**ME 7247: Advanced Control Systems** **Fall 2022–23**

# Lecture 2: Least squares estimation

Tuesday September 13, 2022

*Lecturer: Laurent Lessard* *Scribe: Laurent Lessard*

Least squares problems, optimization formulation, necessary and sufficient conditions for optimality, solution using geometrical intuition and calculus.

## 1 Least squares estimation

Consider the equation $Ax = b$, where $A \in \mathbb{R}^{m\times n}$. The approach we will describe works for any $A$, but for now, imagine a case where $m > n$ ($A$ is a tall matrix), and we are in the *estimation* setup; there is no $x$ satisfying $Ax = b$, so we would like to find an approximate solution $A\hat{x} \approx b$.

Define the *residual* $r := Ax - b$. As mentioned above, we assume there is no way to make $r$ zero. Instead, we try to make $r$ *small.* In the least squares problems, we do this by making $\|r\|$ as small as possible, where $\|\cdot\|$ is the standard Euclidean norm. There are other valid ways of making $r$ small, for example by using a different norm. This will generally produce different answers depending on the choice of norm. "Least squares" refers exclusively to the case where we use the Euclidean norm. We will use optimization notation to write the problem. If we let $\hat{x}$ be the optimal $x$ and we let $\hat{r} := A\hat{x} - b$ be the associated optimal residual, we can write least squares in two ways:

$$\|\hat{r}\|^2 \;=\; \underset{x\in\mathbb{R}^n}{\text{minimize}}\; \|Ax-b\|^2 \qquad \text{or} \qquad \hat{x} \;\in\; \underset{x\in\mathbb{R}^n}{\arg\min}\; \|Ax-b\|^2$$

On the left, "minimize" finds the smallest possible squared residual. The reason for squaring it is that it removes the square root associated with the norm so it's a bit neater. It makes no difference; if the residual is as small as possible, so is the squared residual. On the right, "arg min" finds an $x$ that achieves the minimum value. In general, there could be many $\hat{x}$ that achieve the minimal residual, so the argmin is a *set.* Both formulations solve the same problem.

### 1.1 Geometric intuition

Another way to put this is that we want to find the vector $\hat{b} \in \text{range}(A)$ that is as close to $b$ as possible ($\|\hat{b} - b\|$ as small as possible). In optimization notation, we would write this as:

$$\underset{x\,\in\,\mathbb{R}^n}{\text{minimize}}\; \|Ax-b\|^2 \;=\; \underset{\hat{b}\,\in\,\text{range}(A)}{\text{minimize}}\; \|\hat{b}-b\|^2$$

Note that we changed $x \in \mathbb{R}^n$ to $\hat{b} \in \text{range}(A)$. Geometrically, this looks like Fig. 1. Intuitively, the optimal residual $\hat{r} := \hat{b} - b = A\hat{x} - b$ should be orthogonal to every vector in range$(A)$. In other words, $\hat{r} \in \text{range}(A)^\perp$, so $\hat{r}^\mathsf{T} Ax = 0$ for all $x \in \mathbb{R}^n$. This implies that $\hat{r}^\mathsf{T} A = 0$. Substituting $\hat{r} = A\hat{x} - b$ and taking the transpose, we obtain the *normal equations*:

$$A^\mathsf{T}(A\hat{x} - b) = 0, \quad \text{or equivalently,} \quad A^\mathsf{T}A\hat{x} = A^\mathsf{T}b. \tag{1}$$

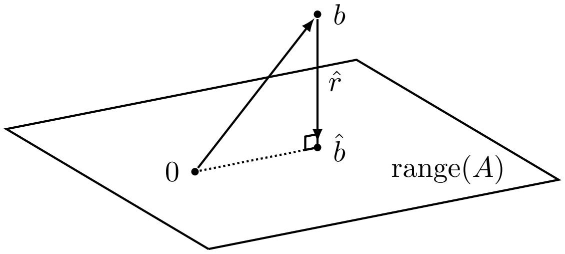

Figure 1: Geometric intuition for minimizing $\|Ax - b\|$: we are looking for the point in range$(A)$ that is closest to $b$. This happens when $\hat{r} \in \text{range}(A)^{\perp}$.

