ROB 101 - Fall 2021

Euclidean Norm, Least Squares, and Linear Regression

October 6, 2021

UNIVERSITY OF MICHIGAN

Learning Objectives

- Learn one way to assign a notion of length to a vector
- The concept of finding approximate solutions to $Ax = b$ when an exact solution does not exist and why this is extremely useful in engineering.

Outcomes

- Euclidean norm and its properties
- If $Ax = b$ does not have a solution, then for any $x \in \mathbb{R}^n$, the vector $Ax - b$ is never zero. We will call $e := Ax - b$ the error vector and search for the value of x that minimizes the norm of the error vector.
- An application of this idea is Linear Regression, one of the "super powers" of Linear Algebra: fitting functions to data.

Euclidean Norm or "Length" of a Vector

Definition

Let $v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ be a vector in $\mathbb{R}^n$. The Euclidean norm of v, denoted $\|v\|$, is defined as

$$\|v\| := \sqrt{(v_1)^2 + (v_2)^2 + \cdots + (v_n)^2}$$

Euclidean Norm or "Length" of a Vector

Definition

Let $v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ be a vector in $\mathbb{R}^n$. The Euclidean norm of v, denoted $\|v\|$, is defined as

$$\|v\| := \sqrt{(v_1)^2 + (v_2)^2 + \cdots + (v_n)^2} = \sqrt{\sum_{i=1}^{n} (v_i)^2} = \sqrt{v^\mathsf{T} v}$$

Example

The length of the vector $v = \begin{bmatrix} \sqrt{2} \\ -1 \\ 5 \end{bmatrix}$ is

The length of the vector $v = \begin{bmatrix} \sqrt{2} \\ -1 \\ 5 \end{bmatrix}$ is

$$\begin{aligned} \|v\| &:= \sqrt{(\sqrt{2})^2 + (-1)^2 + (5)^2} = \sqrt{2+1+25} = \sqrt{28} \\ &= \sqrt{4 \cdot 7} = 2\sqrt{7} \approx 5.29. \end{aligned}$$

Properties of the Norm of a vector

All norms satisfy the following properties

1. For all vectors $v \in \mathbb{R}^n$, $\|v\| \geq 0$ and moreover, $\|v\| = 0 \iff v = 0$.

All norms satisfy the following properties

1. For all vectors $v \in \mathbb{R}^n$, $\|v\| \geq 0$ and moreover, $\|v\| = 0 \iff v = 0$.

2. For any real number $\alpha \in \mathbb{R}$ and vector $v \in \mathbb{R}^n$,

$$\|\alpha v\| = |\alpha| \cdot \|v\|.$$

Properties of the Norm of a vector

All norms satisfy the following properties

1. For all vectors $v \in \mathbb{R}^n$, $\|v\| \geq 0$ and moreover, $\|v\| = 0 \iff v = 0$.

2. For any real number $\alpha \in \mathbb{R}$ and vector $v \in \mathbb{R}^n$,
$$\|\alpha v\| = |\alpha| \cdot \|v\|.$$

3. For any pair of vectors v and w in $\mathbb{R}^n$,
$$\|v + w\| \leq \|v\| + \|w\|.$$

1. The first property says that v has norm zero if, and only if, v is the zero vector.

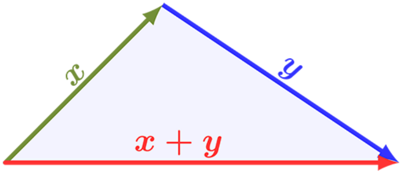
Properties of the Norm of a vector

1. The first property says that v has norm zero if, and only if, v is the zero vector.

2. For the second property, we note that we have to take the absolute value of the constant when we “factor it out” of the norm. This is because $\sqrt{a^2} = |a|$ and not a when $a < 0$. Of course, when $a \geq 0$, $\sqrt{a^2} = a$.

3. The third property is called the triangle inequality. It says that the norm of a sum of vectors is upper bounded by the sum of the norms of the individual vectors.

$$\|x + y\| \leq \|x\| + \|y\|.$$



3. The third property is called the triangle inequality. It says that the norm of a sum of vectors is upper bounded by the sum of the norms of the individual vectors.

$$\|x + y\| \leq \|x\| + \|y\|.$$

What happens if you have $||x - y||$?

