Math Summary

Tags: #fleeting #literature #permanent

Description:

Theme:

github : https://github.com/viniciusrondon/PO-245---Neural-Net-and-LLM---Prof.-Mauri.git

ID: 20250808194614

Brief Description:

This summary reviews the fundamental concepts of Artificial Neural Networks (ANNs) with a focus on the Backpropagation algorithm. Special attention is given to its mathematical foundation in `differential calculus`, which supports the computation of gradients used to optimize network parameters.

`

Derivatives

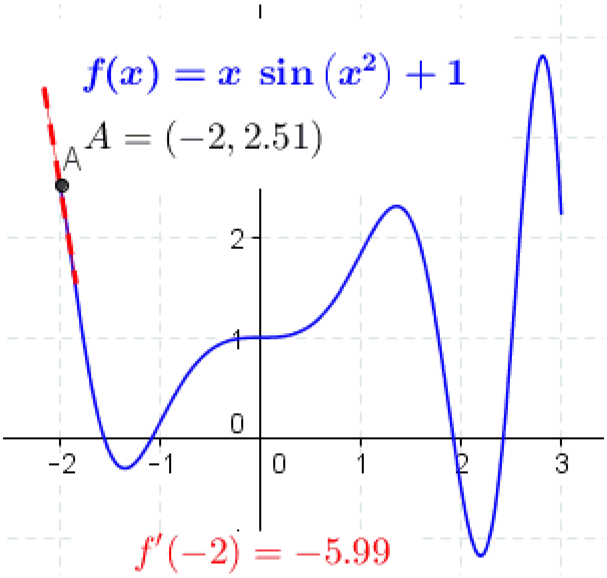


Figure 01: Derivative at Different points. Wikipedia

The derivative is a fundamental concept in calculus that quantifies how sensitive a function's output is to changes in its input. It is defined as the slope of the tangent line to a continuous function at a given point, representing the instantaneous rate of change of the function at that point.

This concept can be demonstrated by determining the tangent line that passes through a point P, as illustrated in figure 02:

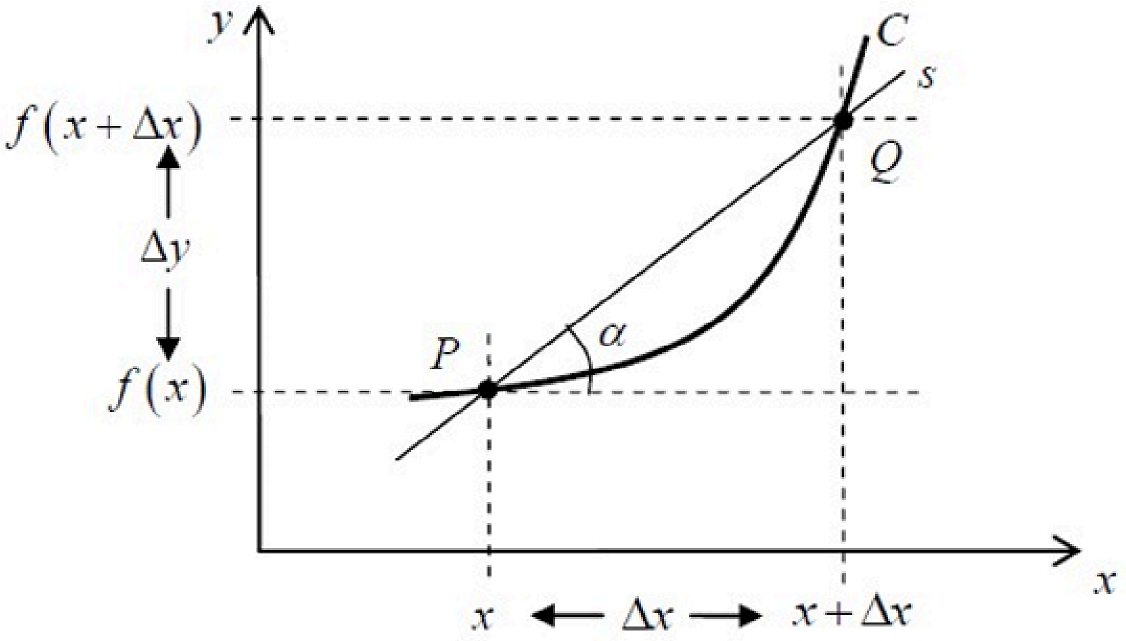


Figure 02: Point P represented on the continuous curve C

The angular coefficient α of the secant line S is given by:

$$\tan(\alpha) = \frac{f(x+\Delta x) - f(x)}{(x+\Delta x) - x}, \tag{01}$$

which simplifies to:

$$\tan(\alpha) = \frac{f(x+\Delta x) - f(x)}{\Delta x}. \tag{02}$$

In order for the secant line to become the tangent line at point P , it is necessary to make Δx approach zero. This ensures that point Q on the curve moves increasingly closer to P, which inplies that $(x+\Delta x) - x$ will be near by x, as illustrated in the figure 03.

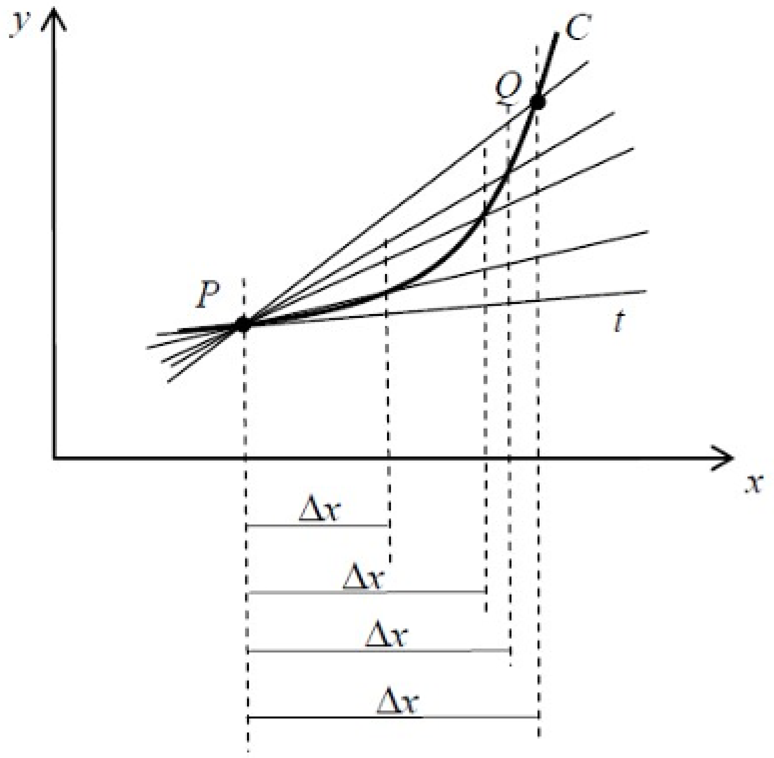


Figure 03: Dinamic representation the point Q approaching point P

Thus, ithe derivative of $f(x)$ at x can be obtained using the limit definition: :

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}. \tag{03}$$

> **Definition (Derivative).**
> Let f be a function defined on an open interval containing x. The derivative of f at x, denoted $f'(x)$, is defined as:
>
> $$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$
>
> provided this limit exists. It represents the instantaneous rate of change of f with respect to xx and is the slope of the tangent line to the graph of f at the point $(x, f(x))$.
>
> **Source:** Stewart, J. *Calculus: Early Transcendentals*, 9th ed., Cengage Learning, 2021, p. 161.

