Math primer for the Neural Networks course

Roberto Esposito

📧 roberto.esposito@unito.it

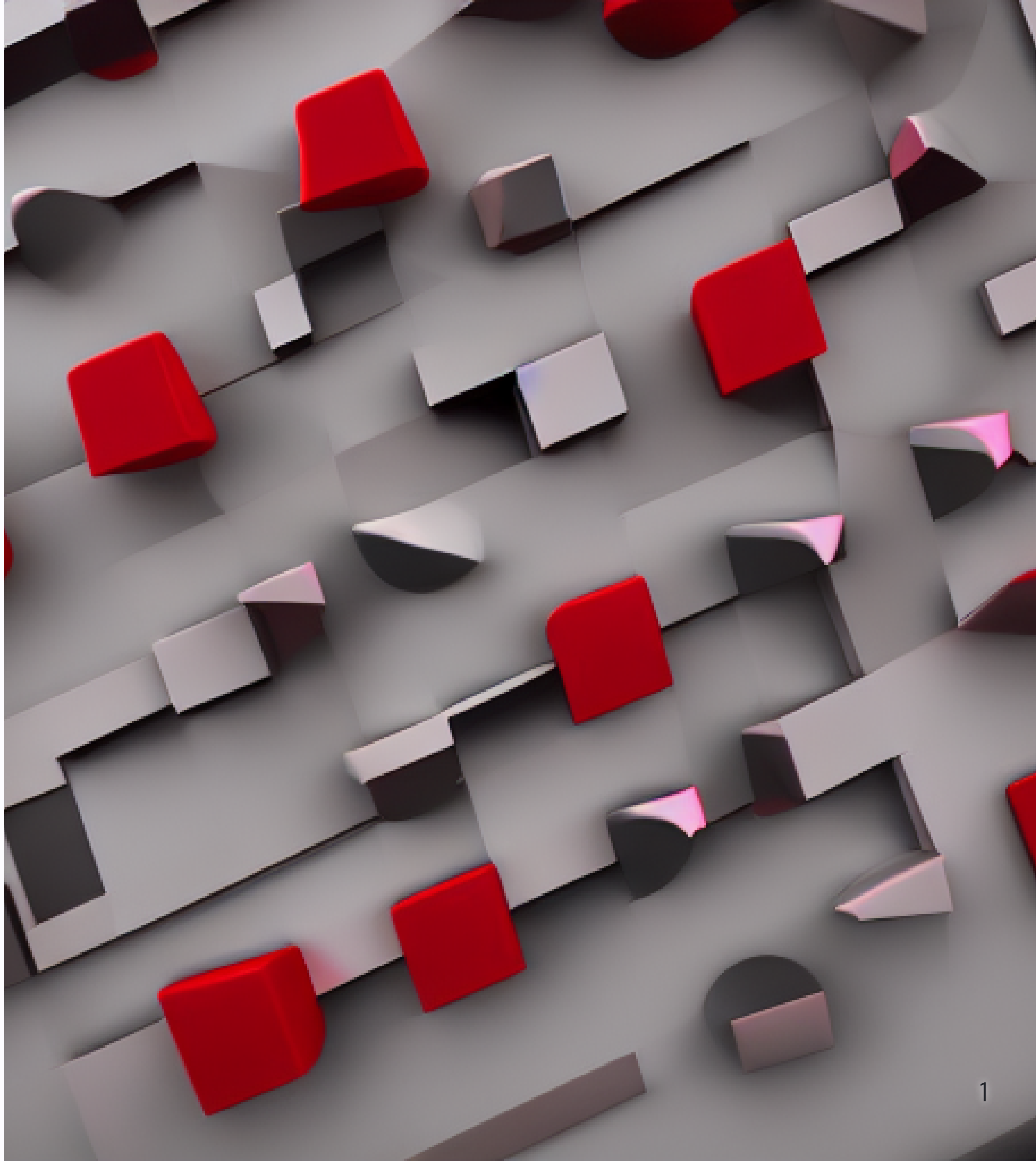


Image dreamed by stable diffusion
Prompt: "abstract, geometric shapes, 3d render, dark dusty colors, dark red, light gray, strokes"

[Calculus knowledge poll]

