

# Lecture 2: Linear algebra

STATS 101: Foundations of Statistics

Linh Tran

ThetaHat.AI@gmail.com

November 21, 2019

# Announcements

- ▶ 30 students submitted homework
- ▶ Next assignment will be posted tonight (due 12/4 @ 11:59pm)
- ▶ Re: prizes for top 3 students
  - ▶ Amazon gift cards
  - ▶ Extra points are awarded for:
    1. Class participation (e.g. asking/answering questions, etc)
    2. Catching & correcting errata/typos
    3. Answering questions / participating in discussions on Piazza
  - ▶ Blinded top 3 scores will be posted to course website
- ▶ No class next week (Happy Thanksgiving!)

## All things linear algebra

- ▶ Basic concepts
- ▶ Matrix multiplication
- ▶ Operations and Properties
- ▶ Matrix Calculus

# Basic concepts

Consider the following equations:

$$4x_1 - 5x_2 = -13 \quad (1)$$

$$-2x_1 + 3x_2 = 9 \quad (2)$$

Let's solve for  $x_1$  and  $x_2$ .

# Basic concepts

Consider the following equations:

$$4x_1 - 5x_2 = -13 \quad (1)$$

$$-2x_1 + 3x_2 = 9 \quad (2)$$

Let's solve for  $x_1$  and  $x_2$ .

We can write this system of equations more compactly in matrix notation, e.g.

$$\mathbf{Ax} = \mathbf{b} \quad (3)$$

$$\text{where } \mathbf{A} = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

# Basic concepts

Some basic notation:

- ▶ We denote a matrix with  $m$  rows and  $n$  columns as  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , where each entry in the matrix is a real number.
- ▶ We denote a vector with  $n$  entries as  $\mathbf{x} \in \mathbb{R}^n$ .
  - ▶ By convention, we typically think of a vector as a 1 column matrix.
- ▶ We denote the  $i^{th}$  element of a vector  $\mathbf{x}$  as  $x_i$ , e.g.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (4)$$

# Basic concepts

Some basic notation:

- We denote each entry in a matrix  $\mathbf{A}$  by  $a_{ij}$ , corresponding to the  $i^{th}$  row and  $j^{th}$  column, e.g.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad (5)$$

- We denote the *transpose* of a matrix as  $\mathbf{A}^\top$ , e.g.

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \quad (6)$$

# Basic concepts

Some basic notation:

- We denote the  $j^{th}$  column of  $\mathbf{A}$  by  $\mathbf{a}_j$  or  $\mathbf{A}_{.j}$ , e.g.

$$\mathbf{A} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \quad (7)$$

- We denote the  $i^{th}$  row of  $\mathbf{A}$  by  $\mathbf{a}_i^\top$  or  $\mathbf{A}_{i..}$ .

$$\mathbf{A} = \begin{bmatrix} \text{---} & \mathbf{a}_1^\top & \text{---} \\ \text{---} & \mathbf{a}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{a}_m^\top & \text{---} \end{bmatrix} \quad (8)$$

n.b. This isn't universal, though should be clear from its presentation and use.



# Matrix multiplication

Given two matrices  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times p}$ , we can multiply them by

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p} : \mathbf{C}_{ij} = \sum_{k=1}^n \mathbf{A}_{ik} \mathbf{B}_{kj} \quad (9)$$

n.b. The dimensions have to be compatible for matrix multiplication to be valid (e.g. the number of columns in  $\mathbf{A}$  must be equal to the number of rows in  $\mathbf{B}$ ).

# Matrix multiplication

Given  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , the quantity  $\mathbf{x}^\top \mathbf{y} \in \mathbb{R}$  (aka *dot product* or *inner product*) is a scalar given by

$$\mathbf{x}^\top \mathbf{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i \quad (10)$$

Note: For vectors, we always have that  $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$ . This is not generally true for matrices.

