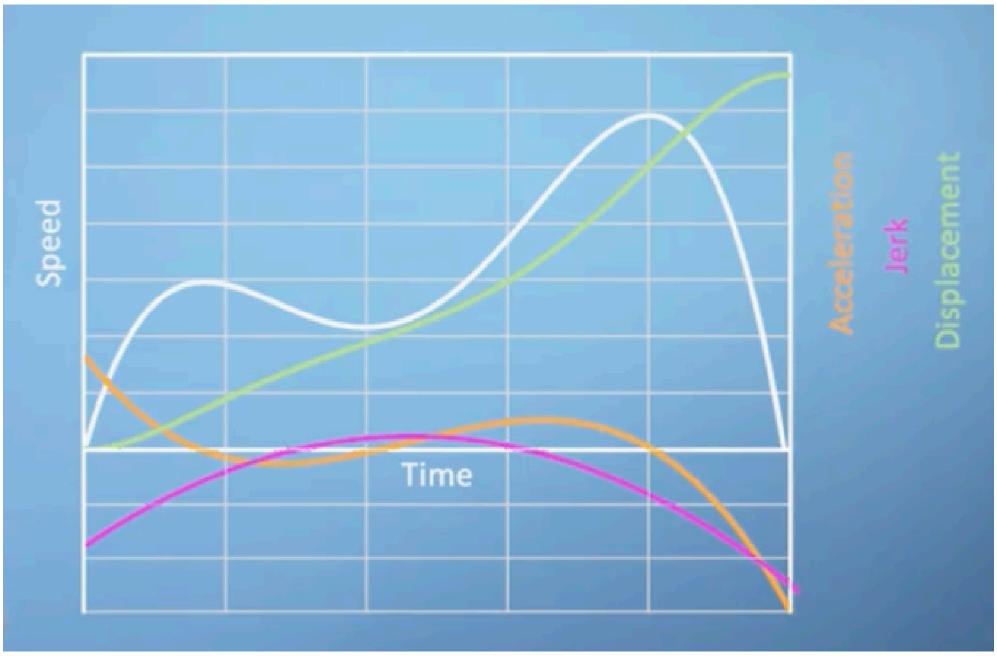
A. Introduction to Calculus

A1. Introduction

- What are functions?
  - A function is a mathematical relationship between inputs and an output. It can be thought of as a machine that takes in one or more variables and produces a single, corresponding result. For example, a function for the temperature of a room might take in the coordinates (x,y,z) and time (t) as inputs and return the temperature at that specific point and time.
  - The notation $f(x)$ represents "f as a function of x", not "f multiplied by x." This can be a point of confusion due to its seemingly arbitrary nature, but it's a standard convention in mathematical language.
- The creative essence of science
  - Selecting a function to model real-world data is a core, creative step in science and machine learning. This process involves formulating a **hypothesis**—a candidate function that could represent the relationship you're observing. Without this initial creative step, there would be nothing to test or investigate.
- Introduction to Calculus
  - **Calculus** is the study of how functions change with respect to their input variables. It provides a set of tools to investigate and manipulate these functions. By understanding calculus, you can analyze the behavior of functions and use them to model complex phenomena in the real world.
- Gradients and Derivatives
  - A great way to visualize this concept is with a **speed-time graph**.
  - 

        - The **gradient** (or slope) of the graph at any point represents the **rate of change** of speed with respect to time, which is the **acceleration**.
        - A positive gradient indicates acceleration, a negative gradient indicates deceleration, and a zero gradient (a flat horizontal line) means constant speed with zero acceleration.
        - The gradient at a single point is called the **local gradient** and can be visualized as the slope of a tangent line that touches the curve at that point.
    - By finding the local gradient at every point on a continuous function, we can create an entirely new function called its **derivative**. The derivative describes the original function's slope at every point.
- Higher-Order Derivatives and Anti-Derivatives
    - This process can be repeated. The derivative of the acceleration function is called the **jerk**, which represents the rate of change of acceleration. This concept is useful for describing the "jerky" motion of a car as it starts and stops. The jerk is the second derivative of the speed-time function.
    - The inverse procedure, finding a function for which our original function is the derivative, is called the **anti-derivative**. For our speed-time example, the anti-derivative would be the **distance-time functio**n, as the rate of change of distance is speed. The anti-derivative is closely related to an integral.

A2. Derivatives (Sum Rule and Power Rule)

- Defining the Derivative
    - The derivative is the formal mathematical notation for the gradient of a function. For a linear function with a constant gradient, the slope is defined as "**rise over run**."
    - For a non-linear function where the gradient changes at every point, we define the derivative at a specific point x by taking the limit of the "rise over run" formula. We consider a second point that is an infinitesimally small distance Δx away from the first point. As this distance approaches zero, the line connecting the two points becomes a perfect approximation of the tangent line at point x.

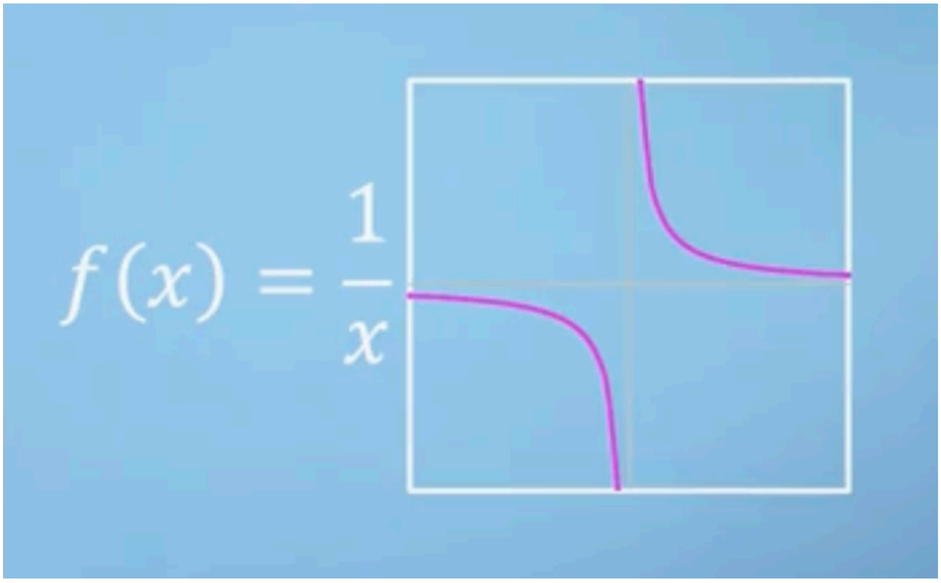
$$\frac{df}{dx} = f'(x) = \lim_{\Delta x \to 0} (\frac{f(x + \Delta x) - f(x)}{\Delta x})$$

        - The notation for the derivative can be either $f'(x)$ (read as "f prime of x")
        - or $\frac{df}{dx}$ (read as "df by dx").
        - The key idea is that we are not dividing by zero, but rather observing the behavior of the expression as Δx gets extremely close to zero.
- Fundamental Rules of Differentiation