### 1.2 Calculus intuition

Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, then at a minimum point, it must be the case that $\frac{\partial f}{\partial x_i} = 0$ for all $i = 1, \ldots, n$. If that were not the case, then we could change $x_i$ by a small amount and $f$ would get smaller, contradicting the minimality of $f$. Another way of putting this is that the *gradient* of $f$ should be zero. The gradient is the vector of all partial derivatives:

$$\nabla f := \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

**Lemma 1.1.** *The gradient of a linear or quadratic function of $x$ are as follows.*

- *For any $c \in \mathbb{R}^n$, we have $\nabla\langle c, x\rangle = \nabla(c^{\mathsf{T}}x) = c$.*
- *For any $A \in \mathbb{R}^{n\times n}$, we have $\nabla\langle Ax, x\rangle = \nabla x^{\mathsf{T}}Ax = (A + A^{\mathsf{T}})x$.*

*Proof.* For an inner product $f(x) = \langle c, x\rangle = c^{\mathsf{T}}x$:

$$\frac{\partial}{\partial x_k}\langle c, x\rangle = \frac{\partial}{\partial x_k}\sum_{i=1}^{n} c_i x_i = c_k. \qquad \text{Therefore, } \nabla\langle c, x\rangle = \nabla(c^{\mathsf{T}}x) = c$$

For a quadratic form $f(x) = x^{\mathsf{T}}Ax = \langle x, Ax\rangle$, we have

$$\frac{\partial}{\partial x_k}x^{\mathsf{T}}Ax = \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\left(\frac{\partial x_i}{\partial x_k}x_j + \frac{\partial x_j}{\partial x_k}x_i\right) = \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} a_{ik}x_i.$$

Therefore, $\nabla(x^{\mathsf{T}}Ax) = (A + A^{\mathsf{T}})x$. Note that in the case where $n = 1$ ($A$ is a scalar), we recover the familiar expression $\frac{\mathrm{d}}{\mathrm{d}x}ax^2 = 2ax$. ■

Using Lemma 1.1, the gradient of the squared norm of the residual is given by

$$\begin{aligned} \nabla\|Ax - b\|^2 &= \nabla(Ax - b)^{\mathsf{T}}(Ax - b) \\ &= \nabla\big(x^{\mathsf{T}}(A^{\mathsf{T}}A)x - 2b^{\mathsf{T}}Ax + b^{\mathsf{T}}b\big) \\ &= 2(A^{\mathsf{T}}A)x - 2A^{\mathsf{T}}b \end{aligned}$$

If the gradient at $x = \hat{x}$ is zero, we immediately recover the normal equations: $A^{\mathsf{T}}A\hat{x} = A^{\mathsf{T}}b$.

### 1.3 Formal proof

The calculus approach and the geometric intuition only provide *necessary conditions* for optimality: "If $\hat{x}$ is an optimal point, *then* the normal equations hold". It turns out the normal equations are sufficient as well: "If $\hat{x}$ satisfies the normal equation, then $\hat{x}$ is an optimal point". But we did not prove it! Setting the derivative equal to zero proves necessity, but not sufficiency. This is because a zero derivative means we have an extremal point, which could be a maximum, a minimum, or a saddle point. All minima have a zero derivative, but not all zero-derivative points are minima.

Here is a more formal statement and proof of necessity and sufficiency.

**Theorem 1.2.** *Suppose $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The following statements are equivalent.*

*(i) $\hat{x} \in \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|^2$. In other words, $\|A\hat{x} - b\| \le \|Ax - b\|$ for all $x \in \mathbb{R}^n$.*

*(ii) $\hat{x}$ satisfies the normal equations: $A^\mathsf{T} A\hat{x} = A^\mathsf{T} b$.*

