CSCI 567 Discussion: Linear Algebra Review I

Jan 16, 2026

(Slides adapted from Nandita Bhaskhar's slides for CS229 at Stanford)

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

Vector space

Set V of vectors equipped with scaling and addition operations, satisfying nice properties, e.g.,

$$1 \cdot v = v$$

$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$$

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$$

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

Vector space

Set V of vectors equipped with scaling and addition operations, satisfying nice properties, e.g.,

$$1 \cdot v = v$$

$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$$

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$$

We will consider vectors $u, v \in \mathbb{R}^d$ and scalars $\alpha, \beta \in \mathbb{R}$.

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

Vector space

Set V of vectors equipped with scaling and addition operations, satisfying nice properties, e.g.,

$$1 \cdot v = v$$

$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$$

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$$

We will consider vectors $u, v \in \mathbb{R}^d$ and scalars $\alpha, \beta \in \mathbb{R}$. Then each vector takes the form

$$v = (v_1, \ldots, v_d),$$

and addition & scaling are entrywise:

$$u + v = (u_1 + v_1, \ldots, u_d + v_d)$$

$$\alpha \cdot v = (\alpha \cdot v_1, \ldots, \alpha \cdot v_d)$$

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

Vector space

Set V of vectors equipped with scaling and addition operations, satisfying nice properties, e.g.,

$$1 \cdot v = v$$

$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$$

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$$

We will consider vectors $u, v \in \mathbb{R}^d$ and scalars $\alpha, \beta \in \mathbb{R}$. Then each vector takes the form

$$v = (v_1, \ldots, v_d),$$

and addition & scaling are entrywise:

$$u + v = (u_1 + v_1, \ldots, u_d + v_d)$$

$$\alpha \cdot v = (\alpha \cdot v_1, \ldots, \alpha \cdot v_d)$$

Linear functions

A linear function $T: \mathbb{R}^d \to \mathbb{R}^k$ is a function satisfying:

1. $T(u + v) = T(u) + T(v)$
2. $T(\alpha \cdot v) = \alpha \cdot T(v)$

What is Linear Algebra?

Linear Algebra is the study of *vector spaces* and *linear functions*.

Vector space

Set V of vectors equipped with scaling and addition operations, satisfying nice properties, e.g.,

$$1 \cdot v = v$$

$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$$

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$$

We will consider vectors $u, v \in \mathbb{R}^d$ and scalars $\alpha, \beta \in \mathbb{R}$. Then each vector takes the form

$$v = (v_1, \ldots, v_d),$$

and addition & scaling are entrywise:

$$u + v = (u_1 + v_1, \ldots, u_d + v_d)$$

$$\alpha \cdot v = (\alpha \cdot v_1, \ldots, \alpha \cdot v_d)$$

Linear functions

A linear function $T\colon \mathbb{R}^d \to \mathbb{R}^k$ is a function satisfying:

1. $T(u + v) = T(u) + T(v)$
2. $T(\alpha \cdot v) = \alpha \cdot T(v)$

Key idea: A linear function is determined by where it maps the vectors $(1,0,\ldots,0)$, $(0,1,\ldots,0)$, $(0,\ldots,0,1)$. For instance,

$$\begin{aligned} T((2,3)) &= T((2,0)) + T((0,3)) \\ &= 2 \cdot T((1,0)) + 3 \cdot T((0,1)) \end{aligned}$$

Basic Notation

By $x \in \mathbb{R}^n$, we denote a vector with n entries.

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

We denote by e_i the vector with 1 in the ith position and 0 elsewhere, e.g.,

$$e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Basic Notation

By $x \in \mathbb{R}^n$, we denote a vector with n entries.

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

We denote by e_i the vector with 1 in the ith position and 0 elsewhere, e.g.,

$$e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

By $A \in \mathbb{R}^{m \times n}$, we denote a matrix with m rows and n columns.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} --- & a_1^T & --- \\ --- & a_2^T & --- \\ & \vdots & \\ --- & a_m^T & --- \end{bmatrix}$$

Matrices

Key point: The matrix $A \in \mathbb{R}^{m \times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i \leq m} A_{ij} e_i$$

In English: the *i*th column of A is the image of the i*th* basis vector.

Matrices

Key point: The matrix $A \in \mathbb{R}^{m \times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i \leq m} A_{ij} e_i$$

In English: the *i*th column of A is the image of the *i*th basis vector.

I think of it like a system of pipes: copies of basis vectors go in, and copies of basis vectors go out.

$$A = \begin{array}{cc} & \begin{array}{cccc} e_1 & e_2 & \cdots & e_n \\ \downarrow & \downarrow & & \downarrow \end{array} \\ \begin{array}{c} e_1 \leftarrow \\ e_2 \leftarrow \\ \vdots \\ e_m \leftarrow \end{array} & \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix} \end{array}$$

Matrices

Key point: The matrix $A \in \mathbb{R}^{m \times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i \leq m} A_{ij} e_i$$

In English: the *i*th column of *A* is the image of the i*th* basis vector.

Consider the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$:

Matrices

Key point: The matrix $A \in \mathbb{R}^{m \times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i \leq m} A_{ij} e_i$$

In English: the *i*th column of *A* is the image of the i*th* basis vector.

Consider the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$:

- The first column says: "Turn each copy of e_1 into a copies of e_1 and c copies of e_2."

Matrices

Key point: The matrix $A \in \mathbb{R}^{m \times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i \leq m} A_{ij} e_i$$

In English: the *i*th column of *A* is the image of the *i*th basis vector.

Consider the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$:

- The first column says: "Turn each copy of e_1 into a copies of e_1 and c copies of e_2."
- The second column says: "Turn each copy of e_2 into b copies of e_1 and d copies of e_2."

Matrices

Key point: The matrix $A \in \mathbb{R}^{m\times n}$ concisely represents the linear function $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by

$$T(e_j) = \sum_{i\leq m} A_{ij} e_i$$

In English: the *i*th column of *A* is the image of the i*th* basis vector.

