Lectures 8-10

Linear algebra: Dot products and Projections

DSC 40A, Fall 2025

Announcements

- Homework 2 was released Friday.
- Groupwork 3 is due **tonight**.
- Check out FAQs page and the tutor-created supplemental resources on the course website.

Agenda

- Recap: Simple linear regression and correlation.
- Connections to related models. ← HW3
- Dot products.
- Spans and projections. ← later this week

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

Simple linear regression

- Model: $H(x) = w_0 + w_1 x$.
- Loss function: squared loss, i.e. $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$.
- Average loss, i.e. empirical risk:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

- Optimal model parameters, found by minimizing empirical risk:

$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2} = r \frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^* \bar{x}$$

The correlation coefficient

- The correlation coefficient, r, is defined as the **average of the product of x and y, when both are in standard units**.
- Let σ_x be the standard deviation of the x_is, and $\bar{x}$ be the mean of the x_is.
- x_i in standard units is $\frac{x_i - \bar{x}}{\sigma_x}$.
- The correlation coefficient, then, is:

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Correlation and mean squared error

- **Claim**: Suppose that w_0^* and w_1^* are the optimal intercept and slope for the regression line. Then,

$$R_{\text{sq}}(w_0^*, w_1^*) = \sigma_y^2(1 - r^2)$$

- That is, the mean squared error of the regression line's predictions and the correlation coefficient, r, always satisfy the relationship above.
- Even if it's true, why do we care?
  - In machine learning, we often use both the mean squared error and r^2 to compare the performances of different models.
  - If we can prove the above statement, we can show that **finding models that minimize mean squared error** is equivalent to **finding models that maximize r^2**.

Proof that $R_{\text{sq}}(w_0^*, w_1^*) = \sigma_y^2(1 - r^2)$

$$R_{sq}(w_0^*, w_1^*) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0^* + w_1^* x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\underbrace{\bar{y} - w_1^*\bar{x}}_{w_0^*} + w_1^* x_i)\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \bar{y}) - \underbrace{r\frac{\sigma_y}{\sigma_x}}_{w_1^*}(x_i - \bar{x})\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \bar{y})^2 + r^2\frac{\sigma_y^2}{\sigma_x^2}(x_i - \bar{x})^2 - 2(y_i - \bar{y})(x_i - \bar{x})\cdot r\frac{\sigma_y}{\sigma_x}\right]$$

$$= \sigma_y^2 + r^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 - 2\frac{\sigma_y}{\sigma_x}r\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}_{* = r\cdot\sigma_x\sigma_y}$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y}$$

$$\sigma_x\sigma_y r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= (1 + r^2)\sigma_y^2 - 2r^2\sigma_x\sigma_y\cdot\frac{\sigma_y}{\sigma_x} = (1 + r^2 - 2r^2)\sigma_y^2 = (1 - r^2)\sigma_y^2$$

We previously showed: $\sum_{i=1}^{n}(y_i-\bar{y})=0$ but $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\sigma_y^2$ ← why is this not 0?

Let's explictly write the two sums:

$$\sum_{i=1}^{n}(y_i-\bar{y})=(y_1-\bar{y})+(y_2-\bar{y})+\ldots+(y_n-\bar{y})$$ sum of differences

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=(y_1-\bar{y})^2+(y_2-\bar{y})^2+\ldots+(y_n-\bar{y})^2$$ sum of squared differences

Therefore the two are different

However note if we square the sum:

$$\Big(\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})}_{0}\Big)^2=0^2=0$$

Connections to related models

Exercise (no slope model) $w_1 = 0$

Suppose we choose the model $H(x) = w_0$ and squared loss.

What is the optimal model parameter, w_0^*?

$H(x) = w_0 = h$ → constant model

loss - squared loss

week 1

$w_0^* = \text{mean}\{y_1, \ldots, y_n\}$

Exercise

(no intercept $w_0 = 0$)

Suppose we choose the model $H(x) = w_1 x$ and squared loss.