  - This definition, while powerful, can be tedious to apply directly to every function. Fortunately, we can derive and use general rules to simplify the process.
  - **The Sum Rule**
    - The derivative of a sum of functions is the sum of their individual derivatives. This means you can differentiate each term in a function separately and then add the results together.
    - Example: The derivative of $f(x) = 3x + 2$ is the derivative of $3x$ plus the derivative of 2.
  - **The Power Rule**
    - For a function in the form of $f(x) = ax^b$, its derivative is:

$$f(x) = ax^b$$

$$f'(x) = abx^{(b-1)}$$

    - The rule is: multiply the coefficient by the original power, and then subtract 1 from the power.
    - Example: For $f(x) = 5x^2$, the derivative is $f'(x) = (5)(2)x^{2-1} = 10x^1 = 10x$
- Special cases
  - The Derivative of $\frac{1}{x}$
    - 

    - The Derivative of $\frac{1}{x}$ has a **discontinuity** at $x = 0$, as division by zero is undefined. However, we can find its derivative using the limit definition of differentiation. After working through the algebra, the derivative is found to be:

$$f(x) = \frac{1}{x}$$

$$f'(x) = \lim_{\Delta x \to 0} \left(\frac{\frac{1}{x+\Delta x} - \frac{1}{x}}{\Delta x}\right)$$

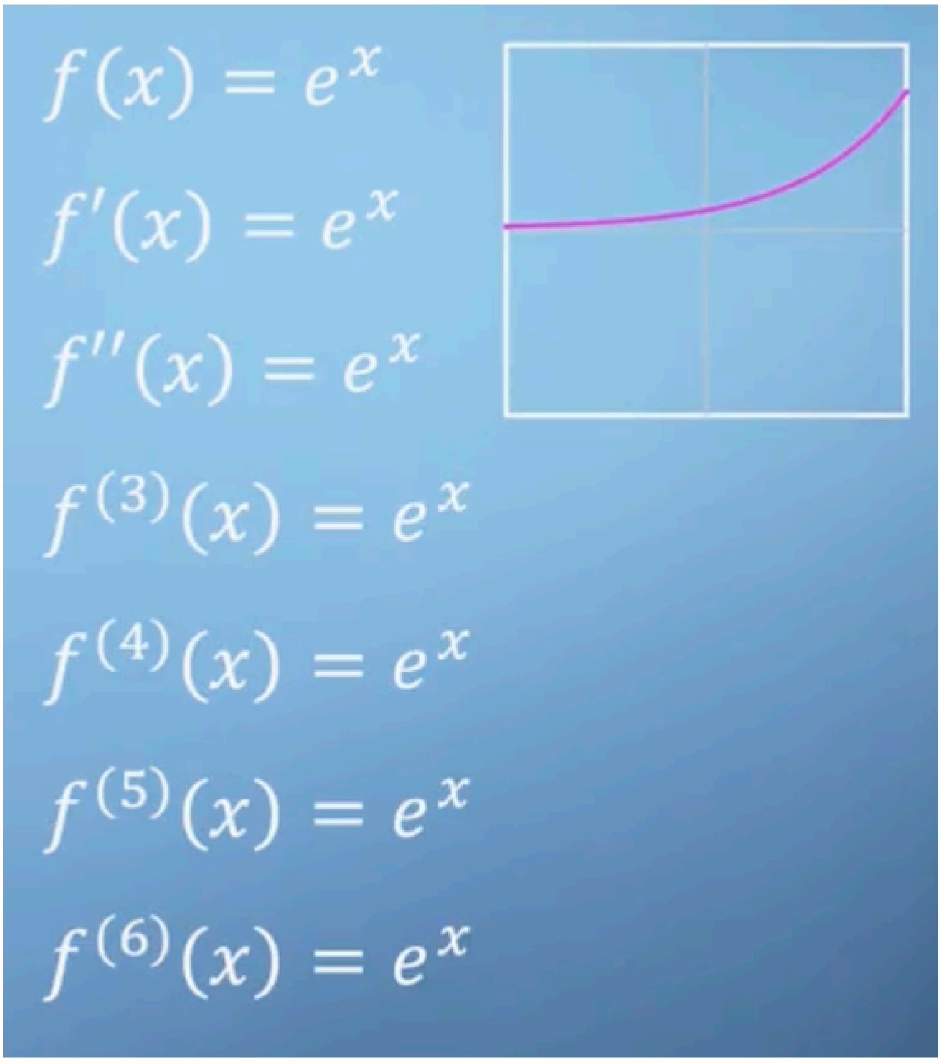
$$= \lim_{\Delta x \to 0} (\frac{\frac{x}{x(x+\Delta x)} - \frac{x+\Delta x}{x(x+\Delta x)}}{\Delta x})$$

$$= \lim_{\Delta x \to 0} (\frac{\frac{-\Delta x}{x(x+\Delta x)}}{\Delta x})$$

$$= \lim_{\Delta x \to 0} (\frac{-1}{x^2 + x\Delta x})$$

$$= -\frac{1}{x^2}$$

  - This derivative function is always negative, matching our visual observation that the original function's slope is always decreasing. Like the original function, the derivative is also undefined at $x = 0$.

- The Exponential Function (e^x)
  - 

  - The exponential function, $f(x) = e^x$, has a unique and powerful property: **its derivative is itself**.

$$\frac{d}{dx}e^x = e^x$$

  - This means the value of the function at any point is equal to its slope at that same point. This self-similarity is incredibly useful in calculus and other areas of mathematics. The constant e (Euler's number), approximately 2.718, is fundamental to this function.

- The "Designed" Nature of the Exponential Function e^x
  - The constant e (Euler's number) is not a random value; it's specifically defined to satisfy a unique and powerful property in calculus. Its entire purpose is to make the derivative of the exponential function $f(x) = e^x$ equal to itself.
- Derivation of the Derivative
  - We can see this by using the formal limit definition of the derivative for a general exponential function $f(x) = a^x$.

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

  - Using the rule of exponents ($a^{x+y} = a^x a^y$), we can factor out a^x:

    $$f'(x) = \lim_{\Delta x \to 0} \frac{a^x \cdot a^{\Delta x} - a^x}{\Delta x} = a^x \lim_{\Delta x \to 0} \frac{a^{\Delta x} - 1}{\Delta x}$$

    .
  - For most values of the base a, the limit part of this expression will evaluate to some constant value other than 1.
  - The number e is precisely the value for the base a that makes this limit exactly 1:

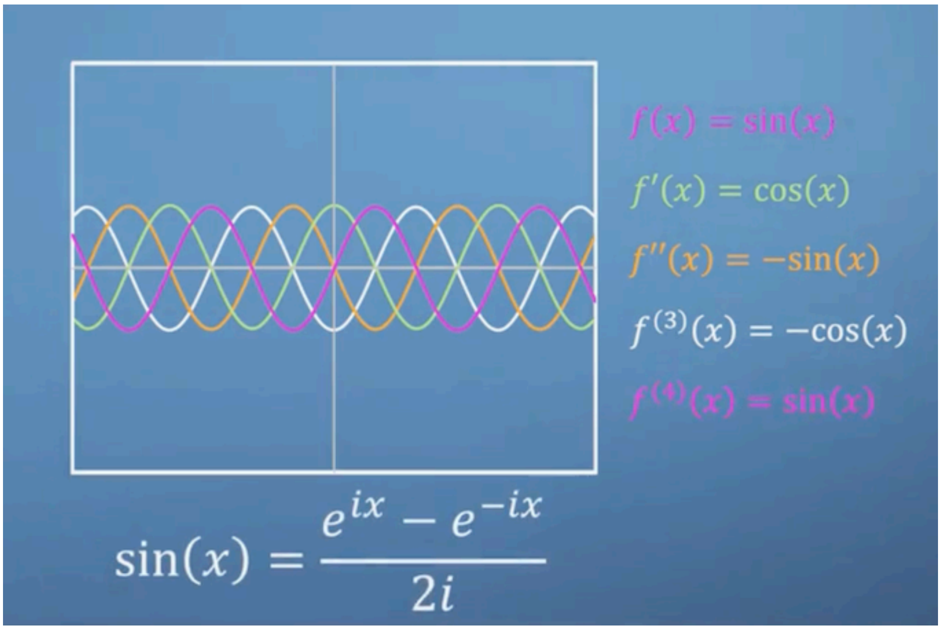
    $$\lim_{\Delta x \to 0} \frac{e^{\Delta x} - 1}{\Delta x}$$

