ENGO 585 Notes: Chapter 2 Estimation for Navigation

In this chapter we will review parametric least-squares, review some common math models for navigation, review accuracy measures, and finally introduce the Kalman filter (KF) as an alternative to parametric least squares (LS) for situations where the parameters being estimated are varying and observations are being made a varying rates.

a) Review of Parametric Least Squares

The standard parametric least squares equations are:

$$l = f(x) \qquad C_l$$

$$\hat{\delta} = -N^{-1}u = -(A^T C_l^{-1} A)^{-1} A^T C_l^{-1} w$$

$$w = f(x_0) - l$$

$$l = f(x) \qquad \hat{x} = x_0 + \hat{\delta}$$

This notation is what is generally seen at the University of Calgary, there are some variations though with minus signs and in terms of whether the misclosure vector is the observations minus the model or the other way around. In principal, parametric least squares can only be used for a batch of observations where there are more observations than there unknown parameters. In the case where there are two independent batches of observations (being used to estimate one set of parameters), it is possible to recast the solution equations as

$$\hat{\delta} = -[N_1 + N_2]^{-1}[u_1 + u_2]$$

This is called the summation of normals method, and it can be easily shown to be correct by taking the parametric least squares solution for all of the observations together and partitioning the design matrix and the covariance matrix of the observations into parts corresponding to the first and second set of observations. The simplification only works if the two sets of observations are independent (in other words the covariance matrix of the combined set of observations must be block diagonal). A second way of using multiple epochs of observations, called sequential least squares is presented later.

b) Math Models for Navigation

i. Location for AOA

In two dimensions, the AOA observation equation can have several forms depending on what assumptions are made about the orientations of the transmitters and receivers. An azimuth at point i to point j (that is what is the azimuth from point i to point j) can be expressed as

$$\alpha_{ij} = r_{ij} + \omega_i$$

where α_{ij} is the azimuth, r_{ij} is the arbitrary direction (ie. measured by the instrument, in our case the instrument is an antenna or antenna array) and ω_i is the arbitrary orientation of the instrument. This equation can be converted into a parametric observation equation by noting the the azimuth from i to j can be expressed using the $\tan$ function. Specifically

$$\alpha_{ij} = \arctan\left(\frac{x_j - x_i}{y_j - y_i}\right)$$

or if the observation is a direction, then

$$r_{ij} = \arctan\left(\frac{x_j - x_i}{y_j - y_i}\right) - \omega_i$$

If you differentiate these equations with respect to x and y, the result are the terms for the design matrix for azimuth or direction observations. The general forms for these will be for azimuth

$$A_i = \begin{bmatrix} \frac{y - y_i}{r^2} & -\frac{x - x_i}{r^2} \end{bmatrix}$$

or for direction

$$A_i = \begin{bmatrix} \frac{y - y_i}{r^2} & -\frac{x - x_i}{r^2} & -1 \end{bmatrix}$$

In both cases, the sign conventions will depend on whether you are differentiating with respect to the i to j coordinates. Note that in the case of the direction observations, the orientation must be either solved for as an additional parameter or cancelled out by differencing observations. If observations are differenced, care must be taken to also

difference the corresponding covariance matrices as well. This process will make the differenced observations mathematically correlated. This is often ignored, but is exactly analogous to the process of mathematical correlation that applies to the case of differencing pseudoranges to get TDOA observations. This is discussed in much more detail in section (e) below. If you do not want to difference, then the process of estimating the orientation is exactly analogous to estimating the receiver clock offset in the pseudoranging problem, also discussed in section (e) below.

ii. Location from TOA

In two dimensions, the TOA observation equation has the form

$$r_i = \sqrt{(x - x_i)^2 - (y - y_i)^2} + \varepsilon$$

where r_i is the observed time of arrival, or range of a signal from reference station i located at (x_i, y_i) to the user located at (x, y) where all errors are represented by ε. A 2-D position estimate can be obtained from two or more such observations using least squares. The two observation solution is ambiguous and for the solution to converge to the correct answer, the initial guess, or point of expansion, in the non-linear least squares solution must be on the correct side of the line between the two base stations. Consider the simplest example, with two base stations located at $(0,1)$ and $(1,0)$ with two ranges (times of arrival) of one measured. In this case the user could be either at $(0,0)$ or $(1,1)$ as shown in the figure below.

Note that the circles surrounding each base station indicating a range of 1 unit are examples of lines of position.

In math, the position solution depicted above can be represented by the covariance of the estimated states $C_{\hat{x}} = (A^T A)^{-1}$ (where for simplicity we are assuming that the covariance matrix of the observations is the identity matrix. To evaluate $C_{\hat{x}}$ the A matrix must be evaluated. For ranging, the ith row of the A matrix is

$$A_i = \left[\frac{x - x_i}{r} \quad \frac{y - y_i}{r} \right]$$

where

$$r = \sqrt{(x - x_i)^2 - (y - y_i)^2}$$

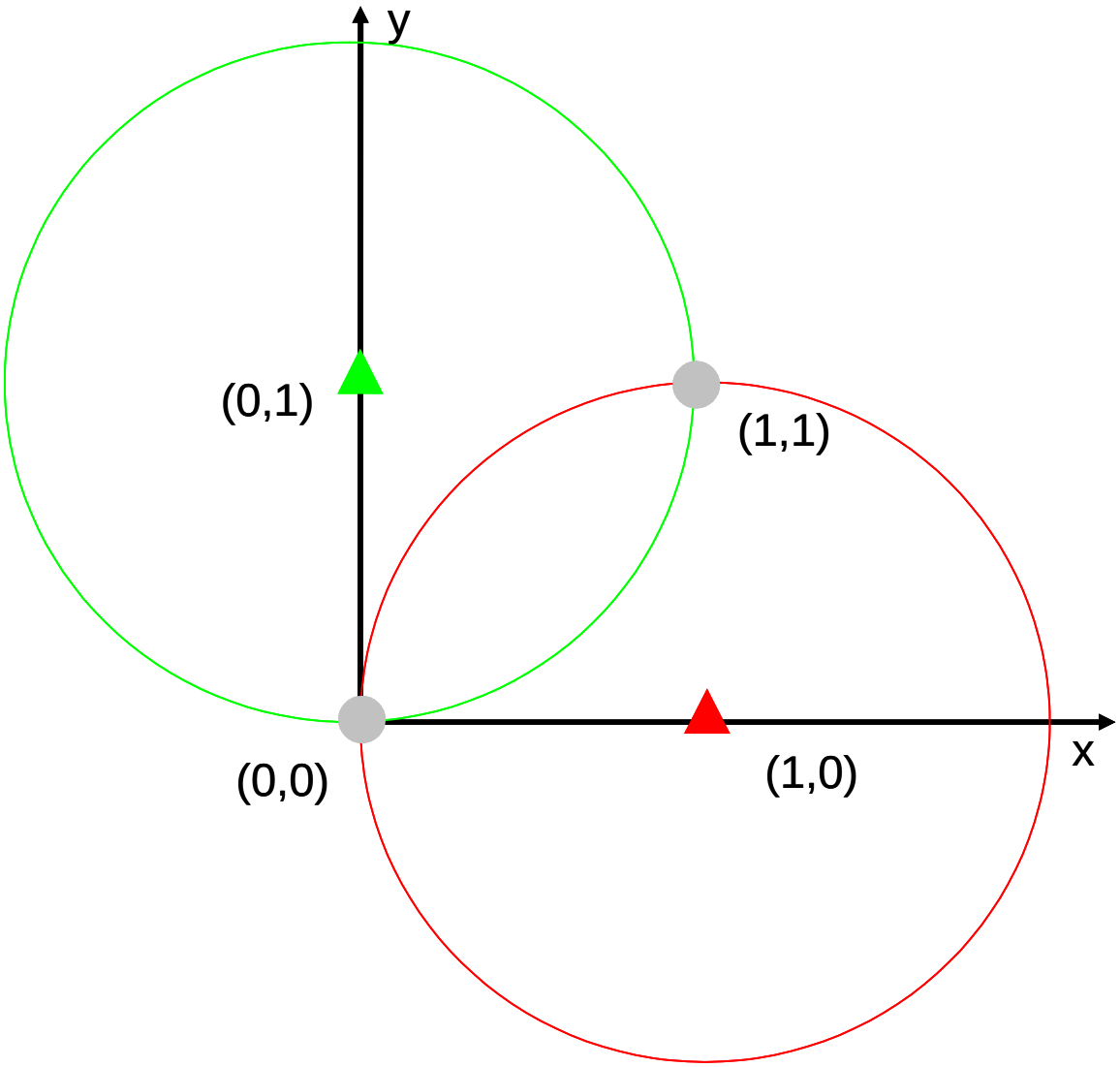


Figure 1: Two range measurements.

The elements of the row are called the direction cosines. They are the partial derivatives of the i th observation equation with respect to x and y and are equivalent to the x and y components of the line-of-sight unit vector.
in this case the A matrix is

$$A = \begin{bmatrix} \frac{1-1}{1} & \frac{1-0}{1} \\ \frac{1-0}{1} & \frac{1-1}{1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and

$$C_{\hat{x}} = (A^T A)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This result makes sense intuitively since two orthogonal observations with variance 1 were used to two parameters, in effect the first observation observes the y component and the second the x component.

Adding a third station will eliminate this ambiguity in addition to providing a redundant observation for a least squares solution.

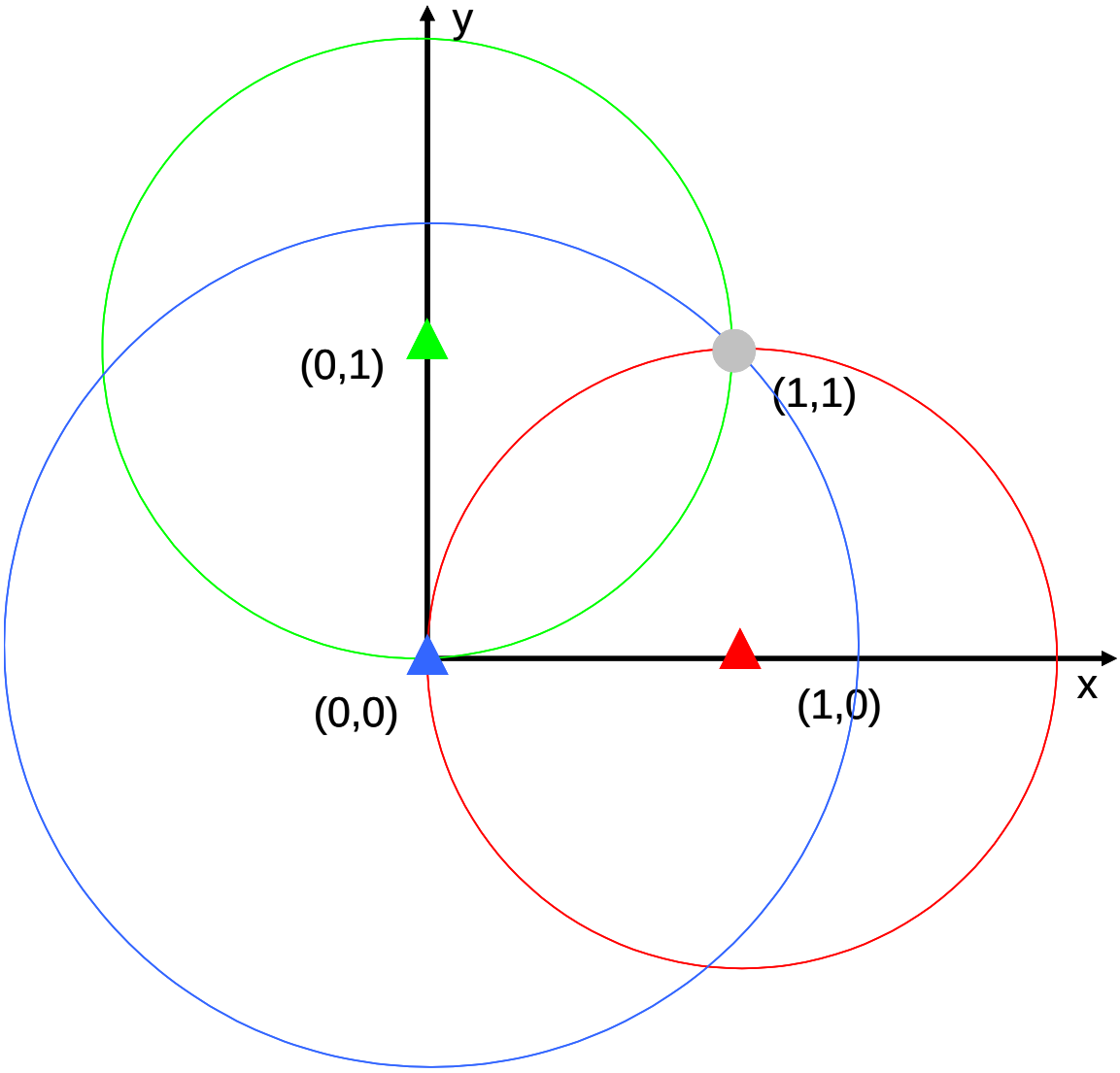


Figure 2: Ranging with three ranges. Note the position of the user is no longer ambiguous.

