Basic Mathematics

The foundations and four pillars of machine learning:

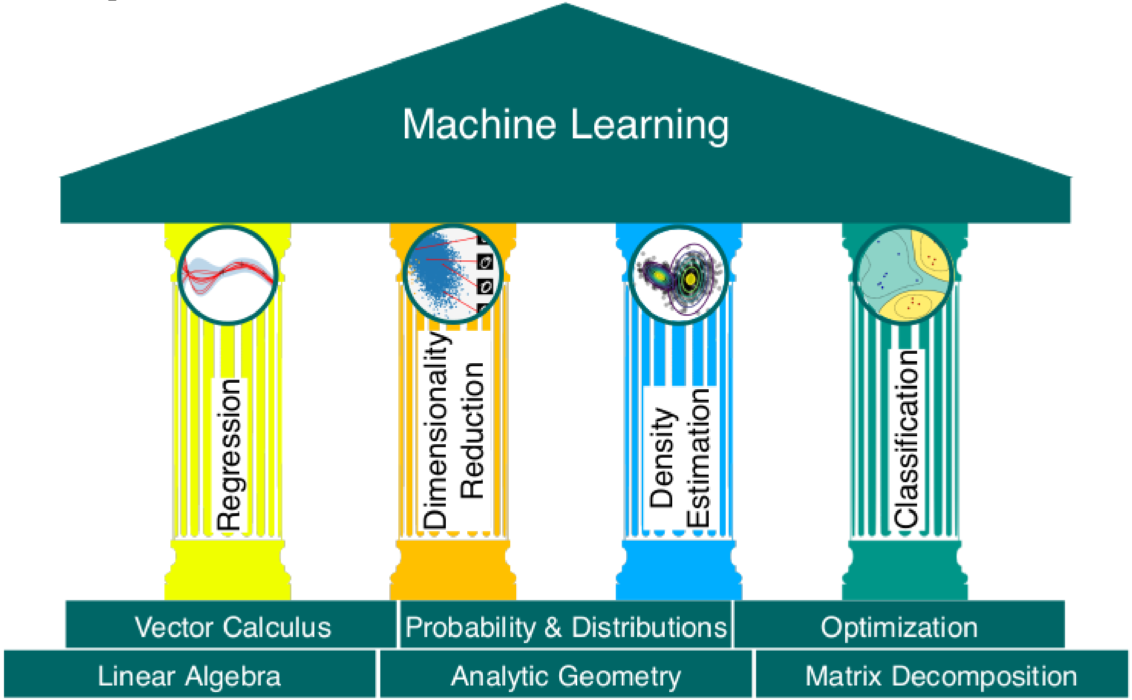


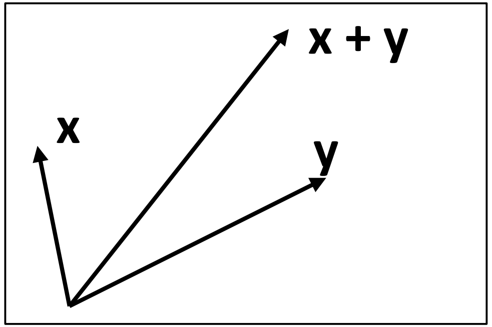
Image from: Mathematics for Machine Learning by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong, Cambridge University Press, 2020.

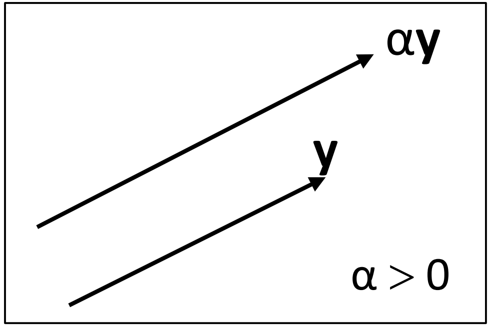
Outline:

1. Linear algebra
2. Analytic geometry
3. Vector calculus
4. Optimization (rest of semester)

1. Linear algebra

- Linear algebra is the study of vectors and certain rules to manipulate vectors.
- Vectors are often denoted by a small arrow above a letter, for example $\vec{x}$ and $\vec{y}$.
- We will use a bold letter to represent vectors, e.g. **x** and **y**.
- Vectors are special objects that can be added together and multiplied by scalars to produce another object of the same kind. For example, geometric vectors **x** and **y:**





1. Linear algebra

- An n-vector or n-dimensional vector **x**, for any positive integer n, is an n-tuple:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{pmatrix}$$

of real numbers. The real number x_i is the i-th component or element of the vector **x**.

- $\mathbb{R}^n$ is the n-dimensional Euclidean space. For a positive integer n, it is the set of all n-vectors.
- For example, $\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \in \mathbb{R}^3$ is an example of a triplet of numbers.
- Another example, $\alpha = 1.5 \in \mathbb{R}$.
- A row vector is a vector with just one row and n elements. For example, $\mathbf{c} = (-1 , 0, 1, -1)$

1. Linear algebra

- Vector addition: Let **x** and **y** $\in \mathbb{R}^n$. The sum of **x** + **y** is defined by:

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ .. \\ x_n + y_n \end{pmatrix}$$

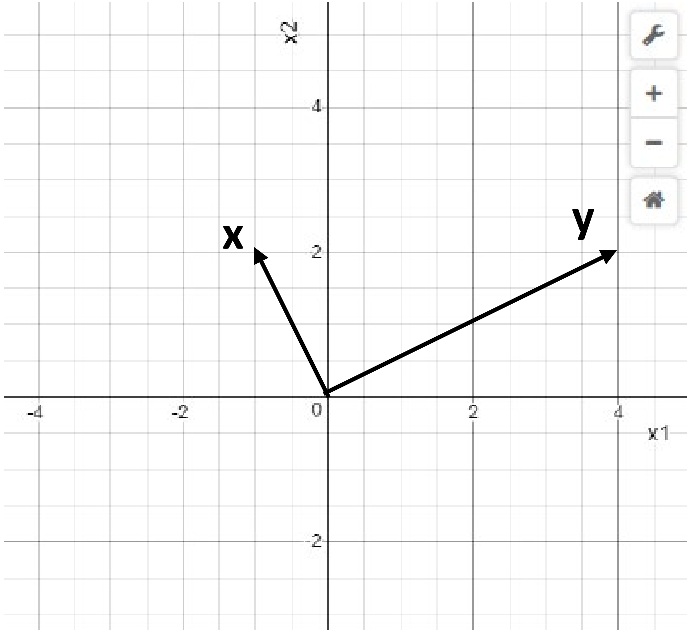
- The multiplication by a real number α is defined by

$$\alpha\mathbf{x} = \alpha \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ .. \\ \alpha x_n \end{pmatrix}$$

1. Linear algebra

- Dot product between two vectors: Let **x** and **y** $\in \mathbb{R}^n$, then the dot product (or scalar product) between **x** and **y** is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = x_1 y_1 + x_2 y_2 + \; \ldots . x_n y_n$$
$$= \sum_{i=1}^{n} x_i y_i$$



