2}(x-\mu)\right\} \tag{1.44}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.45}$$

In the case of independent Gaussian random variables, the i^{th} element of the mean vector remains the same

$$\mathrm{E}\{x_i\} = \mu_i \tag{1.46}$$

The covariance, however, is diagonal with the i^{th} row, j^{th} column being given by

$$P_{ij} = \mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = \sigma_i^2 \delta_{ij} \tag{1.47}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise.

In the independent case, the probability density function becomes

$$p(\boldsymbol{x}) = \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}\right\} \tag{1.48}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2}\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right\} \tag{1.49}$$

As expected, since the random variables are all independent, the joint probability density function reduces down to the product of the individual probability density functions.

Central Limit Theorem If the random variables are independent and their mean and variance are finite, the distribution of the sum of those independent random variables, each having an arbitrary distribution, tends toward a normal distribution as the number of variables in the sum tends toward infinity.

Chapter 2

Catalogs and Data Formats

2.1 USSPACECOM and Two-Line Elements (TLEs)

A publicly available catalog of the known, unclassified, and known origin (launch, shedding from known unclassified spacecraft) of orbital elements is provided by the United States Space Command, short USSPACECOM or SPACECOM (formerly: US Strategic Command (USSTRATCOM)).

USSPACECOM is one of the currently 11 unified combatant commands of the US Department of Defense. The United States Space Force's 18th Space Control Squadron is a space control unit located at Vandenberg Space Force Base, California.

18th Space provides continuous and uninterrupted support to the Space Surveillance Network (SSN). As of Sep. 24, 2020, 18th Space Control Squadron began publicly sharing data for debris-on-debris conjunction predictions via www.Space-Track.org, while previously only collision data messages were sent to owner-operators for conjunction predictions with active assets involved [1]. From the same source in 2020: *The 18th SPCS monitors approximately 3,200 active satellites for close approaches with approximately 24,000 pieces of space debris, and issues an average of 15 high-interest warnings for active near-earth satellites, and ten high-interest warnings for active deep-space satellites, each day.*. Now, 2023, the space-track website lists 8700 active payload, 16'800 analyst objects and 19200 debris objects with a total of 44700 specific objects that are tracked.

The US Space Surveillance Network, shown in Fig.2.1 denotes not only the ground-based radar and electro-optical (EO) sensors but includes its communications links, processing centers, and data distribution channels, and also the space-based assets.

Several Canadian sensors, part of the CSSS (Canadian Space Surveillance System), contribute to the USSPACECOM catalog data.

An integral component of the SSN is the so-called GEODSS (Ground-based, Electro-optical Deep Space Surveillance) and the Space Fence. The former Space Fence, AN/FPS-133 Air Force Space Surveillance System, ceased operation in 2013.

The new Space Fence contracted to Lockheed Martin is operational since March 2020 [42], It is said that the budget for the fence was US $1.594 billion [2]. Gallium Nitride (GaN) powered S-band ground-based radars are operated in the fence: The smaller wavelength allows for the detection of smaller objects. The space fence is suspected to detect well over 200'000 objects, mainly consisting of smaller, that is, less reflective objects in LEO. Á certain fraction of the detected objects from the space-fence are feeding into the USSPACECOM catalog.

USSPACECOM also owns and operates space-based surveillance assets.

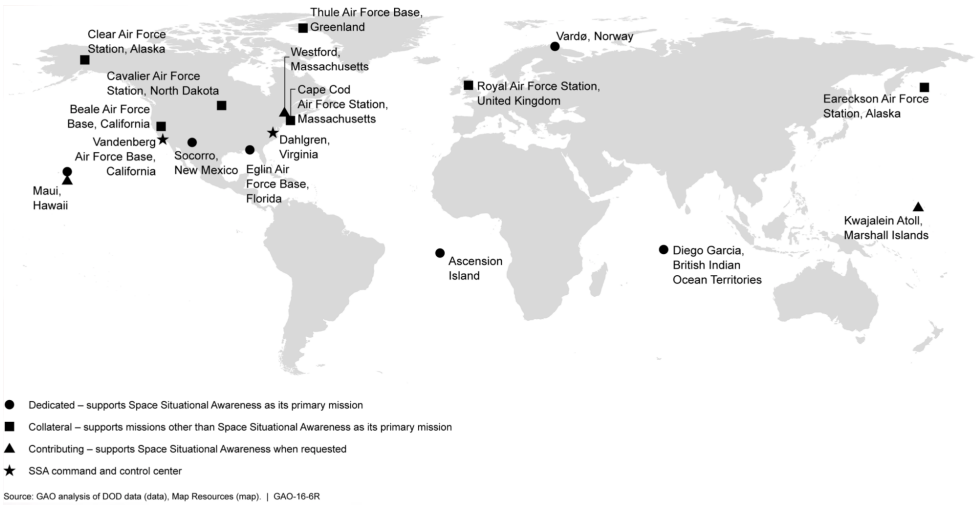


Figure 2.1: Ground-based sensors of the USSPACECOM Space Surveillance network (SSN) [20]; the map is missing the contributing Harold E. Holt (HEH) sensor, Exmouth, Western Australia.

The mission of the Space-Based Space Surveillance satellite (Pathfinder SBSS-1 satellite) has been extended past its predicted end of life in 2016 and seemed still operational Oct 2020 [73].

In the Geosynchronous Space Situational Awareness Program (GSSAP), currently, four satellites are operational [72], two more are scheduled to become operational in 2021 [59]. All of them operate on the electro-optical waveband. The USSPACECOM catalog is considered complete for objects in the geosynchronous region to object sizes of around 1 meter and around 10cm in Low Earth orbit, excluding the space fence data.

Of course, the sizes assume a favorable albedo of the observed objects. Objects with complex dynamics, such as High-Area-to-Mass (HAMR) objects, are independent of their size not generally maintained in the catalog. Objects whose origin cannot be associated with a specific launch are provided in the supplementary dataset on www.space-track.org and are not part of the so-called *TLE catalog*. The USSPACECOM catalog has been provided in the so-called two-line element (TLE) format. but has moved to the more flexible orbital mean message (OMM) data format.

Note: The TLE format provides the orbital element data (can be interpreted as a mean) but no uncertainty information. The OMM format may contain uncertainty information. Conjunctions are provided as so-called conjunction data messages (CDM).

2.1.1 The Historic TLE Format

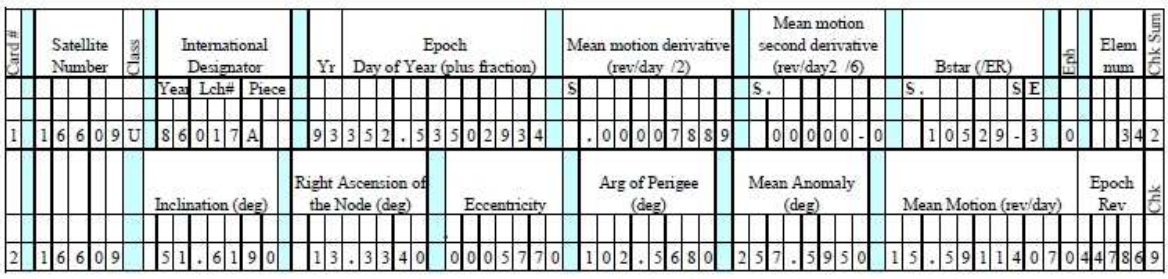


Figure 2.2: The two-line element set (TLE) format [74]. Shaded cells do not contain data. S indicates that the cell is either blank or a sign, either + or −, can be displayed. E is the exponent in base 10. Eccentricity, mean motion derivative, and Bstar have implied decimal points before the first digit. The mean motion derivative is divided by 2, the second derivative by 6. The units of the first and second derivatives of the mean motion are rev/day^2 and rev/day^3.

The TLE format is a fixed format, which was originally developed for punch cards. For every entry, a fixed number of columns is reserved, including decimal points. Subsequently each entry is briefly explained [74], [16]:

1. The first number in each row indicates the row number. The TLE format consists of two rows.
2. The satellite number has been traditionally the NORAD number. NORAD stands for North American Aerospace Defense Command and is a joint organization of the United States and Canada. NORAD assigns continuous numbers to objects according to their first observation date. For a valid two-line element set, the NORAD number has to be repeated in the second line. However, with the increased space traffic and the better detection capabilities for example with the Space Fence, the NORAD numbering scheme has been exhausted. Instead the new Alpha-5 format has been adopted, while being still compatible with the (TLE/3LE) format. Object numbers below 100,000 are unaffected by Alpha-5, but in order to present legacy operations that depend on 5-digit integers, the 1st digit of the 5-digit object number is replaced with an alphanumeric character. Note that only capital letters and numbers are used in Alpha-5. The letters *I* and *O* are omitted to avoid confusion with the numbers one and zero. Thus, Alpha-5 can incorporate 240'000 more numbers. Legacy API Classes *tle*, *tle_latest*, and *tle_publish* have not been changed to Alpha-5, see Fig.2.3 for illustration.
3. The class indicates if the object is classified or unclassified. All publicly available data is unclassified. An empty entry indicates unclassified data.
4. The international launch designator is assigned by the World Data Center-A for rockets and satellites (and parts thereof) in accordance with the international *Convention on Registration of Objects launched into outer space*.

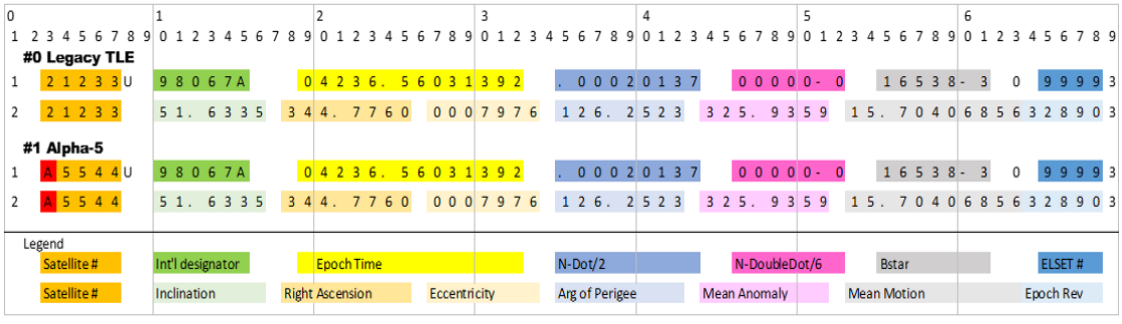


Figure 2.3: The ALPHA-5 TLE format. ALPHA-5 is extending the satellite number to be able to incorporate up to represent 240,000 more numbered objects [74].

The World Data Center-A cooperates with the North American Aerospace Defense Command (NORAD) and the National Space Science Data Center (NSSDC) of the National Aeronautics and Space Administration (NASA). The first two digits of the launch designator represent the year of launch, the launch number of that year, which is counted continuously within one year, and three digits reserved for letters representing the pieces of the same launch.

5. The first two digits of the epoch denote the year. The next three digits denote the day of the year, and the digits after the decimal point indicate the fraction of the day in decimal units. The epoch starts at UT midnight and is measured in UTC.

6. The mean motion derivative has an implicit leading decimal point before the first digit. It can be preceded by a sign (+ or −). It is already divided by two to be used directly in the calculation of the resistance coefficient of the SGP/SDP model. Details on the SGP/SDP models can be found in Section 2.1.4.

7. The second derivative of the mean motion can carry a signed exponent to the base ten (±). It is already divided by six to be used directly in the calculation of the resistance coefficient of the SGP/SDP model. It is not used for the SGP4/SDP4 model; it is only valid for older SGP models. Its value is often displayed as zero. Details on the SGP/SDP models can be found in Section 2.1.4.

8. Bstar is a drag-like coefficient in SGP4. It is an adjustment to the physical quantity of the ballistic coefficient (B_c). Bstar is using a reference value for the atmospheric density, ρ_0, at the height of one Earth radius.

$$B_c := \left(c_D \cdot \frac{A}{m} \right)^{-1} = \frac{R_e \rho_0}{2 \cdot Bstar} \tag{2.1}$$

with: c_D drag coefficient, A effective cross-sectional area, m mass, R_e earth radius, $\rho_0 = 2.461 \times 10^{-5} \mathrm{kg/m^2}$ atmospheric density at one Earth radius.

Bstar is not a physical quantity but a free modeling parameter. The value may not be correlated to drag effects. This is the case in the presence of satellite maneuvers, significant solar radiation pressure, atmospheric perturbations, large third body effects, or mis-modeling of the Earth's gravitational field. Bstar may have a negative value.

9. The ephemeris type determines the model with which the ephemerides were generated. Spacetrack Report Number 3 suggests the following assignments: 1=SGP, 2=SGP4, 3=SDP4, 4=SGP8, 5=SDP8. The field is blank or filled with a zero for all TLEs used outside of Cheyenne Mountain Operations Center (CMOC) of USSPACECOM. All TLE data is generated with SGP4/SDP4 in those cases.

10. The ephemerides number is a continuous data set number incremented each time a new data set is generated. This rule is not strictly followed, however.

11. The checksum number is a modulo 10 checksum. The checksum is calculated by taking the modulus 10 of the sum of all digit entries in the current line, ignoring all letters, plus-signs, and decimal points. A value of 1 is assigned to each minus sign. The majority of errors, which are likely to happen in the TLE generation process, are detected via the checksum.

12. The entries in the second row of the TLEs contain the orbital elements of the satellite orbit: Inclination in degrees, right ascension of ascending node in degrees, eccentricity with a leading decimal point, the argument of perigee in degrees, mean anomaly in degrees at the epoch displayed, mean motion in revolutions per day. Those are mean orbital elements generated with SGP4/SDP4 for publicly available TLE data. The reference frame is a geocentric coordinate system using the **true equator and the mean equinox (TEME)** of the corresponding epoch.

13. The number of revolutions at epoch is represented by five digits. The revolution is counted from the ascending node onwards. In NORAD's convention, which is adapted for the TLE generation, the time period from launch till reaching the first ascending node is counted as revolution zero. Revolution one begins when the first ascending node is reached.

3LE refers to having the two-line format with an additional line preceeding the two lines listing the human readable satelliten name spelled out.

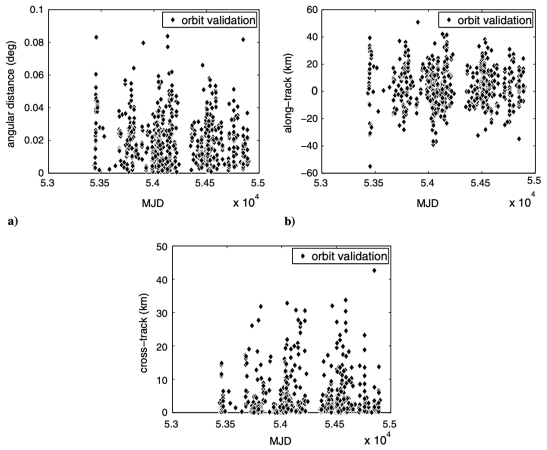


Figure 2.4: Difference between astrometric optical observations (uncertainty 2 arcseconds) and the TLE propagated to the observation epoch using SGP/SDP4 for geosynchronous objects [30].

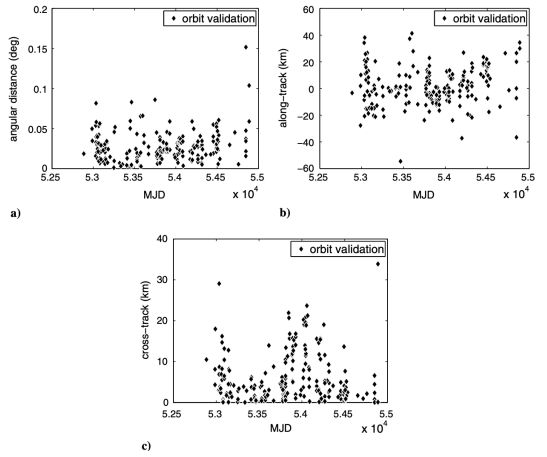


Figure 2.5: Difference between astrometric optical observations (uncertainty 2 arcseconds) and the TLE propagated to the observation epoch using SGP/SDP4 for objects in geostationary transfer orbits [30].

2.1.1.1 Problems and Future Developments

The quality of the TLE data is varying. As no uncertainty information is provided, the quality of a single TLE set is unknown.

A study conducted in 2008 showed a comparison between optical data collected at the Zimmerwald Observatory (Switzerland) with the predicted TLE measurements for objects in geosynchronous orbits and geostationary transfer orbits. Large differences of up to 50km were determined, Fig.2.4,2.5,2.6[30].

The accuracy of the TLE data is limited not only by the observations in the Space Surveillance Network or the orbit determination but also by the number of decimal digits available in each field [75].

With eight decimal places, the accuracy of the epoch is only accurate up to 0.0004 seconds. An object in a circular LEO orbit at an altitude of 400 km has a velocity of 7.6 km/s; it, therefore, moves by about 3 m in 0.0004 seconds. A GEO object in a perfectly geostationary orbit has a velocity of about 2.6 km/s. The error introduced in the position is of the order of one meter.

Parameter	Value
	GEO
angular $\pm \sigma$ deg	$2.02 \cdot 10^{-2} \pm 1.45 \cdot 10^{-2}$
along-track $\pm \sigma$ km	4.05 ± 23.25
cross-track $\pm \sigma$ km	4.96 ± 7.54
	HEO
angular $\pm \sigma$ deg	$2.73 \cdot 10^{-2} \pm 1.86 \cdot 10^{-2}$
along-track $\pm \sigma$ km	1.56 ± 24.25
cross-track $\pm \sigma$ km	6.44 ± 6.45

Figure 2.6: Summary of the results: Difference between TLE and measurements, not crosstrack represents absolute values only [30].

The eccentricity is specified by seven decimal places. This introduces an error of the order of $r \sim a\delta e$ corresponding to two meters for a GEO orbit. The inclination and right ascension of ascending node are only accurate to four decimal places, with a simple estimation of the semi-major axis times the inclination angle, an estimated error of 6 meters in LEO and of around 35 meters in GEO can be calculated.

Such errors are simply introduced by the TLE format.

2.1.2 The Orbital Mean Message Format

Consultative Committee for Space Data Systems (CCSDS) has developed an alternative format for reporting mean orbital data, the orbital mean message (OMM). The OMM is able to display the same information that is stored in the TLE format, without the limitations of significant digits. Furthermore, covariance information can be added at ease; although it is not distributed by space-track currently. Also, implicit assumptions, such as the TEME reference frame are made explicit.

```
GOES 9 [P]
1 23581U 95025A   07064.44075725 -.00000113  00000-0  10000-3 0  9250
2 23581   3.0539  81.7939 0005013 249.2363 150.1602  1.00273272 43169
```

Figure 2.7: Orbital Mean Message Example: A sample object classical TLE [66].

```
CCSDS_OMM_VERS = 2.0
CREATION_DATE  = 2007-065T16:00:00
ORIGINATOR     = NOAA/USA

OBJECT_NAME    = GOES 9
OBJECT_ID      = 1995-025A
CENTER_NAME    = EARTH
REF_FRAME      = TEME
TIME_SYSTEM    = UTC
MEAN_ELEMENT_THEORY = SGP/SGP4


EPOCH             = 2007-064T10:34:41.4264
MEAN_MOTION       = 1.00273272
ECCENTRICITY      = 0.0005013
INCLINATION       =   3.0539
RA_OF_ASC_NODE    =  81.7939
ARG_OF_PERICENTER = 249.2363
MEAN_ANOMALY      = 150.1602
GM                = 398600.8
EPHEMERIS_TYPE    = 0
CLASSIFICATION_TYPE = U
NORAD_CAT_ID      = 23581
ELEMENT_SET_NO    = 0925
REV_AT_EPOCH      = 4316
BSTAR             = 0.0001
MEAN_MOTION_DOT   = -0.00000113
MEAN_MOTION_DDOT  = 0.0
```

Figure 2.8: Orbital Mean Message Example: The corresponding OMM without covariance information [66].

```
CCSDS_OMM_VERS = 2.0
CREATION_DATE  = 2007-065T16:00:00
ORIGINATOR     = NOAA/USA

OBJECT_NAME    = GOES 9
OBJECT_ID      = 1995-025A
CENTER_NAME    = EARTH
REF_FRAME      = TEME
TIME_SYSTEM    = UTC
MEAN_ELEMENT_THEORY = SGP/SGP4

EPOCH             = 2007-064T10:34:41.4264
MEAN_MOTION       = 1.00273272
ECCENTRICITY      = 0.0005013
INCLINATION       =   3.0539
RA_OF_ASC_NODE    =  81.7939
ARG_OF_PERICENTER = 249.2363
MEAN_ANOMALY      = 150.1602
GM                = 398600.8

EPHEMERIS_TYPE    = 0
CLASSIFICATION_TYPE = U
NORAD_CAT_ID   = 23581
ELEMENT_SET_NO = 0925
REV_AT_EPOCH      = 4316
BSTAR             = 0.0001
MEAN_MOTION_DOT   = -0.00000113
MEAN_MOTION_DDOT  = 0.0

COV_REF_FRAME = TEME
CX_X =  3.331349476038534e-04
CY_X =  4.618927349220216e-04
CY_Y =  6.782421679971363e-04
CZ_X =  -3.070007847730449e-04
CZ_Y =  -4.221234189514228e-04
CZ_Z =  3.231931992380369e-04
CX_DOT_X =  -3.349365033922630e-07
CX_DOT_Y =  -4.686084221046758e-07
CX_DOT_Z =  2.484949578400095e-07
CX_DOT_X_DOT =  4.296022805587290e-10
CY_DOT_X =  -2.211832501084875e-07
CY_DOT_Y =  -2.864186892102733e-07
CY_DOT_Z =  1.798098699846038e-07
CY_DOT_X_DOT =  2.608899201686016e-10
CY_DOT_Y_DOT =  1.767514756338532e-10
CZ_DOT_X =  -3.041346050686871e-07
CZ_DOT_Y =  -4.989496988610662e-07
CZ_DOT_Z =  3.540310904497689e-07
CZ_DOT_X_DOT =  1.869263192954590e-10
CZ_DOT_Y_DOT =  1.008862586240695e-10
CZ_DOT_Z_DOT =  6.224444338635500e-10
```

Figure 2.9: Orbital Mean Message Example: The corresponding OMM with covariance information [66].

2.1.3 Conjunction Data Messages

The conjunction data messages (CDM) do entail mean and covariance information for the time of closest approach for the objects involved in a particular conjunction. The orbital data that is used to generate the CDM is of higher precision and accuracy than the catalog TLE data that is provided on all cataloged objects. The CDM format has been developed by the Consultative Committee for Space Data Systems (CCSDS) and is documented in the so-called Blue Book [67]. Some CDMs are publicly available on the space-track website.

CCSDS_CDM_VERS	= 1.0	
CREATION_DATE	= 2010-03-12T22:31:12.000	
ORIGINATOR	= JSPOC	
MESSAGE_ID	= 201113719185	
TCA	= 2010-03-13T22:37:52.618	
MISS_DISTANCE	= 715	[m]
OBJECT	= OBJECT1	
OBJECT_DESIGNATOR	= 12345	
CATALOG_NAME	= SATCAT	
OBJECT_NAME	= SATELLITE A	
INTERNATIONAL_DESIGNATOR	= 1997−030E	
EPHEMERIS_NAME	= EPHEMERIS SATELLITE A	
COVARIANCE_METHOD	= CALCULATED	
MANEUVERABLE	= YES	
REF_FRAME	= EME2000	
X	= 2570.097065	[km]
Y	= 2244.654904	[km]
Z	= 6281.497978	[km]
X_DOT	= 4.418769571	[km/s]
Y_DOT	= 4.833547743	[km/s]
Z_DOT	= −3.526774282	[km/s]
CR_R	= 4.142E+01	[m**2]
CT_R	= −8.579E+00	[m**2]
CT_T	= 2.533E+03	[m**2]
CN_R	= −2.313E+01	[m**2]
CN_T	= 1.336E+01	[m**2]
CN_N	= 7.098E+01	[m**2]
CRDOT_R	= 2.520E−03	[m**2/s]
CRDOT_T	= −5.476E+00	[m**2/s]
CRDOT_N	= 8.626E−04	[m**2/s]
CRDOT_RDOT	= 5.744E−03	[m**2/s**2]
CTDOT_R	= −1.006E−02	[m**2/s]
CTDOT_T	= 4.041E−03	[m**2/s]
CTDOT_N	= −1.359E−03	[m**2/s]
CTDOT_RDOT	= −1.502E−05	[m**2/s**2]
CTDOT_TDOT	= 1.049E−05	[m**2/s**2]
CNDOT_R	= 1.053E−03	[m**2/s]
CNDOT_T	= −3.412E−03	[m**2/s]
CNDOT_N	= 1.213E-02	[m**2/s]
CNDOT_RDOT	= -3.004E-06	[m**2/s**2]

CNDOT_TDOT	= -1.091E-06	[m**2/s**2]
CNDOT_NDOT	= 5.529E-05	[m**2/s**2]
OBJECT	= OBJECT2	
OBJECT_DESIGNATOR	= 30337	
CATALOG_NAME	= SATCAT	
OBJECT_NAME	= FENGYUN 1C DEB	
INTERNATIONAL_DESIGNATOR	= 1999-025AA	
EPHEMERIS_NAME	= NONE	
COVARIANCE_METHOD	= CALCULATED	
MANEUVERABLE	= NO	
REF_FRAME	= EME2000	
X	= 2569.540800	[km]
Y	= 2245.093614	[km]
Z	= 6281.599946	[km]
X_DOT	= -2.888612500	[km/s]
Y_DOT	= -6.007247516	[km/s]
Z_DOT	= 3.328770172	[km/s]
CR_R	= 1.337E+03	[m**2]
CT_R	= -4.806E+04	[m**2]
CT_T	= 2.492E+06	[m**2]
CN_R	= -3.298E+01	[m**2]
CN_T	= -7.5888E+02	[m**2]
CN_N	= 7.105E+01	[m**2]
CRDOT_R	= 2.591E-03	[m**2/s]
CRDOT_T	= -4.152E-02	[m**2/s]
CRDOT_N	= -1.784E-06	[m**2/s]
CRDOT_RDOT	= 6.886E-05	[m**2/s**2]
CTDOT_R	= -1.016E-02	[m**2/s]
CTDOT_T	= -1.506E-04	[m**2/s]
CTDOT_N	= 1.637E-03	[m**2/s]
CTDOT_RDOT	= -2.987E-06	[m**2/s**2]
CTDOT_TDOT	= 1.059E-05	[m**2/s**2]
CNDOT_R	= 4.400E-03	[m**2/s]
CNDOT_T	= 8.482E-03	[m**2/s]
CNDOT_N	= 8.633E-05	[m**2/s]
CNDOT_RDOT	= -1.903E-06	[m**2/s**2]
CNDOT_TDOT	= -4.594E-06	[m**2/s**2]
CNDOT_NDOT	= 5.178E-05	[m**2/s**2]

Figure 2.10: A sample Conjuntion Data Message (CDM) with the required entries [67]. Further explanations [67].

2.1.4 The Propagators: SGP4/SDP4, SGP8/SDP8, and SGP4-XP

The development of the Simplified General Perturbation (SGP) model for orbit determination and propagation started in the 1960s and became operational in 1970 in the Space Detection and Tracking System (SPADATS) Center, located in Colorado Springs, Colorado. Further improvements (SGP4/SDP4, SGP8/SDP8) and adjustments to the different orbital regimes were developed and implemented in the 1980s. The description of the different models are taken from Hoots [35] and Vallado [74].

The first semi-analytical model, called SGP, is based on the two different astrodynamic solutions for the equations of motion of a near-Earth satellite due to Brouwer [12],[13] and Kozai [40], both developed in 1959.

The gravitational field is represented only by the zonal harmonics up to degree five. For the development of the propagator theory the long- and short periodic terms, which do not have the eccentricity as an explicit factor, are adopted from Brouwer's solution.

From Kozai the convention relating mean motion and semi-major axis was adopted. The solutions are transformed into non-singular coordinates to avoid the singularities for small eccentricities and inclinations close to zero degrees; this approach was based on a work by Arsenault et al. [6].

An atmospheric drag model has been included, based on the ideas of King-Hele [39]. In a semi-empirical approach the effect of drag on the mean motion is represented as a quadratic time function, where the coefficients are parameters in the orbit determination. The time rate of the change of eccentricity is based on the assumption that the perigee height remains constant as the semi-major axis diminishes.

A first enhancement was performed in implementing an analytical rather than an empirical drag model. A simplified version of the work by Lane and Cranford [41] was implemented. The simplification consists of modeling only secular effects of drag. The model is known as SGP4. It replaced SGP as the sole model for the US satellite catalogue maintenance since 1979.

In 1977 an extension of the model was implemented for so-called deep space modeling (SDP4) in the existing SGP4 routines. The approach was based on the work by Bowman [11], who modeled the influence of the lunar and solar gravity and the resonance effects of the Earth's tesseral harmonics. It was incorporated as a first order model.

In the 1980s a further development leading to the SGP8/SDP8 was performed. Deficiencies in the re-entry prediction of decaying objects of the SGP4/SDP4 models were mitigated by a closed-form solution based on general trends of orbital element evolution near re-entry. The SGP4/SDP4 models are, however, still used without exception for the generation of publicly available TLEs of USSPACECOM.

The mathematical foundation of the SGP4/SDP4 model and the equations are published in Hoots [35].

A complete reworking of the propagators has been done under the lead of the Aerospace Coorporation, and a new model has been published in 2021, SGP4-XP. No legacy code from the SGP4 models is transferred. Critical expansions include the incorporation of solar radiation pressure, the effect of solar activity is included in the drag model with an improved ballistic coefficient and the propagator is valid for this cislunar domain. The promised performance improvements are two orders of magnitude while keeping the propagation speed of the semi-analytic model. The new model and a TLE catalog are available on the Government SPACETRACK.org web site.

Upon release it was expected that the public TLEs are generated and to be propagated with SGP4-XP by Dec 2021. Because the published elements are mean elements, the propagation models are not interchangable. The key is the ephemeris type in column 63 of the first line of TLE, **Ephtyp = 0 SGP4, and Ephtyp = 4 SGP4-XP**. [65].

2.2 Other Catalogs

The Keldish Institute of Applied Mathematics in collaboration with Roscosmos has been in the past providing data in the so-called the Vimpel catalog; as of 2023, the catalog is not publicly available any more. It contains mostly objects in high altitude orbits. The vimpel catalog has been not in two-line format but has been listing osculating states and did provide one variance in position.

Celestrak.com [71] out of AGI is providing the USSTRATCOM data with some limited additional services at no cost.

The DISCOS database (Database and Information System Characterizing Objects in Space) of the European Space Agency (ESA) is partially based on data supplied by USSPACECOM [28]. It provides the data of USSPACECOM catalogue in TLE format together with additional information, e.g., object type. No additional data with respect to the orbital elements are provided. DISCOS information is available through heavens-above.com [18].

Several private and government agencies own complete or incomplete data sets. As of now, none of them is publicly available. A good resource for data and space related services from any provicer is the Universal Data Library (UDL, *https://unifieddatalibrary.com*).

2.3 Other Data Formats

Other data formats are listed in the blue book and the pink book issued by the Consultative Committee for Space Data Systems (CCSDS). Those are standardized formats to transfer orbital data [66] and tracking data [68], that is observations. Two examples are shown here for the Tracking Data Messages. The first, Fig.2.11 shows an example of a TDM for optical ground-based observations, the second, Fig.2.12 shows an example of radar observations.

```
CCSDS_TDM_VERS = 12.0

COMMENT TDM example created by yyyyy-nnnA Nav Team (NASA/JPL)
COMMENT StarTrek: one minute of launch angles from DSS-16

CREATION_DATE = 2005-157T18:25:00
ORIGINATOR = NASA/JPL
META_START
TIME_SYSTEM = UTC
START_TIME = 2004-216T07:44:00
STOP_TIME =  2004-216T07:45:00
PARTICIPANT_1 = DSS-16
PARTICIPANT_2 = yyyy-nnnA
MODE = SEQUENTIAL
PATH = 2,1
ANGLE_TYPE = XSYE
CORRECTION_ANGLE_1 = -0.09
CORRECTION_ANGLE_2 = 0.18
CORRECTIONS_APPLIED = NO
META_STOP

DATA_START

ANGLE_1 = 2004-216T07:44:00 -23.62012
ANGLE_2 = 2004-216T07:44:00 -73.11035

ANGLE_1 = 2004-216T07:44:10 -23.04004
ANGLE_2 = 2004-216T07:44:10 -72.74316

ANGLE_1 = 2004-216T07:44:20 -22.78125
ANGLE_2 = 2004-216T07:44:20 -72.53027

ANGLE_1 = 2004-216T07:44:30 -22.59180
ANGLE_2 = 2004-216T07:44:30 -72.37598

ANGLE_1 = 2004-216T07:44:40 -22.40527
ANGLE_2 = 2004-216T07:44:40 -72.23730

ANGLE_1 = 2004-216T07:44:50 -22.23047
ANGLE_2 = 2004-216T07:44:50 -72.08887

ANGLE_1 = 2004-216T07:45:00 -22.08984
ANGLE_2 = 2004-216T07:45:00 -71.93750

DATA_STOP
```

Figure 2.11: A sample of a Tracking Data Message (TDM) for ground-based optical observations [68]. Further explanations [68].

```
CCSDS_TDM_VERS = 2.0
COMMENT Test file
CREATION_DATE = 2011-05-12T00:00:00.000
ORIGINATOR = ESA
META_START
COMMENT
TIME_SYSTEM = UTC
PARTICIPANT_1 = CAMRA
PARTICIPANT_2 = CRYOSAT
MODE = SEQUENTIAL
PATH = 1,2,1
EPHEMERIS_NAME = 3203_2013-11-09T23-02-30
RANGE_UNITS = km
ANGLE_TYPE = AZEL
CORRECTION_RANGE =  -1.48
CORRECTIONS_APPLIED = NO
META_STOP
DATA_START
RANGE =          2011-05-11T10:26:33.2613    2808.2696
ANGLE_1 =        2011-05-11T10:26:33.2613     191.40208435
ANGLE_2 =        2011-05-11T10:26:33.2613      25.44166756
CARRIER_POWER =  2011-05-11T10:26:33.2613     -36.73723984
RCS =            2011-05-11T10:26:33.2613       2.984
RANGE =          2011-05-11T10:26:33.7008    2803.1731
ANGLE_1 =        2011-05-11T10:26:33.7008     191.43959045
ANGLE_2 =        2011-05-11T10:26:33.7008      25.51874924
CARRIER_POWER =  2011-05-11T10:26:33.7008     -35.88296509
RCS =            2011-05-11T10:26:33.7008       2.992
RANGE =          2011-05-11T10:26:33.9686    2799.8754
ANGLE_1 =        2011-05-11T10:26:33.9686     191.46458435
ANGLE_2 =        2011-05-11T10:26:33.9686      25.56875038
CARRIER_POWER =  2011-05-11T10:26:33.9686     -36.67897415
RCS =            2011-05-11T10:26:33.7008     2.986
DATA_STOP
```

Figure 2.12: A sample of a Tracking Data Message (TDM) for ground-based radar observations [68]. Further explanations [68].

Chapter 3

Observations

3.1 Introduction

The (night) sky is showing a satellite, how do we identify it?

3.1.1 Sensors and their observables

- optical:
  - two angles α, δ *immediately*
  - two angles and angular rates $\alpha, \delta, \dot{\alpha}, \dot{\delta}$ (from two observations or extraction of relative velocities from a single image)
- radar:
  - range, two angles, r, α, δ (pointing angles, slant range)
  - range, range rate, two angles, $r, \dot{r}, \alpha, \delta$ (Doppler radar: pointing angles, slant range, Doppler range)
- laser ranging:
  - range, two angles, r, α, δ (pointing angles, slant range)

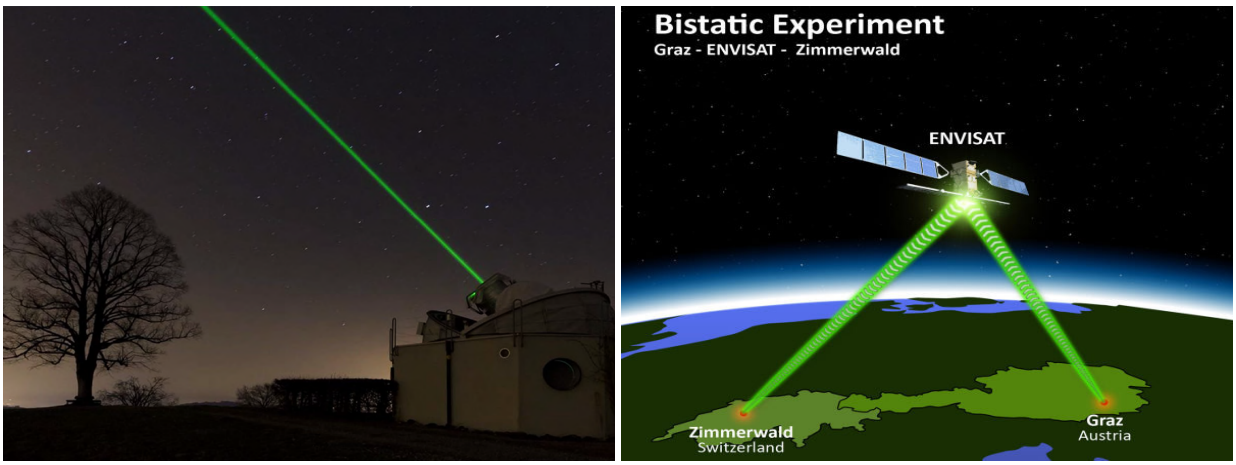


Figure 3.1: Laser Ranging Sensor and Measurement Principle.

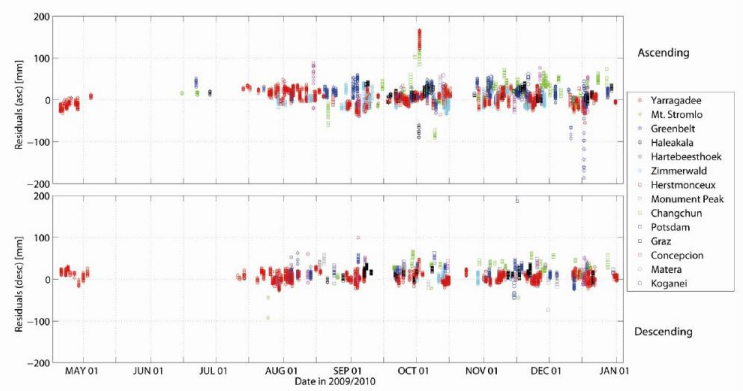


Figure 3.2: Laser Ranging Measurement Sample.

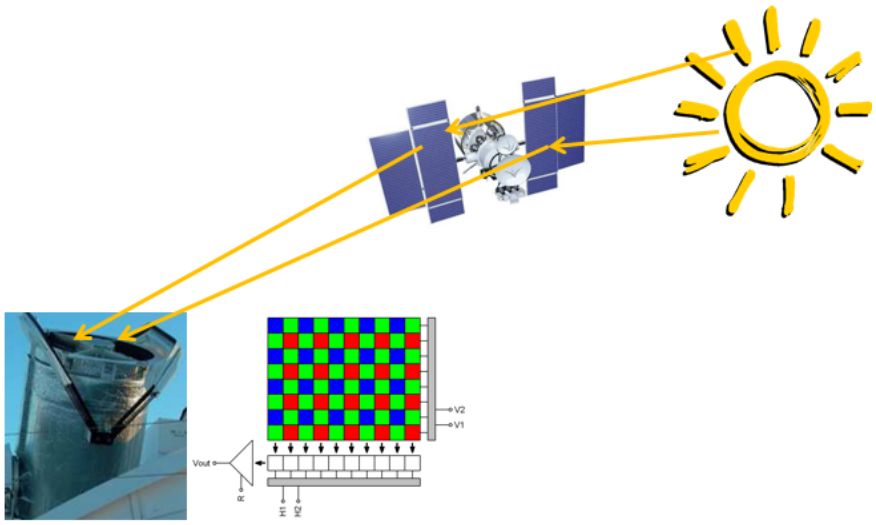


Figure 3.3: Optical Observation Principle.

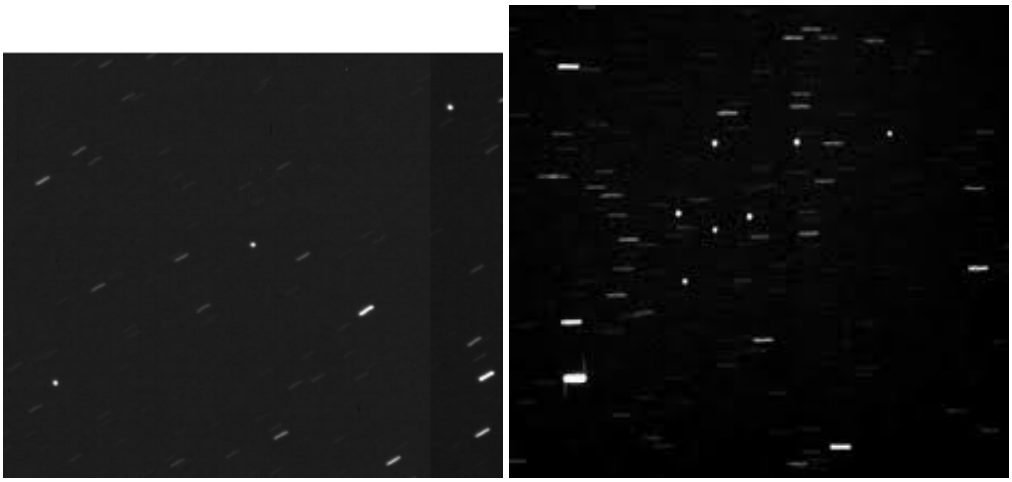


Figure 3.4: Optical Image Samples.

3.2 Electro-Optical (EO) Sensors

An electro-optical sensor consistents of the optic, collecting the light from the direction the sensor is pointed at and the detector. The pixel detector response of a single near-Earth object is illustrated in Fig.3.5 as an illustration.

In the following, the computation of a simulated EO sensor is explicated. It has the following steps:

- the signal emitted from the object arriving at the sensor
- the sensor and detector response

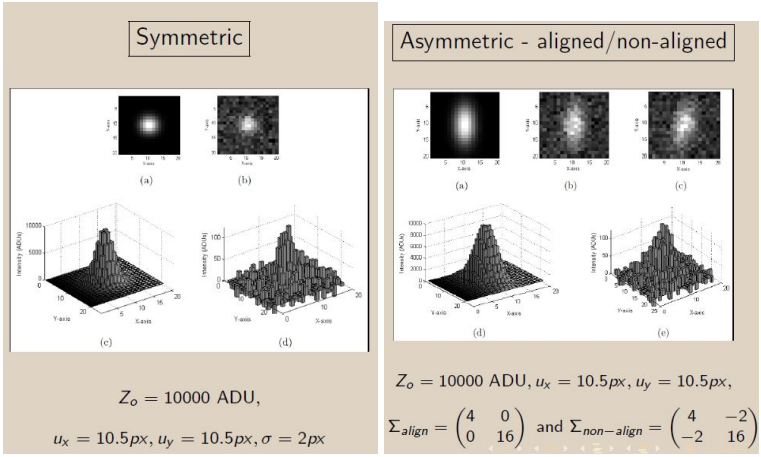


Figure 3.5: Simulated object images on a pixel grid with and without noise sources included; listed are the values for the peak intensity and center location relative to the pixel grid (C. Frueh, R. Manish).

3.2.1 Signal Emitted from the Object Arriving at the Sensor

The amount of light that is arriving at the sensor for electro-optical passive observations depends upon the following quantities:

- the illumination source $\rightarrow I_S$
- the reflection geometry $\rightarrow \Psi$
- the shape and reflection properties of the illuminated object $\rightarrow \Psi$
- potential attenuation sources in the light path and the light travel $\rightarrow \tau$

Reflection geometry:

Traditionally, the phase angle α has been used to determine observation geometry.

It is defined as the angle $\alpha = \sphericalangle$: Sun,Object,Observer.

Check: What is the zero phase angle? What is the phase angle at full moon/half moon?

Further question: Is the phase angle sufficient to determined how much light is reflected towards the observer?

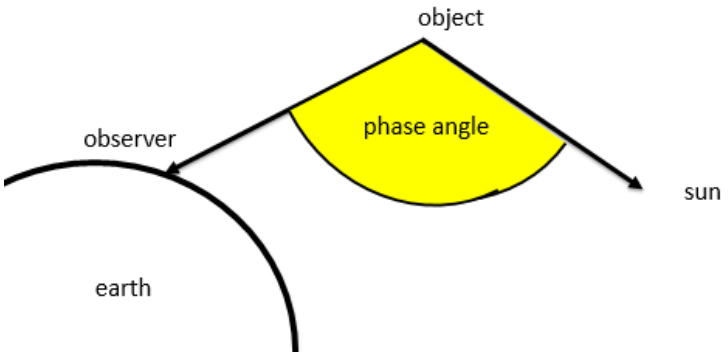


Figure 3.6: Definition of the phase angle

For objects that are not spherical the phase angle is necessary but not sufficient to describe the reflection geometry.

The reflection is governed by the phase angle and the object specific phase function.

The phase function, Ψ describes how light coming from a specific direction is reflected off a specific object.

This allows to compute the irradiation from an object received at the location of the sensor, denoted by I_{obj} to be computed as the following in the simplest convex case:

$$I_{\text{obj}} = I_{\text{S}} \cdot \tau \cdot \Psi, \tag{3.1}$$

I_{S} is the received irradiation at the location of the object,
τ is the travel function over the distance from the object to the observer,
Ψ is the bidirectional reflection function, also called phase function.
For multi-faceted objects with potential concavities, Eq.3.1 needs to be extended to a sum over all n surface parts with areas A_i in the function with Ψ_i:

$$I_{\text{obj}} = \sum_{i=1}^{n} (I_{\text{S}} + I_{\text{2nd},i}) \cdot \tau_i \cdot \Psi_i \cdot \sigma_{\text{ss},i} \sigma_{\text{os},i} \tag{3.2}$$

where $\sigma_{\text{ss},i} \in \{0,1\}$ is the sun-shadowing or self-shadowing term,
determining if a particular facet A_i is blocked by another and does not receive sunlight $\sigma_{\text{ss},i} = 0$
and the observer-shadowing term, $\sigma_{\text{os},i} \in \{0,1\}$,
determining if the facet associated to Phi_i is blocked from view to the observer by another facet in the line of sight.
$I_{\text{2nd},i}$ is secondary reflection that might be received from surrounding facets. In a first approximation, secondary reflection is often neglected.
For satellites with extensions that are small relative to the distance to the observer, the travel function may be defined as facet independent and relative to the center of mass to of the object $\tau_i \approx \tau_{\text{CM}}$.

Position dependence of the illumination source beyond the center of mass can be neglected in case of a solar illumination of the object by the sun other than in the near sun region.

3.3 The Illumination Source: Sun

The light, which is emitted from the Sun, can be quantified in numerous ways.

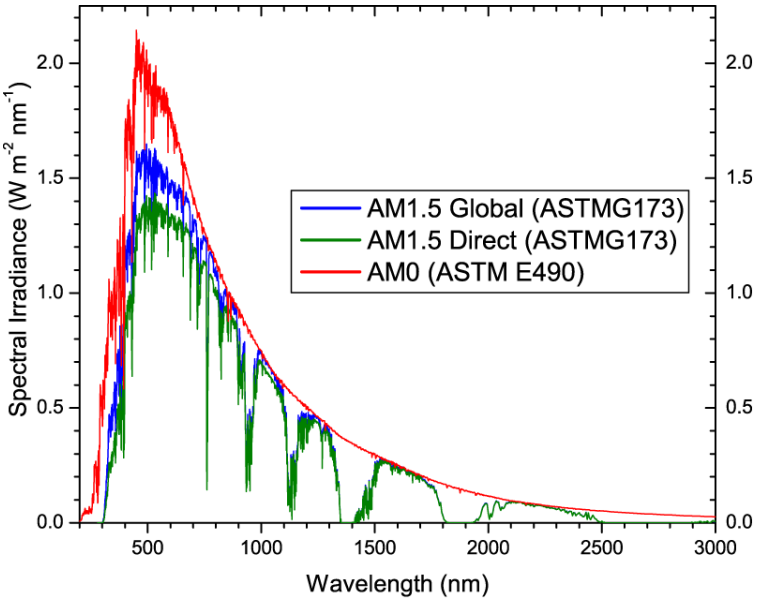


Figure 3.7: Solar spectral irradiance generated by SMARTS [34].

One is the overall luminosity, or also called flux, L_s, measured in Watt or Lumen (W). It is the overall power that a light source emits.

If the flux divided by the area the flux passes through, the so-called flux density is defined. Flux densities are measured in Watt/square meter ($\frac{W}{m^2}$).

The flux density at the distance of one AU is the so-called solar constant I_0.

The radiance R_s, measured in Watt per square meter per steradian ($\frac{W}{m^2 sr}$); it is the flux density in a specific angular segment.

When the radiant flux is incident on the area of a surface, the term irradiation is used for the radiant flux density.

If the radiant flux is emitted from the area of the surface, the radiant flux density is called exitance; however the terminology is not strict.
The aforementioned quantities assume that the solar radiation is already integrated over a specific wavelength band. A radiant flux density split into the different wavelengths, sometimes called spectral irradiance (when incident), or spectral radiant flux density $I_{0,\lambda}$ $(W/(m^2 nm))$ [32]:

$$\int I_{0,\lambda}(\lambda) d\lambda = I_0 \tag{3.3}$$

The same applies anaologous to the flux directly.

Fig.3.7 shows the solar spectrum.

AMO0 cossresponds to the spectral irradiance outside the Earth atmosphere.

For terrestrial use, AM1.5 Global and AM1.5 Direct are in use.

AM1.5 Global for flat plate applications (solar cell simulations), AM1.5 direct includes the sun's circumsolar compo-

nent.
In terms of luminosity L_s, the sun power changes with the 11 year cycle, but such changes are negligible in terms of direct radiation that is received at the Earth distance and near Earth region.

A nominal value for the mean luminosity is $\mathbf{L_s}(\text{mean}) = \mathbf{3.828 \cdot 10^{26} W}$ **[54]**.

The sun can be approximated as emitting uniformly in all directions and at a constant rate.

The power per area (oriented perpendicular to the direction of the incoming radiation), or in other words the flux density I_s, at a given distance r_{sunobj} can be computed as:

$$I_s = \frac{L_s}{4\pi r_{\text{sunobj}}^2} \tag{3.4}$$

If the average distance of the Sun to the Earth is assumed, one Astronomical Unit ($AU = 149597870.7km$), one reaches the definition of the solar constant $I_{0\perp}$:

$$I_0 = \frac{L_s}{4\pi AU^2} \tag{3.5}$$

The solar constant is not actually a constant.

The biggest effect is that the Earth is on an eccentric orbit around the sun. The mean solar constant is reported as $\mathbf{I_0} = \mathbf{1361.0} \frac{\mathbf{W}}{\mathbf{m^2}}$ **[54]**, at the exact 1 AU distance.

One can already note, that the conversion, although physically correct, leads to a slightly different value than the one reported by IAU, because of rounding.

In order to foce one solar constant to the nominal value, the mean luminosity is approximated as $L_s \approx 3.82753185 \cdot 10^{26} W$. The following equation can be used to scale the solar constant directly to a given location r_{sunobj}:

$$I_s = I_0 \frac{AU^2}{r_{\text{sunobj}}^2} \tag{3.6}$$

Another way of computing the solar constant or the flux density at any given distance is via the radiance R_s. Using the mean radius of the Sun $\mathbf{r_s} = \mathbf{6.957 \cdot 10^5 km}$ [54], the half angle of size of the disk the sun has in the sky at one AU is:

$$\rho_{s,AU} = \tan^{-1}\left(\frac{r_s}{AU}\right) \tag{3.7}$$

The value is $\rho_{\mathbf{s,AU}}[\mathbf{deg}] = \mathbf{0.266}\,\text{deg}$, which is in good agreement with the measurements displayed in Fig.3.8.

Fig.3.8 does show the normalized irradiation in various wavelengths indicated by the colors of plotted graphs. One can see the sharp decline in irradiation at 0.266 degrees (at one AU distance) and the second cutoff value at 2.5 degrees that is often used.

The solar constant I_0 can be computed via the volumetric angle of the visible disk (Note that $\rho_{s,AU}$ needs to be expressed in radians!!):

$$I_0 = R_s(2\rho_{s,AU}[rad])^2 \tag{3.8}$$

The nominal value of the radiance is reported as $R_s = 2.009 \cdot 10^7 \frac{W}{m^2 sr}$.

The variation of the solar constant at perihel $1412 \frac{W}{m^2}$ of and at aphel of $1320 \frac{W}{m^2}$.

In order to scale the solar constant to other distances between the sun and the object of interest r_{sunobj}, the following

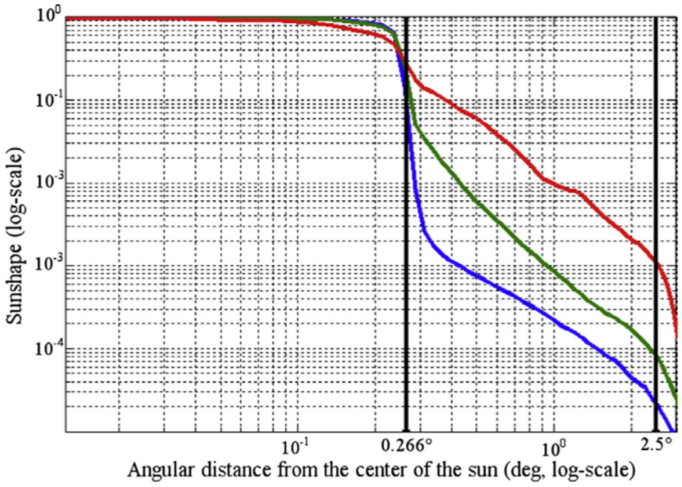


Figure 3.8: Normalized Sun irradiation in different wavelengths [10].

factor can be multipled to the nominal solar constant at one AU:

$$I_s = I_0 \cdot \frac{\rho^2_{s,\text{sunobj}}}{\rho^2_{s,AU}} = I_0 \cdot \frac{\left(\tan^{-1}\left(\frac{2R_s}{r_{\text{sunobj}}}\right)\right)^2}{\left(\tan^{-1}\left(\frac{2R_s}{AU}\right)\right)^2} \tag{3.9}$$

As $\rho_{s,AU}$ can be precomputed and $\rho_{s,r_{\text{sunobj}}}$ is already computed for the specular reflection function, Eq.3.9 is preferred over Eq.3.5 for practical reasons.

The actual irradiation $I_{s,p}$ at the space object that a flat surface reaches that is not perpendicular to the sun is then goverened by the cosine law:

$$I_{\text{s,plate}} = I_s \cos\theta_s, \tag{3.10}$$

where θ_s is the angle between the facet normal direction and the direction to the center of the Sun, see also Fig. 3.15.

3.4 Magnitudes

In astronomy the so-called *magnitude* is used. Magnitudes have been invented as a logarithmic scale, such that a brightness ration of 100 corresponds to a magnitude difference of 5:

$$\frac{I_1}{I_2} = 100^{\frac{m_2-m_1}{5}} \tag{3.11}$$

$$m_2 - m_1 \approxeq -2.5\log_{10}(\frac{I_2}{I_1}) \tag{3.12}$$

This defines so-called relative magnitudes.

A bit of nomenclature: **Apparent magnitudes** are the magnitudes as they appear from Earth, the **absolute magnitudes** is scaled to the standard distance of 10 parsecs (they are often denoted by a capital M). The apparent bolometric magnitude of the Sun is $mag_{\text{Sun}} = -26.832$ [54], the absolute magnitude of the sun is $M_{\text{Sun,absolute}} = 4.74$ [54].

The zero point of absolute bolometric magnitude ($M_{\text{absolute}} = 0$) is defined to correspond to a liminosity of $L_{\text{M=0}} = 3.0128 \cdot 10^2 8W$ [54]. The absolute magnitude M of a source with luminosity L can hence be computed as:

$$M = -2.5\log_{10}(\frac{L}{L_{\text{M=0}}}) \approx -2/5\log_{10}(L) + 71.197425 \tag{3.13}$$

magnitude m	0	1	2	3	4	5	6	7	8	9	10
relative brightness ratios	1	2.5	6.3	16	40	100	250	630	1600	4000	10,000

Figure 3.9: Magnitudes and related irradiation ratios.

Object	m_V
Sun	-26.8
Full Moon	-12.5
Venus at brightest	-4.4
Jupiter at brightest	-2.7
Sirius	-1.47
Vega	0.04
Betelgeuse	0.41
Polaris	1.99
Naked eye limit	6
Pluto	15.1

Figure 3.10: Common apparent magnitudes.

It is chosen to fix the nominal value of the Sun's magnitude at its nominal luminosity of $L_s(\text{mean}) = 3.828 \cdot 10^{26} W$.

Apparent bolometric magnitude zero corresponds to a flux density $I_{\text{mag}=0} = 2.518021002 \cdot 10^{-8} \frac{W}{m^2}$. Hence the apparent magnitude for a space object of irradiance I can be computed one of two ways, either via linking it to the flux density corresponding to $mag = 0$, or via relation to the Sun's magnitude and the Sun's nominal flux density (solar constant) with the mean nominal value $I_0 = 1361.0 \frac{W}{m^2}$:

$$mag = -2.5\log_{10}(\frac{I}{I_{\text{mag}=0}}) \approx -2.5\log_{10}(I) - 18.997351 \tag{3.14}$$

$$= mag_{\text{Sun}} - 2.5\log_{10}(\frac{I}{I_0}) \tag{3.15}$$

$$\tag{3.16}$$

The latter formulation proves advantageous as the irradiance of the object is diretly proportional to the irradiance of the illumination source, the Sun. In Tab.3.11 is a summary of the parameters given by the IAU as nominal values that should be used in all conversion calculations.

SOLAR CONVERSION CONSTANTS		
$1\mathcal{R}_{\odot}^{\text{N}}$	=	6.957×10^8 m
$1\mathcal{S}_{\odot}^{\text{N}}$	=	1361 W m^{-2}
$1\mathcal{L}_{\odot}^{\text{N}}$	=	3.828×10^{26} W
$1\mathcal{T}_{\text{eff}\odot}^{\text{N}}$	=	5772 K
$1(\mathcal{GM})_{\odot}^{\text{N}}$	=	$1.327\,124\,4 \times 10^{20}$ m^{3}s^{-2}

Figure 3.11: Solar conversion constants: solar radius $\mathfrak{R}_{\odot}^{N}$ (r_s), total solar irradiance $\mathfrak{S}_{\odot}^{N}$ (I_0), solar luminosity $\mathfrak{L}_{\odot}^{N}$ (L_S), solar effective temperature $\mathfrak{T}_{\text{eff}\odot}^{N}$, and solar mass parameter $\mathfrak{GM}_{\odot}^{N}$: The nominal values may be used respectively, which are by definition exact and are expressed in SI unit [54].

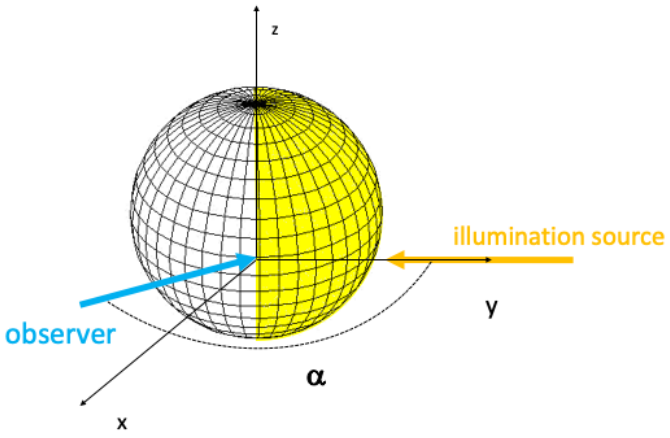


Figure 3.12: Spherical object with illumination along the y-axis and the observer located at angle α from the illumination source.

3.5 Phase Function or Bidirectional Reflection Function (BRDF)

The bidirectional reflection function (BRDF) or also called phase function, here denoted by the letter Ψ.

Ψ governs what fraction of an input radiation from a given direction is reflected towards an observer of a given direction.

Usually, this is refered to as BRDF in the realm of computer graphics and as reflection function in a physics and engineering context, and as phase function in the context of astronomy.

In the following, first the point source model is discussed. Then the extended source, relevant to applications evolving the Sun, is developed. The surface properties of the objects is modeled via a mixture of specular and Lambertian reflection in combination with absorption.

3.5.1 Reflection function: Point Light Source

In the following the reflection from an infinetely far point source are discussed.

For the infinetly far point source, two different model approaches do exist, that in the most cases coincide, however, sometimes lead to subtle differences.

The one is the investigation of a single ray, the other is the use of parallel rays. Both are valid representations of the points source.

The differences in the use of the single versus the parallel ray representation are pointed out in the discription explicitly below to the extent as such they result in a different model interpretation.

3.5.1.1 Sphere

For a spherical object, the computation in spherical coordiantes is most convenient.

The incoming flux density is in the μ_0 direction, μ is the direction of the observer. f_r is the bidirectional reflectance distribution function, or in other words the model of the BRDF.

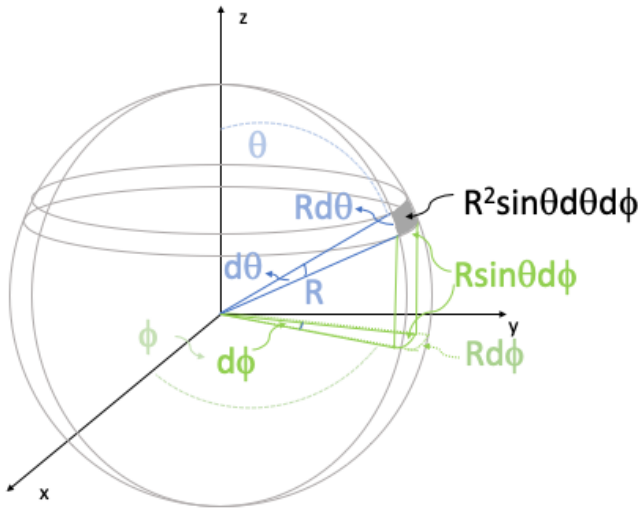


Figure 3.13: Surface element of a sphere.

The exitance, that is the reflected light, can be determined as the following, compare Fig.3.12:

$$\Psi(\bar{\lambda},\alpha)_{\text{sphere}} = \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{0}^{\pi} f_r \mu_0 \mu R^2 \sin\theta d\phi d\theta \tag{3.17}$$

where R as the radius of the sphere, defining the surface element of the sphere as $R^2 \sin\theta d\theta d\phi$, see Fig.3.13.

Using spherical coordinates to define the direction to the light source as $\mu_0 = \sin\theta\cos\phi$.

For a sphere the observer can be placed in the xy-plane without restriction of generality.

This allows defining the direction of the observer as $\mu = \sin\theta\cos(\alpha-\phi)$.

This allows explicating the integral:

$$\Psi(\bar{\lambda},\alpha)_{\text{sphere}} = \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{0}^{\pi} f_r R^2 \sin^3\theta \cos(\alpha-\phi)\cos\phi\, d\phi d\theta \tag{3.18}$$

3.5.1.1.1 Lambertian, Diffuse Reflection For a sphere and the Lambertian reflection, the BRDF function is very simple, $f_{r,\text{lamb}} = \frac{C_d}{\pi}$, as it does not carry the function dependency on the incoming and outgoing directions explicitly:

$$\Psi(\bar{\lambda},\alpha)_{\text{sphere,lamb}} = \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{0}^{\pi} \frac{C_d}{\pi} R^2 \sin^3\theta \cos(\alpha-\phi)\cos\phi\, d\phi d\theta \tag{3.19}$$

$$= \frac{C_d}{\pi} R^2 \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{0}^{\pi} \sin^3\theta \cos(\alpha-\phi)\cos\phi\, d\phi d\theta \tag{3.20}$$

$$= \frac{C_d}{\pi} R^2 \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{4}{3} \cos(\alpha-\phi)\cos\phi\, d\phi \tag{3.21}$$

$$= \frac{C_d}{\pi} R^2 \frac{4}{3}\frac{1}{2}(\sin\alpha + (\pi-\alpha)\cos\alpha) \tag{3.22}$$

$$= \frac{2}{3}\frac{C_d}{\pi} R^2 (\sin\alpha + (\pi-\alpha)\cos\alpha) \tag{3.23}$$

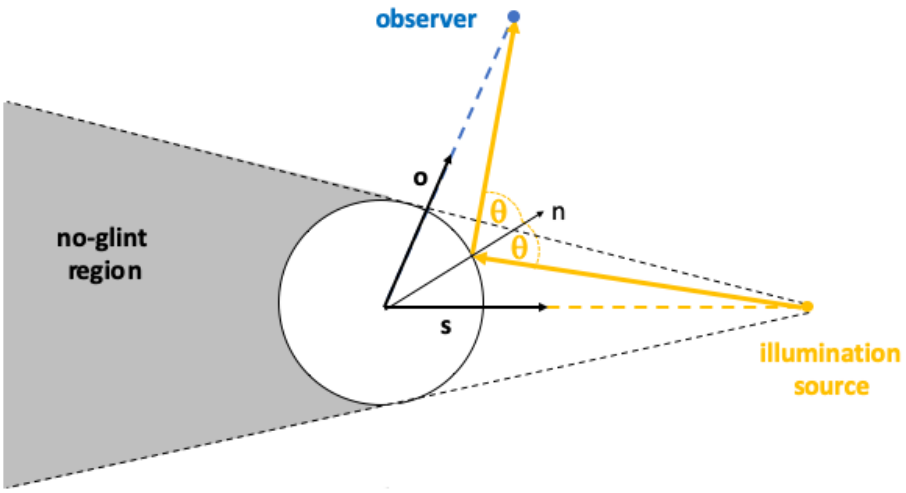


Figure 3.14: Specular reflection on a sphere.

It can easily be seen that the amount of the reflected light depends on the angle α determines the difference between the incoming light flux and the observer direction. Because of the rotation symmetric nature of the problem, the direction of the observer and the direction to the illumination source form a plane at the position of the object. Hence, the angle α is the so-called phase angle.

3.5.1.1.2 Specular Reflection For the specular reflection on a sphere, usually iterative procedures are needed to determine the specific location of a glint.

For our purpose as only non-resolved images are acquired, the specific location of the glint is not of interest but only, if a glint is received in the direction to the observer location.

In contrast to a flat plate, numerous directions lead to a glint in the observer location.

As long as the observer and the illumination source are on the same side, a glint is received, as always a normal direction on the sphere can be found that fulfills the glint condition, as shown in Fig.3.14.

The condition for the avoidance of the no-glint region is simply to be on the same side as the sun (within 180 deg plane).

3.5.1.2 Flat Surface

For a flat object, the computation can be performed analogous.

Because no rotation symmetry is present any more, the compuation is performed in Carthesian space.

Without loss of generality, the area can be placed in the $x-y$-plane.

Unlike the spherical case, the plate is fully illuminated as long as the illuminating source has a positive z-cmponent, or

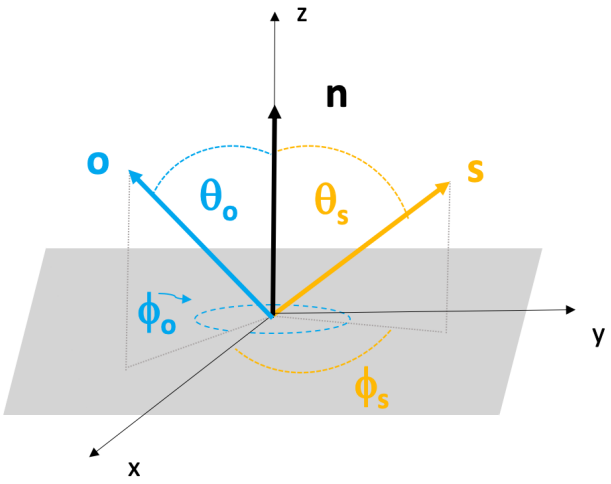


Figure 3.15: Observer unit direction $\boldsymbol{o}$, sun unit direction $\boldsymbol{s}$ and unit normal direction $\boldsymbol{n}$ on the flat plate.

in other words:

$$\text{For}\,\Psi(\bar{\lambda},\alpha)_{\text{plate}} \quad = \quad \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}} \int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} f_r \mu_{0_\text{p}} \mu_\text{p}\, dxdy \tag{3.24}$$

defining

L_1 and L_2 as the length and width of the flat plate surface.

For the flat plate, there are three important directions:

the direction to the light source, defined as the unit vector $\boldsymbol{s}$

the direction to the observer, denoted by the unit vector $\boldsymbol{o}$

and the normal direction of the flat surface, defined as $\boldsymbol{n}$.

Without loss of generality, the flat plate can be placed such that the normal vector and the z-axis coincide.

The light source direction $\boldsymbol{s}$ and the direction to the observer $\boldsymbol{o}$, relative to the normal direction $\boldsymbol{n}$, are illustrated in Fig.3.15 and can be expressed as:

$$\boldsymbol{o} := \boldsymbol{o}(\phi_o, \theta_o) \tag{3.25}$$
$$\boldsymbol{s} := \boldsymbol{s}(\phi_s, \theta_s) \tag{3.26}$$

It has to be noted, that the convention of the zenith angle is used for the spherical coordinates, compared to the more frequently used elevation angle definition; this has the advantage that the second angle is the enclosed angle to the normal direction:

$$\cos\theta_s = \boldsymbol{n}\cdot\boldsymbol{s} \tag{3.27}$$
$$\cos\theta_o = \boldsymbol{n}\cdot\boldsymbol{o} \tag{3.28}$$

For the flat plate, point source reflection is independent of the direction of the source (although that seems counter-intuitive at first), $\boldsymbol{s}$ as long as the point source is above the surface. This means:

$$\mu_{0_\text{p}} = \begin{cases} 1 & \text{for } \theta_s < \frac{1}{2}\pi \\ 0 & \text{for } \theta_s \geq \frac{1}{2}\pi \end{cases} \tag{3.29}$$

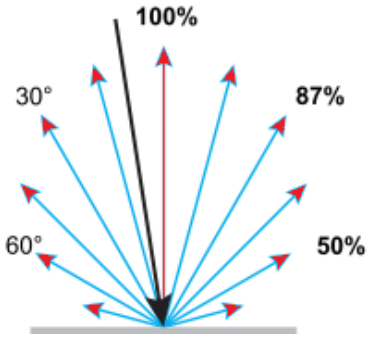


Figure 3.16: Lambertian Reflection off a plate illustrating the cosine-viewing law. The incident light is indicated via the black arrow. The reflection is independent of the specific direction of the illuminating source ($\frac{C_d}{\pi}$), the light is scattered equally in all directions. The percentrages show the perceived reflected irradiation based upon the viewing angle and hence the fraction of the projected area to the observer $\cos\theta_o$ (Light Measurement Handbook © 1998 by Alex Ryer, International Light Inc.).

This seems counter intuitive at first, however, as the surface is flat, it is, in contrast to the sphere either fully illuminated or not. Differences that do occur, are in fact only observer direction dependent. Observer direction dependent factors differ, based on the reflection model.

3.5.1.2.1 Lambertian, Diffuse Reflection Lambertian reflection is defined as the reflection that is equally distributed in all viewing directions, leading $f_{r,\text{lamb}} = \frac{C_d}{\pi}$ as before. Same as for the spherical surface, it hence depends only upon the illuminated area that is projected towards the observer. This leads to the famous Lambertian cosine law (Light Measurement Handbook © 1998 by Alex Ryer, International Light Inc., and numerous other sources):

$$\mu_\text{p} = \mu_{\text{p,l}} = \cos\theta_o \tag{3.30}$$

Fig.3.16 illustrates the Lambertian reflection principle of the cosine law, independent of the incoming light direction as long as Eg.3.29 is fulfilled. As a result, the integration Eq.3.24 is trivial to solve:

$$\begin{aligned}
\Psi(\bar{\lambda},\alpha)_{\text{plate,lamb}} &= \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}}\int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} \frac{C_d}{\pi}\mu_{0_\text{p}}\cos\theta_\text{o}\,dxdy && (3.31)\\
&= \frac{C_d}{\pi}L_1L_2\mu_{0_\text{p}}\cos\theta_\text{o} && (3.32)\\
&= \frac{C_d}{\pi}A\mu_{0_\text{p}}\cos\theta_\text{o} && (3.33)
\end{aligned}$$

defining $A = L_1L_2$ as the area of the plate with μ_{0_p} defined in Eq.3.29.

3.5.1.2.2 Specular Reflection If we are defining the specular reflection, the reflection function f_r is simply defined as $f_{r,\text{spec}} = 1\cdot C_s$. However the formulation of the direction function $\mu_\text{p} = \mu_{\text{p,s}}$ is highly restrictive, as only a glint is produced when the observer $\boldsymbol{o}$ is exactly opposite of the surface plate normal vector compared to the source direction $\boldsymbol{s}$. The reflection BRDF for specular reflection direction function $\mu_\text{p} = \mu_{\text{p,s}}$ can be defined via the Kronecker delta $\delta(\cdot)$:

$$\mu_\text{p} = \mu_{\text{p,s}} = \delta(\theta_\text{s}-\theta_\text{o})\,\delta(\phi_\text{s}+\pi-\phi_\text{o}) \tag{3.34}$$

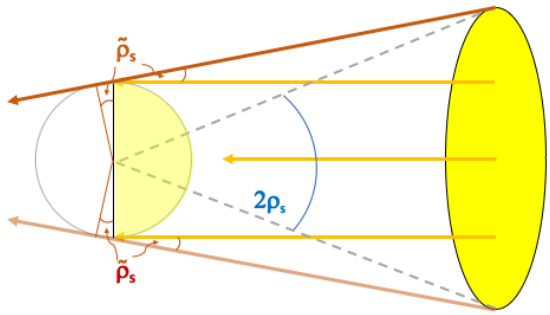


Figure 3.17: Projection of the Lambertian reflection off a spherical object from an extended source.

The Kronecker delta is defined as $\delta(i-j) = 1$ for $i = j$ and zero otherwise. Thus:

$$\Psi(\bar{\lambda},\alpha)_{\text{plate,spec}} = \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}} \int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} C_s \mu_{0_\text{p}} \mu_{\text{p,s}} dxdy \tag{3.35}$$

$$= \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}} \int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} C_s \mu_{0_\text{p}} \delta(\theta_\text{s} - \theta_\text{o})\, \delta(\phi_\text{s} + \pi - \phi_\text{o}) dxdy \tag{3.36}$$

$$= C_s L_1 L_2 \mu_{0_\text{p}} \delta(\theta_\text{s} - \theta_\text{o})\, \delta(\phi_\text{s} + \pi - \phi_\text{o}) \tag{3.37}$$

$$= C_s A \mu_{0_\text{p}} \delta(\theta_\text{s} - \theta_\text{o})\, \delta(\phi_\text{s} + \pi - \phi_\text{o}) \tag{3.38}$$

3.5.2 Reflection function: Extended Light Source

The sun is better approximated with an extended source than an actual point source.

The sun disk, actually does not have sharp edges, but fades out, as limb darkening does occur. The limb darkening is wavelength dependent and various models exist [10], as discussed in the previous section 3.3.

In the following it is assumed that the overall irradiance of the extended source, I_0 is the same for the extended source compared to the irradiance of the point source.

3.5.2.1 Sphere - Extended Source

3.5.2.1.1 Lambertian, Diffuse Reflection Extended Source For the Lambertian sphere, when illuminated by an extended source is very simular to the situation of the point source.

The only difference that depending on the angular extension of the source at the location of the object, $2\rho_s$, more than exactly half the sphere by that exact angle are illuminted, see Fig.3.17.

This is irrelevant for a phase angle α of zero, but gives slightly more illumination in all other phase angles.

As the sun disk, radius R_{sun} is much larger than the radius of the spherical space object of radius R, the two angles ρ denoting the extension of the Sun disk at the object location in the near Earth region and the angle $\tilde{\rho_s}$ are nearly identical.

Their are only differing by the alternation of the angle from the center of the sun disk to is rim, R_{sun} to $R_{\text{sun}} - R$. The expression for the reflection function for the sphere, hence only needs to be extended by the excess area that is illuminated:

$$\text{with} \quad R_{\text{sun}} >> R \quad \rightarrow \quad \tilde{\rho_s} \approx \rho_s \tag{3.39}$$

$$\Psi(\bar{\lambda},\alpha)_{\text{sphere,lamb,e}} = \int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\rho_s}^{\pi+\rho_s}\frac{C_d}{\pi}R^2\sin^3\theta\cos(\alpha-\phi)\cos\phi\,d\phi d\theta \tag{3.40}$$

$$= \frac{C_d}{\pi}R^2\int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-\rho_s}^{\pi+\rho_s}\sin^3\theta\cos(\alpha-\phi)\cos\phi\,d\phi d\theta \tag{3.41}$$

$$= \frac{C_d}{\pi}R^2\int_{\alpha-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos(\alpha-\phi)\cos\phi\frac{1}{6}(9\cos\rho_s-\cos 3\rho_s)\,d\phi \tag{3.42}$$

$$= \frac{C_d}{\pi}R^2\cdot\frac{1}{6}(9\cos\rho_s-\cos 3\rho_s))\cdot\frac{1}{2}(\sin\alpha+(\pi-\alpha)\cos\alpha) \tag{3.43}$$

$$= \frac{1}{6}\frac{C_d}{\pi}R^2\cos\rho_s(5-\cos\rho_s)(\sin\alpha+(\pi-\alpha)\cos\alpha) \tag{3.44}$$

One can easily see that for ρ_s=0, the expression for the point source, Eq.3.23 is obtained.

As the extension of the sun is only half a degree in the near Earth region, the differences to the point source expression are negligible for a sphere.

3.5.2.2 Flat Surface Extended Source

With an extended source and a flat surface, one distinguishing criterion compared to the point source is, that the object can be fully above the surface plane or fully below it, but also partly above the surface plane.

Assuming a spherical extended source, the net irradiation that is received is assumed to be the same as for the point source when the object disk is completely above the image plane.

Fig. 3.18 does illustrate the case, where the disk center is still above the image plane, but a part of the disk is already below it. The fraction of the disk that is contributing to the illumination of the surface, and hence scalling the received overall irradiation I_0 can be directly incorporated in the source direction paramter $\mu_{0,p,e}$:

$$\mu_{0_{\text{p,e}}} = \begin{cases} 1 & \text{for } \theta_s+\rho_s < \frac{1}{2}\pi \\ \frac{(\pi\rho_s^2 - A_{\text{segment}})^2}{\rho_s^2} & \text{for } \theta_s < \frac{1}{2}\pi \wedge \theta_s+\rho_s > \frac{1}{2}\pi \\ \frac{A_{\text{segment}}^2}{\rho_s^2} & \text{for } \theta_s > \frac{1}{2}\pi \wedge \theta_s-\rho_s < \frac{1}{2}\pi \\ 0 & \text{for } \theta_s-\rho_s \geq \frac{1}{2}\pi \end{cases} \tag{3.45}$$

ρ_s is the radius of the extended radiation source.
$\varepsilon_s = \frac{\pi}{2} - \theta_s$ is the distance between the surface and the center of the disk (or 90 degrees minus the center of the disk angle).

The segment of the disk below the plane in Fig.3.18 is computed as the difference between the sector of the circle minus the isosceles triangle formed by the sides ρ_s and the base β_s:

$$A_{\text{segment}} = \frac{2\delta_s}{2\pi}\pi\rho_s^2 - \frac{1}{2}\beta_s\varepsilon_s$$

$$= \cos^{-1}\left(\frac{\varepsilon_s}{\rho_s}\right)\rho_s^2 - \varepsilon_s\sqrt{\rho_s^2-\varepsilon_s^2} \tag{3.46}$$

with $\delta_s = \cos^{-1}(\frac{\varepsilon_s}{\rho_s})$.

As the sun radiance is asusmed to be completely uniform over the disc, but now, only part of the disk is visible, necessitating the scaling of the area used in Eq.3.9.

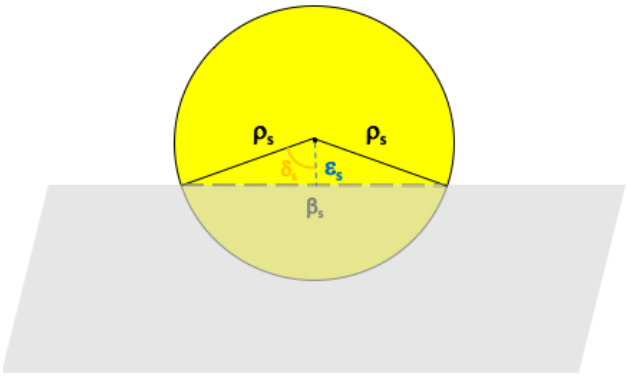


Figure 3.18: Illustration of the case $\theta_s < \frac{1}{2}\pi \wedge \theta_s + \rho_s > \frac{1}{2}\pi$, a fraction of the extended source is above the plane of the flat object.

3.5.2.2.1 Lambertian, Diffuse Reflection Extended Source The expression for the flat plate in the model of the Lambertian reflection phase function for a point source, $\Psi(\bar{\lambda},\alpha)_{\text{plate,lamb}}$, is defined in Eq.3.33.

It can be seen that this expression is independent of the source direction. Thus, only the modified $\mu_{0,p,e}$ representing the fraction of the extended source contributing to the illumination of the flat plate needs to be taken into account:

$$\Psi(\bar{\lambda},\alpha)_{\text{plate,lamb,ext.}} = \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}} \int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} \frac{C_d}{\pi} \mu_{0_{\text{p,e}}} \cos\theta_{\text{o}}\, dxdy \tag{3.47}$$

$$= \frac{C_d}{\pi} L_1 L_2 \mu_{0_{\text{p,e}}} \cos\theta_{\text{o}} \tag{3.48}$$

$$= \frac{C_d}{\pi} A \mu_{0_{\text{p,e}}} \cos\theta_{\text{o}} \tag{3.49}$$

3.5.2.2.2 Specular Reflection Extended Source For the specular reflection of an extended source, the situation is different, as the result is highly dependent upon the direction to the illumination source. As the source is extended, the region in which the specular glint is received is extended.

The fraction of the reflection that is received on the ground (neglecting atmosphere) is illustrated in Fig.3.19, leading to a $\mu_{p,s,e}$ of the following form:

$$\mu_{p,s,e} = \frac{1}{A_{\text{reflection,s}}} \cos\delta_{\text{obs}} \cdot p \tag{3.50}$$

with

$$p = \begin{cases} 1 & \text{for } \sigma < \arctan(\frac{d_{\text{reflection,s}}}{r_{\text{obj\_obs}}} + \tan\rho_s) \\ 0 & \text{for } else \end{cases} \tag{3.51}$$

and

$$\cos\sigma = \boldsymbol{o} \cdot \boldsymbol{o}_{\text{spec\_perfect}} \tag{3.52}$$

With the same argument as before, the two angles ρ denoting the extension of the Sun disk at the object location in the near Earth region and the angle $\tilde{\rho}_s$ are nearly identical in Fig.3.19.

$$R_{\text{sun}} >> R \quad \rightarrow \quad \tilde{\rho}_s \approx \rho_s \tag{3.53}$$

The reflection directional parameter $\mu_{p,s,e}$, consists of a leading part, which scales the incoming intensity via the area the reflection occupies at the location of the observer $A_{\text{reflection,s}}$, see Fig.3.19.

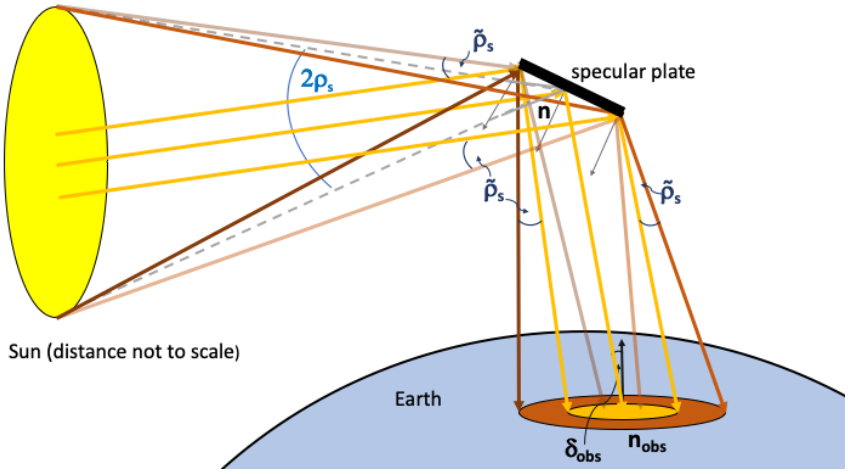


Figure 3.19: Illustration from the specular reflection of the Sun.

It is the original area of the reflecting surface expanded by the effect of the extension of the sun. For a round or rectangular shape, $A_{\text{reflection,s}}$ is computed as the following:

$$\text{round reflection surface area A} \qquad A_{\text{reflection,s}} = \pi(\tan(\rho_s) r_{\text{obj\_obs}} + \sqrt{\frac{A}{\pi}})^2 \tag{3.54}$$

$$\text{square reflection surface area A} \qquad A_{\text{reflection,s}} = (\tan(\rho_s) r_{\text{obj\_obs}} + \sqrt{A})^2 \tag{3.55}$$

with $\tilde{\rho_s} \approx \rho_s$, where $r_{\text{obj\_obs}}$ is the distance between the reflecting surface on the space object and the observer.

It has to be noted, that in literature, as ρ_s is small, of the tangens is replaced by the angle direction $\tan(\rho_s) \approx \rho_s$. The factor $\cos\delta_{\text{obs}}$ in Eq.3.50 is the angle between the observation plane and the opposite direction incoming ray that is reflected in the perfect specular point-source direction $-\boldsymbol{o}_{\text{spec\_perfect}}$.

It should also be noted that a limit is reached at which the size increase of the reflecting facet does NO increase the reflection any more; this is the case, when the whole sun disk is reflected in a single area. The size of the reflecting facet would need to be the size of the apparent sun disk at the distance of the observer.

For a telescope pointing towards the object, the angle is usually zero, leading $\cos\delta_{\text{obs}} = 1$.

The factor p in in Eq.3.50 determines if the observer is located within $A_{\text{reflection,s}}$ and receives a specular glint or not.

The condition therefore is the distance of the direction of the actual observer at direction $\boldsymbol{o}$ compared to the perfect reflection direction of the point source $\boldsymbol{o}_{\text{spec\_perfect}}$, denoted by σ.

The limit is determined as the extension of the reflecting surface in the plane spanned by $\boldsymbol{o}$ and $\boldsymbol{n}$, denoted by $d_{\text{reflection,s}}$.

For a round facet of area A, this is simply $d_{\text{reflection,s}} = \sqrt{\frac{A}{\pi}}$, the radius.

For other facets, not well approximated as round, this can be computed as the extension of the facet in the direction $\boldsymbol{o}$ projected on the 90 degrees rotated normal vector $\boldsymbol{n}$. The perfect reflection direction $\boldsymbol{o}_{\text{spec\_perfect}}$, corresponding to the specular reflection direction of a point source is easily computed:

$$\boldsymbol{o}_{\text{spec\_perfect}} = 2\cos\theta_s \boldsymbol{n} - \boldsymbol{s} \tag{3.56}$$

This leads to the reflection function for the specular reflection on a flat plate from an extended source:

$$\Psi(\bar{\lambda},\alpha)_{\text{plate,spec,ext.}} = \int_{-\frac{L_1}{2}}^{\frac{L_1}{2}}\int_{-\frac{L_2}{2}}^{\frac{L_2}{2}} f_{\text{r,spec}}\mu_{0_{\text{p,e}}}\mu_{p,s,e}\,dxdy \tag{3.57}$$

$$= L_1L_2C_s\mu_{0_{\text{p,e}}}\frac{1}{A_{\text{reflection,s}}}\cos(\delta_{\text{obs}})\cdot p \tag{3.58}$$

$$= C_s\mu_{0_{\text{p,e}}}\frac{A}{A_{\text{reflection,s}}}\cos(\delta_{\text{obs}})\cdot p \tag{3.59}$$

for telescope observing the object $\cos(\delta_{\text{obs}}) = 1$

$$= C_s\mu_{0_{\text{p,e}}}\frac{A}{A_{\text{reflection,s}}}p \tag{3.60}$$

$$\tag{3.61}$$

Side note:Lambertian BRDF

Some confusion is normally occurring in he Lambertian reflection because of the factor π, which is not following from the definition of the BRDF alone, but only from the definition of the BRDF in combination with energy conservation. This latter part can easily be shown (courtesy to Rory Driscoll who also shows this nice and easy proof on his homepage [25], and replaced my long winded one). Assuming all light is reflected diffusely and $C_d = 1$, then the exitant radiation flux density I_{ex} has to be equal to the incident radiation flux density I_{in}:

$$I_{ex} = I_{in} \tag{3.62}$$

For the Lambertian reflection, the exitant radiation flux density follows the cosine law, spreading the incident radiation flux density over all directions:

$$I_{ex} = I_{in}\cdot\int_0^{2\pi}\int_0^{\frac{\pi}{2}}\cos\theta\sin\theta d\theta d\phi \tag{3.63}$$

The cosine originates form the Lambertian reflection law as stated in Eq.3.30 (sine is just from the space integration in spherical coordinates). Evaluating the integral leads to:

$$I_{ex} = I_{in}\pi \tag{3.64}$$

In order for this to be energy conserved and to satisfy Eq.3.63, the factor π has to be introduced as done in Eq.3.33.

3.6 The Travel Function

The travel function τ determines, how light travels towards the observer after it is reflected off the object.

It is distinct but not independent from the BRDF/phase function.

Over very short distances it is usually neglected. In space applications, it is a significant factor.

As it is dependent on the BRDF or phase function, it is different for different shapes and reflection models.

In this work, two fundamental surface areas are used, the sphere and the flat plate; the reflection model is a mixture between Lambertian, specular reflection and absorption.

3.6.1 Spherical Surface

From a spherical surface, the radiation is spread equally in all directions from the object, from the halfspace of the sphere that is illuminated:

$$\tau_{\text{sphere,spec}} = 1 \tag{3.65}$$

$$\tau_{\text{sphere,lamb}} = \frac{1}{2\pi r_{\text{obj\_obs}}^2} \tag{3.66}$$

3.6.2 Flat Surface

For a flat surface the travel function τ_{flat} depends upon the type of reflection function. For the specular reflection as it is a directed radiation, no spread or loss outside the Earth atmosphere is taking place:

$$\tau_{\text{plate,spec}} = 1 \tag{3.67}$$

For a flat surface, the situation can be approximated via also via the scaling with the distance to the object as the radiation spreads out into the half-space, however not equally as with the sphere, but only around the local tangential direction:

$$\tau_{\text{plate,lamb}} = \frac{1}{r_{\text{obj\_obs}}^2} \tag{3.68}$$

3.7 Interaction with the EO Detector

3.7.1 Some Introductory Remarks

The light is received by a ground-based optic and sensor. It is assumed that it is equipped with a charge-coupled device (CCD) or complementary metal-oxide semiconductor (CMOS) sensor.

A typical CCD sensor is composed of a thin layer of photoactive semiconductors (typically silicons) and a transmitter region.

Photon impinging the sensor lead to electron emissions that are collected in the capacitor well; each well is a so-called pixel.

After the exposure, a control circuit leads to the readout of the CCD, in which each capacitor transfers its charge to the neighbouring pixel.

The large capacitor in an array reaches a charge amplifier and the electrons are transferred in a voltage level.

Processed through an analog to digital transformer the voltage levels are then stored.

The readout process it he crucial difference to a CMOS, in which each pixel is read out individually without shifting the charge. This leads to statistically independent pixel noises, of also independent noise levels and very fast readout times.

Electrons are generated proportionally to the amount of photons reaching the detector.

The so-called quantum efficiency is for uncooled sensors in the range of 60 percent, CO2 cooled sensors can reach quantum efficiencies of 97 percent.

The rate how many photons are counting into one analog to digital unit (ADU) is called (quantum) gain.

A perfectly linear gain is desired; in reality, CCD sensors have a range in which the gain is practically linear, linearity is thwarted at very low photon rates and approaching the saturation point.

Saturation is the maximum amount of photons that can be transformed into ADUs. If more than the maximum amount of photons is reaching the pixel, it can overflow, that is transferring charge to the neighboring pixels; this is called bleeding. Bleeding is to be avoided; it is controlled via shortening the exposure time.

CCDs are not perfect sensors, internal noise sources do exist.

Extra undesirable electron emissions are provoked by thermal energy, even in complete darkness, or if the shutter is not even opened.

They are referred to as the dark noise, colling reduces the dark noise levels significantly. Flaws in the CCD can lead to electron losses or spurious emissions (due to traps or recombinations) that affects the number of electrons transmitted. The readout process itself hence introduces through the charge transfer and circuit current, additional electrons; they are normally referred to as read-out noise. This rounding leads to a truncation error. More comprehensive descriptions on CCDs can be found in [36, 37].

Hot (always show the same value) pixels or dead (always show zero value) pixels are easily handeled in the image processing step. They are determined in a dark calibration measurement with closed shutter and than masked (simply omitted) in the image processing step.

In the observation of near-Earth objects, the field of view (FOV) ranges between half a square degree to up to 8×8 or more degrees. Often mosaics of CCDs are used. The pixel scales can vary between 0.5 arcseconds to several arcseconds per pixel.

3.7.2 Object Light Received and Object Image at the Detector

When the irradiation is passing through the optics, ir ia leading to the following expectation value for the signal function:

$$S_{\text{sig,obj}} = \int (D-d)\frac{\lambda}{hc} I_{\lambda\,\text{obj}}(\lambda) \cdot \exp(-\gamma(\lambda)R(\zeta)) \cdot L \cdot d\lambda \tag{3.69}$$

with:
c the speed of light
h the Planck's constant
ζ the zenith angle ($\pi/2-$ elevation)
γ the atmospheric extinction coefficient
R the atmospheric function
L is the loss function, describing a fractional loss as the light goes through the optical system
D is the aperture area, which is the same as the area of the primary mirror
d is the obstruction are relative to the aperture, e.g. the secondary mirror.

The simplest atmospheric model is $R = \frac{1}{cos\zeta}$, the so-called van Rhijn factor.

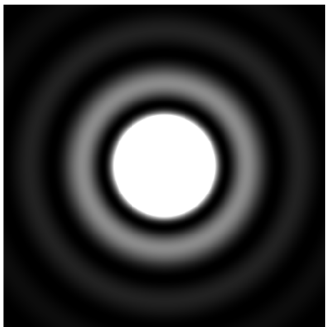
Normally an approximation is used replacing the explicit wavelength integration:

$$S_{\text{sig,obj}} \approx (D-d)\frac{\bar{\lambda}}{hc} \exp(-\gamma(\bar{\lambda})R(\zeta)) \cdot I_{\text{obj}}(\bar{\lambda}) \cdot L \tag{3.70}$$

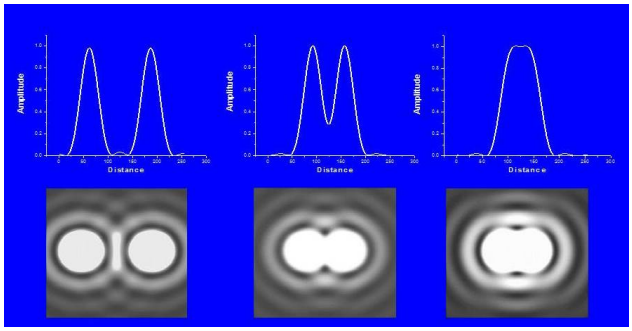
using $I_{\text{obj}}(\bar{\lambda})$ from Eq.3.1.

The count rate $C(S_{\text{obj}})_{\text{all}}$ is derived from the signal via the time integration Δt, during with the sensor is able to catch photons:

$$C(S_{\text{obj}})_{\text{all}} = \int S_{\text{sig,obj}} \cdot Q(\lambda) \cdot dt \approx S_{\text{sig,obj}} Q(\bar{\lambda})\Delta t, \tag{3.71}$$



(a) Diffraction pattern with Airy disk.



(b) Airy disk and resolution (courtesy: D.E.Wolf).

Q is the quantum efficiency.

This is the amount of light in the analog-to-digital unit (ADU). In SI units this is dimensionless.

The conversion into electrons would involve the multiplication with the gain g. Note that the gain is not linear over the whole range of detections.

The approximation neglects the shutter function itself and assumes that the integration time is the same over all the field of view of the sensor.

The signal, however, is spread over several pixels, which report their count rates separately.
Note: Because of the quantization of the sensing process, the signal has become a random variable.

This means, S, C_{all} are actually computed as means or expectation values.

Because the signal is (potentially) spread over several pixels, the diffraction of the circular aperture has to be taken into account.

This is a deviation of geometric optics that have been used to compute I_obj.

In the following, the model of Fraunhofer diffraction on a circular aperture is used.

Fraunhofer diffraction is the limit of the Fresnel diffraction for small Fresnel numbers.

The well established results are stated e.g. here [32].
The Airy disk is defined as the extension of the first maximum of the diffraction pattern on the detector.

The intensity, in our case the count rate at an angular distance θ from the center of the Airy disk is denoted as [32]:

$$C(\theta) = C_0 \cdot \left(\frac{2B_1(k \cdot r_D \sin\theta)}{k \cdot r_D \sin\theta} \right)^2 \quad \rightarrow \quad \sin\theta_{min1} = 1.22 \frac{\lambda}{2r_D} \tag{3.72}$$

$$C_0 = \frac{S_{\text{obj}}^2 D^2}{2f^4} \tag{3.73}$$

with: $k = \frac{2\pi}{\lambda}$ the wavenumber
$r_D = \sqrt{\frac{D}{\pi}}$ radius of the aperture
B_1 the first Bessel functions
f is the focal distance

θ_{min1} the angular distance to the first minimum at the detector
C_0 is the amplitude at the center of the Airy disk.

The larger the aperture, the better two different object images can be resolved.

Please note that here, the obstruction via the secondary mirror has not been taken into account.
Ground-based telescope imaging does suffer from the effects of atmosphere.

Atmosphere not only attenuates the signal, but the turbulent mixing of the atmosphere breaks up the Airy disk in speckle pattern.

The superimposed signal in the integration time during an observation interval leads to an effectively broadened signal at the detector.

This so called seeing leads to the fact that the size of the object image differs from the size of the Airy disk determined by the aperture of the telescope optic.
In general seeing is expressed in the full width of half maximum (FWHM) that the signal disk has on the detector as an angular measure:

$$FWHM_{\text{airy}} = \frac{1.028\lambda}{2r_D} \quad \rightarrow \quad FWHM_{\text{seeing}} = const. \tag{3.74}$$

Depending on the size of the telescope and the specific seeing conditions at the observing site, the FWHM can be dominated by the telescope aperture or limited by seeing.

The seeing is usually limited to around one arcsecond.

The larger value of the two is used for ground-based telescopes.
The diffraction function can be explicitly numerically evaluated, Eq.3.73.

Taking the seeing into account complicates the matter. Often, the signal count on the detector, created by the (smeared) Airy disk, is fitted with a function.

Different profiles are in use, but most often Gaussian functions or Lorentz functions are used. We are using a Gaussian going forward, which corresponds to a non or minimally distorted Airy disk.

The variance σ^2 of the Gaussian can be derived from the FWHM via the following relation:

$$FWHM = 2\sqrt{2\ln 2}\sigma \tag{3.75}$$

The volume underneath the two dimensional Gaussian fitting the Airy disk is adapted to match the volume of the Airy disk.

In an Airy disk, 83.8% of the overall volume is enclosed.

This can be evaluated via integrating the intensity function and the total encircled energy can be calculated by integrating Eq.3.73:

$$V_C = \int_0^\infty 2\pi \cdot C(\theta)\theta d\theta = C_0 \cdot 4\pi \cdot (1 - B_0^2(k \cdot D\sin\phi) - B_1^2(k \cdot D\sin\phi)) = 0.838 \cdot \bar{C}_{\text{all}} \tag{3.76}$$

Using the volume of a Gaussian function, one can derive the amplitude of the Gaussian:

$$V_{\text{Gauss}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A_{\text{Gauss}} \cdot \exp(-(\frac{(x-x_0)^2}{2\sigma^2} + \frac{(y-y_0)^2}{2\sigma^2}))dxdy = 2\pi A_{\text{Gauss}}\sigma^2 := V_C \tag{3.77}$$

$$A_{\text{Gauss}} = \frac{0.838 \cdot \bar{C}_{\text{all}}}{2\pi\sigma^2} \tag{3.78}$$

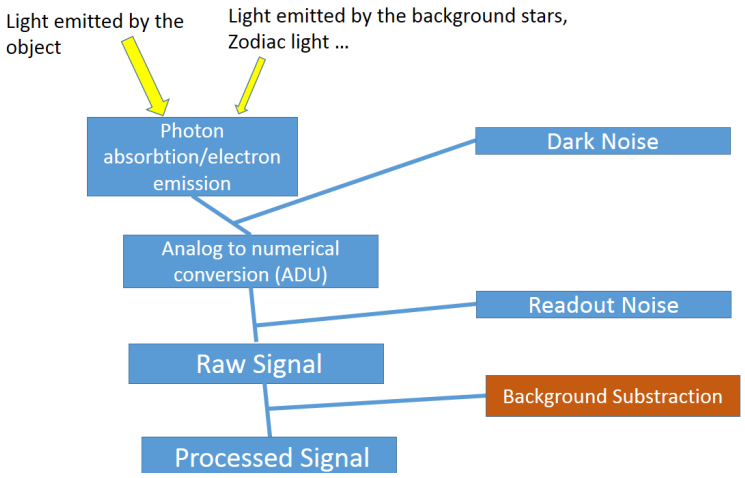


Figure 3.20: Noise generation in a CCD, [60].

Here the assumption is that the image of the object is a symmetric point source.

This is achieved when the telescope is not moving relative to the object during the exposure.

x_0, y_0 is the position of the center of the Gaussian within the pixel grid.

In order to be unit consistent, they should be expressed in units of pixels, same as the square root of the variance. The units are hence pixel widths and fractions thereof. To do this one uses the pixel scale, in units of arcseconds per pixel. In general, the Gaussian center is not in the middle of one pixel.

In order to compute the signal on the exact pixel grid:

$$\bar{S}_{\text{obj,cpix}} = \int_{x_0-\Delta x_1}^{x_0+\Delta x_2} \int_{y_0-\Delta y_1}^{y_0+\Delta y_2} A_{\text{Gauss}} \cdot \exp(-(\frac{(x-x_0)^2}{2\sigma^2} + \frac{(y-y_0)^2}{2\sigma^2}))dxdy \tag{3.79}$$

$\Delta x_i, \Delta y_i$ with i=1,2 are the distances to the edge of the pixel based upon the center of Gaussian.

Note: The center of the Gaussian is a random variable uniformly distributed within the center pixel.

Note: C_{all} is the mean of a Poisson random variable.

3.8 Image Noise

The irradiation of the object is not the only light that is reflected towards the observing sensor. In optical observations several background sources need to be taken into account.

3.8.1 The Internal Noise: Insights into Charged Coupled Devices

A more in depth treatment of the CCD-equation can be found in [60, 61]. Parts of this section are taken from the latter source.
This section may be omitted on a first reading

3.8.2 External Background Light Sources

Various light sources contribute to the so-called external background, that is the background noise that is not due to the object nor the detector. It includes effects from the celestial background, but also includes atmosphere related effects. Published values can be found in [5], [21].

3.8.2.1 Celestial Background Sources

The brightest background sources are the sun and the moon. If observations during daylight are planned, sunlight has to be taken into account thoroughly, only the very brightest objects are detectable away from the Sun direction. Sun stray light also has to be taken into account in space based missions, e.g. asteroid detection missions. In ground based observations, simplifications can be made, using sun set times. Astronomical sunset is at a sun elevation of -9 degrees, and has to be discriminated from nautical and ordinary sunset. In ground based observations, the moon may be treated in a geometrical sense,, too. Stray moon light and no-detection zones are normally expanded up to 15 degrees around the actual moon halo. Alternatively, both sources can be included in the noise calculation alongside other celestial irradiation sources.

In the course of this section, all spectral irradiances are defined in units SI units of $Watt/m^3$, all angles are considered in radians.

The brightest background source is so-called Zodiac light is the sunlight which is scattered by the dust in the ecliptic. It is hence a function of the ecliptic latitude and longitude with the same spectral distribution as the sun as a first order approximation. Zodiac light is obtained using look-up tables for the white light radiance.

$$I_{ZODI}(\lambda) = s^2 \cdot J_{ZODI}(\gamma, \delta) \cdot \frac{J_{\text{Sun}}(\lambda)}{E_{\text{Sun}}}, \tag{3.80}$$

where γ, δ are the longitude and latitude in the ecliptic coordinate system, $J_{ZODI}(\gamma, \delta)$ is the total radiance per unit angle. In general observations in the ecliptic are tried to be avoided, if other options (observing the object of interest in front of a different celestial background) exist. Besides the zodiac light, this is for the reasons of the accumulation of stars.

Stars are beside the zodiac light, the major light source. One way to include stars is to include them at the exact position as they appear in extensive star catalogs. However, this is a very time consuming procedure, if done for all stars. In addition, star catalogs are more imprecise towards the higher magnitudes. As a consequence, exact star positions are only extracted for the brighter stars, all stars around and with higher magnitude than the detection level of the instrument are smeared out as a background over the image. Tables exist with the number of stars of given photographic magnitudes, see e.g. [21, 5]. Using these, they can be converted to radiance values, assuming the spectral distribution of faint stars. The conversion is done in the blue wavelength ($440nm$), to have the best equivalence with the photographic magnitudes m. This leads to the spectral star irradiation:

$$I_{STAR}(\lambda) = n \cdot \frac{s^2 \cdot 32400}{pi^2} \cdot 6.76 \cdot 10^{-12-0.4 \cdot m} \frac{J_{GAL}}{\int J_{GAL} d\lambda} \tag{3.81}$$

where n is the number of stars in the assigned bin. The irradiation values correspond to the irradiation without an atmosphere.

A very faint but sometimes relevant background source is diffuse galactic light is a light source that is concentrated along the galactic plane. Its spectral radiance can be represented as the following:

$$I_{GAL}(\lambda) = s^2 \cdot J_{GAL}(\lambda) \exp(-\beta \cdot 180/(15 \cdot \pi)), \tag{3.82}$$

where J_{GAL} is the spectral radiance at unit angle zero galactic latitude β.

3.8.2.2 Atmosphere Related Background Sources

Two effects that are related to the Earth atmosphere are prominent in the background level of a ground based CCD image. The so-called airglow spectral radiation $I_{AG}(\lambda)$, which is the brightness of the atmosphere itself; it is faint glow if the

atmosphere itself, which is caused by chemiluminescent reactions occurring between 80 and 100 km. Atmospherically scattered light $I_{AS}(\lambda)$, is the sum of all light that is scattered by the atmosphere, excluding Sun and Moonlight. It is a contribution that varies little over the image, but adds to an overall elevated image background and hence should not be neglected.

$$I_{AG,AS}(\lambda) = s^2 \cdot J_i(\lambda) \cdot R(\zeta) \qquad J_i = J_{AG}, J_{AS} \tag{3.83}$$

where s is the angle under consideration, in case of the telescope, e.g. the field of view, or the angle that is fitted into a single pixel, $R(\zeta)$ is the van Rhijn factor, it can be approximated as $\frac{1}{\cos\zeta}$ in first order and describes the deviation from the zenith by angle ζ and the additional air mass and thickness, that has to be accounted for in low elevations [21]. J_{AG} is the spectral radiance of the zenith unit angle airglow in units of $Watts/m^2 ster \mu m$. J_{AS} is the spectral unit zenith angle radiance due to scattered light. It can be assumed that the faint star spectrum is an adequate representation.

3.8.2.3 White Approximation to Background

Sometimes one is not interested in a specific wavelength, but the total radiation, one can integrate or use approximations for the white light, which leads to the following, utilizing also the simple airmass approximation:

$$\mathfrak{I} = \int I(\lambda) d\lambda \approx \tilde{\mathfrak{I}} = I(\bar{\lambda}) \cdot \Delta\lambda \tag{3.84}$$

$$\tilde{\mathfrak{I}}_{AG} = s^2 \cdot \frac{1}{\cos\zeta} \cdot 1.42 \cdot 10^{-14} \quad [W/m^2] \tag{3.85}$$

$$\tilde{\mathfrak{I}}_{AS} = s^2 \cdot \frac{1}{\cos\zeta} \cdot 1.57 \cdot 10^{-15} \quad [W/m^2] \tag{3.86}$$

$$\tilde{\mathfrak{I}}_{GAL} = s^2 \exp(-\beta \cdot 180/(15 \cdot \pi)) \cdot 2.12 \cdot 10^{-15} \quad [W/m^2] \tag{3.87}$$

$$\tilde{\mathfrak{I}}_{ZODI} = s^2 \cdot J_{ZODI}(\gamma, \delta) \cdot 5.0 \cdot 10^{-15} \quad [W/m^2] \tag{3.88}$$

$$\tilde{\mathfrak{I}}_{STAR} = n \cdot s^2 \cdot 10^{-0.4 \cdot m} \cdot 3.0 \cdot 10^{-16} \quad [W/m^2] \tag{3.89}$$

$$\bar{T} = \exp(-0.27 \frac{1}{\cos\zeta}) \quad [-] \tag{3.90}$$

The atmosphere related and celestial background sources are then included in the image the same way as the light form the object itself, using Eq.3.69 and projected onto the pixel grid. For most precise background modeling the center coordinate of each pixel is used to determine the background level at this point and in the one pixel width area around it and integrating the irradiation, same as for the object signal. For low magnitude stars, the Gaussian shape approximation for the Airy disk, same as for the object irradiation should be used.

3.8.3 Signal-to-noise Ratio; the CCD equation

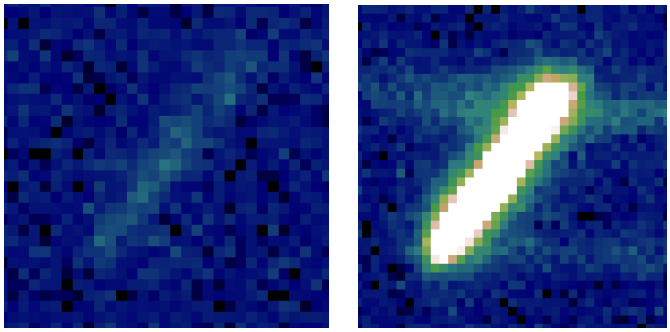
A more in depth treatment of the CCD-equation can be found in [60, 61]. Parts of this section are taken from the latter source.

The signal-to-noise ratio (SNR) is a quantity used in optical images to quantify how *bright* an object appears relative to the image background.

Fig.3.21 shows the comparison between a high SNR and a low SNR image of the same trailed object. Traditionally, the SNR is referred to as the CCD equation.

Classically, two different versions of the CCD equation are in use: the classical as in [69]Merline, Merline's derived in [53, 46]. A third version has been derived by Sanson, Frueh

The CCD equation is an analytical measure for a statistical quantity.



(a) Low SNR. (b) High SNR.

Figure 3.21: Different SNR images of the same object trace.

It is an approximation but saves doing a Monte-Carlo simulation of a large number of image realizations of the same observation frame.

The basis of all three derivations is that the electron emittance after absorption excited by the incoming photons is modelled by a Poisson random variable (hypothesis 1).

This is true for the signals from the object $S_{\text{obj},i}$ and $S_{\text{S},i}$. Furthermore the dark noise, fluctuations on the detector, is also modeled as a Poisson variable $N_{\text{D},i}$.
The following notation is used:

- n_{pix} is the number of pixels the signal is spread over
- $S_{\text{obj},i}$ the number of electrons emitted after absorption of photons emitted or reflected by the object for the pixel i. It is a Poisson random variable of parameter $\lambda_{\text{obj},i}$
- $S_{\text{S},i}$ the number of electrons emitted after absorption of photons emitted by background sources (e.g. stars) for the pixel i. It is a Poisson random variable of parameter $\lambda_{\text{S},i}$
- D_i the number of spurious electrons emitted for the pixel i (dark noise). It is a Poisson random variable of parameter $\lambda_{D,i}$
- R_i the number of electrons introduced by the read out process per pixel i.
- U_i is the number of electrons for the pixel i that are introduced by the limited CCD resolution.

$S_{\text{obj},i}$ is the signal in the different object pixels according to Eq.3.79.

The SNR is defined as the expectation value of the signal of interest divided by the standard deviation of the noise.

$$\text{SNR} = \frac{E\{S_{\text{obj}}\}}{\sqrt{\sigma^2(N)}} \tag{3.91}$$

The signal of interest is in our case, the object signal, that is the trace that the object leaves at the detector.

The object signal is spread over a number of pixels n. It is assumed to be well represented as a Poisson random variable The signal S and its expectation value can hence be written as:

$$S = \sum_{i}^{n_{pix}} S_{\text{obj},i} \qquad S := E\{S\} = \sum_{i}^{n_{pix}} \lambda_{\text{obj},i}, \tag{3.92}$$

In the classical and in the derivation of Merline of the CCD equation is assumed that the number n_{pix} of object pixels is exactly known (hypothesis 8).

The noise is defined as the variance of the sum of the object signal S together with the noise sources.

For the i^{th} pixel, the noise sources are the following:

Celestial and sky background sources $S_{\mathrm{S},i}$, such as stars, and other light sources, such as the zodiac light and other sources, that contribute to a non-zero photo background,

the dark noise, D_i, of the detector,

the read out noise R_i,

the truncation noise introduced U_i, that is introduced by the limited resolution of the sensor readout.

The realizations of dark and readout noise are influenced by the temperature of the detector.

Then the total noisy CCD output is:

$$S_{CCD} = \sum_{i}^{n_{\mathrm{pix}}} S_{obj,i} + \sum_{i}^{n_{\mathrm{pix}}} S_{S,i} + \sum_{i}^{n_{\mathrm{pix}}} D_i + \sum_{i}^{n_{\mathrm{pix}}} R_i + \sum_{i}^{n_{\mathrm{pix}}} U_i \tag{3.93}$$

The classical derivation concludes all these noise terms.

If it is assumed that all noise sources are independent (hypothesis 5), then the noise $N^2 = Var(S_{CCD})$ can be deduced from Eq. 3.93 :

$$N_{\mathrm{classical}} = \sum_{i}^{n_{pix}} S_{\mathrm{obj},i} + \sum_{i}^{n_{pix}} S_{\mathrm{S},i} + \sum_{i}^{n_{pix}} N_{\mathrm{D},i} + \sum_{i}^{n_{pix}} N_{\mathrm{R},i} + \sum_{i}^{n_{pix}} N_{\mathrm{U},i} \tag{3.94}$$

Making the variances explicit:

$$\sigma^2(N_{\mathrm{classical}}) = \sum_{i}^{n_{pix}} \lambda_{\mathrm{obj},i} + \sum_{i}^{n_{\mathrm{pix}}} \lambda_{\mathrm{S},i} + \sum_{i}^{n_{\mathrm{pix}}} N_{D,i}^2 + \sum_{i}^{n_{\mathrm{pix}}} N_{R,i}^2 + \sum_{i}^{n_{\mathrm{pix}}} N_{U,i}^2, \tag{3.95}$$

where $N_{R,i}^2 = \sigma^2(R_i)$, $N_{U,i}^2 = \sigma^2(U_i)$ and $N_{D,i}^2 = \sigma^2(D_i)$.

