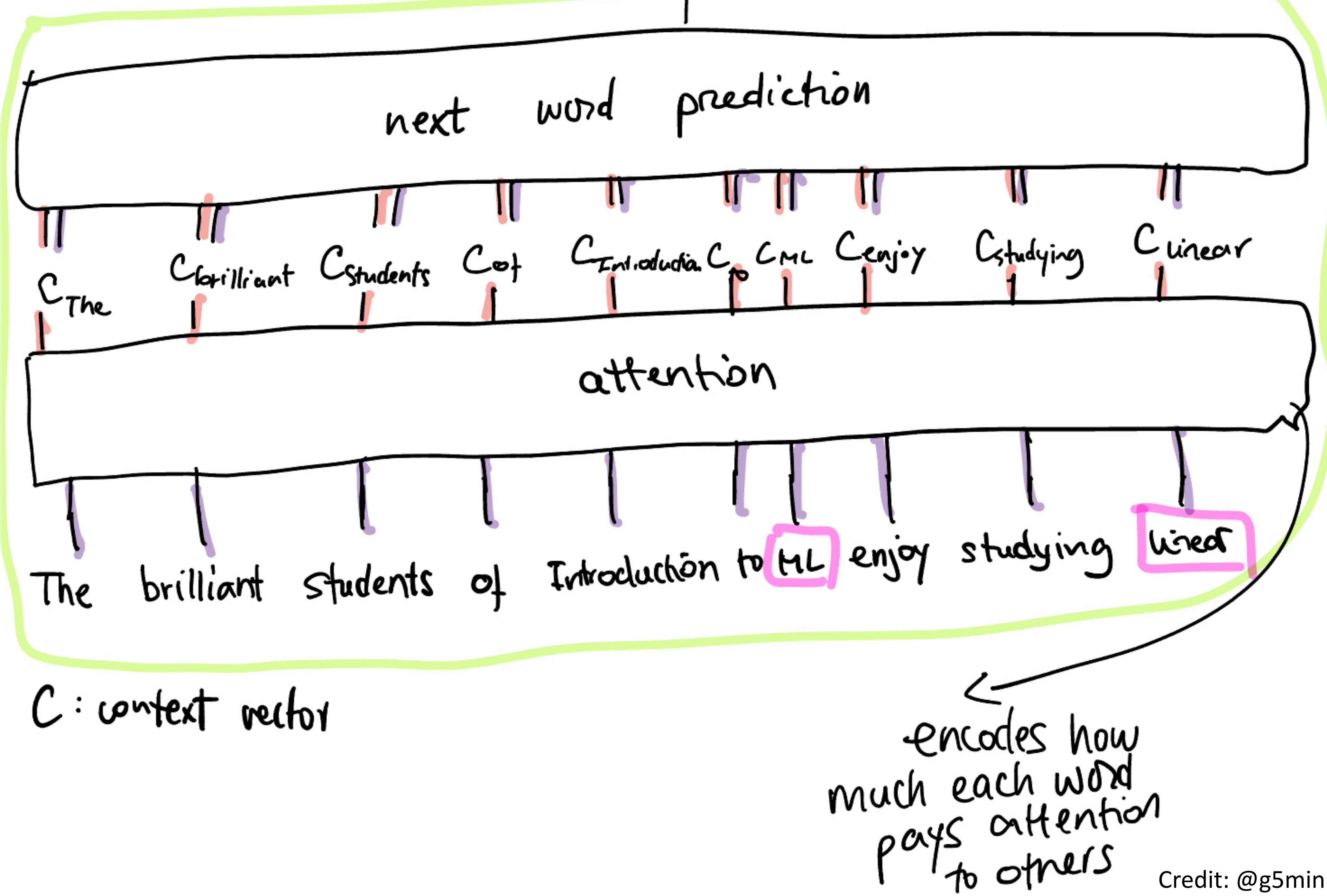
Analytic Geometry 1

Jo Ciucă
Australian National University
comp36706670@anu.edu.au

Attention is all you need

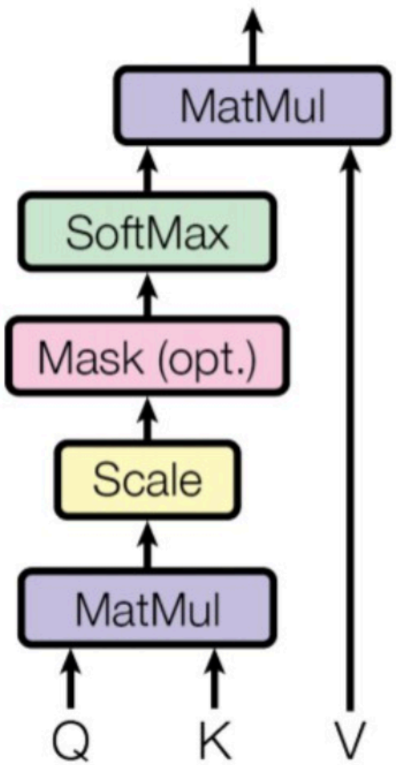
The **brilliant students** of Introduction to **ML** enjoy **studying** **linear algebra.**

