1. **Vector Derivative for Least Squares**

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

\[
\| \vec{b} - A \vec{x} \|^2 = (\vec{b} - A \vec{x})^T (\vec{b} - A \vec{x})
\]

\[
= (\vec{b} - A \vec{x})^T (\vec{b} - A \vec{x})
\]

\[
= \vec{b}^T - \vec{x}^T A^T \vec{b} + \vec{x}^T A \vec{x} - \vec{b}^T A \vec{x} + \vec{A}^T \vec{x} + \vec{A}^T b - \vec{b}^T b
\]

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

\[
\| \vec{b} - A \vec{x} \|^2 = \vec{x}^T A^T A \vec{x} - 2\vec{b}^T A \vec{x} + \vec{b}^T \vec{b}
\]

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:

\[
\frac{dy}{d\vec{x}} = \begin{bmatrix} \frac{dy}{dx_1} \\ \frac{dy}{dx_2} \\ \vdots \\ \frac{dy}{dx_n} \end{bmatrix}
\]

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

\[
\frac{dy}{d\vec{x}} = 2\vec{x}^T A^T A - 2\vec{b}^T A
\]

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

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

\[
2\vec{A}^T \vec{x} - 2\vec{A}^T \vec{b} = 0
\]

\[
2\vec{A}^T \vec{x} = 2\vec{A}^T \vec{b}
\]

\[
\vec{A}^T \vec{x} = \vec{A}^T \vec{b}
\]
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:

\[
\begin{bmatrix}
1 & 4 \\
3 & 8 \\
0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\end{bmatrix}
=
\begin{bmatrix}
3 \\
1 \\
2 \\
\end{bmatrix},
\]

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

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

Answer:
Recall from the earlier linear algebra module that in order for there to be a solution for the matrix system \( A \vec{x} = \vec{b} \), we must have \( \vec{b} \in \text{Col}(A) \).

Here, \( A = \begin{bmatrix}
1 & 4 \\
3 & 8 \\
0 & 0 \\
\end{bmatrix} \) and \( \vec{b} = \begin{bmatrix}
3 \\
1 \\
2 \\
\end{bmatrix} \). However, we can see that in this case we have \( \vec{b} \notin \text{Col}(A) \) because looking at the last row of \( A \), we see that there does not exist a vector \( \vec{x} \), such that \[ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \vec{x} = 2. \]

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

Answer:
Recall the equation to find the linear least squares estimate:

\[ \vec{x} = (A^T A)^{-1} A^T \vec{b} \]

Plugging in \( A = \begin{bmatrix}
1 & 4 \\
3 & 8 \\
0 & 0 \\
\end{bmatrix} \) and \( \vec{b} = \begin{bmatrix}
3 \\
1 \\
2 \\
\end{bmatrix} \), we get \( \vec{x} = \begin{bmatrix}
-5 \\
2 \\
\end{bmatrix} \).

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

Answer:
To determine the least squares estimate, we need to find the vector \( \vec{y} \in \text{Col}(A) \), such that \( \vec{y} \) is as close as possible to \( \vec{b} \). We know that \( \text{Col}(A) = \left\{ \begin{bmatrix} z_1 \\ z_2 \\ 0 \end{bmatrix} \middle| z_1, z_2 \in \mathbb{R} \right\} \), i.e. the xy-plane. Therefore, \( \vec{y} = \begin{bmatrix}
3 \\
1 \\
0 \\
\end{bmatrix} \) will be the vector in \( \text{Col}(A) \) that is as close as possible to \( \vec{b} \).

Why does the above statement make sense? There are two ways to see it:

- To minimize the square of the Euclidean distance between \( \vec{y} \) and \( \vec{b} \), we want to choose \( \vec{y} \), such that the error vector, \( \vec{e} = \vec{b} - \vec{y} \), is perpendicular to any vector in \( \text{Col}(A) \). For \( \vec{y} = \begin{bmatrix}
3 \\
1 \\
0 \\
\end{bmatrix} \), we get \( \vec{e} = \begin{bmatrix}
0 \\
0 \\
2 \\
\end{bmatrix} \). We notice that for any vector \( \vec{z} = \begin{bmatrix} z_1 \\ z_2 \\ 0 \end{bmatrix} \),

\[
\langle \vec{z}, \vec{e} \rangle = z_1 \cdot 0 + z_2 \cdot 0 + 0 \cdot 2 = 0
\]
• Let’s consider any \( \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix} \). Then, the error vector will be \( \vec{e} = \begin{bmatrix} 3 - y_1 \\ 1 - y_2 \\ 2 \end{bmatrix} \). We wish to minimize \( \|\vec{e}\|_2^2 \). Therefore, it makes sense to set \( y_1 = 3 \) and \( y_2 = 1 \), giving us \( \vec{y} = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} \).

Hence, we can simply solve for the system:

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

which simply amounts to solving the following:

\[
\begin{bmatrix} 1 & 4 \\ 3 & 8 \end{bmatrix} \vec{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}
\]

Then, we have \( \vec{x} = \begin{bmatrix} -5 \\ 2 \end{bmatrix} \).

3. Ohm’s Law With Noise

We are trying to measure the resistance of a black box. We apply various \( i_{\text{test}} \) currents and measure the output voltage \( v_{\text{test}} \). 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.

<table>
<thead>
<tr>
<th>Test</th>
<th>( i_{\text{test}} ) (mA)</th>
<th>( v_{\text{test}} ) (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>-8</td>
<td>-15</td>
</tr>
<tr>
<td>6</td>
<td>-5</td>
<td>-11</td>
</tr>
</tbody>
</table>

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

Answer:

Notice that these points do not lie on a line!