Recall that for a Poisson random variable the variance and the expected value are equal.

The truncation noise is modelled by an independent uniform random variable with support $[-\frac{g}{2}, \frac{g}{2}]$, where g is the gain. [53] (hypothesis 4), where U_i are independent and identically distributed (iid) uniform random variables with support $[-\frac{g}{2}, \frac{g}{2}]$; note that the variance is hence $N_{\mathrm{U,i}}^2 = Var(U_i) = \frac{g^2}{24}$.

Inspired by the work of [46, 45], the readout error is chosen modeled by a centered Gaussian distribution with variance $N_{\mathrm{R},i}^2$ in the classical formulation and in the formulation by [46] .
It is assumed that the readout noise, R_i is independent of the other components of the other signals (hypothesis 5).

The signal of the celestial background $S_{\mathrm{S},i}$ is assumed to be a Poisson random variable, with variance $\sigma^2(S_{\mathrm{S},i}) = \lambda_{\mathrm{S},i}$, same as the dark noise, accordingly D_i with $N_{\mathrm{D},i}^2 = \sigma^2(D_i) = \lambda_{\mathrm{D},i}$.

The background is assumed to be constant over the pixels that belong to the object image (hypothesis 6).

The dark noise and readout noise are assumed to be independent and identically distributed (iid) over all the image.

The noise can be written as:

$$\begin{aligned}\sigma^2(N)_{\text{classical}} &= S + n_{\text{pix}} \cdot (S_{\text{S}} + N_{\text{D}}^2 + N_{\text{R}}^2 + N_{\text{U}}^2) \\ &= \sum_i^{n_{\text{pix}}} \lambda_{\text{obj},i} + n_{\text{pix}} \cdot (\lambda_{\text{S}} + \lambda_{\text{D}} + N_R^2 + \frac{g^2}{24})\end{aligned} \tag{3.96}$$

The classical formulation of the CCD equation hence results in the following expression:

$$\begin{aligned}\text{SNR}_{\text{classical}} &= \frac{S}{\sqrt{S + n_{\text{pix}} \cdot (S_{\text{S}} + N_{\text{D},i}^2 + N_{\text{R},i}^2 + N_{\text{U,i}}^2)}} \\ &= \frac{\sum_i^{n_{\text{pix}}} \lambda_{\text{obj},i}}{\sqrt{\sum_i^{n_{\text{pix}}} \lambda_{\text{obj},i} + n_{\text{pix}} \cdot (\lambda_{\text{S},i} + \lambda_{\text{D},i} + N_{R,i}^2 + \frac{g^2}{24})}}\end{aligned} \tag{3.97}$$

The background subtraction is not included in the noise in the classical CCD equation. It is equivalent to assuming that the background is perfectly determined (hypothesis 7).
The CCD equation that is derived by Merline[46] differs in one significant instance from the classical derivation, which is, it takes the background estimation process into account, and leads to an additional term for the background noises. In the case of a constant background the estimated background is:

$$\sigma^2(B) = \frac{1}{n_{\text{B}}} \sum_i^{n_{\text{B}}} (S_{\text{S},i} + D_i + R_i + U_i) := \frac{1}{n_{\text{B}}} N_{b,d}^2 \tag{3.98}$$

Where B is the background subtraction term

n_B is the number of background pixels, which are used to estimate the background.

A common way of estimating the background is the background pixel identification method used is explained in [63]: The CCD image is divided in groups of m cells. In every group, the cell are ranked according their intensity. Then the p lowest intensity cells and the p highest intensity cell are dropped. The background is the mean value of the intensity of the pixels that have not been dropped. The size of the sub-frame should ideally be much larger than the signal.

The noise variance becomes:

$$N_{\text{Merline}}^2 = N_{\text{classical}}^2 + \frac{n_{\text{pix}}}{n_{\text{B}}} N_{b,d}^2 \tag{3.99}$$

This leads to the modified CCD equation of Merline:

$$\text{SNR}_{\text{Merline}} = \tag{3.100}$$

$$\frac{S}{\sqrt{S + n_{\text{pix}} \left(1 + \frac{1}{n_b}\right) (S_{\text{S}} + N_{\text{D},i}^2 + N_{\text{R},i}^2 + N_{\text{U,i}}^2)}}$$

$$= \frac{\sum_i^{n_{\text{pix}}} \lambda_{\text{obj},i}}{\sqrt{\sum_i^{n_{\text{pix}}} \lambda_{\text{obj},i} + n_{\text{pix}} \left(1 + \frac{1}{n_b}\right) (\lambda_{\text{S},i} + \lambda_{\text{D},i} + N_{R,i}^2 + \frac{g^2}{24})}} \tag{3.101}$$

Discussion of the Hypotheses of the Classical and Merline CCD Equation

Hypothesis 1 the number of electrons emitted after the absorption of photons is a Poisson random variable This assumption is plausible and is a classical model for electron emission.

Hypothesis 2 the background, signal and dark noise are independent: Independence is an accurate model since the electron emissions are emitted by different and independent sources, however an intense electric current increases

temperature by Joule dissipation leading to an increase in the dark noise, which is normally not the case in a cooled sensor.

Hypothesis 3 the pixel are uncorrelated: As long as the Poisson parameter λ_{obj} can be modeled as fully deterministic the pixels can be safely viewed as independent. The light reflected upon the object can be modelled using geometric optic macroscopic laws under the assumption that the object and the illumination and observation geometry is known. However, atmospheric disturbance modeling could be viewed as introducing thwarting the fully deterministic nature of the Poisson parameter, depending on the level of accuracy modeling. Furthermore, in a few particular cases with very high pixel intensity, there can lead to bleeding effects and in this case neighbor pixels may be correlated [7].

Hypothesis 4 the truncation noise is an independent additive uniform noise: During the truncation process the signal is converted from electrons into ADU. This conversion leads to lost in resolution: the CCD can only count a number of electrons at the time. This assumption is conceptually wrong and leads to inaccurate estimations of the truncation noise for faint signals (cf section III for more details), besides it entails that the signal remains a Poisson distribution after the round off error.

Hypothesis 5 the read out noise is an independent additive Gaussian noise: The read out error is a sum of independent random variables each accounting for a flaw in the electronics. The almost Gaussian distribution usually obtained [45] can be justified by Lindeberg-Feller theorem. Under mild assumptions on the U_i such as finite second moment, we have [26] $\sum_i^\infty U_i$ is normally distributed.

Hypothesis 6 the background is constant over the signal: Some studies such as [63] propose more complicated models of backgrounds. For instance, due to optical effects the background may be intense at the center of the image and celestial sources such as stars may vary from pixel to pixel, however for signal of reasonable size, the variation of the background are usually negligible.

*Hypothesis 7 the background is perfect determined:*This assumption that is assumed in the classical CCD equation and has been improved upon by [46], is wrong in general since only a limited number of pixels available to evaluate the estimated quantity of the background level.

Hypothesis 8 the number of signal pixels is perfectly known: As with the background estimation, the number of pixels that belong to the object is determined as the number of pixels above the background level. Especially for very faint signals this assumption is problematic. In this case it may be impossible to tell signal pixels from background pixels.

3.9 Optical Instrument Hardware

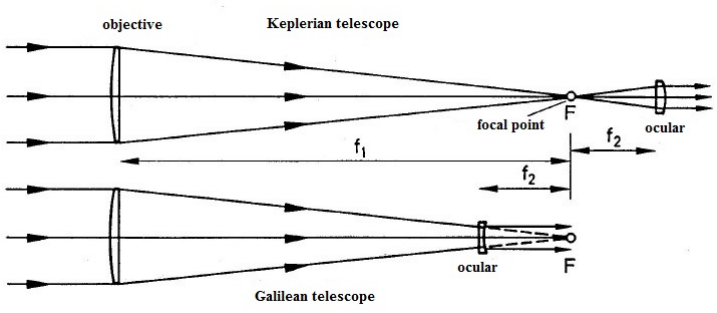


Figure 3.22: Historic Refractors [62].

In terms of the telescope optic one has to distinguish between refractors and reflectors.

Refractors use a correction lens to focus the light rays, reflectors work with mirrors.

Refractors lose a lot of light in the passage through the lense and precision lenses are not stable over time and are difficult to fabricate. The chromatic aberration is significant.

Refractors are not used in professional astronomy any more. The fraction of glass needed for a mirror is significantly less compared to a large lens so the stability of glass over time is less or a problem and in the construction of mirrors huge advances have been made. There is not chromatic aberration, however, astigmatism and coma. Furthermore, the secondary mirror is in the line of sight. Despite those disadvantages all modern telescopes are reflectors.

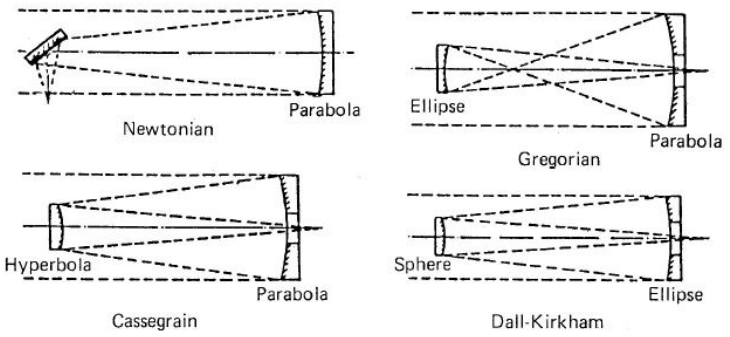


Figure 3.23: Reflectors [62].

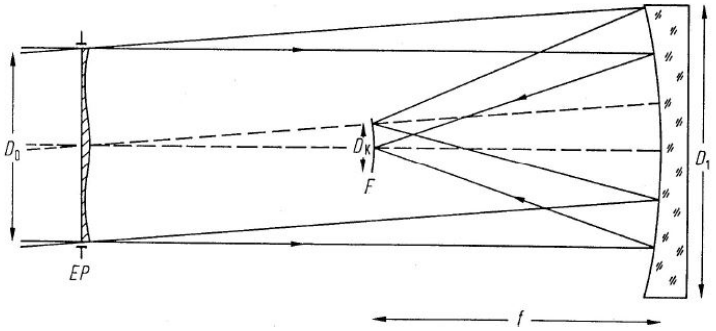


Figure 3.24: Wide-field telescopes [62].

The Schmidt telescope (schematics fig.3.24) was for a long time the state of the art instrument. It consists of a correction lens, a primary mirror and a secondary mirror. It allowed for very large field of views larger than ten degrees.

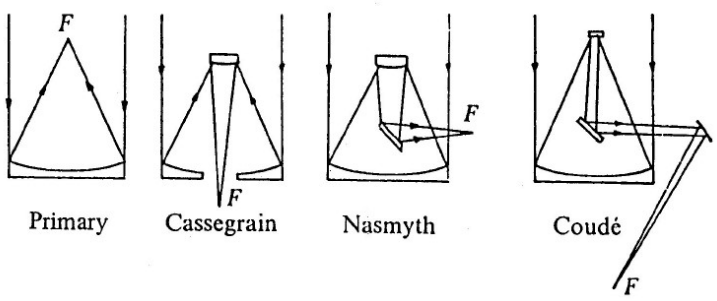


Figure 3.25: Ray path geometries[62].

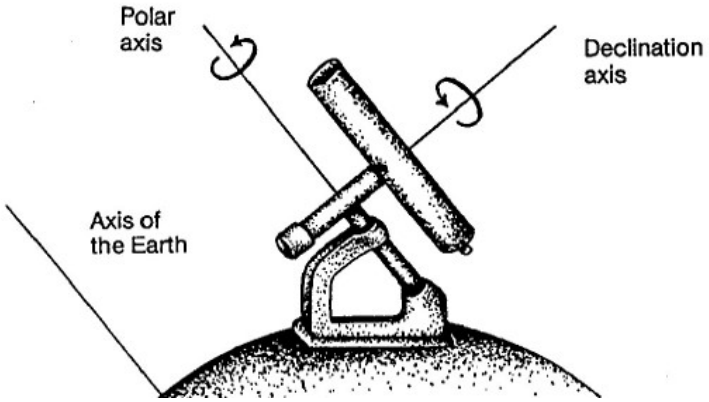


Figure 3.26: Parallactic or also called Equatorial mount geometry[62].

The paralactic of equatorial mount is the classical mount for star observations, as it can follows the star movement without frequent transpositions. However, it is technically more challenging (more stress on the single parts)and needs to be apt for the latitude at which the observer is located. Observations around meridian are tricky.

Alt-azimuth mounts are very simple and also intuitive to use/control. It is the most used mount. However, a movement around both axis is needed at all times and frequent transpositions occur.

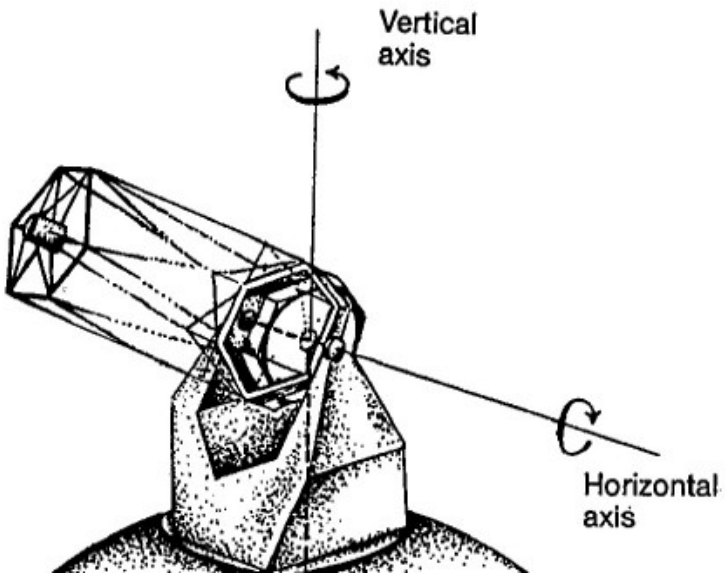


Figure 3.27: Alt-azimuth mount geometry[62].

3.10 A Few Useful Approximations and Expressions Characterizing Optical Systems

Field of View
The field of view (FoV) can be determined approximately by using the focal length of the main lense or mirror, denoted by f_1 and the detector size d in units of length via the opening angle β.

$$\text{FOV}[rad] = 2\beta = 2\tan^{-1}(\frac{d}{2f_1}) \approx \frac{d[length]}{f_1} \tag{3.102}$$

$$\text{FOV}[deg] = \frac{180}{\pi}\frac{d}{f_1} = 57.4\frac{d}{f_1} \tag{3.103}$$

Sometimes it is of interest how much a satellite at a given height h can fit into its FOV, at a given hight. The so-called Ground Sample Distance (GSD) is related to the FOV via:

$$\text{FOV}[rad] = 2\beta = \frac{\text{GSD}}{h} \tag{3.104}$$

Pixel Scale
The pixel scale *PS* is an approximation of what angular value is fitted into a single pixel of an optical system. It is strongly connected to the FOV:

$$\text{PS} = \frac{\text{FOV}}{d[pix]}, \tag{3.105}$$

where $d[pix]$ is the size of the detector in units of pixels. Usually the pixel scale is expressed in units of arcseconds for the sake of easy comprehension. One arcsecond is equal to $\frac{1}{3600}$ of a degree.

Resolution
The resolution has already been covered leading up and including Eq.3.73. Resolution is determined by the Rayleigh criterion. Two sources can be separated, if their angular separation distance is larger than the θ_{min} value determining

the first diffraction maximum, or in other words the size of the Airy disk. For a circular aperture of diameter D, the resolution is hence:

$$\theta_{\min} = 1.22 \sin^{-1}(\frac{\lambda}{D}), \tag{3.106}$$

f-Number
The so-called f-number, is often used to characterize a system, and is written usually as "$f/$"N, as in f/30, e.g., where N=30. The number N is simply defined as:

$$N = \frac{f_1}{D}, \tag{3.107}$$

where D is, as before, the aperture diameter, and f_1 is the focal length of the system.

Magnification
A paramter less relevant in space object tracking, because objects are non-resolved, the magnification is traditionally determnined by the ratio between the focal length f_1 of the system and the focal length of the eye-piece (or secondary) f_2, as illustrated in Fig.3.22:

$$\text{Magnification} = \frac{f_1}{f_2} \tag{3.108}$$

Approximating Limiting Magnitude
The actual limiting magnitude of a telescope system depends on many factors. As illustrated in the previous sections, the faintest object an optical system can detect depends at least on the background-noise, the exposure time, atmospheric turbulence condition, and the image processing methods, besides the optical system itself. A crude approximation can be made using the definition of relative bolometric magnitudes and comparing the telescope aperture diameter D to the pupil diameter $d_{\text{humanpupil}}$ which is about 7mm. Using the approximation of limiting magnitude for the average human eye to be $mag = 6$, the following, very approximate relation can be made, calculating limiting magnitude as:

$$\begin{aligned} mag_{\text{limit}} &\approx 6 + 2.5 \cdot \log_{10}(D^2/d^2_{\text{humanpupil}}) \\ &= 6 + 5 \cdot \log_{10}(D[mm]/d_{\text{humanpupil}}[mm]) = 6 + (5 \cdot \log_{10}(D[mm]) - 5 \cdot \log_{10}(d_{\text{humanpupil}}[mm]) \quad (3.109) \\ &\approx 2 + 5 \cdot \log_{10}(D[mm]) \quad (3.110) \end{aligned}$$

Please note that the units of D and $d_{\text{humanpupil}}$ have to be in agreement, customary millimeters ($[mm]$) are used. Other approximations take the background sky brightness into account and contrast the diameter of the aperture D with that one of the exit pupil of the telescope imaging system D_{exit}, and the limiting sky magnitude (background) mag_{sky}:

$$\begin{aligned} \tilde{m}ag_{\text{limit}} &\approx mag_{\text{sky}} + 2.5 \cdot \log_{10}(D^2/D^2_{\text{exit}}) \\ &= mag_{\text{sky}} + 5 \cdot \log_{10}(D/D_{\text{exit}}) \end{aligned} \tag{3.111}$$

3.11 Some Notes on Image Processing

Once the object images are located, a background level has to be determined.

The background, in general, is not uniform over an astrometric image.

Background determination usually done in an iterative process. An initial threshold is determined and iterated over. This can be done in a mosaic of smaller areas of the image, e.g. via sliding windows technique to avoid sharp changes in the background level. Alternatively or as a secondary step, a polynomial may be fit over the image to determine the background. The detection and centroiding process is sensitive to the background determination.

Background level corrections and callibrations can be made via dark level images (closed shutter) and flat-images of either a reference surface or a mostly uniform night sky at dusk or dawn.
Stacking: In classical astronomy, often many images are stacked and combined into one image to increase signal to noise. This also can be applied for observation of satellites and debris. However, either the stars or the object images can be enhanced, but not both at the same time, because of their relative movement. If objects are not tracked, only so-called blind stacking can be applied.

In order to determine an orbit of the object, two steps are necessary after background determination:

- Centroiding - Location of the object image (Airy disk) center
- Tranformation and mapping of pixel coordinates to celestial coordinates

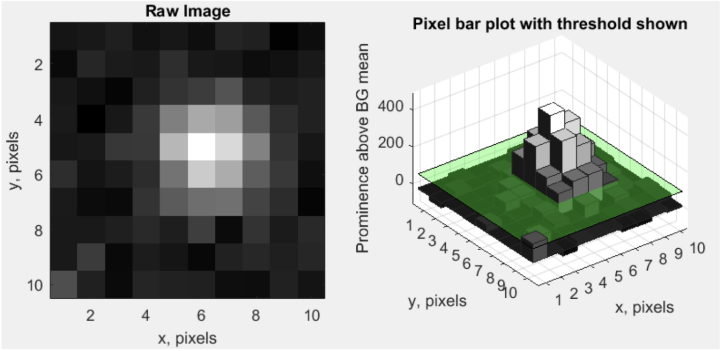


Figure 3.28: Illustration: Object Image and first step background determination.

A centroiding on the sub-pixel level is the aim. The fastest and simplest method of estimating the centroid $\hat{x}_0, \hat{y}_0$ of an object image in pixel coordinates is the so-called center of light method:

$$\hat{x}_0 = \frac{\sum_i^{n_{\text{pix}}} S_{obj,i} \cdot x_i}{\sum_i^{n_{\text{pix}}} S_{obj,i}} \tag{3.112}$$

$$\hat{y}_0 = \frac{\sum_i^{n_{\text{pix}}} S_{obj,i} \cdot y_i}{\sum_i^{n_{\text{pix}}} S_{obj,i}} \tag{3.113}$$

$$\tag{3.114}$$

x_i and y_i, respectively, denote the center location of the $i-th$ pixel in the x and y direction.

Center of light is not the most accurate method, especially when the signal-to-noise ratio is low. More accurate methods use the fitting of a function, such as a Gaussian or a Lorentzian. Another method is border and fill, which is independent of object image shape and is apt for highly distorted images.

For the transformation and mapping of the pixel coordinates to celestial coordinates, the star detections in the image are mapped to the positions listed in a star catalog. This necessitates coordinate and time transformations. Note that coordinates in star catalogs are aberration corrected. This and the fact that objects and/or stars appear streaked in space surveillance imagery is the reason, why many commercial astrometry tools developed for astronomy users, such as IRAF and astronomy.net, are only of limited use and currently do not deliver the actual, time-changing coordinates needed for satellite observation processing.

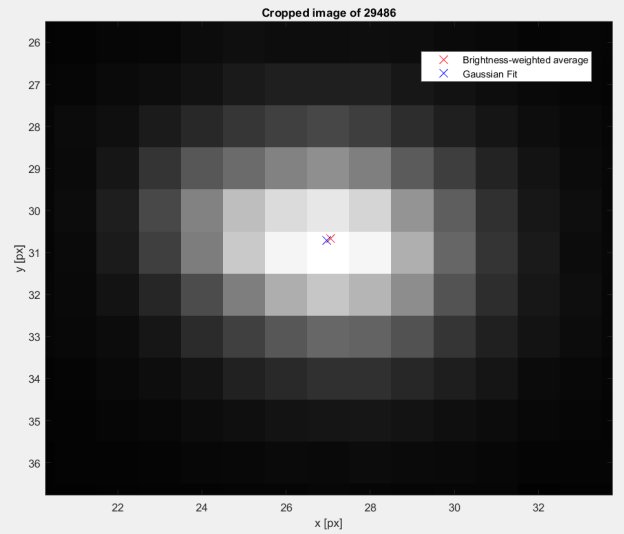


Figure 3.29: Centroiding using center of light and Gaussian fitting on the object image of 29486.

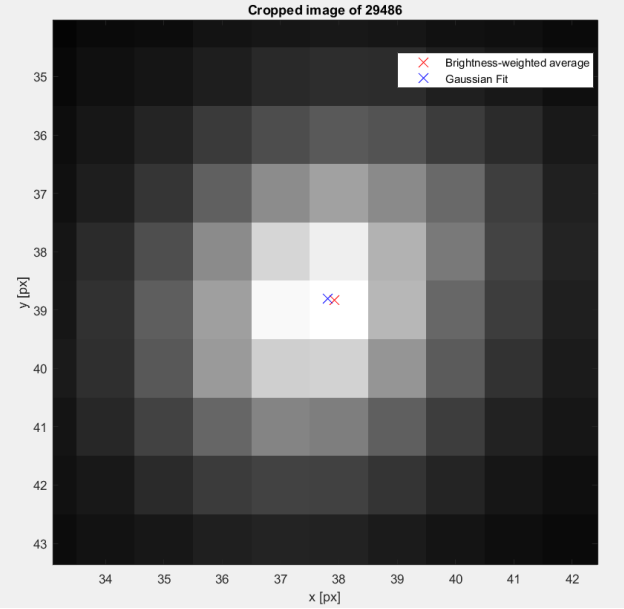


Figure 3.30: Centroiding using center of light and Gaussian fitting on a different object image of 29486.

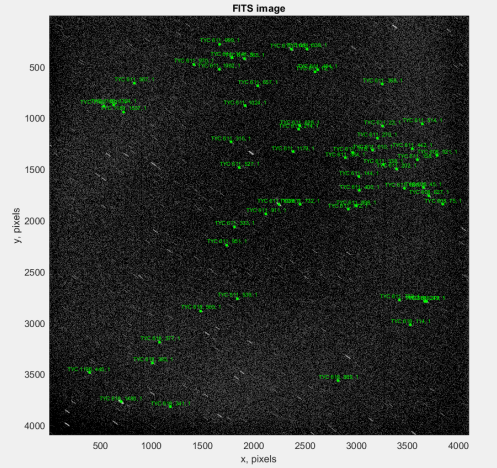


Figure 3.31: Observation Frame from the Purdue Optical Ground Station (POGS) with some matched stars to the Tycho star catalog.

Chapter 4

Coordinate systems and Time

This chapter is heavily based on the books of Dr Oliver Montenbruck [51, 49, 50].

4.1 Time

In order to determine a time system an origin, and a time scale needs to be defined. In space science, there are three different time scales of relevance:

- Earth rotation based (apparent daily motion of the stars and/or Sun)
- celestial mechanics based (orbital motion of moon, planets)
- atom physics based ((sub)atomic oscillations)

4.1.1 Earth Rotation Based Times

4.1.1.1 Solar Time

The Earth rotation based time scales are the solar time (=Earth time) and the sidereal time.

Traditionally, time is measured in days, determined by the subsequent meridian transits of the Sun with traditionally 86400 seconds.

As discussed before, the Sun's right ascension changes by around one degree per day, a solar day (Earth day) is hence about 4 minutes longer than the period of the Earth's rotation as seen from space, as shown in Figure 4.1.

A sidereal day amounts to 23h 56m 4.1sec and is equal to the time between successive meridian passages of the vernal equinox.

The true solar time or local time is defined as the hour angle of the apparent true sun (true sun shifted by aberration (see next section)) plus 12hours.

However, the true sun us not well suited for time measuring purposes and is replaced by the mean Sun, which moves uniformly in right ascension and declination (see Figure 4.2).

The equation of time (EoT) describes the time difference between the true sun and the mean sun, or in other words, the difference between the mean local time and the true local time.

The mean local time is the Greenwich hour angle of the apparent Sun (GHA) and the true local time is the UT, Greenwich hour angle of the mean Sun (GMHA):

$$EoT = GHA - GMHA \tag{4.1}$$

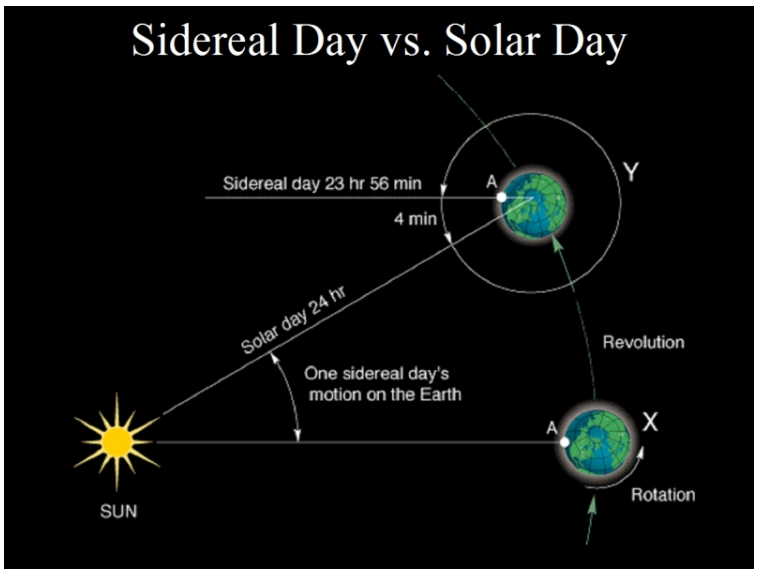


Figure 4.1: Sidereal Day vs. Solar Day [52]

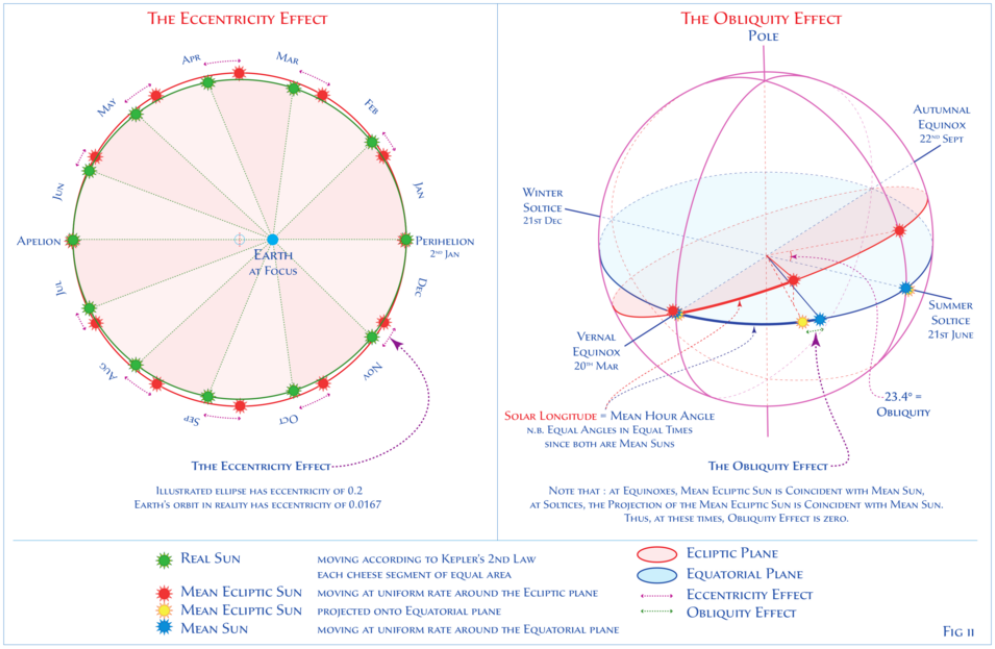


Figure 4.2: Mean sun visualization [44]

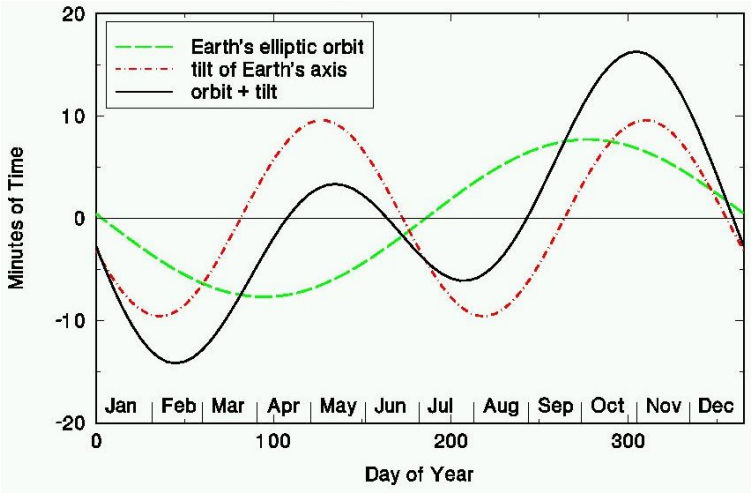


Figure 4.3: Equation of Time.

The difference between the mean and the true time are caused by the elliptic orbit of the Earth around the Sun, and that the Earth axis is not perpendicular to the ecliptic.

- the elliptic shape of the Earth orbit creates a periodic difference (period one year) of ± 7.5 minutes
- parallel translation of the inclined Earth axis (period 0.5 years) of ± 10 minutes

The two periodic changes are phase shifted; hence, the maximum values are around +16 and -14 minutes.

The equation of time is constantly changing by up to 30 seconds every 24 hours.

Tables for the equation of time typically have values given at 12 UT every day for a specific year.

However, for most applications it is sufficient to use this value for all topocenteric positions throughout the same day. Hence, the mean Sun is defined as a fictitious body that goes along the ecliptic with a constant angular velocity, and coincides at aphel and perihel with the true Sun. The mean local time is defined as the mean Sun's hour angle plus 12 hours.

In the 18th century, mean solar time was adopted. The IAU switched to ephemeris time in 1952 and then eventually to atomic time in 1972.

4.1.1.2 Universal Time

The universal time (UT) is a solar time (Earth time) scale, defined via the mean Sun.

The relation is fixed via the Greenwich sidereal time relation, as displayed in Eq.4.15. Hence:

$$UT = \theta - 12h - \alpha_{Sun,mean} - \Lambda, \tag{4.2}$$

with θ the sidereal time of an arbitrary observer,
$\alpha_{Sun,mean}$ the mean Sun's right ascension
Λ the observer's longitude.

The definition of the terrestrial time (TT) has been coordinated with UT such that at the beginning of his century their time difference was nearly zero.

Their time difference is however increasing by about 0.5 to 1 second per year.

This corresponds to the slowing of Earth rotation due to friction from tidal motion.

The time that is directly derived from the true sidereal time is called UT0.

If UT0 is corrected for the polar movement (next section) and short periodic tidal effects, it is called UT1.

If UT is corrected for half yearly tidal induced shifts, it is called UT2.
The Universal coordinated time (UTC) is defined as:

$$UTC = TAI + n(1sec) \tag{4.3}$$
$$|UTC - UT1| < 0.9sec. \tag{4.4}$$

TAI stands for "Temps Atomique Internationale," or international atomic time in English.

In order for both conditions to hold leap seconds are inserted, normally end of June or/and Dec 31.

4.1.1.3 Sidereal Time

The sidereal time is defined as the hour angle of vernal equinox, the relation Eq.4.15 is used to relate the sidereal time to UT and the Julian Date.

The position of the vernal equinox is affected by nutation.

The mean sidereal time is defined as the hour angle of the mean vernal equinox, in which the nutation effect is neglected.

Keep in mind that the sidereal time is also observer dependent (on the longitude of the observer).

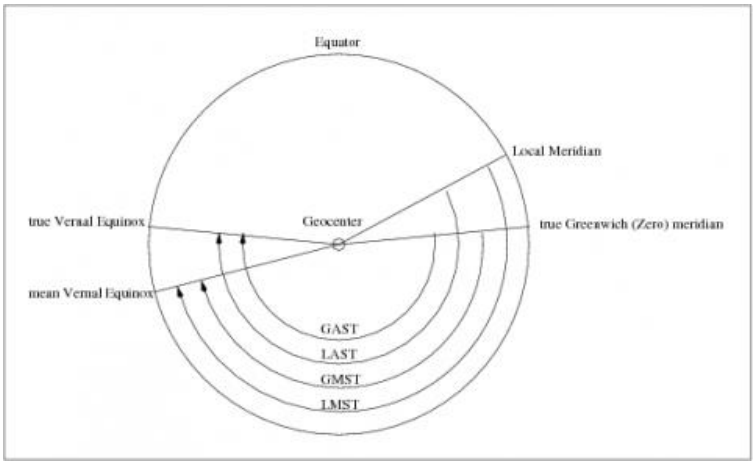


Figure 4.4: Local and mean times.

The Local Apparent Sidereal Time (LAST) and the Local Mean Sidereal Time (LMST) are connected to the Greenwich Apparent Sidereal Time (GAST) and the Greenwich Mean Sidereal Time (GMST) in the following way:

$$GAST = LAST - \Lambda \tag{4.5}$$
$$GMST = LMST - \Lambda, \tag{4.6}$$

where Λ is the longitude of the observer.

The mean sidereal time of Greenwich at 0h Earth time (UT) is given via:

$$\Theta(UT = 0h) = 24110.54841sec + 8640184.812866sec \cdot T + 0.093104 \cdot T^2 - 0.0000062 \cdot T^3 \tag{4.7}$$

T is defined as the time since January 1, 2000, 12h UT (JD 2451545), measured in centuries of 36525 days. This means one can rewrite as the following:

$$\Theta(UT = 0h) = 24110.54841sec + 8640184.812866sec \cdot T_0 + 0.093104 \cdot T_1^2 - 0.0000062 \cdot T_1^3 \tag{4.8}$$

$$T_0 = \frac{JD_0 - 2451545}{36525} \qquad T_1 = \frac{JD - 2451545}{36525}, \tag{4.9}$$

where JD and JD_0 are the Julian date of the time of observation and the Julian Date of 0h on the date of observation. Having the sidereal time at Greenwich at 0h UT, allows us to determine the sidereal time at Greenwich at every arbitrary time:

$$\begin{aligned} \Theta(UT) &= \Theta(UT = 0h)_0 + 1.0027279093 \cdot UT(sec) && (4.10)\\ &= 24110.54841sec + 8640184.812866sec \cdot T_0 + 0.093104 \cdot T_1^2 - 0.0000062 \cdot T_1^3 \\ &\quad + 1.0027279093 \cdot UT[sec] && (4.11)\\ &\quad T_0 = \frac{JD_0 - 2451545}{36525} \qquad T_1 = \frac{JD - 2451545}{36525}, && (4.12) \end{aligned}$$

An approximation that is often used is the following:

$$\begin{aligned} \Theta(UT) &= 6.664520h + 0.0657098244h \cdot (JD_0 - 2451544.5) + 1.0027279093 \cdot UT[hours] && (4.13)\\ & && (4.14) \end{aligned}$$

With this, we can compute the hour angle or sidereal time of any location on Earth, relating it to Greenwich:

$$\theta(UT) = \Theta(UT) + \lambda(1[hour]/15[deg]), \tag{4.15}$$

where λ is the geographic longitude (East of Greenwich).

The difference between the apparent sidereal time and the mean sidereal time is given via the difference between the right ascension of the true (=perturbed) vernal equinox and the mean vernal equinox:

$$\Theta_{\text{app}} = \Theta + \Delta\Psi \cos\varepsilon, \tag{4.16}$$

Θ_{app} is the apparent Greenwich sidereal time,
Θ is the mean Greenwich sidereal time,
$\Delta\Psi$ is the nutation in longitude
ε is the inclination of the ecliptic.

The difference between Θ_{app} and Θ is maximally one second. The numerical values for $\Delta\Psi$ and ε are derived in the section on nutation and precession.

4.1.2 Celestial Mechanics-Based Times

During mid-20th century, the irregularities became apparent with the improvement of Earth stationary clocks.

Hence, an artificial time, the Ephemeris Time (ET) has been determined, which has been calculated a posteriori from the orbits of the planets and the moon.

The theory has been put forth by Newcomb and according to him, the mean vernal equinox of date of mean longitude (mean anomaly plus longitude of the perihel) of the Sun can be determined via:

$$L = 279\deg 41'48.04'' + 129602768.13'' \cdot T + 1.089'' \cdot T^2, \tag{4.17}$$

T denotes the number of centuries since noon January 0 1900 (same as December 31, 1899).

Hence T is an independent variable, bringing forth of the notion of a dynamic time, where T is a solve-for parameter.

The count is hence initialized when the mean longitude of the sun equals $279\,\mathrm{deg}\,41'48.04''$. $\frac{dL}{dT}$ is then $129602768.13''$/century.

If one then defined the unit of T in 100 ephemeris years with 365.25 ephemeris days with 86400 ephemeris seconds, (hence in total 3 155 760 000 ephemeris seconds), then the Sun's longitude would have completed 360 degrees (uniformly) in the following time:

$$\frac{360 \cdot 3600''}{129602768.13''} \cdot 3155760000sec = 31556925.9747sec \tag{4.18}$$

Here the 360 degrees correspond to the time between one vernal equinox to the next, which is called one tropical year. This leads to the following definition of a second by the IAU:

One ephemeris second is the 31 556 925.9747-th part of the length
of the tropical year at Jan 0, 1900, 12 hours ephemeris time.

The ephemeris second is of practically equal length as the SI unit, hence:

$$ET = TAI + 32.184sec \tag{4.19}$$

Since 1984, the ephemeris time has been replaced by the Terrestrial Time (TT) and the Barycentric Dynamic Time (BDT).

Terrestrial Time is based on the SI second but the same relation as in Eq.4.19 holds, when Ephemeris Time is replaced with Terrestrial Time.

The center of mass representations of the solar system are normally done in Barycentric Dynamic Time. It takes relativistic effects of the observer and the difference in the length of time scales (different eigentime) into account.

4.1.3 Atom Physics-Based Times

The international atom time (Temps Atomique Internationale, TAI) currently fulfills best the notion of a continuous time. This led to the definition of the SI unit:
The SI unit is of the same length as 9 192 631 770 oscillations of the radiation that is created by the transition between the two hyperfine levels of the ground state of the cesium 133 atom.
TAI has been introduced in 1972 and replaced the ET as a basis of the definition of a second. TAI is linked to UT1 via:

$$TAI = UT1 \mathrm{on} \mathrm{Jan} 1, 1958, 0\mathrm{h} \tag{4.20}$$

4.1.4 Summary: Time-scales

Nowadays we refer to the following time scales:

- Terrestrial Time (TT) a conceptually uniform time scale that would be measured by an ideal clock on the surface of the geoid. TT is measured in day of 86400 SI seconds and is used as the independent argument of geocentric ephemerides.
- International Atomic Time (TAI), which provides the practical realization of a uniform time scale based on atomic clocks and agrees with TT except for a constant offset of 32.184 seconds and the imperfections of existing clocks.
- GPS time, which is like TAI an atomic time scale but differs in the chosen offset and the choice of atomic clocks used in its realization.

- Greenwich Mean Sidereal Time (GMST), the Greenwich hour angle of the vernal equinox (see previous section).
- Universal time UT1, today's realization of a mean solar time, which is derived from GMST by a conventional relation
- Coordinated Universal Time (UTC), which is tied to the International Atomic Time TAI by an offset of integer seconds that is regularly updated to keep UTC in close agreement with UT1
- for planetary and linear motion: Geocentric and Barycentric coordinate time (TGC and TCB) and Dynamical Barycentric time (TDB)

Dynamic times serve as independent argument in the equations of motion; atomic scales provide the practical realization of a uniform clock, and the non-uniform solar times scales are tied to the motion of the Sun and the rotation of the Earth. In the years from 1993 to 2004, a decrease of the length of the day of 2ms has been determined. There is a significant

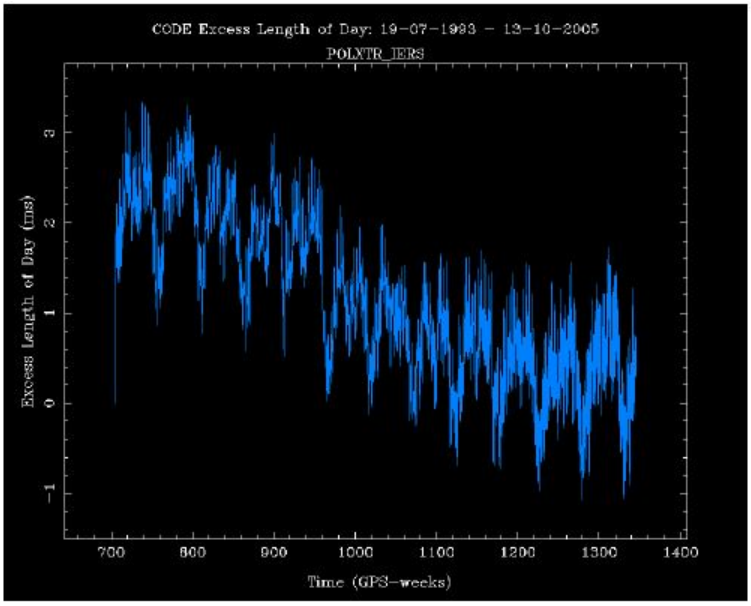


Figure 4.5: Length of one day, measured by CODE (caused by irregularities in the Earth rotation).

yearly change of 1ms. This is 98 percent due to change in the rotational impulse of the Earth's atmosphere and the non-rigidity of Earth. High frequency changes are caused by the moon (solid Earth tides). The Earth's rotation has been slowing down since 1000 before Christ. This means on average the length of the day increases by ms per year, and this causes UT1-UTC to differ by four hours in 2000 years.

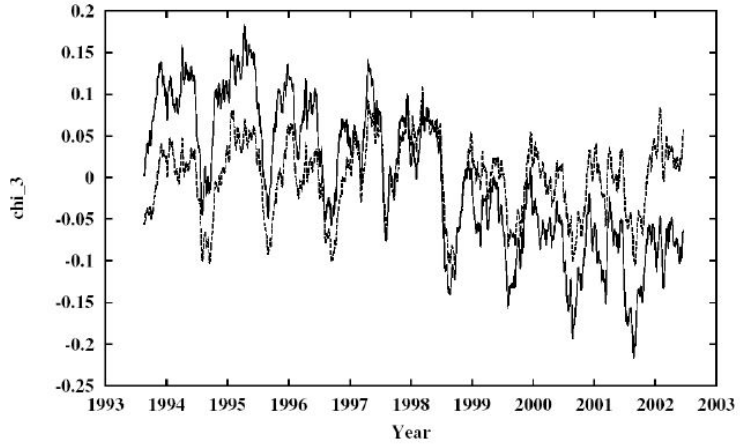


Figure 4.6: Rotational impulse of the rigid Earth and the atmosphere.

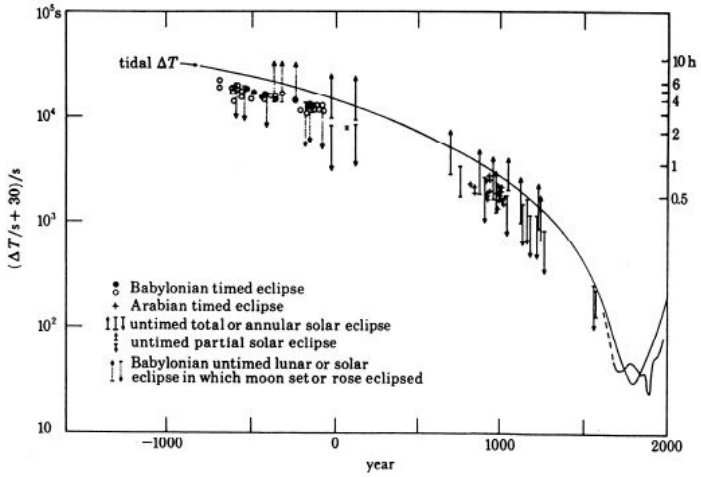


Figure 4.7: Earth rotation rate.

4.1.5 Julian Date

The Julian Date (JD) is defined as the number of days since January 1 4713 before Christ, 12h.

Until 1582 after Christ, the Julian calendar was in place.

In the Julian calendar, one leap day in every year that can be divided by four (without remnant) is inserted.

At the Gregorian reform of the calendar, October 4 was followed by October 15. Since then the leap year rule is the following:

A leap day is inserted at each year, which can be divided by four but not by 100, or can be divided by 400.

Determination of the Julian Date of a date year (Y), month (M), day (D), UT:

$$\begin{aligned}
&y = Y - 1 && m = M + 12 && \text{for} M \leq 2 \\
&y = Y && m = M && \text{for} M > 2 \\
&B = -2 && \text{till(inclusive)} 10/4/1582 \\
&B = floor(y/400) - floor(y/100) && \text{since(inclusive)} 10/15/1582 \\
&JD = floor(365.25y) + floor(30.6001(m+1)) + B + 1720996.5 + D + UT/24
\end{aligned} \tag{4.21}$$

where $floor(x)$ is denoted as the integer number that is less than or equal to x.

Note: During the time between March 1st, 1900 and February 28, 2100, B has the value of -15.

One can retrieve the calendar date from the Julian Date (JD) in the following way:

$$\begin{aligned}
&a = floor(JD + 0.5) \\
&c = a + 1524 && \text{for} a < 2299161 \\
&c = a + b - floor(b/4) + 1525 && \text{for} a \geq 2299161 \\
&b = floor((a - 1867216.25)/36524.25) \\
&d = floor((c - 122.1)/365.25) \\
&e = floor(365.25d) \\
&f = floor((c - e)/30.6001) \\
&D = c - e - floor(30.6001f) + (JD + 0.5 - a) \\
&M = f - 1 - 12floor(f/14) \\
&Y = d - 4715 - floor((7 + M)/10)
\end{aligned} \tag{4.22}$$

Nowadays often the modified Julian date (MJD) is used:

$$MJD := JD - 2400000.5 \tag{4.23}$$

MJD = 0.0 corresponds to November 17, 1858 0h.

4.2 Coordinate Systems

For the definition of a coordinate system, the following quantities need to be defined:

- origin
- fundamental plane

- direction of reference
- handedness (right-handed, left- handed system)
- (Cartesian or non-Cartesian (spherical, cylindrical))

In celestial mechanics, it is not unusual to refer to angles not only in degree (arcseconds) or radians, but also in measures of time. 1h corresponds to 15 degrees.

$$1h \equiv 15\,\text{deg} \tag{4.24}$$
$$1min \equiv 15' \tag{4.25}$$
$$1sec \equiv 15'' \tag{4.26}$$

$$1\,\text{deg} \equiv 4min \tag{4.27}$$
$$1' \equiv 4sec \tag{4.28}$$
$$1'' \equiv 0.067sec \tag{4.29}$$

Some fundamental terms:

- The plane that contains the Earth's orbit around the Sun is called ecliptic.
- The celestial equator, or also often just called shortly equator is the name for the plane that is perpendicular to the Earth rotation axis and expands out from the true Earth equator on the celestial sphere.
- The celestial sphere is a sphere around the Earth, sharing the same pole direction and the equatorial plane. It is a mathematical construct and has radius 1 (unitless).
- The directions on the line, which is defined by the intersection of the ecliptic and the equator, are called vernal or autumn equinox. The vernal equinox is defined to be the direction at which the Sun appears as seen from the Earth at the beginning of spring and her apparent trace shifts to the North side of the Earth hemisphere.
- Over longer times, neither the ecliptic nor the equator are stable (because of precession). Therefore, the equinox has to be defined, e.g. the equinox referring to a specific date, such as J2000.0 (referring to the vernal equinox in the year 2000).
- The topocenter is a reference point on the surface of the Earth, e.g. where the observer is located.

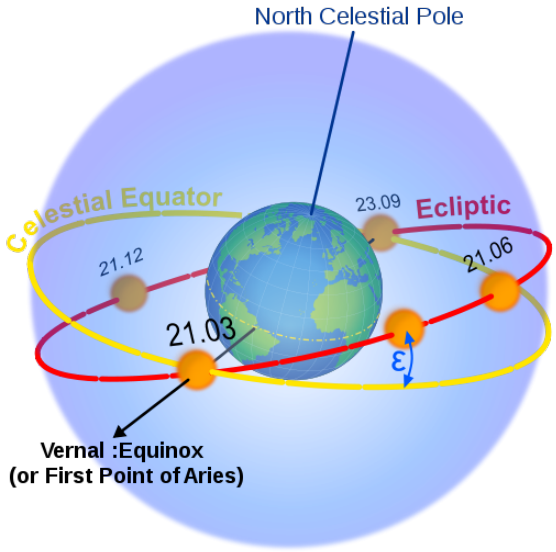


Figure 4.8: Equator and ecliptic, definition of equinox.

Definition of rotation and mirror matrices: The rotation matrices $\boldsymbol{R}_i$ and mirror matrices $\boldsymbol{S}_i$ in astronomy and celestial coordinate frame applications are traditionally defined as the following:

$$\boldsymbol{R_1}(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & cos\alpha \end{bmatrix} \qquad \boldsymbol{R_2}(\alpha) = \begin{bmatrix} \cos\alpha & 0 & -\sin\alpha \\ 0 & 1 & 0 \\ \sin\alpha & 0 & cos\alpha \end{bmatrix} \qquad \boldsymbol{R_3}(\alpha) = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.30}$$

Note that in our applications the rotation matrices are rotating opposite to the mathematically positive direction (and hence might differ from some other definitions)!

$$\boldsymbol{S_1} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{S_2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{S_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{4.31}$$

4.2.1 Coordinate Systems

4.2.1.1 Geocentric Equatorial System

- Origin: Center of the Earth
- Fundamental plane: Equator at a fixed equinox
- Reference direction: vernal equinox at a fixed equinox
- Handedness: right-handed system
- Coordinates: right ascension α', declination δ', (radial distance r).

α is the in-plane angle and defined to be zero for the direction to the vernal equinox. δ defines the angle above or below the equator (South $-\pi/2$, North $\pi/2$) counted from the equator plane. Besides radians or degrees, α is also sometimes measured in hours, minutes and seconds.

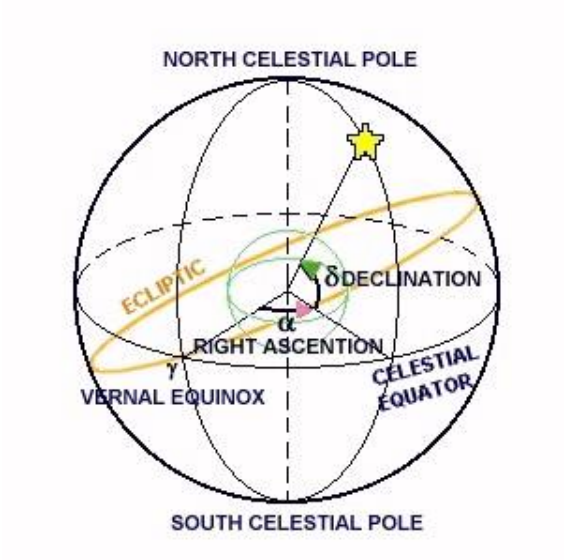


Figure 4.9: Geocentric equatorial coordinate system.

A state x,y,z is transformed via the rules of spherical coordinates:

$$x = r \cdot \cos(\delta') \cos(\alpha') \tag{4.32}$$
$$y = r \cdot \cos(\delta') \sin(\alpha') \tag{4.33}$$
$$z = r \cdot \sin(\delta) \tag{4.34}$$

$$r = \sqrt{x^2 + y^2 + z^2} \tag{4.35}$$
$$\tilde{r} = \sqrt{x^2 + y^2} \tag{4.36}$$
$$\delta = \frac{\pi}{2}, 0, -\frac{\pi}{2} \qquad \text{for} \qquad \tilde{r} = 0 \text{ and } z > 0, z = 0, z < 0 \tag{4.37}$$
$$\delta = \arctan(\frac{z}{\tilde{r}}) \qquad \text{for} \qquad \tilde{r} \neq 0 \tag{4.38}$$

$$\alpha = 0 \qquad \text{for} \qquad x = 0 \text{ and } y = 0 \tag{4.39}$$
$$\alpha = \phi \qquad \text{for} \qquad x \geq 0 \text{ and } y \geq 0 \tag{4.40}$$
$$\alpha = 2\pi + \phi \qquad \text{for} \qquad x \geq 0 \text{ and } y \leq 0 \tag{4.41}$$
$$\alpha = \pi + \phi \qquad \text{for} \qquad x < 0 \tag{4.42}$$
$$\phi = \arctan(\frac{y}{x}) \tag{4.43}$$

Some further questions: What is the definition inertial system? How does one find the center of the earth?

4.2.2 Geodetic and Geocentric Latitude and Earth Radius

For referencing an observer, either Catesian coordiantes can be supplied or spherical coordiantes. In terms of spherical coordinates, so-called *geodetic coordiantes* of latitude ϕ, longitude λ, and altitude h can be provided. The *geodetic* coordinate system is defined in relation to a reference ellipsoid. The Earth's ellipsoid is usually defined as an ellipsoid of revolution, which is axisymmetric , where two of the semimajor axes are equal. Geodetic latitude, ϕ, is the angle formed by the surface normal vector relative to the equatorial plane, the longitude λ is defined as the out-of-plane angle of any point relative to the prime meridian. The geodetic coordinates are usually the ones that are used when referring to locations on a map of a planet.

For our calculations, especially orbit mechanics, we prefer the geocentric latitude and longitude, which are formed formed by the vector pointing from geocenter to the specified point on a sphere. Thus, unless the reference ellipsoid is a sphere, these two types of latitude are not equal.

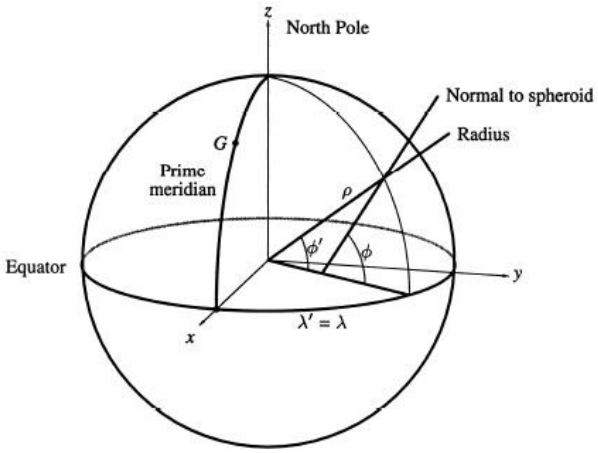


Figure 4.10: Center of Earth corrections, geocentric latitude ϕ', geographic (geodetic) latitude ϕ.

Of course, there are many different reference ellipsoids used for the Earth, so any geodetic latitude value should be accompanied by a specification of the reference ellipsoid used. In this script, we use the World Geodetic System '84 (WGS84) reference ellipsoid is used as the reference ellipsoid [24]:

1. **Origin:** The center of mass of the Earth, including the oceans and atmosphere
2. **Z-axis:** The direction of the IERS Reference Pole
3. **X-axis:** The intersection of the IERS Reference Meridian and the plane passing through the origin and normal to the z-axis

Parameter	Symbol	Value	Units
Semi-major Axis (Equatorial Radius of the Earth)	a	6378137.0	[m]
Flattening Factor of the Earth	1/f	298.257223563	[ ]
Geocentric Gravitational Constant	GM	$3.986004418 \times 10^{+14}$	[m^3 s^{-2}]
Nominal Mean Angular Velocity of the Earth	ω	7.292115×10^{-5}	[rad s^{-1}]

The transformation from geodetic latitude ϕ, longitude λ, and altitude h of a the WGS84 ellipsoid with semi-major axis a and flattening factor $1/f$ to Cartesian ITRS coordinates $\mathbf{r}^{ITRS}$ is given by [47]:

$$e^2 = 2*f - f^2 = 1 - \frac{b^2}{a^2} \tag{4.44}$$

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}} \tag{4.45}$$

$$\mathbf{r}^{ITRS} = \begin{bmatrix} (N+h)\cos(\phi)\cos(\lambda) \\ (N+h)\cos(\phi)\sin(\lambda) \\ (N(1-e^2)+h)\sin(\phi) \end{bmatrix} \tag{4.46}$$

where e^2 is the eccentricity of the ellipsoid squared, b is the semi-minor axis of the ellipsoid (the polar radius of the Earth), which is not needed if f is given, and N is the distance from the surface of the ellipsoid at the given latitude to

the z-axis along the direction perpendicular to the surface. From the Cartesian coordinates, the geocentric longitude λ' and latitude ϕ' can be solved for using the convention of spherical coordinates.

The reverse conversion is more invilved and is often solved iteratively, but the following algorithm by Heikkinen[33] is the most accurate numerically, according to Zhu[77]. Beginning with the known WGS84 ellipsoid parameters shown above, and a set of Cartesian coordinates $\mathbf{r}^{ITRS} = [x,y,z]^T$:

$$
\begin{aligned}
b &= a(1-f) \\
e^2 &= \frac{a^2-b^2}{a^2} \\
e'^2 &= \frac{a^2-b^2}{b^2} \\
F &= 54b^2z^2 \\
G &= r^2 + (1-e^2)z^2 - e^2(a^2-b^2) \\
c &= e^4Fr^2G^3 \\
s &= \sqrt[3]{1+c+\sqrt{c^2+2c}} \\
P &= \frac{F}{3\left(s+(1/s)+1\right)^2G^2} \\
Q &= \sqrt{1+2e^4P} \\
r_0 &= -\frac{Pe^2r}{1+Q} + \sqrt{\frac{a^2}{2}\left(1+\frac{1}{Q}\right) - \frac{P(1-e^2)z^2}{Q(1+Q)} - \frac{Pr^2}{2}} \\
U &= \sqrt{(r-e^2r_0)^2+z^2} \\
V &= \sqrt{(r-e^2r_0)^2+(1-e^2)z^2} \\
z_0 &= \frac{b^2z}{aV} \\
h &= U\left(1-\frac{b^2}{aV}\right) \\
\phi &= \arctan\left(\frac{z+e'^2z_0}{r}\right) \\
\lambda &= \arctan 2(y,x)
\end{aligned}
$$

where h is the altitude in the same units as a and b, ϕ is the geodetic latitude, and λ is the geodetic longitude calculated using a two-argument arc-tangent function to give the result in the correct quadrant.

4.2.2.1 Topocentric Equatorial System

Unfortunately, most observations are not performed at the center of the Earth, but on the surface of the Earth, at the so-called topocenter. This defines our topocentric equatorial system.

- Origin: Topocenter (position of the observer on the Earth surface, time-dependent)
- Fundamental plane: Plane parallel to the equator at a fixed equinox
- Reference direction: vernal equinox at a fixed equinox
- Handedness: right-handed system
- Coordinates: right ascension α, declination δ, (range ρ), sidereal time θ.

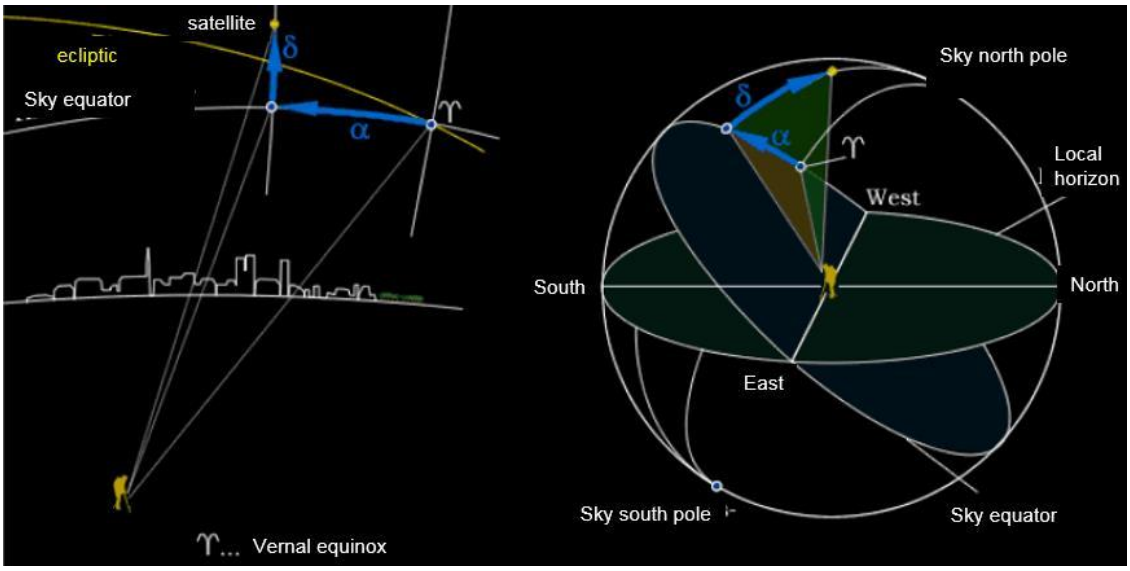


Figure 4.11: Illustration of the topocentric equatorial coordinate system.

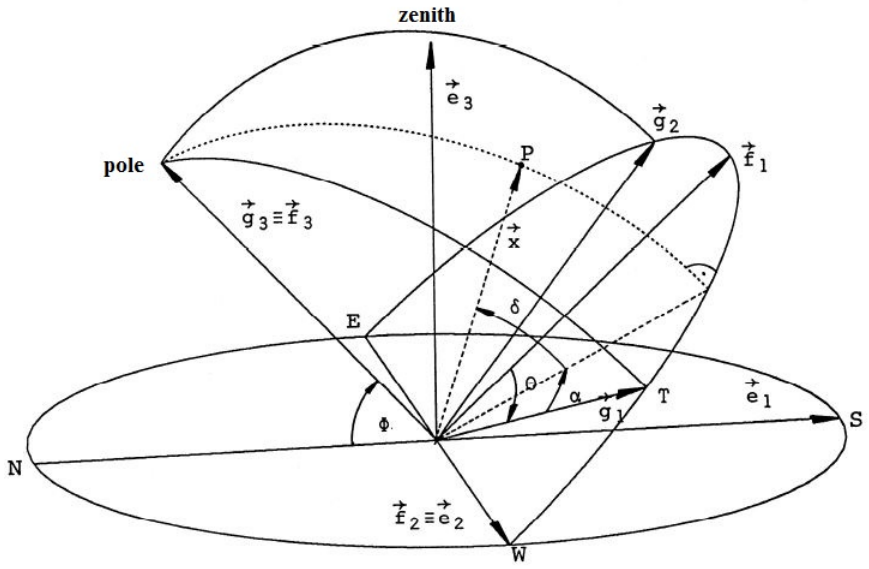


Figure 4.12: Illustration of the topocentric equatorial coordinate system, right ascension α, declination δ, sidereal time θ, and vernal equinox .

Note: The terms right ascension and declination are used for both the topocentric and the geocentric system. The declination of the zenith (the point directly above the observer) is equal to the geographic latitude of the observer. The right ascension of the zenith depends on the geographic longitude and the time, not on the latitude. For the observations of objects that are at stellar distances the difference between the topocentric and geocentric equatorial system are negligible, for earth orbiting satellites, the difference is crucial, because of the relatively small distance to the objects relative to the earth radius.

The sidereal time θ is the right ascension of the zenith at a given time t, the sidereal time of all observers at the same longitude is the same.

The hour angle τ is the difference between the sidereal time and the right ascension of an object: $\tau = \theta - \alpha$. The sidereal time is hence the hour angle of the vernal equinox. Hence, θ, τ, α are often measured in units of time rather than degrees or radians.

4.2.2.2 Direct Transformation from the Geocentric Equatorial to the Topocentric Equatorial Coordinate System

. The conversion is to add the momentary position of the observer to the topocentric observation vector, to obtain the geocentric vector. As we are operating in spherical coordinates this addition is done in spherical coordinates as well.

- geocentric equatorial coordinates of the object: α', δ', r
- topocentric equatorial coordinates of the object: α, δ, ρ
- geocentric geographic latitude of the observer (Earth fixed): ϕ'
- sidereal time of the observer: θ
- distance between the center of the earth and the topocenter, the corrected Earth radius plus the height of the station h_{sta}.

The geocentric vector to the object is:

$$\boldsymbol{r} = \begin{pmatrix} r\cos(\delta')\cos(\alpha') \\ r\cos(\delta')\sin(\alpha') \\ r\sin(\delta') \end{pmatrix} \tag{4.47}$$

Accordingly, the topocentric vector is:

$$\boldsymbol{r}_{topo} = \begin{pmatrix} \rho\cos(\delta)\cos(\alpha) \\ \rho\cos(\delta)\sin(\alpha) \\ \rho\sin(\delta) \end{pmatrix} = \rho \cdot \hat{L} \tag{4.48}$$

The position of the topocenter is computed using the sidereal time:

$$\mathbf{R_{topo}} = \begin{pmatrix} R\cos(\phi')\cos(\theta) \\ R\cos(\phi')\sin(\theta) \\ R\sin(\phi') \end{pmatrix}, \tag{4.49}$$

which is taking the spherical coordinates of the earth fixed position and using the sidereal angle as the in-plane angle, accounting for Earth's rotation. Note, in order for this to work, it has to be the sidereal time relative to this topocentric position. Alternatively, one could also take the sidereal time of Greenwich θ_0, leading to:

$$\mathbf{R_{topo}} = \mathbf{R_3}(\theta_0) \cdot \begin{pmatrix} R\cos(\phi')\cos(\psi) \\ R\cos(\phi')\sin(\psi) \\ R\sin(\phi') \end{pmatrix} \tag{4.50}$$

with ψ being the observer longitude relative to Greenwich. This leads to the full transformation expression:

$$r\cos(\delta')\cos(\alpha') = \rho\cos(\delta)\cos(\alpha) + R\cos(\phi')\cos(\theta) \tag{4.51}$$
$$r\cos(\delta')\sin(\alpha') = \rho\cos(\delta)\sin(\alpha) + R\cos(\phi')\sin(\theta) \tag{4.52}$$
$$r\sin(\delta') = \rho\sin(\delta) + R\sin(\phi') \tag{4.53}$$

The sidereal time of the observer is the angle by which at a given epoch the vector from the center of the Earth. We know that the topocenter makes one turn in 24 hours (Earth time (!)); however, what is needed is the orientation relative to the vernal equinox at a given time. A detailed discussion of different time systems will be provided in the next section. For now, we assume either the hour angle or sidereal time needs to be given to do the transformation. Hence, in general:

$$\boldsymbol{r} = \mathbf{R_{topo}} + \boldsymbol{r_{topo}} = \mathbf{R_{topo}} + \rho \cdot \hat{\mathbf{L}} \tag{4.54}$$

4.2.2.3 Topocentric Local Horizon Coordinate System

- Origin: Topocenter (position of the observer on the Earth's surface)
- Fundamental plane: local horizon
- Reference direction: South (direction, in which places of the same geographic latitude but smaller latitude are located)
- Handedness: left handed system
- coordinates: elevation h, azimuth a, (normally range ρ is not reported in this system). h is the angle above (positive) or below (below) the local horizon, zenith is defined to be $\pi/2$. a is defined from 0 (South) to 2π.

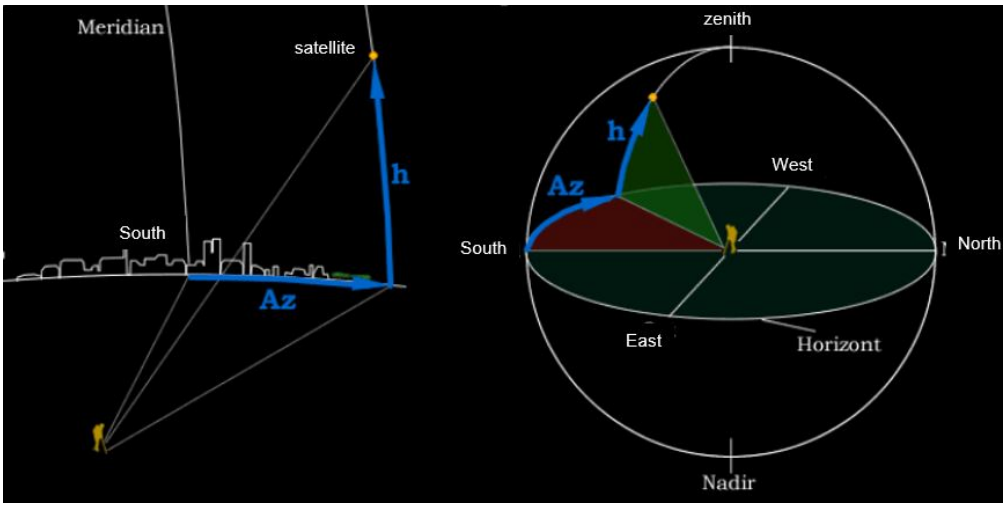


Figure 4.13: Illustration of the local horizon coordinate system.

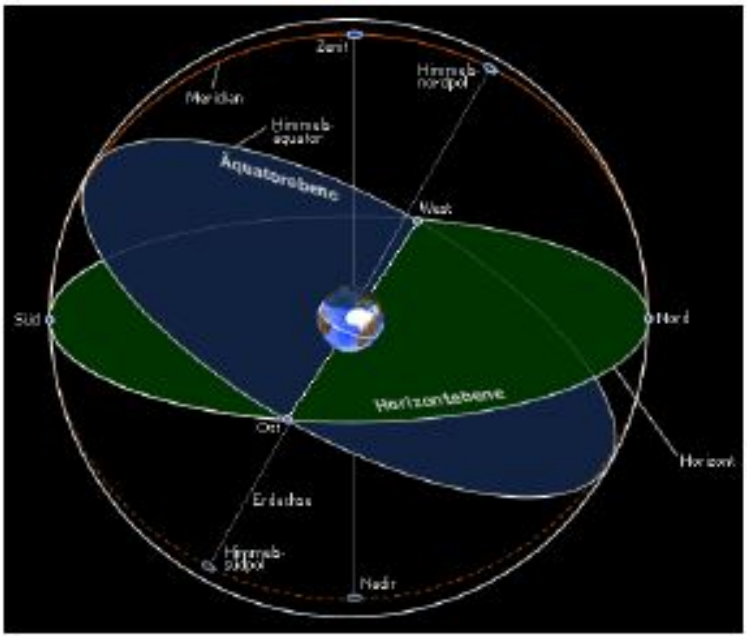


Figure 4.14: Illustration of the local horizon coordinate system.

4.2.2.4 Transformation of the Topocentric Local Horizon and Topocentric Equatorial Coordinate System

Coordinates:

- τ hour angle ($\tau = \theta - \alpha$)
- θ sidereal time of the observer (in angular measure)
- ϕ geographic latitude of the observer
- α, δ topocentric equatorial coordinates at a fixed equinox: right ascension and declination
- a, h azimuth, elevation (local horizon coordinates, Earth fixed)

The transformation is done:

$$\begin{pmatrix} cos(\alpha)cos(\delta) \\ sin(\alpha)cos(\delta) \\ sin(\delta) \end{pmatrix} = \mathbf{S_2 R_3}(\theta)\mathbf{R_2}(-(\frac{\pi}{2} - \phi)) \begin{pmatrix} cos(a)cos(h) \\ sin(a)cos(h) \\ sin(h) \end{pmatrix} \tag{4.55}$$

This means, in order to get from the local horizon system to the topocentric equator system, for one, only the angular values, no distances are needed. The first step is to correct for the orientation of the fundamental plane, making a rotation around the second axis R_2. Then we have to move the reference direction towards the vernal equinox using the rotation around the third axis R_3 by the sidereal time. Lastly, the system is mirrored along the second axis, to fix the handedness S_2.

To write it out explicitly using the hour angle, leads to:

$$\begin{aligned} \cos(\delta)\cos(\tau) &= \cos(\phi)\sin(h) + \sin(\phi)\cos(h)\cos(a) \\ \cos(\delta)\sin(\tau) &= \cos(h)\sin(a) \\ \sin(\delta) &= \sin(\phi)\sin(h) - \cos(\phi)\cos(h)\cos(a) \end{aligned} \tag{4.56}$$

Alternatively, to put it the other way around:

$$\begin{aligned} \cos(h)\cos(a) &= sin(\phi)\cos(\delta)\cos(\tau) - \cos(\phi)\sin(\delta) \\ \cos(h)\sin(a) &= \cos(\delta)\sin(\tau) \\ \sin(h) &= \sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\cos(\tau) \end{aligned} \tag{4.57}$$

4.2.3 Refraction

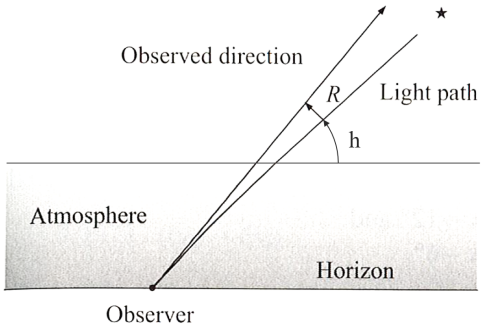


Figure 4.15: Refraction.[48]

When light enters the optically thicker Earth atmosphere, it is refracted towards the zenith direction.

Because of the longer way of the light during the atmosphere, the refraction is most prominent close to horizon, and hence influences rise and set times significantly.

The refraction is dependent on the refraction index of the atmosphere, and hence depends on the temperature and atmospheric pressure.

The approximation of the refraction has been found in fitting observed data. The formula is hence empirically derived:

$$R = \frac{p}{T}[3.430289(z' - \arcsin(0.9986047\sin(0.996714z'))) - 0.01115929z'], \tag{4.58}$$

with:
h true elevation [degrees]
h' observed elevation [degrees]
z'= 90 degrees - h' observed zenith distance [degrees]
p atmospheric pressure [hPa]
R refraction R=h'- h [arc minutes]
T temperature in Kelvin

Table 3.2. Values of refraction near the horizon

h	10°	5°	2°	1°	0°
R	5′31″	10′15″	19′7″	25′36″	34′

Figure 4.16: Effects of the refraction.[48]

4.2.4 Note on commonly used names

The Earth centered equatorial system is often also referred to as the Earth Centered Inertial (space fixed) frame (ECI) with the reference direction vernal equinox. It is normally contrasted by the Earth Centered Earth fixed frame (ECEF), which is a Cartesian frame, fixed on the Earth with reference direction South.

4.2.5 Aberration and Light Travel time

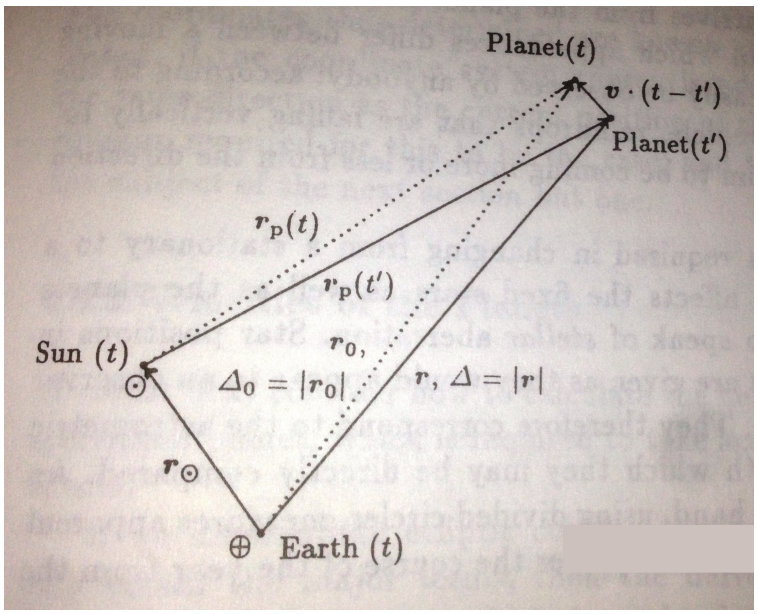


Figure 4.17: Light Travel Time.

The light travel time and aberration effects are an often-overlooked factor in the generation of pseudo-observations and in dealing with real observations.

The light is reflected off the object and takes time to reach the observer, time in which the object has already moved.

When taking optical observations of unknown objects, no range information is observable, and hence light travel times and aberration cannot be determined.

However, it can be inserted into the procedure of first orbit determination and orbit improvement.

4.2.5.1 Angle measurements

The distance from the observer at location $\mathbf{R}$ on the Earth surface and the object at position $\mathbf{r}$ is denoted as $\mathbf{d} = \mathbf{r} - \boldsymbol{R}$.

The time at the reception if the signal is however different from the time at which the light left the object. We write that in the following manner:

$$\boldsymbol{d} = \boldsymbol{r}(t-\tau) - \boldsymbol{R}(t) \tag{4.59}$$

The signal travel time may be computed from the implicit light-time equation:

$$c \cdot \tau = |\boldsymbol{r}(t-\tau) - \boldsymbol{R}(t)|, \tag{4.60}$$

where c is the velocity of light. Starting from an initial guess $\tau^{(0)} = 0$ the light travel time is consecutively determined in the fixed-point iteration:

$$\tau^{(k+1)} = \frac{1}{c}|\boldsymbol{r}(t-\tau) - \boldsymbol{R}(t)|. \tag{4.61}$$

The integration may be continued until a threshold of $\tau^{(k+1)} - \tau^{(k)} < 10^{-7}$ is reached.

For a LEO satellites that would lead to an accuracy in the light correction of around 7".

The solution of the light-time equation leads to the true signal path in the inertial system but is different form the apparent direction from a moving platform (ground station or e.g. on orbit observer).

Aberration is caused by the relative motion of the signal to the observer.

The motion of a ground based observer leads to daily aberration and yearly aberrations, due to Earth daily motion and orbital motion around the Sun.

Neglecting the relativistic corrections and staying in Newtonian motion models the observed direction $\boldsymbol{d}'$ compared to the true direction $\boldsymbol{d}$ can be expressed as the following:

$$\boldsymbol{d}' = \boldsymbol{d} + \tau \boldsymbol{v}_o, \tag{4.62}$$

where $\boldsymbol{v}_o$ is the velocity in the geocentric space fixed inertial frame, of the ground station. This means:

$$\boldsymbol{d}' = \boldsymbol{d} + \tau \boldsymbol{v}_o = \boldsymbol{r}(t-\tau) - \boldsymbol{R}(t) + \tau \boldsymbol{v}_o \approx \boldsymbol{r}(t-\tau) - \boldsymbol{R}(t-\tau), \tag{4.63}$$

matches the true position of the object at time $t - \tau$ to first order.

Aberration itself is of the order of 0.6" for LEO and 0.3" for GEO, contrary to what literature states, it should not be neglected.

Note: In optical observations, positions in the image processing are extracted relative to the star positions, which are listed in star catalogs on the same CCD image.

Star catalogs are corrected for the yearly and daily aberrations, hence report aberration free positions.

If directly compared with the objects in the same frame, leads multiples of the aberration offsets, as positions in two different coordinate frames are compared. Hence, we also look at the aberration equations for the stars:
yearly aberration:

$$\Delta\alpha_j = \frac{A + A_e}{\cos(\delta)} \tag{4.64}$$

$$\Delta\delta_j = D + D_e \tag{4.65}$$

with

$$A = -20.49'' [\sin(L)\sin(\alpha) + \cos(L)\cos(\alpha)\cos(\varepsilon)] \tag{4.66}$$

$$A_e = 0.343 [\sin(p)\sin(\alpha) + \cos(\alpha)\cos(p)\cos(\varepsilon)] \tag{4.67}$$

$$D = -20.49'' [\sin(\delta)\cos(\delta)\sin(L) + (\sin(\varepsilon)\cos(\delta) - \cos(\varepsilon)\sin(\delta)\sin(\alpha))\cos(L)] \tag{4.68}$$

$$D_e = 0.343'' [\sin(\delta)\cos(\alpha)\sin(p) + (\sin(\varepsilon)\cos(\delta) - \cos(\varepsilon)\sin(\delta)\sin(\alpha))\cos(p)] \tag{4.69}$$

Daily aberration:

$$\Delta\alpha_t = 0.32'' \frac{\cos(\phi)\cos(\theta - \alpha)}{\cos(\delta)} \tag{4.70}$$

$$\Delta\delta_t = 0.32'' \sin(\delta)\cos(\phi)\sin(\theta - \alpha) \tag{4.71}$$

with:
$\Delta\alpha_j, \Delta\delta_j$ yearly aberration, correction of the equatorial coordinates for an observer in the geocenter, who moves with

the Earth
$\Delta\alpha_t, \Delta\delta_t$ daily aberration, correction for the equatorial coordinates for an observer, who participates in Earth daily rotation (on the Earth surface, neglecting R as $R << r$.
α, δ right ascension and declination of the observed star
L ecliptic longitude of the sun
p perihel longitude of the apparent sun orbit ($\approx$ 283 degrees)
ε inclination of the ecliptic ($\approx$ 23.44 degrees)
ϕ geographic latitude of the observer
θ sidereal time of the observer at the time of the observation
The corrections A and D correspond to the circular yearly motion of the Earth. They depend on the location of the observer and on the location of the Sun. The eccentricity of the Earth orbit, motivates the inclusion of the terms D_e, A_e. They are time dependent.

4.2.5.2 Range Measurements

The only difference in range measurements compared to optical measurements is comprised in the fact that so-called active illumination, that is, up-link and downlink, takes place.

When the signal is recorded at the ground station at time t, the signal has been received and transmitted back by the satellite at $t - \tau_d$, where τ_d is the downlink light travel time.

The transmission time of the signal at the ground station is then, quite self-explanatory, $t - \tau_d - \tau_u$ with τ_u being the up-link light travel time. This means the light-travel-time-equation changes to:

$$c \cdot \tau_u = |\boldsymbol{r}(t - \tau_d) - \boldsymbol{R}(t - \tau_d - \tau_u)|, \tag{4.72}$$

and

$$\tau_u^{(k+1)} = \frac{1}{c}|\boldsymbol{r}(t - \tau_d) - \boldsymbol{R}(t - \tau_d - \tau_u^{(i)})|, \tag{4.73}$$

it requires one iteration step less than in the optical case as the initial value of $\tau_u^{(0)} = \tau_d$ can be applied. Because of the different velocities of the satellite and the station, the light time correction to the up-links is a factor of around 20 smaller than for the downlink.

The two way range measurement ρ is then modeled:

$$\rho = \frac{1}{2}(\rho_u + \rho_d) = \frac{1}{2c}(\tau_u + \tau_d). \tag{4.74}$$

4.2.5.3 State correction

In case the full state is known at time t, the state at time t of either transmission of the signal from Earth or at the time the reflected light left the object, can be recovered either via full orbit backward propagation or via Taylor series expansion:

$$\boldsymbol{r}(t - \tau) \approx \boldsymbol{r}(t) - \dot{\boldsymbol{r}}(t)\tau + \frac{1}{2}\ddot{\boldsymbol{r}}(t)\tau^2 \tag{4.75}$$

The light travel time ranges for LEO satellites from around 5 ms (2.5 ms one-way), speed around 7.5 km/sec, to around 100 ms (50ms), speed 3km/s, for GEO satellites. This means the linear term is about 400 (200) meters for GEO satellites and around 100 (50) meters for LEOs.

4.3 Standard Epochs and Bessel Year, Reference Systems

In celestial mechanics, the reference to standard epochs has been shown to be beneficial. The standard epochs are discriminated via Julian centuries of 36525 days, and have the prefix "J".

J1900: JD 2 415 020.0 = Jan 0.5 days, 1900
J2000: JD 2 451 545.0 = Jan. 1.5 days, 2000

The system refers to the mean equator and equinox (compare to mean Sun concept).
The Julian standard epochs replace the Bessel year, which is defined as the revolution of the fictitious mean Sun, and starts when the (geocentric) right ascension is 18 hours and 40min.

For practical purposes, the Bessel year can be set equal to the tropical year. Bessel years get the prefix "B". Often, the notation B1950 is used. The Julian date of the start of an arbitrary Bessel year Y_B is:

$$JD = 2415020.31352 + 365.242198781 \cdot (Y_B - 1900) \tag{4.76}$$

and hence:

B1950: JD 2 433 282.423 = Jan 0.923d, 1950

The Earth Mean Equator and Equinox of the J2000 (also called EME2000) is provided by the FK5 star catalog (Fricke et al, 1988), which provides positions and proper motions of 1500 stars for the epoch J2000 as the reference frame.

However, conceptual difficulties in the correct definition of the ecliptic and equinox (Kinoshita and Aoki, 1983) became overburdening.

Consequently, IAU decided in 1991 to establish the International Celestial Reference System (ICRS) and adopted it from 1998 onwards.

For a smooth transition to the new system, the ICRS axes are chosen in such a way as to be consistent with the previous FK5 system within the accuracy of FK5/J2000.0.

The fundamental plane of the ICRS is closely aligned with the mean Equator at J2000.0 and the origin of the right ascension is defined by an adopted right ascension of quasar 3C273.

For all practical purposes, the better definition of ICRS can be taken advantage of as a replacement for J2000. This is often done, when people nevertheless refer to J2000 instead of ICRS.

The origin of the ICRS is defined as the solar system barycenter within a relativistic framework and its axis are fixed with respect to distant extragalactic radio objects using the VLBI. Links to existing optical catalogs are provided by radio stars (Seidelmann 1998).

The practical realization of the ICRS, the ICRF international Celestial Reference Frame is jointly maintained by the IERS and the IAU Working Group on Reference Frames.

The Geocentric Celestial Reference Frame (GCRF) is the counterpart of the ICRF, but with the Earth as its origin. It is our realization of the Earth Centered Inertial (ECI) reference frame.

Complementary to the GCRS, the International Terrestrial Reference System (ITRS) is defined as an Earth centered Earth fixed reference system (ECEF) (McCarthy, 1996).

Its origin is located at the Earth's center of mass (including oceans and atmosphere) and its unit of length is SI meter.

The orientation of the IERS Reference Pole (IRP) and Meridian (IRM) are consistent with the previously adopted BIH system at epoch 1984.0 and the former Conventional Origin (CIO).

The time evolution of the ITRS is such that it exhibits no net rotation with respect to the Earth's crust.

Realizations of the ITRS are given by the International Terrestrial Reference Frame (ITRF) that provides estimated coordinates and velocities of selected observing stations under authority of the IERS.

It is determined using observation techniques of the satellites laser ranging (SLR) and lunar laser ranging (LLR), GPS and VLBI measurements.

The transformation we are interested in is the one from the ITRS, which is our realization of a precise ECEF system, to the GCRS, which for all practical purposes is the same as J2000.0 and our precise realization of an ECI system, with the Earth as a reference point. The transformation takes the following perturbations and classical models for it into account:

- precession (Lieske at al, 1977), describing the secular change in the orientation of the Earth's rotation axis and the equinox. Those variations are caused by the presence of other solar system planets, the Sun and the Moon. They exert a torque on the Earth's rotation axis, leading to a gyroscopic motion of the Earth's rotation axis around the pole of the ecliptic with a period of 26 000 years. As a result, the vernal equinox recedes slowly on the ecliptic.
- nutation (Seidelmann 1982), describing the periodic and short-term variation of the equator and the vernal equinox. Those periodic variations are caused by the variations of the solar and lunar torques on time scales larger than a month.
- Sidereal Time in relation to UT1 (Aoki et al, 1982), describing the Earth's rotation about its axis.
- polar motion

Those models are supplemented by the IERS Earth Observation Parameters (EOP), comprising of observations of the UT1-TAI difference and the measured coordinates of the rotation axis relative to the IERS reference pole. This means:

$$\boldsymbol{r}_{GCRS/J2000.0} = \boldsymbol{P}^T(t)\boldsymbol{N}^T(t)\boldsymbol{\theta}^T(t)\boldsymbol{\Pi}^T(t)\boldsymbol{r}_{ITRS} \tag{4.77}$$

$$\boldsymbol{r}_{ITRS} = \boldsymbol{\Pi}(t)\boldsymbol{\theta}(t)\boldsymbol{N}(t)\boldsymbol{P}(t)\boldsymbol{r}_{GCRS/J2000.0} \tag{4.78}$$

for $\boldsymbol{P}$ describing precession,
$\boldsymbol{N}$ nutation,
$\boldsymbol{\theta}$ Earth rotation (sidereal time of the Greenwich),
$\boldsymbol{\Pi}$ polar motion
coordinate changes, respectively.

Light travel time, refraction, and aberration corrections are applied prior to any reference frame transformations.

Those transformations are required when dealing with real observations.

For more precise modeling, the Earth tidal (solid and tidal) motion and plate motion need to be taken into account.

The first can cause a displacement of the observation station of around 25mm with a daily period, and around 7mm in horizontal direction. Tectonic plates shift around 5cm or more per year.

For proof of concept theoretical calculations in the absence of real observations, the transformation:

$$\boldsymbol{r}_{ECI}(t) = \boldsymbol{\theta}^T(t)\boldsymbol{r}_{ECEF}(t) \tag{4.79}$$

$$\boldsymbol{r}_{ECEF}(t) = \boldsymbol{\theta}(t)\boldsymbol{r}_{ECI}(t) \tag{4.80}$$

where $\boldsymbol{\theta}$ represents the rotation matrix around the Earth rotation axis with the hour angle θ (true or mean). The only decision one has to take is to transform between the true equator and vernal equinox to the mean one. For example, this is relevant when transforming TLEs, which are mean equator, true of date.

4.3.1 Transformations for J2000.0/ICRS

References: [51, 49, 50].

4.3.2 Precession and Nutation

4.3.2.1 Lunisolar Torques and The Motion of the Earth's Rotation Axis

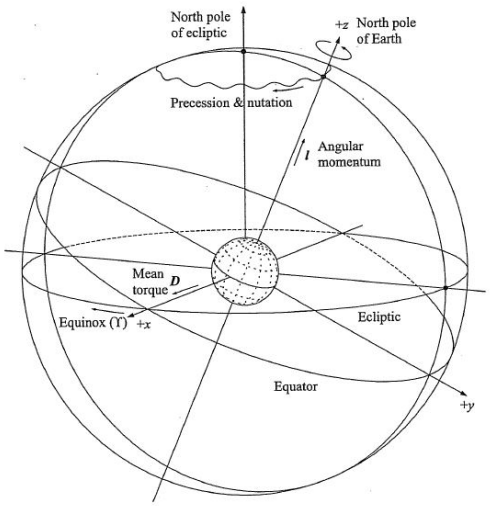


Figure 4.18: Motion of the Earth's axis influenced by solar and lunar torques.

In order to describe the precession of the Earth's rotation axis, the Earth is considered as a rotationally symmetric gyroscope with an angular momentum $\boldsymbol{l}$ that changes with time under the influence of an external torque $\boldsymbol{D}$ according to $\frac{d\boldsymbol{l}}{dt}$. In general, the symmetry axis of the gyroscope and the instantaneous axis of rotation may differ, this difference is neglected for the current purpose. Hence, the angular momentum $\boldsymbol{l}$ is assumed to be parallel to the unit vector $\boldsymbol{e}_z$ that defines the Earth's axis:

$$\boldsymbol{l} = I\omega\boldsymbol{e}_z, \tag{4.81}$$

where $\omega \approx 7.29 \cdot 10^{-5} rad/sec$ is the angular velocity of Earth rotation, and I is the moment of inertia.

For a spherical body with homogeneous mass density of mass M, and the Earth radius R, the moment of inertia is given for:

$$I_{\text{sphere}} = \frac{2}{5}MR^2 = 0.4 \cdot MR^2, \tag{4.82}$$

for arbitrary rotation (spherical symmetry). However, as we know the Earth is not a perfect sphere and the mass distribution is non-uniform. If the mass distribution and the flattening of the Earth is taken into account the moments of inertia for a rotation around an axis in the equatorial plane I_{eq} and rotation around the polar axis I are given by the values.

$$I_{\text{eq}} = 0.329 \cdot MR^2 \qquad I = 0.330 \cdot MR^2 \tag{4.83}$$

Note: Later in this lecture we will talk about the Earth gravitational field expansion, the connection to the termJ_{20} is defined as:

$$I - I_{\text{eq}} = J_{20}MR^2. \tag{4.84}$$

After finding the moments of inertia, the torque $\boldsymbol{D}$ has to be determined. The torque due to a point Mass m (e.g. Sun or moon) at a geocentric position $\boldsymbol{r}$ is given by:

$$\boldsymbol{D} = -m(\boldsymbol{r} \times \ddot{\boldsymbol{r}}). \tag{4.85}$$

$\ddot{\boldsymbol{r}}$ is the acceleration of the mass m by the gravitational force of the Earth. As we will see later in class, the Earth gravitational field can be to first order be expressed as:

$$\ddot{\boldsymbol{r}} = -\frac{GM}{r^3}\boldsymbol{r} - \frac{3}{2}\frac{GMR^2J_{20}}{r^7}\left((5(\boldsymbol{r}\boldsymbol{e}_z)^2 - r^2)\boldsymbol{r} - 2(r^2\boldsymbol{r}\boldsymbol{e}_z)\boldsymbol{e}_z\right). \tag{4.86}$$

The terms $\boldsymbol{r}\boldsymbol{e}_z$ denotes the distance of the attracted mass from the equatorial plane. If plugged back into Eq. 4.85, only the last term does not cancel out. If we also substitute Eq.4.84, we can find:

$$\boldsymbol{D} = Gm(I - I_{\text{eq}})\frac{3z(\boldsymbol{r} \times \boldsymbol{e}_z)}{r^5}, \tag{4.87}$$

with $z = \boldsymbol{r}\boldsymbol{e}_z$.

The Sun appears to move around the Earth in a near-circular orbit that is inclined and the angle ε when the Sun crosses seemingly the equator (z=0). When $\boldsymbol{e}_x$ is pointing towards vernal equinox, the torque created by the Sun at right angles to vernal equinox direction can be expressed by the following:

$$\boldsymbol{D}_{\text{perptoequi,sun}} = GM_{\text{Sun}}(I - I_{\text{eq}})\frac{3\sin\varepsilon\cos\varepsilon}{r^3}\boldsymbol{e}_x. \tag{4.88}$$

If we are integrating for arbitrary directions to the vernal equinox the net expression amounts to:

$$\bar{\boldsymbol{D}}_{\text{sun}} = GM_{\text{Sun}}(I - I_{\text{eq}})\frac{3\sin\varepsilon\cos\varepsilon}{2r^3}\boldsymbol{e}_x, \tag{4.89}$$

because $\int_0^{\pi/4}\sin\varepsilon\cos\varepsilon d\varepsilon = \frac{1}{2}$ and the net direction amounts to $\boldsymbol{e}_x$ as the orthogonal direction cancels out. Using Kelper's third law (mean motion $n = \sqrt{GM/a^3}$ with the semi-major axis of a) under the assumption of the circular motion (r=a at all times), the torque can be expressed as:

$$\bar{\boldsymbol{D}}_{\text{sun}} = \frac{3}{2}(I - I_{\text{eq}})\sin\varepsilon\cos\varepsilon n_{\text{Sun}}^2\boldsymbol{e}_x \tag{4.90}$$

In case of the Moon, the lunar orbit with respect to the Earth is not fixed. In fact, it varies between inclinations of 18 to 28 degrees with a period of 18 years. As an approximation we assume that the moon also moves along the ecliptic. This is justified because the precession motion has a much larger period. This leads to a net torque:

$$\bar{\boldsymbol{D}} = \frac{3}{2}(I - I_{\text{eq}})\sin\varepsilon\cos\varepsilon\left(n_{\text{Sun}}^2 + n_{\text{moon}}^2\right)\boldsymbol{e}_x, \tag{4.91}$$

where n_{moon} is the mean motion of the moon. Because the net torque is in the $\boldsymbol{e}_x$ direction, it changed neither the Earth's total angular momentum nor the obliquity ε but it forces L to move around the pole of the ecliptic with an angular velocity which can be found as:

$$\Omega_{\text{prec}} = \frac{|\bar{\boldsymbol{D}}|}{\sin(\varepsilon)|\boldsymbol{l}|} = \frac{3}{2}\frac{C - A}{C}\cos\varepsilon\frac{n_{\text{Sun}}^2 + n_{\text{moon}}^2(\frac{M_{\text{Moon}})}{M_{\text{Earth}}}}{\omega_{\text{Earth}}} \tag{4.92}$$

which just follows the standard definitions from any gyro system.

4.3.2.2 Coordinate Changes due to Precession

As a net effect, the influence of precession affects the orientation of the ecliptic as seen from the Earth and the equator. As investigated before, the true and the mean equator account for short periodic effects. In addition, it is necessary to

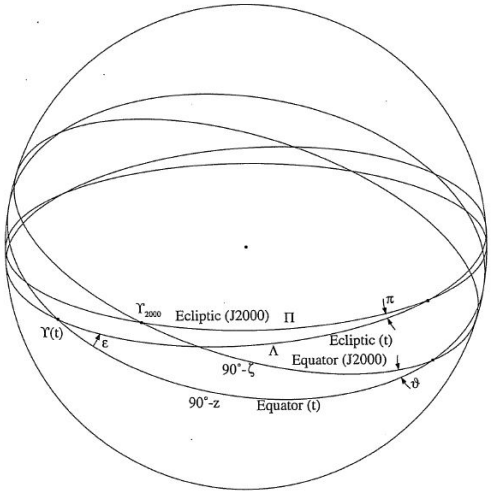


Figure 4.19: mean vernal equinox and equator and reference J2000 equator and equinox [51].

define a reference epoch, such as J2000.0. In Fig.4.19, the motion of the ecliptic and the equator are shown with respect to the mean equator and reference ecliptic at J2000.0

Due to luni-solar precession the intersection of the mean equator of epoch t and the mean ecliptic of J2000 lags behind the vernal equinox of J2000 by the following angle:

$$\psi_{\text{lunisolar-only}} = 5038.8'' \cdot T - 1.1'' \cdot T^2, \tag{4.93}$$

where T is again $T = (JD - 2451545.0)/36525.0$, and is evaluated in terrestrial time (TT), it is the time measured since J2000 TT. It can be seen that the angle ψ increases almost linearly with time. The inclination of the mean equator with respect to the ecliptic if J2000 is nearly constant:

$$\varepsilon_{\text{lunisolar-only}} = 23^\circ 26' 21'' + 0.05'' \cdot T^2. \tag{4.94}$$

The torque from the Sun and the moon do not change the ecliptic but only change the orientation of the Earth. However, the ecliptic itself is also not stable. The orbit of the Earth itself is perturbed by the influence of other plants, leading to small shifts in the ecliptic itself. Relative to J2000 the ecliptic is inclined by:

$$\pi = 47.0029'' \cdot T - 0.03302'' \cdot T^2 + 0.000060'' \cdot T^3 \tag{4.95}$$

The values follow the theory by Lieske et at (1977). This means the ecliptic can be computed as:

$$\varepsilon = 23.43929111^\circ - 46.8150'' \cdot T - 0.00059'' \cdot T^2 + 0.001813'' \cdot T^3 \tag{4.96}$$

The combined precession of the longitude is:

$$P = 5029.0966'' \cdot T + 1.11113 \cdot T^2 + 0.000006'' \cdot T^3 \tag{4.97}$$

For the orientation of the mean equator and equinox of epoch T in relation to mean equator and equinox if J2000 can be defined via the three Euler angles:

$$\zeta = 2306.2181'' \cdot T + 0.30188'' \cdot T^2 + 0.017998'' \cdot T^3 \tag{4.98}$$

$$\theta = 2004.3109'' \cdot T - 0.42665'' \cdot T^2 - 0.041833'' \cdot T^3 \tag{4.99}$$

$$z = \zeta + 0.79280'' \cdot T^2 + 0.000205'' \cdot T^3, \tag{4.100}$$

which rely on the fundamental quantities, π, ψ, and ε. The transformation from the state $\boldsymbol{r}_{\text{GCRF}}$ in coordinates GCRF (same as J2000.0 within the uncertainties of the J2000 system) to the state in the system mean equator, mean equinox of

another epoch (e.g. the epoch of our observations), also called mean of date, $\boldsymbol{r}_{\text{mod}}$ is then:

$$\boldsymbol{r}_{\text{mod}} = \boldsymbol{P}\boldsymbol{r}_{\text{GCRF}}, \tag{4.101}$$

$$\boldsymbol{P} = \boldsymbol{R}_z(-z)\boldsymbol{R}_y(\theta)\boldsymbol{R}_z(-\zeta) \tag{4.102}$$

with the elements:

$$p_{11} = -\sin z \sin\zeta + \cos z \cos\theta \cos\zeta \tag{4.103}$$

$$p_{21} = \cos z \sin\zeta + \sin z \cos\theta \cos\zeta \tag{4.104}$$

$$p_{31} = \sin\theta \cos\zeta \tag{4.105}$$

$$\tag{4.106}$$

$$p_{12} = -\sin z \cos\zeta - \cos z \cos\theta \sin\zeta \tag{4.107}$$

$$p_{22} = \cos z \cos\zeta - \sin z \cos\theta \sin\zeta \tag{4.108}$$

$$p_{32} = -\sin\theta \sin\zeta \tag{4.109}$$

$$p_{13} = -\cos z \sin\theta \tag{4.110}$$

$$p_{23} = -\sin z \sin\theta \tag{4.111}$$

$$p_{33} = \cos\theta \tag{4.112}$$

$\boldsymbol{P}$ is an orthonormal matrix, hence its inverse and transpose are identical, using the product rule it leads to:

$$\boldsymbol{P} = \boldsymbol{R}_z(z)\boldsymbol{R}_y(-\theta)\boldsymbol{R}_z(\zeta) \tag{4.113}$$

The question that now remains open is, how is the transformation between two arbitrary epochs T_1 and T_2 performed, that is between two epochs mean of date. The key is that one transforms the first state vector back to J2000 and then forward to the mean of date of the second epoch again.

$$\boldsymbol{r}_{2,\text{mod}} = \boldsymbol{P}(T_2)\boldsymbol{P}^T(T_1)\boldsymbol{r}_{1,\text{mod}} \tag{4.114}$$

One has to keep in mind that numerical errors accumulate in the subsequent use of $\boldsymbol{P}$ matrices.

It is hence more beneficial to transform every new state into J2000.0 and then to any new epoch in one step (e.g. T_2, T_3) rather than transforming from T_2 directly to T_3 and so on.

This derivation is in agreement with the IAU 1976 theory of precession.

EXAMPLE: Precession
Transform an arbitrary but fixed state r from ICRS/J2000 (mean equator, mean equinox) to the ITRS, true of date at march 4 1999, 0h UTC.

In order to know the transformation we can either look up the time relations here in the script or at IERS Bulletins B (No. 135) and C (No. 16) in order to get: UTC-TAI=-32.0 sec
TT-UTC=64.184 sec
UT1-UTC=0.649232 sec
Step 1: Transforming the time into TT and computing the matrix P results in:

$$\boldsymbol{P} = \begin{pmatrix} 0.99999998 & 0.00018581 & 0.00008074 \\ -0.00018581 & 0.99999998 & -0.00000001 \\ -0.00008074 & -0.00000001 & 1.00000000 \end{pmatrix} \tag{4.115}$$

4.3.2.3 Nutation: Or transformation from mean equator, equinox to true equator, equinox

Besides the secular effects, small periodic motion of the Earth's rotational axis are called nutation.

They are due to monthly and annual variations of the lunar and solar torque. In the treatment of the precession we were referring to the mean equator and mean equinox, mean of date.

Now we want to investigate how we get to the true of date, true equinox, and true equator.

The **"true"** system is the one in which we are observing as we are fixed on the Earth surface (no motion relative to the Earth's crust).

The nutation also allows to compute the true and the mean sidereal time for any position on the Earth surface, see Eq.4.16.

The main contribution to the nutation stems from the varying orientation of the lunar orbit with respect to the Earth equator, that can be expressed via the longitude of the Moon's ascending node Ω.

The nodal period of the moon is 18.6 years. To first order the nutation can be computed as the following:

$$T = (JD - 2451545.0)/36525 \text{ centuries since J2000.0} \tag{4.116}$$
$$l = 357.525deg + 35999.049deg \cdot T \text{ mean anomaly of the Sun} \tag{4.117}$$
$$F = 93.273deg + 483202.019deg \cdot T \text{ mean distance between the nodes of the moon} \tag{4.118}$$
$$D = 297.850deg + 445267.111deg \cdot T \text{ mean distance Sun to moon} \tag{4.119}$$
$$\Omega = 125.045deg - 1934.136deg \cdot T \text{ mean longitude of the moon} \tag{4.120}$$

all quantities are referenced to the vernal equinox of the date.

If we do a Taylor series expansion of the torques that are created, the nutation angles can be extracted:

$$\Delta\psi_{\text{approx}} = -17.200'' \sin(\Omega) + 0.202'' \sin(2\Omega) - 1.319'' \sin(2(F - D + \Omega)) + 0.143'' \sin(l) - 0.227'' \sin(2(F + \Omega)) \tag{4.121}$$
$$\Delta\varepsilon_{\text{approx}} = 9.203'' \cos(\Omega) - 0.090'' \cos(2\Omega) - 0.547'' \cos(2(F - D + \Omega)) + 0.098'' \cos(2(F + \Omega)), \tag{4.122}$$

where $\Delta\Psi$ is the longitude of the mean vernal equinox in relation to the true vernal equinox, and $\Delta\varepsilon$ the difference between the true and the mean obliquity of the ecliptic. These values can be used .e.g. to calculate the true sidereal time and the mean sidereal time as pointed out in Eq.4.16.

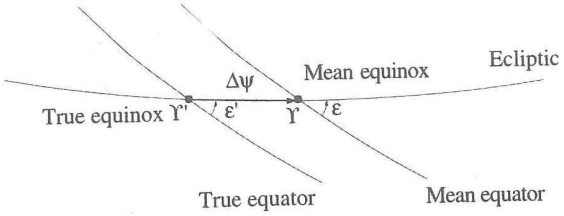


Figure 4.20: True equinox and equator and mean equator and equinox [51].

For higher accuracy and precision, the IAU1980 theory can be adopted, that allows a transformation to J2000. It is based on the theory of Kinoshita (1977) and Wahr (1981). It is a series evolution of 106 terms.

$$\Delta\psi = \sum_{i=1}^{106} (\Delta\psi)_i \sin(\phi_i) \tag{4.123}$$
$$\Delta\varepsilon = \sum_{i=1}^{106} (\Delta\varepsilon)_i \cos(\phi_i) \tag{4.124}$$
$$\phi_i = p_{l,i} \cdot l + p_{l',i} \cdot l' + p_{F,i} \cdot F + p_{D,i} \cdot D + p_{\Omega,i} \cdot \Omega \tag{4.125}$$

p_l	$p_{l'}$	p_F	p_D	p_Ω	$\Delta\Psi$ [0.0001"]	$\Delta\varepsilon$ [0.0001"]	i	p_l	$p_{l'}$	p_F	p_D	p_Ω	$\Delta\Psi$	$\Delta\varepsilon$	i
0	0	0	0	1	−171996−174.2T	+92025+8.9T	1	1	0	2	2	2	−8	3	54
0	0	0	0	2	2062 +0.2T	−895+0.5T	2	1	0	0	0	0	6	0	55
−2	0	2	0	1	46	−24	3	2	0	2−2	2		6	−3	56
2	0	−2	0	0	11	0	4	0	0	0	2	1	−6	3	57
−2	0	2	0	2	−3	1	5	0	0	2	2	1	−7	3	58
1	−1	0	−1	0	−3	0	6	1	0	2−2	1		6	−3	59
0	−2	2	−2	1	−2	1	7	0	0	0−2	1		−5	3	60
2	0	−2	0	1	1	0	8	1	−1	0	0	0	5	0	61
0	0	2	−2	2	−13187 −1.6T	5736−3.1T	9	2	0	2	0	1	−5	3	62
0	1	0	0	0	1426 −3.4T	54−0.1T	10	0	1	0−2	0		−4	0	63
0	1	2	−2	2	−517 +1.2T	224−0.6T	11	1	0	−2	0	0	4	0	64
0	−1	2	−2	2	217 −0.5T	−95+0.3T	12	0	0	0	1	0	−4	0	65
0	0	2	−2	1	129 +0.1T	−70	13	1	1	0	0	0	−3	0	66
2	0	0	−2	0	48	1	14	1	0	2	0	0	3	0	67
0	0	2	−2	0	−22	0	15	1	−1	2	0	2	−3	1	68
0	2	0	0	0	17 −0.1T	0	16	−1	−1	2	2	2	−3	1	69
0	1	0	0	1	−15	9	17	−2	0	0	0	1	−2	1	70
0	2	2	−2	2	−16 +0.1T	7	18	3	0	2	0	2	−3	1	71
0	−1	0	0	1	−12	6	19	0	−1	2	2	2	−3	1	72
−2	0	0	2	1	−6	3	20	1	1	2	0	2	2	−1	73
0	−1	2	−2	1	−5	3	21	−1	0	2−2	1		−2	1	74
2	0	0	−2	1	4	−2	22	2	0	0	0	1	2	−1	75
0	1	2	−2	1	4	−2	23	1	0	0	0	2	−2	1	76
1	0	0	−1	0	−4	0	24	3	0	0	0	0	2	0	77
2	1	0	−2	0	1	0	25	0	0	2	1	2	2	−1	78
0	0	−2	2	1	1	0	26	−1	0	0	0	2	1	−1	79
0	1	−2	2	0	−1	0	27	1	0	0	−4	0	−1	0	80
0	1	0	0	2	1	0	28	−2	0	2	2	2	1	−1	81
−1	0	0	1	1	1	0	29	−1	0	2	4	2	−2	1	82
0	1	2	−2	0	−1	0	30	2	0	0	−4	0	−1	0	83
0	0	2	0	2	−2274 −0.2T	977−0.5T	31	1	1	2	−2	2	1	−1	84
1	0	0	0	0	712 +0.1T	−7	32	1	0	2	2	1	−1	1	85
0	0	2	0	1	−386 −0.4T	200	33	−2	0	2	4	2	−1	1	86
1	0	2	0	2	−301	129−0.1T	34	−1	0	4	0	2	1	0	87
1	0	0	−2	0	−158	−1	35	1	−1	0	−2	0	1	0	88
−1	0	2	0	2	123	−53	36	2	0	2−2		1	1	−1	89
0	0	0	2	0	63	−2	37	2	0	2	2	2	−1	0	90
1	0	0	0	1	63 +0.1T	−33	38	1	0	0	2	1	−1	0	91
−1	0	0	0	1	−58 −0.1T	32	39	0	0	4−2		2	1	0	92
−1	0	2	2	2	−59	26	40	3	0	2−2		2	1	0	93
1	0	2	0	1	−51	27	41	1	0	2−2		0	−1	0	94
0	0	2	2	2	−38	16	42	0	1	2	0	1	1	0	95
2	0	0	0	0	29	−1	43	−1	−1	0	2	1	1	0	96
1	0	2−2		2	29	−12	44	0	0	−2	0	1	−1	0	97
2	0	2	0	2	−31	13	45	0	0	2−1		2	−1	0	98
0	0	2	0	0	26	−1	46	0	1	0	2	0	−1	0	99
−1	0	2	0	1	21	−10	47	1	0	2−2		0	−1	0	100
−1	0	0	2	1	16	−8	48	0	−1	2	0	1	−1	0	101
1	0	0	−2	1	−13	7	49	1	1	0	−2	1	−1	0	102
−1	0	2	2	1	−10	5	50	1	0	−2	2	0	−1	0	103
1	1	0	−2	0	−7	0	51	2	0	0	2	0	1	0	104
0	1	2	0	2	7	−3	52	0	0	2	4	2	−1	0	105
0	−1	2	0	2	−7	3	53	0	1	0	1	0	1	0	106

Figure 4.21: IAU 1980 nutation coefficients [51].

the coefficients can be found in the Table in Fig.4.21. For the parameters l, l', F, D, Ω the higher order terms are included:

$$T = (JD(TT) - 2451545.0)/36525 \tag{4.126}$$
$$l = 134^\circ 57' 46.733'' + 477198^\circ 52' 02.633'' \cdot T + 31.310'' \cdot T^2 + 0.064'' T^3 \tag{4.127}$$
$$l' = 357^\circ 31' 39.804'' + 35999^\circ 03' 01.244'' \cdot T - 0.577'' \cdot T^2 - 0.012'' T^3 \tag{4.128}$$
$$F = 93^\circ 16' 18.877'' + 483202^\circ 01' 03.137'' \cdot T - 13.257'' \cdot T^2 + 0.011'' T^3 \tag{4.129}$$
$$D = 297^\circ 51' 01.307'' + 445267^\circ 06' 41.328'' \cdot T - 6.891'' \cdot T^2 + 0.019'' T^3 \tag{4.130}$$
$$\Omega = 125^\circ 02' 40.280'' - 1934^\circ 08' 10.539'' \cdot T + 7.455'' \cdot T^2 + 0.008'' T^3 \tag{4.131}$$

The transformation of the mean of date coordinates (mean equator, mean equinox) and the true of date coordinates (true equator and true equinox), can be written as, for a state $\boldsymbol{r}_{\text{tod}}$ and $\boldsymbol{r}_{\text{mod}}$, respectively:

$$\boldsymbol{r}_{\text{tod}} = \boldsymbol{N}(t)\boldsymbol{r}_{\text{mod}} \tag{4.132}$$
$$\boldsymbol{N} = \boldsymbol{R}_x(-\varepsilon - \Delta\varepsilon)\boldsymbol{R}_z(-\Delta\psi)\boldsymbol{R}_x(\varepsilon) \tag{4.133}$$

with the following elements

$$n_{11} = \cos\Delta\psi \tag{4.134}$$
$$n_{21} = \cos\varepsilon' \sin\Delta\psi \tag{4.135}$$
$$n_{31} = \sin\varepsilon' \sin\Delta\psi \tag{4.136}$$

$$n_{12} = -\cos\varepsilon \sin\Delta\psi \tag{4.137}$$
$$n_{22} = \cos\varepsilon \cos\varepsilon' \cos\Delta\psi + \sin\varepsilon \sin\varepsilon' \tag{4.138}$$
$$n_{32} = \cos\varepsilon \sin\varepsilon' \cos\Delta\psi - \sin\varepsilon \cos\varepsilon' \tag{4.139}$$

$$n_{13} = -\sin\varepsilon\sin\Delta\psi \tag{4.140}$$
$$n_{23} = \sin\varepsilon\cos\varepsilon'\cos\Delta\psi - \cos\varepsilon\sin\varepsilon' \tag{4.141}$$
$$n_{33} = \sin\varepsilon\sin\varepsilon'\cos\Delta\psi + \cos\varepsilon\cos\varepsilon' \tag{4.142}$$

$$\varepsilon' = \varepsilon + \Delta\varepsilon \tag{4.143}$$

From VLBI and LLR observations it is known that there is an error on the level of several milli-arcseconds in the IAU1980 theory. The IERS1996 sought to remedy these offsets but they can be neglected for practical purposes.