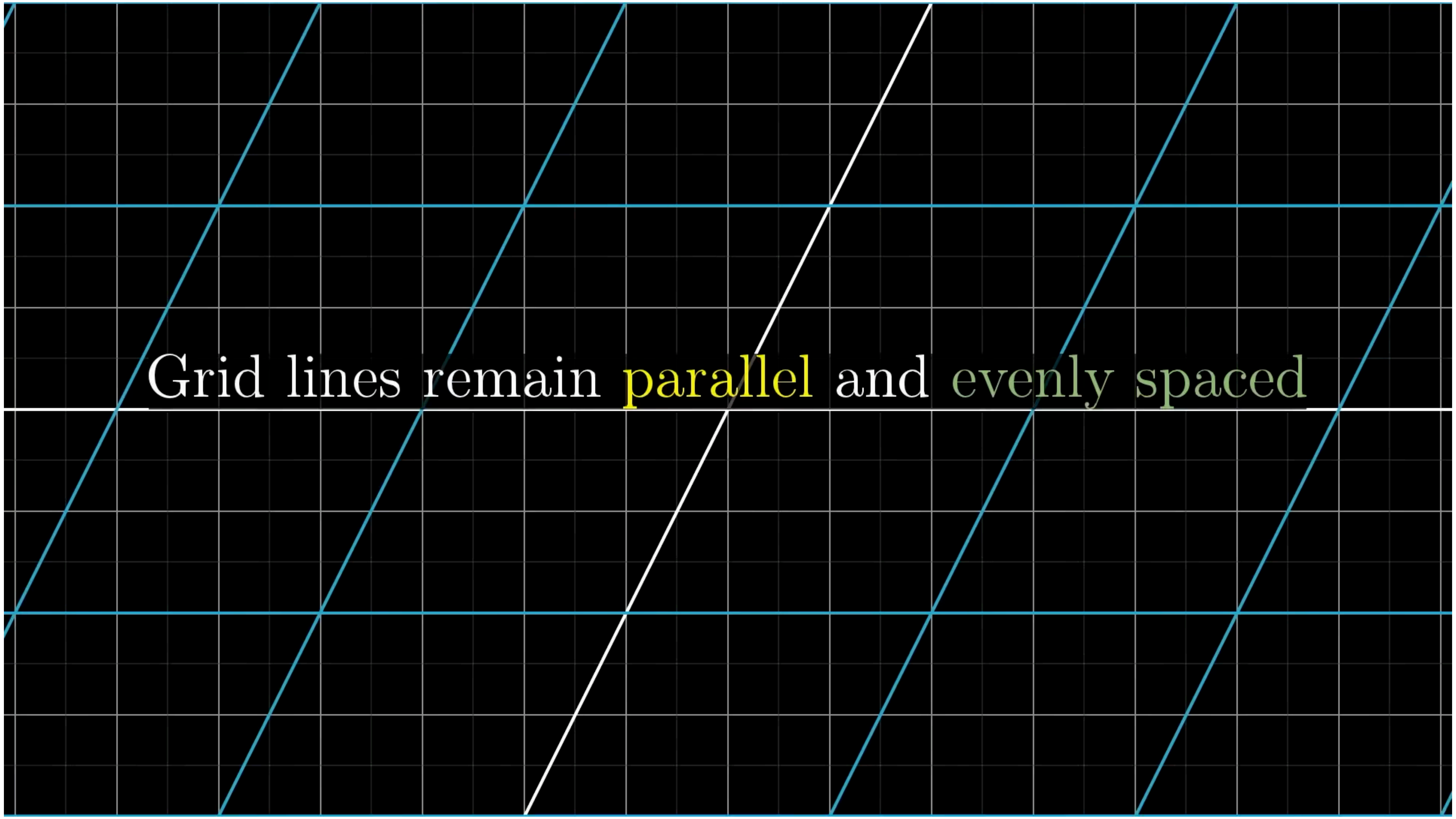
Consider the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$:

- The first column says: "Turn each copy of e_1 into a copies of e_1 and c copies of e_2."
- The second column says: "Turn each copy of e_2 into b copies of e_1 and d copies of e_2."

That's it — now you understand matrices!

Matrices

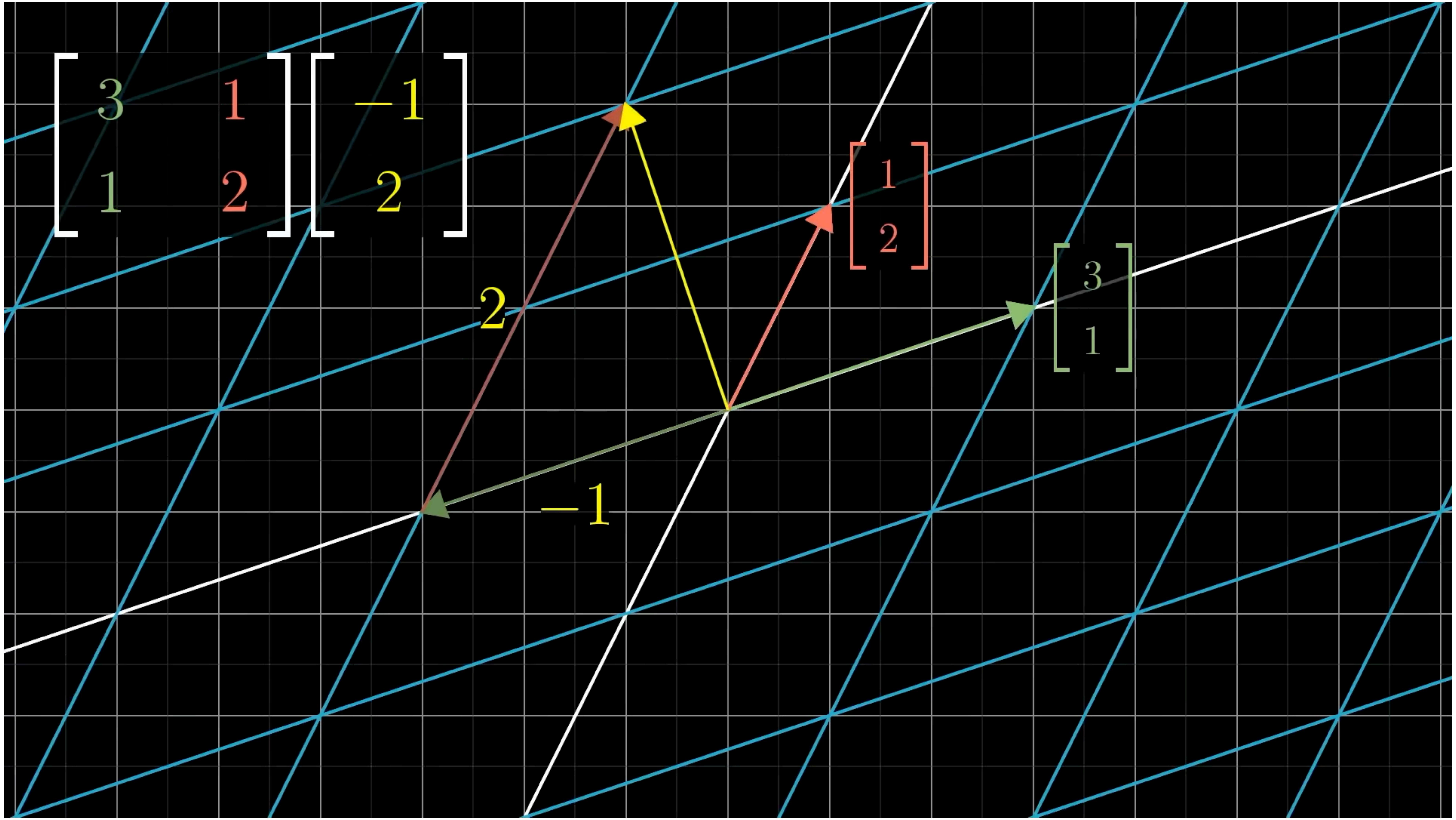
Visualization from 3Blue1Brown, Essence of linear algebra (**3 min**)