Parcial Derivatives

The concept of partial derivatives arises from the need to efficiently compute derivatives of functions with multiple variables. In such cases, the derivative is taken with respect to one variable at a time, while keeping the other variables constant.

The partial derivative of a function $f(x, y, \ldots)$ with respect to the variable x is variously denoted by:

f_{x},&nbsp;f'_{x},&nbsp;\delta_{x}f,&nbsp;&nbsp;D_{x}f,&nbsp;D_{1}f,\frac{\delta}{\delta x}f, \textt

Sometimes, for $z = f(x, y, \ldots)$, the partial derivative of z with respect to x is denoted as $\frac{\delta z}{\delta x}$.

Math intuiton by example, considering some function $f(x, y) = x^3 + x^2y - xy^2$ to determine function f in relation of x, is given by

$$\begin{aligned} f(x+h, y) = (x+h)^3 + (x+h)^2y - (x+h)y^2 = \\ x^3 + 3x^2h + 3xh^2 + h^3 + (x^2 + 2xh + h^2)y - y^2x - y^2h = \end{aligned} \tag{05}$$

$$\begin{aligned} f(x+h, y) - f(x, y) = x^3 + 3x^2h + 3xh^2 + h^3 + (x^2 + 2xh + h^2)y - y^2x - y^2h - x^3 - x^2y + a \\ 3x^2h + 3xh^2 + h^3 + 2xhy + h^2y - \end{aligned}$$

So,

$$\frac{\delta f}{\delta x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h} \tag{08}$$

$$\begin{aligned} \frac{\delta f}{\delta x} = \lim_{h \to 0} \frac{3x^2h + 3xh^2 + h^3 + 2xhy + h^2y - hy^2}{h} = \\ \lim_{h \to 0} 3x^2 + 3xh + h^2 + 2xy + hy - y^2 = \\ \frac{\delta f}{\delta x} = 3x^2 + 2xy - y^2 \end{aligned} \tag{09}$$

Rules for Computing Partial Derivatives

Rule	Statement	Notes / Condition
Linearity	$\frac{\partial}{\partial x}[af(x, y) + bg(x, y)] = a\frac{\partial f}{\partial x} + b\frac{\partial g}{\partial x}$	a, b are constants
Power Rule	$\frac{\partial}{\partial x}[x^n] = nx^{n-1}$	n real cons
Constant Rule	$\frac{\partial}{\partial x}[c] = 0$	c constant
Sum Rule	$\frac{\partial}{\partial x}[f(x, y) + g(x, y)] = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$	Works for subtractior well
Product Rule	$\frac{\partial}{\partial x}[u(x, y) \cdot v(x, y)] = u_x \cdot v + u \cdot v_x$	$u_x = \frac{\partial u}{\partial x}$
Quotient Rule	$\frac{\partial}{\partial x}\left[\frac{u}{v}\right] = \frac{u_x v - u v_x}{v^2}$	$v \neq 0$
Chain Rule	$Ifz = f(u, v), u = g(x, y), v = h(x, y), \mathbf{then}\ \frac{\partial z}{\partial x} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \cdot \frac{\partial v}{\partial x}$	Fundamen for backpropa

Jacobian

Given a function that maps **multiple variables to multiple outputs**:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix} \tag{10}$$

The **Jacobian matrix** is the matrix of **all first-order partial derivatives**:

$$J_{\mathbf{f}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{11}$$

Wherein,

- **Rows** correspond to functions f_i
- **Columns** correspond to variables x_j

Some real aplications

- In **backpropagation**, the Jacobian tells us how **a change in each input variable affects each output variable**.
- It is crucial when applying the **multivariable chain rule** in vector form.

Practical Example 1

Let's take:

$$\mathbf{f}(x, y) = \begin{bmatrix} f_1(x, y) = x^2 + y \\ f_2(x, y) = \sin(xy) \end{bmatrix} \tag{12}$$

The Jacobian is:

$$J_{\mathbf{f}}(x, y) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & 1 \\ y\cos(xy) & x\cos(xy) \end{bmatrix}$$

At $(x, y) = (1, 2)$:

$$J_{\mathbf{f}}(1, 2) = \begin{bmatrix} 2 & 1 \\ 2\cos(2) & 1\cos(2) \end{bmatrix}$$

Practical Example 2 – ANN Layer Transformation

Considering a **fully connected layer** in a neural network:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{13}$$

where:

- $\mathbf{x} \in \mathbb{R}^n$ (input vector)
- $\mathbf{y} \in \mathbb{R}^m$ (output vector)
- $\mathbf{W} \in \mathbb{R}^{m \times n}$ (weights)
- $\mathbf{b} \in \mathbb{R}^m$ (biases)

Since:

$$y_i = \sum_{j=1}^{n} W_{ij} x_j + b_i \tag{13}$$

we have:

$$\frac{\partial y_i}{\partial x_j} = W_{ij} \tag{14}$$

Thus, the **Jacobian** of this transformation is simply:

$$\mathbf{y}(\mathbf{x}) = \mathbf{W} \tag{15}$$

This is why **matrix multiplication** naturally appears in backpropagation — it *is* the Jacobian of a linear transformation.

Directional Derivative

A **directional derivative**:

- Measures the **slope** (rate of change) of a function’s surface **at a specific point**.
- The slope is taken **along a given direction**, specified by a vector (usually a unit vector).
- It tells *how fast* the function increases or decreases if you move **from that fixed point in that specific direction**.

> The directional derivative measures the slope of a function at a fixed point in a specified direction, quantifying how the function changes as we move from that point along that direction.

