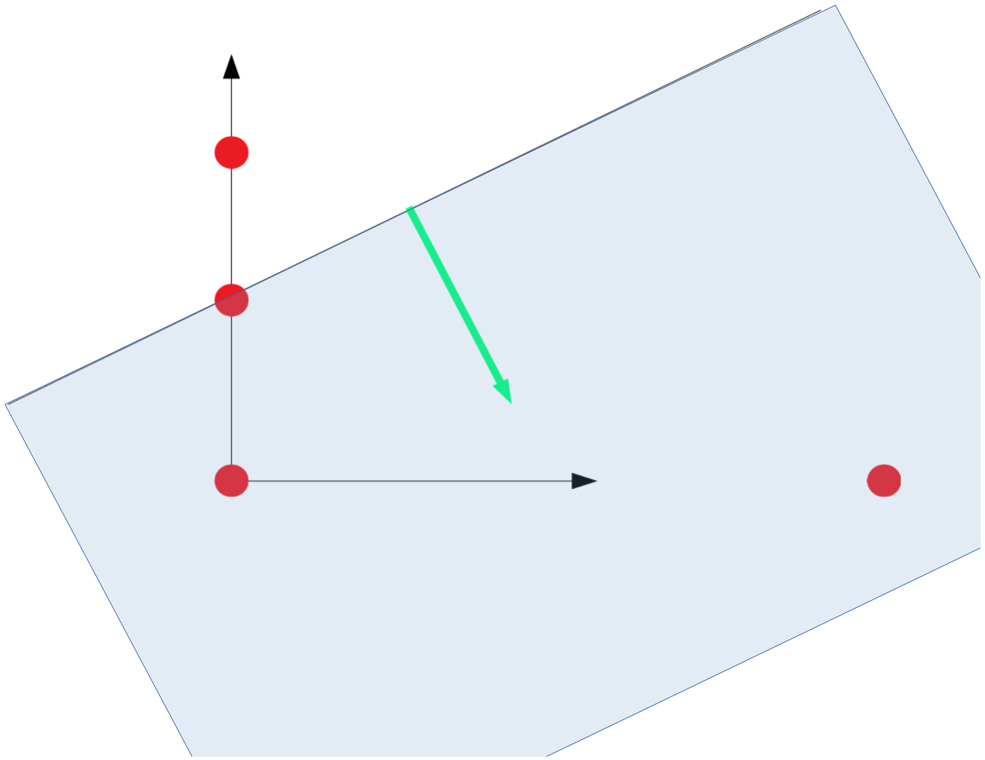
Machine Learning Preliminaries

Jian Liu



Part 1: Calculus & linear algebra

Derivatives

What is the derivative of

$$f(x) = 2x$$

$$f(x) = 2x + x^2$$

Derivatives

What is the derivative of

$f(x) = 2x$ $\quad\quad$ $f'(x) = 2$

$f(x) = 2x + x^2$ $\quad\quad$ $f'(x) = 2 + 2x$

Partial Derivatives

With functions of more than one variable, we can compute the derivative with respect to each variable, treating the others as constants:

$$f(x,y)=2x+3y \qquad \frac{\partial f(x,y)}{\partial x}=2 \qquad \frac{\partial f(x,y)}{\partial y}=3$$

What are the partial derivatives of:

$$f(x,y,z)=2x^2+xy+yz^2$$

Partial Derivatives

With functions of more than one variable, we can compute the derivative with respect to each variable, treating the others as constants:

$$f(x, y) = 2x + 3y \qquad \frac{\partial f(x, y)}{\partial x} = 2 \qquad \frac{\partial f(x, y)}{\partial y} = 3$$

What are the partial derivatives of:

$$f(x, y, z) = 2x^2 + xy + yz^2$$

$$\frac{\partial f(x, y, z)}{\partial x} = 4x + y \qquad \frac{\partial f(x, y, z)}{\partial y} = x + z^2 \qquad \frac{\partial f(x, y, z)}{\partial z} = 2yz$$

Partial Derivatives

The vector of partial derivatives is called the **gradient**:

$$f(x, y, z) = 2x^2 + xy + yz^2$$

$$\nabla f(x, y, z) = \begin{matrix} \dfrac{\partial f(x, y, z)}{\partial x} \\ \dfrac{\partial f(x, y, z)}{\partial y} \\ \dfrac{\partial f(x, y, z)}{\partial z} \end{matrix} = \begin{matrix} 4x + y \\ x + z^2 \\ 2yz \end{matrix}$$

Polynomial

- Polynomial: $p(x) = c_d x^d + c_{d-1} x^{d-1} + \cdots + c_1 x + c_0$

- $(c_d, c_{d-1}, \ldots, c_0)$ completely describes p

- Addition: $(x^2 + 2x - 1) + (3x^3 + 7x) = 3x^3 + x^2 + 9x - 1$

- Multiplication:
$$(x^2 + 2x - 1) \cdot (3x^3 + 7x) = 3x^5 + 4x^3 + 6x^4 + 14x^2 - 7x$$

- Evaluation: p(5) = $c_d 5^d + c_{d-1} 5^{d-1} + \cdots + c_1 5 + c_0$

Polynomial degree

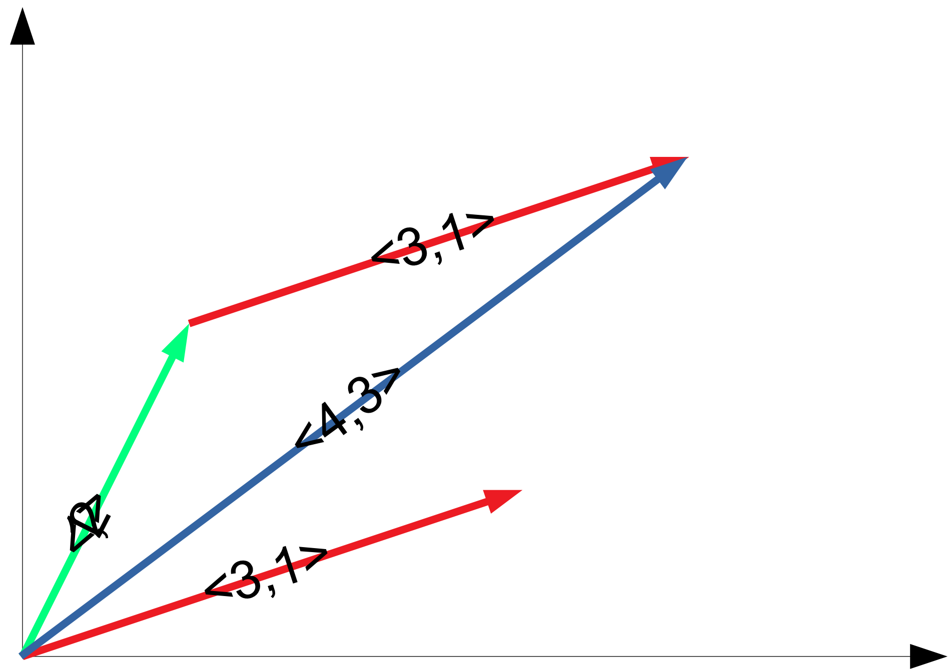
- Polynomial: $p(x) = c_d x^d + c_{d-1} x^{d-1} + \cdots + c_1 x + c_0$
- If $c_d \neq 0$, the degree is d
- If A(x) has degree d and B(x) has degree d, then A(x) + B(x) has degree at most d