  - Therefore, when we substitute $a = e$ back into our derivative expression, we get:

    $$f'(x) = e^x \cdot 1 = e^x$$

- This result shows that the rate of change of the function e^x at any point is simply the value of the function itself at that point. This isn't a coincidence; it's the very reason why e is so fundamental to calculus and the modeling of natural growth and decay processes.

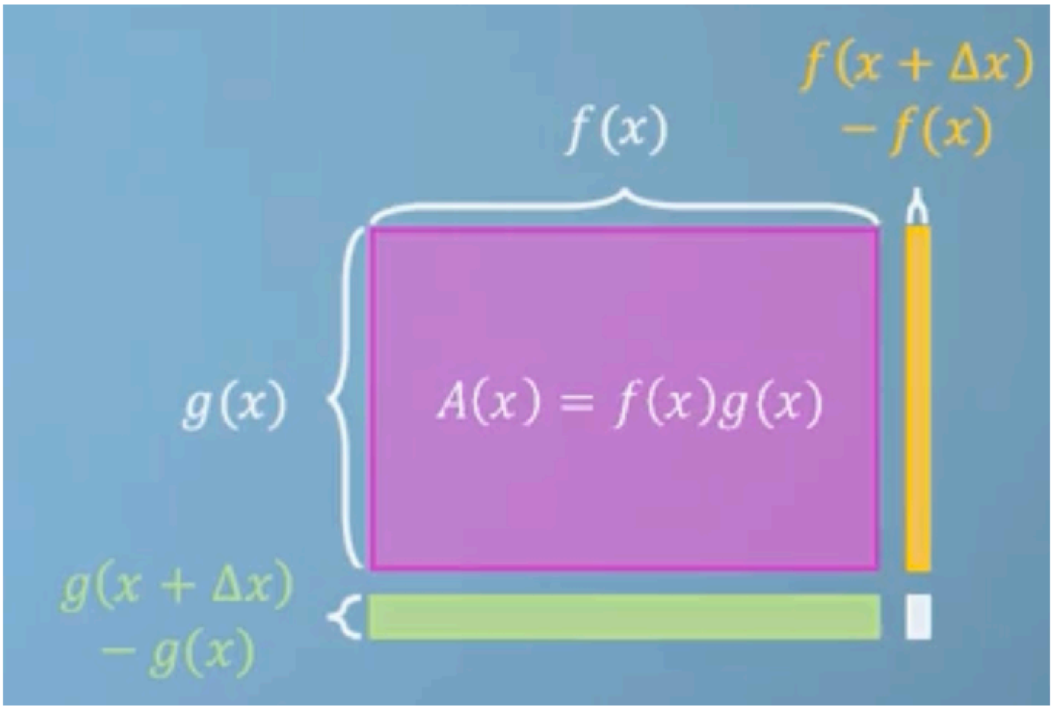
- Trigonometric Functions (Sine and Cosine)

        - 

        - The trigonometric functions sine and cosine have an interesting relationship when differentiated. They follow a cyclical pattern:
            - The derivative of $\sin(x)$ is $\cos(x)$.
            - The derivative of $\cos(x)$ is $-\sin(x)$.
            - The derivative of $-\sin(x)$ is $-\cos(x)$.
            - The derivative of $-\cos(x)$ is $\sin(x)$.
        - After four differentiations, the function returns to its original form. This self-similarity is a hint that these functions are deeply related to the exponential function, although the connection is not immediately obvious.
    - Ultimately, these examples demonstrate that even with complex functions, the core concept of differentiation remains the same: finding the "rise over run" at every point on the curve.

A3. The Product Rule of Differentiation

- The product rule is a shortcut for finding the derivative of a function that is the product of two separate functions, $A(x) = f(x)g(x)$. Instead of using the tedious limit definition, we can visualize the rule by thinking about the change in area of a rectangle with sides $f(x)$ and $g(x)$.

- 

- When we increase x by a small amount Δx, the area of the rectangle changes. The increase in area, ΔA, consists of three parts:
  - A vertical strip with area $f(x)(g(x+\Delta x)-g(x))$.
  - A horizontal strip with area $g(x)(f(x+\Delta x)-f(x))$
  - A small corner rectangle with area $(f(x+\Delta x)-f(x))(g(x+\Delta x)-g(x))$
- So, the whole ΔA will be:

$$\Delta A = f(x)(g(x+\Delta x)-g(x))+$$

$$g(x)(f(x+\Delta x)-f(x))+$$

$$(f(x+\Delta x)-f(x))(g(x+\Delta x)-g(x))$$

- As Δx approaches zero, the area of the smallest corner rectangle $(f(x+\Delta x)-f(x))(g(x+\Delta x)-g(x))$ becomes negligible compared to the other two parts and can be ignored in the limit.

$$\lim_{\Delta x\to 0}(\Delta A(x)) = \lim_{\Delta x\to 0}(f(x)(g(x+\Delta x)-g(x))+g(x)(f(x+\Delta x)-f(x)))$$

$$= \lim_{\Delta x\to 0}(\frac{\Delta A(x)}{\Delta x}) = \lim_{\Delta x\to 0}(\frac{f(x)(g(x+\Delta x)-g(x))+g(x)(f(x+\Delta x)-f(x)}{\Delta x}$$

$$= \lim_{\Delta x\to 0}(\frac{\Delta A(x)}{\Delta x}) = \lim_{\Delta x\to 0}(\frac{f(x)(g(x+\Delta x)-g(x))}{\Delta x}+\frac{g(x)(f(x+\Delta x)-f(x}{\Delta x}$$

$$= \lim_{\Delta x\to 0}(\frac{\Delta A(x)}{\Delta x}) = \lim_{\Delta x\to 0}(f(x)\frac{(g(x+\Delta x)-g(x))}{\Delta x}+g(x)\frac{(f(x+\Delta x)-f(x}{\Delta x}$$

$$= \lim_{\Delta x\to 0}(\frac{\Delta A(x)}{\Delta x}) = \lim_{\Delta x\to 0}(f(x)g'(x)+g(x)f'(x))$$

$$= A'(x) = f(x)g'(x)+g(x)f'(x)$$

- Based on this intuition, we can derive the formal product rule. It states that the derivative of a product of two functions, $f(x)g(x)$, is the sum of two terms: the first function times the derivative of the second, plus the second function times the derivative of the first.