- Example: $\mathbf{x} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$, then

$$\mathbf{x}^T \mathbf{y} = -4 + 4 = 0$$

1. Linear algebra

- Linear dependence and independence: The vectors $\mathbf{x}^1, \mathbf{x}^2, \ \dots . \mathbf{x}^{\mathrm{m}} \in \mathbb{R}^n$ are said to be *linearly independent* if

$$\left.\begin{array}{c} \lambda_1\mathbf{x}^1 + \lambda_2\mathbf{x}^2 + \ \dots . \ \lambda_m\mathbf{x}^{\mathrm{m}} = 0 \\ \lambda_1, \ \lambda_2, \ \dots . , \lambda_m \ \in \mathbb{R} \end{array}\right\} \Rightarrow \ \lambda_1 = \lambda_2 = \ \dots . = \lambda_m = 0$$

Otherwise they are *linearly dependent* (at least one of the vectors $\mathbf{x}^1, \mathbf{x}^2, \ \dots . \mathbf{x}^{\mathrm{m}}$ can be defined as a linear combination of the others).

- Linear combination: The vector $\mathbf{x} \in \mathbb{R}^n$ is a linear combination of $\mathbf{x}^1, \mathbf{x}^2, \ \dots . \mathbf{x}^{\mathrm{m}} \in \mathbb{R}^n$ if $\mathbf{x} = \lambda_1\mathbf{x}^1 + \lambda_2\mathbf{x}^2 + \ \dots . \ \lambda_m\mathbf{x}^{\mathrm{m}}$ for some $\lambda_1, \ \lambda_2, \ \dots . , \lambda_m \ \in \mathbb{R}$.

1. Linear algebra

- Example 1: $n = 2, m = 2$, $\mathbf{x^1} = \begin{pmatrix}1\\0\end{pmatrix}$ and $\mathbf{x^2} = \begin{pmatrix}0\\1\end{pmatrix}$

$$\left.\begin{array}{l} \lambda_1\mathbf{x^1} + \lambda_2\mathbf{x^2} = \begin{pmatrix}\lambda_1\\0\end{pmatrix} + \begin{pmatrix}0\\\lambda_2\end{pmatrix} = \begin{pmatrix}\lambda_1\\\lambda_2\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \\ \lambda_1,\ \lambda_2 \in \mathbb{R} \end{array}\right\} \Rightarrow \lambda_1 = \lambda_2 = 0$$

Hence, the two vectors are linearly independent.

- Example 2: $m = 3$, $\mathbf{x^1} = \begin{pmatrix}1\\0\end{pmatrix}$, $\mathbf{x^2} = \begin{pmatrix}0\\1\end{pmatrix}$ and $\mathbf{x^3} = \begin{pmatrix}3\\5\end{pmatrix}$

Clearly $\mathbf{x^3} = 3\mathbf{x^1} + 5\mathbf{x^2}$, that is:

  - $\mathbf{x^3}$ is a linear combination of $\mathbf{x^1}$ and $\mathbf{x^2}$.
  - The vectors $\mathbf{x^1}$, $\mathbf{x^2}$ and $\mathbf{x^3}$ are not linearly independent.

1. Linear algebra

- <u>Matrix</u>: $\mathbb{R}^{m\times n}$ denotes the space of all m-by-n real matrices

$$\mathbf{A} \in \mathbb{R}^{m\times n} \Leftrightarrow \mathbf{A} = [A_{ij}] = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}$$

- The number of rows in the matrix **A** is m and the number of columns is n.

- The following linear system of equations will be encountered frequently:

$$\begin{array}{ccccccccc} A_{11}x_1 & + & A_{12}x_2 & + & \dots & + & A_{1n}x_n & = & b_1 \\ A_{21}x_1 & + & A_{22}x_2 & + & \dots & + & A_{2n}x_n & = & b_2 \\ A_{31}x_1 & + & A_{32}x_2 & + & \dots & + & A_{3n}x_n & = & b_3 \\ \dots & .. & \dots & .. & \dots & .. & \dots & .. & \dots \\ A_{m1}x_1 & + & A_{m2}x_2 & + & \dots & + & A_{mn}x_n & = & b_m \end{array}$$

A_{ij} and b_i are real numbers
i = 1,2, … m and j = 1,2, … n.

1. Linear algebra

- If we let $\mathbf{A_i}$ denote an n-vector whose n components are A_{ij}, j = 1,2 …n and if let

$\mathbf{x} \in \mathbb{R}^n$, then

$$\begin{array}{ccccccccc} A_{11}x_1 & + & A_{12}x_2 & + & \dots & + & A_{1n}x_n & = & b_1 \\ A_{21}x_1 & + & A_{22}x_2 & + & \dots & + & A_{2n}x_n & = & b_2 \\ A_{31}x_1 & + & A_{32}x_2 & + & \dots & + & A_{3n}x_n & = & b_3 \\ \dots & .. & \dots & .. & \dots & .. & \dots & .. & \dots \\ A_{m1}x_1 & + & A_{m2}x_2 & + & \dots & + & A_{mn}x_n & = & b_m \end{array}$$

is equivalent to $\mathbf{A_i}^T \mathbf{x} = b_i$, i = 1,2, … m.

- The n-vector $\mathbf{A_i}$ is the i-th row of the matrix

$$\mathbf{A_i} = \begin{pmatrix} A_{i1} \\ A_{i2} \\ .. \\ A_{in} \end{pmatrix}$$

1. Linear algebra

- The linear system can be written still in another form as follows:

$$\sum_{j=1}^{n} \boldsymbol{A}_{\bullet\mathbf{j}} x_j = \boldsymbol{b}$$

where the j-th column of matrix **A** is $\mathbf{A}_{\bullet\mathbf{j}} = \begin{pmatrix} A_{1j} \\ A_{2j} \\ \cdot\cdot \\ A_{mj} \end{pmatrix}$, x_j is the j-th component of **x**,