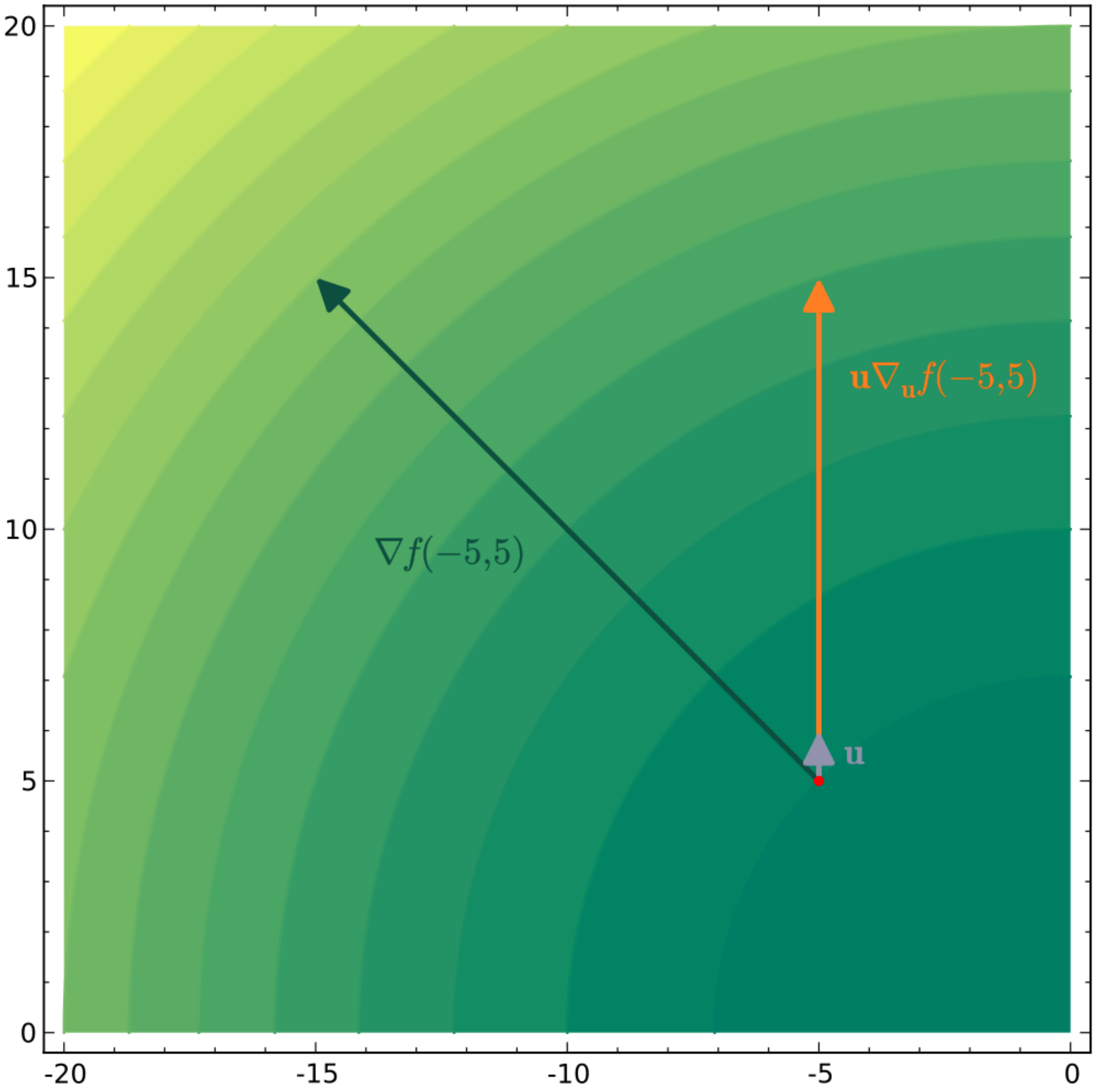


Figure 04: A contour plot of , showing the gradient vector in black, and the unit vector scaled by the directional derivative in the direction of in orange. The gradient vector is longer because the gradient points in the direction of greatest rate of increase of a function. image from [wikipedia] (https://en.wikipedia.org/wiki/Partial_derivative#/media/File:Directional_derivative_contour_plot.svg)

The Figure 04 illustrates contour plot of $f(x, y) = x^2 + y^2$, showing the gradient vector in black, and the unit vector u scaled by the directional derivative in the direction of u in orange. The gradient vector is longer because the gradient points in the direction of greatest rate of increase of a function.

1. The Idea Behind the Directional Derivative

For a function of several variables $f(x, y)$, the **partial derivatives**

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y} \tag{16}$$

tells how f changes **if you move only along the x-axis or only along the y-axis**.

But what if moving in **any arbitrary direction**, not just along the coordinate axes?

That's where the directional derivative comes in — it measures the rate of change of f in the direction of a given vector.

2. Formal Definition

Let $\mathbb{R}^n \to \mathbb{R}$ be differentiable, and let
$\mathbf{u}$ be a **unit vector** (direction of motion).

The **directional derivative** of f at $\mathbf{p}$ in the direction $\mathbf{u}$ is:

$$D_{\mathbf{u}} f(\mathbf{p}) = \lim_{h \to 0} \frac{f(\mathbf{p} + h\mathbf{u}) - f(\mathbf{p})}{h}. \tag{17}$$

3. Shortcut Using the Gradient

If f is differentiable, we can compute the directional derivative quickly using the **gradient**:

$$D_{\mathbf{u}} f(\mathbf{p}) = \nabla f(\mathbf{p}) \cdot \mathbf{u}. \tag{18}$$

Where:

- $\nabla f = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle$ is the gradient vector.
- $\cdot$ is the dot product.
- $\mathbf{u}$ is the **unit vector** in the desired direction.

This formula works because the gradient points in the **direction of steepest increase**, and the dot product extracts the component of that increase in the direction of $\mathbf{u}$.

4. Applying to Some Example

We have:

$$f(x, y) = 2x^2 - 3y^2 + 3x + 2y. \tag{19}$$

Step 1 – Find the partial derivatives:

$$\frac{\partial f}{\partial x} = 4x + 3, \quad \frac{\partial f}{\partial y} = -6y + 2. \tag{20}$$

Step 2 – Gradient vector:

$$\nabla f(x, y) = \langle 4x + 3, \ -6y + 2 \rangle. \tag{21}$$

Step 3 – Direction vector:

$\theta = \frac{\pi}{3}$ means:

$$\mathbf{u} = \langle \cos(\pi/3),\ \sin(\pi/3) \rangle = \left\langle \frac{1}{2},\ \frac{\sqrt{3}}{2} \right\rangle. \tag{22}$$

Step 4 – Directional derivative at a point:

If we pick $P = (x_0, y_0)$, then:

$$D_{\mathbf{u}} f(P) = \nabla f(P) \cdot \mathbf{u}. \tag{23}$$

For example, if $P = (1, 2)$:

$$\begin{gathered} \nabla f(1,2) = \langle 4(1) + 3,\ -6(2) + 2 \rangle = \langle 7, -10 \rangle. \\ D_{\mathbf{u}} f(1,2) = \langle 7, -10 \rangle \cdot \left\langle \frac{1}{2}, \frac{\sqrt{3}}{2} \right\rangle = \frac{7}{2} - 5\sqrt{3}. \end{gathered} \tag{24}$$

This is the **rate of change of** f when moving from P **in the direction of** $\theta = \pi/3$.