Causal Language Modelling Objective



Credit: @g5min

Linear Algebra 😍



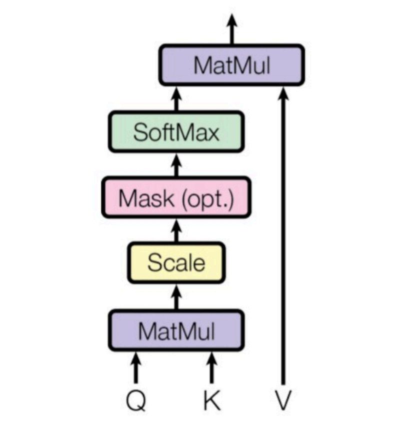
Key

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

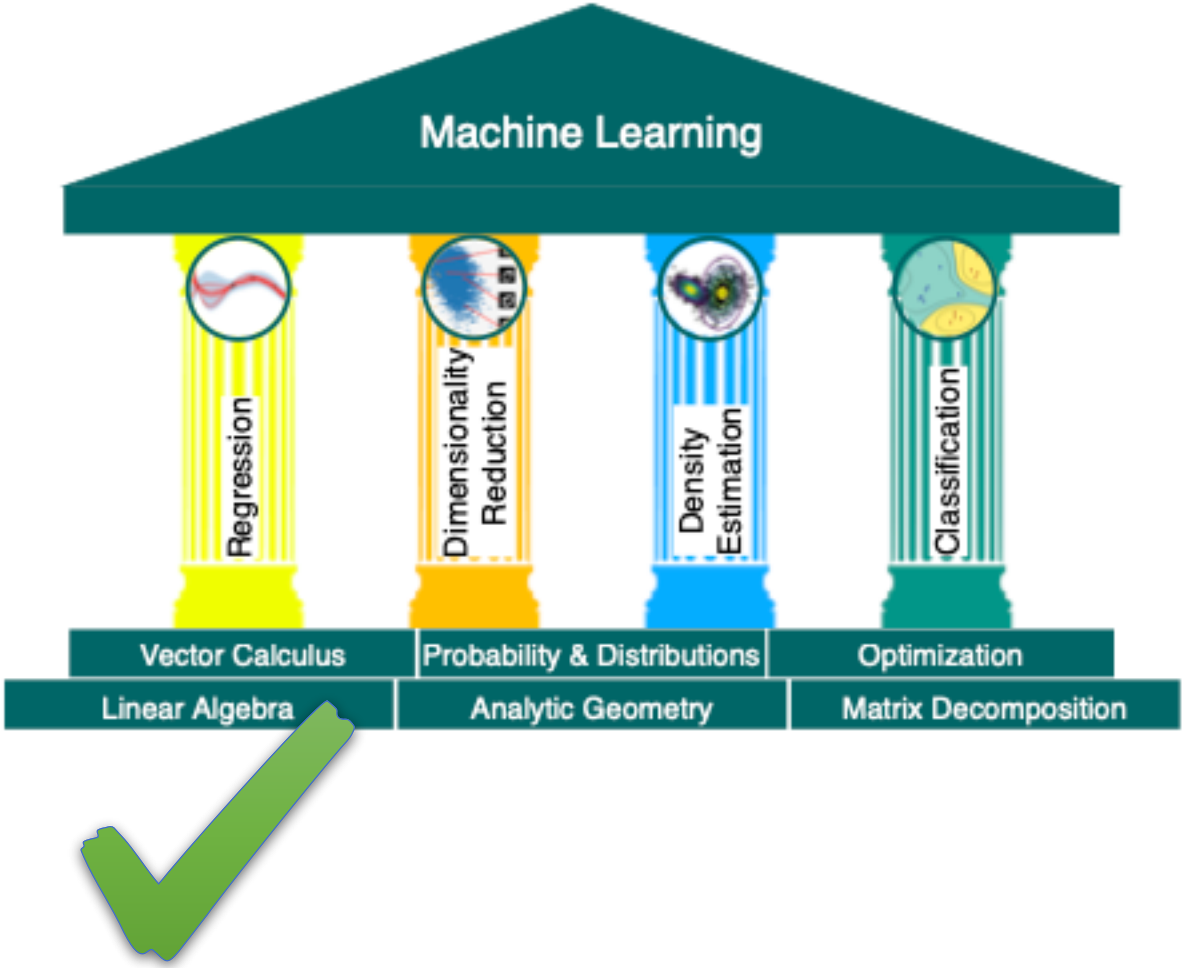
Query

Value

```python
import torch
import torch.nn.functional as F

def scaled_dot_product_attention(query, key, value):
    # Calculate the dot product between query and key
    scores = torch.matmul(query, key.transpose(-2, -1))

    # Scale the scores by square root of 'dk' (the dimension of the keys)
    dk = key.size(-1)  # get the size of the key's last dimension
    scaled_scores = scores / torch.sqrt(torch.tensor(dk).float())

    # Apply the softmax function to the scaled scores to get the attention weights
    attention_weights = F.softmax(scaled_scores, dim=-1)

    # Multiply the weights by the value vectors to get the output
    output = torch.matmul(attention_weights, value)

    return output, attention_weights

# Test the function
device = "cuda" if torch.cuda.is_available() else "cpu"
query = torch.randn(3, 8, device=device)
key = torch.randn(3, 8, device=device)
value = torch.randn(3, 8, device=device)

output, attention_weights = scaled_dot_product_attention(query, key, value)
print("Output shape: ", output.shape)
print("Attention weights shape: ", attention_weights.shape)
```



Thank you, next.



Check your understanding

- Which of the following statements is correct?

(A) In a vector space, any vector can be represented as a linear combination of a certain set of vectors in this space.

(B) The dimension of a vector equals the dimension of the space it is in.

(C) U is a vector subspace of V. Then vectors in U have lower dimension than vectors in V.

(D) Set $\left\{ \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 7 \\ 7 \\ -1 \end{bmatrix} \right\}$ forms a basis for $\mathbb{R}^3$.

(E) $U = \{(x, y) : x = y,\ x \in \mathbb{R},\ y \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^2$.

(F) The vector $\mathbf{0}$ is linearly dependent on any vector in the same vector space.

Outline

- Bilinear Mappings
- Inner Product
- Lengths & distances
- Angles & Orthogonality

3.1 Norms

- A norm on a vector space V is a function

$$\|\bullet\| : V \to \mathbb{R},$$

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|,$$

which assigns each vector $\boldsymbol{x}$ its length $\|\boldsymbol{x}\| \in \mathbb{R}$.

