Machine Learning

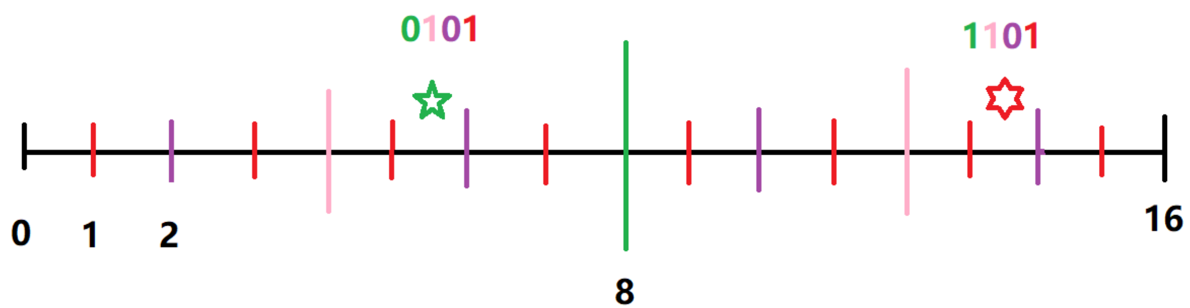
Part 1: Mathematical Foundation of Machine Learning

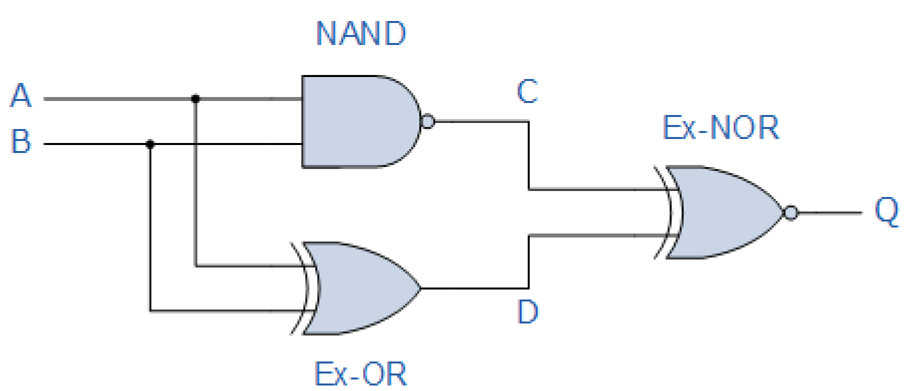
Zengchang Qin (PhD)

Function and Data Generalization

Numbers and Logic

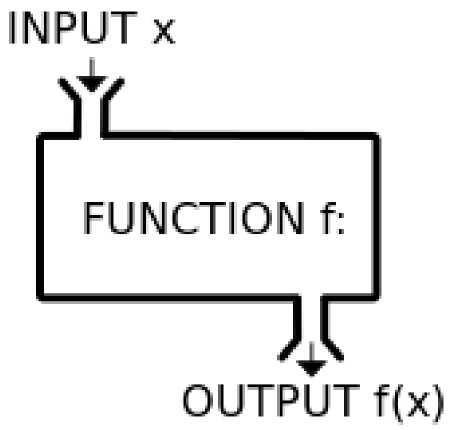
There are 10 kinds of people in the world, the ones who understand binary and those who do not.

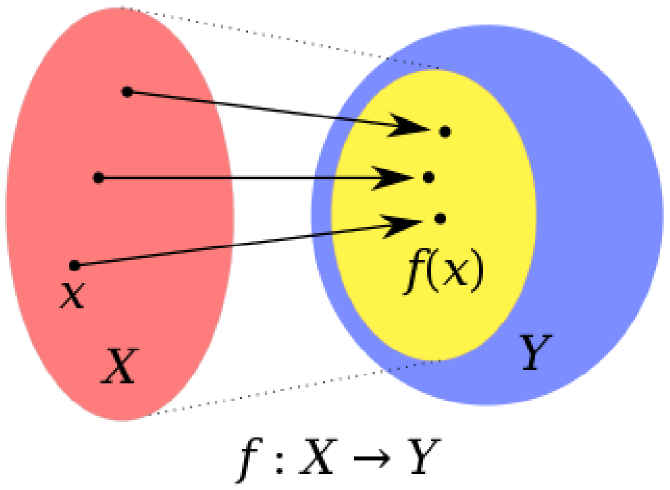




Functions

In mathematics, a function is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output.

