1. **Vector Derivative for Least Squares**

Recall that for least squares, we are trying to minimize $\|\vec{b}-\mathbf{A}\vec{x}\|^2$.

$$\begin{aligned}\|\vec{b}-\mathbf{A}\vec{x}\|^2 &= \langle \vec{b}-\mathbf{A}\vec{x}, \vec{b}-\mathbf{A}\vec{x}\rangle \\ &= \left(\vec{b}-\mathbf{A}\vec{x}\right)^T\left(\vec{b}-\mathbf{A}\vec{x}\right) \\ &= \left(\vec{b}^T-\vec{x}^T\mathbf{A}^T\right)\left(\vec{b}-\mathbf{A}\vec{x}\right) \\ &= \vec{x}^T\mathbf{A}^T\mathbf{A}\vec{x}-\vec{b}^T\mathbf{A}\vec{x}-\vec{x}^T\mathbf{A}\vec{b}+\vec{b}^T\vec{b}\end{aligned}$$

Note that $\vec{b}^T\mathbf{A}\vec{x}=\vec{x}^T\mathbf{A}\vec{b}$ since both sides are scalars. Therefore,

$$\|\vec{b}-\mathbf{A}\vec{x}\|^2 = \vec{x}^T\mathbf{A}^T\mathbf{A}\vec{x}-2\vec{b}^T\mathbf{A}\vec{x}+\vec{b}^T\vec{b}$$

For a column vector $\vec{x}=\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}$, the vector derivative of a scalar y is defined as a row vector:

$$\frac{dy}{d\vec{x}} = \begin{bmatrix}\frac{dy}{dx_1} & \frac{dy}{dx_2} & \cdots & \frac{dy}{dx_n}\end{bmatrix}$$

In this case, $y=\vec{x}^T\mathbf{A}^T\mathbf{A}\vec{x}-2\vec{b}^T\mathbf{A}\vec{x}+\vec{b}^T\vec{b}$. Evaluating $\frac{dy}{d\vec{x}}$ (out-of-scope for this class) gives:

$$\frac{dy}{d\vec{x}} = 2\vec{x}^T\mathbf{A}^T\mathbf{A}-2\vec{b}^T\mathbf{A}$$

To find the minimum, we set $\frac{dy}{d\vec{x}}=\vec{0}^T$.

$$\frac{dy}{d\vec{x}} = 2\vec{x}^T\mathbf{A}^T\mathbf{A}-2\vec{b}^T\mathbf{A}=\vec{0}^T$$

$$2\mathbf{A}^T\mathbf{A}\vec{x}-2\mathbf{A}^T\vec{b}=\vec{0}$$

$$2\mathbf{A}^T\mathbf{A}\vec{x}=2\mathbf{A}^T\vec{b}$$

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

2. Least Squares: A Toy Example

Let's start off by solving a little example of least squares.

We're given the following system of equations:

$$\begin{bmatrix} 1 & 4 \\ 3 & 8 \\ 0 & 0 \end{bmatrix} \vec{x} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix},$$

where $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

(a) Why can we not solve for $\vec{x}$ exactly?

(b) Find $\vec{\hat{x}}$, the *least squares estimate* of $\vec{x}$, using the formula we derived in lecture.

(c) Now, let's try to find $\vec{\hat{x}}$ in a different (geometric) way. How might you do it?

3. Ohm's Law With Noise

We are trying to measure the resistance of a black box. We apply various i_{test} currents and measure the ouput voltage v_{test}. Sometimes, we are quite fortunate to get nice numbers. Oftentimes, our measurement tools are a little bit noisy, and the values we get out of them are not accurate. However, if the noise is completely random, then the effect of it can be averaged out over many samples. Say that we repeat our test many times.

Test	i_{test} (mA)	v_{test} (V)
1	10	21
2	3	7
3	-1	-2
4	5	8
5	-8	-15
6	-5	-11

(a) Plot the measured voltage as a function of the current.

(b) Suppose we stack the currents and voltages to get $\vec{I} = \begin{bmatrix} 10 \\ 3 \\ -1 \\ 5 \\ -8 \\ -5 \end{bmatrix}$ and $\vec{V} = \begin{bmatrix} 21 \\ 7 \\ -2 \\ 8 \\ -15 \\ -11 \end{bmatrix}$. Can you solve for R?

What conditions must $\vec{I}$ and $\vec{V}$ satisfy in order for us to solve for R?

(c) Ideally, we would like to find R such that $\vec{V} = \vec{I}R$. If we cannot do this, we'd like to find a value of R that is the *best* solution possible, in the sense that $\vec{I}R$ is as "close" to $\vec{V}$ as possible. The idea of a best solution is subjective and dependent on the cost function we are using. One way of expressing this cost function in terms of R is to quantify the difference between each component of $\vec{V}$ (V_j) and each component of $\vec{I}R$ (I_jR) and add these "differences" up as follows:

$$\text{cost}(R) = \sum_{j=1}^{6} (V_j - I_jR)^2$$

Do you think this is a good cost function? Why or why not?

(d) Show that you can also express the above cost function in vector form, that is,

$$\text{cost}(R) = \left\langle (\vec{V} - \vec{I}R), (\vec{V} - \vec{I}R) \right\rangle$$

(e) Find $\hat{R}$, the optimal R that minimizes $\text{cost}(R)$.

Hint: Use calculus and minimize the expression in part (c).

(f) On your original IV plot, also plot the line $v = \hat{R}i$. Can you visually see why this line "fits" the data well? What if we had guessed $R = 3$? How well would we have done? What about $R = 1$?

Calculate the cost functions for each of these choices of R to validate your answer.

(g) Now, suppose that we add a new data point: $i_7 = 2\,\text{mA}$, $v_7 = 4\,\text{V}$. Will $\hat{R}$ increase, decrease, or remain the same? Why? What does that say about the line $v = \hat{R}i$?

(h) Let's add another data point: $i_8 = 4\,\text{mA}, v_8 = 11\,\text{V}$. Will $\hat{R}$ increase, decrease, or remain the same? Why? What does that say about the line $v = \hat{R}i$?

(i) Now your mischievous friend has hidden the black box. You want to know what the output voltage would be across the terminals if you applied $5.5\,\text{mA}$ through the black box. What would your best guess be? (This is an example of estimation from machine learning! You have *learned* what is going on inside the black box by making observations, and now you're using what you learned to make estimates.)