02: More Multivariate Chain Rule

Chapter Goal: To "level up" the Multivariate Chain Rule by extending it to longer chains and expressing it elegantly using the language of matrices and vectors.

---

1. A Note on Notation: Jacobian vs. Gradient

- **Recap of the Multivariate Chain Rule:** $\frac{df}{dt} = \nabla f \cdot \frac{d\vec{x}}{dt}$.
- **Observation:**
  - The Gradient ($\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \end{bmatrix}$) is typically written as a **column vector**.
  - In the Chain Rule formula, it is treated like a **row vector** for the dot product.
- **Formal Relationship:** The vector of partial derivatives ($\frac{\partial f}{\partial \vec{x}}$) is the **Transpose** of the Gradient.

$$\frac{\partial f}{\partial \vec{x}} = (\nabla f)^T$$

- **"Aha!" Moment:** The dot product $\vec{a} \cdot \vec{b}$ is computationally identical to the matrix multiplication $\vec{a}^T\vec{b}$.
  - $\nabla f \cdot \frac{d\vec{x}}{dt}$ is the same as $(\nabla f)^T \frac{d\vec{x}}{dt}$.
- **Conclusion:** This shows that the **Jacobian** (which is defined as a row vector for scalar functions) is the most convenient representation for the Chain Rule, as we can write it directly as a matrix multiplication.

$$\frac{df}{dt} = J_f \frac{d\vec{x}}{dt}$$

---

2. Extending the "Chain": More Than Two Links

- **Core Idea:** The Chain Rule works like a "domino effect" and can be extended as long as we need.
- **Univariate (1D) Example:**
  - We have a chain: $t \to u(t) \to x(u) \to f(x)$.
  - To find $\frac{df}{dt}$, we simply multiply all the derivatives along the chain.

$$\frac{df}{dt} = \frac{df}{dx} \cdot \frac{dx}{du} \cdot \frac{du}{dt}$$

- The Leibniz notation intuitively "cancels out": $\frac{df}{\cancel{dx}} \cdot \frac{\cancel{dx}}{\cancel{du}} \cdot \frac{\cancel{du}}{dt}$.

---

3. The Final "Aha!" Moment: The Chain Rule for Vectors and Matrices

- **The Most General (and Powerful) Scenario:**
  Imagine a chain where the intermediate variables are **vectors**.
  - **Input:** A scalar `t`.
  - $t \to \vec{u}(t)$ (Output is a vector $\vec{u}$).
  - $\vec{u} \to \vec{x}(\vec{u})$ (Input is vector $\vec{u}$, output is vector $\vec{x}$).
  - $\vec{x} \to f(\vec{x})$ (Input is vector $\vec{x}$, output is a scalar `f`).
- **Question:** How do we find $\frac{df}{dt}$?
- **Answer:** We use the same multiplication rule, but now the "things" we multiply are vectors and matrices.

$$\frac{df}{dt} = \frac{\partial f}{\partial \vec{x}} \frac{\partial \vec{x}}{\partial \vec{u}} \frac{d\vec{u}}{dt}$$

- **Let's dissect each part:**
  - $\frac{d\vec{u}}{dt}$**:** The derivative of vector $\vec{u}$ with respect to the scalar `t`.
    - The result is a **Column Vector** (`m x 1`).
  - $\frac{\partial \vec{x}}{\partial \vec{u}}$**:** The derivative of vector $\vec{x}$ with respect to vector $\vec{u}$.
    - This is the [Jacobian Matrix] we learned about. Each row is the gradient of a single component of `x`.
    - The result is a **Matrix** (`n x m`).
  - $\frac{\partial f}{\partial \vec{x}}$**:** The derivative of the scalar `f` with respect to the vector `x`.
    - This is the **Gradient** of `f` (written as a **Row Vector**).
    - The result is a **Row Vector** (`1 x n`).
- **The Matrix Multiplication:**

$$\frac{df}{dt} = (\text{Row Vector}_{1 \times n}) \cdot (\text{Matrix}_{n \times m}) \cdot (\text{Column Vector}_{m \times 1})$$

- This multiplication is **valid** (the inner dimensions match).
- The final result is a `(1 x 1)` matrix, which is a single number (a scalar). This is exactly what we expect for $\frac{df}{dt}$.