Why is the degree at most d?

Vectors

Sum of two vectors: $\langle u_1, u_2, \cdots u_n \rangle + \langle v_1, v_2, \cdots, v_n \rangle = \langle u_1 + v_1, u_2 + v_2, \cdots u_n + v_n \rangle$

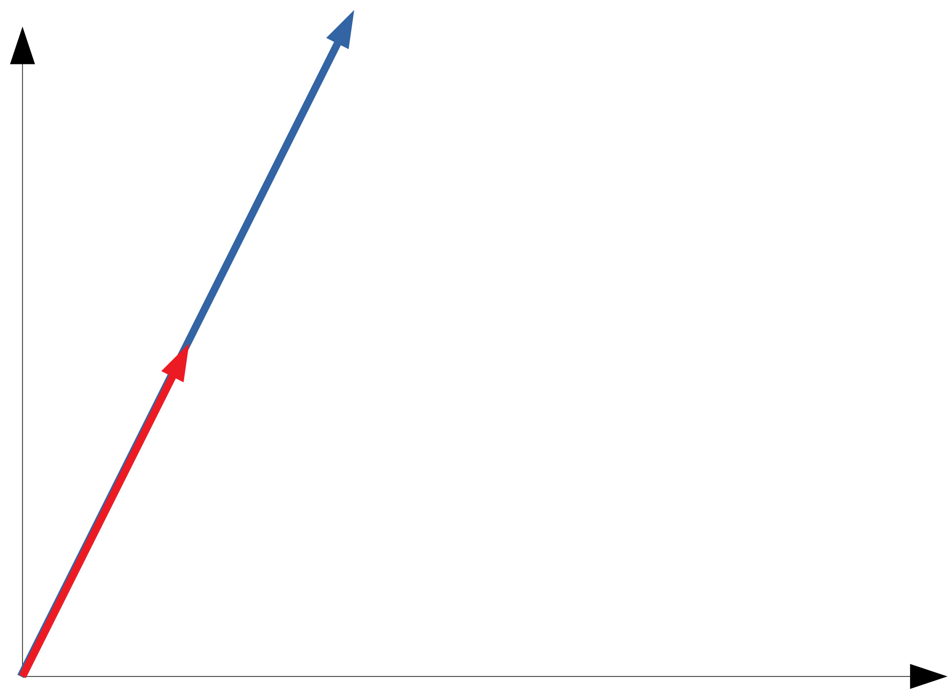
Example: $\langle 1, 2 \rangle + \langle 3, 1 \rangle = \langle 4, 3 \rangle$



Vectors

Multiplication of vector by a constant: $c \cdot \langle v_1, v_2, \cdots, v_n \rangle = \langle cv_1, cv_2, \cdots, cv_n \rangle$

Example: $2 \cdot \langle 1,2 \rangle = \langle 2, 4 \rangle$

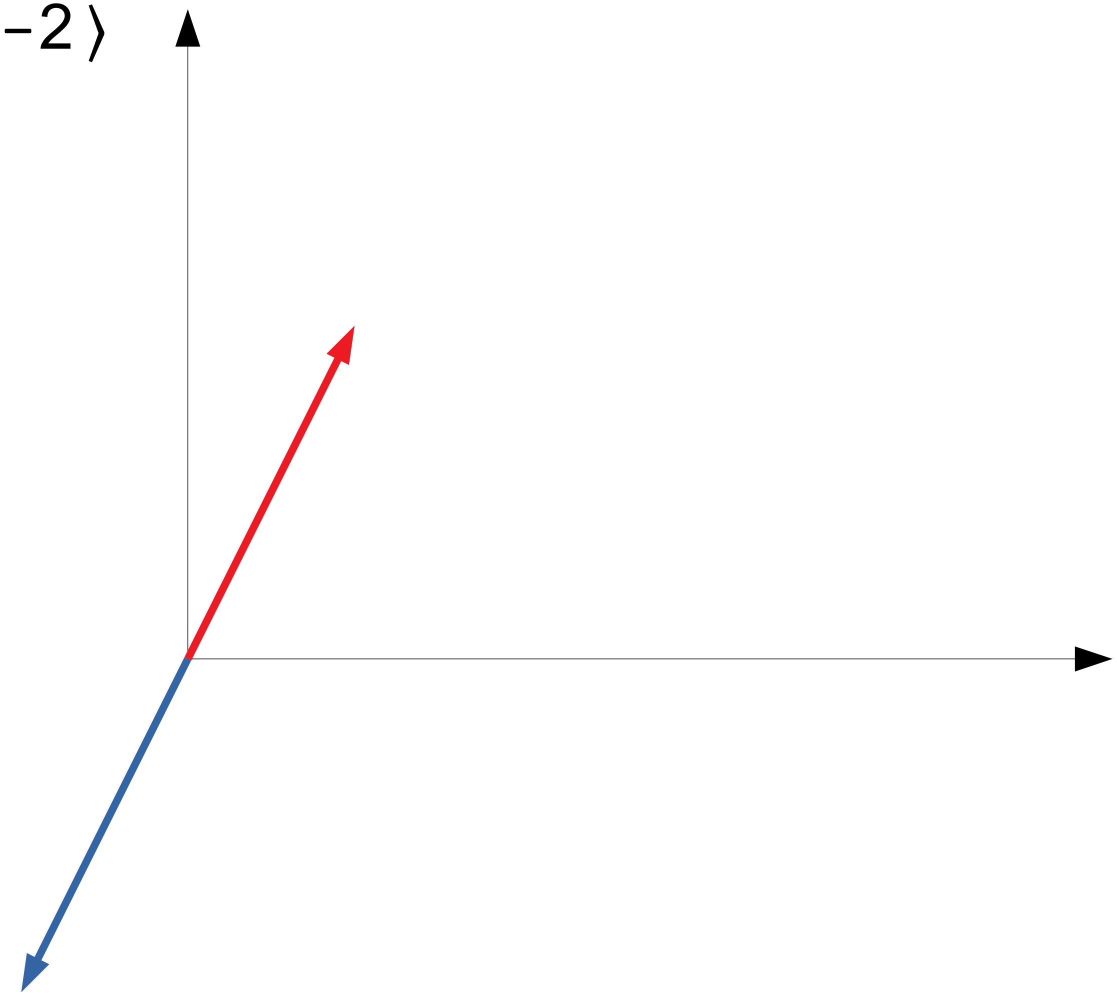


If the constant is different from 1, the length of the vector changes!