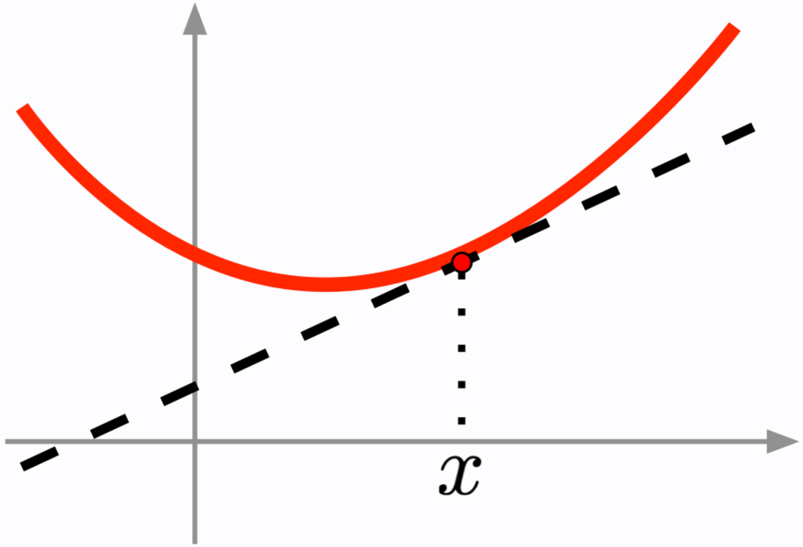
Calculus

Derivatives

Suppose we have a function $y = f(x)$, where both x and y are real numbers.

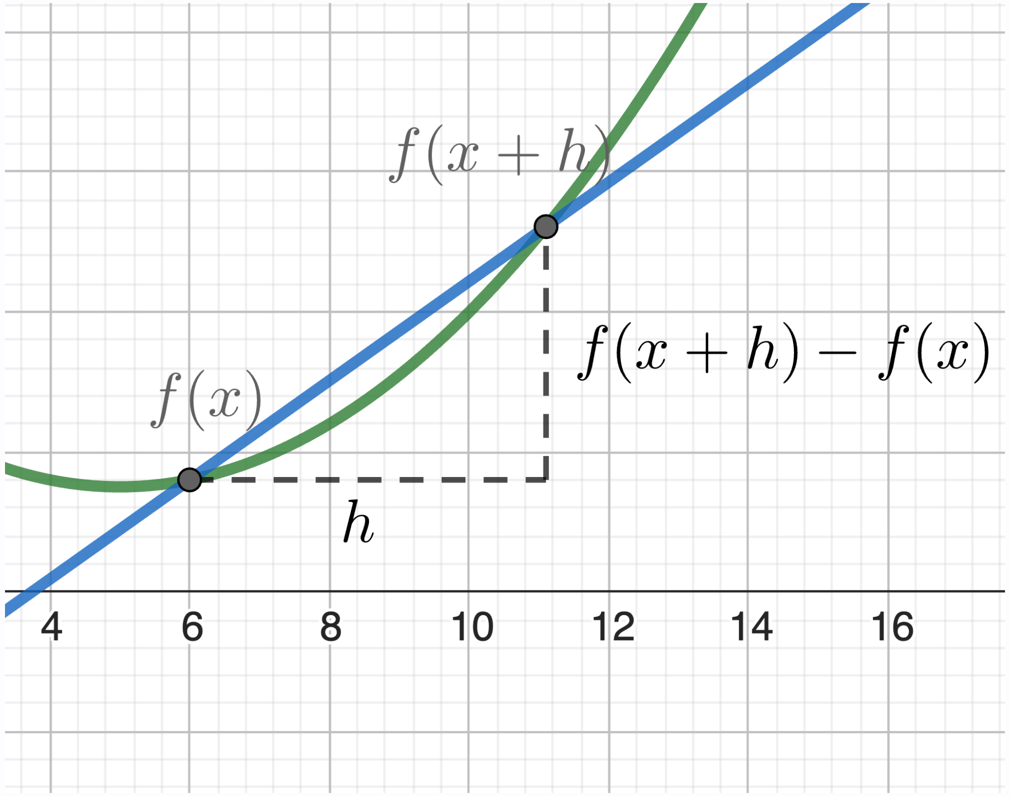
The derivative of f at point x, denoted $f'(x)$ or $\frac{df}{dx}(x)$ is the slope of the tangent line to f at point x.