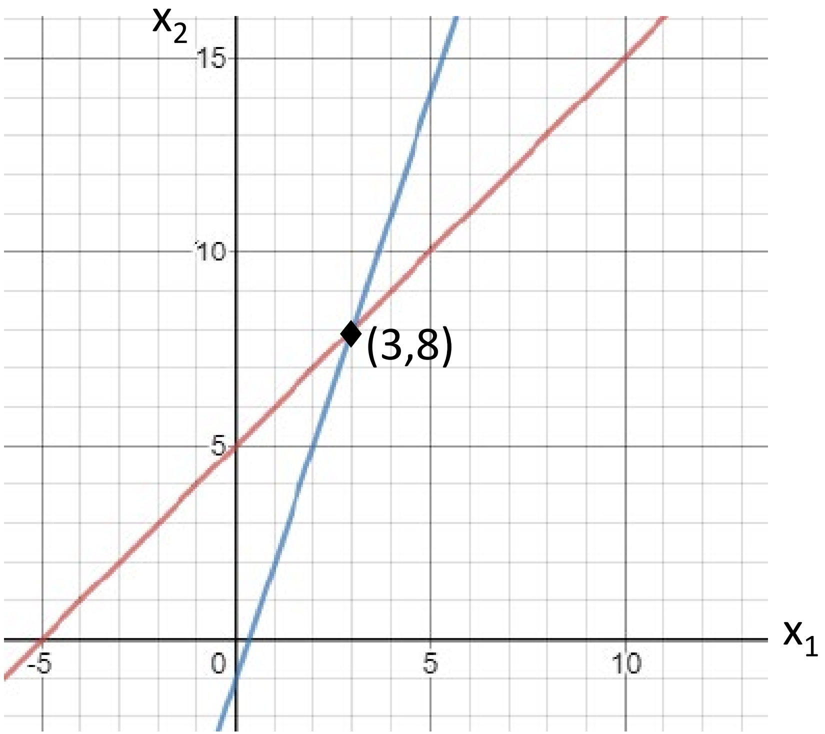
and **b** is the m-vector whose m components are b_1 ,b_2 , …. b_m

$\Rightarrow$ **b** is the linear combination of columns of **A** (if **x** exists).

- The linear system can be simply written as **Ax** = **b**.

1. Linear algebra

- Linear system of equations, example with m = n = 2:



$$x_1 - x_2 = -5$$
$$3x_1 - x_2 = 1$$

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 3 & -1 \end{bmatrix} \quad \mathbf{b} = \begin{pmatrix} -5 \\ 1 \end{pmatrix}$$

- Solution: $x_1 = 3$, $x_2 = 8$
- The solution space of a system of two linear equations with 2 variables is the intersection of two lines.
- Note:

$$\begin{pmatrix} -5 \\ 1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 3 \end{pmatrix} + 8\begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

1. Linear algebra

- The <u>sum</u> of two matrices **A** $\in \mathbb{R}^{m\times n}$ and **B** $\in \mathbb{R}^{m\times n}$ is defined as element-wise

sum, i.e. $\mathbf{A} + \mathbf{B} = \begin{bmatrix} A_{11}+B_{11} & \cdots & A_{1n}+B_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1}+B_{m1} & \cdots & A_{mn}+B_{mn} \end{bmatrix}$

- Multiplication by a scalar α: Given the matrix **A** $\in \mathbb{R}^{m\times n}$, the elements of α**A** are

$$\alpha\,\mathbf{A} = \begin{bmatrix} \alpha A_{11} & \cdots & \alpha A_{1n} \\ \vdots & \ddots & \vdots \\ \alpha A_{m1} & \cdots & \alpha A_{mn} \end{bmatrix}$$

1. Linear algebra

- Given the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times k}$, the elements $\mathbf{C}_{ij}$ of the product matrix $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m\times k}$ are computed as

$$\mathbf{C}_{ij} = \sum_{\ell=1}^{n} \boldsymbol{A}_{i\ell}\mathbf{B}_{\ell j} \quad \text{for all } i = 1,2, \ldots\ m, \quad j = 1,2, \ldots.\ k$$

The number of columns in **A** must be the same as the number of rows in **B** (= n).

- Example: $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$ $\quad$ $B = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}$

Then $\quad \mathbf{AB} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix}$

$$\mathbf{BA} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 4 & 2 \\ -2 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix}$$

Clearly, $\mathbf{AB} \neq \mathbf{BA}$

1. Linear algebra

- Identity matrix: We define the identity matrix in $\mathbb{R}^{n\times n}$ as the n×n matrix

containing 1 on the diagonal and 0 everywhere else.

- For example, when n = 4:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- A square matrix: the number of rows = the number of columns.
- For example, when n = 2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

1. Linear algebra

- Matrix transpose: For $\mathbf{A} \in \mathbb{R}^{n\times m}$, the matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$ with $\mathbf{B}_{ij} = \mathbf{A}_{ji}$ is called the <u>transpose</u> of **A**. We write $\mathbf{B} = \mathbf{A}^T$.
- For example:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 & 5 \\ 0 & 1 & -4 & 0 \end{bmatrix} \qquad \mathbf{A}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & -4 \\ 5 & 0 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 3 \\ 1 \\ 8 \end{bmatrix} \qquad \mathbf{x}^T = \begin{bmatrix} 3 & 1 & 8 \end{bmatrix}$$

- A square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is <u>symmetric</u> if $\mathbf{A} = \mathbf{A}^T$.

1. Linear algebra

- Inverse of a square matrix: Consider matrix **A** $\in \mathbb{R}^{n\times n}$. The matrix **B** $\in \mathbb{R}^{n\times n}$ with the property that **AB** = **I** $^{n\times n}$ = **BA** is called the inverse of **A** and denoted by **A**$^{-1}$.
- Not all matrices have inverse.
- If the inverse of **A** exists, then **A** is called regular/invertible/non-singular.
- Otherwise, **A** is singular or non-invertible.
- For n = 2, let **A** = $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, its inverse is **A**$^{-1}$ = $\frac{1}{D}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$ where the determinant D = $a_{11}\, a_{22}$ - $a_{12}\, a_{21}$ (when the determinant D = 0, **A** is singular).

1. Linear algebra

If the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, then the following statements are equivalent:

- $\mathbf{A}$ is invertible ($\mathbf{A}^{-1}$ exists)
- $\mathbf{A}\mathbf{x} = 0$ has only the trivial solution (that is, $\mathbf{x} = 0$ is the only solution)
- $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent for every n-vector $\mathbf{b}$.
- $\det(\mathbf{A}) \neq 0$ (determinant of matrix A is not 0)
- **A** has rank n (**A** has full rank).
- The rows of **A** are linearly independent.
- The columns of **A** are linearly independent.

1. Linear algebra

- Solving **linear system of equations by elimination**.