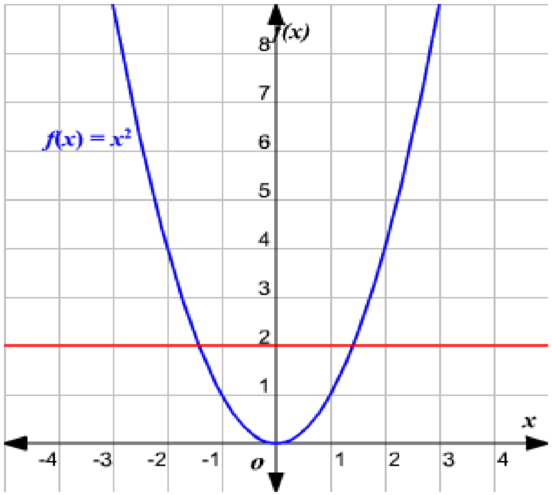


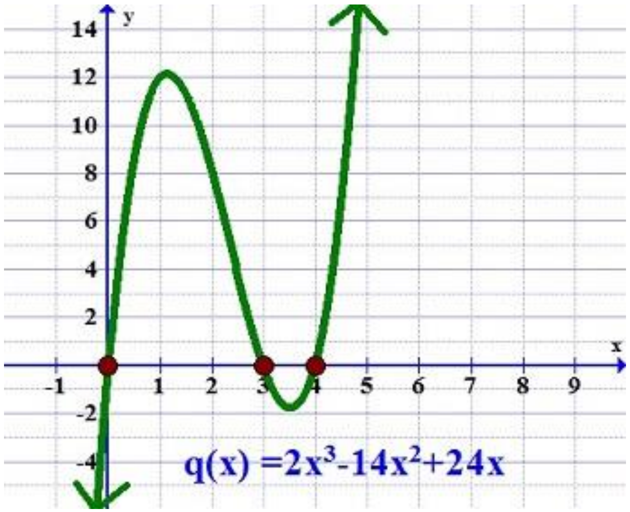


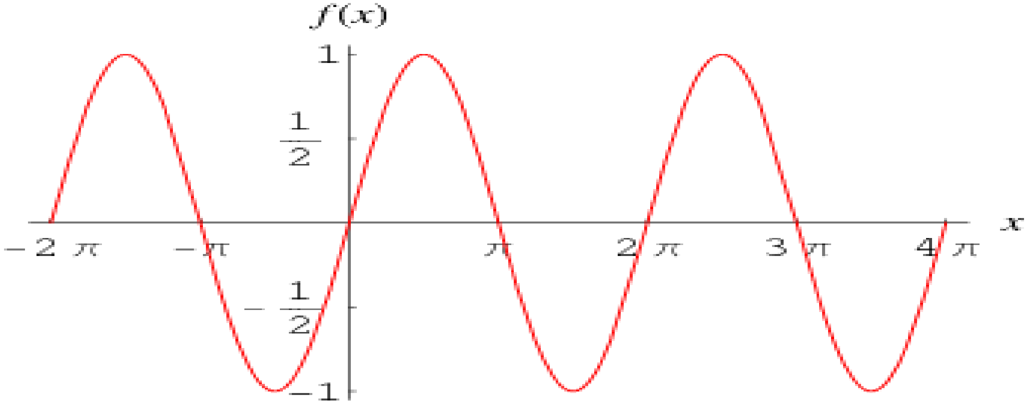
A sample function: f(x) = 2x+3

Functions

We have learned different types of functions.

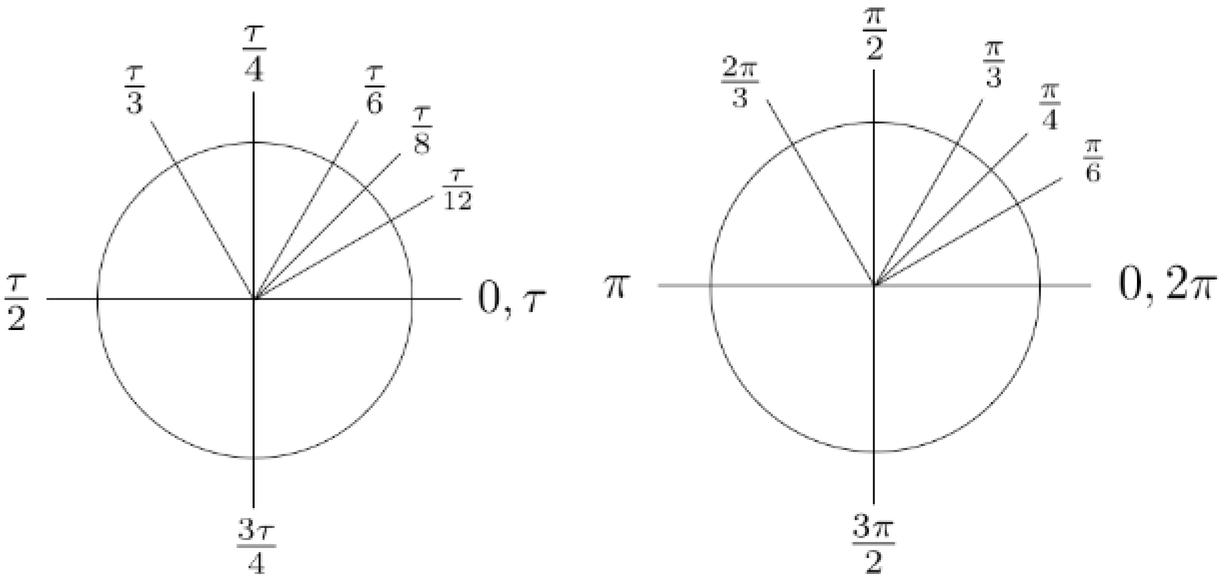






Tau or Pie?

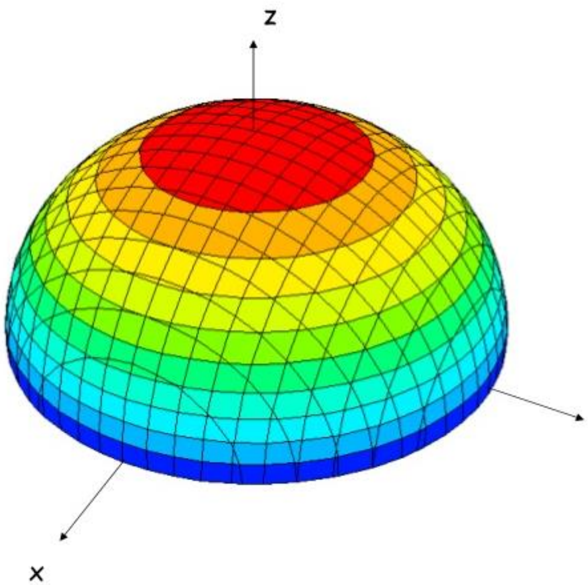
Constant are somehow arbitrary, even the most important one!

