EECS 16A Designing Information Devices and Systems I
Spring 2018 Lecture Notes Note 23

Introduction

In our previous exploration of linear algebra, we used techniques like Row Reduction and Matrix Inversion to solve explicit systems of equations in multiple variables. However, what happens when we introduce noise or imprecision into the system? How do we solve our system of equations in such a way that we minimize error, and get as precise a solution as we possibly can?

In this section we will explore **Least Squares**, a method through which we can minimize error by solving overdetermined systems of equations (more equations than variables). We will also learn how to find the **Minimum Norm Solution** of an underdetermined system of equations (fewer equations than variables).

Key Concepts

By the end of this note, you should be able to do the following:

1. Understand why we use least squares, especially in the context of our GPS Locationing systems.
2. Explain how the Least Squares method minimizes the error term of a system of equations.

Least Squares

Let's use our GPS system, to develop intuition as to why we need a system like Least Squares: Given some number of beacons and the (possibly erroneous) readings of distances between a user and the 3 beacons, how is it possible to determine which distance reading is correct? It turns out that if we add more beacons to get more data, we can make more robust measurements and use the least squares approach to find the answer.

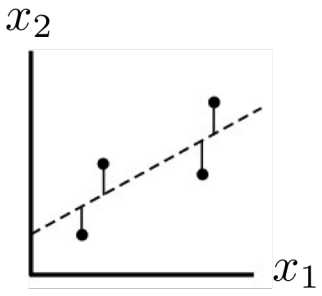
Another example: Suppose you have n points in $\mathbb{R}^2$, $(a_1, b_1), \cdots, (a_n, b_n)$ and want to fit a straight line to these points. If all these points lie on the line, then they will all satisfy the equation $b_i = x_1 a_i + x_2$ for some unknown x_1 and x_2. Written in matrix form:

$$\begin{bmatrix} a_1 & 1 \\ a_2 & 1 \\ \vdots & \vdots \\ a_n & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

However not all points may lie on the same straight line due to measurement error or model mismatch. In this case we have n equations and 2 unknowns. If $n > 2$, the system of equations may have no solutions. A general principle to apply when there's more equations than unknowns is to introduce more unknowns so that the equations can be solved. In this case, we introduce an unknown error term, e_i, for each equation, so that the equations become:

$$b_i + e_i = x_1 a_i + x_2$$

This corresponds to the vertical distance between the best-fit line and points in the following diagram:



We then try to find the line that minimizes the squared vertical distance between each point and the line.

Derivation

Let's think about how we would set up a system in which we have more equations (n) than unknowns (m).

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m = b_1 + e_1 \tag{1}$$
$$a_{21}x_1 + ... + ... + a_{2m}x_m = b_2 + e_2 \tag{2}$$
$$a_{n1}x_1 + ... + ... + a_{nm}x_m = b_n + e_n \tag{3}$$

With the above equations, we are trying to approximate a solution to the system with a range for the error. Essentially, we know b (in the above example it would be the distance readings), but we already know that they are a little bit off, so we capture the error with the e term.

We can put the above equations into a matrix and attempt to generate an $\vec{x}$ such that $\|\vec{e}\|$ is as small as possible. In this case, we know that the b's are noisy, so we try to separate the b's out from the noise:

$$\vec{e} = A\vec{x} - \vec{b}. \tag{4}$$

Hence, we would want to choose an $\vec{x}$ that minimizes $||\vec{e}||^2$, which is equivalent to solving the following optimization problem:

$$\min_{\vec{x}} \|\vec{e}\|^2 = \min_{\vec{x}} \left(\left\| A\vec{x} - \vec{b} \right\|^2 \right). \tag{5}$$

Here, the notation $\min_{\vec{x}} \|e\|^2$ means to find the $\vec{x}$ that gives the minimum $\|e\|^2$. To illustrate, let's use a simple example where $\vec{x}$ is one-dimensional and $n = 2$. In this case, let's use a slight abuse of notation and treat x as a scalar.

$$a_{11}x = b_1 + e_1 \tag{6}$$
$$a_{21}x = b_2 + e_2 \tag{7}$$

Hence, we have the following optimization problem:

$$\min_x [(a_{11}x - b_1)^2 + (a_{21}x - b_2)^2]. \tag{8}$$

To calculate the minimum of a function we use the idea of finding critical points, i.e. points where the slope of the function is zero. For this, we take the derivative of a function and set it equal to zero to find the critical point. Then, we can check whether the point corresponds to a maximum or minimum by using the second derivative. A positive second derivative means that it is a minimum.