Gradient

The **gradient** is closely connected to directional derivatives, so once understanding that, the gradient will feel natural.

1. What the Gradient Is

For a function $f(x, y)$ (or more generally $\mathbb{R}^n \to \mathbb{R}$), the **gradient** is a **vector** that collects all the first partial derivatives:

$$\nabla f(x, y) = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle \tag{25}$$

for two variables, and for n variables.

$$\nabla f(x_1, x_2, \ldots, x_n) = \left\langle \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right\rangle \tag{26}$$

2. What It Represents

- **Direction**:
  The gradient **points in the direction of the steepest increase** of the function's value.
- **Magnitude**:
  The length $||\nabla f||$ tells you **how steep** that increase is.

So, if you think of $f(x, y)$ as a surface, the gradient at a point is like an arrow in the horizontal plane pointing toward the direction where the surface rises fastest.

3. Link to Directional Derivatives

The directional derivative formula:

$$D_{\mathbf{u}}f(P) = \nabla f(P) \cdot \mathbf{u} \tag{28}$$

means that:

- If $\mathbf{u}$ points in the same direction as ∇f, the dot product is maximized → steepest slope.
- If $\mathbf{u}$ points opposite to ∇f, the slope is negative → steepest descent.
- If $\mathbf{u}$ is perpendicular to ∇f, the slope is zero → moving along a contour line.

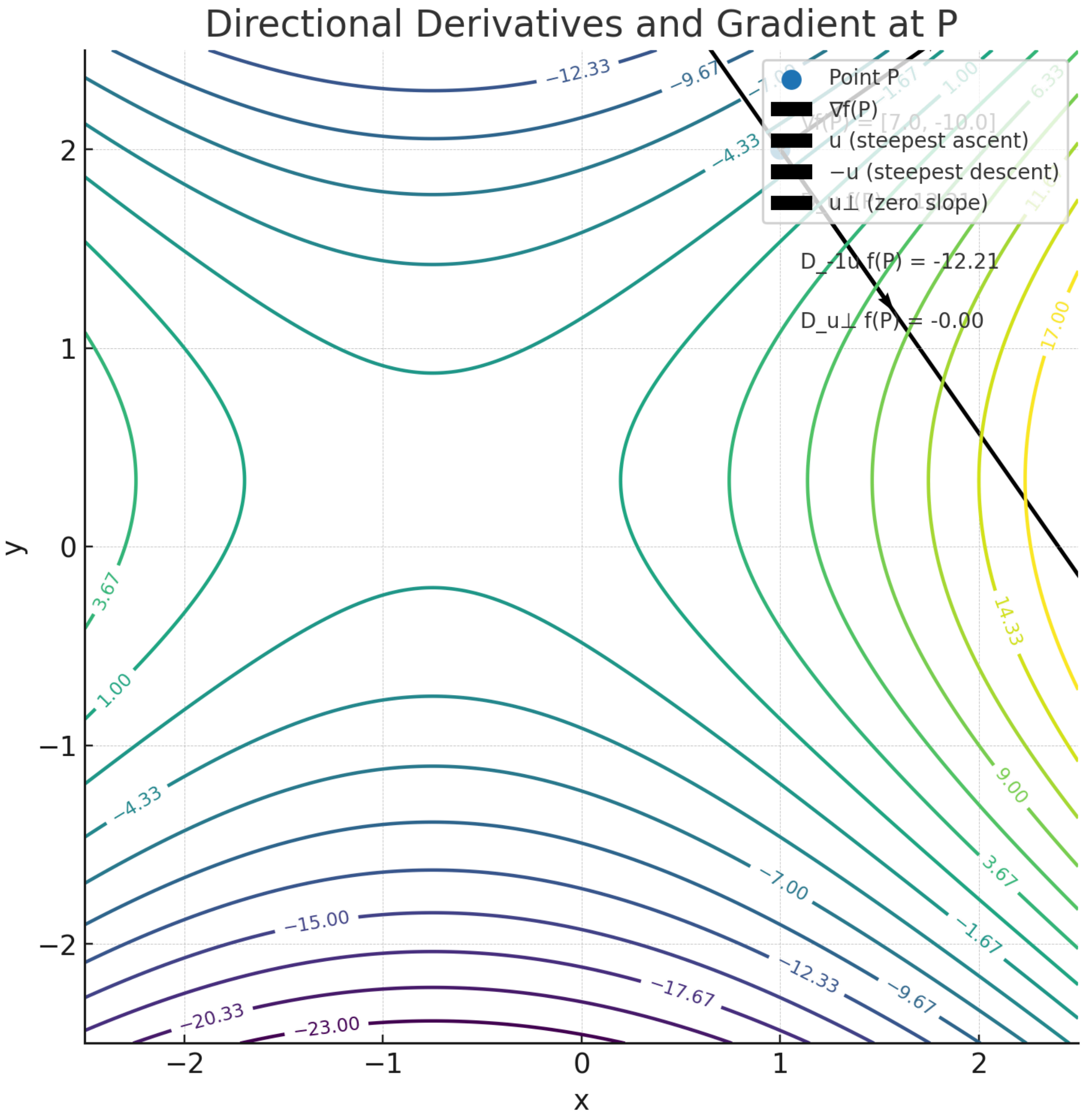


Figure 05: a contour plot about directional derivative and gradient at P

Figure 05 illustrates a contour plot showing $f(x, y) = 2x^2 - 3y^2 + 3x + 2y$ with point $P = (1, 2)$, the gradient $\nabla f(P)$, and three directions from P: along ∇f (steepest ascent), opposite to it (steepest descent), and perpendicular (zero slope).
The figure also annotates the directional derivative values.