*Proof.* Suppose Item (ii) does not hold. Then $A^\mathsf{T}(A\hat{x} - b) \neq 0$. so there must exist some $z \in \mathbb{R}^n$ such that $z^\mathsf{T} A^\mathsf{T}(A\hat{x} - b) \neq 0$. In other words, $\langle Az, A\hat{x} - b\rangle \neq 0$. Define the point

$$\tilde{x} := \hat{x} - \frac{\langle Az, A\hat{x} - b\rangle}{\|Az\|^2} z.$$

We can divide by $\|Az\|$ in the definition of $\tilde{x}$ since $Az \neq 0$. This must be the case because otherwise we would have $\langle Az, A\hat{x} - b\rangle = 0$. Consider using $\tilde{x}$ instead of $\hat{x}$ now:

$$\begin{aligned}
\|A\tilde{x} - b\|^2 &= \left\| A\hat{x} - b - \tfrac{\langle Az, A\hat{x}-b\rangle}{\|Az\|^2} Az \right\|^2 \\
&= \|A\hat{x} - b\|^2 + \left\| \tfrac{\langle Az, A\hat{x}-b\rangle}{\|Az\|^2} Az \right\|^2 - 2\left\langle A\hat{x} - b, \tfrac{\langle Az, A\hat{x}-b\rangle}{\|Az\|^2} Az \right\rangle \\
&= \|A\hat{x} - b\|^2 + \tfrac{\langle Az, A\hat{x}-b\rangle^2}{\|Az\|^2} - 2\tfrac{\langle Az, A\hat{x}-b\rangle^2}{\|Az\|^2} \\
&= \|A\hat{x} - b\|^2 - \tfrac{\langle Az, A\hat{x}-b\rangle^2}{\|Az\|^2} \\
&< \|A\hat{x} - b\|^2.
\end{aligned}$$

We showed that $\tilde{x}$ achieves a strictly better residual than $\hat{x}$, so $\hat{x}$ cannot be the argmin, and (i) does not hold. We just proved that $\neg$(ii) $\implies$ $\neg$(i). Therefore, (i) $\implies$ (ii). This proves necessity in yet another way (we already showed this using an orthogonality argument and calculus argument).

Now let's prove sufficiency. Suppose that Item (ii) holds. We will prove that no matter what $x$ is chosen, we will have $\|A\hat{x} - b\| \le \|Ax - b\|$, which means that $\hat{x}$ is the argmin. Expanding the residual,

$$\begin{aligned}
\|Ax - b\|^2 &= \|(A\hat{x} - b) + (Ax - A\hat{x})\|^2 \\
&= \|A\hat{x} - b\|^2 + \|A(x - \hat{x})\|^2 + 2\langle A\hat{x} - b, Ax - A\hat{x}\rangle \\
&= \|A\hat{x} - b\|^2 + \|A(x - \hat{x})\|^2 \\
&\ge \|A\hat{x} - b\|^2.
\end{aligned}$$

The inner product in the second line vanishes because $\langle A\hat{x} - b, Ax - A\hat{x}\rangle = \langle A^\mathsf{T}(A\hat{x} - b), x - \hat{x}\rangle = 0$. Therefore, (ii) $\implies$ (i), which completes the proof. ■

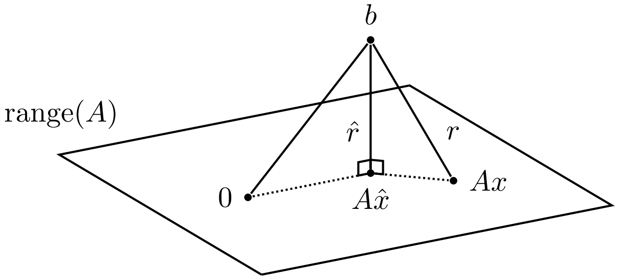

Figure 2: Geometric intuition for sufficiency. If the normal equations are satisfied, i.e. if $\hat{r} \in \text{range}(A)^{\perp}$, then any other point in range($A$) will be farther from $b$ than the chosen point because $\|r\|^2 = \|\hat{r}\|^2 + \|A(x - \hat{x})\|^2 \geq \|\hat{r}\|^2$ (Pythagorean theorem).

### 1.4 Full rank case

When we looked at solutions of $Ax = b$, we saw that when $A$ has full column rank, any solutions are unique. The same is true of least squares problem, except we don't have to worry about existence, since least squares problems always have a solution.