In other words, it specifies how to scale a small change in the input in order to obtain the corresponding change in the output: $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.



Secant line

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



https://www.geogebra.org/calculator/duqnvsxm

Properties of derivatives

Property		
Linearity	$(\alpha f(x) + \beta g(x))'$	$\alpha f'(x) + \beta g'(x)$
Chain rule	$(f(g(x)))'$	$f'(g(x))g'(x)$
Product rule	$(g(x)h(x))'$	$g'(x)h(x) + g(x)h'(x)$
Quotient Rule	$\left(\frac{f(x)}{g(x)}\right)'$	$\frac{f(x)'g(x) - f(x)g'(x)}{(g(x))^2}$
Power rule	$(x^r)'$	rx^{r-1}

...

Examples

- Power rule: $(x^4)'$
- Linearity: $(3\sin(x) + x^2)'$
- Chain rule: $(\sin(x^2))'$
- Product rule: $(x^2 x^3)'$
- Quotient rule: $(\frac{x^5}{x^2})'$

Examples

- Power rule: $(x^4)' = 4x^3$
- Linearity: $(3\sin(x) + x^2)' = 3(\sin(x))' + (x^2)' = 3\cos(x) + 2x$
- Chain rule: $(\sin(x^2))' = \cos(x^2)(x^2)' = 2\cos(x^2)x$
- Product rule: $(x^2x^3)' = 2x(x^3) + x^2(3x^2) = 5x^4 = (x^5)'$
- Quotient rule: $(\frac{x^5}{x^2})' = \frac{5x^4(x^2)-x^5(2x)}{x^4} = \frac{3x^6}{x^4} = 3x^2 = (x^3)'$

Integrals

Consider $f : \mathbb{R} \to \mathbb{R}$ and an interval $[a, b]$ on the real line.

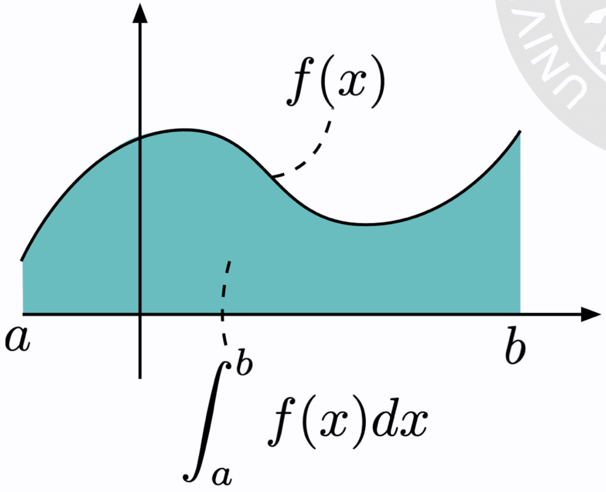
The **integral of f between a and b is the area under f in the given region** (when the function is below 0, the area contributes negatively).

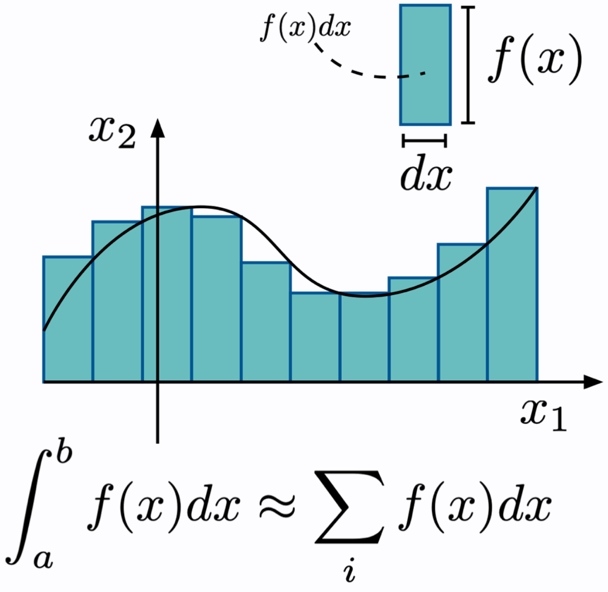
If f admits an antiderivative F, i.e., if it exists F such that $F'(x) = f(x)$, then:

$$\int f(x)dx = F(x) + C$$

and

$$\int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a)$$



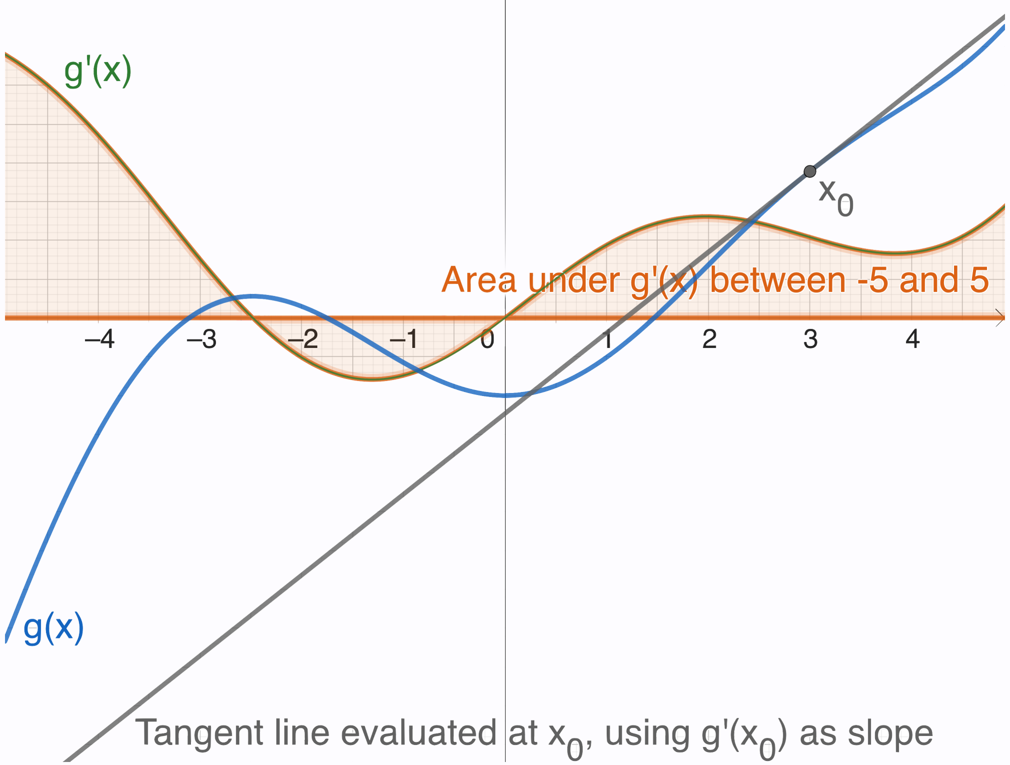


Properties of Integrals

Property		
linearity	$\int \alpha f(x) + \beta g(x) dx$	$\alpha \int f(x) dx + \beta \int g(x) dx$
constant rule	$\int k dx$	$kx + C$
power rule	$\int x^n dx$	$\frac{x^{n+1}}{n+1} + C, n \neq -1$
log rule	$\int \frac{1}{x} dx$	$\ln(\lvert x \rvert) + C$
exponential rule	$\int a^{kx} dx$	$\frac{a^{kx}}{k \ln a} + C, a > 0, a \neq 1$
Sin rule	$\int \sin(x) dx$	$-\cos(x) + C$
Cosin rule	$\int \cos(x) dx$	$\sin(x) + C$

