CH 2

LINEAR ALGEBRA

BOOK NOTES DEEP LEARNING by Ian Goodfellow, Yoshua Bengio, and Aaron Courville.

OTHER READING MATERIAL:
THE MATRIX COOKBOOK by Petersen and Pedersen
LINEAR ALGEBRA by Shilov

SCALARS

(N) lowercase italics

$i = 1, \dots, m$
$j = 1, \dots, n$

- a single number
- specify type: real valued $r \in \mathbb{R}$
  natural number $m \in \mathbb{N}$

VECTORS

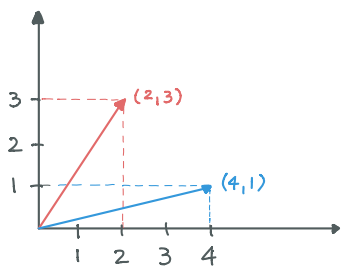
(N) lowercase bold

- array of numbers arranged in order
- identifying points in space
  - each element is a coordinate along a different axis
- defined by magnitude and direction

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$ x_1 → first element, x_2 → second element

EXAMPLE: 2D vector

$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\boldsymbol{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ $\boldsymbol{x} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$



$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

set of elements of $\boldsymbol{x}$: $S = \{1, 2, 4\}$

$$\boldsymbol{x}_S = \begin{bmatrix} x_1 \\ x_2 \\ x_4 \end{bmatrix}$$

complement of S

$$\boldsymbol{x}_{-S} = \begin{bmatrix} x_3 \\ x_5 \end{bmatrix}$$

OPERATIONS:
- addition
- substraction
- scaling
- dot product
- length
- cross product

MATRICES

(N) uppercase bold

- rectangular 2D array of real numbers
- 2 indices: m rows; n columns

▶ need to specify the values of its components

Example 3x2 matrix

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \in \begin{bmatrix} \mathbb{R} & \mathbb{R} \\ \mathbb{R} & \mathbb{R} \\ \mathbb{R} & \mathbb{R} \end{bmatrix} = \mathbb{R}^{3 \times 2}$$

(m rows, n columns)

$$\boldsymbol{A}_{1,:} = \begin{bmatrix} A_{1,1} & A_{1,2} \end{bmatrix} \qquad \boldsymbol{A}_{:,1} = \begin{bmatrix} A_{1,1} \\ A_{2,1} \\ A_{3,1} \end{bmatrix}$$

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$; $A_{i,j}$

$\boldsymbol{A}_{i,:}$ i-th row

$\boldsymbol{A} = \begin{bmatrix} - & a_m^T & - \end{bmatrix}$

$\boldsymbol{A}_{:,j}$ j-th column

$\boldsymbol{A} = \begin{bmatrix} | \\ a_n \\ | \end{bmatrix}$

$f(\boldsymbol{A})_{i,j}$ function on elements of $\boldsymbol{A}$

OPERATIONS:

addition	$A+B$
substraction	$A-B$
scaling	$\propto A$
matrix product	AB
matrix inverse	A^{-1}
trace	$Tr(A)$
determinant	$\lvert A \rvert$

TENSORS

- array of numbers with more than two axes
- $\boldsymbol{A}_{i,j,k}$

@ Ha Cirovic Donev

OPERATIONS

TRANSPOSE

$$(A^T)_{i,j} = A_{j,i}$$

$$\underset{m}{\left[\quad\right]}^T = \underset{n}{\left[\quad\right]}$$

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

- mirror image of the matrix across the main diagonal ➡ flip rows & columns
- vector as matrix with one column
  $x = [x_1, x_2, \dots, x_n]^T$
- scalar : a matrix with one single entry
  $a = a^T$

PROPERTIES

$$(A^T)^T = A \qquad (A+B)^T = A^T + B^T$$
$$(AB)^T = A^T B^T \qquad (\alpha A)^T = \alpha A^T$$

α ↓ scalar (real valued)

ADDITION

$$C_{i,j} = A_{i,j} + B_{i,j}$$

- A and B have to be the same size

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} + \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix} = \begin{bmatrix} A_{1,1}+B_{1,1} & A_{1,2}+B_{1,2} \\ A_{2,1}+B_{2,1} & A_{2,1}+B_{2,2} \end{bmatrix}$$

$$A + B = C$$

SUBSTRACTION → entry-wise as addition

PROPERTIES

$$A + B = B + A$$
$$A + (B + C) = (A + B) + C$$

BROADCASTING matrix-vector addition

$$C = A + b \qquad C_{i,j} = A_{i,j} + b_j$$

- vector b is added to each row of A

EXAMPLES

✓
$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 \\ b_1 & b_2 \end{bmatrix} = \dots$$

$$\begin{bmatrix} A_{1,1} & A_{1,1} \\ A_{2,1} & A_{2,1} \\ A_{3,1} & A_{3,1} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 \\ b_1 & b_2 \\ b_1 & b_2 \end{bmatrix} = \dots$$