**Lemma 1.3.** *Suppose $A \in \mathbb{R}^{m \times n}$. Then* $\text{null}(A) = \text{null}(A^{\mathsf{T}}A)$.

*Proof.* Suppose $x \in \text{null}(A)$. Then, $Ax = 0 \implies A^{\mathsf{T}}Ax = 0 \implies x \in \text{null}(A^{\mathsf{T}}A)$. Therefore, $\text{null}(A) \subseteq \text{null}(A^{\mathsf{T}}A)$. Conversely, suppose $x \in \text{null}(A^{\mathsf{T}}A)$. Then,

$$A^{\mathsf{T}}Ax = 0 \implies x^{\mathsf{T}}A^{\mathsf{T}}Ax = 0 \implies \|Ax\|^2 = 0 \implies Ax = 0 \implies x \in \text{null}(A).$$

Therefore, $\text{null}(A^{\mathsf{T}}A) \subseteq \text{null}(A)$. This completes the proof. ■

**Corollary 1.4.** *Suppose $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. If $A$ has full column rank, then the least squares problem $\hat{x} \in \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|^2$ has a unique solution, and it is given by $\hat{x} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}b$.*

*Proof.* If $A$ has full column rank, then $\text{null}(A) = \{0\}$. By Lemma 1.3, $\text{null}(A^{\mathsf{T}}A) = \{0\}$ as well, so $A^{\mathsf{T}}A$ has full column rank. This matrix is square and full rank, so it is invertible. Therefore, the normal equations $A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$ have a unique solution, and it is given by $(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}b$. ■

When $A$ has full column rank, the matrix $A^{\dagger} := (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$ is called the *pseudoinverse* of $A$ (also known as the *Moore–Penrose pseudoinverse*). There is a more general definition that holds for non-full-rank $A$; we'll see this later. In the full column rank case, we have the following properties:

- If $A \in \mathbb{R}^{m \times n}$, then $A^{\dagger} \in \mathbb{R}^{n \times m}$. So $A^{\dagger}$ has the same shape as $A^{\mathsf{T}}$.
- $A^{\dagger}A = I_n$ (the $n \times n$ identity matrix). In other words, $A^{\dagger}$ is a *left*-inverse of $A$.
- $AA^{\dagger}$ is the projection matrix that maps a vector onto range($A$). In other words, $AA^{\dagger}b$ is the vector in range($A$) that is closest to $b$. Mathematically,

$$(AA^{\dagger})b = \text{proj}_{\text{range}(A)}b = \underset{z \in \text{range}(A)}{\arg\min} \|z - b\|.$$

- If $A$ is square and full rank (invertible), then $A^{\dagger} = A^{-1}$.

## 2 Linear regression

The most popular example of least squares is *linear regression*. For this example, suppose we want to estimate the age of maple trees in a particular forest based on the circumferences of their trunks. Finding the true age of a tree is an invasive and time-consuming procedure, so we only have data for $N = 10$ trees.

Our data consists of points $(x_i, y_i)$ for $i = 1, \ldots, N$. Here, $x_i$ and $y_i$ are the circumference and age of tree $i$, respectively. The data and a scatter plot are shown in Fig. 3.

| $x_i$ (inches) | $y_i$ (years) |
|---|---|
| 18.1 | 83 |
| 8.0 | 42 |
| 16.8 | 79 |
| 2.8 | 16 |
| 3.9 | 24 |
| 12.7 | 73 |
| 11.5 | 60 |
| 9.5 | 44 |
| 8.0 | 47 |
| 13.2 | 67 |

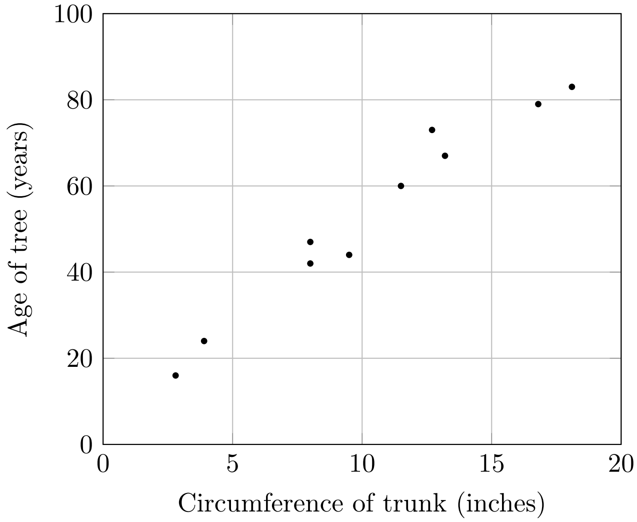

Figure 3: Scatter plot of tree data

We want to predict the age of a tree based on its circumference. Based on the data and our intuition about how trees grow, we suspect a formula of the form