…

Integration/derivatives



$$g(x) = 0.1\frac{x^3}{3} - \cos(x)$$

$$g'(x) = 0.1x^2 + \sin(x)$$

$$\int g'(x)dx = g(x) + C$$

https://www.geogebra.org/calculator/wd7uqns6

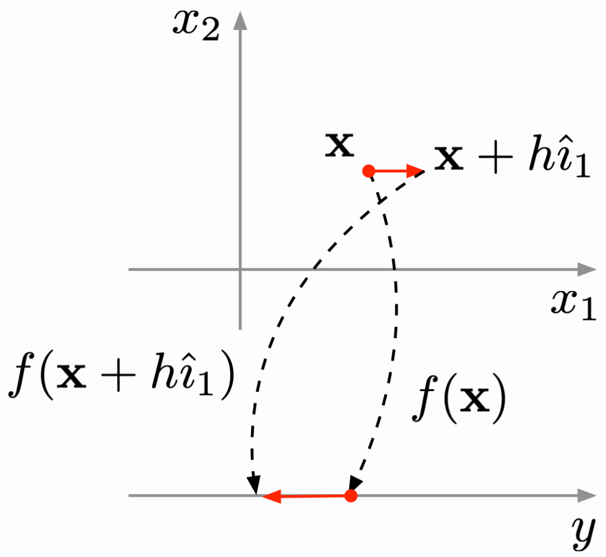
Partial derivatives and Gradients

Assume you have a function $y = f(x_1, \ldots, x_n) = f(\mathbf{x})$, where $y \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^n$.

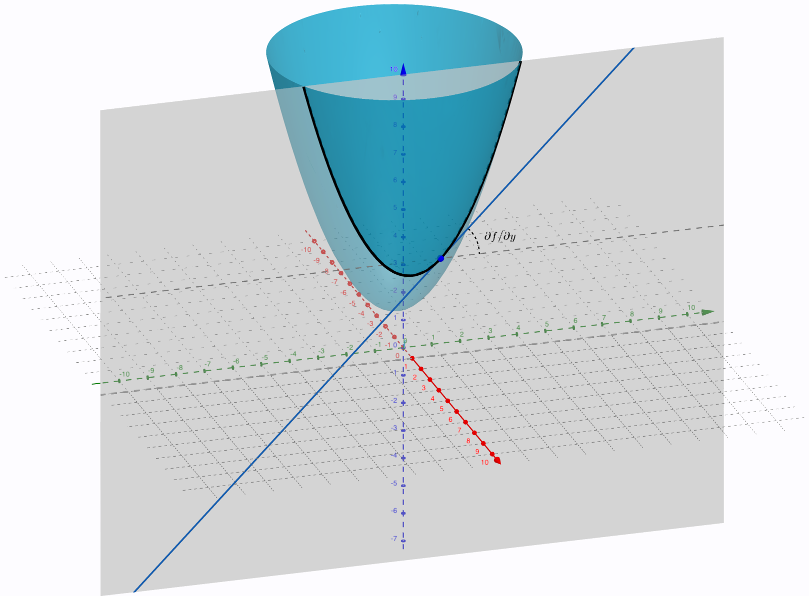
The partial derivative $\frac{\partial}{\partial x_j} f(\mathbf{x})$ measures how f changes as only the x_j variable increases at point $\mathbf{x}$:

$$\frac{\partial}{\partial x_j} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\hat{\imath}_j) - f(\mathbf{x})}{h}$$

$$= \lim_{h \to 0} \frac{f(x_1, \ldots, x_j + h, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}$$



Partial derivative



In this example:

- gray plane: $x = 1$
- black parabola: intersection between $x = 1$ and f
- blu line: tangent to f on the plane $x = 1$ evaluated at $y = 1$.

https://www.geogebra.org/m/cpswyasx

The **gradient** of f, denoted $\nabla_{\mathbf{x}} f$ (or simply ∇f), is the vector collecting all partial derivatives:

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right]^\top$$

Chain rule for multivariate calculus

Assume $z = f(x, y)$ and let x, y depend on an additional variable t, then:

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}.$$

More in general for $f : \mathbb{R}^n \to \mathbb{R}$, when $x_1 \dots x_n$ depend on a variable t:

$$\frac{df}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{dx_i}{dt}$$

Example

Let:

- $(x, y) = (t^2, t)$, i.e., $x(t) = t^2$ and $y(t) = t$.
- $z = f(x, y) = x^2 y^2$.

Evaluate the derivative of z w.r.t. t.

Example

Let:

- $(x, y) = (t^2, t)$, i.e., $x(t) = t^2$ and $y(t) = t$.
- $z = f(x, y) = x^2y^2$.

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} = 2xy^2 \cdot 2t + 2x^2y \cdot 1 = 4t^5 + 2t^5 = 6t^5$$

Note

By noticing that $f(x, y) = (x(t))^2 \cdot (y(t))^2 = (t^2)^2 \cdot (t)^2 = t^4t^2 = t^6$.