(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 \( R \)?

Answer:
We cannot solve for \( R \) because \( \vec{V} \) is not a scalar multiple of \( \vec{I} \). In general, we need \( \vec{V} \) to be a scalar multiple of \( \vec{I} \) to be able to solve for \( R \) exactly (another linear algebraic way of saying this is that \( \vec{V} \) is in the span of \( \vec{I} \)).

We know that the physical reason we are not able to solve for \( R \) is that we have imperfect observations of the voltage across the terminals, \( \vec{V} \). Although these observations are imperfect, they are observations, all the same. Therefore, now that we know we cannot solve for \( R \), a very pertinent goal would be to find a value of \( R \) that approximates the relationship between \( \vec{I} \) and \( \vec{V} \) as closely as possible.

Let’s move on and see how we do this.

(c) Ideally, we would like to find \( R \) such that \( \vec{V} = \vec{IR} \). If we cannot do this, we’d like to find a value of \( R \) that is the best solution possible, in the sense that \( \vec{IR} \) 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 \( \vec{IR} \) \( (I_jR) \) 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?

**Answer:**
For each point \((I_j, V_j)\), we want \(|V_j - I_jR|\) to be as small as possible. We can call this term the individual error term for this point.

One way of looking at the aggregate “error” in our fit is to add up the squares of the individual errors, so that all errors add up. If we did not square the differences, then a positive difference and a negative difference cancel each other out. This is precisely what we’ve done in the cost function.

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

\[
\text{cost}(R) = \langle (\vec{V} - \vec{IR}), (\vec{V} - \vec{IR}) \rangle
\]

**Answer:**
Let’s define the error vector as \( \vec{e} = \vec{V} - \vec{IR} \).

Then, we observe that \( e_j = V_j - I_jR \).

Therefore,

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

\[
= \sum_{j=1}^{6} e_j^2
\]

\[
= \|\vec{e}\|^2
\]

\[
= \langle e, e \rangle
\]

\[
= \langle (\vec{V} - \vec{IR}), (\vec{V} - \vec{IR}) \rangle
\]
(e) Find $\hat{R}$, the optimal $R$ that minimizes $\text{cost}(R)$.

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

**Answer:**
First, note that

$$\frac{d\text{cost}(R)}{dR} = -2 \sum_{j=1}^{6} I_j (V_j - I_j R)$$

For $R = \hat{R}$, we will have $\frac{d\text{cost}(R)}{dR} = 0$. This means that

$$-2 \sum_{j=1}^{6} I_j (V_j - I_j \hat{R}) = 0,$$

which will ultimately give us

$$\hat{R} = \frac{\sum_{j=1}^{6} I_j V_j}{\sum_{j=1}^{6} I_j^2} = \frac{\langle \vec{I}, \vec{V} \rangle}{\|\vec{I}\|^2_2}.$$ 

In our particular example, $\langle \vec{I}, \vec{V} \rangle = 448$ and $\|\vec{I}\|^2_2 = 224$. Therefore, we will get $\hat{R} = 2$ again!

(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 $R = 1$?

**Answer:** When $V = 2I$, we have

$$\text{cost}(R) = (21 - 2 \cdot 10)^2 + (7 - 2 \cdot 3)^2 + (-2 - 2 \cdot (-1))^2 +
(8 - 2 \cdot 5)^2 + (-15 - 2 \cdot (-8))^2 + (-11 - 2 \cdot (-5))^2$$

$$= 8.$$

When $V = 3I$, we have

$$\text{cost}(R) = (21 - 3 \cdot 10)^2 + (7 - 3 \cdot 3)^2 + (-2 - 3 \cdot (-1))^2 +
(8 - 3 \cdot 5)^2 + (-15 - 3 \cdot (-8))^2 + (-11 - 3 \cdot (-5))^2$$

$$= 232.$$

When $V = I$, we have

$$\text{cost}(R) = (21 - 1 \cdot 10)^2 + (7 - 1 \cdot 3)^2 + (-2 - 1 \cdot (-1))^2 +
(8 - 1 \cdot 5)^2 + (-15 - 1 \cdot (-8))^2 + (-11 - 1 \cdot (-5))^2$$

$$= 232.$$

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

**Answer:** We can qualitatively see that $\hat{R}$ will remain 2. This is because we already obtained $\hat{R}$ to fit our previous data in the best way. Now, you should notice that this new piece of data $(i_7, v_7)$ also lies exactly on the line $V = \hat{R}I$. Therefore, you have no reason to change $\hat{R}$. It is the best fit for the old data and will fit the new data anyway.
(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 $v = \hat{R}i$?

**Answer:**

We can qualitatively see that $\hat{R}$ should be something greater than or equal to 2. This is because you have already obtained $\hat{R}$ to fit your previous data in the best way. Now, you notice that this new piece of data $(I_8, V_8)$ also lies above the line $V = \hat{R}I$! Therefore, if you decreased $\hat{R}$, it would be a worse fit for the old data and the new data. You would increase $\hat{R}$ to find a better fit.

(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 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.)

**Answer:**

Hopefully, by now, it makes sense to the class that you will estimate $\hat{V} = 5.5\, \text{mA} \cdot \hat{R} = 5.5\, \text{mA} \cdot 2\, \text{k}\Omega = 11\, \text{V}$. This is an example of estimation from machine learning! You have learned what is going on inside the black box, that is, $\hat{R}$, by making observations of $\vec{I}$ and $\vec{V}$. Now, you are using what you have learned, $\hat{R}$, to estimate $\hat{V}$ for new values of $I$. 

EECS 16A, Summer 2017, Discussion 7A 6