Vectors

If the constant is negative, the direction changes:

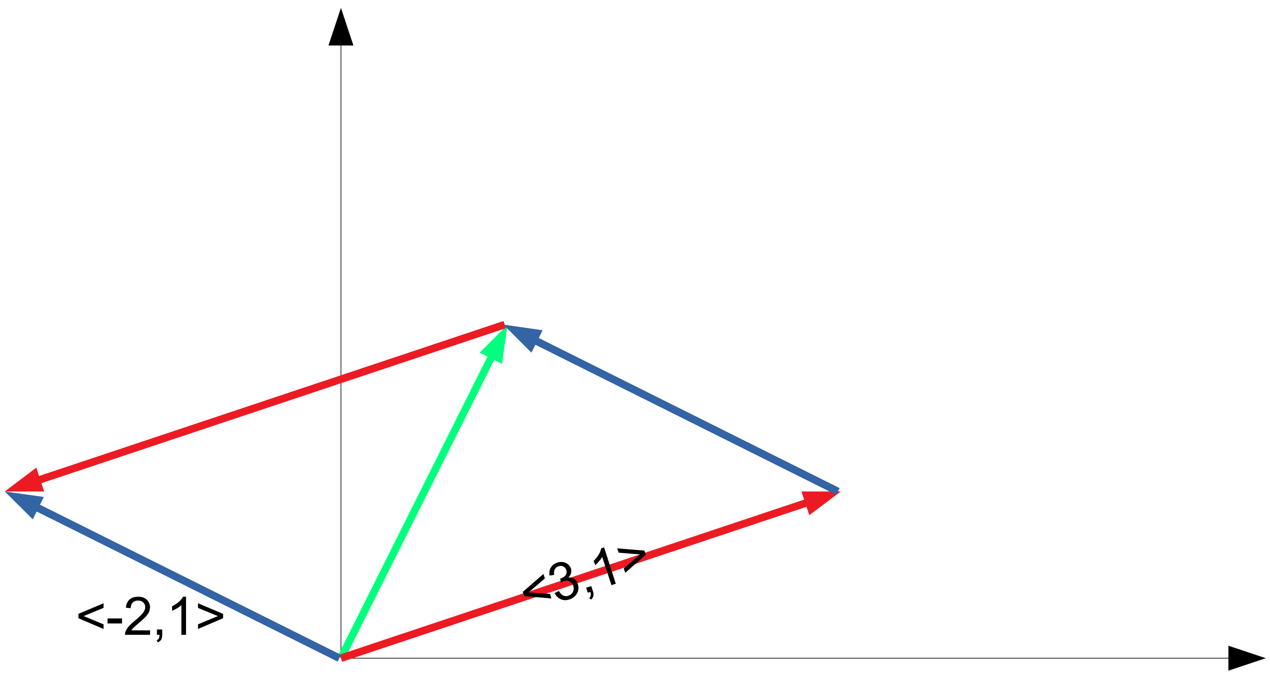
Example: $-1 \cdot \langle 1, 2 \rangle = \langle -1, -2 \rangle$



Vectors

With sum and multiplication by a constant we can subtract two vectors:

Example: $\langle 1,2\rangle + -1 \cdot \langle 3,1\rangle = \langle -2,1\rangle$



Dot product

The dot product of two vectors is:

$$\langle u_1, u_2, \cdots, u_n \rangle \cdot \langle v_1, v_2, \cdots, v_n \rangle = u_1 \cdot v_1 + u_2 \cdot v_2 + \cdots + u_n \cdot v_n = \sum_{i=1}^{n} u_i \cdot v_i$$

Example:

$$\langle 1, 2 \rangle \cdot \langle -1, 3 \rangle = -1 + 6 = 5$$

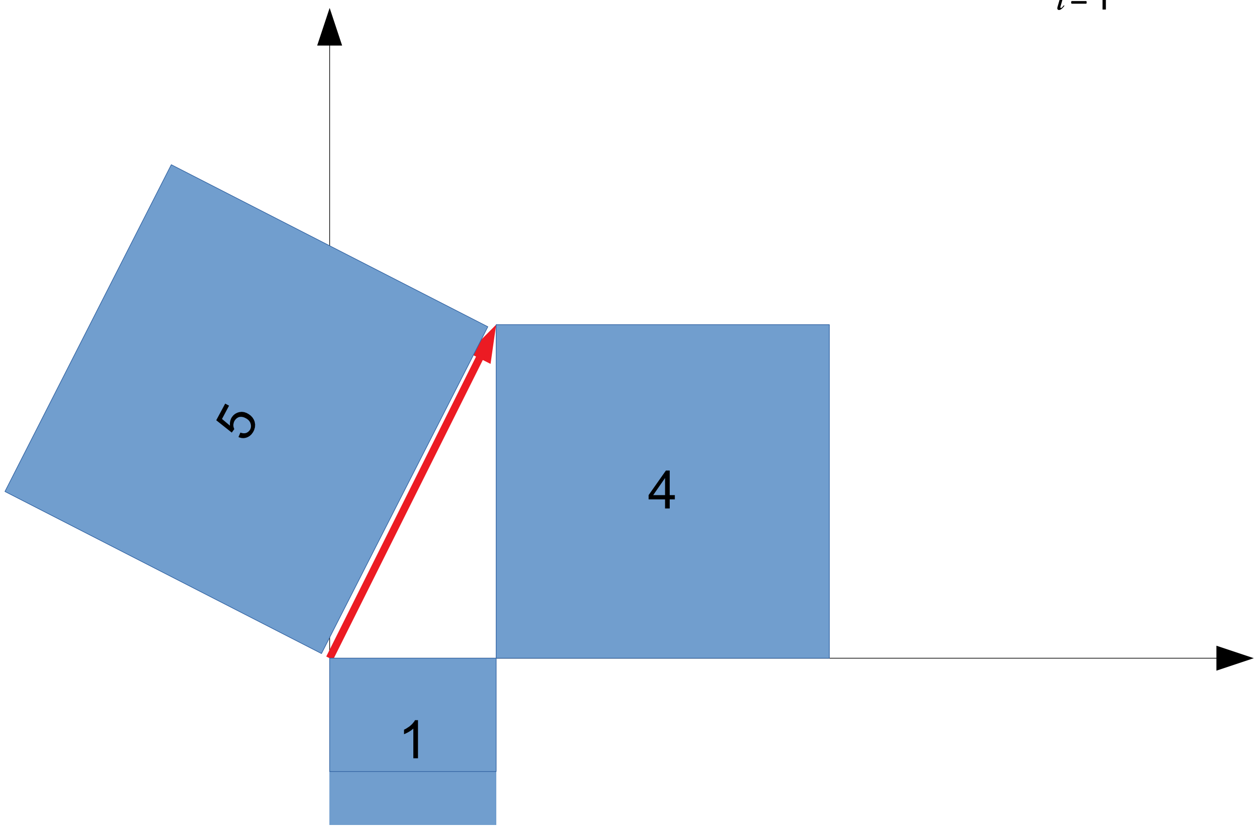
Norm

The **norm** of a vector is the square root of the dot product of the vector with itself:

$$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}\cdot\boldsymbol{v}} = \sqrt{\langle v_1, v_2, \cdots, v_n\rangle \cdot \langle v_1, v_2, \cdots, v_n\rangle} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Example:

$$\|\langle 1, 2\rangle\| = \sqrt{1+4} = \sqrt{5}$$



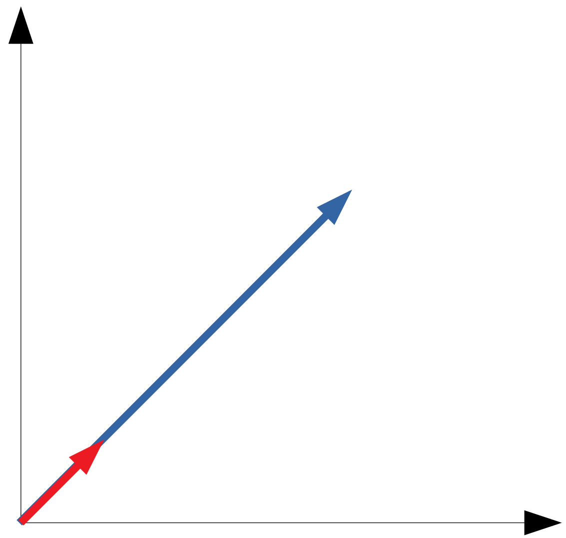
The norm of a vector is the "length" of the vector.

Unit-length vector

If we divide a vector by its norm, we obtain a vector of the same direction, but length 1:

$$\|\boldsymbol{v}\| = \|\langle 2,2\rangle\| = \sqrt{8}$$

$$\boldsymbol{v} = \langle 2,2\rangle \frac{1}{\sqrt{8}} = \langle \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}} \rangle$$



This new vector has norm 1:

$$\|\langle \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}} \rangle\| = \frac{4}{8} + \frac{4}{8} = 1$$

Any vector u parallel to v can be written as:

$$\boldsymbol{u} = d\ \boldsymbol{v}_u$$

Where d is the length of u

For instance: $\langle 5, 5 \rangle = \frac{5\sqrt{8}}{2} \langle \frac{2}{\sqrt{8}}, \frac{2}{\sqrt{8}} \rangle$

What is the norm of <5,5> ?

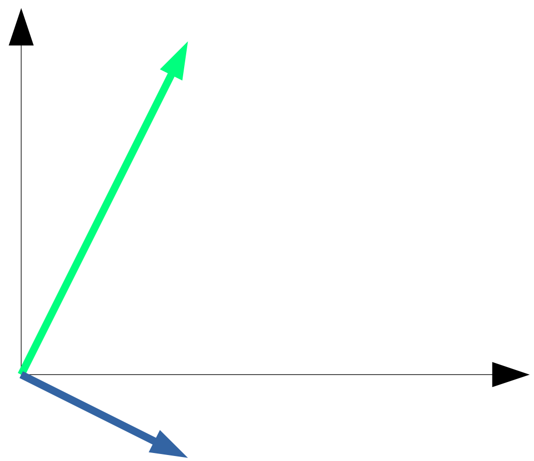
Orthogonal vectors

The dot product of two vectors has the following property:

$\boldsymbol{u} \cdot \boldsymbol{v} = \|u\| \cdot \|v\| \cdot \cos\theta$ Where θ is the angle between the vectors

It follows that the dot product between two non-null vectors is 0 if and only if the vectors are **orthogonal**

$$\langle 1, 2 \rangle \cdot \langle 1, -\frac{1}{2} \rangle = 0$$



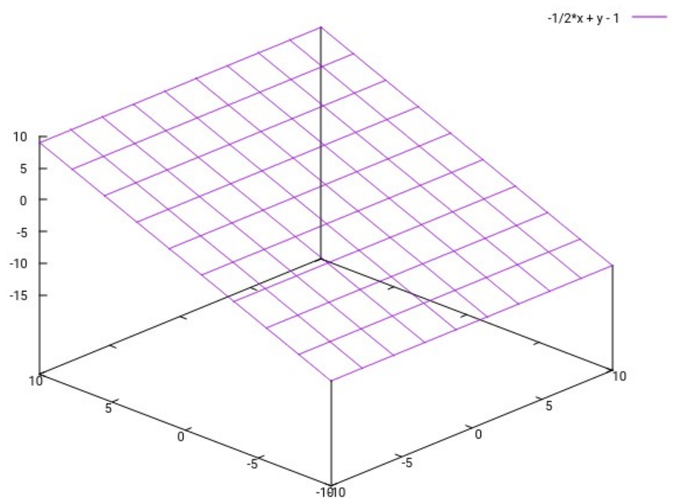
Surfaces

If $f(x_1, x_2, \cdots, x_n)$ is a function of n variables.

The points satisfying the equation: $f(x_{1,} x_{2,} \cdots, x_n) = 0$ lie on a surface in a space of n dimensions.

