CS 6140

MATH REVIEW 1

09-13-2021

JORIO COCOLA

cocola.j@northeastern.edu

References

- Garrett Thomas - "Mathematics of Machile Learning"

- Deisenroth et al. - "Mathematics for Machile Learning"

- Kevin Murphy - "Probabilistic Machine Learning"

<u>Remark</u> Below sections refer to
Garrett Thomas' notes

3.1 VECTOR SPACES

Vector spaces are the basic setting in which linear algebra happens. A vector space V is a set (the elements of which are called **vectors**) on which two operations are defined: vectors can be added together, and vectors can be multiplied by real numbers[1] called **scalars**. V must satisfy

ADDITION: $+ : V \times V \longrightarrow V$

$$(x, y) \longmapsto x + y$$

MULTIPLICATION: $\cdot : (\mathbb{R}, V) \longrightarrow V$

$$(\alpha, x) \longmapsto \alpha \cdot x$$

(i) There exists an additive identity (written $\mathbf{0}$) in V such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in V$

(ii) For each $\mathbf{x} \in V$, there exists an additive inverse (written $-\mathbf{x}$) such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$

(iii) There exists a multiplicative identity (written 1) in $\mathbb{R}$ such that $1\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in V$

(iv) Commutativity: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$

(v) Associativity: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ and $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$

(vi) Distributivity: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$

Remark

Products between vectors $x \cdot y$ are <u>not</u> a priori defined.

Example: Euclidean Space $\mathbb{R}^n$

Tuples of n numbers

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{(column vector)}$$

Then

$$x + y = \begin{bmatrix} x_1 + x_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \quad ; \quad \alpha \cdot x = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

e.g. $\mathbb{R}^2$

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad ; \quad 2 \cdot \begin{bmatrix} \pi \\ 0 \end{bmatrix} = \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$$

Q What are some other examples of vector spaces?

- Complex vectors: $\mathbb{C}^2$

$$x = \begin{bmatrix} 0 \\ i \end{bmatrix} \quad y = \begin{bmatrix} i \\ 0 \end{bmatrix} \Rightarrow x + y = \begin{bmatrix} i \\ i \end{bmatrix}$$

- Vector space of Functions

$$f : \mathbb{R} \to \mathbb{R}$$

ADDITION $\quad f : \mathbb{R} \to \mathbb{R} \;,\; g : \mathbb{R} \to \mathbb{R}$

$$(f + g)(x) = f(x) + g(x)$$

Mult $\quad (\alpha f)(x) = \alpha f(x)$

Matrices

A matrix $A \in \mathbb{R}^{m \times n}$ is a tuple of $m \cdot n$ numbers arranged as follows

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

$\mathbb{R}^{m \times n}$ is a vector space with entrywise sum an product by a scalar.

$$[A + B]_{ij} = A_{ij} + B_{ij}$$

$$[\alpha A]_{ij} = \alpha A_{ij}$$

- $A \in \mathbb{R}^{m \times n}$ then A^T is the transpose of A and

$$A_{ij} = [A^T]_{ji} \qquad \text{for each } (i,j)$$

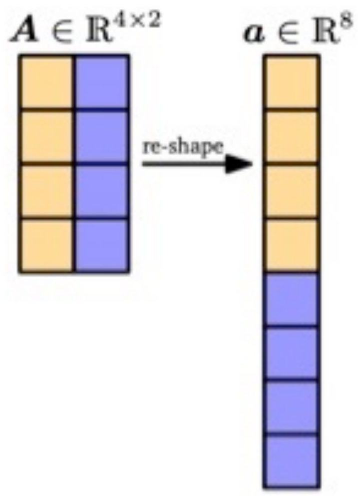
- $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times K}$ then $AB \in \mathbb{R}^{m \times K}$ such that

$$[AB]_{ij} = \sum_{\ell=1}^{n} a_{i\ell}\, b_{\ell j} \qquad \text{for each } (i,j)$$

