4.1 Least Squares

Background

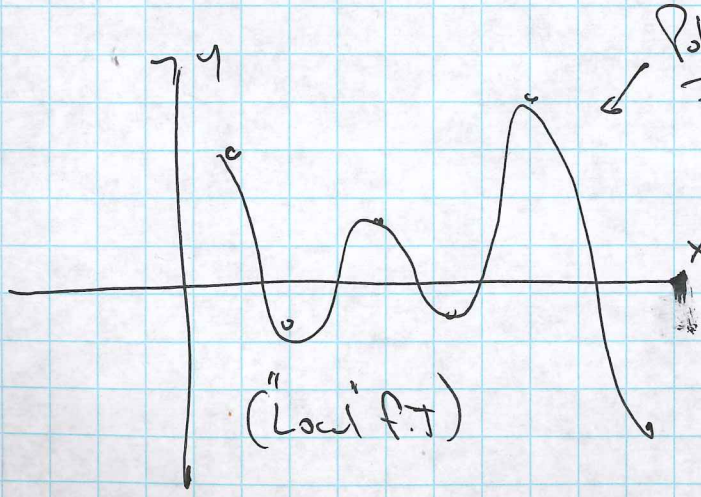
Recall that in Chapter 2 we developed a repertoire of numerical techniques for solving consistent systems of the form: $A\vec{x} = \vec{b}$.

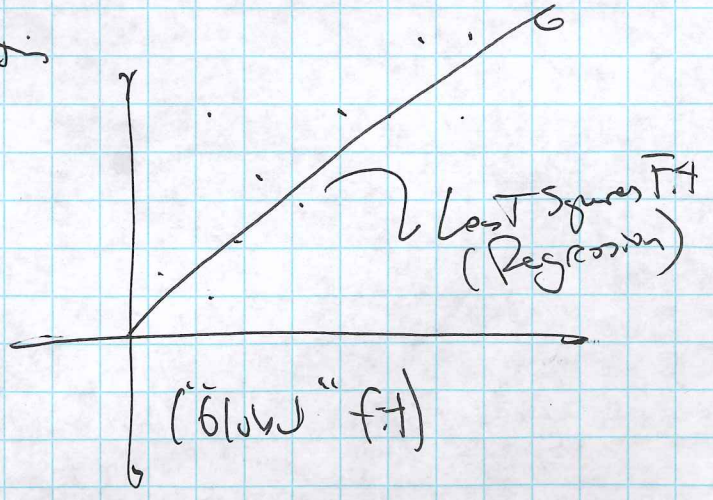
In Chapter 3 we studied techniques of polynomial & Trig function interpolation.

Note that interpolation, in general, provides a model that "locally" fits a data set - however, polynomial interpolation is oftentimes subject to the Runge Phenomenon & "overfitting" issues in general.

With Least Squares we investigate an alternative (of sorts) to function interpolation by building a model that "globally" fits a data set by minimizing "residual error" (e.g. least squares).

In particular, we investigate in detail systems of the form: $A\vec{x} = \vec{b}$ that are inconsistent ($n > p$), where the # of equations (i.e. data points) exceeds the number of model parameters (p).





Ex. Consider the system:

$$\begin{aligned} x_1 + x_2 &= 2 \\ x_1 - x_2 &= 1 \\ x_1 + x_2 &= 3 \end{aligned}$$

$\rightarrow A\vec{x} = \vec{b} \rightarrow \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & -1 & 1 \\ 1 & 1 & 3 \end{array}\right]$. Note that this system is **inconsistent!**

i.e. it has no exact solution. Why? Most obviously, $x_1 + x_2 = 2$ & $x_1 + x_2 = 3$ cannot both hold. Furthermore, using the language of linear algebra, A is singular (i.e. non-invertible) – i.e. the columnspace of A does not span $\mathbb{R}^3$.

Q: why study inconsistent systems? One Answer: Most problems in applied science involve data sets where the number of data points exceeds the number of model parameters used for the model ~~with~~ to with which we want to model the data (or "compress" the data, as we use my use)

So, now do we "solve" a system that is insoluble?