For example, consider the function: $f(x, y, z) = \frac{1}{2}x - y + z + 1$

The points satisfying the equation $\frac{1}{2}x - y + z + 1 = 0$ lie on a plane:



Surfaces

The points satisfying the inequality $f(x_{1,}x_{2,}\cdots,\ x_n)\geq 0$ lie on one side of the surface

Which side? Let's see on hyperplanes, and in particular in this 2D example, on a straight line:

$$f(x,\ y)=\frac{1}{2}x-y+1\geq 0$$

Let's evaluate the function on some points:

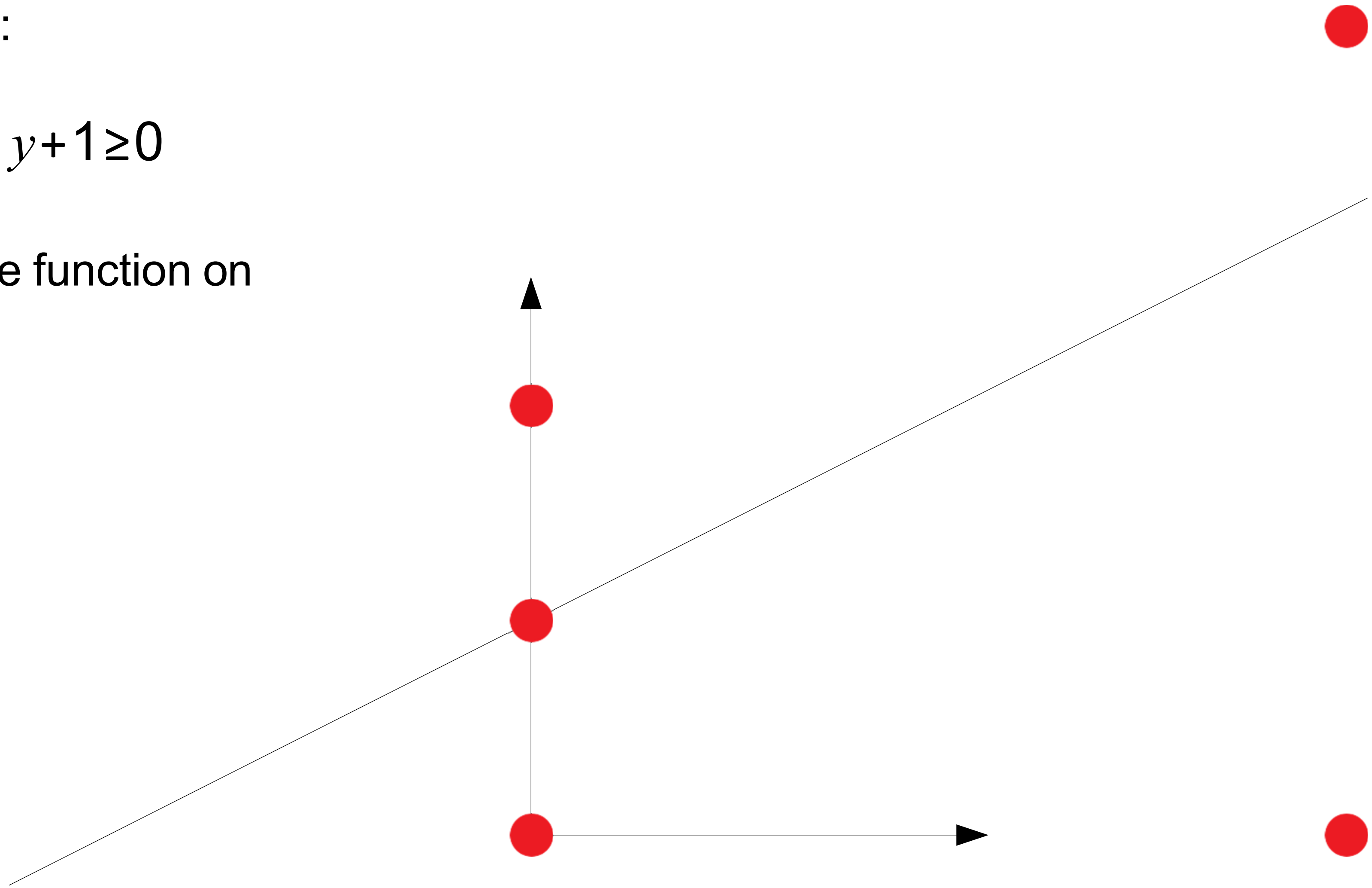
$$f(0,0)=1$$

$$f(0,1)=0$$

$$f(0,2)=-1$$

$$f(4,0)=3$$

$$f(4,4)=-1$$



Surfaces

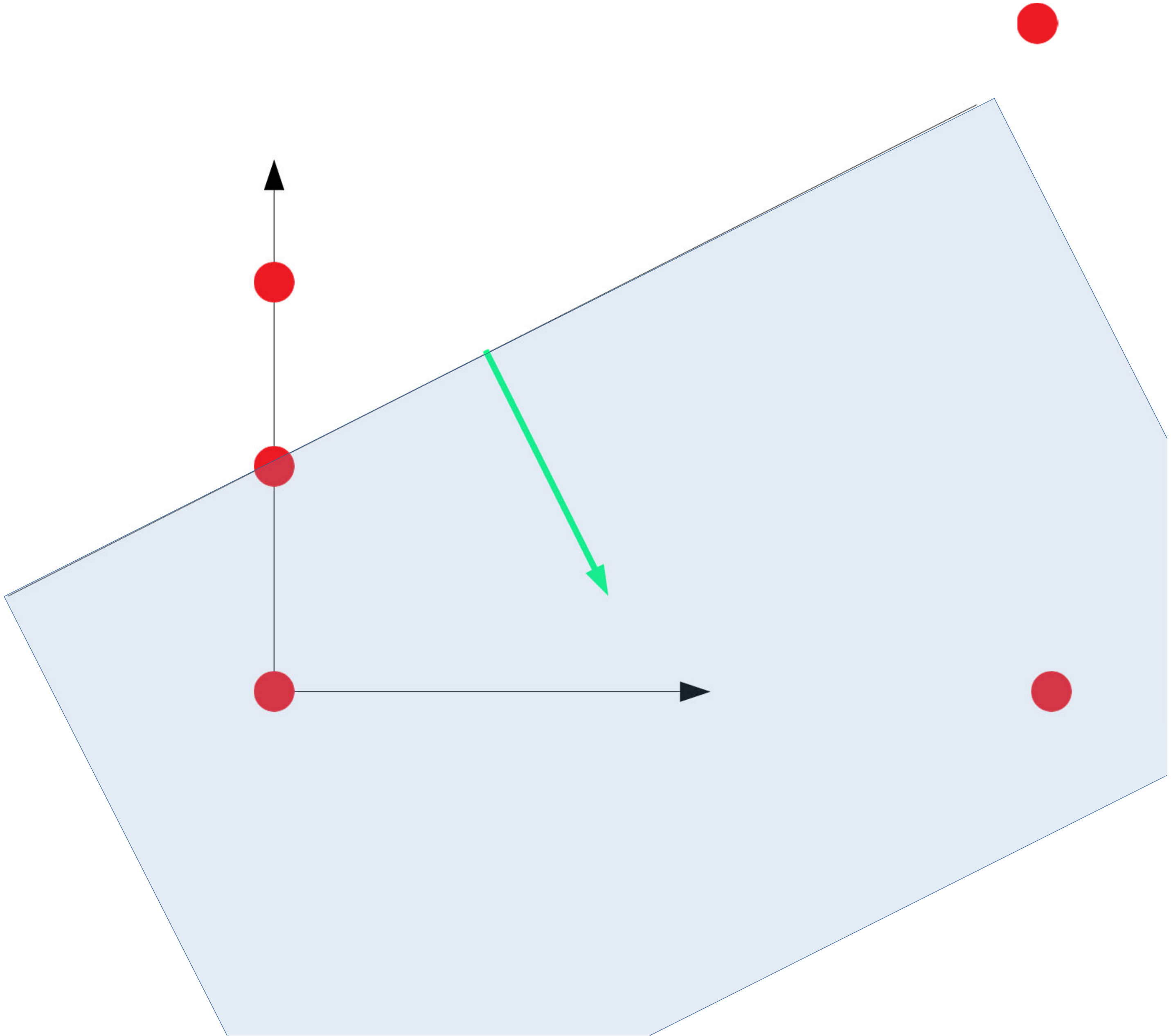
$$f(x,y)=\frac{1}{2}x-y+1\geq 0$$

$f(0,0)=1$ $f(4,0)=3$

$f(0,1)=0$

$f(4,4)=-1$

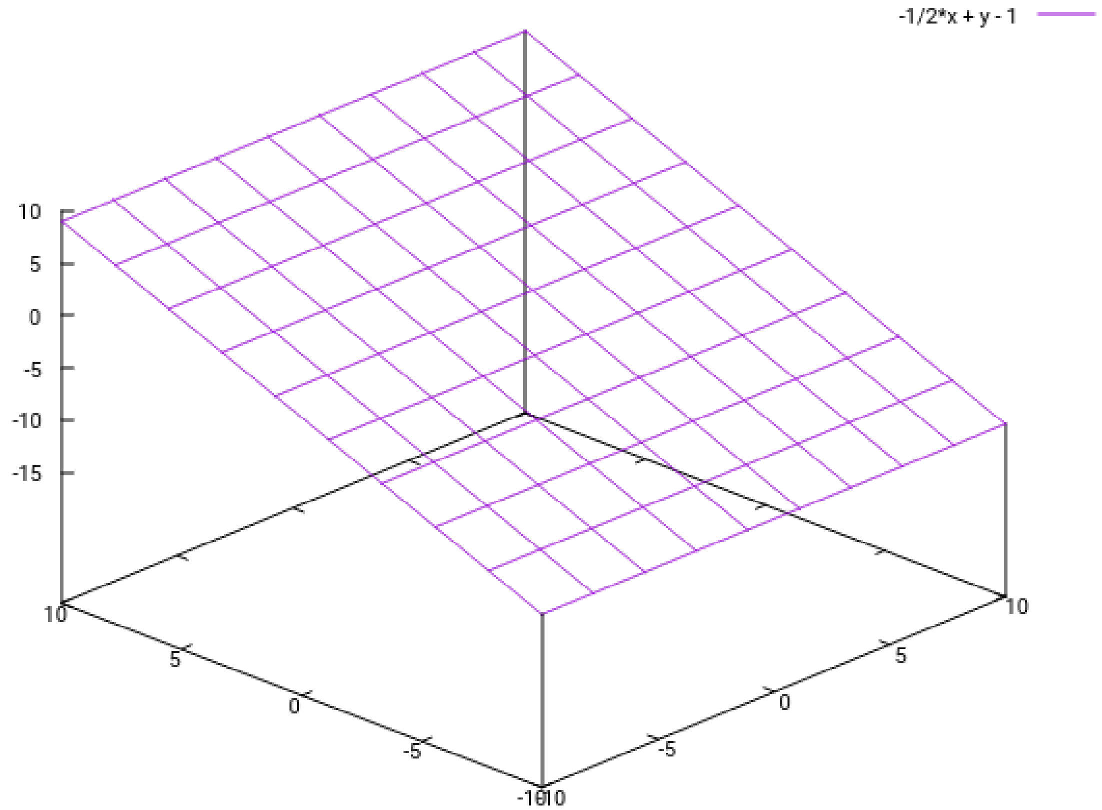
$f(0,2)=-2$



The solution of the inequality is a half-plane, which contains all the points on the same side, with respect to the line, as the vector of parameters multiplied by the variables of the line. The same is true in more dimensions.

Question

Is the inequality $\frac{1}{2}x - y + z + 1 \geq 0$ satisfied by the points above or below the corresponding plane?



Matrix

Matrix notation allows the two equations

$$1x + 1y = b_1$$
$$1x - 1y = b_2$$

to be expressed as

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

or as $\mathbf{Az} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Here $\mathbf{A}, \mathbf{z}, \mathbf{b}$ are respectively: (i) the coefficient matrix; (ii) the vector of unknowns; (iii) the vector of right-hand sides.

Matrices as rectangular arrays

An $m \times n$ matrix $\mathbf{A} = (a_{ij})_{m \times n}$ is a (rectangular) array

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = ((a_{ij})_{i=1}^{m})_{j=1}^{n} = ((a_{ij})_{j=1}^{n})_{i=1}^{m}$$

Note that in a_{ij}, we write the row number i before the column number j.

An $m \times 1$ matrix is a column vector with m rows and 1 column.

A $1 \times n$ matrix is a row vector with 1 row and n columns.