Defining $f(x)$ to be the expression we are trying to minimize, first we calculate the derivative.

$$\begin{aligned} f(x) &= (a_{11}x - b_1)^2 + (a_{21}x - b_2)^2 \\ \frac{df}{dx} &= 2a_{11}(a_{11}x - b_1) + 2a_{21}(a_{21}x - b_2) \end{aligned}$$

Then we set the derivative equal to zero and find the critical point:

$$\begin{aligned} \frac{df}{dx} &= 0 \\ 2a_{11}(a_{11}x - b_1) + 2a_{21}(a_{21}x - b_2) &= 0 \\ 2a_{11}^2 x + 2a_{21}^2 x &= 2a_{11}b_1 + 2a_{21}b_2 \\ x &= \frac{b_1 a_{11} + b_2 a_{21}}{(a_{11})^2 + (a_{21})^2} \end{aligned}$$

Since the second derivate of $f(x)$ is always positive, the following is the minimum of x in the 2D case:

$$x = \frac{b_1 a_{11} + b_2 a_{21}}{(a_{11})^2 + (a_{21})^2} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|^2}$$

Note that in 2D, we know that the orthogonal projection is the shortest distance from a point onto a line. This principle generalizes to higher dimensions. (In 3D the shortest distance from a point to a plane is the orthogonal projection of the point onto the plane.)

Based on this strategy, we can try to generalize this to n dimensions.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1} & \dots & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix} + \vec{e}$$

In order to achieve the same result as with 2 dimensions, we need $\vec{e}$ to be orthogonal to all the columns in the matrix (in order to minimize e).

$$\begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} = 0$$

This can essentially be simplified to:

$$(A^T)(\vec{e}) = 0$$

Since $\vec{e}$ is just $A\vec{x} - \vec{b}$ we have the following:

$$A^T(A\vec{x} - \vec{b}) = 0$$

$$A^T A\vec{x} - A^T\vec{b} = 0$$

$$A^T A\vec{x} = A^T\vec{b}$$

$$\vec{x} = (A^T A)^{-1} A^T \vec{b} \tag{9}$$

We have just derived the least squares algorithm for the general case! The $\vec{x}$ given by Eq. (9) is the optimal solution (in the least squares sense).

Application of Least Squares

It turns out Gauss used this technique to predict where certain planets would be in their orbit. A scientist named Piazzi made 19 observations over the period of a month in regards to the orbit of Ceres (can be viewed as equations). Gauss used some of these observations. He also knew the general shape of the orbit of planets due to Kepler's laws of planetary motion. Gauss set up equations like so:

$$\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = \phi$$

If one divides the whole equation by ϕ, nothing significant happens so we can ignore the denominator and treat the right side of the equation as 1.

$$\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = 1$$

We can set up a matrix like so:

$$\begin{bmatrix} x_1^2 & x_1 y_1 & y_1^2 & x_1 & y_1 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_n^2 & \dots & \dots & \dots & y_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Using this matrix, we can use the least squares formula to figure out the estimated value of $\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix}$ using the least square formula derived earlier:

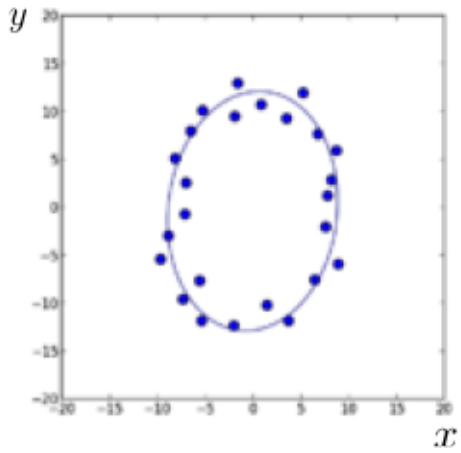
$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix} = (A^T A)^{-1} A^T \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \tag{10}$$

where

$$A = \begin{bmatrix} x_1^2 & x_1 y_1 & y_1^2 & x_1 & y_1 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_n^2 & \dots & \dots & \dots & y_n \end{bmatrix} \tag{11}$$

.

In doing so, a possible result could be the following (the curve represents the function $\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = 1$):



As we can see, the least squares method is useful for fitting noisy or approximate measurements to a curve, provided that we know the general shape of the curve!