Examples

- The Manhattan norm on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

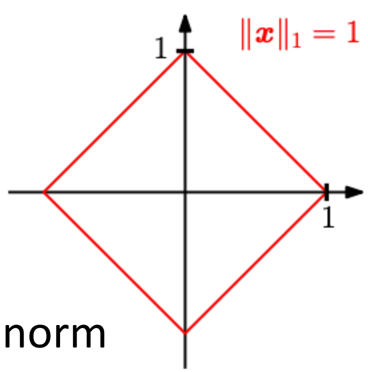
$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|,$$

where $|\bullet|$ is the absolute value. It is also called ℓ_1 norm.

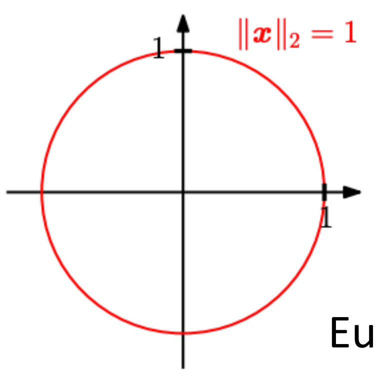
- The Euclidean norm of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$$

It is the Euclidean distance of $\boldsymbol{x}$ from the origin; also called ℓ_2 norm



Manhattan norm

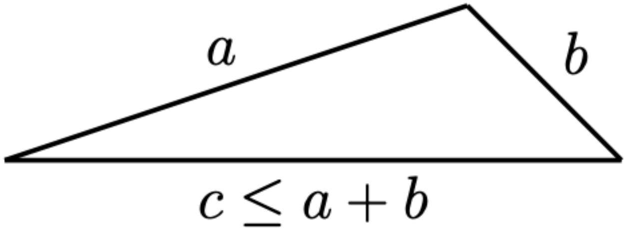


Euclidean norm

3.1 Norms

For all $\lambda \in \mathbb{R}$, and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following holds:

- Absolutely homogeneous: $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
- Triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- Positive definite: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$



3.2. Inner products

Dot Product

- Scalar product/dot product in $\mathbb{R}^n$ is given by

$$\underset{1\times n}{\boldsymbol{x}^{\mathrm{T}}}\underset{\mathrm{n}\times 1}{\boldsymbol{y}} = \sum_{i=1}^{n} x_i y_i$$

Bilinear mapping

- A bilinear mapping Ω is a mapping with two arguments, and it is linear in each argument. Consider a vector space V, for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V,\ \lambda, \varphi \in \mathbb{R}$,

$$\Omega(\lambda \boldsymbol{x} + \varphi \boldsymbol{y},\ \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{z}) + \varphi\Omega(\boldsymbol{y}, \boldsymbol{z})$$

Ω is linear in the first argument

$$\Omega(\boldsymbol{x}, \lambda \boldsymbol{y} + \varphi \boldsymbol{z}) = \lambda\Omega(\boldsymbol{x}, \boldsymbol{y}) + \varphi\Omega(\boldsymbol{x}, \boldsymbol{z})\,.$$

Ω is linear in the second argument

Inner product

- Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ be a bilinear mapping.
- Ω is called symmetric if $\Omega(\boldsymbol{x}, \boldsymbol{y}) = \Omega(\boldsymbol{y}, \boldsymbol{x})$
- Ω is called positive definite if

$$\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \Omega(\boldsymbol{x}, \boldsymbol{x}) > 0, \ \Omega(\boldsymbol{0}, \boldsymbol{0}) = 0$$

- A positive definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an inner product on V. We write $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ instead of $\Omega(\boldsymbol{x}, \boldsymbol{y})$.
- The pair $(V, \langle \bullet, \bullet \rangle)$ is called is called an inner product vector space. If we use the dot product, we call $(V, \langle \bullet, \bullet \rangle)$ a Euclidean vector space.