$$\{\text{Product Rule} = \ \}$$

$$\text{if } A(x) = f(x)g(x)$$

$$\text{then } A'(x) = f(x)g'(x) + g(x)f'(x)$$

- This rule is a powerful tool in calculus and can be added to our toolbox alongside the Sum Rule and the Power Rule. It simplifies the process of differentiating complex functions that are the product of simpler ones.
- Example:
  - To differentiate $A(x) = xe^x \cos(x)$, we apply the product rule for three functions from the previous question.
  - Let $f(x) = x$,$g(x) = e^x$, and $h(x) = \cos(x)$.
    - $f'(x) = 1$
    - $g'(x) = e^x$
    - $h'(x) = -\sin(x)$
  - Applying the three-function product rule $A'(x) = f'(x)g(x)h(x) + f(x)g'(x)h(x) + f(x)g(x)h'(x)$:

$$f'(x)g(x)h(x) = 1e^x \cos(x) = e^x \cos(x)$$

$$f(x)g'(x)h(x) = xe^x \cos(x)$$

$$f(x)g(x)h'(x) = xe^x - \sin(x)$$

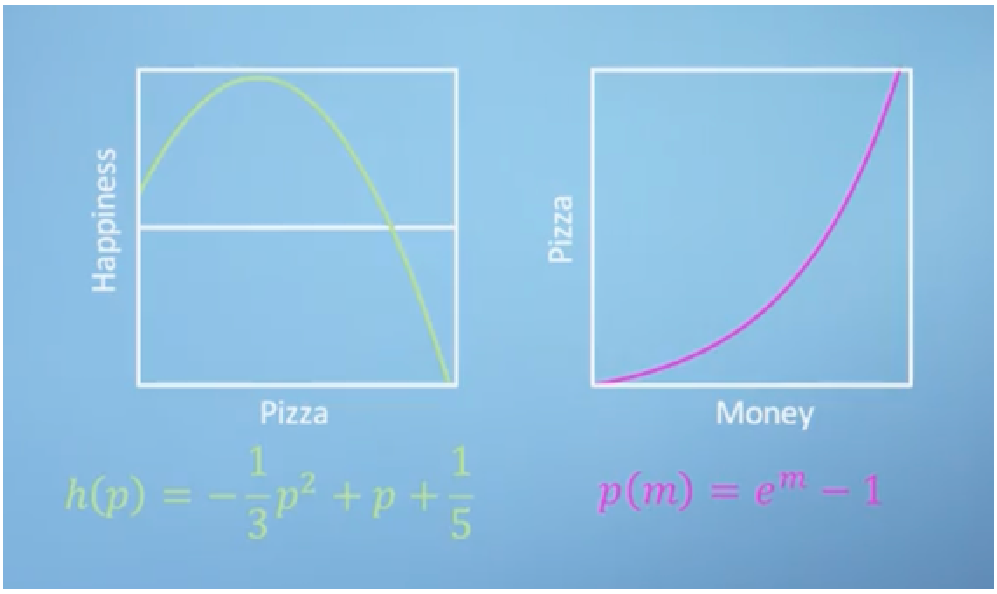
$$A'(x) = e^x \cos(x) + xe^x \cos(x) + xe^x - \sin(x)$$

$$A'(x) = e^x[\cos(x) + x\cos(x) - x\sin(x)]$$

$$A'(x) = e^x[(1 + x)\cos(x) - x\sin(x)]$$

A4. The Chain Rule of Differentiation

- The Chain Rule is the essential tool for differentiating composite functions, where one function is nested inside another (e.g., $h(p(m))$). This rule is the fourth and final tool needed to tackle complex differentiation problems.
  1. Conceptualizing the chain
  - A composite function relates an ultimate output to a final input through a chain of intermediate variables. This structure is common in science and engineering.
  - Example: Relating Happiness (h) to Money (m) via Pizza (p).
    - h is a function of p.
    - p is a function of m.

        - Goal: Find the rate of change of happiness with respect to money, $\frac{dh}{dm}$.
    - Example functions:
        - $h(p) = -\frac{1}{3}p^2 + p + \frac{1}{5}$
        - $p(m) = e^m - 1$
    - 

2. The Chain Rule formula
    - The Chain Rule provides an elegant approach by multiplying the derivatives of the successive functions.
    - The derivative of the composite function is the product of the derivative of the outer function with respect to the intermediate variable, and the derivative of the intermediate variable with respect to the innermost variable.
    - The formula is intuitively represented as a chain of derivative relationships:

$$\frac{dh}{dm} = \frac{dh}{dp}\frac{dp}{dm}$$

3. Application example
    - Step 1: Differentiate the individual functions.
        - $h(p) = -\frac{1}{3}p^2 + p + \frac{1}{5} \rightarrow \frac{dh}{dp} = 1 - \frac{2}{3}p$
        - $p(m) = e^m - 1 \rightarrow \frac{dp}{dm} = e^m$
    - Step 2: Apply the Chain Rule and eliminate the intermediate variable.
        - $\frac{dh}{dm} = \frac{dh}{dp}\frac{dp}{dm} = (1 - \frac{2}{3}p) \cdot e^m$
        - By substituting $p = e^m - 1$ back into the expression, we ensure the final derivative is only a function of m.

$$\frac{dh}{dm} = \left(1 - \frac{2}{3}(e^m - 1)\right) e^m$$

$$\frac{dh}{dm} = \frac{1}{3}e^m(5 - 2e^m)$$

- The magic of the Chain Rule is that it works even when direct substitution is impossible, provided we know the derivatives of the individual functions within the chain.

A5. Combination of the 4 rules (Sum, Power, Product, and Chain)

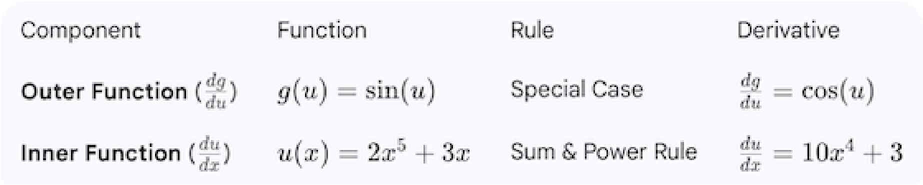
- Applying the Calculus toolbox
  - The function to be differentiated is:

$$f(x) = \frac{\sin(2x^5 + 3x)}{e^{7x}}$$

  - The core strategy is to decompose the scary function into manageable pieces and use the Product Rule as the final step.
    1. Preparation: Rewriting as a product
    - To avoid the Quotient Rule, we rewrite the fraction as a product using a negative exponent:

$$f(x) = \underbrace{\sin(2x^5 + 3x)}_{g(x)} \cdot \underbrace{e^{-7x}}_{h(x)}$$

    - The derivative will be found by the Product Rule: $f'(x) = g'(x)h(x) + g(x)h'(x)$.
    2. Part 1: Differentiating $g(x) = \sin(2x^5 + 3x)$
    - This is a classic Chain Rule scenario, $g(x) = g(u(x))$, where the inner function is $u(x)$.