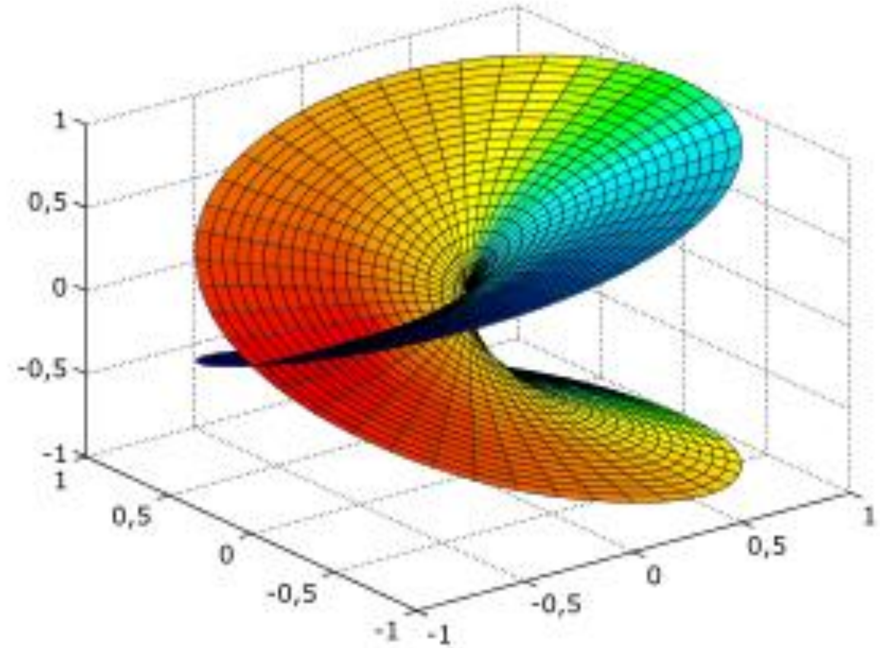


https://www.theverge.com/tldr/2018/3/14/17119388/pi-day-pie-math-tau-circle-constant-mathematics-circumference-diameter-radius-holiday-truth

Functions

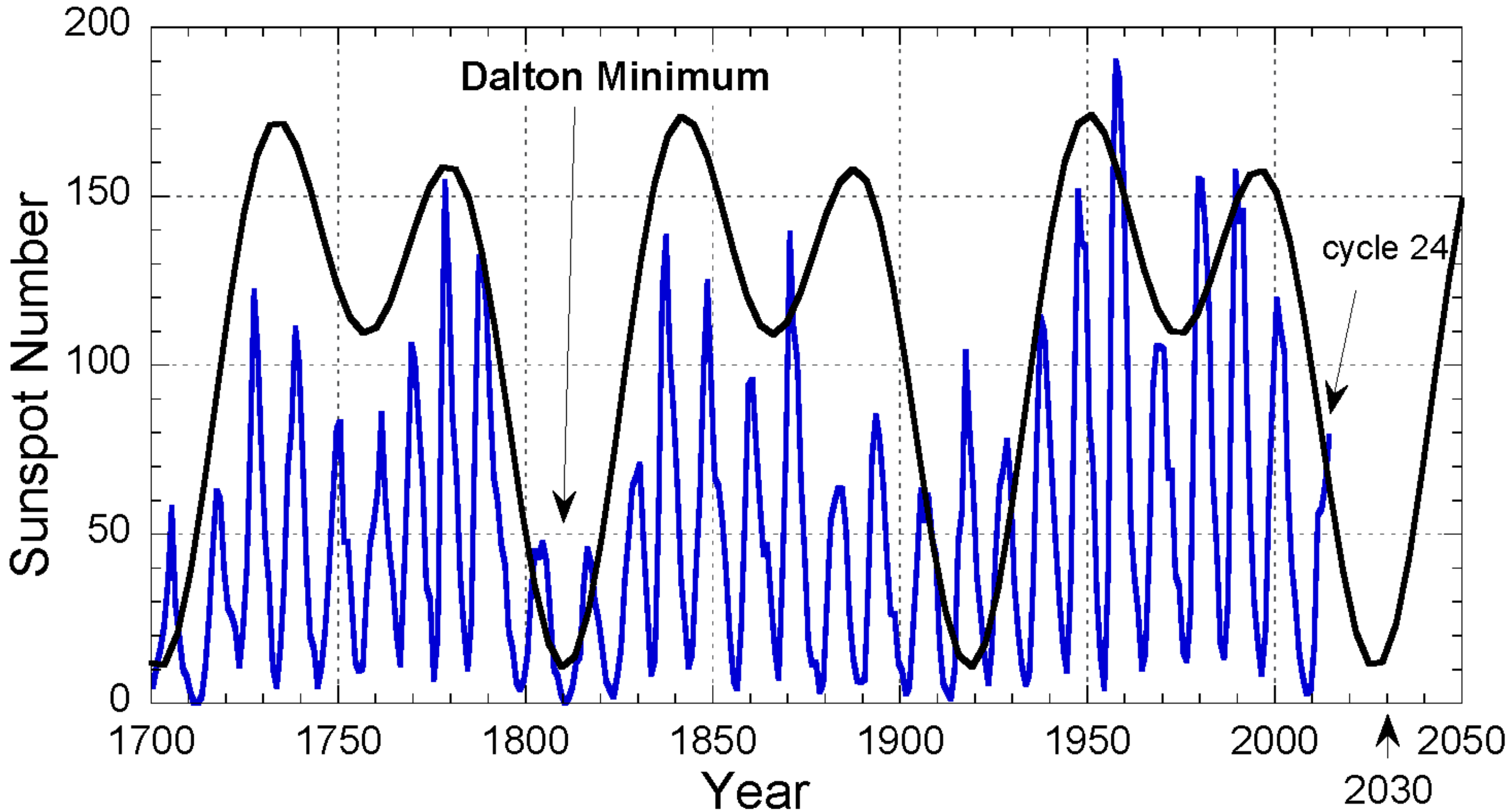
We have learned different types of functions.