In terms of math, an additional row is now added to the design matrix

$$A = \begin{bmatrix} \frac{1-1}{1} & \frac{1-0}{1} \\ \frac{1-0}{1} & \frac{1-1}{1} \\ \frac{1-0}{\sqrt{2}} & \frac{1-0}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

And

$$C_{\hat{x}} = (A^T A)^{-1} = \begin{bmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{bmatrix}$$

Notice that adding the third observation has improved the variance of both states, and has also made the two states estimates correlated. Also notice that the correlation is negative, meaning that a small increase in x would lead to a decrease in y and vice versa. You can visualize this as the solution moving a very small amount either way along the blue circle since the blue observation will try to keep the solution located along the circle. The other two observations of course will try to prevent this.

iii. Location from TDOA

Time difference of arrival measurements can be used in two different ways to obtain a position solution. Either the differences themselves can be combined to generate a solution, or the biased times of arrival, also known as pseudoranges, can be combined to solve for position and local clock offset. Using the same geometry, the two options are illustrated below. Consider the situation where the user has a clock offset of +0.25 units. ie. the user clock is 0.25 units fast.

In the case where the observations are differenced, the receiver clock offset doesn't matter since it cancels in the difference. If three base stations transmit at the same time, then the signal from 3 (blue) will arrive at the rover 0.414 units after signals have arrived from 1 (red) and 2 (green). In this case the rover will observe time differences of arrival of 0.414 units for both the 3-1 and 3-2 differences. The lines of constant difference corresponding to these TDOA observations are shown in below as cyan and magenta hyperbolas.

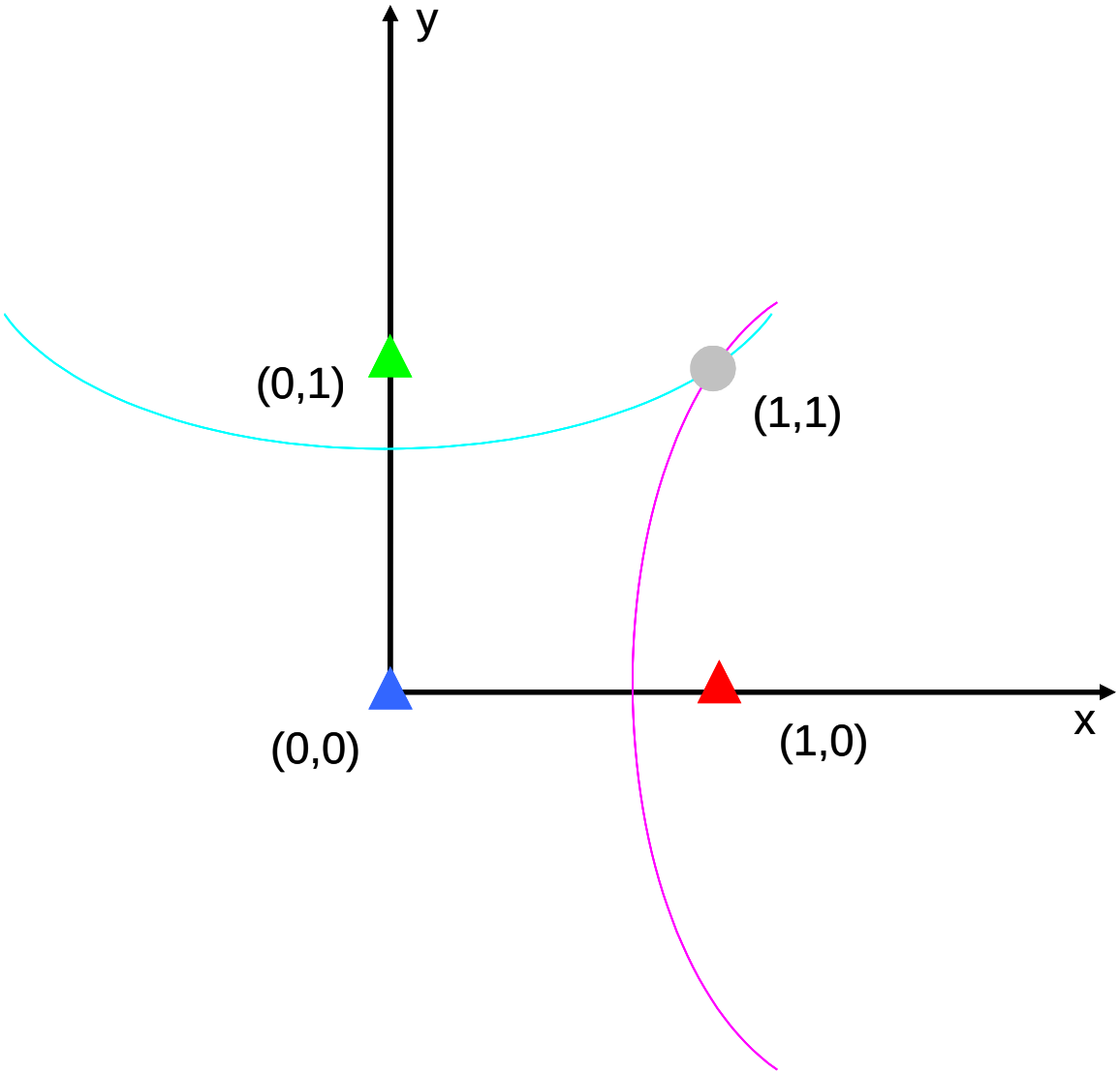


Figure 3: Lines of Position for TDOA using three base stations. Note that there are two other ways to difference the three observations and all three will result in the same solution.

The second option is to use pseudoranging, that is to estimate the receiver clock offset explicitly. This adds a parameter to the adjustment and can be represented geometrically by drawing a circle of radius 0.25 around the user (This is a value that needs to be determined, in this example we already know what the value is). Ranging signals from the synchronized base stations then arrive after 1.0,1.0, and 1.4 units respectively are recorded by the user as having arrived at 0.75, 0.75, and 1.15 units (represented by the three coloured circles, now correspondingly smaller.

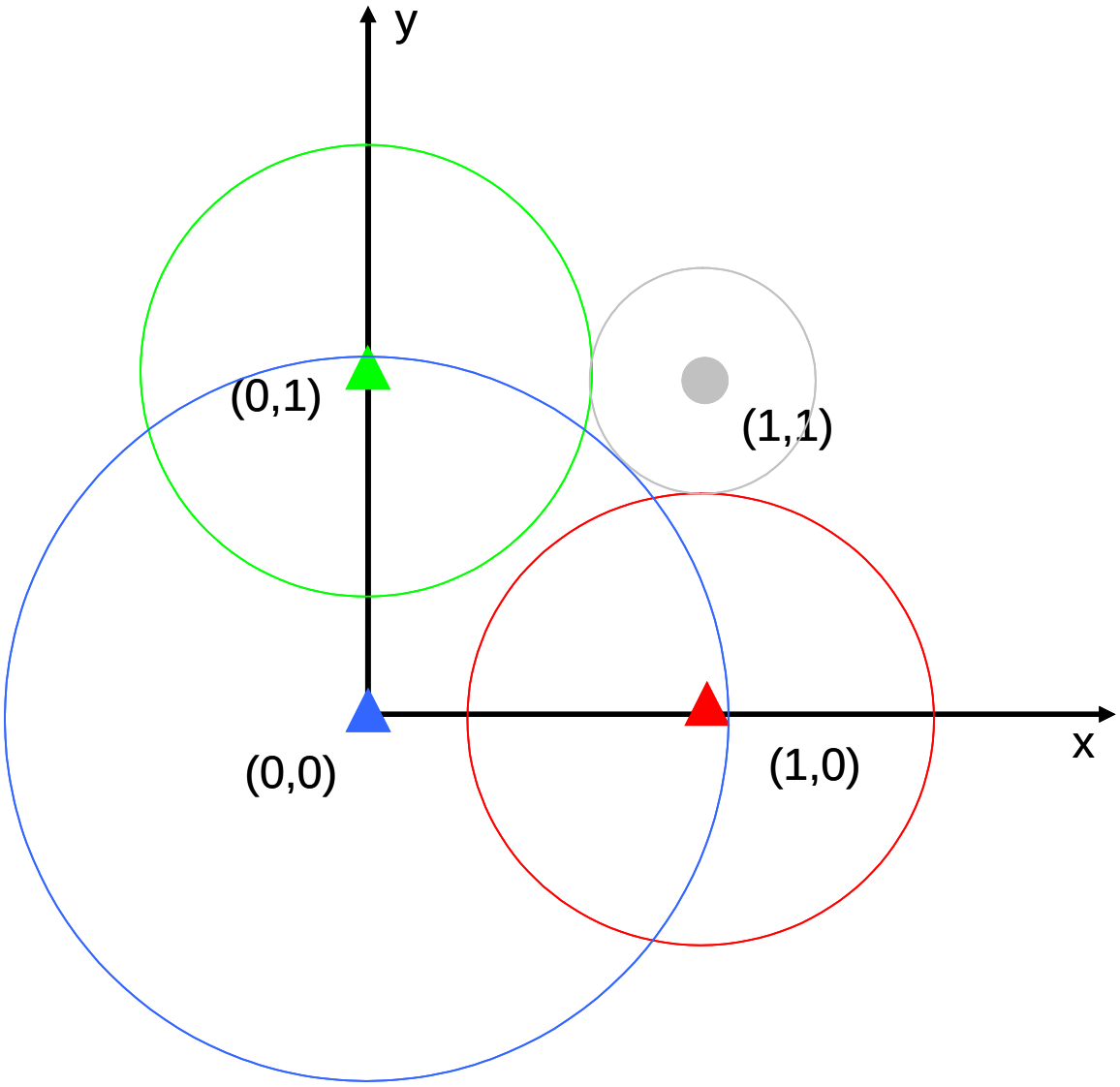


Figure 4: Lines of position for pseudoranging. Each pseudorange is represented by a coloured circle and consists of the sum of the true range and the receiver clock offset (the gray circle).

Both of these methods can be represented in math. It is more intuitive to start by showing the design matrix for pseudoranging.

$$A = \begin{bmatrix} \frac{1-1}{1} & \frac{1-0}{1} & 1 \\ \frac{1-0}{1} & \frac{1-1}{1} & 1 \\ \frac{1-0}{\sqrt{2}} & \frac{1-0}{\sqrt{2}} & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 1 \end{bmatrix}$$

Where the new column has been added to correspond to the new parameter (clock offset) Evaluating gives

$$C_{\hat{x}} = (A^T A)^{-1} = \begin{bmatrix} 9.242 & 8.242 & -9.950 \\ 8.242 & 9.242 & -9.950 \\ -9.950 & -9.950 & 11.657 \end{bmatrix}$$

Note how the accuracy achievable with the uniquely determined pseudoranging case is much poorer than with the overdetermined ranging case.