Concrete Examples

Matrices

Visualization from 3Blue1Brown, Essence of linear algebra (**1 min**)



Matrix Multiplication

Recall that a matrix $A \in \mathbb{R}^{m \times n}$ is a concise representation of a linear function $T_A \colon \mathbb{R}^n \to \mathbb{R}^m$.

Matrix multiplication is **defined** so that $A \times B$ represents the linear function $T_A \circ T_B$, when this composition is legal. (I.e., when the dimension of B's output equals that of A's input.)

Matrix Multiplication

Recall that a matrix $A \in \mathbb{R}^{m \times n}$ is a concise representation of a linear function $T_A \colon \mathbb{R}^n \to \mathbb{R}^m$.

Matrix multiplication is **defined** so that $A \times B$ represents the linear function $T_A \circ T_B$, when this composition is legal. (I.e., when the dimension of B's output equals that of A's input.)

Formally, for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, the matrix product $C = AB \in \mathbb{R}^{m \times p}$ is the matrix with

$$C_{i,j} := \sum_{k=1}^{n} A_{i,k} \, B_{k,j} \, .$$

Matrix Multiplication

Recall that a matrix $A \in \mathbb{R}^{m \times n}$ is a concise representation of a linear function $T_A : \mathbb{R}^n \to \mathbb{R}^m$.

Matrix multiplication is **defined** so that $A \times B$ represents the linear function $T_A \circ T_B$, when this composition is legal. (I.e., when the dimension of B's output equals that of A's input.)

Formally, for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, the matrix product $C = AB \in \mathbb{R}^{m \times p}$ is the matrix with

$$C_{i,j} := \sum_{k=1}^{n} A_{i,k} \, B_{k,j} \, .$$

Intuition:

- $B_{k,j}$ tracks how the *j*th input vector turns into the *k*th "middle vector".
- $A_{i,k}$ tracks how the *k*th "middle vector" turns into the *i*th output vector.
- Together, they track how the *j*th input vector turns into the *i*th output vector.

Matrix Multiplication

Visualization from 3Blue1Brown, Essence of linear algebra (**2 min**)

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}} \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Rotation}} = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Composition}}$$

Matrix Multiplication

Matrix multiplication has very different algebraic properties from multiplication of real numbers.

They can be explained by remembering that matrix multiplication is really composition of linear functions in disguise!

Matrix Multiplication

Matrix multiplication has very different algebraic properties from multiplication of real numbers.

They can be explained by remembering that matrix multiplication is really composition of linear functions in disguise!

- **Not** commutative: It can be that $AB \neq BA$ for square matrices A, B .
  - Why? $T_A \circ T_B$ can be very different from $T_B \circ T_A$!

Matrix Multiplication

Matrix multiplication has very different algebraic properties from multiplication of real numbers.

They can be explained by remembering that matrix multiplication is really composition of linear functions in disguise!

- **Not** commutative: It can be that $AB \neq BA$ for square matrices A, B.
  - Why? $T_A \circ T_B$ can be very different from $T_B \circ T_A$!
- Inverses may not exist: Many matrices A do not have an A^{-1}
  - Why? (Linear) functions can destroy information! Take $T(x) = 0$

Matrix Multiplication

Matrix multiplication has very different algebraic properties from multiplication of real numbers.

They can be explained by remembering that matrix multiplication is really composition of linear functions in disguise!

- **Not** commutative: It can be that $AB \neq BA$ for square matrices A, B .
  - Why? $T_A \circ T_B$ can be very different from $T_B \circ T_A$!
- Inverses may not exist: Many matrices A do not have an A^{-1}
  - Why? (Linear) functions can destroy information! Take $T(x) = 0$
- Multiplication is not always defined: Requires shape compatibility
  - Why? Composition of functions $f \circ g$ requires $\text{codomain}(g) = \text{domain}(f)$

Special Matrices

Identity matrix

$I_n \in \mathbb{R}^{n \times n}$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

For all $A \in \mathbb{R}^{m \times n}$, $AI_n = A = I_m A$.

Special Matrices

Identity matrix

$I_n \in \mathbb{R}^{n \times n}$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

For all $A \in \mathbb{R}^{m \times n}$, $AI_n = A = I_m A$.