# Matrix multiplication

Given  $\mathbf{x} \in \mathbb{R}^m$ ,  $\mathbf{y} \in \mathbb{R}^n$ , the quantity  $\mathbf{x}^\top \mathbf{y} \in \mathbb{R}^{m \times n}$  (aka *outer product*) is a matrix given by

$$\mathbf{xy}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix} \quad (11)$$

# Matrix multiplication

**Example:** Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  be a matrix such that all columns are equal to some vector  $\mathbf{x} \in \mathbb{R}^m$ . Using outer products, we can represent  $\mathbf{A}$  compactly as

$$\mathbf{A} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x} & \mathbf{x} & \cdots & \mathbf{x} \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} \quad (12)$$

$$= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \quad (13)$$

$$= \mathbf{x} \mathbf{1}^\top \quad (14)$$

# Matrix-vector products

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{x} \in \mathbb{R}^n$ , their product is a vector  $\mathbf{y} = \mathbf{Ax} \in \mathbb{R}^m$ .

# Matrix-vector products

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{x} \in \mathbb{R}^n$ , their product is a vector  $\mathbf{y} = \mathbf{Ax} \in \mathbb{R}^m$ .

There are two ways of interpreting this:

$$\mathbf{y} = \mathbf{Ax} = \begin{bmatrix} \text{---} & \mathbf{a}_1^\top & \text{---} \\ \text{---} & \mathbf{a}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{a}_m^\top & \text{---} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix} \quad (15)$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (16)$$

$$= \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \cdots + \mathbf{a}_n x_n \quad (17)$$

# Matrix-vector products

**Example:**

Define  $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$ ,  $\mathbf{x} = \begin{bmatrix} -3 \\ -2 \\ -1 \end{bmatrix}$ .

Calculate  $\mathbf{y} = \mathbf{Ax}$ .

# Matrix-matrix products

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times p}$ , their product is a matrix  $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}$ .



# Matrix-matrix products

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times p}$ , their product is a matrix  $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}$ .

Similar to before, we can think of this in two ways:

## Interpretation # 1

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \text{---} & \mathbf{a}_1^\top & \text{---} \\ \text{---} & \mathbf{a}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{a}_m^\top & \text{---} \end{bmatrix} \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & & | \end{bmatrix} \quad (18)$$

$$= \begin{bmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots & \mathbf{a}_1^\top \mathbf{b}_p \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots & \mathbf{a}_2^\top \mathbf{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_m^\top \mathbf{b}_1 & \mathbf{a}_m^\top \mathbf{b}_2 & \cdots & \mathbf{a}_m^\top \mathbf{b}_p \end{bmatrix} \quad (19)$$

## Interpretation # 2

$$\mathbf{C} = \mathbf{AB} = \mathbf{A} \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & \cdots & | \end{bmatrix} \quad (20)$$

$$= \begin{bmatrix} | & | & \cdots & | \\ \mathbf{Ab}_1 & \mathbf{Ab}_2 & \cdots & \mathbf{Ab}_p \\ | & | & \cdots & | \end{bmatrix} \quad (21)$$

$$= \begin{bmatrix} \text{---} & \mathbf{a}_1^\top & \text{---} \\ \text{---} & \mathbf{a}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{a}_m^\top & \text{---} \end{bmatrix} \mathbf{B} = \begin{bmatrix} \text{---} & \mathbf{a}_1^\top \mathbf{B} & \text{---} \\ \text{---} & \mathbf{a}_2^\top \mathbf{B} & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{a}_m^\top \mathbf{B} & \text{---} \end{bmatrix} \quad (22)$$

# Matrix multiplication properties

- ▶ Associative:  $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- ▶ Distributive:  $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- ▶ Not commutative:  $\mathbf{AB} \neq \mathbf{BA}$

# Matrix multiplication properties

Demonstrating *associativity*:

We just need to show that  $((\mathbf{AB})\mathbf{C})_{ij} = (\mathbf{A}(\mathbf{BC}))_{ij}$ :

$$((\mathbf{AB})\mathbf{C})_{ij} = \sum_{k=1}^p (\mathbf{AB})_{ik} \mathbf{C}_{kj} = \sum_{k=1}^p \left( \sum_{l=1}^n \mathbf{A}_{il} \mathbf{B}_{lk} \right) \mathbf{C}_{kj} \quad (23)$$

$$= \sum_{k=1}^p \left( \sum_{l=1}^n \mathbf{A}_{il} \mathbf{B}_{lk} \mathbf{C}_{kj} \right) = \sum_{l=1}^n \left( \sum_{k=1}^p \mathbf{A}_{il} \mathbf{B}_{lk} \mathbf{C}_{kj} \right) \quad (24)$$

$$= \sum_{l=1}^n \mathbf{A}_{il} \left( \sum_{k=1}^p \mathbf{B}_{lk} \mathbf{C}_{kj} \right) = \sum_{l=1}^n \mathbf{A}_{il} (\mathbf{BC})_{lj} \quad (25)$$

$$= (\mathbf{A}(\mathbf{BC}))_{ij} \quad (26)$$

## The identity matrix:

The *identity matrix*, denoted  $\mathbf{I} \in \mathbb{R}^{n \times n}$  is a square matrix with 1's in the diagonal and 0's everywhere else, i.e.

$$\mathbf{I}_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (27)$$

## The identity matrix:

The *identity matrix*, denoted  $\mathbf{I} \in \mathbb{R}^{n \times n}$  is a square matrix with 1's in the diagonal and 0's everywhere else, i.e.

$$\mathbf{I}_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (27)$$

It has the property

$$\mathbf{A}\mathbf{I} = \mathbf{A} = \mathbf{I}\mathbf{A} \quad \forall \mathbf{A} \in \mathbb{R}^{m \times n} \quad (28)$$

n.b. The dimensionality of  $\mathbf{I}$  is typically inferred (e.g.  $n \times n$  vs  $m \times m$ )

**The diagonal matrix:** The *diagonal matrix*, denoted  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$  is a matrix where all non-diagonal elements are 0, i.e.

$$\mathbf{D}_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases} \quad (29)$$

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

# The transpose

The *transpose* of a matrix results from “*flipping*” the rows and columns, i.e.

$$(\mathbf{A}^\top)_{ij} = \mathbf{A}_{ji} \quad (30)$$

Consequently, for  $\mathbf{A} \in \mathbb{R}^{m \times n}$  we have that  $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$ .

Some properties:

- ▶  $(\mathbf{A}^\top)^\top = \mathbf{A}$
- ▶  $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$
- ▶  $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$



# Symmetry

A square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is *symmetric* if  $\mathbf{A} = \mathbf{A}^\top$ .

It is *anti-symmetric* if  $\mathbf{A} = -\mathbf{A}^\top$ .

A square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is *symmetric* if  $\mathbf{A} = \mathbf{A}^\top$ .

It is *anti-symmetric* if  $\mathbf{A} = -\mathbf{A}^\top$ .

It is easy to show that  $\mathbf{A} + \mathbf{A}^\top$  is symmetric and  $\mathbf{A} - \mathbf{A}^\top$  is anti-symmetric. Consequently, we have that

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top) \quad (31)$$

A square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is *symmetric* if  $\mathbf{A} = \mathbf{A}^\top$ .

It is *anti-symmetric* if  $\mathbf{A} = -\mathbf{A}^\top$ .

It is easy to show that  $\mathbf{A} + \mathbf{A}^\top$  is symmetric and  $\mathbf{A} - \mathbf{A}^\top$  is anti-symmetric. Consequently, we have that

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top) \quad (31)$$

Symmetric matrices tend to be denoted as  $\mathbf{A} \in \mathbb{S}^n$ .