The same result can be obtained by differencing the observations (ie. TDOA mode, or hyperbolic mode). It is helpful to explicitly show the differencing matrix. In the case illustrated, we are forming the differences 3-1 and 3-2. This can be expressed by the differencing matrix

$$B = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

which can be used to operate to transform three pseudoranges into two differences (and cancel the clock offset in the process. The B can also be applied directly to the pseudoranging A matrix to form the differenced A matrix and to the covariance matrix of the observations (in this case identity) to form a covariance matrix of the differenced observations that fully accounts for the mathematical correlation between the differenced observations.

$$A_{TDOA} = BA_{pseudoranging} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 1 \end{bmatrix} = \begin{bmatrix} 0.707 & -0.293 & 0 \\ -0.293 & 0.707 & 0 \end{bmatrix}$$

and

$$C_l = BIB^T = BB^T = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Note that the column of zeros in A shows that in the difference, the clock offset is no longer observable and there are now off diagonal elements in C_l and the diagonals in C_l are two representing the fact that the difference has twice the variance of the original observation.

The estimated covariance of the states

$$C_{\hat{x}} = (A_{TDOA}{}^T C_l^{-1} A_{TDOA})^{-1} = (A^T B^T (BC_l B^T)^{-1} BA)^{-1} = \text{singular}$$

because

$$(A^T B^T (BC_l B^T)^{-1} BA) = \begin{bmatrix} 0.529 & -0.471 & 0 \\ -0.471 & 0.529 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

contains a row and column of zeros (ie. it is rank deficient and singular so it can't be inverted). If you remove the 3rd row and column and invert, the result is

$$C_{\hat{x}} = \begin{bmatrix} 9.242 & 8.242 \\ 8.242 & 9.242 \end{bmatrix}$$

Which is the same result as in the pseudoranging case. An interesting property of this result is that you always get the same result, regardless of the differencing scheme used, provided that the B matrix does not contain any rows that are linearly dependent on the other rows. (This is the same as saying that each observation is used in at least 1 difference, and that no observations are used in a closed loop of differences). This can be shown be evaluating $B^T (BB^T)^{-1} B$ for various B's. In the 3 observations/2 difference example shown here,

$$B^T (BB^T)^{-1} B = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

and in general

$$B^T (BB^T)^{-1} B = I_{n \times n} - \left[\frac{1}{n}\right]_{nxn}$$

where $\left[\frac{1}{n}\right]_{nxn}$ is an nxn matrix with every element equal to $\frac{1}{n}$.

You will get a different result if you ignore the fact that the observations are correlated. Many operational systems do in fact ignore the correlation. In this example, if you replace the correlated observation covariance matrix with the identity matrix, the result is

$$C_{\hat{x}} = (A_{TDOA}{}^T C_l^{-1} A_{TDOA})^{-1} = (A^T B^T BA)^{-1} = \begin{bmatrix} 3.414 & 2.414 \\ 2.414 & 3.414 \end{bmatrix}$$

which is more optimistic that the correct value. This makes sense because a correlated observation contains less observing power than an independent one. Furthermore, you will find that you will get different results for different differencing schemes if you fail to take into account the correlation of the observations.

c) Accuracy Measures

The best accuracy measure is the covariance matrix of the estimated states itself. This matrix contains all the information available about the covariance of the state vector and can be used to generate the error-ellipse (in 2D) or ellipsoid (in 3D) etc. In the 2-D case, the 2x2 covariance matrix contains 3 independent elements, a variance n and e (or y and x) and a covariance. The error-ellipse can be constructed by finding the eigen-values and eigen vectors of the covariance matrix. The eigen vectors will be orthogonal and represent the directions of the semi-major and semi-minor axes of the covariance ellipse. The eigen values are the magnitudes of each semi-axis. This can be generalized to an unlimited number of dimensions. For the 2-D case

$$a = \left[\frac{1}{2}\left(\sigma_e^2 + \sigma_n^2\right) + \left[\frac{1}{4}\left(\sigma_e^2 - \sigma_n^2\right)^2 + \sigma_{en}^2 \right]^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

$$b = \left[\frac{1}{2}\left(\sigma_e^2 + \sigma_n^2\right) - \left[\frac{1}{4}\left(\sigma_e^2 - \sigma_n^2\right)^2 + \sigma_{en}^2 \right]^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

$$\theta = \frac{1}{2}\arctan\left[\frac{2\sigma_{en}}{\sigma_e^2 - \sigma_n^2} \right]$$

where θ is the azimuth of the semi-major axis.

A 2-D error ellipse only encloses 39% of the probability density of the solution. In order to have a higher probability region, the semi-axes must be multiplied by scale factors. For 90%, the factor is 2.15. For 95%, the factor is 2.45, and for 99% the factor is 3.03.

DRMS and CEP

Two other measures are commonly used are the “Distance Root Mean Squared” (DRMS) and “Circular Error Probable” (CEP). Both of these were developed to simply the math in the era when calculations were done by hand. DRMS is the square root of the trace of the 2-D covariance matrix. It is an important measure in GNSS accuracy because of it’s relationship with Horizontal Dilution of Precision (HDOP). CEP is an even more simplified measure of accuracy and is defined as the the radius of a circle containing half of the probability density function of the position estimate. It originated as a measure of accuracy for ballistics, specifically if you fire n shots with a cannon, what is the radius that contains n/2 of the hits. Derivations of both will be presented in class.

d) Sequential Least Squares and the Linear Kalman filter.

There is a third LS formulation called “Sequential Least Squares” that once again rearranges the batch LS solution. The sequential solution is recursive, that is it is expressed in terms of the answer from the first batch and the change to that answer resulting from adding new observations. The standard solution for Sequential LS is given by

$$\hat{\delta}^{(+)} = \hat{\delta}^{(-)} - K[w_2 + A_2\hat{\delta}^{(-)}]$$
$$C_{\hat{\delta}^{(+)}} = N_1^{-1} - KA_2N_1^{-1}]$$

where K is called the gain matrix and is given by

$$K = N_1^{-1}A_2^T[C_l + A_2N_1^{-1}A_2^T]^{-1}$$

What the sequential LS solution represents is a weighted average of the new observations and the previous parameter estimates in a way that produces new parameter estimates that are equivalent to the parameter estimates that would be obtained had all the observations been used together. The biggest advantage of the sequential solution is that it is recursive. Only the previous estimate and its covariance matrix need be stored from epoch to epoch.

It should be noted that sequential LS can process one observation at a time as long as that observation is independent of the other observations and a previous solution exists. In this case the misclosure vector becomes a scalar, the design matrix has only one row, and the inverse in the gain matrix becomes a scalar inverse. The quality is very useful in indoor navigation and/or integrated systems where you might only have one observation available at one time (and not enough for a full solution). This one observation is still better than no observations and can be used in the solution.

The Kalman Filter is an extension of sequential least squares to the case where the parameters being estimated vary from one epoch to the next. By this we mean that they physically vary, not just that their estimates vary. In other words, sequential least squares can only be used to estimate the value of a stationary process, while Kalman filtering can be used on time varying processes.

i. A comparison of common Kalman Filter and Least Squares notation

Most Kalman filter literature uses different notation from what we have seen for Least squares. The matrix formulations are equivalent, but different letters are used to represent each matrix. The equivalent expressions are shown in the table below. For simplicity the linear LS case is shown. Note there are some minus signs missing.

Least Squares	Kalman Filtering
Parametric Batch	
$l = Ax + r$	$z = Hx + v$
C_l	R
$\hat{x} = (A^T C_l^{-1} A)^{-1} A^T C_l^{-1} l$	$\hat{x} = (H^T R^{-1} H)^{-1} A^T R^{-1} z$
$C_{\hat{x}} = (A^T C_l^{-1} A)^{-1}$	$P = (H^T R^{-1} H)^{-1}$
x is called the parameter vector	x is called the state vector
Sequential	
$\hat{x}^{(+)} = \hat{x}^{(-)} - K[w_2 + A_2 \hat{x}^{(-)}]$	$\hat{x}^{(+)} = \hat{x}^{(-)} + K[z - H\hat{x}^{(-)}]$
$C_{\hat{\delta}^{(+)}} = N_1^{-1} - KA_2 N_1^{-1}]$	$P^{(+)} = P^{(-)} - KHP^{(-)} = (I - KH)P^{(-)}$
$K = N_1^{-1} A_2^T [C_l + A_2 N_1^{-1} A_2^T]^{-1}$	$K = P^{(-)} H^T [HP^{(-)} H^T + R]^{-1}$

ii. The Kalman Filter and accounting for changes in the parameters

The reason for the (-) and (+) in the sequential formulas is that (-) represents the state and its covariance before being updated with new observations and the (+) quantities represent the states after they have been updated. The Kalman filter adds one more element to this procedure, the idea of prediction. If there is a model for the behaviour of the state vector as a function of time, it can be used to predict the state vector forward in time to the time of the next observations. The next observations can then be used to update the state vector. The newly updated state vector can then be predicted to the time of the next observations and so on.

The behaviour of the state vector over time can be described by a dynamics model. In continuous time, a dynamics model can be given by a system of first order linear differential equations

$$\dot{x} = Fx$$

That is each element of the first time derivative of the state vector equals some linear combination of the state vector with the linear combination given by the dynamics matrix F. In discrete time, this same relationship can be expressed as

$$x_k = \Phi x_{k-1}$$

where Φ is the transition matrix that represents the linear combination that predicts the current epoch as a function of the state vector from the last epoch.

The covariance matrix can also be propagated forward remembering the error propagation rule that when a linear transformation is applied to a vector, the same linear transformation must be applied to both sides of the corresponding covariance matrix.

$$P_k^{(-)} = \Phi P_{k-1}^{(+)} \Phi^T + Q$$

The additional matrix Q is called process noise and is normally added to represent the fact that predicting the state vector forward in time is based on a less than perfect model. Often the process noise is time dependent to represent the fact that as more time elapses, our knowledge of the state vector decreases. The only way to reduce the covariance is through adding new observations with an sequential update step. Thus the standard procedure for Kalman filtering is, starting with an initial value, predict the state and covariance forward to the time of the observations, then update the state and covariance with the observations and then predict forward to the next observation interval.

This type of filter is very useful since it can incorporate individual observations that occur at different times. It can also output a state value at any prediction point, which means that a Kalman filter can provide the best available state estimate at regular intervals even though observations may be irregularly spaced. The choice of transition matrix and process noise depends on the type of dynamic system that is being modeled by the state vector. The simplest possible case involves an identity transition matrix and zero process noise. In this case the KF reduces back to sequential least squares.

A simple example of a transition matrix is the constant velocity model. In this model, position and velocity are being estimated and the dynamic model assumes constant velocity over time, and a position that depends on previous position and velocity. If the state vector is

$$x = [x, y, z, v_x, v_y, v_z]^T$$

then the transition matrix is

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

assuming that the time interval between the two epochs is Δt . If you multiply this matrix by the state vector, you get three equations each saying position = position + velocity x time and three equations saying velocity = velocity, hence the name “constant velocity model”.

State vectors and transition functions can be chosen to suit the estimation task at hand. Kalman filters are very common in carrier phase GPS, inertial navigation, and sensor

integration, especially when many sensors are being used and these sensors have biases or systematic error that need to be estimated during each run.

The process noise added to the state covariance serves to make the filter slowly forget earlier observations. This is very useful if the state vector is expected to change over time, but in an unpredictable way. Without process noise, the state covariance will get smaller with every observation update, meaning that eventually the state estimate will be so well known that new observations will have no additional effect on the solution, in order to avoid this, process noise is added. Choosing appropriate process noise is not easy and balancing the process noise with the observation noise is known as filter tuning.

ENGO 585 Notes: Chapter 3 Part.

a. Fundamentals of RF propagation

To introduce radio-frequency propagation, we have to start with some basic electromagnetic theory. We will start with Maxwell's equations that describe the relationship between charge, current, electricity and magnetism.