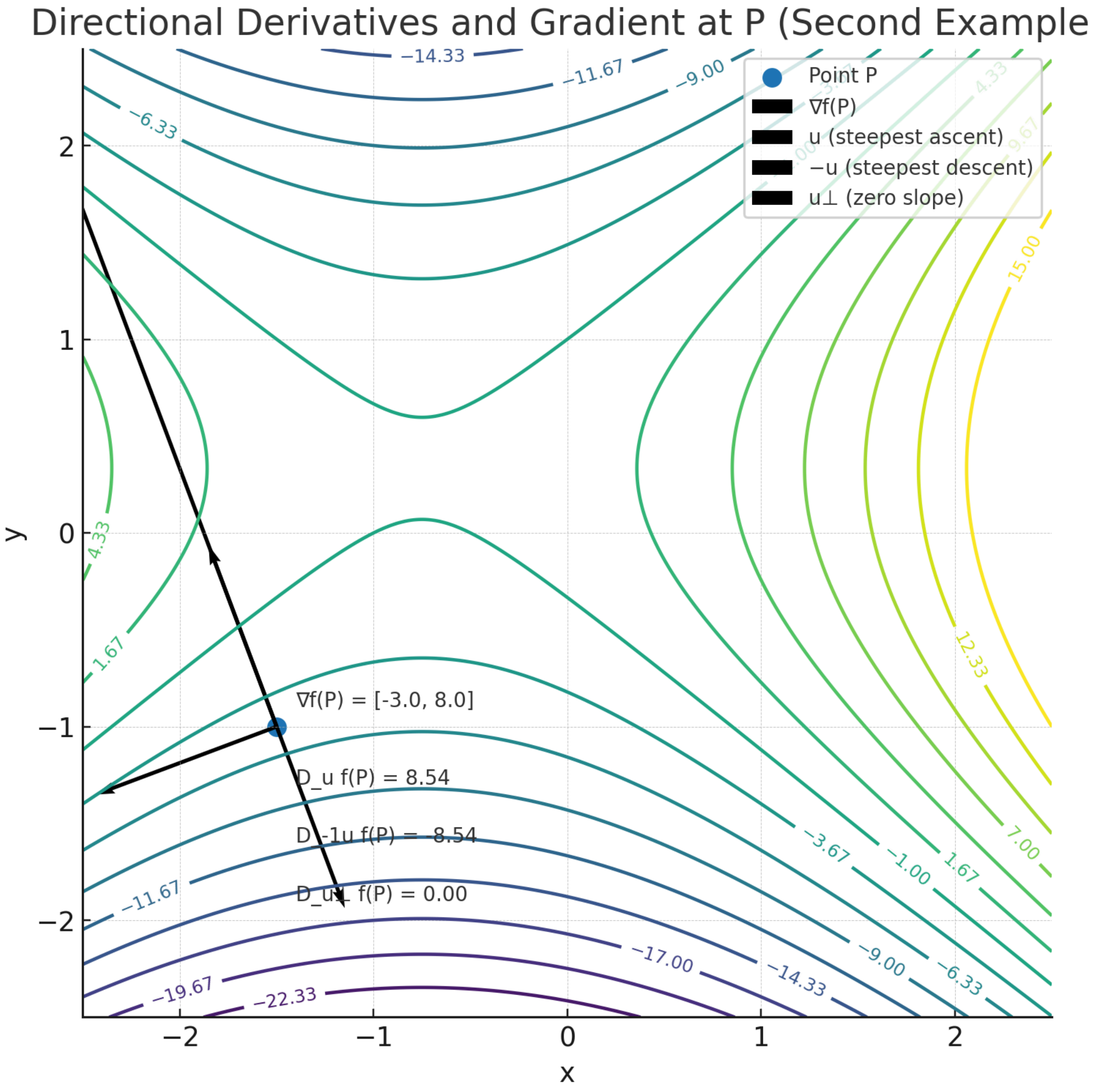


_Figure 06: the second version using the point $P = (-1.5, -1.0)$ with its corresponding gradient and directional derivative vectors

4. Practical Example

Let’s take your function again:

$$f(x, y) = 2x^2 - 3y^2 + 3x + 2y \tag{29}$$

Gradient:

$$\begin{gathered} \frac{\partial f}{\partial x} = 4x + 3, \\ \frac{\partial f}{\partial y} = -6y + 2 \end{gathered} \tag{30}$$

So:

$$\nabla f(x, y) = \langle 4x + 3, \; -6y + 2 \rangle \tag{31}$$

At $P = (1, 2)$:

$$\nabla f(1, 2) = \langle 7, \;\; -10 \rangle \tag{32}$$

This means:

- The steepest increase occurs if you move in direction $\langle 7, -10 \rangle$.
- The magnitude $\sqrt{7^2 + (-10)^2}$ tells you how steep that climb is.

5. Intuition

Think of hiking:

- **Gradient direction** = “Point your feet toward the steepest uphill.”
- **Gradient magnitude** = “How steep the hill is in that direction.”
- If you go exactly that way, you increase altitude fastest; if you go the opposite way, you descend fastest; if you go perpendicular, you stay at the same height.

Gradient descend

In the practical example it should be noticed that at $P = (1, 2)$:

$$\nabla f(1, 2) = \langle 7, \;\; -10 \rangle \tag{32}$$

This means:

1. **Steepest ascent/descent at $P = (1, 2)$.**

$$\nabla f(1, 2) = \langle 7, -10 \rangle. \tag{33}$$

- The **steepest increase** direction is **along the gradient**, i.e. the **unit** vector

$$\mathbf{u}_{\max} = \frac{\langle 7, -10 \rangle}{\|\langle 7, -10 \rangle\|}. \tag{34}$$

- The **steepest descent** direction is the opposite unit vector

$$-\mathbf{u}_{\max} = \frac{\langle -7, \, 10 \rangle}{\|\langle 7, -10 \rangle\|}. \tag{35}$$

(Using the raw vector $\langle 7, -10 \rangle$ expresses the direction correctly, but to plug into the directional-derivative formula $D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u}$ you must use a **unit** vector.)

But how can be found the minimun in the entire curve, it should been able calculating and than finding the point ?

“Minimum on the entire curve?” (Global extrema).

$$f(x, y) = 2x^2 - 3y^2 + 3x + 2y \tag{29}$$

has gradient

$$\nabla f(x,y) = \langle 4x + 3, \ -6y + 2 \rangle \tag{31}$$

so the **critical point** solves $\nabla f = 0$:

$$4x + 3 = 0 \Rightarrow x^* = -\tfrac{3}{4}, \qquad -6y + 2 = 0 \Rightarrow y^* = \tfrac{1}{3}. \tag{36}$$

So the **critical point** is:

$$P^* = \left(-\frac{3}{4}, \frac{1}{3}\right) \tag{37}$$

At this point,

$$f(x^*, y^*) = \ 2\left(\tfrac{9}{16}\right) - 3\left(\tfrac{1}{9}\right) + 3\left(-\tfrac{3}{4}\right) + 2\left(\tfrac{1}{3}\right) = -\tfrac{19}{24}. \tag{38}$$

The **Hessian matrix** is:

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & -6 \end{pmatrix} \tag{39}$$

- First diagonal entry $4 > 0$ → convex in x-direction.
- Second diagonal entry $-6 < 0$ → concave in y-direction.

Because we have **one positive** and **one negative** second derivative, the Hessian is **indefinite**.

The surface curves **upward** in one direction and **downward** in another — that's the definition of a **saddle point**.

Why there's no global min or max

- **Unbounded below:**
  If x is fixed and $|y| \to \infty$,
  the term $-3y^2 \to -\infty$ dominates → $f(x,y) \to -\infty$.
  So there's no **lowest** value.
- **Unbounded above:**
  If y is fixed and $x| \to \infty$,
  the term $2x^2 \to +\infty$ dominates → $f(x,y) \to +\infty$.
  So there's no **highest** value.

When would a minimum exist?