Diagonal matrix

$D = \text{diag}(d_1, d_2, \ldots, d_n)$

$$\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & d_n \end{bmatrix}$$

Clearly, $I = \text{diag}(1,1,\ldots,1)$.

Vector-Vector Product

Inner Product or Dot Product

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i \, .$$

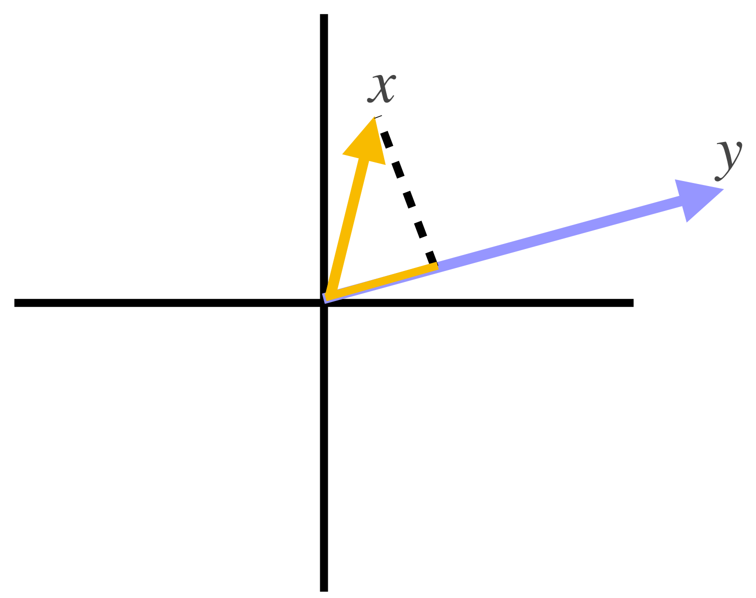
Vector-Vector Product

Inner Product or Dot Product

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i \,.$$

Geometric Intuition

$x^T y =$ (Length of projected x)·(Length of y)



Vector-Vector Product

Outer Product

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

$$\begin{bmatrix} & x_1 & & x_1 & \cdots & & x_1 \\ y_1 & \vdots & y_2 & \vdots & \ddots & y_n & \vdots \\ & x_m & & x_m & \cdots & & x_m \end{bmatrix} \quad \begin{bmatrix} x_1 & (\cdots & y^T & \cdots) \\ x_2 & (\cdots & y^T & \cdots) \\ \vdots & \vdots & \vdots & \vdots \\ x_m & (\cdots & y^T & \cdots) \end{bmatrix}$$

Vector-Vector Product

Outer Product

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

$$\begin{bmatrix} & x_1 & & x_1 & \cdots & & x_1 \\ y_1 & \vdots & y_2 & \vdots & \ddots & y_n & \vdots \\ & x_m & & x_m & \cdots & & x_m \end{bmatrix} \qquad \begin{bmatrix} x_1 & (\cdots & y^T & \cdots) \\ x_2 & (\cdots & y^T & \cdots) \\ \vdots & \vdots & \vdots & \vdots \\ x_m & (\cdots & y^T & \cdots) \end{bmatrix}$$

Geometric Intuition

xy^T is the linear map that measure how much an input aligns with y, then outputs that amount in direction x.

(Applications to attention, covariance matrices, PCA, etc.)

Matrix-Vector Product

View 1: Write A by rows

$$y = Ax = \begin{bmatrix} --- & a_1^T & --- \\ --- & a_2^T & --- \\ & \vdots & \\ --- & a_m^T & --- \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

This is function evaluation! Ax is the vector $T_A(x)$

Set of inner products with each row vector

Intuition: $a_i^T x$ is how much of e_i gets "produced" by x, across all of its entries.

Matrix-Vector Product

View 2: Write A by columns

$$y = Ax = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ a^1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ a^2 \\ | \end{bmatrix} x_2 + \ldots + \begin{bmatrix} | \\ a^n \\ | \end{bmatrix} x_n .$$

Linear combination of column vectors

Intuition: $a^1 x_1$ is the full vector produced by $(x_1, 0, \ldots, 0) = x_1 e_1$

Key corollary: Ax is restricted to the "column space" of A

Vector-Matrix Product

View 1: Write A by rows

$$y^T = x^T A = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix} \begin{bmatrix} \text{---} & a_1^T & \text{---} \\ \text{---} & a_2^T & \text{---} \\ & \vdots & \\ \text{---} & a_m^T & \text{---} \end{bmatrix}$$

$$= x_1 \begin{bmatrix} \text{---} & a_1^T & \text{---} \end{bmatrix} + x_2 \begin{bmatrix} \text{---} & a_2^T & \text{---} \end{bmatrix} + \ldots + x_m \begin{bmatrix} \text{---} & a_m^T & \text{---} \end{bmatrix}$$

Intuition: $x^T A$ expresses linear combination of A's rows, whereas Ax expresses linear combination of A's columns

Vector-Matrix Product

View 2: Write A by columns

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a^1 & x^T a^2 & \cdots & x^T a^n \end{bmatrix}$$

Set of inner products with each column vector

Intuition: Combining rows of A one dimension at a time, rather than in one shot.

Matrix-Matrix Multiplication

View 1: Set of inner products

$$C = AB = \begin{bmatrix} \text{---} & a_1^T & \text{---} \\ \text{---} & a_2^T & \text{---} \\ & \vdots & \\ \text{---} & a_m^T & \text{---} \end{bmatrix} \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b^1 & a_1^T b^2 & \cdots & a_1^T b^n \\ a_2^T b^1 & a_2^T b^2 & \cdots & a_2^T b^n \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b^1 & a_m^T b^2 & \cdots & a_m^T b^n \end{bmatrix}$$

Matrix of all possible row/column inner products

Intuition: b^i measures "intermediate" output of e_i. a_i^T measures how "intermediate" vectors produce final output e_j. Dot product glues them together!