Suppose

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 3 \\ 1 & 3 & -4 \\ 3 & 1 & 2 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ -7 \\ 5 \end{bmatrix}$$

- Find $\mathbf{x} \in \mathbb{R}^3$ such that $\mathbf{Ax} = \mathbf{b}$

$$\begin{array}{ccccccc} 4\,x_1 & - & 2\,x_2 & + & 3\,x_3 & = & 1 \\ x_1 & + & 3\,x_2 & - & 4\,x_3 & = & -7 \\ 3\,x_1 & + & x_2 & + & 2\,x_3 & = & 5 \end{array}$$

1. Linear algebra

Solving **linear system of equations by elimination**

- Pivot row = r, pivot column = c
- Pivot number = $\mathbf{A}_{r,c}$
- New entries:
  - Pivot element new($\mathbf{A}_{r,c}$) = 1
  - All in row r, new($\mathbf{A}_{r,j}$) = $\mathbf{A}_{r,j}$ / $\mathbf{A}_{r,c}$ for all column j
  - All in column c, new($\mathbf{A}_{i,c}$) = 0 for all row i, except new($\mathbf{A}_{r,c}$) = 1
  - All other entries:

    new($\mathbf{A}_{i,j}$) = $\mathbf{A}_{i,j}$ - $\mathbf{A}_{i,c}$ × $\mathbf{A}_{r,j}$/$\mathbf{A}_{r,c}$

    Example 1: i = 1, j = 4, r = 2, c = 1

    1 – [(4) × (-7)]/1 = 29

    Example 2: i = 3, j = 3, r = 2, c = 1

    2 – [(3) × (-4)]/1 = 14

$$\begin{bmatrix} 4 & -2 & 3 & \| & 1 \\ 1 & 3 & -4 & \| & -7 \\ 3 & 1 & 2 & \| & 5 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} 0 & -14 & 19 & \| & 29 \\ 1 & 3 & -4 & \| & -7 \\ 0 & -8 & 14 & \| & 26 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} 0 & 0 & -5.5 & \| & -16.5 \\ 1 & 0 & 1.25 & \| & 2.75 \\ 0 & 1 & -1.75 & \| & -3.25 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0 & 1 & \| & 3 \\ 1 & 0 & 0 & \| & -1 \\ 0 & 1 & 0 & \| & 2 \end{bmatrix}$$

Solution: x_1 = -1, x_2 = 2, x_3 = 3

1. Linear algebra

Finding the inverse of matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ by Gauss-Jordan method:

- For example, $\mathbf{A} = \begin{bmatrix} 2 & 4 & 1 \\ -1 & 1 & -1 \\ 1 & 4 & 0 \end{bmatrix}$
- Start with $[\mathbf{A} \parallel \mathbf{I}_3] = \left[\begin{array}{ccc|ccc} 2 & 4 & 1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 0 & 1 \end{array}\right]$
- After pivoting with pivot row r = 1, pivot column = 1:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & ½ & ½ & 0 & 0 \\ 0 & 3 & -½ & ½ & 1 & 0 \\ 0 & 2 & -½ & -½ & 0 & 1 \end{array}\right]$$

1. Linear algebra

Finding the inverse of matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ by Gauss-Jordan method:

- After pivoting with pivot row = 2, pivot column = 2:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 3 & -\frac{1}{2} & \frac{1}{2} & 1 & 0 \\ 0 & 2 & -\frac{1}{2} & -\frac{1}{2} & 0 & 1 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & \frac{5}{6} & \frac{1}{6} & -\frac{2}{3} & 0 \\ 0 & 1 & -\frac{1}{6} & \frac{1}{6} & \frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{6} & -\frac{5}{6} & -\frac{2}{3} & 1 \end{array}\right]$$

- Finally, after pivoting with pivot row = 3, pivot column = 3:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -4 & -4 & 5 \\ 0 & 1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 1 & 5 & 4 & -6 \end{array}\right] \qquad A^{-1} = \begin{bmatrix} -4 & -4 & 5 \\ 1 & 1 & -1 \\ 5 & 4 & -6 \end{bmatrix}$$

2. Analytic geometry

- The *p-norm* of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined by

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \ldots.. + |x_n|^p)^{1/p}$$

of which

- The Manhattan (or ℓ_1) norm: $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots.. + |x_n|$
- The Euclidean norm: $\|\mathbf{x}\|_2 = (|x_1|^2 + |x_2|^2 + \ldots.. + |x_n|^2)^{1/2}$
- The infinity norm: $\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \ldots., |x_n|)$

are the most important.

- When the subscript is dropped from the norm, we mean the 2 norm (Euclidean).

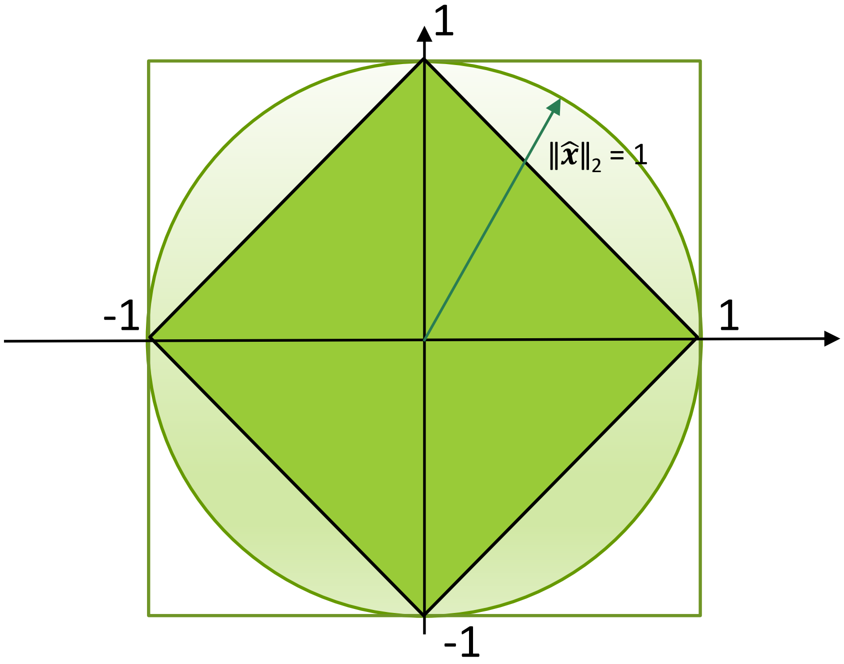
2. Analytic geometry

- Unit vector: Given a non-zero vector $\mathbf{x} \in \mathbb{R}^n$, we can define the unit vector

$$\hat{\boldsymbol{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

then $\hat{\boldsymbol{x}}$ has the same direction as $\mathbf{x}$

and its norm $\|\hat{\boldsymbol{x}}\|$ is 1.



Circle: $\|\hat{\boldsymbol{x}}\|_2 = 1$

Square: $\|\hat{\boldsymbol{x}}\|_\infty = 1$

Diamond: $\|\hat{\boldsymbol{x}}\|_1 = 1$

2. Analytic geometry

- Symmetric, positive (semi-) definite matrix: we say that the symmetric matrix

$\mathbf{A} \in \mathbb{R}^{n\times n}$ is <u>positive definite</u> if for all vector $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$, the following holds:

$$\mathbf{x}^T \mathbf{A}\mathbf{x} > 0.$$

- The symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is <u>positive semidefinite</u> if for all vector $\mathbf{x} \in \mathbb{R}^n$

$$\mathbf{x}^T \mathbf{A}\mathbf{x} \geq 0.$$

- For example: $\mathbf{A} = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}$, then $\mathbf{x}^T \mathbf{A}\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 9x_1^2 + 12\, x_1x_2 + 5\, x_2^2$

$= (3\, x_1 + 2x_2)^2 + x_2^2 > 0$ for any $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \neq \mathbf{0}$. Hence, $\mathbf{A}$ is positive definite.