If **restrict** the domain — for example, $x^2 + y^2 = 1 (circle)$ — then can be use **Lagrange multipliers** to find the min/max *on that region*.
Because the region is **closed and bounded** (compact set), the **Extreme Value Theorem** guarantees a min and max will exist.

Gradient Descent in ANN context

In an Artificial Neural Network, we define a **loss function** $L(\mathbf{W}, \mathbf{b})$ that measures how far the network's predictions are from the target outputs.

- $\mathbf{W}$ = weights matrix (and $\mathbf{b}$ = biases).
- **Goal:** Find values of $\mathbf{W}$ and $\mathbf{b}$ that **minimize** L.

  Why gradient? The gradient

$$\nabla_{\mathbf{W}} L \tag{40}$$

tells us the direction in weight space where L increases the fastest.

Why descent?
We step **opposite** to the gradient so we move toward decreasing L — ideally toward a minimum.

How it relates to the "minimum" idea

In the previous earlier $2D$ example, the gradient $\nabla f(x, y)$ told us the direction of steepest increase.
Gradient descent is like walking **downhill** on that surface until you reach a low point (a local minimum).

In ANN training:

- The "surface" = **loss function** over many weight parameters.
- The "point" = current weight values.
- The "descent path" = successive weight updates in the negative gradient direction.

Practical example

Taking a **very small ANN**: one input x, one weight w, no bias.
The network output is:

$$\hat{y} = w \cdot x \tag{41}$$

We'll use **Mean Squared Error (MSE)** for one training example (x, y):

$$L(w) = \frac{1}{2}(y - wx)^2 \tag{42}$$

Step 1 – Compute derivative (gradient):

$$\frac{dL}{dw} = -x(y - wx) \tag{43}$$

Step 2 – Update rule (gradient descent):

$$w_{\text{new}} = w_{\text{old}} - \eta \cdot \frac{dL}{dw} \tag{44}$$

where η is the learning rate.

Example calculation
Let's say:

- $x = 2$
- $y = 4$
- initial $w = 0$
- learning rate $\eta = 0.1$

1. Compute loss:
   $L(0) = \frac{1}{2}(4 - 0)^2 = 8$
2. Gradient:
   $\frac{dL}{dw} = -2 \cdot (4 - 0) = -8$
3. Update:
   $w_{\text{new}} = 0 - 0.1 \cdot (-8) = 0.8$

Next iteration:

1. $L(0.8) = \frac{1}{2}(4 - 1.6)^2 = 2.88$ (loss decreased ✅)
2. Gradient: $-2 \cdot (4 - 1.6) = -4.8$
3. Update: $w_{\text{new}} = 0.8 - 0.1 \cdot (-4.8) = 1.28$

Repeating this, w converges toward 2.0, which is the value that makes predictions perfect ($\hat{y} = wx = 4$) and gives the **minimum loss**.

Partial derivative

A **partial derivative** is the derivative of a function with multiple variables **with respect to one variable at a time**, treating the other variables as constants.

If we have:

$$f(x, y) = x^2 y + 3xy^3 \tag{45}$$

- $\frac{\partial f}{\partial x}$ means **differentiate with respect to** x, treat y as constant.
- $\frac{\partial f}{\partial y}$ means **differentiate with respect to** y, treat x as constant.

In **multivariable functions**, the output depends on **all inputs**. To know how sensitive the output is to changes in just one variable, we need partial derivatives.

Formal definition

If $f(x_1, x_2, \ldots, x_n)$ is differentiable, then the **partial derivative** with respect to x_i is:

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h} \tag{46}$$

Practical Example 1

Let:

$$f(x, y) = 2x^2 y + 5y \tag{46}$$

- **Partial derivative x:**
  Treat y as constant:

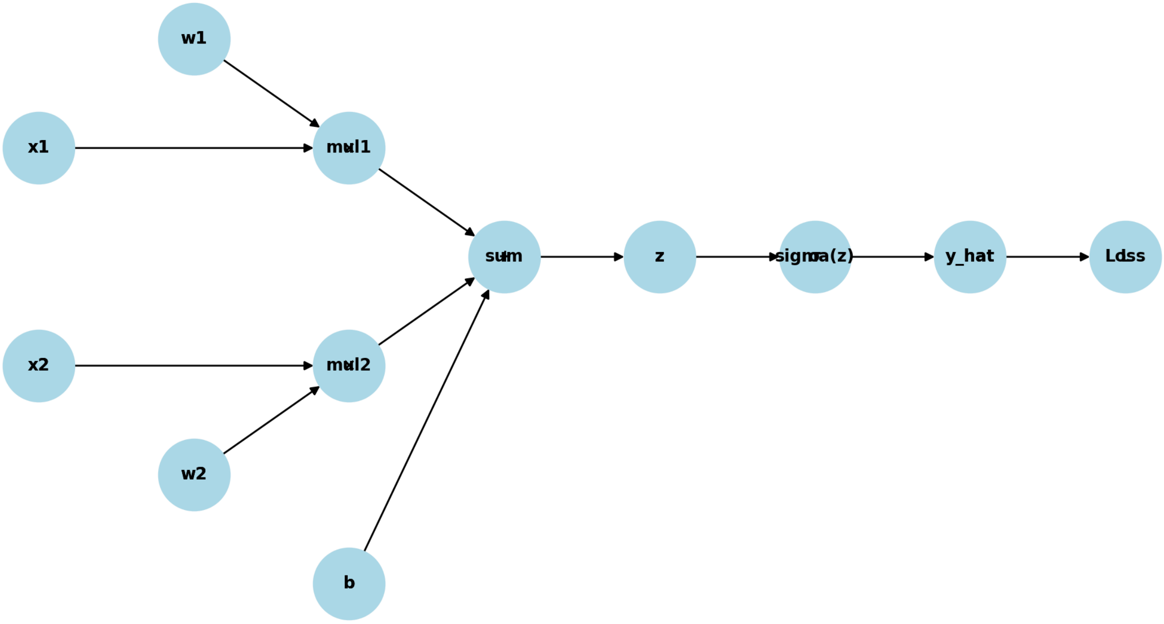
$$\frac{\partial f}{\partial x} = 4xy \tag{47}$$

- **Partial derivative y:**
  Treat x as constant:

$$\partial f \partial y = 2x2 + 5 \frac{\partial f}{\partial y} = 2x^2 + 5 \tag{48}$$

These tell you how f changes if you change x while keeping y fixed, or vice versa.