✗
$$\begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & z \\ b_1 & b_2 & z \end{bmatrix} = \dots$$

$(2,3) \neq (1,2)$

SCALING

$$B = \alpha A_{i,j} = B_{i,j} \qquad \alpha \in \mathbb{R}$$

$$B_{i,j} = \begin{bmatrix} \alpha A_{1,1} & \dots & \alpha A_{1,n} \\ \alpha A_{2,1} & \dots & \alpha A_{2,n} \\ \dots & \ddots & \dots \\ \alpha A_{m,1} & \dots & \alpha A_{m,n} \end{bmatrix}$$

- scaling A by α

PROPERTIES

$$\alpha(\beta A) = (\alpha\beta) A$$
$$(\alpha + \beta) A = \alpha A + \beta A$$
$$\alpha(A + B) = \alpha A + \alpha B$$
$$A(\alpha B) = \alpha(AB) = (\alpha A) B$$

PRODUCT

$$C_{i,j} = A_{i,k} B_{k,j}$$

A B
$i \times k$! $k \times j$
equal
dimension of C

$$C_{i,j} = \text{row}_i(A)\ \text{col}_j(B) = \sum_{k=1}^{p} A_{i,k} B_{k,j}$$

- **dot** product $x^T y$
  x, y : same dimensions

$$\begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,p} \\ A_{2,1} & A_{2,2} & \dots & A_{2,p} \\ A_{3,1} & A_{3,2} & \dots & A_{3,p} \\ \dots & \dots & \ddots & \dots \\ A_{m,1} & A_{m,2} & \dots & A_{m,p} \end{bmatrix} \begin{bmatrix} B_{1,1} & B_{1,2} & \dots & B_{1,n} \\ B_{2,1} & B_{2,2} & \dots & B_{2,n} \\ \dots & \dots & \ddots & \dots \\ B_{p,1} & B_{p,2} & \dots & B_{p,n} \end{bmatrix} = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,n} \\ C_{2,1} & C_{2,2} & \dots & C_{2,n} \\ \dots & \dots & \ddots & \dots \\ C_{m,1} & C_{m,2} & \dots & C_{m,n} \end{bmatrix}$$

$$C_{2,2} = A_{2,1} B_{1,2} + A_{2,2} B_{2,1} + \dots + A_{2,p} B_{p,2}$$

PROPERTIES

$$A(B + C) = AB + AC$$ ✓ DISTRIBUTIVE
$$A(BC) = (AB)C$$ ✓ ASSOCIATIVE
$$AB \overset{\times}{=} BA$$ ✗ COMMUTATIVE
$$x^T y = y^T x$$ ✓ COMMUTATIVE

HADAMARD product

- element-wise
  $A \odot B$

@ Ha Cirovic Donev

IDENTITY matrix

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

└ of order n

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

└ preserves n-dim vectors $I_n x = x \; ; \; \forall x \in \mathbb{R}^n$

└ preserves $m \times n$ matrices $I_m A = A I_n = A \; ; \; A \in \mathbb{R}^{m \times n}$

INVERSE

$$A^{-1} \; ; \; A^{-1} A = I_n$$

└ NON SINGULAR

└ not all matrices have inverses (e.g. non-square matrices)

└ inverse exists ▶ full rank matrix

PROPERTIES

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

TRACE

$$Tr(A) = \sum_{i=1}^{n} A_{ii}$$

└ $A \in \mathbb{R}^{n \times n}$

└ sum of diagonal elements of a matrix

USES:

└ easier specification of some operations, without the need for summation notation

example: FROBENIUS NORM

$$\|A\|_F = \sqrt{Tr(AA^T)}$$

PROPERTIES

$Tr(A) = Tr(A^T) \; ; \; A \in \mathbb{R}^{n \times n}$

$Tr(A+B) = Tr(A) + (B) \; ; \; A, B \in \mathbb{R}^{n \times n}$

$Tr(\alpha A) = \alpha Tr(A) \; ; \; A \in \mathbb{R}^{n \times n}, \; \alpha \in \mathbb{R}$