Solving underdetermined sets of equations

So far, we've learned how to solve overdetermined systems of equations (where there are more equations than variables) using least-squares. In this lecture, we solve under-determined system of equations using a slightly different approach.

Suppose $A\vec{x} = b$, where A is an $n \times m$ matrix, b is an unknown n-vector, and x is an unknown m-vector. Assume $n < m$ — there are fewer constraints than unknowns.

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} = \begin{bmatrix} \\ b \\ \\ \end{bmatrix}$$

We also assume that A has full row rank, i.e., rank(A) = n. (Rank is the number of linearly independent columns.)

Example 1: How do you represent the line $x_1 + x_2 = 1$ as $A\vec{x} = b$?

We can formulate this as an underdetermined set of equations.

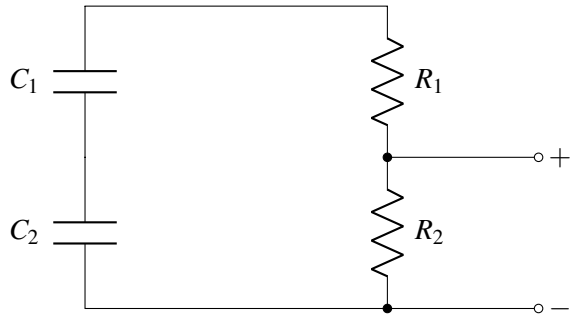
$$A\vec{x} = b$$
$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}$$

In general, in this case, $A\vec{x} = b$ has an infinite number of solutions. We can pick one of these solutions by finding the one with the minimum norm.

$$\min_{\vec{x}} \|\vec{x}\|^2 \text{ such that } A\vec{x} = b$$

In our particular example, the point is $(\frac{1}{2}, \frac{1}{2})$.

Example 2: Suppose C_1, C_2, R_1, R_2 are the components given to you, and they are organized in the setting below. You would like to have voltage b across resistor R_2 in the figure. We would like to charge the capacitors to voltages x_1 and x_2 to get b. Find the capacitor voltages x_1 and x_2 to minimize $\|x\|^2$. b **is the voltage between the 2 unconnected terminals to the right of the circuit.**



Solution: Notice here that what matters to determine b is just the total volatage $x_1 + x_2$. So we only really have one constraint, but again we have two variables that we can choose. We know from the equation for a voltage divider that if the voltages on C_1 and C_2 are x_1 and x_2 then

$$b = (x_1 + x_2)\frac{R_2}{R_1 + R_2}$$

Now, we can set:

$$A = \begin{bmatrix} \frac{R_2}{R_1+R_2} \\ \frac{R_2}{R_1+R_2} \end{bmatrix}.$$

Then, we have

$$\begin{bmatrix} \frac{R_2}{R_1+R_2} \\ \frac{R_2}{R_1+R_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b = (x_1 + x_2)\frac{R_2}{R_1 + R_2}$$

in the desired format $A\vec{x} = b$.

We solve this constrained optimization problem using the method of Lagrange multipliers, where we add a term to the quantity to be minimized. The vector $\vec{\lambda}$ is the vector of the Lagrange multipliers.

$$\min_{\vec{x},\vec{\lambda}} \|\vec{x}\|^2 + \vec{\lambda}^T(b - A\vec{x}) \tag{12}$$

Differentiating with respect to $\vec{x}$ and setting the result to 0 gives

$$\begin{aligned}\frac{\partial}{\partial \vec{x}}(\vec{x}^T\vec{x} + \vec{\lambda}^T(b - A\vec{x})) &= 0\\ 2\vec{x}^T - \vec{\lambda}^T A &= 0\\ 2\vec{x} - A^T\vec{\lambda} &= 0\end{aligned}$$

Left-multiplying by A:

$$\begin{aligned}2A\vec{x} - AA^T\vec{\lambda} &= 0\\ \therefore \vec{\lambda} &= (AA^T)^{-1}2A\vec{x}\end{aligned}$$

Differentiating (12) with respect to $\vec{\lambda}$ and setting the result to zero:

$$\begin{aligned}A\vec{x} &= b\\ \vec{\lambda} &= (AA^T)^{-1}2b\end{aligned}$$

Since $2\vec{x} - A^T\vec{\lambda} = 0$,

$$\vec{x} = A^T(AA^T)^{-1}b$$

This is the Minimum Norm Solution to $A\vec{x} = b$.