Answer: Find the "solution" $\bar{x} \in \text{Col}(A)$ so that $\|A\bar{x} - \vec{b}\|$ is (*) as small as possible!

Recall that the system above, $A\vec{x} = \vec{b}$ can be expressed in vector form: $\left[\begin{array}{cc} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{array}\right]\left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right] \rightarrow x_1\left[\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right] + x_2\left[\begin{array}{c} 1 \\ -1 \\ 1 \end{array}\right] = \left[\begin{array}{c} 2 \\ 1 \\ 3 \end{array}\right]$.

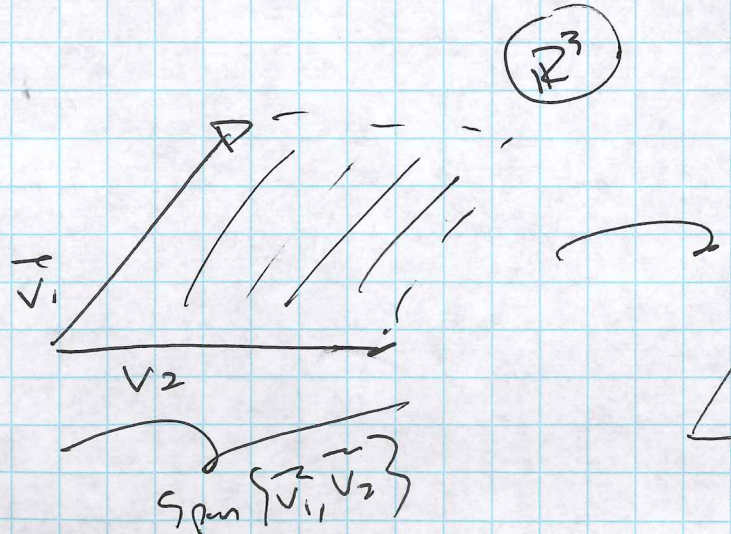
In fact, any $m \times n$ system: $A\vec{x} = \vec{b}$ can be viewed in the vector equation:

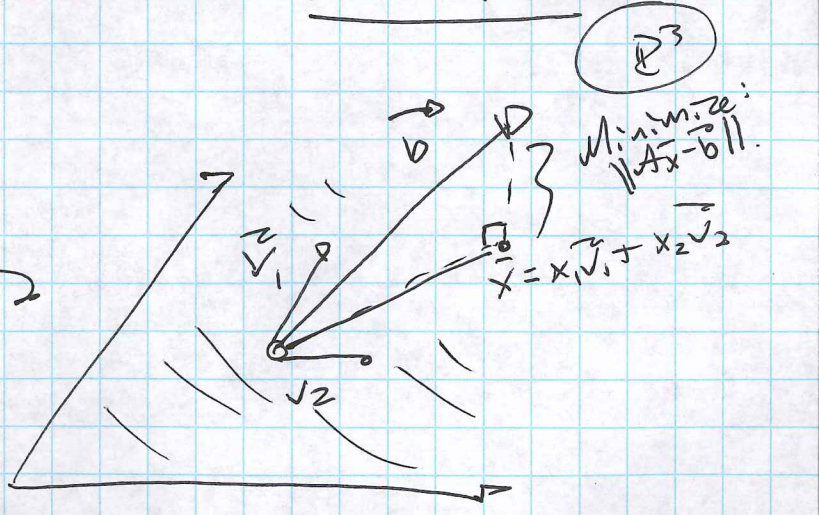
$$x_1\vec{v}_1 + x_2\vec{v}_2 + \cdots + x_n\vec{v}_n = \vec{b} \quad \left(\text{where } A = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} \right)$$

which expresses $\vec{b}$ as a linear combination of the column $\vec{v}_i$ of A, with coefficients: $x_1, \ldots, x_n$.

In our case, we are attempting to "hit" the target vector $\vec{b}$ as a linear combination of (2) (only) 3-dimensional vectors! Since a linear combination of 2 3-dim vectors forms a plane in $\mathbb{R}^3$, the vector equation above has a solution only if $\vec{b}$ lies in the aforementioned plane. Too many equations makes this problem "overspecified."