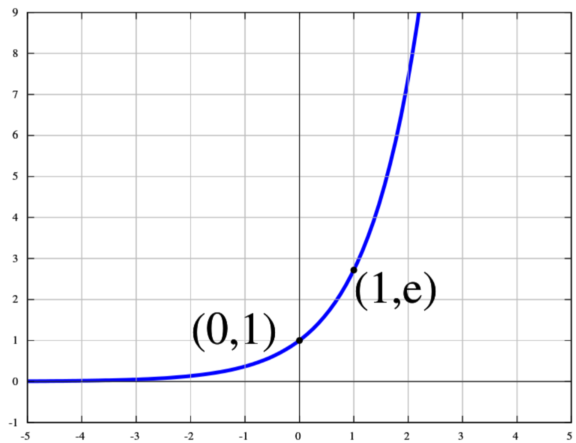


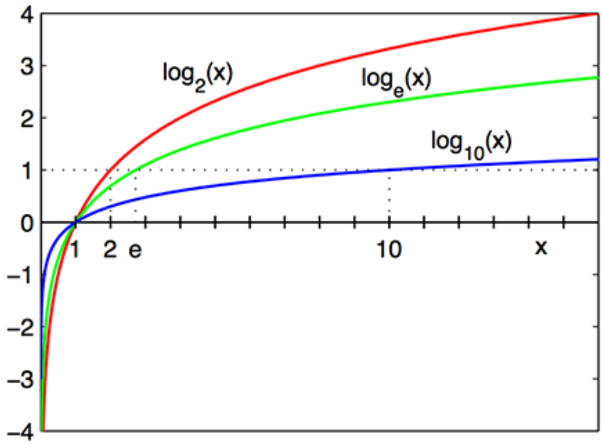
The Real-World Data

In the real-world, when we are investigating relations, we may find the following:

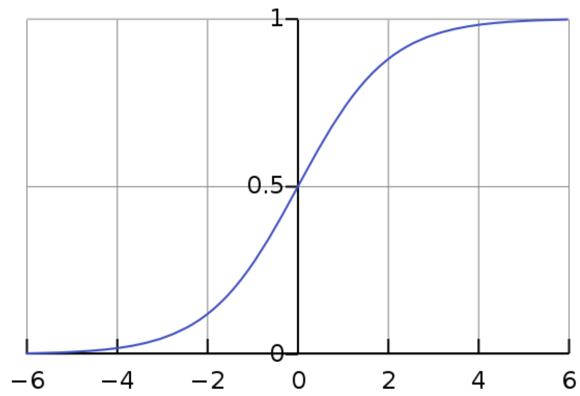


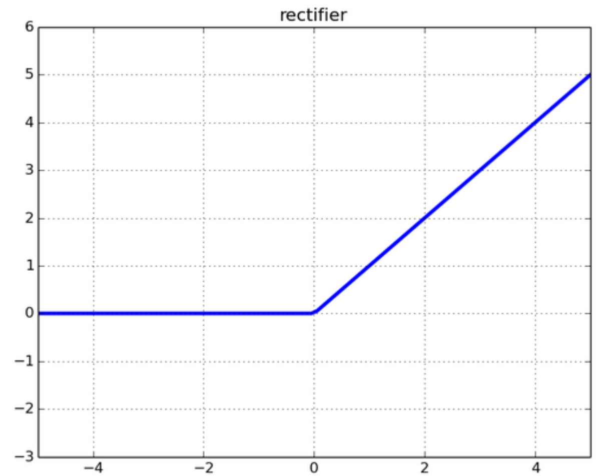
Some Functions





$y = e^x$ http://setosa.io/ev/exponentiation/





$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

$$f(x) = x^+ = \max(0, x)$$

Function Decomposition

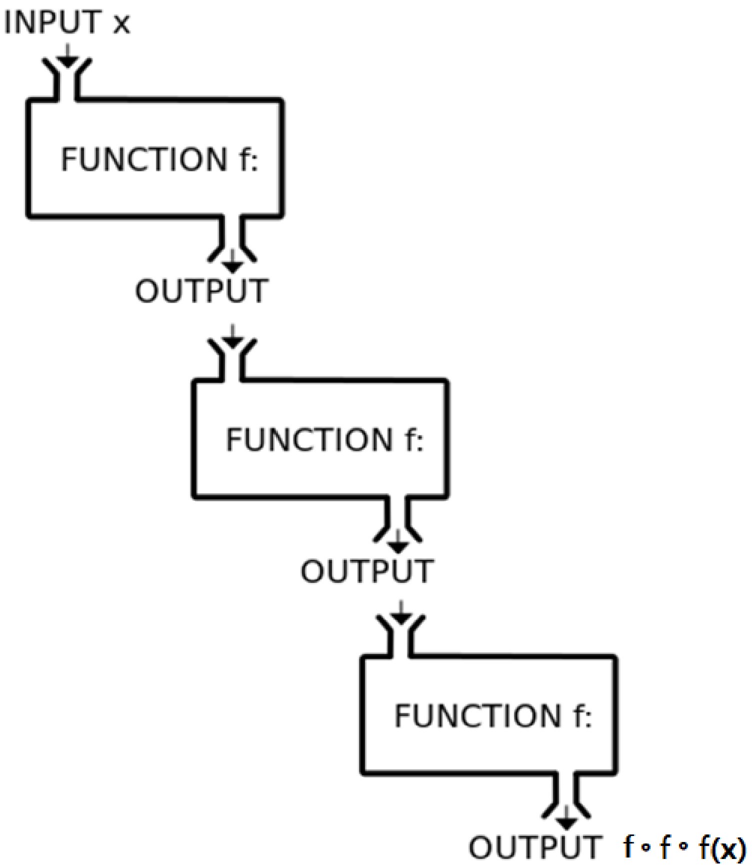
"Function Composition" is applying one function to the results of another:
The result of f() is sent through g()

