

Math review

COMP 4630 | Winter 2026

Charlotte Curtis

Math review

- MATH 1203: Linear algebra
- MATH 1200: Differential calculus
- MATH 2234: Statistics

Further reading:

- Linear algebra: [notebook](#), [deep learning book](#)
- Calculus: [notebook](#)

Linear algebra

Vectors are multidimensional quantities
(unlike **scalars**):

$$\vec{v} = \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

A common **vector space** is \mathbb{R}^2 , or the 2D Euclidean plane. Example:

$$\mathbf{v}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

Vector operations

- **Addition:** $\mathbf{v}_1 + \mathbf{v}_2 = \begin{bmatrix} v_{11} + v_{21} \\ v_{12} + v_{22} \end{bmatrix}$
- **Scalar multiplication:** $c\mathbf{v} = \begin{bmatrix} cv_1 \\ cv_2 \end{bmatrix}$
- **Dot product:** $\mathbf{v}_1 \cdot \mathbf{v}_2 = v_{11}v_{21} + v_{12}v_{22}$ (yields a scalar)
 - Can be thought of as the **projection** of one vector onto another, or how much two vectors are aligned in the same direction

Vector norms

- The **norm** of a vector is a measure of its length
- Most common is the **Euclidean norm** (or L^2 norm):

$$\|\mathbf{v}\|_2 = \|\mathbf{v}\| = \sqrt{\left(\sum_{i=1}^n v_i^2\right)}$$

- You might also see the L^1 norm, particularly as a regularization term:

$$\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$$

Useful vectors

- **Unit vector:** A vector with a norm of 1, e.g. $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- **Normalized vector:** A vector divided by its norm, e.g. $\mathbf{v} = \hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$
- Dot product can also be written as $\mathbf{v}_1 \cdot \mathbf{v}_2 = \|\mathbf{v}_1\| \|\mathbf{v}_2\| \cos(\theta)$

Yes, a normalized vector is also a unit vector, main difference is in context and notation

Matrices

A **matrix** is a 2D array of numbers:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Notation: Element a_{ij} is in row i , column j , also written as A_{ij} .

Rows then columns! $M \times N$ matrix has M rows and N columns

Matrix operations

- **Addition:** element-wise *if* dimensions match. $A + B = B + A$
- **Scalar multiplication:** just like vectors
- **Matrix multiplication:** $C = AB$ where the elements of C are:

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}$$

- Multiply and sum rows of A with columns of B
- Usually, $AB \neq BA$

Matrix multiplication examples

Matrix times a matrix:

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 3 \\ -4 & 5 \end{bmatrix}, B = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 3 & 7 \end{bmatrix}$$

Matrix times a vector:

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Where we left off on January 14

Matrix transpose

- **Transpose:** A^T swaps rows and columns

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, A^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

- **Inverse:** just as $\frac{1}{x} \cdot x = 1$, $A^{-1}A = I$, where I is the identity matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, A^{-1} = \begin{bmatrix} -2 & 1 \\ 1.5 & -0.5 \end{bmatrix}, A^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Not every matrix is invertible!