$Tr(AB) = Tr(BA) \; ;$ s.t. AB is square

$Tr(ABC) = Tr(BCA) = Tr(CAB)$ s.t. ABC is square

$\alpha = Tr(\alpha)$ scalar is its own trace

SYSTEM OF LINEAR EQUATIONS

$A x = b \; ; \; A \in \mathbb{R}^{m \times n} \quad b = \mathbb{R}^m$

$x \in \mathbb{R}^n$ unknown

- using matrix inverse to solve

$$A x = b$$
$$A^{-1} A x = A^{-1} b$$
$$I_n x = A^{-1} b$$
$$x = A^{-1} b$$

! if A^{-1} exists, i.e. A is nonsingular

- compute A^{-1} once, so that whenever we change b we find the corresponding x
- represented with only limited precision → should not be used in practice

system of m linear equations in n unkowns

$$A_{1,1} x_1 + A_{1,2} x_2 + \cdots + A_{1,n} x_n = b_1$$
$$A_{2,1} x_1 + A_{2,2} x_2 + \cdots + A_{2,n} x_n = b_2$$
$$\cdots$$
$$A_{m,1} x_1 + A_{m,2} x_2 + \cdots + A_{m,n} x_n = b_m$$

each row of A and each element of b provide a constraint

LINEAR DEPENDENCE and Span

set of vectors $\{v^{(1)}, v^{(2)}, \dots, v^{(n)}\}$

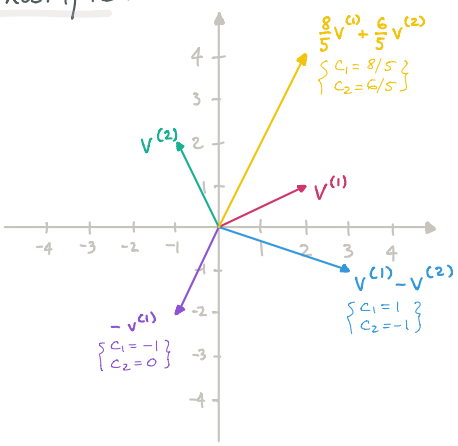
LINEAR COMBINATION

$\sum_i c_i v^{(i)}$

- scale vectors $v^{(i)}$ by some scalar c_i
- $c_1, \dots, c_n \in \mathbb{R}$
- $v^{(1)}, \dots, v^{(n)} \in \mathbb{R}^n$

Solutions to $Ax = b$
- ✓ $= 1$ $\Rightarrow$ A^{-1} exists; A is square; all columns are linearly independent
- ✓ $= \infty$
- ✓ $= 0$
- ✗ $1 < \dots < \infty$

Example:



$v^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ $v^{(2)} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$c_1 v^{(1)} + c_2 v^{(2)}$

$(1)v^{(1)} + (-1)v^{(2)} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$

$(-1)v^{(1)} + (0)v^{(2)} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$

$(8/5)v^{(1)} + (6/5)v^{(2)} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$

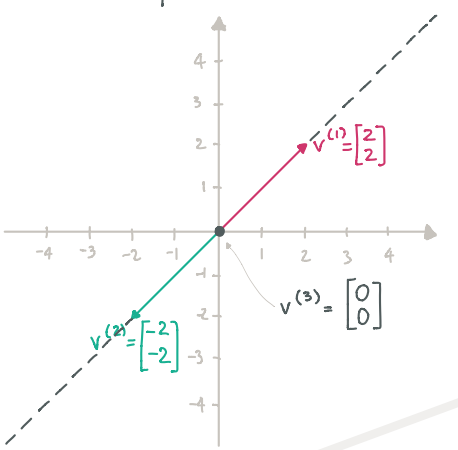
What is the set of all the vectors that we can compute by linear combinations?

↳ all $\mathbb{R}^2$ can be represented by $v^{(1)}$ and $v^{(2)}$.

⇒ **SPAN**

$\text{span}(v^{(1)}, v^{(2)}) = \mathbb{R}^2$

Example: $\text{span}(v^{(1)}, v^{(2)}) =$ line
$\text{span}(v^{(1)}) =$ line
$\text{span}(v^{(3)}) = 0$



$\text{span}(\{v^{(1)}, \dots, v^{(n)}\}) = \{u : u = \sum_{i=1}^{n} c_i v^{(i)}, c_i \in \mathbb{R}\}$

$Ax = \sum_i x_i A_{:,i}$

↳ if b is in the span of the columns of A then $Ax = b$ has a solution. ⇒ **COLUMN SPACE** **RANGE of A**

↳ to have solutions for all values of $b \in \mathbb{R}^m$, then the column space of A has to be all of $\mathbb{R}^m$.