We proceed to find a pair: x_1, x_2 describing a vector in the plane spanned by $\{\vec{v}_1, \vec{v}_2\}$ that is closest to $\vec{b}$!





Recall that two vectors $\vec{u}, \vec{v}$ are said to be orthogonal if $\vec{u} \cdot \vec{v} = 0$ (Geometrically: $\vec{u} \perp \vec{v}$).

From the previous picture: the residual vector $(\vec{b} - A\vec{x}) \perp$ to the plane $\{A\vec{x} \mid \vec{x} \in \mathbb{R}^n\}$ (i.e. $Col(A)$).

$$\Downarrow$$

$$(\vec{b} - A\vec{x}) \perp \{A\vec{x} \mid \vec{x} \in \mathbb{R}^n\}$$

Note: A dot product $(\vec{u} \cdot \vec{v})$ can be expressed as matrix multiplication: $\vec{u}^T\vec{v} = [u_1 \cdots u_n]\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \vec{u} \cdot \vec{v}$.

We use this fact + the notion of orthogonality to write:

$$(\vec{b} - A\vec{x}) \perp \{A\vec{x} \mid \vec{x} \in \mathbb{R}^n\} \iff (A\vec{x})^T(\vec{b} - A\vec{x}) = 0 \quad \forall \vec{x} \in \mathbb{R}^n.$$

(Using that $(AB)^T = B^TA^T$) this yields:

$$\Downarrow$$

$$\vec{x}^TA^T(\vec{b} - A\vec{x}) = 0 \quad \forall \vec{x} \in \mathbb{R}^n$$

This means that $A^T(\vec{b} - A\vec{x}) \perp \vec{x}$ for any choice of $\vec{x} \in \mathbb{R}^n$.

The only way for this to hold is if: $A^T(\vec{b} - A\vec{x}) = 0$

Simplifying this equation produces the following:

$$A^TA\bar{x} = A^T\bar{b}$$ → This system is known as the "normal equations."

The solution to the normal equations, $\bar{x}$, is known as the least squares solution of the system: $A\bar{x} = \bar{b}$.

Normal Equations for Least Squares

(1) Given the inconsistent system: $A\bar{x} = \bar{b}$

solve: (2) $A^TA\bar{x} = A^T\bar{b}$ for the least squares solution $\bar{x}$, that minimizes the Euclidean length of the residual: $\bar{r} = \bar{b} - A\bar{x}$.

Ex. Use the normal eqns to find the least squares solutn of the inconsistent system.

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \qquad A^TA = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$A^T\bar{b} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix} \longrightarrow \text{Normal Eqns: } \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\longrightarrow \bar{x} = \langle \bar{x}_1, \bar{x}_2 \rangle = \langle 3/4, 3/4 \rangle$$

Substituting the least squares solution into the orginal problem yields:

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{7}{4} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2.5 \end{bmatrix} \neq \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

The residual: $\vec{r} = \vec{b} - A\vec{x} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2.5 \\ 1 \\ 2.5 \end{bmatrix} = \begin{bmatrix} -.5 \\ 0 \\ .5 \end{bmatrix}$

Note if $\vec{r} = \vec{0}$, we have found the solution $A\vec{x} = \vec{b}$ exactly.

Here we can measure our degree of success for finding a "best" approximate solution to the inconsistent system: $A\vec{x} = \vec{b}$ in at least (3) ways.

(1) Using Euclidean distance

$$\|\vec{r}\|_2 = \sqrt{r_1^2 + \cdots + r_m^2}$$

(2) The Squared Error

$$SE = r_1^2 + \cdots + r_m^2$$

(3) The Root Mean Squared Error

$$RMSE = \sqrt{SE/m} = \sqrt{\frac{r_1^2 + \cdots + r_m^2}{m}}$$

Note:

$$RMSE = \frac{\sqrt{SE}}{\sqrt{m}} = \frac{\|\vec{r}\|_2}{\sqrt{m}}$$

→

Thus, finding the $\bar{x}$ that minimizes each type of error measure above, minimizes all.