EXAMPLE: Nutation
For the reference date of date at March 4 1999, 0h UTC (true date), the nutation matrix takes the following form:

$$\boldsymbol{N} = \begin{bmatrix} 1.00000000 & 0.00004484 & 0.00001944 \\ -0.00004484 & 1.00000000 & 0.00003207 \\ -0.00001944 & -0.00003207 & 1.00000000 \end{bmatrix} \tag{4.144}$$

4.3.2.4 Polar Motion

Until now we defined the transformation between the true of date system (without beloved nutation and precession theory) and the Earth fixed system via the sidereal time only.

Hence, it is assumed the z-axis through the pole of the Earth and the ephemeris pole do coincide at all time and is fixed relative to the Earth crust.

This is however, not the case. The Earth pole performs relative to the projected celestial pole a periodic motion around its position, which differs up to 10m.

This is known as polar motion and can be understood by assuming a rotationally symmetric gyroscope, in which the rotation axis moves around the axis of the figure in the absence of external torques.

4.3.2.5 Free Eulerian Motion

Assume a body fixed system $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$, that is aligned with the principal axis of inertia. The angular momentum $\boldsymbol{l}'$ of a symmetric gyroscope is given by:

$$\boldsymbol{l}' = \begin{bmatrix} A & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & C \end{bmatrix} \cdot \boldsymbol{\omega}, \tag{4.145}$$

where $\boldsymbol{\omega}$ is the instantaneous rotation axis and where A and C are the moments of inertia for a rotation around the $\boldsymbol{e}_i$ axis.

Without external torques (we already have taken care of those), the angular momentum $\boldsymbol{l}$ is constant in an inertial reference frame, but since $\boldsymbol{l}'$ refers to a rotating system, we know the famous relation to transport theorem or basic kinematic equations (BKE):

$$\frac{d\boldsymbol{l}}{dt} = \frac{d\boldsymbol{l}'}{dt} + \boldsymbol{\omega} \times \boldsymbol{l}' = \boldsymbol{0}, \tag{4.146}$$

because there is no change in angular momentum. In turn that means:

$$A\frac{d\omega_1}{dt} + (C-A)\omega_2\omega_3 = 0, \tag{4.147}$$

$$A\frac{d\omega_2}{dt} + (C-A)\omega_1\omega_3 = 0, \tag{4.148}$$

$$A\frac{d\omega_3}{dt} = 0. \tag{4.149}$$

The last equation just implies a constant rotation around the $\boldsymbol{E}_3$ axis, the first equation the first and second equations lead to:

$$\omega_1 = a\cos(\frac{C-A}{A}\omega_3 t + b), \tag{4.150}$$

$$\omega_1 = a\sin(\frac{C-A}{A}\omega_3 t + b). \tag{4.151}$$

Hence, referring to an instantaneous circle around the $\boldsymbol{e}_3$ axis. The period is:

$$P = \frac{2\pi \cdot A}{\omega_3 \cdot (C-A)}. \tag{4.152}$$

It depends on the difference of the moments. The earth has a dynamical flattening, which leads to a period of 305 days.

Observations show that the polar motion is actually a superposition of two periods, one free precession with a period of around 435 days (the so-called Chandler period), which can only be explained by a non-rigid Earth models.

The second period is an annual motion that is induced by seasonal changes of the Earth's mass distribution due to air and water flows. It leads to a beat period of around 6 years. In contrast to nutation and precession, no exact mathematical model exists and we heavily rely on observations.

The matrix for the polar motion can be defined as:

$$\mathbf{\Pi} = \boldsymbol{R}_y(-x_p)\boldsymbol{R}_x(-y_p) \approx \begin{pmatrix} 1 & 0 & x_p \\ 0 & 1 & -y_p \\ -x_p & y_p & 1 \end{pmatrix}. \tag{4.153}$$

Only first order terms were taken into account. For the parameters x_p and y_p look-up tables do exist, provided by the IERS bulletins and EOP parameter files (e.g. from JPL).

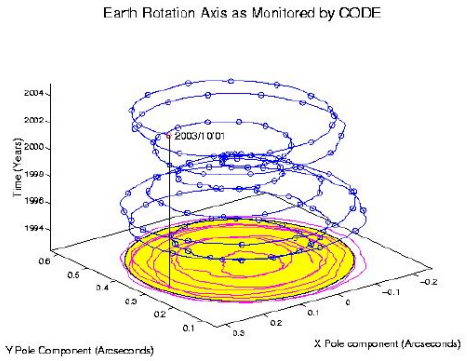


Figure 4.22: Polar motion around the IERS reference celestial pole.

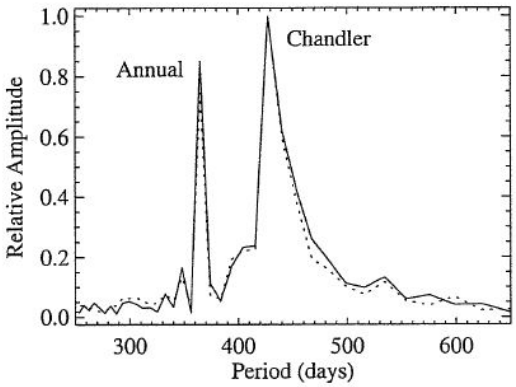


Figure 4.23: Polar motion frequency spectrum [51].

Chapter 5

Probability of Collision

An object impacting at 3 km/sec delivers kinetic energy equal to its mass in TNT. stated Rick Robinson in his so-called First Law of Space Combat [19].

5.1 Problem Setup: Probability Density Functions and Hardbody Radius

Collisions are, traditionally, computed on the two object level. Consider two objects, A and B, which may be defined in carthesian coordiantes via their probability density functions. Often, the probability density function (pdf) is represented via their first two moments, mean 6×1 state vectors $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$

$$\boldsymbol{x}_A = \begin{bmatrix} \boldsymbol{r}_A \\ \boldsymbol{v}_A \end{bmatrix} = \begin{bmatrix} x_A \\ y_A \\ z_A \\ \dot{x}_A \\ \dot{y}_A \\ \dot{z}_A \end{bmatrix} \qquad \boldsymbol{x}_B = \begin{bmatrix} \boldsymbol{r}_B \\ \boldsymbol{v}_B \end{bmatrix} = \begin{bmatrix} x_B \\ y_B \\ z_B \\ \dot{x}_B \\ \dot{y}_B \\ \dot{z}_B \end{bmatrix}, \tag{5.1}$$

and covariance 6×6, $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$:

$$\boldsymbol{P}_A = \begin{bmatrix} \sigma^2_{x_A} & \sigma_{x_A y_A} & \sigma_{x_A z_A} & \sigma_{x_A \dot{x}_A} & \sigma_{x_A \dot{y}_A} & \sigma_{x_A \dot{z}_A} \\ \sigma_{y_A x_A} & \sigma^2_{y_A} & \sigma_{y_A z_A} & \sigma_{y_A \dot{x}_A} & \sigma_{y_A \dot{y}_A} & \sigma_{y_A \dot{z}_A} \\ \sigma_{z_A x_A} & \sigma_{z_A y_A} & \sigma^2_{z_A} & \sigma_{z_A \dot{x}_A} & \sigma_{z_A \dot{y}_A} & \sigma_{z_A \dot{z}_A} \\ \sigma_{\dot{x}_A x_A} & \sigma_{\dot{x}_A y_A} & \sigma_{\dot{x}_A z_A} & \sigma^2_{\dot{x}_A} & \sigma_{\dot{x}_A \dot{y}_A} & \sigma_{\dot{x}_A \dot{z}_A} \\ \sigma_{\dot{y}_A x_A} & \sigma_{\dot{y}_A y_A} & \sigma_{\dot{y}_A z_A} & \sigma_{\dot{y}_A \dot{x}_A} & \sigma^2_{\dot{y}_A} & \sigma_{\dot{y}_A \dot{z}_A} \\ \sigma_{\dot{z}_A x_A} & \sigma_{\dot{z}_A y_A} & \sigma_{\dot{z}_A z_A} & \sigma_{\dot{z}_A \dot{x}_A} & \sigma_{\dot{z}_A \dot{y}_A} & \sigma^2_{\dot{z}_A} \end{bmatrix} \tag{5.2}$$

$$\boldsymbol{P}_B = \begin{bmatrix} \sigma^2_{x_B} & \sigma_{x_B y_B} & \sigma_{x_B z_B} & \sigma_{x_B \dot{x}_B} & \sigma_{x_B \dot{y}_B} & \sigma_{x_B \dot{z}_B} \\ \sigma_{y_B x_B} & \sigma^2_{y_B} & \sigma_{y_B z_B} & \sigma_{y_B \dot{x}_B} & \sigma_{y_B \dot{y}_B} & \sigma_{y_B \dot{z}_B} \\ \sigma_{z_B x_B} & \sigma_{z_B y_B} & \sigma^2_{z_B} & \sigma_{z_B \dot{x}_B} & \sigma_{z_B \dot{y}_B} & \sigma_{z_B \dot{z}_B} \\ \sigma_{\dot{x}_B x_B} & \sigma_{\dot{x}_B y_B} & \sigma_{\dot{x}_B z_B} & \sigma^2_{\dot{x}_B} & \sigma_{\dot{x}_B \dot{y}_B} & \sigma_{\dot{x}_B \dot{z}_B} \\ \sigma_{\dot{y}_B x_B} & \sigma_{\dot{y}_B y_B} & \sigma_{\dot{y}_B z_B} & \sigma_{\dot{y}_B \dot{x}_B} & \sigma^2_{\dot{y}_B} & \sigma_{\dot{y}_B \dot{z}_B} \\ \sigma_{\dot{z}_B x_B} & \sigma_{\dot{z}_B y_B} & \sigma_{\dot{z}_B z_B} & \sigma_{\dot{z}_B \dot{x}_B} & \sigma_{\dot{z}_B \dot{y}_B} & \sigma^2_{\dot{z}_B} \end{bmatrix} \tag{5.3}$$

Often, a Gaussian assumption is made, although we have already seen that for longer propagation times this assumption can be problematic. Each object also has a physical extension, which is crucial in defining a collision. Usualy, the assumption is made that the object extension is defined via the radius of a sphere encompassing the entire object, defined as the so-called **Hard Body Radius (HBR)**, ρ_A and ρ_B. It is possible to use other shapes (see Patera [57]), but the spherical approximation is normally used for the sake of simplicity.

Obviously if there was no uncertainty in the state vectors it would be easy to compute PC. We would just propagate the

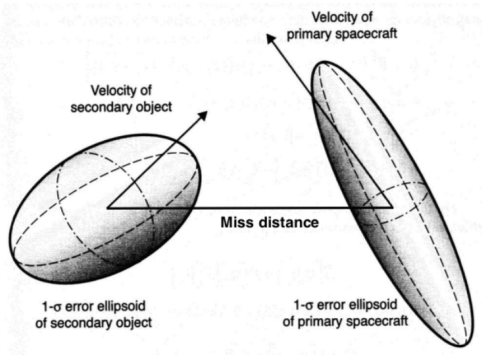


Figure 5.1: Two objects means with Gaussian one sigma error ellipsoids Chan (2008).

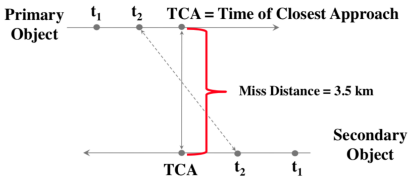


Figure 5.2: Illustration of TCA, PCA, Miss Distance, and Conjuction, credits NASA CARA.

states of A and B forward, and if $||\boldsymbol{r}_A - \boldsymbol{r}_B|| \leq \sqrt{\rho_A^2 + \rho_B^2}$ at any time then a collision will occur and PC = 1. In reality we often do not know the positions of objects in space this precisely, so we need a method that takes uncertainty into account.
In the presence of uncertainty, only a probability of collision can be defined. But first, we need to clarify terminology.

Time of Closest Approach (TCA): The time at which the means of the two object pdfs are closest.

Point of Closest Approach (PCA): The position of the closest approach of the means of the two object pdfs (at TCA).

Miss Distance: The distance between the positions of the means of the two object pdfs.

Probability of Collision: Statistical measure of the likelihood that the objects are within each other's hard body radius. Thresholds for the probability of collision that are currently in use are 10^{-6} for uncrewed missions and 10^{-8} for crewed missions.

Conjunction: When the predicted miss distance (defined above) is less than a specified threshold OR the probability of collision exceeds the defined threshold.

Short encouters: Two encountering objects are moving quickly relative to each other, this means either a large relative velocity difference or that the velocity of two objects are close to or in the vicinity of a 90 degree angle. The time for one object to pass through the one sigma region of the combined covariance is very short relative to the orbital periods of both objects involved.

Long Encouters: Two encountering objects move slow or not at all, relative to each other, this means either a small relative velocity difference or the velocity vectors of the two objects are at a significantly acute angle. The time for one object to pass through the one sigma region of the combined covariance is long relative to the orbital periods of both objects involved.

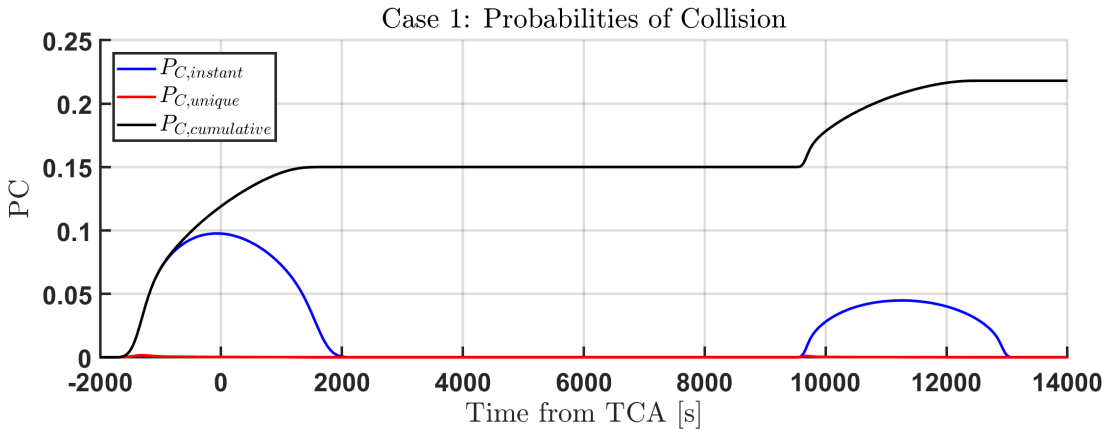


Figure 5.3: Illustrates the relations of three different types of PC. Taken from [14].

5.2 Three Kinds of Probability of Collisions

When discussing probabilities of collision we can use PC to refer to three different quantities. The first is the probability that the two objects are currently in contact at a given point in time. Call this the instantaneous probability of collision, PC_{inst}. This is subtly different from the second quantity, which is the probability that the two objects are just comining into contact at this point in time. This second type of PC is the "unique" PC or PC_{unique}. The name comes from how PC_{unique} is estimated in the Monte Carlo method.

The final quantitiy is the probability that a collision will occur at some point during the encounter. This is cumulative probability of collision, PC_{cum}. Usually when the term "probability of collision" is used it refer to PC_{cum}, since it is the most practically useful of the three.

5.2.1 An Illustration

Figure 5.3 is a plot of the relations betwen the three kinds of PC over the course of an encounter in GEO. The black line is the cumulative PC up to that point in the encounter. Notice that PC_{cum} never decreases, since it is the probability that a collision has occured prior to time t. The blue line is the instantaneous PC, which is always equal to or less than the cumulative PC. By inspection, we can see that PC_{cum} is not the integral of PC_{inst}, since the black curve levels out well before $t = 2000sec$, when PC_{inst} goes to zero.

This leaves us with the red line which plots the unique probability of collision. The PC_{unique} values are much smaller than PC_{inst}, because it counts each possible collision event only at one time step, rather than over an extended period. We get PC_{cum} by integrating PC_{unique}.

5.2.2 Monte Carlo Simulations

Another way to explain the three kinds of PC is from how they are computed in Monte Carlo simulations. For the simulation we populate the state-space distributions for object A and object B with $n_{\text{sample,A}}$ and $n_{\text{sample,B}}$, respectively. Each particle is propagated forward using standard dynamics equations for orbital motion. At each time step we test to see how far each particle from the set corresponding to object A is from each particle in set B. If the distance is less than $\rho_A + \rho_B$ then the two particles are colliding.

Depending on how we handle the colliding particles we can compute either PC_{inst} or PC_{unique} at the time step. If we simply count up the number of colliding particle pairs, $n_{\text{collide,pair}}$, and divide by the number of possible colliding pairs, then we will get PC_{inst}.

$$PC_{inst} = \frac{n_{collide,pair}}{n_{\text{sample,A}} \cdot n_{\text{sample,B}}} \tag{5.4}$$

We can also get PC_{unique} from the Monte Carlo method. We do this by tracking which particles have collided with each other in previous time steps, and excluding those combinations from our PC calculations. For example, suppose that at

t_k particle 1 from set A is colliding with particles 2 and 3 from set B. For the PC_{inst} calculation we would say that two collisions are occuring, $n_{collide,pairs} = 2$. However if we know that particle 1 had already collided with particle 2 at t_{k-1}, then for the PC_{unique} calculation we would count only one new (or unique) collision occuring at t_k. So $n_{unique} = 1$ and

$$PC_{unique} = \frac{n_{unique,pairs}}{n_{\text{sample,A}} \cdot n_{\text{sample,B}}} \tag{5.5}$$

Finally, we compute PC_{cum} by summing PC_{unique} over the course of the simulation.

5.3 Computing the PC

The Monte Carlo method accurately captures the probability of collision for all kinds of encounters, but requires a large number of particles to be useful. Therefore, it is generally more practical to use analytic methods to approximate the cumulative PC.

5.3.1 Useful Simplifications

In order to make the problem more solvable, some common simplifications are made. First, we will be looking at the relative motion of B with respect to A.

Second, all of the position and velocity uncertainty is put on object A, and the state of B is assumed to be perfectly known. This keeps the uncertainty in the relative position and velocity of the objects the same, but means that we only need to keep track of one probability distribution rather than two. The covariance matrix $\boldsymbol{P}$ of the new distribution, centered on A, is given below.

$$\boldsymbol{P} = \boldsymbol{P}_A + \boldsymbol{P}_B \tag{5.6}$$

Finally, we reduce object A to a point and make B a sphere of radius $\rho = \rho_A + \rho_B$. This sphere is called the "combined hardbody" (often shortened to "hardbody") and ρ is the combined hardbody radius. A collision will occur if A is ever within ρ of B. This does not change the probability of collision results, but allows us to do computations over only one spherical volume rather than two.

5.3.2 Exact Cumulative PC

The most direct way to get the cumulative PC for an encounter is to integrate the density function of the distribution centered on A over the volume that the hardbody at B passes through. This is shown in (5.7) [4], where V is the volume swept out by the hardbody over the course of the encouter. Expressing the system in the eigenvalue space to diagonalize the combined covariance matrix:

$$\begin{aligned} x &= x_B - x_A \\ y &= y_B - y_A \\ z &= z_B - z_A \\ PC &= \frac{1}{\sigma_x \sigma_y \sigma_z \sqrt{8\pi^3}} \int\int\int_V \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}\right)\right] \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z \end{aligned} \tag{5.7}$$

The full equation is not usually used however, mostly because it is extremely difficult to set the integration bounds. This expression also does not take the evolution of the covariance matrix over time into account, which can be an important consideration for longer encounters.

5.3.3 Assuming a Linear Encounter

It can be easier to handle the problem if we assume that the encounter is linear. "Linear" encounters occur when the relative trajectory of B with respect to A does not turn. Chan [17] recommends looking at the trajectory of B as long as it remains within either 3σ (probably good enough) or 8.5σ (the limits of double-precision represenation) of A.

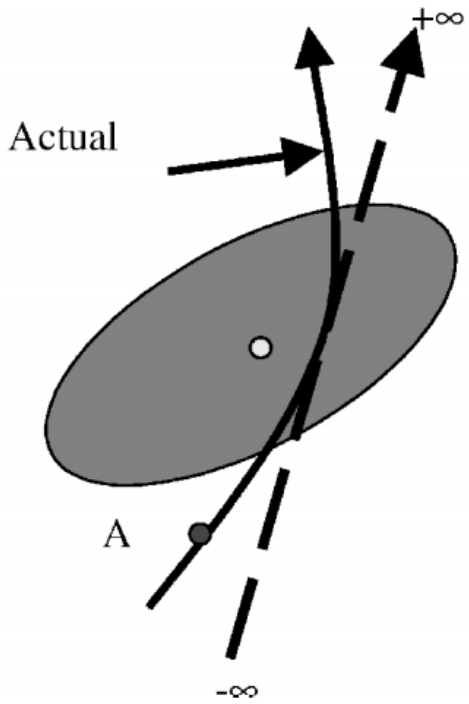


Figure 5.4: Illustrates the difference between a linear relative trajectory (dashed line) and a nonlinear trajectory (solid line). Taken from [56].

This generally only happens for relatively short encounters (on the order of a few seconds to a few minutes), when the satellites are moving quickly relative to each other. Longer encounters occur between slower-moving objects (say in geosynchronous orbit) and can last for several hours to over a day. In long encounters the relative trajectory of B curves and may self-intersect. Figure 5.4 illustrates the difference in the relative trajectory for linear and nonlinear encounters.

5.3.4 2D Approach

Since the full equation is hard to integrate, we often use a simpler approach to get PC_{cum}. The simplified approach is called the 2D PC approximation, because it projets the entire encounter onto a two-dimensional plane. In order for this method to work, we first assume that the encounter is linear, as described above, and short. Both of these assumptions are necessary. If the encounter is nonlinear then there is no single plane that is perpendicular to the entire trajector of B with respect to A, and the projection will not fully capture the motion of the satellite. If the encounter is long, then the covariance $\boldsymbol{P}$ will change significantly over the course of the encounter. This is important because for this method we focus on a single moment during the encounter, when A and B are closest to each other. This is the time of closest approach (TCA).

At TCA we project everything onto the encounter plane, which is the plane whose normal vector is the relative velocity vector $\boldsymbol{v} = \boldsymbol{v}_B - \boldsymbol{v}_A$. We start by defining a new reference frame using the U matrix, with z-direction $\hat{k}$ against the relative velocity vector, and the x-direction $\hat{i}$ pointing from A to B. Figure 5.5 shows how the encounter plane will

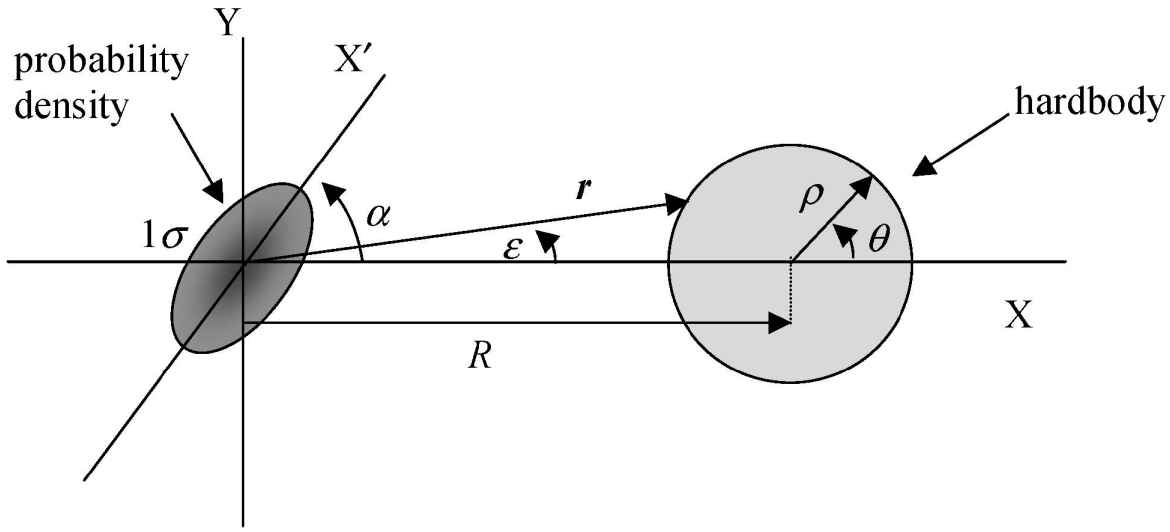


Figure 5.5: A diagram of the encounter plane. Taken from [57].

look in the new coordinate system.

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{r}_B - \boldsymbol{r}_A \\ \boldsymbol{v} &= \boldsymbol{v}_B - \boldsymbol{v}_A \\ \hat{k} &= \frac{\boldsymbol{v}}{|\boldsymbol{v}|} \\ \hat{i} &= \frac{\boldsymbol{r}}{|\boldsymbol{r}|} \\ -\hat{j} &= \hat{i} \times \hat{k} \\ U &= \begin{bmatrix} \hat{i} & \hat{j} & \hat{k} \end{bmatrix} \end{aligned} \tag{5.8}$$

Once we have defined the U matrix we can use it to transform the covariance matrix into the frame of the encounter plane. Note that we are only interested in the covariance matrix of the position states, $\boldsymbol{P}_{pos}$.

$$\boldsymbol{P}_{enc} = U^T \boldsymbol{P}_{pos} U \tag{5.9}$$

To make everything two-dimensional we reduce $\boldsymbol{P}_{enc}$ down to a 2×2 matrix. Just cut out the third row and third column of $\boldsymbol{P}_{enc}$.

$$\begin{aligned} \boldsymbol{P}_{enc} &= \begin{bmatrix} p_{ij} \end{bmatrix}_{i,j=1,2,3} \\ \boldsymbol{P}_{2x2} &= \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \end{aligned} \tag{5.10}$$

Now we have only two dimensions to integrate over. Patera [57, 55] reduced this to a one-dimensional line intregral around the edges of the hardbody. For a circular projection of radius ρ the integral is

$$PC_{cum} = \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{f\rho^2 + Rf\rho\cos\theta}{r^2} \right] \times \left[1 - \exp\left(-\frac{r^2}{2\sigma^2} \right) \right] \mathrm{d}\theta \tag{5.11}$$

where

$$r^2 = [R+\rho\cos\theta]^2\left[\cos^2\alpha + f^2\sin^2\alpha\right] + \rho^2\sin^2\theta\left[\sin^2\alpha + f^2\cos^2\alpha\right] + 2\rho(1-f^2)\cos\alpha\sin\alpha\sin\theta\,[R+\rho\cos\theta] \tag{5.12}$$

and $R = |\boldsymbol{r}| = |\boldsymbol{r}_B - \boldsymbol{r}_A|$ is the distance from the A to B at TCA.

We find f, α and σ by diagonalizing $\boldsymbol{P}_{2x2}$. The diagonalizing T matrix satisfying (5.13) can be found using the eigenvectors/eigenvalues of $\boldsymbol{P}_{2x2}$, or directly from α (the angle between the closest axis of $\boldsymbol{P}_{2x2}$ and $\boldsymbol{r} = \hat{i}$). The direct method is given in (5.14) and (5.15). If the eigenvector method is used to get T, then α can be easily computed either from T or the eigenvector.

$$\boldsymbol{P}_{diag} = T\boldsymbol{P}_{2x2}T^T = \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \tag{5.13}$$

$$T = \begin{bmatrix}\cos\alpha & \sin\alpha\\ -\sin\alpha & \cos\alpha\end{bmatrix} \tag{5.14}$$

$$\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right) \tag{5.15}$$

Note that if $\rho_{xy} > 0$ then α must be in the first quadrant ($\alpha \in [0, \frac{\pi}{2}]$). If $\rho_{xy} < 0$ then α must be in the fourth quadrant ($\alpha \in [-\frac{\pi}{2}, 0]$). We get f and σ from $\boldsymbol{P}_{diag}$.

$$\begin{aligned} f &= \frac{\sigma_1}{\sigma_2}\\ \sigma &= \sigma_1 \end{aligned} \tag{5.16}$$

Once we know R, α, f, and σ we can numerically integrate (5.11) to get an approximate PC_{cum} for the encounter. Note that there are other ways of solving the integral in (5.7) for a linear encounter. Alfano's [4, 3] and Carpenter et al.'s [15] versions of the 2D method are particularly worth noting .

5.3.5 Accuracy of the 2D Method

As mentioned previously, the 2D PC method relies on the assumption that an encounter between two satellites is linear. If this assumption is violated then it can give very innacurate results. Despite this, it is still the standard method used to estimate the PC for an encounter between satellites because of its computational efficiency and easy implementation.

Alfano tested the accuracy of several methods for approximating the cumulative PC for an encounter in *Satellite Conjunction Monte Carlo Analysis* [4]. Using Monte Carlo simulations to establish a baseline, he found that the 2D approximation had an error of less than 1%[1] for linear encounters. However, for nonlinear encounters the error could reach 60%. For example, in the long encounter from Figure 5.3 the 2D PC methods returned a PC_{cum} value of about 0.147, or 33% below the true probability of 0.217.

5.3.6 Voxels

There are several other ways to estimate PC_{cum} that do not rely on computationally-expensive Monte Carlo simulations (see Alfano [4]). However, we will only briefly touch on the method of voxels here.

In the method of voxels we transform the relative position of object B with respect to object A from regular Cartesian space into Mahalnobis space. In Mahalanobis space the distance between two points is measured not in units of length, but in the standard deviations of some probability distribution (in our case the distribution described by the covariance $\boldsymbol{P}$). This Mahalanobis distance is calculated using (5.17) [22].

$$D_M = \sqrt{(\boldsymbol{r}_B - \boldsymbol{r}_a)^T\boldsymbol{P}_{pos}(\boldsymbol{r}_B - \boldsymbol{r}_A)} \tag{5.17}$$

[1] That is $\left|\frac{PC_{2D} - PC_{MC}}{PC_{MC}}\right| \leq 0.01$, not $|PC_{2D} - PC_{MC}| \leq 0.01$.

Once B's trajectory has been transformed into Mahalanobis space, we divide the space up into discrete volumes called "voxels", and track which voxels the hardbody occupies at any given time step. Because the voxels are fixed in Mahalanobis space, each has a specific PC value associated with it. We get PC_{inst} using the sum of the PC values for each voxel occupied by the hardbody at a given moment. Similarly, we can find PC_{unique} by checking how many of those voxels have not been inside the hardbody before. Integrate the PC_{unique} curve to get PC_{cum}.

The reason why this works is that Mahalanobis space uses the covariance matrix $\boldsymbol{P}$ as its fixed reference, rather than regular 3D space. Because of this the hardbody returning to the same point in Mahalanobis space indicates that object B is passing through some point in the covariance ellipsoid that it has passed through before. It is very difficult to keep track of such events in Cartesian space. However since the distribution does not change in Mahalanobis sapce, we are able to create "holes" in that space, representing where B has already been, without worrying about how to propagate them forward in time. These holes serve the same purpose as tracking which pairs of particles have already collided in the Monte Carlo simulation.

5.4 Two Examples

5.4.1 A Short Encounter

We will start by looking at one of the linear encounters that Alfano [4] uses as a case study. Alfano's case 5 is a short encounter between two satellites in low earth orbit. Because their relative velocity is high the relative trajectory of the objects does not bend (much) during the encounter. We begin by looking at the two satellite's mean states and covariance matrices at the time of closest approach (TCA). For object A, the mean and covariance are shown below. Note that to keep the numbers relatively neat, the number of significant digits has been greatly reduced from Alfano. The coordinates are in the Earth-fixed equatorial frame.

$$\boldsymbol{\mu}_A = \begin{bmatrix} 6878090.1623\text{ m} \\ -17948.6786\text{ m} \\ -17948.6786\text{ m} \\ 28.0938\text{ m/s} \\ 5382.8902\text{ m/s} \\ 5382.8902\text{ m/s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_A \\ \boldsymbol{v}_A \end{bmatrix} \tag{5.18}$$

$$\boldsymbol{P}_A = \begin{bmatrix} 0.6421 & -18.9907 & -18.9907 & 0.02971 & (-1.6880\cdot 10^{-4}) & (-1.6880\cdot 10^{-4}) \\ -18.9907 & 790.40447 & 790.39662 & -12.3804 & 0.06420 & 0.06417 \\ -18.9907 & 790.39662 & 790.40447 & -12.3804 & 0.06417 & 0.06420 \\ 0.02971 & -12.3804 & -12.3804 & 0.01939 & (-1.0051\cdot 10^{-4}) & (-1.0051\cdot 10^{-4}) \\ (-1.6880\cdot 10^{-4}) & 0.06420 & 0.06417 & (-1.0051\cdot 10^{-4}) & (5.4473\cdot 10^{-7}) & (5.2059\cdot 10^{-4}) \\ (-1.6880\cdot 10^{-4}) & 0.06417 & 0.06420 & (-1.0051\cdot 10^{-4}) & (5.2059\cdot 10^{-7}) & (5.4473\cdot 10^{-7}) \end{bmatrix} \tag{5.19}$$

And for object B we have the following mean and covariance.

$$\boldsymbol{\mu}_B = \begin{bmatrix} 6878089.1620\text{ m} \\ -17946.6789\text{ m} \\ -17947.6783\text{ m} \\ 28.3938\text{ m/s} \\ 5383.1902\text{ m/s} \\ 5382.5902\text{ m/s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_B \\ \boldsymbol{v}_B \end{bmatrix} \tag{5.20}$$

$$\boldsymbol{P}_B = \begin{bmatrix} 0.6211 & -18.5521 & -18.5500 & 0.02902 & (-1.6522\cdot 10^{-4}) & (-1.6522\cdot 10^{-4}) \\ -18.5521 & 790.53549 & 790.43953 & -12.3818 & 0.06421 & 0.06417 \\ -18.5500 & 790.43953 & 790.35928 & -12.3804 & 0.06417 & 0.06420 \\ 0.02902 & -12.3817 & -12.3804 & 0.01939 & (-1.0051\cdot 10^{-4}) & (-1.0051\cdot 10^{-4}) \\ (-1.6522\cdot 10^{-4}) & 0.06421 & 0.06417 & (-1.0051\cdot 10^{-4}) & (5.4473\cdot 10^{-7}) & (5.2058\cdot 10^{-7}) \\ (-1.6522\cdot 10^{-4}) & 0.06417 & 0.06420 & (-1.0051\cdot 10^{-4}) & (5.2058\cdot 10^{-7}) & (5.4470\cdot 10^{-7}) \end{bmatrix} \tag{5.21}$$

The two objects have a combined hardbody radius, ρ, of 10 meters. The miss distance, R, between the two objects at TCA is only 2.83 m, so we expect that the cumulative PC for the encounter will be significant. We want to use the linear method to get estimate the PC_{cum} for the encounter, so we start by getting the combined position covariance matrix. We do this by adding the upper left 3×3 entries of $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$.

$$\boldsymbol{P}_{pos} = \begin{bmatrix} 0.1263 & -37.5428 & -37.5407 \\ -37.5428 & 1580.9400 & 1580.8362 \\ -37.5407 & 1580.8362 & 1580.7637 \end{bmatrix} \tag{5.22}$$

Now we need the U matrix to rotate $\boldsymbol{P}_{pos}$ into the encounter plane's frame. We start by computing the unit vectors $\hat{i}$, $\hat{j}$, and $\hat{k}$ for the frame. Recall that $\hat{i}$ points from A to B.

$$\hat{i} = \frac{\boldsymbol{r}_B - \boldsymbol{r}_A}{|\boldsymbol{r}_B - \boldsymbol{r}_A|} = \frac{1}{2.4495} \begin{bmatrix} -1.0003 \\ 1.9997 \\ 1.0003 \end{bmatrix} = \begin{bmatrix} -0.4083 \\ 0.8165 \\ 0.4083 \end{bmatrix} \tag{5.23}$$

We also know that $\hat{k}$ is perpendicular to the encounter plane, i.e. along the relative velocity vector.

$$\hat{k} = \frac{\boldsymbol{v}_B - \boldsymbol{v}_A}{|\boldsymbol{v}_B - \boldsymbol{v}_A|} = \frac{1}{0.5196} \begin{bmatrix} 0.3000 \\ 0.3000 \\ -0.3000 \end{bmatrix} = \begin{bmatrix} 0.57774 \\ 0.5774 \\ -0.5773 \end{bmatrix} \tag{5.24}$$

We complete the coordinate system with $\hat{j} = -\hat{i} \times \hat{k}$.

$$\hat{j} = -\hat{i} \times \hat{k} = \begin{bmatrix} 0.7071 \\ -3.0527 \cdot 10^{-6} \\ 0.7071 \end{bmatrix} \tag{5.25}$$

Use the unit vectors to construct U.

$$U = \begin{bmatrix} \hat{i} & \hat{j} & \hat{k} \end{bmatrix} = \begin{bmatrix} -0.4083 & 0.7071 & -0.57774 \\ 0.8165 & (-3.0527 \cdot 10^{-6}) & -0.5774 \\ 0.4083 & 0.7071 & 0.5773 \end{bmatrix} \tag{5.26}$$

Now we can rotate $\boldsymbol{P}_{pos}$ into the encounter frame.

$$\boldsymbol{P}_{enc} = U^T \boldsymbol{P}_{pos} U = \begin{bmatrix} 23750.6175 & 13668.5572 & -25.4594 \\ 13668.5572 & 7866.4015 & -14.7172 \\ -25.4594 & -14.7172 & 0.1444 \end{bmatrix} \tag{5.27}$$

We are not interested in the part of the covariance ellipsoid that lies outside the encounter plane, so we can discard the last row and column of $\boldsymbol{P}_{enc}$.

$$\boldsymbol{P}_{2x2} = \begin{bmatrix} 23750.6175 & 13668.5572 \\ 13668.5572 & 7866.4015 \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \tag{5.28}$$

Now that we have $\boldsymbol{P}_{2x2}$, we need to find the T that diagonalizes it so that we can get f and σ. We start by finding the angle α.

$$\alpha = \frac{1}{2} \tan^{-1} \left(\frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \right) \tag{5.29}$$

By examining (5.28) we can get the values of ρ_{xy}, σ_x, and σ_y to plut into (5.29). This gives us an alpha value of 0.5222. Note that since $\rho_{xy}\sigma_x\sigma_y = 13668.5572 > 0$, α must be in the first quadrant. Now we can compute T.

$$T = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} = \begin{bmatrix} 0.8667 & 0.4988 \\ -0.4988 & 0.8667 \end{bmatrix} \tag{5.30}$$

After using T to diagonalize $\boldsymbol{P}_{2x2}$, we get

$$\boldsymbol{P}_{diag} = T\boldsymbol{P}_{2x2}T^T = \begin{bmatrix} 31616.9419 & 0 \\ 0 & 0.07708 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \tag{5.31}$$

This gives us our values of f and σ.

$$\begin{aligned} f &= \frac{\sigma_1}{\sigma_2} = 640.4724 \\ \sigma &= \sigma_1 = 177.8115 \end{aligned} \tag{5.32}$$

We already know that $R = |\boldsymbol{r}_B - \boldsymbol{r}_A| = 2.4495$ m. Using all of the above we can now integrate the line integral.

$$PC_{cum} = \frac{1}{2\pi}\int_0^{2\pi}\left[\frac{f\rho^2 + Rf\rho\cos\theta}{r^2}\right] \times \left[1 - \exp\left(-\frac{r^2}{2\sigma^2}\right)\right] \mathrm{d}\theta \tag{5.33}$$

where

$$\begin{aligned} r^2 = [R+\rho\cos\theta]^2\left[\cos^2\alpha + f^2\sin^2\alpha\right] + \rho^2\sin^2\theta\left[\sin^2\alpha + f^2\cos^2\alpha\right] \\ + 2\rho(1-f^2)\cos\alpha\sin\alpha\sin\theta\,[R+\rho\cos\theta] \end{aligned} \tag{5.34}$$

Since (5.33) is extremely difficult to integrate analytically, we will be integrating numerically. The following segment of code shows how this can be done in MATLAB. Note that the angle θ goes from zero to 2π.

```
function [ q ] = pateraInt( theta,~)

   global rho alpha F sig R

   % precompute trig functions
   ct = cos(theta);
   st = sin(theta);
   ca = cos(alpha);
   sa = sin(alpha);

   % Compute r2
   r2 = (R + rho * ct)^2 * (ca^2 + F^2 * sa^2) + rho^2 * st^2 * ...
      (sa^2 + F^2 * ca^2) + 2 * rho * (1 - F^2) * ca * sa * st * (R + rho * ct);

   % Output the argument of integral at angle theta
   q = 1 / (2 * pi) * (F * rho^2 + R * F * rho * ct) / r2 * (1 - exp(-r2 / (2 * sig^2)));

end
```

After the numeric integration we estimate that the cumulative probability of collision for this encounter is 0.04477. This value closesly matches the 0.04509 PC found using a Monte Carlo simulation. The two values are compared in Figure 5.6. Although this PC may appear small at first glance it is actually quite high, since NASA will often maneuver satellites to avoid collisions with probabilities as low as 10^{-4} [38].

5.4.2 A Long Encounter

We can make a similar comparison for case 2 from Alfano's paper. This encounter is between two satellites in geosynchronous orbits. It is a long encounter (over three and a half hours), so the linearity assumption no longer holds.

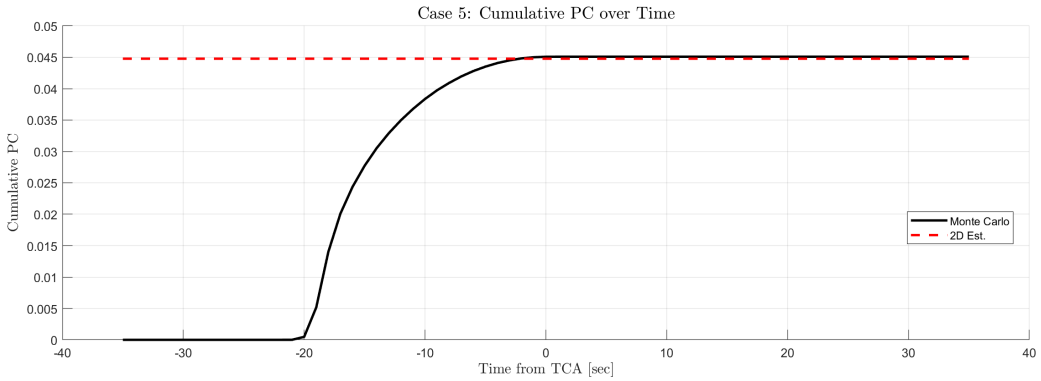


Figure 5.6: Compares the growth of cumulative PC over time found using a Monte Carlo simulation to the estimated PC using the linear method at TCA. The linear estimate has an error of 0.69% compared to the Monte Carlo results. The Monte Carlo simulation used 2500 particles per satellite to populate the covariance ellipsoids.

The mean states and covariance matrices of objects A and B are given below.

$$\boldsymbol{\mu}_A = \begin{bmatrix} 153446.180 \text{ m} \\ 41874155.872 \text{ m} \\ 0 \text{ m} \\ 3066.875 \text{ m/s} \\ -11.374 \text{ m/s} \\ 0 \text{ m/s} \end{bmatrix} \tag{5.35}$$

$$\boldsymbol{P}_A = \begin{bmatrix} 6494.080 & -376.139 & 0 & 0.0160 & -0.494 & 0 \\ -376.139 & 22.560 & 0 & (-9.883 \cdot 10^{-4}) & 0.0286 & 0 \\ 0 & 0 & 1.205 & 0 & 0 & (-6.071 \cdot 10^{-5}) \\ 0.0160 & (-9.883 \cdot 10^{-4}) & 0 & (4.437 \cdot 10^{-8}) & (-1.212 \cdot 10^{-6}) & 0 \\ -0.494 & 0.0286 & 0 & (-1.212 \cdot 10^{-6}) & (3.762 \cdot 10^{-5}) & 0 \\ 0 & 0 & (-6.071 \cdot 10^{-5}) & 0 & 0 & (3.390 \cdot 10^{-9}) \end{bmatrix} \tag{5.36}$$

$$\boldsymbol{\mu}_B = \begin{bmatrix} 153446.679 \text{ m} \\ 41874156.372 \text{ m} \\ 5.000 \text{ m} \\ 3066.865 \text{ m/s} \\ -11.364 \text{ m/s} \\ -1.358 \cdot 10^{-6} \text{ m/s} \end{bmatrix} \tag{5.37}$$

$$\boldsymbol{P}_B = \begin{bmatrix} 6494.224 & -376.156 & (-4.492 \cdot 10^{-5}) & 0.0160 & -0.494 & (-5.902 \cdot 10^{-8}) \\ -376.156 & 22.561 & (2.550 \cdot 10^{-6}) & (-9.885 \cdot 10^{-3}) & 0.0286 & (3.419 \cdot 10^{-9}) \\ (-4.492 \cdot 10^{-5}) & (2.550 \cdot 10^{-6}) & 1.205 & (-1.180 \cdot 10^{-10}) & (3.419 \cdot 10^{-9}) & (-6.072 \cdot 10^{-5}) \\ 0.0160 & (-9.885 \cdot 10^{-3}) & (-1.180 \cdot 10^{-10}) & (4.438 \cdot 10^{-8}) & (-1.212 \cdot 10^{-6}) & (-1.448 \cdot 10^{-13}) \\ -0.494 & 0.0286 & (3.419 \cdot 10^{-9}) & (-1.212 \cdot 10^{-6}) & (3.762 \cdot 10^{-5}) & (4.492 \cdot 10^{-12}) \\ (-5.902 \cdot 10^{-8}) & (3.419 \cdot 10^{-9}) & (-6.072 \cdot 10^{-5}) & (-1.448 \cdot 10^{-13}) & (4.492 \cdot 10^{-12}) & (3.392 \cdot 10^{-9}) \end{bmatrix} \tag{5.38}$$

If we decide to use the linear PC estimate despite the lack of linearity, then after following the same steps as in the short encounter we get an estimated cumulative PC of 0.0062. From Monte Carlo simulations, we know that the true PC should be about 0.0157. Unsurprisingly this is a much larger error than in the linear case. To be precise, the error has gone up from 0.69% to 60.3%. The difference between the two values is illustrated in Figure 5.7.

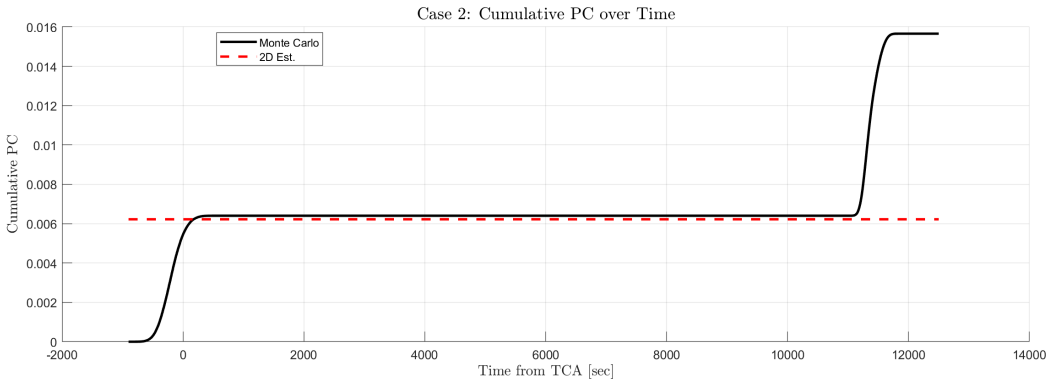


Figure 5.7: Compares the growth of cumulative PC over time found using a Monte Carlo simulation to the estimated PC using the linear method at TCA for the longer encounter. The linear estimate has an error of 60.3% compared to the Monte Carlo results. Note that most of the error comes from the two satellites making a second pass at each other after the initial TCA, which the linear method could not predict. The Monte Carlo simulation used 2500 particles per satellite to populate the covariance ellipsoids.

Chapter 6

Initial Orbit Determination

Initial orbit determination is the art to determine a full state, or a full set of orbital elements from observations. It is discriminated from Lambert's problem by the fact that the epochs of each observation are known. A full state has six parameters (position and velocity), the same number of quantities in the orbital element space (six orbital elements), fully define a unique dynamical situation.

6.1 Precursor - Orbit Parameters and Orbital Coordinate Systems

6.1.1 Keplerian Elements

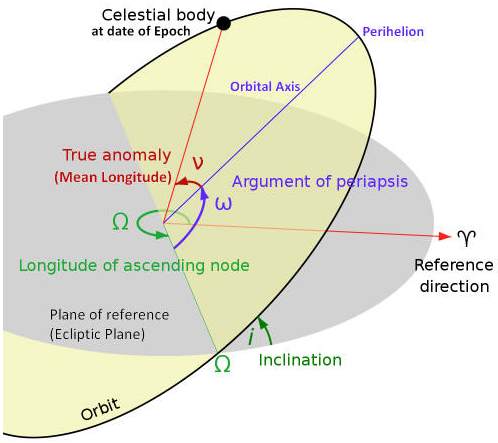


Figure 6.1: As a reminder: The definition of Keplerian orbital elements.

As a reminder, one set of orbital elements that is classically used is the semi-major axis a, the eccentricity e that describe the shape of the orbit, the two angles inclination i and right ascension of the ascending node (RAAN) Ω that describe the orientation of the orbital plane relative to the ECI frame, the argument of perigee ω (perigee, when it is referenced to the Earth, perihel when it is referenced to the Sun as central body, and periapsis when the central body is not explicitly defined) that describes how the orbit is oriented in the orbital plane, and the anomaly (true ν or mean M) that describe where on the orbit the object is at the moment.

Alternatively, the quantity of the passing time of the perigee, T_0. T_0 can have positive or negative values. It is understood as the time (in the future or the past) at which perigee is passed by the object, ideally T_0 (same as the anomalies) is understood to be defined within the same orbital period. A quantity that is often used is the argument of

latitude u, which is defined as:

$$u = \omega + \nu \tag{6.1}$$

6.1.2 The Orbital Coordinate System

Four different coordinate systems can be defined, all having the orbital plane as the fundamental plane. Those are also called the four systems of the two body problem. All systems share the same third axis, $\boldsymbol{h}$ the angular momentum axis. Enforcing that all coordinate systems are right handed and orthogonal,the coordinate systems hence can be uniquely defined, defining only one more axis, let's assume axis 1.

With the object being placed in the point P (where the small dot is drawn) and Π being the perigee, ω being the perigee axis and Ω being the right ascension of the ascending node, like before, the four different coordinate systems are defined via their four different first axis, $\boldsymbol{e}_\Omega$, , $\boldsymbol{e}_\Pi$, , $\boldsymbol{e}_R$, $\boldsymbol{e}_I$. Taking $\boldsymbol{r}$ to be the vector of the position of the object in the inertial ECI system and $\dot{\boldsymbol{r}}$ its velocity at the time t, the transformations leading to the first axis of the coordinate system and the transformation of the vector $\boldsymbol{r}$ in the new coordinate system can be found in Fig.6.8, corresponding to the notation in Fig.6.7. The angle ξ is defined as the angle between the Laplace vector $\boldsymbol{q}$ that is pointing towards the perigee and the velocity vector of the object $\dot{\boldsymbol{r}}$. It can also be defined as: The angle ξ is defined as:

$$\xi = 3\sqrt{\frac{\mu}{p^3}}(t - T_0), \tag{6.2}$$

with p being the orbital parameter and $T-0$ being the time of perigee passage. The vector $\boldsymbol{q}$ is the Laplace vector pointing to the perigee, defined as:

$$\boldsymbol{q} = (\dot{\boldsymbol{r}}^2 - \frac{\mu}{r})\boldsymbol{r} - (\boldsymbol{r}\cdot\dot{\boldsymbol{r}})\dot{\boldsymbol{r}}) \tag{6.3}$$

Note that $\boldsymbol{q} = \boldsymbol{0}$ for circular orbits.
The orbital element systems can be used for computationally efficient formulation of Kepler's equation and in the

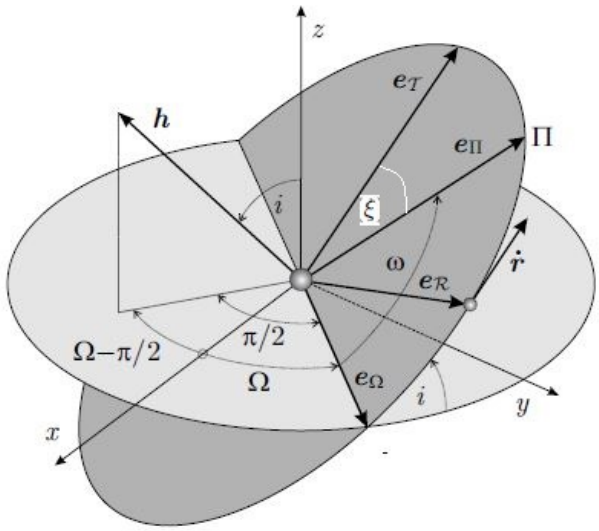


Figure 6.2: Illustration of the orbital element and the ECI coordinate system [9].

formulation of the integrals of motion. We use one of the systems, the one corresponding to axis e_Ω as the first axis to compute the true anomaly for the restricted orbit determination.

System	First unit vector	Transformation from Inertial System $\mathcal{I}$
Ω	$e_\Omega = \frac{e_3 \times h}{h}$	$r_\Omega = \mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
Π	$e_\Pi = \frac{q}{q}$	$r_\Pi = \mathbf{R}_3(\omega)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
$\mathcal{R}$	$e_\mathcal{R} = \frac{r}{r}$	$r_\mathcal{R} = \mathbf{R}_3(u)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
$\mathcal{T}$	$e_\mathcal{T} = \frac{\dot{r}}{\lvert\dot{r}\rvert}$	$r_\mathcal{T} = \mathbf{R}_3(\xi)\,\mathbf{R}_3(\omega)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$

Figure 6.3: Definition of the orbital element coordinate systems [9].

6.1.3 Orbital elements and the Angular Momentum Vector

References [9]. Fig. 6.8 shows the angular momentum vector $\boldsymbol{h}$. x,y,z are the axis of the ECI coordinate system with x pointing in the direction of the vernal equinox. Π denotes the perigee, where ω is the angle of perigee passage, also called argument of perigee. i is the inclination, and Ω is the right ascension of the ascending node. Because, the angular momentum vector is perpendicular on the orbital plane, the angle between $\boldsymbol{h}$ and the z-axis is again i. The projection of the angular momentum vector then forms an angle π with the ascending node, allowing it to be expressed via the following relation:

$$\boldsymbol{h} = |\boldsymbol{h}| \begin{bmatrix} \cos(\Omega - \frac{\pi}{2}) \sin i \\ \sin(\Omega - \frac{\pi}{2}) \sin i \\ \cos i \end{bmatrix} = |\boldsymbol{h}| \begin{bmatrix} \sin\Omega \sin i \\ -\cos\Omega \sin i \\ \cos i \end{bmatrix} \tag{6.4}$$

with

$$\Omega = \arctan\left(\frac{h_1}{-h_2}\right) \qquad i = \arccos\left(\frac{h_3}{|\boldsymbol{h}|}\right) \tag{6.5}$$

6.1.4 Kepler's Equation

[75] and [9].

6.1.5 Deriving the Orbital Elements from the State

Position $\boldsymbol{r} = [r_1, r_2, r_3]^T$, with $|\boldsymbol{r}| := r$
Velocity $\boldsymbol{v} = [v_1, v_2, v_3]^T$, with $|\boldsymbol{v}| := v$
In the inertial coordinate frame $\hat{i}, \hat{j}, \hat{k}$

Specific angular momentum:

$$\boldsymbol{h} = \boldsymbol{r} \times \boldsymbol{v} = [h_1, h_2, h_3]^T \tag{6.6}$$

$$|\boldsymbol{h}| := h \tag{6.7}$$

Inclination i:

$$i = \arccos(\frac{h_3}{h}) \tag{6.8}$$

$$\tag{6.9}$$

Right ascension of the ascending node:

$$\Omega = \begin{cases} \arccos(\frac{N_1}{N}) & \text{for } N_2 \geq 0 \\ 2\pi - \arccos(\frac{N_1}{N}) & \text{for } N_2 < 0 \end{cases} \tag{6.10}$$

$$\boldsymbol{N} = \hat{k} \times \boldsymbol{h} = [N_1, N_2, N_3]^T, \qquad |\boldsymbol{N}| := N \tag{6.11}$$

Eccentricity:

$$e := |\boldsymbol{e}| \tag{6.12}$$

$$\boldsymbol{e} = \frac{1}{\mu}[\boldsymbol{v}\times\boldsymbol{h} - \mu\frac{\boldsymbol{r}}{r}] = \frac{1}{\mu}[(v^2 - \frac{\mu}{r})\boldsymbol{r} - rv_r\boldsymbol{v}] \tag{6.13}$$

$$v_r = \frac{\boldsymbol{rv}}{r} \tag{6.14}$$

Argument of perigee:

$$\omega = \begin{cases} \arccos(\frac{\boldsymbol{Ne}}{Ne}) & \text{for } e_3 \geq 0 \\ 2\pi - \arccos(\frac{\boldsymbol{Ne}}{Ne}) & \text{for } e_3 < 0 \end{cases} \tag{6.15}$$

True anomaly:

$$\nu = \begin{cases} \arccos(\frac{\boldsymbol{er}}{er}) & \text{for } v_r \geq 0 \\ 2\pi - \arccos(\frac{\boldsymbol{er}}{er}) & \text{for } v_r < 0 \end{cases} \tag{6.16}$$

$$\nu = \begin{cases} \arccos(\frac{1}{e}(\frac{h^2}{\mu r} - 1)) & \text{for } v_r \geq 0 \\ 2\pi - \arccos(\frac{1}{e}(\frac{h^2}{\mu r} - 1)) & \text{for } v_r < 0 \end{cases} \tag{6.17}$$

Semi-major axis via perigee and apogee distance, for $0 < e < 1$:

$$a = \frac{1}{2}(r_p + r_a) \tag{6.18}$$

$$r_p = \frac{h^2}{\mu}\frac{1}{1+e} \tag{6.19}$$

$$r_a = \frac{h^2}{\mu}\frac{1}{1-e} \tag{6.20}$$

or:

$$a = -\frac{\mu}{2\varepsilon} \tag{6.21}$$

$$\varepsilon = \frac{v^2}{2} - \frac{\mu}{r} \tag{6.22}$$

Orbital parameter:

$$p = \begin{cases} a(1-e^2) & \text{for } e \neq 1 \\ \frac{h^2}{\mu} & else \end{cases} \tag{6.23}$$

6.1.6 Deriving the State from Orbital Elements

In the perifocal frame Π, position r_Π and velocity v_Π:

$$r_{1,\Pi} = \frac{p\cos(\nu)}{1+e\cos(\nu)} \tag{6.24}$$

$$r_{2,\Pi} = \frac{p\sin(\nu)}{1+e\cos(\nu)} \tag{6.25}$$

$$r_{3,\Pi} = 0 \tag{6.26}$$

$$v_{1,\Pi} = -\sqrt{\frac{\mu}{p}}\sin(\nu) \tag{6.27}$$

$$v_{2,\Pi} = \sqrt{\frac{\mu}{p}}(e + \cos(\nu)) \tag{6.28}$$

$$v_{3,\Pi} = 0 \tag{6.29}$$

$$\boldsymbol{r} = \boldsymbol{R}_3(-\Omega)\boldsymbol{R}_1(-i)\boldsymbol{R}_3(-\omega)\boldsymbol{r}_\Pi \tag{6.30}$$
$$\boldsymbol{v} = \boldsymbol{R}_3(-\Omega)\boldsymbol{R}_1(-i)\boldsymbol{R}_3(-\omega)\boldsymbol{v}_\Pi \tag{6.31}$$

Circular equatorial orbit, Ω, ω and ν is ill-defined in Keplerian elements, use true longitude $\lambda = \Omega + \omega + \nu$:

$$\nu = \lambda \tag{6.32}$$
$$\omega = 0.0 \tag{6.33}$$

Circular inclined orbit, ν is ill-defined, use argument of latitude $u = \nu + \omega$:

$$\nu = u \tag{6.34}$$
$$\omega = 0.0 \tag{6.35}$$

Elliptical equatorial orbit, Ω, ω are ill-defined, use the longitude or periapsis $\bar{\omega} = \Omega + \omega$:

$$\Omega = 0.0 \tag{6.36}$$
$$\omega = \bar{\omega} \tag{6.37}$$

6.2 Classical Methods

References [9, 75, 27, 8].

6.2.1 Two Astrometric (Angle-only) Meaurements - Restricted Orbit Determination

References [9]. If only two measurements are available, a restricted orbit can be determined. For a circular orbit, two of the orbital elements are known or restricted, that is the eccentricity e is set to zero, and the argument of periapsis is set to zero as well, because it is not defined for a circular orbit:

$$e = 0 \qquad \omega = 0 \tag{6.38}$$

This leaves four orbital elements to be determined. The time of pericenter passage T_0 is therefore necessarily defined as the time of passage through the ascending node.

It is assumed two angle-observations are available, α_{t1}, δ_{t1}, α_{t2}, δ_{t2} at times t_1 and t_2. Furthermore we know our station vector $\boldsymbol{R}_{topo,t1}$ and $\boldsymbol{R}_{topo,t2}$. We can define our two unit vectors in the direction of the object at the time of observations:

$$\hat{\boldsymbol{L}}_{t1,t2} = \begin{bmatrix} \cos\alpha_{t1,t2}\cos\delta_{t1,t2} \\ \sin\alpha_{t1,t2}\cos\delta_{t1,t2} \\ \sin\delta_{t1,t2} \end{bmatrix} \tag{6.39}$$

The station vector might be given in Cartesian Earth fixed (ECEF) coordinates. With the definition of the sidereal time of the observer θ, however, it is not a problem to transform it into the space fixed frame (ECI). If we would like to take nutation, precession and polar movement into account we could do so, and e.g. express our station vector relative to J2000. This has the advantage that we would define our orbit in J2000 as well and the pseudo observations could be readily compared with further real observations.

$$\boldsymbol{R}_{\text{ECI,topo,t}} = \boldsymbol{R}_3(-\theta(t))\tilde{\boldsymbol{R}}_{\text{ECEF}} \tag{6.40}$$
$$\boldsymbol{R}_{ICRS/J2000.0,topot,t} = \boldsymbol{P}^T(t)\boldsymbol{N}^T(t)\boldsymbol{\theta}^T(t)\boldsymbol{\Pi}^T(t)\tilde{\boldsymbol{R}}_{ITRS} \tag{6.41}$$
$$\text{Note: } \boldsymbol{R}_3(-\theta(t)) = \boldsymbol{R}_3^T(\theta(t)) := \boldsymbol{\theta}(t) \tag{6.42}$$

What we are missing are the range ρ and all angular rates. We can express the position $\boldsymbol{r}$ of the object at each observation time as the following:

$$\boldsymbol{r}_{t1,t2} = \rho_{t1,t2}\hat{\boldsymbol{L}}_{t1,t2} + \boldsymbol{R}_{topo;t1,t2}, \tag{6.43}$$

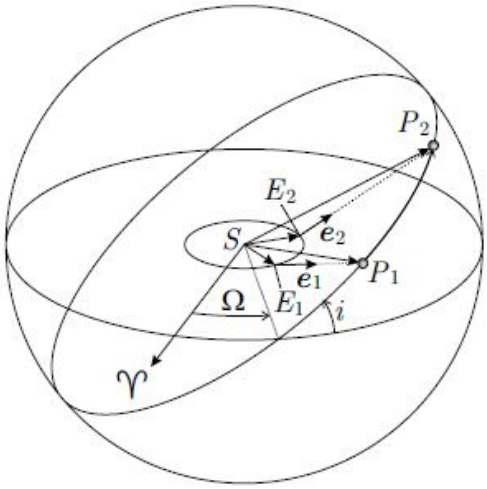


Figure 6.4: Two angular observations.

with the range ρ and the position vector of the observation station in the ECI system $\boldsymbol{R}$. In the above equation the only missing quantity is the range. Solving for the range leads to the following expression:

$$r_{t1,t2}^2 = \rho_{t1,t2}^2 + R_{topo;t1,t2}^2 + 2\hat{\boldsymbol{L}}_{t1,t2}\boldsymbol{R}_{topo;t1,t2}\rho_{t1,t2}, \tag{6.44}$$

$$\text{because } \hat{\boldsymbol{L}}_{t1,t2}^2 = 1 \tag{6.45}$$

$$\rho_{t1;1,2} = -\hat{\boldsymbol{L}}_{t1}\boldsymbol{R}_{topo;t1} \pm \sqrt{(\hat{\boldsymbol{L}}_{t1}\boldsymbol{R}_{topo;t1})^2 - (R_{topo;t1}^2 - r_{t1}^2)} \tag{6.46}$$

$$\rho_{t2;1,2} = -\hat{\boldsymbol{L}}_{t2}\boldsymbol{R}_{topo;t2} \pm \sqrt{(\hat{\boldsymbol{L}}_{t2}\boldsymbol{R}_{topo;t2})^2 - (R_{topo;t1}^2 - r_{t2}^2)} \tag{6.47}$$

But we are not out of luck, we can discard the negative solutions for ρ as non-physical right away!
Furthermore, the absolute values, $r_{t1,2}$, at both times have to be equal to the radius of the circular orbit, which is nothing else than the semi-major axis a. Hence:

$$\rho_{t1} = -\hat{\boldsymbol{L}}_{t1}\boldsymbol{R}_{topo;t1} + \sqrt{(\hat{\boldsymbol{L}}_{t1}\boldsymbol{R}_{topo;t1})^2 - (R_{topo;t1}^2 - a^2)} \tag{6.48}$$

$$\rho_{t2} = -\hat{\boldsymbol{L}}_{t2}\boldsymbol{R}_{topo;t2} + \sqrt{(\hat{\boldsymbol{L}}_{t2}\boldsymbol{R}_{topo;t2})^2 - (R_{topo;t2}^2 - a^2)} \tag{6.49}$$

Once we know the semi-major axis, we readily have two full position vectors via the range or vice versa.

In order to find the semi-major axis, we use a trick: We now have two choices to express the angle between $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Either geometrically using the dot product of the two vectors or via considerations of orbital mechanics. We call the two angles ϕ_g for geometric and ϕ_{CM} for celestial mechanics.

$$\cos\phi_g = \frac{\boldsymbol{r}_1 \cdot \boldsymbol{r}_2}{|\boldsymbol{r}_1||\boldsymbol{r}_2|} \tag{6.50}$$

$$\phi_g = \arccos\frac{\boldsymbol{r}_1 \cdot \boldsymbol{r}_2}{|\boldsymbol{r}_1||\boldsymbol{r}_2|} \tag{6.51}$$

Given the expressions in Eq.6.43, 6.48 and 6.49, the angle ϕ_g only depends on the unknown ρ and hence the semi-major axis a as the only single unknown.

Alternatively, the angle can be expressed via orbital mechanics considerations, as determined by the time difference Δt and the mean motion $n = \sqrt{\frac{\mu}{a^3}}$:

$$\phi_{CM} = n\Delta t = \sqrt{\frac{\mu}{a^3}} \cdot [t_2 - \frac{\rho_2}{c} - (t_1 - \frac{\rho_1}{c})] \tag{6.52}$$

Note that we are correcting for the light travel time here. Again, ϕ_{CM} depends only on the semi-major and on the range. But with Eq.6.49 and Eq.6.48 the ranges can be expressed in terms of the semi-major axis as well.

Of course, in reality, the two angles need to be identical, this leads to a root finding problem for the only remaining free variable, the semi-major axis a:

$$F := \phi_{CM}(a) - \phi_g(a) = 0 \tag{6.53}$$

For the root finding problem, a Newton method or equivalent can be applied.

Unsurprisingly, Eq.6.53 in general has more than one solution, but three. One is the orbit of the Earth itself, and then two which are equally likely from a mathematical point of view. One of them can sometimes be excluded because it contradicts the time evolution constraint $t2 > t1$. This leaves us with the task to find the remaining three orbital

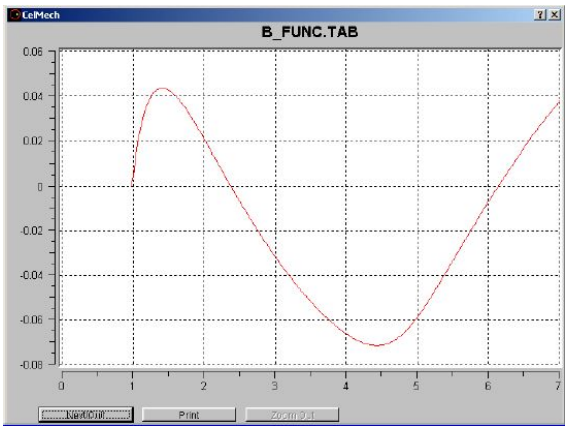


Figure 6.5: Example of the function F, X axis in Earth radii.

elements. However we do have the advantage now that we know the orbital plane defined by $\boldsymbol{r}_1, \boldsymbol{r}_2$. We can define the angular moment vector $\boldsymbol{h}$ such that:

$$\boldsymbol{h} = [h_1, h_2, h_3]^T = \boldsymbol{r}_1 \times \boldsymbol{r}_2 \tag{6.54}$$

Which directly allows to determine the RAAN Ω and the inclination i which are the orbital elements linked to the definition of the orbital plane:

$$\Omega = \arctan\left(\frac{h_1}{-h_2}\right) \qquad i = \arccos\left(\frac{h_3}{|\boldsymbol{h}|}\right) \tag{6.55}$$

The tricky part that is left then is, how do we find either the mean/true anomaly or the time of perigee passage. As mentioned before because the argument of perigee is set to zero, the perigee and the RAAN coincide. The angle of the RAAN Ω however we already know. The position vector of the object expressed in the coordinate system of the orbital plane is:

$$\boldsymbol{r}_{\Omega,t1} = [r_{\Omega,1,t1}, r_{\Omega,2,t1}, r_{\Omega,3,t1}]^T = \boldsymbol{R}_1(i)\boldsymbol{R}_3(\Omega)\boldsymbol{r}_{t1} \tag{6.56}$$

For the definition of the orbital element coordinate system see Figs.6.8 and 6.7. This leads to the definition of the angle defining the argument of latitude u_1 at time t_1:

$$u_1 = \arctan(\frac{r_{\Omega,2,t1}}{r_{\Omega,1,t1}}) := \omega + \nu, \tag{6.57}$$

with the true anomaly ν. However, we already know that ω, the argument of perigee, was defined to be zero.

$$\nu = \arctan(\frac{r_{\Omega,2}}{r_{\Omega,1}}) \tag{6.58}$$

Alternatively, we can derived the argument of perigee in units of time, or in other words the time since the perigee passage:

$$T_0 = t_1 - \frac{\rho_1}{c} - \frac{u_1}{n} = t_1 - \frac{\rho_1}{c} - \frac{\nu}{n}, \tag{6.59}$$

where n is the mean motion of the object.

6.2.2 Three Astrometric (Angle-Only) Observations - Geometrically Constrained: Gauss Method

References [9, 75, 27, 8]. It is assumed we have three astrometric observations, right ascension and declination at three difference epochs:

$$t_1, \alpha_1, \delta_1$$
$$t_2, \alpha_2, \delta_2$$
$$t_3, \alpha_3, \delta_3$$

That provides sufficient information to determine an orbit, the question as always is, how do we do it. With the three observations we can express the position vector of the object at the time of the observation as the following:

$$\boldsymbol{r}_i = \rho_i \hat{\boldsymbol{L}}_i + \boldsymbol{R}_i \tag{6.60}$$

$$\hat{\boldsymbol{L}}_i = \begin{bmatrix} \cos\alpha_i \cos\delta_i \\ \sin\alpha_i \cos\delta_i \\ \sin\delta_i \end{bmatrix} \tag{6.61}$$

$$\boldsymbol{R}_i := \boldsymbol{R}_{\text{ECI,topo,t}} = \boldsymbol{R}_3(-\theta(t))\tilde{\boldsymbol{R}}_{\text{ECEF}} \tag{6.62}$$

$$[\text{or } \boldsymbol{R}_i := \boldsymbol{R}_{ICRS/J2000.0,topot,t} = \boldsymbol{P}^T(t)\boldsymbol{N}^T(t)\boldsymbol{\theta}^T(t)\boldsymbol{\Pi}^T(t)\tilde{\boldsymbol{R}}_{ITRS}] \tag{6.63}$$

Note that the shorthanded notation is used here, the index i corresponds to the time t_i, i=1,2,3. Note $\boldsymbol{R}_3(-\theta(t)) = \boldsymbol{R}_3^T(\theta(t)) := \boldsymbol{\theta}(t)$.

Gauss' method runs the following train of thoughts:

- with three observations, it is probably best to project everything on the middle observation using the information of the other two observations as boundaries
- all three positions of the object lie in the same plane
- a Taylor series expansion is always a good thing especially if I have some knowledge of orbital mechanics and can use that in the expansion
- love algebra!

So, now let's do the math:

Gauss' method is called a geometrically constrained method, because it relies on the assumption, that all observations are in the same plane:

$$a\boldsymbol{r}_1 + b\boldsymbol{r}_2 + c\boldsymbol{r}_3 = \boldsymbol{0}, \tag{6.64}$$

where $\boldsymbol{r}_i$ is the position of the object at time t_i and a,b,c are real numbers (definition equation for a plane!), note that a=b=c=0 is not valid here.

What assumption does that correspond to and what might be a problem with that assumption?

Of course, the problem is, the states $\boldsymbol{r}_i$ are not known, otherwise we would be half done already. But, let's see how far we get. What can be done is to directly solve for the state at the middle time t_2:

$$\boldsymbol{r}_2 = -(a/b)\boldsymbol{r}_1 - (c/b)\boldsymbol{r}_3 \tag{6.65}$$

In order to declutter the representation, it is defined $c_1 := -(a/b)$ and $c_3 := -(c/b)$, such that:

$$c_1\boldsymbol{r}_1 - \boldsymbol{r}_2 + c_3\boldsymbol{r}_3 = \boldsymbol{0}, \tag{6.66}$$

or

$$c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + c_3\boldsymbol{r}_3 = \boldsymbol{0}, \tag{6.67}$$

with $c_2 = -1$. Note, one could also just define $b = -1$, because one of the parameters is free in Eq.6.64 and keep using a and c; one will find this approach used in some representations of the method.

Using Eq.6.66 we can solve for the position at time t_2:

$$\boldsymbol{r}_2 = c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3. \tag{6.68}$$

If the coefficients would be known we could use the equation to tackle the unknown range of the middle position, ρ_2. Hence, in order to make a step towards solving for the coefficients c_1 and c_3 the cross products can be formed:

$$\boldsymbol{r}_1 \times \boldsymbol{r}_2 = \boldsymbol{r}_1 \times (c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3) \; = c_3\boldsymbol{r}_1 \times \boldsymbol{r}_3 \tag{6.69}$$

$$\boldsymbol{r}_3 \times \boldsymbol{r}_2 = \boldsymbol{r}_3 \times (c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3) \; = c_1\boldsymbol{r}_3 \times \boldsymbol{r}_1 \tag{6.70}$$

It is sought now, to express $\boldsymbol{r}_1$ and $\boldsymbol{r}_3$ in terms of $\boldsymbol{r}_2$, this would allow to express the above relations all for the middle observation and there could be a chance of solving the system of equations. The problem is approached using so-called f and g functions, centered around the time t_2:

$$\boldsymbol{r}_1 = f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2 \tag{6.71}$$

$$\boldsymbol{r}_3 = f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2 \tag{6.72}$$

This effectively says that if the position and velocity of an object are known at a given instant, then the position and velocity at any later time are found in terms of the known position and velocity. The f and g coefficients are known as the Lagrange coefficients. f and g function expressions can be found via comparison with the well known Taylor series expansion:

$$\begin{aligned} \boldsymbol{r}(t_i) = \boldsymbol{r}_i &= \boldsymbol{r}(t_2) + \dot{\boldsymbol{r}}(t_2)\tau_i + \frac{\ddot{\boldsymbol{r}}(t_2)\tau_i^2}{2!} + \mathscr{O}(\tau_i^3), \\ \boldsymbol{r}_i &= \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_{\boldsymbol{2}}\tau_i + \frac{\ddot{\boldsymbol{r}}_2\tau_i^2}{2!} + \mathscr{O}(\tau_i^3) \end{aligned} \tag{6.73}$$

with $\tau_i = t_i - t_2$. Because the dynamical system is known it is already known that:

$$\dot{\boldsymbol{r}}_i = \boldsymbol{v}_i \tag{6.74}$$

$$\ddot{\boldsymbol{r}}_i = -\frac{\mu}{r_i^3}\boldsymbol{r}_i := -u_i\boldsymbol{r}_i \tag{6.75}$$

$$\dddot{\boldsymbol{r}}_i = -\dot{u}_i\boldsymbol{r}_i - u_i\dot{\boldsymbol{r}}_i \tag{6.76}$$

This means for the definition of the f and g series centered at the second time t_2

$$f_i = 1 - \frac{1}{2}u_2\tau_i^2 + \mathscr{O}(\tau_i^3) \tag{6.77}$$

$$g_i = \tau_i - \frac{1}{6}u_2\tau_i^3 + \mathscr{O}(\tau_i^4) \tag{6.78}$$

where $u_2 = \mu\|\boldsymbol{r}_2\|^{-3}$. It is also important to note that the coefficients of the f and g functions are truncated. If the τ_i^3 and higher terms for f and the τ_i^4 and higher terms for g would have been kept, then the values would contain the

unknown velocity at t_2.

Substituting the f and g series expansions for the first and third positions into the cross product relationship of Eq.6.69 and Eq.6.70:

$$(f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2)\times\boldsymbol{r}_2 = c_3(f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2)\times(f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2) \tag{6.79}$$
$$(f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2)\times\boldsymbol{r}_2 = c_1(f_3\boldsymbol{r}_2 + g_3\boldsymbol{v}_2)\times(f_1\boldsymbol{r}_2 + g_1\boldsymbol{v}_2) \tag{6.80}$$

Expanding out all of the cross products, and then rearranging and reducing the resulting expression, leads to:

$$-g_1(\boldsymbol{r}_2\times\boldsymbol{v}_2) = c_3(f_1g_3 - g_1f_3)(\boldsymbol{r}_2\times\boldsymbol{v}_2), \tag{6.81}$$
$$-g_3(\boldsymbol{r}_2\times\boldsymbol{v}_2) = c_1(f_3g_1 - g_3f_1)(\boldsymbol{r}_2\times\boldsymbol{v}_2). \tag{6.82}$$

c_1 and c_3 can be found by equating the leading coefficients from the left-hand and right-hand sides of the preceding equations; this gives

$$c_1 = \frac{g_3}{f_1g_3 - f_3g_1} \qquad \text{and} \qquad c_3 = -\frac{g_1}{f_1g_3 - f_3g_1} \tag{6.83}$$

Hence, if the f and g series expressions are known, the coefficients can be calculated.
Now, if the higher-order terms are neglected, that is to truncated at terms $\mathscr{O}(\tau_i^3)$ in the f series and $\mathscr{O}(\tau_i^4)$ in the g series, the common denominator of the c_1 and c_3 coefficients is given by

$$f_1g_3 - f_3g_1 \approx \left[1 - \frac{1}{2}u_2\tau_1^2\right]\left[\tau_3 - \frac{1}{6}u_2\tau_3^3\right] - \left[1 - \frac{1}{2}u_2\tau_3^2\right]\left[\tau_1 - \frac{1}{6}u_2\tau_1^3\right] \tag{6.84}$$
$$= \tau_3 - \frac{1}{2}u_2\tau_1^2\tau_3 - \frac{1}{6}u_2\tau_3^3 + \frac{1}{12}u_2^2\tau_1^2\tau_3^3 \tag{6.85}$$
$$- \tau_1 + \frac{1}{2}u_2\tau_1\tau_3^2 + \frac{1}{6}u_2\tau_1^3 - \frac{1}{12}u_2^2\tau_1^3\tau_3^2 \tag{6.86}$$

In order to be consistent 5$^{\text{th}}$-order terms (mixed products), which are also the terms that are $\mathscr{O}(u_2^2)$, are truncated, such that the denominator can be manipulated to yield

$$f_1g_3 - f_3g_1 \approx \left[\tau_3 - \tau_1\right] - \frac{1}{6}u_2\left[\tau_3^3 + 3\tau_1^2\tau_3 - 3\tau_1\tau_3^2 - \tau_1^3\right] \tag{6.87}$$
$$= (\tau_3 - \tau_1) - \frac{1}{6}u_2(\tau_3 - \tau_1)^3 \tag{6.88}$$
$$= \tau_{13} - \frac{1}{6}u_2\tau_{13}^3 \tag{6.89}$$
$$= \tau_{13}\left[1 - \frac{1}{6}u_2\tau_{13}^2\right] \tag{6.90}$$

where we note that $\tau_{13} = \tau_3 - \tau_1$. That's the denominator. It can be inverted using the generalized binomial theorem, which is given by

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \tag{6.91}$$

and apply it to find $(f_1g_3 - f_3g_1)^{-1}$, yielding

$$\frac{1}{f_1g_3 - f_3g_1} = \frac{1}{\tau_{13}}\left[1 - \frac{1}{6}u_2\tau_{13}^2\right]^{-1} \tag{6.92}$$
$$= \frac{1}{\tau_{13}}\left[1 + \frac{1}{6}u_2\tau_{13}^2 + \frac{1}{36}u_2^2\tau_{13}^4 + \cdots\right] \tag{6.93}$$
$$\approx \frac{1}{\tau_{13}}\left[1 + \frac{1}{6}u_2\tau_{13}^2\right] \tag{6.94}$$

Note that 4$^{\text{th}}$-order and higher terms, or equivalently terms that are $\mathcal{O}(u_2^2)$ and higher, in the expansion generated by the application of the binomial theorem have been neglected.

Substituting Eq.6.94 back into the definition of the coefficients in Eq.6.83, yields:

$$c_1 \approx \frac{1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[\tau_3-\frac{1}{6}u_2\tau_3^3\right] \tag{6.95}$$

$$c_3 \approx -\frac{1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[\tau_1-\frac{1}{6}u_2\tau_1^3\right] \tag{6.96}$$

Now, we manipulate and reduce the expression for c_1:

$$c_1 = \frac{1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[\tau_3-\frac{1}{6}u_2\tau_3^3\right] \tag{6.97}$$

$$= \frac{\tau_3}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[1-\frac{1}{6}u_2\tau_3^2\right] \tag{6.98}$$

$$= \frac{\tau_3}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2-\frac{1}{6}u_2\tau_3^2-\frac{1}{36}u_2^2\tau_{13}^2\tau_3^2\right] \tag{6.99}$$

$$\approx \frac{\tau_3}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2-\frac{1}{6}u_2\tau_3^2\right] \tag{6.100}$$

$$= \frac{\tau_3}{\tau_{13}}\left[1+\frac{1}{6}u_2\left[\tau_{13}^2-\tau_3^2\right]\right] \tag{6.101}$$

Next, we manipulate and the expression for c_3:

$$c_3 = -\frac{1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[\tau_1-\frac{1}{6}u_2\tau_1^3\right] \tag{6.102}$$

$$= -\frac{\tau_1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2\right]\left[1-\frac{1}{6}u_2\tau_1^2\right] \tag{6.103}$$

$$= -\frac{\tau_1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2-\frac{1}{6}u_2\tau_1^2-\frac{1}{36}u_2^2\tau_1^2\tau_{13}^2\right] \tag{6.104}$$

$$\approx -\frac{\tau_1}{\tau_{13}}\left[1+\frac{1}{6}u_2\tau_{13}^2-\frac{1}{6}u_2\tau_1^2\right] \tag{6.105}$$

$$= -\frac{\tau_1}{\tau_{13}}\left[1+\frac{1}{6}u_2\left[\tau_{13}^2-\tau_1^2\right]\right] \tag{6.106}$$

Now approximations for the c_1 and c_3 coefficients are reached(remember also that $c_2 = -1$) in terms of the time differences between observations, the gravitational parameter, and the unknown value of $\|\boldsymbol{r}_2\|$.

Repeating Eq.6.66

$$c_1\boldsymbol{r}_1 - \boldsymbol{r}_2 + c_3\boldsymbol{r}_3 = \boldsymbol{0}, \tag{6.107}$$

the coefficients are now known, in case the times are known to solve for τ and the expression u_2 is known. However, $u_2 := u_2(|\boldsymbol{r}_2|)$. Because the range is missing, an expression for $|\boldsymbol{r}_2|$ is not readily available.

As defined in Eq.6.60 it is known:

$$\boldsymbol{r}_i = \boldsymbol{R}_i + \rho_i\hat{\boldsymbol{L}}_i \quad i \in \{1,2,3\} \tag{6.108}$$

We can now substitute the object positions into the planar condition, such that

$$c_1\left[\boldsymbol{R}_1+\rho_1\hat{\boldsymbol{L}}_1\right]-\left[\boldsymbol{R}_2+\rho_2\hat{\boldsymbol{L}}_2\right]+c_3\left[\boldsymbol{R}_3+\rho_3\hat{\boldsymbol{L}}_3\right]=\boldsymbol{0} \tag{6.109}$$

The problem is, only one equation is given for the three unknown ranges. The planar condition can be rewritten in order to group all observations together and to group all of the observer positions together

$$c_1\rho_1\hat{\boldsymbol{L}}_1-\rho_2\hat{\boldsymbol{L}}_2+c_3\rho_3\hat{\boldsymbol{L}}_3=-c_1\boldsymbol{R}_1+\boldsymbol{R}_2-c_3\boldsymbol{R}_3 \tag{6.110}$$

Our objective now is to isolate the slant ranges ρ_1, ρ_2, and ρ_3.

The dot products will be of great use again at this enterprise. To isolate ρ_1, we will take the dot product of the Eq.6.110 with the term $(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)$, which yields

$$c_1\rho_1\hat{\boldsymbol{L}}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)-\rho_2\hat{\boldsymbol{L}}_2\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)+c_3\rho_3\hat{\boldsymbol{L}}_3\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.111}$$

$$=-c_1\boldsymbol{R}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)+\boldsymbol{R}_2\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)-c_3\boldsymbol{R}_3\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.112}$$

Since

$$\hat{\boldsymbol{L}}_2\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)=\hat{\boldsymbol{L}}_3\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)=\boldsymbol{0} \tag{6.113}$$

this can be reduced to

$$c_1\rho_1\hat{\boldsymbol{L}}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.114}$$

$$=-c_1\boldsymbol{R}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)+\boldsymbol{R}_2\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3)-c_3\boldsymbol{R}_3\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.115}$$

We will define the terms D_0, D_{11}, D_{21}, and D_{31} to be

$$D_0=\hat{\boldsymbol{L}}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \qquad D_{11}=\boldsymbol{R}_1\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.116}$$

$$D_{21}=\boldsymbol{R}_2\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \qquad D_{31}=\boldsymbol{R}_3\cdot(\hat{\boldsymbol{L}}_2\times\hat{\boldsymbol{L}}_3) \tag{6.117}$$

which gives

$$c_1\rho_1D_0=-c_1D_{11}+D_{21}-c_3D_{31} \tag{6.118}$$

It is worth noting that each of the D terms can be completely computed based on the available data. Then, provided that $D_0\neq 0$, which will only happen if $\hat{\boldsymbol{L}}_1$, $\hat{\boldsymbol{L}}_2$, and $\hat{\boldsymbol{L}}_3$ lie in a plane, we can solve for the slant range ρ_1 as

$$\rho_1=\frac{1}{D_0}\left[-D_{11}+\frac{1}{c_1}D_{21}-\frac{c_3}{c_1}D_{31}\right] \tag{6.119}$$

This is the solution for ρ_1 in terms of the c_1 and c_3 coefficients.

Now, let's find similar solutions for ρ_2 and ρ_3. If we take the planar condition equation and dot it with $(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_3)$, then follow a similar process, we can show that

$$\rho_2=\frac{1}{D_0}\left[-c_1D_{12}+D_{22}-c_3D_{32}\right] \tag{6.120}$$

where

$$D_{12}=\boldsymbol{R}_1\cdot(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_3) \quad D_{22}=\boldsymbol{R}_2\cdot(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_3) \quad D_{32}=\boldsymbol{R}_3\cdot(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_3) \tag{6.121}$$

It is also important to note that $\hat{\boldsymbol{L}}_2\cdot(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_3)=-D_0$ was used to arrive in the preceding relationship for ρ_2. For the relationship for ρ_3, we take the planar condition equation and dot it with $(\hat{\boldsymbol{L}}_1\times\hat{\boldsymbol{L}}_2)$, which leads to

$$\rho_3=\frac{1}{D_0}\left[-\frac{c_1}{c_3}D_{13}+\frac{1}{c_3}D_{23}-D_{33}\right] \tag{6.122}$$

where

$$D_{13} = \boldsymbol{R}_1 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) \quad D_{23} = \boldsymbol{R}_2 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) \quad D_{33} = \boldsymbol{R}_3 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) \tag{6.123}$$

Here, we have used the fact that $\hat{\boldsymbol{L}}_3 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) = D_0$. We now have solutions for the three ranges in terms of the D-coefficients and the c-coefficients.

The D-coefficients can be found completely in terms of the known station locations at the times of the measurements and the line-of-sight measurements.

The c coefficients, however, depend on the times of the measurements and the unknown $\|\boldsymbol{r}_2\|$. However, now a solution for the range, ρ_2 is available, and we substitute for the c-coefficients:

$$\rho_2 = \frac{1}{D_0}\left[-c_1 D_{12} + D_{22} - c_3 D_{32}\right] \tag{6.124}$$

$$= \frac{1}{D_0}\left\{-\frac{\tau_3}{\tau_{13}}\left[1 + \frac{1}{6}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\left(\tau_{13}^2 - \tau_3^2\right)\right] D_{12} + D_{22}\right. \tag{6.125}$$

$$\left. + \frac{\tau_1}{\tau_{13}}\left[1 + \frac{1}{6}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\left(\tau_{13}^2 - \tau_1^2\right)\right] D_{32}\right\} \tag{6.126}$$

$$= \frac{1}{D_0}\left[-\frac{\tau_3}{\tau_{13}} D_{12} + D_{22} + \frac{\tau_1}{\tau_{13}} D_{32}\right] \tag{6.127}$$

$$+ \mu \frac{1}{6D_0}\left[-\left(\tau_{13}^2 - \tau_3^2\right)\frac{\tau_3}{\tau_{13}} D_{12} + \left(\tau_{13}^2 - \tau_1^2\right)\frac{\tau_1}{\tau_{13}} D_{32}\right]\frac{1}{\|\boldsymbol{r}_2\|^3} \tag{6.128}$$

Now, define A and B as

$$A = \frac{1}{D_0}\left[-\frac{\tau_3}{\tau_{13}} D_{12} + D_{22} + \frac{\tau_1}{\tau_{13}} D_{32}\right] \tag{6.129}$$

$$B = \frac{1}{6D_0}\left[-\left(\tau_{13}^2 - \tau_3^2\right)\frac{\tau_3}{\tau_{13}} D_{12} + \left(\tau_{13}^2 - \tau_1^2\right)\frac{\tau_1}{\tau_{13}} D_{32}\right] \tag{6.130}$$

such that ρ_2 can be written as

$$\rho_2 = A + \mu B \|\boldsymbol{r}_2\|^{-3} \tag{6.131}$$

The A and B coefficients can be computed based on information that we know. If we substitute for the c-coefficients into our solutions for ρ_1 and ρ_3, we can show that

$$\rho_1 = \frac{1}{D_0}\left[\frac{6\left(D_{31}\frac{\tau_1}{\tau_3} + D_{21}\frac{\tau_{13}}{\tau_3}\right)\|\boldsymbol{r}_2\|^3 + \mu D_{31}\left(\tau_{13}^2 - \tau_1^2\right)\frac{\tau_1}{\tau_3}}{6\|\boldsymbol{r}_2\|^3 + \mu\left(\tau_{13}^2 - \tau_3^2\right)} - D_{11}\right] \tag{6.132}$$

$$\rho_3 = \frac{1}{D_0}\left[\frac{6\left(D_{13}\frac{\tau_3}{\tau_1} - D_{23}\frac{\tau_{13}}{\tau_1}\right)\|\boldsymbol{r}_2\|^3 + \mu D_{13}\left(\tau_{13}^2 - \tau_3^2\right)\frac{\tau_3}{\tau_1}}{6\|\boldsymbol{r}_2\|^3 + \mu\left(\tau_{13}^2 - \tau_1^2\right)} - D_{33}\right] \tag{6.133}$$

Each of our ranges now depends upon the times of the observations, which we know, the D-coefficients, which we can compute, and the magnitude of the position vector at t_2, which we do not know.

So how do we solve for $\|\boldsymbol{r}_2\|$? We need this value to find any of the ranges.

Recall that

$$\boldsymbol{r}_2 = \boldsymbol{R}_2 + \rho_2 \hat{\boldsymbol{L}}_2 \tag{6.134}$$

Let's compute the squared magnitude of the geocentric position:

$$\|\boldsymbol{r}_2\|^2 = \|\boldsymbol{R}_2 + \rho_2 \hat{\boldsymbol{L}}_2\|^2 \tag{6.135}$$

$$= \rho_2^2 + 2\rho_2(\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2) + \|\boldsymbol{R}_2\|^2 \tag{6.136}$$

Now, we can substitute for our solution for ρ_2 in terms of A, B, and $\|\boldsymbol{r}_2\|$

$$\|\boldsymbol{r}_2\|^2 = \left[A + \mu B \|\boldsymbol{r}_2\|^{-3}\right]^2 + 2\left[A + \mu B \|\boldsymbol{r}_2\|^{-3}\right](\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2) + \|\boldsymbol{R}_2\|^2 \tag{6.137}$$

$$= \left[A^2 + 2A(\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2) + \|\boldsymbol{R}_2\|^2\right] \tag{6.138}$$

$$+ \left[2\mu B\big(A + (\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2)\big)\right]\|\boldsymbol{r}_2\|^{-3} + \left[\mu^2 B^2\right]\|\boldsymbol{r}_2\|^{-6} \tag{6.139}$$

Let's define

$$a = -A^2 - 2A(\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2) - \|\boldsymbol{R}_2\|^2 \tag{6.140}$$

$$b = -2\mu B\big(A + (\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2)\big) \tag{6.141}$$

$$c = -\mu^2 B^2 \tag{6.142}$$

such that we can write

$$\|\boldsymbol{r}_2\|^2 + a + b\|\boldsymbol{r}_2\|^{-3} + c\|\boldsymbol{r}_2\|^{-6} = 0 \tag{6.143}$$

Now, multiply through by $\|\boldsymbol{r}_2\|^6$, which yields

$$\|\boldsymbol{r}_2\|^8 + a\|\boldsymbol{r}_2\|^6 + b\|\boldsymbol{r}_2\|^3 + c = 0 \tag{6.144}$$

We have an 8^{th}-order polynomial in terms of $\|\boldsymbol{r}\|$. All that remains, then, is to find a real root of the octic. Once we have a real root of the octic, we can compute each of the slant ranges, ρ_1, ρ_2, and ρ_3.

Then, with each of the ranges, we can find the three position vectors as

$$\boldsymbol{r}_i = \boldsymbol{R}_i + \rho_i \hat{\boldsymbol{L}}_i \quad i \in \{1,2,3\} \tag{6.145}$$

This is typically the formal ending point of Gauss' method.

We have three positions that should form a plane, but we don't have an orbit...we don't have a velocity. A common approach is to apply something like Gibbs' method (observations are far apart form each other) or the Herrick-Gibbs method (observations are closely spaced) to convert three position vectors into a position and velocity. These methods are common as an end to Gauss' method or in the use of radar-based initial orbit determination.

We can, as a quick (and not as accurate) alternative, compute the velocity as well. Recall that the f and g series can be used to determine the position at some other time given the position and velocity at a given time, i.e.

$$\boldsymbol{r}_1 = f_1 \boldsymbol{r}_2 + g_1 \boldsymbol{v}_2 \tag{6.146}$$

$$\boldsymbol{r}_3 = f_3 \boldsymbol{r}_2 + g_3 \boldsymbol{v}_2 \tag{6.147}$$

If we solve the first equation for the position at time t_2, we have

$$\boldsymbol{r}_2 = \frac{1}{f_1}\boldsymbol{r}_1 - \frac{g_1}{f_1}\boldsymbol{v}_2 \tag{6.148}$$

Now, we substitute this position into the equation for the position at time t_3 to yield

$$\boldsymbol{r}_3 = f_3\left[\frac{1}{f_1}\boldsymbol{r}_1 - \frac{g_1}{f_1}\boldsymbol{v}_2\right] + g_3\boldsymbol{v}_2 \tag{6.149}$$

$$= \frac{f_3}{f_1}\boldsymbol{r}_1 + \left[\frac{f_1 g_3 - f_3 g_1}{f_1}\right]\boldsymbol{v}_2 \tag{6.150}$$

From this, we can solve for the velocity at time t_2, which gives

$$\boldsymbol{v}_2 = \frac{1}{f_1 g_3 - f_3 g_1}\left[f_1\boldsymbol{r}_3 - f_3\boldsymbol{r}_1\right] \tag{6.151}$$

That is, if we know the positions at times t_1 and t_3, we can determine the velocity at time t_2.

6.2.2.1 Algorithm for Gauss' Method

- Given: t_i, $\hat{\boldsymbol{L}}_i$, and $\boldsymbol{R}_i$ for $i \in \{1,2,3\}$

- Calculate the time intervals

$$\tau_1 = t_1 - t_2 \qquad \tau_3 = t_3 - t_2 \qquad \tau_{13} = \tau_3 - \tau_1$$

- Compute the ten triple products

$$D_0 = \hat{\boldsymbol{L}}_1 \cdot (\hat{\boldsymbol{L}}_2 \times \hat{\boldsymbol{L}}_3)$$

$$\begin{aligned} D_{11} &= \boldsymbol{R}_1 \cdot (\hat{\boldsymbol{L}}_2 \times \hat{\boldsymbol{L}}_3) & D_{21} &= \boldsymbol{R}_2 \cdot (\hat{\boldsymbol{L}}_2 \times \hat{\boldsymbol{L}}_3) & D_{31} &= \boldsymbol{R}_3 \cdot (\hat{\boldsymbol{L}}_2 \times \hat{\boldsymbol{L}}_3) \\ D_{12} &= \boldsymbol{R}_1 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_3) & D_{22} &= \boldsymbol{R}_2 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_3) & D_{32} &= \boldsymbol{R}_3 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_3) \\ D_{13} &= \boldsymbol{R}_1 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) & D_{23} &= \boldsymbol{R}_2 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) & D_{33} &= \boldsymbol{R}_3 \cdot (\hat{\boldsymbol{L}}_1 \times \hat{\boldsymbol{L}}_2) \end{aligned}$$

- Calculate the coefficients A and B

$$A = \frac{1}{D_0}\left[-\frac{\tau_3}{\tau_{13}}D_{12} + D_{22} + \frac{\tau_1}{\tau_{13}}D_{32}\right] \tag{6.152}$$

$$B = \frac{1}{6D_0}\left[-\left(\tau_{13}^2 - \tau_3^2\right)\frac{\tau_3}{\tau_{13}}D_{12} + \left(\tau_{13}^2 - \tau_1^2\right)\frac{\tau_1}{\tau_{13}}D_{32}\right] \tag{6.153}$$

- Calculate the polynomial coefficients a, b, and c

$$a = -A^2 - 2A(\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2) - \|\boldsymbol{R}_2\|^2 \tag{6.154}$$

$$b = -2\mu B\big(A + (\hat{\boldsymbol{L}}_2 \cdot \boldsymbol{R}_2)\big) \tag{6.155}$$

$$c = -\mu^2 B^2 \tag{6.156}$$

- Position vector magnitude: solve

$$0 = \|\boldsymbol{r}_2\|^8 + a\|\boldsymbol{r}_2\|^6 + b\|\boldsymbol{r}_2\|^3 + c \tag{6.157}$$

to obtain the applicable real root $\|\boldsymbol{r}_2\|$.

- Determine the slant ranges

$$\rho_1 = \frac{1}{D_0}\left[\frac{6\left(D_{31}\frac{\tau_1}{\tau_3}+D_{21}\frac{\tau_{13}}{\tau_3}\right)\|\boldsymbol{r}_2\|^3+\mu D_{31}\left(\tau_{13}^2-\tau_1^2\right)\frac{\tau_1}{\tau_3}}{6\|\boldsymbol{r}_2\|^3+\mu\left(\tau_{13}^2-\tau_3^2\right)}-D_{11}\right] \tag{6.158}$$

$$\rho_2 = A+\mu B\|\boldsymbol{r}_2\|^{-3} \tag{6.159}$$

$$\rho_3 = \frac{1}{D_0}\left[\frac{6\left(D_{13}\frac{\tau_3}{\tau_1}-D_{23}\frac{\tau_{13}}{\tau_1}\right)\|\boldsymbol{r}_2\|^3+\mu D_{13}\left(\tau_{13}^2-\tau_3^2\right)\frac{\tau_3}{\tau_1}}{6\|\boldsymbol{r}_2\|^3+\mu\left(\tau_{13}^2-\tau_1^2\right)}-D_{33}\right] \tag{6.160}$$

- Compute the position vectors at t_1, t_2, and t_3:

$$\boldsymbol{r}_i = \boldsymbol{R}_i+\rho_i\hat{\boldsymbol{L}}_i \quad i\in\{1,2,3\} \tag{6.161}$$

Insert now into Gibbs (large times between observations) or Herrick-Gibbs (short times between observations) method (depending on the spacing of the observations relative to the orbital period),
or a (not as accurate) short cut:

- *Calculate the Lagrange coefficients*

$$f_1 = 1-\frac{1}{2}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\tau_1^2 \qquad f_3 = 1-\frac{1}{2}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\tau_3^2$$

$$g_1 = \tau_1-\frac{1}{6}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\tau_1^3 \qquad g_3 = \tau_3-\frac{1}{6}\frac{\mu}{\|\boldsymbol{r}_2\|^3}\tau_3^3$$

- *Output the position and velocity:*

$$\boldsymbol{r}_2 = \boldsymbol{r}_2 \qquad \boldsymbol{v}_2 = \frac{1}{f_1g_3-f_3g_1}\left[f_1\boldsymbol{r}_3-f_3\boldsymbol{r}_1\right]$$

6.2.3 Three Astrometric (Angle-Only) Observations - Two-Body Constrained: Laplace's Method

References [9, 75, 27, 8]. Laplace's method begins by assuming that we have a set of unit vector observations (from angles) taken from a site

$$\boldsymbol{L}_i = \begin{bmatrix}\cos\delta_i\cos\alpha_i\\ \cos\delta_i\sin\alpha_i\\ \sin\delta_i\end{bmatrix} \tag{6.162}$$

The position of the object is the position of the observer combined with the position of the object with respect to the observer

$$\boldsymbol{r} = \boldsymbol{R}+\rho\boldsymbol{L} \tag{6.163}$$

Note that the magnitude of the object position is given by

$$\|\boldsymbol{r}\| = \sqrt{\boldsymbol{r}\cdot\boldsymbol{r}} = \left[\rho^2+2\rho\boldsymbol{L}\cdot\boldsymbol{R}+\|\boldsymbol{R}\|^2\right]^{1/2} \tag{6.164}$$

We can take the time derivative of the object position to find that

$$\boldsymbol{r} = \boldsymbol{R} + \rho \boldsymbol{L} \tag{6.165}$$
$$\dot{\boldsymbol{r}} = \dot{\boldsymbol{R}} + \dot{\rho}\boldsymbol{L} + \rho\dot{\boldsymbol{L}} \tag{6.166}$$
$$\ddot{\boldsymbol{r}} = \ddot{\boldsymbol{R}} + \ddot{\rho}\boldsymbol{L} + \rho\ddot{\boldsymbol{L}} + 2\dot{\rho}\dot{\boldsymbol{L}} \tag{6.167}$$

Now, we make the assumption that the object obeys two-body motion, such that

$$\ddot{\boldsymbol{r}} = -\mu \|\boldsymbol{r}\|^{-3}\boldsymbol{r} \tag{6.168}$$

Let's substitute for $\boldsymbol{r}$ and $\ddot{\boldsymbol{r}}$ in terms of the site location and the relative position into the two-body equation

$$\ddot{\boldsymbol{R}} + \ddot{\rho}\boldsymbol{L} + \rho\ddot{\boldsymbol{L}} + 2\dot{\rho}\dot{\boldsymbol{L}} = -\mu\|\boldsymbol{r}\|^{-3}(\boldsymbol{R} + \rho\boldsymbol{L}) \tag{6.169}$$

We can rearrange terms to isolate the observer terms on the right-hand side of the equation

$$\ddot{\rho}\boldsymbol{L} + 2\dot{\rho}\dot{\boldsymbol{L}} + \rho(\ddot{\boldsymbol{L}} + \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L}) = -(\ddot{\boldsymbol{R}} + \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}) \tag{6.170}$$

This can be formed into a fundamental system of equations as

$$\left[\begin{array}{c|c|c} \boldsymbol{L} & 2\dot{\boldsymbol{L}} & (\ddot{\boldsymbol{L}} + \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L}) \end{array}\right] \begin{bmatrix} \ddot{\rho} \\ \dot{\rho} \\ \rho \end{bmatrix} = -(\ddot{\boldsymbol{R}} + \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}) \tag{6.171}$$

Note that the observer's position and acceleration are known.

Values of $\boldsymbol{L}$ are known at several times (at each of the measurement times); there may be observation errors in these values, but we cannot do anything about that.

The values of $\boldsymbol{L}$ will be used to determine $\dot{\boldsymbol{L}}$ and $\ddot{\boldsymbol{L}}$ later.

The magnitude of the object's position, $\|\boldsymbol{r}\|$, is unknown; this is very important to keep in mind, as it will come back up.

The objective is to solve for ρ and $\dot{\rho}$. If these values can be found, then the position and velocity of the object can be determined.

To solve for ρ and $\dot{\rho}$, we are going to use Cramer's rule.

That is, in using Cramer's rule, we will assume that there is a 3×3 system of the form

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \tag{6.172}$$

or, in order to obtain a model that is directly related to our exact formulation, we can define the system in terms of vectors as

$$\left[\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{array}\right] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \boldsymbol{d} \tag{6.173}$$

Solutions for the 3×3 system are obtained from Cramer's rule

$$x = \frac{\left|\left|\begin{array}{c|c|c} \boldsymbol{d} & \boldsymbol{b} & \boldsymbol{c} \end{array}\right|\right|}{\left|\left|\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{array}\right|\right|}, \quad y = \frac{\left|\left|\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{d} & \boldsymbol{c} \end{array}\right|\right|}{\left|\left|\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{array}\right|\right|}, \quad \text{and} \quad z = \frac{\left|\left|\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{d} \end{array}\right|\right|}{\left|\left|\begin{array}{c|c|c} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{array}\right|\right|} \tag{6.174}$$

All of the solutions to the 3×3 system are obtained as ratios of determinants of 3×3 matrices.

The denominator is always the determinant of the original left-hand side matrix.

The numerator for the i^{th} solution is the determinant of the original left-hand side matrix, but with the i^{th} column replaced with the right-hand side vector.

If we now apply Cramer's rule to the fundamental system that we cast previously, we can solve for ρ and $\dot{\rho}$ to yield

$$\rho = \frac{\left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ \right|\right|}{\left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right|} \tag{6.175}$$

$$\dot{\rho} = \frac{\left|\left|\ \boldsymbol{L}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right|}{\left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right|} \tag{6.176}$$

These look pretty ugly; can we do something to make them look a bit better?

More than just looking better, can we remove the dependence on the unknown magnitude of the object's position, $\|\boldsymbol{r}\|$?

Let us begin the process by defining Δ_0 to be the common denominator

$$\Delta_0 = \left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right| \tag{6.177}$$

Now, it might help if we know some neat properties of the determinant; these will help us to reduce our expressions down to things that we can compute.

Properties of the determinant:

1. if two columns (or rows) in a matrix are interchanged, the determinant changes sign

$$\left|\left|\ \boldsymbol{b}\ |\quad \boldsymbol{a}\ |\quad \boldsymbol{c}\ \right|\right| = -\left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}\ |\quad \boldsymbol{c}\ \right|\right| \tag{6.178}$$

2. the value of the determinant is unchanged if any scalar multiple of any column (or row) is added to any other column (or row)

$$\left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}\ |\quad \boldsymbol{c}+k\boldsymbol{a}\ \right|\right| = \left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}\ |\quad \boldsymbol{c}\ \right|\right| \tag{6.179}$$

3. if any column (or row) of a matrix is identically zero, the value of the determinant is zero

$$\left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}\ |\quad \boldsymbol{0}\ \right|\right| = 0 \tag{6.180}$$

4. multiplying any column (or row) of a matrix by a nonzero scalar multiplies the determinant by the same scalar

$$\left|\left|\ \boldsymbol{a}\ |\quad k\boldsymbol{b}\ |\quad \boldsymbol{c}\ \right|\right| = k\left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}\ |\quad \boldsymbol{c}\ \right|\right| \tag{6.181}$$

5. if a column (or row) of a matrix is given by the sum of two vectors, the determinant may be split into two determinants

$$\left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}_1+\boldsymbol{b}_2\ |\quad \boldsymbol{c}\ \right|\right| = \left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}_1\ |\quad \boldsymbol{c}\ \right|\right| + \left|\left|\ \boldsymbol{a}\ |\quad \boldsymbol{b}_2\ |\quad \boldsymbol{c}\ \right|\right| \tag{6.182}$$

Remember, we are trying to solve for ρ and $\dot{\rho}$, but we do not know $\|\boldsymbol{r}\|$.

Applying Property #2 to Δ_0 by adding $-\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L}$ (scalar times first column) to the third column yields

$$\Delta_0 = \left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.183}$$

Next, applying Property #4 to Δ_0 allows the factor of 2 in the second column to be brought outside of the determinant, such that

$$\Delta_0 = 2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.184}$$

Now, the denominator determinant, Δ_0, simply depends upon $\boldsymbol{L}$ and its rates.
We have a series of $\boldsymbol{L}$ values, which we will use to approximate the derivatives later on.
For now, let's look at the range/range-rate solutions.
Using the definition of Δ_0, the solutions of the fundamental system are

$$\Delta_0\rho = \left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ \right|\right| \tag{6.185}$$

$$\Delta_0\dot{\rho} = \left|\left|\ \boldsymbol{L}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right| \tag{6.186}$$

Let's use the properties of the determinant to manipulate the range solution

$$\Delta_0\rho = \left|\left|\ \boldsymbol{L}\ |\quad 2\dot{\boldsymbol{L}}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ \right|\right| \tag{6.187}$$

$$\overset{\text{P4}}{=} -2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}\ \right|\right| \tag{6.188}$$

$$\overset{\text{P5}}{=} -2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}\ \right|\right| - 2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}\ \right|\right| \tag{6.189}$$

$$\overset{\text{P4}}{=} -2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}\ \right|\right| - 2\mu\|\boldsymbol{r}\|^{-3}\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \boldsymbol{R}\ \right|\right| \tag{6.190}$$

So, the determinants in the range solution now depend upon the $\boldsymbol{L}$ terms and the observer terms, but the $\|\boldsymbol{r}\|$ dependence is now removed from the determinants.
We can now define Δ_1 and Δ_2 as

$$\Delta_1 = \left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}\ \right|\right| \quad \text{and} \quad \Delta_2 = \left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \boldsymbol{R}\ \right|\right| \tag{6.191}$$

such that the range solution can be written as

$$\Delta_0\rho = -2\Delta_1 - 2\mu\|\boldsymbol{r}\|^{-3}\Delta_2 \tag{6.192}$$

We will leave this solution alone for now and turn back to the range-rate solution.
Let us once again make use of the properties of the determinant, but this time in the cause of manipulating the range-rate solution:

$$\Delta_0\dot{\rho} = \left|\left|\ \boldsymbol{L}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ |\quad (\ddot{\boldsymbol{L}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{L})\ \right|\right| \tag{6.193}$$

$$\overset{\text{P2}}{=} \left|\left|\ \boldsymbol{L}\ |\quad -(\ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R})\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.194}$$

$$\overset{\text{P4}}{=} -\left|\left|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}+\mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.195}$$

$$\overset{\text{P5}}{=} -\left|\left|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| - \left|\left|\ \boldsymbol{L}\ |\quad \mu\|\boldsymbol{r}\|^{-3}\boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.196}$$

$$\overset{\text{P4}}{=} -\left|\left|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| - \mu\|\boldsymbol{r}\|^{-3}\left|\left|\ \boldsymbol{L}\ |\quad \boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.197}$$

As with the range solution, the determinants now depend upon the $\boldsymbol{L}$ and $\boldsymbol{R}$ terms, but not upon the unknown term $\|\boldsymbol{r}\|$.
Therefore, we define Δ_3 and Δ_4 as

$$\Delta_3 = \left|\left|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \quad \text{and} \quad \Delta_4 = \left|\left|\ \boldsymbol{L}\ |\quad \boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \tag{6.198}$$

and then the range-rate solution can be expressed as

$$\Delta_0\dot{\rho} = -\Delta_3 - \mu\|\boldsymbol{r}\|^{-3}\Delta_4 \tag{6.199}$$

Alright...where are we?
The solutions to the fundamental system are now written as

$$\Delta_0\rho = -2\Delta_1 - 2\mu\|\boldsymbol{r}\|^{-3}\Delta_2 \tag{6.200}$$

$$\Delta_0\dot{\rho} = -\Delta_3 - \mu\|\boldsymbol{r}\|^{-3}\Delta_4 \tag{6.201}$$

which means that, provided that $\Delta_0 \neq 0$, solutions for the range and range-rate are given by

$$\rho = -2(\Delta_1/\Delta_0) - 2\mu\|\boldsymbol{r}\|^{-3}(\Delta_2/\Delta_0) \tag{6.202}$$

$$\dot{\rho} = -(\Delta_3/\Delta_0) - \mu\|\boldsymbol{r}\|^{-3}(\Delta_4/\Delta_0) \tag{6.203}$$

Each of the five determinants depends only on the observer's position and acceleration and the line-of-sight and its rates

$$\Delta_0 = 2\Big|\Big|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{L}}\ \Big|\Big|,\quad \Delta_1 = \Big|\Big|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}\ \Big|\Big|,\quad \Delta_2 = \Big|\Big|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \boldsymbol{R}\ \Big|\Big|,$$
$$\Delta_3 = \Big|\Big|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}\ |\quad \ddot{\boldsymbol{L}}\ \Big|\Big|,\quad \text{and}\quad \Delta_4 = \Big|\Big|\ \boldsymbol{L}\ |\quad \boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \Big|\Big|$$

Therefore, if we know $\boldsymbol{L}$, $\dot{\boldsymbol{L}}$, $\ddot{\boldsymbol{L}}$, $\boldsymbol{R}$, and $\ddot{\boldsymbol{R}}$, we can directly compute the five determinants.