Component	Function	Rule	Derivative
Outer Function ($\frac{dg}{du}$)	$g(u) = \sin(u)$	Special Case	$\frac{dg}{du} = \cos(u)$
Inner Function ($\frac{du}{dx}$)	$u(x) = 2x^5 + 3x$	Sum & Power Rule	$\frac{du}{dx} = 10x^4 + 3$

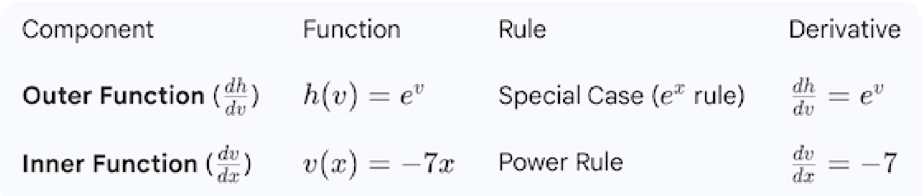
    - Applying the Chain Rule: $\frac{dg}{dx} = \frac{dg}{du} \cdot \frac{du}{dx}$:

$$\frac{dg}{dx} = \cos(u) \cdot (10x^4 + 3)$$

    - Substituting $u = 2x^5 + 3x$ back:

$$g'(x) = \cos(2x^5 + 3x)(10x^4 + 3)$$

    3. Part 2: Differentiating $h(x) = e^{-7x}$
    - This is also a Chain Rule scenario, $h(x) = h(v(x))$, where the inner function is $v(x)$.



Component	Function	Rule	Derivative
Outer Function ($\frac{dh}{dv}$)	$h(v) = e^v$	Special Case (e^x rule)	$\frac{dh}{dv} = e^v$
Inner Function ($\frac{dv}{dx}$)	$v(x) = -7x$	Power Rule	$\frac{dv}{dx} = -7$

    - Applying the Chain Rule: $\frac{dh}{dx} = \frac{dh}{dv} \cdot \frac{dv}{dx}$:

$$\frac{dh}{dx} = e^v \cdot (-7)$$

- Substituting $v = -7x$ back:

$$h'(x) = -7e^{-7x}$$

4. Final step: Applying the Product Rule

- Finally, apply the Product Rule: $f'(x) = g'(x)h(x) + g(x)h'(x)$.

$$f'(x) = \left[\cos(2x^5 + 3x)(10x^4 + 3)\right] \cdot \left[e^{-7x}\right] + \left[\sin(2x^5 + 3x)\right] \cdot \left[-7e^{-}\right.$$

- The expression can be slightly rearranged by factoring out e^{-7x}:

$$f'(x) = e^{-7x}\left[(10x^4 + 3)\cos(2x^5 + 3x) - 7\sin(2x^5 + 3x)\right]$$

- This single example utilized all four differentiation rules: Power Rule, Sum Rule, Chain Rule, and Product Rule. As is common in coding, further algebraic simplification (optimization) is often deferred until necessary.

B. Multivariate Calculus

B1. Partial Differentiation

- Variables, Constants, and Parameters

  In multivariate calculus, understanding the role of each term in a function is crucial for differentiation.

  - Dependent Variable (y): A variable whose value depends on the values of others (e.g., speed depends on time).
  - Independent Variable (x): A variable that is controlled or chosen freely, and on which the dependent variable relies (e.g., time).
  - Constants: Values that are fixed in the context of the problem (e.g., π).
  - Parameters: Variables that are considered constants during a standard differentiation but are often varied by an engineer or designer to explore a family of similar functions (e.g., car mass or drag coefficient in a specific context).

  Key Takeaway: What is a constant or a variable is context-dependent. In calculus, you can differentiate any term with respect to any other, provided the context makes sense.

- Introduction to Partial Differentiation

  Partial differentiation is the method of applying the familiar rules of single-variable calculus to functions of multiple variables.

  - The core rule is:

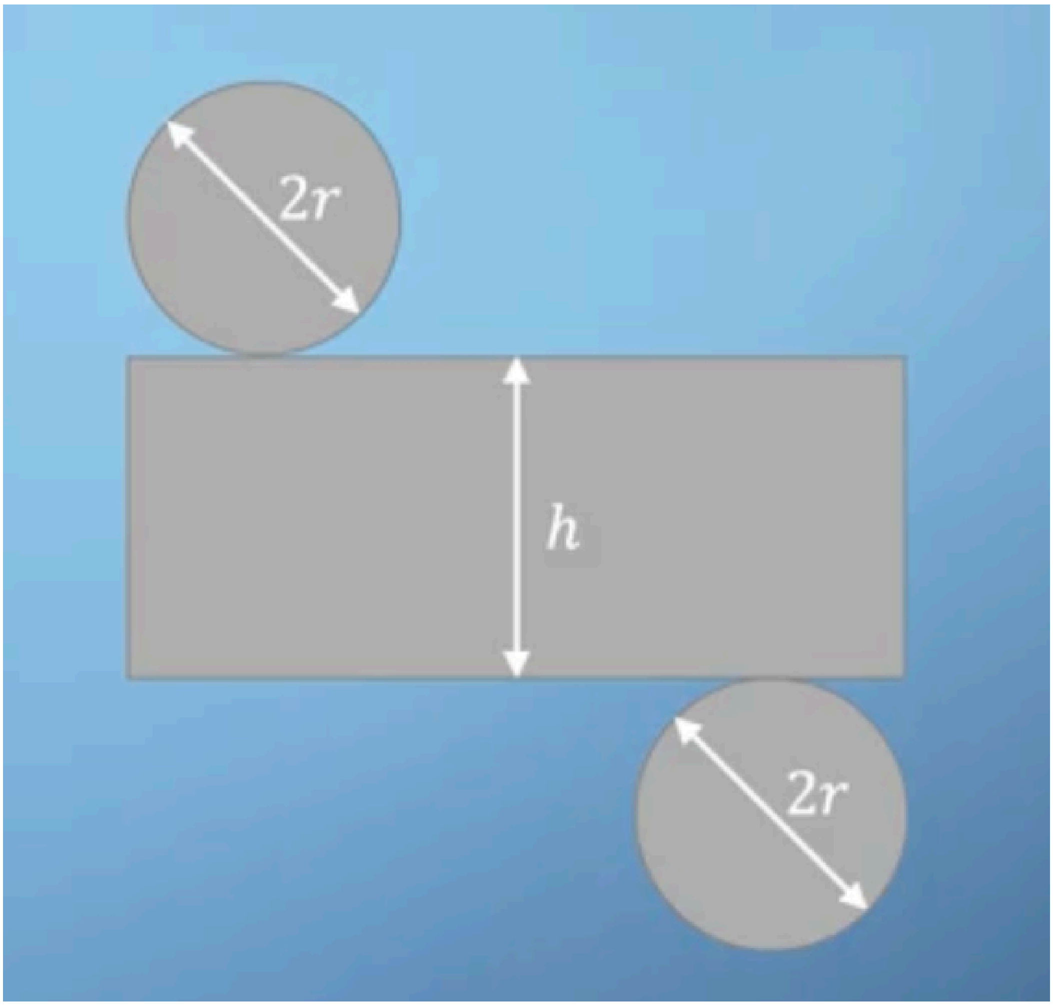
When differentiating a function with respect to a specific variable, treat all other variables as constants.