For mine, $SE = (.5)^2 + 0^2 + (-.5)^2 = \boxed{.5}$; $\|\bar{r}\|_2 = \sqrt{.5} \approx .707$;

$RMSE = \sqrt{\frac{.5}{3}} \approx .408$.

Ex. Solve the least squares problem.

$$\begin{bmatrix} 1 & -4 \\ 2 & 3 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -3 \\ 15 \\ 9 \end{bmatrix}$$

→ Normal Equations: $A^T A \bar{x} = A^T \bar{b}$ →

$$\begin{bmatrix} 9 & 6 \\ 6 & 29 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 45 \\ 75 \end{bmatrix}$$

→ $\bar{x_1} = 3.8$, $\bar{x_2} = 1.8$.

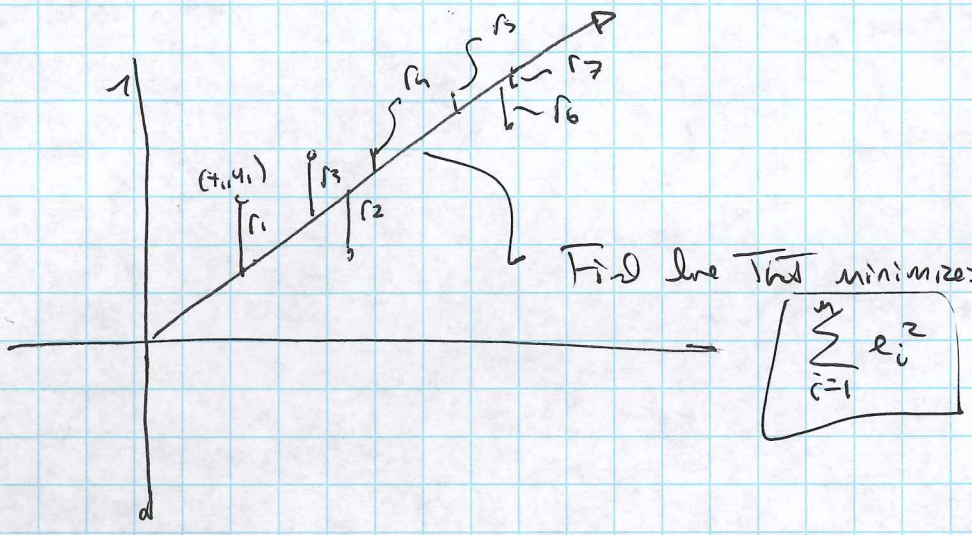
$$\bar{r} = \bar{b} - A\bar{x} = \begin{bmatrix} -3 \\ 15 \\ 9 \end{bmatrix} - \begin{bmatrix} 1 & -4 \\ 2 & 3 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 3.8 \\ 1.8 \end{bmatrix} = \begin{bmatrix} -3 \\ 15 \\ 9 \end{bmatrix} - \begin{bmatrix} -3.4 \\ 13 \\ 11.2 \end{bmatrix} = \begin{bmatrix} .4 \\ 2 \\ -2.2 \end{bmatrix}$$

$\|\bar{e}\|_2 = \sqrt{(.4)^2 + 2^2 + (-2.2)^2} = \boxed{3}$

Fitting Models To Data (Regression)

Let $\{(x_1,y_1)\ldots,(x_m,y_m)\}$ be a data set. Given a <u>fixed</u> "class of models":

$\hat{y} = c_1 + c_2 x$, find the model that "<u>best fits</u>" the data by minimizing the sum of residual errors squared.



Find the line that minimizes: $\sum_{i=1}^{m} e_i^2$

Ex. Given data: $\{(1,2),(-1,1),(1,3)\}$ Find the <u>best fit</u> linear model.