Then, only the lack of knowledge of $\|\boldsymbol{r}\|$ stands in the way of computing ρ and $\dot{\rho}$.

We know, at least the position of the observer, $\boldsymbol{R}$.

If we take the Earth to be a rigid body that is rotating at a constant angular velocity and the observer to be fixed to the Earth, the observer's velocity and acceleration may be found as

$$\dot{\boldsymbol{R}} = \boldsymbol{\omega} \times \boldsymbol{R} \qquad\qquad \ddot{\boldsymbol{R}} = \boldsymbol{\omega} \times \boldsymbol{\omega} \times \boldsymbol{R} = \boldsymbol{\omega} \times \dot{\boldsymbol{R}} \tag{6.204}$$

where $\boldsymbol{\omega}$ is the angular velocity of the Earth.

This takes care of the observer-related quantities in the determinant calculations.

What about the line-of-sight-dependent quantities? We already calculated those, and can derive those directly, as we have seen. A step between finite differences and the full solution are to use Lagrange's interpolation formula, which is given by

$$\boldsymbol{L}(t) = \sum_{j=1}^{n} \boldsymbol{L}_j \prod_{\ell \neq j} \frac{t - t_\ell}{t_j - t_\ell} \tag{6.205}$$

When applied to three observations, Lagrange interpolation yields

$$\boldsymbol{L}(t) = \frac{(t-t_2)(t-t_3)}{(t_1-t_2)(t_1-t_3)}\boldsymbol{L}_1 + \frac{(t-t_1)(t-t_3)}{(t_2-t_1)(t_2-t_3)}\boldsymbol{L}_2 + \frac{(t-t_1)(t-t_2)}{(t_3-t_1)(t_3-t_2)}\boldsymbol{L}_3 \tag{6.206}$$

But we also want derivatives, so we can differentiate twice with respect to time to find

$$\dot{\boldsymbol{L}}(t) = \frac{2t-t_2-t_3}{(t_1-t_2)(t_1-t_3)}\boldsymbol{L}_1 + \frac{2t-t_1-t_3}{(t_2-t_1)(t_2-t_3)}\boldsymbol{L}_2 + \frac{2t-t_1-t_2}{(t_3-t_1)(t_3-t_2)}\boldsymbol{L}_3 \tag{6.207}$$

$$\ddot{\boldsymbol{L}}(t) = \frac{2}{(t_1-t_2)(t_1-t_3)}\boldsymbol{L}_1 + \frac{2}{(t_2-t_1)(t_2-t_3)}\boldsymbol{L}_2 + \frac{2}{(t_3-t_1)(t_3-t_2)}\boldsymbol{L}_3 \tag{6.208}$$

There is *no restriction* to using only three observations, but this is the minimal set to get a non-zero second time rate-of-change of the line-of-sight.

Given some time (usually taken as the time of the middle observation), each of the five determinants can be determined.

Therefore, all that is left unknown in the solutions for range and range-rate is the magnitude of the position vector, $\|\boldsymbol{r}\|$.

Recall the range and range-rate solutions

$$\rho = -2(\Delta_1/\Delta_0) - 2\mu\|\boldsymbol{r}\|^{-3}(\Delta_2/\Delta_0) \tag{6.209}$$

$$\dot{\rho} = -(\Delta_3/\Delta_0) - \mu\|\boldsymbol{r}\|^{-3}(\Delta_4/\Delta_0) \tag{6.210}$$

Also, recall that the magnitude of the position vector (of the object) must satisfy

$$\|\boldsymbol{r}\|^2 = \rho^2 + 2\rho\boldsymbol{L}\cdot\boldsymbol{R} + \|\boldsymbol{R}\|^2 \tag{6.211}$$

Now, substitute for the range solutions in terms of the unknown $\|\boldsymbol{r}\|$

$$\|\boldsymbol{r}\|^2 = \left[-2(\Delta_1/\Delta_0) - 2\mu\|\boldsymbol{r}\|^{-3}(\Delta_2/\Delta_0)\right]^2 \tag{6.212}$$

$$+2\left[-2(\Delta_1/\Delta_0) - 2\mu\|\boldsymbol{r}\|^{-3}(\Delta_2/\Delta_0)\right]\boldsymbol{L}\cdot\boldsymbol{R} + \|\boldsymbol{R}\|^2 \tag{6.213}$$

If we rearrange terms, it follows that we find an 8^{th}-order polynomial in terms of $\|\boldsymbol{r}\|$ that must be satisfied:

$$0 = \|\boldsymbol{r}\|^8 + \left[-4(\Delta_1/\Delta_0)^2 + 4(\Delta_1/\Delta_0)\boldsymbol{L}\cdot\boldsymbol{R} - \|\boldsymbol{R}\|^2\right]\|\boldsymbol{r}\|^6 \tag{6.214}$$

$$+\left[-8\mu(\Delta_1/\Delta_0)(\Delta_2/\Delta_0) + 4\mu(\Delta_2/\Delta_0)\boldsymbol{L}\cdot\boldsymbol{R}\right]\|\boldsymbol{r}\|^3 - 4\mu^2(\Delta_2/\Delta_0)^2 \tag{6.215}$$

It is noted that we know (can compute) all of the coefficients in the polynomial.

All that remains, then, is to find a real root of the octic.

With this value, the range and range-rate can be determined, which then allows for the position and velocity of the object to be found.

Recall that it was assumed that $\Delta_0 \neq 0$. This determinant only vanishes when the three positions of the object as seen from the observer lie on the arc of a great circle. In this case, another observation should be used to remove the indeterminacy.

6.2.3.1 Algorithm for Laplace's Method

- Given: t_i, $\boldsymbol{L}_i$, and $\boldsymbol{R}_i$ for $i \in \{1,2,3\}$ and $\boldsymbol{\omega}$

- Observer position at middle observation: $\boldsymbol{R} = \boldsymbol{R}_2$

- Observer velocity and acceleration: $\dot{\boldsymbol{R}} = \boldsymbol{\omega}\times\boldsymbol{R}$ and $\ddot{\boldsymbol{R}} = \boldsymbol{\omega}\times\dot{\boldsymbol{R}}$

- Lagrange interpolation coefficients:

$$s_1 = \frac{t_2 - t_3}{(t_1-t_2)(t_1-t_3)}, \quad s_2 = \frac{2t_2 - t_1 - t_3}{(t_2-t_1)(t_2-t_3)}, \quad s_3 = \frac{t_2 - t_1}{(t_3-t_1)(t_3-t_2)},$$
$$s_4 = \frac{2}{(t_1-t_2)(t_1-t_3)}, \quad s_5 = \frac{2}{(t_2-t_1)(t_2-t_3)}, \quad s_6 = \frac{2}{(t_3-t_1)(t_3-t_2)}$$

- Line-of-sight and associated rates:

$$\boldsymbol{L} = \boldsymbol{L}_2 \tag{6.216}$$

$$\dot{\boldsymbol{L}} = s_1\boldsymbol{L}_1 + s_2\boldsymbol{L}_2 + s_3\boldsymbol{L}_3 \tag{6.217}$$

$$\ddot{\boldsymbol{L}} = s_4\boldsymbol{L}_1 + s_5\boldsymbol{L}_2 + s_6\boldsymbol{L}_3 \tag{6.218}$$

- Determinants:

$$\Delta_0 = 2\left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \qquad \Delta_1 = \left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \ddot{\boldsymbol{R}}\ \right|\right|$$
$$\Delta_2 = \left|\left|\ \boldsymbol{L}\ |\quad \dot{\boldsymbol{L}}\ |\quad \boldsymbol{R}\ \right|\right| \qquad \Delta_3 = \left|\left|\ \boldsymbol{L}\ |\quad \ddot{\boldsymbol{R}}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right| \qquad \Delta_4 = \left|\left|\ \boldsymbol{L}\ |\quad \boldsymbol{R}\ |\quad \ddot{\boldsymbol{L}}\ \right|\right|$$

- Polynomial coefficients:

$$a = -4(\Delta_1/\Delta_0)^2 + 4(\Delta_1/\Delta_0)\boldsymbol{L}\cdot\boldsymbol{R} - \|\boldsymbol{R}\|^2 \tag{6.219}$$
$$b = -8\mu(\Delta_1/\Delta_0)(\Delta_2/\Delta_0) + 4\mu(\Delta_2/\Delta_0)\boldsymbol{L}\cdot\boldsymbol{R} \tag{6.220}$$
$$c = -4\mu^2(\Delta_2/\Delta_0)^2 \tag{6.221}$$

- Position vector magnitude: solve

$$0 = \|\boldsymbol{r}\|^8 + a\|\boldsymbol{r}\|^6 + b\|\boldsymbol{r}\|^3 + c \tag{6.222}$$

  to obtain the applicable real root $\|\boldsymbol{r}\|$.

- Range and range-rate:

$$\rho = -2(\Delta_1/\Delta_0) - 2\mu\|\boldsymbol{r}\|^{-3}(\Delta_2/\Delta_0) \tag{6.223}$$
$$\dot{\rho} = -(\Delta_3/\Delta_0) - \mu\|\boldsymbol{r}\|^{-3}(\Delta_4/\Delta_0) \tag{6.224}$$

- Object position and velocity:

$$\boldsymbol{r} = \boldsymbol{R} + \rho\boldsymbol{L} \tag{6.225}$$
$$\dot{\boldsymbol{r}} = \dot{\boldsymbol{R}} + \dot{\rho}\boldsymbol{L} + \rho\dot{\boldsymbol{L}} \tag{6.226}$$

- Output: $\boldsymbol{r}_2 = \boldsymbol{r}$ and $\boldsymbol{v}_2 = \dot{\boldsymbol{r}}$

6.2.4 Three Position Vectors - Orbit-Based: Gibbs' Method

Let's suppose that we have three successive radar observations, giving us the line of sight measurements and the range. If we do not have a Doppler radar, no range rate measurements are available. This allows us to readily determine the position of an orbiting object at three times: $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$.

We know from previous discussions that these three positions must lie within a single plane if we assume a Keplerian orbit, which is a constraint from the conservation of angular momentum.

If this is the case, then the unit vector along $\boldsymbol{r}_1$ should be perpendicular to the plane defined by $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$, i.e.

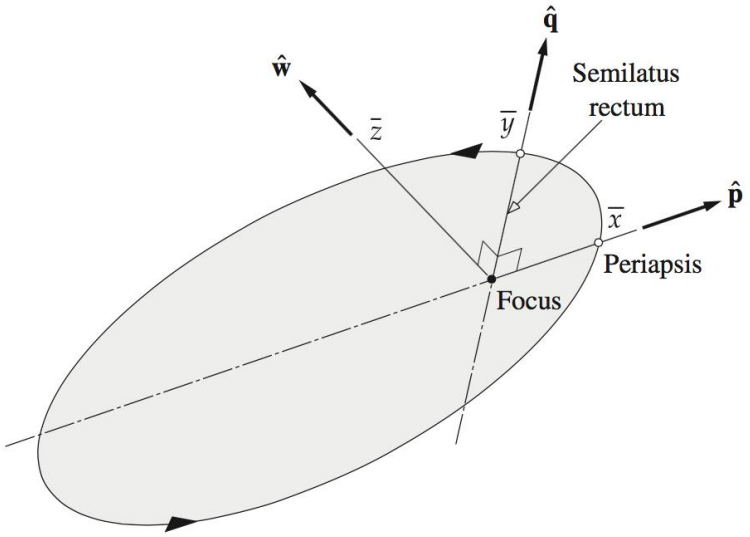
$$\frac{\boldsymbol{r}_1}{\|\boldsymbol{r}_1\|}\cdot\frac{\boldsymbol{r}_2\times\boldsymbol{r}_3}{\|\boldsymbol{r}_2\times\boldsymbol{r}_3\|} = 0 \tag{6.227}$$

Additionally, as we discussed in our treatment of Gauss' method, we should be able to find scalar factors c_1 and c_3 so that

$$\boldsymbol{r}_2 = c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3 \tag{6.228}$$

Gibbs' method makes use of some concepts and relationships that are standard in orbital mechanics, so let's briefly review those.

The first is the perifocal coordinate system, which is illustrated below.



It is a Cartesian coordinate system fixed in space and centered at the focus of the orbit.

The $x-y$ plane is in the plane of the orbit with the x-axis directed through the periapsis.

The z-axis is normal to the plane of the orbit and points along the direction of the (specific) angular momentum vector $\boldsymbol{h} = \boldsymbol{r} \times \boldsymbol{v}$.

The y-axis completes the triad.

The eccentricity vector and angular momentum vector are integrally linked to this coordinate system:

$$\boldsymbol{e} = e\hat{\boldsymbol{p}} \qquad \text{and} \qquad \boldsymbol{h} = h\hat{\boldsymbol{w}}$$

where e is the eccentricity (magnitude of the eccentricity vector) and h is the magnitude of the angular momentum vector. Another relationship that we will need is the orbit equation, which is given by

$$\|\boldsymbol{r}\| + \boldsymbol{r} \cdot \boldsymbol{e} = \frac{h^2}{\mu} \tag{6.229}$$

This is perhaps more commonly seen as

$$\|\boldsymbol{r}\| = \frac{h^2}{\mu} \frac{1}{1 + e\cos\nu} \tag{6.230}$$

where ν is the true anomaly, which is the angle between the fixed vector $\boldsymbol{e}$ and the varying vector $\boldsymbol{r}$. The other element from orbital mechanics that we will need is a relationship that links the position, velocity, eccentricity vector, and angular moment vector:

$$\boldsymbol{v} \times \boldsymbol{h} = \mu \left[\frac{\boldsymbol{r}}{\|\boldsymbol{r}\|} + \boldsymbol{e} \right] \tag{6.231}$$

This relationship holds at any time, so we can relate the position at t_1 to the velocity at t_1, and so on for the other times of interest.

Particularly, we are interested in the velocity, which can be isolated by crossing the angular momentum vector with the preceding relationship to yield

$$\boldsymbol{h} \times (\boldsymbol{v} \times \boldsymbol{h}) = \mu \left[\frac{\boldsymbol{h} \times \boldsymbol{r}}{\|\boldsymbol{r}\|} + \boldsymbol{h} \times \boldsymbol{e} \right] \tag{6.232}$$

We will now make use of the triple product identity

$$\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c}) = \boldsymbol{b}(\boldsymbol{a}\cdot\boldsymbol{c}) - \boldsymbol{c}(\boldsymbol{a}\cdot\boldsymbol{b}) \tag{6.233}$$

This is commonly referred to as the "bac-cab rule."

Applying this to the left-hand side of our previous relationship, it follows that

$$\begin{aligned}\boldsymbol{h}\times(\boldsymbol{v}\times\boldsymbol{h}) &= \boldsymbol{v}(\boldsymbol{h}\cdot\boldsymbol{h}) - \boldsymbol{h}(\boldsymbol{v}\cdot\boldsymbol{h}) &(6.234)\\ &= h^2\boldsymbol{v} - \boldsymbol{h}(\boldsymbol{v}\cdot\boldsymbol{h}) &(6.235)\\ &= h^2\boldsymbol{v} &(6.236)\end{aligned}$$

where the last equality follows from the fact that the angular momentum vector is perpendicular to the velocity vector.

At this point, we have

$$h^2\boldsymbol{v} = \mu\left[\frac{\boldsymbol{h}\times\boldsymbol{r}}{\|\boldsymbol{r}\|} + \boldsymbol{h}\times\boldsymbol{e}\right] \tag{6.237}$$

or, solving for $\boldsymbol{v}$,

$$\boldsymbol{v} = \frac{\mu}{h^2}\left[\frac{\boldsymbol{h}\times\boldsymbol{r}}{\|\boldsymbol{r}\|} + \boldsymbol{h}\times\boldsymbol{e}\right] \tag{6.238}$$

Recall that the eccentricity vector and the angular momentum vector can be expressed in terms of the perifocal coordinate system as

$$\boldsymbol{e} = e\hat{\boldsymbol{p}} \qquad \text{and} \qquad \boldsymbol{h} = h\hat{\boldsymbol{w}}$$

and substitute these expressions into the solution for the velocity to get

$$\begin{aligned}\boldsymbol{v} &= \frac{\mu}{h^2}\left[\frac{h\hat{\boldsymbol{w}}\times\boldsymbol{r}}{\|\boldsymbol{r}\|} + he\hat{\boldsymbol{w}}\times\hat{\boldsymbol{p}}\right] &(6.239)\\ &= \frac{\mu}{h}\left[\frac{\hat{\boldsymbol{w}}\times\boldsymbol{r}}{\|\boldsymbol{r}\|} + e\hat{\boldsymbol{w}}\times\hat{\boldsymbol{p}}\right] &(6.240)\end{aligned}$$

Since $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{p}}$ are perpendicular to one another, their cross product is simply the vector $\hat{\boldsymbol{q}}$, which completes the triad of the perifocal coordinate system, so

$$\boldsymbol{v} = \frac{\mu}{h}\left[\frac{\hat{\boldsymbol{w}}\times\boldsymbol{r}}{\|\boldsymbol{r}\|} + e\hat{\boldsymbol{q}}\right] \tag{6.241}$$

Now, if we can use the vectors $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ to compute h, e, $\hat{\boldsymbol{w}}$, and $\hat{\boldsymbol{q}}$, then we can compute the velocity vector for any of the positions.

Recall the planar orbit condition

$$\boldsymbol{r}_2 = c_1\boldsymbol{r}_1 + c_3\boldsymbol{r}_3 \tag{6.242}$$

and take the dot product of this equation with the eccentricity vector to get

$$\boldsymbol{e}\cdot\boldsymbol{r}_2 = c_1\boldsymbol{e}\cdot\boldsymbol{r}_1 + c_3\boldsymbol{e}\cdot\boldsymbol{r}_3 \tag{6.243}$$

From the orbit equation, i.e.

$$\|\boldsymbol{r}\| + \boldsymbol{r}\cdot\boldsymbol{e} = \frac{h^2}{\mu} \tag{6.244}$$

it follows that

$$\boldsymbol{r}_1\cdot\boldsymbol{e} = \frac{h^2}{\mu} - \|\boldsymbol{r}_1\| \qquad \boldsymbol{r}_2\cdot\boldsymbol{e} = \frac{h^2}{\mu} - \|\boldsymbol{r}_2\| \qquad \boldsymbol{r}_3\cdot\boldsymbol{e} = \frac{h^2}{\mu} - \|\boldsymbol{r}_3\| \tag{6.245}$$

Therefore, the planar orbit condition (dotted with the eccentricity vector) becomes

$$\frac{h^2}{\mu} - \|\boldsymbol{r}_2\| = c_1\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_1\|\right] + c_3\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_3\|\right] \tag{6.246}$$

We need to eliminate the c_1 and c_3 coefficients somehow.
To do this, let's first multiply our modified planar orbit condition by $(\boldsymbol{r}_3 \times \boldsymbol{r}_1)$ to get

$$\frac{h^2}{\mu}(\boldsymbol{r}_3 \times \boldsymbol{r}_1) - \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) = c_1(\boldsymbol{r}_3 \times \boldsymbol{r}_1)\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_1\|\right] + c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_1)\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_3\|\right] \tag{6.247}$$

This seems to be a step backward at first, but now we can take the original planar orbit condition and cross it with $\boldsymbol{r}_1$ or with $\boldsymbol{r}_3$, which gives the two equations

$$\boldsymbol{r}_2 \times \boldsymbol{r}_1 = c_1\boldsymbol{r}_1 \times \boldsymbol{r}_1 + c_3\boldsymbol{r}_3 \times \boldsymbol{r}_1 \qquad \text{and} \qquad \boldsymbol{r}_2 \times \boldsymbol{r}_3 = c_1\boldsymbol{r}_1 \times \boldsymbol{r}_3 + c_3\boldsymbol{r}_3 \times \boldsymbol{r}_3 \tag{6.248}$$

These equations can be reduced to yield

$$(\boldsymbol{r}_2 \times \boldsymbol{r}_1) = c_3(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \qquad \text{and} \qquad -(\boldsymbol{r}_2 \times \boldsymbol{r}_3) = c_1(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \tag{6.249}$$

The right-hand sides appear exactly in our modified planar orbit condition!
Now, we substitute the preceding relationships back into our modified planar orbit condition to find

$$\frac{h^2}{\mu}(\boldsymbol{r}_3 \times \boldsymbol{r}_1) - \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) = -(\boldsymbol{r}_2 \times \boldsymbol{r}_3)\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_1\|\right] + (\boldsymbol{r}_2 \times \boldsymbol{r}_1)\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_3\|\right] \tag{6.250}$$

We have now eliminated any appearance of c_1 and c_3 in our equation.
Let's now rearrange the terms to collect all of the terms multiplying h^2/μ on the left-hand side and all of the other terms on the right-hand side. This gives

$$\frac{h^2}{\mu}\left[(\boldsymbol{r}_3 \times \boldsymbol{r}_1) + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) - (\boldsymbol{r}_2 \times \boldsymbol{r}_1)\right] \tag{6.251}$$

$$= \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) + (\boldsymbol{r}_2 \times \boldsymbol{r}_3)\|\boldsymbol{r}_1\| - (\boldsymbol{r}_2 \times \boldsymbol{r}_1)\|\boldsymbol{r}_3\| \tag{6.252}$$

or, after reversing cross products and rearranging terms,

$$\frac{h^2}{\mu}\left[(\boldsymbol{r}_1 \times \boldsymbol{r}_2) + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) + (\boldsymbol{r}_3 \times \boldsymbol{r}_1)\right] \tag{6.253}$$

$$= \|\boldsymbol{r}_1\|(\boldsymbol{r}_2 \times \boldsymbol{r}_3) + \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) + \|\boldsymbol{r}_3\|(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \tag{6.254}$$

To simplify this expression, let's define a few terms:

$$\boldsymbol{d} = (\boldsymbol{r}_1 \times \boldsymbol{r}_2) + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) + (\boldsymbol{r}_3 \times \boldsymbol{r}_1) \tag{6.255}$$

$$\boldsymbol{n} = \|\boldsymbol{r}_1\|(\boldsymbol{r}_2 \times \boldsymbol{r}_3) + \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) + \|\boldsymbol{r}_3\|(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \tag{6.256}$$

Note that, given the three position vectors, $\boldsymbol{n}$ and $\boldsymbol{d}$ can be readily determined.
Now, we can write the modified orbit condition as

$$\frac{h^2}{\mu}\boldsymbol{d} = \boldsymbol{n} \tag{6.257}$$

or, by taking the norm of each side,

$$\frac{h^2}{\mu}\|\boldsymbol{d}\| = \|\boldsymbol{n}\| \tag{6.258}$$

Now, we can solve for the magnitude of the angular momentum vector as

$$h = \sqrt{\mu \frac{\|\boldsymbol{n}\|}{\|\boldsymbol{d}\|}} \tag{6.259}$$

Remember that we're trying to find h, e, $\hat{\boldsymbol{w}}$, and $\hat{\boldsymbol{q}}$ in terms of $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ so that we can solve for the velocity via

$$\boldsymbol{v} = \frac{\mu}{h}\left[\frac{\hat{\boldsymbol{w}} \times \boldsymbol{r}}{\|\boldsymbol{r}\|} + e\hat{\boldsymbol{q}}\right] \tag{6.260}$$

We now have h in terms of the positions!

Since $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ all lie in a single plane, the cross products $(\boldsymbol{r}_1 \times \boldsymbol{r}_2)$, $(\boldsymbol{r}_2 \times \boldsymbol{r}_3)$, and $(\boldsymbol{r}_3 \times \boldsymbol{r}_1)$ lie in the same direction, which is normal to the orbital plane.

Therefore, $\boldsymbol{d}$ must be normal to the orbital plane; thus,

$$\hat{\boldsymbol{w}} = \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \tag{6.261}$$

Now we have h and $\hat{\boldsymbol{w}}$ in terms of the positions!

To find $\hat{\boldsymbol{q}}$, first note that

$$\hat{\boldsymbol{q}} = \hat{\boldsymbol{w}} \times \hat{\boldsymbol{p}} = \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \times \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} = \frac{1}{e\|\boldsymbol{d}\|}(\boldsymbol{d} \times \boldsymbol{e}) \tag{6.262}$$

Let's substitute for our definition of $\boldsymbol{d}$, which gives

$$\hat{\boldsymbol{q}} = \frac{1}{e\|\boldsymbol{d}\|}\left[(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} + (\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e}\right] \tag{6.263}$$

We can applying the bac-cab rule by noting

$$(\boldsymbol{a} \times \boldsymbol{b}) \times \boldsymbol{c} = -\boldsymbol{c} \times (\boldsymbol{a} \times \boldsymbol{b}) \tag{6.264}$$

$$= \boldsymbol{b}(\boldsymbol{a} \cdot \boldsymbol{c}) - \boldsymbol{a}(\boldsymbol{b} \cdot \boldsymbol{c}) \tag{6.265}$$

From this form of the bac-cab rule, we find that

$$(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} = \boldsymbol{r}_2(\boldsymbol{r}_1 \cdot \boldsymbol{e}) - \boldsymbol{r}_1(\boldsymbol{r}_2 \cdot \boldsymbol{e}) \tag{6.266}$$

$$(\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} = \boldsymbol{r}_3(\boldsymbol{r}_2 \cdot \boldsymbol{e}) - \boldsymbol{r}_2(\boldsymbol{r}_3 \cdot \boldsymbol{e}) \tag{6.267}$$

$$(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e} = \boldsymbol{r}_1(\boldsymbol{r}_3 \cdot \boldsymbol{e}) - \boldsymbol{r}_3(\boldsymbol{r}_1 \cdot \boldsymbol{e}) \tag{6.268}$$

Recall from the orbit equation that we can express the dot product of the position with the eccentricity vector as

$$\boldsymbol{r}_i \cdot \boldsymbol{e} = \frac{h^2}{\mu} - \|\boldsymbol{r}_i\| \tag{6.269}$$

for $i \in \{1,2,3\}$.

Using the orbit equation relationship, we can write

$$(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} = \boldsymbol{r}_2\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_1\|\right] - \boldsymbol{r}_1\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_2\|\right] \tag{6.270}$$

$$(\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} = \boldsymbol{r}_3\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_2\|\right] - \boldsymbol{r}_2\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_3\|\right] \tag{6.271}$$

$$(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e} = \boldsymbol{r}_1\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_3\|\right] - \boldsymbol{r}_3\left[\frac{h^2}{\mu} - \|\boldsymbol{r}_1\|\right] \tag{6.272}$$

A simple rearrangement then gives

$$(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} = \frac{h^2}{\mu}\left[\boldsymbol{r}_2 - \boldsymbol{r}_1\right] + \|\boldsymbol{r}_2\|\boldsymbol{r}_1 - \|\boldsymbol{r}_1\|\boldsymbol{r}_2 \tag{6.273}$$

$$(\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} = \frac{h^2}{\mu}\left[\boldsymbol{r}_3 - \boldsymbol{r}_2\right] + \|\boldsymbol{r}_3\|\boldsymbol{r}_2 - \|\boldsymbol{r}_2\|\boldsymbol{r}_3 \tag{6.274}$$

$$(\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e} = \frac{h^2}{\mu}\left[\boldsymbol{r}_1 - \boldsymbol{r}_3\right] + \|\boldsymbol{r}_1\|\boldsymbol{r}_3 - \|\boldsymbol{r}_3\|\boldsymbol{r}_1 \tag{6.275}$$

Now, we can add up the three preceding equations to find that

$$(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} + (\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e} \tag{6.276}$$

$$= \boldsymbol{r}_1\left[\|\boldsymbol{r}_2\| - \|\boldsymbol{r}_3\|\right] + \boldsymbol{r}_2\left[\|\boldsymbol{r}_3\| - \|\boldsymbol{r}_1\|\right] + \boldsymbol{r}_3\left[\|\boldsymbol{r}_1\| - \|\boldsymbol{r}_2\|\right] \tag{6.277}$$

Recall that we're working on an expression for $\hat{\boldsymbol{q}}$ and that we previously had shown that

$$\hat{\boldsymbol{q}} = \frac{1}{e\|\boldsymbol{d}\|}\left[(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \times \boldsymbol{e} + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) \times \boldsymbol{e} + (\boldsymbol{r}_3 \times \boldsymbol{r}_1) \times \boldsymbol{e}\right] \tag{6.278}$$

We just arrived at a new expression for the term in square brackets on the right-hand side that did not contain the eccentricity vector, so we can write $\hat{\boldsymbol{q}}$ as

$$\hat{\boldsymbol{q}} = \frac{1}{e\|\boldsymbol{d}\|}\boldsymbol{s} \tag{6.279}$$

where

$$\boldsymbol{s} = \boldsymbol{r}_1\left[\|\boldsymbol{r}_2\| - \|\boldsymbol{r}_3\|\right] + \boldsymbol{r}_2\left[\|\boldsymbol{r}_3\| - \|\boldsymbol{r}_1\|\right] + \boldsymbol{r}_3\left[\|\boldsymbol{r}_1\| - \|\boldsymbol{r}_2\|\right] \tag{6.280}$$

It is important to note that $\boldsymbol{s}$, much like $\boldsymbol{n}$ and $\boldsymbol{d}$, can be found based solely upon the known position vectors.
We still have the eccentricity in our solution for $\hat{\boldsymbol{q}}$, but this is going to turn out to work just fine for us.
Ultimately, we're still trying to find the velocity from

$$\boldsymbol{v} = \frac{\mu}{h}\left[\frac{\hat{\boldsymbol{w}} \times \boldsymbol{r}}{\|\boldsymbol{r}\|} + e\hat{\boldsymbol{q}}\right] \tag{6.281}$$

and we have shown that

$$h = \sqrt{\mu\frac{\|\boldsymbol{n}\|}{\|\boldsymbol{d}\|}} \qquad \hat{\boldsymbol{w}} = \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \qquad \hat{\boldsymbol{q}} = \frac{1}{e\|\boldsymbol{d}\|}\boldsymbol{s} \tag{6.282}$$

Therefore, substituting h, $\hat{\boldsymbol{w}}$, and $\hat{\boldsymbol{q}}$ into the expression for $\boldsymbol{v}$ yields

$$\boldsymbol{v} = \frac{\mu}{\sqrt{\mu\frac{\|\boldsymbol{n}\|}{\|\boldsymbol{d}\|}}}\left[\frac{\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \times \boldsymbol{r}}{\|\boldsymbol{r}\|} + e\frac{1}{e\|\boldsymbol{d}\|}\boldsymbol{s}\right] \tag{6.283}$$

After we simplify the expression, we find the velocity to be given by

$$\boldsymbol{v} = \sqrt{\frac{\mu}{\|\boldsymbol{n}\| \cdot \|\boldsymbol{d}\|}}\left[\frac{\boldsymbol{d} \times \boldsymbol{r}}{\|\boldsymbol{r}\|} + \boldsymbol{s}\right] \tag{6.284}$$

If we want to find the velocity at a certain time, say t_2, then we just evaluate this expression at the corresponding position; this means that

$$\boldsymbol{v}_2 = \sqrt{\frac{\mu}{\|\boldsymbol{n}\| \cdot \|\boldsymbol{d}\|}}\left[\frac{\boldsymbol{d} \times \boldsymbol{r}_2}{\|\boldsymbol{r}_2\|} + \boldsymbol{s}\right] \tag{6.285}$$

It is worth mentioning and remembering that $\boldsymbol{n}$, $\boldsymbol{d}$, and $\boldsymbol{s}$ are *all* functions solely of our three position vectors, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$.

This means that all of the terms appearing in the expression for the velocity can be found from the position data that is known from the beginning.

Gibbs' method did not actually make use of the time information.

6.2.4.1 Algorithm for Gibbs' Method

- Given: $\boldsymbol{r}_i$ for $i \in \{1,2,3\}$

- Verify that $(\boldsymbol{r}_1/\|\boldsymbol{r}_1\|) \cdot (\boldsymbol{r}_2 \times \boldsymbol{r}_3/\|\boldsymbol{r}_2 \times \boldsymbol{r}_3\|) = 0$.

- Calculate $\boldsymbol{n}$, $\boldsymbol{d}$, and $\boldsymbol{s}$ via

$$\boldsymbol{n} = \|\boldsymbol{r}_1\|(\boldsymbol{r}_2 \times \boldsymbol{r}_3) + \|\boldsymbol{r}_2\|(\boldsymbol{r}_3 \times \boldsymbol{r}_1) + \|\boldsymbol{r}_3\|(\boldsymbol{r}_1 \times \boldsymbol{r}_2) \tag{6.286}$$

$$\boldsymbol{d} = (\boldsymbol{r}_1 \times \boldsymbol{r}_2) + (\boldsymbol{r}_2 \times \boldsymbol{r}_3) + (\boldsymbol{r}_3 \times \boldsymbol{r}_1) \tag{6.287}$$

$$\boldsymbol{s} = \boldsymbol{r}_1\big[\|\boldsymbol{r}_2\| - \|\boldsymbol{r}_3\|\big] + \boldsymbol{r}_2\big[\|\boldsymbol{r}_3\| - \|\boldsymbol{r}_1\|\big] + \boldsymbol{r}_3\big[\|\boldsymbol{r}_1\| - \|\boldsymbol{r}_2\|\big] \tag{6.288}$$

- Output: $\boldsymbol{v}_2$, which is calculated as

$$\boldsymbol{v}_2 = \sqrt{\frac{\mu}{\|\boldsymbol{n}\| \cdot \|\boldsymbol{d}\|}}\left[\frac{\boldsymbol{d} \times \boldsymbol{r}_2}{\|\boldsymbol{r}_2\|} + \boldsymbol{s}\right] \tag{6.289}$$

6.2.5 Three (Close) Position Vectors - Averaging: Herrick-Gibbs' Method

We assume three radar observations (angles and range) are given, which let's us readily compute three position vectors, $\boldsymbol{r}_i, i \in 1,2,3$ at times $t_i, i \in 1,2,3$.

When we can determine the velocity of one of the observations, we have a full state and are finished with our orbit determination. Herrick-Gibbs Method relies on Taylor series expansion and on the assumption of co-planar observations. A real orbit is never a Keplerian orbit, this assumption is hence always violated. The question is to which extentwith and that condition is violated. When we only have three optical observations and are using Gauss' method, we have no means to check on this condition a priori but only a posteriori. Now that we have the three position vectors we can state the limits under which the method performs well a priori:

- Deviation from Keplerian orbit

$$\boldsymbol{z}_{23} = \boldsymbol{r}_2 \times \boldsymbol{r}_3 \qquad \boldsymbol{z}_{23} \cdot \boldsymbol{r}_1 = \boldsymbol{0} \qquad \text{exact for Keplerian orbit} \tag{6.290}$$

$$\phi = \frac{\pi}{2} - \arccos\left(\frac{\boldsymbol{z}_{23} \cdot \boldsymbol{r}_1}{\|\boldsymbol{z}_{23}\| \cdot \|\boldsymbol{r}_1\|}\right) \leq 3^\circ \text{for HG IOD} \tag{6.291}$$

- small angular separation

$$\cos\phi_{12} = \frac{\boldsymbol{r}_1 \cdot \boldsymbol{r}_2}{\|\boldsymbol{r}_1\| \cdot \|\boldsymbol{r}_2\|} \qquad ;\cos\phi_{23} = \frac{\boldsymbol{r}_2 \cdot \boldsymbol{r}_3}{\|\boldsymbol{r}_2\| \cdot \|\boldsymbol{r}_3\|} \tag{6.292}$$

$$\phi_{12}, \phi_{23} < 20^\circ \tag{6.293}$$

First step: Taylor series expansion around the middle position:

$$\boldsymbol{r}_i(t_i) = \boldsymbol{r}_2(t_i) + \dot{\boldsymbol{r}}_2(t_i - t_2) + \frac{\ddot{\boldsymbol{r}}_2(t_i - t_2)^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2(t_i - t_2)^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2(t_i - t_2)^4}{4!} + ... \tag{6.294}$$

Denoting the time differences with $\Delta_{ij} = t_i - t_j$ we can expand both r_1 and r_3:

$$\boldsymbol{r}_1 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{12} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{12}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{12}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{12}^4}{4!} \tag{6.295}$$

$$\boldsymbol{r}_3 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{32} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{32}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{32}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{32}^4}{4!} \tag{6.296}$$

The trick is now to multiply both equations subsequently with multipliers of the time difference and add them in order to subsequently eliminate different orders of the derivatives. Single derivatives can be computed using two-body equations.

Step 2: Multiply Eq.6.295 and E.6.296 with

$$\boldsymbol{r}_1 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{12} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{12}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{12}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{12}^4}{4!} \qquad |\cdot(-\Delta t_{32}^2) \tag{6.297}$$

$$\boldsymbol{r}_3 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{32} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{32}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{32}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{32}^4}{4!} \qquad |\cdot\Delta t_{12}^2 \tag{6.298}$$

and addition of both equations leads to:

$$\begin{aligned} -\boldsymbol{r}_1\Delta t_{32}^2 + \boldsymbol{r}_3\Delta t_{12}^2 = \quad & \boldsymbol{r}_2(\Delta t_{12}^2 - \Delta t_{32}^2) \\ & \dot{\boldsymbol{r}}_2(\Delta t_{12}^2\Delta t_{32} - \Delta t_{12}\Delta t_{32}^2) \\ & \ddot{\boldsymbol{r}}_2(\Delta t_{12}^2\Delta t_{32}^2 - \Delta t_{12}^2\Delta t_{32}^2)\cdot\frac{1}{2} \\ & \dddot{\boldsymbol{r}}_2(\Delta t_{12}^2\Delta t_{32}^3 - \Delta t_{12}^3\Delta t_{32}^2)\cdot\frac{1}{6} \\ & \ddddot{\boldsymbol{r}}_2(\Delta t_{12}^2\Delta t_{32}^4 - \Delta t_{12}^4\Delta t_{32}^2)\cdot\frac{1}{24} \end{aligned} \tag{6.299}$$

Lets take a look at the single terms:

$$(\Delta t_{12}^2\Delta t_{32} - \Delta t_{12}\Delta t_{32}^2) = \quad \Delta t_{12}\Delta t_{32}(\Delta t_{12} - \Delta t_{32}) = \Delta t_{12}\Delta t_{32}\Delta t_{13} \tag{6.300}$$

$$(\Delta t_{12}^2\Delta t_{32}^2 - \Delta t_{12}^2\Delta t_{32}^2) = \quad 0 \tag{6.301}$$

$$(\Delta t_{12}^2\Delta t_{32}^3 - \Delta t_{12}^3\Delta t_{32}^2) = \quad \Delta t_{12}^2\Delta t_{32}^3(\Delta t_{32} - \Delta t_{12}) = \Delta t_{12}^2\Delta t_{32}^3\Delta t_{31} \tag{6.302}$$

$$\begin{aligned} (\Delta t_{12}^2\Delta t_{32}^4 - \Delta t_{12}^4\Delta t_{32}^2) = \quad & \Delta t_{12}^2\Delta t_{32}^2(\Delta t_{32}^2 - \Delta t_{12}^2) \\ & = \Delta t_{12}^2\Delta t_{32}^2(t_3^2 - t_2^2 - 2t_3t_2 - t_1^2 + t_2^2 + 2t_1t_2) \\ & = \Delta t_{12}^2\Delta t_{32}^2(t_3^2 - 2t_3t_2 + 2t_1t_2 - t_1^2) \\ & \text{completing the square, adding } +t_1t_3 - t_1t_3 \\ & = \Delta t_{12}^2\Delta t_{32}^2\Delta t_{31}(\Delta t_{12} + \Delta t_{32}) \end{aligned} \tag{6.303}$$

Plugging everything back in:

$$\begin{aligned} \dot{\boldsymbol{r}}_2\Delta t_{12}\Delta t_{32}\Delta t_{31} = \quad & \boldsymbol{r}_1\Delta t_{32}^2 + \boldsymbol{r}_2(\Delta t_{12}^2 - \Delta t_{32}^2) - \boldsymbol{r}_3\Delta t_{12}^2 \\ & \frac{\dddot{\boldsymbol{r}}_2}{6}(\Delta t_{12}^2\Delta t_{32}^2\Delta t_{31}) \\ & \frac{\ddddot{\boldsymbol{r}}_2}{24}\Delta t_{12}^2\Delta t_{32}^2\Delta t_{31}(\Delta t_{12} + \Delta t_{32}) \end{aligned} \tag{6.304}$$

The first term (magenta color) in the equation above is already known, what is missing is the third and the fourth derivative of the middle vector in order to determine the velocity at time t_2.

Step 3: In order to compute the other derivatives the same procedure is repeated again, eliminating the third and the fourth derivative in taking the derivative of the initial equations, while preserving only terms to fourth order. Afterwards, multiply with the appropriate time differences again. Hence starting again with Eq.6.295 and 6.296:

$$\boldsymbol{r}_1 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{12} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{12}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{12}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{12}^4}{4!} \tag{6.305}$$

$$\boldsymbol{r}_3 = \boldsymbol{r}_2 + \dot{\boldsymbol{r}}_2\Delta t_{32} + \frac{\ddot{\boldsymbol{r}}_2\Delta t_{32}^2}{2!} + \frac{\dddot{\boldsymbol{r}}_2\Delta t_{32}^3}{3!} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{32}^4}{4!} \tag{6.306}$$

Differentiating twice and preserving fourth order derivatives (!):

$$\ddot{\boldsymbol{r}}_1 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2\Delta t_{12} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{12}^2}{2!} + .. \tag{6.307}$$

$$\ddot{\boldsymbol{r}}_3 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2\Delta t_{32} + \frac{\ddddot{\boldsymbol{r}}_2\Delta t_{32}^2}{2!} + ... \tag{6.308}$$

Multiplying the first equation with $-\Delta t_{32}$ and the second with Δt_{12} and adding them:

$$\ddot{\boldsymbol{r}}_1 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{12} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{12}^2}{2!} \qquad |\cdot(-\Delta t_{32}) \tag{6.309}$$

$$\ddot{\boldsymbol{r}}_3 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{32} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{32}^2}{2!} \qquad |\cdot \Delta t_{12} \tag{6.310}$$

leads to:

$$\begin{aligned} \ddot{\boldsymbol{r}}_3 \Delta t_{12} - \ddot{\boldsymbol{r}}_1 \Delta t_{32} = \quad & \ddot{\boldsymbol{r}}_2 (\Delta t_{12} - \Delta t_{32}) \\ & \dddot{\boldsymbol{r}}_2 (\Delta t_{12}\Delta t_{32} - \Delta t_{12}\Delta t_{32}) \\ & \ddddot{\boldsymbol{r}}_2 (\Delta t_{12}\Delta t_{32}^2 - \Delta t_{12}^2 \Delta t_{32}) \cdot \frac{1}{2} \end{aligned} \tag{6.311}$$

The second term is obviously zero, the rest lead to:

$$(\Delta t_{12} - \Delta t_{32}) = \quad \Delta t_{13} \tag{6.312}$$

$$(\Delta t_{12}\Delta t_{32}^2 - \Delta t_{12}^2 \Delta t_{32}) = \quad \Delta t_{12}\Delta t_{32}(\Delta t_{32} - \Delta t_{12}) = \Delta t_{12}\Delta t_{32}\Delta t_{31} \tag{6.313}$$

substitute everything back into Eq.6.311:

$$\ddddot{\boldsymbol{r}}_2 = \frac{2}{\Delta t_{12}\Delta t_{32}\Delta t_{31}} (-\ddot{\boldsymbol{r}}_1 \Delta t_{32} + \ddot{\boldsymbol{r}}_2 \Delta t_{31} + \ddot{\boldsymbol{r}}_3 \Delta t_{12}) \tag{6.314}$$

The right (magenta colored) term is completely known using two body dynamics:

$$\ddot{\boldsymbol{r}}_i = -\frac{\mu}{\|\boldsymbol{r}_i\|^3} \boldsymbol{r}_i \tag{6.315}$$

and the left hand side is the quantity we needed to solve for the velocity $\dot{\boldsymbol{r}}_2$.

Step 4: ok, all we are missing now is an expression for the third derivative. And guess what, we are going back to equations 6.295 and 6.296, differentiate twice, keep only the terms up to order four and choose the appropriate factors. Repeating Eq.6.307 and 6.308:

$$\ddot{\boldsymbol{r}}_1 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{12} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{12}^2}{2!} + .. \tag{6.316}$$

$$\ddot{\boldsymbol{r}}_3 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{32} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{32}^2}{2!} + ... \tag{6.317}$$

Multiplying by $-\Delta t_{32}^2$ and Δt_{12}^2:

$$\ddot{\boldsymbol{r}}_1 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{12} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{12}^2}{2!} \qquad |\cdot(-\Delta t_{32}^2) \tag{6.318}$$

$$\ddot{\boldsymbol{r}}_3 = \ddot{\boldsymbol{r}}_2 + \dddot{\boldsymbol{r}}_2 \Delta t_{32} + \frac{\ddddot{\boldsymbol{r}}_2 \Delta t_{32}^2}{2!} \qquad |\cdot \Delta t_{12}^2 \tag{6.319}$$

and adding the equations leads to:

$$\begin{aligned} \ddot{\boldsymbol{r}}_3 \Delta t_{12}^2 - \ddot{\boldsymbol{r}}_1 \Delta t_{32}^2 = \quad & \ddot{\boldsymbol{r}}_2 (\Delta t_{12}^2 - \Delta t_{32}^2) \\ & \dddot{\boldsymbol{r}}_2 (\Delta t_{12}^2 \Delta t_{32} - \Delta t_{12}\Delta t_{32}^2) \\ & \ddddot{\boldsymbol{r}}_2 (\Delta t_{12}^2 \Delta t_{32}^2 - \Delta t_{12}^2 \Delta t_{32}^2) \cdot \frac{1}{2} \end{aligned} \tag{6.320}$$

The last term is obviously zero. The others can be modified according to the following:

$$\begin{aligned} (\Delta t_{12}^2 - \Delta t_{32}^2) \quad & = (t_1 - t_2)(t_1 - t_2) - (t_3 - t_2)(t_3 - t_2) \\ & = (t_1 - t_2 + t_3 - t_3)(t_1 - t_2) - (t_3 - t_2 + t_1 - t_1)(t_3 - t_2) \\ & = (t_1 - t_2)((t_3 - t_2) - (t_3 - t_1)) - (t_3 - t_2)((t_3 - t_1) + (t_1 - t_2)) \\ & = -(t_3 - t_1)((t_3 - t_2) + (t_1 - t_2)) \\ & = -\Delta t_{31}(\Delta t_{32} + \Delta t_{12}) \end{aligned} \tag{6.321}$$

$$\begin{aligned} (\Delta t_{12}^2 \Delta t_{32} - \Delta t_{12}\Delta t_{32}^2) \quad & = \Delta t_{32}\Delta t_{12}(\Delta t_{12} - \Delta t_{32}) \\ & = \Delta t_{12}\Delta t_{32}\Delta t_{13} \end{aligned} \tag{6.322}$$

Substituting everything back leads to:

$$\dddot{\boldsymbol{r}}_2 = \frac{1}{\Delta t_{12}\Delta t_{32}\Delta t_{13}}(-\ddot{\boldsymbol{r}}_1\Delta t_{32}^2 + \ddot{\boldsymbol{r}}_2\Delta t_{31}(\Delta t_{12}+\Delta t_{32}) + \ddot{\boldsymbol{r}}_3\Delta t_{12}^2) \tag{6.323}$$

where again the right side is completely known.

Now we are ready to substitute everything back into our original equation for $\dot{\boldsymbol{r}}_2$ Eq.6.304 the quantity that we are missing in order to have one full state at time t_2.

$$\begin{aligned}\dot{\boldsymbol{r}}_2 = \quad & -\Delta t_{32}\left(\frac{1}{\Delta t_{21}\Delta t_{31}} + \frac{\mu}{12\|\boldsymbol{r}_1\|^3}\right)\boldsymbol{r}_1 \\ & +(\Delta t_{32}-\Delta t_{21})\left(\frac{1}{\Delta t_{21}\Delta t_{32}} + \frac{\mu}{12\|\boldsymbol{r}_2\|^3}\right))\boldsymbol{r}_2 \\ & \Delta t_{21}\left(\frac{1}{\Delta t_{32}\Delta t_{31}} + \frac{\mu}{12\|\boldsymbol{r}_3\|^3}\right)\boldsymbol{r}_3\end{aligned} \tag{6.324}$$

6.3 Probabilistic Methods

6.3.1 Admissible Regions

In the previous sections we determined a full orbit based on a number of observations. It was either all six orbital elements (full state of the object), or a restricted orbit, when not enough information was available. In recent years, with the raise of probabilistic tracking methods another method of dealing with the initial orbit determination problem has become popular. It was initially invented and brought forth by Andrea Milani (University of Pisa) [70, 29]. As a reference also [23] has been used in this section.

The underlying thought is the following: If we have just a single observation, a so-called attributable, e.g. angles and angular rates $(\alpha, \dot{\alpha}, \delta, \dot{\delta})$ we cannot determine a full orbit, because we do not have enough information. However, can constraint the possible orbits? Or in other words, can we put constraints on possible ranges and range rates $(\rho, \dot{\rho})$ under certain premises? If we have the angles, angular rates, range, range rates we have a full state/orbit. The region of possible ranges and ranges rates for an optical attributable, and the possible angular rates for a Doppler radar attributable is called admissible region.

We focus on the optical admissible region and assume, the topocentric right ascension, α, the declination, δ, the time rate of change of the right ascension, $\dot{\alpha}$, and the time rate of change of the declination, $\dot{\delta}$, are made available via an optical observation. This forms our attributable $\alpha, \delta, \dot{\alpha}, \dot{\delta}$.

If we constrain ourselves to Earth orbiting objects (**First Assumption**), the two-body internal energy is given as:

$$\mathscr{E} = \frac{\|\dot{\boldsymbol{r}}\|^2}{2} - \frac{\mu}{\|\boldsymbol{r}\|}$$

where μ is the gravitational parameter of the central body, $\boldsymbol{r}$ is the inertial position of the object with respect to the Earth center (ECI), and $\dot{\boldsymbol{r}}$ is the inertial velocity of the object in ECI. Since the optical observation is made from a ground station, the position of the object with respect to the Earth center is given as the sum of the position of the ground station and the position of the object with respect to the station, and likewise for the velocities:

$$\boldsymbol{r} = \boldsymbol{R} + \boldsymbol{\rho} \qquad \text{and} \qquad \dot{\boldsymbol{r}} = \dot{\boldsymbol{R}} + \dot{\boldsymbol{\rho}}$$

where $\boldsymbol{R}$ is the inertial position of the ground station, $\dot{\boldsymbol{r}}$ is the inertial velocity of the ground station, $\boldsymbol{\rho}$ is the topocentric of the object with respect to the station, and $\dot{\boldsymbol{\rho}}$ is the velocity of the object with respect to the station, both in topocentric equatorial system. Now, let the position and the velocity of the object with respect to the station be given in the spherical coordinates of range, ρ, right ascension, α, declination, δ, and their time rates of change, such that

$$\boldsymbol{\rho} = \rho \boldsymbol{u}_\rho \qquad \text{and} \qquad \dot{\boldsymbol{\rho}} = \dot{\rho}\boldsymbol{u}_\rho + \rho\dot{\alpha}\boldsymbol{u}_\alpha + \rho\dot{\delta}\boldsymbol{u}_\delta$$

where the vectors of $\boldsymbol{u}_\rho$, $\boldsymbol{u}_\alpha$, and $\boldsymbol{u}_\delta$ are given by

$$\boldsymbol{u}_\rho = \begin{bmatrix} \cos\alpha\cos\delta \\ \sin\alpha\cos\delta \\ \sin\delta \end{bmatrix}, \quad \boldsymbol{u}_\alpha = \begin{bmatrix} -\sin\alpha\cos\delta \\ \cos\alpha\cos\delta \\ 0 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{u}_\delta = \begin{bmatrix} -\cos\alpha\sin\delta \\ -\sin\alpha\sin\delta \\ \cos\delta \end{bmatrix}$$

Assuming right ascension and the declination are known from our measurement, as part of the attributable. This means that we can compute the vectors; however, we do not know the range or the range-rate. Let's define a set of scalar values as

$$w_0 = \|\boldsymbol{R}\|^2, \quad w_1 = 2\left(\dot{\boldsymbol{R}}\cdot\boldsymbol{u}_\rho\right), \quad w_2 = \dot{\alpha}^2\cos^2\delta + \dot{\delta}^2,$$
$$w_3 = 2\dot{\alpha}\left(\dot{\boldsymbol{R}}\cdot\boldsymbol{u}_\alpha\right) + 2\dot{\delta}\left(\dot{\boldsymbol{R}}\cdot\boldsymbol{u}_\delta\right), \quad w_4 = \left\|\dot{\boldsymbol{R}}\right\|^2, \quad \text{and} \quad w_5 = 2\left(\boldsymbol{R}\cdot\boldsymbol{u}_\rho\right)$$

With these scalar values, the squared Euclidean norms of the position and velocity of the object with respect to the Earth center can be written as

$$\|\boldsymbol{r}\|^2 = \rho^2 + w_5\rho + w_0$$
$$\|\dot{\boldsymbol{r}}\|^2 = \dot{\rho}^2 + w_1\dot{\rho} + w_2\rho^2 + w_3\rho + w_4$$

Substituting the squared norms of the position and velocity into the two-body energy equation, it follows that we can express twice the energy as

$$2\mathscr{E} = \dot{\rho}^2 + w_1\dot{\rho} + F(\rho)$$

where

$$F(\rho) = w_2\rho^2 + w_3\rho + w_4 - \frac{2\mu}{\sqrt{\rho^2 + w_5\rho + w_0}}$$

We can rewrite the energy equation in standard quadratic form by subtracting $2\mathscr{E}$ from both sides, such that

$$\dot{\rho}^2 + w_1\dot{\rho} + F(\rho) - 2\mathscr{E} = 0$$

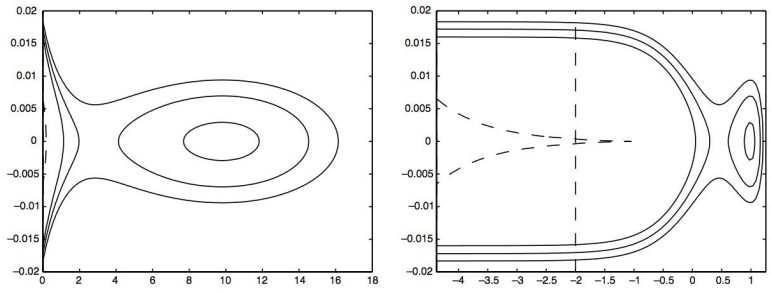


Figure 6.6: Mapping of the possible admissible regions in range and range rate for two different orbits. In the optical case there are at most two connected components.

Therefore, given a value of ρ, we can solve the preceding equation for $\dot{\rho}$ yielding two solutions as

$$\dot{\rho} = -\frac{w_1}{2} \pm \sqrt{\left(\frac{w_1}{2}\right)^2 - F(\rho) + 2\mathscr{E}}$$

If we specify that $\mathscr{E} = 0$, we obtain the zero-energy curve.Since all Earth-orbiting objects must have negative orbital energy, the zero-energy curve in range/range-rate space describes the region of all location of range and range-rate that, when paired with the measurements of the right ascension, declination, and their rates, leads to orbits bound to the Earth.

There are at most two connected components in the range/range-rate plane for curves of constant energy. Following is an example by Milani that illustrates the appearance of two connected components, see Fig6.6.
The admissible region is a rather large region that maps out the possible values. Every point within the region represents one possible orbit.

6.3.1.1 Constrained Admissible Region

It is sometimes desired to add constraints to the admissible region in order to reduce the possible combinations of range/range-rate pairs that lead to permissible orbit solutions. A wide variety of constraints can be considered, such as minimum periapse altitude or minimum range. We will focus on two constraints here: constraints on the semi-major axis of the orbit and constraints on the eccentricity of the orbit.

6.3.1.1.1 Semi-Major Axis The first constraint we will consider is a constraint on the semi-major axis, or equivalently energy since the two are related by

$$\mathscr{E} = -\frac{\mu}{2a}$$

where a is the semi-major axis. By setting a value for the semi-major axis, an equivalent energy value may be determined. Then, by using this value of energy, the admissible region procedure may be used to solve for range-rate given range, which yields a curve of constant semi-major axis in the range/range-rate space.

6.3.1.1.2 Eccentricity Another potential constraint is that of the orbit eccentricity. To develop the eccentricity constraint, first consider the specific angular momentum as

$$\boldsymbol{h} = \boldsymbol{r} \times \dot{\boldsymbol{r}}$$

We will define some vector parameters as

$$\boldsymbol{h}_1 = \boldsymbol{R} \times \boldsymbol{u}_\rho \,, \quad \boldsymbol{h}_2 = \boldsymbol{u}_\rho \times (\dot{\alpha}\boldsymbol{u}_\alpha + \dot{\delta}\boldsymbol{u}_\delta) \,,$$
$$\boldsymbol{h}_3 = \boldsymbol{u}_\rho \times \dot{\boldsymbol{R}} + \boldsymbol{R} \times (\dot{\alpha}\boldsymbol{u}_\alpha + \dot{\delta}\boldsymbol{u}_\delta) \,, \quad \text{and} \quad \boldsymbol{h}_4 = \boldsymbol{R} \times \dot{\boldsymbol{R}}$$

Then, it can be shown that the specific angular momentum is given by

$$\boldsymbol{h} = \boldsymbol{h}_1\dot{\rho} + \boldsymbol{h}_2\rho^2 + \boldsymbol{h}_3\rho + \boldsymbol{h}_4$$

Next, we define a set of scalar parameters as

$$c_0 = \|\boldsymbol{h}_1\|^2 \,, \quad c_1 = 2\boldsymbol{h}_1 \cdot \boldsymbol{h}_2 \,, \quad c_2 = 2\boldsymbol{h}_1 \cdot \boldsymbol{h}_3 \,, \quad c_3 = 2\boldsymbol{h}_1 \cdot \boldsymbol{h}_4 \,, \quad c_4 = \|\boldsymbol{h}_2\|^2 \,,$$
$$c_5 = 2\boldsymbol{h}_2 \cdot \boldsymbol{h}_3 \,, \quad c_6 = 2\boldsymbol{h}_2 \cdot \boldsymbol{h}_4 + \|\boldsymbol{h}_3\|^2 \,, \quad c_7 = 2\boldsymbol{h}_3 \cdot \boldsymbol{h}_4 \,, \quad \text{and} \quad c_8 = \|\boldsymbol{h}_4\|^2$$

With these scalar parameters, it is possible to show that the squared Euclidean norm of the specific angular momentum is given by

$$\|\boldsymbol{h}\|^2 = c_0\dot{\rho}^2 + P(\rho)\dot{\rho} + U(\rho)$$

where

$$P(\rho) = c_1\rho^2 + c_2\rho + c_3$$
$$U(\rho) = c_4\rho^4 + c_5\rho^3 + c_6\rho^2 + c_7\rho + c_8$$

The eccentricity is related to both the specific angular momentum and specific energy by

$$e = \sqrt{1 + \frac{2\mathscr{E}\|\boldsymbol{h}\|^2}{\mu^2}}$$

which may be rearranged as

$$2\mathscr{E}\|\boldsymbol{h}\|^2 = -\mu^2(1 - e^2)$$

We have already found an equation for $2\mathscr{E}$ from our original development of the admissible region. And, we have just found an expression for $\|\boldsymbol{h}\|^2$.

Therefore, if we substitute for $2\mathscr{E}$ and for $\|\boldsymbol{h}\|^2$, it follows that

$$(\dot{\rho}^2 + w_1\dot{\rho} + F(\rho))(c_0\dot{\rho}^2 + P(\rho)\dot{\rho} + U(\rho)) = -\mu^2(1 - e^2)$$

which may be rewritten as

$$a_4\dot{\rho}^4 + a_3\dot{\rho}^3 + a_2\dot{\rho}^2 + a_1\dot{\rho} + a_0 = 0$$

where

$$a_4 = c_0 \,, \quad a_3 = P(\rho) + c_0 w_1 \,, \quad a_2 = U(\rho) + c_0 F(\rho) + w_1 P(\rho) \,,$$
$$a_1 = F(\rho)P(\rho) + w_1 U(\rho) \,, \quad \text{and} \quad a_0 = F(\rho)U(\rho) + \mu^2(1 - e^2)$$

If we are given a value of eccentricity, a curve of constant eccentricity may be determined by solving for the roots of the quartic equation that we have developed, where the equation is quartic in $\dot{\rho}$ given a value of ρ.

Since we are solving for the roots of a quartic, four solutions will be obtained.

Any imaginary solutions which result are discarded and only the real solutions are considered when determining the curve of constant eccentricity.

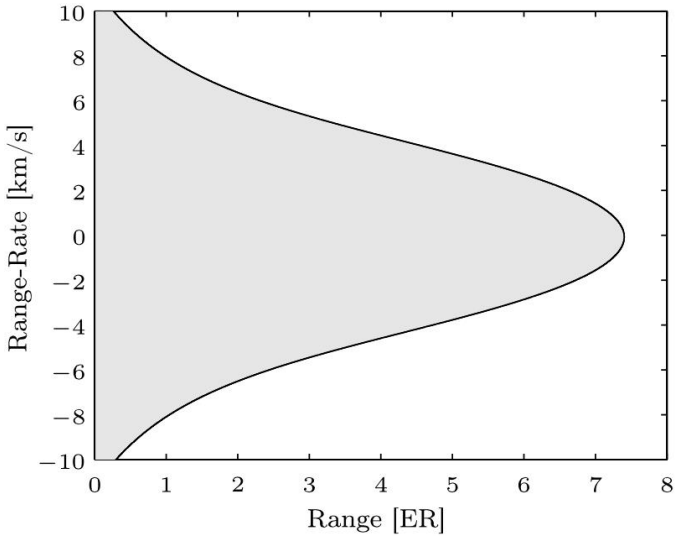
6.3.1.2 Example of the Admissible Region

To illustrate the determination of the admissible region and the semi-major axis and eccentricity constraints, consider an optical observation of $\alpha = 10$ [deg], $\delta = -2$ [deg], $\dot{\alpha} = 15$ [deg/hr], and $\dot{\delta} = 3$ [deg/hr].

In this example, the inertial ground station is taken to be on the surface of the Earth (i.e. $\|\boldsymbol{R}\| = R_e$) and to have spherical coordinates of $\phi = 30$ [deg] (as measured from the equatorial plane) and $\lambda = 0$ [deg] (as measured from the inertial x-axis), and the inertial velocity is given by $\dot{\boldsymbol{R}} = \boldsymbol{\omega} \times \boldsymbol{R}$, where $\boldsymbol{\omega} = [0 \;\; 0 \;\; \omega_e]^T$ is the angular velocity vector of the Earth.

Then, the admissible region is determined by setting $\mathscr{E} = 0$ and solving for $\dot{\rho}$ given values of the range, ρ.

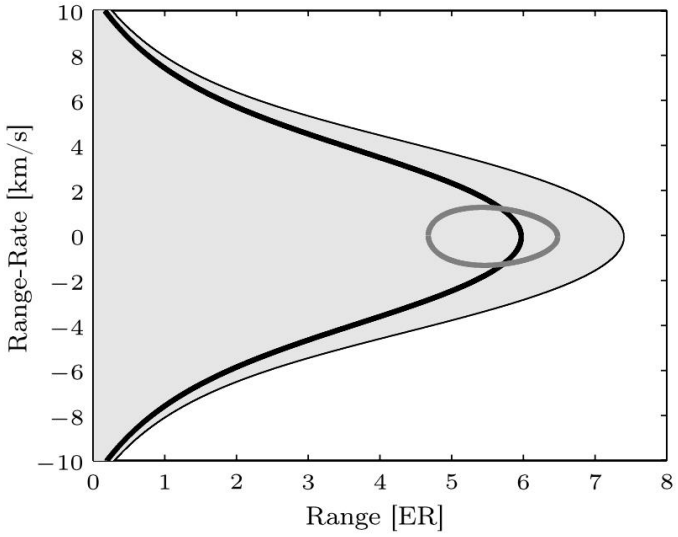
This then yields the admissible region shown below; the shaded region describes the possible combinations of range and range-rate which permit the object to be in an Earth captured orbit (negative orbital energy).



The determination of the constrained admissible region results by imposing a semi-major axis constraint along with an eccentricity constraint (and any other desired constraints) on the unconstrained admissible region.

In this example, the semi-major axis constraint is chosen to be $a \leq 50000$ [km], and the eccentricity constraint is chosen to be $e \leq 0.4$.

The preceding procedure for determining the curves of constant semi-major axis and constant eccentricity are applied, resulting in the curves shown below, which are overlaid on the unconstrained admissible region.

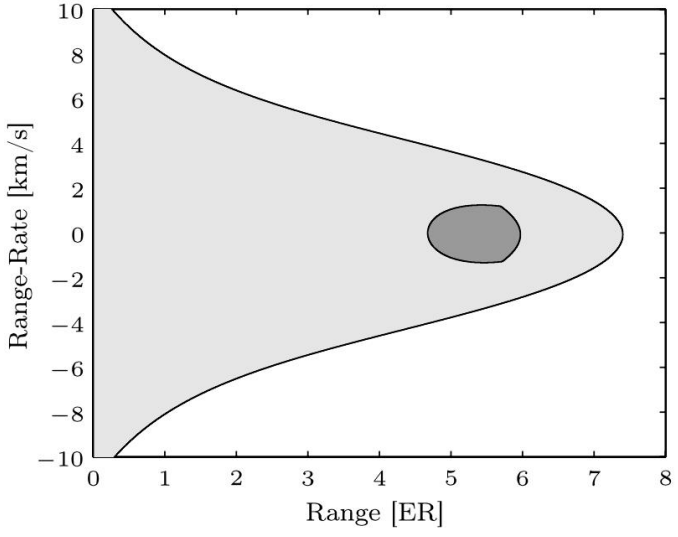


The curve of constant semi-major axis is shown in black and the curve of constant eccentricity is shown in dark gray.

The region associated with orbits that have semi-major axis less than the constraint value is the region which encompasses the $(0,0)$ point of the range/range-rate plane and which does not go beyond the curve of constant semi-major axis.

Similarly, the region associated with orbits that have eccentricity less than the constraint value is the region inside of the closed curve of constant eccentricity.

By taking the intersection of these two regions, the constrained admissible region may be obtained, and this region is shown by the darker region in the following figure.



6.3.2 Gaussian Mixture Admissible Region

Given the (constrained) admissible region and no additional information, no single combination of range and range-rate is more likely than any other, meaning that the admissible region can be interpreted probabilistically as describing a uniform distribution with the support defined by the boundaries of the (constrained) admissible region.

For this reason, a method for describing a uniform distribution via a Gaussian mixture model is now developed.

First, the problem of approximating a univariate uniform distribution is discussed, followed by the application of the method to generating a Gaussian mixture approximation of the admissible region.

The problem of approximating a univariate uniform distribution using a Gaussian mixture pdf is now examined. It is desired to approximate the uniform pdf

$$p(x) = \begin{cases} \dfrac{1}{b-a} & , \quad a \leq x \leq b \\ 0 & , \quad \text{otherwise} \end{cases}$$

by a Gaussian mixture pdf of the form

$$q(x) = \sum_{\ell=1}^{L} w_\ell p_g(x \mid m_\ell,\, P_\ell)$$

where $p_g(x \mid a,\, A)$ represents a Gaussian pdf for the random variable x with mean a and covariance A.

To facilitate the approximation, the distance between $p(x)$ and $q(x)$ is taken as the L_2 norm via

$$L_2[p||q] = \int_{\mathbb{R}} (p(x) - q(x))^2 dx$$

and the optimization problem

$$\min J = L_2[p||q] \quad \text{subject to} \quad w_\ell \geq 0 \,\forall\, \ell \quad \text{and} \quad \sum_{\ell=1}^{L} w_\ell = 1$$

is considered in order to determine the weights, means, and variances of the components in the Gaussian mixture approximation.

It should be noted that the number of components in the Gaussian mixture is assumed to be specified so that it is not a parameter of the optimization problem

In order to make the optimization problem computationally feasible, the L_2 distance between the uniform and Gaussian mixture pdfs needs to be cast in closed-form.

Noting that $p(x) = 0 \; \forall \, x \in \mathbb{R} \setminus \{[a,b]\}$, it follows that the L_2 norm can be written as

$$L_2[p||q] = \int_a^b p^2(x)dx + \int_{-\infty}^{\infty} q^2(x)dx - 2\int_a^b p(x)q(x)dx$$

Substituting for $p(x)$ (the uniform pdf) from its definition then yields

$$L_2[p||q] = \frac{1}{b-a} + \int_{-\infty}^{\infty} q^2(x)dx - \frac{2}{b-a}\int_a^b q(x)dx$$

Now, we consider the second term in the L_2 norm.

It can be shown that the multiplication of two Gaussian pdfs is given by an unnormalized Gaussian pdf as

$$p_g(x;a,A)p_g(x;b,B) = \Gamma(a,b,A,B)p_g(x;c,C)$$

where

$$c = C(A^{-1}a + B^{-1}b), \quad C = (A^{-1} + B^{-1})^{-1}, \quad \text{and}$$
$$\Gamma(a,b,A,B) = |2\pi(A+B)|^{-1/2} \exp\left\{-\frac{1}{2}(a-b)^T(A+B)^{-1}(a-b)\right\}$$

Using this property, it is then straightforward to show that

$$\int_{-\infty}^{\infty} q^2(x)dx = \sum_{i=1}^{L}\sum_{j=1}^{L} w_i w_j \Gamma(m_i, m_j, P_i, P_j).$$

This is the second term in the L_2 norm between the GM pdf and the uniform pdf, but we still need the cross term.

To consider the cross term, we first define the error function as

$$\text{erf}\{z\} = \frac{2}{\sqrt{\pi}} \int_0^z \exp\{-t^2\}\, dt$$

Then, the cumulative distribution function (cdf) for a Gaussian pdf can be expressed in terms of the error function as

$$\int_{-\infty}^{z} p_g(x \mid \mu,\, \Sigma) = \frac{1}{2}\left[1 + \text{erf}\left\{\frac{z-\mu}{\sqrt{2\Sigma}}\right\}\right]$$

By applying the cdf of a Gaussian distribution to the cross term in the L_2 norm, it follows that

$$\int_a^b q(x)dx = \frac{1}{2}\sum_{\ell=1}^{L} w_\ell \left[\text{erf}\left\{\frac{b-m_\ell}{\sqrt{2P_\ell}}\right\} - \text{erf}\left\{\frac{a-m_\ell}{\sqrt{2P_\ell}}\right\}\right]$$

Now, we substitute our preceding individual developments of the integral terms in the L_2 norm to the L_2 norm, such that

$$\begin{aligned} L_2[p||q] = \frac{1}{b-a} &+ \sum_{i=1}^{L}\sum_{j=1}^{L} w_i w_j \Gamma(m_i, m_j, P_i, P_j) \\ &- \frac{1}{b-a}\sum_{\ell=1}^{L} w_\ell \left[\text{erf}\left\{\frac{b-m_\ell}{\sqrt{2P_\ell}}\right\} - \text{erf}\left\{\frac{a-m_\ell}{\sqrt{2P_\ell}}\right\}\right] \end{aligned}$$

Except for the appearance of the error function, this is a closed-form expression for the L_2 distance between an arbitrary uniform pdf and an arbitrary Gaussian mixture pdf.

However, even with this form of the L_2 norm, the optimization problem is still ill-conditioned?

Why?

Effectively, we can swap any two components of the GM pdf and not change the cost function.

There are just too many parameters and there is not a clear unique solution or a clear way to stop an optimizer from just running in circles.

After all, there are $3L$ parameters that need to be found via the optimization, which even for small L yields a larger than

desired search space.

What can we do? We restrict our search space...

To help reduce the number of parameters and create a better conditioned optimization problem, it is assumed that

- the weights are equal for all components,
- the means are evenly distributed across the support of $p(x)$, and
- the Gaussian mixture is homoscedastic.

If we make the preceding assumptions, we can reduce the L_2 distance to

$$L_2[p||q] = \frac{1}{b-a} + \frac{w^2}{2\sqrt{\pi}\sigma}\sum_{i=1}^{L}\sum_{j=1}^{L}\exp\left\{-\frac{1}{4}\left(\frac{m_i - m_j}{\sigma}\right)^2\right\} - \frac{w}{b-a}\sum_{\ell=1}^{L}\left[\text{erf}\left\{B_\ell\right\} - \text{erf}\left\{A_\ell\right\}\right]$$

where

$$A_\ell = \left(\frac{a-m_\ell}{\sqrt{2}\sigma}\right) \quad \text{and} \quad B_\ell = \left(\frac{b-m_\ell}{\sqrt{2}\sigma}\right)$$

Now, we only need to optimize the common standard deviation parameters, σ.

Furthermore, the constraints in the optimization problem, namely that the weights all be positive and sum to one, are automatically satisfied.

The optimization problem is now readily cast as a root finding problem by satisfaction of the first-order optimality conditions, i.e.

$$\frac{dL_2[p||q]}{d\sigma} = 0$$

Now, we just take the derivative of the L_2 distance with respect to the standard deviation parameter, which gives

$$\frac{dL_2[p||q]}{d\sigma} = \frac{w^2}{2\sqrt{\pi}\sigma^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left[\frac{1}{2}\left(\frac{m_i-m_j}{\sigma}\right)^2 - 1\right]\exp\left\{-\frac{1}{4}\left(\frac{m_i-m_j}{\sigma}\right)^2\right\} - \frac{2w}{(b-a)\sqrt{\pi}\sigma}\sum_{\ell=1}^{L}\left[A_\ell \exp\left\{-A_\ell^2\right\} - B_\ell \exp\left\{-B_\ell^2\right\}\right]$$

If we find a solution to the first-order optimality condition, we have a candidate minimum, so we should check the second derivative.

For brevity, the second derivative is not given, but it readily follows from the first derivative.

It is also interesting to note that the derivative removes the error function and instead now contains exponentials.

Therefore, the derivative is actually found completely in closed form.

The procedure for generating a library of solutions (i.e. an optimized value of σ given a number of components, L) is now summarized.

Without loss of generality, let the parameters of the univariate uniform pdf be $a = 0$ and $b = 1$.

Specifying the number of components, L, and following the previously discussed rules regarding the weights and means, it follows that

$$\tilde{w} = \frac{1}{L} \qquad \text{and} \qquad \tilde{m}_\ell = \frac{\ell}{L+1} \;\; \forall \;\; \ell \in \{1,2,\ldots,L\}$$

The designation of $\tilde{w}$ and $\tilde{m}_\ell$ is to denote that the solutions are to be obtained for the case of $a = 0$ and $b = 1$.

To determine the optimal standard deviation, the root finding problem summarized by is solved, and the output solution that satisfies the first-order condition is denoted by $\tilde{w}$.

The library of solutions is then given by the pairs of L and $\tilde{w}$.

For instance, the values of $\tilde{w}$ for $L = 1$ to $L = 15$ obtained via this procedure are summarized in the following table.

L	$\tilde{\sigma}$
1	0.3467
2	0.2903
3	0.2466
4	0.2001
5	0.1531
6	0.1225
7	0.1026
8	0.0884
9	0.0778
10	0.0696
11	0.0629
12	0.0575
13	0.0529
14	0.0490
15	0.0456

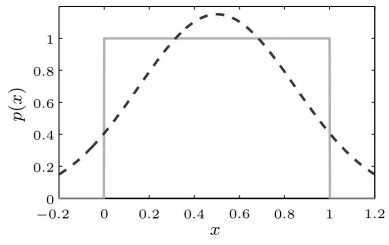
By specifying L, we can directly compute the weight and mean parameters and the library of solutions is used to determine the optimal value of $\tilde{w}$.

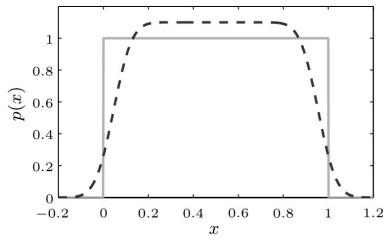
Given the values of the Gaussian mixture (i.e. $\tilde{w}$, $\tilde{m}_\ell$, and $\tilde{\sigma}$) that approximate the uniform distribution with $a = 0$ and $b = 1$, an arbitrary uniform distribution approximation can then be obtained via

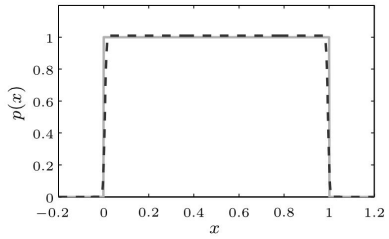
$$w = \tilde{w}, \quad m_\ell = a + (b-a)\tilde{m}_\ell, \quad \text{and} \quad \sigma = (b-a)\tilde{\sigma}$$

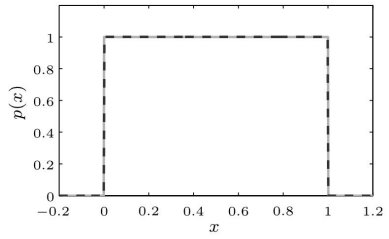
Let's look at what happens when we apply this technique to the problem of approximating a uniform distribution.

In particular, let's look at the cases of determining a GM approximation to a uniform pdf using 1, 10, 100, and 1000 components in the GM.









As we might expect, we observe that by increasing the number of components in the mixture, we can better approximate the uniform pdf with the GM pdf.

Now, we have the ability to approximate a uniform distribution, but how can we translate this into something that's useful?

Remember that we have the concept of the admissible region.

Recall also that we have interpreted the admissible region probabilistically as a bivariate uniform distribution.

Since the admissible region represents a uniform pdf in two dimensions, the problem of approximating the admissible region with a Gaussian mixture must be addressed by breaking the two-dimensional approximation into two one-dimensional approximations so that the developed Gaussian mixture approximation strategy may be applied.

The first of the two approximations deals with the range coordinate of the admissible region.

The first step is to determine the range-marginal pdf that results from interpreting the admissible region as a uniform distribution in two dimensions.

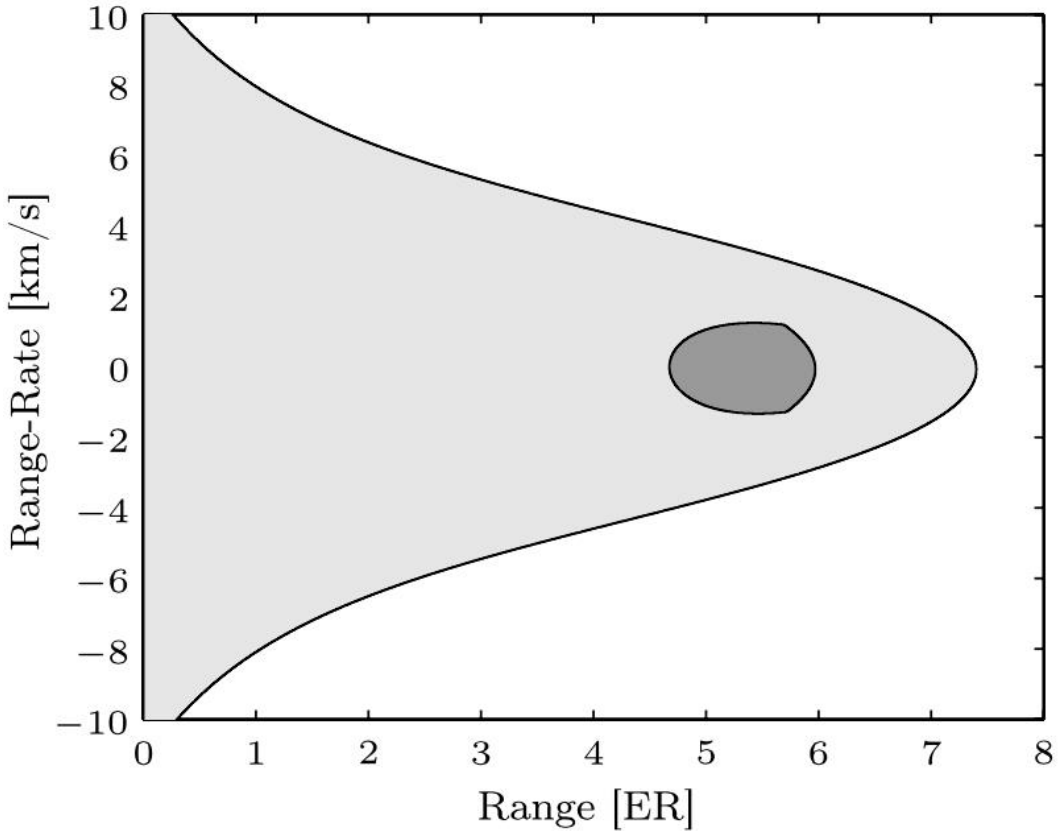
That is, if $p_{\rho,\dot{\rho}}(\rho,\dot{\rho})$ represents the total uniform pdf of the admissible region, then

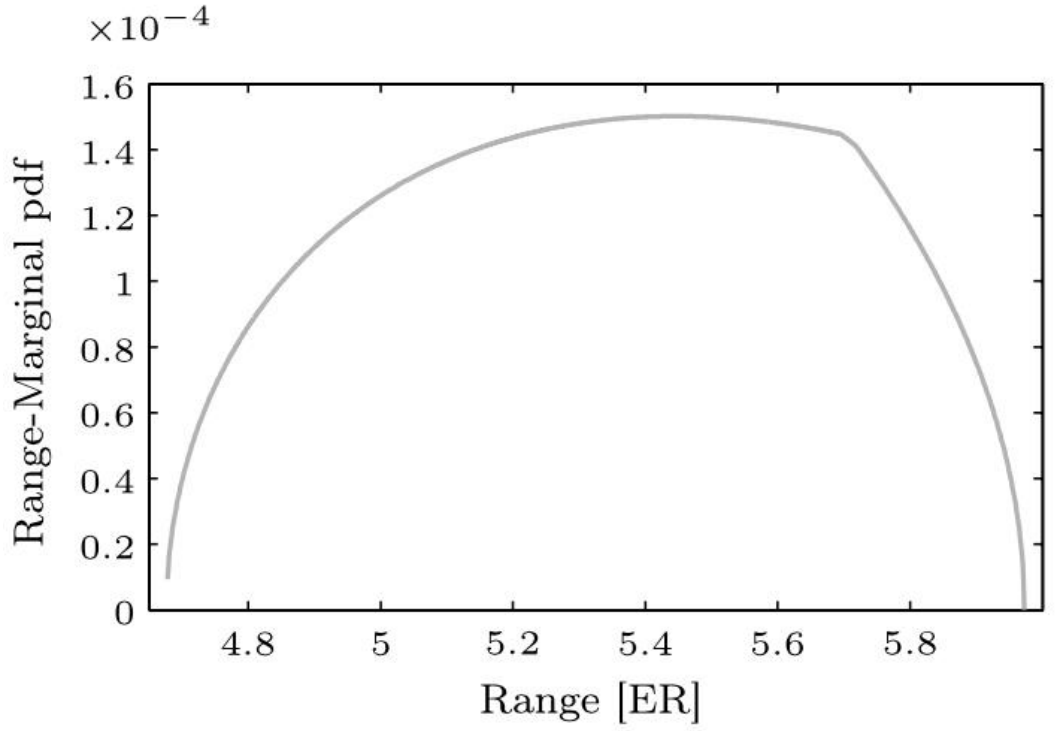
$$p_{\rho}(\rho) = \int_{-\infty}^{\infty} p_{\rho,\dot{\rho}}(\rho,\nu)d\nu$$

is the range-marginal pdf.

In general, the admissible region boundaries are determined numerically, so a simple numerical procedure can be employed to determine the range-marginal pdf.

Recall the constrained admissible region. Below, we have the constrained admissible region from before and then the range-marginal pdf.


Range-Rate [km/s]
10
8
6
4
2
0
−2
−4
−6
−8
−10
0
1
2
3
4
5
6
7
8
Range [ER]


×10−4
Range-Marginal pdf
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
4.8
5
5.2
5.4
5.6
5.8
Range [ER]

The support of the range-marginal pdf, defined by the scalar values a and b with $a < b$, is readily determined by the extremal range values for which the admissible region exists.

In order to apply a set of predetermined libraries which contain the optimized GMM parameters, a maximum standard deviation for the range direction, $\sigma_{\rho,\max}$, is specified.

Then, the precomputed libraries are searched to find L, the number of components, such that $\tilde{\sigma}(b-a) \leq \sigma_{\rho,\max}$.

For example, let $\sigma_{\rho,\max} = 2$ and $(b-a) = 30$.

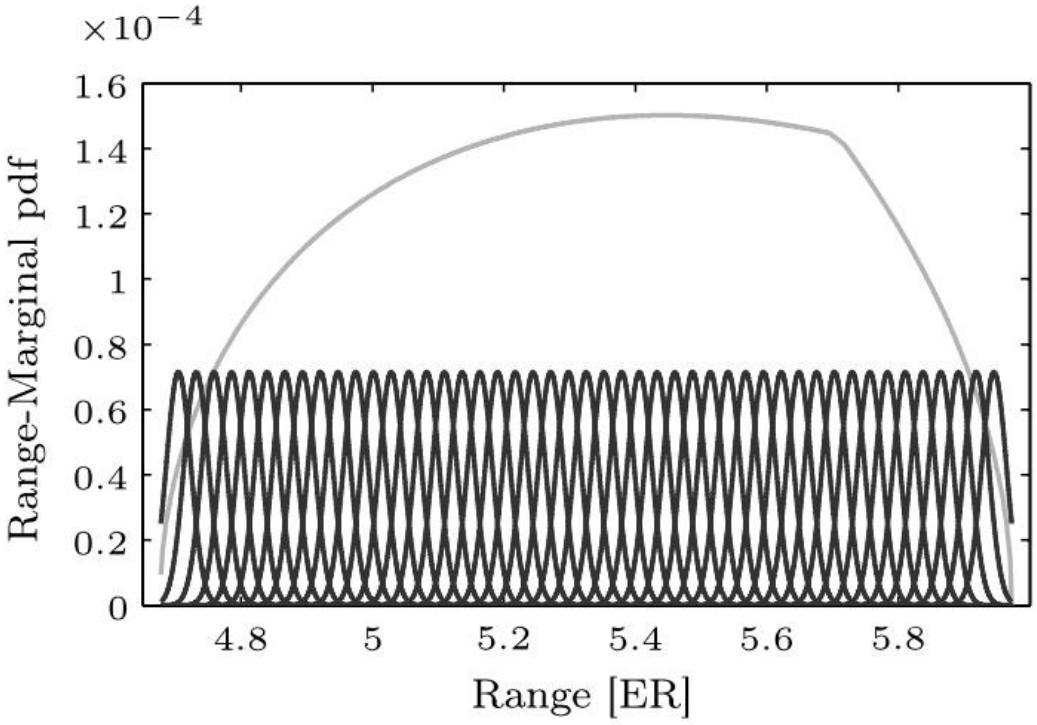
In this case, it is desired to find $\tilde{\sigma}$ such that $30\tilde{\sigma} \leq 2$, or $\tilde{\sigma} \leq 1/15 = 0.0667$.

From the table of optimized values, it follows that the smallest number of components which satisfies this condition is $L = 11$.

Given the determination of the number of components and hence $\tilde{\sigma}$, the parameters of the GM approximation are

$$w_\ell = \frac{1}{L}, \quad m_\ell = a + \frac{(b-a)\ell}{L+1}, \quad \text{and} \quad P_\ell = ((b-a)\tilde{\sigma})^2 \quad \forall\ \ell \in \{1,2,\ldots,L\}$$

These parameters then produce a GM approximation to a uniform distribution with the same support set as the range-marginal pdf; they do not, however, produce an approximation to the range-marginal pdf since the range-marginal pdf is not, in general, represented by a uniform distribution.



Our GM range-marginal pdf is, as of now, given by

$$p(\rho) = \sum_{\ell=1}^{L} w_\ell p_g(\rho \mid m_\ell,\, P_\ell)$$

In order to account for the fact that the range-marginal pdf is not a uniform distribution, first note that the GM pdf is linear in the weight parameters, such that if the weights are concatenated into a vector $\boldsymbol{w}$, the L-component GM pdf may be expressed as

$$q(x) = \boldsymbol{h}^T(x)\boldsymbol{w}$$

where $\boldsymbol{w} \in \mathbb{R}^L$ and $\boldsymbol{h}(x) \in \mathbb{R}^L$ with the ℓ^{th} element given by

$$h_\ell(x) = p_g(x \mid m_\ell,\, P_\ell)$$

Now, let M values of the range that lay in the support of the range-marginal pdf be taken, yielding $\rho_i \; \forall \, i \in \{1, 2, \ldots, M\}$.

Evaluate the range-marginal pdf at each value of the range, and set $p_i = p_\rho(\rho_i)$.

Collect the evaluations of the range-marginal pdf in a vector, $\boldsymbol{p}$, where the i^{th} element is p_i.

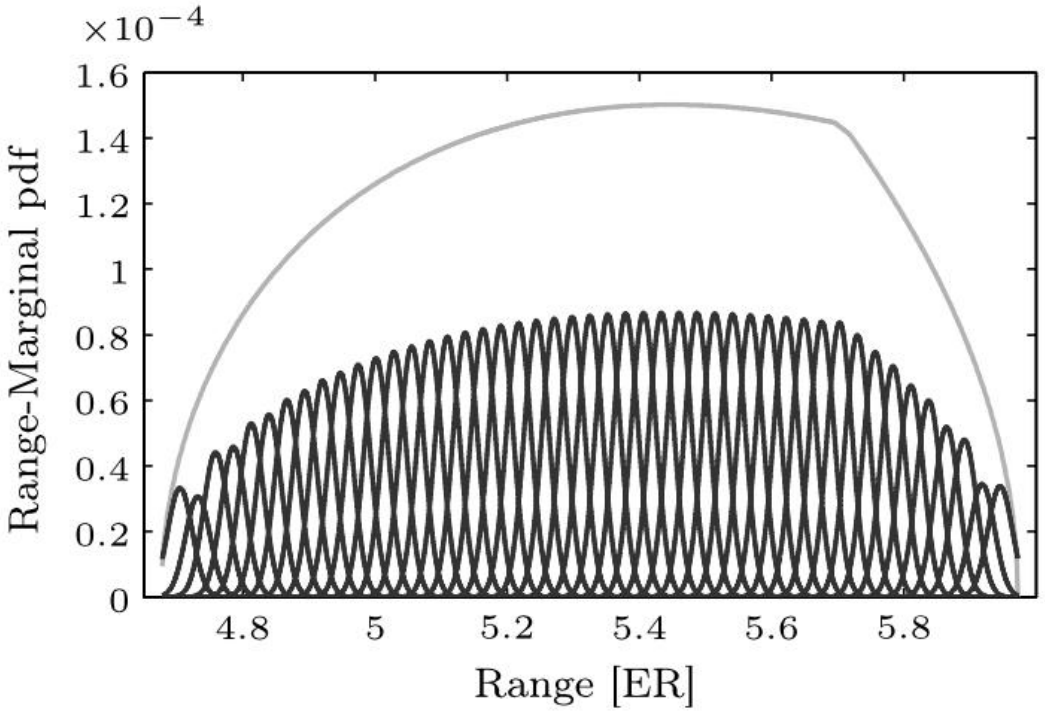
Then, the weights of GM pdf are found by solving the following least-squares problem subject to linear equality/inequality constraints:

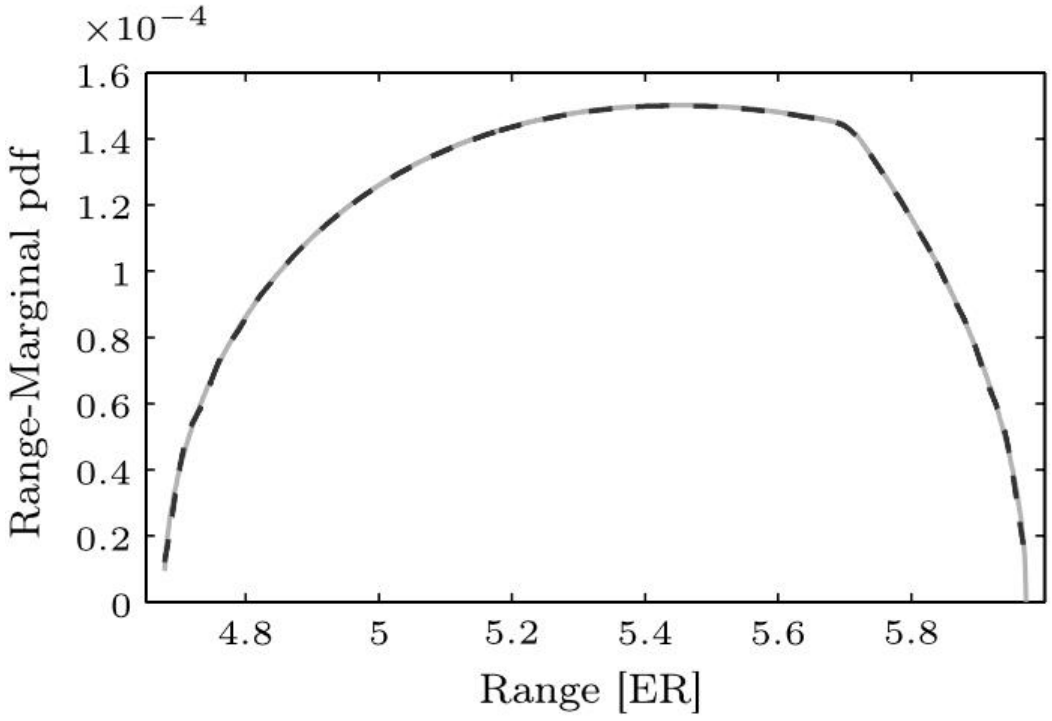
$$\min J = \|\boldsymbol{p} - \boldsymbol{H}\boldsymbol{w}\| \quad \text{subject to} \quad \boldsymbol{w} \geq \boldsymbol{0} \quad \text{and} \quad \boldsymbol{1}^T\boldsymbol{w} = 1$$

where $\boldsymbol{1} \in \mathbb{R}^L$ is a vector of ones, and $\boldsymbol{H} \in \mathbb{R}^{M\times L}$ with the element in the i^{th} row and ℓ^{th} column given by

$$H_{i,\ell} = p_g(\boldsymbol{\rho}_i \mid m_\ell,\, P_\ell)$$

Obtaining the solution to the least-squares problem yields a set of weights such that when combined with the means and covariances of the GM pdf produce a GM approximation to the range-marginal pdf.



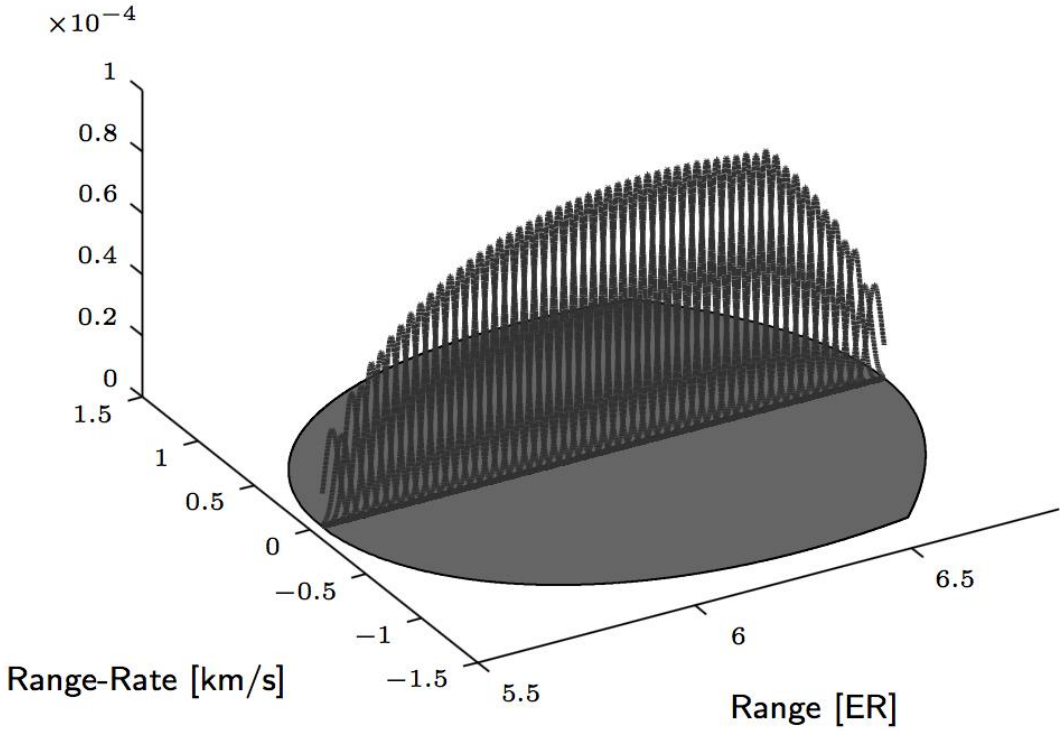


Where are we at?

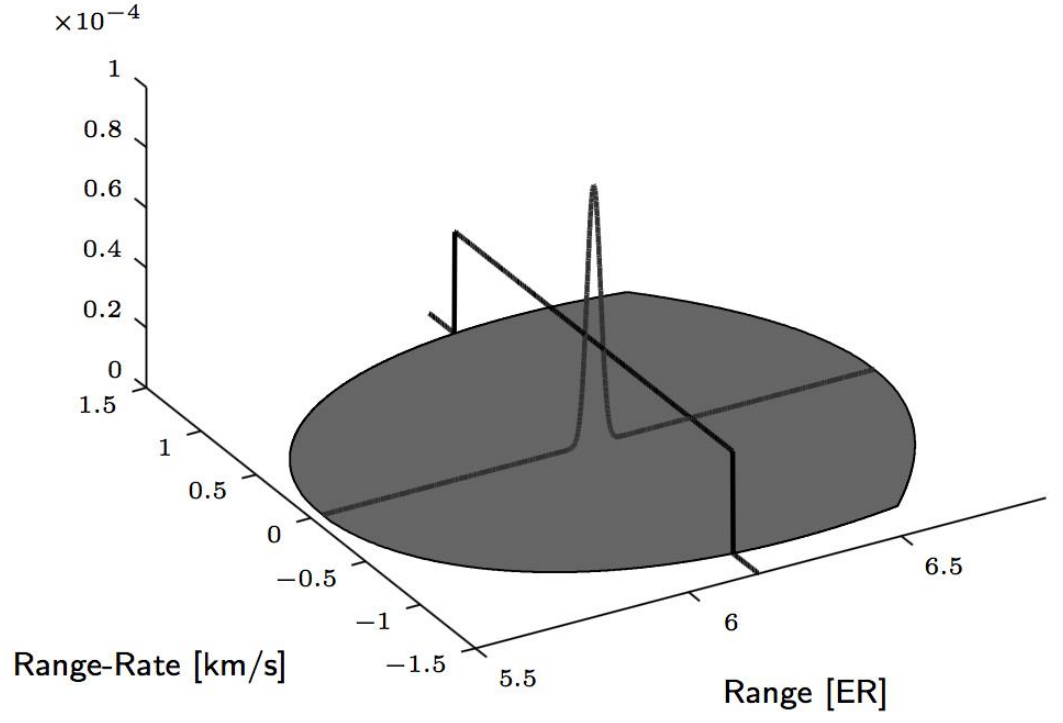
We're trying to generate a GM approximation to the constrained admissible region.

Up to this point, we have obtained an approximation for the range-marginal pdf, but no consideration of the range-rate direction has been given.

Before proceeding, consider a single component of the range-marginal GM approximation.



The range-rate ambiguity associated with this component is uniform, meaning that given a value of range no associated value of range-rate within the constrained admissible region is more or less likely.



Therefore, the idea behind extending the range-marginal GM pdf to the constrained admissible region is to apply the previously discussed GM approximation of the uniform pdf to the range-rate direction for each component of the range-marginal GM approximation in order to develop a bivariate GM pdf which describes the constrained admissible region.

The first step to including the range-rate direction is to relabel the GM approximation of the range-marginal pdf.

As such, let the number of components, means, and covariances be given by L_ρ, $m_{\rho,\ell}$, and $P_{\rho,\ell}$, respectively.

Similarly, let the weights that were found via the least-squares problem be given by $w_{\rho,\ell}$.

When put together, these values form a GM approximation of the range-marginal pdf, which is given by

$$p_\rho(\rho) \approx \sum_{\ell=1}^{L_\rho} w_{\rho,\ell} p_g(\rho \mid m_{\rho,\ell},\, P_{\rho,\ell})$$

The next step is to determine the total number of components that will be in the bivariate GM representing the constrained admissible region.

Recalling that $p_{\rho,\dot\rho}(\rho,\dot\rho)$ represents the total uniform pdf of the constrained admissible region, the extremal range-rate values are computed at each range value that is dictated by the means of the range-marginal GM.

That is, for each $m_{\rho,\ell}$, compute

$$a_\ell = \arg\min_{\dot\rho} p_{\rho,\dot\rho}(m_{\rho,\ell},\dot\rho) \qquad \text{and} \qquad b_\ell = \arg\max_{\dot\rho} p_{\rho,\dot\rho}(m_{\rho,\ell},\dot\rho)$$

Then, we define $\Delta_\ell = (b_\ell - a_\ell)$.

In order to apply the predetermined libraries (recall the table of solutions) of optimized GM parameters that approximate a uniform pdf, a maximum standard deviation for the range-rate direction, $\sigma_{\dot\rho,\max}$, is specified.

Then, the precomputed libraries are searched to find $L_{\dot\rho,\ell}$, the number of components, such that $\tilde\sigma\Delta_\ell \leq \sigma_{\dot\rho,\max}$.

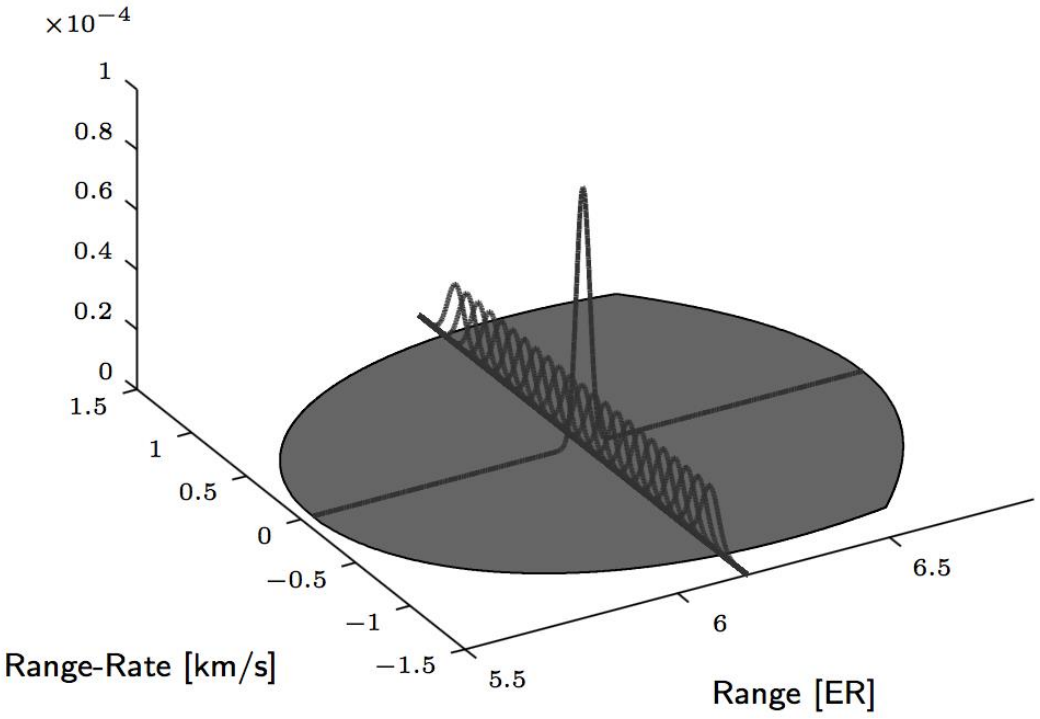
Then, the total number of components of the bivariate GM approximation of the constrained admissible region is given by $L = \sum_{\ell=1}^{L_\rho} L_{\dot\rho,\ell}$.

The final step is to compute the parameters of the L-component bivariate GM derived from the GM parameters of the range-marginal pdf and application of the GM approximation of a uniform pdf to the range-rate direction.

Recall that for each component of the range-marginal GM pdf, the range-rate direction ambiguity is uniform.

Therefore, for the ℓ^{th} component of the range-marginal pdf, the range-rate uniform pdf is approximated by an $L_{\dot\rho,\ell}$-component GM with parameters

$$w_{\dot\rho,k} = \frac{1}{L_{\dot\rho,\ell}}, \quad m_{\dot\rho,k} = a_\ell + \frac{\Delta_\ell k}{L_{\dot\rho,\ell}+1}, \quad \text{and} \quad P_{\dot\rho,k} = (\Delta_\ell \tilde\sigma)^2 \quad \forall\ k \in \{1,2,\ldots,L_{\dot\rho,\ell}\}$$



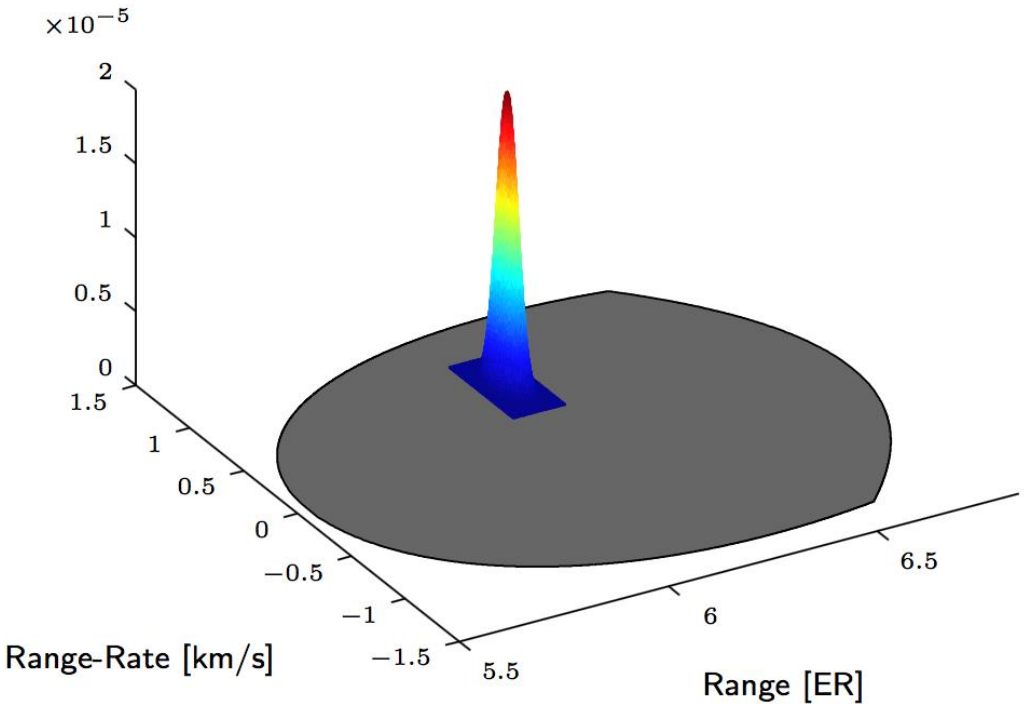
Let the indexing variable i_ℓ be defined as $i_\ell = \sum_{j=1}^{\ell-1} L_{\dot\rho,j}$.

Then, given the parameters of the ℓ^{th} component of the range-marginal GM approximation and the k associated

components of the range-rate GM approximation, the corresponding components of the bivariate GM approximation of the constrained admissible region are

$$w_{\rho,\dot{\rho},i_\ell+k} = w_{\rho,\ell} w_{\dot{\rho},k}\,, \quad \boldsymbol{m}_{\rho,\dot{\rho},i_\ell+k} = \begin{bmatrix} m_{\rho,\ell} \\ m_{\dot{\rho},k} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{P}_{\rho,\dot{\rho},i_\ell+k} = \begin{bmatrix} P_{\rho,\ell} & 0 \\ 0 & P_{\dot{\rho},k} \end{bmatrix}$$

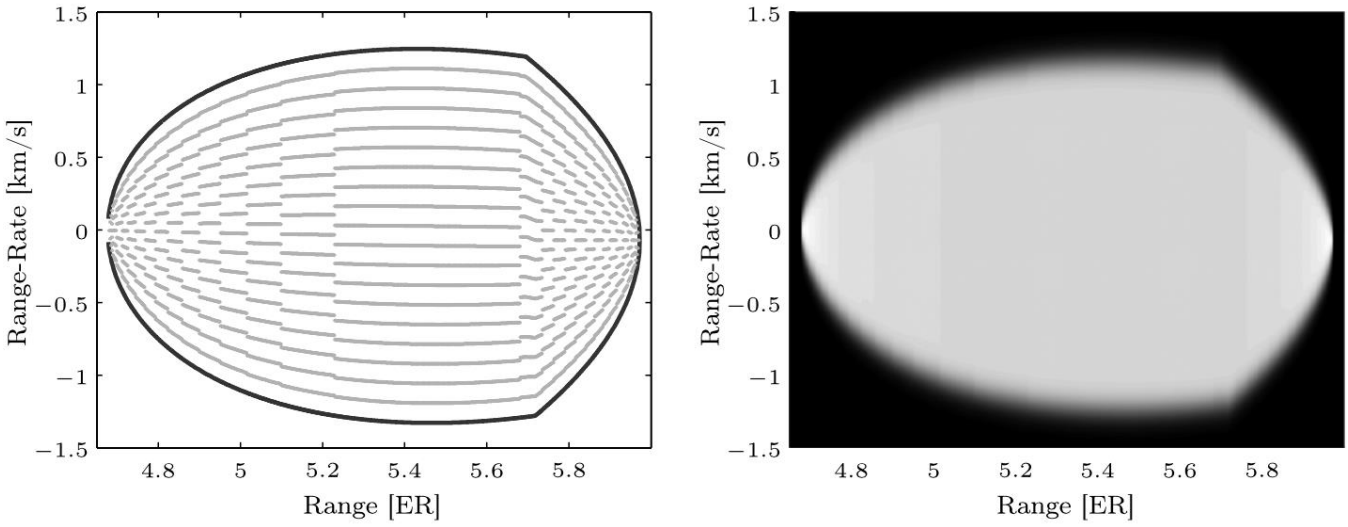
where $k \in \{1,2,\ldots,L_{\dot{\rho},\ell}\}$.



The above procedure is then repeated for every component of the range-marginal pdf, such that the end result is a GM approximation of the constrained admissible region of the form

$$p_{\rho,\dot{\rho}}(\rho,\dot{\rho}) \approx \sum_{i=1}^{L} w_{\rho,\dot{\rho},i}\, p_g(\rho,\dot{\rho} \mid \boldsymbol{m}_{\rho,\dot{\rho},i},\ \boldsymbol{P}_{\rho,\dot{\rho},i})$$

Applying the entire procedure for generating a Gaussian mixture approximation of the admissible region to the constrained admissible region yields the pdf approximation illustrated below.



That's the bulk of IOD process!

We've taken a set of optical data (including the line-of-sight rates), determined the region where Earth-bound orbits can exist subject to some semi-major axis and eccentricity constraints, and then determined a GM approximation.

The result is then a GM pdf in the range/range-rate space describing the uniform bivariate distribution of the admissible region.

So, we have a set of weights, means, and covariances describing the admissible region.

How do we get to a true IOD solution? After all, we need weights, means, and covariances in a usual state space (Cartesian coordinates, Keplerian elements, equinoctial orbital elements, etc.).

For instance, if we consider the state-space to be range/right-ascension/declination and their rates, we could write the pdf in this space as

$$p(\boldsymbol{x}_0) = \sum_{i=1}^{L} w_{i,0} p_g(\boldsymbol{x}_0 \mid \boldsymbol{m}_{i,0},\, \boldsymbol{P}_{i,0})$$

where

$$w_{i,0} = w_{\rho,\dot{\rho},i}\,, \quad \boldsymbol{m}_{i,0} = \boldsymbol{W} \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{m}_{\rho,\dot{\rho},i} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{P}_{i,0} = \boldsymbol{W} \begin{bmatrix} \boldsymbol{R}_0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{\rho,\dot{\rho},i} \end{bmatrix} \boldsymbol{W}^T$$

That is, we combine the GM of the admissible region with the measurement noise to get a pdf for the 6-dimensional state.

The matrix $\boldsymbol{W}$ which is a permutation matrix which maps $[\alpha\ \delta\ \dot{\alpha}\ \dot{\delta}\ \rho\ \dot{\rho}]^T$ into $[\rho\ \alpha\ \delta\ \dot{\rho}\ \dot{\alpha}\ \dot{\delta}]^T$, and is given by

$$\boldsymbol{W} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

If it is desired to map into any other coordinate system, then we can view this as the transformation

$$\boldsymbol{x}' = \boldsymbol{g}(\boldsymbol{x})$$

where $\boldsymbol{x}$ is our range/right-ascension/declination coordinate system and $\boldsymbol{x}'$ is our other coordinate system, say Cartesian coordinates, for example.

Supplement

References [9].

6.4 Orbital elements and the Angular Momentum Vector

Fig. 6.8 shows the angular momentum vector $\boldsymbol{h}$. x,y,z are the axis of the ECI coordinate system with x pointing in the direction of the vernal equinox. Π denotes the perigee, where ω is the angle of perigee passage, also called argument of perigee. i is the inclination, and Ω is the right ascension of the ascending node. Because, the angular momentum vector is perpendicular on the orbital plane, the angle between $\boldsymbol{h}$ and the z-axis is again i. The projection of the angular momentum vector then forms an angle π with the ascending node, allowing it to be expressed via the following relation:

$$\boldsymbol{h} = |\boldsymbol{h}| \begin{bmatrix} \cos(\Omega - \frac{\pi}{2}) \sin i \\ \sin(\Omega - \frac{\pi}{2}) \sin i \\ \cos i \end{bmatrix} = |\boldsymbol{h}| \begin{bmatrix} \sin\Omega \sin i \\ -\cos\Omega \sin i \\ \cos i \end{bmatrix} \tag{6.325}$$

with

$$\Omega = \arctan\left(\frac{h_1}{-h_y}\right) \qquad i = \arccos\left(\frac{h_3}{|\boldsymbol{h}|}\right) \tag{6.326}$$

6.5 The Orbital Coordinate System

Four different coordinate systems can be defined, all having the orbital plane as the fundamental plae. Those are also called the four systems of the two body problem. All systems share the same third axis, $\boldsymbol{h}$ the angular momentum axis. Enforcing that all coordinate systems are right handed and orthogonal,the coordinate systems hence can be uniquely defined, defining only one more axis, let's assume axis 1.