It is written: $(g \circ f)(x)$
Which means: $g(f(x))$

$f(x) = 2x+3$

$f \circ f(x) = ?$
$f \circ f \circ f\ (x) = ?$

$f \circ f \circ f\ (2) =$

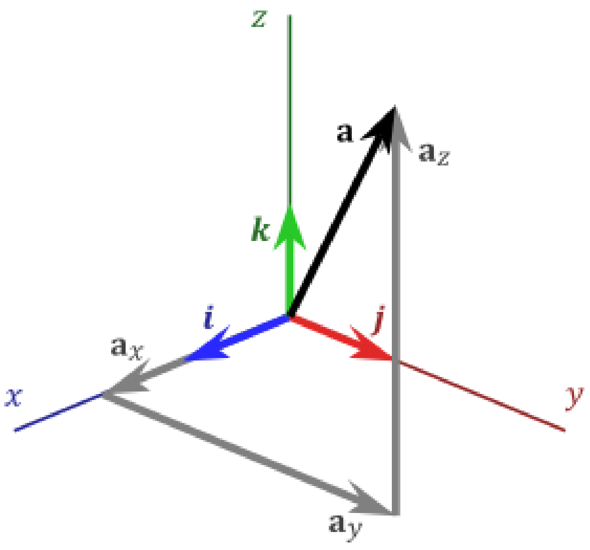


Linear Algebra

Vector

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 called scalars.

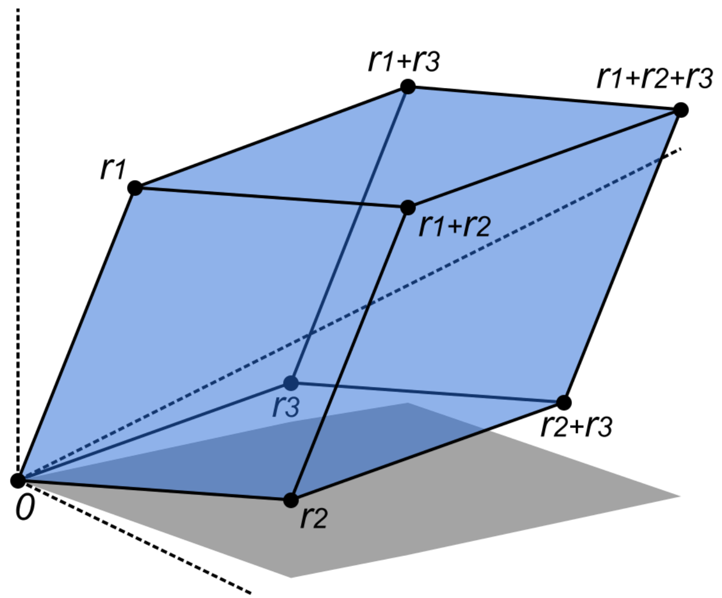
Can be written in column form or row form – Column form is conventional!



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \mathbf{x}+\mathbf{y} = \begin{bmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{bmatrix}, \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

Vector Space

- Euclidean space is used to mathematically represent physical space, with notions such as distance, length, and angles.
- Although it becomes hard to visualize for n > 3, these concepts generalize mathematically in obvious ways.
-
- Linear relations hold in high dimensional space.



Norm of Vectors

A **norm** on a real vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies

(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)

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

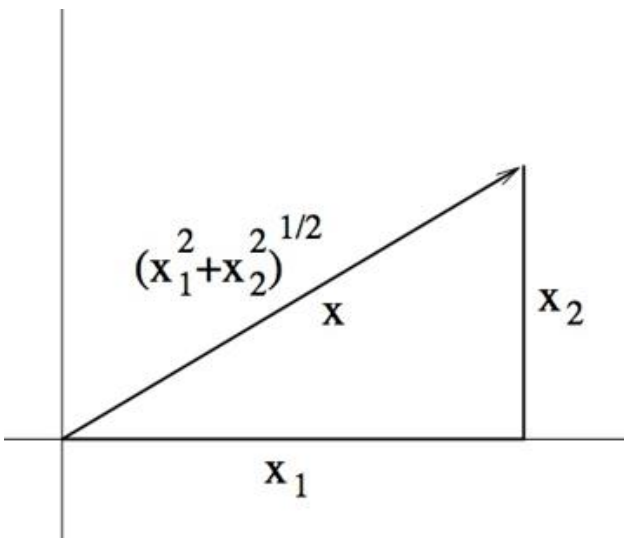
$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad (p \geq 1)$$

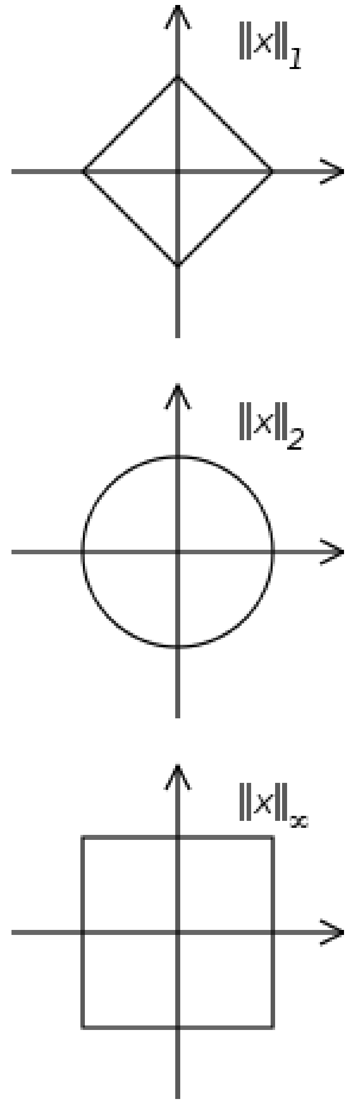
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$$