What is the optimal model parameter, w_1^*?

Groupwork 3!

Comparing mean squared errors

- With both:
  - the constant model, $H(x) = h$, and
  - the simple linear regression model, $H(x) = w_0 + w_1 x$,

  when we chose squared loss, we minimized mean squared error to find optimal parameters:

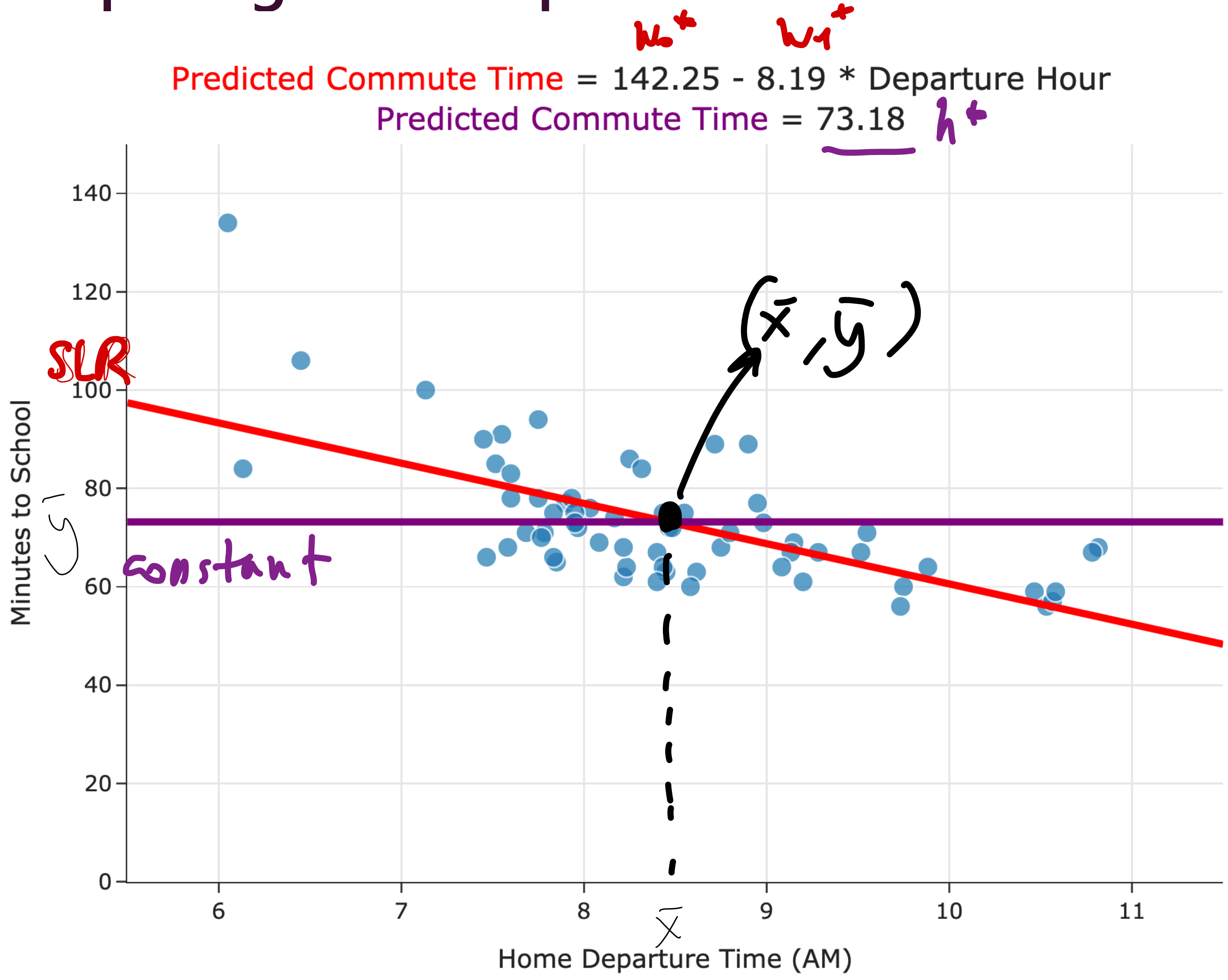
$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$$