With the object being placed in the point P (where the small dot is drawn) and Π being the perigee, ω being the perigee axis and Ω being the right ascension of the ascending node, like begore, the four different coordinate systems are defined via their four different first axis, $\boldsymbol{e}_\Omega$, , $\boldsymbol{e}_\Pi$, , $\boldsymbol{e}_R$, , $\boldsymbol{e}_I$. Taking $\boldsymbol{r}$ to be the vector of the position of the object in the inertial ECI system and $\dot{\boldsymbol{r}}$ its velocity at the time t, the transformations leading to the first axis of the coordinate system and the transformation of the vector $\boldsymbol{r}$ in the new coordinate system can be found in Fig.6.8, corresponding to the notation in Fig.6.7. THe angle ξ is defined as the agnel between the Laplace vector $\boldsymbol{q}$ that is pointing towards the perigee and the velocity vector of the object $\dot{\boldsymbol{r}}$. It can also be defined as: The angle ξ is defined as:

$$\xi = 3\sqrt{\frac{\mu}{p^3}}(t - T_0), \tag{6.327}$$

with p being the orbital parameter and $T - 0$ being the time of perigee passage. The vector $\boldsymbol{q}$ is the Laplace vector pointing to the perigee, defined as:

$$\boldsymbol{q} = (\dot{\boldsymbol{r}}^2 - \frac{\mu}{r})\boldsymbol{r} - (\boldsymbol{r} \cdot \dot{\boldsymbol{r}})\dot{\boldsymbol{r}}) \tag{6.328}$$

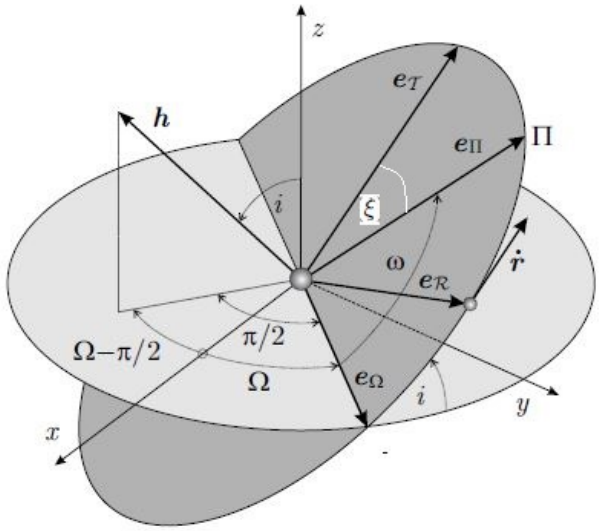


Figure 6.7: Illustration of the orbital element and the ECI coordinate system.

System	First unit vector	Transformation from Inertial System $\mathcal{I}$
Ω	$e_\Omega = \frac{e_3 \times h}{h}$	$r_\Omega = \mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
Π	$e_\Pi = \frac{q}{q}$	$r_\Pi = \mathbf{R}_3(\omega)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
$\mathcal{R}$	$e_\mathcal{R} = \frac{r}{r}$	$r_\mathcal{R} = \mathbf{R}_3(u)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$
$\mathcal{T}$	$e_\mathcal{T} = \frac{\dot{r}}{\lvert\dot{r}\rvert}$	$r_\mathcal{T} = \mathbf{R}_3(\xi)\,\mathbf{R}_3(\omega)\,\mathbf{R}_1(i)\,\mathbf{R}_3(\Omega)\,r$

Figure 6.8: Definition of the orbital element coordinate systems.

Note that $\boldsymbol{q} = \mathbf{0}$ for circular orbits.

The orbital element systems can be used for computationally efficient formulation of Kepler's equation and in the formulation of the integrals of motion. We use one of the systems, the one corresponding to axis e_Ω as the first axis to compute the true anomaly for the restricted orbit determination.

Chapter 7

Orbit Propagation and Perturbations in the Near Earth Space

7.1 A Few Words on Orbit Propagation

The focus of this chapter is on integration of a reference trajectory in three degrees of freedom, that is the position and velocity of the center of mass for a fixed or known orientation (including material properties and shape) of the object at all times. In general, the orbit propagation of a space object (Earth orbiting) can be characterized as the following:

$$\ddot{\mathbf{x}}(t) = \mathbf{a}_{\text{body}-\text{indep.}}(\mathbf{x}(t)) + \mathbf{a}_{\text{body}-\text{dep.}}(\mathbf{x}(\mathbf{t}), \mathbf{b}, \mathbf{q}(\mathbf{t}))(+\mathbf{a}_{\text{unmodeled}}(\mathbf{x}(\mathbf{t}), \mathbf{b}, \mathbf{q}(\mathbf{t}), \mathbf{p}(\mathbf{t}))) \tag{7.1}$$

were $\mathbf{x}$ is the geocentric object state (position and velocity), $a_{\text{body}-\text{indep.}}$ are the accelerations that only depend on the center of mass, the accelerations $a_{\text{body}-\text{dep.}}$ are the non-conservative accelerations that depend on the body parameters $\mathbf{b}$, the body orientation $\mathbf{q}(t)$ and the state of the object $(\mathbf{x}(t)$. $a_{\text{unmodeled}}$ are the accelerations that remain unmodeled. They are either higher orders of magnitudes or fidelity than the force models that are included, or physical effects that have been ignored, or inaccuracies and mismodeling in the force models, body parameters etc. In general the body independent forces are conservative forces and the body-dependent one non-conservative forces. Classically, for objects in Earth orbits the dominant perturbations are:

$$\mathbf{a}_{\text{body}-\text{indep.}}(\mathbf{x}(t)) = \boldsymbol{a}_{Earth-grav} + \boldsymbol{a}_{ThirdBody} + ... \tag{7.2}$$

$$\mathbf{a}_{\text{body}-\text{dep.}}(\mathbf{x}(t)) = \boldsymbol{a}_{SRP} + \boldsymbol{a}_{drag} + ... \tag{7.3}$$

In the numerical integration, the three second order differential equations of Eg.7.1 is transformed into six first order differential equation and integrated step wise:

$$\boldsymbol{y}(t) = [\boldsymbol{r}(t), \boldsymbol{v}(t)]^T \tag{7.4}$$

with:

$$\dot{\boldsymbol{r}}^i = \boldsymbol{v}^i \tag{7.5}$$

$$\dot{\boldsymbol{v}}^i = \sum \boldsymbol{a}^i, \tag{7.6}$$

$$\boldsymbol{f}(t, \boldsymbol{y}(t)) = [\dot{\boldsymbol{r}}(t), \dot{\boldsymbol{v}}(t)]^T, \tag{7.7}$$

using the following scheme:

$$\boldsymbol{y}(t) = \boldsymbol{y}(t - \Delta t) + \int_{t-\Delta t}^{t} \boldsymbol{f}(\tau, \boldsymbol{y}(\tau)) d\tau \tag{7.8}$$

Normally, all quantities are defined in the inertial frame.

Table 3.7. Accelerations acting on LEOs

Perturbation	Acceleration	Orbit Error after one Day		
	[m/s^2]	Radial [m]	Along Track [m]	Out of Plane [m]
$\frac{1}{r^2}$-Term	8.42	"∞"	"∞"	"∞"
Oblateness	$1.5 \cdot 10^{-2}$	60000	400000	900000
Atmospheric Drag	$7.9 \cdot 10^{-7}$	150	8900	1.5
Higher Terms of the Earth's Grav. Field	$2.5 \cdot 10^{-4}$	550	3400	820
Lunar Attraction	$5.4 \cdot 10^{-6}$	2	45	2
Solar Attraction	$5.0 \cdot 10^{-7}$	1	38	15
Direct Rad. Pressure	$9.7 \cdot 10^{-8}$	10	24	0
Solid Earth Tides	$1.1 \cdot 10^{-7}$	0.2	13	1
y-bias	$1.0 \cdot 10^{-9}$	0.1	4.7	0.0

Figure 7.1: Courtesy, Methods of Celestial Mechanics, G. Beutler

Table 3.8. Accelerations acting on GPS satellites

Perturbation	Acceleration	Orbit Error after one Day		
	[m/s^2]	Radial [m]	Along Track [m]	Out of Plane [m]
$\frac{1}{r^2}$-Term	0.57	"∞"	"∞"	"∞"
Oblateness	$5.1 \cdot 10^{-5}$	2750	32000	15000
Lunar Attraction	$4.5 \cdot 10^{-6}$	400	1800	30
Solar Attraction	$2 \cdot 10^{-6}$	200	1200	400
Higher Terms of the Earth's Grav. Field	$4.2 \cdot 10^{-7}$	60	440	10
Direct Rad. Pressure	$9.7 \cdot 10^{-8}$	75	180	5
y-bias	$1.0 \cdot 10^{-9}$	0.9	8.1	0.3
Solid Earth Tides	$5.0 \cdot 10^{-9}$	0.0	0.4	0.0
Atmospheric Drag	—	—	—	—

Figure 7.2: Courtesy, Methods of Celestial Mechanics, G. Beutler

7.2 Earth Gravity

The gravity field can be discriminated in the central term of the point mass of the Earth, and higher order terms.

7.2.1 Point Mass Model

In the point mass model of the gravitational field, the potential is given by

$$U = \frac{\mu}{r}, \tag{7.9}$$

where μ is the gravitational parameter of the body and $r = \left\|\boldsymbol{r}^i\right\| = \left\|\boldsymbol{r}^f\right\|$ is the magnitude of the position vector of the satellite with respect to the center of the body. It is then straightforward to show that the gravitational acceleration vector described in Eq. (7.15) is given by

$$\boldsymbol{a}_g(\boldsymbol{r}) = \frac{\partial U}{\partial r} = -\frac{\mu}{r^3}\boldsymbol{r}, \tag{7.10}$$

7.2.1.1 Jacobian

For the state transition matrix we need to linearize and hence the partial derivatives need to be computed:

$$\boldsymbol{G}(\boldsymbol{r}) = \frac{\partial \boldsymbol{a}_g}{\partial \boldsymbol{r}} = \frac{\mu}{r^5}\left(3\boldsymbol{r}(\boldsymbol{r})^T - r^2\boldsymbol{I}\right), \tag{7.11}$$

$$\boldsymbol{G}(\boldsymbol{v}) = \frac{\partial \boldsymbol{a}_g}{\partial \boldsymbol{v}} = \boldsymbol{0} \tag{7.12}$$

both of which are seen to be orientation independent.

7.2.2 Spherical Harmonics Model: Preliminaries

Any other models except the point mass model are defined in Earth fixed Earth centered coordinates for obvious reasons. Of we start with the representation of the gravitational potential:

$$U = U(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}), \tag{7.13}$$

where $\boldsymbol{r}^{ECEF} = \boldsymbol{T}_{ECI}^{ECEF}\boldsymbol{r}^{ECI}$ is the fixed-frame position of the satellite, $\boldsymbol{T}_{ECI}^{ECEF}$ is the transformation of the inertial reference frame to the fixed reference frame, $\boldsymbol{r}^{ECI}$ is the inertial position of the satellite, and $\boldsymbol{\theta}$ is the collection of the model parameters (e.g. the gravitational parameter of the central body) into a parameter vector.

The first expression of interest is that of the gravitational acceleration. By taking the gradient of Eq. (7.13) with respect to the inertial position, it is readily observed that the inertial gravitational acceleration vector is given by

$$\boldsymbol{a}_g^{ECI} = \boldsymbol{T}_{ECEF}^{ECI}\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}), \tag{7.14}$$

where

$$\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = \left[\frac{\partial U(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})}{\partial \boldsymbol{r}^{ECEF}}\right]^T. \tag{7.15}$$

7.2.2.1 Jacobian

Furthermore, by taking the gradient of Eq. (7.14) with respect to the inertial position, it is found that

$$\boldsymbol{G}_g = \frac{\partial \boldsymbol{a}_g^{ECI}}{\partial \boldsymbol{r}^{ECI}} = \boldsymbol{T}_{ECEF}^{ECI}\boldsymbol{G}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})\boldsymbol{T}_{ECI}^{ECEF}, \tag{7.16}$$

where

$$\boldsymbol{G}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = \frac{\partial \boldsymbol{a}_g(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})}{\partial \boldsymbol{r}^{ECEF}}. \tag{7.17}$$

In the following developments, the form of $\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})$ and $\boldsymbol{G}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})$ will be derived for each of the associated models of the gravitational potential.

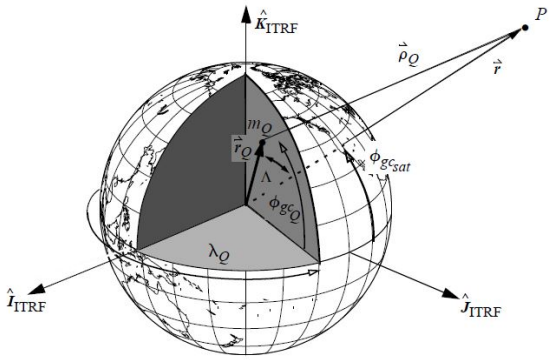


Figure 7.3: Schematics of the Gravitational Field Derivation

7.2.3 Spherical Harmonics Model

$$dU = G \cdot \frac{dm_q}{\rho_q} \tag{7.18}$$

where m_Q is the infinitesimal element of mass, at distance ρ_q to the satellite. The total potential hence corresponds to the integral:

$$U = G \cdot \int_{\text{body}} \frac{1}{\rho_q} dm_q \tag{7.19}$$

Defining r, the ECEF position of the satellite, and r_q, the ECEF vector to the infinitesimal mass element:

$$r = r_{ECEF} = \sqrt{x^2+y^2+z^2} \qquad r_q = r_{q,ECEF} = \sqrt{\xi^2+\eta^2+\zeta^2}, \tag{7.20}$$

Using the law of cosines the range to the satellite ρ_q, from the infinitesimal mass element can be expressed as:

$$\rho_q^2 = r^2 + r_q^2 - 2rr_q \cos\Lambda, \tag{7.21}$$

where Λ is the angle between r and r_q

$$\cos\Lambda = \frac{\boldsymbol{r}\boldsymbol{r}_q}{rr_q} \tag{7.22}$$

The range can hence be defined:

$$\rho_q = r\sqrt{1-2\frac{r_q}{r}\gamma+(\frac{r_q}{r})^2} = r\sqrt{1-2\alpha\gamma+\alpha^2} \tag{7.23}$$

$$\alpha = \frac{r_q}{r} \qquad \gamma = \cos\Lambda \tag{7.24}$$

The potential then can be written as:

$$U = G \cdot \int_{\text{body}} \frac{dm_q}{r\sqrt{1-2\alpha\gamma+\alpha^2}} \tag{7.25}$$

α will always be less than 1.0 for a point outside the central body, and then the absolute value of γ will always be less than or equal one. Using again our beloved binominal theorem to expand the denominator we find:

$$\frac{1}{r\sqrt{1-2\alpha\gamma+\alpha^2}} = \frac{1}{\sqrt{1+\tilde{x}}} = \sum_{l=0} \alpha^l P_l[\gamma] \tag{7.26}$$

So, here our Legendre polynomial series pops up, P_l. The argument of the polynomial is written in []. The so-called Rodriguez formula gives the conventional Legendre polynomials.

$$P_l[\gamma] = \frac{1}{2^l l!}\frac{d^l(\gamma^2-1)^l}{d\gamma^l} \tag{7.27}$$

$$P_l[\gamma] = \frac{1}{2^l}\sum_{j=0}^{\lfloor \frac{l}{2} \rfloor}\frac{(-1)^j(2l-2j)!}{j!(l-j)!(l-2j)!}\gamma^{l-2j} \tag{7.28}$$

Higher order Legendre polynomials are most efficiently computed via the recursive formula:

$$P_l[\gamma] = \frac{2l-1}{l}\gamma P_{l-1}[\gamma] - \frac{l-1}{l}P_{l-2}[\gamma] \tag{7.29}$$

$$\frac{dP_{l+1}[\gamma]}{d\gamma} = (l+1)P_l[\gamma] + u\frac{dP_l(\gamma)}{d\gamma}. \tag{7.30}$$

The potential can hence be rewritten in the following way:

Table 7.1: Legendre Polynomials and their Derivatives

Degree	Legendre Polynomial
0	1
1	γ
2	$\frac{1}{2}(3\gamma^2-1)$
3	$\frac{1}{2}(5\gamma^3-3\gamma)$
4	$\frac{1}{8}(35\gamma^4-30\gamma^2+3)$
5	$\frac{1}{8}(63\gamma^5-70\gamma^3+15\gamma)$

$$U = \frac{G}{r}\cdot\int_{\text{body}}\sum_{l=0}\alpha^l P_l[\gamma]dm_q \tag{7.31}$$

This form is of limited use because we cannot directly find Λ. Spherical geometry allows to develop an equation in the following way. So using the spherical geometry illustrated in Fig.7.3 the following relation can be found:

$$\cos\Lambda = \cos(\pi/2-\phi_{gc,q})\cos(\pi/2-\phi_{gc,sat}) + \sin(\pi/2-\phi_{gc,q})\sin(\pi/2-\phi_{gc,sat})\cos(\lambda_q-\lambda_{sat}) \tag{7.32}$$

$$\cos\Lambda = \sin(\phi_{gc,q})\sin(\phi_{gc,sat}) + \cos(\phi_{gc,q})\cos(\phi_{gc,sat})\cos(\lambda_q-\lambda_{sat}) \tag{7.33}$$

Applying the decomposition formula of spherical harmonics, leads to the following substitions of Λ:

$$P_l[\gamma] = P_l[\cos(\Lambda)] = P_l[\sin(\phi gc,q)]P_l[\sin(\phi gc,sat)] + \sum_{m=1}^{l}\frac{(l-m)!}{(l+m)!}(A_{l,m}A'_{l,m} + B_{l,m}B'_{l,m}), \tag{7.34}$$

with

$$A_{l,m} = P_{l,m}[\sin(\phi gc,q)]\cos(m\lambda_q) \tag{7.35}$$

$$A'_{l,m} = P_{l,m}[\sin(\phi gc,sat)]\cos(m\lambda_{sat}) \tag{7.36}$$

$$B_{l,m} = P_{l,m}[\sin(\phi gc,q)]\sin(m\lambda_q) \tag{7.37}$$

$$B'_{l,m} = P_{l,m}[\sin(\phi gc,sat)]\sin(m\lambda_{sat}) \tag{7.38}$$

$P_{0,0}$	1	$P_{3,2}$	$15\cos^2(\phi_{gc_{sat}})\sin(\phi_{gc_{sat}})$
$P_{1,0}$	$\sin(\phi_{gc_{sat}})$	$P_{3,3}$	$15\cos^3(\phi_{gc_{sat}})$
$P_{1,1}$	$\cos(\phi_{gc_{sat}})$	$P_{4,0}$	$\frac{1}{8}\{35\sin^4(\phi_{gc_{sat}})-30\sin^2(\phi_{gc_{sat}})+3\}$
$P_{2,0}$	$\frac{1}{2}\{3\sin^2(\phi_{gc_{sat}})-1\}$	$P_{4,1}$	$\frac{5}{2}\cos(\phi_{gc_{sat}})\{7\sin^3(\phi_{gc_{sat}})-3\sin(\phi_{gc_{sat}})\}$
$P_{2,1}$	$3\sin(\phi_{gc_{sat}})\cos(\phi_{gc_{sat}})$	$P_{4,2}$	$\frac{15}{2}\cos^2(\phi_{gc_{sat}})\{7\sin^2(\phi_{gc_{sat}})-1\}$
$P_{2,2}$	$3\cos^2(\phi_{gc_{sat}})$	$P_{4,3}$	$105\cos^3(\phi_{gc_{sat}})\sin(\phi_{gc_{sat}})$
$P_{3,0}$	$\frac{1}{2}\{5\sin^3(\phi_{gc_{sat}})-3\sin(\phi_{gc_{sat}})\}$	$P_{4,4}$	$105\cos^4(\phi_{gc_{sat}})$
$P_{3,1}$	$\frac{1}{2}\cos(\phi_{gc_{sat}})\{15\sin^2(\phi_{gc_{sat}})-3\}$		

Figure 7.4: Associated Legendre functions

l and m are normally called degree and order, respectively. This gives us the opportunity to introduce the associated Legendre polynomials:

$$P_{l,m}[\gamma] = \frac{1}{2^l l!}(1-\gamma^2)^{m/2}\frac{d^{l+m}}{d\gamma^{l+m}}(\gamma^2-1)^l \tag{7.39}$$

or alternatively

$$P_{l,m}[\gamma] = (1-\gamma^2)^{m/2}\frac{d^m}{d\gamma^m}P_l[\gamma] \tag{7.40}$$

Note: for zero order (m=0) the associated Legendre polynomials are simple the conventional Legendre polynomials. An important trick is now to separate all terms that are independent of the satellite's location and those which are dependent. This allows to isolate terms that only depend on the central body and can be precomputed:

$$C'_{l,m} = \int_{body} r_q^l \frac{(l-m)!}{(l+m)l!}P_{l,m}[\sin\phi_{gc,q}]\cos(m\lambda_q)dm_q \tag{7.41}$$

$$S'_{l,m} = \int_{body} r_q^l \frac{(l-m)!}{(l+m)l!}P_{l,m}[\sin\phi_{gc,q}]\sin(m\lambda_q)dm_q \tag{7.42}$$

The coefficients represent the mathematical model for the Earth's shape in spherical harmonics.A special case for the zonal harmonics is:

$$C'_{l,0} = \int_{body} r_q^l P_l[\sin\phi_{gc,q}]dm_q \tag{7.43}$$

$$\tag{7.44}$$

which can be represented with the conventional Legendre polynomials. For convenience, the gravitational coefficients

C and S can be normalized to be dimensioneless.

$$C'_{l,m} = C_{l,m}R_E m_E, \tag{7.45}$$

$$S'_{l,m} = S_{l,m}R_E m_E, \tag{7.46}$$

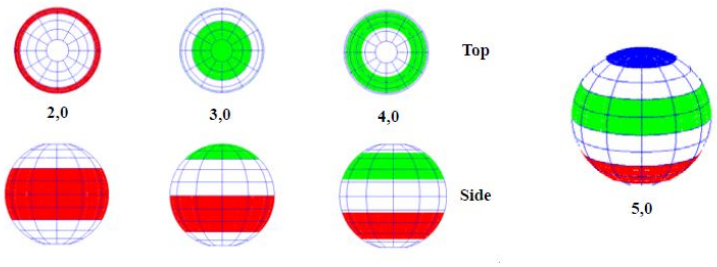


Figure 7.5: Zonal harmonics, courtesy Vallado.

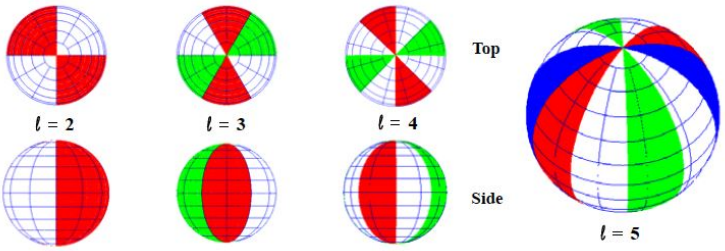


Figure 7.6: Sectorial harmonics, courtesy Vallado.

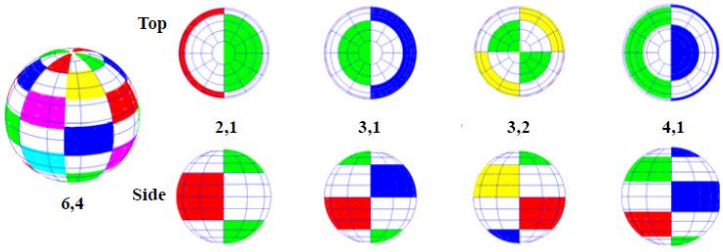


Figure 7.7: Tesseral harmonics, courtesy Vallado.

where R_E is the mean Earth's radius and m_E is the mean Earth's mass. The Earth gravitational potential can hence be represented as:

$$U = \frac{\mu}{r}\sum_{l=0}^{\inf}\sum_{m=0}^{l}(\frac{R_E}{r})^l P_{l,m}[\sin\phi_{gc,sat}]\big(C_{l,m}\cos(m\lambda_{sat}) + S_{l,m}\sin(m\lambda_{sat})\big) \tag{7.47}$$

This is one possible representation. However, $S_{1,0}$ is zero per definition, if the center of the coordinate system coincides with the center of mass of the attracting body, $C_{1,0} = C_{1,1} = S_{1,1} = 0$. Hence the summations can be adapted leading to the following slightly different representation of the potential:

$$U = \frac{\mu}{r} + \frac{\mu}{r}\sum_{l=2}^{\inf}\sum_{m=0}^{l}(\frac{R_E}{r})^l P_{l,m}[\sin\phi_{gc,sat}]\big(C_{l,m}\cos(m\lambda_{sat}) + S_{l,m}\sin(m\lambda_{sat})\big) \tag{7.48}$$

Sometimes the *J-notation* is used, for the zonal harmonics:

$$-C_{l,0} = J_l \tag{7.49}$$

Another alternative is to separate the zonal harmonics from the tesseral terms:

$$U = \frac{\mu}{r} - \frac{\mu}{r}\sum_{l=2}^{\inf} J_l(\frac{R_E}{r})^l P_l[\sin\phi_{gc,sat}] + \frac{\mu}{r}\sum_{l=2}^{\inf}\sum_{m=1}^{l}(\frac{R_E}{r})^l P_{l,m}[\sin\phi_{gc,sat}]\big(C_{l,m}\cos(m\lambda_{sat}) + S_{l,m}\sin(m\lambda_{sat})\big) \tag{7.50}$$

7.2.3.1 Computationally Efficient Methods

Another representation is to use the so-called normalized representation, this is computationally advantageous.

$$U = \frac{\mu}{r}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{R_E}{r}\right)^l \bar{P}_{l,m}(u)\left[\bar{C}_{l,m}r_m(s,t) + \bar{S}_{l,m}i_m(s,t)\right], \tag{7.51}$$

where μ is the gravitational parameter, and $\bar{C}_{n,m}$ and $\bar{S}_{n,m}$ are the normalized spherical harmonics mass coefficients of the gravitating body. Furthermore, r is the magnitude of the position vector from the center of mass of the gravitating body to the spacecraft, and s, t, and u make up the directions of the unit vector pointing to the spacecraft from the center of the body, such that the position unit-vector (expressed in planet-fixed coordinates) is given by

$$\frac{\boldsymbol{r}^{ECEF}}{r} = \begin{bmatrix} s \\ t \\ u \end{bmatrix} = \begin{bmatrix} \cos\phi_{sat}\cos\lambda_{sat} \\ \cos\phi_{sat}\sin\lambda_{sat} \\ \sin\phi_{sat} \end{bmatrix},$$

where ϕ_{sat} and λ_{sat} are the body-centric spherical latitude and longitude, respectively. $\bar{P}_{l,m}(u)$ is the set of normalized derived Legendre polynomials given by

$$\bar{P}_{l,m}(u) = N_{l,m}P_{l,m}(u) \qquad \text{where} \qquad P_{l,m}(u) = \frac{1}{2^l l!}\frac{d^{l+m}}{du^{l+m}}\left(u^2-1\right)^l. \tag{7.52}$$

Here, $N_{l,m}$ is a normalizing factor which serves to aid in the numerical computation of the spherical harmonics expansion, and is given by

$$N_{l,m} = \left[\frac{(l-m)!\,(2l+1)\,(2-\delta_{0,m})}{(l+m)!}\right]^{1/2}, \quad \delta_{0,m} = \begin{cases} 1 & , \quad m = 0 \\ 0 & , \quad m > 0 \end{cases}.$$

Finally, the terms $r_m(s,t)$ and $i_m(s,t)$ are

$$r_m(s,t) = \mathrm{Re}\{(s+jt)^m\} \qquad \text{and} \qquad i_m(s,t) = \mathrm{Im}\{(s+jt)^m\}, \tag{7.53}$$

where $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ indicate the real and imaginary parts of the input complex-valued number and $j = \sqrt{-1}$ is the imaginary number. In practical implementations, the infinite sum in Eq. (7.51) is replaced by a finite sum. In subsequent developments we leave this as an infinite sum with the understanding that the sum will be truncated.

7.2.3.1.1 Recursion Relationships In order for the uniform representation of the gravitational potential to be utilized via computational means, it is necessary to formulate recursion relationships for quantities such as $\bar{P}_{n,m}(u)$, $r_m(s,t)$, and $i_m(s,t)$. These recursions then allow for faster, more reliable computation of the desired parameters for use in simulation.

7.2.3.1.1.1 Recursions for $\bar{P}_{l,m}(u)$ A more detailed development of the recursion formulas for the non-normalized derived Legendre polynomials is given by Pines [58] and a development of the recursion formulas for the normalized derived Legendre polynomials is given by Lundberg [43]. We can think of the terms $\bar{P}_{n,m}(u)$ as the elements of a lower-triangular matrix. It is a lower-triangular matrix because all elements which would lie along the diagonal do not involve the parameter u and hence all elements to the right of diagonal will be zero as seen by the definition of the derived Legendre polynomial. This helps in establishing recursions as "diagonal," "off-diagonal," or "column." Thus $\bar{P}_{0,0}$ would be the upper leftmost element, increasing n would increase the row index, and increasing m would increase the column index. A numerically stable recursion for a column (fixed m and varying n) is given by [43]

$$\begin{aligned}\bar{P}_{l,m}(u) = &\left[\frac{(2l+1)(2l-1)}{(l+m)(l-m)}\right]^{1/2} u\bar{P}_{l-1,m}(u) \\ &- \left[\frac{(2l+1)(l+m-1)(l-m-1)}{(2l-3)(l+m)(l-m)}\right]^{1/2} \bar{P}_{l-2,m}(u)\,.\end{aligned} \tag{7.54}$$

Note that this recursion requires the terms $\bar{P}_{l-1,m}(u)$ and $\bar{P}_{l-2,m}(u)$ in order to calculate the term $\bar{P}_{l,m}(u)$. This means that the two previous elements of the column must be present in order to calculate the current element, such that if given the diagonal element and the element immediately below it, one entire column of the "matrix" may be determined. Assuming that the diagonal element is known, it can be shown that the element immediately below the diagonal element is given by

$$\bar{P}_{l+1,l}(u) = [(2l+3)]^{1/2} u\bar{P}_{l,l}(u)\,. \tag{7.55}$$

Therefore, if the diagonal of the matrix can be populated then the first off-diagonal can be populated and the above column recursion can be utilized to complete the matrix one column at a time. It can be shown that the diagonal elements of the matrix are determined via the recursion

$$\bar{P}_{l,l}(u) = \left[\mathscr{S}_l\left(1+\frac{1}{2l}\right)\right]^{1/2} \bar{P}_{l-1,l-1}(u) \quad , \quad \mathscr{S}_l = \begin{cases} 2 & , \quad l=1 \\ 1 & , \quad l>1 \end{cases} , \tag{7.56}$$

which is initialized with $\bar{P}_{0,0}(u) = 1$. Given the value of $\bar{P}_{0,0}(u)$, the diagonal terms may be populated using Eq. (7.56), the first off-diagonal terms may be populated using Eq. (7.55) and the columns may be populated one at a time using Eq. (7.54), and therefore the entire set of the normalized derived Legendre polynomials can be obtained for a given value of u.

7.2.3.1.1.2 Recursions for $r_m(s,t)$ and $i_m(s,t)$ From the definitions of $r_m(s,t)$ and $i_m(s,t)$ given in Eq. (7.53) and manipulation to relate the m^{th} terms to the previous terms, it can be shown that $r_m(s,t)$ and $i_m(s,t)$ satisfy the recursions

$$r_m(s,t) = sr_{m-1}(s,t) - ti_{m-1}(s,t) \qquad \text{and} \qquad i_m(s,t) = si_{m-1}(s,t) + tr_{m-1}(s,t)\,,$$

which are initialized via

$$r_0(s,t) = 1 \qquad \text{and} \qquad i_0(s,t) = 0\,.$$

7.2.3.1.2 Derivative Relationships Before computing the actual derivatives of the potential, it is convenient to establish relationships on the derivatives of the terms $\bar{P}_{l,m}(u)$, $r_m(s,t)$, and $i_m(s,t)$. These relationships will then be used to establish more general derivatives in the subsequent developments.

7.2.3.1.2.1 Derivatives of $\bar{P}_{l,m}(u)$ The set of normalized derived Legendre polynomials is functionally dependent on the parameter u alone; therefore, the only derivative which will be required is the derivative of the normalized polynomials with respect to the parameter u. From the definition of the derived Legendre polynomials in Eq. (7.52), it is seen that

$$\frac{\partial}{\partial u}\left\{P_{l,m}(u)\right\} = P_{l,m+1}(u)\,.$$

Therefore, utilizing the normalization factor to find the derivative of the normalized derived Legendre polynomials yields

$$\frac{\partial}{\partial u}\left\{\bar{P}_{l,m}(u)\right\} = \frac{\partial}{\partial u}\left\{N_{l,m}P_{l,m}(u)\right\} = N_{l,m}P_{l,m+1}(u) = \frac{N_{l,m}}{N_{l,m+1}}\bar{P}_{l,m+1}(u)\,.$$

Define a parameter $\lambda_{l,m}$ to be the ratio of the $N_{l,m}$ normalization factor to the $N_{l,m+1}$ normalization factor. Thus,

$$\lambda_{l,m} = \frac{N_{l,m}}{N_{l,m+1}} = \left[\mathscr{S}_m(l-m)(l+m+1)\right]^{1/2}\,, \qquad \mathscr{S}_m = \begin{cases} \frac{1}{2} & , \quad m=0 \\ 1 & , \quad m>0 \end{cases}\,,$$

and the derivative may be written as

$$\frac{\partial}{\partial u}\left\{\bar{P}_{l,m}(u)\right\} = \lambda_{l,m}\bar{P}_{l,m+1}(u)\,. \tag{7.57}$$

7.2.3.1.2.2 Derivatives of $r_m(s,t)$ and $i_m(s,t)$ The terms $r_m(s,t)$ and $i_m(s,t)$ depend functionally only on the parameters s and t, and so each terms derivative with respect to the parameters s and t must be obtained. From the definition of $r_m(s,t)$ in Eq. (7.53), it is seen that

$$\frac{\partial\, r_m(s,t)}{\partial s} = \frac{\partial}{\partial s}\left\{\mathrm{Re}\left\{(s+jt)^m\right\}\right\} = \mathrm{Re}\left\{m(s+jt)^{m-1}\right\} = m r_{m-1}(s,t)\,. \tag{7.58}$$

Similarly, the remaining derivative relationships can be found as

$$\frac{\partial\, r_m(s,t)}{\partial t} = -m i_{m-1}(s,t)\,, \quad \frac{\partial\, i_m(s,t)}{\partial s} = m i_{m-1}(s,t)\,, \tag{7.59a}$$

$$\text{and} \quad \frac{\partial\, i_m(s,t)}{\partial t} = m r_{m-1}(s,t)\,. \tag{7.59b}$$

7.2.3.1.3 The Gravitational Acceleration Vector Following the process of Pines [58], it can be shown that the gravitational acceleration vector of Eq. (7.15) is given by

$$\boldsymbol{g}(\boldsymbol{r}^f, \boldsymbol{\theta}) = \begin{bmatrix} g_1 + s g_4 \\ g_2 + t g_4 \\ g_3 + u g_4 \end{bmatrix}\,. \tag{7.60}$$

Define a set of combined mass coefficients as

$$\begin{aligned}
\bar{D}_{l,m}(s,t) &= \bar{C}_{l,m} r_m(s,t) + \bar{S}_{l,m} i_m(s,t) \\
\bar{E}_{l,m}(s,t) &= \bar{C}_{l,m} r_{m-1}(s,t) + \bar{S}_{l,m} i_{m-1}(s,t) \\
\bar{F}_{l,m}(s,t) &= \bar{S}_{l,m} r_{m-1}(s,t) - \bar{C}_{l,m} i_{m-1}(s,t) \\
\bar{G}_{l,m}(s,t) &= \bar{C}_{l,m} r_{m-2}(s,t) + \bar{S}_{l,m} i_{m-2}(s,t) \\
\bar{H}_{l,m}(s,t) &= \bar{S}_{l,m} r_{m-2}(s,t) - \bar{C}_{l,m} i_{m-2}(s,t)\,.
\end{aligned}$$

Then, making use of the derivative relationships described by Eqs. (7.57)–(7.59), it can be shown that the gravity coefficients are

$$g_1 = \frac{\mu}{r^2}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m\bar{P}_{n,m}(u)\bar{E}_{l,m}(s,t) \tag{7.61a}$$

$$g_2 = \frac{\mu}{r^2}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m\bar{P}_{n,m}(u)\bar{F}_{l,m}(s,t) \tag{7.61b}$$

$$g_3 = \frac{\mu}{r^2}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l \lambda_{l,m}\bar{P}_{n,m}(u)\bar{D}_{l,m}(s,t) \tag{7.61c}$$

$$-g_4 = \frac{\mu}{r^2}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l \left[(l+m+1)\bar{P}_{l,m}(u)+\lambda_{l,m}u\bar{P}_{l,m+1}(u)\right]\bar{D}_{l,m}(s,t)\,, \tag{7.61d}$$

where we recall that

$$\lambda_{l,m} = \left[\mathscr{S}_m(l-m)(l+m+1)\right]^{1/2}\,, \quad \mathscr{S}_m = \begin{cases} \frac{1}{2} & , \quad m=0 \\ 1 & , \quad m>0 \end{cases}\,.$$

Note that the g_1 and g_2 are the same as shown by Pines [58] due to the fact that the normalization procedure affects only the derivatives of terms involving the parameter u. Therefore, while g_1 and g_2 remain the same (modulo the difference caused by normalization) the terms g_3 and g_4 are different.

7.2.3.1.4 The Gravitational Jacobian Matrix Similar to the development of the gravitational acceleration vector, following the method described in Pines [58], it can be shown that the gravitational Jacobian of Eq. (7.17) is given by

$$\boldsymbol{G}(\boldsymbol{r}^f,\boldsymbol{\theta}) = \left[\begin{array}{c|c|c} g_{11}+2sg_{41}+s^2g_{44}+g_4/r & g_{12}+tg_{41}-sg_{42}+stg_{44} & g_{13}+ug_{41}+sg_{43}+sug_{44} \\ g_{12}+tg_{41}-sg_{42}+stg_{44} & -g_{11}+2tg_{42}+t^2g_{44}+g_4/r & g_{23}+ug_{42}+tg_{43}+tug_{44} \\ g_{13}+ug_{41}+sg_{43}+sug_{44} & g_{23}+ug_{42}+tg_{43}+tug_{44} & g_{33}+2ug_{43}+u^2g_{44}+g_4/r \end{array}\right]. \tag{7.62}$$

Again, making use of the derivative relationships in Eqs. (7.57)–(7.59), it can be shown that

$$\begin{aligned} g_{11} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m(m-1)\bar{P}_{l,m}(u)\bar{G}_{l,m}(s,t) \\ g_{12} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m(m-1)\bar{P}_{l,m}(u)\bar{H}_{l,m}(s,t) \\ g_{13} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m\lambda_{l,m}\bar{P}_{l,m+1}(u)\bar{E}_{l,m}(s,t) \\ g_{23} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m\lambda_{l,m}\bar{P}_{l,m+1}(u)\bar{F}_{l,m}(s,t) \\ g_{33} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l m\zeta_{l,m}\bar{P}_{l,m+2}(u)\bar{D}_{l,m}(s,t) \end{aligned} \tag{7.63}$$

$$
\begin{aligned}
-g_{41} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l\left[m(l+m+1)\bar{P}_{l,m}(u)+m\lambda_{l,m}u\bar{P}_{l,m+1}\right]\bar{E}_{l,m}(s,t)\\
-g_{42} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l\left[m(l+m+1)\bar{P}_{l,m}(u)+m\lambda_{l,m}u\bar{P}_{l,m+1}\right]\bar{F}_{l,m}(s,t)\\
-g_{43} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l\left[(l+m+1)\lambda_{l,m}\bar{P}_{l,m+1}(u)+\zeta_{l,m}u\bar{P}_{l,m+2}\right]\bar{D}_{l,m}(s,t)\\
g_{44} &= \frac{\mu}{r^3}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left(\frac{a_e}{r}\right)^l\left[(l+m+1)(l+m+3)\bar{P}_{l,m}(u)\right.\\
&\qquad\left.+(2l+2m+4)\lambda_{l,m}u\bar{P}_{l,m+1}(u)+\zeta_{l,m}u^2\bar{P}_{l,m+2}(u)\right]\bar{D}_{l,m}(s,t)\,,
\end{aligned}
$$

where

$$
\zeta_{l,m} = \left[\mathscr{S}_m(l-m)(l-m-1)(l+m+1)(l+m+2)\right]^{1/2}\,,\quad \mathscr{S}_m = \begin{cases} \frac{1}{2} & , \quad m=0 \\ 1 & , \quad m>0 \end{cases}\,.
$$

7.2.3.2 Zonal Harmonics Gravitational Acceleration

For a gravitational field modeled with zonal harmonics, the gravitational potential is given by [64]

$$
U = -\frac{\mu}{r}\sum_{l=0}^{\infty}\left(\frac{R_E}{r}\right)^l P_l(u)J_l\,, \tag{7.64}
$$

where μ is the gravitational parameter of the body, R_E is the mean Earth radius, r is the distance from the center of the body to the satellite, $u=\sin\phi_{sat}$, ϕ_{sat} is the spherical latitude of the satellite, J_l is the l^{th} zonal harmonic of the body in the J notation, and $P_l(u)$ is the conventional Legendre polynomial of degree l. As a reminder, the conventional Legendre polynomials are defined as

$$
P_l(u) = \frac{1}{2^l l!}\frac{d^l}{du^l}(u^2-1)^l\,,
$$

and can be shown to satisfy the recursions [64, 58]

$$
P_l(u) = \frac{2n-1}{l}uP_{l-1}(u) = \frac{n-1}{l}P_{l-2}(u) \tag{7.65a}
$$

$$
\frac{dP_{l+1}(u)}{du} = (l+1)P_l(u)+u\frac{dP_l(u)}{du}\,. \tag{7.65b}
$$

In practical applications, the infinite summation is truncated to enable computation.

Typically, low degree representations of the zonal harmonics potential are implemented so as to capture the dominant effects due to asphericity of the body without involving overburdening computation.

In the sequel, we will restrict our treatment of the zonal harmonics model to a maximum degree of 4, that is we truncate the infinite summation at 4 to develop equations for the gravitational acceleration vector and Jacobian matrix.

However, it should be noted that truncation at a higher degree is merely an extension of the given treatment. In the subsequent developments we leave this as an infinite sum with the understanding that the sum is to be truncated for implementation.

Having established the form of the gravitational potential in Eq. (7.64), we now turn towards developing a relationship for

$$
\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF},\boldsymbol{\theta}) = \left[\frac{\partial U(\boldsymbol{r}^{ECEF},\boldsymbol{\theta})}{\partial \boldsymbol{r}^{ECEF}}\right]^T\,. \tag{7.66}
$$

Table 7.2: Legendre Polynomials and their Derivatives

Degree	Legendre Polynomial	Derivative
0	1	0
1	u	1
2	$\frac{1}{2}(3u^2-1)$	$3u$
3	$\frac{1}{2}(5u^3-3u)$	$\frac{3}{2}(5u^2-1)$
4	$\frac{1}{8}(35u^4-30u^2+3)$	$\frac{5}{2}(7u^3-3u)$
5	$\frac{1}{8}(63u^5-70u^3+15u)$	$\frac{5}{8}(63u^4-42u^2+3)$

Let the fixed-frame position vector be given by $\boldsymbol{r}^{ECEF} = [x\ y\ z]^T$, which yields the relationship that $u = z/r$, and define

$$g_1 = \frac{\partial U(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})}{\partial x}, \qquad g_2 = \frac{\partial U(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})}{\partial y}, \qquad \text{and} \qquad g_3 = \frac{\partial U(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta})}{\partial z},$$

such that Eq. (7.66) becomes

$$\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}. \tag{7.67}$$

Differentiating the potential in Eq. (7.64) with respect to x, y, and z then yields

$$g_1 = \frac{\mu x}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \left((l+1)P_l(u) + u\frac{dP_l(u)}{du}\right) J_l \tag{7.68a}$$

$$g_2 = \frac{\mu y}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \left((l+1)P_l(u) + u\frac{dP_l(u)}{du}\right) J_l \tag{7.68b}$$

$$g_3 = \frac{\mu z}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \left((l+1)P_l(u) + u\frac{dP_l(u)}{du}\right) J_l - \frac{\mu}{r^2} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \frac{dP_l(u)}{du} J_l. \tag{7.68c}$$

Utilizing the recursion relationship in Eq. (7.65b), Eqs. (7.68) may be rewritten more compactly as

$$g_1 = \frac{\mu x}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_e}{r}\right)^l \frac{dP_{l+1}(u)}{du} J_l \tag{7.69a}$$

$$g_2 = \frac{\mu y}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \frac{dP_{l+1}(u)}{du} J_l \tag{7.69b}$$

$$g_3 = \frac{\mu z}{r^3} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \frac{dP_{l+1}(u)}{du} J_l - \frac{\mu}{r^2} \sum_{l=0}^{\infty} \left(\frac{R_E}{r}\right)^l \frac{dP_l(u)}{du} J_l. \tag{7.69c}$$

Substituting for the Legendre polynomial derivatives from Table 7.2 into Eqs. (7.69), noting that for all gravitational fields $J_0 = -1$, and that $J_1 = 0$ provided that the center of mass coincides with the origin of the coordinate system, it can be shown that the acceleration vector in Eq. (7.67) is given by

$$\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = -\frac{\mu}{r^3}\boldsymbol{r}^f + \sum_{l=2}^{\infty} \frac{\mu R_E^l J_l}{r^{2l+3}} \boldsymbol{r}_{J_l}, \tag{7.70}$$

where, for $l=2$, $l=3$, and $l=4$, we have

$$\boldsymbol{r}_{J_2} = \frac{3}{2}\begin{bmatrix} 5xz^2 - xr^2 \\ 5yz^2 - yr^2 \\ 5z^3 - 3zr^2 \end{bmatrix}, \qquad \boldsymbol{r}_{J_3} = \frac{1}{2}\begin{bmatrix} 35xz^3 - 15xzr^2 \\ 35yz^3 - 15yzr^2 \\ 35z^4 - 30z^2r^2 + 3r^4 \end{bmatrix} \qquad \text{and}$$

$$\boldsymbol{r}_{J_4} = \frac{5}{8}\begin{bmatrix} 63xz^4 - 42xz^2r^2 + 3xr^4 \\ 63yz^4 - 42yz^2r^2 + 3yr^4 \\ 63z^5 - 70z^3r^2 + 15zr^4 \end{bmatrix}.$$

Similar to the acceleration vector, the gravity Jacobian matrix may be found as

$$\boldsymbol{G}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = \frac{\mu}{r^5}\left(3\boldsymbol{r}^{ECEF}(\boldsymbol{r}^{ECEF})^T - r^2\boldsymbol{I}\right) - \sum_{l=2}^{\infty}\frac{\mu R_E^l J_l}{r^{2l+5}}\left((2l+3)\boldsymbol{r}_{J_l}(\boldsymbol{r}^{ECEF})^T - r^2\boldsymbol{G}_{J_l}\right), \tag{7.71}$$

where, for $l=2$, $l=3$, and $l=4$, it can be shown that

$$\boldsymbol{G}_{J_2} = \frac{3}{2}\left[\begin{array}{c|c|c} 5z^2 - 2x^2 - r^2 & -2xy & 8xz \\ -2xy & 5z^2 - 2y^2 - r^2 & 8yz \\ -6xz & -6yz & 9z^2 - 3r^2 \end{array}\right]$$

$$\boldsymbol{G}_{J_3} = \frac{1}{2}\left[\begin{array}{c|c|c} 35z^3 - 30x^2z - 15zr^2 & -30xyz & 75xz^2 - 15xr^2 \\ -30xyz & 35z^3 - 30y^2z - 15zr^2 & 75yz^2 - 15yr^2 \\ -60xz^2 + 12xr^2 & -60yz^2 + 12yr^2 & 80z^3 - 48zr^2 \end{array}\right]$$

$$\boldsymbol{G}_{J_4} = \frac{5}{8}\left[\begin{array}{c|c|c} 63z^4 - 84x^2z^2 - 42z^2r^2 + 12x^2r^2 + 3r^4 & & \\ -84xyz^2 + 12xyr^2 & & \\ -140xz^3 + 60xzr^2 & & \\ \hline -84xyz^2 + 12xyr^2 & & 168xz^3 - 72xzr^2 \\ 63z^4 - 84y^2z^2 - 42z^2r^2 + 12y^2r^2 + 3r^4 & & 168yz^3 - 72yzr^2 \\ -140yz^3 + 60yzr^2 & & 175z^4 - 150z^2r^2 + 15r^4 \end{array}\right].$$

Therefore, given the fixed-frame position of the satellite, the determination of the gravitational acceleration vector is accomplished via Eq. (7.70) and the gravity Jacobian via Eq. (7.71).

7.2.3.2.1 Numerical Considerations The appearance of r^{2n+3} in the denominator of $\boldsymbol{a}_g^f(\boldsymbol{r}^f, \boldsymbol{\theta})$ in Eq. (7.70) and r^{2n+5} in the denominator of $\boldsymbol{G}(\boldsymbol{r}^f, \boldsymbol{\theta})$ in Eq. (7.71) can potentially present numerical issues when r is large. As such, it is desirable to reformulate Eqs. (7.70) and (7.71) to avoid this situation. Let us define $s = x/r$, $t = y/r$, and recall that $u = z/r$. The gravity vector can be written as

$$\boldsymbol{a}_g^{ECEF}(\boldsymbol{r}^{ECEF}, \boldsymbol{\theta}) = -\frac{\mu}{r^2}\boldsymbol{u}^{ECEF} + \sum_{l=2}^{\infty}\frac{\mu J_l}{r^2}\left(\frac{R_E}{r}\right)^l \boldsymbol{u}_{J_l}, \tag{7.73}$$

where

$$\boldsymbol{u}^{ECEF} = \begin{bmatrix} s \\ t \\ u \end{bmatrix},$$

and, for $l=2$, $l=3$, and $l=4$, we have

$$\boldsymbol{u}_{J_2} = \frac{3}{2}\begin{bmatrix} 5su^2 - s \\ 5tu^2 - t \\ 5u^3 - 3u \end{bmatrix}, \qquad \boldsymbol{u}_{J_3} = \frac{1}{2}\begin{bmatrix} 35su^3 - 15su \\ 35tu^3 - 15tu \\ 35u^4 - 30u^2 + 3 \end{bmatrix} \qquad \text{and}$$

$$\boldsymbol{u}_{J_4} = \frac{5}{8}\begin{bmatrix} 63su^4 - 42su^2 + 3s \\ 63tu^4 - 42tu^2 + 3t \\ 63u^5 - 70u^3 + 15u \end{bmatrix}.$$

Similarly, the gravity Jacobian matrix of Eq. (7.71) may be rewritten as

$$\boldsymbol{G}(\boldsymbol{r}^{ECEF},\boldsymbol{\theta}) = \frac{\mu}{r^3}\left(3\boldsymbol{u}^{ECEF}(\boldsymbol{u}^{ECEF})^T - \boldsymbol{I}\right) - \sum_{l=2}^{\infty}\frac{\mu J_l}{r^3}\left(\frac{R_E}{r}\right)^l\left((2l+3)\boldsymbol{u}_{J_l}(\boldsymbol{u}^{ECEF})^T - \boldsymbol{U}_{J_l}\right), \tag{7.74}$$

where, for $l=2$, $l=3$, and $l=4$, it can be shown that

$$\boldsymbol{U}_{J_2} = \frac{3}{2}\left[\begin{array}{c|c|c} 5u^2-2s^2-1 & -2st & 8su \\ -2st & 5u^2-2t^2-1 & 8tu \\ -6su & -6tu & 9u^2-3 \end{array}\right]$$

$$\boldsymbol{U}_{J_3} = \frac{1}{2}\left[\begin{array}{c|c|c} 35u^3-30s^2u-15u & -30stu & 75su^2-15s \\ -30stu & 35u^3-30t^2u-15u & 75tu^2-15t \\ -60su^2+12s & -60tu^2+12t & 80u^3-48u \end{array}\right]$$

$$\boldsymbol{U}_{J_4} = \frac{5}{8}\left[\begin{array}{c|c|c} 63u^4-84s^2u^2-42u^2+12s^2+3 & -84stu^2+12st & \\ -84stu^2+12st & 63u^4-84t^2u^2-42u^2+12t^2+3 & \\ -140su^3+60su & -140tu^3+60tu & \end{array}\right.$$

$$\left.\begin{array}{c} 168su^3-72su \\ 168tu^3-72tu \\ 175u^4-150u^2+15 \end{array}\right].$$

Thus, to avoid the potential numerical difficulties associated with computation of the gravitational acceleration vector of Eq. (7.70) and the gravity Jacobian of Eq. (7.71), it is recommended to use Eqs. (7.73) and (7.74) instead.

7.3 Third Body Perturbations

Third body perturbations are the gravitational perturbations from the sun and moon that affect the satellite dynamics. The gravitational field of the Sun and the moon can savely be modeled as point masses for Earth orbiting objects. Exceptions are of course lunar orbiting missions where the moon is the central body (e.g. GRAIL).

Let the position of the satellite in an inertially-referenced frame be given by $\boldsymbol{r}_S$, the position of the third body (moon, or sun) be given by $\boldsymbol{r}_M$, and the position of the earth be given by $\boldsymbol{r}_E$. Thus, the position of the satellite relative to the earth is given by

$$\boldsymbol{r}_{ES} = \boldsymbol{r}_S - \boldsymbol{r}_E := \boldsymbol{r}. \tag{7.76}$$

The acceleration is found by taking the derivative:

$$\ddot{\boldsymbol{r}}_{ES} = \ddot{\boldsymbol{r}}_S - \ddot{\boldsymbol{r}}_E. \tag{7.77}$$

Now, we define the forces acting on the earth to be using Newton's second law:

$$\sum F_E = m_E\ddot{r}_E = \frac{Gm_Em_S\boldsymbol{r}_{ES}}{r_{ES}^3} + \frac{Gm_Em_M\boldsymbol{r}_{EM}}{r_{EM}^3}, \tag{7.78}$$

where $\boldsymbol{r}_{EM}$ is the vector from the earth to the third body. This means, the satellite pulling on the Earth and the Sun pulling on the Earth. The acceleration $\ddot{r}_E$, is the one an observer at the origin of the inertial coordinate system would see as acting on the Earth. However, we are interested in the forces on the satellite, we can find those from above's equation:

$$\sum F_S = m_S\ddot{r}_S = -\frac{Gm_Em_S\boldsymbol{r}_{ES}}{r_{ES}^3} - \frac{Gm_Mm_S\boldsymbol{r}_{MS}}{r_{MS}^3}. \tag{7.79}$$

The forces are negative because they are in the direction opposite to that of the vectors of the satellite. Now, recalling that the acceleration of a point is given by $\ddot{\boldsymbol{r}} = \sum F/m$, and subsequently substituting Eq.s (7.78) and (7.79) into Eq. (7.77) gives

$$\ddot{\boldsymbol{r}}_{ES} = -\frac{Gm_E\boldsymbol{r}_{ES}}{r_{ES}^3} - \frac{Gm_M\boldsymbol{r}_{MS}}{r_{MS}^3} - \frac{Gm_S\boldsymbol{r}_{ES}}{r_{ES}^3} - \frac{Gm_M\boldsymbol{r}_{EM}}{r_{EM}^3}. \tag{7.80}$$

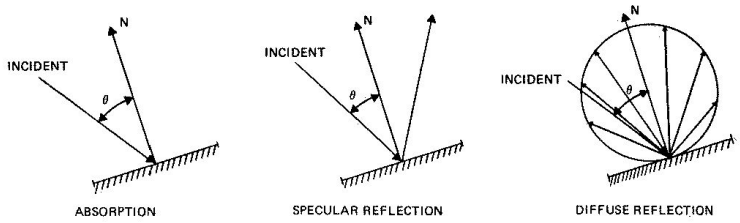


Figure 7.8: Courtesy, Wertz [76]

With some rearranging, using $\boldsymbol{r}_{ES} = -\boldsymbol{r}_{SE}$, we arrive at

$$\ddot{\boldsymbol{r}}_{ES} = -\frac{G(m_E + m_S)\boldsymbol{r}_{ES}}{r_{ES}^3} + Gm_M\Big[\frac{\boldsymbol{r}_{SM}}{r_{SM}^3} - \frac{\boldsymbol{r}_{EM}}{r_{EM}^3}\Big]. \tag{7.81}$$

Now that we have the acceleration of the satellite due to third body forces expressed relative to the earth. This form can be generalized to any number of third bodies. To write it in the most general form and using the definition of $\boldsymbol{r} = \boldsymbol{r}_{ES}$:

$$\ddot{\boldsymbol{r}} = -\frac{G(m_E + m_S)\boldsymbol{r}}{r^3} - G\sum_{i=1}^{n} m_i\big(\frac{\boldsymbol{r} - \boldsymbol{r}_i}{|\boldsymbol{r} - \boldsymbol{r}_i|^3} + \frac{\boldsymbol{r}_i}{r_i^3}\big) \tag{7.82}$$

setting the satellite mass to zero (because it is negligible relative to the other masses involved), we arrive at:

$$\ddot{\boldsymbol{r}} = -\frac{G(m_E)\boldsymbol{r}}{r^3} - G\sum_{i=1}^{n} m_i\big(\frac{\boldsymbol{r} - \boldsymbol{r}_i}{|\boldsymbol{r} - \boldsymbol{r}_i|^3} + \frac{\boldsymbol{r}_i}{r_i^3}\big) \tag{7.83}$$

The first term is clearly the point mass acceleration of the Earth, and the second terms in the sum are the ones of potential third bodies. This leads to the following expression for the third body acceleration from the Sun and the Moon:

$$\boldsymbol{a}_{ThirdBody} = -\mu_{sun}\big(\frac{\boldsymbol{r} - \boldsymbol{r}_{sun}}{|\boldsymbol{r} - \boldsymbol{r}_{sun}|^3} + \frac{\boldsymbol{r}_{sun}}{r_{sun}^3}\big) - \mu_{moon}\big(\frac{\boldsymbol{r} - \boldsymbol{r}_{moon}}{|\boldsymbol{r} - \boldsymbol{r}_{moon}|^3} + \frac{\boldsymbol{r}_{moon}}{r_{moon}^3}\big) \tag{7.84}$$

7.3.0.1 Jacobian

The Jacobian is readily derived from the Eq.7.84:

$$\frac{\partial \boldsymbol{a}_{ThirdBody}}{\partial \boldsymbol{r}} = \sum_{i=1,2} \mu_i\big(\frac{1}{|\boldsymbol{r} - \boldsymbol{r}_i|^3}\big)\boldsymbol{I} + 3\frac{(\boldsymbol{r} - \boldsymbol{r}_i)(\boldsymbol{r} - \boldsymbol{r}_i)^T}{|\boldsymbol{r} - \boldsymbol{r}_i|^5}\big) \tag{7.85}$$

$$\frac{\partial \boldsymbol{a}_{ThirdBody}}{\partial \boldsymbol{v}} = \boldsymbol{0} \tag{7.86}$$

7.4 Direct Solar Radiation Pressure

7.4.1 Flat Surface

The interaction of light with a surface material may be described in a number of ways. One option is a full bidirectional reflection function (BRDF). However, a simple approximation is to represent all materials as a mixture of three different processes, specular reflection, Lambertian diffuse reflection and absorption, which are weighted against each other according to material properties with the coefficients C_a, C_s, C_d. If the material is opaque those coefficients add up to one. The flux on a surface is given by the solar flux, which is equal to the solar constant E at Earth surface, divided by the speed of light c. To find the flux at the object's position it has to be scaled to the appropriate distance. The force acting on a surface is given by the following expression:

$$\boldsymbol{F}_{\text{rad}} = \frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x} - \boldsymbol{x}_{Sun}|^2} \cdot \boldsymbol{f}(A) \tag{7.87}$$

m is the total mass of the object, A_{Earth} the astronomical unit, $\boldsymbol{x}_{Sun}$ the geocentric position of the sun, c velocity of light, $\boldsymbol{S}$ the unit direction of the radiation source, $\boldsymbol{x}$ is the position vector of the object and $\boldsymbol{f}(A)$ the area dependent acceleration function.It is sometimes referred to as force function, which is strictly speaking not correct, as this would neglect the mass term.

The reflection function consists of three different parts according to the different kinds of reflection. The absorption exerts the following acceleration on an infinitesimal surface dA:

$$\boldsymbol{d}f(A)_{\text{abs}} = -C_a \cos\theta \boldsymbol{S} dA, \tag{7.88}$$

where θ is the angle between the unit face normal $\boldsymbol{N}$ and the sun vector $\boldsymbol{S}$. The specular reflection is reflected back in the direction $(\cos\theta \cdot (\boldsymbol{S} - 2\boldsymbol{N}\cos\theta))$, leading to:

$$\boldsymbol{d}f(A)_{\text{spec}} = C_s(\cos\theta \boldsymbol{S} - 2\cos^2\theta \boldsymbol{N} - \cos\theta \boldsymbol{S})dA = -2C_s \cos^2\theta \boldsymbol{N}. \tag{7.89}$$

The diffuse reflection is of a Lambertian surface is distributed proportional to $\cos\phi$, where ϕ is the angle between the reflected radiation and $\boldsymbol{N}$. Integrating over all reflection directions leads to:

$$\boldsymbol{d}f(A)_{\text{dif}} = C_d(-\frac{2}{3}\cos\theta \boldsymbol{N} - \cos\theta \boldsymbol{S})dA. \tag{7.90}$$

Taking all terms together leads to the following expression:

$$\boldsymbol{f}(A) = -\int_A \boldsymbol{SN}[(1-C_{\text{s}})\boldsymbol{S} + 2(C_{\text{s}} \cdot \boldsymbol{SN} + \frac{1}{3}C_{\text{d}})\boldsymbol{N}]dA \qquad \text{for:} \quad 0 < \arccos(\boldsymbol{SN}_i) < \pi/2,$$

where the convention $\cos\theta = \boldsymbol{SN}$ and $C_d + C_s + C_a = 1$ are used. Analytic solutions can be found for some simple shapes.

For a flat surface with area A, the acceleration function is:

$$\boldsymbol{f}_{\text{rad,flat}} = -A\boldsymbol{SN}[(1-C_{\text{s}})\boldsymbol{S} + 2(C_{\text{s}} \cdot \boldsymbol{SN} + \frac{1}{3}C_{\text{d}})\boldsymbol{N}] \qquad \text{for:} \quad 0 < \arccos(\boldsymbol{SN}_i) < \pi/2,$$

note that in contrary to the sphere a visibility constraint has to be applied.

$$\begin{aligned} \boldsymbol{F}_{\text{rad,flat}} &= -\frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x} - \boldsymbol{x}_{Sun}|^2} \cdot A\boldsymbol{SN}[(1-C_{\text{s}})\boldsymbol{S} + 2(C_{\text{s}} \cdot \boldsymbol{SN} + \frac{1}{3}C_{\text{d}})\boldsymbol{N}] && (7.91) \\ \boldsymbol{a}_{\text{rad,flat}} &= \frac{\boldsymbol{F}_{\text{rad,flat}}}{m} \\ &= -\frac{A}{m}\frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x} - \boldsymbol{x}_{Sun}|^2} \cdot \boldsymbol{SN}[(1-C_{\text{s}})\boldsymbol{S} + 2(C_{\text{s}} \cdot \boldsymbol{SN} + \frac{1}{3}C_{\text{d}})\boldsymbol{N}] && (7.92) \end{aligned}$$

7.4.2 Sphere

For a spherical surface of radius r we get:

$$\boldsymbol{f}_{\text{rad,sphere}} = -4\pi \text{r}^2(\frac{1}{4} + \frac{1}{9}C_{\text{d}})\hat{\boldsymbol{S}} \tag{7.93}$$

For a sphere sometimes the parameter C_d is replaced by a single value $\tilde{C} = (\frac{1}{4} + \frac{1}{9}C_{\text{d}})$.

$$\boldsymbol{f}_{\text{rad,sphere}} = -4\pi \text{r}^2\tilde{C}\hat{\boldsymbol{S}} \tag{7.94}$$

Satellite	A/m [m^2/kg]
Lageos 1 and 2	0.0007
Starlette	0.001
GPS(Block II)	0.02
Moon	$1.3 \cdot 10^{-10}$

Figure 7.9: AMR example values, courtesy Beutler

Leading to the following expression for the solar radiation pressure force for a sphere:

$$\boldsymbol{f}_{\text{rad,sphere}} = -4\pi r^2 \tilde{C}\hat{\boldsymbol{S}} \tag{7.95}$$

$$\boldsymbol{F}_{\text{rad,sphere}} = -\frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x}-\boldsymbol{x}_{Sun}|^2}\cdot 4\pi r^2 \tilde{C}\hat{\boldsymbol{S}} \tag{7.96}$$

$$\begin{aligned}\boldsymbol{a}_{\text{rad,sphere}} &= \frac{\boldsymbol{F}_{\text{rad,sphere}}}{m} \\ &= -\frac{4\pi r^2}{m}\frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x}-\boldsymbol{x}_{Sun}|^2}\cdot \tilde{C}\hat{\boldsymbol{S}} \qquad (7.97)\\ &= -\frac{A}{m}\frac{E}{c}\frac{A_{Earth}^2}{|\boldsymbol{x}-\boldsymbol{x}_{Sun}|^2}\cdot \tilde{C}\hat{\boldsymbol{S}} \qquad (7.98)\end{aligned}$$

This is the most frequently used equation for the direct solar radiation pressure, the so-called canon ball model.

7.4.3 Cylinder

For a cylinder barrel the following acceleration function can be found:

$$\begin{aligned}\boldsymbol{f}_{\text{rad,cyl}} &= (\sin\phi(1+\frac{1}{3}C_s)\cdot 2rh + (1-C_s)\cos\phi\cdot\pi r^2)\boldsymbol{S} \\ &+((-\frac{3}{4}C_s\sin\phi - \frac{\pi}{6}C_d)\cos\phi\cdot 2rh + 2(C_s\cos\phi+\frac{1}{3}C_d)\cos\phi\cdot\pi r^2)\boldsymbol{Z}\end{aligned} \tag{7.99}$$

The angle ϕ is the angle between the symmetry axis of the cylinder $\boldsymbol{Z}$ and the sun vector $\boldsymbol{S}$. The cylinder is assumed to have a radius r and the height h.

7.5 Atmospheric Drag

Above a height of 50km the density of the neutral atmosphere can be modeled as laminar air current, due to the low density. If we neglect the thermal motion of the molecules, the linear momentum that is transfered by the atmospheric particles to the surface of the satellite can be easily determined.

During a short time interval Δt the velocity $\boldsymbol{v}'$ of the satellite relative to the atmosphere can be assumed to be constant. If we assume that all particles are absorbed, the linear momentum $\Delta\boldsymbol{p}$ lost by the satellite equals the product of the volume that has been passed $v'\Delta tA$, where A is the area of the satellite, with the density $\rho(\boldsymbol{r}$ and the velocity $-\boldsymbol{v}'$ of the molecules relative to the satellite at the current position of the space craft. Hence:

$$\Delta\boldsymbol{p} = -\rho(\boldsymbol{r})A\boldsymbol{v}'^2\Delta t\frac{\boldsymbol{v}'}{|\boldsymbol{v}'|} \tag{7.100}$$

The acceleration can be computed taking the limit of the time difference to zero:

$$\boldsymbol{a}_{drag} = -\rho(\boldsymbol{r})\frac{A}{m}\boldsymbol{v}'^2\frac{\boldsymbol{v}'}{|\boldsymbol{v}'|} \tag{7.101}$$

The acceleration is hence anti-parallel to the velocity of the satellite relative to the Earth fixed system.

However, Eq.7.101 needs to be modified: Only a fraction of the particles are absorved by the satellite surface, and of course the force is surface dependent. Again normally a canon ball model is assumed, leading to the following modification of the formula:

$$\boldsymbol{a}_{drag} = -\frac{C}{2}\rho(\boldsymbol{r})\frac{A}{m}\boldsymbol{v}'^2\frac{\boldsymbol{v}'}{|\boldsymbol{v}'|} \tag{7.102}$$

For a spherical satellite C=2 (independent of the fraction of absorbed versus reflected particles), and full absorption of the particles. In general one can assume a values of:

$$2 \leq C \leq 2.5 \tag{7.103}$$

For the velocity relative to the atmosphere, we do assume that the atmosphere rotates with the Earth, hence leading to our beloved BKE:

$$\boldsymbol{v}' = \dot{\boldsymbol{r}} - \boldsymbol{\omega} \times \boldsymbol{r} \tag{7.104}$$

where $\boldsymbol{r}$ is the ECI position of the satellite and $\boldsymbol{\omega}$ the rotation rate of the Earth. More advanced models take into account that the atmosphere is far from being fixed on the Earth's surface, leading to the following expression (Escobal 1965):

$$\boldsymbol{v}' = \dot{\boldsymbol{r}} - \boldsymbol{\omega} \times \boldsymbol{r} + \begin{bmatrix} v_w(\cos\alpha\sin\delta\cos\beta_w - \sin\alpha\sin\beta_w) \\ v_w(-\sin\alpha\sin\delta\cos\beta_w + \cos\alpha\sin\beta_w) \\ v_w(-\cos\delta\cos\beta_w) \end{bmatrix} \tag{7.105}$$

where v_w is the wind speed, and wind azimuth β_w, and the satellite's right ascension and declination, α, δ, respectively.

In order to determine the change of a satellite over one orbital revolution, it can be helpful to resort to express the drag force in orbital elements, and then integrate over one period. This approach is also known as Gauss variational equations (can be done for all orbital elements and forces). For drag the following expressions in change in semi-major axis a for one orbital revolution (denoted as Δa) and the eccentricitiy e, denoted as Δe, respectively **??**:

$$\Delta a = -2\pi(\frac{C_D A}{m})a^2\rho(\boldsymbol{r}_{\text{peri}})\exp(-\frac{a\cdot e}{H})[B_0 + 2eB_1] \tag{7.106}$$

$$\Delta e = -2\pi(\frac{C_D A}{m})a\rho(\boldsymbol{r}_{\text{peri}})\exp(-\frac{a\cdot e}{H})[B_1 + \frac{e(B_0+B_2)}{2}], \tag{7.107}$$

with H being the scale height, $\rho(\boldsymbol{r}_{\text{peri}})$ is the atmospheric density at the perigee height of the orbit, B_i are the modified Bessel functions of order i with the argument $\frac{a\cdot e}{H}$, $B_i = B_i(\frac{a\cdot e}{H})$. For nearly circular orbits, the expressions simplify to **??**:

$$\Delta a = -2\pi\frac{C_D A}{m}\rho(\boldsymbol{r}_{\text{peri}})a^2 \tag{7.108}$$

Accodingly for nearly-circular orbits, expressions for the change of the orbital period P and eccentricity can be found:

$$\Delta P = -6\pi^2\frac{C_D A}{m}\rho(\boldsymbol{r}_{\text{peri}})\frac{a^2}{\boldsymbol{v}'} \tag{7.109}$$

$$\Delta \boldsymbol{v}' = \pi\frac{C_D A}{m}\rho(\boldsymbol{r}_{\text{peri}})a\boldsymbol{v}' \tag{7.110}$$

$$\Delta e = 0 \tag{7.111}$$

A rought approximation for the lifetime of a satellite is:

$$L \approx -\frac{H}{\Delta a} \tag{7.112}$$

To model the atmospheric neutral density is not an easy task and depends not only on the height but also solar activity etc. The most precise model is MSIS (Mass Spectrometer and Incoherent Scatter), determined by by NASA Goddard Space Flight Center. Simpler models are e.g. the GHOST model, developed by the Russian Academy of Space, or the Jaccia atmospheric model.

Local variations can be accounted for via the barometric height formula at a hright h:

$$\rho(h) \approx \rho_0 \exp(-\frac{h - h_{\text{alt}}}{H_0}) \tag{7.113}$$

where ρ_0 is the density at the references height h_{alt}, and H_0 is the scaling height. This can be understood as a coarse approximation, see Fig.7.10 for values.

Earth Satellite Parameters

	25	26	27	28	29	30	31	32
		ATMOSPHERIC DENSITY			ΔV TO MAINTAIN ALTITUDE			
Alt (km)	Atm. Scale Ht. (km)	Minimum (kg/m^3)	Mean (kg/m^3)	Maximum (kg/m^3)	Solar Min 50 kg/m^2 (m/s)/yr	Solar Max 50 kg/m^2 (m/s)/yr	Solar Min 200 kg/m^2 (m/s)/yr	Solar Max 200 kg/m^2 (m/s)/yr
0	8.4	1.2	1.2	1.2	2.37×10^{13}	2.37×10^{13}	5.92×10^{12}	5.92×10^{12}
100	5.9	4.61×10^{-7}	4.79×10^{-7}	5.10×10^{-7}	6.95×10^{6}	9.90×10^{6}	2.24×10^{6}	2.47×10^{6}
150	25.5	1.65×10^{-9}	1.81×10^{-9}	2.04×10^{-9}	3.17×10^{4}	3.94×10^{4}	7.93×10^{3}	9.85×10^{3}
200	37.5	1.76×10^{-10}	2.53×10^{-10}	3.52×10^{-10}	3.40×10^{3}	6.72×10^{3}	8.51×10^{2}	1.68×10^{3}
250	44.8	3.35×10^{-11}	6.24×10^{-11}	1.06×10^{-10}	6.36×10^{2}	2.02×10^{3}	1.59×10^{2}	5.04×10^{2}
300	50.3	8.19×10^{-12}	1.95×10^{-11}	3.96×10^{-11}	1.54×10^{2}	7.47×10^{2}	3.86×10^{1}	1.87×10^{2}
350	54.8	2.34×10^{-12}	6.98×10^{-12}	1.66×10^{-11}	4.37×10^{1}	3.11×10^{2}	1.09×10^{1}	7.78×10^{1}
400	58.2	7.32×10^{-13}	2.72×10^{-12}	7.55×10^{-12}	1.36×10^{1}	1.40×10^{2}	3.40×10^{0}	3.50×10^{1}
450	61.3	2.47×10^{-13}	1.13×10^{-12}	3.61×10^{-12}	4.55×10^{0}	6.66×10^{1}	1.14×10^{0}	1.66×10^{1}
500	64.5	8.96×10^{-14}	4.89×10^{-13}	1.80×10^{-12}	1.64×10^{0}	3.29×10^{1}	4.11×10^{-1}	8.23×10^{0}
550	68.7	3.63×10^{-14}	2.21×10^{-13}	9.25×10^{-13}	6.59×10^{-1}	1.68×10^{1}	1.65×10^{-1}	4.20×10^{0}
600	74.8	1.68×10^{-14}	1.04×10^{-13}	4.89×10^{-13}	3.03×10^{-1}	8.81×10^{0}	7.58×10^{-2}	2.20×10^{0}
650	84.4	9.14×10^{-15}	5.15×10^{-14}	2.64×10^{-13}	1.64×10^{-1}	4.73×10^{0}	4.09×10^{-2}	1.18×10^{0}
700	99.3	5.74×10^{-15}	2.72×10^{-14}	1.47×10^{-13}	1.02×10^{-1}	2.61×10^{0}	2.55×10^{-2}	6.52×10^{-1}
750	121	3.99×10^{-15}	1.55×10^{-14}	8.37×10^{-14}	7.04×10^{-2}	1.48×10^{0}	1.76×10^{-2}	3.69×10^{-1}
800	151	2.96×10^{-15}	9.63×10^{-15}	4.39×10^{-14}	5.19×10^{-2}	8.63×10^{-1}	1.30×10^{-2}	2.16×10^{-1}
850	188	2.28×10^{-15}	6.47×10^{-15}	3.00×10^{-14}	3.97×10^{-2}	5.23×10^{-1}	9.94×10^{-3}	1.31×10^{-1}
900	226	1.80×10^{-15}	4.66×10^{-15}	1.91×10^{-14}	3.11×10^{-2}	3.30×10^{-1}	7.78×10^{-3}	8.25×10^{-2}
950	263	1.44×10^{-15}	3.54×10^{-15}	1.27×10^{-14}	2.48×10^{-2}	2.18×10^{-1}	6.19×10^{-3}	5.45×10^{-2}
1,000	296	1.17×10^{-15}	2.79×10^{-15}	8.84×10^{-15}	1.99×10^{-2}	1.51×10^{-1}	4.98×10^{-3}	3.77×10^{-2}
1,250	408	4.67×10^{-16}	1.11×10^{-15}	2.59×10^{-15}	7.69×10^{-3}	4.27×10^{-2}	1.92×10^{-3}	1.07×10^{-2}
1,500	516	2.30×10^{-16}	5.21×10^{-16}	1.22×10^{-15}	3.68×10^{-3}	1.95×10^{-2}	9.20×10^{-4}	4.88×10^{-3}
2,000	829	—	—	—	—	—	—	—
2,500	1,220	—	—	—	—	—	—	—
3,000	1,590	—	—	—	—	—	—	—
3,500	1,900	—	—	—	—	—	—	—
4,000	2,180	—	—	—	—	—	—	—
4,500	2,430	—	—	—	—	—	—	—
5,000	2,690	—	—	—	—	—	—	—
6,000	3,200	—	—	—	—	—	—	—
7,000	3,750	—	—	—	—	—	—	—
8,000	4,340	—	—	—	—	—	—	—
9,000	4,970	—	—	—	—	—	—	—
10,000	5,630	—	—	—	—	—	—	—
15,000	9,600	—	—	—	—	—	—	—
20,000	14,600	—	—	—	—	—	—	—
20,184	14,600	—	—	—	—	—	—	—
25,000	20,700	—	—	—	—	—	—	—
30,000	27,800	—	—	—	—	—	—	—
35,000	36,000	—	—	—	—	—	—	—
35,786	37,300	—	—	—	—	—	—	—

Figure 7.10: Atmospheric density values with scale heights, **??**.

	33	34	35	36	37	38	39	40
	ORBIT DECAY RATE				ESTIMATED ORBIT LIFETIME			
Alt (km)	Solar Min 50 kg/ m² (km/yr)	Solar Max 50 kg/ m² (km/yr)	Solar Min 200 kg/ m² (km/yr)	Solar Max 200 kg/ m² (km/yr)	Solar Min 50 kg/ m² (days)	Solar Max 50 kg/ m² (days)	Solar Min 200 kg/ m² (days)	Solar Max 200 kg/ m² (days)
0	3.82×10^{13}	3.82×10^{13}	9.55×10^{12}	9.55×10^{12}	0.00	0.00	0.00	0.00
100	1.48×10^{7}	1.64×10^{7}	3.70×10^{6}	3.96×10^{6}	0.06	0.06	0.06	0.06
150	5.30×10^{4}	6.58×10^{4}	1.32×10^{4}	1.57×10^{4}	0.24	0.18	0.54	0.48
200	5.75×10^{3}	1.14×10^{4}	1.44×10^{3}	2.67×10^{3}	1.65	1.03	5.99	3.6
250	1.09×10^{3}	3.45×10^{3}	2.72×10^{2}	7.99×10^{2}	10.06	3.82	40.21	14.98
300	2.67×10^{2}	1.29×10^{3}	6.67×10^{1}	2.95×10^{2}	49.9	11.0	196.7	49.2
350	7.64×10^{1}	5.44×10^{2}	1.91×10^{1}	1.23×10^{2}	195.6	30.9	615.9	140.3
400	2.40×10^{1}	2.48×10^{2}	6.01×10^{0}	5.50×10^{1}	552.2	77.4	1024.5	346.9
450	8.12×10^{0}	1.19×10^{2}	2.03×10^{0}	2.60×10^{1}	872	181	1,497	724
500	2.97×10^{0}	5.95×10^{1}	7.42×10^{-1}	1.28×10^{1}	1,205	393	2,377	3,310
550	1.20×10^{0}	3.07×10^{1}	3.01×10^{-1}	6.53×10^{0}	1,638	801	5,470	4,775
600	5.60×10^{-1}	1.63×10^{1}	1.40×10^{-1}	3.41×10^{0}	2,580	3,430	14,100	13,400
650	3.05×10^{-1}	8.83×10^{0}	7.64×10^{-2}	1.83×10^{0}	5,560	4,550	28,500	27,900
700	1.92×10^{-1}	4.92×10^{0}	4.81×10^{-2}	1.00×10^{0}	13,400	12,600	53,400	52,700
750	1.34×10^{-1}	2.82×10^{0}	3.36×10^{-2}	5.67×10^{-1}	24,400	24,300	98,500	97,700
800	1.00×10^{-1}	1.66×10^{0}	2.50×10^{-2}	3.30×10^{-1}	42,000	41,000	175,200	174,200
850	7.74×10^{-2}	1.02×10^{0}	1.93×10^{-2}	1.99×10^{-1}	76,600	76,200	307,400	306,700
900	6.12×10^{-2}	6.49×10^{-1}	1.53×10^{-2}	1.26×10^{-1}	127,000	128,000	521,000	520,000
950	4.92×10^{-2}	4.33×10^{-1}	1.23×10^{-2}	8.26×10^{-2}	21,1000	210,000	853,000	852,000
1,000	4.00×10^{-2}	3.03×10^{-1}	9.99×10^{-3}	5.70×10^{-2}	341,000	340,000	1,361,000	1,362,000
1,250	1.62×10^{-2}	9.00×10^{-2}	4.05×10^{-3}	1.59×10^{-2}	1,700,000	1,700,000	6,800,000	6,800,000
1,500	8.15×10^{-3}	4.32×10^{-3}	2.04×10^{-3}	7.17×10^{-3}	4,810,000	4,810,000	19,250,000	19,250,000
2,000	—	—	—	—	—	—	—	—
2,500	—	—	—	—	—	—	—	—
3,000	—	—	—	—	—	—	—	—
3,500	—	—	—	—	—	—	—	—
4,000	—	—	—	—	—	—	—	—
4,500	—	—	—	—	—	—	—	—
5,000	—	—	—	—	—	—	—	—
6,000	—	—	—	—	—	—	—	—
7,000	—	—	—	—	—	—	—	—
8,000	—	—	—	—	—	—	—	—
9,000	—	—	—	—	—	—	—	—
10,000	—	—	—	—	—	—	—	—
15,000	—	—	—	—	—	—	—	—
20,000	—	—	—	—	—	—	—	—
20,184	—	—	—	—	—	—	—	—
25,000	—	—	—	—	—	—	—	—
30,000	—	—	—	—	—	—	—	—
35,000	—	—	—	—	—	—	—	—
35,786	—	—	—	—	—	—	—	—

Figure 7.11: Overview over orbital lifetimes, **??**.

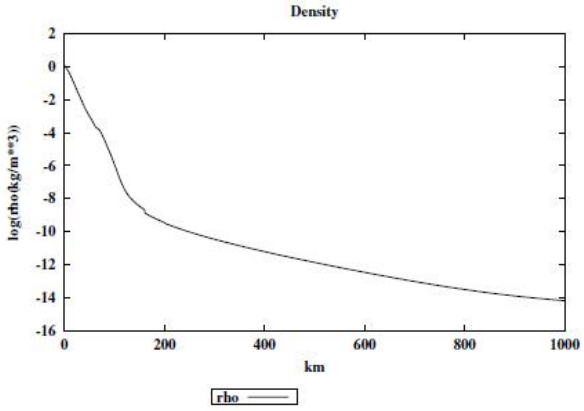


Figure 7.12: MSIS90e example density, courtesy Beutler

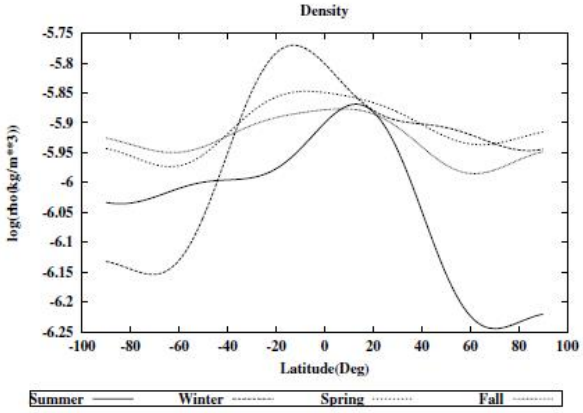


Figure 7.13: Density of the atmosphere in a height of 100km as a function of latitude, courtesy Beutler

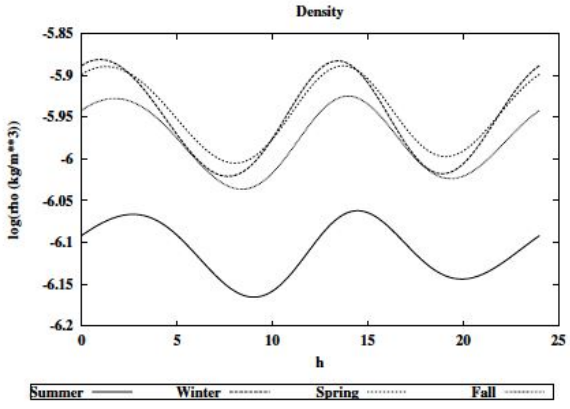


Figure 7.14: Daily variations of the density of the atmosphere in a height if 100km at mid-latitude (northern hemisphere), courtesy Beutler

7.6 Further Perturbations

Further perturbations that are relevant for precise orbit determination/propagation that we do not cover and are of smaller magnitude than the ones we named here are:

- indirect radiation pressure, self-shadowing effects
- Earth shadow passages
- thermal effects (remember the pioneer anomaly)
- Earth albedo radiation pressure
- drag due to electrically neutral atmoshphere
- drag due to charged particles on the satellite's surface
- induced Lorentz forces through charging and movement in the Earth magnetosphere
- ...

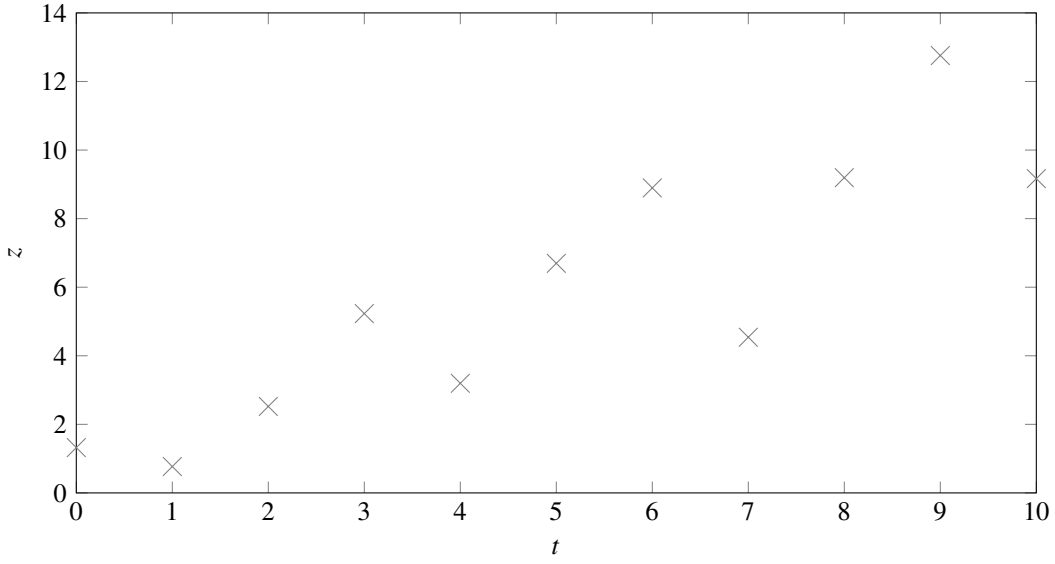
Chapter 8

First Orbit Improvement

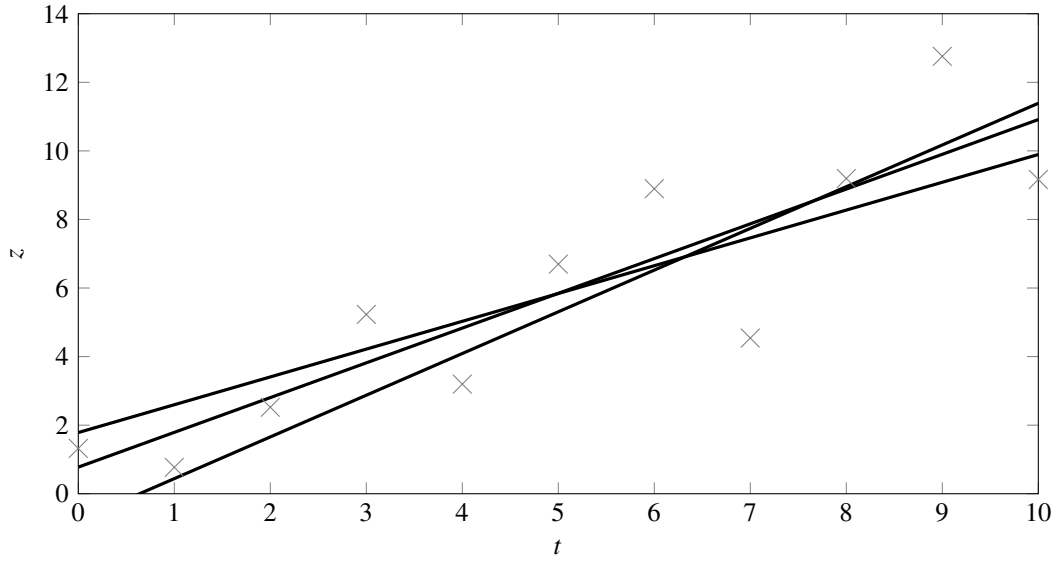
Once we have a first orbit determined (with the classical techniques), the question naturally comes up: What now? We have seen that the first orbit is not perfect, depending on our measurement scenario, we already know that our initial guess might bears large errors. The logical next step is, to seek for new measurements soon and provide an orbit improvement step. We have an initial guess, but no covariance information yet. Hang on a second, covariance? .. ok, ok, before diving into it, maybe a few words and some recap on some probability theory would be in place. As a reference [31, 23, 64] have been used.

8.1 Least Squares Estimation: Introduction Parameter Estimation

Given observations z_i of "something" at "times" t_i for $i \in \{1,2,\ldots,m\}$,....



..... it is desired to fit a model, g, to the data. For instance, if it is desired to fit a line to the data, then the model would be $g = a + bt$, such that the model-predicted observation is $g_i = a + bt_i$.

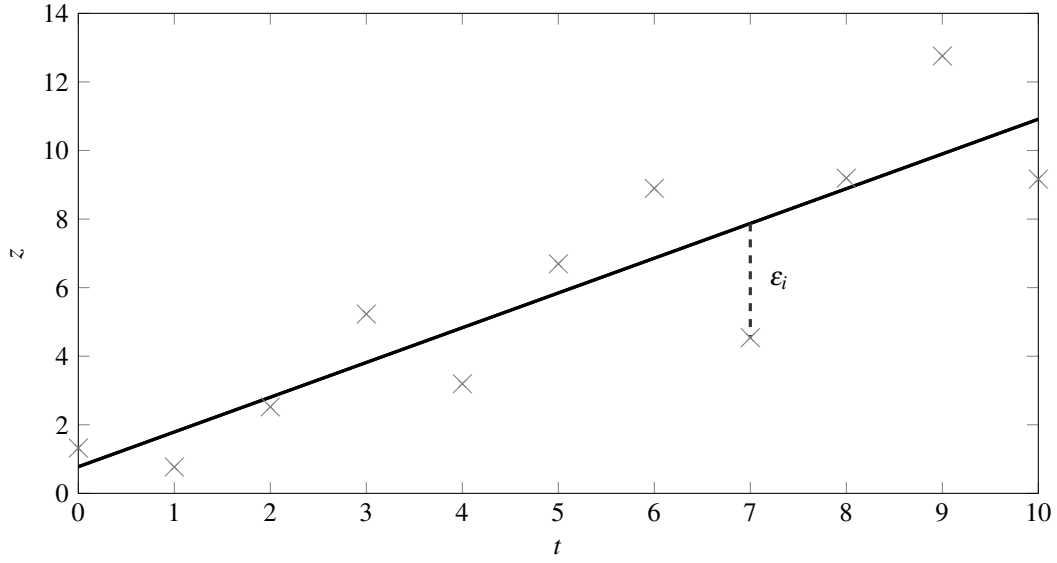


The objective then becomes to determine a method by which the parameters of the model, which are, in the case of fitting a line, the numbers a and b, can be selected. However, we do not want to fit any line, but as such that a performance index is minimized.

What should the performance index be?

The performance index should measure the cumulative error between the observations that were taken (e.g. z_i taken at time t_i) and the values predicted by the model (e.g. $g_i = a + bt_i$).

Define this difference between the actual and predicted observations to be the residual, $\varepsilon_i = z_i - g_i$.



If the performance index is taken to be the sum of the errors, there could be a model such that the i^{th} residual is opposite in sign and equal in magnitude to the j^{th} residual causing no change in the performance index. This does not

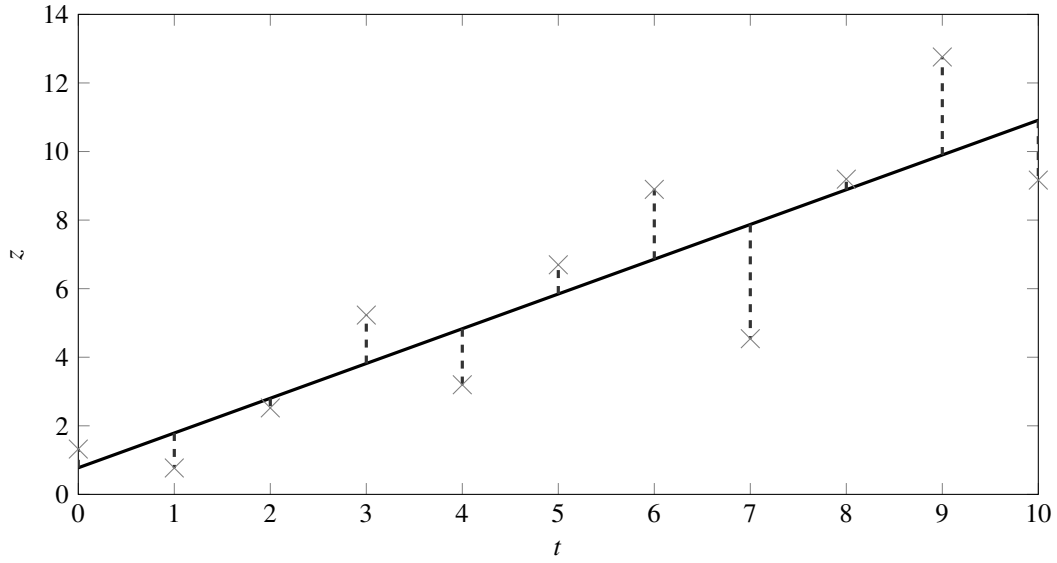
yield a desirable performance index since observations can effectively nullify one another. Instead, let's consider the performance index to be the sum of the squares of the residuals

$$J = \sum_{i=1}^{m} \varepsilon_i^2 \tag{8.1}$$

Note that sometimes a factor of $1/2$ is used to scale the performance index. This is only a matter of convenience, and it will not influence the final result. For m measurements and the line-fitting problem, choosing the performance index in this manner yields

$$J(a,b) = \sum_{i=1}^{m} \varepsilon_i^2 \tag{8.2}$$

$$= \sum_{i=1}^{m} \left[z_i - \left(a + bt_i\right)\right]^2 \tag{8.3}$$



For each choice of (a,b), i.e. for each choice of the model, J can have a different value. We want to choose the model that minimizes the performance index, which yields a parameter optimization problem. The steps are:

- Set the first derivatives equal to zero, and solve for the parameters.
- If the matrix of second derivatives is positive definite at the solution, this is a minimum.

8.2 Linear Least Squares

8.2.1 Original Least Squares

Assume that q observations, $\boldsymbol{z}_i \in \mathbb{R}^p$, are acquired at times t_i. Note that p is the dimension of a single vector observation. The first step is to identify the parameters we wish to estimate. We will call this the "state." As an example, the state of the line-fitting problem is $\boldsymbol{x}^T = [a \;\; b]$. The next step is to express the model as linear combinations of the state, such that

$$\boldsymbol{g}_i = \boldsymbol{H}_i \boldsymbol{x} \tag{8.4}$$

We have extended our model to a vector, or modeled measurements $\boldsymbol{g}_i \in \mathbb{R}^p$.

The state is taken to be n-dimensional, or $\boldsymbol{x} \in \mathbb{R}^n$, which means that the model matrix or measurement matrix, $\boldsymbol{H}_i$, is an

$p \times n$ matrix.
As before, we define the residual to be the difference between the actual true observations $\boldsymbol{z}_i \in \mathbb{R}^p$ and the model-predicted values $\boldsymbol{g}_i$, which gives

$$\boldsymbol{\varepsilon}_i = \boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x} \tag{8.5}$$

The least-squares performance index is now the sum of the squares of the residuals:

$$J = \sum_{i=1}^{q} \boldsymbol{\varepsilon}_i^T \boldsymbol{\varepsilon}_i \tag{8.6}$$

Note that, since we are dealing with vector observations, we cannot simply square the individual residuals, but the inner product produces the equivalent formulation since each scalar element of the residual is squared.
Substituting for the residual into the performance index gives

$$J = \sum_{i=1}^{q} \left[\boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x}\right]^T \left[\boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x}\right] \tag{8.7}$$

In order to proceed, we want to re-express the performance index in a more convenient fashion. To that end, define a concatenated measurement, a concatenated model matrix, and a concatenated residual as

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_q \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \vdots \\ \boldsymbol{H}_q \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_q \end{bmatrix} \tag{8.8}$$

Note that

$$J = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_1^T \boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2^T \boldsymbol{\varepsilon}_2 + \cdots + \boldsymbol{\varepsilon}_q^T \boldsymbol{\varepsilon}_q = \sum_{i=1}^{q} \boldsymbol{\varepsilon}_i^T \boldsymbol{\varepsilon}_i \tag{8.9}$$

That is, using the concatenated residual results in the same performance index with which we began. Define $m = pq$, such that $\boldsymbol{\varepsilon} \in \mathbb{R}^m$, $\boldsymbol{z} \in \mathbb{R}^m$, and $\boldsymbol{H} \in \mathbb{R}^{m \times n}$. Now, substitute for the residual in terms of the observations and the model, to get

$$J = \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right] \tag{8.10}$$

At this point, we are ready to apply our conditions for minimizing the performance index; namely, setting the first derivative of the performance index with respect to the state equal to zero and verifying that the second derivative of the performance index with respect to the state is positive definite. The first derivative is

$$\frac{\partial J}{\partial \boldsymbol{x}} = -2\left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \boldsymbol{H} \tag{8.11}$$

Setting this equal to zero yields

$$\left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \boldsymbol{H} = \boldsymbol{0}^T \quad \rightarrow \quad \boldsymbol{z}^T \boldsymbol{H} - \boldsymbol{x}^T \boldsymbol{H}^T \boldsymbol{H} = \boldsymbol{0}^T \tag{8.12}$$

Transposing the preceding result gives us

$$\boldsymbol{H}^T \boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}^T \boldsymbol{z} \tag{8.13}$$

Solving for the state produces the (potential) least-squares estimate as

$$\hat{\boldsymbol{x}} = \left[\boldsymbol{H}^T \boldsymbol{H}\right]^{-1} \boldsymbol{H}^T \boldsymbol{z} \tag{8.14}$$

This result is the solution to the so-called normal equation:

$$\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{x}} = \boldsymbol{H}^T\boldsymbol{z} \tag{8.15}$$

Whenever we see an inverted matrix, we should always wonder if the inverse exists. We will show that we expect it to exist.

Let's move on to the second derivative. The first thing we do is to transpose the result of the first derivative (before we made any manipulations to solve the equation). This gives us

$$\left[\frac{\partial J}{\partial \boldsymbol{x}}\right]^T = -2\boldsymbol{H}^T\left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right] \tag{8.16}$$

Now, we can take the derivative of this expression with respect to $\boldsymbol{x}$:

$$\frac{\partial}{\partial \boldsymbol{x}}\left[\frac{\partial J}{\partial \boldsymbol{x}}\right]^T = 2\boldsymbol{H}^T\boldsymbol{H} \tag{8.17}$$

It is important to note that this is (except for scaling) precisely the matrix we need to invert in order to find the (potential) least-squares estimate.

If this matrix is positive definite, we know that it can be inverted and we can solve the system. Additionally, we know that the solution does indeed minimize the performance index, so the solution is also the least-squares estimate.

How do we know that the matrix is positive definite?

First, note that the matrix $\boldsymbol{H}$ is an $m \times n$ matrix with $m > n$ (in order to have an over-determined system).

Therefore, $\boldsymbol{H}$ is at most rank n. Provided that we have n linearly independent observations, $\boldsymbol{H}$ will be rank n; that is, it is full (column) rank.

For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ that is rank n ($n < m$), then $\boldsymbol{A}^T\boldsymbol{A}$ is rank n (full rank) and is therefore positive definite.

Applying the preceding result to the matrix $\boldsymbol{H}$ means that as long as we have n linearly independent observations (where n is the number of states we are estimating), a least-squares solution exists.

8.2.2 Example line fitting

Let's take a look at the two-dimensional line-fitting problem. For this problem, we had m observations z_i taken at times t_i, and we wanted to fit the model $g_i = a + bt_i$ to this data in the least-squares sense.

Previously, we came up with a solution through a specialized treatment for the line-fitting problem; now, we want to show that our general solution procedure produces the same solution.

The state that we want to estimate is defined to be $\boldsymbol{x}^T = [a \;\; b]$.

Next, we write the model as a linear combination of the states, which gives

$$g_i = \begin{bmatrix} 1 & t_i \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \boldsymbol{H}_i \boldsymbol{x} \tag{8.18}$$

Note that this is exactly the same as

$$g_i = a + bt_i \tag{8.19}$$

The least-squares solution is

$$\hat{\boldsymbol{x}} = \left[\boldsymbol{H}^T \boldsymbol{H}\right]^{-1} \boldsymbol{H}^T \boldsymbol{z} \tag{8.20}$$

where

$$\boldsymbol{H} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} \tag{8.21}$$

Now, form the products $\boldsymbol{H}^T\boldsymbol{H}$ and $\boldsymbol{H}^T\boldsymbol{z}$:

$$\boldsymbol{H}^T\boldsymbol{H} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_m \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m 1 & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{bmatrix} \tag{8.22}$$

and

$$\boldsymbol{H}^T\boldsymbol{z} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_m \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m z_i \\ \sum_{i=1}^m z_i t_i \end{bmatrix} \tag{8.23}$$

If we define

$$\beta = \sum_{i=1}^m t_i\,, \quad \alpha = \sum_{i=1}^m t_i^2\,, \quad e_1 = \sum_{i=1}^m z_i\,, \quad \text{and} \quad e_2 = \sum_{i=1}^m z_i t_i$$

then, it follows that the least-squares solution may be expressed as

$$\hat{\boldsymbol{x}} = \begin{bmatrix} m & \beta \\ \beta & \alpha \end{bmatrix}^{-1} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{8.24}$$

Now, this time let's use MATLAB to perform line fitting to the data in a least-squares sense.

Since we will be generating our observations randomly, we will start by setting the random number seed.

```
% Set random seed
rng(100)
```

Create a vector of times at which you wish to have observations.

```
% Times for observations
t = (0:1:10)';
```

Determine the number of observations that we will have. There will be one scalar measurement at each time step, so the number of measurements is equal to the number of time steps.

```
% Number of measurements
m = length(t);
```

Set the true parameters of the line; in this case, we will use $a = 1$ and $b = 1$. It is important to note that this information is not known to the least-squares method that we will use soon.

```
% Set true parameters of the line
a = 1;
b = 1;
```

Create a set of nominal measurements from the true line.

```
% Generate data on a line
y  = a + b.*t;
```

Create a set of noisy observations from this line according to a Gaussian distribution of mean 0 and standard deviation 2.

```
% Generate noisy observations
s = 2;
z = y + s.*randn(m,1);
```

Now for the fun part: least-squares! Construct the model matrix $\boldsymbol{H}$ by appending a column vector of ones with the times of the observations.

```
% Form the measurement mapping matrix
H = [ones(m,1), t];
```

Now we build our least-squares estimate, and we're done.

```
% Compute least-squares line fit
x = (H'*H)\(H'*z);
```

Acknowledging that in MATLAB, $\mathrm{x} = [\hat{a} \;\; \hat{b}]^T$, we can obtain our estimates for the linear fit as

$$\hat{a} = 0.7761 \qquad \hat{b} = 1.0137, \tag{8.25}$$

8.2.3 Example: Polynomial Fitting

We can now very easily extend the least-squares solution to more than simple linear fits.

Consider the polynomial model

$$g_i = a_0 + a_1 t_i + \frac{1}{2} a_2 t_i^2 + \frac{1}{6} a_3 t_i^3 + \cdots + \frac{1}{k!} a_k t_i^k \tag{8.26}$$

where the scaled a_i's are our states to be determined.

Define the state to be

$$\boldsymbol{x}^T = \begin{bmatrix} a_0 & a_1 & \frac{1}{2}a_2 & \frac{1}{6}a_3 & \cdots & \frac{1}{k!}a_k \end{bmatrix} \tag{8.27}$$

The model mapping, or measurement matrix is then given by

$$\boldsymbol{H}_i = \begin{bmatrix} 1 & t_i & t_i^2 & t_i^3 & \cdots & t_i^k \end{bmatrix} \tag{8.28}$$

which gives the concatenated mapping matrix as the Vandermonde matrix

$$\boldsymbol{H} = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & t_2^3 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & t_m^3 & \cdots & t_m^k \end{bmatrix} \tag{8.29}$$

From here, we can simply apply the least-squares solution to estimate the model parameters:

$$\hat{\boldsymbol{x}} = \left[\boldsymbol{H}^T\boldsymbol{H}\right]^{-1}\boldsymbol{H}^T\boldsymbol{z} \tag{8.30}$$

Note that, to retain an over-determined system, we must ensure that the number of parameters in the model is less than the number of measurements.

This is precisely what the MATLAB routine `polyfit` does.

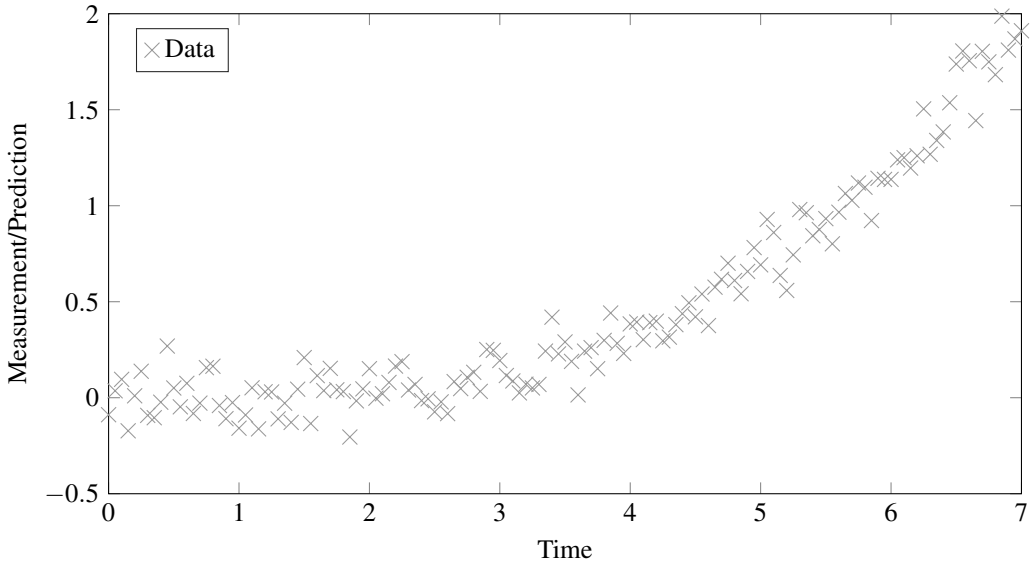
Let's look at an example:

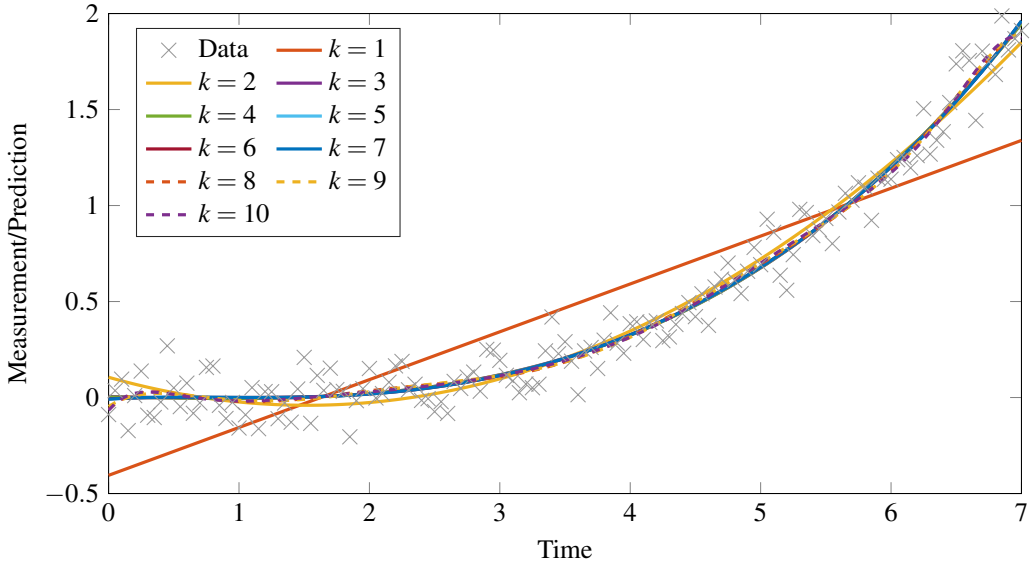
- The true data are generated from a degree four polynomial.
- Measurement errors are added to the data.
- We use the least-squares polynomial fitting method to fit degree k polynomials.
- The quality of the fit is analyzed by computing the post-fit residuals as

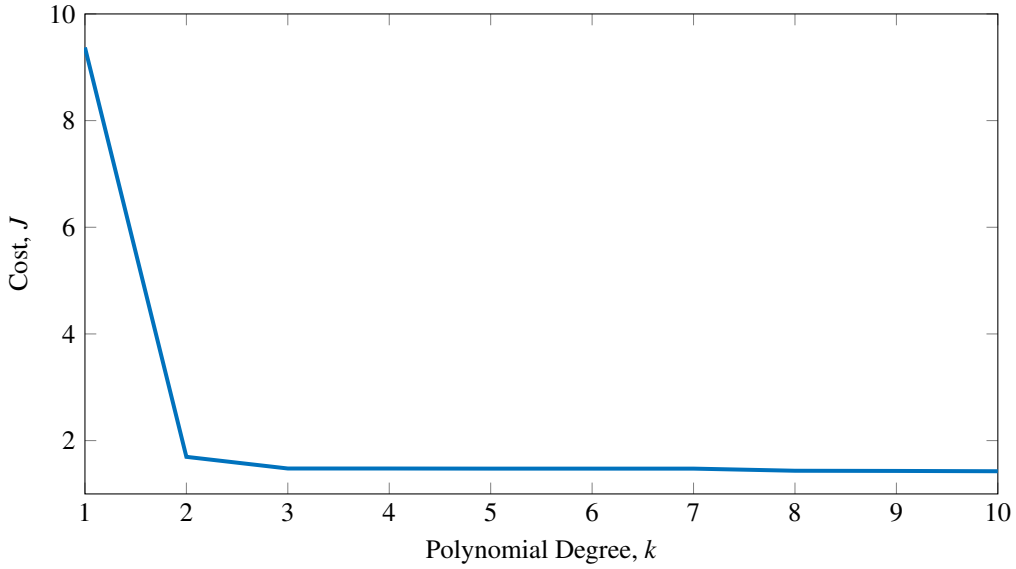
$$\hat{\boldsymbol{\varepsilon}}_i = \boldsymbol{z}_i - \boldsymbol{H}_i\hat{\boldsymbol{x}} \tag{8.31}$$

  and then finding the cost via

$$J = \sum_{i=1}^{m} \hat{\boldsymbol{\varepsilon}}_i^T\hat{\boldsymbol{\varepsilon}}_i \tag{8.32}$$


Data
Measurement/Prediction
Time
2
1.5
1
0.5
0
−0.5
0
1
2
3
4
5
6
7


Data
$k = 1$
$k = 2$
$k = 3$
$k = 4$
$k = 5$
$k = 6$
$k = 7$
$k = 8$
$k = 9$
$k = 10$
Measurement/Prediction
Time
2
1.5
1
0.5
0
−0.5
0
1
2
3
4
5
6
7



8.2.4 Example: Dynamical System

What if our state is not a static set of parameters? What if our state obeys some dynamical system? In this case, let's assume that our state evolves in continuous time as

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t)\,, \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{8.33}$$

To give an example, consider a robot moving in the horizontal plane at constant velocity. If the position is described by x and y and the velocity is described by u and v, then it follows that the dynamics of the robot are

$$\dot{x} = u \tag{8.34}$$
$$\dot{y} = v \tag{8.35}$$
$$\dot{u} = 0 \tag{8.36}$$
$$\dot{v} = 0 \tag{8.37}$$

By defining the state vector to be

$$\boldsymbol{x}(t) = \begin{bmatrix} x \\ y \\ u \\ v \end{bmatrix} \tag{8.38}$$

we can express the set of dynamics for the robot as

$$\dot{\boldsymbol{x}}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \boldsymbol{x}(t) \tag{8.39}$$

This is equivalent to the model $\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t)$ with

$$\boldsymbol{F}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{8.40}$$

In conjunction with the dynamical system, we have vector observations $\boldsymbol{z}_k$ of the state at times t_k, and an appropriate model of the observations given by

$$\boldsymbol{g}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k \quad \text{where} \quad \boldsymbol{x}_k = \boldsymbol{x}(t_k) \tag{8.41}$$

For the robot case, let's assume that we can observe the position of the robot.

This gives us

$$\boldsymbol{g}_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \boldsymbol{x}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k \tag{8.42}$$

How do we handle this type of problem using least squares?

Least squares provides us with the machinery to estimate a static state, or a static set of parameters.

If we can relate the state at any arbitrary time back to the epoch state $\boldsymbol{x}_0$, then we can apply the least squares method to estimate the epoch state.

The solution of the continuous time dynamical system is given by

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}(t, t_0)\boldsymbol{x}_0 \tag{8.43}$$

$\boldsymbol{\Phi}(t, t_0)$ is the state transition matrix, which satisfies the properties:

1. $\boldsymbol{\Phi}(t_i, t_i) = \boldsymbol{I}$

2. $\boldsymbol{\Phi}(t_i, t_\ell) = \boldsymbol{\Phi}(t_i, t_j)\boldsymbol{\Phi}(t_j, t_\ell)$

3. $\boldsymbol{\Phi}(t_i, t_j) = \boldsymbol{\Phi}^{-1}(t_j, t_i)$

4. $\dot{\boldsymbol{\Phi}}(t,t_i) = \boldsymbol{F}(t)\boldsymbol{\Phi}(t,t_i)\,, \quad \boldsymbol{\Phi}(t_i,t_i) = \boldsymbol{I}$

There is an easy representation for autonomous systems where $\boldsymbol{F}(t) = \boldsymbol{F}$.