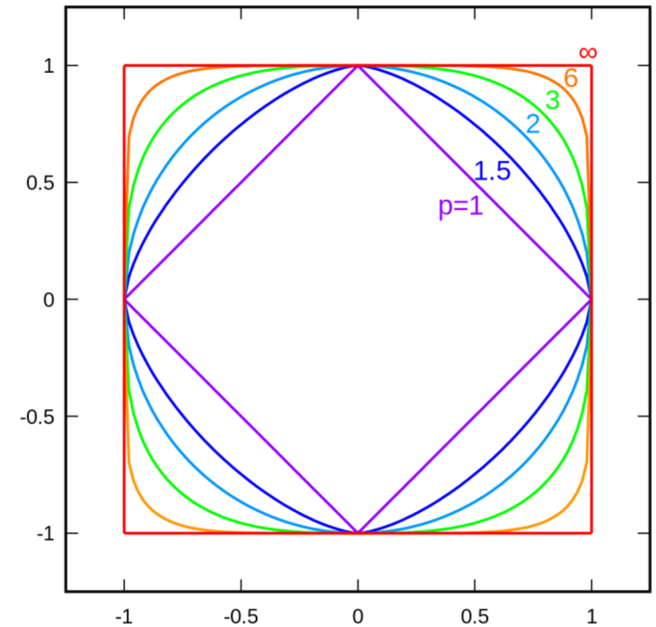
L-0 to L-infinity Norms

a **norm** is a function that assigns a strictly *positive length* to a vector。

A simple example is two dimensional Euclidean space R2 equipped with the "Euclidean norm"







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

Matrix

A vector can be regarded as special case of a matrix, where one of matrix dimensions = 1.

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

Matrix transpose (denoted T)

$$\mathbf{A} = \begin{pmatrix} 2 & 7 & -1 & 0 & 3 \\ 4 & 6 & -3 & 1 & 8 \end{pmatrix} \qquad \mathbf{A}^{\mathrm{T}} = \begin{pmatrix} 2 & 4 \\ 7 & 6 \\ -1 & -3 \\ 0 & 1 \\ 3 & 8 \end{pmatrix}$$

$$C = AB \qquad \Leftrightarrow \qquad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

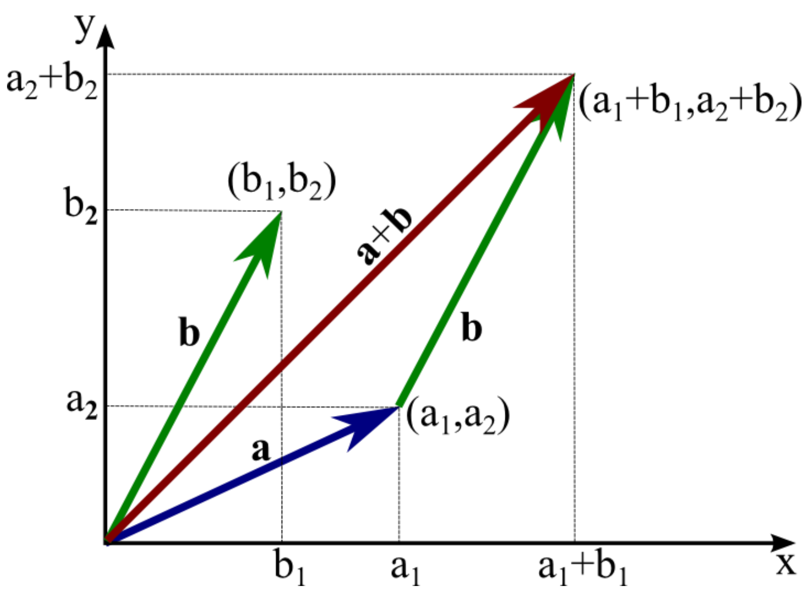
$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} \end{bmatrix}$$

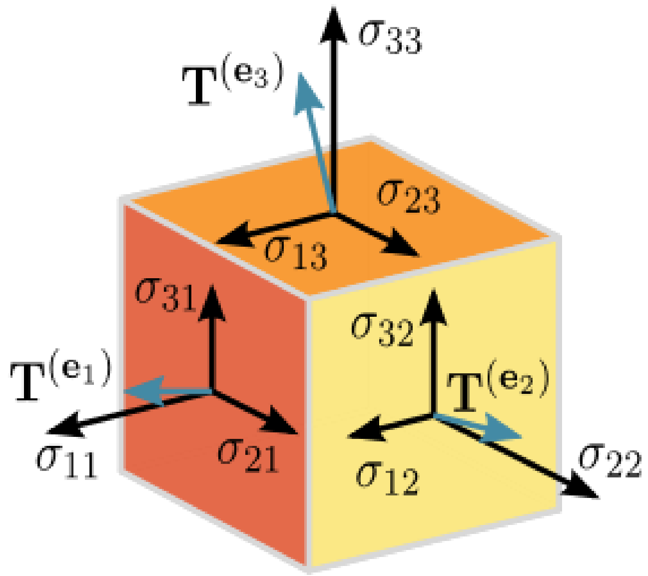
Vector to Tensor

https://www.quora.com/What-is-a-tensor

Columns are the stresses (forces per unit area) acting on the $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ faces of the cube.