- **Which model minimizes mean squared error more?**

h^* and w_0^* not necessarily the same

MSE(SLR) ≤ MSE(constant)

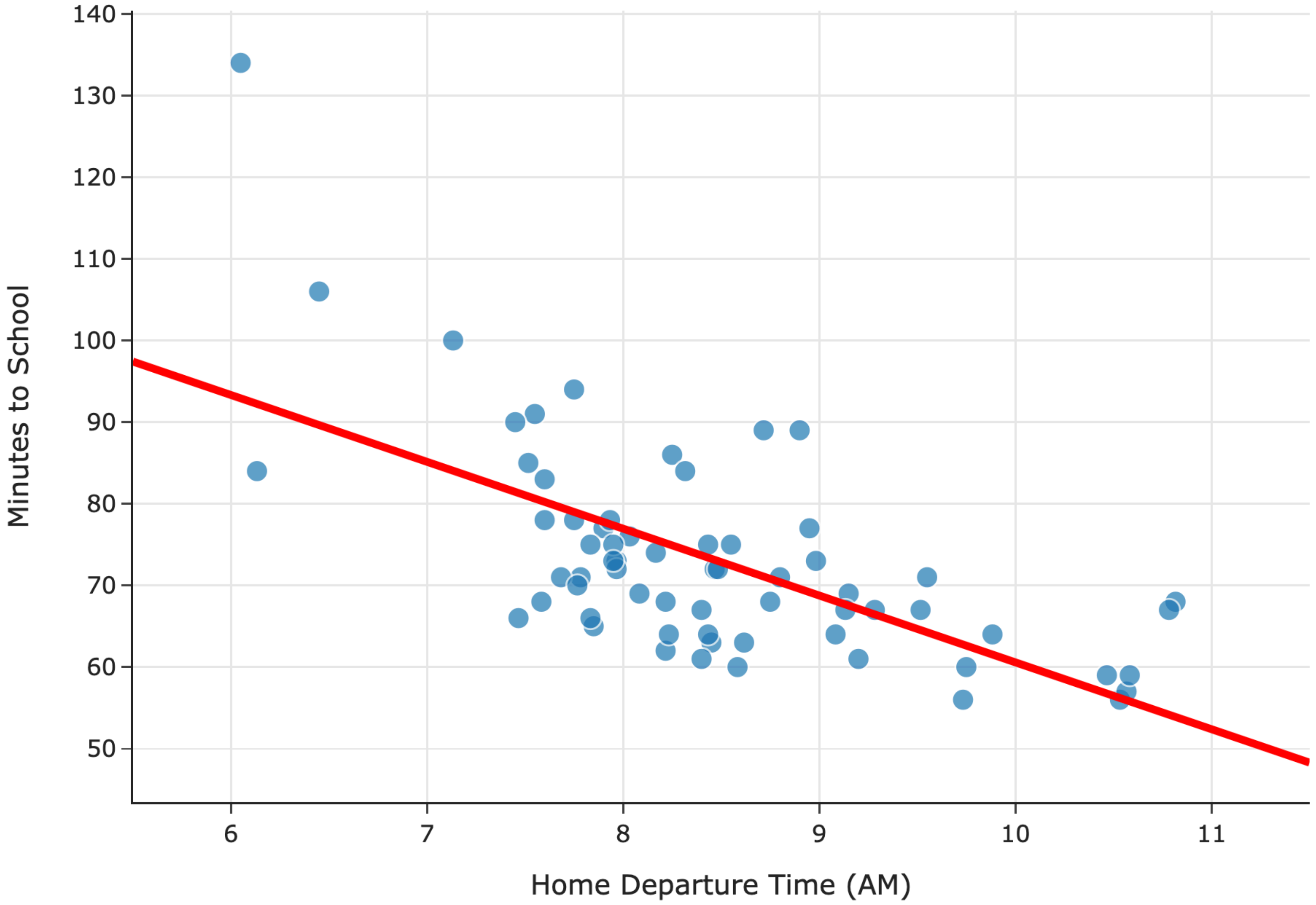
Comparing mean squared errors



$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$$

- The MSE of the best simple linear regression model is ≈ 97
- The MSE of the best constant model is ≈ 167
- The simple linear regression model is a more flexible version of the constant model.

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



Linear algebra

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature.
  - Example: Predicting commute times using departure hour and temperature.
- Thinking about linear regression in terms of **matrices and vectors** will allow us to find hypothesis functions that:
  - Use multiple features (input variables).
  - Are nonlinear in the features, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.

Wait... why do we need linear algebra?

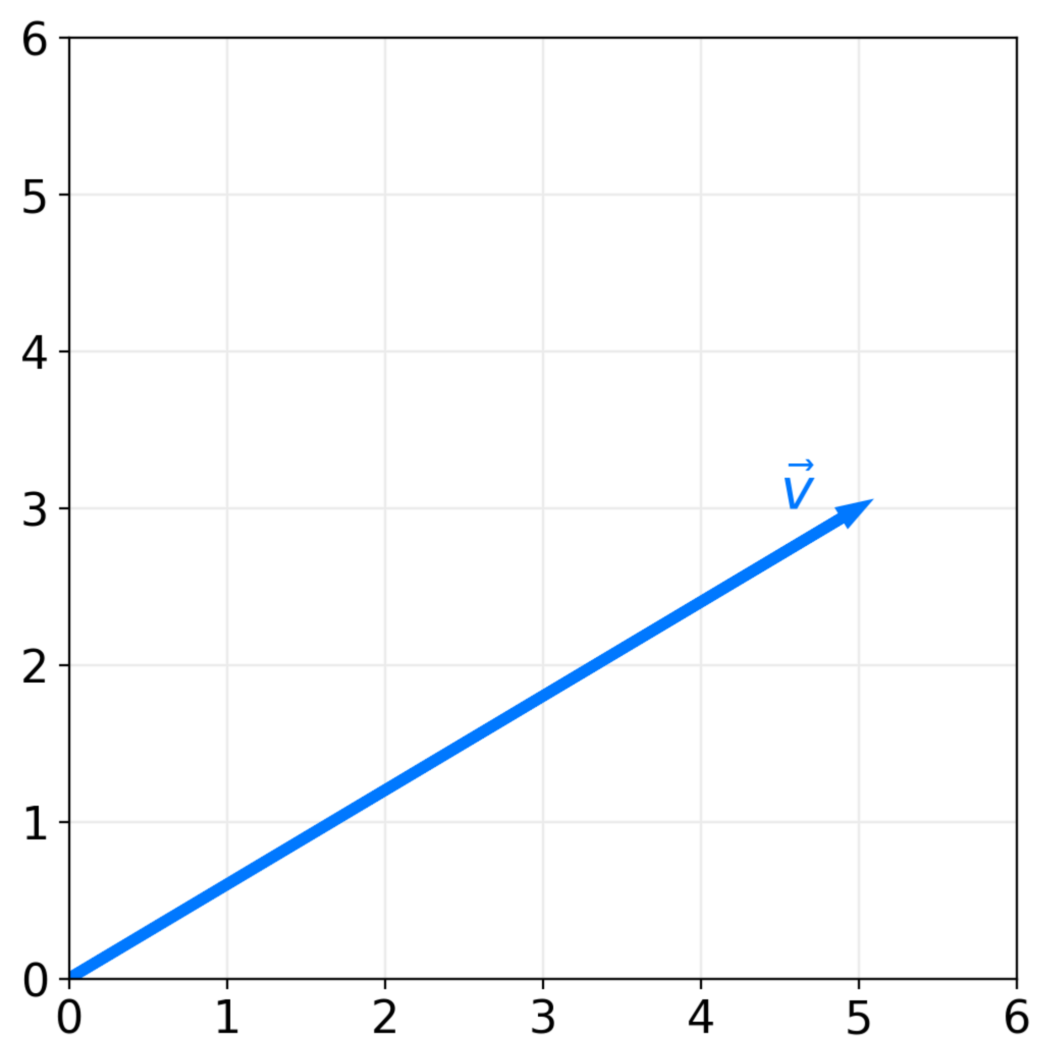
- Soon, we'll want to make predictions using more than one feature.
  - Example: Predicting commute times using departure hour and temperature.
- Thinking about linear regression in terms of **matrices and vectors** will allow us to find hypothesis functions that:
  - Use multiple features (input variables).
  - Are nonlinear in the features, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.
- Before we dive in, let's do a quick knowledge assessment.
- Go to https://forms.gle/LXBXydpsX8rtJQPz7