2. Analytic geometry

- The length of vector $\mathbf{x} \in \mathbb{R}^n$ can be computed as the square root of the dot product:

$$\|\mathbf{x}\| = (\mathbf{x}^T \mathbf{x})^{1/2} = (|x_1|^2 + |x_2|^2 + ... + |x_n|^2)^{1/2}$$

- The Euclidean distance between $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ is computed as follows:

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = (|x_1 - y_1|^2 + |x_2 - y_2|^2 + ... + |x_n - y_n|^2)^{1/2}$$

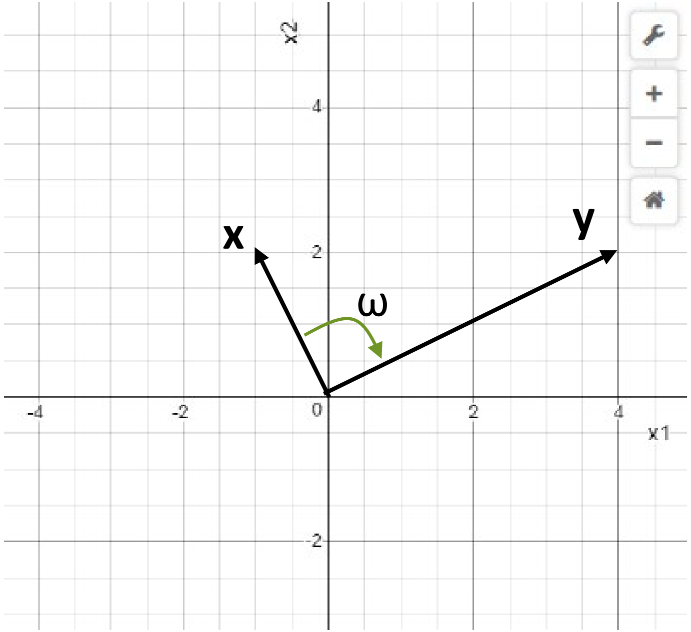
- The following properties hold for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

1. Non-negativity: $d(\mathbf{x},\mathbf{y}) \geq 0$
2. Symmetry: $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$
3. Triangular inequality: $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$

2. Analytic geometry

- Angle between two vectors: Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the angle ω between these vectors is such that:

$$\cos \omega = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \; \|\boldsymbol{y}\|}$$



- Example: $\mathbf{x} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$, then

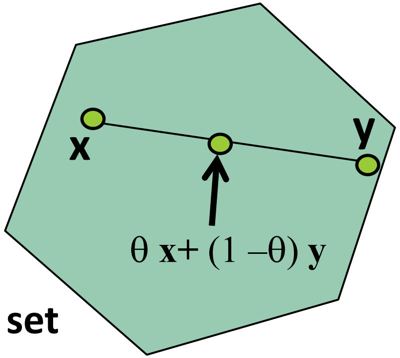
$$\mathbf{x}^T \mathbf{y} = -4 + 4 = 0$$

$\|\mathbf{x}\| = \sqrt{1+4} = \sqrt{5}$ and $\|\mathbf{y}\| = \sqrt{16+4} = \sqrt{20}$

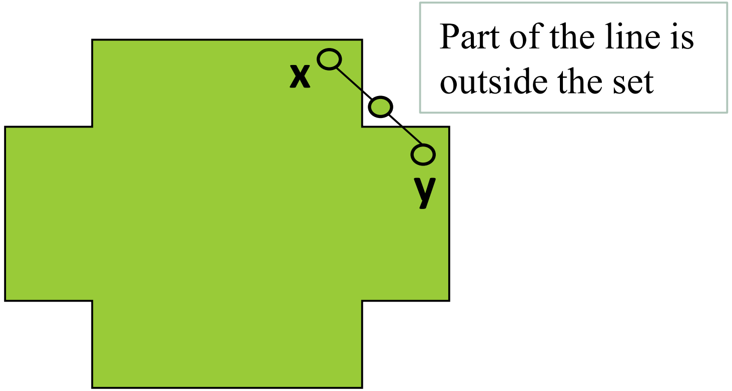
$\cos \omega = 0$, or $\omega = \pi/2 = 90^0$ (**x** and **y** are orthogonal)

2. Analytic geometry

- Convex set: A set C is a <u>*convex set*</u> if for any $\mathbf{x}, \mathbf{y} \in C$ and for any scalar θ with $0 \leq \theta \leq 1$, we have $\theta\mathbf{x} + (1 - \theta)\mathbf{y} \in C$.
- Convex sets are sets such that a straight line connecting any two elements of the set lie inside the set.



Convex set



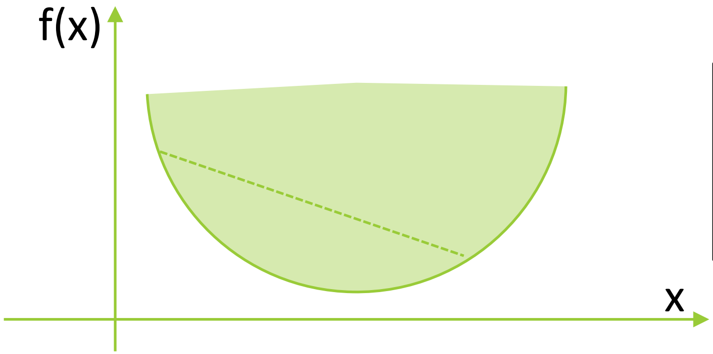
Non-convex set

2. Analytic geometry

- Convex function: let function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function whose domain is a convex set. The function f is a <u>*convex function*</u> if for all **x**, **y** in the domain and for any scalar θ with $0 \leq \theta \leq 1$, we have $f(\theta \mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y})$

A straight line between any two points of the function lie <u>above</u> the function.

- A convex function is a bowl-like object. The *epigraph* of a convex function is a convex set.



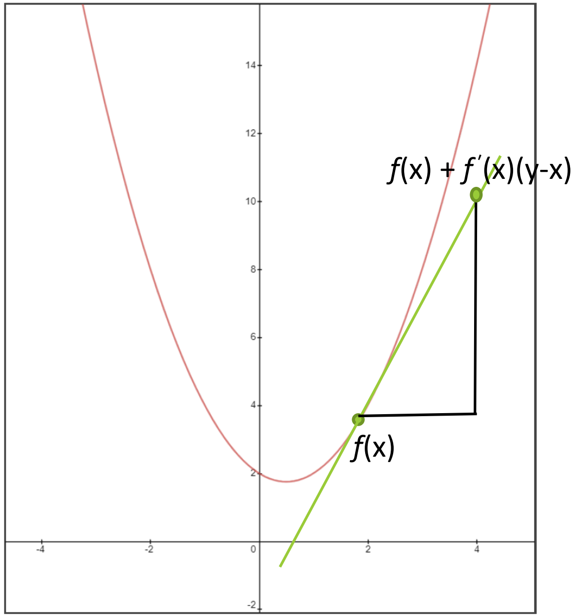
Epi f = $\{(\mathbf{x},\mu) : \mathbf{x} \in \mathbb{R}^n, \mu \in \mathbb{R}, \mu \geq f(x)\} \subseteq \mathbb{R}^{n+1}$

2. Analytic geometry

- If the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, then it is convex if and only if for any two points **x**, **y** it holds that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$$

- Example: $f(x) = x^2 - x + 2$
  - $\nabla_x f(x) = f'(x) = 2x - 1$
  - At $x = 2$, $f(x) = 4$ and $f'(x) = 3$
  - $f(x) + \nabla_x f(x)^T(y - x) = 4 + 3(y - 2) = 3y - 2$
  - Check that $f(y) = y^2 - y + 2 \geq 3y - 2$ for all y.