Maxwell's equations in a vacuum are:

$$\nabla \cdot \mathbf{E} = \frac{1}{\varepsilon_0}\rho \qquad \text{Gauss' Law}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{Faraday's Law}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{No name}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \qquad \text{Ampere's Law + Maxwell's displacement current}$$

Where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields, ε_0 and μ_0 are the electric permittivity and magnetic permeability of free space, ρ is the electric charge density, and $\mathbf{J}$ is the current density. Sometimes, Maxwell's equations are written slightly differently by defining auxiliary fields $\mathbf{D} = \varepsilon \mathbf{E}$ and $\mathbf{H} = \frac{1}{\mu}\mathbf{B}$ where ε and μ are the permittivity and permeability of the material as opposed to free space. In general, the magnetic permeability does not vary much in non-ferromagnetic materials while the permittivity varies substantially between materials. These two definitions also change if the material in question is electrically polarized, but this topic will not be discussed in this course. With these substitutions, the equations become:

$$\nabla \cdot \mathbf{D} = \rho \qquad \text{Gauss' Law}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{Faraday's Law}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{No name}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{Ampere's Law + Maxwell's displacement current}$$

Some authors call $\mathbf{D}$ the electric flux density (as opposed to the electric field $\mathbf{E}$) and conversely call $\mathbf{H}$ the magnetic field (as opposed to the magnetic flux density $\mathbf{B}$). My preference is to simply call them the $\mathbf{E}$, $\mathbf{D}$, $\mathbf{B}$, and $\mathbf{H}$ fields.

We will not go into how these equations were derived in this course, but we are interested in what they mean and how the can be used to describe electromagnetic wave propagation.

Gauss' Law states that the divergence of electric flux equals the charge density. This equivalent to saying that charges are sources of electric field lines, or that the total electric flux through a closed surface is equal to the total charge inside the surface. The equation with no name simply affirms that there are no "magnetic charges". Another interpretation is that all magnetic field lines form closed loops. Whether or not magnetic charges called magnetic monopoles is still a matter of debate among particle physicists, but in our part of the universe and normal energies none have ever been observed. Finally, and most importantly, Faraday's law means that change in the $\mathbf{B}$ creates an $\mathbf{E}$ field and Ampere's law states that a changing $\mathbf{E}$ (or $\mathbf{D}$) field or a moving charge (ie a current) will create a $\mathbf{B}$ field. This effect can be observed when a $\mathbf{B}$ field is created by a current moving in a wire, for example.

Ok, so what is the point of this? Maxwell's equations govern how electric and magnetic fields and charges and currents interact. Now consider the case where and electric and magnetic field exist in region space with no charges or currents. In this case, Maxwell's equations reduce to

$\nabla \cdot \mathbf{E} = 0$ Gauss' Law

$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$ Faraday's Law

$\nabla \cdot \mathbf{B} = 0$ No name

$\nabla \times \mathbf{B} = \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}$ Ampere's Law

These are a set of four coupled first order differential equations. By applying the $\nabla \times$ operator to both sides, using the vector identity $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$, and moving the time derivative outside of the curl operator Faraday's Law becomes

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla \times (-\frac{\partial \mathbf{B}}{\partial t})$$
$$\nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{B})$$

Substituting Gauss' Law on the left results in

$$\nabla^2\mathbf{E} = \frac{\partial}{\partial t}(\nabla \times \mathbf{B})$$

And substituting Ampere's Law on the right gives the final result

$$\nabla^2\mathbf{E} = \mu_0\varepsilon_0 \frac{\partial^2\mathbf{E}}{\partial^2 t}$$

A similar procedure can be applied to obtain a similar result for the $\mathbf{B}$ field

$$\nabla^2\mathbf{B} = \mu_0\varepsilon_0 \frac{\partial^2\mathbf{B}}{\partial^2 t}$$

These two results are 2nd order differential equations that describe wave motion with a speed of

$$v = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

and if you plug in experimentally observed values of ε_0 and μ_0 you get $v = 3.0\times10^8$ m/s.

So Maxwell's equations show that an $\mathbf{E}$ and $\mathbf{B}$ fields travel naturally as waves with the speed of light. If ε and μ are substituted for ε_0 and μ_0, a slower speed is obtained, representing the fact that EM waves travel slower in media than they do in a vacuum. The ratio of the speeds $n = \sqrt{\frac{\mu\varepsilon}{\mu_0\varepsilon_0}}$ is called the refractive index.

EM waves obtained by this method can be shown to be transverse waves, that is the amplitude of the wave is perpendicular to the direction of travel of the wave. Also, the $\mathbf{E}$ and $\mathbf{B}$ are perpendicular to each other an in phase. For details about this, see any introductory electrodynamics textbook.

So far we have seen that electromagnetic waves exist and travel at the speed of light in a vacuum and slower in media. Now how do we make them? The simple answer to this is that electromagnetic waves are created by accelerating charges. Imagine a wire with a DC current. We know, from Faraday's Law that a wire carrying a DC current will set up a magnetic field that curls around the wire according to the right hand rule. Current is simply charges moving, a time varying current will cause a time varying magnetic field. A time varying magnetic field will lead to a time varying electric field, and the two fields will continue to propagate in this manner. This simple model also introduces the simplest type of antenna, a piece of wire that when conducting a current that is time varying at RF frequencies, radiates some of the energy of that current in the form of RF radiation. In general, wireless communications systems consist of transmitter electronics that generate a signal with a particular shape that is passed through an antenna and is radiated. At the other end (or after reflection and return to the transmitter in the case of radar for example)

another antenna receives the radiation and converts it into an electrical signal that can be interpreted. Antennas are often described by their gain and gain pattern. The gain of an antenna is ratio of the peak power transmitted in a particular direction to the power transmitted by a theoretical isotropic antenna (one that transmits power equally in all directions.) It is usually expressed in decibels, where 3dB represents a factor of 2 and 10dB a factor of 10.

The example of an antenna formed by a single wire also introduces the concept of EM wave polarization. In an EM wave, the E-field may point in any direction that is perpendicular to the direction of propagation. In the case of a vertical wire, the E filed is parallel to the wire and the B-field perpendicular. In order to receiver this energy most efficiently, the receive antenna must be polarized in the same direction.

In narrow band RF communication systems, the waveform that is transmitted is typically a single frequency cosine. This part of the signal is generally called the carrier. For the system to be useful, the carrier must also carry information. This is done by signal modulation where information is added to the signal. At the receiver end, the signal must be demodulated in order to recover the information in the signal.

i. Amplitude, Power, Frequency, Phase

A single frequency carrier can be described by

$$f(t) = A\cos(\omega t + \varphi)$$

where A is the amplitude, ω is the angular frequency (in rad/s), and φ is the instantaneous phase.

When a signal is transmitted by RF radiation between a transmitter and a receiver, the received power is related to the transmit power by

$$P_R = P_T \left(\frac{\lambda}{4\pi R} \right)^2 G_R G_T$$

where λ is the wavelength of the signal, R is the distance between the transmitter and receiver and G_R and G_T are gains of the two antennas respectively. The effectiveness of a transmitter-receiver pair is also affected by noise. Noise arises both from thermal noise of the components of the transmitter and receiver circuits and due to random EM disturbances that enter the receiver antenna. The ratio of signal power to noise power is called signal-to-noise ratio (SNR). For the received signal to be useful, it must be significantly more powerful than the noise in the system. Though we will see later that spread spectrum systems find an interesting way around this problem.

In the dipole antenna example, the resulting RF wave consisted of an oscillating E-field that was in the plane of the antenna and an oscillating B-field in the plane perpendicular to the antenna. The direction of propagation and the energy density of the wave can be expressed by the Poynting vector.

$$\mathbf{S} = \frac{1}{\mu_0}(\mathbf{E} \times \mathbf{B})$$

Another important property both antennas and of the RF-waves is called polarization. It is defined by the plane that contains the electric field vector. It is particularly important because in order to most efficiently receive RF-waves, the receive antenna must match the polarization of the incoming RF-wave. In the simple dipole example, the transmit antenna and the transmitted wave are vertically polarized (because the E component of the RF wave is in the vertical plane). This wave will be best received by another dipole antenna that is also oriented vertically.

AM radio is a good example of a system that uses vertical polarization. AM transmit antenna are vertical towers. North American analog TV on the other hand is transmitted using horizontal polarization, and a typical outdoor VHF TV antenna consists of horizontal elements.

It is also possible to generate signals with more complex polarization by using antennas with multiple elements. For example, circular polarization can be generated with a vertical and horizontal element where both elements are driven by the same source, but one of the two elements is ¼ wavelength delayed with respect to the other. In the case, the E-fields of the two components will be 90 degrees out of phase and if the two components are vector added, the resulting E-vector will be one that traces out a circular pattern. If the amplitudes of the two components are not equal, the result will be elliptical polarization.

Polarization is important to be aware of because it affects the efficiency of receiver antennas. It is also exploited by some systems to select the direct signal path and reject reflected signals (multipaths). For example, GPS has a right handed circularly polarized signal. This is advantageous because a circularly polarized user antenna does not need to be properly oriented (in the same sense that a vertical linear polarized antenna should be held vertically). It is also useful because satellite signals are affected by a phenomena called Faraday rotation in the ionosphere. Finally, signal reflection multipaths can be shown to have left hand circular polarization and these signals are rejected by right hand circularly polarized antennas.

Gain, decibels and logarithms

In addition to polarization, antennas can be designed to select signals arriving from certain directions. This is known as antenna gain. An ideal antenna would accept signals equally from all directions. This is called an isotropic antenna. In reality these do not

exist, but the behaviour of real antennas is quantified by comparing them to isotropic antennas.

The pattern of signal strength transmitted (or received) as a function of direction is called the radiation pattern. The gain of an antenna is defined as the ratio of the peak value of the radiation pattern to a reference antenna. The reference antenna is typically an isotropic one.

It should be noted that antenna gain does not create more power, it just redistributes it. For example if an isotropic antenna is transmitting 10 W, then an antenna that transmits that 10 W only over half the surface of the antenna (ie in one direction) would be transmitting twice as much power in that direction (because no power is being transmitted in the other direction). The total power transmitted would still only be 10 W.

Antenna gain, like many quantities in electronics is usually expressed logarithmically using the decibel. A decibel is a tenth of Bel, which is a unit of gain defined as

$$1\ \text{Bel} = \log_{10}\left(\frac{P}{P_{ref}}\right)$$

Once this unit was invented, it was found to be too large to be practical, so the decibel is defined as

$$1\ \text{dB} = 10\log_{10}\left(\frac{P}{P_{ref}}\right)$$

Before moving on, a note about exponentials and logarithms. The rule is that their arguments must never have units. For example exponential decay is usually written something like

$$A(t) = A_0 e^{-bt}$$

where b is the time constant that has units 1/time.

This means that when expressing something in decibels, it is always a ratio, and if the denominator is not really easily defined, then a reference value must invented and applied. An example of this is in acoustics, where sound is expressed as decibels with respect to one micro-pascal at one metre (or yard). Another common example are the dBW and the dBm. These are common non-SI units that mean decibel w.r.t. 1 Watt and decibel w.r.t. 1 milliwatt respectively. Note that a quantity expressed in dBW should be 30 less than the same quanity in dBm since a factor of 1000 is exactly 30 dB.

It should also be noted that sometimes antenna gains are listed with the unit dBi. This just means dB w.r.t. isotropic.

The expression 3 dB is also very common since it represents approximately a factor of 2. 0 dB is the same as a factor of 1, and 10 dB is equivalent to a factor of 10.

Decibel quantities are most often used to compare powers, however in electronics they are also used to compare voltages and currents. Remember that

$$P = IV = I^2R = \frac{V^2}{R}$$

When the ratio of two powers is taken, this is equivalent to the ratio of currents squared or voltage squared so sometimes the decibel formula is expressed as

$$1\ \text{dB} = 10\log_{10}\left(\frac{P}{P_{ref}}\right) = 10\log_{10}\left(\frac{V^2}{V_{ref}^2}\right) = 10\log_{10}\left(\frac{V}{V_{ref}}\right)^2 = 20\log_{10}\left(\frac{V}{V_{ref}}\right)$$

ii. Signal Structures

In the frequency domain, the carrier occupies only one frequency. If the carrier is modulated, the frequency spectrum will spread around the central carrier frequency depending on the data rate of the information in the modulated signal and the type of modulation used. The modulation may be either analog or digital. In most modern wireless communications systems, digital modulation schemes are used, however, we will start by looking at one analog scheme called amplitude modulation or AM.