Matrices as column vectors

By stacking each of the n columns of a matrix $A \in \mathbb{R}^{m \times n}$ we obtain a column vector

$$a = \mathrm{vec}(A) \in \mathbb{R}^{m \cdot n}$$



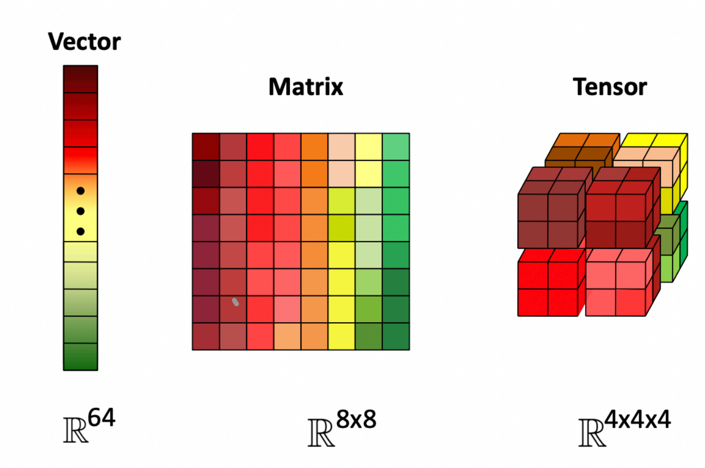
Note the size of the matrices.

```
C =
np.einsum('il,
lj', A, B)
```

from the mml-book

Tensors

Generalizes matrices to more than 2 dimensions



from Murphy - Probabilistic Machine Learning

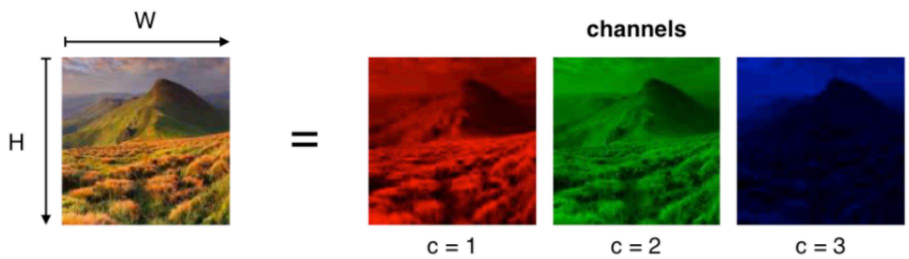
(we can flatten tensors)

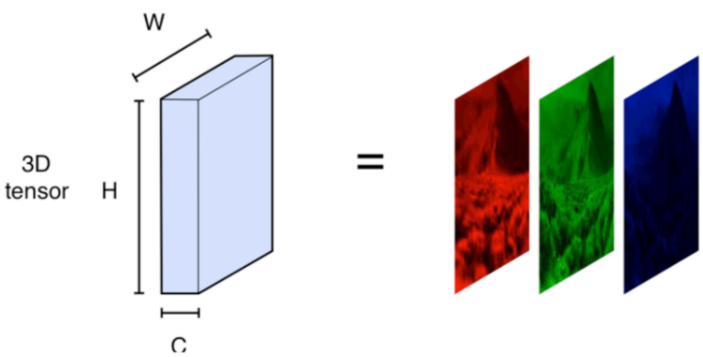
Examples pictures, videos, etc.

Examples

In ML we often work with high dimensional vectors

$\in \mathbb{R}^{150 \times 200}$





$\in \mathbb{R}^{256 \times 256 \times 3}$

Linear Independence and Bases

Def Linear Combination

V vector space and $x_1, \dots, x_K \in V$

Then every $y \in V$ of the form

$$y = \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_K x_K$$

with $\lambda_1, \dots, \lambda_K \in \mathbb{R}$,

is a LINEAR COMBINATION of $x_1, \dots, x_K$.