Example

- Consider $V = \mathbb{R}^2$. If we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := x_1 y_1 - \left(x_1 y_2 + x_2 y_1\right) + 2 x_2 y_2$$

- then $\langle \bullet, \bullet \rangle$ is an inner product but different from the dot product.

This mapping is symmetric: it is easy to derive $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$
Is it positive definite?
$\forall \boldsymbol{x} \in V \backslash \{\boldsymbol{0}\}, \langle \boldsymbol{x}, \boldsymbol{x} \rangle = x_1^2 - \left(x_1 x_2 + x_2 x_1\right) + 2 x_2^2 = \left(x_1 - x_2\right)^2 + x_2^2 > 0$

3.2.3 Symmetric, Positive Definite Matrices

- Consider an n-dimensional vector space V with an inner product $\langle \bullet , \bullet \rangle$: $V \times V \to \mathbb{R}$, and a basis $B = \left(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n\right)$ of V.

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \left\langle \sum_{i=1}^{n} \varphi_i \boldsymbol{b}_i, \sum_{j=1}^{n} \lambda_j \boldsymbol{b}_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_i \left\langle \boldsymbol{b}_i, \boldsymbol{b}_j \right\rangle \lambda_j = \hat{\boldsymbol{x}}^T \boldsymbol{A} \hat{\boldsymbol{y}}$$

where $A_{ij} := \left\langle \boldsymbol{b}_i, \boldsymbol{b}_j \right\rangle$ and $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$ are the coordinates of $\boldsymbol{x}, \boldsymbol{y}$ with respect to the basis $\boldsymbol{B}$.

- The inner product $\langle \bullet , \bullet \rangle$ is uniquely determined through $\boldsymbol{A}$. The symmetry of the inner product also means that $\boldsymbol{A}$ is symmetric.
- The positive definiteness of the inner product implies that

$$\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} > 0$$

3.2.3 Symmetric, Positive Definite Matrices

- A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ that satisfies $\forall \boldsymbol{x} \in V \setminus \{\boldsymbol{0}\} : \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} > 0$ is called symmetric, positive definite, or just positive definite. If only $\geq$ holds, then $\boldsymbol{A}$ is called symmetric, positive semidefinite.
- Example

$$\boldsymbol{A}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 1 & 3 \\ 3 & 3 \end{bmatrix}$$

- $\boldsymbol{A}_1$ is positive definite because it is symmetric and

$$\boldsymbol{x}^T \boldsymbol{A}_1 \boldsymbol{x} = [x_1 \quad x_2] \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 3x_1^2 + 2x_1x_2 + 4x_2^2 = (x_1 + x_2)^2 + 2x_1^2 + 3x_2^2 > 0$$

for all $\boldsymbol{x} \in V \setminus \{\boldsymbol{0}\}$.

- $\boldsymbol{A}_2$ is symmetric but not positive definite

$\boldsymbol{x}^T \boldsymbol{A}_2 \boldsymbol{x} = x_1^2 + 6x_1x_2 + 3x_2^2 = (x_1 + 3x_2)^2 - 6x_2^2$ can be less than 0

3.2.3 Symmetric, Positive Definite Matrices

- For a real-valued, finite-dimensional vector space V and a basis B of V, it holds that $\langle \bullet, \bullet \rangle: V \times V \to \mathbb{R}$ is an inner product if and only if there exists a symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^{T} \boldsymbol{A} \hat{\boldsymbol{y}}$$

- If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric and positive definite,

  the diagonal elements a_{ii} of $\boldsymbol{A}$ are positive because $a_{ii} = \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{A} \boldsymbol{e}_i = \langle \boldsymbol{e}_i, \boldsymbol{e}_i \rangle > 0$, where $\boldsymbol{e}_i$ is the ith vector of the standard basis in $\mathbb{R}^n$.