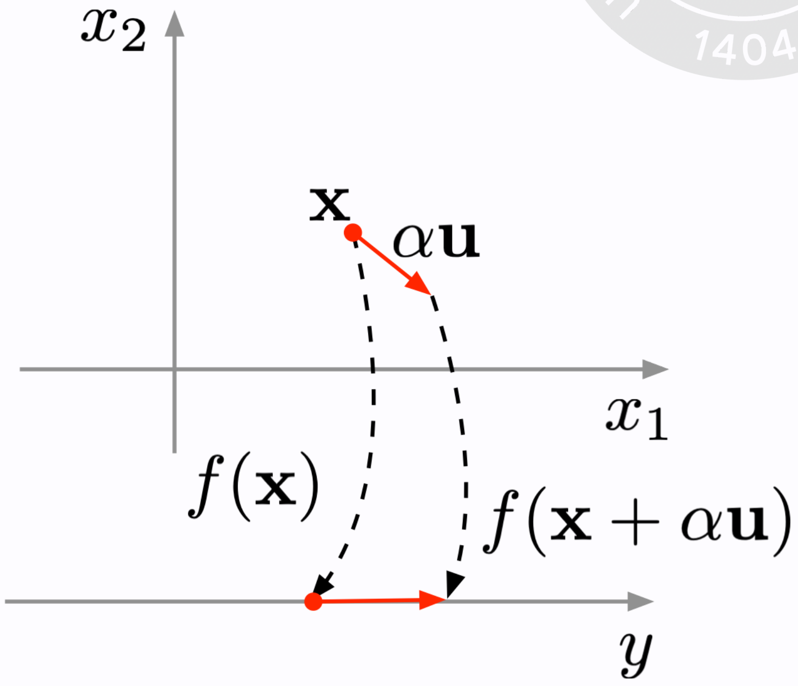
The same results could have been obtained simply by evaluating $\frac{d}{dt}t^6 = 6t^5$.

Directional derivatives

Assume $\mathbf{u}$ to be a unit vector. The directional derivative of f at $\mathbf{x}$ in $\mathbf{u}$ direction is the rate of change in the direction given by vector $\mathbf{u}$.

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}$$

In other words, the directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{u}$ is the derivative of $f(\mathbf{x} + \alpha\mathbf{u})$ w.r.t. α evaluated at $\alpha = 0$.

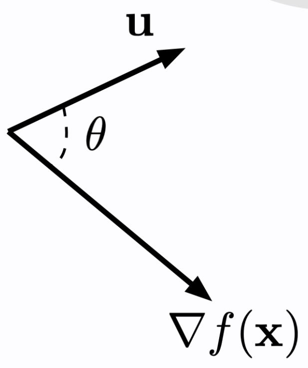


Using the chain rule, we can easily compute an expression for $D_{\mathbf{u}} f(\mathbf{x})$:

$$D_{\mathbf{u}} f(\mathbf{x}) = \frac{d}{d\alpha} f(\mathbf{x} + \alpha\mathbf{u})\Big|_{\alpha=0} = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x} + \alpha\mathbf{u})}{\partial x_i}\Big|_{\alpha=0} \frac{dx_i}{d\alpha} = \nabla f(\mathbf{x}) \cdot \mathbf{u} = \mathbf{u}^\top \nabla f(\mathbf{x})$$

Let assume we want to find the direction in which the function increases the most, i.e., we want to find $\mathbf{u}$ such that $\nabla_{\mathbf{u}} f$ is largest. We want to solve:

$$\max_{\mathbf{u},\mathbf{u}^\top\mathbf{u}=1} D_{\mathbf{u}} f(\mathbf{x}) = \max_{\mathbf{u},\mathbf{u}^\top\mathbf{u}=1} \mathbf{u}^\top \nabla f(\mathbf{x}) = \max_{\mathbf{u},\mathbf{u}^\top\mathbf{u}=1} |\mathbf{u}||\nabla f(\mathbf{x})| \cos(\theta)$$



Since $|\mathbf{u}| = 1$ and since $\nabla f(\mathbf{x})$ does not depend on $\mathbf{u}$, we are left with finding $\mathbf{u}$ that maximizes $\cos\theta$. Which implies that the maximum is attained when $\mathbf{u}$ is in the same direction as $\nabla f(\mathbf{x})$.