2. Analytic geometry

- If the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice-differentiable, then it is convex if and only if its second derivative matrix (Hessian) is positive semidefinite.
- Example: $f(x) = x^2 - x + 2$
  - $\nabla_x f(x) = f'(x) = 2x - 1$
  - $H = \nabla^2_x f(x) = f''(x) = 2 > 0$, hence $f(x)$ is (strictly-) convex.
- A *concave function* is the negative of a convex function.
- A linear function is both convex and concave.

2. Analytic geometry

- Example 1: $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$ ⇦ function is not convex
- $f'(x) = 4x^3 + 21x^2 + 10x - 17$
- The derivative $f'(x) = 0$ when

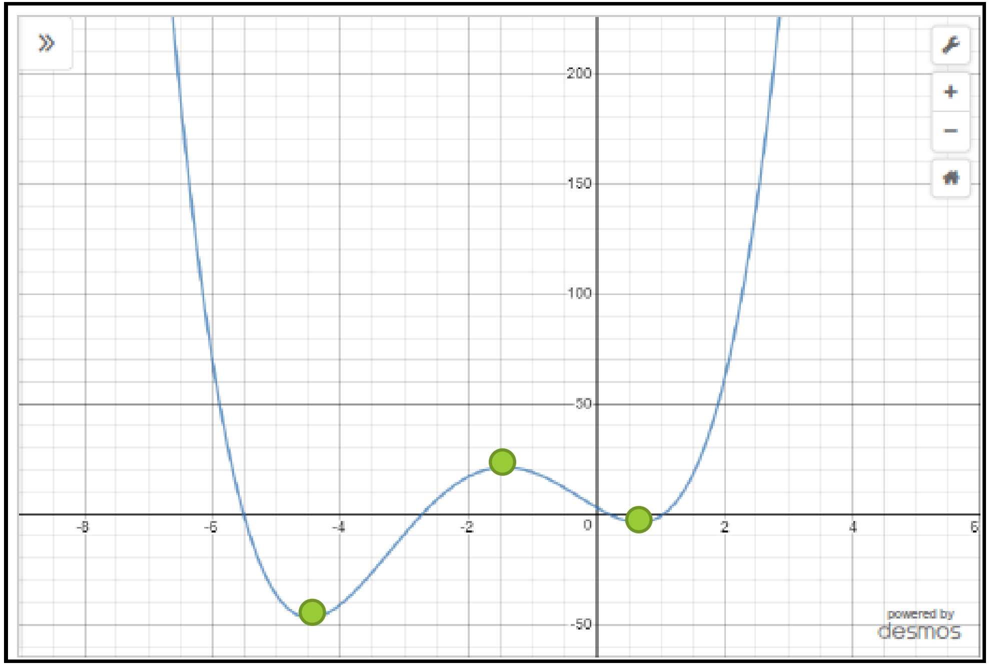
  $x \approx -4.48, -1.43, 0.66$
- The second derivative:

  $f''(x) = 12x^2 + 42x + 10$

At $x = -4.48$, $f''(x) = 62.70 > 0$ ⇨ local minimum

At $x = -1.43$, $f''(x) = -25.60 < 0$ ⇨ local maximum

At $x = 0.66$, $f''(x) = 42.94 > 0$ ⇨ local minimum



3. Vector calculus

- A function f is a single-valued mapping from a set X into a set Y. That is for each $\mathbf{x} \in X$, the image set $f(\mathbf{x})$ consists of a single element of Y.
- The domain of f is X and we say that f is defined on X.
- The range of f is $f(X) = \cup_{\mathbf{x} \in X} f(\mathbf{x})$.
- A numerical function θ is a function from a set X into $\mathbb{R}$.
- An m-dimensional vector function f is a function from a set X into $\mathbb{R}^m$. The m components of the vector $f(\mathbf{x})$ are denoted by $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})$. Each f_i is a numerical function on X.

3. Vector calculus

- An m-dimensional <u>vector function</u> f defined on $\mathbb{R}^n$ is said to be <u>linear</u> if

$$f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

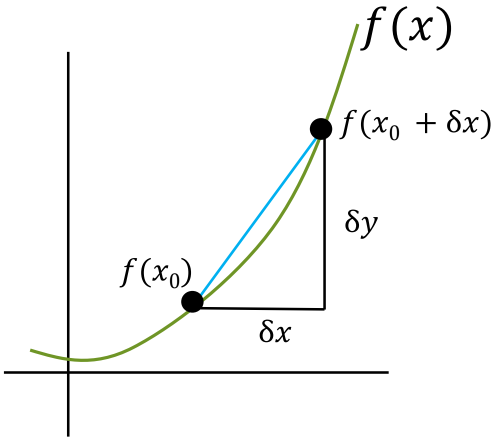
where **A** is some fixed m×n matrix and **b** is some fixed vector in $\mathbb{R}^m$.

- The <u>difference quotient</u> of a *univariate function* y = f(x), x,y $\in$ $\mathbb{R}$ is

$$\frac{\delta y}{\delta x} := \frac{f(x+\delta x) - f(x)}{\delta x}$$

- For h > 0, the derivative of f at x is defined as the limit

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



3. Vector calculus

- The Taylor series is a representation of a function *f* as an infinite sum of terms determined using derivatives of *f* evaluated at x_0.
- The Taylor polynomial of degree n of *f*: $\mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

where $f^{(k)}(x_0)$ is the k^{th} derivative of *f* at x_0 and $f^{(k)}(x_0)/k!$ are the coefficients of the polynomial.

- When $x_0 = 0$, we obtain the Maclaurin series.

3. Vector calculus

- Taylor polynomial example: consider the function

$$f(x) = x^4$$

- To obtain the Taylor polynomial of T_6 evaluated at $x_0 = 1$, compute:
  - $f(1) = 1$, $f^{(1)}(x_0) = f'(x_0) = 4\,x_0^3 = 4$, $f^{(2)}(1) = f''(1) = 12\,x_0^2 = 12$,
  - $f^{(3)}(x_0) = 24\,x_0 = 24$, $f^{(4)}(x_0) = 24$, $f^{(5)}(x_0) = f^{(6)}(x_0) = 0$

- The polynomial is $T_6(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$

$$= 1 + 4(x-1) + (12/2!)(x-1)^2 + (24/3!)(x-1)^3 + (24/4!)(x-1)^4 + 0 + 0$$

$$= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0 + 0 = x^4$$

3. Vector calculus

- Differentiation rules: Let $f'(x)$ be the derivative of the function $f(x)$ and $g'(x)$ be the derivative of the function $g(x)$.