- The Partial Derivative Symbol

  We use the curly symbol, ∂ (read as "partial"), instead of the standard d, to signify that the function being differentiated has more than one variable.For a function $f(x, y, z)$, the partial derivative with respect to x is written as $\frac{\partial f}{\partial x}$.

- Example: Mass of a Can (m)

  The mass of a metal can is a function of its design parameters: radius (r), height (h), wall thickness (t), and density (ρ).



The mass function is:

$$m(r, h, t, \rho) = (2\pi r^2 t\rho) + (2\pi rht\rho)$$

- Partial Derivative with respect to Height (h)

  When calculating $\frac{\partial m}{\partial h}$, we treat r, t, and ρ as constants.

$$\frac{\partial m}{\partial h} = \frac{\partial}{\partial h}[2\pi r^2 t\rho] + \frac{\partial}{\partial h}[2\pi rht\rho]$$

  - The first term $(2\pi r^2 t\rho)$ does not contain h, so its derivative is 0.

- The second term ($2\pi rt\rho$ is the constant multiplier of h), so its derivative is $2\pi rt\rho$.

$$\frac{\partial m}{\partial h} = 0 + 2\pi rt\rho$$

$$\frac{\partial m}{\partial h} = 2\pi rt\rho$$

The result no longer contains h, as mass varies linearly with height (all else being equal).

- Partial Derivative with respect to Radius (r)

When calculating $\frac{\partial m}{\partial r}$, we treat h, t, and ρ as constants.

$$\frac{\partial m}{\partial r} = \frac{\partial}{\partial r}[2\pi r^2 t\rho] + \frac{\partial}{\partial r}[2\pi rht\rho]$$

- For the first term, $\frac{\partial}{\partial r}[2\pi t\rho \cdot r^2] = 2\pi t\rho \cdot (2r) = 4\pi rt\rho$.
- For the second term, $\frac{\partial}{\partial r}[2\pi ht\rho \cdot r] = 2\pi ht\rho$.

$$\frac{\partial m}{\partial r} = 4\pi rt\rho + 2\pi ht\rho$$

- Partial Derivative with respect to Thickness (t)

$$\frac{\partial m}{\partial t} = \frac{\partial}{\partial t}[2\pi r^2 \rho \cdot t] + \frac{\partial}{\partial t}[2\pi rh\rho \cdot t]$$

$$\frac{\partial m}{\partial t} = 2\pi r^2 \rho + 2\pi rh\rho$$

- Partial Derivative with respect to Density (ρ)

$$\frac{\partial m}{\partial \rho} = \frac{\partial}{\partial \rho}[2\pi r^2 t \cdot \rho] + \frac{\partial}{\partial \rho}[2\pi rht \cdot \rho]$$

$$\frac{\partial m}{\partial \rho} = 2\pi r^2 t + 2\pi rht$$

B2. Total Derivative

- Partial differentiation is the process of finding the derivative of a multivariate function with respect to one variable while treating all others as constants.
  - Consider the function: $f(x, y, z) = \sin(x)e^{yz^2}$
    - With respect to x ($\frac{\partial f}{\partial x}$)
      - Treat e^{yz^2} as a constant. Differentiate $\sin(x)$ to $\cos(x)$.
      - $\frac{\partial f}{\partial x} = \cos(x)e^{yz^2}$
    - With respect to y ($\frac{\partial f}{\partial y}$)

        - Treat $\sin(x)$ as a constant. Apply the Chain Rule to e^{yz^2} by multiplying by the derivative of the exponent w.r.t y, which is z^2.
        - $\frac{\partial f}{\partial y} = \sin(x)e^{yz^2}(z^2)$
    - With respect to z ($\frac{\partial f}{\partial z}$)
        - Treat $\sin(x)$ as a constant. Apply the Chain Rule to e^{yz^2} by multiplying by the derivative of the exponent w.r.t. z, which is $2yz$
        - $\frac{\partial f}{\partial z} = \sin(x)e^{yz^2}(2yz)$
- The Total Derivative
    - The Total Derivative is used when the function f depends on multiple variables (x, y, z), and those variables are themselves functions of a single other parameter, t (e.g., $x(t), y(t), z(t)$). The goal is to find the overall rate of change of f with respect to this single parameter, $\frac{df}{dt}$.
    - The total derivative is the sum of the changes contributed by each variable. It is a direct extension of the Chain Rule:

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

    - The power of this rule is that it works even when direct substitution of all t-expressions into f is algebraically too complex or impossible (if an analytical expression for f is unavailable).
    - Application Example

$$f(x, y, z) = \sin{(x)}e^{yz^2}$$

        - Given the dependencies: $x(t) = t - 1$, $y(t) = t^2$, $z(t) = 1/t = t^{-1}$

$$f(t) = \sin{(t-1)}e^{t^2(\frac{1}{t})^2}$$

$$f(t) = \sin{(t-1)}e$$

$$\frac{df}{dt} = \cos{(t-1)}e$$

        - Now we know that:
            - $\frac{\partial f}{\partial x} = \cos(x)e^{yz^2}$, and $\frac{dx}{dt} = 1$
            - $\frac{\partial f}{\partial y} = \sin(x)e^{yz^2}(z^2)$, and $\frac{dy}{dt} = 2t$
            - $\frac{\partial f}{\partial z} = \sin(x)e^{yz^2}(2yz)$, and $\frac{dz}{dt} = -t^{-2}$

$$\frac{df(x,y,z)}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt}$$

$$\frac{df(x,y,z)}{dt} = \cos(x)e^{yz^2} \cdot 1 + \sin(x)e^{yz^2}(z^2) \cdot 2t + \sin(x)e^{yz^2}(2yz)$$

        - We substitute x, y, z all interms of t, $x(t) = t - 1$, $y(t) = t^2$, $z(t) = 1/t = t^{-1}$

$$\frac{df(x,y,z)}{dt} = \cos(t-1)e + t^{-2}\sin(t-1)e \cdot 2t + 2t\sin(t-1)e \cdot (-t^{-}$$

$$\frac{df(x,y,z)}{dt} = \cos(t-1)e + 2t^{-1}\sin(t-1)e - 2t^{-1}\sin(t-1)e$$

$$\frac{df(x,y,z)}{dt} = \cos(t-1)e$$

B3. Jacobian

- The Jacobian is a vector that brings together the partial derivatives of a multivariate function into a single, useful structure. This concept is fundamental to optimization and machine learning, particularly when dealing with functions of many variables.

1. Definition of the Jacobian Vector (J)

   For a single function f that depends on multiple variables, $f(x_1, x_2, x_3, \ldots)$, the Jacobian is simply a row vector where each entry is the partial derivative of f with respect to each variable in turn.

   The Jacobian J is formally defined as:

$$J = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \frac{\partial f}{\partial x_3} & \ldots \end{bmatrix}$$

2. Worked Example

   Consider the function: $f(x,y,z) = x^2y + 3z$.