↳ A must have at least m columns, i.e. $n \geq m$ else, the dimensionality of the column space would be less than m.
- ↳ only a necessary condition for every b to have a solution
- ↳ not sufficient condition

Examples:

$v^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$; $v^{(2)} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$ → $v^{(1)} + 2 \cdot v^{(1)}$

LINEAR DEPENDENCE - one of the vectors in a set can be represented by some combination of other vectors in the set
↳ decreases the column space

Set of vectors $\{v^{(1)}, v^{(2)}, \dots, v^{(n)}\}$ are **linearly dependent iff**

$c_1 v^{(1)} + c_2 v^{(2)} + \dots + c_n v^{(n)} = 0$

for some c_i's where at least one is nonzero.

! If each vector adds another dimension to the span then they are linearly dependent

SINGULAR
→ a matrix with lin. de endent columns

LINEAR INDEPENDENCE

- no vector in the set is a linear combination of other vectors
- A contains at least one set of exactly m linearly independent columns
  ⟹ the span or column space of A is $\mathbb{R}^m$

necessary & sufficient condition for $Ax = b$ to have solution for every b.

- **COLUMN RANK** of A $\quad A \in \mathbb{R}^{m \times n}$
  - dimension of the column space of A
  - size of the largest number of columns of A that form a linearly independent set.
- column rank = row rank ; rank(A)

PROPERTIES $\quad A \in \mathbb{R}^{m \times n}$

$\text{rank}(A) \leq \min(m, n)$
$\text{rank}(A) = \min(m, n)$ $\quad$ FULL RANK
$\text{rank}(A) = \text{rank}(A^T)$
$\text{rank}(BA) \leq \min(\text{rank}(A), \text{rank}(B))$; $B \in \mathbb{R}^{n \times p}$
$\text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B)$; $B \in \mathbb{R}^{m \times n}$

NORMS

- measure of vector size or length
  - norm is any function $f : \mathbb{R}^n \to \mathbb{R}$

PROPERTIES $\quad x, y \in \mathbb{R}^n$; $\alpha \in \mathbb{R}$

$f(x) \geq 0$
$f(x) = 0 \iff x = 0$
$f(x+y) \leq f(x) + f(y)$ $\quad$ TRIANGLE INEQUALITY
$\forall \alpha \in \mathbb{R}$ $\quad f(\alpha x) = |\alpha| f(x)$ $\quad$ HOMOGENEITY

L^2 : Euclidean norm

$\|x\|$

- distance from the origin to the point x.

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

Squared L^2 norm

$x^T x$

- more convenient mathematically and computationally
- it increases very slowly near the origin

L^p norm

$p \in \mathbb{R}$; $p \geq 1$

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

L^0 norm

- number of nonzero elements
- ✗ not ok terminology ; scaling the vector does not change the number of nonzero elements
- ✓ use L^1 instead

L^1 norm

$$\|x\|_1 = \sum_i |x_i|$$

- grows at the same rate at all locations
- used when the difference between zero and nonzero elements is very important.

L^∞ norm (max norm)

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

- largest magnitude in the vector

FROBENIUS norm

- size of a matrix

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}}$$

DOT PRODUCT of two vectors

$$x^T y = \|x\|_2 \, \|y\|_2 \cos(\theta)$$

- where θ is the angle between x and

SPECIAL KIND OF MATRICES and VECTORS

DIAGONAL matrices

▷ entries only along the main diagonal
▷ example : Identity matrix
▷ multiplying by a diagonal matrix is computationally efficient
$\text{diag}(\boldsymbol{v})\boldsymbol{x} = \boldsymbol{v} \odot \boldsymbol{x}$

▷ it doesn't have to be square
↳ no inverse

$\boldsymbol{D}$ is diagonal $\Leftrightarrow$ $D_{i,j} = 0 \quad \forall i \neq j$

$\text{diag}(\boldsymbol{v})$ ▷ square diagonal matrix with vector $\boldsymbol{v}$ entries

▷ $\text{diag}(\boldsymbol{v})^{-1}$ exists only if every diagonal entry is nonzero.

$$\text{diag}(\boldsymbol{v})^{-1} = \text{diag}\left[\left(\frac{1}{v_1}, \dots, \frac{1}{v_n}\right)\right]^T$$

SYMMETRIC

$$A^T = A \; ; \; A_{i,j} = A_{j,i}$$

ANTI-SYMMETRIC $A = -A^T$

if $A \in \mathbb{R}^{n \times n}$ then
↳ $A + A^T$ is symmetric
↳ $A - A^T$ is anti-symmetric

ORTHOGONAL VECTORS

↳ $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal to each other if

$$\boldsymbol{x}^T\boldsymbol{y} = 0$$

↳ orthogonal matrix
▷ square
▷ rows and columns are mutually orthonormal

$$A^T A = A A^T = I$$
$$\Downarrow$$
$$A^{-1} = A^T$$

↳ inverse is cheap to compute
↳ $\det(\boldsymbol{A}) = \pm 1$ if $\boldsymbol{A}$ is orthogonal

UNIT VECTOR

▷ vector with unit norm
$\|\boldsymbol{x}\|_2 = 1$

and

ORTHONORMAL

If $\boldsymbol{A}$ is symmetric, $\exists$ orthogonal matrix $\boldsymbol{P}$ s.t. $\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{D}$
▷ the eigenvalues of $\boldsymbol{A}$ are on the main diagonal of $\boldsymbol{D}$.