Def Span

$$\operatorname{span}\{x_1, \dots, x_K\}$$

$$= \{\text{set of linear comb. of } x_1, \dots, x_K\}$$

$$= \{y \in V : y = \alpha_1 x_1 + \dots + \alpha_K x_K \text{ for some } \alpha_1, \dots, \alpha_K \in \mathbb{R}\}$$

Example

$A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ then

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} x_2 + \dots + \begin{bmatrix} | \\ a_n \\ | \end{bmatrix} x_n$$

- y is a linear combination of the columns of A
- $y \in \text{span}\{a_1, a_2, \dots, a_n\}$

Def Linear Independence

- $x_1, \dots, x_K \in V$ are <u>Linearly dependent</u>

  if $\exists\ \lambda_1, \dots, \lambda_K \in \mathbb{R}$ such that

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_K x_K = 0$$

  with at least one $\lambda_i \neq 0$

- $x_1, \dots, x_K \in V$ are <u>Linearly independent</u>

  if $\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_K x_K = 0$

  implies $\lambda_1 = \lambda_2 = \dots = \lambda_K = 0$

ex

- $V = \mathbb{R}^2$

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad ; \quad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad ; \quad z = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- Are x, y, z linearly (in)dependent?
- Is $z \in \operatorname{span}\{x, y\}$
- Are x, y linearly (in)dependent?

Answers

- $x + y = z$ or $x + y - z = 0$

$$2x + 2y - 2z = 0$$

x, y, z are linearly dependent

- $z \in \operatorname{span}\{x, y\}$ bc. $z = x + y$

- If they were dependent

$\lambda_1 x + \lambda_2 y = 0$ *

$\lambda_1 \neq 0$ or $\lambda_2 \neq 0$

$$\lambda_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\Rightarrow \lambda_1 = 0$ and $\lambda_2 = 0$

(contradiction)

$\text{span}\{x, y\} = \mathbb{R}^2$

Def. Bases

If a set of vectors is linearly independent and its span is the whole of V, those vectors are said to be a **basis** for V. In fact, every linearly independent set of vectors forms a basis for its span.

If a vector space is spanned by a finite number of vectors, it is said to be **finite-dimensional**. Otherwise it is **infinite-dimensional**. The number of vectors in a basis for a finite-dimensional vector space V is called the **dimension** of V and denoted $\dim V$.

ex The standard basis in $\mathbb{R}^n$ is

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} ; \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} ; \quad \cdots ; \quad e_n = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Q Find a basis for

- $\mathbb{R}^{2\times 2}$
- $P_2(\mathbb{R}) = \{$ polynomials in $x \in \mathbb{R}$ of degree at most $2\}$

Basis for $\mathbb{R}^{2\times 2}$

$$E_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad E_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$E_3 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad E_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} = 2 \cdot E_1 + 1 \cdot E_4$$

$$\mathrm{Vec}(E_1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathrm{Vec}(E_4) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$\mathbb{R}^{m \times n}$ = $m \times n$ matrices

$\Rightarrow \dim(\mathbb{R}^{m \times n}) = m \times n$

3.1.2 Subspaces

Vector spaces can contain other vector spaces. If V is a vector space, then $S \subseteq V$ is said to be a **subspace** of V if

(i) $\mathbf{0} \in S$

(ii) S is closed under addition: $\mathbf{x}, \mathbf{y} \in S$ implies $\mathbf{x} + \mathbf{y} \in S$

(iii) S is closed under scalar multiplication: $\mathbf{x} \in S, \alpha \in \mathbb{R}$ implies $\alpha \mathbf{x} \in S$

Q Is S a subspace :

- $S_1 = \left\{ y \in \mathbb{R}^n : y = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ for some } \alpha \in \mathbb{R} \right\} \subseteq \mathbb{R}^n$

- $S_2 = \left\{ y \in \mathbb{R}^n : y_1 \geq 0 \right\} \subseteq \mathbb{R}^n$

- $S_3 = \left\{ 128 \times 128 \text{ grayscale image of a cat} \right\} \subseteq \mathbb{R}^{128 \times 128}$

3.2 Linear Maps

A **linear map** is a function $T : V \to W$, where V and W are vector spaces, that satisfies

(i) $T(\mathbf{x}+\mathbf{y}) = T\mathbf{x} + T\mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in V$

(ii) $T(\alpha\mathbf{x}) = \alpha T\mathbf{x}$ for all $\mathbf{x} \in V, \alpha \in \mathbb{R}$

Examples

Image filters, convolutional layers, etc.