   To construct the Jacobian, we find each partial derivative:

   - $\frac{\partial f}{\partial x}$ (Treat y, z as constants): $2xy$
   - $\frac{\partial f}{\partial y}$ (Treat x, z as constants): x^2
   - $\frac{\partial f}{\partial z}$ (Treat x, y as constants): 3

   Combining these results gives the Jacobian vector J:

$$J = \begin{bmatrix} 2xy & x^2 & 3 \end{bmatrix}$$

3. Geometric Interpretation

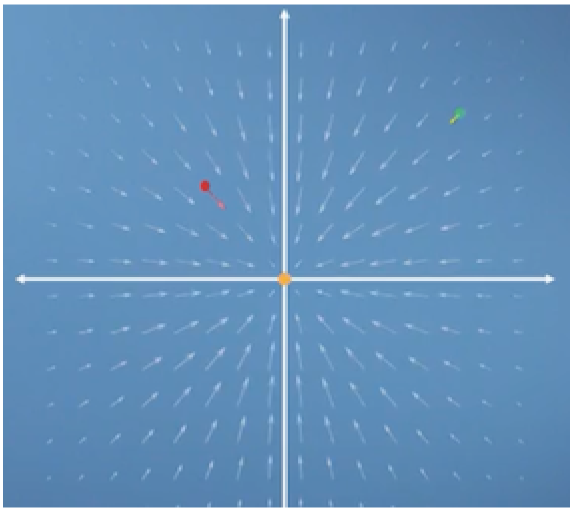
   The Jacobian vector is crucial because it provides the local properties of the function at any given point (x, y, z):

   - Direction of Steepest Slope: When evaluated at a specific point, the Jacobian vector points in the direction of the steepest uphill slope (or gradient) of the function.
   - Magnitude (Steepness): The magnitude (length) of the Jacobian vector at that point is equal to the steepness of the function at that location. The

    tighter the contour lines are packed, the larger the magnitude of the Jacobian.
  - Flat Regions: At the peaks, valleys, or flat plains of a function, the gradient is zero, and the magnitude of the Jacobian vector will be small (or zero).

- Understanding the Jacobian graphically (e.g., on a contour plot) allows us to visualize the gradient in 2D or 3D, providing the necessary intuition for tackling higher-dimensional problems later in the course.

4. Recap: The Jacobian Vector and its Meaning

- The Jacobian of a single function $f(x_1, x_2, ...)$ is a vector composed of its partial derivatives. It provides critical information about the function's local behavior:
  - Direction: The Jacobian vector, $\mathbf{J}$, points in the direction of the steepest uphill slope (the gradient).
  - Magnitude: The magnitude of $\mathbf{J}$ represents the steepness of the function at that specific point.
  - Example: $f(x, y) = e^{-(x^2+y^2)}$
    - The Jacobian points toward the origin $(0, 0)$.
    - $\mathbf{J} = [-2xe^{-(x^2+y^2)}, -2ye^{-(x^2+y^2)}]$
    - $\mathbf{J}(-1, 1) = [2e^{-2}, -2e^{-2}] = [0.27, -0.27]$
    - $\mathbf{J}(2, 2) = [-0.001, -0.001]$
    - At the origin, $\mathbf{J}(0, 0) = \mathbf{0}$, indicating the function is flat (a maximum, minimum, or saddle point). Visualizing the function confirms the origin is a maximum.
    - 

5. Extending to the Jacobian Matrix

- The Jacobian Matrix allows us to describe the rates of change for a vector-valued function—a function that takes a vector as an input and returns a vector as an output.

  If we have two output functions, u and v, that are both functions of x and y:

$$\mathbf{F}(x, y) = \begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix}$$

The Jacobian Matrix, $\mathbf{J}$, is formed by stacking the Jacobian vectors of the individual output functions (u and v) as rows:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix}$$

- Example: Linear Vector Field

- Consider the linear vector-valued function with components:

  - $u(x, y) = x - 2y$
  - $v(x, y) = 3y - 2x$
- The Jacobian matrix $\mathbf{J}$ is:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ -2 & 3 \end{bmatrix}$$

- Since u and v are linear functions, the partial derivatives (gradients) are constant everywhere. This Jacobian matrix represents the linear transformation from the (x, y) space to the (u, v) space.

- Application: Non-Linear Transformations

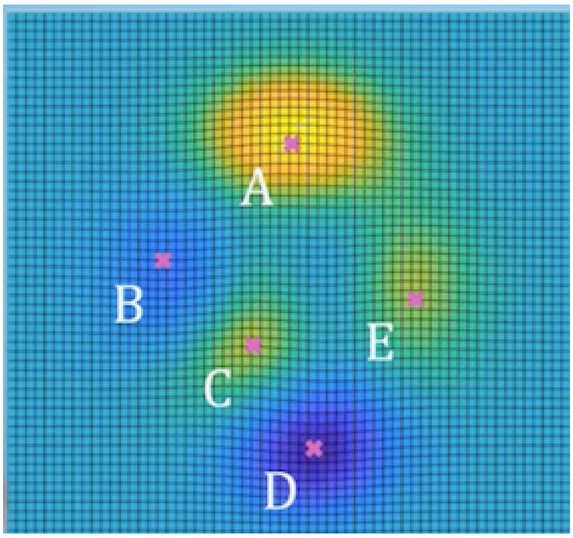
- For highly nonlinear functions, the Jacobian matrix is a crucial tool because:

  - Linear Approximation: If the function is smooth, we can zoom in close enough to consider a small region of space to be approximately linear. The Jacobian matrix provides the best linear approximation of the function's behavior at that point.
  - Transformation Scaling: The determinant of the Jacobian matrix, $|\mathbf{J}|$, quantifies how a small region of space changes in size (scales) when it is transformed by the function.
- A classic example is transforming between coordinate systems, such as Cartesian (x, y) and Polar (r, θ) coordinates, where the Jacobian determinant $\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = r$ is used to correctly calculate areas and volumes.

B4. Optimization

- Optimization in mathematics refers to the process of finding the input values that correspond to the maximum or minimum output values of a function. This is vital in fields like city planning, scheduling, and machine learning.
- The Jacobian serves as our primary tool in optimization, acting as a "torch" in the dark.

1. Optimization Targets

- Functions often have multiple locations with zero gradient ($J = \mathbf{0}$). These critical points are categorized based on their functional value across the entire domain:
  - 

  - Global Maximum: The single highest peak of the function (e.g., Point A).
  - Local Maxima: Peaks that are higher than their immediate surrounding area, but not the highest overall (e.g., Points C and E).
  - Global Minimum: The single deepest trough of the function (e.g., Point D).
  - Local Minima: Troughs that are lower than their immediate surrounding area, but not the lowest overall (e.g., Point B).

2. The Jacobian as a Guide (The "Night Walk" Analogy)

- The Jacobian vector is essential when we do not have an analytical expression for the function (the "night-time scenario," common when outputs come from complex simulations or experiments).
  - Guidance: The Jacobian vector always points uphill (in the direction of the steepest ascent).
  - Limitation: While the Jacobian points uphill, it does not guarantee the path leads to the Global Maximum. Following the local gradient often leads only to the nearest Local Maximum.
  - At Critical Points: At any maximum or minimum (local or global), the gradient is zero, meaning the Jacobian vector is the zero vector ($\mathbf{J} = \mathbf{0}$).

3. The Sandpit Analogy

- While the "night-time hill walk" helps visualize the local direction, a better analogy for the process is the sandpit with an uneven base exercise:
  - The Task: Find the deepest point (Global Minimum) of the pit by only measuring the depth at various points using a long stick.
  - The Constraint: You cannot see the overall landscape (no analytical expression/plot). You must rely on measurements at isolated points.
  - The Reality: Unlike walking, calculating $f(x)$ at x_1 and then jumping to x_2 has the same cost, regardless of the distance between them (no "walking"

time is incurred).