$$y \approx mx + b$$

should work, where $m$ and $b$ are the slope and intercept of a line of best fit. One way to formulate this is as a least-squares problem. So the task is to find $m$ and $b$ so that we

$$\underset{m,b}{\text{minimize}} \sum_{i=1}^{N} (mx_i + b - y_i)^2 .$$

Our variables are $m$ and $b$. This is a least squares problem to minimize$_x$ $\|Az - y\|^2$, where

$$A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}, \qquad z = \begin{bmatrix} m \\ b \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

This is a least squares problem where $A \in \mathbb{R}^{N\times 2}$ has full column rank. The normal equations are

$A^{\mathsf{T}} A\hat{z} = A^{\mathsf{T}} y$. Let's divide both sides by $N$ (does not change anything), and evaluate:

$$\frac{1}{N}A^{\mathsf{T}}A = \frac{1}{N}\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N x_i^2 & \frac{1}{N}\sum_{i=1}^N x_i \\ \frac{1}{N}\sum_{i=1}^N x_i & 1 \end{bmatrix} = \begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix},$$

$$\frac{1}{N}A^{\mathsf{T}}y = \frac{1}{N}\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N x_i y_i \\ \frac{1}{N}\sum_{i=1}^N y_i \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix},$$

where we used the notation $\overline{x}$ as a short-hand to denote averaging. The normal equations are

$$\begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix}. \qquad \text{With numerical values:} \quad \begin{bmatrix} 131.893 & 10.45 \\ 10.45 & 1 \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 659.94 \\ 53.5 \end{bmatrix}.$$

This is always a 2 × 2 system of equations, no matter how many data points $N$ we have! Solving these equations, we obtain $\hat{m} = 4.445$ and $\hat{b} = 7.047$. We can plot the line $y = \hat{m}x + \hat{b}$ and we get the following:

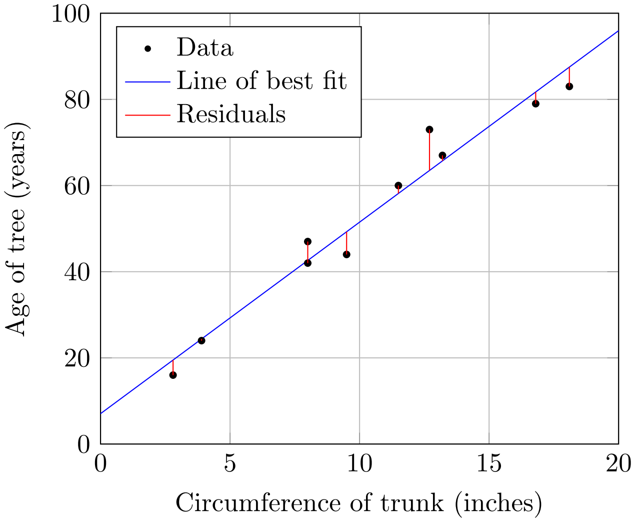

Figure 4: Scatter plot of tree data with line of best fit.

The line of best fit minimizes the sum of the squares of the residuals, which are the vertical lines between each data point and the line of best fit shown in Fig. 4. This "best fit" line is exactly the same as what is produced if you use the "linear trendline" function in Excel; it's nothing more than least squares.

**Degenerate cases.** $A^{\mathsf{T}}A$ is non-invertible precisely when the columns are linearly dependent, which happens when $\bar{x} = x_1 = \cdots = x_N$. In this case, each point has the form $(\bar{x}, y_i)$ (they're on the same vertical line). Here, the normal equations have infinitely many solutions; the line of best fit is not unique. If $\bar{y} = y_1 = \cdots = y_N$ instead, the data lie on a horizontal line. Here, $y$ is a multiple of the vector of all 1's, i.e. $y \in \text{range}(A)$, so we get an exact solution (the line passes through all the points and the residual is zero).