Matrix-Matrix Multiplication

View 2: Set of matrix-vector products

$$C = AB = \boxed{A}\begin{bmatrix} | & | & & | \\ b^1 & b^2 & \cdots & b^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab^1 & Ab^2 & \cdots & Ab^n \\ | & | & & | \end{bmatrix}$$

Intuition: b^i is B's output from e_i. So Ab^i is $AB's$ output from e_i. I.e.,

$$(AB)e_i = A(Be_i) = Ab^i$$

Matrix-Matrix Multiplication

Properties

- Associative: $(AB)C = A(BC)$.
- Distributive: $A(B + C) = AB + AC$.
- In general, **not** commutative; it can be the case that $AB \neq BA$.

Transpose

The transpose of a matrix results from 'flipping' the rows and columns.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

Transpose

The transpose of a matrix results from 'flipping' the rows and columns.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

- Properties:
  - $(A^T)^T = A$.
  - $(AB)^T = B^T A^T$.
  - $(A + B)^T = A^T + B^T$
- If $A = A^T$, then A is a symmetric matrix
- If $A = -A^T$, then A is an anti-symmetric matrix

Exercise

- Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are all D-dimensional vectors, and $X \in \mathbb{R}^{N \times D}$ is a matrix where the n-th row is $\mathbf{x}_n^\top$. Then which of the following identities are correct?

A. $X^\top X = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top$

B. $X^\top X = \sum_{n=1}^{N} \mathbf{x}_n^\top \mathbf{x}_n$

C. $X X^\top = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top$

D. $X X^\top = \sum_{n=1}^{N} \mathbf{x}_n^\top \mathbf{x}_n$

Exercise

- Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are all D-dimensional vectors, and $X \in \mathbb{R}^{N \times D}$ is a matrix where the n-th row is $\mathbf{x}_n^\top$. Then which of the following identities are correct?

A. $X^\top X = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top$

B. $X^\top X = \sum_{n=1}^{N} \mathbf{x}_n^\top \mathbf{x}_n$

C. $XX^\top = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top$

D. $XX^\top = \sum_{n=1}^{N} \mathbf{x}_n^\top \mathbf{x}_n$

Trace

The trace of a square matrix is the sum of its diagonal elements

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}\,.$$

- Properties ($A, B, C \in \mathbb{R}^{n\times n}$):
  - $\mathrm{tr}A = \mathrm{tr}A^T$.
  - $\mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$.
  - $\mathrm{tr}(tA) = t\ \mathrm{tr}A$
  - $\mathrm{tr}AB = \mathrm{tr}BA$
  - $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on.

Norms

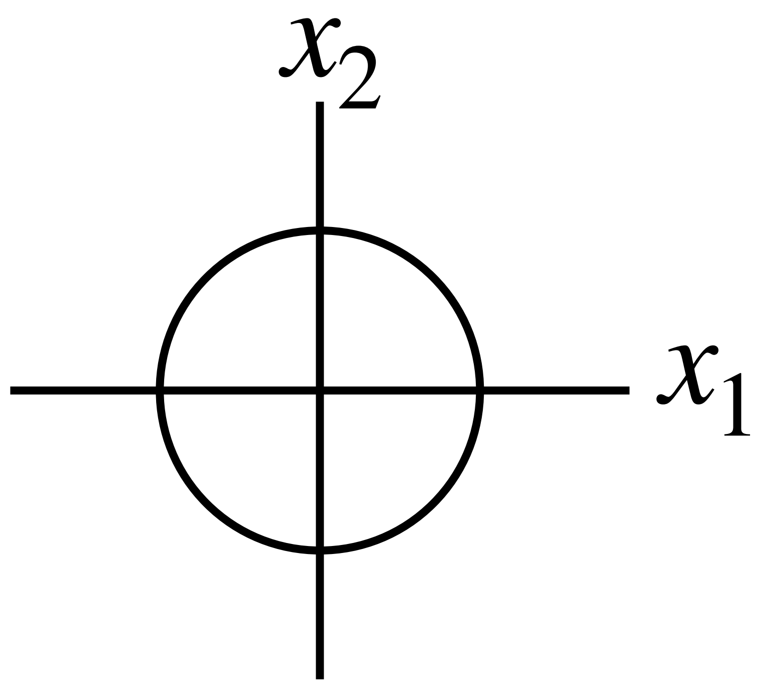
- Informally, norm of a vector measures the 'length' of the vector.

- Formally, any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties for $x, y \in \mathbb{R}^n$:
  - Non-negativity: $f(x) \geq 0$
  - Definiteness: $f(x) = 0$ iff $x = 0$
  - Homogeneity: $f(tx) = |t| f(x)$
  - Triangle inequality: $f(x + y) \leq f(x) + f(y)$

Examples of Norms

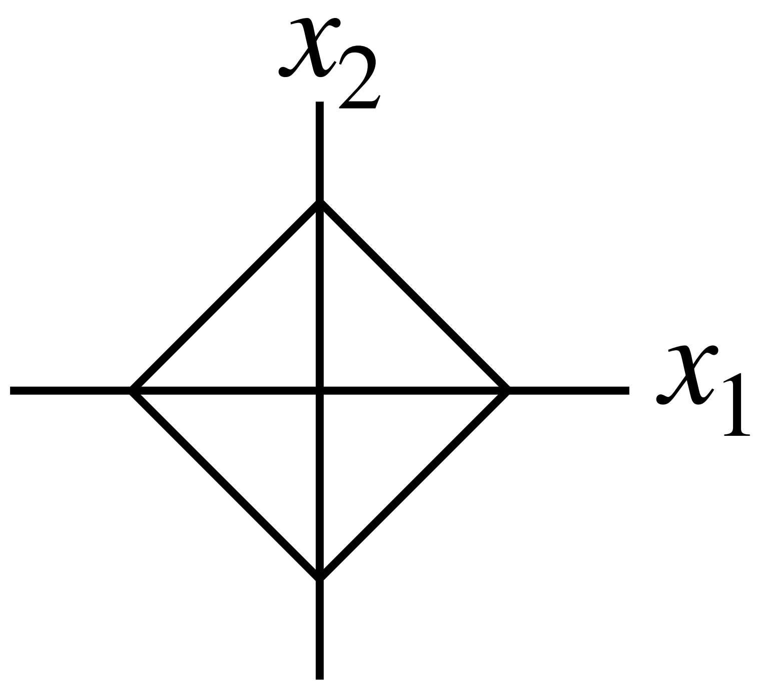
- Euclidean or ℓ_2−norm:

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



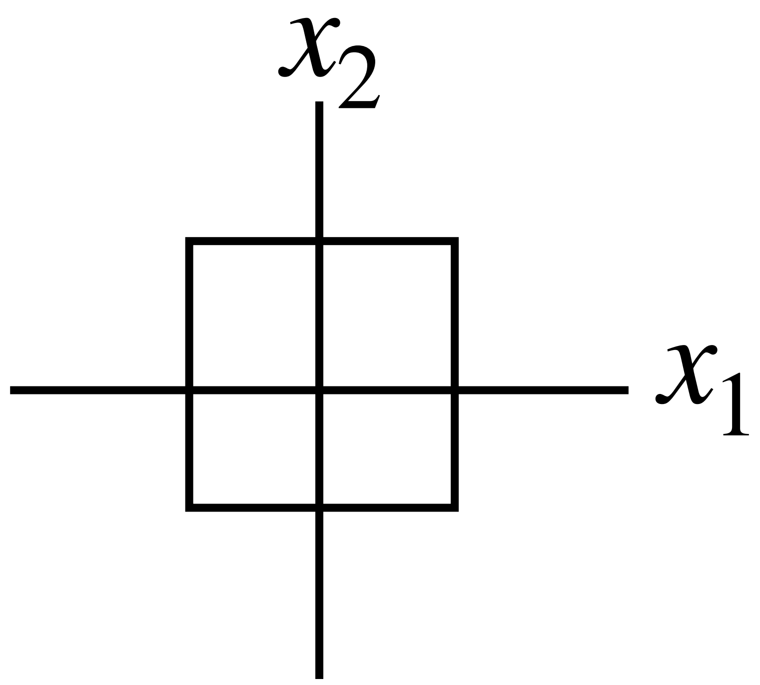
- ℓ_1−norm:

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



- ℓ_∞−norm:

$$\|x\|_\infty = \max_i |x_i|$$

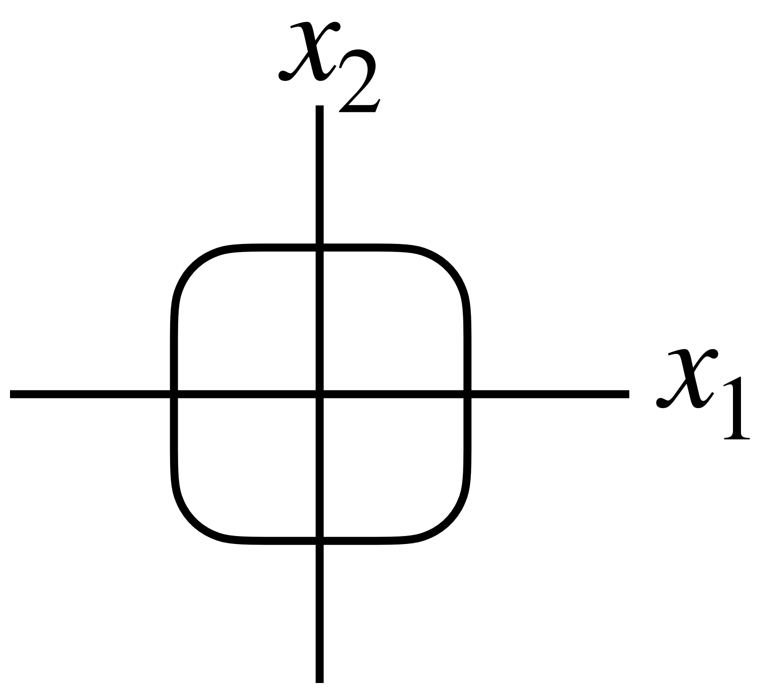


ℓ_p−Norms

- Family of ℓ_p−norms, parameterized by a real number $p \geq 1$:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

- For $p \geq 2$:



Matrix Norms

- Frobenius norm:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$$

$$= \sqrt{\sum_{i=1}^{m} \|a_i\|_2^2} = \sqrt{\sum_{j=1}^{n} \|a^j\|_2^2}$$

$$= \sqrt{\operatorname{tr}(A^T A)}$$

Linear Combinations and Span

- The span of a set of vectors $\{x_1, x_2, \ldots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \ldots, x_n\}$. That is,

$$\text{span}(\{x_1, \ldots x_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i x_i, \;\; \alpha_i \in \mathbb{R} \right\}$$

- The span of column vectors of a matrix is known as the column space.

- Similarly, the span of row vectors is known as the row space.

Linear Independence

- A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be (linearly) dependent if one vector belonging to the set can be represented as a linear combination of the remaining vectors; that is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$.

- Otherwise, the vectors are (linearly) independent.

Rank

- Column rank: largest number of columns that constitute a linearly independent set.
- Row rank: largest number of rows that constitute a linearly independent set.
- Column rank of any matrix is equal to its row rank.
- Both quantities collectively referred to as the rank of the matrix.

Rank

- Column rank: largest number of columns that constitute a linearly independent set.
- Row rank: largest number of rows that constitute a linearly independent set.
- Column rank of any matrix is equal to its row rank.
- Both quantities collectively referred to as the rank of the matrix.

- Properties ($A \in \mathbb{R}^{m \times n}$):
  - $\mathrm{rank}(A) \leq \min(m, n)$. If $\mathrm{rank}(A) = \min(m, n)$, A is said to be full rank.
  - $\mathrm{rank}(A) = \mathrm{rank}(A^T)$.
  - For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$.
  - For $A, B \in \mathbb{R}^{m \times n}$, $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$.

Inverse of a Square Matrix

- The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted A^{-1}, is the unique matrix such that $A^{-1}A = I_n = AA^{-1}$.
- A must be full rank for its inverse to exist.
- A is invertible or non-singular if A^{-1} exists and non-invertible or singular otherwise.