We can expand the state of the system in a Taylor series as

$$\boldsymbol{x}(t) = \boldsymbol{x}(t_0) + \dot{\boldsymbol{x}}(t_0)(t-t_0) + \frac{1}{2}\ddot{\boldsymbol{x}}(t_0)(t-t_0)^2 + \cdots \tag{8.44}$$

From our system dynamics model, it follows that

$$\dot{\boldsymbol{x}}(t_0) = \boldsymbol{F}\boldsymbol{x}(t_0) \tag{8.45}$$

Similarly,

$$\ddot{\boldsymbol{x}}(t_0) = \boldsymbol{F}\dot{\boldsymbol{x}}(t_0) = \boldsymbol{F}\boldsymbol{F}\boldsymbol{x}(t_0) = \boldsymbol{F}^2\boldsymbol{x}(t_0) \tag{8.46}$$

Then, from the Taylor series, we see that

$$\begin{aligned}
\boldsymbol{x}(t) &= \boldsymbol{x}(t_0) + \dot{\boldsymbol{x}}(t_0)(t-t_0) + \frac{1}{2}\ddot{\boldsymbol{x}}(t_0)(t-t_0)^2 + \cdots && (8.47)\\
&= \boldsymbol{x}(t_0) + \boldsymbol{F}\boldsymbol{x}(t_0)(t-t_0) + \frac{1}{2}\boldsymbol{F}^2\boldsymbol{x}(t_0)(t-t_0)^2 + \cdots && (8.48)\\
&= \left[\boldsymbol{I} + \boldsymbol{F}(t-t_0) + \frac{1}{2}\boldsymbol{F}^2(t-t_0)^2 + \cdots\right]\boldsymbol{x}(t_0) && (8.49)
\end{aligned}$$

The matrix in brackets is identically the matrix exponential, such that

$$\boldsymbol{x}(t) = \mathrm{e}^{\boldsymbol{F}\cdot(t-t_0)}\boldsymbol{x}(t_0) \tag{8.50}$$

where

$$\mathrm{e}^{\boldsymbol{F}\cdot(t-t_0)} = \left[\boldsymbol{I} + \boldsymbol{F}(t-t_0) + \frac{1}{2}\boldsymbol{F}^2(t-t_0)^2 + \cdots\right] = \sum_{k=0}^{\infty}\frac{1}{k!}\boldsymbol{F}^k(t-t_0)^k \tag{8.51}$$

This is only true for stationary dynamics. Implemented as `expm` in MATLAB.

What happens when we apply the matrix exponential to the robot's dynamics?

Recall that

$$\boldsymbol{F} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{8.52}$$

Computing $\texttt{expm}(\boldsymbol{F} \cdot (t - t_0))$ yields

$$\boldsymbol{\Phi}(t, t_0) = \begin{bmatrix} 1 & 0 & t - t_0 & 0 \\ 0 & 1 & 0 & t - t_0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.53}$$

We can also arrive at this solution by noting that $\boldsymbol{F}^2 = \boldsymbol{0}$, such that only the first two terms of the infinite series for the matrix exponential are required.

From the solution of the linear system, we can write the state at time t_k as

$$\boldsymbol{x}_k = \boldsymbol{\Phi}(t_k, t_0)\boldsymbol{x}_0 \tag{8.54}$$

Therefore, we can express the model in terms of the epoch state, which gives

$$\begin{aligned} \boldsymbol{g}_k &= \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k && (8.55) \\ &= \widetilde{\boldsymbol{H}}_k \boldsymbol{\Phi}(t_k, t_0)\boldsymbol{x}_0 && (8.56) \\ &= \boldsymbol{H}_k \boldsymbol{x}_0 \qquad \text{where} \qquad \boldsymbol{H}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{\Phi}(t_k, t_0) && (8.57) \end{aligned}$$

We're done! We have expressed the individual models as functions of a single epoch state, so we can directly apply the work we've done to develop the least-squares solution.

To summarize:

1. We have a dynamical system with some known $\boldsymbol{F}(t)$

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) \tag{8.58}$$

2. Accompanying the dynamical system are observations $\boldsymbol{z}_k$ at times t_k

3. Our model of the observations at time t_k is

$$\boldsymbol{g}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k \tag{8.59}$$

4. For each observation, we integrate the state transition matrix to the time t_k

$$\boldsymbol{\Phi}(t_k, t_0) = \int_{t_0}^{t_k} \boldsymbol{F}(\tau)\boldsymbol{\Phi}(\tau, t_0) d\tau, \qquad \boldsymbol{\Phi}(t_0, t_0) = \boldsymbol{I} \tag{8.60}$$

(note that in some cases we have an explicit representation of the state transition matrix, but we can also use numerical integration methods in the case that we don't have an explicit representation)

5. Next, we determine the mapped observation model matrix, such that

$$\boldsymbol{H}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{\Phi}(t_k, t_0) \tag{8.61}$$

6. We assemble the concatenated mapped observation model matrices and the concatenated observations

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{H}}_1 \boldsymbol{\Phi}(t_1, t_0) \\ \widetilde{\boldsymbol{H}}_2 \boldsymbol{\Phi}(t_2, t_0) \\ \vdots \end{bmatrix} \quad \text{and} \quad \boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \end{bmatrix} \tag{8.62}$$

7. Finally, we compute the least-squares estimate of the epoch state

$$\hat{\boldsymbol{x}}_0 = \left[\boldsymbol{H}^T \boldsymbol{H}\right]^{-1} \boldsymbol{H}^T \boldsymbol{z} \tag{8.63}$$

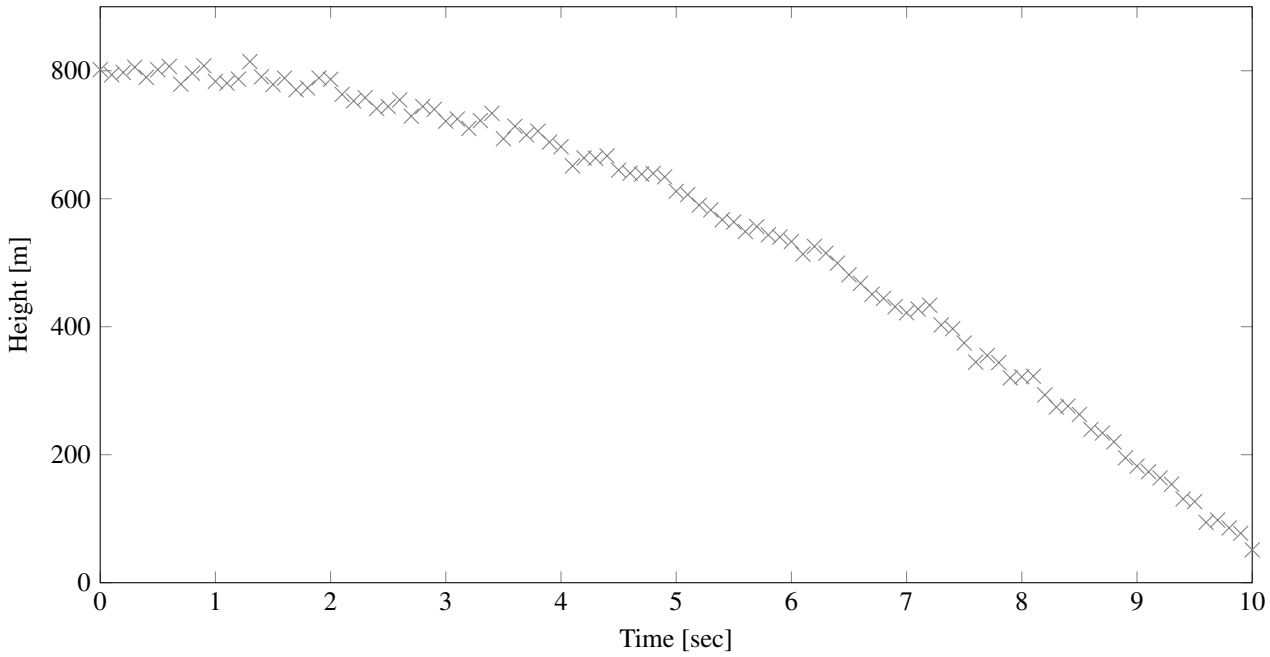
8.2.4.1 Example: Falling Body Problem

Imagine that we are conducting an experiment on a planet with an unknown constant gravitational acceleration g.

We will drop a ball from the top of a building at an unknown altitude. Additionally, the ball is "dropped" with an unknown initial velocity.

Our objective is to determine the height and velocity at which the ball is released. Additionally, we will determine the gravitational acceleration on this planet.

To produce these estimates, we will employ the least-squares approach making use of observations of the ball's height at a rate of 10Hz. These measurements can be seen in the following figure



Using these measurements, we seek estimates of the local gravitational acceleration $\hat{g}$, the height of the building $\hat{h}_0$, and the velocity with which we released the ball $\hat{\dot{h}}_0$.

From Newtonian mechanics, we know that the height of the ball obeys the second-order differential equation

$$\ddot{h}(t) = g \tag{8.64}$$

We will have measurements through time, and we want an estimate for the initial state, right? We can do this with least-squares for dynamic systems!

Define a state corresponding to height and its rates of change

$$\boldsymbol{x}_k = [h_k \quad \dot{h}_k \quad \ddot{h}_k]^T \tag{8.65}$$

We know that, acknowledging that the gravitational acceleration is constant, the second-order dynamical system can be expressed in first-order form as

$$\frac{dh_k}{dt} = \dot{h}_k \qquad \frac{d\dot{h}_k}{dt} = \ddot{h}_k \qquad \frac{d\ddot{h}_k}{dt} = 0 \tag{8.66}$$

In matrix-vector form, this gives us the model for the dynamical system as

$$\frac{d}{dt}\begin{bmatrix} h_k \\ \dot{h}_k \\ \ddot{h}_k \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} h_k \\ \dot{h}_k \\ \ddot{h}_k \end{bmatrix} \tag{8.67}$$

This means that we can write the Jacobian of the system dynamics as

$$\boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{8.68}$$

We also need to establish our measurement model. Since we are taking measurements of the height of the ball, we have

$$\boldsymbol{z}_k = h_k \tag{8.69}$$

This can then be written as a linear combination of our states as

$$\boldsymbol{g}_k = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} h_k \\ \dot{h}_k \\ \ddot{h}_k \end{bmatrix} \tag{8.70}$$

Therefore, our model matrix is

$$\widetilde{\boldsymbol{H}}_k = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \tag{8.71}$$

We then use the fact that by definition

$$\boldsymbol{\Phi}(t_k, t_0) = \mathrm{e}^{\boldsymbol{F}\cdot(t_k - t_0)} \tag{8.72}$$

to construct the concatenated observation model and observations to perform the least-squares fit

This procedure results in estimates for the initial gravitational acceleration and height as

$$\hat{g} = -14.9269 \ [\mathrm{m/s^2}] \qquad \hat{h}_0 = 800.4682 \ [\mathrm{m}] \tag{8.73}$$

corresponding to true values of

$$g = -15 \text{ [m/s}^2\text{]} \qquad h_0 = 800 \text{ [m]} \tag{8.74}$$

Let's look at how we would implement this in MATLAB.

Since we will be generating measurements based on a random generator, let's set the random number generation seed so we can exactly replicate these results.

```
% Set random seed
rng(100)
```

Now, we have to generate our observations. First, define a sequence of times at which observations will occur and the *true* gravitational acceleration.

```
% Times of observations
t = 0:0.1:10; % We'll generate data every 0.1 [sec] for 10 [sec]
g = -15;      % True gravitational acceleration on said planet [m/s^2]
```

Define the true motion of the ball so we can generate noisy observations of its motion. We know that its motion (under constant acceleration) obeys $x(t) = x_0 + v_0 t + \frac{1}{2} g t^2$.

```
% True motion of the ball
x0 = 800;     % Let's say the building is 800 [m] tall
v0 = 0;       % We drop the ball with no initial velocity
a  = g;       % The acceleration is only due to gravity
x  = x0 + v0.*t + 0.5.*a.*t.^2;
```

Generate measurements of the altitude of the ball by adding noise to the true altitude of the ball.

```
% Generate observations
s = 10;                     % Standard deviation of our measurements [m]
z = x + s.*randn(size(x)); % Generate corrupted height measurements [m]
z = z';
```

In order to construct a least-squares estimate, we need the concatenated observational model matrix. We have to assemble this in a loop.

```
% Dynamic state estimation with least squares
F  = [0, 1, 0; 0, 0, 1; 0, 0, 0];
Ht = [1, 0, 0];
H  = [];
for i = 1:length(z)
    zk  = z(i);
    Phi = expm(F*(t(i) - t(1)));
    Hk  = Ht*Phi;
    H   = [H; Hk];
end
```

Finally, we perform the least-squares fit.

```
% Least-squares estimate
xh0 = (H'*H)\(H'*z);
```

This gives us our estimate.

8.2.4.2 Example: Coding a Robot Estimation Problem

Let's code up a least-squares solver for estimating the position and velocity of a robot.

We've already discussed most of the elements that we need to solve this problem.

We'll walk through the generation of each of the parts of a MATLAB code to generate the true path of a robot, the measurements of the position of the robot, and finally the construction of a least-squares estimate for the initial position and velocity of the robot.

In this case, we are assuming that we have a robot operating in a planar environment under a constant acceleration model. We are also assuming that we can acquire observations of the position of the robot, say by mounting a camera above the environment and taking images.

Since we would like to be able to replicate any experiments that we do, we start off our code by setting the seed on a random number generator.

```
% Set random seed
rng(150)
```

For this problem, we will assume that the position is described by x and y and that the velocity is described by u and v. We will combine these four quantities into a vector to form our state; that is

$$\boldsymbol{x}(t) = \begin{bmatrix} x(t) & y(t) & u(t) & v(t) \end{bmatrix}^T \tag{8.75}$$

where we note that $x(t)$ is one element of the state and $\boldsymbol{x}(t)$ is the full state. To simulate the true path that the robot takes, we need to specify the true initial position and velocity of the robot.

```
% Specify the initial true state of the robot
x0 = [0; 0; 0.1; 0.2];
```

We will also need to fix a range of times for which we will simulate the true path of the robot. This will also give us the ability to simulate our measurements of the position of the robot.

```
% Set up our timing variables
%  - assume measurements every 0.1 [sec]
%  - assume the measurements continue for 10 [sec]
t0 =  0.0;
dt =  0.1;
tf = 10.0;
tv = (t0:dt:tf)';
```

The last piece that we need to simulate the true path of the vehicle is the dynamical system description. Previously, we noted that the constant velocity model gives us the dynamics of the states as

$$\dot{x}(t) = u(t) \qquad \dot{y}(t) = v(t) \qquad \dot{u}(t) = 0 \qquad \text{and} \qquad \dot{v}(t) = 0 \tag{8.76}$$

From the definition of our state vector, these dynamics give us the matrix-vector system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) \qquad \text{where} \qquad \boldsymbol{F}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{8.77}$$

We can now specify the dynamics in our code.

```
% Dynamics of the robot (continuous time)
F = [0, 0, 1, 0; 0, 0, 0, 1; 0, 0, 0, 0; 0, 0, 0, 0];
```

Now, we are ready to simulate the true path of our robot. We will do this using numerical integration to solve the initial value problem

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) \quad \text{s.t. } \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{8.78}$$

In MATLAB, the easiest (and most general) way is to use `ode45` to numerically integrate the state forward in time.

```
% Integrate the state for the true object
opts  = odeset('AbsTol',1e-9,'RelTol',1e-9);
[~,X] = ode45(@eom_robot,tv,x0,opts,F);
```

Now that we have the true state as a function of time, we can generate measurements of the position. Since we are

only considering measurements of the position, we will take the first two states at each time and we will add on some measurement noise to simulate the effects of the observation process. Here, we have assumed that the camera system can determine the position to within about 0.1 meters of the actual position.

```
% Create measurements of position
%  - add noise with a standard deviation of 0.1 [m]
z = X(:,1:2)' + 0.1*randn(2,length(tv));
```

Up to this point, we've just been constructing the information that we need to synthesize measurements. We have not done anything related to determining a least-squares estimate. We will now move to that process.

We need two more pieces of information: the state transition matrix and the measurement mapping matrix.

For this problem, we've already determined these two elements:

$$\boldsymbol{\Phi}(t_k,t_0)=\begin{bmatrix}1&0&t_k-t_0&0\\0&1&0&t_k-t_0\\0&0&1&0\\0&0&0&1\end{bmatrix}\qquad\text{and}\qquad\widetilde{\boldsymbol{H}}_k=\begin{bmatrix}1&0&0&0\\0&1&0&0\end{bmatrix}\tag{8.79}$$

The state transition matrix can be found via the matrix exponential, and the measurement mapping matrix can be found by recalling that our measurements are of the position of the robot.

Before computing the least-squares estimate, we need to accumulate all of the observations and all of the measurement mapping matrices (accounting for the state transition matrix). These are given by

$$\boldsymbol{z}=\begin{bmatrix}\boldsymbol{z}_1\\\boldsymbol{z}_2\\\vdots\end{bmatrix}\qquad\text{and}\qquad\boldsymbol{H}=\begin{bmatrix}\widetilde{\boldsymbol{H}}_1\boldsymbol{\Phi}(t_1,t_0)\\\widetilde{\boldsymbol{H}}_2\boldsymbol{\Phi}(t_2,t_0)\\\vdots\end{bmatrix}\tag{8.80}$$

We will do this inside of a loop.

1. Add in the code to accumulate the observations.
2. Add in the code to compute $\widetilde{\boldsymbol{H}}_k$.
3. Add in the code to compute $\boldsymbol{\Phi}(t_k,t_0)$.
4. Add in the code to accumulate the measurement mapping matrices.

The following is what the template code should look like where you'll be adding in the preceding elements.

```
% Assemble the concatenated measurement vector and model matrix
Z = [];
H = [];
for k = 1:length(tv)
    % time at kth observation
    tk = tv(k);

    % concatenated measurements
    Z =

    % concatenated model matrix
    Htilde =
    Phik0  =
    H      =
end
```

Finally, we compute the least-squares estimate of the epoch state using

$$\hat{\boldsymbol{x}}_0 = \left[\boldsymbol{H}^T\boldsymbol{H}\right]^{-1}\boldsymbol{H}^T\boldsymbol{z} \tag{8.81}$$

Add in the code to compute the estimated initial position and velocity.

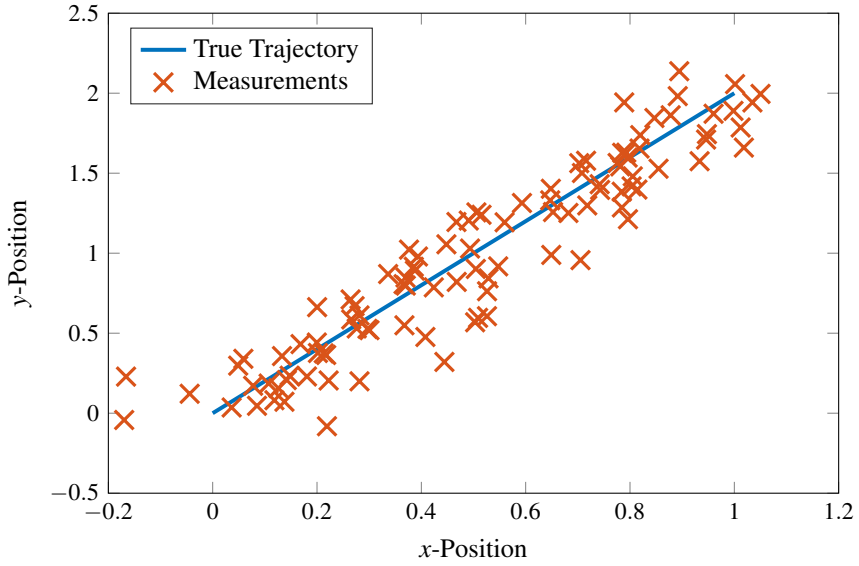
```
% Least-squares estimate
xhat0 =
```

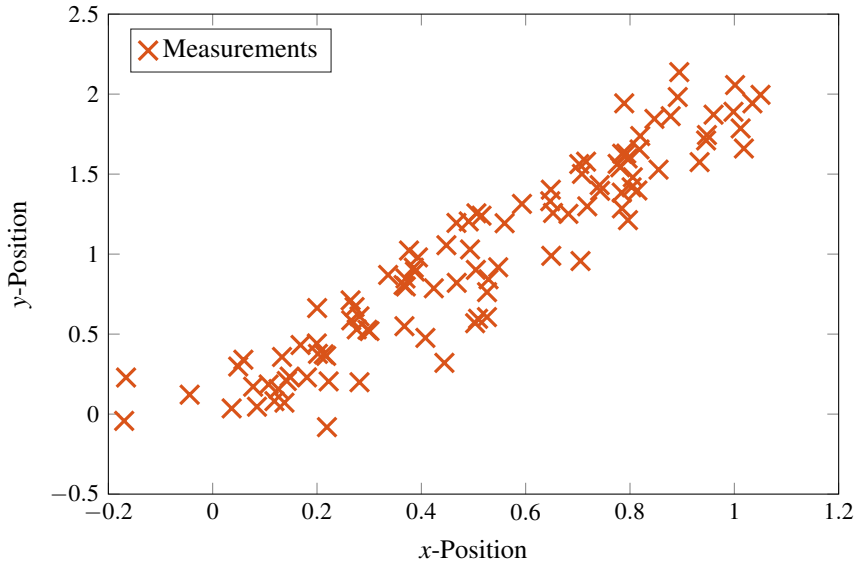
You should find that the estimated state is

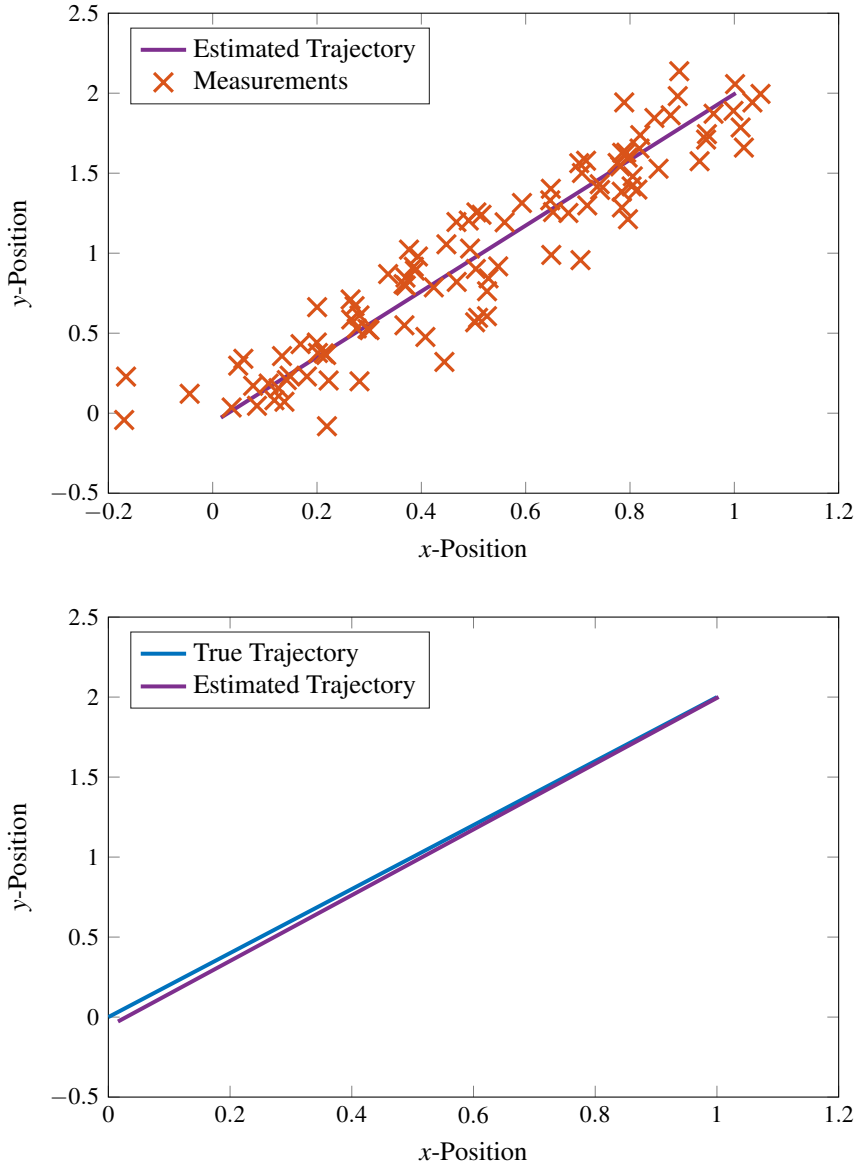
$$\hat{\boldsymbol{x}}_0 = \begin{bmatrix} 0.0158 \\ -0.0277 \\ 0.0987 \\ 0.2028 \end{bmatrix} \tag{8.82}$$

Comparing this to our initial true state, we see that we have a pretty good estimate of the initial position and velocity of our robot!

Let's look at a few plots to see what we got from the least squares process.


True Trajectory
Measurements
y-Position
x-Position
2.5
2
1.5
1
0.5
0
−0.5
−0.2
0
0.2
0.4
0.6
0.8
1
1.2


Measurements
y-Position
x-Position
2.5
2
1.5
1
0.5
0
−0.5
−0.2
0
0.2
0.4
0.6
0.8
1
1.2



8.2.5 Weighted Least Squares

This section may be omitted in a first reading; it will not be covered in class

A shortcoming of the least-squares method is that it does not provide a mechanism for weighting certain observations more heavily (or less heavily) than others.

In cases where we may have more confidence in some measurements, due to their inherent accuracy, we would like to be able to place more emphasis on the information provided by these measurements.

Consider the case where we have measurements $\boldsymbol{z}_i$ at times t_i and where our predicted observations are modeled by

$$\boldsymbol{g}_i = \boldsymbol{H}_i \boldsymbol{x} \tag{8.83}$$

Note that if we are estimating the state of a dynamic system that we would write our predicted observations as

$$\boldsymbol{g}_i = \widetilde{\boldsymbol{H}}_i \boldsymbol{\Phi}(t_i, t_0)\boldsymbol{x}_0 \tag{8.84}$$

We will use the first formulation of the predicted observations for ease of notation with the understanding that the second formulation is equivalent.

The i^{th} residual is given by

$$\boldsymbol{\varepsilon}_i = \boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x} \tag{8.85}$$

In the standard least squares formulation, each of the residuals is given the same weight and the least-squares performance index for q measurements is taken to be

$$J = \sum_{i=1}^{q} \boldsymbol{\varepsilon}_i^T \boldsymbol{\varepsilon}_i \tag{8.86}$$

Now, we wish to extend this so that we can weight individual residuals (or individual measurements) differently.

To accomplish this, let each residual be accompanied by a weight $w_i > 0$. In the vector-measurement case, this weight will be a weight matrix $\boldsymbol{W}_i = \boldsymbol{W}_i^T > \boldsymbol{0}$.

This gives us a set of q residuals with associated weights as

$$\boldsymbol{\varepsilon}_i = \boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x} \quad \text{with weight} \quad \boldsymbol{W}_i \tag{8.87}$$

The performance index is then modified to be

$$J = \sum_{i=1}^{q} \boldsymbol{\varepsilon}_i^T \boldsymbol{W}_i \boldsymbol{\varepsilon}_i \tag{8.88}$$

$$= \sum_{i=1}^{q} \left[\boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x}\right]^T \boldsymbol{W}_i \left[\boldsymbol{z}_i - \boldsymbol{H}_i \boldsymbol{x}\right] \tag{8.89}$$

As before, we will define concatenated terms, but now we also have a concatenated weight, such that

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 \\ \boldsymbol{H}_2 \\ \vdots \\ \boldsymbol{H}_q \end{bmatrix}, \quad \boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_q \end{bmatrix}, \quad \text{and} \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & & & \\ & \boldsymbol{W}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{W}_q \end{bmatrix} \tag{8.90}$$

With these definitions, the weighted least-squares performance index may be expressed as

$$J = \big[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\big]^T \boldsymbol{W} \big[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\big] \tag{8.91}$$

which is completely equivalent to the summation representation.

The first derivative condition for an optimal is

$$\frac{\partial J}{\partial \boldsymbol{x}} = \boldsymbol{0}^T \tag{8.92}$$

where

$$\frac{\partial J}{\partial \boldsymbol{x}} = -2\big[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\big]^T \boldsymbol{W}\boldsymbol{H} \tag{8.93}$$

Applying the first derivative condition to solve for $\hat{\boldsymbol{x}}$ yields

$$\boldsymbol{0}^T = -2\big[\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}\big]^T \boldsymbol{W}\boldsymbol{H} \tag{8.94}$$

Manipulation of the preceding equation gives

$$\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H}\hat{\boldsymbol{x}} = \boldsymbol{H}^T\boldsymbol{W}\boldsymbol{z} \tag{8.95}$$

This is the normal equation for the weighted least-squares problem, which has the solution

$$\hat{\boldsymbol{x}} = \big[\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H}\big]^{-1}\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{z} \tag{8.96}$$

provided that the inverse exists.

Note that if the measurements (residuals) are all equally weighted, then $\boldsymbol{W} = w\boldsymbol{I}$, and the solution becomes

$$\begin{aligned} \hat{\boldsymbol{x}} &= \big[\boldsymbol{H}^T w\boldsymbol{I}\boldsymbol{H}\big]^{-1}\boldsymbol{H}^T w\boldsymbol{I}\boldsymbol{z} && (8.97)\\ &= w\big[w\boldsymbol{H}^T\boldsymbol{H}\big]^{-1}\boldsymbol{H}^T\boldsymbol{z} && (8.98)\\ &= \big[\boldsymbol{H}^T\boldsymbol{H}\big]^{-1}\boldsymbol{H}^T\boldsymbol{z} && (8.99)\\ & && (8.100) \end{aligned}$$

That is, we recover the original least-squares solution.

Is our solution a minimum? We need the second derivative to figure this out.

Recall that the first derivative (before any manipulations) is

$$\frac{\partial J}{\partial \boldsymbol{x}} = -2\big[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\big]^T \boldsymbol{W}\boldsymbol{H} \tag{8.101}$$

such that

$$\frac{\partial}{\partial \boldsymbol{x}}\left[\frac{\partial J}{\partial \boldsymbol{x}}\right]^T = \frac{\partial}{\partial \boldsymbol{x}}\Big\{-2\boldsymbol{H}^T\boldsymbol{W}\big[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\big]\Big\} \tag{8.102}$$

$$= 2\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H} \tag{8.103}$$

Is the second derivative condition satisfied? Is this matrix positive definite? That is

$$2\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H} \overset{?}{>} \boldsymbol{0} \tag{8.104}$$

Note that we can factor $\boldsymbol{W} = \boldsymbol{V}^T\boldsymbol{V}$, such that

$$\frac{\partial}{\partial \boldsymbol{x}}\left[\frac{\partial J}{\partial \boldsymbol{x}}\right]^T = 2\big[\boldsymbol{V}\boldsymbol{H}\big]^T\big[\boldsymbol{V}\boldsymbol{H}\big] \tag{8.105}$$

The presence of $\boldsymbol{V}$ will not alter the rank of the matrix in brackets, so we can use the previous arguments to conclude that if rank $\boldsymbol{H} = n$, then rank $\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H} = n$, so the second derivative condition is satisfied.

Since we can weight each observation differently, one might ask if we can also add some prior information regarding the state of the system.

That is to say, if we have some information about the state, call it $\bar{\boldsymbol{x}}$, can we include this in our weighted least-squares approach? In addition, can we include it with some weight $\overline{\boldsymbol{W}}$?

We still have the sequence of observations $\boldsymbol{z}_i$ at times t_i that are modeled as $\boldsymbol{g}_i = \boldsymbol{H}_i\boldsymbol{x}$.

Lets also add a zeroth observation of the state, or

$$\boldsymbol{z}_0 = \bar{\boldsymbol{x}} \tag{8.106}$$

and model this observation as

$$\boldsymbol{g}_0 = \boldsymbol{x} \qquad \text{i.e. } \boldsymbol{H}_0 = \boldsymbol{I} \tag{8.107}$$

and assign it a weight of $\boldsymbol{W}_0 = \overline{\boldsymbol{W}}$.

Then, our weighted least-squares performance index is

$$J = \sum_{i=0}^{q} \left[\boldsymbol{z}_i - \boldsymbol{H}_i\boldsymbol{x}\right]^T \boldsymbol{W}_i \left[\boldsymbol{z}_i - \boldsymbol{H}_i\boldsymbol{x}\right] \tag{8.108}$$

Extracting out the zeroth observation yields

$$J = \left[\boldsymbol{z}_0 - \boldsymbol{H}_0\boldsymbol{x}\right]^T \boldsymbol{W}_0 \left[\boldsymbol{z}_0 - \boldsymbol{H}_0\boldsymbol{x}\right] + \sum_{i=1}^{q} \left[\boldsymbol{z}_i - \boldsymbol{H}_i\boldsymbol{x}\right]^T \boldsymbol{W}_i \left[\boldsymbol{z}_i - \boldsymbol{H}_i\boldsymbol{x}\right] \tag{8.109}$$

Substitute for the definitions of $\boldsymbol{z}_0$, $\boldsymbol{H}_0$, and $\boldsymbol{W}_0$ and write the summation in our usual concatenated form to give

$$J = \left[\bar{\boldsymbol{x}} - \boldsymbol{x}\right]^T \overline{\boldsymbol{W}} \left[\bar{\boldsymbol{x}} - \boldsymbol{x}\right] + \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \boldsymbol{W} \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right] \tag{8.110}$$

Now we are ready to apply the derivative conditions to solve for the least-squares estimate.

The first derivative condition for an optimal is

$$\frac{\partial J}{\partial \boldsymbol{x}} = \boldsymbol{0}^T \tag{8.111}$$

where

$$\frac{\partial J}{\partial \boldsymbol{x}} = \frac{\partial}{\partial \boldsymbol{x}} \left\{ \left[\bar{\boldsymbol{x}} - \boldsymbol{x}\right]^T \overline{\boldsymbol{W}} \left[\bar{\boldsymbol{x}} - \boldsymbol{x}\right] + \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \boldsymbol{W} \left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right] \right\} \tag{8.112}$$

$$= -2\left[\bar{\boldsymbol{x}} - \boldsymbol{x}\right]^T \overline{\boldsymbol{W}} - 2\left[\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}\right]^T \boldsymbol{W}\boldsymbol{H} \tag{8.113}$$

Applying the first derivative condition to solve for $\hat{\boldsymbol{x}}$ yields

$$\boldsymbol{0} = -2\overline{\boldsymbol{W}} \left[\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}\right] - 2\boldsymbol{H}^T \boldsymbol{W} \left[\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}\right] \tag{8.114}$$

Manipulation of the preceding equation gives

$$\left[\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H}+\overline{\boldsymbol{W}}\right]\hat{\boldsymbol{x}}=\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{z}+\overline{\boldsymbol{W}}\bar{\boldsymbol{x}} \tag{8.115}$$

which has the solution

$$\hat{\boldsymbol{x}}=\left[\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{H}+\overline{\boldsymbol{W}}\right]^{-1}\left[\boldsymbol{H}^T\boldsymbol{W}\boldsymbol{z}+\overline{\boldsymbol{W}}\bar{\boldsymbol{x}}\right] \tag{8.116}$$

This is the weighted least-squares solution with the inclusion of some prior knowledge of the state of the system.

If we have no prior knowledge of the state, then $\overline{\boldsymbol{W}}=\boldsymbol{0}$, and we are left with the standard weighted least-squares solution.

8.2.6 The Minimum Variance Estimate

Least-squares provides a very powerful framework for estimating the state of a static or dynamic system.

The naive formulation of least squares gave us no ability to embed more or less confidence in individual observations of the system, so we developed the weighted least squares method.

The weighted least squares method also provided us with a natural mechanism for including information that we may have regarding the state of the system prior to acquiring measurements.

Neither of these approaches, however, allows us to include any information on the statistical characteristics of the measurements or of the prior state.

The weightings are simply values corresponding to how much we believe one quantity over another.

The minimum variance approach will provide us with a mechanism for obviating this limitation.

This approach is statistical in nature, but only requires the first and second moments (the mean and covariance) of the measurement errors and the prior state error.

Instead of taking a residual to be the difference between the actual and predicted observations, it is assumed that the measurements follow the model with the addition of a measurement error that is random with zero mean and known covariance.

The objective of the minimum variance estimator is stated as: given the dynamical system

$$\boldsymbol{x}_i = \boldsymbol{\Phi}(t_i, t_k)\boldsymbol{x}_k \tag{8.117}$$

and the observational system

$$\boldsymbol{z}_i = \widetilde{\boldsymbol{H}}_i \boldsymbol{x}_i + \boldsymbol{v}_i \tag{8.118}$$

where

$$\mathrm{E}\{\boldsymbol{v}_i\} = \boldsymbol{0} \;\; \forall \, i \quad \text{and} \quad \mathrm{E}\{\boldsymbol{v}_i \boldsymbol{v}_j^T\} = \boldsymbol{R}_{ij} \tag{8.119}$$

find the linear, unbiased, minimum variance estimate (LUMVE), $\hat{\boldsymbol{x}}_k$ of the state $\boldsymbol{x}_k$.

First, using the methods employed previously, we reduce this problem to one of estimating $\boldsymbol{x}_k$ through the use of the state transition matrix.

That is, we express the collection of measurements as

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v} \tag{8.120}$$

where

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \widetilde{\boldsymbol{H}}_1 \boldsymbol{\Phi}(t_1, t_k) \\ \widetilde{\boldsymbol{H}}_2 \boldsymbol{\Phi}(t_2, t_k) \\ \vdots \end{bmatrix} \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \end{bmatrix} \tag{8.121}$$

It is important to remember that we are now describing the measurements as being the model subjected to measurement noise. We assume that the concatenated measurement noise has first and second central moments of

$$\mathrm{E}\{\boldsymbol{v}\} = \begin{bmatrix} \mathrm{E}\{\boldsymbol{v}_1\} \\ \mathrm{E}\{\boldsymbol{v}_2\} \\ \vdots \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \end{bmatrix} \qquad \text{and} \qquad \mathrm{E}\{\boldsymbol{v}\boldsymbol{v}^T\} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \cdots \\ \boldsymbol{R}_{12}^T & \boldsymbol{R}_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \boldsymbol{R} \tag{8.122}$$

Generally, the terms $\boldsymbol{R}_{ij}$ for $i \neq j$ are zero, implying that the measurement noise is a white sequence (uncorrelated in time). Also, it is common to find that the terms $\boldsymbol{R}_{ii}$ are all equal. Neither of these conditions, however, is required to be imposed in order to continue. In fact, the case of $\boldsymbol{R}_{ij} \neq \boldsymbol{0}$ corresponds to the case of time-correlated observation errors.

Now, we begin to apply the conditions of LUMVE (linear, unbiased, and minimum variance) to determine an estimate, $\hat{\boldsymbol{x}}_k$.

Linear We require that our estimate is a linear combination of the measurement data that we received. That is, we form the estimate as

$$\hat{\boldsymbol{x}}_k = \boldsymbol{M}\boldsymbol{z} \tag{8.123}$$

where the matrix $\boldsymbol{M} \in \mathbb{R}^{n \times m}$ is to be found.

It is important to note that the estimated state is indeed random, but that we will assume that the true state of the system is not random.

Unbiased We require that our estimate is unbiased. That is, the expected value of our estimate should be the true state, which gives

$$\mathrm{E}\{\hat{\boldsymbol{x}}_k\} = \boldsymbol{x}_k \tag{8.124}$$

From the condition of the estimate being a linear combination of the measurement data, it follows that

$$\mathrm{E}\{\boldsymbol{M}\boldsymbol{z}\} = \boldsymbol{x}_k \tag{8.125}$$

Then, from our assumed model for the measurement data, we have

$$\mathrm{E}\{\boldsymbol{M}\left[\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}\right]\} = \mathrm{E}\{\boldsymbol{M}\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{M}\boldsymbol{v}\} = \boldsymbol{x}_k \tag{8.126}$$

The expected value is a linear operator, so we may break the summation apart. Moreover, since $\boldsymbol{H}$ and $\boldsymbol{M}$ are deterministic matrices, they can be brought outside of the expectation, such that

$$\boldsymbol{M}\boldsymbol{H}\mathrm{E}\{\boldsymbol{x}_k\} + \boldsymbol{M}\mathrm{E}\{\boldsymbol{v}\} = \boldsymbol{x}_k \tag{8.127}$$

Finally, recalling that the measurement noise is taken to be zero mean and that the true state is not random, we find that

$$\boldsymbol{M}\boldsymbol{H}\boldsymbol{x}_k = \boldsymbol{x}_k \tag{8.128}$$

or that $\boldsymbol{M}\boldsymbol{H} = \boldsymbol{I}$ in order to obtain an unbiased estimate.

This effectively imposes a constraint on the matrix $\boldsymbol{M}$.

This condition requires that the rows of $\boldsymbol{M}$ are orthogonal to the columns of $\boldsymbol{H}$.

Minimum Variance This is the heart of LUMVE, so it will take a bit of work to get through the minimum variance condition.

We want our estimate to be linear and we want it to be unbiased, but we also want it to be good in that we want to

extract as much information as possible regarding the state of the system from our data.

Let's define the error in our estimate to be the deviation of our estimate away from the truth:

$$\boldsymbol{e}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k \tag{8.129}$$

We can show that the unbiased condition gives us an expected error that is zero, which is good.

$$\mathrm{E}\{\boldsymbol{e}_k\} = \mathrm{E}\{\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\} = \mathrm{E}\{\boldsymbol{x}_k\} - \mathrm{E}\{\hat{\boldsymbol{x}}_k\} = \boldsymbol{x}_k - \boldsymbol{x}_k = \boldsymbol{0} \tag{8.130}$$

From here, we can define the covariance of the error. We simply take the expected value of the product of the deviation of the error from its mean and the transpose of this quantity to give

$$\begin{aligned} \boldsymbol{P}_k &= \mathrm{E}\{\big[\boldsymbol{e}_k - \mathrm{E}\{\boldsymbol{e}_k\}\big]\big[\boldsymbol{e}_k - \mathrm{E}\{\boldsymbol{e}_k\}\big]^T\} && (8.131)\\ &= \mathrm{E}\{\boldsymbol{e}_k \boldsymbol{e}_k^T\} && (8.132) \end{aligned}$$

We expect that our error is zero. We also want to have an error that is small in some sense. This is where our minimum variance condition comes in to play. The unbiased condition centers our error on zero, but we also want the spread about the center to be small.

From the preceding covariance equation and the definition of the error, it follows that the error covariance is

$$\boldsymbol{P}_k = \mathrm{E}\{\big[\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\big]\big[\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\big]^T\} \tag{8.133}$$

We can manipulate this expression a bit by recalling the conditions of our estimator.

First of all, the estimate is linear, so we can replace the estimated state with its linear mapping to give

$$\boldsymbol{P}_k = \mathrm{E}\{\big[\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{z}\big]\big[\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{z}\big]^T\} \tag{8.134}$$

Next, we know that the data are related to the state by the observational equation, so we can substitute this back into the covariance expression

$$\begin{aligned} \boldsymbol{P}_k &= \mathrm{E}\{\big[\boldsymbol{x}_k - \boldsymbol{M}\big(\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}\big)\big]\big[\boldsymbol{x}_k - \boldsymbol{M}\big(\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}\big)\big]^T\} && (8.135)\\ &= \mathrm{E}\{\big[\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{H}\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{v}\big]\big[\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{H}\boldsymbol{x}_k - \boldsymbol{M}\boldsymbol{v}\big]^T\} && (8.136) \end{aligned}$$

Now, we recall the condition for an unbiased estimator is $\boldsymbol{MH} = \boldsymbol{I}$, which gives

$$\boldsymbol{P}_k = \mathrm{E}\{\left[\boldsymbol{x}_k - \boldsymbol{x}_k - \boldsymbol{Mv}\right]\left[\boldsymbol{x}_k - \boldsymbol{x}_k - \boldsymbol{Mv}\right]^T\} \tag{8.137}$$

A simple elimination of the true state produces a covariance expression that is only dependent upon the linear mapping of the estimator and the measurement noise

$$\begin{aligned}\boldsymbol{P}_k &= \mathrm{E}\{\left[\boldsymbol{Mv}\right]\left[\boldsymbol{Mv}\right]^T\} &(8.138)\\ &= \mathrm{E}\{\boldsymbol{Mvv}^T\boldsymbol{M}^T\} &(8.139)\end{aligned}$$

Since the linear mapping is deterministic, we have

$$\boldsymbol{P}_k = \boldsymbol{M}\mathrm{E}\{\boldsymbol{vv}^T\}\boldsymbol{M}^T \tag{8.140}$$

The expected value here is just the covariance of the concatenated measurement, such that

$$\boldsymbol{P}_k = \boldsymbol{MRM}^T \tag{8.141}$$

This is the covariance we wish to minimize, but we also have to keep in mind the active constraint on the matrix $\boldsymbol{M}$.

To account for the constraint on $\boldsymbol{M}$, we add "zero" to the covariance matrix:

$$\boldsymbol{P}_k = \boldsymbol{MRM}^T + \boldsymbol{\Lambda}^T\left[\boldsymbol{I} - \boldsymbol{MH}\right]^T \tag{8.142}$$

However, since this is a covariance matrix, we must ensure that it remains symmetric, so we add "zero" in another way:

$$\boldsymbol{P}_k = \boldsymbol{MRM}^T + \boldsymbol{\Lambda}^T\left[\boldsymbol{I} - \boldsymbol{MH}\right]^T + \left[\boldsymbol{I} - \boldsymbol{MH}\right]\boldsymbol{\Lambda} \tag{8.143}$$

The matrix $\boldsymbol{\Lambda}$ is a matrix of unspecified Lagrange multipliers that account for the constraint on the matrix $\boldsymbol{M}$

while still retaining a symmetric covariance matrix. The advantage is that we can proceed with unconstrained minimization techniques.

To find a minimum, we need to set the first variation of the covariance matrix equal to zero

$$\delta \boldsymbol{P}_k = \boldsymbol{0} \tag{8.144}$$

We have two parameters to solve for: $\boldsymbol{M}$ and $\boldsymbol{\Lambda}$.

The variation of $\boldsymbol{P}_k$ with respect to $\boldsymbol{\Lambda}$ will simply return the constraint $\boldsymbol{MH} = \boldsymbol{I}$.

The variation of $\boldsymbol{P}_k$ with respect to $\boldsymbol{M}$ is what we're really after

$$\delta \boldsymbol{P}_k = \delta \boldsymbol{M} \cdot \boldsymbol{R}\boldsymbol{M}^T + \boldsymbol{M}\boldsymbol{R} \cdot \delta \boldsymbol{M}^T + \boldsymbol{\Lambda}^T \left[-\delta \boldsymbol{M} \cdot \boldsymbol{H} \right]^T + \left[-\delta \boldsymbol{M} \cdot \boldsymbol{H} \right] \boldsymbol{\Lambda} \tag{8.145}$$

After a bit of rearranging, we find that the first variation is

$$\delta \boldsymbol{P}_k = \delta \boldsymbol{M} \left[\boldsymbol{R}\boldsymbol{M}^T - \boldsymbol{H}\boldsymbol{\Lambda} \right] + \left[\boldsymbol{M}\boldsymbol{R} - \boldsymbol{\Lambda}^T \boldsymbol{H}^T \right] \delta \boldsymbol{M}^T \tag{8.146}$$

Now, we set this equal to zero, which will happen if

$$\boldsymbol{M}\boldsymbol{R} - \boldsymbol{\Lambda}^T \boldsymbol{H}^T = \boldsymbol{0} \tag{8.147}$$

Note that it will also occur if $\delta \boldsymbol{M}$ and/or $\boldsymbol{M}\boldsymbol{R} - \boldsymbol{\Lambda}^T \boldsymbol{H}^T$ are not full rank. We will focus on the first condition in order to find the requirements to obtain a minimum of the covariance matrix.

We need to solve a set of two simultaneous equations

$$\boldsymbol{M}\boldsymbol{R} - \boldsymbol{\Lambda}^T \boldsymbol{H}^T = \boldsymbol{0} \qquad \text{and} \qquad \boldsymbol{I} - \boldsymbol{M}\boldsymbol{H} = \boldsymbol{0} \tag{8.148}$$

Since $\boldsymbol{R}$ is taken to be positive definite, its inverse is guaranteed to exist, and we can solve the first equation for the matrix $\boldsymbol{M}$ as

$$\boldsymbol{M} = \boldsymbol{\Lambda}^T \boldsymbol{H}^T \boldsymbol{R}^{-1} \tag{8.149}$$

We can substitute this result into the second equation to find that

$$\boldsymbol{I} = \boldsymbol{\Lambda}^T \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H} \tag{8.150}$$

Provided that we have more observations than parameters, the inverse of $\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}$ will exist, and we can solve for the matrix of Lagrange multipliers as

$$\boldsymbol{\Lambda}^T = \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} \tag{8.151}$$

Finally, we take the now known Lagrange multipliers and substitute them back into the solution for the matrix $\boldsymbol{M}$, which gives

$$\boldsymbol{M} = \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \tag{8.152}$$

This is the value of $\boldsymbol{M}$ that minimizes the (co)variance $\boldsymbol{P}_k$ while simultaneously yielding an unbiased estimate $\hat{\boldsymbol{x}}_k$.

The covariance matrix can then be found as

$$\begin{aligned}
\boldsymbol{P}_k &= \boldsymbol{M}\boldsymbol{R}\boldsymbol{M}^T && (8.153)\\
\text{subst. for } \boldsymbol{M} \qquad &= \left\{\left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1}\right\} \boldsymbol{R} \left\{\boldsymbol{R}^{-1} \boldsymbol{H} \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1}\right\} && (8.154)\\
\text{eliminate and group} \qquad &= \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right] \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} && (8.155)\\
\text{cancel} \qquad &= \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} && (8.156)
\end{aligned}$$

An alternative way of finding the covariance matrix depends on the matrix of Lagrange multipliers:

$$\begin{aligned}
\boldsymbol{P}_k &= \boldsymbol{M}\boldsymbol{R}\boldsymbol{M}^T && (8.157)\\
\text{subst. for } \boldsymbol{M}\boldsymbol{R} = \boldsymbol{\Lambda}^T\boldsymbol{H}^T \qquad &= \boldsymbol{\Lambda}^T \boldsymbol{H}^T \boldsymbol{M}^T && (8.158)\\
\text{regroup terms} \qquad &= \boldsymbol{\Lambda}^T \left[\boldsymbol{M}\boldsymbol{H}\right]^T && (8.159)\\
\text{subst. for } \boldsymbol{M}\boldsymbol{H} = \boldsymbol{I} \qquad &= \boldsymbol{\Lambda}^T && (8.160)\\
\text{def. of } \boldsymbol{\Lambda} \qquad &= \left[\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right]^{-1} && (8.161)
\end{aligned}$$

Even though we have found a solution to the first variational equation, it is not obvious that this solution actually

minimizes the estimation error covariance.

To show that this solution does indeed yield a minimum variance estimate, we first recall that the estimate is given by

$$\hat{\boldsymbol{x}}_k = \boldsymbol{M}\boldsymbol{z} = \left[\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z} \tag{8.162}$$

Consider then, without loss of generality, another estimate

$$\tilde{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z} \tag{8.163}$$

where $\boldsymbol{B}$ is not a null matrix.

This estimate is still a linear estimate. What is the condition for the estimate to be unbiased?

Take the expected value of the new estimate and apply the known relationships to find that

$$\begin{aligned}
\mathrm{E}\{\tilde{\boldsymbol{x}}_k\} &= \mathrm{E}\{\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z}\} && (8.164)\\
\text{separate terms} \qquad &= \mathrm{E}\{\hat{\boldsymbol{x}}_k\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{z}\} && (8.165)\\
\text{unbiased estimate} \qquad &= \boldsymbol{x}_k + \mathrm{E}\{\boldsymbol{B}\boldsymbol{z}\} && (8.166)\\
\text{meas. model} \qquad &= \boldsymbol{x}_k + \mathrm{E}\{\boldsymbol{B}\left[\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}\right]\} && (8.167)\\
\text{separate terms} \qquad &= \boldsymbol{x}_k + \mathrm{E}\{\boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\} && (8.168)\\
\text{pull out deterministic terms} \qquad &= \boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{H}\mathrm{E}\{\boldsymbol{x}_k\} + \boldsymbol{B}\mathrm{E}\{\boldsymbol{v}\} && (8.169)\\
\text{unbiased est. and zero-mean noise} \qquad &= \boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k && (8.170)
\end{aligned}$$

That is, $\boldsymbol{B}\boldsymbol{H} = \boldsymbol{0}$ is the condition for this new estimate to be unbiased.

We have already excluded the case that $\boldsymbol{B} = \boldsymbol{0}$, and we know that $\boldsymbol{H}$ is full rank, so it must be that the matrix $\boldsymbol{B}$ cannot be full rank in order to have an unbiased estimate.

Define the error in the estimate to be

$$\tilde{\boldsymbol{e}}_k = \boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k \tag{8.171}$$

and take the expected value of the error

$$\mathrm{E}\{\tilde{\boldsymbol{e}}_k\} = \mathrm{E}\{\boldsymbol{x}_k\} - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\} = \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k = \boldsymbol{0} \tag{8.172}$$

The covariance of the estimation error is then given by

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k\right]\left[\boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k\right]^T\} \tag{8.173}$$

Note that we can write the error also in terms of the state estimate and its expectation, which gives

$$\begin{aligned} \boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k &= \left[\boldsymbol{x}_k - \tilde{\boldsymbol{x}}_k\right] + \left[\mathrm{E}\{\tilde{\boldsymbol{x}}_k\} - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right] && (8.174)\\ &= \left[\mathrm{E}\{\tilde{\boldsymbol{x}}_k\} - \tilde{\boldsymbol{x}}_k\right] && (8.175)\\ &= -\left[\tilde{\boldsymbol{x}}_k - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right] && (8.176) \end{aligned}$$

From the new expression of the error, we have the covariance as

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\tilde{\boldsymbol{x}}_k - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right]\left[\tilde{\boldsymbol{x}}_k - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right]^T\} \tag{8.177}$$

Substitute for the new estimate in terms of the original estimate

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\left(\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z}\right) - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right]\left[\left(\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z}\right) - \mathrm{E}\{\tilde{\boldsymbol{x}}_k\}\right]^T\} \tag{8.178}$$

Substitute for the expected value of the new estimate without restricting $\boldsymbol{B}$

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\left(\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z}\right) - \left(\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k\right)\right]\left[\left(\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{z}\right) - \left(\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k\right)\right]^T\} \tag{8.179}$$

Rearrange terms

$$\begin{aligned} \boldsymbol{P}_{\tilde{\boldsymbol{x}}} &= \mathrm{E}\{\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \left(\boldsymbol{B}\boldsymbol{z} - \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k\right)\right]\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \left(\boldsymbol{B}\boldsymbol{z} - \boldsymbol{B}\boldsymbol{H}\boldsymbol{x}_k\right)\right]^T\} && (8.180)\\ &= \mathrm{E}\{\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \boldsymbol{B}\left(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}_k\right)\right]\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \boldsymbol{B}\left(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}_k\right)\right]^T\} && (8.181) \end{aligned}$$

Recall that the observations are assumed to follow $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}$:

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \boldsymbol{B}\boldsymbol{v}\right]\left[\left(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right) + \boldsymbol{B}\boldsymbol{v}\right]^T\} \tag{8.182}$$

Expand the product and distribute the expectation to each of the resulting terms

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \mathrm{E}\{\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} + \mathrm{E}\{\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]\boldsymbol{v}^T\boldsymbol{B}^T\} \tag{8.183}$$

$$+ \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{B}^T\} \tag{8.184}$$

The first term is simply the error covariance of the original estimate

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \boldsymbol{P}_k + \mathrm{E}\{\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]\boldsymbol{v}^T\boldsymbol{B}^T\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{B}^T\} \tag{8.185}$$

The $\boldsymbol{B}$ matrices can be pulled outside of the expectation in the last term, leaving the expectation as simply the measurement noise covariance, such that

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \boldsymbol{P}_k + \mathrm{E}\{\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]\boldsymbol{v}^T\boldsymbol{B}^T\} + \mathrm{E}\{\boldsymbol{B}\boldsymbol{v}\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} + \boldsymbol{B}\boldsymbol{R}\boldsymbol{B}^T \tag{8.186}$$

The matrix $\boldsymbol{B}$ is deterministic, so we pull it out of the middle two terms

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \boldsymbol{P}_k + \mathrm{E}\{\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]\boldsymbol{v}^T\}\boldsymbol{B}^T + \boldsymbol{B}\mathrm{E}\{\boldsymbol{v}\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} + \boldsymbol{B}\boldsymbol{R}\boldsymbol{B}^T \tag{8.187}$$

Now we just need to determine what the form of the middle two expectations is. Note that they are just transposes of one another, so it will suffice to focus on the second of the two.

$$\mathrm{E}\{\boldsymbol{v}\left[\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\right]^T\} = \mathrm{E}\{\boldsymbol{v}\left[\boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z} - \boldsymbol{x}_k\right]^T\} \tag{8.188}$$

$$\text{(meas. model)} \quad = \mathrm{E}\{\boldsymbol{v}\left[\boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\left(\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}\right) - \boldsymbol{x}_k\right]^T\} \tag{8.189}$$

$$\text{(distribute)} \quad = \mathrm{E}\{\boldsymbol{v}\left[\boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{v} - \boldsymbol{x}_k\right]^T\} \tag{8.190}$$

$$\text{(cov. def.)} \quad = \mathrm{E}\{\boldsymbol{v}\left[\boldsymbol{x}_k + \boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{v} - \boldsymbol{x}_k\right]^T\} \tag{8.191}$$

$$\text{(cancel)} \quad = \mathrm{E}\{\boldsymbol{v}\left[\boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{v}\right]^T\} \tag{8.192}$$

$$\text{(rearrange)} \quad = \mathrm{E}\{\boldsymbol{v}\boldsymbol{v}^T\}\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{P}_k \tag{8.193}$$

$$\text{(noise cov.)} \quad = \boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{P}_k \tag{8.194}$$

$$\text{(cancel)} \quad = \boldsymbol{H}\boldsymbol{P}_k \tag{8.195}$$

Applying the preceding result to the covariance for the alternative update yields

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \boldsymbol{P}_k + \boldsymbol{P}_k\boldsymbol{H}^T\boldsymbol{B}^T + \boldsymbol{B}\boldsymbol{H}\boldsymbol{P}_k + \boldsymbol{B}\boldsymbol{R}\boldsymbol{B}^T \tag{8.196}$$

For the new update to be unbiased we required that $\boldsymbol{BH} = \boldsymbol{0}$, so we arrive at the covariance of the new update

$$\boldsymbol{P}_{\tilde{\boldsymbol{x}}} = \boldsymbol{P}_k + \boldsymbol{BRB}^T \tag{8.197}$$

Since $\boldsymbol{R}$ is positive definite and $\boldsymbol{B}$ is not full rank, $\boldsymbol{BRB}^T \geq \boldsymbol{0}$ and $\boldsymbol{P}_{\tilde{\boldsymbol{x}}} \geq \boldsymbol{P}_k$. That is, the alternate estimate of the state has a larger covariance than the original estimate. Our original estimate is therefore a minimum variance estimate.

The development of the linear, unbiased, minimum variance estimator is now complete.

To summarize, the estimate is given by

$$\hat{\boldsymbol{x}}_k = \left[\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z} \tag{8.198}$$

and the covariance of this estimate is

$$\boldsymbol{P}_k = \left[\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.199}$$

It is interesting to note that we would arrive at the same estimate by applying $\boldsymbol{W} = \boldsymbol{R}^{-1}$ in the weighted least squares approach.

However, now we have an approach to least-squares estimation that allows us to weight each measurement by our statistical confidence in each measurement.

There is no appearance of a prior estimate (or its associated statistical confidence), but we had a method for including a prior estimate in weighted least squares.

We can follow the same approach used for weighted to least squares to include prior information in LUMVE.

Let's define a "zeroth" measurement to be a measurement of the state

$$\boldsymbol{z}_0 = \boldsymbol{x}_k + \boldsymbol{\eta} \tag{8.200}$$

We have used a "measurement noise" represented by $\boldsymbol{\eta}$ to denote that this is not the same as the standard measurement noise.

We assume that $\boldsymbol{\eta}$ is zero mean with covariance $\bar{\boldsymbol{P}}_k$.

Additionally, we represent this zeroth measurement symbolically as $\bar{\boldsymbol{x}}_k$.

Given the data, measurement mapping, and measurement noise as $\boldsymbol{z}$, $\boldsymbol{H}$, and $\boldsymbol{v}$, respectively, from the LUMVE development, we define augmented data, measurement mapping, and measurement noise as

$$\bar{\boldsymbol{z}} = \begin{bmatrix} \bar{\boldsymbol{x}}_k \\ \boldsymbol{z} \end{bmatrix}, \quad \bar{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{H} \end{bmatrix}, \quad \text{and} \quad \bar{\boldsymbol{v}} = \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{bmatrix} \tag{8.201}$$

Since $\boldsymbol{\eta}$ and $\boldsymbol{v}$ are both zero mean, it follows that $\bar{\boldsymbol{v}}$ is also zero mean.

We take the prior data to be uncorrelated with the measurement data, such that the covariance of $\bar{\boldsymbol{v}}$ is

$$\bar{\boldsymbol{R}} = \begin{bmatrix} \bar{\boldsymbol{P}}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \tag{8.202}$$

where $\boldsymbol{R}$ is the covariance of $\boldsymbol{v}$.

Using $\bar{\boldsymbol{z}}$, $\bar{\boldsymbol{H}}$, and $\bar{\boldsymbol{R}}$, LUMVE gives the estimated state and its associated covariance as

$$\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{H}}^T \bar{\boldsymbol{R}}^{-1} \bar{\boldsymbol{H}}\right]^{-1} \bar{\boldsymbol{H}}^T \bar{\boldsymbol{R}}^{-1} \bar{\boldsymbol{z}} \tag{8.203}$$

$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{H}}^T \bar{\boldsymbol{R}}^{-1} \bar{\boldsymbol{H}}\right]^{-1} \tag{8.204}$$

Now we want to substitute for the augmented quantities and express this estimate in terms of the original variables.

$$\bar{\boldsymbol{H}}^T \bar{\boldsymbol{R}}^{-1} \bar{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{H} \end{bmatrix}^T \begin{bmatrix} \bar{\boldsymbol{P}}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{H} \end{bmatrix} \tag{8.205}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^T \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{P}}_k^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{H} \end{bmatrix} \tag{8.206}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^T \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{P}}_k^{-1} \\ \boldsymbol{R}^{-1}\boldsymbol{H} \end{bmatrix} \tag{8.207}$$

$$= \bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H} \tag{8.208}$$

$$\bar{\boldsymbol{H}}^T \bar{\boldsymbol{R}}^{-1} \bar{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{H} \end{bmatrix}^T \begin{bmatrix} \bar{\boldsymbol{P}}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}^{-1} \begin{bmatrix} \bar{\boldsymbol{x}}_k \\ \boldsymbol{z} \end{bmatrix} \tag{8.209}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^T \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{P}}_k^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}^{-1} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{x}}_k \\ \boldsymbol{z} \end{bmatrix} \tag{8.210}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^T \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k \\ \boldsymbol{R}^{-1}\boldsymbol{z} \end{bmatrix} \tag{8.211}$$

$$= \bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{z} \tag{8.212}$$

Putting it all back together, the LUMVE with prior information and its covariance are given by

$$\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z}\right] \tag{8.213}$$

$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.214}$$

8.2.6.1 Example: Revisiting the Robot Problem

Let's revisit our robot tracking problem.

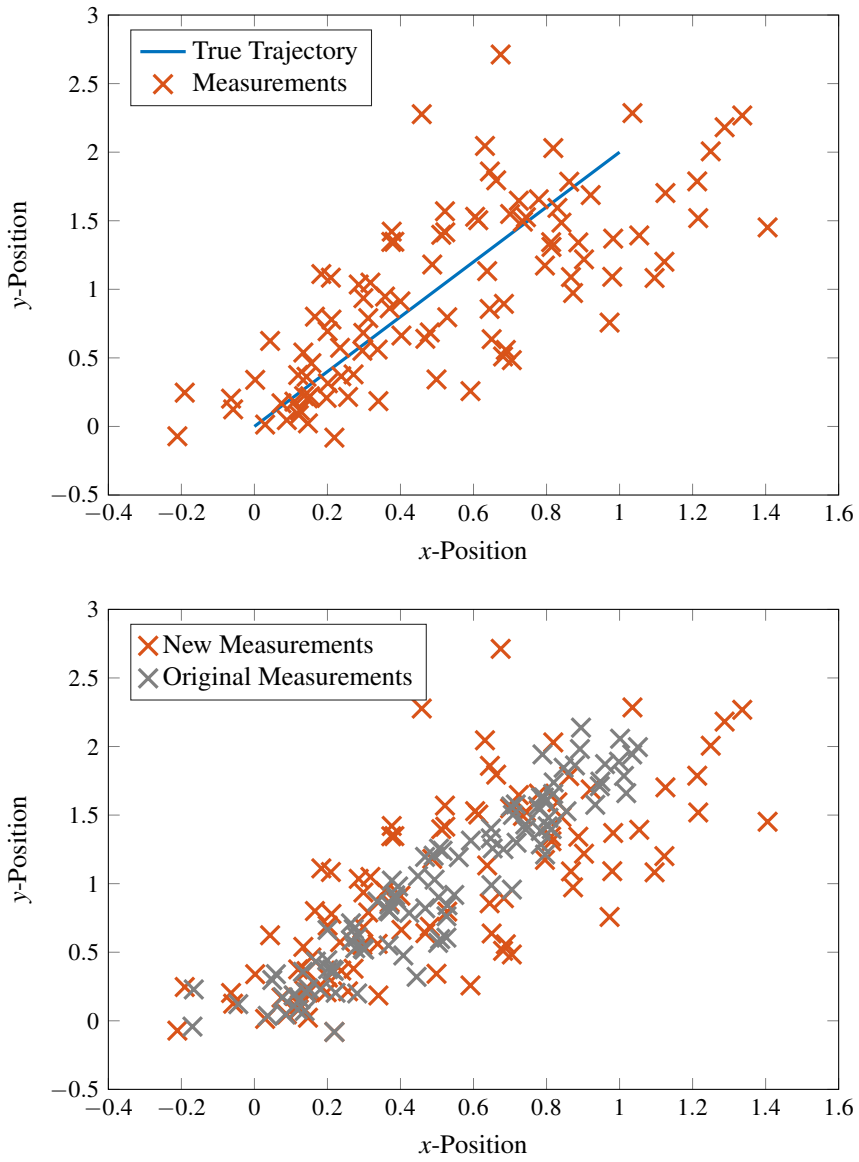
This time, we want to apply the least squares, weighted least squares, and LUMVE estimates to see any differences.

The configuration is identical to the previous time we dealt with the robot problem, except now our measurements will get progressively less accurate.

We are still assuming that the robot moves in a plane under a constant velocity model, starting from the origin.

The measurements are still formed from the position of the robot with noise taken from a Gaussian distribution. As time progresses, the noise increases from a standard deviation of 0.1 meters to 0.4 meters.

```
% Create measurements of position
sig = linspace(0.1,0.4,length(tv));
z   = zeros(2,length(tv));
for k = 1:length(tv)
    z(:,k) = X(k,1:2)' + sig(k)*randn(2,1);
end
```



For the least squares estimate, we do not have any mechanism to deal with the changing measurement noise, so this approach remains the same.

For the weighted least squares estimate, we need to select the weighting matrices $\boldsymbol{W}_i$ that accompany the observations.

These weighting matrices only reflect our rough confidence in the measurements and are not necessarily selected based on statistics.

Therefore, we arbitrarily claim that the first half of the measurements are twice as believable as the second half of the

measurements.

For LUMVE, we use the actual time-varying statistics of the measurement noise to accumulate our $\boldsymbol{R}$ matrix from the $\boldsymbol{R}_i$ matrices, which are chosen as $\boldsymbol{R}_i = \sigma_i^2 \boldsymbol{I}_{2\times 2}$.

Additionally, we assume that the measurement noise is white, such that the off-diagonal terms in $\boldsymbol{R}$ are zero.

Let's look at how the accumulation loop would be coded in MATLAB

```
% Assemble the concatenated terms
Z = []; H = []; W = []; R = [];
for k = 1:length(tv)
    % time at kth observation
    tk = tv(k);

    % concatenate measurements
    Z = [Z;z(1,k);z(2,k)];

    % concatenate measurement-mapping matrices
    Htilde = [1,0,0,0;0,1,0,0];
    Phik0  = [1,0,tk-t0,0;0,1,0,tk-t0;0,0,1,0;0,0,0,1];
    H      = [H;Htilde*Phik0];




    % concatenate weighting matrices for WLS
    if k < round(length(tv)/2)
        W = blkdiag(W,2.0*eye(2));
    else
        W = blkdiag(W,1.0*eye(2));
    end

    % concatenate LUMVE inverse weighting matrix
    R = blkdiag(R, sig(k)^2*eye(2));
end
% LUMVE weighting matrix
Ri = inv(R);
```

Now, we can apply our three estimation techniques.

```
% Least-squares estimate
xhatLS = (H'*H)\(H'*Z);

% Weighted least-squares estimate
xhatWLS = (H'*W*H)\(H'*W*Z);

% LUMVE estimate
xhatLUMVE = (H'*Ri*H)\(H'*Ri*Z);
```

The results from the estimates are

$$\hat{\boldsymbol{x}}_0^{(\text{LS})} = \begin{bmatrix} 0.0182 \\ -0.0533 \\ 0.1003 \\ 0.2053 \end{bmatrix} \quad \hat{\boldsymbol{x}}_0^{(\text{WLS})} = \begin{bmatrix} 0.0189 \\ -0.0491 \\ 0.0996 \\ 0.2040 \end{bmatrix} \quad \hat{\boldsymbol{x}}_0^{(\text{LUMVE})} = \begin{bmatrix} 0.0244 \\ -0.0241 \\ 0.0989 \\ 0.1986 \end{bmatrix} \tag{8.215}$$

8.2.7 Sequential Least Squares

This section may be omitted on a first reading; it will not be covered in class.

The batch processor gives us a method for accumulating all of the data and processing them together in order to formulate a statistical estimate of the system state at some fixed time.

To reiterate the batch processor, it is given by

$$\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z}\right] \tag{8.216}$$

$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.217}$$

where $\hat{\boldsymbol{x}}_k$ is our state estimate and $\boldsymbol{P}_k$ is the covariance (a measure of the confidence) of the estimation error.

Note that the computation of the estimate $\hat{\boldsymbol{x}}_k$ *always* requires the inverse of the matrix $\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}$, and also requires the inverse of $\bar{\boldsymbol{P}}_k$ when we are including prior information on the state.

Assuming that there are n states, $\boldsymbol{H} \in \mathbb{R}^{n\times m}$ implies that we must compute two $n \times n$ inverses.

If the state dimension is large, these inverse calculations may be quite cumbersome.

Additionally, we require $\boldsymbol{R}^{-1}$, which is an $m \times m$ inverse. For the least squares techniques to work, $m > n$ means that the computation of $\boldsymbol{R}^{-1}$ may also be challenging.

This last challenge is usually easily overcome when the measurements are uncorrelated. In this case, it is easy to see that

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{R}_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{8.218}$$

where $\boldsymbol{R}_i \in \mathbb{R}^{q\times q}$ (provided that each measurement is q-dimensional).

Even if the individual measurements are not of the same dimension, the dimension of an individual measurement is substantially smaller than the dimension of the concatenated set of measurements.

If the concatenated measurement noise covariance matrix is block diagonal (which is the case for uncorrelated measurement noises), then the inverse is simply given by the block diagonal matrix

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{R}_1^{-1} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{R}_2^{-1} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{8.219}$$

and requires only the smaller inverse computations.

The objective of the sequential least squares method is to alleviate some of the computational burden associated with the inverse of large-scale matrices in favor of inverting smaller matrices.

First, let us assume that we have applied the batch processor to a set of data to arrive at an estimate and its covariance at time t_j:

$$\hat{\boldsymbol{x}}_j = \left[\bar{\boldsymbol{P}}_j^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\left[\bar{\boldsymbol{P}}_j^{-1}\bar{\boldsymbol{x}}_j + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z}\right] \tag{8.220}$$

$$\boldsymbol{P}_j = \left[\bar{\boldsymbol{P}}_j^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.221}$$

Further, let us assume that we then acquire a new measurement at time $t_k > t_j$ that is of the form

$$\boldsymbol{z}_k = \widetilde{\boldsymbol{H}}_k\boldsymbol{x}_k + \boldsymbol{v}_k \tag{8.222}$$

where $\boldsymbol{v}_k$ is zero mean with covariance $\boldsymbol{R}_k$.

In order to apply the batch processor, we would have to append this new data to the previous data set (along with the measurement mapping and covariance) and then recompute the **entire** state estimate, which means that we have lost all of the work we did in computing the estimate $\hat{\boldsymbol{x}}_j$.

This is not an ideal scenario.

Instead, it would be better if we could *propagate* our previous estimate and the covariance to time t_k and then *update* our state estimate and covariance.

If we continue to acquire more data, we would then continue the cycle of propagating and updating our estimate and its

covariance; this is the idea behind the sequential estimation approach.

We need to "sequentialize" the LUMVE.

To achieve this, we first need to propagate the state estimate and the covariance.

Recall that our dynamical system is of the form

$$\boldsymbol{x}_k = \boldsymbol{\Phi}(t_k, t_j)\boldsymbol{x}_j \tag{8.223}$$

Taking the expected value of both sides, and noting that the state transition matrix is deterministic, it follows that the state estimate is propagated as

$$\bar{\boldsymbol{x}}_k = \boldsymbol{\Phi}(t_k, t_j)\hat{\boldsymbol{x}}_j \tag{8.224}$$

where $\hat{\boldsymbol{x}}_j$ is our state estimate at time t_j and $\bar{\boldsymbol{x}}_k$ is the propagated state estimate at time t_k.

Define the estimation error at time t_j as

$$\boldsymbol{e}_j = \boldsymbol{x}_j - \hat{\boldsymbol{x}}_j \tag{8.225}$$

and at time t_k as

$$\boldsymbol{e}_k = \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k \tag{8.226}$$

Pre-multiply the state estimation error at time t_j by the state transition matrix:

$$\begin{aligned}
\boldsymbol{\Phi}(t_k, t_j)\boldsymbol{e}_j &= \boldsymbol{\Phi}(t_k, t_j)\left[\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j\right] && (8.227)\\
&= \boldsymbol{\Phi}(t_k, t_j)\boldsymbol{x}_j - \boldsymbol{\Phi}(t_k, t_j)\hat{\boldsymbol{x}}_j && (8.228)\\
&= \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k && (8.229)\\
&= \boldsymbol{e}_k && (8.230)
\end{aligned}$$

The covariance of the state estimation error at time t_k is

$$\bar{\boldsymbol{P}}_k = \mathrm{E}\{\boldsymbol{e}_k \boldsymbol{e}_k^T\} \tag{8.231}$$

Since the error evolves as

$$\boldsymbol{e}_k = \boldsymbol{\Phi}(t_k, t_j)\boldsymbol{e}_j \tag{8.232}$$

we have

$$\bar{\boldsymbol{P}}_k = \mathrm{E}\{\boldsymbol{e}_k \boldsymbol{e}_k^T\} \tag{8.233}$$
$$= \mathrm{E}\{\boldsymbol{\Phi}(t_k,t_j)\boldsymbol{e}_j \boldsymbol{e}_j^T \boldsymbol{\Phi}^T(t_k,t_j)\} \tag{8.234}$$
$$= \boldsymbol{\Phi}(t_k,t_j)\mathrm{E}\{\boldsymbol{e}_j \boldsymbol{e}_j^T\}\boldsymbol{\Phi}^T(t_k,t_j) \tag{8.235}$$
$$= \boldsymbol{\Phi}(t_k,t_j)\boldsymbol{P}_j\boldsymbol{\Phi}^T(t_k,t_j) \tag{8.236}$$

Therefore, given the state estimate and covariance at time t_j, the propagated state estimate and covariance at time t_k are

$$\bar{\boldsymbol{x}}_k = \boldsymbol{\Phi}(t_k,t_j)\hat{\boldsymbol{x}}_j \tag{8.237}$$
$$\bar{\boldsymbol{P}}_k = \boldsymbol{\Phi}(t_k,t_j)\boldsymbol{P}_j\boldsymbol{\Phi}^T(t_k,t_j) \tag{8.238}$$

Now, we proceed to the update stage of the sequential estimation algorithm.

At time t_k, we have the additional measurement

$$\boldsymbol{z}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k + \boldsymbol{v}_k \tag{8.239}$$

along with the prior information contained in $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{P}}_k$.

The fused estimate is then given by LUMVE as

$$\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1} \widetilde{\boldsymbol{H}}_k\right]^{-1} \left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1}\boldsymbol{z}_k\right] \tag{8.240}$$
$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1} \widetilde{\boldsymbol{H}}_k\right]^{-1} \tag{8.241}$$

It is important to note that this application of LUMVE is only for fusing the prior information with the single measurement $\boldsymbol{z}_k$, as opposed to previous applications where we collected the measurement data and concatenated the terms.

However, LUMVE still requires the computation of $n \times n$ inverses, and this is what we want to avoid.

We will use the Matrix Inversion Lemma to reduce the $n \times n$ inverse into a $q \times q$ inverse.

Matrix Inversion Lemma

$$\left[\boldsymbol{A} + \boldsymbol{UCV}\right]^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left[\boldsymbol{VA}^{-1}\boldsymbol{U} + \boldsymbol{C}^{-1}\right]^{-1}\boldsymbol{VA}^{-1} \tag{8.242}$$

Let

$$\boldsymbol{A} = \bar{\boldsymbol{P}}_k^{-1}, \quad \boldsymbol{U} = \widetilde{\boldsymbol{H}}_k^T, \quad \boldsymbol{C} = \boldsymbol{R}_k^{-1}, \quad \text{and} \quad \boldsymbol{V} = \widetilde{\boldsymbol{H}}_k$$

Then, from the Matrix Inversion Lemma, it follows that

$$\left[\bar{\boldsymbol{P}}_k^{-1} + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1} \widetilde{\boldsymbol{H}}_k\right]^{-1} = \bar{\boldsymbol{P}}_k - \bar{\boldsymbol{P}}_k \widetilde{\boldsymbol{H}}_k^T \left[\widetilde{\boldsymbol{H}}_k \bar{\boldsymbol{P}}_k \widetilde{\boldsymbol{H}}_k^T + \boldsymbol{R}_k\right]^{-1} \widetilde{\boldsymbol{H}}_k \bar{\boldsymbol{P}}_k \tag{8.243}$$

Note that the Matrix Inversion Lemma has converted an $n \times n$ inverse on the left-hand side to a $q \times q$ inverse on the right-hand side. Provided that $q < n$, which is often the case, this represents an easier inversion. Additionally, there are fewer inverses required.

For ease of notation, it is common to define

$$\boldsymbol{K}_k = \bar{\boldsymbol{P}}_k \widetilde{\boldsymbol{H}}_k^T \left[\widetilde{\boldsymbol{H}}_k \bar{\boldsymbol{P}}_k \widetilde{\boldsymbol{H}}_k^T + \boldsymbol{R}_k\right]^{-1} \tag{8.244}$$

such that

$$\left[\bar{\boldsymbol{P}}_k^{-1} + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1} \widetilde{\boldsymbol{H}}_k\right]^{-1} = \bar{\boldsymbol{P}}_k - \boldsymbol{K}_k \widetilde{\boldsymbol{H}}_k \bar{\boldsymbol{P}}_k \tag{8.245}$$

$$= \left[\boldsymbol{I} - \boldsymbol{K}_k \widetilde{\boldsymbol{H}}_k\right] \bar{\boldsymbol{P}}_k \tag{8.246}$$

Substituting this result into the update equations gives

$$\hat{\boldsymbol{x}}_k = \left[\boldsymbol{I} - \boldsymbol{K}_k \widetilde{\boldsymbol{H}}_k\right] \bar{\boldsymbol{P}}_k \left[\bar{\boldsymbol{P}}_k^{-1} \bar{\boldsymbol{x}}_k + \widetilde{\boldsymbol{H}}_k^T \boldsymbol{R}_k^{-1} \boldsymbol{z}_k\right] \tag{8.247}$$

$$\boldsymbol{P}_k = \left[\boldsymbol{I} - \boldsymbol{K}_k \widetilde{\boldsymbol{H}}_k\right] \bar{\boldsymbol{P}}_k \tag{8.248}$$

While the covariance update looks good, we still have a bit of a mess in the state estimate update.

Specifically, we still see a matrix inverse for a matrix with the same dimension as the state.

To clean this up, we first note that

$$\hat{\boldsymbol{x}}_k = \left[\boldsymbol{I} - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\right]\bar{\boldsymbol{P}}_k\left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k\right] \tag{8.249}$$

$$\text{upd. cov.} \quad = \boldsymbol{P}_k\left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k\right] \tag{8.250}$$

$$\text{distribute} \quad = \boldsymbol{P}_k\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k \tag{8.251}$$

$$\text{upd. cov.} \quad = \left[\boldsymbol{I} - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\right]\bar{\boldsymbol{P}}_k\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k \tag{8.252}$$

$$\text{cancel} \quad = \left[\boldsymbol{I} - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\right]\bar{\boldsymbol{x}}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k \tag{8.253}$$

$$\text{distribute} \quad = \bar{\boldsymbol{x}}_k - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{x}}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k \tag{8.254}$$

The first two terms are ok, but what about that last term? We need another relationship...

Let's go back to the LUMVE covariance equation and manipulate that

$$\boldsymbol{P}_k^{-1} = \bar{\boldsymbol{P}}_k^{-1} + \widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\widetilde{\boldsymbol{H}}_k \tag{8.255}$$

1. Pre-multiply by $\boldsymbol{P}_k$

$$\boldsymbol{I} = \boldsymbol{P}_k\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\widetilde{\boldsymbol{H}}_k \tag{8.256}$$

2. Post-multiply by $\bar{\boldsymbol{P}}_k$

$$\bar{\boldsymbol{P}}_k = \boldsymbol{P}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k \tag{8.257}$$

3. Post-multiply by $\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}$

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} \tag{8.258}$$

4. Factor out $\boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}$ on the right-hand side

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\left[\boldsymbol{I} + \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\right] \tag{8.259}$$

5. Write $\boldsymbol{I}$ as $\boldsymbol{R}_k\boldsymbol{R}_k^{-1}$

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\left[\boldsymbol{R}_k\boldsymbol{R}_k^{-1} + \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\right] \tag{8.260}$$

6. Pull $\boldsymbol{R}_k^{-1}$ out of the bracketed term

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\left[\boldsymbol{R}_k + \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\right]\boldsymbol{R}_k^{-1} \tag{8.261}$$

7. Post-multiply by $\boldsymbol{R}_k$

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\left[\boldsymbol{R}_k + \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\right] \tag{8.262}$$

8. Solve for $\boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}$

$$\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\left[\boldsymbol{R}_k + \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\right]^{-1} = \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} \tag{8.263}$$

9. Rearrange

$$\boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\left[\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T + \boldsymbol{R}_k\right]^{-1} \tag{8.264}$$

10. Recall the definition of $\boldsymbol{K}_k$

$$\boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1} = \boldsymbol{K}_k \tag{8.265}$$

Now we have the expression that we need, so let's go back to the state update equation.

Previously, we found that

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{x}}_k + \boldsymbol{P}_k\widetilde{\boldsymbol{H}}_k^T\boldsymbol{R}_k^{-1}\boldsymbol{z}_k \tag{8.266}$$

Substitute our new equation into the last term to find

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{x}}_k + \boldsymbol{K}_k\boldsymbol{z}_k \tag{8.267}$$

We just need to rearrange some terms to get the updated estimate as

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{K}_k\left[\boldsymbol{z}_k - \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{x}}_k\right] \tag{8.268}$$

Note that the bracketed term is what we've previously called the residual; that is, it is the difference between our actual measurement and our prediction of the observation.

The residual is also known as the innovation, the new information being added.

The update takes the previous best estimate and adds a term that is a gain multiplied by the innovation.

To summarize...

- at time t_j, we have (by some means) a state estimate and its covariance

$$\hat{\boldsymbol{x}}_j \quad \text{and} \quad \boldsymbol{P}_j$$

- at time t_k, a new measurement, $\boldsymbol{z}_k$, is received; it is modeled as

$$\boldsymbol{z}_k = \widetilde{\boldsymbol{H}}_k \boldsymbol{x}_k + \boldsymbol{v}_k \tag{8.269}$$

  where $\boldsymbol{v}_k$ is zero mean, white noise, with covariance $\boldsymbol{R}_k$

- we propagate the estimate and covariance to time t_k

$$\bar{\boldsymbol{x}}_k = \boldsymbol{\Phi}(t_k, t_j)\hat{\boldsymbol{x}}_j \tag{8.270}$$

$$\bar{\boldsymbol{P}}_k = \boldsymbol{\Phi}(t_k, t_j)\boldsymbol{P}_j\boldsymbol{\Phi}^T(t_k, t_j) \tag{8.271}$$

- we then update the estimate to incorporate the new measurement

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{K}_k\left[z_k - \widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{x}}_k\right] \tag{8.272}$$

$$\boldsymbol{P}_k = \left[\boldsymbol{I} - \boldsymbol{K}_k\widetilde{\boldsymbol{H}}_k\right]\bar{\boldsymbol{P}}_k \tag{8.273}$$

  where

$$\boldsymbol{K}_k = \bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T\left[\widetilde{\boldsymbol{H}}_k\bar{\boldsymbol{P}}_k\widetilde{\boldsymbol{H}}_k^T + \boldsymbol{R}_k\right]^{-1} \tag{8.274}$$

- for any additional data, we cycle the propagation and update steps

This is the sequential version of LUMVE.

It only requires inverses of matrices with the same dimension as the measurement.

8.2.7.1 Example: Revisiting the Robot Problem Again

Let's revisit our robot tracking problem one more time.

This time, we want to apply the batch and sequential LUMVE approaches to show that they yield the same results.

The configuration is identical to the previous time we dealt with the robot problem.

We are still assuming that the robot moves in a plane under a constant velocity model, starting from the origin.

The measurements are of the position of the robot with noise taken from a Gaussian distribution. As time progresses,

the noise increases from a standard deviation of 0.1 meters to 0.4 meters.

The one difference is that we will make use of some prior information, which is given by

$$\bar{\boldsymbol{x}}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \bar{\boldsymbol{P}}_0 = \begin{bmatrix} 10^2 & 0 & 0 & 0 \\ 0 & 10^2 & 0 & 0 \\ 0 & 0 & 2^2 & 0 \\ 0 & 0 & 0 & 2^2 \end{bmatrix} \tag{8.275}$$

In code, we have the initial information

```
% initial information
xbar0 = [0.0;0.0;0.0;0.0];
Pbar0 = diag([10.0;10.0;2.0;2.0].^2);
```

We also have the initial true state and the continuous-time dynamics

```
% initial true state of the robot
x0 = [0;0;0.1;0.2];

% dynamics of the robot (continuous time)
F = [0,0,1,0;0,0,0,1;0,0,0,0;0,0,0,0];
```

We will simulate the truth for 10 sec at 10 Hz so that we can generate true data

```
% timing variables
t0 =  0.0;
dt =  0.1;
tf = 10.0;
tv = (t0:dt:tf)';
```

The truth is propagated using ODE45

```
% integrate the eoms for the true object
opts  = odeset('AbsTol',1e-9,'RelTol',1e-9);
[T,X] = ode45(@eom_robot,tv,x0,opts,F);
```

This allows us to create measurements of the position with a time-varying measurement noise

```
% create measurements of position
sig = linspace(0.1,0.4,length(tv));
z   = zeros(2,length(tv));
for k = 1:length(tv)
    z(:,k) = X(k,1:2)' + sig(k)*randn(2,1);
end
```

First, we will compute the estimate using LUMVE (batch) for the estimate at t_0.

```
% assemble the concatenated terms
Z = []; H = []; R = [];
for k = 1:length(tv)
    % time at kth observation
    tk = tv(k);
```

and

```
        % concatenate measurements
        Z = [Z;z(1,k);z(2,k)];

        % concatenate measurement-mapping matrices
        Htilde = [1,0,0,0;0,1,0,0];
        Phik0  = [1,0,tk-t0,0;0,1,0,tk-t0;0,0,1,0;0,0,0,1];
        H      = [H;Htilde*Phik0];

        % concatenate LUMVE inverse weighting matrix
        R = blkdiag(R,sig(k)^2*eye(2));
    end
    % LUMVE weighting matrix
    Ri = inv(R);

    % LUMVE estimate
    xhatB = (inv(Pbar0) + H'*Ri*H)\(Pbar0\xbar0 + H'*Ri*Z);
    PB    = inv(inv(Pbar0) + H'*Ri*H);
```

Then, so that we can compare to a sequential implementation, we propagate the estimate and the covariance to t_f:

```
    % map the LUMVE estimate to the final time
    Phif0  = [1,0,tf-t0,0;0,1,0,tf-t0;0,0,1,0;0,0,0,1];
    xhatBf = Phif0*xhatB;
    PBf    = Phif0*PB*Phif0';
```

Now, we apply the sequential form of LUMVE

```
    % redo the process using sequential LUMVE
    tj    = t0;
    xhatj = xbar0;
    Pj    = Pbar0;
    for k = 1:length(tv);
        % time at kth observation
        tk = tv(k);

        % kth observations
        zk = z(:,k);

        % state transition matrix from tj to tk
        Phikj = [1,0,tk-tj,0;0,1,0,tk-tj;0,0,1,0;0,0,0,1];

        % propagate estimate and covariance
        xbark = Phikj*xhatj;
        Pbark = Phikj*Pj*Phikj';
```

and

```
        % update estimate and covariance
        Htildek = [1,0,0,0;0,1,0,0];
        Rk      = sig(k)^2*eye(2);
        Kk      = Pbark*Htildek'/(Htildek*Pbark*Htildek' + Rk);
        xhatk   = xbark + Kk*(zk - Htildek*xbark);
        Pk      = Pbark - Kk*Htildek*Pbark;

        % reset for next step
        tj    = tk;
        xhatj = xhatk;
        Pj    = Pk;
    end
```

The following figures illustrate the application of the batch and sequential forms of LUMVE applied to this problem. We will look at the batch estimates mapped to the final time and plotted as constant in time. This is *only* for visualization as these estimates do not move in time.

The thing to notice is that the estimates and covariances from both forms of LUMVE converge to the same solutions after all of the data have been processed.

You may also notice that we've added a covariance contour onto the first two plots. This is to graphically represent our prior estimate. Because the covariance is quite large, which indicates low confidence in our initial estimate, the covariance contour associated with 0.1σ is shown.

How can you plot covariance contours?

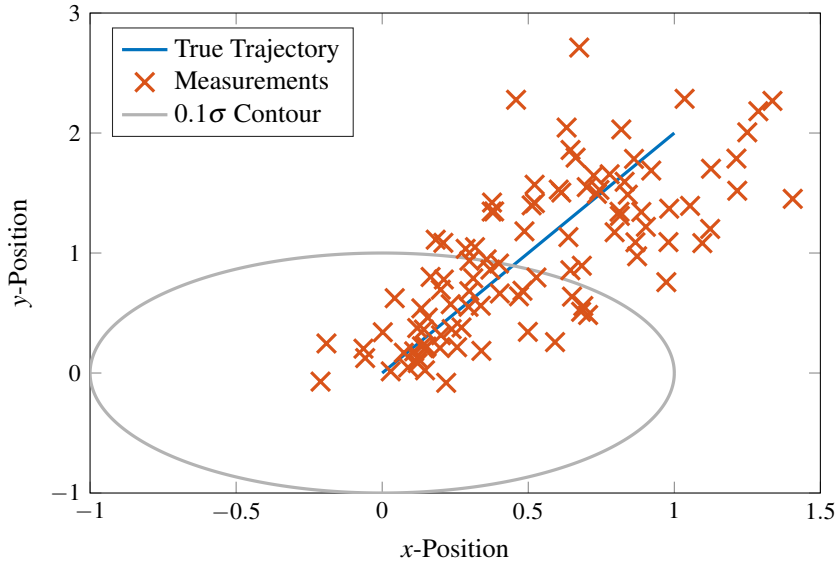
Assume that $\xi = \cos\theta$ and $\eta = \sin\theta$ for $0 \leq \theta \leq 2\pi$.

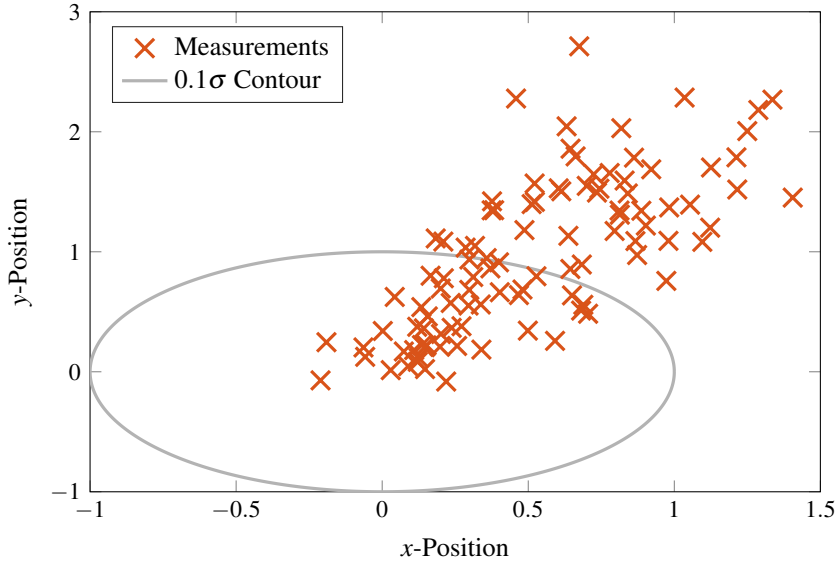
Given a mean, $\boldsymbol{m} \in \mathbb{R}^2$, and a covariance, $\boldsymbol{P} \in \mathbb{R}^{2\times 2}$, compute the Cholesky factor, $\boldsymbol{S}$, such that $\boldsymbol{P} = \boldsymbol{S}\boldsymbol{S}^T$.

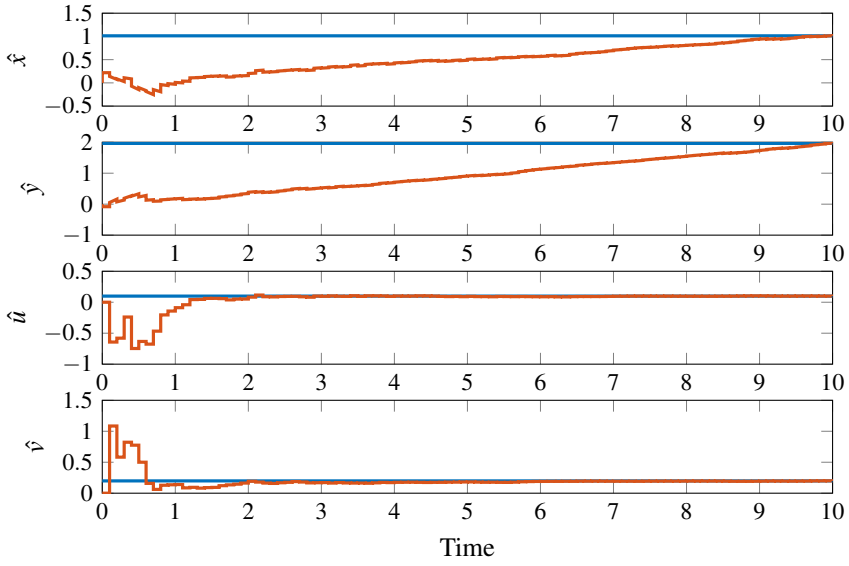
Then, the points on an s-σ level contour are found for each θ as

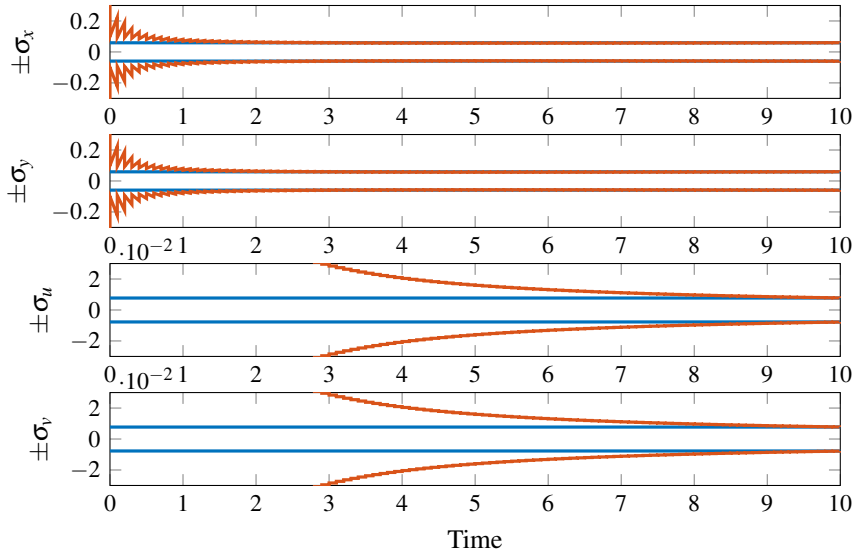
$$\begin{bmatrix} x \\ y \end{bmatrix} = \boldsymbol{m} + s\boldsymbol{S}\begin{bmatrix} \xi \\ \eta \end{bmatrix} \tag{8.276}$$

It is important to note that the mean and covariance only represent the first two central moments of the prior distribution even though we are illustrating them in the typical Gaussian manner.









Key point: Prior information is *required* for the sequential form of LUMVE, as we have currently posed it, but it is not required for the batch form of LUMVE.

If there is no prior information, what do we propagate? If we could propagate "zero information," what would the update look like?

What problems, if any, do we encounter in trying to process data against the propagated mean and covariance?

There is a way to overcome this deficiency in the sequential form of LUMVE that is called the *information filter*.

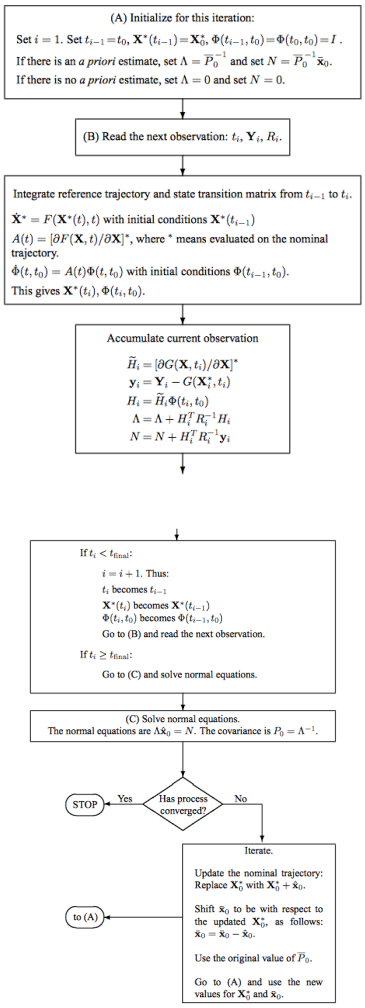
8.3 Non-linear Least Squares

What if the system is nonlinear? This is clearly the case for all orbits in the Cartesian space and for all orbits but the Keplerian one in orbital element space.

The answer is a bit anticlimactic, but perhaps expected: linearize and iterate.

The resulting method is an iterative application of what we called the batch processor.

The following schematic is taken from Tapley, Schutz, and Born. We will go further into the development of this method, but we will transition over to our notation.



You should be able to see LUMVE at work in the heart of the batch processor.

The LUMVE solution is in block (C) and the building blocks of LUMVE are two blocks below the (B) block.

You probably also see the linearization at work for the dynamics in the block below (B) and for the observations in the block that is two down from (B).

To reiterate the batch processor, it is given by

$$\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\left[\bar{\boldsymbol{P}}_k^{-1}\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{z}\right] \tag{8.277}$$

$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.278}$$

where $\hat{\boldsymbol{x}}_k$ is our state estimate and $\boldsymbol{P}_k$ is the covariance of the estimation error.

Recall that the index k is an arbitrary fixed time at which the solution is computed.

It is important to remember that everything up to this point is for *linear* systems, only.

So how do we handle the nonlinear case?

Nonlinearities can be present in the dynamics or the measurements, which means that we need to deal with both, ultimately.

Moreover, we also want a method that enables us to work with continuous-time dynamics.

Alright, so now we have a system described by the dynamics and measurements of the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) \tag{8.279}$$

$$\boldsymbol{z}_i = \boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{v}_i \tag{8.280}$$

where $\mathrm{E}\{\boldsymbol{v}_i\} = \boldsymbol{0}$ and $\mathrm{E}\{\boldsymbol{v}_i\boldsymbol{v}_j^T\} = \boldsymbol{R}_{ii}\delta_{ij}$.

We have assumed here that the noise is white. This will help establish a computational algorithm later on.

Since we have a linear method and a nonlinear system, we're going to linearize these equations.

Therefore, consider a reference described by

$$\dot{\boldsymbol{x}}^*(t) = \boldsymbol{f}(\boldsymbol{x}^*(t)) \tag{8.281}$$

$$\boldsymbol{z}_i^* = \boldsymbol{h}(\boldsymbol{x}_i^*) \tag{8.282}$$

Note that we compute the reference measurement *without* noise.

Now, we look at deviations away from the reference as

$$\delta\boldsymbol{x}(t) = \boldsymbol{x}(t) - \boldsymbol{x}^*(t) \tag{8.283}$$
$$\delta\boldsymbol{z}_i = \boldsymbol{z}_i - \boldsymbol{z}_i^* \tag{8.284}$$

Differentiating the first equation with respect to time and using our nonlinear dynamics and measurements yields

$$\delta\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) - \boldsymbol{f}(\boldsymbol{x}^*(t)) \tag{8.285}$$
$$\delta\boldsymbol{z}_i = \boldsymbol{h}(\boldsymbol{x}_i) - \boldsymbol{h}(\boldsymbol{x}_i^*) + \boldsymbol{v}_i \tag{8.286}$$

This is where we apply linearization via a first-order Taylor series expansion of the first terms about the reference state

$$\delta\dot{\boldsymbol{x}}(t) \approxeq \boldsymbol{f}(\boldsymbol{x}^*(t)) + \boldsymbol{F}(\boldsymbol{x}^*(t))\delta\boldsymbol{x}(t) - \boldsymbol{f}(\boldsymbol{x}^*(t)) \tag{8.287}$$
$$\delta\boldsymbol{z}_i \approxeq \boldsymbol{h}(\boldsymbol{x}_i^*) + \widetilde{\boldsymbol{H}}(\boldsymbol{x}_i^*)\delta\boldsymbol{x}_i - \boldsymbol{h}(\boldsymbol{x}_i^*) + \boldsymbol{v}_i \tag{8.288}$$

We can eliminate the similar terms to find

$$\delta\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{x}^*(t))\delta\boldsymbol{x}(t) \tag{8.289}$$
$$\delta\boldsymbol{z}_i = \widetilde{\boldsymbol{H}}(\boldsymbol{x}_i^*)\delta\boldsymbol{x}_i + \boldsymbol{v}_i \tag{8.290}$$

One more step: we need to convert the continuous time evolution of the deviation into a discrete time evolution.

How? We will use the state transition matrix to replace the differential equation, which gives us the system

$$\delta\boldsymbol{x}_i = \boldsymbol{\Phi}(t_i, t_k)\delta\boldsymbol{x}_k \tag{8.291}$$
$$\delta\boldsymbol{z}_i = \widetilde{\boldsymbol{H}}(\boldsymbol{x}_i^*)\delta\boldsymbol{x}_i + \boldsymbol{v}_i \tag{8.292}$$

where, from the properties of the state transition matrix, we know that $\boldsymbol{\Phi}(t_i, t_k)$ is found by integrating the matrix differential equation

$$\dot{\boldsymbol{\Phi}}(t, t_k) = \boldsymbol{F}(\boldsymbol{x}^*(t))\boldsymbol{\Phi}(t, t_k) \tag{8.293}$$

from $t = t_k$ to $t = t_i$ with the initial condition $\boldsymbol{\Phi}(t_k, t_k) = \boldsymbol{I}$.

The "deviation" equations are now in the same form that we started with in the development of the LUMVE solution, except that we've assumed that the measurement noise is white.

We don't have to make that assumption, but it's a pretty standard assumption, and it leads to a more efficient solution.

The important element to note is that we are dealing with *deviations* now.

That is, when we apply the LUMVE solution to our deviations, we need to use measurements of the form

$$\delta\boldsymbol{z}_i = \boldsymbol{z}_i - \boldsymbol{z}_i^* \tag{8.294}$$
$$= \boldsymbol{z}_i - \boldsymbol{h}(\boldsymbol{x}_i^*) \tag{8.295}$$

where $\boldsymbol{z}_i$ are the *actual* measurements.

Then, the LUMVE solution is given by

$$\delta\hat{\boldsymbol{x}}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\left[\bar{\boldsymbol{P}}_k^{-1}\delta\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T\boldsymbol{R}^{-1}\delta\boldsymbol{z}\right] \tag{8.296}$$

$$\boldsymbol{P}_k = \left[\bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1} \tag{8.297}$$

Because we are working in deviations, the prior estimate in the form of a deviation from the reference, i.e.

$$\delta\bar{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k - \boldsymbol{x}_k^* \tag{8.298}$$

However, we will provide $\delta\bar{\boldsymbol{x}}_k$ as our prior estimate and not $\bar{\boldsymbol{x}}_k$.

Moreover, from the definition of the state deviation, the update is of the form

$$\delta\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k^* \tag{8.299}$$

Using the update obtained from LUMVE, the estimate of the full state is given by

$$\hat{\boldsymbol{x}}_k = \boldsymbol{x}_k^* + \delta\hat{\boldsymbol{x}}_k \tag{8.300}$$

We still need the $\boldsymbol{R}$ matrix and the $\boldsymbol{H}$ matrix. These are still concatenations of the individual terms, where the elements of $\boldsymbol{R}$ don't change, but now the $\boldsymbol{H}$ matrix is assembled using the reference state to compute the Jacobian.

The concatenation of terms, therefore, looks like

$$\delta\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 - \boldsymbol{h}(\boldsymbol{x}_1^*) \\ \boldsymbol{z}_2 - \boldsymbol{h}(\boldsymbol{x}_2^*) \\ \vdots \end{bmatrix} \qquad \boldsymbol{H} = \begin{bmatrix} \widetilde{\boldsymbol{H}}(\boldsymbol{x}_1^*)\boldsymbol{\Phi}(t_1,t_k) \\ \widetilde{\boldsymbol{H}}(\boldsymbol{x}_2^*)\boldsymbol{\Phi}(t_2,t_k) \\ \vdots \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{R}_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{8.301}$$

Let's develop a more computationally friendly version of LUMVE that exploits the white-noise property.

First, we define a few terms:

$$\boldsymbol{\lambda} = \bar{\boldsymbol{P}}_k^{-1}\delta\bar{\boldsymbol{x}}_k + \boldsymbol{H}^T\boldsymbol{R}^{-1}\delta\boldsymbol{z} \tag{8.302}$$

$$\boldsymbol{\Lambda} = \bar{\boldsymbol{P}}_k^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} \tag{8.303}$$

Now, we can write the estimate as the solution to the normal equations

$$\boldsymbol{\Lambda}\delta\hat{\boldsymbol{x}}_k = \boldsymbol{\lambda} \tag{8.304}$$

and the covariance is given by

$$\boldsymbol{P}_k = \boldsymbol{\Lambda}^{-1} \tag{8.305}$$

We will now use the assumption that the noise is time-wise uncorrelated, such that

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{R}_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{8.306}$$

As we previously pointed out in the sequential form of LUMVE, the inverse of the block diagonal covariance matrix is just the block diagonal matrix formed from the individual inverses, or

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{R}_{11}^{-1} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{R}_{22}^{-1} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{8.307}$$

When we have a time-wise uncorrelated noise, it is straightforward to see that

$$\boldsymbol{H}^T\boldsymbol{R}^{-1}\delta\boldsymbol{z} = \sum_{\ell=1}^{m}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right]^T\boldsymbol{R}_{\ell\ell}^{-1}\delta\boldsymbol{z}_\ell \tag{8.308}$$

$$\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} = \sum_{\ell=1}^{m}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right]^T\boldsymbol{R}_{\ell\ell}^{-1}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right] \tag{8.309}$$

This allows us to compute $\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$ through accumulation by summation as

$$\boldsymbol{\lambda} = \bar{\boldsymbol{P}}_k^{-1}\delta\bar{\boldsymbol{x}}_k + \sum_{\ell=1}^{m}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right]^T\boldsymbol{R}_{\ell\ell}^{-1}\delta\boldsymbol{z}_\ell \tag{8.310}$$

$$\boldsymbol{\Lambda} = \bar{\boldsymbol{P}}_k^{-1} + \sum_{\ell=1}^{m}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right]^T\boldsymbol{R}_{\ell\ell}^{-1}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_k)\right] \tag{8.311}$$

Then, our LUMVE solution is

$$\delta\hat{\boldsymbol{x}}_k = \boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda} \tag{8.312}$$

$$\boldsymbol{P}_k = \boldsymbol{\Lambda}^{-1} \tag{8.313}$$

Since we have a nonlinear system, we want to iterate. We do this by resetting some of our starting values: the reference state, the prior estimate (as an estimated deviation), and the prior covariance.

Note that these are the only values we have starting off.

Let $(\boldsymbol{x}_k^*)_{n-1}$ represent the reference state of iteration $n-1$.

The reference state is reset using the estimated value of the deviation:

$$(\boldsymbol{x}_k^*)_n = (\boldsymbol{x}_k^*)_{n-1} + \delta\hat{\boldsymbol{x}}_k \tag{8.314}$$

This is the same as saying that our reference state of iteration n is equal to the full estimated state after iteration $n-1$ is completed.

We must also reset the prior information, and we must be careful not to change the prior information.

Therefore, if the prior estimate and covariance at iteration $n-1$ are given by

$$(\delta\bar{\boldsymbol{x}}_k)_{n-1} \qquad \text{and} \qquad (\bar{\boldsymbol{P}}_k)_{n-1} \tag{8.315}$$

then the prior estimate and covariance at iteration n are

$$(\delta\bar{\boldsymbol{x}}_k)_n = (\delta\bar{\boldsymbol{x}}_k)_{n-1} - \delta\hat{\boldsymbol{x}}_k \tag{8.316}$$
$$(\bar{\boldsymbol{P}}_k)_n = (\bar{\boldsymbol{P}}_k)_{n-1} \tag{8.317}$$

Why? To preserve the prior information, we need to ensure that the full estimated state between iterations remains constant. That is

$$(\boldsymbol{x}_k^*)_n + (\delta\bar{\boldsymbol{x}}_k)_n = (\boldsymbol{x}_k^*)_{n-1} + (\delta\bar{\boldsymbol{x}}_k)_{n-1} \tag{8.318}$$

The left-hand side is the full estimated state at iteration n, and the right-hand side is the full estimated state at iteration $n-1$.

We know that the reference state for iteration n is updated from iteration $n-1$ by

$$(\boldsymbol{x}_k^*)_n = (\boldsymbol{x}_k^*)_{n-1} + \delta\hat{\boldsymbol{x}}_k \tag{8.319}$$

Therefore, the condition for preserving the full estimated state becomes

$$(\boldsymbol{x}_k^*)_{n-1} + \delta\hat{\boldsymbol{x}}_k + (\delta\bar{\boldsymbol{x}}_k)_n = (\boldsymbol{x}_k^*)_{n-1} + (\delta\bar{\boldsymbol{x}}_k)_{n-1} \tag{8.320}$$

We can now cancel like terms and solve for the estimate (deviation) at iteration n that preserves the full state estimate from iteration $n-1$ to iteration n. This gives us the equation that we started with for the prior estimate update, i.e.

$$(\delta\bar{\boldsymbol{x}}_k)_n = (\delta\bar{\boldsymbol{x}}_k)_{n-1} - \delta\hat{\boldsymbol{x}}_k \tag{8.321}$$

We also need to preserve the uncertainty information. This is readily accomplished by starting iteration n with the same covariance used to start iteration $n-1$.

Let's enumerate the steps for the iterative batch processor algorithm (remember that the index k in the LUMVE solution is arbitrary, so we will use $k=0$ for convenience in this listing):

1. Input a set of data described by values of $\boldsymbol{z}_i$ at times t_i with associated measurement noise covariances $\boldsymbol{R}_{ii}$, where $t_i \geq t_0 \; \forall \; i \in \{1,2,\ldots,m\}$.

2. Begin with a reference state $\boldsymbol{x}_0^*$, a prior estimate $\delta\bar{\boldsymbol{x}}_0$, and a prior covariance $\bar{\boldsymbol{P}}_0$, all at time t_0.

3. Set $n=1$, and initialize an iteration loop

   (a) Set $\ell = 1$, and initialize an accumulation loop with

   $$t_{\ell-1} = t_0 \qquad \boldsymbol{x}^*(t_{\ell-1}) = \boldsymbol{x}_0^* \qquad \boldsymbol{\Phi}(t_{\ell-1}, t_0) = \boldsymbol{I}$$
   $$\boldsymbol{\lambda} = \bar{\boldsymbol{P}}_0^{-1}\delta\bar{\boldsymbol{x}}_0 \qquad \boldsymbol{\Lambda} = \bar{\boldsymbol{P}}_0^{-1}$$

   Note that $\boldsymbol{\lambda} = \boldsymbol{0}$ and $\boldsymbol{\Lambda} = \boldsymbol{0}$ if there is no prior estimate.

   i. Parse the ℓ^{th} measurement to get t_ℓ, $\boldsymbol{z}_\ell$, and $\boldsymbol{R}_{\ell\ell}$.

ii. Integrate the reference trajectory and state transition matrix from $t_{\ell-1}$ to t_ℓ

$$\begin{aligned} \dot{\boldsymbol{x}}^*(t) &= \boldsymbol{f}(\boldsymbol{x}^*(t)) && \text{s.t.i.c.} \quad \boldsymbol{x}^*(t_{\ell-1}) \\ \dot{\boldsymbol{\Phi}}(t,t_0) &= \boldsymbol{F}(\boldsymbol{x}^*(t))\boldsymbol{\Phi}(t,t_0) && \text{s.t.i.c.} \quad \boldsymbol{\Phi}(t_{\ell-1},t_0) \end{aligned}$$

This gives $\boldsymbol{x}_\ell^* = \boldsymbol{x}^*(t_\ell)$ and $\boldsymbol{\Phi}(t_\ell,t_0)$.

iii. Accumulate the current observation

$$\boldsymbol{\lambda} = \boldsymbol{\lambda} + \left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_0)\right]^T \boldsymbol{R}_{\ell\ell}^{-1}\left[\boldsymbol{z}_\ell - \boldsymbol{h}(\boldsymbol{x}_\ell^*)\right] \tag{8.322}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda} + \left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_0)\right]^T \boldsymbol{R}_{\ell\ell}^{-1}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{x}_\ell^*)\boldsymbol{\Phi}(t_\ell,t_0)\right] \tag{8.323}$$

iv. If $\ell < m$, set $\ell = \ell + 1$ and return to Step 3(a)i with $\boldsymbol{x}^*(t_{\ell-1}) = \boldsymbol{x}^*(t_\ell)$ and $\boldsymbol{\Phi}(t_{\ell-1},t_0) = \boldsymbol{\Phi}(t_\ell,t_0)$. Otherwise, exit the accumulation loop.