Practical Example 2 — Backpropagation in ANN



Imagine a single neuron:

$$\begin{aligned} z &= w_1 x_1 + w_2 x_2 + b \\ \hat{y} &= \sigma(z) \end{aligned} \tag{49}$$

where:

- w_1, w_2 = weights
- x_1, x_2 = inputs

- b = bias
- σ = activation function
- Loss:

$$L = \frac{1}{2}(\hat{y} - y)^2 \tag{50}$$

We want:

$$\frac{\partial L}{\partial w_1} \quad and \quad \frac{\partial L}{\partial w_2}$$

Step-by-step using the chain rule:

1. **Loss derivative output:**

$$\frac{\partial L}{\partial \hat{y}} = \hat{y} - y \tag{51}$$

2. **Output derivative z:**
   If σ is sigmoid:

$$\frac{\partial \hat{y}}{\partial z} = \hat{y}(1 - \hat{y}) \tag{52}$$

3. **z derivative w_1:**

$$\frac{\partial z}{\partial w_1} = x_1 \tag{53}$$

and similarly:

$$\frac{\partial z}{\partial w_2} = x_2 \tag{54}$$

4. **Chain rule to get $\frac{\partial L}{\partial w_1}$:**

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_1} = (\hat{y} - y) \cdot \hat{y}(1 - \hat{y}) \cdot x_1 \tag{55}$$

Similarly:

$$\frac{\partial L}{\partial w_2} = (\hat{y} - y) \cdot \hat{y}(1 - \hat{y}) \cdot x_2 \tag{56}$$

Interpretation:

- The **partial derivative** w_1 tells us how much the loss would change **if we only tweak** w_1, keeping all other weights constant.
- Gradient descent uses **all these partial derivatives together** to update every weight simultaneously toward reducing loss.

Summary of Main Differentiation Rules for Multivariable Functions

1. Product Rule

For

$$u(x,y) \quad and \quad v(x,y) \tag{57}$$

$$\frac{\partial}{\partial x}[u \cdot v] = \frac{\partial u}{\partial x} \cdot v + u \cdot \frac{\partial v}{\partial x} \tag{58}$$

Differentiate each factor in turn, keeping the other fixed.

2. Chain Rule (Single variable composition)

If

$$z = f(g(x)) \tag{59}$$

$$\frac{dz}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx} \tag{60}$$

3. Multivariable Chain Rule

If:

$$z = f(u,v), \quad u = u(x,y), \quad v = v(x,y) \tag{61}$$

then:

$$\frac{\partial z}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} \tag{62}$$

(similarly for $\frac{\partial z}{\partial y}$).

4. Exponential Function Derivative

For

$$e^{g(x,y)} \tag{63}$$

$$\frac{\partial}{\partial x} e^{g(x,y)} = e^{g(x,y)} \cdot \frac{\partial g}{\partial x} \tag{64}$$

Always multiply the exponential by the derivative of its exponent.

5. Sum Rule

$$\frac{\partial}{\partial x}[u + v] = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x} \tag{65}$$

Practical Example

Let’s take:

$$f(x,y) = xy \cdot e^{-(x^2+y^2)} \tag{66}$$

Here:

$$u(x,y) = xy$$

$$v(x,y) = e^{-(x^2+y^2)}$$

Step 1 — Partial derivative x (Product Rule)

$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial (xy)}{\partial x}}_{y} \cdot e^{-(x^2+y^2)} + (xy) \cdot \underbrace{\frac{\partial}{\partial x} e^{-(x^2+y^2)}}_{\text{Chain Rule}} \tag{67}$$

Step 2 — Chain Rule on the exponential

Let

$$g(x,y) = -(x^2+y^2) \tag{68}$$

$$\frac{\partial}{\partial x} = e^{g(x,y)} \cdot \frac{\partial g}{\partial x} = e^{-(x^2+y^2)} \cdot (-2x) \tag{69}$$

Step 3 — Put it together

$$\begin{aligned}\frac{\partial f}{\partial x} &= ye^{-(x^2+y^2)} + (xy) \cdot \left[e^{-(x^2+y^2)}(-2x) \right] \\ &= ye^{-(x^2+y^2)} - 2x^2ye^{-(x^2+y^2)} \\ &= y(1-2x^2)e^{-(x^2+y^2)}\end{aligned}$$

• *Derivadas*

Sejam u e v funções deriváveis de x e n constante.

1. $y = u^n \;\Rightarrow y' = n\,u^{n-1}u'$.
2. $y = uv \;\Rightarrow y' = u'v + v'u$.
3. $y = \frac{u}{v} \;\Rightarrow y' = \frac{u'v - v'u}{v^2}$.
4. $y = a^u \;\Rightarrow y' = a^u(\ln a)\,u',\; (a > 0,\; a \neq 1)$.
5. $y = e^u \;\Rightarrow y' = e^u u'$.
6. $y = \log_a u \;\Rightarrow y' = \frac{u'}{u}\log_a e$.
7. $y = \ln u \;\Rightarrow y' = \frac{1}{u}u'$.
8. $y = u^v \;\Rightarrow y' = v\,u^{v-1}\,u' + u^v(\ln u)\,v'$.
9. $y = \text{sen}\; u \;\Rightarrow y' = u' \cos\; u$.
10. $y = \cos\; u \;\Rightarrow y' = -u'\text{sen}\; u$.
11. $y = \text{tg}\; u \;\Rightarrow y' = u' \sec^2 u$.
12. $y = \text{cotg}\; u \;\Rightarrow y' = -u'\text{cosec}^2 u$.
13. $y = \sec\; u \;\Rightarrow y' = u' \sec\; u\,\text{tg}\; u$.
14. $y = \text{cosec}\; u \;\Rightarrow y' = -u'\text{cosec}\; u\;\text{cotg}\; u$.
15. $y = arc\;\text{sen}\; u \;\Rightarrow y' = \frac{u'}{\sqrt{1-u^2}}$.
16. $y = arc\;\cos\; u \;\Rightarrow y' = \frac{-u'}{\sqrt{1-u^2}}$.
17. $y = arc\;\text{tg}\; u \;\Rightarrow y' = \frac{u'}{1+u^2}$.
18. $y = arc\;\text{cot}\,g\; u \;\Rightarrow \frac{-u'}{1+u^2}$.
19. $y = arc\;\sec\; u,\; |u| \geqslant 1$
    $\Rightarrow y' = \frac{u'}{|u|\sqrt{u^2-1}},\; |u| > 1$.
20. $y = arc\;\text{cosec}\; u,\; |u| \geqslant 1$
    $\Rightarrow y' = \frac{-u'}{|u|\sqrt{u^2-1}},\; |u| > 1$.

• *Identidades Trigonométricas*

1. $\text{sen}^2 x + \cos^2 x = 1$.
2. $1 + \text{tg}^2 x = \sec^2 x$.
3. $1 + \text{cotg}^2 x = \text{cosec}^2 x$.
4. $\text{sen}^2 x = \frac{1-\cos\,2x}{2}$.
5. $\cos^2 x = \frac{1+\cos\,2x}{2}$.
6. $\text{sen}\;2x = 2\;\text{sen}\;x\;\cos\;x$.
7. $2\;\text{sen}\;x\;\cos\;y = \text{sen}\,(x-y) + sen\,(x+y)$.
8. $2\;\text{sen}\;x\;\text{sen}\;y = \cos\,(x-y) - \cos\,(x+y)$.
9. $2\;\cos\;x\;\cos\;y = \cos\,(x-y) + \cos\,(x+y)$.
10. $1 \pm \text{sen}\;x = 1 \pm \cos\left(\frac{\pi}{2} - x\right)$.