The $m \times n$ matrix $\mathbf{A}$ consists of:

n columns in the form of m-vectors
$\mathbf{a}_j = (a_{ij})_{i=1}^{m} \in \mathbb{R}^m$ for $j = 1, 2, \dots, n$;

m rows in the form of n-vectors
$\mathbf{a}_i^\top = (a_{ij})_{j=1}^{n} \in \mathbb{R}^n$ for $i = 1, 2, \dots, m$.

Transpose

The transpose of the $m \times n$ matrix $\mathbf{A} = (a_{ij})_{m \times n}$ is defined as the $n \times m$ matrix

$$\mathbf{A}^\top = (a_{ij}^\top)_{n \times m} = (a_{ji})_{n \times m} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix}$$

Thus the transposed matrix $\mathbf{A}^\top$ results from transforming each column m-vector $\mathbf{a}_j = (a_{ij})_{i=1}^m$ $(j = 1, 2, \dots, n)$ of $\mathbf{A}$ into the corresponding row m-vector $\mathbf{a}_j^\top = (a_{ji}^\top)_{i=1}^m$ of $\mathbf{A}^\top$.

Equivalently, for each $i = 1, 2, \dots, m$,
the ith row n-vector $\mathbf{a}_i^\top = (a_{ij})_{j=1}^n$ of $\mathbf{A}$
is transformed into the ith column n-vector $\mathbf{a}_i = (a_{ji})_{j=1}^n$ of $\mathbf{A}^\top$.

Either way, one has $a_{ij}^\top = a_{ji}$ for all relevant pairs i, j.

Transpose

$$(\boldsymbol{A}^\top)_{i,j} = A_{j,i}. \tag{2.3}$$

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \boldsymbol{A}^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

Figure 2.1: The transpose of the matrix can be thought of as a mirror image across the main diagonal.

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top. \tag{2.9}$$

Inversion

- Matrix inverse:
$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n. \tag{2.21}$$
- Solving a system using an inverse:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.22}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.23}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.24}$$

- Numerically unstable, but useful for abstract analysis

Invertibility

- Matrix can't be inverted if...
  - More rows than columns
  - More columns than rows
  - Redundant rows/columns ("linearly dependent", "low rank")

Rows & Columns

VERY Important Rule: Rows before columns!

This order really matters.

Reversing it gives a transposed matrix.

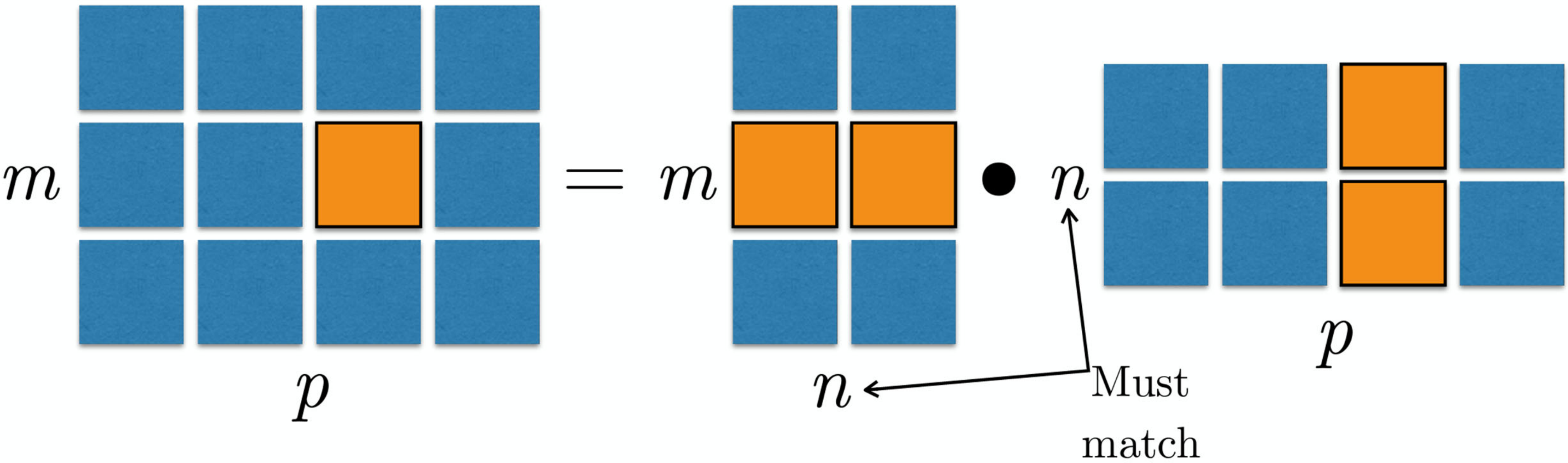
Exercise

Verify that the double transpose of any $m \times n$ *matrix* $\mathbf{A}$ *satisfies* $(\mathbf{A}^\top)^\top = \mathbf{A}$
— i.e., transposing a matrix twice recovers the original matrix.

Matrix (Dot) product

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}. \tag{2.4}$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}. \tag{2.5}$$



Diagonal matrices

A square matrix $\mathbf{A} = (a_{ij})^{n \times n}$ is diagonal
just in case all of its off diagonal elements are 0
— i.e., $i \neq j \Longrightarrow a_{ij} = 0$.

A diagonal matrix of dimension n can be written in the form

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix} = \mathbf{diag}(d_1, d_2, d_3, \dots, d_n) = \mathbf{diag}\,\mathbf{d}$$

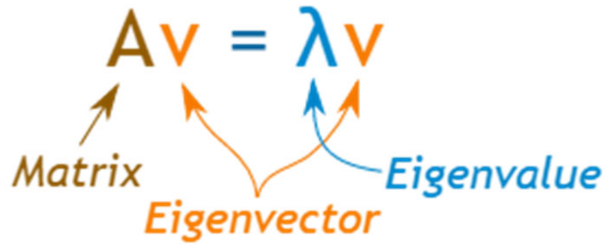
where the n-vector $\mathbf{d} = (d_1, d_2, d_3, \dots, d_n) = (d_i)_{i=1}^n$
consists of the diagonal elements of $\mathbf{D}$.