(b) Solve the normal equations to find the estimate and compute the covariance:

$$\boldsymbol{\Lambda}\delta\hat{\boldsymbol{x}}_0 = \boldsymbol{\lambda} \qquad \boldsymbol{P}_0 = \boldsymbol{\Lambda}^{-1}$$

(c) If the process has converged, exit the iteration loop. Otherwise, set $n = n + 1$ and return to Step 3a with

$$\boldsymbol{x}_0^* = \boldsymbol{x}_0^* + \delta\hat{\boldsymbol{x}}_0 \tag{8.324}$$

$$\delta\bar{\boldsymbol{x}}_0 = \delta\bar{\boldsymbol{x}}_0 - \delta\hat{\boldsymbol{x}}_0 \tag{8.325}$$

$$\bar{\boldsymbol{P}}_0 = \bar{\boldsymbol{P}}_0 \tag{8.326}$$

4. Output the converged reference trajectory $\boldsymbol{x}_0^*$ and the covariance $\boldsymbol{P}_0$.

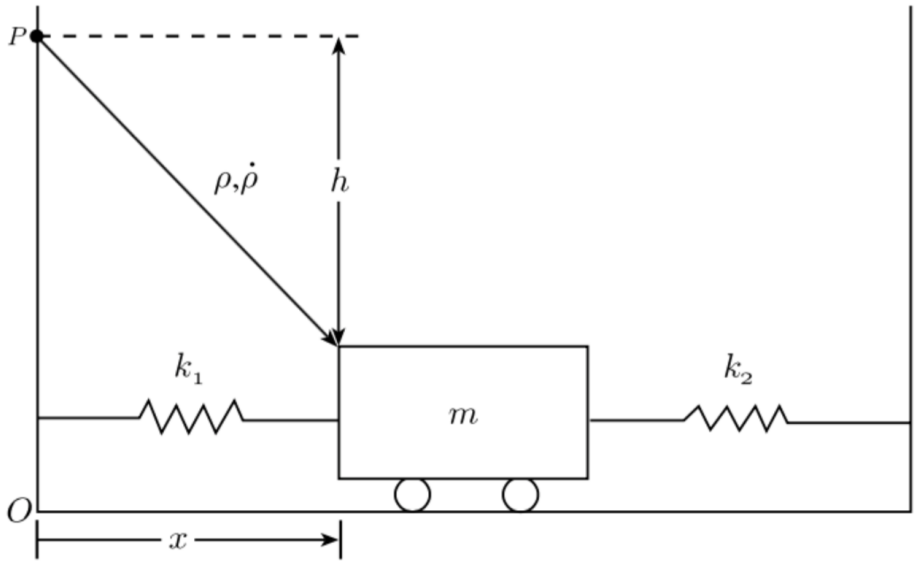
The previous steps should correlate ($\rho > 0.8$) with the schematic that we looked at earlier from Tapley, Schutz, and Born.

The main difference is in notation, and we have also explicitly added an iteration loop where theirs is a bit implied.

However, the process is the same: linearize and iterate. Iterate until you converge to a solution and your updates to the reference state become small.

8.3.0.1 A Spring-Mass Problem

Let's take a look at a problem from Tapley, Schutz, and Born and apply the iterative batch processor to see if we can replicate the results that they give.



A block of mass m is attached to two parallel vertical walls by two springs. The spring constants are k_1 and k_2.

An observer is placed at a height h on the left wall (at position P). This observer measures the range ρ and the range-rate $\dot{\rho}$ of the mass (taken to be a point mass).

If the horizontal distance of the mass from the left wall is denoted by x, the objective is to use the range and range-rate information to estimate the position x and the velocity $\dot{x}$.

Let's put together all of the pieces that we need to use the iterative batch processor.

First, we'll consider all of the dynamics-related quantities.

The equation of motion governing the position of the mass is

$$\ddot{x} = -\frac{k_1+k_2}{m}x \tag{8.327}$$

If we define states to be x and $v = \dot{x}$, and define $\omega^2 = (k_1+k_2)/m$, then the first-order form of the dynamics is

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ -\omega^2 x \end{bmatrix} \tag{8.328}$$

This is the nonlinear system form of the dynamics, i.e. $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$, but these dynamics are linear, so we can also write them as

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) \qquad \text{where} \qquad \boldsymbol{F}(t) = \boldsymbol{F} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \tag{8.329}$$

If you take the partials of the "nonlinear" system, you will find that the Jacobian matrix is state-independent and that it is identical to the linear system dynamics matrix.

We also need the state transition matrix to complete our description of the dynamical system.

We can compute this by numerically integrating the dynamics of the state transition matrix or, since this is a linear time-invariant system, we can also find the state transition matrix using the matrix exponential.

The matrix exponential option is certainly viable here, but it is not quite as clean as the cases we had before since $\boldsymbol{F}^2 \neq \boldsymbol{0}$.

However, we have a system that is a harmonic oscillator, so we would expect the solution to the second-order system to be of the form

$$x(t) = A\cos\omega t + B\sin\omega t \tag{8.330}$$

This also gives us a velocity solution by differentiation as

$$v(t) = B\omega\cos\omega t - A\omega\sin\omega t \tag{8.331}$$

If the mass is at position x_0 and velocity v_0 at time $t_0 = 0$, then the coefficients are found to be

$$A = x_0 \qquad \text{and} \qquad B = v_0/\omega \tag{8.332}$$

and the solution is

$$\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} \cos\omega(t-t_0) & \frac{1}{\omega}\sin\omega(t-t_0) \\ -\omega\sin\omega(t-t_0) & \cos\omega(t-t_0) \end{bmatrix} \begin{bmatrix} x_0 \\ v_0 \end{bmatrix} \tag{8.333}$$

Therefore, the state transition matrix can be determined exactly as

$$\boldsymbol{\Phi}(t,t_0) = \begin{bmatrix} \cos\omega(t-t_0) & \frac{1}{\omega}\sin\omega(t-t_0) \\ -\omega\sin\omega(t-t_0) & \cos\omega(t-t_0) \end{bmatrix} \tag{8.334}$$

This is everything related to the dynamics that we need to determine a batch LUMVE solution.

So, we now turn towards the measurement-related information.

In particular, we need a nonlinear function representing the measurements, and we need a measurement Jacobian for the linearization of the nonlinear function.

For this problem, we are considering range and range-rate measurements from an observer that is located at the point P (refer back to the diagram). These are given as functions of the position and velocity of the mass, as well as the height of the observer, as

$$\rho = \sqrt{x^2 + h^2} \tag{8.335}$$

$$\dot{\rho} = \frac{xv}{\sqrt{x^2 + h^2}} \tag{8.336}$$

The range-rate equation can be confirmed by differentiating the range equation with respect to time.

Putting these two equations together, we have a nonlinear function of the state describing the measurements as

$$\boldsymbol{h}(\boldsymbol{x}) = \begin{bmatrix} \sqrt{x^2+h^2} \\ \dfrac{xv}{\sqrt{x^2+h^2}} \end{bmatrix} \tag{8.337}$$

Remember, we're modeling the data as

$$\boldsymbol{z}_i = \boldsymbol{h}(\boldsymbol{x}_i) + \boldsymbol{v}_i \tag{8.338}$$

where $\mathrm{E}\{\boldsymbol{v}_i\} = \boldsymbol{0}$ and $\mathrm{E}\{\boldsymbol{v}_i \boldsymbol{v}_i^T\} = \boldsymbol{R}_{ii}$.

The only thing left in order to be able to determine the batch LUMVE solution is the measurement Jacobian, which is defined as

$$\widetilde{\boldsymbol{H}}(\boldsymbol{x}) = \left[\frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right] \tag{8.339}$$

Taking the derivatives, it follows that

$$\widetilde{\boldsymbol{H}}(\boldsymbol{x}) = \begin{bmatrix} \dfrac{x}{\sqrt{x^2+h^2}} & 0 \\ \dfrac{v}{\sqrt{x^2+h^2}} - \dfrac{x^2 v}{(\sqrt{x^2+h^2})^3} & \dfrac{x}{\sqrt{x^2+h^2}} \end{bmatrix} \tag{8.340}$$

We now have all of the pieces to put together a batch LUMVE solution; moreover, since we have a nonlinear measurement function, we are going to apply the batch LUMVE solution iteratively.

We just need numbers!

Let's take the mass, spring constants, observer altitude, and true position and velocity of the mass to be

$$\begin{array}{lll} m = 1.5 \text{ kg} & k_1 = 2.5 \text{ N/m} & k_2 = 3.7 \text{ N/m} \\ h = 5.4 \text{ m} & x_0 = 3.0 \text{ m} & v_0 = 0.0 \text{ m/s} \end{array}$$

We will also use an initial reference, a prior estimate, and a prior covariance of

$$\boldsymbol{x}_0^* = \begin{bmatrix} 4.0 \\ 0.2 \end{bmatrix} \qquad \delta\bar{\boldsymbol{x}}_0 = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \qquad \bar{\boldsymbol{P}}_0 = \begin{bmatrix} 1000 & 0 \\ 0 & 100 \end{bmatrix}$$

The range and range-rate data, along with the times at which they are acquired, are shown in the following table.

Time [s]	Range [m]	Range-Rate [m/s]
0.00	6.17737808459220	0.000000000000000
1.00	5.56327661282686	1.312858634955140
2.00	5.69420161397342	-1.544881143816120
3.00	6.15294262127432	0.534923988815733
4.00	5.46251322092491	0.884698415328368
5.00	5.83638064328625	-1.561232489180540
6.00	6.08236452736002	1.009799431575470
7.00	5.40737619817037	0.317051170392150
8.00	5.97065615746125	-1.374530709756060
9.00	5.97369258835895	1.367681694432360
10.00	5.40669060248179	-0.302111588503166

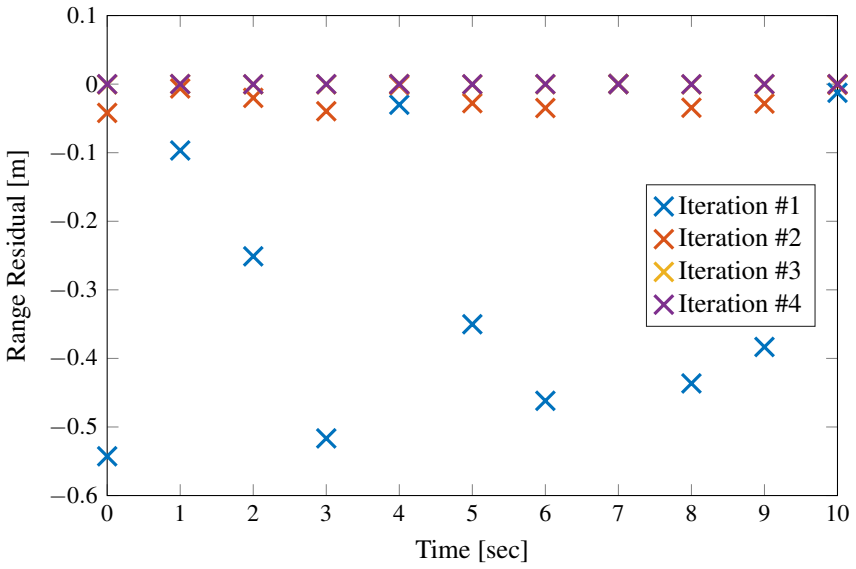
Finally, the measurement noise covariance is taken to be $\boldsymbol{R}_{ii} = \boldsymbol{I}$.

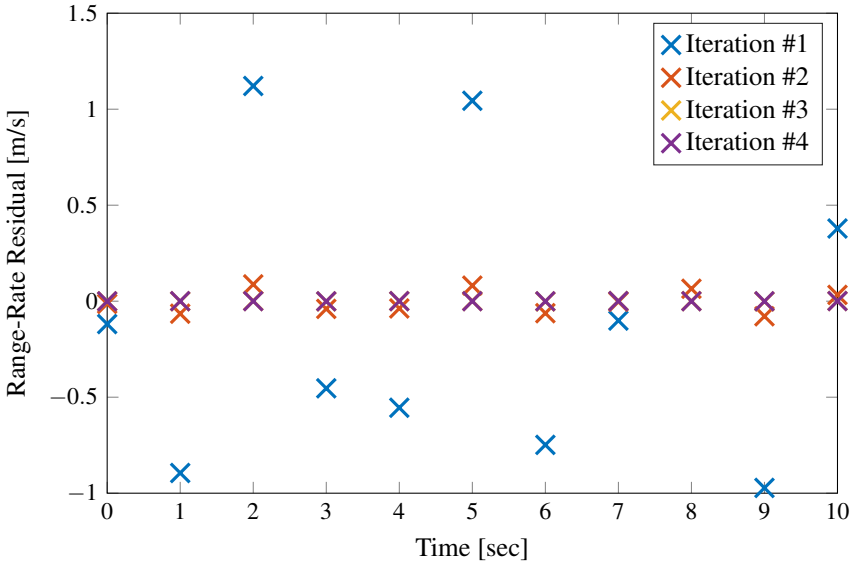
We will apply four iterations to this data. Usually, you would continue until some stopping condition, but you can also apply a fixed number of iterations.

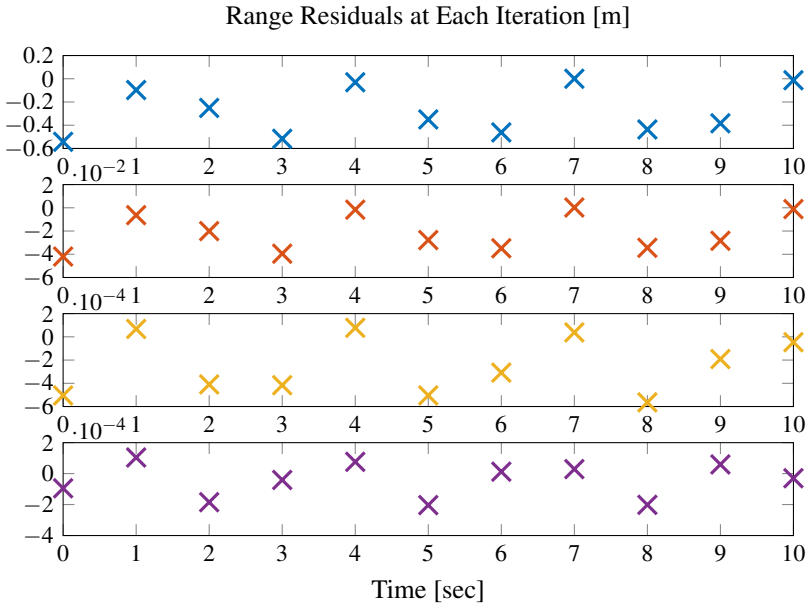
For each iteration, we will plot the range and range-rate residuals, which are given by

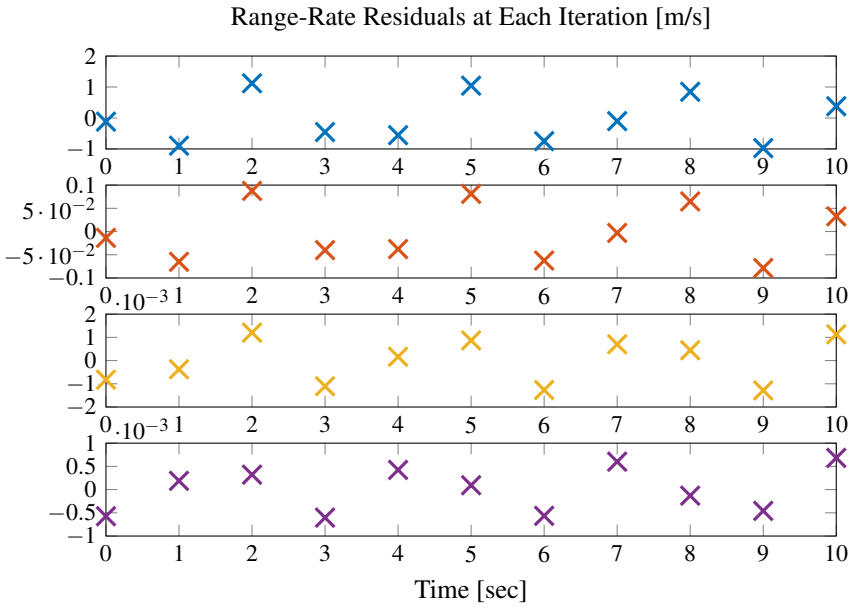
$$\delta \boldsymbol{z}_i = \boldsymbol{z}_i - \boldsymbol{h}(\boldsymbol{x}_i^*) \tag{8.341}$$

For the fourth iteration, we will then analyze statistics of the residuals and provide the output solution from the iterative batch processor.




Iteration #1
Iteration #2
Iteration #3
Iteration #4
Range-Rate Residual [m/s]
Time [sec]


Range Residuals at Each Iteration [m]
Time [sec]



The mean, root-mean-square, and standard deviation of the range residuals at the fourth iteration are, respectively,

$$\mathrm{E}\{\delta\rho\} = -4.3000 \times 10^{-5} \text{ m} \tag{8.342}$$

$$\sqrt{\mathrm{E}\{\delta\rho^2\}} = 1.1628 \times 10^{-4} \text{ m} \tag{8.343}$$

$$\sqrt{\mathrm{E}\{(\delta\rho - \mathrm{E}\{\delta\rho\})^2\}} = 1.0804 \times 10^{-4} \text{ m} \tag{8.344}$$

The mean, root-mean-square, and standard deviation of the range-rate residuals at the fourth iteration are, respectively,

$$\mathrm{E}\{\delta\dot{\rho}\} = -1.7577 \times 10^{-6} \text{ m/s} \tag{8.345}$$

$$\sqrt{\mathrm{E}\{\delta\dot{\rho}^2\}} = 4.6661 \times 10^{-4} \text{ m/s} \tag{8.346}$$

$$\sqrt{\mathrm{E}\{(\delta\dot{\rho} - \mathrm{E}\{\delta\dot{\rho}\})^2\}} = 4.6661 \times 10^{-4} \text{ m/s} \tag{8.347}$$

The reference state (now our estimated full state) and the covariance after the fourth iteration are

$$\boldsymbol{x}_0^* = \begin{bmatrix} 3.0002 \text{ m} \\ 1.1818 \times 10^{-3} \text{ m/s} \end{bmatrix} \tag{8.348}$$

$$\boldsymbol{P}_0 = \begin{bmatrix} 1.6935 \times 10^{-1} \text{ m}^2 & 1.2775 \times 10^{-2} \text{ m}^2/\text{s} \\ 1.2775 \times 10^{-2} \text{ m}^2/\text{s} & 5.8448 \times 10^{-1} \text{ m}^2/\text{s}^2 \end{bmatrix} \tag{8.349}$$

We can also represent the information in the covariance matrix by the standard deviation of the position, the standard deviation of the velocity, and the correlation, which, after the fourth iteration, are

$$\sigma_{x_0} = 4.1152 \times 10^{-1} \text{ m} \tag{8.350}$$

$$\sigma_{v_0} = 7.6451 \times 10^{-1} \text{ m/s} \tag{8.351}$$

$$\rho_{x_0 v_0} = 4.0607 \times 10^{-2} \tag{8.352}$$

8.3.1 Example of IOD and Batch First Orbit Improvement

We're going to take a look at an example of applying Gauss' method for initial orbit determination and the iterative batch processor for orbit improvement.

This example will assume that the data have already been generated and will instead focus on the arrangement of the IOD and batch processor elements.

Since we've already generated the data, here's a short listing of what we have available moving forward.

```
% DATA PROVIDED ARE:
%   Tm = (1 x n) array of observation times [sec]
%   Zm = (2 x n) array of right-ascension and declination observations [rad]
%   Rm = (2 x 2 x n) array of measurement noise covariances [arcsec^2]
%   Xt = (6 x n) array of true object position and velocity [km] and [km/s]
%   Rt = (3 x n) array of observer position in inertial frame [km]
```

To apply IOD via Gauss' method, we need to select three observations. For simplicity, we'll just use the first, "middle," and last measurements of right-ascension and declination, and extract the data that we need.

```
% ------------------------------------------------------------------------
% Perform initial orbit determination via Gauss' method
% extract three measurements (time, RA, DEC, inertial position of station)
% we're using the first, "middle", and last measurements for IOD
iIOD = [1;90;181];
tIOD = Tm(iIOD);
aIOD = Zm(1,iIOD)*rad2deg; % [rad] -> [deg]
dIOD = Zm(2,iIOD)*rad2deg; % [rad] -> [deg]
```

We can now apply Gauss' method to generate an IOD solution at the middle time.

```
% apply Gauss' method
```

Everything for Gauss' method is embedded in the previous function call, but we're simply using the standard approach outlined in the notes with the f and g series (not Gibbs' method) to determine the velocity at the middle time.

We might ask ourselves how well Gauss' method performed.

```
% compute pos/vel err at t2
rerr_gauss_t2 = norm(x2(1:3) - Xt(1:3,iIOD(2)));
```

As it turns out the position error is just 48.672 km and the velocity error is 8.756 m/s. Not bad!

To interface with the iterative refinement, however, we want a reference state at the first measurement time, so we can directly map the output from Gauss' method back to the time of the first measurement.

```
% propagate Gauss solution at t2 back to t1
opt     = odeset('AbsTol',1e-9,'RelTol',1e-9);
[~,XX] = ode45(@eom_car,[t2,tIOD(1)],x2,opt,GM);
t1      = tIOD(1);
```

And again, we can check on the quality of our solution, which shows us that the position error at the time of the first measurement is 50.006 km and the velocity error is 9.058 m/s. Note that the errors are slightly worse here, but that's not surprising at all.

```
% compute pos/vel err at t1
rerr_gauss_t1 = norm(x1(1:3) - Xt(1:3,iIOD(1)));
verr_gauss_t1 = norm(x1(4:6) - Xt(4:6,iIOD(1)));
% End of initial orbit determination via Gauss' method
```

So now, we're done with IOD. This has allowed us to generate an initial guess (and a pretty good one) for the orbit. We will now use that guess as our initial reference in the application of the iterative batch processor to refine our orbit solution (hopefully) and to provide us with an estimate of the uncertainty as well.

The first part of the batch processor is to set our reference time, reference state, and provide any prior information (estimated deviation and covariance). Our reference time and state come from IOD, but we have no prior information ($\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$ will both be zero in a little bit).

```
% ---------------------------------------------------------------------
% Perform iterative improvement via batch least squares (LUMVE)
% set the epoch time, reference state, initial deviation, and covariance
t0    = t1;
x0ref = x1;
x0    = zeros(6,1);
```

Now, we really should iterate until we converge, but we'll keep it simple and just apply a sequence of four iterations here.

```
% for simplicity, we'll do a fixed number of iterations
```

At this point, we're ready to begin the iteration loop, and this is where we set that prior information to zero.

```
% begin the iteration loop
for loop = 1:iter
    % no prior information, so Lambda and lambda are both zero
    Lam  = zeros(6,6);
```

We're going to have to propagate the reference state and the state transition matrix, so let's go ahead and initialize those along with a timing variable that will help us cycle through the observation times correctly. Note that the state transition matrix is an identity matrix at the beginning of *each* iteration because it maps everything back to the reference time.

```
% initialize a time variable, the reference state, and the STM
tkm1 = t0;
xref = x0ref;
```

Let's also declare some storage for the residuals so that we can analyze the performance of the batch processor.

```
% declare storage for the residual and the time of the residual
rest = zeros(length(Tm),1);
```

We're inside of the iteration loop, but now we need to move inside of a time loop so that we can accumulate all of our data. When we start the time loop, we're going to go ahead and extract the time, measurement, measurement noise covariance, and the station position (in the inertial frame) for the k^{th} observation.

```
resm = zeros(length(Tm),2);
for k = 1:length(Tm)
    % extract the time, measurement, covariance, and station position
    % for the kth observation
    tk = Tm(k);
    zk = Zm(:,k)*rad2asc;
    Rk = Rm(:,:,k);
```

We have the measurement noise covariance, but LUMVE uses the inverse, so we can go ahead and compute that, too.

```
    % determine the LUMVE weighting matrix
```

The next step is to actually propagate our reference state and our state transition matrix. We'll handle this using numerical integration, as it's the most general method. This just applies ODE45 to equations of motion governing the evolution of the reference and the state transition matrix; here, we're just using two-body dynamics.

```
    % propagate the reference state and STM, but only if we're past t0
    if(tk > t0)
        opts    = odeset('AbsTol',1e-9,'RelTol',1e-9);
        [~,XX] = ode45(@eom_tbp_ref,[tkm1,tk],[xref;Phi(:)],opts,GM);
        xref    = XX(end,1:6)';
        Phi     = reshape(XX(end,7:end)',6,6);
```

We need a "reference" measurement. This is just the measurement we would compute from the reference trajectory, which means that we take our reference trajectory's position, subtract the station position, and convert that to right-ascension and declination. We'll process the data in arcseconds, so make sure to convert that over!

```
% compute the reference state measurement (RA and DEC)
%   form the relative position
%   change the units of the reference measurement to [arcsec]
rosi  = xref(1:3) - rk;
x     = rosi(1);
y     = rosi(2);
z     = rosi(3);
wsq   = x*x + y*y;
w     = sqrt(wsq);
rhosq = wsq + z*z;
zref  = [atan2(y,x); atan2(z,w)];
```

It's time to compute the measurement Jacobian $\widetilde{\boldsymbol{H}}$. This follows directly from the definition of the right-ascension and declination, but don't forget to take into account units here, too.

```
% compute the measurement mapping matrix (Htilde)
%   change the units of the Jacobian to [arcsec]
Ht = [       -y/wsq,          x/wsq,     0.0, 0.0, 0.0, 0.0;
      -x*z/(w*rhosq), -y*z/(w*rhosq), w/rhosq, 0.0, 0.0, 0.0];
```

Now, multiply the measurement Jacobian by the state transition matrix to get the time-mapped measurement Jacobian.

```
% time-mapping of the observation matrix
```

Finally! The LUMVE accumulation step, where we add in the contribution of the k^{th} observation to $\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$. Remember, we're dealing with *linearized* systems, so the actual measurement processed by LUMVE is the deviation away from the reference measurement.

```
% accumulate the lambdas for LUMVE
lam = lam + H'*Ri*(zk - zref);
```

We can store the measurement residual (actual minus reference) and the time associated with the residual so that we can plot them later on.

```
% store the time and the residual for plotting later
rest(k)   = tk;
```

We need to reset our timing variable so that the next trip through the time loop will be reference accordingly. This allows to integrate sequentially through time instead of integrating further and further with each step of the time loop.

```
% reset the timing variable
tkm1 = tk;
```

And, of course, we can't forget to actually compute the least squares solution.

```
end
% get the least squares solution
```

To end the all-critical iteration loop, we also need to reset our reference state by adding the least squares solution, and, to preserve the prior information contained in the estimated deviation, we need to subtract the same thing from the estimated deviation.

```
% perform the iteration by shifting the reference and the estimated deviation
x0ref = x0ref + delx;
```

If desired, we can plot some residuals.

```
% plot the measurement residuals for right-ascension on one plot
figure(1)
C = get(gca,'colororder');
plot(rest,resm(:,1),'x','Color',C(loop,:),'LineWidth',1.2,'MarkerSize',5)
```

This can be done for the right-ascension residuals or the declination residuals and can be plotted on a single plot or single plots per iteration (this is usually more useful to look at).

Now, we end the iteration loop.

```
plot(rest,resm(:,2),'x','Color',C(loop,:),'LineWidth',1.2,'MarkerSize',5)
```

After all of our iterations are complete, the estimated state is given by the reference state from LUMVE, and the covariance is given by the inverse of the $\mathbf{\Lambda}$ matrix.

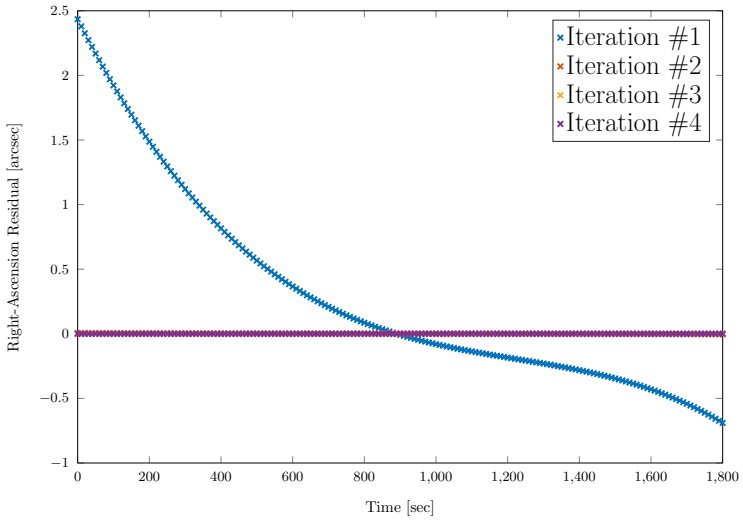
```
% determine the estimated state and covariance
xhat = x0ref;
```

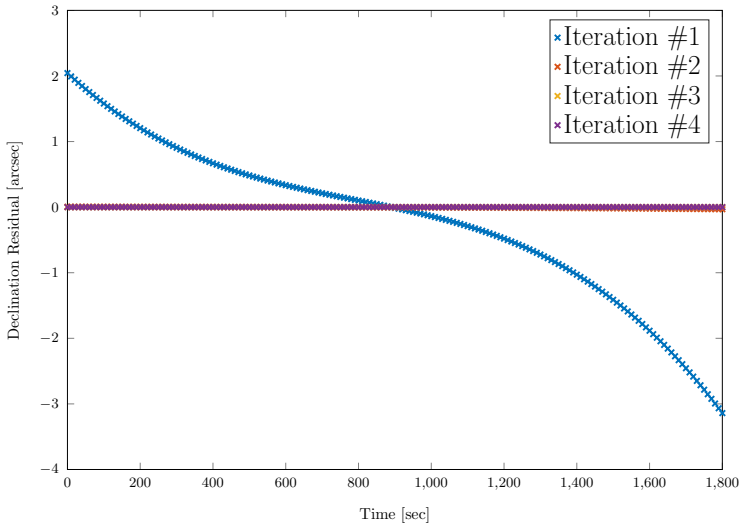
We should ask how well we did (again, and for the last time in this example).

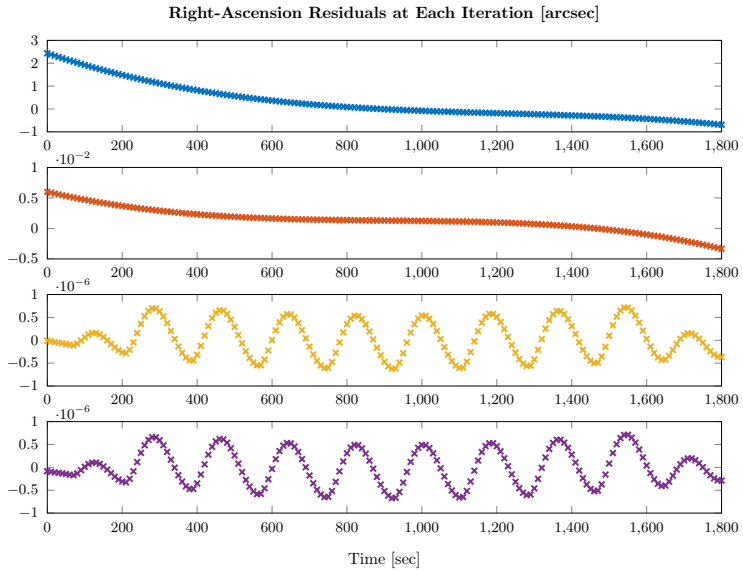
```
% compute pos/vel err at t1
rerr_LSQ_t1 = norm(xhat(1:3) - Xt(1:3,iIOD(1)));
verr_LSQ_t1 = norm(xhat(4:6) - Xt(4:6,iIOD(1)));
% End of iterative improvement via batch least squares (LUMVE)
```

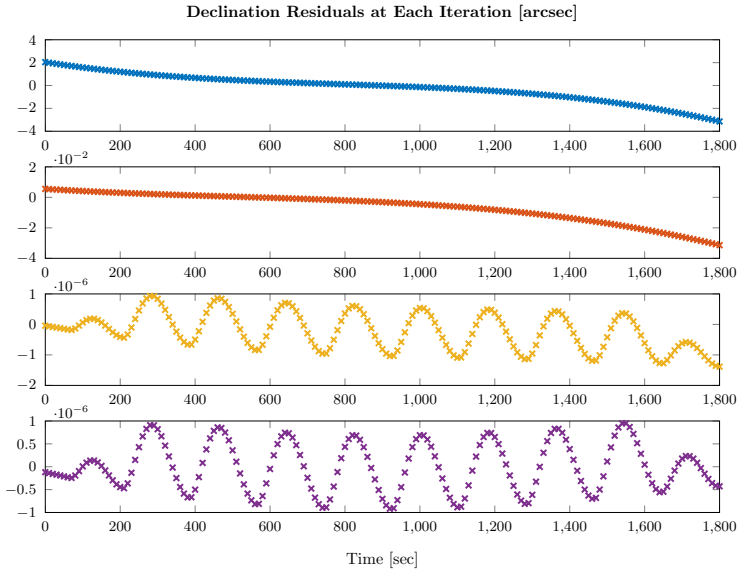
The position error is 5.298 mm and the velocity error is 1.073 μm/s. Well, that's not too bad at all...

What is important, however, is that we have now embedded all of the information from the 181 measurements into our estimate of the state.


Iteration #1
Iteration #2
Iteration #3
Iteration #4
Right-Ascension Residual [arcsec]
Time [sec]


Iteration #1
Iteration #2
Iteration #3
Iteration #4
Declination Residual [arcsec]
Time [sec]


Right-Ascension Residuals at Each Iteration [arcsec]
Time [sec]


Declination Residuals at Each Iteration [arcsec]
4
2
0
−2
−4
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
·10−2
2
0
−2
−4
·10−6
1
0
−1
−2
·10−6
1
0.5
0
−0.5
−1
Time [sec]

Chapter 9

Orbit Improvement/Filtering: Minimum Mean Square Error Estimation

Once we have a relatively good orbit (first orbit determination and first orbit improvement) it is advantageous to switch to a sequential filter that can incorporate new measurements as soon as they come in and does not force to fit all observations at the same time. The errors in the orbit propagation can built up, hence longer observation time spans cannot be fitted. In principal, several options are available to handle this case: a) *delete* old observations (automatically or by hand) and use a standard least squares approach, such as the LUMVE, or LUMVE with the special observation incorporated b) use the sequential version of the LUMVE (not discussed in class but in the notes), or c) use a (extended/unscented) Kalman filter for for minimum mean square error (MMSE) estimation.

That is, the standard process is to

1. Use IOD on a set of data to produce a reference position and velocity.
2. Reprocess the data (and maybe more) with the iterative batch method to obtain an improved estimate and a covariance.
3. Process subsequent data using a Kalman filter starting from the batch estimate and covariance.

The following is a paraphrasing of the preface to Bill Lear's unpublished book on Kalman filtering: Kalman Filtering Techniques.

> A Kalman filter is a computer algorithm that is used to process error corrupted measurement data. The purpose of the processing is to better determine the parameters or variables associated with the process that generates the measurements. For example, a radar station tracking the Space Shuttle makes range, azimuth and elevation angle measurements. Using a Kalman filter to process these measurements, one can determine the position, velocity and acceleration of the Shuttle, and also determine what bias errors are adding to the measurements.
>
> Kalman filters are useful in modern navigation problems, particularly those problems requiring instantaneous real-time solutions. [For instance, a Kalman filter navigation algorithm] processed the Earth-based Doppler data of the Lunar Module (LM) as it descended and ascended from the surface of the moon. Based on this program, a real-time navigation position correction was voice-linked to the astronauts as they descended to the surface of the moon. This enabled pinpoint landing accuracy.
>
> Kalman filters are useful in many areas other than navigation. They can be used to determine irregularities in the Earth's gravity field. They can be used to determine the density of the atmosphere from altitude measurements of a falling sphere. They can be used to analyze the stock market (don't get your hopes up, they don't predict well). They can process calibration measurements to better determine the state of a chemical process.
>
> Their are many variations of Kalman filters to which various people attach their names. But that is all they are, variations. The man who started it all is Rudy Kalman.

Just a note from my site, his full name is Rudolf Emil Kálmán.

9.1 The Kalman Filter (Linear Dynamics)

The system dynamics are given by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t) \tag{9.1}$$

where

$$\mathrm{E}\{\boldsymbol{w}(t)\} = \boldsymbol{0} \quad \text{and} \quad \mathrm{E}\left\{\boldsymbol{w}(t)\boldsymbol{w}^T(\tau)\right\} = \boldsymbol{Q}_s(t)\delta(t-\tau) \tag{9.2}$$

We also assume that the initial state has mean, $\boldsymbol{m}(t_0) = \boldsymbol{m}_0$, and covariance $\boldsymbol{P}(t_0) = \boldsymbol{P}_0$.

The state of the system is denoted by $\boldsymbol{x}(t)$, and $\boldsymbol{F}(t)$ are the dynamics of the state.

The term $\boldsymbol{w}(t)$ is a white-noise process providing stochastic excitation to the deterministic dynamics. This is *roughly* the corollary to the measurement noise that we have previously dealt with, but it is now a continuous-time process.

White noise is physically unrealizable as it is characterized by constant power across all frequencies, but it is a useful mathematical model of stochastic perturbations. The power spectral density, $\boldsymbol{Q}_s(t)$, is actually constant for white-noise processes, but there is nothing that will stop us from assuming it is time-varying in the following developments.

Finally, $\boldsymbol{M}(t)$ is a shape matrix that maps the noise into the dynamics.

We will proceed by developing time-wise evolutionary equations for the mean and covariance of the state.

The mean of the state as a function of time is given by

$$\boldsymbol{m}(t) = \mathrm{E}\{\boldsymbol{x}(t)\} \tag{9.3}$$

Taking the time rate of change and interchanging the order of differentiation and expectation yields

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\dot{\boldsymbol{x}}(t)\} \tag{9.4}$$

Applying the system dynamics within the expectation, it follows that

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\boldsymbol{F}(t)\boldsymbol{x}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.5}$$
$$= \mathrm{E}\{\boldsymbol{F}(t)\boldsymbol{x}(t)\} + \mathrm{E}\{\boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.6}$$

From the fact that $\boldsymbol{F}(t)$ and $\boldsymbol{M}(t)$ are deterministic and recalling that the process noise is taken to be zero-mean, the mean satisfies

$$\dot{\boldsymbol{m}}(t) = \boldsymbol{F}(t)\boldsymbol{m}(t) \tag{9.7}$$

This is our forward evolution equation for the mean; we now turn to developing a similar equation for the covariance.

Define the error to be the difference of the truth from the mean, i.e.

$$\boldsymbol{e}(t) = \boldsymbol{x}(t) - \boldsymbol{m}(t) \tag{9.8}$$

which gives the error dynamics as

$$\dot{\boldsymbol{e}}(t) = \dot{\boldsymbol{x}}(t) - \dot{\boldsymbol{m}}(t) \tag{9.9}$$

Substitute for the true dynamics of the state and the dynamics of the mean to get

$$\dot{\boldsymbol{e}}(t) = \dot{\boldsymbol{x}}(t) - \dot{\boldsymbol{m}}(t) \tag{9.10}$$
$$= \big[\boldsymbol{F}(t)\boldsymbol{x}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t)\big] - \big[\boldsymbol{F}(t)\boldsymbol{m}(t)\big] \tag{9.11}$$
$$= \boldsymbol{F}(t)\big[\boldsymbol{x}(t) - \boldsymbol{m}(t)\big] + \boldsymbol{M}(t)\boldsymbol{w}(t) \tag{9.12}$$
$$= \boldsymbol{F}(t)\boldsymbol{e}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t) \tag{9.13}$$

The solution of the linear differential equation for the error is

$$\boldsymbol{e}(t) = \boldsymbol{\Phi}(t, t_{k-1})\boldsymbol{e}(t_{k-1}) + \int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)d\tau \tag{9.14}$$

where $\boldsymbol{\Phi}(t, t_{k-1})$ is the state transition matrix which satisfies

$$\dot{\boldsymbol{\Phi}}(t, t_{k-1}) = \boldsymbol{F}(t)\boldsymbol{\Phi}(t, t_{k-1}), \qquad \boldsymbol{\Phi}(t_{k-1}, t_{k-1}) = \boldsymbol{I} \tag{9.15}$$

It is important to note that the time t_{k-1} is purely arbitrary. That is, t_{k-1} can represent any starting condition off of which the evolution of the error is based.

The state estimation error covariance is found via

$$\boldsymbol{P}(t) = \mathrm{E}\left\{\boldsymbol{e}(t)\boldsymbol{e}^T(t)\right\} \tag{9.16}$$

By forming the product of $\boldsymbol{e}(t)$ with its transpose, we find

$$\begin{aligned}
\boldsymbol{e}(t)\boldsymbol{e}^T(t) &= \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) && (9.17)\\
&+ \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\int_{t_{k-1}}^{t}\boldsymbol{w}^T(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau && (9.18)\\
&+ \left[\int_{t_{k-1}}^{t}\boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)d\tau\right]\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) && (9.19)\\
&+ \int_{t_{k-1}}^{t}\boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)d\tau\int_{t_{k-1}}^{t}\boldsymbol{w}^T(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau && (9.20)
\end{aligned}$$

For now, we are going to focus on this product. Later, we will take its expected value.

Now, let's pull everything inside of the integral in the middle terms and relabel the second integral in the final term

$$\begin{aligned}
\boldsymbol{e}(t)\boldsymbol{e}^T(t) &= \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) && (9.21)\\
&+ \int_{t_{k-1}}^{t}\boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\boldsymbol{w}^T(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau && (9.22)\\
&+ \int_{t_{k-1}}^{t}\boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1})d\tau && (9.23)\\
&+ \int_{t_{k-1}}^{t}\boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)d\tau\int_{t_{k-1}}^{t}\boldsymbol{w}^T(\sigma)\boldsymbol{M}^T(\sigma)\boldsymbol{\Phi}^T(t,\sigma)d\sigma && (9.24)
\end{aligned}$$

The dummy variable of time, τ, that is used in the first three integrals can also be represented by σ, as we have done in the last integral since it is simply a dummy variable of time.

Rewrite the final term as a double integral

$$\boldsymbol{e}(t)\boldsymbol{e}^T(t) = \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) \tag{9.25}$$

$$+\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{e}(t_{k-1})\boldsymbol{w}^T(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.26}$$

$$+\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)\boldsymbol{e}^T(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1})d\tau \tag{9.27}$$

$$+\int_{t_{k-1}}^{t}\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)\boldsymbol{w}^T(\sigma)\boldsymbol{M}^T(\sigma)\boldsymbol{\Phi}^T(t,\sigma)d\sigma d\tau \tag{9.28}$$

At this point, we are ready to take the expected value of the product, which distributes as an expected value of all four terms, keeping in mind that $\boldsymbol{\Phi}(\cdot,\cdot)$ and $\boldsymbol{M}(\cdot)$ are deterministic

$$\boldsymbol{P}(t) = \mathrm{E}\left\{\boldsymbol{e}(t)\boldsymbol{e}^T(t)\right\} \tag{9.29}$$

$$= \boldsymbol{\Phi}(t,t_{k-1})\mathrm{E}\left\{\boldsymbol{e}(t_{k-1})\boldsymbol{e}^T(t_{k-1})\right\}\boldsymbol{\Phi}^T(t,t_{k-1}) \tag{9.30}$$

$$+\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,t_{k-1})\mathrm{E}\left\{\boldsymbol{e}(t_{k-1})\boldsymbol{w}^T(\tau)\right\}\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.31}$$

$$+\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\mathrm{E}\left\{\boldsymbol{w}(\tau)\boldsymbol{e}^T(t_{k-1})\right\}\boldsymbol{\Phi}^T(t,t_{k-1})d\tau \tag{9.32}$$

$$+\int_{t_{k-1}}^{t}\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\mathrm{E}\left\{\boldsymbol{w}(\tau)\boldsymbol{w}^T(\sigma)\right\}\boldsymbol{M}^T(\sigma)\boldsymbol{\Phi}^T(t,\sigma)d\sigma d\tau \tag{9.33}$$

From the covariance definition

$$\boldsymbol{P}(t) = \mathrm{E}\left\{\boldsymbol{e}(t)\boldsymbol{e}^T(t)\right\} \tag{9.34}$$

it follows that, by setting $t = t_{k-1}$,

$$\boldsymbol{P}(t_{k-1}) = \mathrm{E}\left\{\boldsymbol{e}(t_{k-1})\boldsymbol{e}^T(t_{k-1})\right\} \tag{9.35}$$

Assume that the process noise is uncorrelated to the state at time t_{k-1}, such that

$$\mathrm{E}\left\{\boldsymbol{e}(t_{k-1})\boldsymbol{w}^T(\tau)\right\} = \boldsymbol{0} \tag{9.36}$$

From the definition of the process noise power spectral density

$$\mathrm{E}\left\{\boldsymbol{w}(\tau)\boldsymbol{w}^T(\sigma)\right\} = \boldsymbol{Q}_s(\tau)\delta(\tau-\sigma) \tag{9.37}$$

Applying the previous three relationships to the covariance equation gives

$$\boldsymbol{P}(t) = \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) \tag{9.38}$$

$$+ \int_{t_{k-1}}^{t}\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\delta(\tau-\sigma)\boldsymbol{M}^T(\sigma)\boldsymbol{\Phi}^T(t,\sigma)d\sigma d\tau \tag{9.39}$$

$$= \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) \tag{9.40}$$

$$+ \int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\left[\int_{t_{k-1}}^{t} \boldsymbol{M}^T(\sigma)\boldsymbol{\Phi}^T(t,\sigma)\delta(\tau-\sigma)d\sigma\right]d\tau \tag{9.41}$$

Finally, by applying the sifting property of the Dirac delta to the inner integral of the second term,

$$\boldsymbol{P}(t) = \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) \tag{9.42}$$

$$+ \int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.43}$$

The second term is what we call the process noise covariance matrix, i.e.

$$\boldsymbol{P}(t) = \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{Q}_c(t) \tag{9.44}$$

where

$$\boldsymbol{Q}_c(t) = \int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.45}$$

Consider temporal differentiation of $\boldsymbol{Q}_c(t)$ via

$$\dot{\boldsymbol{Q}}_c(t) = \frac{d}{dt}\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.46}$$

Since the upper limit of integration is a function of time (it is t itself), we must apply Leibniz' rule in order to take the derivative of the integral. This gives us

$$\dot{\boldsymbol{Q}}_c(t) = \int_{t_{k-1}}^{t} \frac{d}{dt}\left\{\boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)\right\}d\tau \tag{9.47}$$

$$+ \boldsymbol{\Phi}(t,t)\boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t)\boldsymbol{\Phi}^T(t,t) \tag{9.48}$$

Recall the properties of the state transition matrix:

$$\dot{\boldsymbol{\Phi}}(t,\tau) = \boldsymbol{F}(t)\boldsymbol{\Phi}(t,\tau) \qquad \text{and} \qquad \boldsymbol{\Phi}(t,t) = \boldsymbol{I} \tag{9.49}$$

Applying the above properties

$$\dot{\boldsymbol{Q}}_c(t) = \boldsymbol{F}(t)\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau \tag{9.50}$$

$$+\int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{Q}_s(\tau)\boldsymbol{M}^T(\tau)\boldsymbol{\Phi}^T(t,\tau)d\tau\boldsymbol{F}^T(t) \tag{9.51}$$

$$+\boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.52}$$

Finally, using the definition of $\boldsymbol{Q}_c(t)$ to substitute for the two integral terms, it is found that the process noise covariance matrix satisfies

$$\dot{\boldsymbol{Q}}_c(t) = \boldsymbol{F}(t)\boldsymbol{Q}_c(t) + \boldsymbol{Q}_c(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.53}$$

with the initial condition of $\boldsymbol{Q}_c(t_{k-1}) = \boldsymbol{0}$.

This is one method for propagating the covariance matrix.

Another method comes from differentiating $\boldsymbol{P}(t)$ with respect to time

$$\dot{\boldsymbol{P}}(t) = \frac{d}{dt}\left\{\boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{Q}_c(t)\right\} \tag{9.54}$$

Carrying out the differentiation, we find

$$\dot{\boldsymbol{P}}(t) = \dot{\boldsymbol{\Phi}}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\dot{\boldsymbol{\Phi}}^T(t,t_{k-1}) + \dot{\boldsymbol{Q}}_c(t) \tag{9.55}$$

where $\boldsymbol{P}(t_{k-1})$ is a fixed initial condition.

Applying the properties of the state transition matrix and the equation derived for $\dot{\boldsymbol{Q}}_c(t)$, it follows that

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(t)\boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1})\boldsymbol{F}^T(t) \tag{9.56}$$

$$+ \boldsymbol{F}(t)\boldsymbol{Q}_c(t) + \boldsymbol{Q}_c(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.57}$$

We can collect all terms that are pre-multiplied by $\boldsymbol{F}(t)$, and we can collect terms that are post-multiplied by $\boldsymbol{F}^T$; this give us

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(t)\left[\boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{Q}_c(t)\right] \tag{9.58}$$

$$+ \left[\boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{Q}_c(t)\right]\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.59}$$

The terms in square brackets are simply the covariance at time t, such that the covariance satisfies

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(t)\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.60}$$

with an initial condition of $\boldsymbol{P}(t_{k-1}) = \boldsymbol{P}_{k-1}$.

Now, we have two methods for propagating our covariance.

First method for covariance propagation:

- Propagate state transition matrix

$$\dot{\boldsymbol{\Phi}}(t,t_{k-1}) = \boldsymbol{F}(t)\boldsymbol{\Phi}(t,t_{k-1}), \quad \boldsymbol{\Phi}(t_{k-1},t_{k-1}) = \boldsymbol{I} \tag{9.61}$$

- Propagate process noise covariance matrix

$$\dot{\boldsymbol{Q}}_c(t) = \boldsymbol{F}(t)\boldsymbol{Q}_c(t) + \boldsymbol{Q}_c(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t), \quad \boldsymbol{Q}_c(t_{k-1}) = \boldsymbol{0} \tag{9.62}$$

- Calculate the propagated covariance matrix

$$\boldsymbol{P}(t) = \boldsymbol{\Phi}(t,t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t,t_{k-1}) + \boldsymbol{Q}_c(t) \tag{9.63}$$

Second method for covariance propagation:

- Propagate the covariance matrix

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(t)\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t), \quad \boldsymbol{P}(t_{k-1}) = \boldsymbol{P}_{k-1} \tag{9.64}$$

In either case, we begin with initial conditions on the mean and covariance from the previous update, i.e.

$$\boldsymbol{m}(t_{k-1}) = \boldsymbol{m}_{k-1}^{+} \qquad \text{and} \qquad \boldsymbol{P}(t_{k-1}) = \boldsymbol{P}_{k-1}^{+} \tag{9.65}$$

Then, we propagate our equations for the mean and covariance from $t = t_{k-1}$, when the previous update was made, to the time of the next measurement, $t = t_k$. At this point, the propagated mean and covariance are now called the *a priori* mean and covariance, and are given by

$$\boldsymbol{m}_k^{-} = \boldsymbol{m}(t_k) \qquad \text{and} \qquad \boldsymbol{P}_k^{-} = \boldsymbol{P}(t_k) \tag{9.66}$$

At time t_k a measurement is made available, which is given by $\boldsymbol{z}_k$. This measurement is a function of the state and is imperfect (noisy).

This measurement is taken to be of the form

$$\boldsymbol{z}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.67}$$

where

$$\mathrm{E}\{\boldsymbol{v}_k\} = \boldsymbol{0} \qquad \text{and} \qquad \mathrm{E}\{\boldsymbol{v}_k \boldsymbol{v}_\ell^T\} = \boldsymbol{R}_k \delta_{k\ell} \tag{9.68}$$

The measurement noise is represented by $\boldsymbol{v}_k$, which is assumed to be a zero mean white-noise sequence with covariance $\boldsymbol{R}_k$.

The mean and covariance prior to incorporation of this new information are given by

$$\boldsymbol{m}_k^- = \mathrm{E}\{\boldsymbol{x}_k\} \tag{9.69}$$

$$\boldsymbol{P}_k^- = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \tag{9.70}$$

We want to find a way to use this new information to *update* the mean and covariance of our state, to update our estimated state and our confidence in the estimated state.

Assume that the *a posteriori* mean is given by a linear combination of the *a priori* mean and the new measurement data via

$$\boldsymbol{m}_k^+ = \boldsymbol{N}_k \boldsymbol{m}_k^- + \boldsymbol{K}_k \boldsymbol{z}_k \tag{9.71}$$

Define the *a priori* and *a posteriori* estimation errors as

$$\boldsymbol{e}_k^- = \boldsymbol{x}_k - \boldsymbol{m}_k^- \quad \text{and} \quad \boldsymbol{e}_k^+ = \boldsymbol{x}_k - \boldsymbol{m}_k^+ \tag{9.72}$$

The linear update equation can then be written as

$$\boldsymbol{x}_k - \boldsymbol{e}_k^+ = \boldsymbol{N}_k \boldsymbol{x}_k - \boldsymbol{N}_k \boldsymbol{e}_k^- + \boldsymbol{K}_k \boldsymbol{z}_k \tag{9.73}$$

Now, we solve for the posterior error and substitute for the measurement model to find

$$\boldsymbol{e}_k^+ = \boldsymbol{x}_k - \boldsymbol{N}_k\boldsymbol{x}_k + \boldsymbol{N}_k\boldsymbol{e}_k^- - \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{x}_k - \boldsymbol{K}_k\boldsymbol{L}_k\boldsymbol{v}_k \tag{9.74}$$

$$= \big[\boldsymbol{I} - \boldsymbol{N}_k - \boldsymbol{K}_k\boldsymbol{H}_k\big]\boldsymbol{x}_k + \boldsymbol{N}_k\boldsymbol{e}_k^- - \boldsymbol{K}_k\boldsymbol{L}_k\boldsymbol{v}_k \tag{9.75}$$

If we take the expected value of the preceding relationship, we have

$$\mathrm{E}\{\boldsymbol{e}_k^+\} = \big[\boldsymbol{I} - \boldsymbol{N}_k - \boldsymbol{K}_k\boldsymbol{H}_k\big]\mathrm{E}\{\boldsymbol{x}_k\} + \boldsymbol{N}_k\mathrm{E}\{\boldsymbol{e}_k^-\} - \boldsymbol{K}_k\boldsymbol{L}_k\mathrm{E}\{\boldsymbol{v}_k\} \tag{9.76}$$

Provided that the prior estimate is unbiased and that the measurement noise is unbiased, it follows that the condition for an unbiased estimator is

$$\boldsymbol{0} = \big[\boldsymbol{I} - \boldsymbol{N}_k - \boldsymbol{K}_k\boldsymbol{H}_k\big]\mathrm{E}\{\boldsymbol{x}_k\} \tag{9.77}$$

This must hold regardless of the value of $\mathrm{E}\{\boldsymbol{x}_k\}$, so we can conclude that the matrix in brackets must be the zero matrix, or, after solving for $\boldsymbol{N}_k$, we find that

$$\boldsymbol{N}_k = \boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k \tag{9.78}$$

Substituting this result into our equation for the linear update, it follows that the posterior mean is

$$\boldsymbol{m}_k^+ = \big[\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k\big]\boldsymbol{m}_k^- + \boldsymbol{K}_k\boldsymbol{z}_k \tag{9.79}$$

$$= \boldsymbol{m}_k^- + \boldsymbol{K}_k\big[\boldsymbol{z}_k - \boldsymbol{H}_k\boldsymbol{m}_k^-\big] \tag{9.80}$$

This is the equation for updating the mean, but it required us to make use of the fact that the measurement is linear with respect to the state.

Is there another approach that does not require us to make this assumption?

Let's take inspiration from our work on LUMVE and consider an update of the form

$$\boldsymbol{m}_k^+ = \boldsymbol{a}_k + \boldsymbol{K}_k\boldsymbol{z}_k \tag{9.81}$$

This form effectively replaces our linear function of the prior mean with a constant vector, but still keeps a portion of the update as being a linear function of the data.

Now, recall the definitions of the prior and posterior error as

$$\boldsymbol{e}_k^- = \boldsymbol{x}_k - \boldsymbol{m}_k^- \quad \text{and} \quad \boldsymbol{e}_k^+ = \boldsymbol{x}_k - \boldsymbol{m}_k^+ \tag{9.82}$$

From these equations, it is straightforward to show that

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{e}_k^- - \boldsymbol{e}_k^+ \tag{9.83}$$

which allows us to write the new update equation as

$$\boldsymbol{m}_k^- + \boldsymbol{e}_k^- - \boldsymbol{e}_k^+ = \boldsymbol{a}_k + \boldsymbol{K}_k \boldsymbol{z}_k \tag{9.84}$$

If we take the expected value of this equation under the assumption of an unbiased prior and enforce the condition that we want an unbiased posterior, it follows that

$$\boldsymbol{m}_k^- = \boldsymbol{a}_k + \boldsymbol{K}_k \hat{\boldsymbol{z}}_k \tag{9.85}$$

where

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{z}_k\} \tag{9.86}$$

is the expected value of the measurement with respect to any stochastic inputs.

Now, we can solve for $\boldsymbol{a}_k$ such that we guarantee an unbiased posterior estimate; this yields

$$\boldsymbol{a}_k = \boldsymbol{m}_k^- - \boldsymbol{K}_k \hat{\boldsymbol{z}}_k \tag{9.87}$$

Our update equation can now be expressed as

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- - \boldsymbol{K}_k\hat{\boldsymbol{z}}_k + \boldsymbol{K}_k\boldsymbol{z}_k \tag{9.88}$$
$$= \boldsymbol{m}_k^- + \boldsymbol{K}_k\left[\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k\right] \tag{9.89}$$

Note that no specification of linearity of the measurement process needs to be made for this equation to hold. $\hat{\boldsymbol{z}}_k$ is simply the mean of the measurement with respect to the state and noise distributions.

However, if we make the specification that the measurement is linear, then

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k\} = \boldsymbol{H}_k\boldsymbol{m}_k^- \tag{9.90}$$

and the update collapses back to the first expression that we had, which is

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k\left[\boldsymbol{z}_k - \boldsymbol{H}_k\boldsymbol{m}_k^-\right] \tag{9.91}$$

In the future, we will make use of the more general form of the update that did not require the linear measurement condition.

With this update, we now want to examine the characteristics of the estimation error, where the goal is to determine the posterior estimation error covariance.

The *a posteriori* state estimation error is

$$\boldsymbol{e}_k^+ = \boldsymbol{e}_k^- - \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.92}$$

or, by substituting for the definitions of the errors,

$$(\boldsymbol{x}_k - \boldsymbol{m}_k^+) = (\boldsymbol{x}_k - \boldsymbol{m}_k^-) - \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.93}$$

Let $\boldsymbol{P}_k^-$ and $\boldsymbol{P}_k^+$ be defined as

$$\boldsymbol{P}_k^- = \mathrm{E}\left\{(\boldsymbol{e}_k^-)(\boldsymbol{e}_k^-)^T\right\} \quad \text{and} \quad \boldsymbol{P}_k^+ = \mathrm{E}\left\{(\boldsymbol{e}_k^+)(\boldsymbol{e}_k^+)^T\right\} \tag{9.94}$$

Substituting from the estimation error, it follows that

$$(\boldsymbol{e}_k^+)(\boldsymbol{e}_k^+)^T = (\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T - (\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T \boldsymbol{K}_k^T \tag{9.95}$$
$$- \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T \tag{9.96}$$
$$+ \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T \boldsymbol{K}_k^T \tag{9.97}$$

Now, we take the expected value of this outer product to get the posterior estimation error covariance

$$\boldsymbol{P}_k^+ = \mathrm{E}\left\{(\boldsymbol{e}_k^+)(\boldsymbol{e}_k^+)^T\right\} \tag{9.98}$$
$$= \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} - \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T \boldsymbol{K}_k^T\right\} \tag{9.99}$$
$$- \mathrm{E}\left\{\boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \tag{9.100}$$
$$+ \mathrm{E}\left\{\boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T \boldsymbol{K}_k^T\right\} \tag{9.101}$$

Assuming that the gain matrix is deterministic, it follows that the *a posteriori* covariance is

$$\boldsymbol{P}_k^+ = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} - \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \boldsymbol{K}_k^T \tag{9.102}$$
$$- \boldsymbol{K}_k \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \tag{9.103}$$
$$+ \boldsymbol{K}_k \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \boldsymbol{K}_k^T \tag{9.104}$$

Let the prior state covariance, cross-covariance (with the measurement), and measurement covariance be defined as

$$\boldsymbol{P}_k^- = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \tag{9.105}$$
$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.106}$$
$$\boldsymbol{W}_k = \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.107}$$

Substituting for the above relationships, the covariance update becomes

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k \boldsymbol{K}_k^T - \boldsymbol{K}_k \boldsymbol{C}_k^T + \boldsymbol{K}_k \boldsymbol{W}_k \boldsymbol{K}_k^T \tag{9.108}$$

Note that no specification of linearity of the measurement process needs to be made for this equation to hold.

We did, however, have to specify that the gain is deterministic, and this will become important later on. For now, just keep this in mind.

Up to this point, no form has been given for the gain matrix, $\boldsymbol{K}_k$.

$\boldsymbol{K}_k$ is found such that the mean square of the *a posteriori* state estimation error is minimized, i.e. the performance index is given by

$$J = \mathrm{E}\left\{(\boldsymbol{e}_k^+)^T(\boldsymbol{e}_k^+)\right\} = \mathrm{trace}\,\mathrm{E}\left\{(\boldsymbol{e}_k^+)(\boldsymbol{e}_k^+)^T\right\} = \mathrm{trace}\,\boldsymbol{P}_k^+ \tag{9.109}$$

Now, we substitute for our form of the posterior covariance matrix to find that

$$J = \mathrm{trace}\,\{\boldsymbol{P}_k^-\} - \mathrm{trace}\,\{\boldsymbol{C}_k\boldsymbol{K}_k^T\} - \mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{C}_k^T\} + \mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T\} \tag{9.110}$$

$$= \mathrm{trace}\,\{\boldsymbol{P}_k^-\} - 2\,\mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{C}_k^T\} + \mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T\} \tag{9.111}$$

$$\tag{9.112}$$

To proceed, we need to know how to take the derivative of the trace of a matrix.

It can be shown that

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}\,\{\boldsymbol{BAC}\} = \boldsymbol{B}^T\boldsymbol{C}^T \tag{9.113}$$

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}\,\{\boldsymbol{ABA}^T\} = \boldsymbol{A}\left[\boldsymbol{B}+\boldsymbol{B}^T\right] \tag{9.114}$$

Now, we can take the derivative of our performance index in a term-by-term fashion. The derivative terms are

$$\frac{\partial}{\partial \boldsymbol{K}_k}\mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{C}_k^T\} = \boldsymbol{C}_k \tag{9.115}$$

$$\frac{\partial}{\partial \boldsymbol{K}_k}\mathrm{trace}\,\{\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T\} = \boldsymbol{K}_k\left[\boldsymbol{W}_k+\boldsymbol{W}_k^T\right] \tag{9.116}$$

Now, we can put the pieces together to get the derivative of the performance index with respect to the gain matrix

$$\frac{\partial J}{\partial \boldsymbol{K}_k} = \boldsymbol{0} - 2\boldsymbol{C}_k + \boldsymbol{K}_k\left[\boldsymbol{W}_k+\boldsymbol{W}_k^T\right] \tag{9.117}$$

Since the matrix $\boldsymbol{W}_k$ is symmetric, we find that

$$\frac{\partial J}{\partial \boldsymbol{K}_k} = -2\boldsymbol{C}_k + 2\boldsymbol{K}_k\boldsymbol{W}_k \tag{9.118}$$

Therefore, the gain which renders the performance index stationary is given by

$$\frac{\partial J}{\partial \boldsymbol{K}_k} = -2\boldsymbol{C}_k + 2\boldsymbol{K}_k\boldsymbol{W}_k = \boldsymbol{0} \tag{9.119}$$

which yields the Kalman gain as

$$\boldsymbol{K}_k = \boldsymbol{C}_k\boldsymbol{W}_k^{-1} \tag{9.120}$$

Does this gain minimize the cost function?

To show this, let's consider another gain that is $\bar{\boldsymbol{K}}_k = \boldsymbol{K}_k + \Delta\boldsymbol{K}_k$. In this case, we know that the posterior covariance (remember that our posterior covariance equation is valid for *any* linear gain) is

$$\bar{\boldsymbol{P}}_k = \boldsymbol{P}_k^- - \boldsymbol{C}_k\bar{\boldsymbol{K}}_k^T - \bar{\boldsymbol{K}}_k\boldsymbol{C}_k^T + \bar{\boldsymbol{K}}_k\boldsymbol{W}_k\bar{\boldsymbol{K}}_k^T \tag{9.121}$$

Now, we simply apply the new gain matrix and expand out terms to find

$$\begin{aligned}
\bar{\boldsymbol{P}}_k &= \boldsymbol{P}_k^- - \boldsymbol{C}_k\big[\boldsymbol{K}_k + \Delta\boldsymbol{K}_k\big]^T - \big[\boldsymbol{K}_k + \Delta\boldsymbol{K}_k\big]\boldsymbol{C}_k^T && (9.122)\\
&\quad + \big[\boldsymbol{K}_k + \Delta\boldsymbol{K}_k\big]\boldsymbol{W}_k\big[\boldsymbol{K}_k + \Delta\boldsymbol{K}_k\big]^T && (9.123)\\
&= \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T && (9.124)\\
&\quad - \boldsymbol{C}_k\Delta\boldsymbol{K}_k^T - \Delta\boldsymbol{K}_k\boldsymbol{C}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T && (9.125)
\end{aligned}$$

Recognizing the top line of the last equation to be our posterior covariance when the gain $\boldsymbol{K}_k$ is used, and denoting it as $\boldsymbol{P}_k^+$ still, it follows that

$$\bar{\boldsymbol{P}}_k = \boldsymbol{P}_k^+ - \boldsymbol{C}_k\Delta\boldsymbol{K}_k^T - \Delta\boldsymbol{K}_k\boldsymbol{C}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T \tag{9.126}$$

Now, let's apply the equation for the Kalman gain, which yields

$$\begin{aligned}
\bar{\boldsymbol{P}}_k &= \boldsymbol{P}_k^+ - \boldsymbol{C}_k\Delta\boldsymbol{K}_k^T - \Delta\boldsymbol{K}_k\boldsymbol{C}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T && (9.127)\\
&\quad + \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T && (9.128)\\
&= \boldsymbol{P}_k^+ + \Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T && (9.129)
\end{aligned}$$

To show that this leads to a higher cost, we take the trace

$$\text{trace}\,\bar{\boldsymbol{P}}_k = \text{trace}\,\boldsymbol{P}_k^+ + \text{trace}\,\{\Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T\} \tag{9.130}$$

Since $\boldsymbol{W}_k > \boldsymbol{0}$, $\Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T \geq \boldsymbol{0}$, which means that

$$\text{trace}\,\{\Delta\boldsymbol{K}_k\boldsymbol{W}_k\Delta\boldsymbol{K}_k^T\} > 0 \tag{9.131}$$

which means that

$$\text{trace}\,\bar{\boldsymbol{P}}_k > \text{trace}\,\boldsymbol{P}_k^+ \tag{9.132}$$

Thus, any gain other than the Kalman gain leads to a higher cost than the Kalman gain, so we can conclude that the Kalman gain does indeed minimize the posterior mean square error.

To apply the Kalman filter, the measurement-dependent quantities

$$\hat{\boldsymbol{z}}_k = \text{E}\,\{\boldsymbol{z}_k\} \tag{9.133}$$

$$\boldsymbol{C}_k = \text{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.134}$$

$$\boldsymbol{W}_k = \text{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.135}$$

are needed.

Consider the case where the measurement is linear in the state and subjected to additive measurement noise via

$$\boldsymbol{z}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{L}_k\boldsymbol{v}_k \tag{9.136}$$

where the first- and second-moment statistics of the measurement noise are

$$\text{E}\,\{\boldsymbol{v}_k\} = \boldsymbol{0} \quad \text{and} \quad \text{E}\left\{\boldsymbol{v}_k\boldsymbol{v}_\ell^T\right\} = \boldsymbol{R}_k\delta_{k\ell} \tag{9.137}$$

Taking the expected value of both sides of the measurement model yields

$$\hat{\boldsymbol{z}}_k = \text{E}\,\{\boldsymbol{z}_k\} = \text{E}\,\{\boldsymbol{H}_k\boldsymbol{x}_k\} + \text{E}\,\{\boldsymbol{L}_k\boldsymbol{v}_k\} \tag{9.138}$$

Since $\boldsymbol{H}_k$ and $\boldsymbol{L}_k$ are deterministic,

$$\hat{\boldsymbol{z}}_k = \boldsymbol{H}_k \mathrm{E}\{\boldsymbol{x}_k\} + \boldsymbol{L}_k \mathrm{E}\{\boldsymbol{v}_k\} \tag{9.139}$$

Recalling that the measurement noise is taken to be zero mean,

$$\hat{\boldsymbol{z}}_k = \boldsymbol{H}_k \mathrm{E}\{\boldsymbol{x}_k\} \tag{9.140}$$

Therefore, the expected value of the measurement is given by

$$\hat{\boldsymbol{z}}_k = \boldsymbol{H}_k \boldsymbol{m}_k^- \tag{9.141}$$

Now, consider the cross-covariance

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(z_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.142}$$

Looking first at the term $(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)$ and substituting from the measurement model and expected measurement, it follows that

$$\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k = \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.143}$$

Thus, the cross-covariance becomes

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T \boldsymbol{H}_k^T\right\} + \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T \boldsymbol{L}_k^T\right\} \tag{9.144}$$

Since $\boldsymbol{H}_k$ and $\boldsymbol{L}_k$ are deterministic

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \boldsymbol{H}_k^T + \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T\right\} \boldsymbol{L}_k^T \tag{9.145}$$

Assuming that the state is not correlated to the measurement noise, i.e.

$$\mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T\right\} = \boldsymbol{0} \tag{9.146}$$

it follows that the cross-covariance for linear measurements with additive noise is

$$\boldsymbol{C}_k = \boldsymbol{P}_k^- \boldsymbol{H}_k^T \tag{9.147}$$

Finally, consider the measurement covariance (also known as the residual covariance or the innovations covariance)

$$\boldsymbol{W}_k = \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.148}$$

Using the previously developed result of

$$\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k = \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.149}$$

and recalling the previous properties/assumptions that

- $\boldsymbol{H}_k$ and $\boldsymbol{L}_k$ are deterministic
- the state is not correlated with the measurement noise
- the covariance of the measurement noise is given by $\boldsymbol{R}_k$

gives the innovations covariance for linear measurements as

$$\boldsymbol{W}_k = \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T + \boldsymbol{L}_k \boldsymbol{R}_k \boldsymbol{L}_k^T \tag{9.150}$$

This completes the Kalman filter!

To summarize, we put everything together in a single table

System Model	$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(t)\boldsymbol{x}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t)$
Meas. Model	$\boldsymbol{z}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{L}_k\boldsymbol{v}_k$
Init. Cond.	$\boldsymbol{m}_0 = \mathrm{E}\{\boldsymbol{x}(t_0)\}$
	$\boldsymbol{P}_0 = \mathrm{E}\left\{(\boldsymbol{x}(t_0) - \boldsymbol{m}_0)(\boldsymbol{x}(t_0) - \boldsymbol{m}_0)^T\right\}$
Mean Prop.	$\dot{\boldsymbol{m}}(t) = \boldsymbol{F}(t)\boldsymbol{m}(t)$
Cov. Prop.	$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(t)\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(t) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t)$
Exp. Meas.	$\hat{\boldsymbol{z}}_k = \boldsymbol{H}_k\boldsymbol{m}_k^-$
Innov. Cov.	$\boldsymbol{W}_k = \boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T + \boldsymbol{L}_k\boldsymbol{R}_k\boldsymbol{L}_k^T$
Cross Cov.	$\boldsymbol{C}_k = \boldsymbol{P}_k^-\boldsymbol{H}_k^T$
Kalman Gain	$\boldsymbol{K}_k = \boldsymbol{C}_k\boldsymbol{W}_k^{-1}$
Mean Upd.	$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)$
Cov. Upd.	$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T$

What the update stage of the Kalman filter framework **does not do**

- makes no requirement that the distribution be Gaussian
- makes no requirement that the measurement function be linear
  - note that this means that we compute $\hat{\boldsymbol{z}}_k$, $\boldsymbol{C}_k$, and $\boldsymbol{W}_k$ with expectations, not solely by the equations stemming from the measurement equation

What the Kalman filter update **does** do

- works with first- and second-moment statistics
- employs a linear update law, i.e. a linear gain
- forces an unbiased posterior estimate (can be relaxed)
- minimizes the posterior mean square error (minimum variance)

9.1.1 Sports Car Example

Suppose you are at a drag strip with a laser rangefinder, and a company is testing their new "constant acceleration" electric motor in a sports car.

You are interested in taking range measurements of the vehicle as it starts from rest to use the previously discussed Kalman filter techniques to acquire an estimate of the car's range and its first two associated rates (with respect to your rangefinder) at any point during its 10 second test.

Your range-finder is right at the starting line and can generate noisy range measurements with a measurement noise standard deviation of 10 meters at a rate of 2 Hz.

Define the state vector

$$\boldsymbol{x} = [\rho \;\; \dot{\rho} \;\; \ddot{\rho}]^T \tag{9.151}$$

The company's engineer tells you that the engine should accelerate at 7 m/s^2, so we can define our initial estimate

$$\boldsymbol{m}_0 = [0 \;\; 0 \;\; 7]^T \tag{9.152}$$

According to the system model for the Kalman filter, we will add some error to this mean to generate the initial truth in simulation.

As such, we need to have some initial uncertainty. We will take the truth to be Gaussian-distributed with the mean as given previously and the covariance to be (with appropriate units)

$$\boldsymbol{P}_0 = \begin{bmatrix} 14^2 & 0 & 0 \\ 0 & 5^2 & 0 \\ 0 & 0 & 1^2 \end{bmatrix} \tag{9.153}$$

Constant acceleration of the vehicle (remember that this is a "constant acceleration" motor) tells us that

$$\boldsymbol{F}(t) = \boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{9.154}$$

We also know that we are only taking range measurements, so

$$\boldsymbol{H}_k = [1 \;\; 0 \;\; 0] \tag{9.155}$$

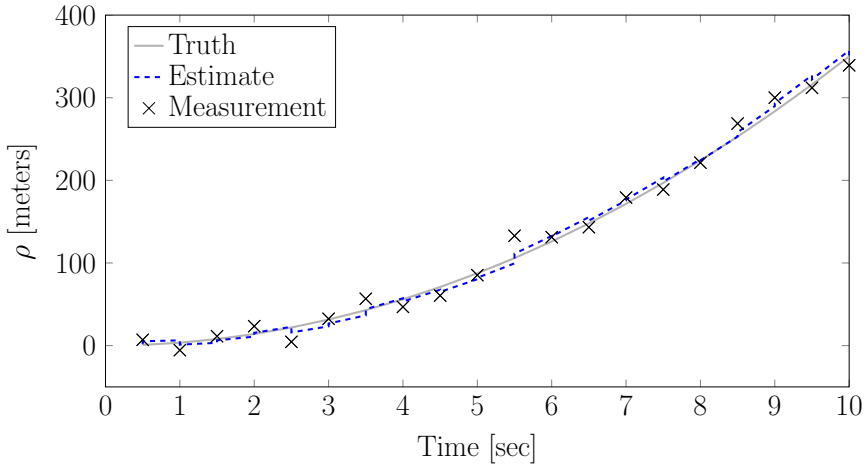
Additionally, to account for noise in our dynamic model, we define the process noise power spectral density to be

$$\boldsymbol{Q}_s(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.01 \end{bmatrix} \tag{9.156}$$

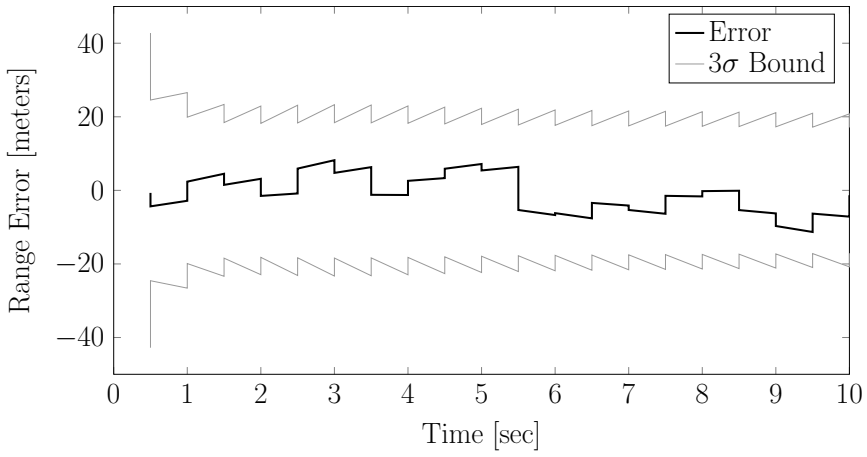
It is important to note that we will not use the process noise in our truth generation. Instead, we use it as it is most often used...to inflate the covariance matrix during propagation.

The car takes off, and we take our measurements while processing the data with a Kalman filter.

The measured, true, and estimated range of the car can be seen in the figure below. It seems like our Kalman filter did its job! It tracks along the trajectory according to our measurements.



More appropriate for evaluating the performance of the filter, perhaps, is the error in our estimate. This can be seen plotted below.



Note that the error in our estimate tends to stay within about 10 meters. Here σ is the square root of the $(1,1)$ element of the covariance matrix at each step (that is, the element of the state estimation covariance which corresponds to the variance in ρ).

A property of the Gaussian distribution, the distribution we have modeled our random variable (the state) as, is that 3σ contains 99.7% of the outcomes generated by that distribution with standard deviation σ. This is a fairly good look at the "worst" we could do at any given time.

How did we code this Kalman filter in MATLAB?
First, set the random number seed and define the relevant system and simulation parameters.

```
% Set random number seed
rng(100)

% Relevant system factors
dt = 0.5;                       % Time step
%tv = 0:dt:10;                   % Time vector
tv = 0:dt:100;                  % Time vector
Rk = 10^2;                      % Measurement noise covariance [m^2]
Qs = diag([1,.1,.01]);          % Process noise PSD
F  = [0, 1, 0;                  % Dynamics
      0, 0, 1;
      0, 0, 0];
Hk = [1, 0, 0];                 % Measurements
M  = eye(3);                    % Process noise mapping
```

Define our initial estimates for mean and covariance and the truth at the starting time.

```
% Truth and estimates at t=0
m0  = [0; 0; 7.0];              % Initial estimate [m; m/s; m/s^2]
P0  = diag([14^2,5^2,1^2]);     % Initial covariance
```

Initialize the Kalman filter variables for truth and estimates. Also, initialize a counting variable (which we will simply use to help us store things along the way).

```
% Initialize
xkm1 = xt0;
mkm1 = m0;
Pkm1 = P0;
```

We will be using MATLAB's ODE45 (a Runge-Kutta 4/5 integrator) to integrate the differential equations required for prediction, so let's define its tolerances.

```
cnt  = 1;
```

Now we can start our Kalman filter! We will use a "for loop" over all the times we expect to receive a measurement from our laser rangefinder. Start by propagating the truth (this is the car driving). Note that the Kalman filter we use won't actually *see* this truth, but we will use it to generate noisy measurements.

```
opts = odeset('AbsTol',1e-6,'RelTol',1e-6);
% Kalman filter
for i = 2:length(tv)
    % Propagate
    [~,X] = ode45(@car_eoms,[tv(i-1),tv(i)],xkm1,opts,F,Qs,M);
```

Now we can generate those noisy measurements we've been talking about.

```
xk     = X(end,1:3)';
% Generate a measurement
```

Predict according to the Kalman filter equations (and store the *a priori* mean and covariance to look at later). We'll detail the equations of motion file `car_eoms` later.

```
% Predict
[~,X] = ode45(@car_eoms,[tv(i-1),tv(i)],[mkm1;Pkm1(:)],opts,F,Qs,M);
mkm    = X(end,1:3)';
Pkm    = reshape(X(end,4:end),3,3);
% Store a priori mean and covariance
zkp(:,cnt)    = zk;
xtp(:,cnt)    = xk;
mpt(:,cnt)    = mkm;
Ppt(:,:,cnt) = Pkm;
```

Correct according to the Kalman filter equations (and store the *a posteriori* mean and covariance to look at later, too).

```
% Correct
zht = Hk*mkm;
Wk  = Hk*Pkm*Hk' + Lk*Rk*Lk';
Ck  = Pkm*Hk';
Kk  = Ck/Wk;
mkp = mkm + Kk*(zk - zht);
Pkp = Pkm - Ck*Kk' - Kk*Ck' + Kk*Wk*Kk';
% Store a posteriori mean and covariance
zkp(:,cnt)    = zk;
xtp(:,cnt)    = xk;
mpt(:,cnt)    = mkp;
Ppt(:,:,cnt) = Pkp;
```

Set up the recursion for the next time step (and don't forget to close that loop).

```
% Cycle
xkm1 = xk;
mkm1 = mkp;
Pkm1 = Pkp;
```

The equations of motion file that ODE45 calls looks like this:

```
function [dxdt] = car_eoms(t,x,F,Q,M)

dxdt = F*x(1:3);

% Stop here if we are simply propagating truth.
% If we aren't (i.e. we are propagating our
% estimate which includes a covariance), propagate
% that covariance!
if length(x) > 3
    P  = reshape(x(4:end),3,3);
    dP = F*P + P*F' + M*Q*M';
    % Add a vectorized rate of change to the rate of change vector
    dxdt = [dxdt; dP(:)];
end
```

9.1.2 Variations on the Covariance Update

Let's look at a few alternative forms of the covariance update.

Our original covariance update equation is

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.157}$$

It is important to remember that no form of the gain, $\boldsymbol{K}_k$, was specified, other than it be a linear gain, in arriving at the preceding equation.

Recall that

$$\boldsymbol{C}_k = \boldsymbol{P}_k^-\boldsymbol{H}_k^T \tag{9.158}$$

$$\boldsymbol{W}_k = \boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T + \boldsymbol{L}_k\boldsymbol{R}_k\boldsymbol{L}_k^T \tag{9.159}$$

Substitute for $\boldsymbol{C}_k$ into our covariance update

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{P}_k^-\boldsymbol{H}_k^T\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{P}_k^- + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.160}$$

Substitute for $\boldsymbol{W}_k$

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{P}_k^-\boldsymbol{H}_k^T\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{P}_k^- \tag{9.161}$$

$$+ \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T\boldsymbol{K}_k^T + \boldsymbol{K}_k\boldsymbol{L}_k\boldsymbol{R}_k\boldsymbol{L}_k^T\boldsymbol{K}_k^T \tag{9.162}$$

This can be rearranged to yield

$$\boldsymbol{P}_k^+ = \left[\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k\right]\boldsymbol{P}_k^-\left[\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k\right]^T + \boldsymbol{K}_k\boldsymbol{L}_k\boldsymbol{R}_k\boldsymbol{L}_k^T\boldsymbol{K}_k^T \tag{9.163}$$

This is known as the Joseph form of the covariance update. We still have not specifically given a value of the linear gain for this equation to be valid. However, we have required the use of linear measurements to arrive at this equation.

Let's go back to our original equation

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.164}$$

Flip the second and third terms

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{C}_k^T - \boldsymbol{C}_k\boldsymbol{K}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.165}$$

Recall that the Kalman gain is

$$\boldsymbol{K}_k = \boldsymbol{C}_k\boldsymbol{W}_k^{-1} \tag{9.166}$$

Substitute the Kalman gain into the first $\boldsymbol{K}_k$ of the last term in the covariance update

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{C}_k^T - \boldsymbol{C}_k\boldsymbol{K}_k^T + \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.167}$$

Cancel the residual covariance and its inverse

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{C}_k^T - \boldsymbol{C}_k\boldsymbol{K}_k^T + \boldsymbol{C}_k\boldsymbol{K}_k^T \tag{9.168}$$

Now, we can cancel the last two terms

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{C}_k^T \tag{9.169}$$

Substitute for the cross-covariance

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{P}_k^- \tag{9.170}$$

Finally, factor our the prior covariance matrix

$$\boldsymbol{P}_k^+ = \big[\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k\big]\boldsymbol{P}_k^- \tag{9.171}$$

This is the standard form of the covariance update that is usually seen in most papers/books. This form does require the

Kalman gain to be the gain that is used, and it also requires a linear system.

For our last alternative form of the covariance update, let's start again from our original equation

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.172}$$

and substitute everywhere for $\boldsymbol{K}_k$, such that

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T - \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T + \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{W}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T \tag{9.173}$$

After reducing terms, we find

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T \tag{9.174}$$

At this point, we can put an identity matrix into the last term in the form of $\boldsymbol{W}_k\boldsymbol{W}_k^{-1}$, which gives us

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\boldsymbol{W}_k\boldsymbol{W}_k^{-1}\boldsymbol{C}_k^T \tag{9.175}$$

Now, we recognize that the Kalman gain can be used twice in the second term, such that

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.176}$$

This form relies on the Kalman gain, but does not require the assumption of a linear measurement.

All of these forms are algebraically equivalent, but they have slightly different numerical properties.

9.1.3 A Property of the Residual

Let's define the residual (innovation) to be

$$\boldsymbol{r}_k = \boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^- \tag{9.177}$$

The residual is zero mean, which is easily shown by taking the expected value and substituting for the measurement model; that is,

$$\begin{aligned}
\mathrm{E}\{\boldsymbol{r}_k\} &= \mathrm{E}\{\boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\} && (9.178)\\
&= \mathrm{E}\{\boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{v}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\} && (9.179)\\
&= \mathrm{E}\{\boldsymbol{H}_k \boldsymbol{e}_k^- + \boldsymbol{v}_k\} && (9.180)\\
&= \boldsymbol{H}_k \mathrm{E}\{\boldsymbol{e}_k^-\} + \mathrm{E}\{\boldsymbol{v}_k\} && (9.181)\\
&= \boldsymbol{0} && (9.182)
\end{aligned}$$

The last equality follows from the fact that we have constructed an unbiased estimator and the fact that the measurement noise is zero mean.

Note that we are considering the case of $\boldsymbol{L}_k = \boldsymbol{I}$ for simplicity.

We have also previously defined the residual (innovations) covariance as

$$\boldsymbol{W}_k = \mathrm{E}\{\boldsymbol{r}_k \boldsymbol{r}_k^T\} = \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T + \boldsymbol{R}_k \tag{9.183}$$

This relationship is obtained by assuming that the state and measurement noise are uncorrelated.

Now, let's define the post-fit residual (sometimes this is just called the residual) as

$$\boldsymbol{r}_k^+ = \boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^+ \tag{9.184}$$

From the update equation for the mean, it follows that

$$\begin{aligned}
\boldsymbol{r}_k^+ &= \boldsymbol{z}_k - \boldsymbol{H}_k \left[\boldsymbol{m}_k^- + \boldsymbol{K}_k \left(\boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\right)\right] && (9.185)\\
&= \left[\boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\right] - \boldsymbol{H}_k \boldsymbol{K}_k \left[\boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\right] && (9.186)\\
&= \left[\boldsymbol{I} - \boldsymbol{H}_k \boldsymbol{K}_k\right] \left[\boldsymbol{z}_k - \boldsymbol{H}_k \boldsymbol{m}_k^-\right] && (9.187)\\
&= \left[\boldsymbol{I} - \boldsymbol{H}_k \boldsymbol{K}_k\right] \boldsymbol{r}_k && (9.188)
\end{aligned}$$

Recalling that the residual is zero mean and assuming that $\boldsymbol{H}_k$ and $\boldsymbol{K}_k$ are deterministic, it follows directly that the post-fit residual is zero mean, i.e.

$$\mathrm{E}\{\boldsymbol{r}_k^+\} = \boldsymbol{0} \tag{9.189}$$

It then follows from the expression for the post-fit residual that the post-fit residual covariance is

$$\boldsymbol{W}_k^+ = \mathrm{E}\{\left(\boldsymbol{r}_k^+\right)\left(\boldsymbol{r}_k^+\right)^T\} = \left[\boldsymbol{I} - \boldsymbol{H}_k\boldsymbol{K}_k\right]\boldsymbol{W}_k\left[\boldsymbol{I} - \boldsymbol{H}_k\boldsymbol{K}_k\right]^T \tag{9.190}$$

This is one possible expression for the post-fit residual covariance, but we can also come up with a slightly more useful expression.

Let's go back to our expression for the post-fit residual and substitute for the Kalman gain

$$\begin{aligned}
\boldsymbol{r}_k^+ &= \left[\boldsymbol{I} - \boldsymbol{H}_k\boldsymbol{K}_k\right]\boldsymbol{r}_k && (9.191)\\
&= \left[\boldsymbol{I} - \boldsymbol{H}_k\boldsymbol{C}_k\boldsymbol{W}_k^{-1}\right]\boldsymbol{r}_k && (9.192)\\
&= \left[\boldsymbol{W}_k\boldsymbol{W}_k^{-1} - \boldsymbol{H}_k\boldsymbol{C}_k\boldsymbol{W}_k^{-1}\right]\boldsymbol{r}_k && (9.193)\\
&= \left[\boldsymbol{W}_k - \boldsymbol{H}_k\boldsymbol{C}_k\right]\boldsymbol{W}_k^{-1}\boldsymbol{r}_k && (9.194)\\
&= \left[\boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T + \boldsymbol{R}_k - \boldsymbol{H}_k\boldsymbol{P}_k^-\boldsymbol{H}_k^T\right]\boldsymbol{W}_k^{-1}\boldsymbol{r}_k && (9.195)\\
&= \boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{r}_k && (9.196)\\
& && (9.197)
\end{aligned}$$

Now, we can establish another relationship for the post-fit residual covariance; namely,

$$\boldsymbol{W}_k^+ = \boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{R}_k \tag{9.198}$$

Note that we have made use of the fact that $\boldsymbol{R}_k$ is symmetric in the preceding equation.

We are now ready for the residual property that we want to establish.

Consider the Mahalanobis distance using the post-fit residual and its covariance as

$$\left(\boldsymbol{r}_k^+\right)^T\left(\boldsymbol{W}_k^+\right)^{-1}\left(\boldsymbol{r}_k^+\right) \tag{9.199}$$

If we substitute for our expressions for the post-fit residual and its covariance in terms of the prior residual and its covariance, it follows that

$$\left(\boldsymbol{r}_k^+\right)^T\left(\boldsymbol{W}_k^+\right)^{-1}\left(\boldsymbol{r}_k^+\right) = \left[\boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{r}_k\right]^T\left[\boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{R}_k\right]^{-1}\left[\boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{r}_k\right] \tag{9.200}$$

$$= \boldsymbol{r}_k^T\boldsymbol{W}_k^{-1}\boldsymbol{R}_k\boldsymbol{R}_k^{-1}\boldsymbol{W}_k\boldsymbol{R}_k^{-1}\boldsymbol{R}_k\boldsymbol{W}_k^{-1}\boldsymbol{r}_k \tag{9.201}$$

$$= \boldsymbol{r}_k^T\boldsymbol{W}_k^{-1}\boldsymbol{r}_k \tag{9.202}$$

The posterior Mahalanobis distance is exactly equal to the prior Mahalanobis distance!

9.1.4 Singular Measurement Noise

An interesting situation arises when the measurement noise is *so* small that the measurement noise covariance is zero.

Theoretically, this cannot occur since the measurement noise covariance is a symmetric, positive definite matrix, but numerically it occurs when you have very precise measurements.

At first, it would appear that $\boldsymbol{R}_k = \boldsymbol{0}$ causes no problems since $\boldsymbol{R}_k^{-1}$ does not appear in the Kalman filter, but let's dig a bit deeper.

Consider the covariance update given by

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{K}_k \boldsymbol{H}_k \boldsymbol{P}_k^- \tag{9.203}$$

This is the "textbook" form of the covariance update that relies on the optimal gain.

When the measurement noise is zero, the Kalman gain becomes

$$\boldsymbol{K}_k = \boldsymbol{P}_k^- \boldsymbol{H}_k^T \left[\boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T\right]^{-1} \tag{9.204}$$

and the covariance update is

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{P}_k^- \boldsymbol{H}_k^T \left[\boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T\right]^{-1} \boldsymbol{H}_k \boldsymbol{P}_k^- \tag{9.205}$$

Now, consider

$$\boldsymbol{H}_k \boldsymbol{P}_k^+ \boldsymbol{H}_k^T \tag{9.206}$$

From the covariance update, we have

$$\begin{aligned} \boldsymbol{H}_k \boldsymbol{P}_k^+ \boldsymbol{H}_k^T &= \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T - \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T \left[\boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T\right]^{-1} \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T && (9.207)\\ &= \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T - \boldsymbol{H}_k \boldsymbol{P}_k^- \boldsymbol{H}_k^T && (9.208)\\ &= \boldsymbol{0} && (9.209) \end{aligned}$$

Interesting! The posterior covariance *must be* singular!

What is the ramification of this singularity?

The covariance at the next time step is propagated according to

$$\boldsymbol{P}_{k+1}^{-} = \boldsymbol{\Phi}(t_{k+1}, t_k)\boldsymbol{P}_k^{+}\boldsymbol{\Phi}^T(t_{k+1}, t_k) + \boldsymbol{Q}_c(t_{k+1}) \tag{9.210}$$

If the time steps are close together, then

$$\boldsymbol{\Phi}(t_{k+1}, t_k) \approx \boldsymbol{I} \qquad \text{and} \qquad \boldsymbol{Q}_c(t_{k+1}) \approx \boldsymbol{0} \tag{9.211}$$

This means that the covariance update may begin to fail. That is, the residual (innovations) covariance may come out to be numerically close to zero, and the Kalman gain cannot be computed.

Ultimately, it means that we have to be very careful when we have precise measurements.

We'll come back to this subject later on, but keep this in mind always.

We have an algorithm for the Kalman filter; however, that does not mean that it solves every problem or that numerical issues cannot degrade the performance of our filter.

9.2 The Extended Kalman Filter (Nonlinear Dynamics)

The Kalman filter operates on linear dynamical systems, but oftentimes we must deal with nonlinear dynamics, nonlinear measurements, or both.

For instance, an object under the influence of two-body dynamics obeys a nonlinear differential equation in Cartesian coordinates. All objects on orbits follow nonlinear dynamics, when perturbations beyond the two body motion are taken into account (both, Cartesian and e.g. orbital elements coordinates).

We often cannot observe the actual state of the object either. Instead, we tend to observe some nonlinear function of the state, such as range.

We therefore want to modify our Kalman filter to be able to handle these nonlinearities.

The extended Kalman filter (EKF) handles nonlinearities through the use of linearization.

As with the Kalman filter, the filter is comprised of two stages: propagation and update.

We will proceed to develop the EKF in the same fashion as the Kalman filter, by first developing the evolutionary equations for the propagation of the mean and covariance and then developing update relationships for the mean and covariance.

Consider the nonlinear dynamical system subjected to random excitations

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{M}(t)\boldsymbol{w}(t) \tag{9.212}$$

where

$$\mathrm{E}\{\boldsymbol{w}(t)\} = \boldsymbol{0} \quad \text{and} \quad \mathrm{E}\left\{\boldsymbol{w}(t)\boldsymbol{w}^T(\tau)\right\} = \boldsymbol{Q}_s(t)\delta(t-\tau) \tag{9.213}$$

9.2.0.1 Propagation Step

The mean of the state as a function of time is given by

$$\boldsymbol{m}(t) = \mathrm{E}\{\boldsymbol{x}(t)\} \tag{9.214}$$

Taking the time rate of change and interchanging the order of differentiation and expectation yields

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\dot{\boldsymbol{x}}(t)\} \tag{9.215}$$

Applying the system dynamics within the expectation, it follows that

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.216}$$

$$= \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}(t))\} + \mathrm{E}\{\boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.217}$$

Express $\boldsymbol{f}(\boldsymbol{x}(t))$ as a first-order Taylor series expansion (FOTSE) about the current mean as

$$\boldsymbol{f}(\boldsymbol{x}(t)) = \boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))(\boldsymbol{x}(t) - \boldsymbol{m}(t)) + \text{H.O.T.} \tag{9.218}$$

where the dynamics Jacobian, $\boldsymbol{F}(\boldsymbol{m}(t))$, is defined as

$$\boldsymbol{F}(\boldsymbol{m}(t)) = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{x}(t))}{\partial \boldsymbol{x}(t)} \bigg|_{\boldsymbol{x}(t)=\boldsymbol{m}(t)} \right] \tag{9.219}$$

Substitute the FOTSE into the expected value of the dynamics

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))(\boldsymbol{x}(t) - \boldsymbol{m}(t))\} + \mathrm{E}\{\boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.220}$$

Define the error with respect to a true state to be

$$\boldsymbol{e}(t) = \boldsymbol{x}(t) - \boldsymbol{m}(t) \tag{9.221}$$

The expected value of the dynamics can then be written as (separating terms within the expectation)

$$\dot{\boldsymbol{m}}(t) = \mathrm{E}\{\boldsymbol{f}(\boldsymbol{m}(t))\} + \mathrm{E}\{\boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{e}(t)\} + \mathrm{E}\{\boldsymbol{M}(t)\boldsymbol{w}(t)\} \tag{9.222}$$

Assuming that $\boldsymbol{f}(\boldsymbol{m}(t))$, $\boldsymbol{F}(\boldsymbol{m}(t))$, and $\boldsymbol{M}(t)$ are deterministic

$$\dot{\boldsymbol{m}}(t) = \boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))\mathrm{E}\{\boldsymbol{e}(t)\} + \boldsymbol{M}(t)\mathrm{E}\{\boldsymbol{w}(t)\} \tag{9.223}$$

This assumption is complicit with the assumption that $\boldsymbol{m}(t)$ is deterministic.

Recalling that the process noise is taken to be zero-mean

$$\dot{\boldsymbol{m}}(t) = \boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))\mathrm{E}\{\boldsymbol{e}(t)\} \tag{9.224}$$

Finally, assuming that the estimate is unbiased (equivalently, assuming that $\boldsymbol{e}(t)$ is a zero-mean process), it follows that the mean satisfies the differential equation

$$\dot{\boldsymbol{m}}(t) = \boldsymbol{f}(\boldsymbol{m}(t)) \tag{9.225}$$

This is the differential equation governing the forward evolution of the mean. We now turn to developing a method for

propagating the covariance.

From the definition of the error, the error dynamics are

$$\dot{\boldsymbol{e}}(t) = \dot{\boldsymbol{x}}(t) - \dot{\boldsymbol{m}}(t) \tag{9.226}$$

Substitute for the dynamics of the truth and the dynamics of the mean to get

$$\dot{\boldsymbol{e}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{M}(t)\boldsymbol{w}(t) - \boldsymbol{f}(\boldsymbol{m}(t)) \tag{9.227}$$

Expand the dynamics of the truth about the mean via

$$\boldsymbol{f}(\boldsymbol{x}(t)) = \boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))(\boldsymbol{x}(t) - \boldsymbol{m}(t)) + \text{H.O.T.} \tag{9.228}$$

Substitute the FOTSE into the error dynamics

$$\begin{aligned} \dot{\boldsymbol{e}}(t) &= \boldsymbol{f}(\boldsymbol{m}(t)) + \boldsymbol{F}(\boldsymbol{m}(t))(\boldsymbol{x}(t) - \boldsymbol{m}(t)) + \boldsymbol{M}(t)\boldsymbol{w}(t) - \boldsymbol{f}(\boldsymbol{m}(t)) && (9.229)\\ &= \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{e}(t) + \boldsymbol{M}(t)\boldsymbol{w}(t) && (9.230) \end{aligned}$$

The solution of the linear differential equation for the error is

$$\boldsymbol{e}(t) = \boldsymbol{\Phi}(t, t_{k-1})\boldsymbol{e}(t_{k-1}) + \int_{t_{k-1}}^{t} \boldsymbol{\Phi}(t,\tau)\boldsymbol{M}(\tau)\boldsymbol{w}(\tau)d\tau \tag{9.231}$$

where $\boldsymbol{\Phi}(t, t_{k-1})$ is the state transition matrix which satisfies

$$\dot{\boldsymbol{\Phi}}(t, t_{k-1}) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{\Phi}(t, t_{k-1}), \qquad \boldsymbol{\Phi}(t_{k-1}, t_{k-1}) = \boldsymbol{I} \tag{9.232}$$

The state estimation error covariance is found via

$$\boldsymbol{P}(t) = \text{E}\left\{\boldsymbol{e}(t)\boldsymbol{e}^T(t)\right\} \tag{9.233}$$

Now, we "simply" follow the same procedure as used for the Kalman filter derivation but with $\boldsymbol{F}(\boldsymbol{m}(t))$ in place of $\boldsymbol{F}(t)$.

This will lead us, as with the Kalman filter, to two separate methods for propagating the covariance.

First method for covariance propagation:

- Propagate state transition matrix with $\boldsymbol{\Phi}(t_{k-1}, t_{k-1}) = \boldsymbol{I}$

$$\dot{\boldsymbol{\Phi}}(t, t_{k-1}) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{\Phi}(t, t_{k-1}) \tag{9.234}$$

- Propagate process noise covariance matrix with $\boldsymbol{Q}_c(t_{k-1}) = \boldsymbol{0}$

$$\dot{\boldsymbol{Q}}_c(t) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{Q}_c(t) + \boldsymbol{Q}_c(t)\boldsymbol{F}^T(\boldsymbol{m}(t)) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.235}$$

- Calculate the propagated covariance matrix

$$\boldsymbol{P}(t) = \boldsymbol{\Phi}(t, t_{k-1})\boldsymbol{P}(t_{k-1})\boldsymbol{\Phi}^T(t, t_{k-1}) + \boldsymbol{Q}_c(t) \tag{9.236}$$

Second method for covariance propagation:

- Propagate the covariance matrix with $\boldsymbol{P}(t_{k-1}) = \boldsymbol{P}_{k-1}$

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(\boldsymbol{m}(t)) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.237}$$

To summarize the prediction stage of the extended Kalman filter, we numerically integrate

$$\dot{\boldsymbol{m}}(t) = \boldsymbol{f}(\boldsymbol{m}(t)) \tag{9.238}$$

$$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(\boldsymbol{m}(t)) + \boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t) \tag{9.239}$$

across some interval $t \in [t_{k-1} \quad t_k]$ starting with the initial conditions

$$\boldsymbol{m}(t_{k-1}) = \boldsymbol{m}_{k-1}^+ \qquad \text{and} \qquad \boldsymbol{P}(t_{k-1}) = \boldsymbol{P}_{k-1}^+ \tag{9.240}$$

We can also use the other method for covariance propagation.

The values obtained after integrating become our *a priori* mean and covariance, $\boldsymbol{m}_k^-$ and $\boldsymbol{P}_k^-$, when we encounter new measurement data.

9.2.0.2 Measurement Update

At time t_k a measurement is made available, which is given by $\boldsymbol{z}_k$. This measurement is a function of the state and is imperfect (noisy).

This measurement is taken to be of the form

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.241}$$

where

$$\mathrm{E}\{\boldsymbol{v}_k\} = \boldsymbol{0} \qquad \text{and} \qquad \mathrm{E}\{\boldsymbol{v}_k \boldsymbol{v}_\ell^T\} = \boldsymbol{R}_k \delta_{k\ell} \tag{9.242}$$

The measurement noise is represented by $\boldsymbol{v}_k$, which is assumed to be a zero mean white-noise sequence with covariance $\boldsymbol{R}_k$.

The mean and covariance prior to incorporation of this new information are given by

$$\boldsymbol{m}_k^- = \mathrm{E}\{\boldsymbol{x}_k\} \tag{9.243}$$

$$\boldsymbol{P}_k^- = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \tag{9.244}$$

We want to find a way to use this new information to *update* the mean and covariance of our state, to update our estimated state and our confidence in the estimated state.

Before proceeding, let's recall what the Kalman filter update does and does not do.

What the framework for the Kalman filter update **does not** do

- makes no requirement that the distribution be Gaussian
- makes no requirement that the measurement function be linear

What the Kalman filter update **does** do

- works with first- and second-moment statistics
- employs a linear update law, i.e. a linear gain
- forces an unbiased posterior estimate (can be relaxed)

- minimizes the posterior mean square error (minimum variance)

As with the Kalman filter, we assume that the *a posteriori* mean is given by

$$\boldsymbol{m}_k^+ = \boldsymbol{a}_k + \boldsymbol{K}_k \boldsymbol{z}_k \tag{9.245}$$

Let the expected value of the measurement (with respect to any stochastic inputs) be given by

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{z}_k\} \tag{9.246}$$

Following the same procedure as for the Kalman filter without any alterations, it can be shown that

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.247}$$

Note that no specification of linearity of the measurement process needs to be made for this equation to hold. $\hat{\boldsymbol{z}}_k$ is simply the mean of the measurement with respect to the state and noise distributions.

The difference between a linear and a nonlinear system is in how we compute the expected value.

Associated with the mean update, we also want to be able to describe how the covariance is updated, and this follows from the manner in which the error gets updated.

The *a posteriori* state estimation error is

$$\boldsymbol{e}_k^+ = \boldsymbol{e}_k^- - \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.248}$$

If we define $\boldsymbol{P}_k^-$ and $\boldsymbol{P}_k^+$ to be

$$\boldsymbol{P}_k^- = \mathrm{E}\left\{(\boldsymbol{e}_k^-)(\boldsymbol{e}_k^-)^T\right\} \qquad \text{and} \qquad \boldsymbol{P}_k^+ = \mathrm{E}\left\{(\boldsymbol{e}_k^+)(\boldsymbol{e}_k^+)^T\right\} \tag{9.249}$$

then, from our previous Kalman filter developments, we know that

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.250}$$

The cross-covariance (with the measurement) and the residual (innovations) covariance are defined as

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.251}$$

$$\boldsymbol{W}_k = \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.252}$$

Note that again, no specification of linearity of the measurement process needs to be made for these equations to hold; the changes will again be in how the expected values are calculated.

This is exactly the same set of relationships obtained for the Kalman filter.

Up to this point, no form has been given for the gain matrix, $\boldsymbol{K}_k$, but we already solved exactly this problem for the Kalman filter developments.

The Kalman gain is the gain that minimizes the mean square of the posterior state estimation error and is

$$\boldsymbol{K}_k = \boldsymbol{C}_k \boldsymbol{W}_k^{-1} \tag{9.253}$$

Ultimately, the following measurement-dependent (model-dependent) quantities are required to apply the extended Kalman filter (or the Kalman filter):

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{z}_k\} \tag{9.254}$$

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.255}$$

$$\boldsymbol{W}_k = \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.256}$$

Consider the case where the measurement is nonlinear in the state and subjected to additive measurement noise via

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.257}$$

where the first- and second-moment statistics of the measurement noise are

$$\mathrm{E}\{\boldsymbol{v}_k\} = \boldsymbol{0} \qquad \text{and} \qquad \mathrm{E}\left\{\boldsymbol{v}_k \boldsymbol{v}_\ell^T\right\} = \boldsymbol{R}_k \delta_{k\ell} \tag{9.258}$$

Taking the expected value of both sides of the measurement model yields

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{z}_k\} = \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\} + \mathrm{E}\{\boldsymbol{L}_k \boldsymbol{v}_k\} \tag{9.259}$$

Expand the measurement function about the *a priori* mean via

$$\boldsymbol{h}(\boldsymbol{x}_k) = \boldsymbol{h}(\boldsymbol{m}_k^-) + \boldsymbol{H}(\boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-) + \text{H.O.T.} \tag{9.260}$$

where the measurement Jacobian, $\boldsymbol{H}(\boldsymbol{m}_k^-)$, is defined as

$$\boldsymbol{H}(\boldsymbol{m}_k^-) = \left[\left. \frac{\partial \boldsymbol{h}(\boldsymbol{x}_k)}{\partial \boldsymbol{x}_k} \right|_{\boldsymbol{x}_k = \boldsymbol{m}_k^-} \right] \tag{9.261}$$

Substitute the FOTSE into the expected value of the measurement

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\left\{\boldsymbol{h}(\boldsymbol{m}_k^-)\right\} + \mathrm{E}\left\{\boldsymbol{H}(\boldsymbol{m}_k^-)\boldsymbol{e}_k^-\right\} + \mathrm{E}\left\{\boldsymbol{L}_k\boldsymbol{v}_k\right\} \tag{9.262}$$

Assuming that $\boldsymbol{h}(\boldsymbol{m}_k^-)$, $\boldsymbol{H}(\boldsymbol{m}_k^-)$ and $\boldsymbol{L}_k$ are deterministic,

$$\hat{\boldsymbol{z}}_k = \boldsymbol{h}(\boldsymbol{m}_k^-) + \boldsymbol{H}(\boldsymbol{m}_k^-)\mathrm{E}\left\{\boldsymbol{e}_k^-\right\} + \boldsymbol{L}_k\mathrm{E}\left\{\boldsymbol{v}_k\right\} \tag{9.263}$$

Recalling that the measurement noise is taken to be zero-mean and that the prediction error was assumed to be zero-mean (unbiased), the expected value of the measurement is

$$\hat{\boldsymbol{z}}_k = \boldsymbol{h}(\boldsymbol{m}_k^-) \tag{9.264}$$

Now, consider the cross-covariance

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.265}$$

Looking first at the term $(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)$ and substituting from the measurement model and expected measurement, it follows that

$$\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k = \boldsymbol{h}(\boldsymbol{x}_k) - \boldsymbol{h}(\boldsymbol{m}_k^-) + \boldsymbol{L}_k\boldsymbol{v}_k \tag{9.266}$$

Applying the FOTSE for $\boldsymbol{h}(\boldsymbol{x}_k)$ gives

$$\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k = \boldsymbol{H}(\boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.267}$$

The cross-covariance is then

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T \boldsymbol{H}^T(\boldsymbol{m}_k^-)\right\} + \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T \boldsymbol{L}_k^T\right\} \tag{9.268}$$

Since $\boldsymbol{H}(\boldsymbol{m}_k^-)$ and $\boldsymbol{L}_k$ are taken to be deterministic

$$\boldsymbol{C}_k = \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-)^T\right\} \boldsymbol{H}^T(\boldsymbol{m}_k^-) + \mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T\right\} \boldsymbol{L}_k^T \tag{9.269}$$

Assuming that the state is not correlated to the measurement noise, i.e.

$$\mathrm{E}\left\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)\boldsymbol{v}_k^T\right\} = \boldsymbol{0} \tag{9.270}$$

it follows that the cross-covariance for nonlinear measurements with additive noise is

$$\boldsymbol{C}_k = \boldsymbol{P}_k^- \boldsymbol{H}^T(\boldsymbol{m}_k^-) \tag{9.271}$$

Finally, consider the residual (innovations) covariance, which is defined to be

$$\boldsymbol{W}_k = \mathrm{E}\left\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\right\} \tag{9.272}$$

Using the previously developed result of

$$\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k = \boldsymbol{H}(\boldsymbol{m}_k^-)(\boldsymbol{x}_k - \boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{v}_k \tag{9.273}$$

and recalling the previous properties/assumptions that

- $\boldsymbol{H}(\boldsymbol{m}_k^-)$ and $\boldsymbol{L}_k$ are deterministic
- the state is not correlated with the measurement noise
- the covariance of the measurement noise is given by $\boldsymbol{R}_k$

gives the innovations covariance for nonlinear measurements as

$$\boldsymbol{W}_k = \boldsymbol{H}(\boldsymbol{m}_k^-)\boldsymbol{P}_k^- \boldsymbol{H}^T(\boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{R}_k \boldsymbol{L}_k^T$$

To summarize, we put everything together in a single table

System Model	$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{M}(t)\boldsymbol{w}(t)$
Meas. Model	$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{L}_k \boldsymbol{v}_k$
Init. Cond.	$\boldsymbol{m}_0 = \mathrm{E}\{\boldsymbol{x}(t_0)\}$
	$\boldsymbol{P}_0 = \mathrm{E}\left\{(\boldsymbol{x}(t_0) - \boldsymbol{m}_0)(\boldsymbol{x}(t_0) - \boldsymbol{m}_0)^T\right\}$
Mean Prop.	$\dot{\boldsymbol{m}}(t) = \boldsymbol{f}(\boldsymbol{m}(t))$
Cov. Prop.	$\dot{\boldsymbol{P}}(t) = \boldsymbol{F}(\boldsymbol{m}(t))\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(\boldsymbol{m}(t))$
	$+\boldsymbol{M}(t)\boldsymbol{Q}_s(t)\boldsymbol{M}^T(t)$
Exp. Meas.	$\hat{\boldsymbol{z}}_k = \boldsymbol{h}(\boldsymbol{m}_k^-)$
Innov. Cov.	$\boldsymbol{W}_k = \boldsymbol{H}(\boldsymbol{m}_k^-)\boldsymbol{P}_k^- \boldsymbol{H}^T(\boldsymbol{m}_k^-) + \boldsymbol{L}_k \boldsymbol{R}_k \boldsymbol{L}_k^T$
Cross Cov.	$\boldsymbol{C}_k = \boldsymbol{P}_k^- \boldsymbol{H}^T(\boldsymbol{m}_k^-)$
Kalman Gain	$\boldsymbol{K}_k = \boldsymbol{C}_k \boldsymbol{W}_k^{-1}$
Mean Upd.	$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)$
Cov. Upd.	$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k \boldsymbol{K}_k^T - \boldsymbol{K}_k \boldsymbol{C}_k^T + \boldsymbol{K}_k \boldsymbol{W}_k \boldsymbol{K}_k^T$

9.2.1 Example: Falling Body

Here we will look at a classic example of the EKF as presented by Gelb (1974).

Consider the problem of tracking a body falling freely through the atmosphere.

The motion is modeled in one dimension by assuming the body falls in a straight line, directly toward a tracking radar.

A radar return is received every 0.1 [sec].

The state is defined as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \\ \beta \end{bmatrix} \tag{9.274}$$

where x is the height of the falling body above the earth and β is the ballistic coefficient of the object.

The equations of motion for the body are given by

$$\dot{\boldsymbol{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ d-g \\ 0 \end{bmatrix} = \boldsymbol{f}(\boldsymbol{x}) \tag{9.275}$$

with

$$d = \frac{\rho x_2^2}{2x_3} \qquad \text{and} \qquad \rho = \rho_0 \exp\left\{-\frac{x_1}{k_\rho}\right\} \tag{9.276}$$

where d is drag acceleration, g is the acceleration of gravity, ρ is atmospheric density (with ρ_0 as the atmospheric density at sea level), and k_ρ is a decay constant.

The differential equation governing x_2 (velocity) is nonlinear through the dependence of drag on velocity, air density, and ballistic coefficient.

We will assume to take linear measurements of height which are corrupted according to $p_g(0,R)$.