  - Product rule: $(f(x)\ g(x))' = f'(x)\ g(x) + f(x)\ g'(x)$

  - Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)\ g(x) - f(x)\ g'(x)}{(g(x))^2}$

  - Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$

  - Chain rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))\, f'(x)$

3. Vector calculus

- Chain rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))\,f'(x)$
- Example 1:
  - $h(x) = (2x + 1)^4 = g(f(x))$
  - $f(x) = 2x + 1$
  - $g(f) = f^4$

  We obtain the derivatives of *f* and *g* as

  - $f'(x) = 2$
  - $g'(f) = 4f^3$

  Hence, $h'(x) = g'(f(x))\,f'(x) = (4f^3)\,2 = 4(2x + 1)^3\,2 = 8(2x + 1)^3$

3. Vector calculus

- Partial derivative: For a function f: $\mathbb{R}^n \rightarrow \mathbb{R}$ of n variables $x_1, x_2 \ldots x_n$, we define partial derivatives as follows:

  - $\frac{\partial f}{\partial x_1} := \lim_{h \to 0} \frac{f(x_1+h, x_2, \ldots, xn) - f(\boldsymbol{x})}{h}$
  - $\ldots.$
  - $\frac{\partial f}{\partial x_n} := \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_n+h) - f(\boldsymbol{x})}{h}$

- The vector $\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \ldots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$ is called the <u>gradient</u> of f.

3. Vector calculus

- Example 2: let the function $f(\mathbf{x}) = f(x_1,x_2) = x_1^2 + 2x_2^2 - 2x_1x_2 - 2x_1$, then
  - $\frac{\partial f}{\partial x_1} = 2x_1 - 2x_2 - 2$
  - $\frac{\partial f}{\partial x_2} = 4x_2 - 2x_1$

Quadratic function: $f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{p}^T\mathbf{x}$

then $\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{p}$

Here $\mathbf{A} = \begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$ $\mathbf{p} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}$

- The gradient of $f(\mathbf{x})$ is $\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2x_2 - 2 \\ 4x_2 - 2x_1 \end{pmatrix}$
- Example 3: let $f(x,y) = (x + 2y^3)^2$, then its gradient is

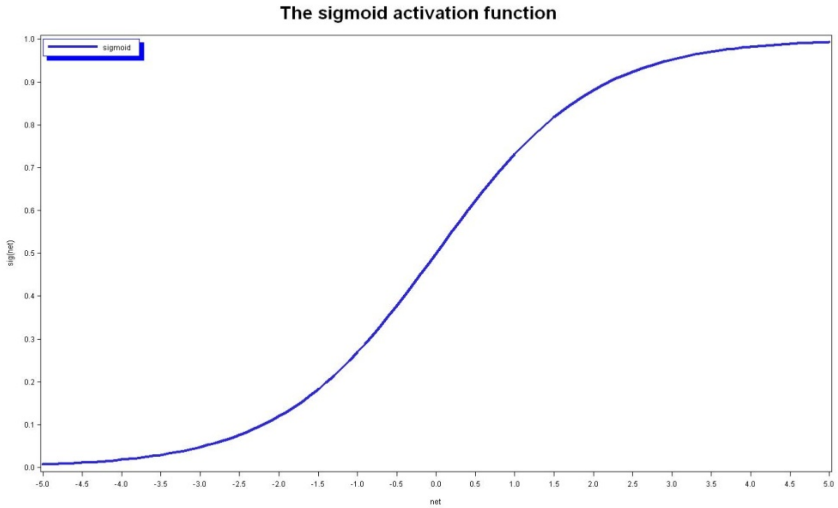
$$\nabla f(x,y) = \begin{pmatrix} 2(x + 2y^3) \\ 2(x + 2y^3)(6y^2) \end{pmatrix} = \begin{pmatrix} 2(x + 2y^3) \\ 12(x + 2y^3)y^2 \end{pmatrix}$$

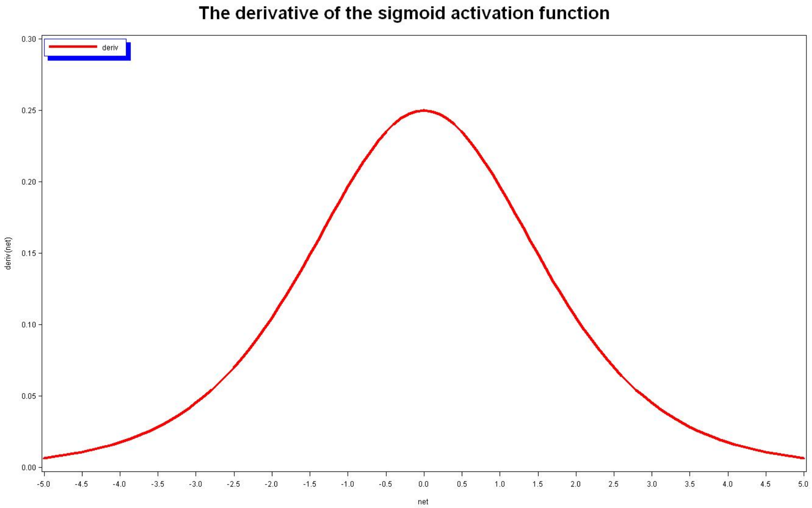
3. Vector calculus

- Example 4: The sigmoid function is defined as

$$\sigma (y) = 1/(1 + \exp (-y)) = 1/(1 + e^{-y})$$

Its derivative is: $\sigma '(y) = \sigma (y) (1 - \sigma (y))$





3. Vector calculus

Example 4 (continued):

- Suppose $y = w_0 + 2w_1 - 3w_2$ and

$$y = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}^{T} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$$

$E(w_0, w_1, w_2) = \frac{1}{2}\,[1 - \sigma(y)]^2$

$= \frac{1}{2}\,[1 - \sigma(w_0 + 2w_1 - 3w_2)]^2$, then its gradient is:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \begin{pmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \\ \frac{\partial E}{\partial w_2} \end{pmatrix} = \frac{1}{2} \times [2] \times [1 - \sigma(y)] \times [-1] \times [\sigma(y)\,(1 - \sigma(y)] \times \begin{pmatrix} \frac{\partial y}{\partial w_0} \\ \frac{\partial y}{\partial w_1} \\ \frac{\partial y}{\partial w_2} \end{pmatrix}$$

$$= [-1] \times \sigma(y) \times [1 - \sigma(y)]^{2} \times \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}$$

3. Vector calculus

- Gradients of vector functions: For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a vector $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$,

we define its ***Jacobian*** to be the collection of first-order derivatives:

$$J = \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

- Note: J is a matrix with m rows and n columns, $J \in \mathbb{R}^{m \times n}$
- Recall: The m components of the vector $f(\mathbf{x})$ are denoted by $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})$.