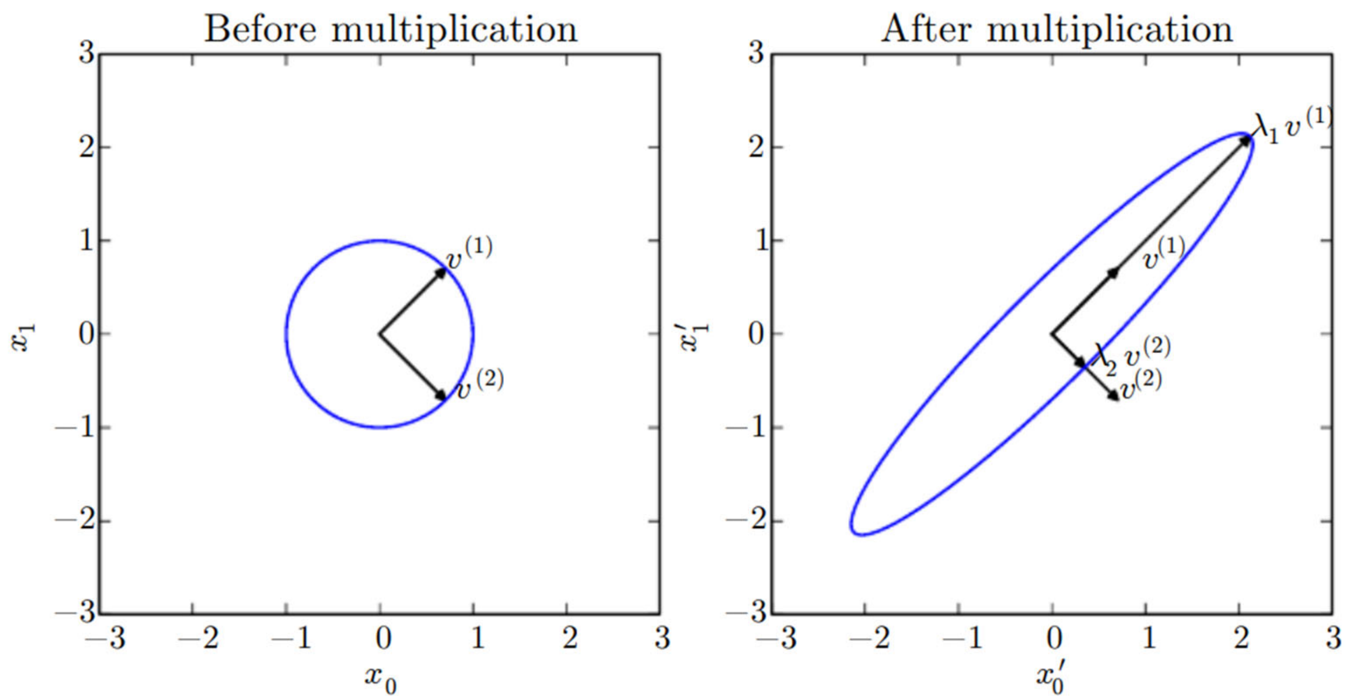
Note that $\mathbf{diag}\,\mathbf{d} = (d_{ij})_{n \times n}$ where each $d_{ij} = \delta_{ij} d_{ii} = \delta_{ij} d_{jj}$.

Obviously, any diagonal matrix is symmetric.

Eigendecomposition

Decompose a matrix into a set of eigenvectors and eigenvalues.





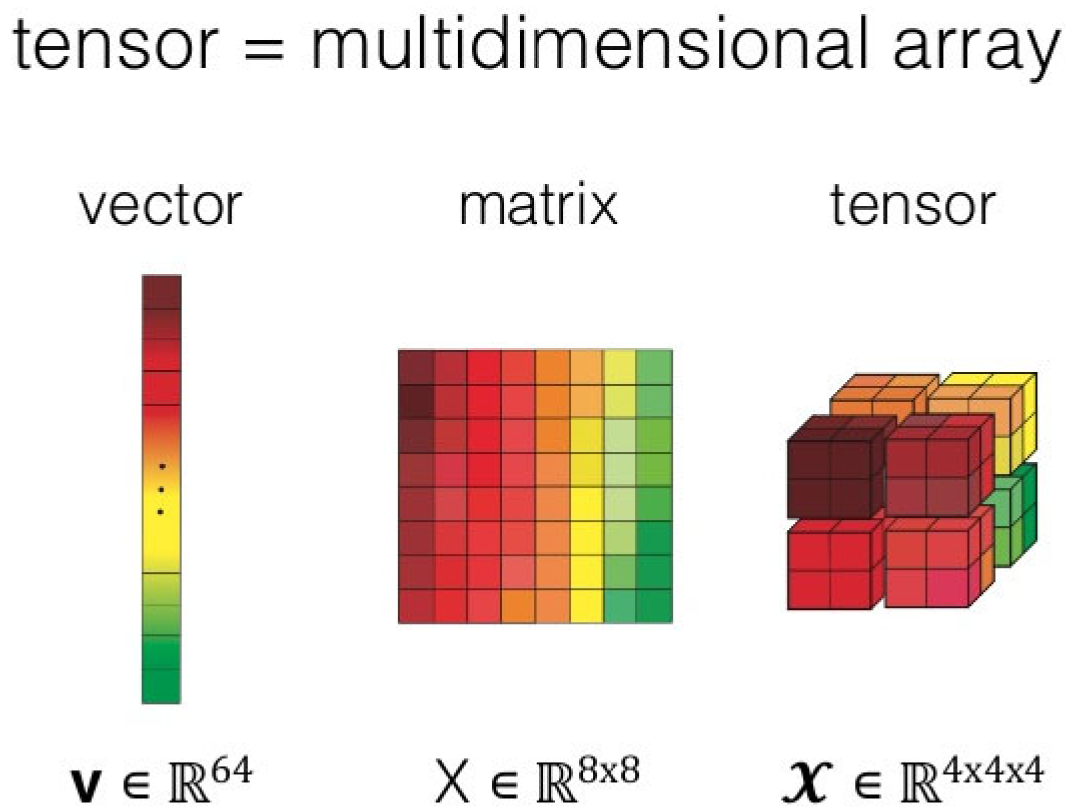
Singular Value Decomposition

- Similar to eigendecomposition
- More general; matrix need not be square

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}. \tag{2.43}$$

Tensors

In some cases we will need an array with more than two axes. In the general case, an array of numbers arranged on a regular grid with a variable number of axes is known as a tensor



Tensors

In some cases we will need an array with more than two axes. In the general case, an array of numbers arranged on a regular grid with a variable number of axes is known as a tensor

- A tensor is an array of numbers, that may have
  - zero dimensions, and be a scalar
  - one dimension, and be a vector
  - two dimensions, and be a matrix
  - or more dimensions.

Learning linear algebra

- Do a lot of practice problems
- Start out with lots of summation signs and indexing into individual entries
- Eventually you will be able to mostly use matrix and vector product notation quickly and easily