The *trace* of a square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , denoted  $tr(\mathbf{A})$  or  $tr\mathbf{A}$  is the sum of the diagonal elements, i.e.

$$tr\mathbf{A} = \sum_{i=1}^n \mathbf{A}_{ii} \quad (32)$$

The trace has the following properties:

- ▶ For  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $tr\mathbf{A} = tr\mathbf{A}^T$
- ▶ For  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ ,  $tr(\mathbf{A} + \mathbf{B}) = tr\mathbf{A} + tr\mathbf{B}$
- ▶ For  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $c \in \mathbb{R}$ ,  $tr(c\mathbf{A}) = c tr\mathbf{A}$
- ▶ For  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n} \ni \mathbf{AB} \in \mathbb{R}^{n \times n}$ ,  $tr\mathbf{AB} = tr\mathbf{BA}$
- ▶ For  $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times n} \ni \mathbf{ABC} \in \mathbb{R}^{n \times n}$ ,  
 $tr\mathbf{ABC} = tr\mathbf{BCA} = tr\mathbf{CAB}$ , and so on for more matrices

**Example:** Proving that  $tr\mathbf{AB} = tr\mathbf{BA}$

$$tr\mathbf{AB} = \sum_{i=1}^m (\mathbf{AB})_{ii} = \sum_{i=1}^m \left( \sum_{j=1}^n \mathbf{A}_{ij} \mathbf{B}_{ji} \right) \quad (33)$$

$$= \sum_{i=1}^m \sum_{j=1}^n \mathbf{A}_{ij} \mathbf{B}_{ji} = \sum_{i=1}^m \sum_{j=1}^n \mathbf{B}_{ji} \mathbf{A}_{ij} \quad (34)$$

$$= \sum_{i=1}^m \left( \sum_{j=1}^n \mathbf{B}_{ji} \mathbf{A}_{ij} \right) = \sum_{j=1}^n (\mathbf{BA})_{jj} \quad (35)$$

$$= tr\mathbf{BA} \quad (36)$$

# Norms

A *norm* of a vector  $\mathbf{x}$ , denoted  $\|\mathbf{x}\|$  is a measure of the “*length*” of the vector. For example, the  $\ell_2$ -norm (aka Euclidean norm) is

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \quad (37)$$

n.b.  $\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}$ , i.e. the squared norm of a vector is the dot product with itself.

A **norm** of a vector  $\mathbf{x}$ , denoted  $\|\mathbf{x}\|$  is a measure of the “length” of the vector. For example, the  $\ell_2$ -norm (aka Euclidean norm) is

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \quad (37)$$

n.b.  $\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}$ , i.e. the squared norm of a vector is the dot product with itself.

## Other norms:

- ▶  $\ell_1$ -norm, i.e.  $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ .
- ▶  $\ell_\infty$ -norm, i.e.  $\|\mathbf{x}\|_\infty = \max_i |x_i|$ .
- ▶  $\ell_p$ -norm, i.e.  $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ .

Formally, a norm is any function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfying four properties:

1.  $\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \geq 0$  (non-negativity).
2.  $f(\mathbf{x}) = 0$  iff  $\mathbf{x} = 0$  (definiteness).
3.  $\forall \mathbf{x} \in \mathbb{R}^n, c \in \mathbb{R}, f(c\mathbf{x}) = |c|f(\mathbf{x})$  (homogeneity).
4.  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$  (triangle inequality).



Formally, a norm is any function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfying four properties:

1.  $\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \geq 0$  (non-negativity).
2.  $f(\mathbf{x}) = 0$  iff  $\mathbf{x} = 0$  (definiteness).
3.  $\forall \mathbf{x} \in \mathbb{R}^n, c \in \mathbb{R}, f(c\mathbf{x}) = |c|f(\mathbf{x})$  (homogeneity).
4.  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$  (triangle inequality).

Norms can also be defined for matrices, e.g. The Frobenius norm,

$$\|\mathbf{A}\|^F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n \mathbf{A}_{ij}^2} = \sqrt{\text{tr}(\mathbf{A}^\top \mathbf{A})} \quad (38)$$

# Linear independence

A set of vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathbb{R}^m$  is *(linearly) dependent* if one of the vectors  $\mathbf{x}_i$  can be represented as a linear combination of the remaining vectors, i.e.