Each f_i is a numerical function on $\mathbb{R}^n$.

3. Vector calculus

- Gradients of vector functions: Example 1.
- Let f(**x**) = $\begin{pmatrix} 2x_1 + 5x_2 \\ x_1 - 8\, x_2 \\ 3x_2 \end{pmatrix}$ = $\begin{bmatrix} 2 & 5 \\ 1 & -8 \\ 0 & 3 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$
- Then J = $\nabla_{\mathbf{x}} f(\mathbf{x})$ = $\begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} \\ \frac{\partial f_3(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_3(\boldsymbol{x})}{\partial x_2} \end{bmatrix}$ = $\begin{bmatrix} 2 & 5 \\ 1 & -8 \\ 0 & 3 \end{bmatrix}$
- In general, for a linear function $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ where A is some fixed m×n matrix

and **b** is some fixed vector in $\mathbb{R}^m$, its derivative (= Jacobian) $\nabla_{\mathbf{x}} f(\mathbf{x}) = A$

3. Vector calculus

- Higher order derivatives: Consider a function: $\mathbb{R}^2 \to \mathbb{R}$ of two variables x and y.
- Let: $\frac{\partial^2 f}{\partial x^2}$ is the second partial derivative of *f* with respect to x.
- $\frac{\partial^n f}{\partial x^n}$ is the n-th partial derivative of *f* with respect to x.
- $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(\frac{\partial f}{\partial x})$ is the partial derivative obtained by first partial differentiating *f*

with respect to x and then with respect to y.

- $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(\frac{\partial f}{\partial y})$ is the partial derivative obtained by first partial differentiating *f*

with respect to y and then with respect to x.

3. Vector calculus

- Higher order derivatives: Consider a function: $\mathbb{R}^2 \rightarrow \mathbb{R}$ of two variables x and y.
- The Hessian is the collection of all second order derivatives.

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

- Example 1: let the function $f(\mathbf{x}) = f(x_1,x_2) = x_1^2 + 2x_2^2 - 2x_1x_2 - 2x_1$, then
  - The gradient of $f(\mathbf{x})$ is $\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2x_2 - 2 \\ 4x_2 - 2x_1 \end{pmatrix}$
  - The Hessian H = $\begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$ is a symmetric matrix.

3. Vector calculus

- Hessian matrix: Example 2.

- Let $f(x_1,x_2,x_3) = -x_1 - x_2 + \frac{1}{2}(x_1)^2 + (x_2)^2 - x_1x_2 + x_3$

- The gradient is $\nabla_x f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{pmatrix} = \begin{pmatrix} -1 + x_1 - x_2 \\ -1 + 2x_2 - x_1 \\ 1 \end{pmatrix}$

- The Hessian is $H = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a symmetric matrix.

3. Vector calculus

- Hessian matrix: Example 3. Suppose we have the following data from n = 10 apartments:

$$X = \begin{pmatrix} 1 & 500 & 4 & 8 \\ 1 & 550 & 7 & 50 \\ 1 & 620 & 9 & 7 \\ 1 & 630 & 5 & 24 \\ 1 & 665 & 8 & 100 \\ 1 & 700 & 4 & 8 \\ 1 & 770 & 10 & 7 \\ 1 & 880 & 12 & 50 \\ 1 & 920 & 14 & 8 \\ 1 & 1000 & 9 & 24 \end{pmatrix}$$

Constant	Size	Floor	Broadband rate
1	500	4	8
1	550	7	50
1	620	9	7
1	630	5	24
1	665	8	100
1	700	4	8
1	770	10	7
1	880	12	50
1	920	14	8
1	1000	9	24

$$Y = \begin{pmatrix} 320 \\ 380 \\ 400 \\ 390 \\ 385 \\ 410 \\ 480 \\ 600 \\ 570 \\ 620 \end{pmatrix}$$

Rental price
320
380
400
390
385
410
480
600
570
620

3. Vector calculus

- Hessian matrix: Example 3 (continued).
- The *linear regression* model for predicting **Y** as a linear function of **X** can be expressed as:

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{e}$$

where **Y** is an n×1 vector of target feature, **X** is n×p matrix of descriptive features,

$\boldsymbol{\beta}$ is an p by 1 vector of parameters (weights) and **e** is an n×1 vector of errors,

n is the number of instances, p is the number of parameters.

- The least squares estimators for $\boldsymbol{\beta}$: minimize the sum of squared errors

$$\min S(\boldsymbol{\beta}) = \min \tfrac{1}{2}(e_1^2 + e_2^2 + \ldots\ldots e_n^2) = \tfrac{1}{2}\,\mathbf{e}^T\mathbf{e} = \tfrac{1}{2}(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})$$

Take the derivative of $S(\boldsymbol{\beta})$ and set it to **0**:

$$\nabla_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = -\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}) = \mathbf{0} \Rightarrow \mathbf{X}^T\mathbf{X}\,\boldsymbol{\beta} - \mathbf{X}^T\mathbf{Y} = \mathbf{0},$$

we have $\mathbf{X}^T\mathbf{X}\,\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$ and $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ if $(\mathbf{X}^T\mathbf{X})^{-1}$ exists.

The Hessian of $S(\boldsymbol{\beta})$ is $\nabla_{\boldsymbol{\beta}}^2 S(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{X}$

3. Vector calculus

- Hessian matrix: Example 3 (continued).

$$\mathbf{X} = \begin{pmatrix} 1 & 500 & 4 & 8 \\ 1 & 550 & 7 & 50 \\ 1 & 620 & 9 & 7 \\ 1 & 630 & 5 & 24 \\ 1 & 665 & 8 & 100 \\ 1 & 700 & 4 & 8 \\ 1 & 770 & 10 & 7 \\ 1 & 880 & 12 & 50 \\ 1 & 920 & 14 & 8 \\ 1 & 1000 & 9 & 24 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 320 \\ 380 \\ 400 \\ 390 \\ 385 \\ 410 \\ 480 \\ 600 \\ 570 \\ 620 \end{pmatrix}$$

n = 10, p = 4

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 10 & 7235 & 82 & 286 \\ 7235 & 5479725 & 62840 & 203810 \\ 82 & 62840 & 772 & 2395 \\ 286 & 203810 & 2395 & 16442 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 4555 \\ 3447725 \\ 39770 \\ 128300 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 19.5615 \\ 0.5487 \\ 4.9635 \\ -0.0621 \end{pmatrix}$$

Regression model:

$\hat{Y}$ = 19.5615 + 0.5487 Size + 4.9635 Floor – 0.0621 Broadband

Reference:

Mathematics for Machine Learning by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong, Cambridge University Press, 2020, Chapters 2,3,5.