3 The third property is called the triangle inequality. It says that the norm of a sum of vectors is upper bounded by the sum of the norms of the individual vectors.

$$\|x + y\| \leq \|x\| + \|y\|.$$

What happens if you have $||x - y||$?

$$\begin{aligned} \|x - y\| &= \|x + (-y)\| \\ &\leq \|x\| + \|-y\| \\ &= \|x\| + |-1| \cdot \|y\| \\ &= \|x\| + \|y\|. \end{aligned}$$

Nothing happens! ✨

We'll take three vectors in $\mathbb{R}^4$ and check the "triangle inequality."

$$u = \begin{bmatrix} 2 \\ -1 \\ 5 \\ 2 \end{bmatrix}, v = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 3 \end{bmatrix}, w = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 4 \end{bmatrix}$$

We'll take three vectors in $\mathbb{R}^4$ and check the "triangle inequality."

$$u = \begin{bmatrix} 2 \\ -1 \\ 5 \\ 2 \end{bmatrix}, v = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 3 \end{bmatrix}, w = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 4 \end{bmatrix}$$

We first compute the norms of the three vectors

$$||u|| = \sqrt{34} \approx 5.83, ||v|| = \sqrt{14} \approx 3.74, ||w|| = \sqrt{17} \approx 4.12.$$

We then form a few sums against which to test the triangle inequality

$$u + v = \begin{bmatrix} 3 \\ 1 \\ 5 \\ 5 \end{bmatrix}, u + v + w = \begin{bmatrix} 3 \\ 2 \\ 5 \\ 9 \end{bmatrix}, v + w = \begin{bmatrix} 1 \\ 3 \\ 0 \\ 7 \end{bmatrix}.$$

Then we can check that

$$\|u+v\| = \sqrt{60} \approx 7.75 \leq 9.5 \leq \|u\| + \|v\|$$
$$\|u+v+w\| = \sqrt{119} \approx 10.91 \leq 11.8 \leq \|u+v\| + \|w\|$$
$$\|u+v+w\| = \sqrt{119} \approx 10.91 \leq 13.6 \leq \|u\| + \|v\| + \|w\|$$

Unfortunately, in many interesting (and real-world) problems, the exact solution does not exist.

https://www.youtube.com/watch?v=E2evC2xTNWg

Least Squared Error Solutions to Linear Equations

Consider a system of linear equations $A_{n\times m} \cdot x_{m\times 1} = b_{n\times 1}$.

- Define the vector

  $$e(x) := Ax - b$$

  as the *error* in the solution for a given value of x.

Least Squared Error Solutions to Linear Equations

Consider a system of linear equations $A_{n\times m} \cdot x_{m\times 1} = b_{n\times 1}$.

- Define the vector

  $$e(x) := Ax - b$$

  as the *error* in the solution for a given value of x.

- We can simply write $e := Ax - b$. We write $e(x)$ to emphasize that the error is a function of x.

Remark

If the system of equations $Ax = b$ has a solution, then it is possible to make the error zero.

The norm squared of the error vector $e(x)$ is then

$$\begin{aligned}\|e(x)\|^2 := \sum_{i=1}^{n} (e_i(x))^2 &= e(x)^{\mathsf{T}} e(x) \\ &= (Ax-b)^{\mathsf{T}}(Ax-b) = \|Ax-b\|^2.\end{aligned}$$

The norm squared of the error vector $e(x)$ is then

$$\begin{aligned}\|e(x)\|^2 := \sum_{i=1}^{n}(e_i(x))^2 &= e(x)^{\mathsf{T}}e(x) \\ &= (Ax-b)^{\mathsf{T}}(Ax-b) = \|Ax-b\|^2.\end{aligned}$$

We note that $\|e(x)\|^2 \geq 0$ for any $x \in \mathbb{R}^m$ and hence zero is a lower bound on the norm squared of the error vector.

- A value $x^* \in \mathbb{R}^m$ is a *Least Squared Error Solution* to $Ax = b$ if it satisfies

$$\|e(x^*)\|^2 = \min_{x \in \mathbb{R}^m} \|Ax - b\|^2$$

- A value $x^* \in \mathbb{R}^m$ is a *Least Squared Error Solution* to $Ax = b$ if it satisfies

$$\|e(x^*)\|^2 = \min_{x \in \mathbb{R}^m} \|Ax - b\|^2$$

- If such an $x^* \in \mathbb{R}^m$ exists and is unique, we will write it as

$$x^* := \arg\min_{x \in \mathbb{R}^m} \|Ax - b\|^2.$$

Remark

With this notation, the value of x *that minimizes the error in the solution is what is returned by the function* $\arg\min$, *while the minimum value of the error is what is returned by the function* $\min$,