Calculus: Notation

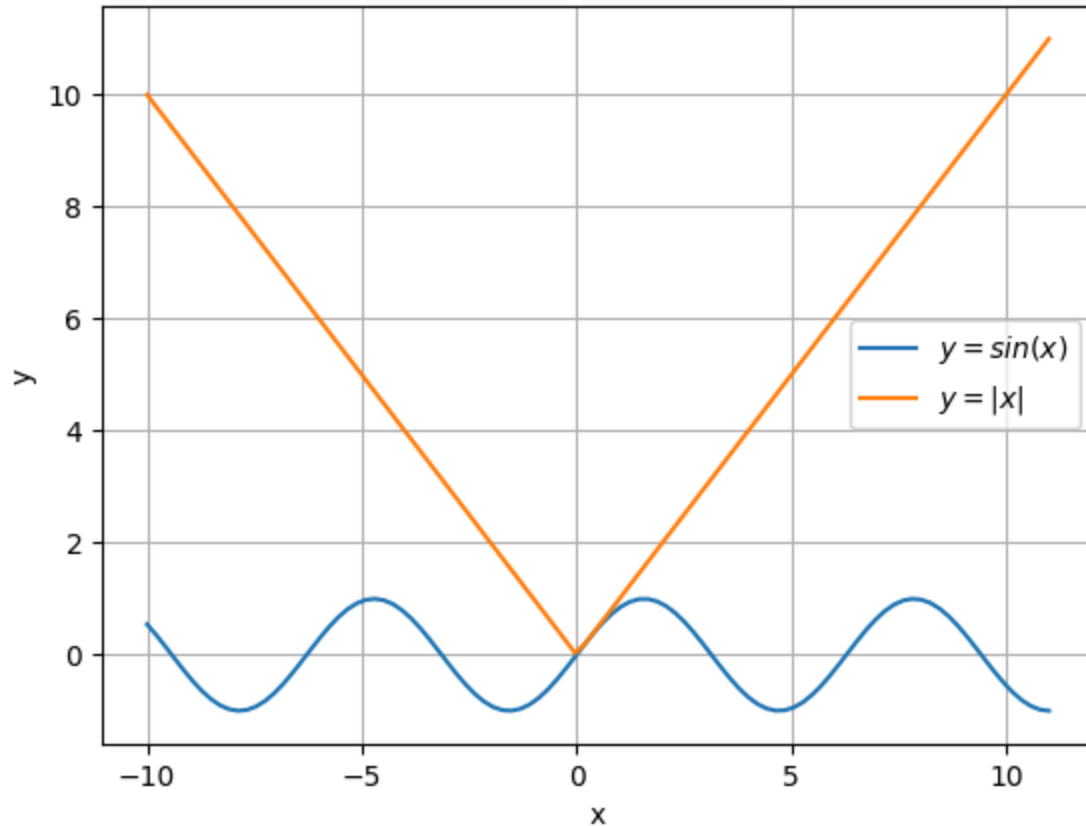
The **derivative** of a function $y = f(x)$ is represented as:

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

The **second derivative** is denoted:

$$f''(x) = \frac{d^2y}{dx^2} = \frac{d}{dx} \left(\frac{dy}{dx} \right)$$

and so on.



Differentiability

For a function to be **differentiable** at a point x_A , it must be:

- **Defined** at x_A
- **Continuous** at x_A
- **Smooth** at x_A
- **Non-vertical** at x_A

Select rules of differentiation

	Function f	Lagrange	Leibniz
Constant	$f(x) = c$	$f'(x) = 0$	$\frac{df}{dx} = 0$
Power	$f(x) = x^r$ with $r \neq 0$	$f'(x) = rx^{r-1}$	$\frac{df}{dx} = rx^{r-1}$
Sum	$f(x) = g(x) + h(x)$	$f'(x) = g'(x) + h'(x)$	$\frac{df}{dx} = \frac{dg}{dx} + \frac{dh}{dx}$
Exponential	$f(x) = e^x$	$f'(x) = e^x$	$\frac{df}{dx} = e^x$
Chain Rule	$f(x) = g(h(x))$	$f'(x) = g'(h(x))h'(x)$	$\frac{df}{dx} = \frac{dg}{dh} \frac{dh}{dx}$

Chain rule example

1. Find $\frac{df}{dx}$ for $f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$

2. Now, let, $y = \sigma(x_1)$, where $x_1 = wx$. What is $\frac{dy}{dx}$?

Partial derivatives

For a scalar valued function $y = f(x_1, x_2)$, there are two partial derivatives:

$$\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}$$

These are computed by holding the "other" variable(s) **constant**. For example, if $y = 2x_1 + x_2 + x_1x_2$, then:

$$\frac{\partial y}{\partial x_1} = 2 + x_2, \frac{\partial y}{\partial x_2} = 1 + x_1$$

A brief introduction to vector calculus

Putting together partial derivatives with vectors and matrices we get:

Scalar-valued $f(\mathbf{x})$:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Vector-valued $\mathbf{f}(\mathbf{x})$:

$$\mathbf{J}_{\mathbf{f}} = \begin{bmatrix} \nabla^T f_1 \\ \nabla^T f_2 \\ \vdots \\ \nabla^T f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Most of the time we'll just be working with the gradient

Statistics: Notation

- A **random variable** $\mathbf{x} \sim P$ is a variable that can take on random variables according to some probability distribution P
- \mathbf{x} may take on **discrete** (e.g. dice rolls) or **continuous** (e.g. age) values
- X or \mathbf{x} for the random variable and x or x_i for a specific value
- $P(\mathbf{x})$ for a discrete distribution and $p(\mathbf{x})$ for continuous
- $\mathbf{x}_P \equiv \mathbf{x} \sim P$ and $\mathbf{x}_p \equiv \mathbf{x} \sim p$

Some textbooks/papers/websites use different notation!

Discrete random variables

- A discrete **probability mass function** describes the probability of x taking on a specific value
- Example: for a balanced 6-sided die, $P(x = 1) = \frac{1}{6}$
- You can add together probabilities, e.g. $P(x \leq 3) = \sum_{i=1}^3 P(x = i)$
- $\sum_x P(x) = 1$ and $P(x_i) \geq 0$ for any valid distribution

Continuous random variables

- A continuous **probability density function** gives the probability of being in some tiny interval δx given by $p(x)\delta x$
- Example: the **uniform distribution**, $p(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- $p(x = x_i) = 0$ for any specific value x_i
- Need to integrate to get a concrete value, e.g. $p(x \leq a) = \int_a^b p(x)dx$
- $\int_{-\infty}^{\infty} p(x)dx = 1$ and $\int_a^b p(x)dx \geq 0$ for any valid distribution

Expectation and variance

- The **expectation** or **expected value** is its average value $\mathbb{E}[\mathbf{x}]$

- $\mathbb{E}[\mathbf{x}_P] = \sum_x xP(\mathbf{x})$ and $\mathbb{E}[\mathbf{x}_p] = \int_{-\infty}^{\infty} xp(x)dx$

- More generally, for any function $f(\mathbf{x})$:

$$\mathbb{E}[f(\mathbf{x})] = \sum_x f(x)P(\mathbf{x}) \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)p(x)dx$$

- The **variance** describes how much the values vary from their mean:

$$\text{Var}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])^2]$$

Multiple random variables

- **Joint probability** $P(\mathbf{x}, y)$ is the probability of \mathbf{x} and y occurring together
- **Conditional probability** $P(\mathbf{x} = x \mid y = y)$ is the probability that \mathbf{x} takes on value x given that $y = y$ has already happened
- In general,
$$P(\mathbf{x} = x \mid y = y) = \frac{P(\mathbf{x} = x, y = y)}{P(y = y)}$$
- For **independent** variables, $P(\mathbf{x} = x \mid y = y) = P(\mathbf{x} = x)$

Covariance

- The **covariance** between $f(\mathbf{x})$ and $g(\mathbf{y})$ gives a sense of how linearly related they are and how much they vary together:

$$\text{Cov}(f(\mathbf{x}), g(\mathbf{y})) = \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])(g(\mathbf{y}) - \mathbb{E}[g(\mathbf{y})])]$$

- Related to correlation as $\text{Corr}(f(\mathbf{x}), g(\mathbf{y})) = \frac{\text{Cov}(f(\mathbf{x}), g(\mathbf{y}))}{\sqrt{\text{Var}(f(\mathbf{x}))\text{Var}(g(\mathbf{y}))}}$
- The **covariance matrix** of a random vector \mathbf{x} is a square matrix where the (i, j) element is the covariance between x_i and x_j
- The diagonal of the covariance matrix gives $\text{Var}(x_i)$