• *Integrais*

1. $\int du = u + c$.
2. $\int u^n du = \frac{u^{n+1}}{n+1} + c,\; n \neq -1$.
3. $\int \frac{du}{u} = \ln|u| + c$.
4. $\int a^u du = \frac{a^u}{\ln\;a} + c,\; a > 0,\; a \neq 1$.
5. $\int e^u du = e^u + c$.
6. $\int \text{sen}\; u\; du = -\cos\; u + c$.
7. $\int \cos\; u\; du = \text{sen}\; u + c$.
8. $\int \text{tg}\; u\; du = \ln|\sec\; u| + c$.
9. $\int \text{cotg}\; u\; du = \ln|\text{sen}\; u| + c$.
10. $\int \sec\; u\; du = \ln|\sec\; u + \text{tg}\; u| + c$.
11. $\int \text{cosec}\; u\; du = \ln|\text{cosec}\; u - \text{cotg}\; u| + c$.
12. $\int \sec\; u\;\text{tg}\; u\; du = \sec\; u + c$.
13. $\int \text{cosec}\; u\;\text{cotg}\; u\; du = -\text{cosec}\; u + c$.
14. $\int \sec^2 u\; du = \text{tg}\; u + c$.
15. $\int \text{cosec}^2 u\; du = -\text{cotg}\; u + c$.
16. $\int \frac{du}{u^2+a^2} = \frac{1}{a} arc\;\text{tg}\frac{u}{a} + c$.
17. $\int \frac{du}{u^2-a^2} = \frac{1}{2a}\ln\left|\frac{u-a}{u+a}\right| + c,\; u^2 > a^2$.
18. $\int \frac{du}{\sqrt{u^2+a^2}} = \ln\left|u + \sqrt{u^2+a^2}\right| + c$.
19. $\int \frac{du}{\sqrt{u^2-a^2}} = \ln\left|u + \sqrt{u^2-a^2}\right| + c$.
20. $\int \frac{du}{\sqrt{a^2-u^2}} = arc\;\text{sen}\frac{u}{a} + c,\; u^2 < a^2$.
21. $\int \frac{du}{u\sqrt{u^2-a^2}} = \frac{1}{a} arc\;\sec\left|\frac{u}{a}\right| + c$.

• *Fórmulas de Recorrência*

1. $\int sen^n au\; du = -\frac{sen^{n-1}au\;\cos\;au}{an}$
   $+ \left(\frac{n-1}{n}\right)\int sen^{n-2}au\; du$.

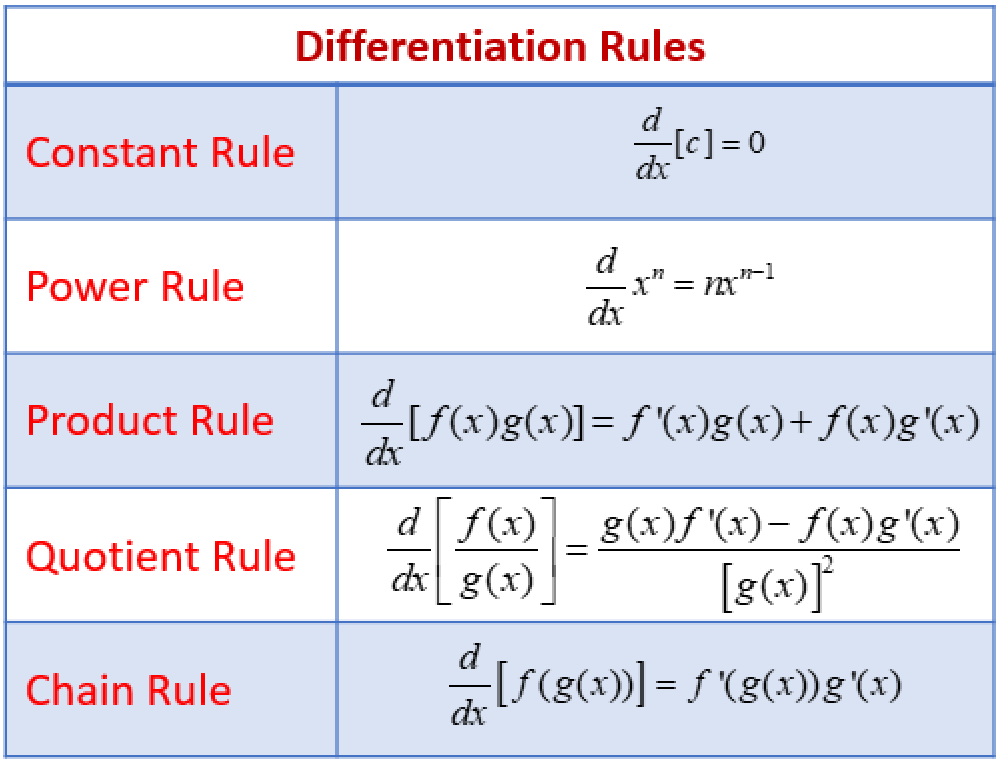
2. $\int \cos^n au\; du = \frac{sen\;au\;\cos^{n-1}au}{an}$
   $+ \left(\frac{n-1}{n}\right)\int \cos^{n-2} au\; du$.

3. $\int tg^n au\; du = \frac{tg^{n-1}au}{a(n-1)} - \int tg^{n-2}au\; du$.

4. $\int cotg^n au\; du = -\frac{cotg^{n-1}au}{a(n-1)} - \int cotg^{n-2}au\; du$.

5. $\int \sec^n au\; du = \frac{\sec^{n-2} au\; tg\; au}{a(n-1)}$
   $+ \left(\frac{n-2}{n-1}\right)\int \sec^{n-2} au\; du$.

6. $\int cosec^n au\; du = -\frac{cosec^{n-2}au\; cotg\; au}{a(n-1)}$
   $+ \left(\frac{n-2}{n-1}\right)\int cosec^{n-2}au\; du$.



Differentiation Rules	
Constant Rule	$\frac{d}{dx}[c] = 0$
Power Rule	$\frac{d}{dx}x^n = nx^{n-1}$
Product Rule	$\frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$
Quotient Rule	$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{\left[g(x)\right]^2}$
Chain Rule	$\frac{d}{dx}[f(g(x))] = f'(g(x))g'(x)$

Contexto

- **Situação**:
- **Fonte**:

Próximos Passos

- **Ação 1**:
- **Ação 2**:

Referências

- Fonte 1 - Nota breve sobre a fonte
- Fonte 2 - Nota breve sobre a fonte