GAUSSIAN BLUR

from Wiki

Q Show that for every linear map $T : V \to W$

$$T(0) = 0$$

ex $A \in \mathbb{R}^{m \times n}$ then

the map $x \mapsto Ax$ is linear

$A : \mathbb{R}^n \to \mathbb{R}^m$

$x \in \mathbb{R}^n$ $\qquad Ax \in \mathbb{R}^m$

ex Consider $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_2 \\ 0 \end{bmatrix}$$

show that it is a linear map.

- $A \in \mathbb{R}^{m \times n}$ you can define

$T : \mathbb{R}^n \to \mathbb{R}^m$ linear map

such that

$T(x) = Ax \in \mathbb{R}^m$ for $x \in \mathbb{R}^n$

3.2.1 The matrix of a linear map

Every linear map is <u>completely</u> <u>determined</u> by specifying its action on the basis vectors:

- $v_1, \dots, v_n$ is a basis for V
- $T: V \to W$ is a linear transformation
- $x \in V$ then for some $\lambda_1, \dots, \lambda_n \in \mathbb{R}$

$$x = \lambda_1 v_1 + \dots + \lambda_n v_n$$

therefore

$$T(x) = \lambda_1 T(v_1) + \dots + \lambda_n T(v_n)$$

3.2.2 Nullspace, range

$T : V \longrightarrow W$ linear transformation

- Nullspace or Kernel is the subset of V that is mapped to zero

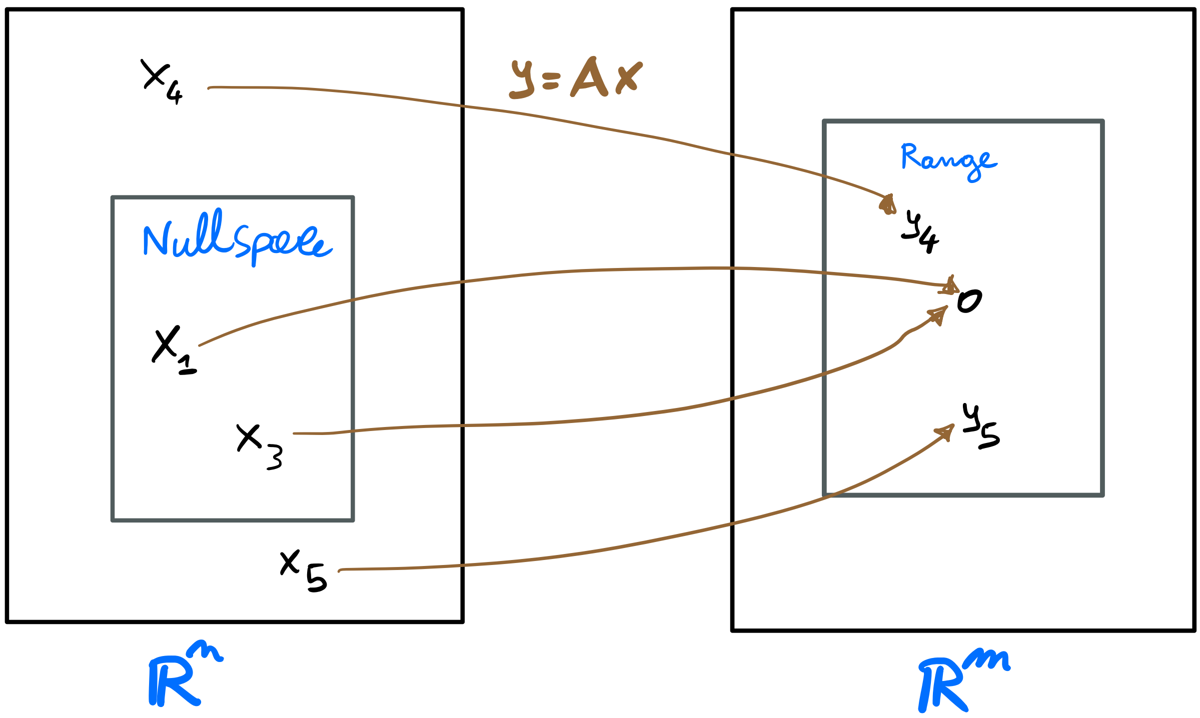
$$Ker(T) = null(A) = \{v \in V : Tv = 0\}$$