Amplitude Modulation

Consider a carrier represented by
$c(t) = C\cos(\omega_c t + \theta)$

Here ω_c is the carrier angular frequency and θ is the phase of the carrier. Now simply add modulation to the amplitude of the carrier by replacing the original amplitude C with

$$A = C + M\cos(\omega_s t + \varphi)$$

In this example, we are added a second sinusoidal signal with a lower frequency ω_s. It helps to keep in mind that in practice A could be any signal made up of a series cos and sin terms (ie. a Fourier series) and the math would be the same, but in this case we are only going to examine the simplest case, where there is a DC term C (DC here stands for direct current, which is another way of saying a component with a frequency of zero) and

a single additional low frequency term $M\cos(\omega_s t+\varphi)$ The modulated signal can now be represented by

$$f(t)=[C+M\cos(\omega_s+\varphi)]\cos(\omega_c t+\theta)$$

Which is equal to

$$f(t)=C\cos(\omega_c t+\theta)+\frac{M}{2}\cos(-\omega_c t-\theta+\omega_s t+\varphi)-\frac{M}{2}\cos(\omega_c t+\theta+\omega_s t+\varphi)$$

Notice that there are now three frequencies present, ω_c, $\omega_c+\omega_s$, and $\omega_c-\omega_s$. The latter two are called sidebands. Note that the two sidebands carry identical information. In practice the modulation is a complex waveform consisting of the superposition of many different frequencies in the range of ω_s. The larger the range of frequencies desired, the larger the bandwidth of the modulated signal. When an AM signal is received, the signal must be demodulated, that is the information in $M\cos(\omega_s t+\varphi)$ must be removed from the carrier so that it can be used (ie. heard in the case of audio).

When designing the AM broadcast radio system, the AM channels had to be spaced far enough apart so that the sidebands of a given station did not interfere with each other. The space in the frequency spectrum occupied by the modulated signal is called a band. The bandwidth literally is the width (highest frequency minus lowest frequency). When modulating a carrier with information it is important that you not exceed the allocated bandwidth. To accomplish this, sometimes the input signal must be filtered to suppress higher frequency components. Other times this filtering happens simply as a result of the equipment used to process the signal. Because the bandwidth limits the detail (ie highest frequency component) that can be carried, the term has now become a synonym with the term data rate when used in digital contexts (ie. high bandwidth internet connection).

In North America, broadcast AM radio stations are nominally separated by 10 kHz but adjacent stations will interfere with each other since the bandwidth being modulated is typically on the order of 10 kHz. As a result, powerful stations are not located in the same region only 10 kHz apart. Even with good planning of station assignments, AM is generally not good at carrying high frequencies and this is one reason that AM radio is typically used for talk and news radio and not for music. It is also possible to transmit digital data using amplitude modulation by simply turning the modulation, or the whole signal, on or off to represent a 0 or 1 state. A simple example of this is transmitting Morse code.

There are many other analog modulation schemes. SSB (AM where one of the two sidebands is suppressed), FM, analog TV (which is AM for the picture and FM for the sound)

Binary Phase-shift keying

A common form of digital modulation is phase-shift keying. The simplest example of this is binary phase shift keying which is a system where the phase of carrier is shifted by 180 degrees to represent a change from the 0 to 1 binary state. The mathematical model to represent this involves multiplying the carrier by a series of 1s and -1s to represent the binary information. When such a signal is received, the receiver has to decide whether a 0 or 1 was received. Instead of distortions caused by noise and interference in the case of analog modulation schemes, in BPSK what will result is a bit error (if the wrong decision about which bit was received is made. The bit error rate is dependent on the signal to noise ratio. The faster the sequence of bits, the more bandwidth the modulated signal will occupy, in practice though, the data rate of the modulation is usually much slower than the carrier frequency.

Direct sequence spread spectrum (DSSS) and code division multiple access (CDMA)

An interesting application of phase shift keying is the creation of direct sequence spread spectrum signals. If a carrier is phase shift modulated by a high frequency pseudorandom sequence, the result will be a very high bandwidth signal but one that will have a low power level spread across a large spectrum. When this kind of signal is received, it can be demodulated by multiplying it by a copy, or replica of the pseudorandom sequence. This will result in de-spreading, or the re-concentration of the original signal energy in a narrow band. This kind of system also makes multiple access possible since different signals occupying the same spectrum can be distinguished based on their pseudorandom code. If you multiply one signal with one pseudorandom code by a different pseudorandom code, the result will be just noise, provided the two codes were selected to have low cross-correlation. This kind of multiplexing or multiple access is known as Code Division Multiple Access (CDMA). CDMA systems include some digital cellular phones and many satellite communication and navigation systems. Data in addition to pseudorandom codes is modulated onto a carrier with the only restriction that the data bit rate must be significantly slower than the code modulation rate which in turn is much slower than the carrier frequency. In Geomatics engineering, the GPS L1 C/A code is the prime example of a data transmitting CDMA system.

Other forms of multiple access

There are two other common forms of multiple access. Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). FDMA is just another way of saying “give each user a different frequency channel” so that multiple users don’t interfere with each other. In many ways, this is the easiest to implement since it requires no codes or time scheduling and this is what is done in analog broadcasting and with the Russian GNSS GLONASS. On the other hand, antennas and receiver hardware are often frequency selective which is a disadvantage of FDMA since complicated hardware might be required to properly receive channels spread a variety of frequencies. The other method, TDMA, is when multiple users share a channel and either have a schedule for time sharing, or simply take their chances and try not to all be “on” at the same time. A

really simple example of TDMA is the party phone line. When the line is in use, the other users can’t use it.

The $70 FRS radio actually represents all three common multiple access schemes. These low power unlicensed radios can tune 14 UHF frequency channels (FDMA) and have “privacy codes” which sound like CDMA, but are in fact just a way of not hearing other users on the same channel where you and other users are sharing the channel using a random TDMA.

ENGO 585 Notes: Chapter 3 Part B.

b. Measurements

There are three general types of measurements used in wireless location systems. They are angle of arrival (AOA), time of arrival (TOA) and time difference of arrival (TDOA). A fourth measurement principle, using received signal strength (RSS) is really just another method of TOA, but will be treated separately below. First we will discuss how these measurements are made, and then we will deal with how to convert each type of measurement into a position solution.

We will consider a mobile user with a supporting network of stationary base stations. In general each of these measurements can be made either by the user or the base stations, though some are more suited for the mobile station implementation than others

(i) Angle or Arrival.

Angle of arrival can be measured in one of three different ways: Using mechanical scanning, beamforming, or estimation. The conceptually simplest way is to mechanically scan with a high-gain (ie. very directional) antenna and attempt to locate the direction of maximum received signal strength. The target could either be transmitting or passive and the rotating antenna could either be on the base station or the mobile user.

The problem with this approach is that it requires mechanical parts. To overcome this the phase array antenna was developed and is used for beamforming. The simplest antenna array consists of several identical antennas lined up in a row. Beamforming can be accomplished by phasing the output of the antennas in such a way that they constructively interfere for signals that arrive from a desired direction.

Consider an example with several antennas lined up along the x-axis, with the output of each connected to a time-delay element before being combined with the other and fed to a receiver of some sort. To “point” the antenna in the direction of the x-axis, all that is required is that the output of each antenna be combined with zero delay. Then, assuming the source if “far” away, the incoming wave fronts will all arrive at the antennas at the same time and constructively add.

Note that such an antenna would not be able to distinguish signals arriving from the north from those arriving from the south (assume each individual antenna has the same gain in all directions in the x-y plane.

No consider a signal arriving from the positive x-direction (ie. the East). For this signal to be rejected, the spacing of the antennas must not be an integer multiple of the wavelength.

So far we haven’t used the delay elements. To select a signal from another direction, the delay elements can be adjusted.

We will compute what the delay needs to be as an exercise in class.

One final thing to consider (without proof). As more and more elements are added, the directivity of the array will be increased. Think of it in terms of adding more and more waves constructively. Also, a 2d array would be required if you wanted to direct the antenna vertically as well. This entire system is also completely analogous to the towed sonar array.

It is also possible to invert this procedure provided that you have lots of electronics. Instead of beamforming, you could simple track the signal from each antenna and then try to determine the relative delay between antennas. Using a correlation, or correlation analysis techniques, you can then determine the delay and then invert the delay to determine the angle of arrival of the signal. This method is much more effective since you are not steering the antenna (and in the process ignoring possible signals arriving from other directions).

(ii) Time of Arrival

Time of arrival is perhaps conceptually the easiest type of wireless location observation to understand. Simply measure one-way or two-way travel time and multiply by the speed of light to get a one-way or two-way distance. In general two-way methods are easier to implement since a lower precision clock can be used at the transmitter/receiver than would be required by both the transmitter and receiver for one-way methods. First, we will review 3 measurement principals that can be used for two-way measurements. In principal all of these can be used in one-way mode as well.

Pulse Time of Arrival.

This conceptually the simplest. A transmitter is turned on and a short duration pulse is transmitted. The pulse is turned off and the antenna is switched to the receiver. When the reflection of the pulse returns, the time between the leading edge of the transmit pulse and received pulse is measured.

In this method, the remote reflector can be either passive or active. A passive reflector simply needs to be large enough and reflective enough to return enough power to be detected. Another alternative is to have an active receiver that receives the pulse and this reception triggers the transmission of a response pulse. If this is the case, the turn-around time, that is the time delay between reception and transmission, must be calibrated in advance and the electronics must be such that this time is more or less constant. It is also possible to have a semi-passive reflector, that is not powered internally, but actually uses the received power to generate a reflected signal with identifiable characteristics. We will discuss this principal later in the course as it is used in a technology called RFID.

We will not discuss the electronics required to detect pulses in this course, but keep in mind that the sharpness of the pulse is proportional to the bandwidth of the signal. In

other words, a very sharp pulse has many high frequency components so modulating one onto either a carrier or onto the baseband requires a lot of bandwidth.

Another thing to consider with such a system is that the range is limited by the pulse width and pulse repetition frequency since the pulse it typically completely sent before it returns, and it must have finished returning before the next pulse is sent. If this is not done, the there will be a pulse ambiguity and you will have to determine which of the pulses is in fact returning.

Frequency modulation (Modulated continuous-wave radar).

Measuring a two-way travel time of a pulse is not easy. A simpler method, especially with analog electronics involves transmitting a frequency modulated carrier. The transmitted and received signals are then compared and the difference in frequency between them represents the two way travel time, assuming that the rate of frequency modulation is known. This frequency modulation could either be sinusoidal or linear.

Correlation Methods

Perhaps the most effective method is one that involves correlating the transmitted and received signals in order to determine the delay between the two. In principal this can be done with any signal, provided that is it sampled as it is sent and then these samples can be compared to the received signal when it arrives. In practice this is usually done by transmitting a known pseudorandom noise sequence (often called a PRN code) and then comparing the received signal to a replica of the code. Note that PRN codes were introduced to provide code division multiple access, which is a form or multiplexing, but here they have the side affect of being useful for determining a range as well.

Correlation Example: Was shown in class in Chapter 2.

From two-way to one-way

All of the above methods work in principal in one way mode as well, provided that the timing issues can be resolved. Pulse detection and correlation work the same way, while frequency modulation requires that the receiver have both a synchronized clock and frequency since the local frequency must be used to determine the frequency modulation state of the received signal. The main difference with two way methods, in addition to the timing issues, is that the receiver must have enough information to produce replicas of the received signals for comparison. This means it must know when pulses are being sent, how the frequency is being modulated, or when and what codes are being transmitted.

(ii) Time Difference of Arrival

Most of the timing issues that limit TOA in one-way mode can be overcome by switching to a time difference of arrival (TDOA) method. Both pulsed and correlation based systems can be easily used in TDOA mode.

The basic principal of TDOA is that now instead of observing the travel time of the pulse (or delay in the correlation), the difference between the arrival of two pulses (or between two correlation delays) is used instead. The methods for making the observations are the same as for one-way TOA, but how they are dealt with changes since the receiver (user) clock is now unknown. There are two ways of dealing with the unknown clock offset. It can be eliminated by differencing (hence the "differencing" in TDOA) or it can be estimated by pseudoranging. Both methods are mathematically equivalent (if the math is done properly and mathematical correlation is accounted for) and this will be demonstrated in class in Chapter 3 Part (e) when the mathematical model for converting TDOA measurements into position solutions is discussed.