The initial truth for the simulation is drawn according to

$$x_0 = p_g(10^5 \text{ ft}, 500 \text{ ft}^2) \tag{9.277}$$
$$\dot{x}_0 = p_g(-6000 \text{ ft/sec}, 2\times 10^4 \text{ ft}^2/\text{sec}^2) \tag{9.278}$$
$$\beta = p_g(2000 \text{ lb/ft}^2, 2.5\times 10^5 \text{ lb}^2/\text{ft}^4) \tag{9.279}$$

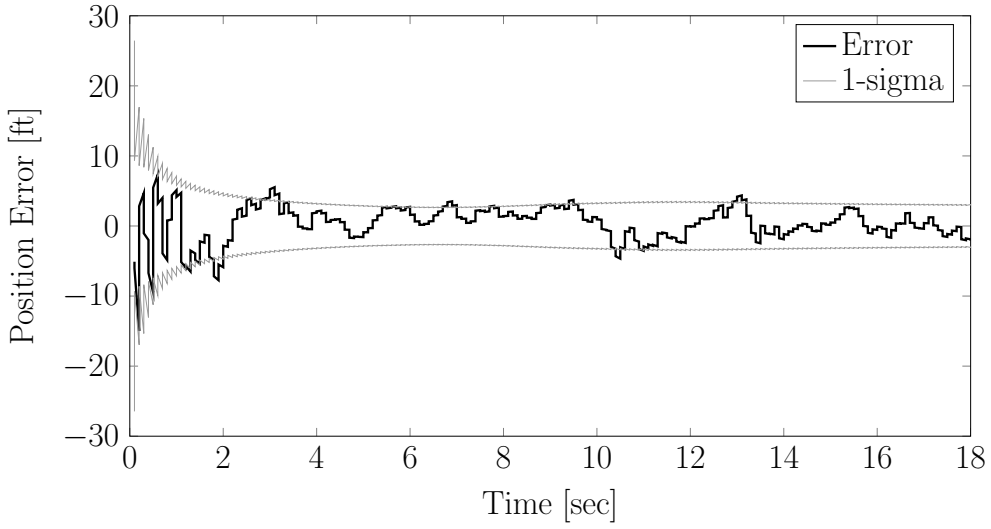
The initial mean and covariance are taken to be

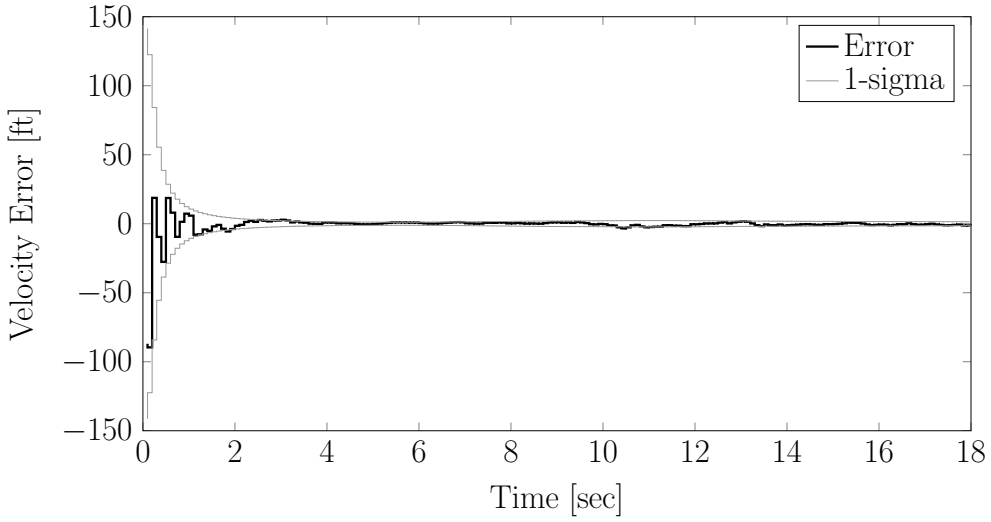
$$\boldsymbol{m}_0 = \begin{bmatrix} 10^5 \text{ ft} \\ -6000 \text{ ft/sec} \\ 2000 \text{ lb/ft}^2 \end{bmatrix} \tag{9.280}$$
$$\boldsymbol{P}_0 = \begin{bmatrix} 500 \text{ ft}^2 & 0 & 0 \\ 0 & 2\times 10^4 \text{ ft}^2/\text{sec}^2 & 0 \\ 0 & 0 & 2.5\times 10^5 \text{ lb}^2/\text{ft}^4 \end{bmatrix} \tag{9.281}$$

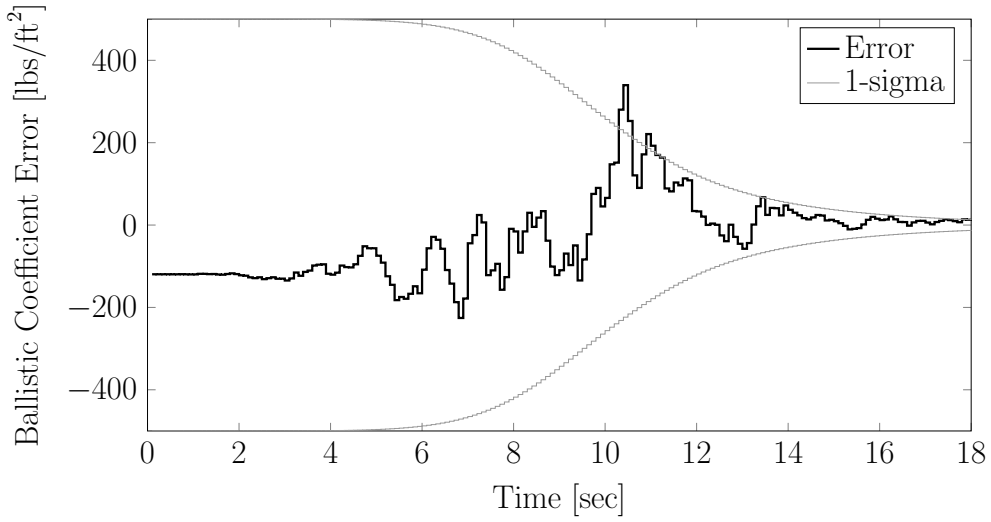
The system parameters are taken to be

$$\rho_0 = 3.4\times 10^{-3} \text{ lb sec}^2/\text{ft}^4 \qquad g = 32.2 \text{ ft/sec}^2 \tag{9.282}$$
$$k_\rho = 22000 \text{ ft} \qquad R = 100 \text{ ft}^2 . \tag{9.283}$$

The resulting error for each state variable and their associated 1σ bounds (from the square root of the corresponding entry in the covariance matrix) can be seen below.







Directing our attention to the errors in estimating the ballistic coefficient, we note that the EKF does not track β accurately in the early stages of tracking.

Physically, this is due to the fact that the thin atmosphere at high altitude produces a small drag force on the body.

The thicker atmosphere, creating an increased drag force, enables the EKF to achieve lower estimation error for β.

How did we code this extended Kalman filter in MATLAB? First, set the random number seed and define the relevant system and simulation parameters.

```
% Set the random number seed
rng(100);

% System parameters
rho0 = 3.4e-3;
g    = 32.2;
krho = 22000;
Rk   = 100;
Lk   = 1;
```

Define the initial truth and estimates.

```
% Initial diagonal entries of covariance matrix
P110 = 500;
P220 = 2e4;
P330 = 2.5e5;

% Initial truth and estimate
x0 = [1e5+sqrt(P110)*randn; -6000+sqrt(P220)*randn; 2000+sqrt(P330)*randn];
m0 = [1e5; -6000; 2000];
P0 = diag([500, 2e4, 2.5e5]);
```

We know we have linear measurements of x_1, so construct the measurement mapping matrix $\boldsymbol{H}_k$.

```
% We have linear measurements, mapped according to Hk
Hk = [1, 0, 0];
```

Create a time vector with intervals corresponding to the desired measurement frequency.

```
% Time vector
tv = 0:0.1:18;
```

Initialize variables for the truth and mean/covariance to iterate over. Also, define a counting variable as before (just so we can save stuff to look at later).

```
% The extended Kalman filter
mkm1 = m0;
Pkm1 = P0;
xkm1 = x0;
cnt  = 1;
```

We have a differential equation governing the motion of the object. We will use ODE45 again to propagate the truth and our estimates forward in time. This requires settings for the integration tolerances.

```
opts = odeset('AbsTol',1e-6,'RelTol',1e-6);
```

Start the loop and propagate the truth with our integrator. The equations of motion are defined in a file called `one_dim_eoms.m` that we will describe a bit later.

```
for i = 2:length(tv)
    % Propagate truth
    [~,X] = ode45(@one_dim_eoms,[tv(i-1),tv(i)],xkm1,opts,g,rho0,krho);
    xtk   = X(end,1:3)';
```

Generate an observation from this truth by corrupting it with measurement noise.

```
        % Generate a measurement
        zk  = xtk(1) + sqrt(Rk)*randn;
```

Use the EKF to predict (again, the equations of motion file will be outlined later). Then, store the results so we can look at it once the loop completes. Note that the mean and covariance propagation equations are coupled, so the dynamics Jacobian *must* be computed online and inside of the equations of motion function. Also, note that we do not have process noise here, so there's no process noise PSD passed into the integrator.

```
% Predict
[~,X] = ode45(@one_dim_eoms,[tv(i-1),tv(i)],[mkm1;Pkm1(:)],opts,g,rho0,krho);
mkm   = X(end,1:3)';
Pkm   = reshape(X(end,4:end),3,3);
% Store a priori mean and covariance
zkp(:,cnt)   = zk;
xtp(:,cnt)   = xtk;
mpt(:,cnt)   = mkm;
Ppt(:,:,cnt) = Pkm;
cnt          = cnt + 1;
```

Perform an update according to the Kalman filter equations and our measurement and store.

```
        % Correct
        zht = Hk*mkm;
        Wk  = Hk*Pkm*Hk' + Lk*Rk*Lk';
        Ck  = Pkm*Hk';
        Kk  = Ck/Wk;
        mkp = mkm + Kk*(zk - zht);
        Pkp = Pkm - Ck*Kk' - Kk*Ck' + Kk*Wk*Kk';
        % Store a posteriori mean and covariance
        zkp(:,cnt)   = zk;
        xtp(:,cnt)   = xtk;
        mpt(:,cnt)   = mkp;
        Ppt(:,:,cnt) = Pkp;
        cnt          = cnt + 1;
```

Finally, cycle the variables and end the loop.

```
        % Cycle
        xkm1 = xtk;
        mkm1 = mkp;
        Pkm1 = Pkp;
    end
```

The equations of motion file can be seen below.

```
function [dxdt] = one_dim_eoms(t,x,g,rho0,krho)
% Density
rho = rho0*exp(-x(1)/krho);
% Rates of change
dx1 = x(2);
dx2 = rho*x(2)^2/(2*x(3)) - g;
dx3 = 0;
% Concatenate
dxdt = [dx1; dx2; dx3];
```

and:

```
% Stop here if we are simply propagating truth.
if length(x) > 3
    % Evaluate dynamics Jacobian
    F = [0, 1, 0;
        -(rho0*x(2)^2*exp(-x(1)/krho))/(2*krho*x(3)), ...
         (rho0*x(2)*exp(-x(1)/krho))/x(3),            ...
        -(rho0*x(2)^2*exp(-x(1)/krho))/(2*x(3)^2);
          0, 0, 0];

    P  = reshape(x(4:end),3,3);
    dP = F*P + P*F';
    % Add a vectorized rate of change to the rate of change vector
    dxdt = [dxdt; dP(:)];
end
```

9.2.2 A Few Important Points

There is a very important and often overlooked difference between the Kalman filter and the extended Kalman filter.

The gain, $\boldsymbol{K}_k$, employed in the EKF is actually a random variable, whereas the gain in the Kalman filter is not.

This stems from the fact that we have chosen to linearize our dynamics and our measurement about the current conditional mean.

Once we update the mean and covariance with a single measurement, which has some random noise included in it, the mean becomes a random variable.

In the Kalman filter, this is equally true, but there is no linearization performed.

In the EKF, linearizing about the mean yields Jacobian matrices that are functions of the mean.

Since the mean is random, the Jacobians become random. Because of this, the Kalman gain becomes random.

It is important to note that we assumed that the Jacobians were deterministic, and we did this several times:

- computing the covariance evolution (not explicitly shown)
- computing the expected measurement
- computing the cross-covariance
- computing the residual covariance
- computing the posterior covariance (not explicitly shown)
- computing the Kalman gain

Additionally, the covariance becomes random as well.

This implies that the EKF is trajectory-dependent, meaning that its inherent accuracy is functionally dependent upon the trajectory and the sequence of measurements employed.

The Kalman filter, by contrast, is not trajectory dependent. The Kalman gain can be computed off-line, and the covariance is not a function of the trajectory.

For the EKF, all calculations must be performed on-line.

9.2.3 Example of an EKF to an Orbit Problem

In this example, we will look at the coding and operation of an extended Kalman filter (EKF) to the problem of a satellite determining its position and velocity using an altimeter.

We'll keep the modeling pretty simple by using two-body mechanics for the motion of the vehicle and using a simple measurement of the altitude to a spherical Earth.

That is, we will take the dynamics to be

$$\begin{aligned}\dot{\boldsymbol{r}}(t) &= \boldsymbol{v}(t)\\ \dot{\boldsymbol{v}}(t) &= -\frac{\mu}{\|\boldsymbol{r}(t)\|^3}\boldsymbol{r}+\boldsymbol{w}(t)\end{aligned}$$

where $\boldsymbol{r}$ is the position of the vehicle in the inertial frame, $\boldsymbol{v}$ is the velocity of the vehicle in the inertial frame, μ is the gravitational parameter, and $\boldsymbol{w}$ represents some process noise that is injected into the dynamics. We assume that the process noise is zero mean with power spectral density $\boldsymbol{Q}_s$.

Additionally, we take the measurements to be

$$z_k = \|\boldsymbol{r}_k\| - R_e + v_k$$

where the subscript k denotes the discrete time of the measurements, R_e is the spherical radius of the Earth, and v_k is zero-mean measurement noise with covariance R_k.

We'll be applying an EKF here, so we will ultimately need to propagate the mean and the covariance, which are given by

$$\begin{aligned}\dot{\boldsymbol{m}}(t) &= \boldsymbol{f}(\boldsymbol{m}(t)) \\ \dot{\boldsymbol{P}}(t) &= \boldsymbol{F}(\boldsymbol{m}(t)\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}^T(\boldsymbol{m}(t)) + \boldsymbol{M}(t)\boldsymbol{Q}_s\boldsymbol{M}^T(t)\end{aligned}$$

where $\boldsymbol{f}(\cdot)$ follows from our definition of the dynamics, $\boldsymbol{F}(\cdot)$ is the dynamics Jacobian, and $\boldsymbol{M}(t)$ is the noise-mapping matrix.

For the update stage, we apply the EKF as

$$\begin{aligned}\boldsymbol{m}_k^+ &= \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \\ \boldsymbol{P}_k^+ &= \boldsymbol{P}_k^- - \boldsymbol{K}_k\boldsymbol{H}(\boldsymbol{m}_k^-)\boldsymbol{P}_k^-\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{z}}_k^- &= \boldsymbol{h}(\boldsymbol{m}_k^-) \\ \boldsymbol{C}_k &= \boldsymbol{P}_k^-\boldsymbol{H}^T(\boldsymbol{m}_k^-) \\ \boldsymbol{W}_k &= \boldsymbol{H}(\boldsymbol{m}_k^-)\boldsymbol{P}_k^-\boldsymbol{H}^T(\boldsymbol{m}_k^-) + \boldsymbol{R}_k \\ \boldsymbol{K}_k &= \boldsymbol{C}_k\boldsymbol{W}_k^{-1}\end{aligned}$$

Here, $\boldsymbol{h}(\cdot)$ follows from our definition of the measurement and $\boldsymbol{H}(\cdot)$ is the measurement Jacobian.

Let's start our code by clearing out everything and setting a random number seed so that we can repeat our run.

```
clear all
close all
clc

% set a random number seed
rng(200)
```

It also helps to establish the constants that we'll use in this example, specifically the gravitational parameter and the radius of the Earth.

```
GM = 3.986004415e5;
Re = 6378.136;
```

We'll run the EKF for a duration of one hour, and we'll generate data twice a minute, so we create some timing information.

```
% timing parameters
t0 =    0.0;
dt =   30.0;
tf = 3600.0;
tv = t0:dt:tf;
```

Of course, we don't have to have regularly spaced times; this is just for a matter of convenience here.

We need an initial mean and covariance. First of all, we're going to work on a planar case to reduce the dimensionality of the problem and to simplify our analysis; this, too, is for convenience.

Let's take the initial mean to be in a circular orbit of altitude 700 km, and let's take our initial position standard deviations to be 100 meters and our initial velocity standard deviations to be 1 m/s.

```
% specify initial mean/covariance and generate random truth
nx = 4;
h0 = 700.0;
m0 = [Re+h0;0.0;0.0;sqrt(GM/(Re+h0))];
P0 = diag([0.1;0.1;1e-3;1e-3].^2);
x0 = m0 + chol(P0)'*randn(nx,1); % truth is random
```

The final line draws an initial true state under the assumption that our mean and covariance describe a Gaussian distribution.

We need to specify the power of the process noise, which is done via the power spectral density matrix. Note that this is a 2×2 matrix in this case since noise is only injected into the velocity dynamics and we are working in a planar problem.

```
% specify the power spectral density of the process noise
nq = 2;
Qs = (1e-9)^2*eye(nq);
```

Furthermore, the measurement noise covariance is taken to have a standard deviation of 10 m, so we will have fairly accurate altitude data to process.

```
% specify the measurement noise covariance
nr = 1;
Rk = (10.0e-3)^2*eye(nr);
```

Ultimately, we're going to want to analyze the performance of our EKF, so we need to declare some storage space for saving the estimation error and the standard deviation (from the filter covariance) of the estimation error. We can also go ahead and compute and store the initial error and standard deviation.

```
% storage for error and standard deviation plotting
tplot        = zeros( 1,2*length(tv)-1);
eplot        = zeros(nx,2*length(tv)-1);
splot        = zeros(nx,2*length(tv)-1);
rplot        = zeros(nr,  length(tv)-1);
wplot        = zeros(nr,  length(tv)-1);
ctr          = 1;
tplot(:,ctr) = t0;
eplot(:,ctr) = x0 - m0;
splot(:,ctr) = sqrt(diag(P0));
```

We'll be using ODE45 to integrate our true state and to propagate the mean and covariance for our EKF, so we can set some absolute/relative tolerances for the integrator.

```
% integrator options
opts = odeset('AbsTol',1e-9,'RelTol',1e-9);
```

The last item before starting the propagation and update for the EKF is to initialize the filter. We will also have to generate the truth online, so we set the initial time, true state, mean, and covariance before beginning our time loop. The `km1` description is to be read as "k minus 1".

```
% initial time, state, mean, and covariance
tkm1 = t0;
xkm1 = x0;
mkm1 = m0;
Pkm1 = P0;
```

Now, we start the time loop. We start the index at 2 since we already know everything at the index 1 and extract the time at the current index.

```
% begin the time loop
for k = 2:length(tv)
    % extract the time
    tk = tv(k);
```

First, we need to propagate our true state from time t_{k-1} to time t_k. This will tell us where the true object is and its velocity at time t_k. Don't forget to add noise into the dynamics!

```
    % propagate true state (with noise) from tkm1 to tk
    wkm1    = chol(Qs)'*randn(nq,1);
    [~,XX] = ode45(@eom_ptbp,[tkm1,tk],xkm1,opts,GM,wkm1);
    xk      = XX(end,:)';
```

Next, we generate a true measurement of the altitude of the vehicle and add some measurement noise to the true altitude.

```
    % generate true measurement (with noise) at time tk
    zk = norm(xk(1:2)) - Re + chol(Rk)'*randn(nr,1);
```

We can now perform our EKF propagation stage. The mean and a column vector form of the covariance are concatenated to be passed into the integrator, and the power spectral density is also sent into the integrator. We then pull the first 4 states from the result of the integrator as the mean and reshape the last 16 elements from the integrator as the covariance. The final line is a brute-force command to ensure that the propagated covariance matrix is symmetric, which is just to ensure proper conditioning of the covariance matrix.

```
        % propagate mean and covariance from time tkm1 to tk
        [~,XX] = ode45(@eom_ptbp_ekf,[tkm1,tk],[mkm1;Pkm1(:)],opts,GM,Qs);
        mkm    = XX(end,1:nx)';
        Pkm    = reshape(XX(end,nx+1:end)',nx,nx);
        Pkm    = 0.5*(Pkm + Pkm');
```

Now, we update the state using our measurement data. This requires computing the expected measurement, the measurement Jacobian, the cross-covariance, the residual covariance, and the Kalman gain before updating the mean and covariance. Once again, we enforce symmetry in a brute-force manner to ensure that the posterior covariance matrix is symmetric.

```
        % update mean and covariance at time tk
        zhatk = norm(mkm(1:2)) - Re;
        Hk    = [mkm(1:2)'./norm(mkm(1:2)),zeros(1,2)];
        Ck    = Pkm*Hk';
        Wk    = Hk*Pkm*Hk' + Rk;
        Kk    = Ck/Wk;
        mkp   = mkm + Kk*(zk - zhatk);
        Pkp   = Pkm - Kk*Hk*Pkm;
        Pkp   = 0.5*(Pkp + Pkp');
```

We're really done with the EKF now, but we also want to be able to analyze the results, so we compute and store the prior and posterior estimation errors (the truth minus the mean) and the standard deviations reported by the filter covariance.

```
        % store the a prior/a posteriori error and standard deviation
        ctr          = ctr + 1;
        tplot(:,ctr) = tk;
        eplot(:,ctr) = xk - mkm;
        splot(:,ctr) = sqrt(diag(Pkm));
        ctr          = ctr + 1;
        tplot(:,ctr) = tk;
        eplot(:,ctr) = xk - mkp;
        splot(:,ctr) = sqrt(diag(Pkp));
```

We can do something similar with the residual and the residual covariance. For a filter applied in "real life," the residual and the residual covariance are known to us, whereas the estimation error is not. Therefore, this is an *essential* step in verifying that our EKF is functioning properly.

```
        % store the residual and its standard deviation
        rplot(:,k-1) = zk - zhatk;
        wplot(:,k-1) = sqrt(Wk);
```

The final step before ending the time loop is to cycle our time, state, mean, and covariance to prepare for the next time step.

```
        % cycle the time, state, mean, and covariance
        tkm1 = tk;
        xkm1 = xk;
        mkm1 = mkp;
        Pkm1 = Pkp;
    end
```

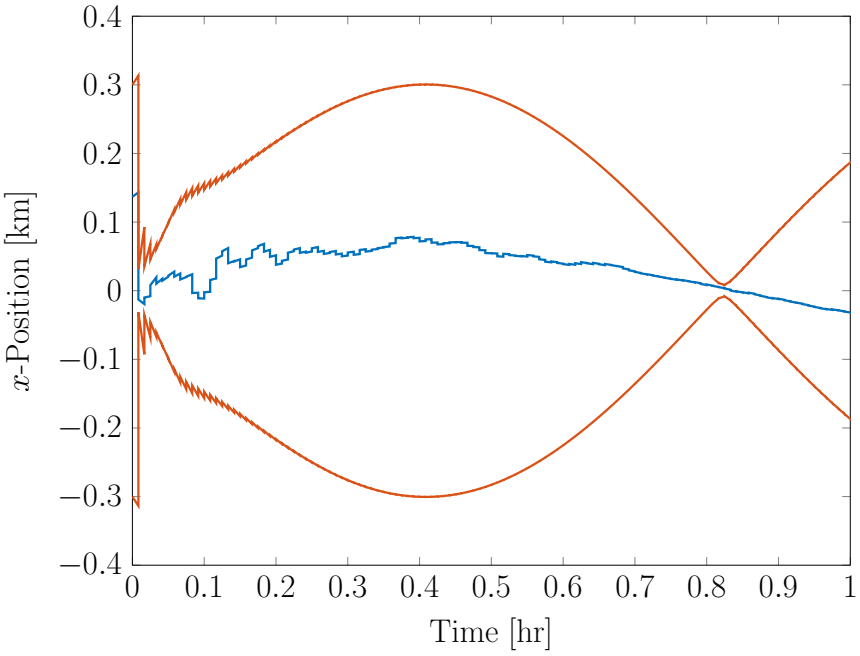
To analyze the filter, we're going to plot the estimation errors and the 3σ "bound" from the standard deviations that we computed from the filter covariance. These aren't really bounds, since there is some probability (albeit, not much) that the error can be outside of the 3σ interval.

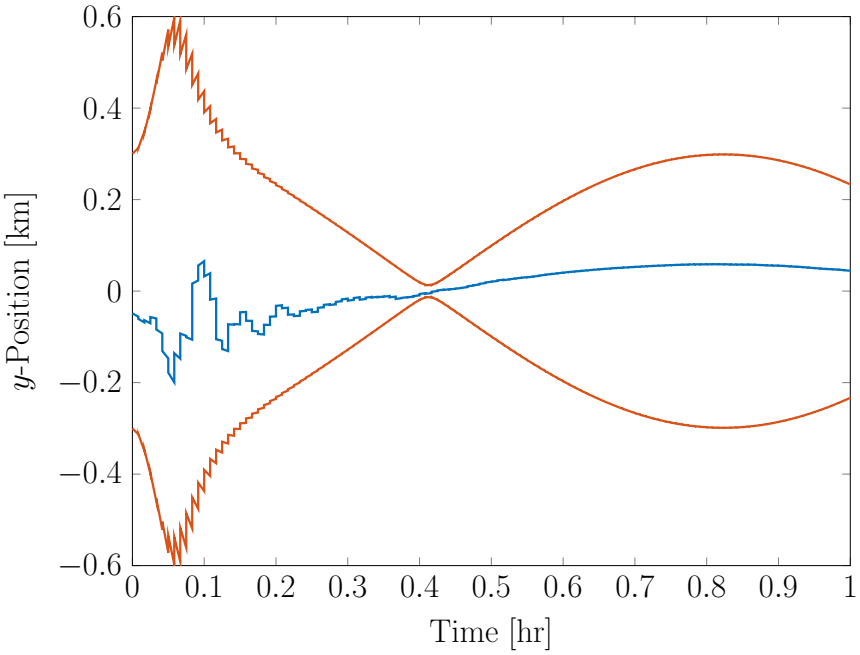
```
% plot the estimation errors and 3sigma "bounds"
xlab = {'Time [hr]','Time [hr]','Time [hr]','Time [hr]'};
ylab = {'$x$-Position [km]','$y$-Position [km]','$x$-Velocity [km/s]','$y$-Velicity [km
for i = 1:nx
    figure
    C = get(gca,'ColorOrder');
    plot(tplot(1,:)./3600.0,      eplot(i,:),'Color',C(1,:),'LineWidth',1.2);
    hold on
    plot(tplot(1,:)./3600.0,+3.0*splot(i,:),'Color',C(2,:),'LineWidth',1.2);
    plot(tplot(1,:)./3600.0,-3.0*splot(i,:),'Color',C(2,:),'LineWidth',1.2);
    xlabel(xlab{i})
    ylabel(ylab{i})
end
```

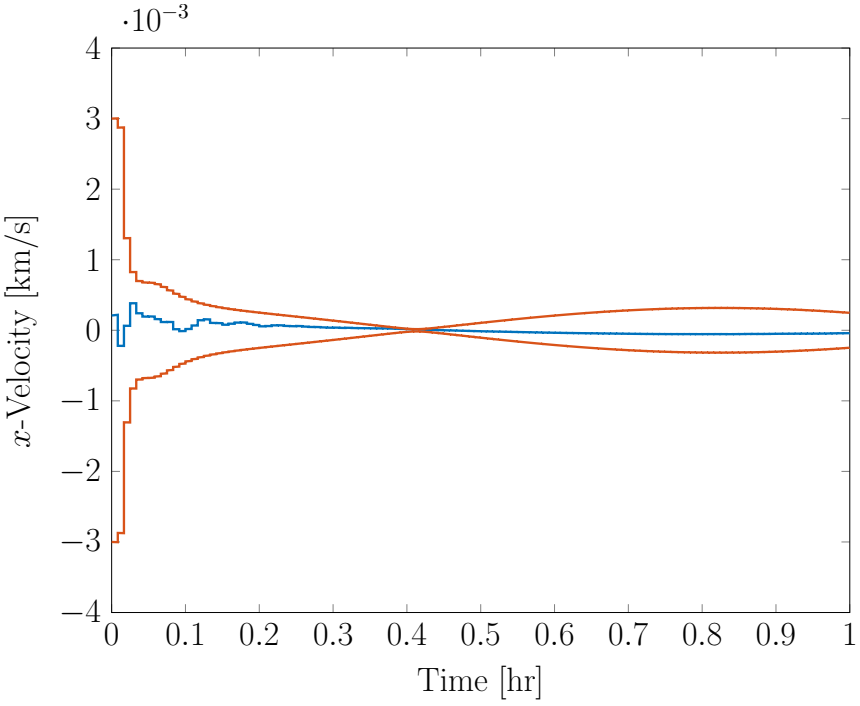
And finally, we do the same thing for the measurement residual and its standard deviation.

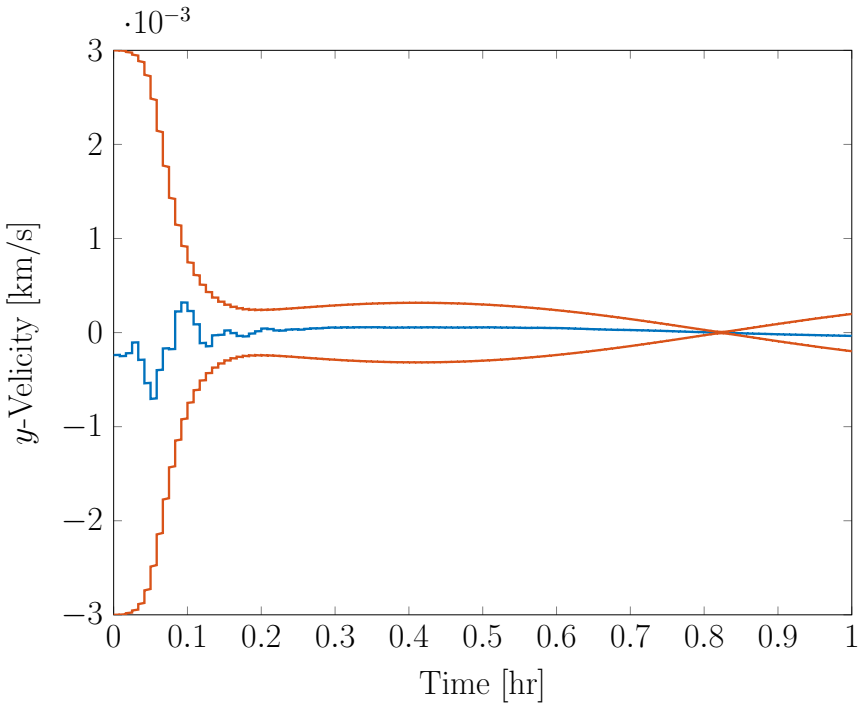
```
% plot the measurement residuals and 3sigma "bounds"
figure
C = get(gca,'ColorOrder');
plot(tv(2:end)./3600.0,      rplot(1,:),'x','Color',C(1,:))
hold on
plot(tv(2:end)./3600.0,+3.0*wplot(1,:),'Color',C(2,:),'LineWidth',1.2)
plot(tv(2:end)./3600.0,-3.0*wplot(1,:),'Color',C(2,:),'LineWidth',1.2)
xlabel('Time [hr]')
ylabel('Residual [km]')
```

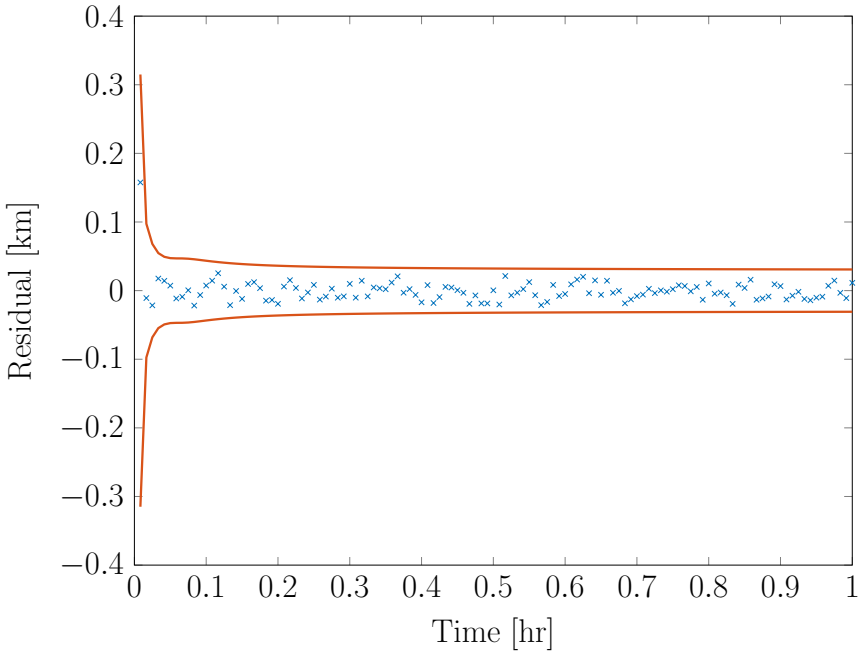
These results are summarized in the following plots.








$\cdot 10^{-3}$
3
2
1
0
−1
−2
−3
y-Velicity [km/s]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time [hr]


0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
Residual [km]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time [hr]

A few things are clear from the preceding plots.

Altitude measurements alone are not sufficient to provide precise tracking of the position and velocity.

At the same time, the altimeter allows a great improvement to the velocity tracking and it keeps all of the errors from just growing ever-larger.

We also see a cyclic behavior in the covariance (standard deviation), which is caused by the periodic nature of the two-body problem.

Finally, the residual plot illustrates that we are properly extracting the information out of the data since the residual covariance comes down to approximately 10 meters.

This is the value that we set for the measurement noise covariance.

The higher value of the residual covariance in the beginning is due to the combined effects of the measurement noise and the uncertainty in our state.

By the end of the run, we've basically removed any excess uncertainty in the measurements that is due to the uncertainty in the states and we are limited by the measurement noise itself.

9.3 Unscented Kalman Filter

9.3.1 Introduction

We briefly recall the dynamical and observational systems that we considered in the development of the Kalman filter.

For simplicity moving forward, we consider only the case of discrete dynamics accompanied by discrete measurements.

In this case, the dynamics and measurements are taken to have the form

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}) + \boldsymbol{w}_{k-1} \tag{9.284}$$

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \tag{9.285}$$

We have omitted the consideration of $\boldsymbol{M}_{k-1}$ and $\boldsymbol{L}_k$ just to simplify the notation moving forward, but it is straightforward to revise the following developments to include these mapping matrices.

Given initial values of the mean and covariance, $\boldsymbol{m}_0$ and $\boldsymbol{P}_0$, along with a sequence of data, $\boldsymbol{z}_k$, for $k = 1, 2, \ldots$, the objective is to sequentially propagate the mean and covariance between measurements and update the mean and covariance using the measurement data.

The prediction step operates on the mean and covariance at step $k-1$, given by $\boldsymbol{m}_{k-1}$ and $\boldsymbol{P}_{k-1}$ in order to determine the mean and covariance at time k.

By definition, the prediction stage is

$$\boldsymbol{m}_k = \mathrm{E}\{\boldsymbol{x}_k\} \tag{9.286}$$

$$\boldsymbol{P}_k = \mathrm{E}\{(\boldsymbol{x}_k - \mathrm{E}\{\boldsymbol{x}_k\})(\boldsymbol{x}_k - \mathrm{E}\{\boldsymbol{x}_k\})^T\} \tag{9.287}$$

If it is assumed that the process noise is zero mean with covariance $\boldsymbol{Q}_{k-1}$ and that the process noise is uncorrelated to the state, then it follows that

$$\boldsymbol{m}_k = \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\} \tag{9.288}$$

$$\boldsymbol{P}_k = \mathrm{E}\{(\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\})(\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\})^T\} + \boldsymbol{Q}_{k-1} \tag{9.289}$$

If we then perform a first-order Taylor series expansion of the nonlinear function about $\boldsymbol{m}_{k-1}$, we can arrive at the propagation equations for the mean and covariance that are used by the extended Kalman filter.

When considering measurement data, the mean and covariance update equations for a linear, unbiased estimator are

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.290}$$

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.291}$$

As a reminder, the terms involved here are

- $\hat{\boldsymbol{z}}_k$ is the predicted measurement
- $\boldsymbol{C}_k$ is the cross-covariance
- $\boldsymbol{W}_k$ is the residual covariance
- $\boldsymbol{K}_k$ is any linear gain

If we take the gain to be the gain that minimizes the mean square *a posteriori* error, then we get the Kalman gain as

$$\boldsymbol{K}_k = \boldsymbol{C}_k\boldsymbol{W}_k^{-1} \tag{9.292}$$

The expected value of the measurement, cross-covariance (with the measurement), and measurement covariance are defined as

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{z}_k\} \tag{9.293}$$
$$\boldsymbol{C}_k = \mathrm{E}\{(\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\} \tag{9.294}$$
$$\boldsymbol{W}_k = \mathrm{E}\{(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k)^T\} \tag{9.295}$$

If the measurement noise is taken to be zero mean with covariance $\boldsymbol{R}_k$ and to be uncorrelated with the state, then it follows that

$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\} \tag{9.296}$$
$$\boldsymbol{C}_k = \mathrm{E}\{(\boldsymbol{x}_k - \mathrm{E}\{\boldsymbol{x}_k\})(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})^T\} \tag{9.297}$$
$$\boldsymbol{W}_k = \mathrm{E}\{(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})^T\} + \boldsymbol{R}_k \tag{9.298}$$

Note that we have replaced $\boldsymbol{m}_k^-$ with $\mathrm{E}\{\boldsymbol{x}_k\}$ with the understanding that this is the expectation prior to the incorporation of the measurement data.

Thus, minimum mean square error estimation requires the calculation of five expectations:

$$\boldsymbol{m}_k = \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\} \tag{9.299}$$
$$\boldsymbol{P}_k = \mathrm{E}\{(\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\})(\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \mathrm{E}\{\boldsymbol{f}(\boldsymbol{x}_{k-1})\})^T\} + \boldsymbol{Q}_{k-1} \tag{9.300}$$
$$\hat{\boldsymbol{z}}_k = \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\} \tag{9.301}$$
$$\boldsymbol{C}_k = \mathrm{E}\{(\boldsymbol{x}_k - \mathrm{E}\{\boldsymbol{x}_k\})(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})^T\} \tag{9.302}$$
$$\boldsymbol{W}_k = \mathrm{E}\{(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})(\boldsymbol{h}(\boldsymbol{x}_k) - \mathrm{E}\{\boldsymbol{h}(\boldsymbol{x}_k)\})^T\} + \boldsymbol{R}_k \tag{9.303}$$

When looking at the EKF, all five of these expectations are computed by linearizing the nonlinear function about the mean.

In general, however, recalling the definition of the expected value, we can view these five expectations as the integral equations

$$\boldsymbol{m}_k = \int \boldsymbol{f}(\boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1}) d\boldsymbol{x}_{k-1} \tag{9.304}$$
$$\boldsymbol{P}_k = \int (\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \boldsymbol{m}_k)(\boldsymbol{f}(\boldsymbol{x}_{k-1}) - \boldsymbol{m}_k)^T p(\boldsymbol{x}_{k-1}) d\boldsymbol{x}_{k-1} + \boldsymbol{Q}_{k-1} \tag{9.305}$$
$$\hat{\boldsymbol{z}}_k = \int \boldsymbol{h}(\boldsymbol{x}_k) p(\boldsymbol{x}_k) d\boldsymbol{x}_k \tag{9.306}$$
$$\boldsymbol{C}_k = \int (\boldsymbol{x}_k - \boldsymbol{m}_k^-)(\boldsymbol{h}(\boldsymbol{x}_k) - \hat{\boldsymbol{z}}_k)^T p(\boldsymbol{x}_k) d\boldsymbol{x}_k \tag{9.307}$$

$$\boldsymbol{W}_k = \int (\boldsymbol{h}(\boldsymbol{x}_k) - \hat{\boldsymbol{z}}_k)(\boldsymbol{h}(\boldsymbol{x}_k) - \hat{\boldsymbol{z}}_k)^T p(\boldsymbol{x}_k) d\boldsymbol{x}_k + \boldsymbol{R}_k \tag{9.308}$$

The Kalman filter framework can be applied, in general, provided that we can evaluate these integrals.

Each of the integrals required is of the form

$$I = \int \boldsymbol{g}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.309}$$

A "simple" procedure for approximating these expectation integrals is to use Monte Carlo integration

$$I \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{x}^{(i)}) \tag{9.310}$$

where there are N statistically sampled points, $\boldsymbol{x}^{(i)}$, drawn from the density $p(\boldsymbol{x})$.

This method is very general in nature, but it converges as $\sqrt{N}$.

A general rule of thumb is to use $N = 10^n$ sample points. Even just working with position and velocity would require 10^6 points.

This curse of dimensionality is what we wish to avoid with the class of Gaussian nonlinear filters.

Effectively, we want to find a way to use a point-based approach, but we want to choose our points intelligently.

We also note that each of the five expectations/integrals may be viewed as computing statistics of the nonlinear transformation

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) \tag{9.311}$$

where the mean and covariance of $\boldsymbol{x}$ are known and we want to compute

1. the mean of $\boldsymbol{y}$

2. the covariance of $\boldsymbol{y}$

3. the cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$

How does this apply to the propagation and update stages?

For the propagation stage, our nonlinear function is $\boldsymbol{f}(\boldsymbol{x}_{k-1})$, and the mean and covariance of $\boldsymbol{y}$ are the propagated mean and covariance.

For the update stage, our nonlinear function is $\boldsymbol{h}(\boldsymbol{x}_k)$, and the mean, covariance, and cross-covariance of $\boldsymbol{y}$ are required to compute the Kalman gain, covariance update, and mean update.

Thus, if we can compute statistics through this general nonlinear function, then we can implement the Kalman filter framework directly.

9.3.2 The Unscented Kalman Filter

The unscented transform (UT) is a relatively recent numerical method that can also be used for approximating the joint distribution of random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ defined as

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}) \tag{9.312}$$
$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) \tag{9.313}$$

where $\mathcal{N}(\boldsymbol{m}, \boldsymbol{P})$ denotes the multivariate Gaussian distribution of mean $\boldsymbol{m}$ and covariance $\boldsymbol{P}$, and $\boldsymbol{g}(\boldsymbol{x})$ denotes a potentially nonlinear function that maps the random variable $\boldsymbol{x}$ into the random variable $\boldsymbol{y}$.

Note that in the following discussion, we will consider some generic nonlinear transformation $\boldsymbol{g}(\boldsymbol{x})$ to approach the UT in general.

Once we start applying the technique to estimation for dynamic systems, this nonlinear transformation will take the form of our system dynamics and measurement model!

Where methods such as linearization and statistical linearization attempt to approximate the behavior of $\boldsymbol{g}(\boldsymbol{x})$ (via Taylor Series), the UT instead attempts to match the first and second moments of the target distribution (i.e. the target mean and covariance).

That is, instead of approximating the nonlinear function, we are attempting to approximate moments of the distribution.

The whole idea of the UT is to deterministically choose a fixed number of so-called "sigma points" to capture the mean and covariance of the original distribution of $\boldsymbol{x}$ exactly.

These sigma points are then subjected to the nonlinear function, and mean and covariance are then extracted from these transformed points.

Note that while this may feel very reminiscent to a sequential Monte Carlo approach, they are in fact very different approaches!

The difference is in that the points in the UT are chosen deterministically and *not* randomly.

A note of importance: the following discussion will be in the development of a Gaussian approximation (utilizing the UT) to the previously discussed Kalman filtering equations. The assumption of Gaussianity is *not* required! It is simply an approach that makes the interpretation of the UT much easier.

First, however, we will need to define the matrix square root factor.

The concept of the square root of a number extends to matrices, and filtering approaches that employ these matrix square roots enjoy many benefits both in utility and computational stability.

We will define a matrix square root factor for some square matrix $\boldsymbol{A}$ as any matrix that satisfies

$$\sqrt{\boldsymbol{A}}\sqrt{\boldsymbol{A}}^T = \boldsymbol{A} \tag{9.314}$$

The question then is, how does one compute the square root of a matrix?

At first, it may be tempting to use the MATLAB command `sqrtm`, but beware!

This returns a matrix $\sqrt{\boldsymbol{A}}$ which satisfies $\sqrt{\boldsymbol{A}}\sqrt{\boldsymbol{A}} = \boldsymbol{A}$ and **not** $\sqrt{\boldsymbol{A}}\sqrt{\boldsymbol{A}}^T = \boldsymbol{A}$.

There are several methods for finding the square root factor of a matrix, but we will discuss the two most common methods.

The first method that is commonly employed is eigen-decomposition (or spectral decomposition) of $\boldsymbol{A}$.

That is, the matrix $\boldsymbol{A}$ is decomposed into the form

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{9.315}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{V}$ is an orthogonal matrix containing its corresponding eigenvectors.

As we are looking for the square root factor of $\boldsymbol{A}$, we are interested in finding $\sqrt{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T}$.

It turns out that

$$\sqrt{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T} = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}} \tag{9.316}$$

This can be proven by simply squaring both sides (but let's just look at the right hand side):

$$\left[\boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\right]\left[\boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\right]^T = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}^T\boldsymbol{V}^T \tag{9.317}$$

$$= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{9.318}$$

$$= \boldsymbol{A} \tag{9.319}$$

This is exactly the result we expect.

Another interesting thing about matrix square root factors is that there are an infinite number of them.

For instance,

$$\sqrt{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T} = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{V}^T \tag{9.320}$$

is also a valid square root factor of the matrix $\boldsymbol{A}$.

In fact, for any matrix $\boldsymbol{U}$ such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ (meaning that $\boldsymbol{U}$ is orthonormal)

$$\sqrt{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T} = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T \tag{9.321}$$

is a valid square root factor of the matrix $\boldsymbol{A}$.

This is shown as follows:

$$\left[\boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T\right]\left[\boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T\right]^T = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^T\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}^T\boldsymbol{V}^T \tag{9.322}$$

$$= \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}^T\boldsymbol{V}^T \tag{9.323}$$

$$= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{9.324}$$

$$= \boldsymbol{A} \tag{9.325}$$

This means that we can obtain a valid square root factor for $\boldsymbol{A}$ as

$$\sqrt{\boldsymbol{A}} = \sqrt{\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T} \tag{9.326}$$

$$= \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}} \tag{9.327}$$

Since $\boldsymbol{\Lambda}$ is a diagonal matrix, its square root is the matrix which contains the square roots of its diagonal entries on its own diagonal. This is easy to compute!

Note that if we require that $\boldsymbol{A}$ is positive-definite and symmetric (as is always the case with a covariance matrix), we should see no issues with negative or complex eigenvalues.

This method can be implemented in MATLAB with the `eig` function.

The second method that is commonly used by many is known as the Cholesky factor of a matrix $\boldsymbol{A}$.

As opposed to the previous method, right away we assume that $\boldsymbol{A}$ is a Hermitian (which here simply requires it is square and self-adjoint), positive-definite matrix.

The Cholesky factor $\boldsymbol{L}$ is said to be the lower triangular matrix which satisfies

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T \tag{9.328}$$

This is a result which can be easily obtained in MATLAB via the command `chol`.

Be careful, however, as MATLAB will naturally return an *upper* triangular matrix, and we are interested in the *lower* triangular version.

This is remedied by simply transposing what is provided by the command or by issuing `chol(A,'lower')`.

As with the spectral factorization method, we see that there are an infinite number of square root factors by simply post-multiplying the Cholesky factor by any square orthonormal matrix of appropriate dimension.

Note that for each method, depending on the properties of $\boldsymbol{A}$, the specific resulting square root factor may be unique; square root factors, however, are not, in general, unique!

Let's look at some numerical examples of these two methods.

Say we are interested in finding a square root factor of some matrix $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix} \tag{9.329}$$

If we were to find a square root factor via eigen-decomposition, we would get

$$\sqrt{\boldsymbol{A}}_{\text{eig}} = \boldsymbol{V}\sqrt{\boldsymbol{\Lambda}} = \begin{bmatrix} -0.2931 & -0.4491 & -0.8440 \\ -0.1560 & 1.2931 & 0.5509 \\ -0.1018 & -0.6881 & 1.5863 \end{bmatrix} \tag{9.330}$$

If we were to find a square root factor via Cholesky decomposition, we would get

$$\sqrt{\boldsymbol{A}}_{\text{chol}} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \tag{9.331}$$

Both of these can be easily checked by simply checking to see if $\sqrt{\boldsymbol{A}}\sqrt{\boldsymbol{A}}^T = \boldsymbol{A}$!

Now that we have a good feeling for matrix square root factors, let's go back to the unscented transform and discuss how we generate points and transform them in order to approximate expectation integrals.

Just as a reminder, given $\boldsymbol{x}$ to have mean, $\boldsymbol{m}$, and covariance, $\boldsymbol{P}$, and a transformation of $\boldsymbol{x}$ through the nonlinear function $\boldsymbol{g}$ into $\boldsymbol{y}$, we want to find the mean and covariance of $\boldsymbol{y}$ and the cross-covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

That is, given

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}) \tag{9.332}$$
$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) \tag{9.333}$$

we want to approximate

$$\hat{\boldsymbol{y}} = \int \boldsymbol{g}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.334}$$
$$\boldsymbol{Y} = \int (\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})(\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})^T p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.335}$$
$$\boldsymbol{C} = \int (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})^T p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.336}$$

As a reminder,

- $\hat{\boldsymbol{y}}$ is the mean of $\boldsymbol{y}$
- $\boldsymbol{Y}$ is the covariance of $\boldsymbol{y}$
- $\boldsymbol{C}$ is the cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$

Now, we are ready to present the unscented transform.

The first step is to draw sigma points for the input, $\boldsymbol{x}$.

For a random variable $\boldsymbol{x} \in \mathbb{R}^n$ given by $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P})$, form a set of $2n+1$ sigma points as

$$\mathscr{X}^{(0)} = \boldsymbol{m} \tag{9.337}$$

$$\mathscr{X}^{(i)} = \boldsymbol{m} + \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}}\right]_i \tag{9.338}$$

$$\mathscr{X}^{(i+n)} = \boldsymbol{m} - \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}}\right]_i \tag{9.339}$$

for $i = 1, \ldots, n$ where

- $[\cdot]_i$ denotes the i^{th} column of the matrix
- λ is a scaling parameter defined as

$$\lambda = \alpha^2(n+\kappa) - n \tag{9.340}$$

- α and κ are parameters that determine the spread of the sigma points around the mean
- the matrix square root denotes a matrix such that $\sqrt{\boldsymbol{P}}\sqrt{\boldsymbol{P}}^T = \boldsymbol{P}$

We also associate mean and covariance weights with each of the sigma points for $i = 1, \ldots, 2n$ as

$$w_0^{(m)} = \frac{\lambda}{n+\lambda} \tag{9.341}$$

$$w_0^{(c)} = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta) \tag{9.342}$$

$$w_i^{(m)} = \frac{1}{2(n+\lambda)} \tag{9.343}$$

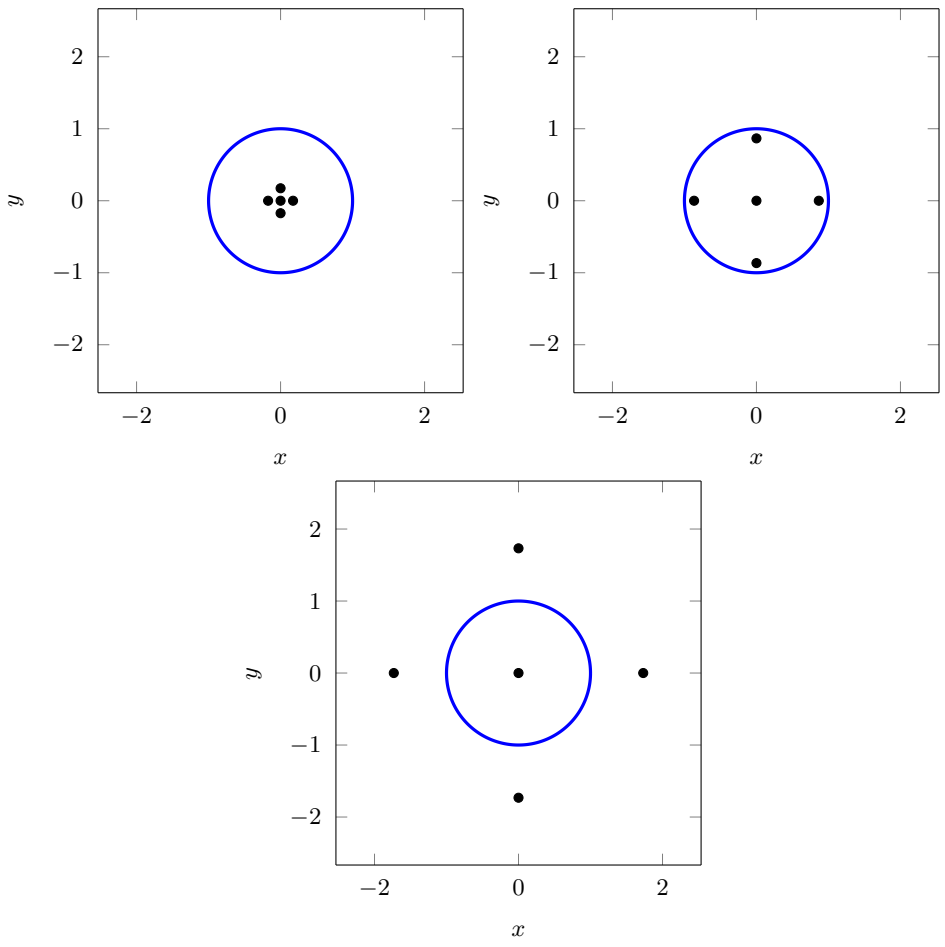
$$w_i^{(c)} = \frac{1}{2(n+\lambda)} \tag{9.344}$$

The value β is an additional nonnegative algorithm parameter that can used for incorporating prior information on the non-Gaussian distribution of $\boldsymbol{x}$ (see Wan and van der Merwe[1] for details regarding how to obtain these weights and techniques for selecting β if that is of interest).

In the case that no prior information is desired to be added, β should be set to zero.

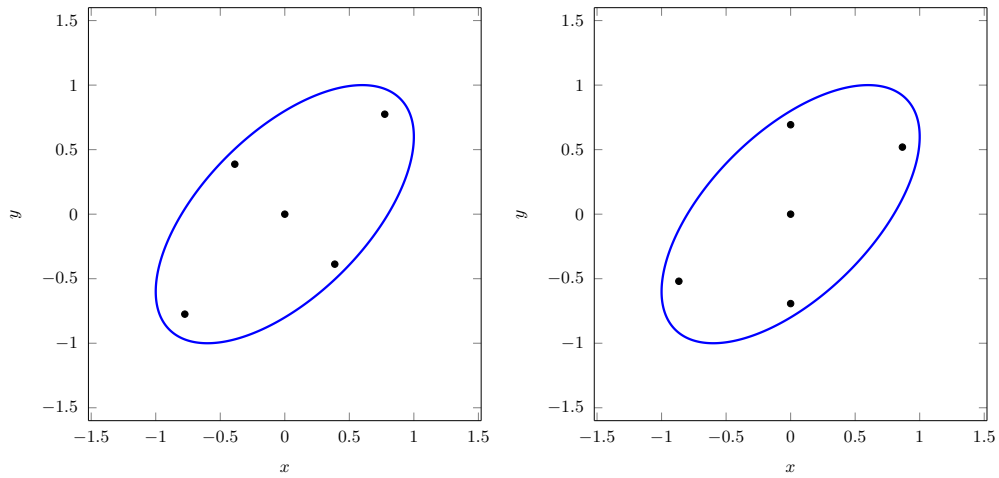
In the case that the prior is Gaussian, the optimal choice turns out to be $\beta = 2$.

Let's look at a few cases of the sigma point generation for different values of α to show how this parameters controls the spread of the sigma points. We will use a zero-mean identity-covariance input and α of 0.1, 0.5, and 1.0.



Additionally, we can look at how the choice of the square root factor influences the determination of the sigma points for the input. Here, we use both a spectral factorization and a Cholesky decomposition, both using $\alpha = 0.5$.

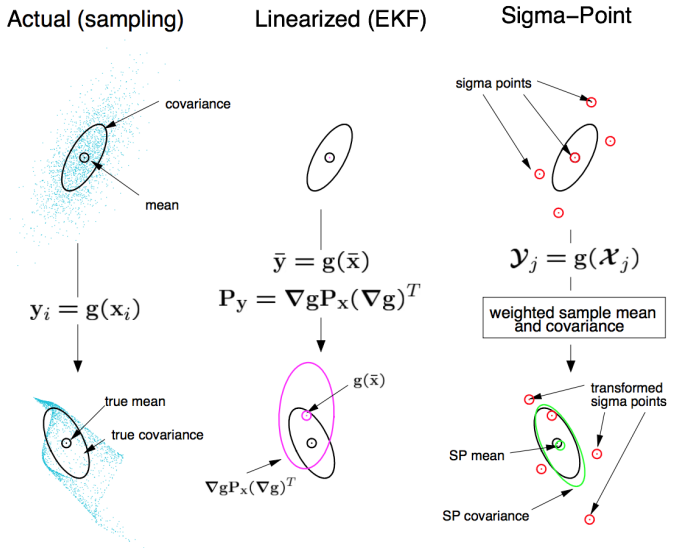
[1] Wan, E.A. and van der Merwe, R., *The unscented Kalman filter*, Ch. 7 of Haykin, S. (ed.), **Kalman Filtering and Neural Networks**, Wiley, 2001.



The second step of the UT is to transform the points through the nonlinear function. That is, for each input sigma point $\mathscr{X}^{(i)}$, we apply the nonlinear function $\boldsymbol{g}(\cdot)$ as

$$\mathscr{Y}^{(i)} = \boldsymbol{g}(\mathscr{X}^{(i)}) , \qquad i = 0, \ldots, 2n \tag{9.345}$$

This is visually depicted in the following figure (taken from van der Merwe[2])



The third and final step in the UT is to approximate the mean, covariance, and cross-covariance as

$$\hat{\boldsymbol{y}} \approx \sum_{i=0}^{2n} w_i^{(m)} \mathscr{Y}^{(i)} \tag{9.346}$$

$$\boldsymbol{Y} \approx \sum_{i=0}^{2n} w_i^{(c)} (\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T \tag{9.347}$$

$$\boldsymbol{C} \approx \sum_{i=0}^{2n} w_i^{(c)} (\mathscr{X}^{(i)} - \boldsymbol{m})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T \tag{9.348}$$

[2] van der Merwe, R., *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*, Ph.D. thesis, Oregon Health and Science University, 2004.

Consider the nonlinear transformation from polar coordinates to Cartesian coordinates.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} \tag{9.349}$$

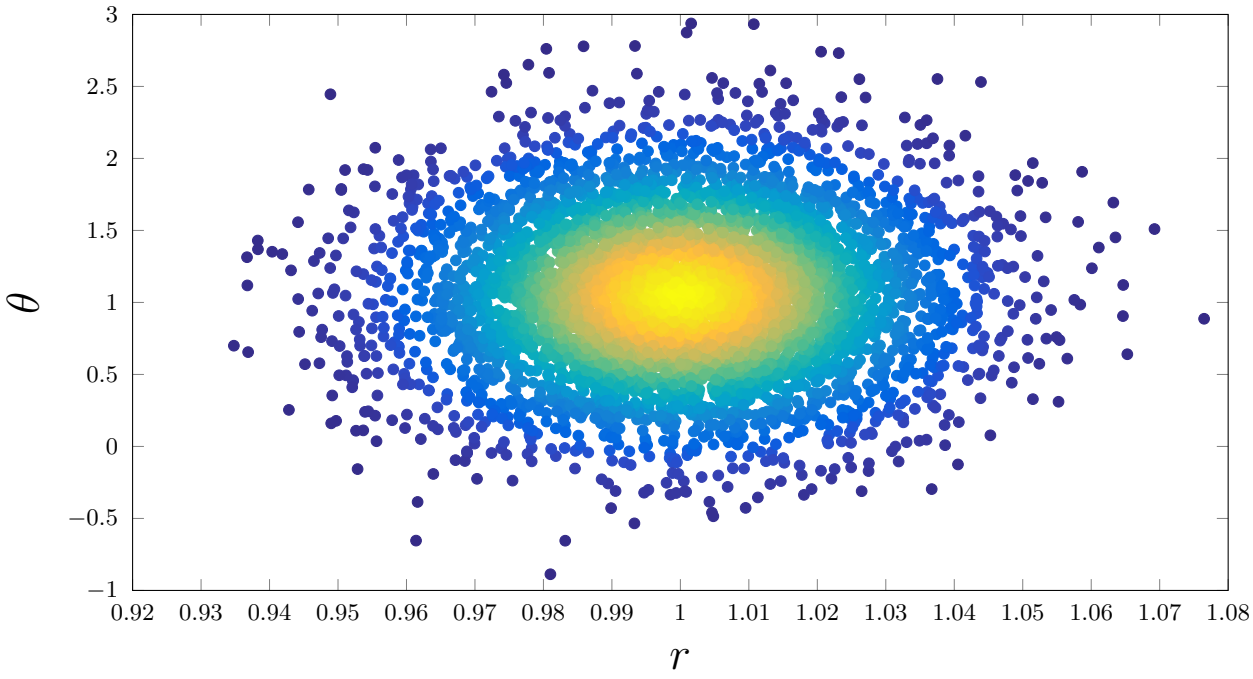
We can illustrate the performance of the UT by defining some mean and covariance and mapping these statistics with the UT.

What we will do is perform a Monte Carlo analysis to offer an understanding of the resulting "true" statistics, and see how both linearization and the UT perform in comparison.

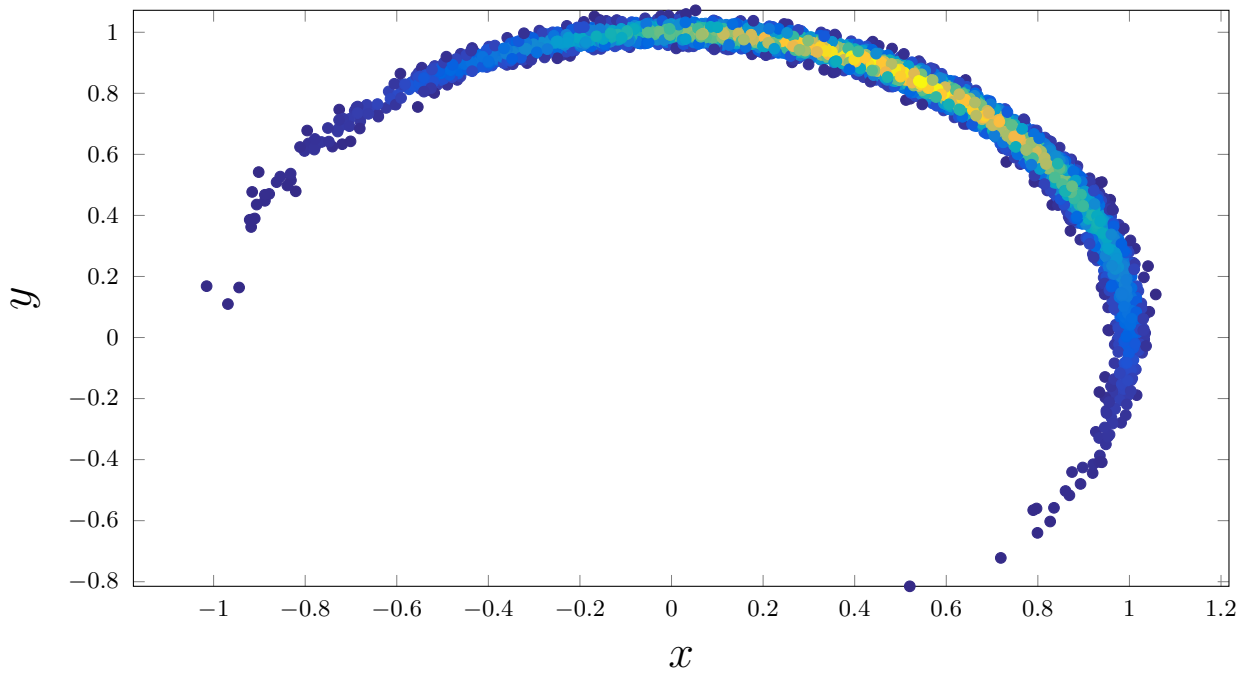
First, let's define a mean and a covariance.

$$\boldsymbol{m} = \begin{bmatrix} r \\ \theta \end{bmatrix} = \begin{bmatrix} 1 \\ 60^\circ \end{bmatrix} \quad \text{and} \quad \boldsymbol{P} = \begin{bmatrix} (0.02)^2 & 0 \\ 0 & (30^\circ)^2 \end{bmatrix} \tag{9.350}$$

Now let's draw 5000 samples from a Gaussian with this mean and covariance.



Now, let's map these according to the above transformation to see what the resulting pdf looks like.

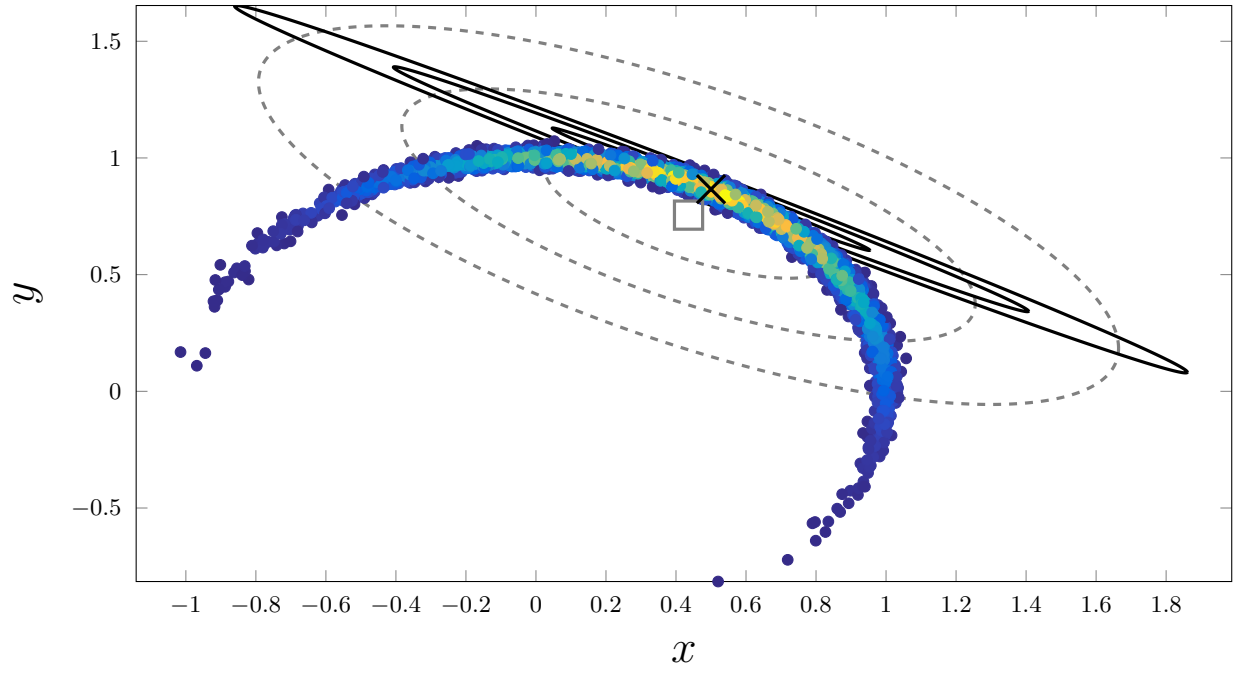


This "banana" shape is common in many problems, especially problems which have periodic or semi-periodic motion (like orbits).

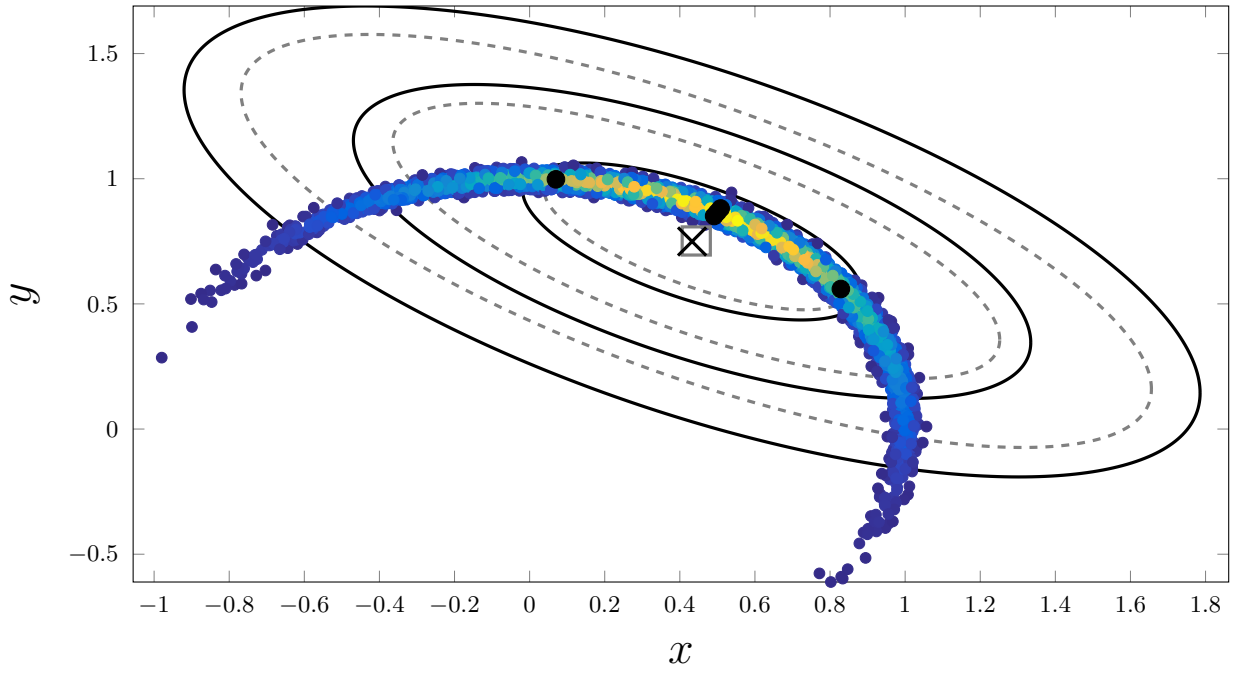
First, let's see how linearization performs when compared to the mean and covariance of the resulting Monte Carlo samples.

For illustration, we have 1, 2, and 3σ curves drawn. This is a Gaussian illustration, though the resulting pdf is clearly non-Gaussian. This is simply to illustrate differences in mean and covariance approximations and, since the Gaussian distribution is fully characterized by a mean and covariance, it makes a convenient visualization tool.

In the below figure, the dashed curves belong to the Monte Carlo sample mean and covariance, and the solid lines belong to the mean and covariance resulting from a linear mapping. Furthermore, their corresponding means are marked by a square and an $\times$ respectively.



Now lets see the same thing, but this time with the UT. This plot is built in the same way as the last one, except now we also have the $2n+1$ sigma points drawn as black dots.



In this case, the UT outperforms linearization in both the transformed mean *and* covariance in the sense that it is much more close to the Monte Carlo results.

Note, however, that this is not to say it will *always* outperform linearization. This problem is simply an illustration.

Alright, so we have considered the statistics of the transformation problem

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}) \tag{9.351}$$
$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) \tag{9.352}$$

where the statistics of the input are known.

But this doesn't quite match up to our filtering problems. There's no noise added in.

It is often the case that we model the inclusion of noise with an additive model such that the noise is added to the transformation $\boldsymbol{g}(\boldsymbol{x})$.

Our transformation is then given by

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{q} \tag{9.353}$$

where the noise is represented by $\boldsymbol{q}$ and is often taken to be zero mean with covariance $\boldsymbol{Q}$.

This model holds for both our dynamical system with $\boldsymbol{q}$ representing the process noise and for our observational system with $\boldsymbol{q}$ representing the measurement noise.

Remember that we do not require this noise to be Gaussian distributed in general, but it is an assumption we will make as it affords a convenient and straightforward explanation.

So, how can we apply the UT method to compute the statistics with an additive noise?

Well, we've actually already solved this problem!

When the noise is additive and uncorrelated to the state, we have that

$$\hat{\boldsymbol{y}} = \int \boldsymbol{g}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.354}$$
$$\boldsymbol{Y} = \int (\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})(\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})^T p(\boldsymbol{x}) d\boldsymbol{x} + \boldsymbol{Q} \tag{9.355}$$
$$\boldsymbol{C} = \int (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{g}(\boldsymbol{x}) - \hat{\boldsymbol{y}})^T p(\boldsymbol{x}) d\boldsymbol{x} \tag{9.356}$$

We've already developed the method for estimating each of the preceding integrals, so we can directly state the result of the UT in the presence of additive noise.

The first step is to draw our state sigma points, $\mathscr{X}^{(i)}$, and to determine the mean and covariance weights $w_i^{(m)}$ and $w_i^{(c)}$, for $i = 0, \ldots, 2n$.

We then determine the transformed sigma points as

$$\mathscr{Y}^{(i)} = \boldsymbol{g}(\mathscr{X}^{(i)}) \,, \qquad i = 0, \ldots, 2n \tag{9.357}$$

Finally, we compute the mean, covariance, and cross-covariance as

$$\hat{\boldsymbol{y}} \approxeq \sum_{i=0}^{2n} w_i^{(m)} \mathscr{Y}^{(i)} \tag{9.358}$$

$$\boldsymbol{Y} \approxeq \sum_{i=0}^{2n} w_i^{(c)} (\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T + \boldsymbol{Q} \tag{9.359}$$

$$\boldsymbol{C} \approxeq \sum_{i=0}^{2n} w_i^{(c)} (\mathscr{X}^{(i)} - \boldsymbol{m})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T \tag{9.360}$$

In a Kalman filtering sense, this enables us to perform both propagation (or prediction) and measurement updates (or correction) in the presence of additive noise!

Recall that we need the mean and covariance equations for propagation and that we need the mean, covariance, and cross-covariance equations for the update.

This is the model we've been working under all along, so we can use this to formulate an unscented Kalman filter (UKF) for additive noise models.

One of the awesome things about the UT is it allows us to fairly easily extend our treatment of the noise.

For instance, what if the noise does not obey a linear, additive model?

What if the noise is included in the nonlinear transformation?

In this case, our transformation takes the form

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{q}) \tag{9.361}$$

where we assume that we still know the statistics of the inputs, $\boldsymbol{x}$ and $\boldsymbol{q}$.

That is, we now ask the question: if we know

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}) \qquad \text{and} \qquad \boldsymbol{q} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}) \tag{9.362}$$

what are the mean, covariance, and cross-covariance (with $\boldsymbol{x}$) of $\boldsymbol{y}$?

Well, let's augment our state with our noise to form an augmented state

$$\boldsymbol{x}_{\text{aug}} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{q} \end{bmatrix} \tag{9.363}$$

We are still assuming that the mean and covariance of $\boldsymbol{x}$ are $\boldsymbol{m}$ and $\boldsymbol{P}$.

We are also still assuming that the mean and covariance of $\boldsymbol{q}$ are $\boldsymbol{0}$ and $\boldsymbol{Q}$.

Finally, we assume that the noise is independent of the state; then, the mean and covariance of the augmented state are

$$\boldsymbol{m}_{\text{aug}} = \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{0} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{P}_{\text{aug}} = \begin{bmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q} \end{bmatrix} \tag{9.364}$$

It is worth noting that non-zero mean or non-independent noises could easily be handled in this augmented formulation.

Now, our transformation takes the form

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}_{\text{aug}}) \tag{9.365}$$

How do we get the mean, covariance, and cross-covariance of $\boldsymbol{y}$?

Well, we already dealt with exactly this problem. The only difference will be in the number of sigma points and in determining the cross-covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Let the dimension of $\boldsymbol{x}$ be n and the dimension of $\boldsymbol{q}$ be n_q. Then, the dimension of $\boldsymbol{x}_{\text{aug}}$ is $n_{\text{aug}} = n + n_q$.

Now, form sigma points for the augmented state, $\boldsymbol{x}_{\text{aug}}$, as

$$\mathscr{X}_{\text{aug}}^{(0)} = \boldsymbol{m}_{\text{aug}} \tag{9.366}$$

$$\mathscr{X}_{\text{aug}}^{(i)} = \boldsymbol{m}_{\text{aug}} + \sqrt{n_{\text{aug}} + \lambda_{\text{aug}}} \left[\sqrt{\boldsymbol{P}_{\text{aug}}}\right]_i \tag{9.367}$$

$$\mathscr{X}_{\text{aug}}^{(i+n_{\text{aug}})} = \boldsymbol{m}_{\text{aug}} - \sqrt{n_{\text{aug}} + \lambda_{\text{aug}}} \left[\sqrt{\boldsymbol{P}_{\text{aug}}}\right]_i \tag{9.368}$$

for $i = 1, \ldots, n_{\text{aug}}$.

The parameter λ_{aug} is defined as it was earlier but with n replaced by n_{aug}.

Associated with each of the sigma points is also the set of mean and covariance weights; these are computed just as before but with n and λ replaced with n_{aug} and λ_{aug}.

These are now represented symbolically by $w_{i,\text{aug}}^{(m)}$ and $w_{i,\text{aug}}^{(c)}$, for $i = 0, \ldots, 2n_{\text{aug}}$.

Now, we simply transform the sigma points through the nonlinear function

$$\mathscr{Y}^{(i)} = \boldsymbol{g}(\mathscr{X}_x^{(i)}, \mathscr{X}_q^{(i)}) \,, \quad \text{for} \quad i = 0, 1, \ldots, 2n_{\text{aug}} \tag{9.369}$$

where $\mathscr{X}_x^{(i)}$ and $\mathscr{X}_q^{(i)}$ denote the parts of the sigma point i which correspond to $\boldsymbol{x}$ and $\boldsymbol{q}$ respectively; that is

$$\mathscr{X}_{\text{aug}}^{(i)} = \begin{bmatrix} \mathscr{X}_x^{(i)} \\ \\ \mathscr{X}_q^{(i)} \end{bmatrix} \tag{9.370}$$

The final step is then to compute the approximate mean, covariance, and cross-covariance as

$$\hat{\boldsymbol{y}} \approxeq \sum_{i=0}^{2n_{\text{aug}}} w_{i,\text{aug}}^{(m)} \mathscr{Y}^{(i)} \tag{9.371}$$

$$\boldsymbol{Y} \approxeq \sum_{i=0}^{2n_{\text{aug}}} w_{i,\text{aug}}^{(c)} (\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T \tag{9.372}$$

$$\boldsymbol{C} \approxeq \sum_{i=0}^{2n_{\text{aug}}} w_{i,\text{aug}}^{(c)} (\mathscr{X}_x^{(i)} - \boldsymbol{m})(\mathscr{Y}^{(i)} - \hat{\boldsymbol{y}})^T \tag{9.373}$$

What about the accuracy of the method?

The unscented transform is a third-order method in the sense that the estimate of the mean of $\boldsymbol{g}(\boldsymbol{x})$ is exact for polynomials up to order three.

However, the covariance approximation is exact only for first order polynomials, because the square of a second order polynomial is already a polynomial of order four, and the UT does not compute the exact result for fourth order polynomials.

In this sense, the UT is only a first order method.

With suitable selection of parameters ($\kappa = 3 - n$) it is possible to get some of the fourth order terms appearing in the covariance computation correct also for quadratic functions, but not all of them.

For more details, see Särkkä,[3] or for an in-depth derivation and discussion, see van der Merwe.[4]

At this point, we have all of the tools required to formulate the unscented Kalman filter.

The unscented Kalman filter works within the Kalman framework to propagate and update the mean and covariance for the state.

In doing so, the UKF leverages the unscented transform to approximate the mean, covariance, and cross-covariance (with the input) of the output of a nonlinear transformation.

In particular, the mean and covariance calculations are used in the propagation stage. The mean, covariance, and cross-covariance are used in the update stage.

[3] Särkkä, S., **Bayesian Filtering and Smoothing**, Cambridge, 2013.

[4] van der Merwe, R., *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*, Ph.D. thesis, Oregon Health and Science University, 2004.

We will consider the system to be described via

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}) + \boldsymbol{w}_{k-1} \tag{9.374}$$
$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \tag{9.375}$$

Given initial values of the mean and covariance, $\boldsymbol{m}_0$ and $\boldsymbol{P}_0$, along with a sequence of data, $\boldsymbol{z}_k$, for $k = 1, 2, \ldots$, the objective is to sequentially propagate the mean and covariance between measurements and update the mean and covariance using the measurement data.

We are also assuming that the process noise and measurement noise are both zero-mean, white-noise sequences that are both uncorrelated with the state.

- **Initialize**
  initialize the mean and covariance

$$\boldsymbol{m}_{k-1}^{+} = \boldsymbol{m}_0 \tag{9.376}$$
$$\boldsymbol{P}_{k-1}^{+} = \boldsymbol{P}_0 \tag{9.377}$$

- **Propagate**
  propagate from the time $k-1$ to time k, the time of a new measurement

  1. Form the sigma points from the posterior mean and covariance at $k-1$

$$\mathscr{X}_{k-1}^{(0)} = \boldsymbol{m}_{k-1}^{+} \tag{9.378}$$
$$\mathscr{X}_{k-1}^{(i)} = \boldsymbol{m}_{k-1}^{+} + \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}_{k-1}^{+}}\right]_i \qquad i = 1, \ldots, n \tag{9.379}$$
$$\mathscr{X}_{k-1}^{(i+n)} = \boldsymbol{m}_{k-1}^{+} - \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}_{k-1}^{+}}\right]_i \qquad i = 1, \ldots, n \tag{9.380}$$

  2. Determine the associated mean and covariance weights as

$$w_0^{(m)} = \frac{\lambda}{n+\lambda} \tag{9.381}$$
$$w_0^{(c)} = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta) \tag{9.382}$$
$$w_i^{(m)} = \frac{1}{2(n+\lambda)} \qquad i = 1, \ldots, 2n \tag{9.383}$$
$$w_i^{(c)} = \frac{1}{2(n+\lambda)} \qquad i = 1, \ldots, 2n \tag{9.384}$$

3. Transform the sigma points through the dynamic model

$$\mathscr{X}_k^{(i)} = \boldsymbol{f}(\mathscr{X}_{k-1}^{(i)}) \,, \qquad i = 0, \ldots, 2n \tag{9.385}$$

4. Compute the predicted (prior) mean $\boldsymbol{m}_k^-$ and covariance $\boldsymbol{P}_k^-$

$$\boldsymbol{m}_k^- = \sum_{i=0}^{2n} w_i^{(m)} \mathscr{X}_k^{(i)} \tag{9.386}$$

$$\boldsymbol{P}_k^- = \sum_{i=0}^{2n} w_i^{(c)} (\mathscr{X}_k^{(i)} - \boldsymbol{m}_k^-)(\mathscr{X}_k^{(i)} - \boldsymbol{m}_k^-)^T + \boldsymbol{Q}_{k-1} \tag{9.387}$$

- **Update**
  utilize incoming measurement at time k to improve our mean and covariance

  1. Form the sigma points

$$\mathscr{X}_k^{-(0)} = \boldsymbol{m}_k^- \tag{9.388}$$

$$\mathscr{X}_k^{-(i)} = \boldsymbol{m}_k^- + \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}_k^-}\right]_i \qquad i = 1,\ldots,n \tag{9.389}$$

$$\mathscr{X}_k^{-(i+n)} = \boldsymbol{m}_k^- - \sqrt{n+\lambda}\left[\sqrt{\boldsymbol{P}_k^-}\right]_i \qquad i = 1,\ldots,n \tag{9.390}$$

  2. Determine the associated mean and covariance weights as

$$w_0^{(m)} = \frac{\lambda}{n+\lambda} \tag{9.391}$$

$$w_0^{(c)} = \frac{\lambda}{n+\lambda} + (1-\alpha^2+\beta) \tag{9.392}$$

$$w_i^{(m)} = \frac{1}{2(n+\lambda)} \qquad i = 1,\ldots,2n \tag{9.393}$$

$$w_i^{(c)} = \frac{1}{2(n+\lambda)} \qquad i = 1,\ldots,2n \tag{9.394}$$

  3. Transform the sigma points through the measurement model

$$\mathscr{Z}_k^{(i)} = \boldsymbol{h}(\mathscr{X}_k^{-(i)}), \qquad i = 0,\ldots,2n \tag{9.395}$$

  4. Compute the predicted measurement $\hat{\boldsymbol{z}}_k$, the predicted covariance of the measurement $\boldsymbol{W}_k$, and the cross-covariance of the state and the measurement $\boldsymbol{C}_k$

$$\hat{\boldsymbol{z}}_k = \sum_{i=0}^{2n} w_i^{(m)}\,\mathscr{Z}^{(i)} \tag{9.396}$$

$$\boldsymbol{W}_k = \sum_{i=0}^{2n} w_i^{(c)}(\mathscr{Z}^{(i)} - \hat{\boldsymbol{z}}_k)(\mathscr{Z}^{(i)} - \hat{\boldsymbol{z}}_k)^T + \boldsymbol{R}_k \tag{9.397}$$

$$\boldsymbol{C}_k = \sum_{i=0}^{2n} w_i^{(c)}(\mathscr{X}_k^{-(i)} - \boldsymbol{m}_k^-)(\mathscr{Z}^{(i)} - \hat{\boldsymbol{z}}_k)^T \tag{9.398}$$

  5. Compute the Kalman gain $\boldsymbol{K}_k$, the filtered (posterior) state mean $\boldsymbol{m}_k^+$ and covariance $\boldsymbol{P}_k^+$, conditional on the measurement $\boldsymbol{z}_k$

$$\boldsymbol{K}_k = \boldsymbol{C}_k\boldsymbol{W}_k^{-1} \tag{9.399}$$

$$\boldsymbol{m}_k^+ = \boldsymbol{m}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k) \tag{9.400}$$

$$\boldsymbol{P}_k^+ = \boldsymbol{P}_k^- - \boldsymbol{C}_k\boldsymbol{K}_k^T - \boldsymbol{K}_k\boldsymbol{C}_k^T + \boldsymbol{K}_k\boldsymbol{W}_k\boldsymbol{K}_k^T \tag{9.401}$$

The advantage of the UKF over methods such as the EKF is that it is not based on a linear approximation at a single point, but instead uses further points in approximating the nonlinearity.

Additionally, the UKF does not require that derivatives of the system dynamics and the measurement model be taken,

whereas the EKF does require these derivatives.

A disadvantage of the UKF when compared to the EKF is that it requires slightly more computational operations than the EKF

We have focused on the form of the UKF when we have both additive process noise and measurement noise.

As shown in the discussion of the UT, it is straightforward to treat non-additive noises and to treat state-correlated noises.

This case will not be discussed, but it essentially just comes down to augmenting the state vector and proceeding in the usual fashion.

Note, however, that augmenting the state vector increases the number of sigma points used in the UKF.

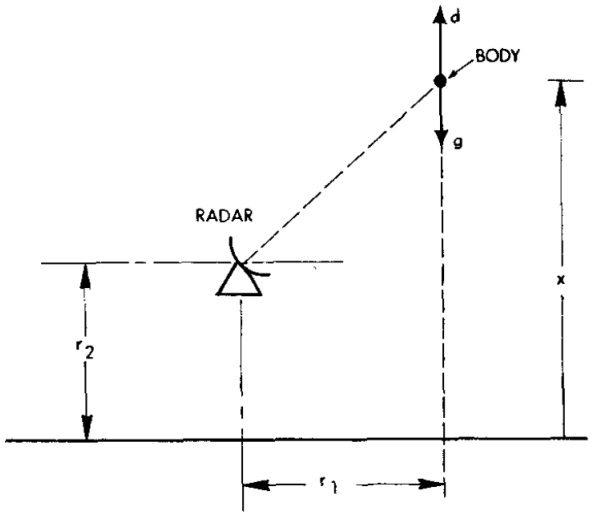
When additive noises are present, it is more computationally efficient to use the additive noise form of the UKF instead of augmenting the state.

9.3.3 Example of the UKF

The following example is taken and modified from Gelb (1974).

Consider the problem of tracking a body falling freely through the atmosphere.

The object falls in a straight line, but a radar observing it is offset from the object horizontally by r_1 and is r_2 off the ground (see the figure below).



A radar return is received every 0.1 [sec].

The state is defined as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \\ 1/\beta \end{bmatrix} \tag{9.402}$$

where x is the height of the falling body above the earth and β is the ballistic coefficient of the object.

Note that our last state vector element is the *inverse* ballistic coefficient, not simply β.

The equations of motion for the body are given by

$$\dot{\boldsymbol{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ d-g \\ 0 \end{bmatrix} = \boldsymbol{f}(\boldsymbol{x}) \tag{9.403}$$

with

$$d = \frac{\rho x_2^2 x_3}{2} \tag{9.404}$$

$$\rho = \rho_0 \exp\left\{-\frac{x_1}{k_\rho}\right\} \tag{9.405}$$

where d is drag acceleration, g is the acceleration of gravity, ρ is atmospheric density (with ρ_0 as the atmospheric density at sea level), and k_ρ is a decay constant.

The differential equation governing x_2 (velocity) is nonlinear through the dependence of drag on velocity, air density, and ballistic coefficient.

Due to the tracking radar's offset, we have nonlinear measurements of range, ρ, corrupted according to $\mathcal{N}(0,R)$.

$$\boldsymbol{h}(\boldsymbol{x}) = \rho = \sqrt{r_1^2 + (x - r_2)^2} \tag{9.406}$$

The initial truth for the simulation is drawn according to

$$x_0 = p_g(10^5\ [\text{ft}], 500\ [\text{ft}^2]) \tag{9.407}$$

$$\dot{x}_0 = p_g(-6000\ [\text{ft/sec}], 2\times 10^4\ [\text{ft}^2/\text{sec}^2]) \tag{9.408}$$

$$\beta = p_g(2000\ [\text{lb/ft}^2], 2.5\times 10^5\ [\text{lb}^2/\text{ft}^4]) \tag{9.409}$$

and the initial mean and covariance are taken to be

$$\boldsymbol{m}_0 = \begin{bmatrix} 10^5\ [\text{ft}] \\ -6000\ [\text{ft/sec}] \\ 1/2000\ [\text{lb/ft}^2]^{-1} \end{bmatrix} \tag{9.410}$$

$$\boldsymbol{P}_0 = \begin{bmatrix} 1000\ [\text{ft}^2] & 0 & 0 \\ 0 & 2\times 10^3\ [\text{ft}^2/\text{sec}^2] & 0 \\ 0 & 0 & 1/(2.5\times 10^5)\ [\text{lb}^2/\text{ft}^4]^{-1} \end{bmatrix}. \tag{9.411}$$

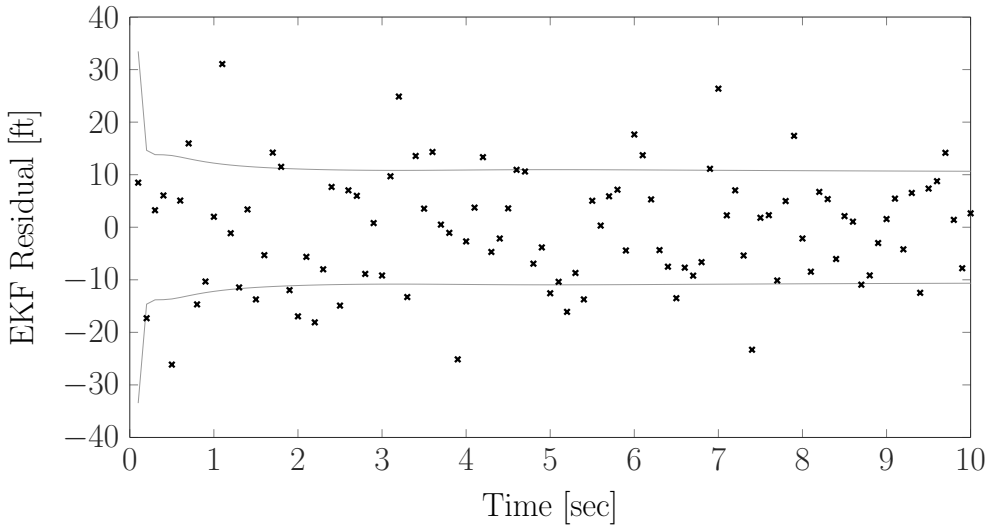
The system parameters are taken to be

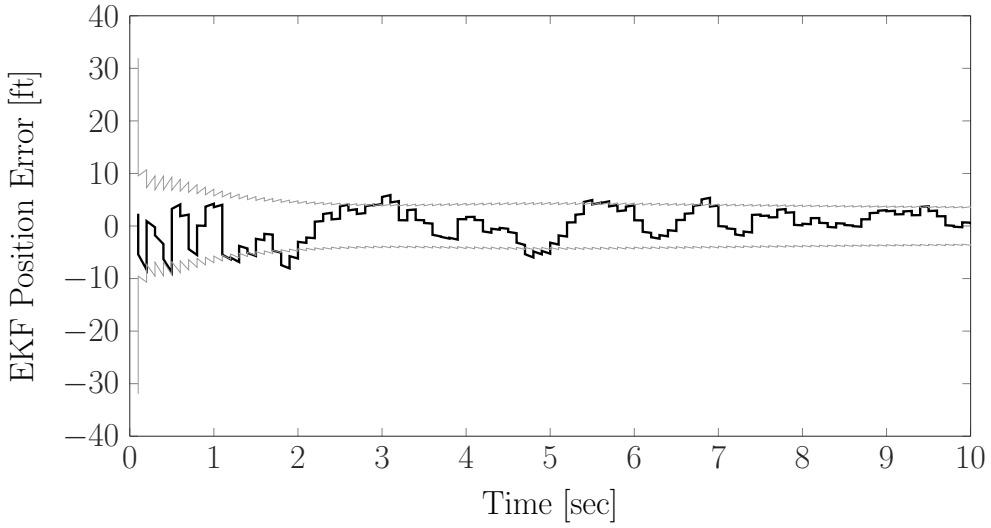
$$\rho_0 = 3.4\times 10^{-3}\ [\text{lb sec}^2/\text{ft}^4] \qquad g = 32.2\ [\text{ft/sec}^2] \tag{9.412}$$

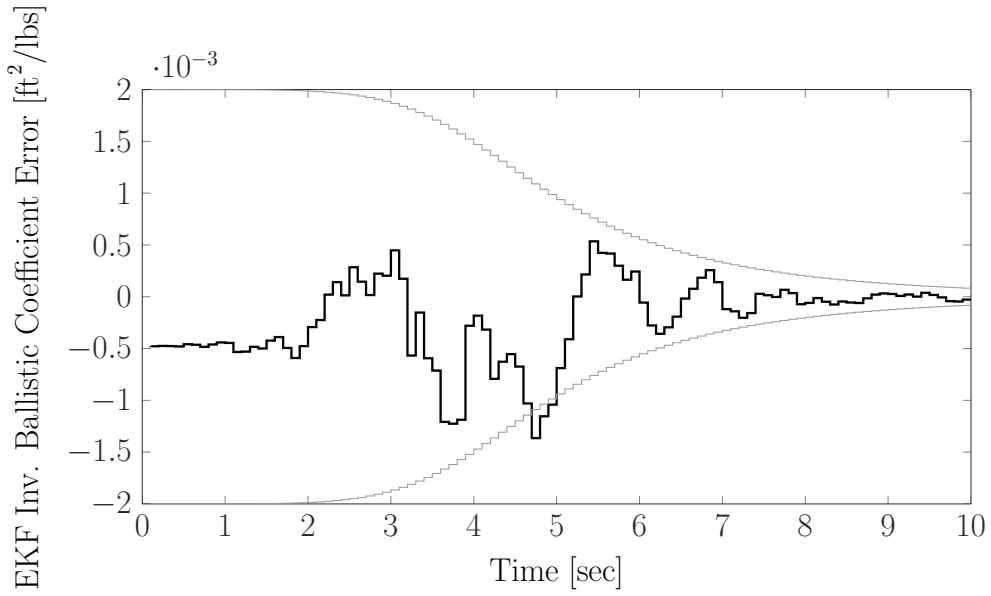
$$k_\rho = 22000\ [\text{ft}] \qquad R = 100\ [\text{ft}^2] \tag{9.413}$$

$$r_1 = 1000\ [\text{ft}] \qquad r_2 = 10\ [\text{ft}] \tag{9.414}$$

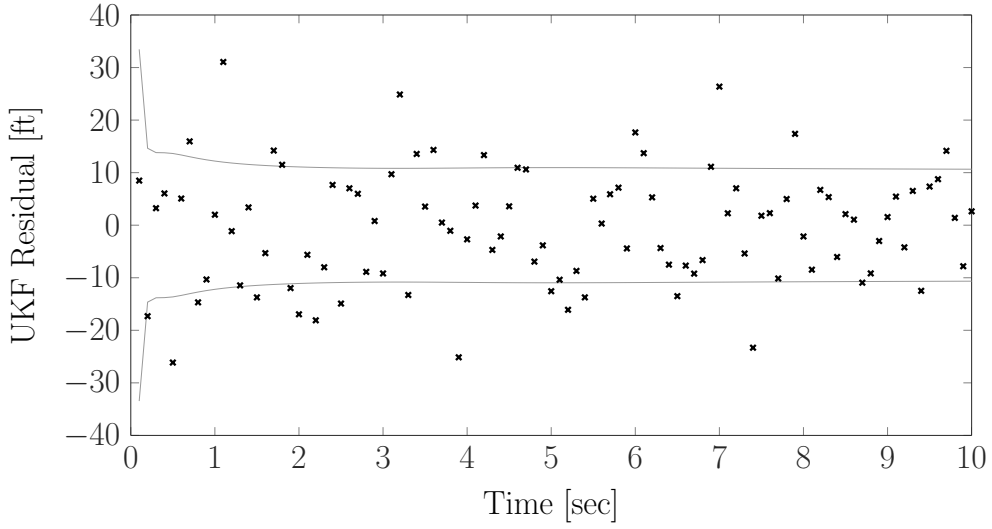
The position residuals, position (altitude) errors, velocity errors, and inverse ballistic coefficient errors produced by running an EKF (for comparison with the UKF) can be seen below.

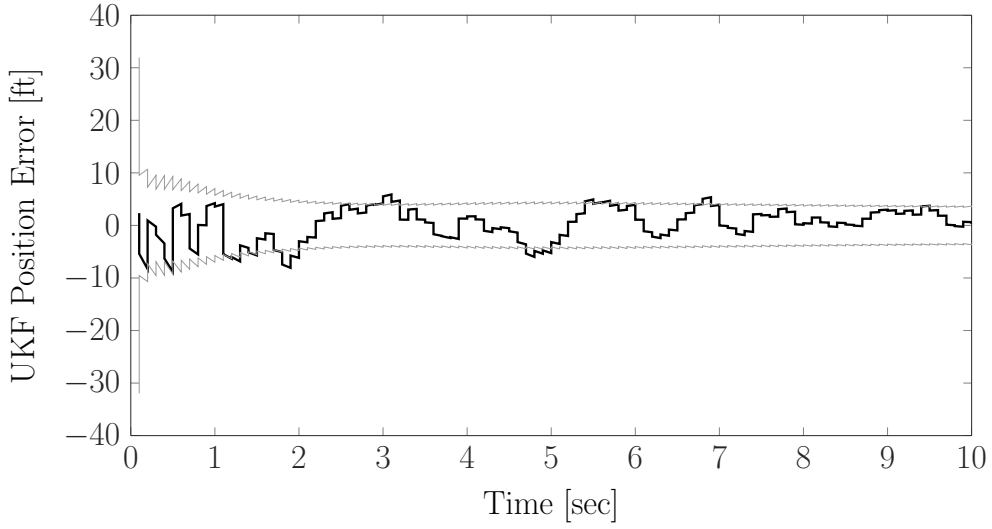


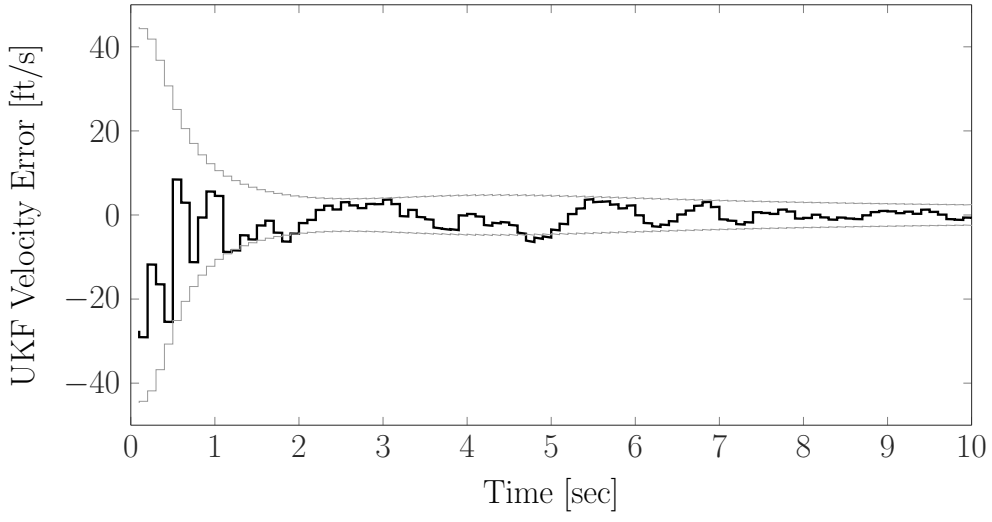


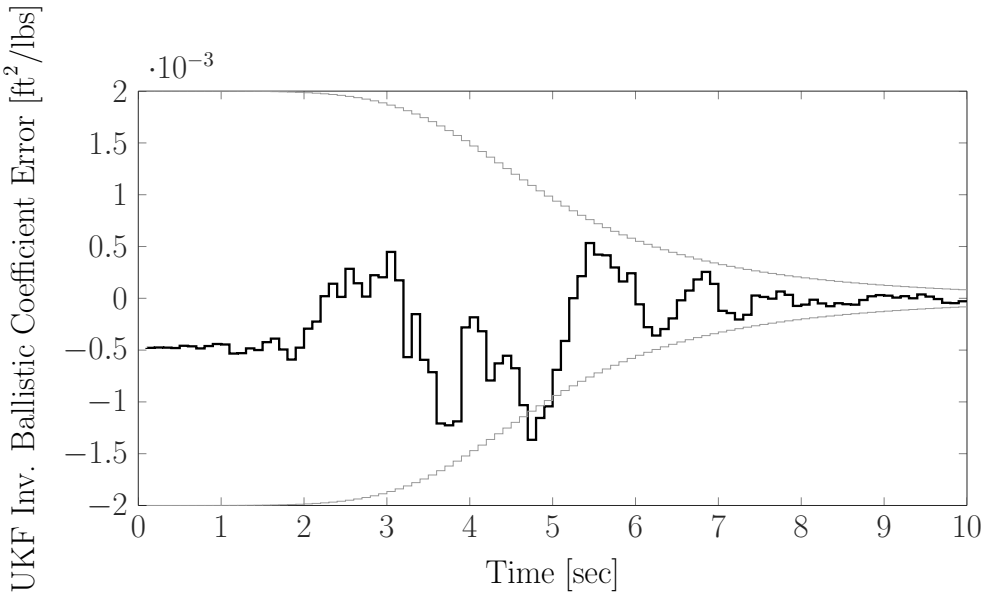


The same plots, though this time produced by the UKF with UT parameters of $\alpha = 1/2$, $\beta = 2$ (not ballistic coefficient), and $\kappa = 3 - n$ can be seen below.









Note how similar the plots are. They are almost identical (for this problem)!

Bibliography

[1] . http://www.stratcom.mil/factsheets/jspoc/.

[2] DEFENSE ACQUISITIONS. Assessments of selected weapon programs. United States Government Accountability Office, 2015.

[3] Salvatore Alfano. A numerical implementation of spherical object collision probability. *Journal of Astronautical Sciences*, 53(1):103–109, 2005.

[4] Salvatore Alfano. Satellite conjunction monte carlo analysis. *AAS Spaceflight Mechanics Mtg, Pittsburgh, PA., Paper*, pages 09–233, 2009.

[5] C.W. Allen. *Astrophysical Quantities*. The Athlone Press, University of London, London, 1973. ISBN:0 485 11150 0.

[6] J.L. Arsenault, L. Chaffee, and J.R. Kuhlmann. General Ephemeris Routine Formulation Document. In *Rept. ESD-TDR-64-522, Aeronutronic Publ. U-2731*, Aug., 1964.

[7] Harrison H Barrett, Christopher Dainty, and David Lara. Maximum-likelihood methods in wavefront sensing: stochastic models and likelihood functions. *JOSA A*, 24(2):391–414, 2007.

[8] Richard H. Battin. *An introduction to the mathematics and methods of astrodynamics*. AIAA education series. American Institute of Aeronautics and Astronautics, Reston, VA, 1999.

[9] G. Beutler. *Methods of Celestial Mechanics*. Two Volumes. Springer-Verlag, Heidelberg, 2005. ISBN: 3-540-40749-9 and 3-540-40750-1.

[10] P. Blanc, Bella Espinar, Norbert Geuder, Christian Gueymard, Richard Meyer, and et al. Direct normal irradiance related definitions and applications: The circumsolar issue. *Solar Energy, Elsevier*, 110:561 – 577, 2014.

[11] B. Bowman. A First Order Semi-Analytical Perturbation Theory for Highly Eccentric 12 Hour Resonating Satellite Orbits. In *1st Aerospace Control Squadron Repost, Colorado Springs, CO*, Nov., 1971.

[12] D. Brouwer. *Solution of the Problem of Artificial Satellite Theory Without Drag*. U.S. Air Force Cambridge Research Center, Geophysics Research Directorate, 1959. AFCRC-TN-59-638, Bedford, MA.

[13] D. Brouwer. Solution of the Problem of Artificial Satellite Theory Without Drag. *Astronomical Journal*, (64, 1274):378 – 397, 1959.

[14] Alex Burton, Mitch Zielinski, Carolin Frueh, Alinda Mashiku, and Nargess Memarsadeghi. Assessing measures to reliably predict collisions in the presence of uncertainty. 2018 AAS/AIAA Astrodynamics Specialist Confererence.

[15] J. Russell Carpenter, Salvatore Alfano, Doyle Hall, Matthew Hejduk, John Gaebler, Moriba Jah, Syed Hasan, Rebecca Besser, Russell DeHart, Matthew Duncan, Marissa Herron, and William Guit. Relevance of the american statistical association's warning on p-values for conjunction assessment. 2017 AAS/AIAA Astrodynamics Specialist Conference.

[16] CelesTrak. CelesTrak Homepage. http://celestrak.com, 2010.

[17] Ken Chan. Short-term vs. long-term spacecraft encounters. In *AIAA/AAS astrodynamics specialist conference and exhibit*, page 5460, 2004.

[18] Chris Peat, DLR Germany. Heavens Above Webpage. http://heavens-above.com, last accessed Sep. 2021.

[19] Winchell Chung. Atomic rockets: Rocket cat's glossary. `http://www.projectrho.com/public_html/rocket/glossary.php`.

[20] GAO (Government Accountability Office) Cristina T. Chaplain. Space situational awareness: Status of efforts and planned budgets. http://www.gao.gov/assets/680/672987.pdf, Oct.8, 2015, last accessed Sep 2021.

[21] G.M. Daniels. A Night Sky Model for Satellite Search Systems. *Optical Engineering No 1*, 16:66 – 71, 1977.

[22] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

[23] K. DeMars. Summer lecture series on orbit determination. In *NASA Houston*, 2015.

[24] National Geospatial-Intelligence Agency (NGA) Standardization Document. World geodetic system 1984, its definition and relationships with local geodetic systems. 2014.

[25] Rory Driscoll. Energy conservation in games, 2009.

[26] R Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[27] Pedro R. Escobal. *Methods of Orbit Determination*. Krieger Publishing Company, Malabar, FL, 1965.

[28] European Space Agency. European Space Agency Website. http://www.esa.int, 2005.

[29] D. Farnocchia, G. Tommei, A. Milani, and A. Rossi. Innovative methods of correlation and orbit determination for space debris. *Celestial Mechanics and Dynamical Astronomy*, 107(1-2):169–185, 2010.

[30] C. Früh and T. Schildknecht. Accuracy of Two-Line-Element Data for Geostationary and High-Eccentricity Orbits. *Journal of Guidance, Control, and Dynamics*, 35(5):1483 – 1491, 2012.

[31] Arthur Gelb, editor. *Applied Optimal Estimation*. The MIT Press, Cambridge, MA, 1974.

[32] E. Hecht. *Optics*. 4 edition. Addison-Wesley, August 12, 2001.

[33] M. Heikkinen. Geschlossene formeln zur berechnung räumlicher geodätischer koordinaten aus rechtwinkligen koordinaten. *Z. Vermess*, 107:207–211, 1982.

[34] Christiana Honsberg and Stuart Bowden. Standard solar spectra.

[35] F.R. Hoots, P.W. Schuhmacher, and R.A. Glover. History of Analytical Orbit Modelling in the U.S. Space Surveillance System. *Journal of Guidance, Control and Dynamics*, 27(2): 174 – 185, 2004.

[36] S.B. Howell. *Handbook of CCD Astronomy*. Cambridge University Press, 2001. ISBN 0-521-64834-3.

[37] James R. Janesick, Tom Elliott, Stewart Collins, Morley M. Blouke, and Jack Freeman. *Scientific charge-coupled devices*. 1987.

[38] Brandon A Jones and Alireza Doostan. Satellite collision probability estimation using polynomial chaos expansions. *Advances in Space Research*, 52(11):1860–1875, 2013.

[39] D. King-Hele. *Theory of Satellite Orbits in an Atmosphere*. Butterworths London, 1964. Chapter 4, pp. 40 – 77.

[40] Y. Kozai. The Motion of a Close Earth Satellite. *Astronomical Journal*, 1274(64):367 – 377, 1959.

[41] M.H. Lane and K.H. Cranford. An Improved Analytical Drag Theory for the Atrificial Satellite Problem. In *AIAA Paper 69 – 925*, August, 1969.

[42] Lockheed Martin. Space fence overview. https://www.lockheedmartin.com/en-us/products/space-fence.html, last accessed Sep 2021.

[43] John B. Lundberg and Bob E. Schutz. Recursion formulas of Legendre functions for use with nonsingular geopotential models. *Journal of Guidance, Control, and Dynamics*, 11(1):31–38, January-February 1988.

[44] Christopher Maes. The heliochronometer, 2008.

[45] George H. Massey, Philip; Jacoby. CCD Data: The Good, The Bad, and The Ugly. In Steve B. Howell, editor, *Astronomical CCD observing and reduction techniques*, volume 23, page 240, San Francisco, 1992. Astronomical Society of the Pacific.

[46] W J Merline and Steve B Howell. A realistic model for point-sources imaged on array detectors: The model and initial results. *Experimental Astronomy*, 6(1-2):163–210, 1995.

[47] P Misra and P Enge. *Global Positioning System: Signals, Measurements, and Performance, revised second ed*. Ganga-Jumana Press, USA, Lincoln, Massachusetts, 2012.

[48] Oliver Montenbrook and Thomas Pfleger. *Astronomy on the Personal Computer*. Springer, 2000.

[49] O. Montenbruck. *Grundlagen der Ephemeridenrechnung*. Spektrum Sterne und Weltraum, Heidelberg, 7th edition, 2005.

[50] O. Montenbruck and T. Fleger. *Astronomy on the Personal Computer*. Springer-Verlag, Berlin, 2nd edition, 2003.

[51] O. Montenbruck and E. Gill. *Satellite Orbits: Models, Methods, and Applications*. Springer-Verlag, Berlin, 3rd edition, 2005.

[52] Lambert E. Murray. An introduction to astronomy part ii: Historical development of astronomy. University Lecture.

[53] M.V. Newberry. Signal to noise considerations for sky-subtracted ccd data. *Publications of the Astronomical Society of the Pacific*, (130):122–130, 1991.

[54] Inter-Division A-G Working Group on Nominal Units for Stellar & Planetary Astronomy. Resolution b3 on recommended nominal conversion constants for selected solar and planetary properties, 2015.

[55] Russell P Patera. General method for calculating satellite collision probability. *Journal of Guidance, Control, and Dynamics*, 24(4):716–722, 2001.

[56] Russell P Patera. Satellite collision probability for nonlinear relative motion. *Journal of Guidance, Control, and Dynamics*, 26(5):728–733, 2003.

[57] Russell P Patera. Calculating collision probability for arbitrary space vehicle shapes via numerical quadrature. *Journal of guidance, control, and dynamics*, 28(6):1326–1328, 2005.

[58] Samuel Pines. Uniform representation of the gravitational potential and its derivatives. *AIAA Journal*, 11(11):1508–1511, November 1973.

[59] Pole Star Publications Ltd. Space flight now - launch schedule. https://spaceflightnow.com/launch-schedule/, last accessed Sep 2021.

[60] F. Sanson and C. Frueh. Noise Quantification in Optical Observations of Resident Space Objects for Probability of Detection and Likelihood. In *AAS/AIAA Astrodynamic Specialist Conference, Vail, CO*, 2015. 15-634.

[61] F. Sanson and C. Frueh. Improved ccd equation and probability of detection. *Journal of Astronautical Sciences*, 2016. submitted.

[62] T. Schildknecht. Vorlesung astronomie i. In *Astronomical Institute University of Bern, Switzerland*, 2012.

[63] T. Schildknecht, U. Hugentobler, A. Verdun, and G. Beutler. CCD Algorithms for space debris detection. Technical report, University of Berne, 1995.

[64] Byron D. Tapley, Bob E. Schutz, and George H. Born. *Statistical Orbit Determination*. Elselvier Academic Press, New York, NY, 2004.

[65] The Aerospace Corporation, Felix Hoots. SGP4-XP Informational Briefing, Aerospace Report NO. TOR-2021-00780, Feb 7, 2021.

[66] The Consultative Committee for Space Data Systems (CCSDS). Recommendation for Space Data System Standards ORBITAL DATA MESSAGE - The Blue Book. https://public.ccsds.org/Pubs/502x0b2c1e2.pdf, 2009, includes all updates up to 2012, last accessed Sep. 2021.

[67] The Consultative Committee for Space Data Systems (CCSDS). Recommendation for Space Data System Standards CONJUNCTION DATA MESSAGE - The Blue Book. https://public.ccsds.org/Pubs/508x0b1e2c1.pdf, 2013, includes all updates up to 2018, last accessed Sep. 2021.

[68] The Consultative Committee for Space Data Systems (CCSDS). Recommendation for Space Data System Standards Tracking Data Messages - The Pink Book. https://public.ccsds.org/Lists/CCSDSlast accessed Sep. 2021.

[69] Heinz Tiersch. S.B. Howell (ed.): Astronomical CCD observing and reduction techniques. Astronomical Society of the Pa-cific 1992, ASP Conference Series 23, 339 s., Preis: 55,-ISBN 0-937707-42-4. *Astronomische Nachrichten*, 314(6):398, 1993.

[70] G. Tommei, A. Milani, and A. Rossi. Orbit determination of space debris: admissible regions. *Celestial Mechanics and Dynamical Astronomy*, 97(4):289–304, 2007.

[71] TS Kelso, CelesTrak. CelesTrak Homepage. http://celestrak.com, last accessed Sep. 2021.

[72] US Space Force. Facts gssap. https://www.spaceforce.mil/About-Us/Fact-Sheets/Article/2197772/geosynchronous-space-situational-awareness-program/, last accessed Sep 2021.

[73] US Space Force. Facts sbss. https://www.spoc.spaceforce.mil/About-Us/Fact-Sheets/Display/Article/2381700/space-based-space-surveillance, last accessed Sep 2021.

[74] D. Valado, R. Crawford, R. Hujsak, and T.S. Kelso. Revisiting Space Track Report #3. *AIAA 2006-6753*, American Institute of Aeronautics und Astronautics, 2006.

[75] D. Vallado and W. McCain. *Fundamentals of Astrodynamics and Applications*. Microcosm Press, El Segundo, California, 2001. ISBN 0-7923-6903-3.

[76] J.R. Wertz. *Spacecraft Attitude Determination and Control*. Volume 73. D.Reidel Publishing Company, Dordrecht: Holland, 1978. ISBN: 90-277-0959-9.

[77] Jijie Zhu. Conversion of earth-centered earth-fixed coordinates to geodetic coordinates. *IEEE Transactions on Aerospace and Electronic Systems*, 30(3):957–961, 1994.