- Range is the subset of W that is reachable by T.

$$range(T) = \{w \in W : Tv = w \text{ for some } v\}$$

if $w \in range(T) \Rightarrow \exists v : Tv = w$

Consider a linear transformation given by $A \in \mathbb{R}^{m \times n}$



columnspace = span{columns of A} = range(A)

rowspace = span{rows of A}

$y = Ax \Rightarrow y = x_1 A_1 + x_2 A_2 + \dots + x_n A_n$

where A_i cols of A

Rank of a matrix

$$\text{rank}(A) = \dim \text{range}(A)$$
$$= \dim \text{rowspace}(A)$$
$$= \# \text{ independent columns}$$
$$= \# \text{ independent rows}$$

If $A \in \mathbb{R}^{m \times n}$ then

$$\text{rank}(A) \leq \min(m, n)$$

A is **full rank** if

$$\text{rank}(A) = \min(m, n)$$

Ex Consider $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ given by

$$T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_2 \\ 0 \end{bmatrix}$$

- Find $A \in \mathbb{R}^{2\times 2}$ that represent T with respect to the standard basis
- Find range(T) and null(T)
- Is T full rank?

Solutions of Linear Systems

A system of linear equations can be written as

$$\begin{cases} a_{11} x_1 + \dots + a_{1n} x_n = b_1 \\ \vdots \\ a_{m1} x_1 + \dots + a_{mn} x_n = b_m \end{cases}$$

That is $Ax = b$ for $A \in \mathbb{R}^{m \times n}$

- A solution exists if and only if

$$b \in \text{range}(A)$$

- If x_p is a particular solution of $Ax = b$ all solutions can be written as

$$x = x_p + x_N$$

for some $x_N \in \text{null}(A)$

3.4 Normed Spaces

A norm on a vector space V

is a nonlinear function $\|\cdot\| : V \to \mathbb{R}$:

(i) $\|\mathbf{x}\| \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{0}$

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$

(iii) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (the **triangle inequality** again)

for all $\mathbf{x}, \mathbf{y} \in V$ and all $\alpha \in \mathbb{R}$. A vector space endowed with a norm is called a **normed vector space**, or simply a **normed space**.

Norms on $\mathbb{R}^n$

We will typically only be concerned with a few specific norms on $\mathbb{R}^n$:

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

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

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} \qquad (p \geq 1)$$

$$\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$$

Norms on $\mathbb{R}^{n \times m}$

Norms can be defined also on $\mathbb{R}^{m \times n}$

For example the **FROBENIUS NORM:**

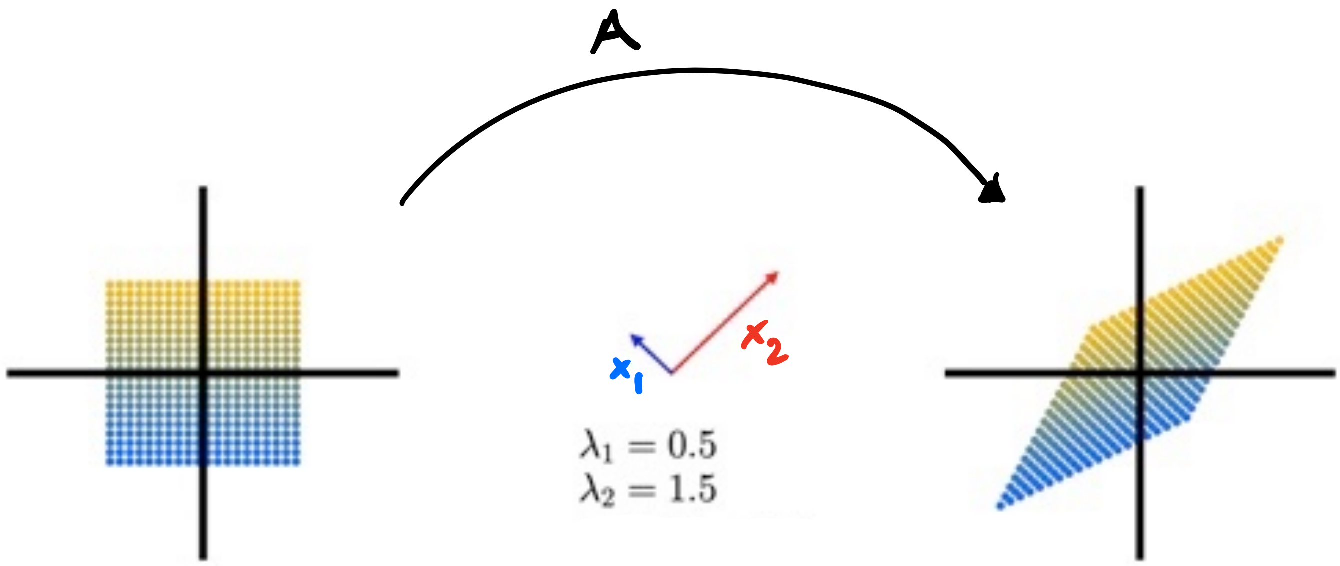
$$\|A\|_F := \left(\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2 \right)^{\frac{1}{2}} = \| \mathrm{vec}(A) \|_2$$

or the **p-NORMS**

$$\|A\|_p := \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \qquad p \geq 1$$

3.6 Eigenthings

For a square matrix $A \in \mathbb{R}^{n \times n}$
the eigenvectors of A are
those vectors that A simply scales



$A \in \mathbb{R}^{2 \times 2}$ and x_1, x_2 eigenvectors

A nonzero $x \in \mathbb{R}^n$ is an **eigenvector** of A with **eigenvalue** λ if

$$Ax = \lambda x$$

$$A(\alpha x) = \alpha A x = \alpha \lambda x = \lambda (\alpha x)$$

Remark

If x is an eigenvector than also αx for any $\alpha \in \mathbb{R}$ is an eigenv. (why?)

When we talk about "the" eigenvector associated with λ we usually mean the eigenvector with length 1:

$$\|x\|_2 = 1$$

3.10 Symmetric Matrices

A matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if

$$A^T = A$$

3.11 Positive (semi-)definite matrices

Consider a symmetric matrix $A \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \dots, \lambda_n$.

Then A is

- positive semi-definite if

$$\lambda_i \geq 0 \qquad \text{for any } i = 1, \dots, n$$

or equivalently

$$x^T A x \geq 0 \qquad \text{for any } x \in \mathbb{R}^n$$

- positive definite if

$$\lambda_i > 0 \qquad \text{for any } i = 1, \dots, n$$

or equivalently

$$x^T A x > 0 \qquad \text{for any } x \in \mathbb{R}^n$$

4. Calculus and Optimization

Derivatives

- Consider $f: \mathbb{R} \to \mathbb{R}$. The derivative of f at x is

$$\frac{df}{dx} = f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

when the limit exists.

- When $f: \mathbb{R}^n \to \mathbb{R}$, the partial derivatives are

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x + h e_i) - f(x)}{h}$$

where $e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ is the i-th standard basis vector.

4.2 Gradient

When $f: \mathbb{R}^n \longrightarrow \mathbb{R}$, the gradient of f at x is:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Note Sometimes this is written as a row vector.