1:40

Question 1: Norm

What is the length of $\vec{v}$?

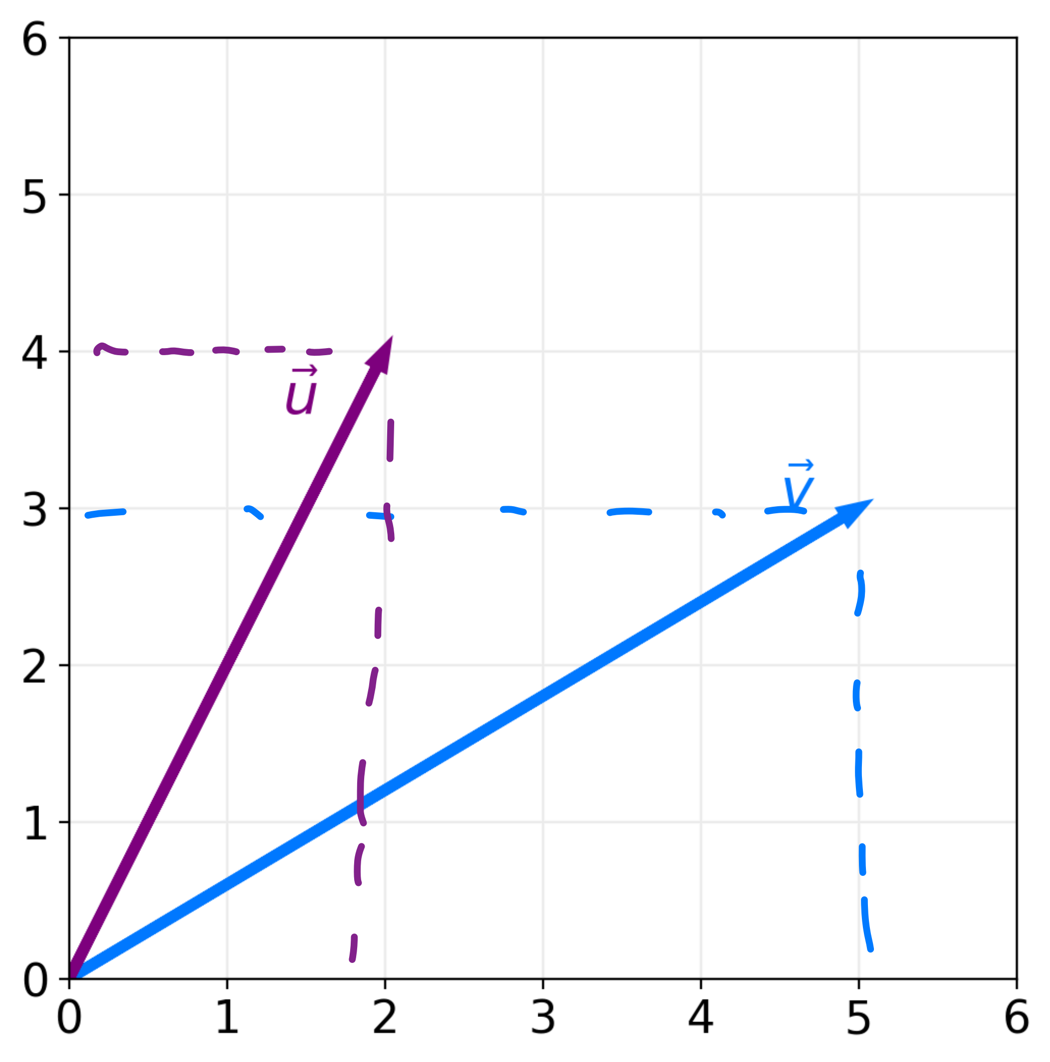
- A. 8
- B. $\sqrt{34}$
- C. $\sqrt{38}$
- D. 34