Remark

With this notation, the value of x *that minimizes the error in the solution is what is returned by the function* $\arg\min$*, while the minimum value of the error is what is returned by the function* $\min$*,*

- $x^* = \arg\min_{x \in \mathbb{R}^m} ||Ax - b||^2$ *is the value of* x *that achieves the minimum value of the squared norm of the error,* $||Ax - b||^2$*, while*

Remark

With this notation, the value of x that minimizes the error in the solution is what is returned by the function $\arg\min$*, while the minimum value of the error is what is returned by the function* $\min$*,*

- $x^* = \arg\min_{x \in \mathbb{R}^m} ||Ax - b||^2$ *is the value of x that achieves the minimum value of the squared norm of the error, $||Ax - b||^2$, while*
- $||e(x^*)||^2 = ||Ax^* - b||^2 = \min_{x \in \mathbb{R}^m} ||Ax - b||^2$ *is the minimum value of the "squared approximation error".*

Least Squares Solutions to Linear Equations

- Assume $A^{\mathsf{T}}A$ is invertible, i.e., the columns of A are linearly independent.

Least Squares Solutions to Linear Equations

- Assume $A^\mathsf{T} A$ is invertible, i.e., the columns of A are linearly independent.

- Then there is a unique vector $x^* \in \mathbb{R}^m$ achieving $\min_{x \in \mathbb{R}^m} ||Ax - b||^2$ and it satisfies the equation (called *the normal equations*)
$$\left(A^\mathsf{T} A\right) x^* = A^\mathsf{T} b.$$

- $$x^* = (A^\mathsf{T} A)^{-1} A^\mathsf{T} b \iff x^* = \underset{x \in \mathbb{R}^m}{\arg\min} \|Ax - b\|^2 \iff \left(A^\mathsf{T} A\right) x^* = A^\mathsf{T} b.$$

Consider a system of linear equations, with more equations than unknowns. The extra equations (rows) provide more conditions that a solution must satisfy, making non-existence of a solution a common occurrence!

$$\underbrace{\begin{bmatrix} 1.0 & 1.0 \\ 2.0 & 1.0 \\ 4.0 & 1.0 \\ 5.0 & 1.0 \\ 7.0 & 1.0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 4 \\ 8 \\ 10 \\ 12 \\ 18 \end{bmatrix}}_{b},$$

The columns of A are linearly independent. If a regular solution exists, find it. If not, then a least squared solution will be fine.

- We check that $\det(A^{\mathsf{T}} \cdot A) = 114$.

- We check that $\det(A^\mathsf{T} \cdot A) = 114$.
- We'll compute a least squared error solution to the equations, and then we'll evaluate the error; if the error is zero, we'll also have an exact solution.

$$A^\mathsf{T} \cdot A = \begin{bmatrix} 95.0 & 19.0 \\ 19.0 & 5.0 \end{bmatrix}, \; A^\mathsf{T} \cdot b = \begin{bmatrix} 246.0 \\ 52.0 \end{bmatrix} \implies x^* = \begin{bmatrix} 2.12 \\ 2.33 \end{bmatrix}$$

Example

- We check that $\det(A^\mathsf{T} \cdot A) = 114$.
- We'll compute a least squared error solution to the equations, and then we'll evaluate the error; if the error is zero, we'll also have an exact solution.

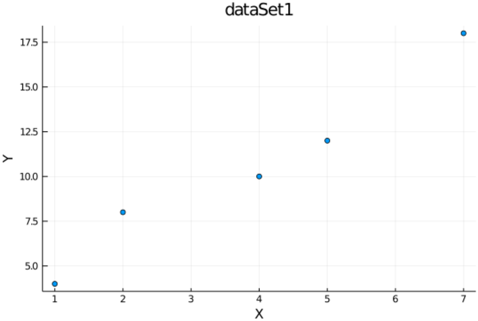
$$A^\mathsf{T} \cdot A = \begin{bmatrix} 95.0 & 19.0 \\ 19.0 & 5.0 \end{bmatrix}, \; A^\mathsf{T} \cdot b = \begin{bmatrix} 246.0 \\ 52.0 \end{bmatrix} \implies x^* = \begin{bmatrix} 2.12 \\ 2.33 \end{bmatrix}$$