(iv) Received Signal Strength as a Wireless Location Measurement Principle

A final possible wireless location measurement is to simply measure the signal power that is received. This is conceptually easy, but has certain limitations. First, for this method to be useful, you must have a very good idea of the propagation environment and of the gain patterns of the both the transmit and receive antennas. If both are isotropic and free space loss is assumed, then provided you know the transmit power, received power can be converted directly into a range (ie. the power decreased by the inverse square of the range). If it is a two-way system, then you additionally need to know the reflection coefficient of the reflector.

In practice though, antennas are not isotropic and the propagation loss does not follow a free space model. In this case, it is still possible to extract information from the received signal strength, however to use it to determine a location will require that the area to be served be pre-surveyed and a map of signal strength as a function of position developed that could then be used for signal strength map matching. This method also assumes that there will be many transmitters and that the propagation environment does not change as a function of time. Positioning using this method is often referred to as RF fingerprinting, since in principal each location would have a unique signal strength fingerprint.

(v) Error Sources

As with all observations in Geomatics Engineering, wireless location measurements are affected by error sources. It is important to understand these error sources so that they can be avoided, corrected, and accounted for when determining estimated accuracies. There are three errors that must be dealt with: noise, interference, and multipath. There are also several less serious error sources that can be mitigated relatively easily: Timing errors, atmospheric errors, and base station location errors

Timing Errors and Propagation Delays

To use a ground based RF system, the speed of light must be correct for the presence of the atmosphere. In relative terms, this correction is larger that it is for satellite navigation since with ground based systems the entire signal path is through the thickest part of the

atmosphere. Also, differential correction as an error mitigation scheme will be unreliable since the total path from the transmitter to a reference receiver may be significantly different that the path to the users.

A standard value for the correction (from EDM surveying) is 340 PPM. A local value may also be computed using either the Essen-Froome or Weirtraub formulas (or other). These formulas were developed empirically.

Long range ground based systems that rely on ground wave propagation require additional corrections for the ground (and sea) conductivity. These are generally not required for line-of-sight propagation systems used in wireless location. Furthermore, refraction (path bending, as observed in long range EDM for example) can be neglected in short range systems.

Base station positioning error

Base station position error will result in user position errors. The magnitude of the error will be the projection of the base station position error onto the line-of-sight vector (in the case of TOA) What would it be in the case of TDOA? For AOA systems, small position errors can be tolerated however orientation errors will become a serious issue unless angle observations are differenced.

In general a base station error's effect on an observation can be determined by multiplying the error vector by the row of the design matrix corresponding to the observations. The effect of a series of range errors on the solution can be obtained by propagation the error vector through the Least-Squares solution since the errors are additive and the equations a linear.

Diffraction

Diffraction will cause errors when LOS conditions do not exist. To correct for this, a detailed map is required as well as enough transmitters such that service with LOS transmitters will be available everywhere in the service area.

Multipath and Fading

Multipath and Fading are perhaps the most significant and limiting error sources in urban wireless location.

Static Multipath environments: Unlike satellite navigation, where the transmitters are slowly moving, with ground based wireless location, the base stations are static, meaning that if they are transmitters, they will produce static multipath distributions in the service area. If the base stations are receiver, this is still the case. There will be static multipath channels from each possible user location to the base stations. When line of sight conditions exist, there are three strategies to mitigate multipath.

1) Use a directional antenna (possibly at both ends)
2) Use a higher bandwidth (or super resolution method) to try to identify the direct signal and time tag it before the arrival of the other multipath signals. In electrical engineering, the focus is on “Channel Characterization” and receivers are designed to estimate the delay of theca multipath component. There are called “rake receivers” where each “Finger” has an estimatable delay. In communications engineering, often the goal is not ranging, but constructive addition of all the components to get more received power.
3) Calibration/Survey: This involves attempting to map the whole service area. The resulting database is then available, however in order to use it, you have to already have a position estimate, so positioning and correcting becomes an iterative process.

In addition to multipath, there is the issue of Fading. Fading results from movement through a multipath environment. As the user moves between constructive and destructive interference areas, the received signal power will vary significantly. This is called fading. Receiver must be carefully designed to be able to continuously track through fading. Fading environments can be modeled (though only to simulate, not to predict) using two statistical distributions, the Rayleigh distribution and Ricean distribution. The Rayleigth distribution results from the complex addition of a large number of identical frequency sine waves with normally distributed amplitudes and uniformly distributed phases. (Rayleigh’s original paper described the resulting signal from a large number of violins playing the same note). NLOS multipath will have a Rayleigh distribution. The Ricean distribution is a Rayleigh distribution except that one signal with a significantly higher amplitude is present. This model us used for multipath with a LOS signal present.

Echo Only Mathematical Models for Positioning

There are several ways to estimate position using multipath other. One is to estimate the location of each “last reflector” and a delay for each signal. This requires a detailed model of the transmitter and reflector locations in addition to having a good idea about the location of the users. A second method, called finger printing, is to record what the signal looks like, either in terms of power, or in terms of the shape of multiple arrivals, and store this information in a searchable database, when another user returns, the user could match the stored finger print. Both of these methods assume the multipath environment is not changing which is a major limitation, especially in crowed or industrial applications. A final method is to measure angle and time of arrival at the user and then estimate a range bias for each for each observation. Consider a two dimensional example where in LOS condition 2 TOA or 2 AOA measurements would be sufficient to located a user. With 4 observations (2 TOA and 2 AOA) it is possible to estimate 4 parameters (two positions and two biases). This method has suggested in several academic studies but has not been implemented commercially.

ENGO 585 Notes: Chapter 3 Part C.

i. Historical Systems:

Historical radio navigation systems are interesting to study because they show how angle and time/time difference of arrive combined with timing and dead reckoning could be used to obtain a position. The advantage of studying historical systems is that they are usually simple both conceptually and in implementation compared to modern systems which may be conceptually simple but complex in implementation. Most of the systems discussed below were developed during the second world war, though many are based on technologies that were being developed in the 1930's

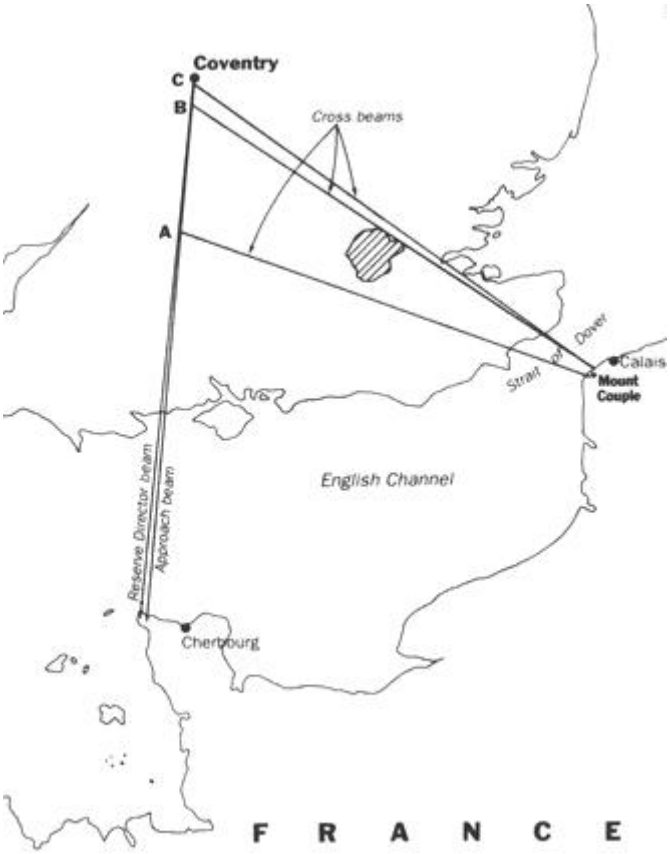
(a) Knickebein

Knickebein was an early German beam navigation system that was based on an instrument landing system that had a switched directional antenna. The antenna was a three element directional antenna where the middle element was always on and the outer two elements were switched. The system stayed "on" on one side longer that the other and resulted in two beams on either side of a centre line. If you were on one side, you heard a series of dots, on the other a series of dashes and if you were traveling right down the intersection of the two beams, you heard a constant tone. Two transmitters were set up in Germany on in the extreme north and the other in the extreme west. Each transmitter sent a beam and the target was located at the intersection of the two beams.

The British code name for the system was "head ache" and the countermeasure was called "aspirin". This involved transmitting extra, time synchronized dot signals so that when the bomber was in the dash area, they would hear the equi-signal and not correct their course. This could be used to steer the path. Also, the British deployed randomly flying planes to figure out what the target was if they were able to find both beams.

(b) X-Gerät

Once Knickebein had been defeated, the Germans developed a more advanced system consisting of four directional beams: One to fly along, and three cross beams. 30, 10, and 5 km from the target. The system consisted of a 60 MHz carrier amplitude modulated at 2kHz. The first cross beam informed the crew to turn on their targeting equipment, the second and when the bomber passed the second beam, a clock was started, when the third beam was passed, the direction of the clock was reversed. Assuming constant airspeed, the bomber would reach the target (5 km after the third beam) just as the clock returned to zero at which points the bombers were dropped. The system was reportedly accurate to 100 m.



X-Gerat beam pattern: (Oxford Companion to WWII)
http://www.valourandhorror.com/BC/Tactics/X_gerat.php

(c) Y-Gerät:

The final German attempt consisted of an amplitude modulated two way ranging system with direct phase comparison. (ie. this is a combination of both two way ranging and angle of arrival). It operated briefly on 45 MHz but was defeated by transmitting back a second signal from a British TV station. The phase multipath error was enough to make it useless.

(d) Gee

Gee was a British hyperbolic system at 30 MHz (very similar to Loran, but at higher frequency.) A Gee chain consisted of one master and two slaves stations: The master sent 1 pulse, then waited 2 ms and sent two pulses. The first slave sent one pulse 1 ms after receiving the first. The second slave sent 1 pulse 1 ms after receiving the second. Then after 4 ms the system repeats. The pulses are displayed on an oscilloscope triggered on the first pulse. The differences between all of them were recorded manually by the navigator who had a chart with hyperbolas on it.

(i) LORAN (Long range navigation)

The specifics of LORAN can be found in the notes or text of ENGO 545 Hydrography. For the purposes of this course, it is sufficient to say that it is the major example of an operational hyperbolic positioning system. Very similar to the British Gee system, but operating at lower frequencies using the ground wave instead of the line-of-sight signal.

ILS (Instrument Landing System)

ILS is still in use, though it was planned to be replaced by the Microwave Landing System and will more likely be replaced by GPS. An ILS consists of two beams, on each side of the runway transmitting around 108 MHz. One beam is amplitude modulated at 90 Hz, the other at 150Hz. Which one is predominant shows what side of the beam you are one. A similar scheme is done with two more beams 330 Hz with the centerline being at a glideslope of 3 degrees. Each runway has a different channel (for each beam) so on approach you simply dial in your channel depending on which runway you are approaching. The beams are transmitted from directional antennas. The horizontal beam's antenna is located on the ground at the foot of the runway, the glide slope beam is transmitted from a small tower located just to one side of the end of the runway. Like all safety of life systems, they also have monitor receivers that continuously test the signal. There are other components to ILS as well, including marker beacons to help guide aircraft during approach.

This map shows the location of the 2 Knickbein base stations in Germany (Stollberg and Kleeves), the two X-Gerat stations in France (Calais and Cherborg), and the three Gee base stations in (Ventnor, Daventry, and Steingot). Note that Knickbein has very poor angle of arrival geometry over England and likewise Gee has very poor TDOA (hyperbolic) geometry over Germany.

ENGO 585 Notes: Chapter 3 Part C.

ii. GPS: HS-GPS, A-GPS

It is well known that GPS does not work well indoors or in urban environments. As a result, GPS can provide coverage to most of the Earth, but not to most of the users in the developed world, who are generally located in cites and spend most of their time indoors.

Two general approaches have been developed to allow GPS to track highly attenuated signals that are present indoors. They are High-sensitivity GPS and Aided or Assisted GPS.

HS-GPS works by designing receivers with several features not present in conventional GPS receivers. First, GPS receivers have many more correlators to allow for larger searches in frequency and delay. Some are implemented in the frequency domain using FFT methods to allow for massive numbers of parallel correlations.