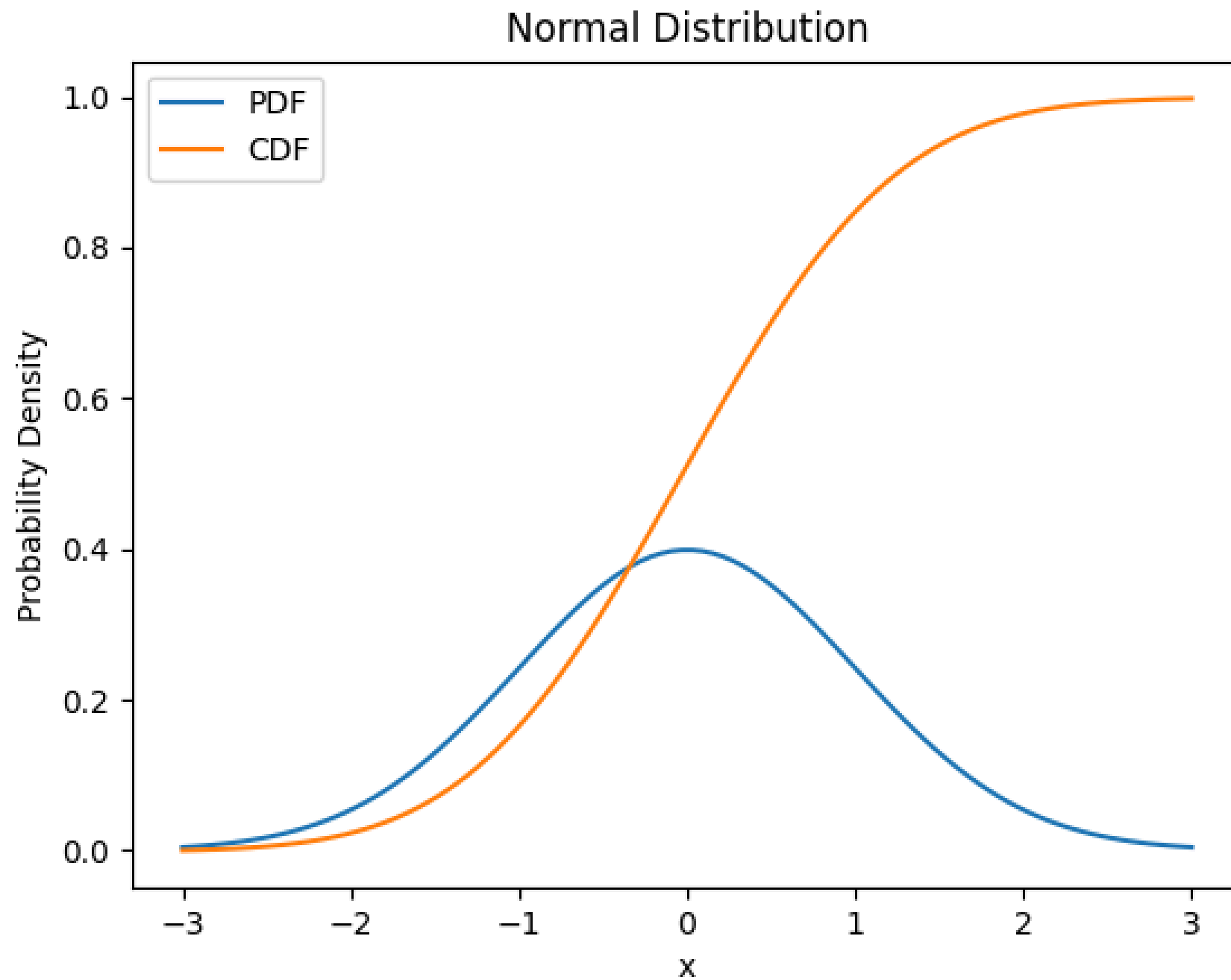
The Normal distribution

$$N(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Good "default choice" for two reasons:

- The **central limit theorem** shows that the sum of many (> 30 ish) independent random variables is normally distributed
- Has the most **uncertainty** of any distribution with the same variance

We can't easily integrate $N(x)$, so numerical approximations are used



Coming up next

- Training (regression) models
 - Linear regression
 - Gradient descent
- References and suggested reading:
 - [Scikit-learn book](#):
 - Chapter 4: Training Models
 - [Deep Learning Book](#)
 - Section 5.1.4: Linear Regression