Question 2: Dot product

What is $\vec{u} \cdot \vec{v}$?

- A. 22
- B. 24
- C. $\sqrt{680}$
- D. $\begin{bmatrix} 10 \\ 12 \end{bmatrix}$



Question 3: Norm

Which of these is another expression for the length of $\vec{v}$?

- A. $\vec{v} \cdot \vec{v}$ $= \|\vec{v}\|^2$
- B. $\sqrt{\vec{v}^2}$
- C. $\sqrt{\vec{v} \cdot \vec{v}}$
- D. $\vec{v}^2$
- E. More than one of the above.

$$= \|\vec{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2} = \sqrt{\sum_{i=1}^{n} v_i v_i}$$

$$= \sqrt{\vec{v} \cdot \vec{v}}$$

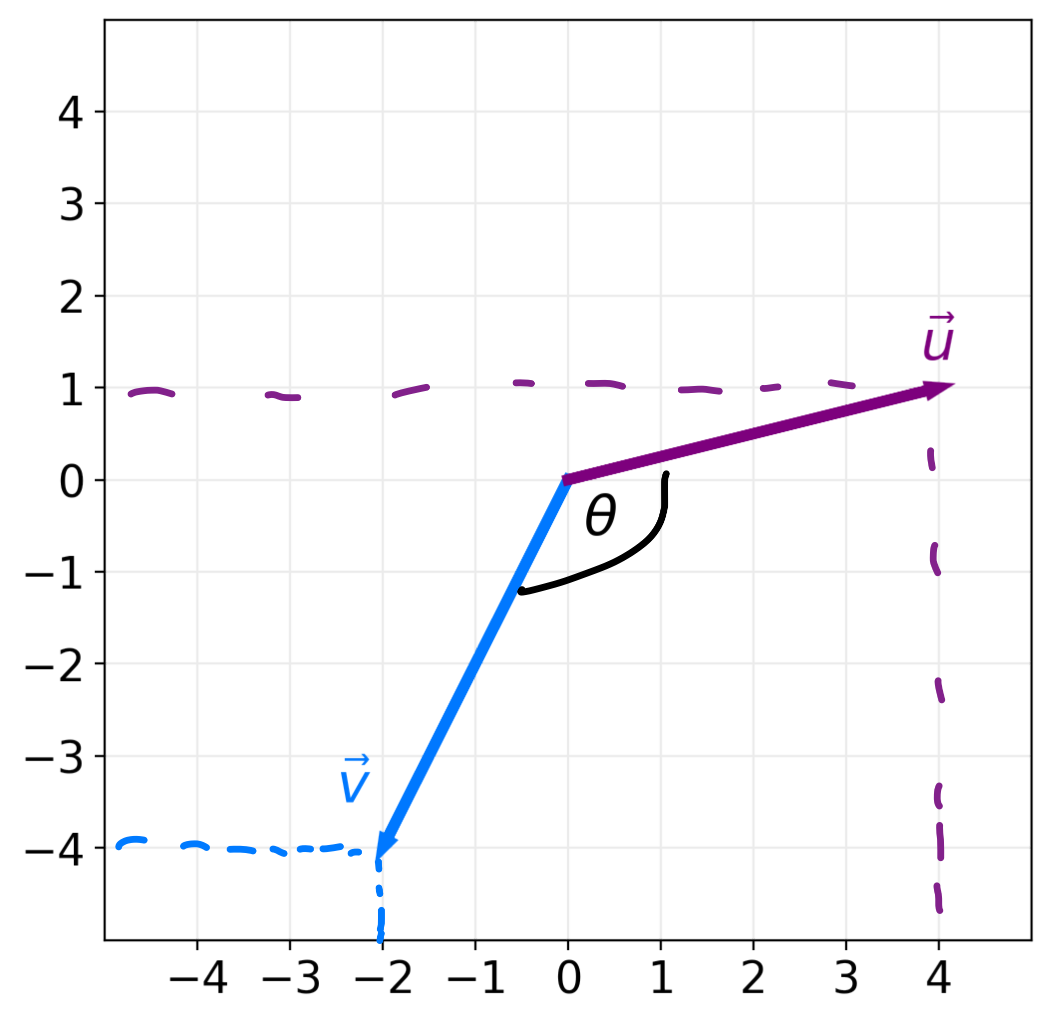
$\vec{v}^2$ is undefined → you can square a vector