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i \quad (39)$$

for some scalar values  $\alpha_1, \alpha_2, \dots, \alpha_{n-1} \in \mathbb{R}$

# Linear independence

A set of vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathbb{R}^m$  is *(linearly) dependent* if one of the vectors  $\mathbf{x}_i$  can be represented as a linear combination of the remaining vectors, i.e.

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i \quad (39)$$

for some scalar values  $\alpha_1, \alpha_2, \dots, \alpha_{n-1} \in \mathbb{R}$

**Example:** Let

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} \quad (40)$$

Is  $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$  linearly independent?

# Rank

The *column rank* of  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is the largest subset of columns of  $\mathbf{A}$  that are linearly independent.

- ▶ The column rank is always  $\leq n$ .

The *row rank* of  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is the largest subset of rows of  $\mathbf{A}$  that are linearly independent.

- ▶ The row rank is always  $\leq m$ .

# Rank

The *column rank* of  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is the largest subset of columns of  $\mathbf{A}$  that are linearly independent.

- ▶ The column rank is always  $\leq n$ .

The *row rank* of  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is the largest subset of rows of  $\mathbf{A}$  that are linearly independent.

- ▶ The row rank is always  $\leq m$ .

n.b. Column rank is always equal to row rank. Thus, we refer to both as the *rank* of the matrix.

- ▶ For  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , if  $\text{rank}(\mathbf{A}) = \min(m, n)$ , then  $\mathbf{A}$  is said to be of *full rank*.
- ▶ For  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$ .
- ▶ For  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times p}$ ,  
 $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$ .
- ▶ For  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ ,  $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$

# Matrix inverse

The *inverse* of a square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is denoted  $\mathbf{A}^{-1}$ , and is unique such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \quad (41)$$

# Matrix inverse

The *inverse* of a square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is denoted  $\mathbf{A}^{-1}$ , and is unique such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \quad (41)$$

n.b. Not all matrices have inverses (e.g.  $m \times n$  matrices).

## Def:

A is *invertible* or *non-singular* if  $\mathbf{A}^{-1}$  exists.

Otherwise, it is *non-invertible* or *singular*.

1.  $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2.  $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
3.  $(\mathbf{A}^{-1})^{\top} = (\mathbf{A}^{\top})^{-1}$

► This matrix is sometimes denoted  $\mathbf{A}^{-\top}$

# Orthogonal Matrices

## Def:

- ▶ A vector  $\mathbf{x} \in \mathbb{R}^n$  is *normalized* if  $\|\mathbf{x}\|_2 = 1$
- ▶ Two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  are *orthogonal* if  $\mathbf{x}^\top \mathbf{y} = 0$
- ▶ A square matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$  is *orthogonal* or *orthonormal* if all its columns are:
  1. Orthogonal to each other
  2. Normalized

We therefore have that

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U} \mathbf{U}^\top \quad (42)$$



# Orthogonal Matrices

## Def:

- ▶ A vector  $\mathbf{x} \in \mathbb{R}^n$  is *normalized* if  $\|\mathbf{x}\|_2 = 1$
- ▶ Two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  are *orthogonal* if  $\mathbf{x}^\top \mathbf{y} = 0$
- ▶ A square matrix  $\mathbf{U} \in \mathbb{R}^{n \times n}$  is *orthogonal* or *orthonormal* if all its columns are:
  1. Orthogonal to each other
  2. Normalized

We therefore have that

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U} \mathbf{U}^\top \quad (42)$$

Another nice property:

$$\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{U} \in \mathbb{R}^{n \times n} \text{ orthogonal} \quad (43)$$

**Def:**

The *span* of a set of vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is

$$\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \left\{ \mathbf{v} : \mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{x}_i, \alpha_i \in \mathbb{R} \right\} \quad (44)$$

**Def:**

The *span* of a set of vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is

$$\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \left\{ \mathbf{v} : \mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{x}_i, \alpha_i \in \mathbb{R} \right\} \quad (44)$$

n.b. If  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is linearly independent, then  $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \mathbb{R}^n$ .