- Properties ($A, B \in \mathbb{R}^{n \times n}$ are non-singular):
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$, denoted by A^{-T}

Determinant

Intuition

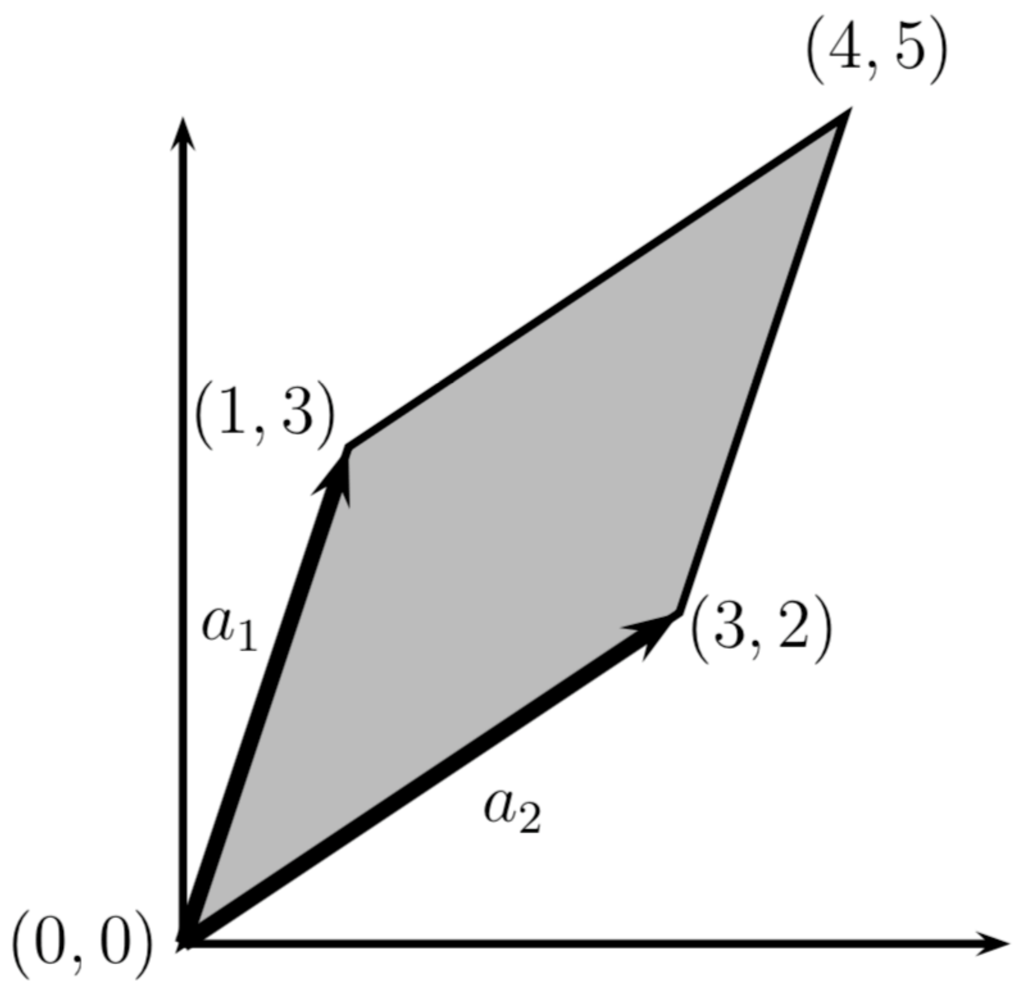
- Let $A \in \mathbb{R}^{n \times n}$, a_i denotes its ith column; consider the set of points $S \subset \mathbb{R}^n$:

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i a_i \ \ (0 \leq \alpha_i \leq 1; \ \ i = 1, \ldots, n)\}$$

- The absolute value of the determinant of A gives the 'volume' of the set S

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$



Determinant

(Recursive) Formula

- Let $A \in \mathbb{R}^{n \times n}$, $A_{\backslash i, \backslash j} \in \mathbb{R}^{(n-1)\times(n-1)}$ be the matrix that results from deleting the ith row and jth column from A

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \qquad (\forall\ j \in 1,\ldots,n)$$

$$= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \qquad (\forall\ i \in 1,\ldots,n)$$

- Equations for small matrices:

$$\left| [a_{11}] \right| = a_{11} \qquad \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \right| = a_{11}a_{22} - a_{12}a_{21}$$

Determinant

Properties

- Properties ($A, B \in \mathbb{R}^{n \times n}$):
  - $|A| = |A^T|$
  - $|AB| = |A||B|$
  - $|A| = 0$ iff A is singular
  - For non-singular A, $|A^{-1}| = 1/|A|$

Exercise

- Which identities are **NOT** correct for real-valued matrices A, B, and C? Assume that inverses exist and multiplications are legal.

A. $(AB)^{-1} = B^{-1}A^{-1}$

B. $(I + A)^{-1} = I - A$

C. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

D. $(AB)^\top = A^\top B^\top$

Exercise

- Consider some vector $x \in \mathbb{R}^n$. What is the rank of the matrix xx^T?

Matrix Calculus

Gradient

- Suppose $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a scalar function that takes as input a matrix $A \in \mathbb{R}^{m\times n}$
- The gradient of f with respect to A is the $(m \times n)$ matrix of partial derivatives:

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

Gradient

- If the input is just a vector $x \in \mathbb{R}^n$,

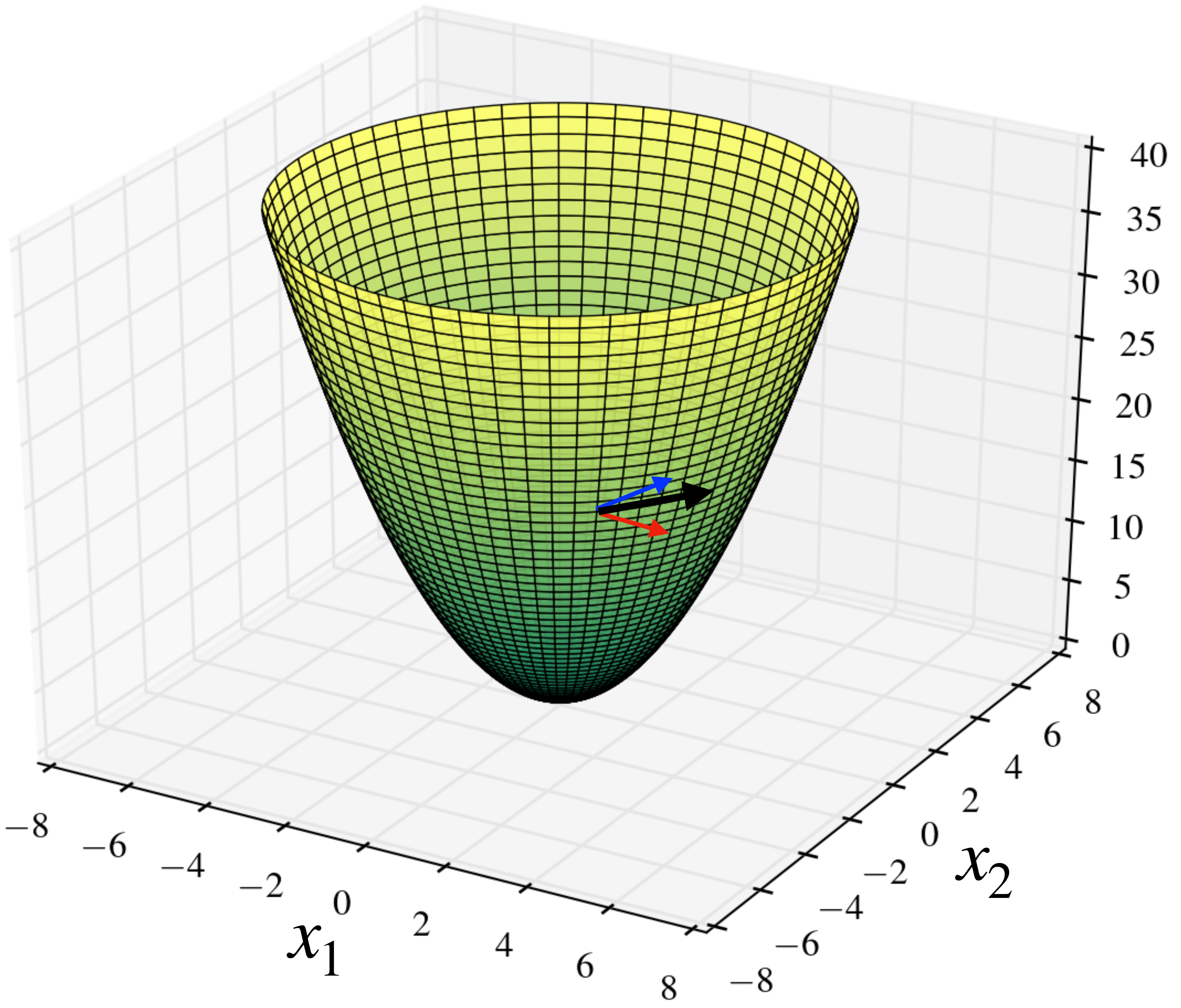
$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

- Properties of partial derivatives extend here:
  - $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
  - For $t \in \mathbb{R}$, $\nabla_x(t\, f(x)) = t \nabla_x f(x)$.

Gradient

Visual Example

$$\nabla_x f(x) = \begin{bmatrix} \dfrac{\partial f(x)}{\partial x_1} \\ \dfrac{\partial f(x)}{\partial x_2} \end{bmatrix}$$



Hessian

- Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a scalar function that takes as input a vector $x \in \mathbb{R}^n$
- The Hessian of f with respect to x is the $(n \times n)$ matrix of partial derivatives:

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

- It is symmetric (provided the second partial derivatives are continuous).

Jacobian

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is a vector function that takes as input a vector $x \in \mathbb{R}^n$
- The Jacobian of f with respect to x is the $(m \times n)$ matrix of partial derivatives:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_x^T f_1(x) \\ \nabla_x^T f_2(x) \\ \vdots \\ \nabla_x^T f_m(x) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}$$

Gradient of a Linear Function

- For $x \in \mathbb{R}^n$, let $f(x) = b^T x \ (= x^T b)$ for some known vector $b \in \mathbb{R}^n$. Then,

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

- This gives:

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$$

$$\nabla_x b^T x = b$$

- Analogous to single variable calculus, where $\dfrac{\partial\,(ax)}{\partial x} = a$

Jacobian of a Linear Function

- For $x \in \mathbb{R}^n$, let $f(x) = Ax$ for some known matrix $A \in \mathbb{R}^{m \times n}$. Then,

$$f_i(x) = a_i^T x \;\; \forall i = 1, \cdots, m$$

- This gives:

$$\nabla_x f_i(x) = a_i$$

$$\nabla_x f(x) = \begin{bmatrix} --- & a_1^T & --- \\ --- & a_2^T & --- \\ & \vdots & \\ --- & a_m^T & --- \end{bmatrix} = A$$

Gradient of a Quadratic Function

- For $x \in \mathbb{R}^n$, let $f(x) = x^T A x$ for some known matrix $A \in \mathbb{R}^{n \times n}$. Then,

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

- Using previous slides, product rule for $f(x) = g(x)^T x$, with $g(x) = A^T x$, we get:

$$\begin{aligned} \nabla_x f(x) &= \nabla_x^T g(x) x + \nabla_x^T x g(x) \\ &= (A^T)^T x + I^T A^T x \\ &= (A + A^T) x \end{aligned}$$

- This gives the Hessian:

$$\nabla_x^2 f(x) = A + A^T$$

Exercise

- A function $f : \mathbb{R}^{n\times 1} \to \mathbb{R}$ is defined as $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ for some $\mathbf{b} \in \mathbb{R}^{n\times 1}$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$. What is the derivative $\dfrac{\partial f}{\partial \mathbf{x}}$ (also called the gradient $\nabla f(\mathbf{x})$)?

Exercise

- A function $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ is defined as $f(\mathbf{A}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^{n \times 1}$. What is the derivative $\frac{\partial f}{\partial \mathbf{A}}$?

Questions?

Next Week: Probability Review