-

$$e^* := Ax^* - b = \begin{bmatrix} 0.456 \\ -1.421 \\ 0.825 \\ 0.947 \\ -0.807 \end{bmatrix}, \; \|e\| = 2.111, \text{ and } \|e\|^2 = 4.456$$

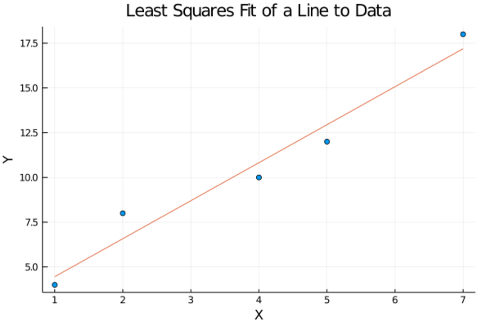
Q. How can we fit a line to the data?

i	x_i	y_i
1	1	4
2	2	8
3	4	10
4	5	12
5	7	18



Q. How can we fit a line to the data approximately?

i	x_i	y_i
1	1	4
2	2	8
3	4	10
4	5	12
5	7	18



- A linear model (here a model for a straight line) takes the form $\hat{y} = mx + b$.

- A linear model (here a model for a straight line) takes the form $\hat{y} = mx + b$.

- We set up the linear equations

$$y_i = mx_i + b = \begin{bmatrix} x_i & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}, \quad 1 \leq i \leq N,$$

- A linear model (here a model for a straight line) takes the form $\hat{y} = mx + b$.

- We set up the linear equations

$$y_i = mx_i + b = \begin{bmatrix} x_i & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix}, \quad 1 \leq i \leq N,$$

- Write it out in matrix form

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & 1 \\ x_N & 1 \end{bmatrix}}_{\Phi} \cdot \underbrace{\begin{bmatrix} m \\ b \end{bmatrix}}_{\alpha}.$$

Remark

In $y = \Phi\alpha$,

- y *is the vector of* y*-data (measurements or observations),*

Remark

In $y = \Phi\alpha$,

- y *is the vector of* y*-data (measurements or observations),*
- Φ *is called the regressor matrix (called design matrix in statistics),*

Remark

In $y = \Phi\alpha$,

- y *is the vector of* y*-data (measurements or observations),*
- Φ *is called the regressor matrix (called design matrix in statistics),*
- *and* α *is the vector of unknown coefficients (what we solve for) that parameterize the model.*

From the data table, we have

$$y = \begin{bmatrix} 4 \\ 8 \\ 10 \\ 12 \\ 18 \end{bmatrix}, \ \Phi = \begin{bmatrix} 1.0 & 1.0 \\ 2.0 & 1.0 \\ 4.0 & 1.0 \\ 5.0 & 1.0 \\ 7.0 & 1.0 \end{bmatrix}, \text{ and } \alpha = \begin{bmatrix} m \\ b \end{bmatrix}.$$

The fitting error will be $e_i = y_i - (mx_i + b)$, which when written as a vector gives

$$\underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}}_{e} = \underbrace{\begin{bmatrix} 4 \\ 8 \\ 10 \\ 12 \\ 18 \end{bmatrix}}_{Y} - \underbrace{\begin{bmatrix} 1.0 & 1.0 \\ 2.0 & 1.0 \\ 4.0 & 1.0 \\ 5.0 & 1.0 \\ 7.0 & 1.0 \end{bmatrix}}_{\Phi} \cdot \underbrace{\begin{bmatrix} m \\ b \end{bmatrix}}_{\alpha},$$

that is, $e := y - \Phi\alpha$.

Least Squares Fit to Data for Linear Regression

We propose to chose the coefficients in α so as to minimize the total squared error

$$E_{tot} = \sum_{i=1}^{5} (e_i)^2 = e^{\mathsf{T}} e = \|e\|^2 = \|y - \Phi\alpha\|^2.$$

Least Squares Fit to Data for Linear Regression

Fact

If the columns of Φ are linearly independent, or equivalently, $\Phi^{\mathsf{T}}\Phi$ is invertible, then the following are equivalent

1. $\alpha^* = \left(\Phi^{\mathsf{T}}\Phi\right)^{-1}\Phi^{\mathsf{T}}y,$
2. $\alpha^* = \arg\min_{\alpha}\|y - \Phi\alpha\|^2,$
3. $\left(\Phi^{\mathsf{T}}\Phi\right)\alpha^* = \Phi^{\mathsf{T}}y.$