EIGENDECOMPOSITION

- shows functional properties of matrices
- decompose matrix into

Only the scale of v is altered

EIGENVECTORS & **EIGENVALUES**

- EIGENVECTORS => vectors that remain on their span after the transformation => axis of rotation
- EIGENVALUES => the factor by which vectors are scaled (stretch or squash). => each eigenvector has an associated eigenvalue

$$Av = \lambda v$$

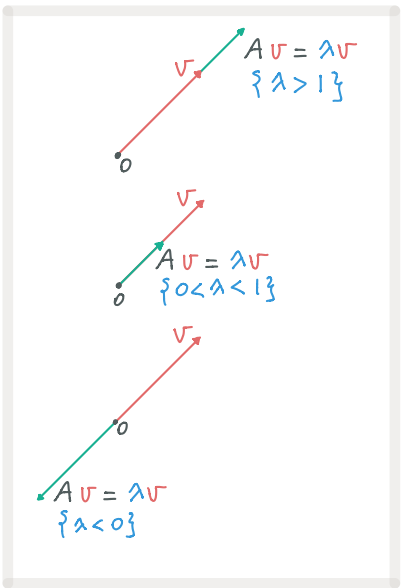
A : matrix of transformation; A is square

v : eigenvector ----------> sv, $s \in \mathbb{R}$, $s \neq 0$ => also eigenvector, with the same eigenvalue

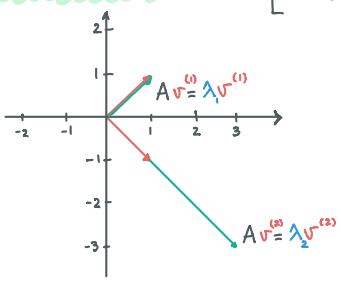
λ : eigenvalue

$$v^T A = \lambda v^T \quad \text{LEFT EIGENVECTOR}$$

Scaling space:



Example: $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$



$$Av^{(1)} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$Av^{(2)} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Find eigenvalues:

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} = (2-\lambda)(2-\lambda) - 1$$

$$= 4 - 4\lambda + \lambda^2 - 1 = (\lambda - 1)(\lambda - 3)$$

$$\Rightarrow \lambda_1 = 1 \quad \text{and} \quad \lambda_2 = 3$$

Find eigenvectors:

$\lambda_1 = 1$ $\quad (A - \lambda I)v = 0$

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Big\} \; v_1 - v_2 = 0 \Rightarrow v^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$\lambda_2 = 3$ $\quad (A - \lambda I)v = 0$

$$\begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Big\} \; -v_1 - v_2 = 0 \Rightarrow v^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

EIGENDECOMPOSITION

$\{v^{(1)}, v^{(2)}, \dots, v^{(n)}\}$ linearly indep. eigenvectors

$\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ eigenvalues

$$V = \begin{bmatrix} v^{(1)} & v^{(2)} & \dots & v^{(n)} \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_n \end{bmatrix}^T$$

$$A = V \, \mathrm{diag}(\lambda) \, V^{-1}$$

! not every matrix can be decomposed

! decomposition can contain complex numbers

! every $A \in \mathbb{R}^{n \times n}$ can be decomposed as

$$A = Q \Lambda Q^T$$

where Q is orthogonal composed of eigenvectors of A; Λ is diagonal with eigenvalues $\Lambda_{i,i} \leftrightarrow Q_{:,i}$ (descending order)

- A is scaling the space by λ_i in the direction of $v^{(i)}$.