$\hat{y} = c_1 + c_2 x \longrightarrow$

$$\begin{cases} c_1 + c_2(1) = 2 \\ c_1 + c_2(-1) = 1 \\ c_1 + c_2(1) = 3 \end{cases}$$

$\downarrow$ Matrix Form

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \rightarrow$$ Has No pure solutions! (points aren't collinear)

From our previous result. $\bar{c}_1 = 7/4$, $\bar{c}_2 = 3/4 \rightarrow$ $\hat{y} = \frac{7}{4} + \frac{3}{4}x$

The RMSE is $\frac{1}{\sqrt{6}}$.

(3) Step Procedure for fitting data (the least squares)

Given m data points: $\{(T_i, y_i)\}_{i=1}^{m}$

(1) Choose a model, e.g. $\vec{y} = c_1 + c_2 T$,

or $\vec{y} = c_1 + c_2 T + c_3 T^2$, or $\vec{y} = e^{c_1 + c_2 T}$ ("logistic regression"), etc.

(2) Force the model to fit the data.

Substitute points into model. This gives system: $A\vec{x} = \vec{b}$ with

$A(m \times p)$

(3) Solve the normal equations: $\boxed{A^T A \vec{x} = A^T \vec{b}}$

Ex. Find both the best line & best parabola to fit:

$\{(-1,1), (0,0), (1,0), (2,-2)\}$.

(1) $\vec{y} = c_1 + c_2 T \rightarrow$

$$\left.\begin{array}{l} c_1 + c_2(-1) = 1 \\ c_1 + c_2(0) = 0 \\ c_1 + c_2(1) = 0 \\ c_1 + c_2(2) = -2 \end{array}\right\} \rightarrow \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}}_{A} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ -2 \end{bmatrix}}_{b}$$

$$\rightarrow A^T A \vec{x} = A^T \vec{b} \rightarrow \begin{bmatrix} 4 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -1 \\ -5 \end{bmatrix} \rightarrow \boxed{\vec{y} = 0.2 - 0.9T}$$

$SE(\text{line}) = (-.1)^2 + (.2)^2 + (.7)^2 + (-.4)^2 = \boxed{.7}$

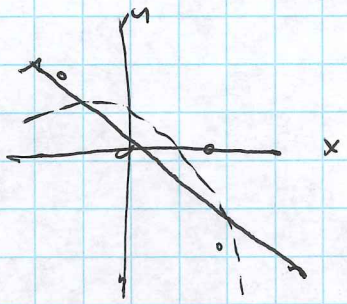
$RMSE = \frac{\sqrt{.7}}{\sqrt{4}} \approx \boxed{.418}$

(2) $\vec{y} = c_1 + c_2 T + c_3 T^2 \rightarrow$

$$\left.\begin{array}{l} c_1 + c_2(-1) + c_3(-1)^2 = 1 \\ c_1 + c_2(0) + c_3(0)^2 = 0 \\ c_1 + c_2(1) + c_3(1)^2 = 0 \\ c_1 + c_2(2) + c_3(2)^2 = -2 \end{array}\right\} \rightarrow \text{by hand}$$

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -2 \end{bmatrix} \rightarrow A^TA\bar{x} = A^T\bar{b} \rightarrow \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} -1 \\ -5 \\ -7 \end{bmatrix}$$

$$\hat{y} = 0.45 - 0.65t - 0.25t^2$$



$SE(\text{quad}) = .45$, $RMSE(\text{quad}) \approx .335$

(Quad fit appears to be superior).

Compression Least Squares is a classic example of data compression, where we fit a data set with a minimally complex model (i.e. a model with few parameters). Oftentimes, in practice least squares is used to replace "noisy" data with a plausible underlying model (see: signal processing)

Conditioning Note that in general, $\text{cond}(A^TA) \approx \text{cond}(A)^2$, so least squares can greatly increase the possibility of an ill-conditioned problem. Other methods such as QR factorization (4.3) compute least squares without need for A^TA.

In this section we encountered the so-called **Vandermonde Matrix**.

(Used in polynomial curve fitting).

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & & & & \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}\begin{bmatrix} c_1 \\ \vdots \\ c_{n+1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$