The second important improvement of HS-GPS is using bit prediction or squaring to try to remove the navigation message. This is important because the navigation message, that has 20 ms long data bits, limits the amount of time that the C/A code can be integrated to a maximum of 20 full codes (each full codes lasts on ms). If you integrate over a bit transition (from +1 to -1) then any additional integration after the bit transition will actually subtract from your integrated signal. In order to integrate for more than 20 ms, the receiver needs to know where the bit transitions are. One way to do this is to square the signal, but this also squares the noise. Another way it to try to predict the bit, either using some knowledge of the structure of the navigation message, or just by trying all possible bit values and then picking the combination that gives the largest total power.

The length of time that you can integrate in HS-GPS is generally limited by the ability to predict the navigation bits, and by the quality of the receiver clock.

Aided and Assisted GPS are techniques, where the network is used to provide the receiver (generally in a cell phone) with additional information to make acquisition and tracking easier. Aided GPS refers to using the network to give almanac and ephemeris information to the receiver. With the almanac, a receiver can cold start more quickly and know what Doppler frequencies to search, having the ephemeris data allows to receiver to reconstruct a local replica of the navigation data bits and thus allows for unlimited integration (limited only by the quality of the receiver clock).

Assisted GPS goes one step further, and in this implementation the receiver is also provided with time and/or frequency synchronization to GPS time through the network. This has the effect of reducing the frequency error limitation on long integration times. Assisted GPS networks may also provide the receiver with initial position estimates (for example of the nearest cell tower) to speed up initial acquisition.

With A-GPS, integration is generally possible up to as much as a whole second. With HS-GPS, 200 to 300 ms is a typical integration times (compare this with 5 to 10 for regular GPS up to a maximum of 20).

ENGO 585 Notes: Chapter 3 Part C.

(Note the subsection numbering for Chapter 4 differs from the course outline. I am trying to improve this section of the course by presenting the topics in a different order with more emphasis on certain topics as suggested by last year's class)

iii. GPS Pseudolites

GPS Pseudolites are ground based transmitters that send a GPS formatted RF signal that can be received and used by GPS receivers. They differ from other types of ranging radios in that they specifically mimic the signal structure of a GPS satellite, even if this signal structure is not ideal for a ground based system.

The first GPS experiments were conducted using GPS pseudolites. Prototype GPS signal transmitters and receivers were constructed and tested on the ground to figure out if the signal structure worked before having to pay for the expensive launch costs. A similar system is currently under development in a valley in the Bavarian Alps for Galileo equipment testing.

GPS pseudolites became commercially available in the late 1990's and there was a brief flurry of interest in developing pseudolite augmentation systems (for augmenting GPS) and for developing autonomous pseudolite positioning systems.

Unfortunately there are several limitations to using pseudolites for wireless location:

a) Near-Far

The 32 GPS L1 PRN codes were chosen from a much larger set of 1023 chip long PRN codes to be the 32 that had the lowest cross-correlation values. The worst-case peak cross correlation between them is approximately -20 dB. Recall the free space loss is a $1/r^2$ process, which means that you observe 20 dB of free space loss when you move 10 times further way from the source. What this means is that if any user is 10 times closer to one transmitter than they are to another (and the two transmitters are transmitting at the same power level) then the near transmitter will be 20 dB "louder" than the far one, making the near transmitter jam attempts to track the far transmitters PRN code.

In practice the results of near jamming can be seen when the far receiver channel begins to appear to track the far transmitter with exactly the same Doppler frequency as the near channel. When GPS is being used in conjunction with pseudolites, the effect is more significant since as soon as the user enters the near radius with respect to the nominal signal strength of GPS satellite singles, all other signals will be lost. Because of this, a pseudolite can be used to deny GPS to a small radius area, however it is not a real military threat since the source of the jamming is easily tracked.

b) Multipath

Pseudolites are typically deployed in urban environments where multipath is severe. To add to this limitation, pseudolite multipath is static since the transmitters are generally stationary. (Unlike GPS where multipath slowly changes, even for static users since the satellite-reflector-receiver geometry slowly changes). This could be used as an advantage, if the service provider were willing to create a multipath map of the pseudolite service area and provide this data to the users. However to be effective, it would have to be transmitted back to the users in some sort of differential correction message that depends on the user location. (Of course if the user knows his or her location, they don't need a pseudolite positioning system to determine their location then).

c) Ephemeris

It is difficult to transmit the coordinates of a pseudolite to a GPS receiver since the GPS interface control document defines the satellite coordinates in terms of a set of modified Keplerian elements. There is no way to model a stationary transmitter antenna on the surface of the earth as a set of Keplerian elements. It is possible to have Keplerian elements that correspond to a particular position near the surface at a particular time, but that set of elements inevitably represents a very high speed orbit (so the pseudolite will move very quickly away from that particular location if it is modeled as a ephemeris record. It is possible to get a pseudolite to transmit an ephemeris message of a geostationary satellite, this at least will allow the pseudolite to be tracked, however all ranges measured will be biased by the difference in distance between the actual location of the pseudolite and the modeled location at geostationary orbit.

d) Receiver compatibility issues

Pseudolites very rarely can be used with off-the-shelf (OTS) GPS receivers. Not only will an OTS receiver not known how to decode the pseudolite ephemeris data (if there is any), but it may also attempt to exclude pseudolite measurements as a safety feature. NovAtel OEM3 receiver, for example, as a default behavior will not log pseudolite pseudoranges without first giving the receiver an undocumented command to accept the pseudolite's signal and the fact that it is using a PRN code beyond the range of 1 to 32. Also, since receivers use the GPS almanac to determine PRN code search behavior, the use will have to manually instruct the receiver to look for a particular pseudolite.

e) Time-transfer and synchronization

Finally there is the issue of timing. To use a pseudolite to augment a single point GPS user (eg a cell-phone or a Garmin handheld), the pseudolite must be time synchronized to GPS time or at least be able to observe the difference and transmit this value to the user as a "satellite clock" correction. Otherwise, any pseudolite derived pseudorange will be biased by the pseudolite-to-GPS clock offset. It is also difficult for receivers to choose the right Doppler frequency to search for the pseudolite since pseudolite Doppler is entirely due to transmitter clock drift (as opposed to satellite Doppler which is mainly due

to satellite motion and only somewhat due to satellite clock drift). A simple approach is to co-locate a GPS receiver with the pseudolite to provide access to GPS time, however, this is only effective if the timing GPS receiver is not jammed by the transmitting pseudolite. Effectively the only way to do this is to have a GPS reference receiver nearby, but not too near, monitoring GPS time and pseudolite time and feeding the correction back to the pseudolite for transmission to the users. An alternative method is to have the reference receiver provide differential corrections (DGPS) including differential corrections for the pseudolite. This method is very effective provided that the user receiver does not get too confused by the possibly very small or very large pseudolite pseudorange.

More recent development in the area of pseudolites has focused on new custom signal structures and frequencies more suitable for ground based and urban use. Locata, and Australian company, currently markets a complete solution including pseudolites (transmitters) and receivers that uses a proprietary signal structure. However, most test results presented by the company are for very ideal conditions, for example a vehicle traveling along a straight flat road with 3 or 4 equidistant transmitters (thus avoiding both near-far and multipath). Other pseudolite research have moved directly into new and complex signal structures (Worcester Polytechnic for example) that resemble UWB signals more than wideband CDMA ranging signals such as GPS

ENGO 585 Notes for Chapter 3 Part C

v. Ultra-wideband (UWB)

Ultra-wideband is an upcoming wireless technology designed for short range high data rate wireless communication. The technology is still under intensive research and development and is not a mature technology. In general a radio system is considered UWB if its fractional bandwidth is greater than 20% or it occupies a bandwidth greater than 500 MHz. In the FCCs literature, the band limits are defined as points 10 dB lower than the peak power.

A theoretical UWB system would cover from DC to several GHz, however there are serious concerns about interference of UWB systems with existing narrowband and spread-spectrum (wideband) systems such as radios, GPS, etc.

The US FCC has set a UWB emission mask which limits intentional UWB transmission to between 3.1 and 10.6 GHz with very tight limits at lower frequencies (in the GPS bands particularly).

Most work on UWB relates to how the system will be used for communication. It is seen as the replacement for Bluetooth and is hoped that it will someday be used for very short range very high data rate applications. For example, replacing the cabling connecting a home theatre system. However, because of the incredibly large bandwidth, UWB systems also have an amazing potential for very high resolution range measurements.

The limit on time resolution of an RF system is given by the Cramer-Rao lower bound:

$$\sigma_\tau \geq \frac{1}{2\pi\beta\sqrt{2 \cdot SNR}}$$

where σ_τ is the standard deviation of the time measurement and β is the bandwidth of the signal. Of course multipath is still a problem but in high multipath environments, a UWB system could be used to resolve the first signal arrival and thus effectively solve the multipath problem (since software could be used to select and reject the multipaths). This ability to resolve individual multipath components is also seen as an advantage from a communications systems perspective since it then become possible to combine the power from each multipath to ensure a more reliable signal.

UWB systems can operate using a principle known as Impulse Radio. In theory this can be described as radio transmission without a carrier frequency by simply sending very short duration pulses (0.5 to 1 ns). These pulses, because of their very short duration, contain a very wide range of frequencies intrinsically without the need to have a separate carrier being modulated with a high data rate signal. The lack of a carrier also suggests that it could be possible to define an UWB radio using only digital components as opposed to carrier radios that require at least an analog RF front end.

There are two main impulse radio modulation schemes that have been demonstrated. The first, Pulse position modulation (PPM) modulates data onto an UWB signal by changing the time position of a periodically repeating pulse. This method is used by the radios used in Lab 5. The second method, Pulse Amplitude Modulation (PAM) changes the power level of the pulse depending on the data stream. Both modulation schemes can be multiplexed by time-hopping (applying a pseudorandom sequence to the pulse positions, known by both the transmitter and the receiver, so that the receiver will be looking when the pulse should be present (or not present depending modulation scheme and the data being sent).

UWB impulse radios can typically resolve all multipath components with differential delays in excess of 30 cm. (Lottici, D'Andrea, and Mengalli, 2002).

There is also a competing technology called Direct Sequence UWB, which is more like conventional wide-band systems where a fast PRN sequence is applied to a carrier, spreading the carrier power over as much of the UWB allocated spectrum as possible.

A final UWB technology is called OFDM (orthogonal frequency division multiplexing). In this technique, several subcarriers are present on either size of the main carrier frequency. The details of how this type of signal is created and demodulated are far beyond the scope of this course. The important aspect to remember is that this method is very effective at fully utilizing the spectrum allocated by the FCC (since the amplitude of each subcarrier can be controlled to maximize power in the allowed band.

Power issues:

Range Multispectral claims 20 metres through two concrete block walls 100 metres through 14 office walls (ie. drywall walls). Time Domain does not make any claims about operational range and accuracy. You will discover this in Lab 4.

Applications: Military, firefighting, possibly asset management, but expensive and you are going to need a dense network of equipment.

Notes Chapter 3 Part C: RFID and WiFi

vi) WiFi Positioning

WiFi (IEEE 802.11a,b,g) can be used for wireless location techniques. If you have a propagation model, and knowledge about the antennas involved (gain patterns etc), received signal strength can be converted directly into range. For example free space loss is given by: path gain = 20 log (lambda/4π distance)

More advanced models have been developed for various environments (and terrain types).

A more useful technique is called RSS Finger Printing (or database comparison). To do this you must first create a database by moving the user through the network to regularly spaced points and making observations of all the WiFi access points that can be heard at that location. You might have to observe various orientations of the user antenna at each survey point to cancel out gain pattern effects. When you do a real observation, you have to find the nearest neighbours in the database. (By nearest we mean the points with the most similar power levels). If the user is going to be using an identical device to the one used for the survey, it is just a matter of looking up values in the database and choosing some cost function. Once you have found the "closest" neighbour, you could either just decide this is your location, or you could find the nearest 3 or 4 survey points and then interpolate between them.