**Example:**

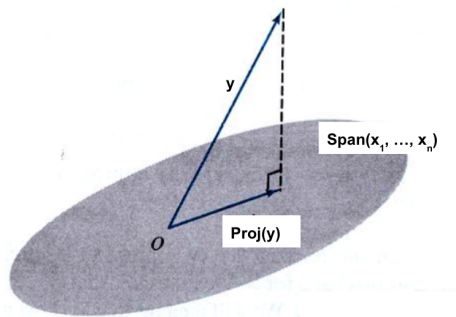
$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (45)$$

# Projection

## Def:

The *projection* of a vector  $\mathbf{y} \in \mathbb{R}^m$  onto  $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \mathbb{R}^n$  is

$$\text{Proj}(\mathbf{y}; \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \arg \min_{\mathbf{v} \in \text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\})} \|\mathbf{y} - \mathbf{v}\|_2 \quad (46)$$



**Def:**

The *range* of a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , denoted  $\mathcal{R}(\mathbf{A})$  is the span of the columns of  $\mathbf{A}$ , i.e.

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^m : \mathbf{v} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\} \quad (47)$$

Assuming that  $\mathbf{A}$  is full rank and  $n < m$ , the projection of  $\mathbf{y} \in \mathbb{R}^m$  onto  $\mathcal{R}(\mathbf{A})$  is

$$\text{Proj}(\mathbf{y}; \mathbf{A}) = \arg \min_{\mathbf{v} \in \mathcal{R}(\mathbf{A})} \|\mathbf{v} - \mathbf{y}\|_2 \quad (48)$$

$$= \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \quad (49)$$

**Def:**

The *nullspace* of a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , denoted  $\mathcal{N}(\mathbf{A})$  is the set of all vectors that equal 0 when multiplied by  $\mathbf{A}$ , i.e.

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}\} \quad (50)$$

Some properties:

- ▶  $\{w : w = u + v, u \in \mathcal{R}(\mathbf{A}^\top), v \in \mathcal{N}(\mathbf{A})\} = \mathbb{R}^n$
- ▶  $\mathcal{R}(\mathbf{A}^\top) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$

This is referred to as *orthogonal complements*, denoted as  $\mathcal{R}(\mathbf{A}^\top) = \mathcal{N}(\mathbf{A})^\perp$

**Def:**

The *determinant* of a square matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , denoted  $|\mathbf{A}|$  or  $\det \mathbf{A}$  is a function  $\det: \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ .

Let  $\mathbf{A}_{\setminus i, \setminus j} \in \mathbb{R}^{(n-1) \times (n-1)}$  be the matrix that results from deleting the  $i^{th}$  row and  $j^{th}$  column. The general (recursive) formula for the determinant is

$$\begin{aligned} |\mathbf{A}| &= \sum_{i=1}^n (-1)^{i+j} a_{ij} |\mathbf{A}_{\setminus i, \setminus j}| \quad (\forall j \in 1, \dots, n) \\ &= \sum_{j=1}^n (-1)^{i+j} a_{ij} |\mathbf{A}_{\setminus i, \setminus j}| \quad (\forall i \in 1, \dots, n) \end{aligned} \quad (51)$$

# Determinant

Given a matrix

$$\mathbf{A} = \begin{bmatrix} \text{—} & \mathbf{a}_1^\top & \text{—} \\ \text{—} & \mathbf{a}_2^\top & \text{—} \\ & \vdots & \\ \text{—} & \mathbf{a}_n^\top & \text{—} \end{bmatrix} \quad (52)$$

and a set  $\mathbf{S} \subset \mathbb{R}^n$ ,

$$\mathbf{S} = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{a}_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \dots, n\} \quad (53)$$

$|\mathbf{A}|$  is the volume of  $\mathbf{S}$ .