3.3 Lengths and Distances

- Any inner product induces a norm

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

- Cauchy-Schwarz Inequality
- For an inner product vector space $(V, \langle \bullet, \bullet \rangle)$ the induced norm $\| \bullet \|$ satisfies the Cauchy-Schwarz inequality

$$\left| \langle \boldsymbol{x}, \boldsymbol{y} \rangle \right| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

Example - Lengths of Vectors Using Inner Products

- We can now use an inner product to compute vector lengths, using $\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. Consider $\boldsymbol{x} = [1,1]^{\mathrm{T}} \in \mathbb{R}^2$. If we use the dot product as the inner product, we obtain

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$\boldsymbol{x}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}$ is dot product

as the length of $\boldsymbol{x}$. Let us now choose a different inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^T \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \boldsymbol{y} = x_1 y_1 - \frac{1}{2}(x_1 y_2 + x_2 y_1) + x_2 y_2$$

With this inner product, we obtain

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = x_1^2 - x_1 x_2 + x_2^2 = 1 - 1 + 1 = 1 \implies \|\boldsymbol{x}\| = \sqrt{1} = 1$$

$\boldsymbol{x}$ is "shorter" with this inner product than with the dot product.

3.3 Lengths and Distances

- Consider an inner product space $\left(V, \langle \bullet, \bullet \rangle\right)$, then

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) := \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$$

  is called the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y} \in V$.
- If we use the dot product as the inner product, then the distance is called Euclidean distance.

3.3 Lengths and Distances

- The mapping
$$d : V \times V \to \mathbb{R}$$
$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto d(\boldsymbol{x}, \boldsymbol{y})$$
is called a metric.
- A metric d satisfies the following:
- d is positive definite, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$ and $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
- d is symmetric, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$
- Triangle inequality: $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$
- Very similar $\boldsymbol{x}$ and $\boldsymbol{y}$ will result in a large value for the inner product and a small value for the metric.

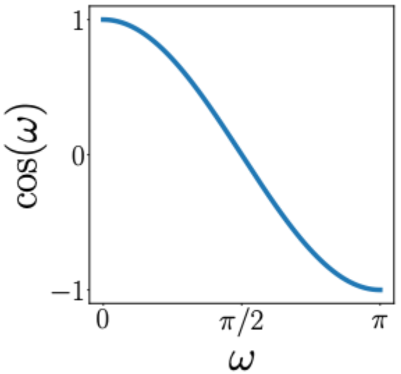
3.4 Angles and Orthogonality $\left|\langle \boldsymbol{x}, \boldsymbol{y}\rangle\right| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$

- According to Cauchy-Schwarz inequality, assume $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{y} \neq \boldsymbol{0}$. Then,

$$-1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \leq 1$$

Therefore, there exists a unique $\omega \in [0,\pi]$, with

$$cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$



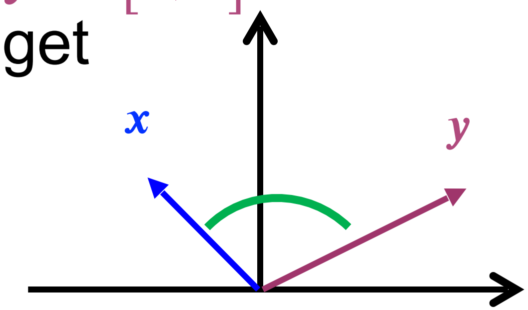
The number ω is the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

- The angle between two vectors tells us how similar their orientations are.
- Using the dot product, the angle between $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{4x}$ is 0, so their orientation is the same.

$$\boxed{cos\omega = \frac{\langle \boldsymbol{x}, 4\boldsymbol{x}\rangle}{\|\boldsymbol{x}\|\|4\boldsymbol{x}\|} = \frac{4\langle \boldsymbol{x}, \boldsymbol{x}\rangle}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}}\sqrt{(4\boldsymbol{x})^T(4\boldsymbol{x})}} = \frac{4\langle \boldsymbol{x}, \boldsymbol{x}\rangle}{4\|\boldsymbol{x}\|\|\boldsymbol{x}\|} = \frac{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{x}\|}}$$