- This analogy highlights the central challenge in optimization: how to efficiently explore a high-dimensional space to find the global optimum without getting stuck at a local optimum.

B5. Hessian

The Hessian is the matrix of all second-order partial derivatives of a multivariate function. It is a fundamental tool used in optimization to determine the nature of critical points (maxima, minima, or saddle points).

1. Definition of the Hessian Matrix

- The Hessian Matrix, $\mathbf{H}$, is formed by collecting all possible second-order partial derivatives of a scalar-valued function f with n variables $(x_1, x_2, \ldots, x_n)$.
- For a function $f(x_1, x_2, \ldots, x_n)$, the Hessian is an $n \times n$ square matrix where the entry in the i-th row and j-th column is:

$$H_{ij} = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

- The Hessian Structure (for $n = 3$ variables x, y, z):

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

- Key Property: Symmetry
- If the function f is continuous (has no sudden step changes), the Hessian matrix will always be symmetrical across the leading diagonal. This means the mixed partial derivatives are equal:

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$$

2. Worked Example: Building the Hessian

- Consider the function $f(x, y, z) = x^2 yz$.
- Step 1: Find the Jacobian (First-Order Derivatives)

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \end{bmatrix} = \begin{bmatrix} 2xyz & x^2 z & x^2 y \end{bmatrix}$$

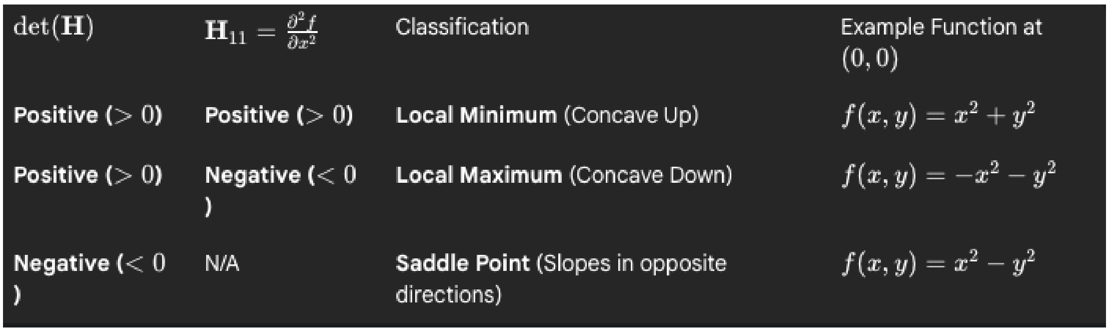
- Step 2: Differentiate the Jacobian terms againThe elements of the Jacobian become the rows of the Hessian:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial}{\partial x}(2xyz) & \frac{\partial}{\partial y}(2xyz) & \frac{\partial}{\partial z}(2xyz) \\ \frac{\partial}{\partial x}(x^2z) & \frac{\partial}{\partial y}(x^2z) & \frac{\partial}{\partial z}(x^2z) \\ \frac{\partial}{\partial x}(x^2y) & \frac{\partial}{\partial y}(x^2y) & \frac{\partial}{\partial z}(x^2y) \end{bmatrix} = \begin{bmatrix} 2yz & 2xz & 2xy \\ 2xz & 0 & x^2 \\ 2xy & x^2 & 0 \end{bmatrix}$$

(Note the symmetry: $H_{12} = H_{21}$, $H_{13} = H_{31}$, $H_{23} = H_{32}$)

3. Using the Hessian for OptimizationThe Hessian is essential for classifying critical points where the Jacobian (gradient) is zero ($\mathbf{J} = \mathbf{0}$).For a critical point, we use the Hessian to determine if it is a maximum, a minimum, or a saddle point. For simplicity, we can look at a 2D function, $f(x, y)$:

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



$\det(\mathbf{H})$	$\mathbf{H}_{11} = \frac{\partial^2 f}{\partial x^2}$	Classification	Example Function at $(0, 0)$
Positive (> 0)	**Positive** (> 0)	**Local Minimum** (Concave Up)	$f(x, y) = x^2 + y^2$
Positive (> 0)	**Negative** (< 0)	**Local Maximum** (Concave Down)	$f(x, y) = -x^2 - y^2$
Negative (< 0)	N/A	**Saddle Point** (Slopes in opposite directions)	$f(x, y) = x^2 - y^2$

B6. Real-World Challenges in Optimization

1. Real-World Challenges in Optimization

- The theoretical tools (Jacobian, Hessian) are powerful, but their application to real systems introduces several challenges:
  - High Dimensionality: Many problems, like training neural networks, involve hundreds or thousands of dimensions. We must trust the math and rely on our 2D/3D intuition to guide us, as we can no longer visualize the function as a landscape.
  - Unknown or Expensive Functions:
    - The function $f(x)$ may not have a nice analytical expression (a simple formula).
    - Calculating a single function value (a single "depth measurement" in the sandpit) might be computationally expensive (requiring supercomputer time or extensive lab work), making it impossible to fully plot the surface.
  - Non-Smoothness: Real-world functions may contain sharp features or discontinuities, making traditional differentiation rules invalid at those points.
  - Noisy Data: If the function measurements are noisy, the calculated Jacobian vectors can become highly unreliable unless carefully handled (e.g.,

averaging).

2. The Challenge of the Unknown Jacobian

- The central question for real-world optimization is: If we don't have an analytical expression for the function $f(x)$, how can we calculate the Jacobian (partial derivatives)?
- The answer lies in Numerical Methods, which provide approximate solutions when exact ones are intractable or unknown.
- The Finite Difference Method
  - The finite difference method takes us back to the original definition of the derivative (rise over run) over a finite interval. Since we cannot calculate the limit as the interval approaches zero, we use an approximation based on available function values.
  - To approximate the partial derivative $\frac{\partial f}{\partial x_i}$ at a starting location $\mathbf{x}$, we:
    - Take a small step size, Δx_i.
    - Calculate the function value at the starting point, $f(\mathbf{x})$.
    - Calculate the function value after taking a small step in the x_i direction, $f(\mathbf{x} + \Delta x_i)$.
  - The partial derivative is approximated by the slope over this finite interval:

$$\frac{\partial f}{\partial x_i} \approx \frac{f(\mathbf{x} + \Delta x_i) - f(\mathbf{x})}{\Delta x_i}$$

  - The Jacobian is then constructed by approximating each partial derivative in turn (taking a small step in x, then a small step in y, etc.).
- Practical Considerations for Step Size (Δx_i)
  - Choosing the right step size is a balance:
    - Too Big: A large step size leads to a poor approximation of the instantaneous slope (high approximation error).
    - Too Small: A too-small step size can lead to numerical issues. Computers store values with limited significant figures; if the change in the function value ($f(\mathbf{x} + \Delta x_i) - f(\mathbf{x})$) is smaller than the computer's precision, the calculated change might be zero, yielding an inaccurate derivative.
- Dealing with Noise
  - If the data is noisy, a simple approach to approximate the gradient robustly is to calculate the finite difference using several different step sizes and then average the resulting gradient approximations.