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \equiv \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}$$



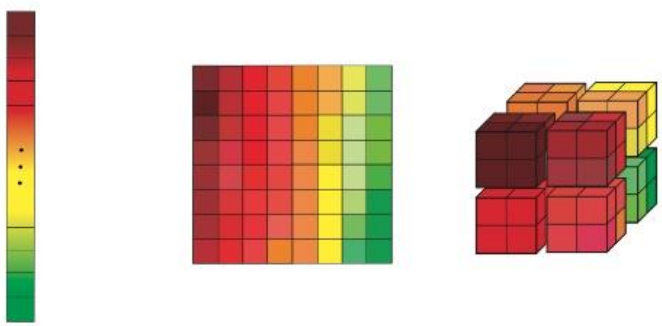


https://www.wukong.com/question/6531498435785261325/
https://www.bilibili.com/video/av10852829/

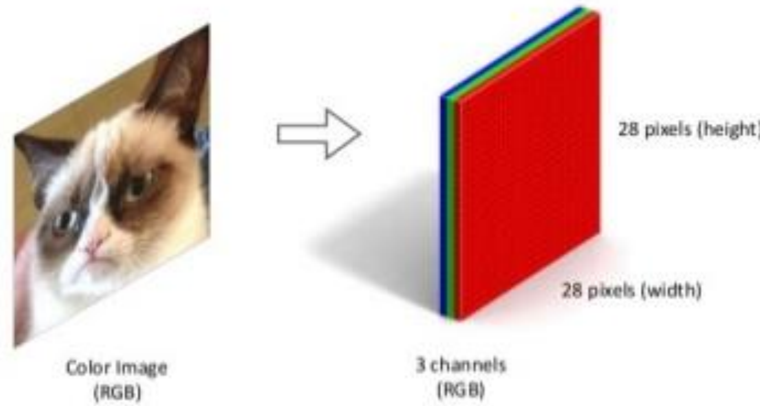
Tensor

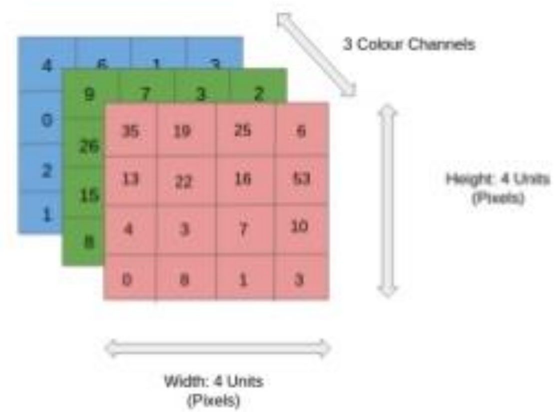
tensor = multidimensional array

vector matrix tensor



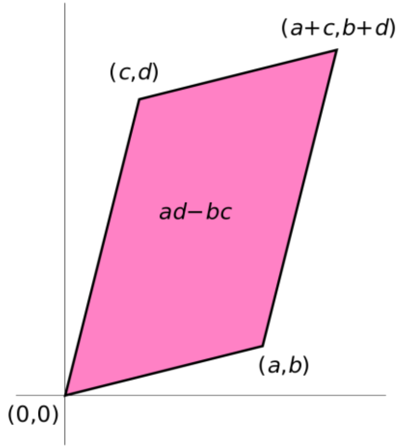
$\mathbf{v} \in \mathbb{R}^{64}$ $\quad X \in \mathbb{R}^{8x8}$ $\quad \boldsymbol{\mathcal{X}} \in \mathbb{R}^{4x4x4}$



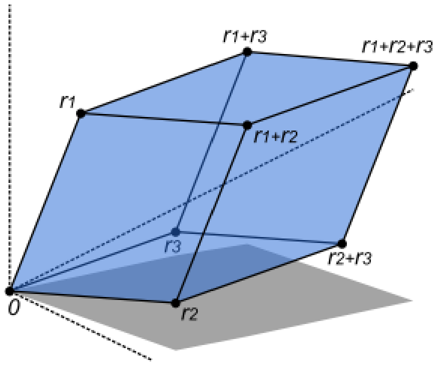


Determinant

In linear algebra, the determinant is a useful value that can be computed from the elements of a square matrix. The determinant of a matrix A is denoted det(A), det A, or |A|. It can be viewed as the scaling factor of the transformation described by the matrix.



$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$



$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
$$= aei + bfg + cdh - ceg - bdi - afh.$$

Eigenvector and Eigenvalue

For a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, there may be vectors which, when $\mathbf{A}$ is applied to them, are simply scaled by some constant. We say that a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ is an **eigenvector** of $\mathbf{A}$ corresponding to **eigenvalue** λ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

The zero vector is excluded from this definition because $\mathbf{A0} = \mathbf{0} = \lambda\mathbf{0}$ for every λ.

We now give some useful results about how eigenvalues change after various manipulations.

The **trace** of a square matrix is the sum of its diagonal entries:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

http://setosa.io/ev/eigenvectors-and-eigenvalues/

Eigen Decomposition

Assume square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with n linearly independent eigenvectors $\mathbf{q}_i, i = 1, \ldots, n$ and n eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $\mathbf{A}$ can be factorised as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

where $\mathbf{Q} = [\mathbf{q}_1 \ldots \mathbf{q}_n]$ and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e. $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$.
Using the eigen-decomposition we can compute various powers of $\mathbf{A}$ as

$$\mathbf{A}^k = \mathbf{Q}\mathbf{\Lambda}^k\mathbf{Q}^{-1}.$$

We can easily verify the above for $k = 2$ as $\mathbf{A}^2 = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^{-1}$. Then, we can easily prove the general case using induction.
In case $k = -1$ we can compute the inverse as