**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \quad (54)$$

**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \quad (54)$$

The matrix rows are:

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad (55)$$

And  $|\mathbf{A}| = -7$

# Determinant

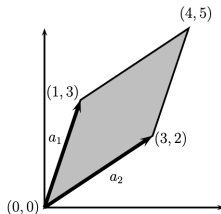
**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \quad (54)$$

The matrix rows are:

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad (55)$$

And  $|\mathbf{A}| = -7$



Properties of determinants:

- ▶ For  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $|\mathbf{A}| = |\mathbf{A}^\top|$
- ▶ For  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ ,  $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- ▶ For  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $|\mathbf{A}| = 0$  iff  $\mathbf{A}$  is singular (i.e. non-invertible).
- ▶ For  $\mathbf{A} \in \mathbb{R}^{n \times n}$  and  $\mathbf{A}$  non-singular,  $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$

# Quadratic form

Given  $\mathbf{A} \in \mathbb{R}^{n \times n}$  and a vector  $\mathbf{x} \in \mathbb{R}^n$ , the *quadratic form* is the scalar value

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{i=1}^n x_i (\mathbf{A} \mathbf{x})_i = \sum_{i=1}^n x_i \left( \sum_{j=1}^n \mathbf{A}_{ij} x_j \right) = \sum_{i=1}^n \sum_{j=1}^n \mathbf{A}_{ij} x_i x_j \quad (56)$$

# Quadratic form

Some properties involving quadratic form:

- ▶ A symmetric matrix  $\mathbf{A} \in \mathbb{S}^n$  is *positive definite* if for a non-zero  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$
- ▶ A symmetric matrix  $\mathbf{A} \in \mathbb{S}^n$  is *positive semi-definite* if for a non-zero  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$
- ▶ A symmetric matrix  $\mathbf{A} \in \mathbb{S}^n$  is *negative definite* if for a non-zero  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0$
- ▶ A symmetric matrix  $\mathbf{A} \in \mathbb{S}^n$  is *negative semi-definite* if for a non-zero  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top \mathbf{A} \mathbf{x} \leq 0$
- ▶ A symmetric matrix  $\mathbf{A} \in \mathbb{S}^n$  is *indefinite* if it is neither positive nor negative semidefinite

n.b. Positive definite and negative definite matrices always have full rank.

# Eigenvalues & eigenvectors

Given  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\lambda \in \mathbb{C}$  is an *eigenvalue* of  $\mathbf{A}$  with corresponding *eigenvector*  $\mathbf{x} \in \mathbb{C}^n$  if

$$\mathbf{Ax} = \lambda \mathbf{x} : \mathbf{x} \neq 0 \quad (57)$$

n.b. The eigenvector is (usually) normalized to have length 1

# Eigenvalues & eigenvectors

Given  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\lambda \in \mathbb{C}$  is an *eigenvalue* of  $\mathbf{A}$  with corresponding *eigenvector*  $\mathbf{x} \in \mathbb{C}^n$  if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} : \mathbf{x} \neq \mathbf{0} \quad (57)$$

n.b. The eigenvector is (usually) normalized to have length 1

We can write all of the eigenvector equations simultaneously as

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \quad (58)$$

where

$$\mathbf{X} \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix}, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n) \quad (59)$$



## Some properties:

- ▶  $\text{tr}\mathbf{A} = \sum_{i=1}^n \lambda_i$
- ▶  $|\mathbf{A}| = \prod_{i=1}^n \lambda_i$
- ▶ The rank of  $\mathbf{A}$  is equal to the number of non-zero eigenvalues of  $\mathbf{A}$ .
- ▶ If  $\mathbf{A}$  is non-singular, then  $1/\lambda_i$  is an eigenvalue of  $\mathbf{A}^{-1}$  with corresponding eigenvector  $\mathbf{x}_i$ , i.e.  $\mathbf{A}^{-1}\mathbf{x}_i = (1/\lambda_i)\mathbf{x}_i$
- ▶ The eigenvalues of a diagonal matrix  $D = \text{diag}(d_1, \dots, d_n)$  are just its diagonal entries  $d_1, \dots, d_n$