this is not equal to $\vec{v} \cdot \vec{v}$ and it is not equal to element-wise squaring

Question 4: $\cos\theta$

What is $\cos\theta$?

- A. $\frac{6}{\sqrt{85}}$
- B. $\frac{-6}{\sqrt{85}}$
- C. $\frac{-3}{85}$
- D. $\frac{-2}{3}$



Question 5: Orthogonality

Which of these vectors in $\mathbb{R}^3$ orthogonal to:

$$\vec{v} = \begin{bmatrix} 2 \\ 5 \\ -8 \end{bmatrix}?$$

- A. $\begin{bmatrix} -2 \\ -5 \\ 8 \end{bmatrix}$
- B. $\begin{bmatrix} 5 \\ -8 \\ 2 \end{bmatrix}$
- C. $\begin{bmatrix} 8 \\ 0 \\ 2 \end{bmatrix}$
- D. All of the above

Warning ⚠️

- We're **not** going to cover every single detail from your linear algebra course.
- There will be facts that you're expected to remember that we won't explicitly say.
  - For example, if A and B are two matrices, then $AB \neq BA$.
  - This is the kind of fact that we will only mention explicitly if it's directly relevant to what we're studying.
  - But you still need to know it, and it may come up in homework questions.
- We **will** review the topics that you really need to know well.

(video on course website 3blue1brown)

Dot Products

Vectors

- A **vector** in $\mathbb{R}^n$ is an **ordered collection of n numbers**.
- We use lower-case letters with an arrow on top to represent vectors, and we usually write vectors as **columns**.

$$\vec{v} = \begin{bmatrix} 8 \\ 3 \\ -2 \\ 5 \end{bmatrix}$$

In general
$\vec{v} \in \mathbb{R}^n$
$n \times 1$

- Another way of writing the above vector is $\vec{v} = [8, 3, -2, 5]^T$. → transpose
- Since $\vec{v}$ has four **components**, we say $\vec{v} \in \mathbb{R}^4$.