! eigendecomposition of $A \in \mathbb{R}^{m \times m}$ always exists but may not be unique. (symmetric)

λ's:

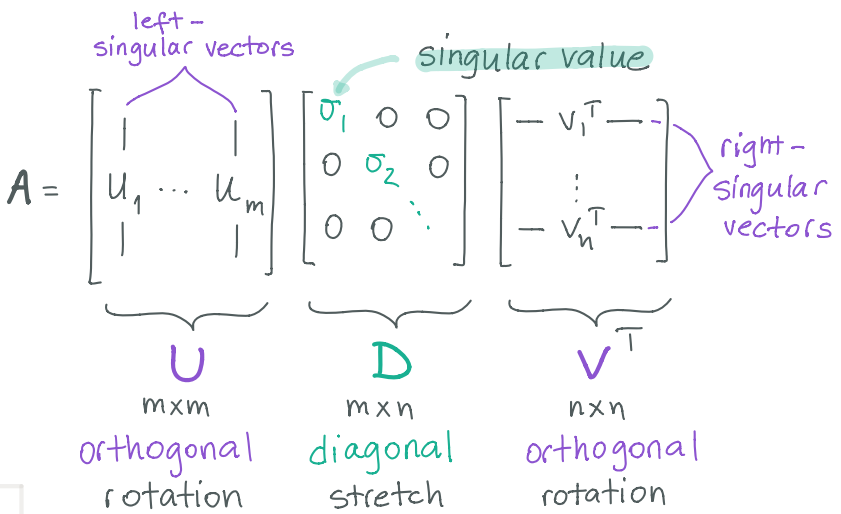
- {iff} any are zero -> SINGULAR
- all positive -> POSITIVE DEFINITE; guarantee -> $x^T A x = 0 \Rightarrow x = 0$
- all positive or zero -> POSITIVE SEMIDEFINITE; guarantee -> $\forall x, \; x^T A x \geq 0$
- all negative -> NEGATIVE DEFINITE
- all negative or zero -> NEGATIVE SEMIDEFINITE

SINGULAR VALUE DECOMPOSITION (SVD)

- factorize matrix into singular value and singular vectors
- every real matrix has SVD.
- used quite a lot in ML applications => e.g. PCA
- getting around the square matrix requirement but with not as many nice properties as eigendecomposition.

! MOST USEFUL : partially generalize matrix inversion for nonsquare matrices

$$A = UDV^T$$



$\sigma_i = \sqrt{\lambda_i}$ where $\sigma_i > 0$

$\{\lambda_i\}$ = eigenvalues of (AA^T)
= eigenvalues of (A^TA)

σ_i
- refers to the strenght of A on the i^{th} subspace
- how much distortion can occur under transformation by A.

biggest σ value ⇓ most information

$\{u_1, \dots, u_m\}$ → left-singular vector → eigenvectors of AA^T

$\{v_1, \dots, v_m\}$ → right-singular vector → eigenvectors of A^TA

THE MOORE-PENROSE PSEUDOINVERSE

! non square matrices don't have inverse defined. ---> $A =$ [ ] possibility of no solution ---- $A =$ [ ] multiple solutions possible

PSEUDOINVERSE ---> gives x for which Ax is as close as possible to y in terms of $\|Ax - y\|_2$. ---- provides one of the many solutions $x = A^+y$ with $\|x\|_2$ among all possible solutions

$$A^+ = VD^+U^T$$

where U, D, V are SVD of A.

D^+ - pseudoinverse of D
- take reciprocal nonzero elements and take transpose

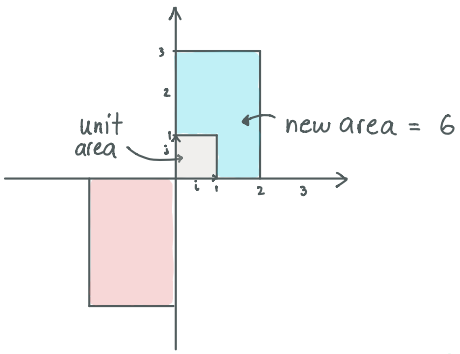
DETERMINANT

$\det : \mathbb{R}^{n \times n} \to \mathbb{R}$

(N) $|A|$; $\det(A)$

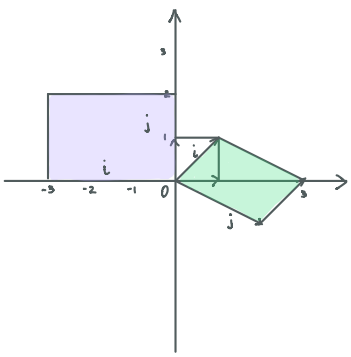
How much area or volume gets scaled?

- exists for every square matrix
- function that maps matrices to real scalars
- equal to the products of all the eigenvalues of a matrix



$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \Rightarrow \det(A) = 6$

$A = \begin{bmatrix} -2 & 0 \\ 0 & -3 \end{bmatrix} \Rightarrow \det(A) = 6$



$A = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \Rightarrow \det(A) = -3$

$A = \begin{bmatrix} -3 & 0 \\ 0 & 2 \end{bmatrix} \Rightarrow \det(A) = -6$

$\det(A) < 0$ when orientation space is inverted

$j \rightarrow$

$i \leftarrow$

$\det(A) = 0 \Rightarrow$ plane $\Rightarrow$ line $\Rightarrow$ point

columns of the matrix are lin. dependent.