> ***Important:*** *the* ***gradient*** *points in the direction in which f increases the most.*

Jacobian Matrix

The **Jacobian** of a multi-valued, multi-variable function:

$$f : \mathbb{R}^n \to \mathbb{R}^m, \quad f(\mathbf{x}) = [f(\mathbf{x})_1, \dots, f(\mathbf{x})_m]^\top$$

is the matrix $\mathbf{J} \in \mathbb{R}^{m \times n}$ containing the partial derivatives of all $f(\mathbf{x})_i, (1 \le i \le m)$ for all variables $x_j, (1 \le j \le n)$:

$$\mathbf{J}_{i,j} = \frac{\partial}{\partial x_j} f(\mathbf{x})_i$$

or, equivalently, the Jacobian is the matrix containing $\nabla[f(\mathbf{x})_i]$ in row i:

$$\mathbf{J} = \left[\nabla[f(\mathbf{x})_i]^\top\right]_{i=1}^{m}$$

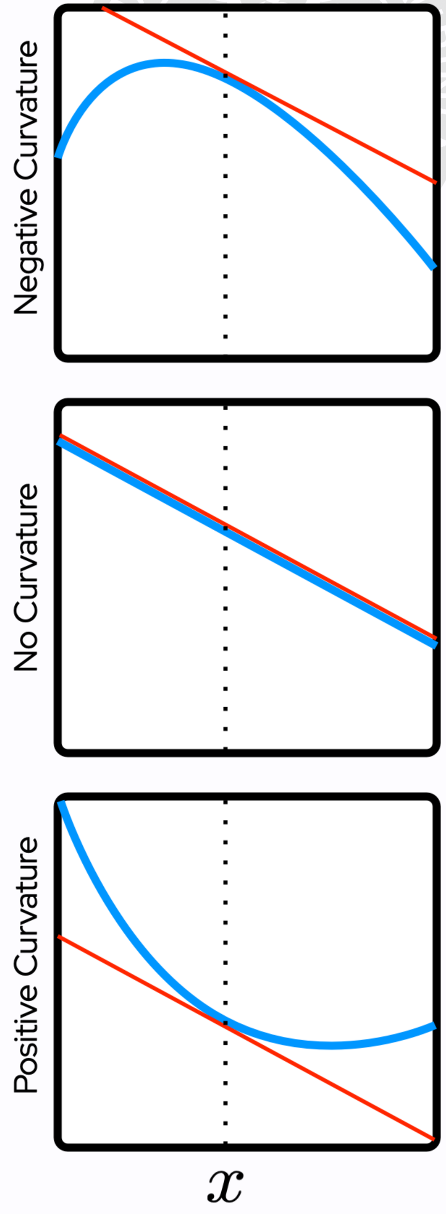
Second derivatives

The **second derivative** is a derivative of a derivative. For instance, if $f : \mathbb{R}^n \to \mathbb{R}$, we can compute n^2 second derivatives:

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

The second derivative tells us how the first derivative will change as we vary the input. We can think of the second derivative as **measuring curvature**.

The matrix $H(f)$ containing all these partial derivatives is called the **Hessian** of the function f. **Note:** $H(f) = \mathbf{J}(\nabla f)$



Properties of the Hessian matrix

Anywhere that the second partial derivatives are continuous, the differential operators are commutative, i.e. their order can be swapped:

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) = \frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x})$$

This implies that **the Hessian is symmetric** at such points.

When $\nabla f(\mathbf{x}_0) = \mathbf{0}$ **the Hessian helps us to understand if we are on a minimum** (true if the Hessian is positive definite, i.e., all eignvalues are > 0), **a maximum** (true if the Hessian is negative definite, i.e., all eigenvalues are < 0). If the Hessian is neither positive nor negative definite (we have at least one zero eigenvalue):

- **we are on a saddle point** if there is at least 1 positive eigenvalue and 1 negative eigenvalue;
- **the test is inconclusive** otherwise.

[Calculus knowledge poll]