4×1

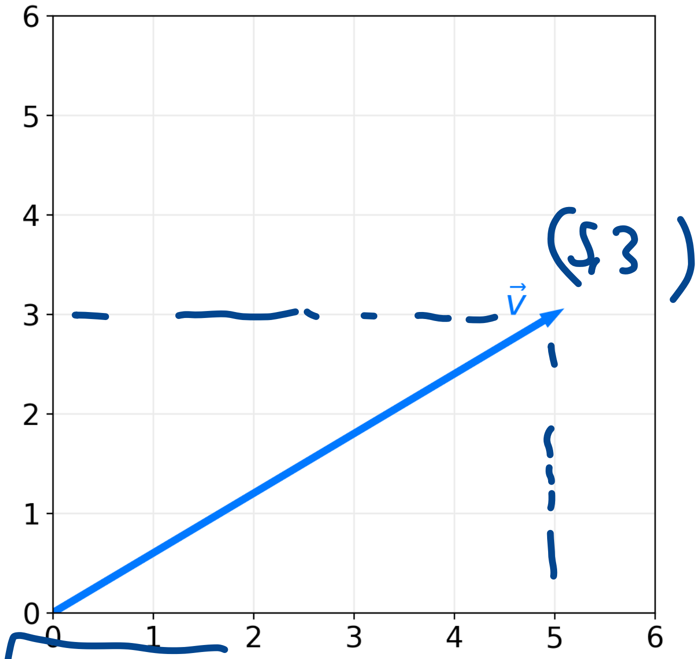
The geometric interpretation of a vector

- A vector $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ is an arrow to the point $(v_1, v_2, \ldots, v_n)$ from the origin.
- The **length**, or L_2 **norm**, of $\vec{v}$ is:

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2} = \sqrt{\sum_{i=1}^{n} v_i^2} = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\vec{v}^T \vec{v}}$$

- A vector is sometimes described as an object with a **magnitude/length** and **direction**.

$$\|\vec{v}\| = \sqrt{5^2 + 3^2} = \sqrt{25 + 9} = \sqrt{34}$$



Dot product: coordinate definition

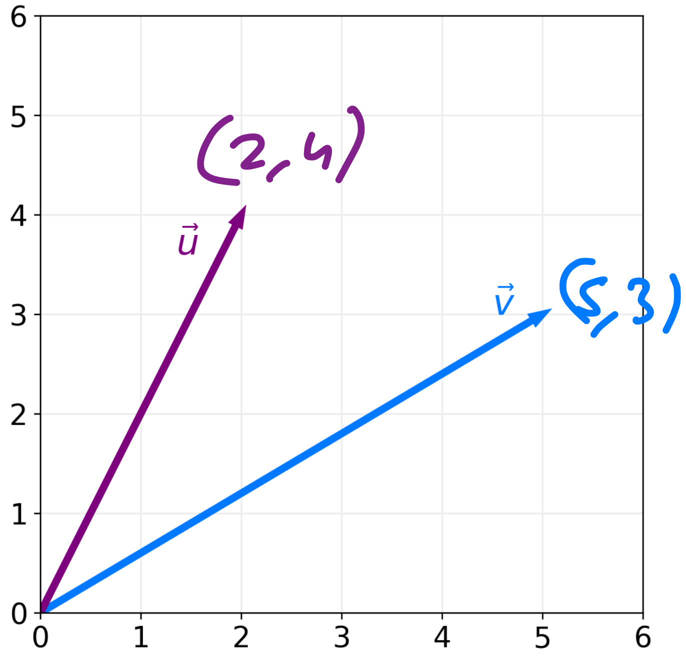
- The **dot product** of two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ is written as:

$$\vec{u} \cdot \vec{v} = \vec{u}^\top \vec{v} = \langle \vec{u}, \vec{v} \rangle$$

- The computational definition of the dot product:

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

- The result is a **scalar**, i.e. a single number.



$$\vec{u} \cdot \vec{v} = 5 \cdot 2 + 3 \cdot 4 = 10 + 12 = 22 \in \mathbb{R}$$

$$\vec{u}^\top \vec{v} = \begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \end{bmatrix} = 10 + 12 = 22 \in \mathbb{R}$$

$$f(\vec{u}, \vec{v}) \in \mathbb{R}$$

$\mathbb{R}^n$ $\mathbb{R}^n$