$$\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}.$$

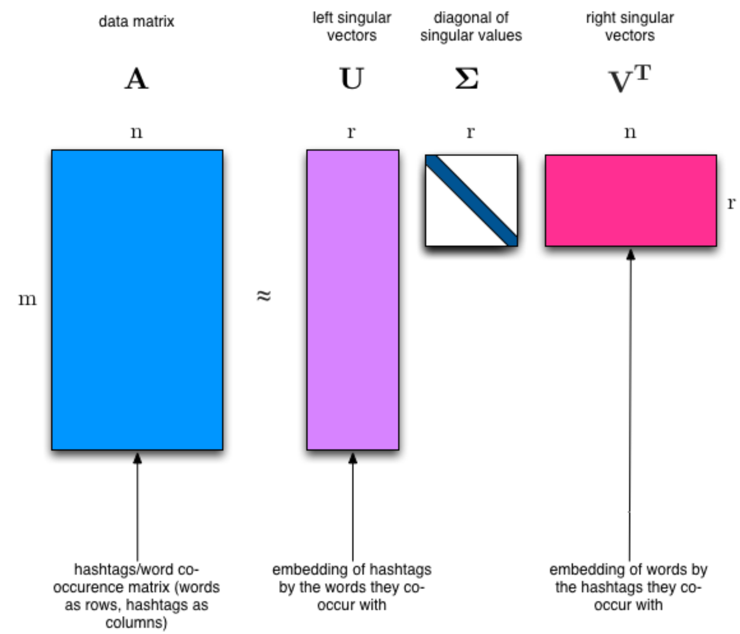
Singular Value Decomposition

Singular Value Decomposition:

Formally, the SVD of a real $m \times n$ matrix A is a factorization of the form $A = U \Sigma V^T$, where U is an $m \times m$ orthogonal matrix of left singular vectors, Σ is an $m \times n$ diagonal matrix of singular values, and V^T is an $n \times n$ orthogonal matrix of right singular vectors.

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$



Jacobian and Hessian Matrices

The **Jacobian** of $f : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix of first-order partial derivatives:

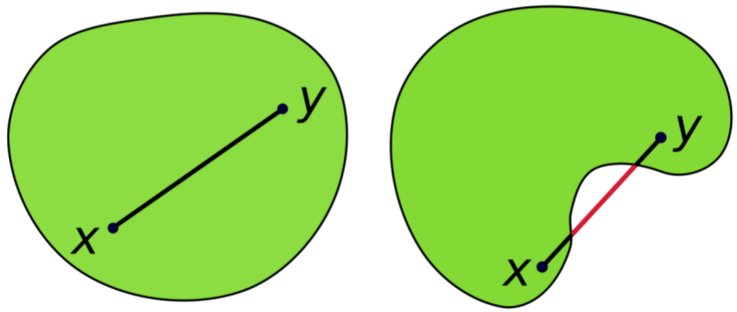
$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad \text{i.e.} \quad [\mathbf{J}_f]_{ij} = \frac{\partial f_i}{\partial x_j}$$

Note the special case $m = 1$, where $\nabla f = \mathbf{J}_f^\top$.

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

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

Convex Set and Function



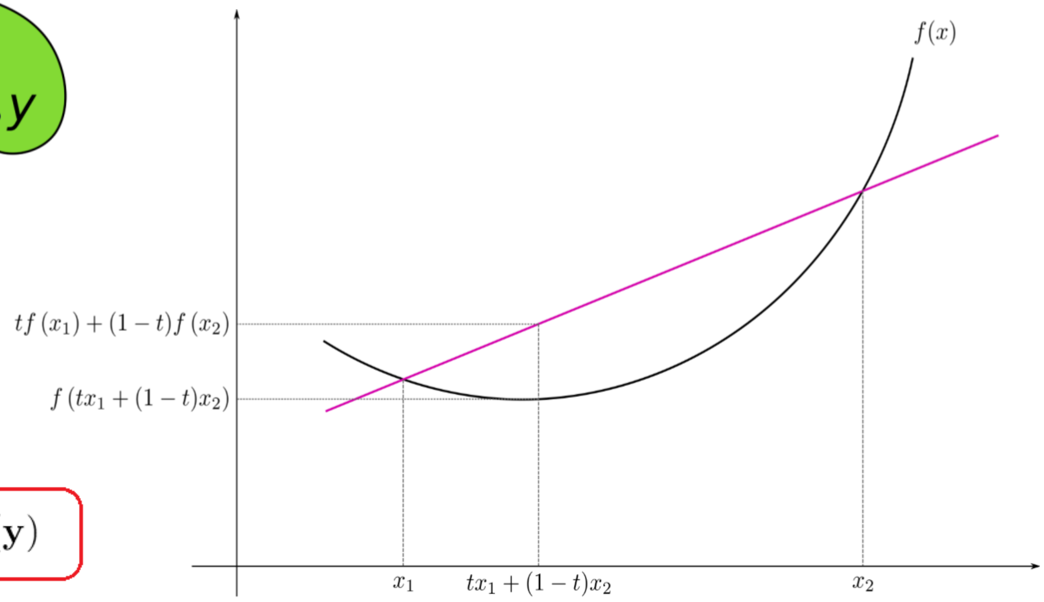


Figure 2: What convex functions look like

A function f is **convex** if

$$\boxed{f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})}$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and all $t \in [0, 1]$.

References

This slide of this class is modified from Lecture Notes of Dimitri P. Bertsekas and John N. Tsitsiklis – Introduction to Probability, MIT, 2000. & Wikipedia.
UNC Lecture Notes on Ecological Stats:
https://www.unc.edu/courses/2008fall/ecol/563/001/docs/lectures/lecture3.htm
Jeff Howbert Introduction to Machine Learning Winter 2012
Mathematics for Machine Learning Garrett Thomas
http:// gwthomas.github.io/docs/math4ml.pdf
https://rorasa.wordpress.com/2012/05/13/l0-norm-l1-norm-l2-norm-l-infinity-norm/
https://www.bilibili.com/video/av10852829/