Example (Angle between Vectors)

- Let us compute the angle between $\boldsymbol{x} = [-1,1]^{\mathrm{T}} \in \mathbb{R}^2$ and $\boldsymbol{y} = [2,1]^{\mathrm{T}} \in \mathbb{R}^2$. We use the dot product as the inner product. We get



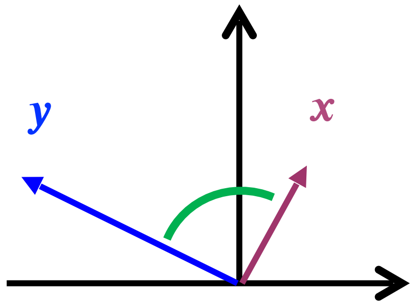
$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle \langle \boldsymbol{y}, \boldsymbol{y} \rangle}} = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\sqrt{\boldsymbol{x}^T \boldsymbol{x} \boldsymbol{y}^T \boldsymbol{y}}} = \frac{-1}{\sqrt{10}}$$

- and the angle between the two vectors is $\arccos\left(\frac{-1}{\sqrt{10}}\right) \approx 1.89\text{rad}$, which corresponds to about 108.4°.

- We then use inner product to characterize orthogonality.

3.4 Angles and Orthogonality

- Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, and we write $\boldsymbol{x} \perp \boldsymbol{y}$. If additionally $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$, i.e., the vectors are unit vectors, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthonormal.
- $\boldsymbol{0}$-vector is orthogonal to every vector in the vector space



- Example (Orthogonal Vectors)
- Consider $\boldsymbol{x} = [1,2]^{\mathrm{T}}$ and $\boldsymbol{y} = [-4, 2]^{\mathrm{T}}$
- Using dot product as inner product, we have
  - $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, so $\boldsymbol{x} \perp \boldsymbol{y}$.
- if we choose the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \boldsymbol{y}$$

- the angle ω between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \left| \boldsymbol{y} \right|} = -\frac{2}{\sqrt{17 \times 12}} \qquad \Longrightarrow \qquad \omega \approx 1.43\text{rad} \approx 81.95^{\circ}$$

3.4 Angles and Orthogonality

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if and only if its columns are orthonormal, such that

$$\boldsymbol{A}\,\boldsymbol{A}^T = \boldsymbol{I} = \boldsymbol{A}^T\boldsymbol{A}$$

which implies that

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

i.e., the inverse is obtained by simply transposing the matrix

Properties - length

- The length of a vector $\boldsymbol{x}$ is not changed when transforming it using an orthogonal matrix $\boldsymbol{A}$. For dot product, we obtain

$$\|\boldsymbol{A}\boldsymbol{x}\|^{2} = (\boldsymbol{A}\boldsymbol{x})^{T}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^{T}\boldsymbol{A}^{T}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^{T}\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^{T}\boldsymbol{x} = \|\boldsymbol{x}\|^{2}$$

Properties - angle

- The angle between any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as measured by their inner product, is also unchanged when transforming both of them using an orthogonal matrix $\boldsymbol{A}$. We use the dot product as inner product

$$cos\omega = \frac{(\boldsymbol{Ax})^T(\boldsymbol{Ay})}{\|\boldsymbol{Ax}\|\|\boldsymbol{Ay}\|} = \frac{\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ay}}{\sqrt{\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Axy}^T\boldsymbol{A}^T\boldsymbol{Ay}}} = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$

- Orthogonal matrices $\boldsymbol{A}$ with $\boldsymbol{A}^{-1} = \boldsymbol{A}^{T}$ preserve both angles and distances.

- Orthogonal matrices define transformations that are rotations.