PROPERTIES

$\det(A^T) = \det(A)$ $\quad \forall$ nxn matrices

$\det(A) \neq 0 \iff A$ is nonsingular; or

$\det(A) = 0 \iff A$ is singular

$\det(AB) = \det(A)\det(B)$

$\det\begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det(A)\det(B)$ if A, D are square

HOW TO COMPUTE?

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc$$ → how much is the area squished or streched.

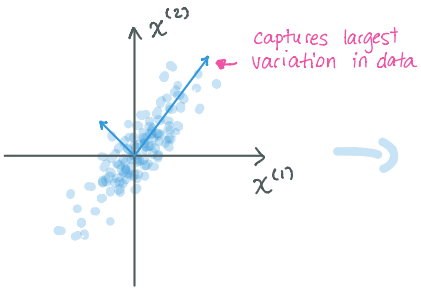
$$\det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) = \det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) = a\det\left(\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right) - b\det\left(\begin{bmatrix} d & f \\ g & i \end{bmatrix}\right) + c\det\left(\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right)$$

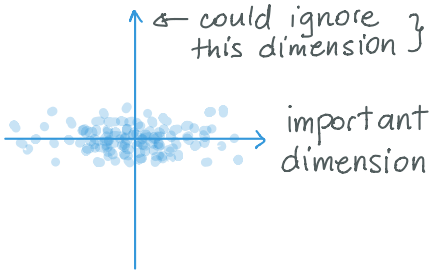
PRINCIPAL COMPONENTS ANALYSIS (PCA)

- unsupervised linear transformation
  - data compression ; dimension reduction
    remove unnecessary information

PRINCIPAL COMPONENTS

=> directions where there is most variance
- data is most spread out

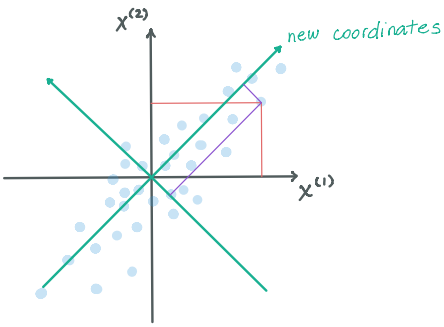




! largest variance doesn't lie along the basis of data measurement

- Rotate the data to the new coordinate system
- removes correlation
- orthonormal transformation

! Assumptions :
- linearity
  -> non-linear features are transformed |centered
- Principal components are orthogonal.



- points are represented by their projected values on to the original coordinates
- new pair of values since the coordinate system changed.

How to find PCA subspace?

$x^{(i)} \in \mathbb{R}^n$; need to find l-dimensional subspace ($l<n$) that would represent the data well => projecting original vectors onto l dimensions.

- need a function that would transform x into c and another function that would reconstruct an approximation of x

encoding function $f(x) = c$
decoding function $x \approx g(f(x))$
$x^{(i)} \in \mathbb{R}^n$; $c^{(i)} \in \mathbb{R}^l$ s.t. $l < n$

- since non-informative dimensions are not used (reduced to zero) x can be only approximated. To capture what is lost we compute the reconstruction error

$$c^* = \arg\min_c \| x - g(c) \|_2$$

! minimize since we don't want to lose too much information

dimensions ↓ information ~

$$x \approx g(c)$$

$$g(c) = Dc \quad ; \quad D \in \mathbb{R}^{n \times l}$$

matrix of the decoding
- orthogonal
- columns have unit norm

- PCA is defined by the choice of the decoding function.

Note maximizing variance of the components is the same as minimizing the reconstruction error

@ Ha Cirovic Donev

STEP 1 How do we find the optimal c^* for every input point x ?

$$c^* = \arg\min_c \| x - g(c) \|_2$$

$$\arg\min_c \| x - g(c) \|_2^2$$ squared L^2 norm and L^2 norm minimize the same c

$$\arg\min_c \left((x - g(c))^T (x - g(c)) \right)$$

$$\arg\min_c \left((x^T - g(c)^T)(x - g(c)) \right)$$ $(x+y)^T = x^T + y^T$

$$\arg\min_c \left(x^T x - x^T g(c) - g(c)^T x + g(c)^T g(c) \right)$$

$$\arg\min_c \left(x^T x - 2 x^T g(c) + g(c)^T g(c) \right)$$ $x^T y = y^T x$

$$\arg\min_c \left(-2 x^T g(c) + g(c)^T g(c) \right)$$ $x^T x$ doesn't depend on c