Least Squares Fit to Data for Linear Regression

Fact

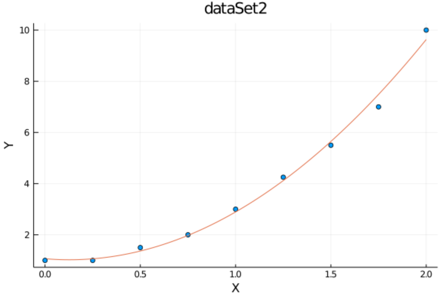
If the columns of Φ are linearly independent, or equivalently, $\Phi^\top \Phi$ is invertible, then the following are equivalent

1. $\alpha^* = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top y$,
2. $\alpha^* = \arg\min_\alpha \|y - \Phi\alpha\|^2$,
3. $\left(\Phi^\top \Phi\right) \alpha^* = \Phi^\top y$.

$$\alpha^* = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top y \iff \alpha^* = \arg\min_\alpha \|y - \Phi\alpha\|^2 \iff \left(\Phi^\top \Phi\right) \alpha^* = \Phi^\top y.$$

Q. Fit a function (quadratic!?) to the data.

i	x_i	y_i
1	0	1.0
2	0.25	1.0
3	0.5	1.5
4	0.75	2.0
5	1.0	3.0
6	1.25	4.25
7	1.5	5.5
8	1.75	7.0
9	2.0	10.0



Example

Let's choose a model of the form

$$y = c_0 + c_1 x + c_2 x^2 = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}.$$

Let's choose a model of the form

$$y = c_0 + c_1 x + c_2 x^2 = \begin{bmatrix} 1 & x & x^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}.$$

Remark

Note that even though the model is nonlinear in x, it is linear in the unknown coefficients c_0, c_1, c_2.

- Define $\hat{y}_i = c_0 + c_1 x_i + c_2 x_i^2$.

- Define $\hat{y}_i = c_0 + c_1 x_i + c_2 x_i^2$.

- The i-th term of the error vector is then

$$e_i := y_i - \hat{y}_i = y_i - (c_0 + c_1 x_i + c_2 x_i^2).$$

- Define $\hat{y}_i = c_0 + c_1 x_i + c_2 x_i^2$.
- The i-th term of the error vector is then

$$e_i := y_i - \hat{y}_i = y_i - (c_0 + c_1 x_i + c_2 x_i^2).$$

- The total squared error is

$$E_{tot} = \sum_{i=1}^{N} e_i^2.$$

- Writing out the equation $y_i = c_0 + c_1 x_i + c_2 x_i^2$, $i = 1, \cdots, N$ in matrix form yields

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & (x_1)^2 \\ 1 & x_2 & (x_2)^2 \\ \vdots & \vdots & \\ 1 & x_N & (x_N)^2 \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}}_{\alpha},$$

- Writing out the equation $y_i = c_0 + c_1 x_i + c_2 x_i^2$, $i = 1, \cdots, N$ in matrix form yields

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & (x_1)^2 \\ 1 & x_2 & (x_2)^2 \\ \vdots & \vdots & \\ 1 & x_N & (x_N)^2 \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}}_{\alpha},$$

- which gives us the equation $y = \Phi\alpha$.

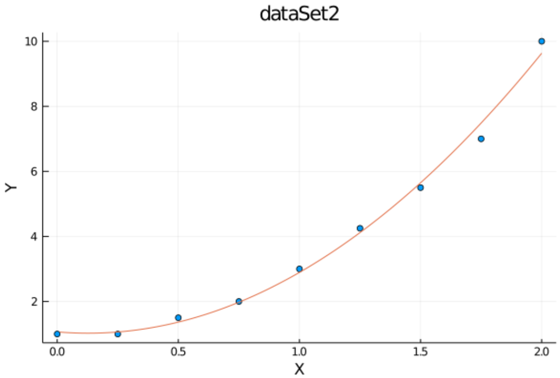
- Writing out the equation $y_i = c_0 + c_1 x_i + c_2 x_i^2$, $i = 1, \cdots, N$ in matrix form yields

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & (x_1)^2 \\ 1 & x_2 & (x_2)^2 \\ \vdots & \vdots & \\ 1 & x_N & (x_N)^2 \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}}_{\alpha},$$

- which gives us the equation $y = \Phi\alpha$.

- We plug in our numbers and check that $\det(\Phi^\top \cdot \Phi) = 40.6 \neq 0$.

Example



Next Time

- The Vector Space $\mathbb{R}^n$: Part 2
- Read Chapter 9 of ROB 101 Book