**Example:** For  $\mathbf{A} \in \mathbb{S}^n$  with ordered eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ ,

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{A} \mathbf{x} \text{ subject to } \|\mathbf{x}\|_2^2 = 1 \quad (60)$$

is solved with  $\mathbf{x}_1$  corresponding to  $\lambda_1$ . Similarly,

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{A} \mathbf{x} \text{ subject to } \|\mathbf{x}\|_2^2 = 1 \quad (61)$$

is solved with  $\mathbf{x}_n$  corresponding to  $\lambda_n$ .

Given  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ , the *gradient* of  $f$  wrt  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is

$$\nabla_{\mathbf{A}} f(\mathbf{A}) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{11}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{12}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{1n}} \\ \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{21}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{22}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{m1}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{m2}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{mn}} \end{bmatrix} \quad (62)$$

Some properties

- ▶  $\nabla_{\mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \nabla_{\mathbf{x}}f(\mathbf{x}) + \nabla_{\mathbf{x}}g(\mathbf{x})$
- ▶ For  $c \in \mathbb{R}$ ,  $\nabla_{\mathbf{x}}(c f(\mathbf{x})) = c \nabla_{\mathbf{x}}(f(\mathbf{x}))$

# The Hessian

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , the *Hessian* of  $f$  wrt  $\mathbf{x} \in \mathbb{R}^n$  is

$$\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \quad (63)$$

n.b. The Hessian is always symmetric, since  $\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}$

# Least squares

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^m \ni \mathbf{b} \notin \mathcal{R}(\mathbf{A})$ , we want to find  $\mathbf{x} \in \mathbb{R}^n$  as close as possible to  $\mathbf{b}$  (via the Euclidean norm),

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \quad (64)$$

$$= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \quad (65)$$

# Least squares

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^m \ni \mathbf{b} \notin \mathcal{R}(\mathbf{A})$ , we want to find  $\mathbf{x} \in \mathbb{R}^n$  as close as possible to  $\mathbf{b}$  (via the Euclidean norm),

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \quad (64)$$

$$= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \quad (65)$$

Taking the gradient wrt  $\mathbf{x}$ , we have

$$\begin{aligned} \nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b}) &= \nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - \nabla_{\mathbf{x}} 2\mathbf{b}^\top \mathbf{Ax} + \nabla_{\mathbf{x}} (\mathbf{b}^\top \mathbf{b}) \\ &= \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{A}^\top \mathbf{b} \end{aligned} \quad (67)$$

# Least squares

Given  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^m \ni \mathbf{b} \notin \mathcal{R}(\mathbf{A})$ , we want to find  $\mathbf{x} \in \mathbb{R}^n$  as close as possible to  $\mathbf{b}$  (via the Euclidean norm),

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \quad (64)$$

$$= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \quad (65)$$

Taking the gradient wrt  $\mathbf{x}$ , we have

$$\begin{aligned} \nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b}) &= \nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - \nabla_{\mathbf{x}} 2\mathbf{b}^\top \mathbf{Ax} + \nabla_{\mathbf{x}} (\mathbf{b}^\top \mathbf{b}) \\ &= \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{A}^\top \mathbf{b} \end{aligned} \quad (67)$$

Setting this expression equal to zero and solving for  $\mathbf{x}$  gives the normal equations,

$$\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} \quad (68)$$

TODO: go into this.



Some textbooks on linear algebra:

- ▶ *Linear Algebra (Jim Hefferon)*
- ▶ *Introduction to Applied Linear Algebra (Boyd & Vandenberghe)*
- ▶ *Linear Algebra (Cherney, Denton et al.)*
- ▶ *Linear Algebra (Hoffman & Kunze)*
- ▶ *Fundamentals of Linear Algebra (Carrell)*
- ▶ *Linear Algebra (S. Friedberg A. Insel L. Spence)*