$$\arg\min_c \left(-2 x^T D c + (Dc)^T Dc \right)$$ $g(c) = Dc$

$$\arg\min_c \left(-2 x^T D c + c^T D^T Dc \right)$$ $D^T c^T = c^T D^T$

$$\arg\min_c \left(-2 x^T D c + c^T I_L c \right)$$ orthogonality ; unit norm constraint on D

$$\arg\min_c \left(-2 x^T D c + c^T c \right)$$

$$\nabla_c \left(-2 x^T D c + c^T c \right) = 0$$ gradient descent

$$-2 x D^T + 2c = 0$$

$$c = D^T x$$

$$f(x) = D^T x$$ ENCODING

$$r(x) = g(f(x)) = g(c) = Dc = DD^T x$$ DECODING

STEP 2 — How to choose matrix D ?

Note:
- use matrix D to decode all points
  - minimize the Frobenius norm of the matrix errors computed over all dimensions and all points

$$D^* = \arg\min_{D} \sqrt{\sum_{i,j} \left(x_j^{(i)} - r(\boldsymbol{x}^{(i)})_j \right)^2} \qquad \text{subject to} \quad D^T D = I$$

Let $l = 1$ => 1st principal component
- D is a single vector $\boldsymbol{d}$

$$\boldsymbol{d}^* = \arg\min_{\boldsymbol{d}} \sum_i \| \boldsymbol{x}^{(i)} - \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{x}^{(i)} \|_2^2 \qquad \text{subject to} \quad \|\boldsymbol{d}\|_2 = 1$$

$$= \arg\min_{\boldsymbol{d}} \sum_i \| \boldsymbol{x}^{(i)T} - \boldsymbol{x}^{(i)T} \boldsymbol{d}\boldsymbol{d}^T \|_2^2 \qquad \text{subject to} \quad \|\boldsymbol{d}\|_2 = 1 \qquad \text{scalar rearrangement}$$

$$= \arg\min_{\boldsymbol{d}} \sum_i \| X - X\boldsymbol{d}\boldsymbol{d}^T \|_F^2 \qquad \text{subject to} \quad \boldsymbol{d}^T\boldsymbol{d} = 1 \qquad X_{i,:} = \boldsymbol{x}^{(i)T}$$

$$= \arg\min_{\boldsymbol{d}} Tr\left(\left(X - X\boldsymbol{d}\boldsymbol{d}^T\right)^T \left(X - X\boldsymbol{d}\boldsymbol{d}^T\right) \right) \qquad \|A\|_F = \sqrt{Tr(AA^T)}$$

$$= \arg\min_{\boldsymbol{d}} Tr\left(X^TX - X^TX\boldsymbol{d}\boldsymbol{d}^T - \boldsymbol{d}\boldsymbol{d}^TX^TX + \boldsymbol{d}\boldsymbol{d}^TX^TX\boldsymbol{d}\boldsymbol{d}^T \right)$$

$$= \arg\min_{\boldsymbol{d}} Tr(X^TX) - Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) - Tr(\boldsymbol{d}\boldsymbol{d}^TX^TX) + Tr(\boldsymbol{d}\boldsymbol{d}^TX^TX\boldsymbol{d}\boldsymbol{d}^T)$$

$$= \arg\min_{\boldsymbol{d}} - Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) - Tr(\boldsymbol{d}\boldsymbol{d}^TX^TX) + Tr(\boldsymbol{d}\boldsymbol{d}^TX^TX\boldsymbol{d}\boldsymbol{d}^T)$$

$$= \arg\min_{\boldsymbol{d}} -2Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) + Tr(\boldsymbol{d}\boldsymbol{d}^TX^TX\boldsymbol{d}\boldsymbol{d}^T)$$

$$= \arg\min_{\boldsymbol{d}} -2Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) + Tr(X^TX\,\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{d}\boldsymbol{d}^T)$$

$$= \arg\min_{\boldsymbol{d}} - Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) \qquad \text{subject to} \quad \boldsymbol{d}^T\boldsymbol{d} = 1 \qquad \boldsymbol{d}^T\boldsymbol{d} = 1$$

$$= \arg\max_{\boldsymbol{d}} Tr(X^TX\boldsymbol{d}\boldsymbol{d}^T) \qquad \text{subject to} \quad \boldsymbol{d}^T\boldsymbol{d} = 1$$

$$= \arg\max_{\boldsymbol{d}} Tr(\boldsymbol{d}X^TX\boldsymbol{d}^T) \qquad \text{subject to} \quad \boldsymbol{d}^T\boldsymbol{d} = 1 \qquad \text{cycling property}$$

Maximum is obtained by computing eigenvectors of X^TX

X^TX → Covariance matrix } if data are centered } ↳ deduct the mean

covariance – redundancy