If the user is using an arbitrary WiFi device as opposed to an identical one, it is most effective to normalize the power levels with respect to one of the access points. This makes the database easier to compare to the user observations. Remember that normalizing (dividing all observations by one observation) when applied to power measurements in decibels, is a subtraction). In a way this is analogous to between satellite single differencing to cancel receiver clock (here we are doing between access point single differencing to cancel receiver power level offset)

An example of this method was done in class

vii) RFID

Radio frequency identification is not really a location technology, but it can be used to provide location information. RFID systems consist readers and "tags" that are either active or passive and communication via RF with each other, mainly for identification purposes (hence the name RFID). Active tags have internal batteries, passive tags do not. RFID location can be implemented using time of flight (TOF or TOA), RSSI, AOA, and proximity methods).

RFID systems have mainly been developed for inventory control, but some specific location based applications have been developed. One is management of toll roads. (A device for doing this was patented in 1973.)

The simplest device is the tag commonly found in clothing. It is basically an antenna with a small passive electronic circuit that has a particular impulse response. It is deactivated by destroying the circuitry with a very strong RF pulse. Then when you walk through the gate (the reader), the gate doesn't detect the impulse response its looking for. This system is the most common on North America and is called the "swept RF system" because it scans the tag at multiple frequencies and measures it power response in the frequency domain. In Europe, the "EM system" is more common. In this case, a magnet is inserted in the article, just like in a library book. If the magnet is properly magnetized, it appears invisible to the gate, however if it is demagnetized, it disturbs the EM signal being transmitted by the gate and causes an alarm. Finally there is the system that is used on CDs for example. In this system, a circuit is created with particular resonant frequency. The gate transmits a pulse that causes the tag to emit a tone, unless it has been deactivated by demagnetization.

These three methods aren't really RFID, because they only provide the user with a single bit of information (deactivated or not) True RFID tags are similar but the particular impulse response of the tag carriers more information than just true/false. These tags are being included in more and more consumer products for anti-theft purposes and for inventory control, with the hope that one day they will replace the common UPC barcode, however some companies are having implementation difficulties, especially in terms of reader collisions, where a user attempts to read an entire pallet of inventory for example and can't handle the large number of responses.

RFID tags can also be put into capsules and implanted with a large gauge hypodermic needle. An example of this: Implanted VIP passes for some Spanish and Dutch nightclubs. After you get in, the waitresses and bartenders can just scan your forearm to bill you for drinks. RFID wristbands have also been developed for hospitals to identify patients

RFID Purse. A purse with an imbedded RFID reader, scans items as you insert them. If an item is missing, the purse tells you.

RFID security cards (like your UCID card) operate on 130 khz and can only be read at short ranges. Higher frequencies get absorbed by water and reflected by thin layers of metal, but are required for high data rate operations. This has been a problem for Walmart for example, who would like to have RFID checkouts, but this is limited by people having cans of pop for example in their shopping carts.

There is also a security problem with RFID since the tag often remains on the item after you have purchased it. This can cause embarrassment when you set of alarms in stores but also could potentially allow thieves to scan the contents of your house before deciding to break int.

How to use RFID for location:

Proximity:

Fixed readers, tags on people or objects. Readers located at doors for example. The opposite is also possible, with fixed tags and a mobile reader. The tags would then have their geographic coordinates installed on them along with other information about the location.

Passive tags may have a “range of 50 cm”. So if you detect a tag, you read a tag, then you must be 50 cm from the coordinates of the reader.

AOA/TOA

“RFID radar” A two antenna system (reader) querries tags and records the two way time of flight and determine the angle. The tags also allow the system to identify the user. This is different than normal radar since normal radar only detects targets, but it is effectively identical to “secondary surveillance radar”, the descendant of “Identification Friend or Foe” discussed at the beginning of this chapter.

ENGO 585 Notes: Chapter 4

a. Inertial Measurements and Models

Inertial navigation involves the use of accelerometers and gyroscopes to make observations of acceleration and rotation and to integrate these observation to estimate a trajectory. In principal accelerometers are used to observe velocity and position (by integrating the output twice) and gyroscopes are used to keep track of the orientation of the coordinates system where the accelerations are being measured.

i) Gyroscopes and Inertial Sensors

A gyroscope is a device that can measure rotation. The simplest is the mechanical gyro, which is simply a spinning mass that has a tendency to maintain the direction of its axis of rotation. If a gyro is fully gimbaled, the rotation axis will not change as the vehicle rotates around it. If this is only partly gimbaled, the gyro will react to external torque by precessing. A rate gyro operates by observing the precession due to an external torque and then applying a compensating torque to drive the gyro back to a non-precessing state. This apply torque can be measured and its integral is proportional to rotation about one particular sensitive access.

The other major gyro technology is the ring laser gyro, which relies on the Sagnac principle. In consists of a ring of fiber optics where a laser is split and sent in both directions around the ring. Because special relativity only applies in inertial frames, any rotation of the ring will cause the optical path in one direction to be shorter and longer in the other. The two lasers can then be recombined and the interference pattern between the two can be observed to determine the rotation during the travel time.

The simplest accelerometer is a mass hanging from a spring with the mass and spring constrained to a particular axis. An external force will cause a displacement of the spring which can be measured. More advanced implementations include the vibrating beam accelerometer where a mass is suspending between two wires that vibrate at resonant frequencies dependant on the tension in each wire. The change in frequencies can be measured used to determine the force. The MEMS (micro electro-mechanical systems) accelerometer works on the principle of having miniature proof masses suspended on a silicon chip with their movement measured electostatically.

Accelerometers are subject to many errors. The most important being bias and linear and non-linear scale factor error. Both of which may change with time and temperature. There is also non-orthogonality error when accelerometer triads are being used.

The observation equation for a single accelerometer output is:

$$l = f + b + (S_1 + S_2 f) f + Nf + \gamma + \delta g + n$$

where S_1 and S_2 are the scale factors, N is the non-orthogonality matrix, γ is normal gravity, δg is uncompensated gravity and n is noise.

Note that accelerometers measure specific force that included gravity. So when an accelerometer is at rest, it will observe a gravity signal depending on its orientation with respect to the gravity vector. This needs to be corrected if the system is to be used for navigation.

Inertial Navigation Systems:

Three accelerometers and three gyros can be combined to form an inertial measurement unit. When these are combined with some kind of processing equipment, the combined system in an inertial measurement system (INS).

Classification of INS depends on the author. Terms that are used are Low-cost, Tactical and Navigation grade, or low, medium and high performance. Definitions vary but typical performances are 10,1, and 0.1 nautical miles per hour.

These errors can be limited by error state modeling, and by updates from other sensors or known states such as the zero velocity update and the coordinate update.

There are some important points to consider when using inertial sensors. The most important thing to remember is that the errors on inertial sensors are unbounded, meaning they can, and depending on the cost of the sensor will, grow with time if you don't do something about it. Take for example an accelerometer, position is obtained by integrating the output twice, so any bias in the accelerometer output will be amplified by the integration process. Usually Kalman filters are used to estimate the error states, and these Kalman filters are generally augmented by position updates from GPS or another source. The level of coupling with GPS may vary and different names are applied to the various levels. They are generally loose, tight, and ultra-tight/deep. These will be briefly explained in class. In the absence of a source of position updates, it is still possible to control INS errors with a zero-velocity update, and this has for example been implemented in backpack mounted systems for surveying in thick forest conditions and in inertial sensor based pedometers among other applications.

The details of inertial navigation, mechanization equations, and error modeling are far beyond the scope of this course.

b. Other Sensors

i) Barometers and Magnetometers

Barometers (altimeters) and Magnetometers are interesting augmentation sensors because though they are self-contained, the measure an external quantity and in each case the external quantity is relative.

Barometers measure atmospheric pressure, which can easily be converted by a very simple model into a height. They are useful in wireless location since they can be used to provide an independent measurement of the z-component of position, for example, a

barometer could be integrated into a GPS enabled cell phone and used to determine what floor the user is on in a tall building (the location of the building being modeled as the last outdoor position recorded by the GPS chip. However, the conversion between pressure and height depends the prevailing atmospheric pressure. Just because the pressure increases, doesn't necessarily mean your altitude has decreased since you could be stationary and the weather may have changed.

$$P(h) = P_0 e^{-\frac{gh\rho_0}{P_0}}$$

where P_0 is the pressure at sea level (101.3 kPa) and ρ_0 is the density at sea level (1.225 kg m^{-3}).

There are two methods of dealing with the relative nature of barometer observations. The simplest is to calibrate the barometer at a known height, for example at the trailhead when using a barometer to measure elevation change while hiking. (the "sea level" values can be replaced with any values and the relationship still holds) This can be further refined by integrating barometer measurements with other location information in a Kalman filter. In addition to states for x, y, z (or lat, lon, height) and additional state for barometer offset could be modeled. Then when there are sufficient observations to determine position (from GPS for example), the barometer observation can be used exclusively to estimate the barometer offset. The barometer offset can be given a small process noise so that later on when the GPS observations are absent, future barometer observations contribute mainly to estimating altitude and then when external observations are again available it will again be possible to update the barometer offset state.

The second approach is to use differenced barometer observations to obtain changes in altitude. This approach is more common, though as we have seen earlier in this course, the two are really equivalent.

Magnetometers, or digital compasses, are another method of obtaining orientation. They are often integrated into low cost GPS receivers to provide an orientation when the user is stationary. To be useful, they must be installed in a device that does not cause magnetic interference and they must be used away from strong magnetic fields. Like gyros, they only work when some assumptions are made about the orientation of the sensor. They have to be horizontal (ie. the sensitive rotation axis should be perpendicular to the magnetic field lines that you want to measure). As a result, errors are introduced when the sensor is tilted and this needs to be corrected with other sensors. Also there is the issue of magnetic declination, which either has to be modeled, or calibrated out. One method of calibration would be to compare the output against the GPS derived velocity vector when the user is mobile and then use this to estimate the local declination. The device could also come with a model to apply this correction as a function of GPS derived position.

ii) Odometers

Odometers (wheel rotation counters) can be used to provide either a range traveled observations or velocity constraint to a navigation solution. For a range observation, this is typically accomplished by observing the change in odometer reading from one epoch to the next. This “observation” is then a range from the last estimated position to the current one. Note that to be any use, such an observation either has to be integrated into the solution in a Kalman filter, or it must be used in conjunction with other observations (either other ranges from external RF sources) or heading information from other sensors.

The odometer rate is a measure of the norm of the vehicle velocity which gives an observations equation of the form

$$|v| = \sqrt{v_x^2 + v_y^2 + v_z^2}$$

which, in a Kalman filter, can be linearized about the current velocity vector. Interestingly this gives a design matrix row of the same form as a range, but this time with entries in the velocity columns as opposed to in position columns.

There is one other way to use odometry, which is called differential odometry. In this application, two odometers are used on two wheels and the change of direction can be obtained by comparing the difference between the two.

In all cases, odometers are affected by scale errors (due mainly to incorrecty measured wheel sizes), and the effects of wheels slipping. There is also a quantization error since some odometers output is simply a pulse for every 3 to 20 cm traveled. Other types of odometers allow for continous output that can tracked with a phase-locked loop, in which case the output is theoretically good to a few percent of the wheel circumference.

iii) Map Aiding and Map Matching

Map aiding consists of using a map edge as an additional line of position. This is usually accomplished by introducing pseudo-observations. The simplest conceptually is to use a map to provide a height pseudo-observation equation.

$$h_{map} = h$$

The design matrix for this observation is simple (a one in the height columm). Using a map to provide a constraint is a little more difficult and usually involves constraining the heading of the user to the direction of the line segment (ie. a piece of road) that the user is traveling on. The observation equation for this type of constraint depends on what states are being estimated. In the case of positions only

$$\text{heading}_{map} = \arctan \frac{y_k - y_{k-1}}{x_k - x_{k-1}}$$

If velocity is being estimated, then the above equation can be expressed in terms of velocities, an obviously if heading is an explicit state then the constraint can be applied directly to this state.

As with any other pseudo-observation, the trick is to determine the appropriate variance to be associated with the constraint. It should be small enough that the constraint information is used, but not so small that the constraint overpowers all other observations and prior estimates.

When map aiding is applied with a small or no variance, it is called map matching. In this case it is not usually done in a Kalman filter, but after the fact the filter estimated position is snapped to the nearest map element. For example, a GPS receiver on a vehicle in a downtown area may have a position solution in the middle of a block and for the output position to be on the nearest street. (This approach may be helpful for navigation, but not necessarily for located parked or stolen vehicles).