4.3 Jacobian

When $f : \mathbb{R}^n \to \mathbb{R}^m$ we define the Jacobian of f at x as the $m \times n$ matrix

$$\mathbf{J}_{\boldsymbol{f}}(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}^\top} \triangleq \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla f_1(\boldsymbol{x})^\top \\ \vdots \\ \nabla f_m(\boldsymbol{x})^\top \end{pmatrix}$$

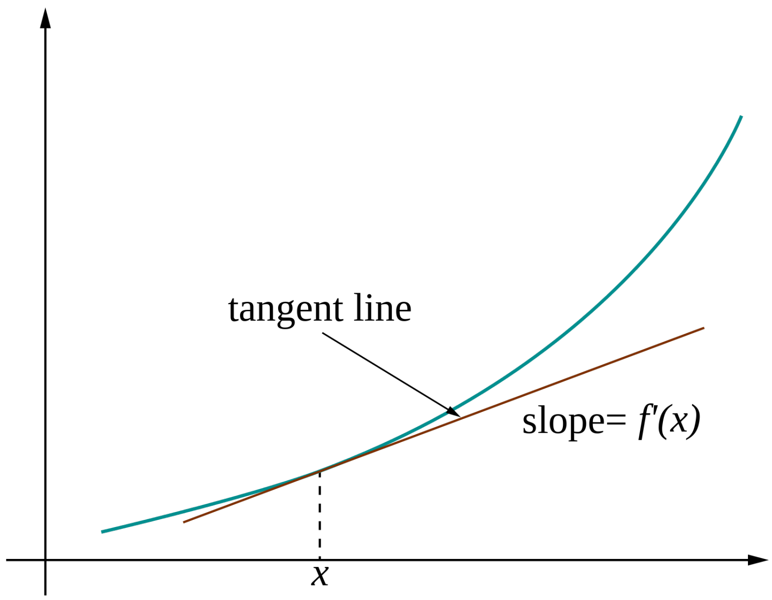
where

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

Tangent and Approximation

If $f: \mathbb{R} \to \mathbb{R}$, $f'(x)$ gives the slope of the tangent line at x

$$f(x+h) \approx f(x) + f'(x)\,h$$



Source : wiki

This generalizes to $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ using hyperplanes.

What about when $f : \mathbb{R}^n \to \mathbb{R}^m$?

$J_f(x)$ is the linear map that approximates f locally around x

$$f(x+h) \approx f(x) + J_f(x)\, h$$

Useful to remember the dimensions

- $f : \mathbb{R}^n \to \mathbb{R}^m$
- $J_f(x) \in \mathbb{R}^{m \times n}$

4.4 Hessian

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

For f with continuous partial derivatives

$$\nabla^2 f = (\nabla^2 f)^T$$

4.6 Taylor's Theorem

We can use the Hessian for computing a quadratic approximation of f at x

$$f(x+h) \approx f(x) + \nabla f(x)^{\top} h + \frac{1}{2} h^{\top} \nabla^2 f(x) h$$

This can also be made exact!

Theorem 6. *(Taylor's theorem) Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and let $\mathbf{h} \in \mathbb{R}^d$. Then there exists $t \in (0,1)$ such that*

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x}+t\mathbf{h})^{\top}\mathbf{h}$$

Furthermore, if f is twice continuously differentiable, then

$$\nabla f(\mathbf{x}+\mathbf{h}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}\,\mathrm{d}t$$

and there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^{\top}\mathbf{h} + \frac{1}{2}\mathbf{h}^{\top}\nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}$$

4.1 Extrema

A large part of Machine Learning is about minimizing/ maximizing (loss) functions:

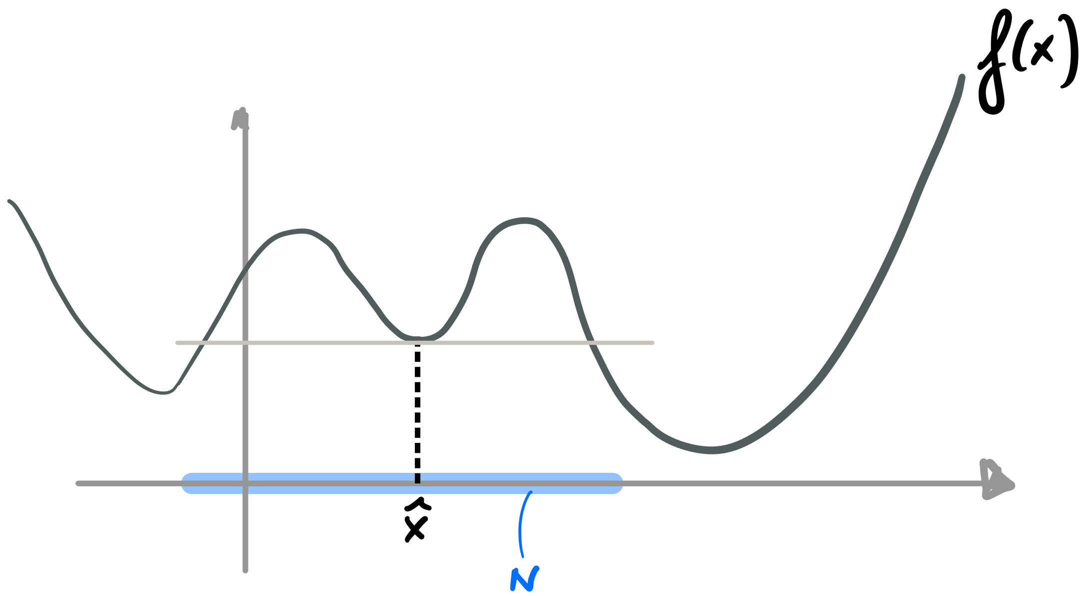
For example for

$$\min_{x \in \mathbb{R}^n} f(x)$$

We define

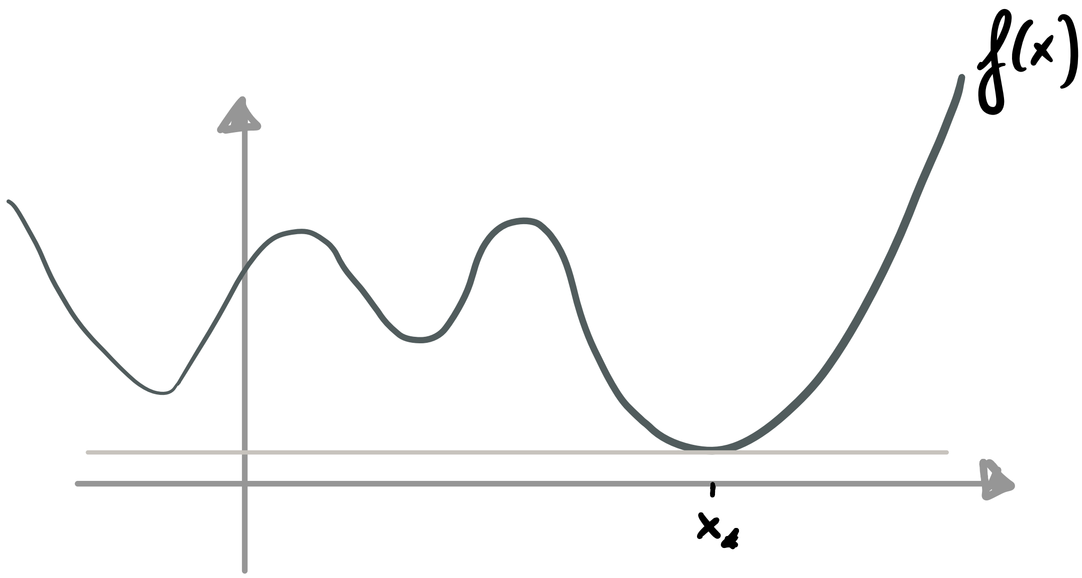
- $\hat{x}$ a local minimum of f if $\exists N \subseteq \mathbb{R}^n$:

$$f(\hat{x}) \leq f(x) \qquad \forall x \in N$$



- x_* a global minimum of f if

$$f(x_*) \leq f(x) \qquad \forall x \in \mathbb{R}^n$$



Similarly we can define local/global minima.

Remark

Maximizing f is equivalent to minimizing $-f$:

if x is a (local) maximum of f

then x is a (local) minimum of $-f$

and vice versa.