---

4. Final Conclusion

- The [Chain Rule] can be extended for longer chains.

- In the multivariate world, the Chain Rule becomes a **product of Gradient Vectors, Jacobian Matrices, and Derivative Vectors**.
- This shows how Linear Algebra (matrix multiplication) and Calculus (derivatives) merge perfectly to solve very complex problems. This is the mathematical foundation of **Backpropagation** in Neural Networks.

---

5. A Concrete Worked Example

A step-by-step example is the best way to see how all the parts (Gradient, Jacobian, Chain Rule) work together.

We will create a "chain" example similar to the one in the video, and then we will calculate $\frac{df}{dt}$ in two ways: the "brute force" way (direct substitution) and the "elegant" way (Multivariate Chain Rule).

Concrete Example Setup

Imagine a chain of relationships like this:

- **Initial Input:** A scalar `t`.
- **Function `u(t)` (from $\mathbb{R}^1 \to \mathbb{R}^2$):** `t` is transformed into a vector `u`.
  - $u_1 = t^2$
  - $u_2 = 3t$
  - So, $\vec{u}(t) = \begin{bmatrix} t^2 \\ 3t \end{bmatrix}$
- **Function `x(u)` (from $\mathbb{R}^2 \to \mathbb{R}^2$):** Vector `u` is transformed into vector `x`.
  - $x_1 = u_1 + 2u_2$
  - $x_2 = u_1 - u_2$
- **Function `f(x)` (from $\mathbb{R}^2 \to \mathbb{R}^1$):** Vector `x` is transformed into a scalar `f`.
  - $f = x_1 \cdot x_2$

Our Goal: Find $\frac{df}{dt}$ (How fast does `f` change with respect to `t`?).

Method #1: Direct Substitution ("Brute Force")

This is the method that does not use the Chain Rule. We substitute everything backwards until we get a single formula for `f` in terms of `t`.

1. **Substitute `x` into `f`:**
   $f = (u_1 + 2u_2) \cdot (u_1 - u_2)$
   $f = u_1^2 - u_1u_2 + 2u_2u_1 - 2u_2^2 = u_1^2 + u_1u_2 - 2u_2^2$
2. **Substitute `u` into `f`:**

- Replace u_1 with t^2.
- Replace u_2 with $3t$.
  $f(t) = (t^2)^2 + (t^2)(3t) - 2(3t)^2$
  $f(t) = t^4 + 3t^3 - 2(9t^2) = t^4 + 3t^3 - 18t^2$

3. **Find the derivative $\frac{df}{dt}$ (Now it's easy):**
   We now have a standard 1D function. Use the Power Rule.

$$\frac{df}{dt} = 4t^3 + 9t^2 - 36t$$

   This is the final answer we will use for comparison.

Method #2: The Multivariate Chain Rule ("Elegant")

Recipe: $\frac{df}{dt} = \frac{\partial f}{\partial \vec{x}} \frac{\partial \vec{x}}{\partial \vec{u}} \frac{d\vec{u}}{dt}$

Let's find each "ingredient" one by one.

- **Ingredient 1: $\frac{d\vec{u}}{dt}$ (Column Vector)**
  - Derivative of vector `u` with respect to scalar `t`.
  - $\vec{u}(t) = \begin{bmatrix} t^2 \\ 3t \end{bmatrix}$
  - $\frac{d\vec{u}}{dt} = \begin{bmatrix} \frac{d}{dt}(t^2) \\ \frac{d}{dt}(3t) \end{bmatrix} = \begin{bmatrix} 2t \\ 3 \end{bmatrix}$
- **Ingredient 2: $\frac{\partial \vec{x}}{\partial \vec{u}}$ (Jacobian Matrix)**
  - Derivative of vector `x` with respect to vector `u`.
  - $x_1 = u_1 + 2u_2$
  - $x_2 = u_1 - u_2$
  - Row 1 (derivatives of x_1): $[\frac{\partial x_1}{\partial u_1}, \frac{\partial x_1}{\partial u_2}] = [1, 2]$
  - Row 2 (derivatives of x_2): $[\frac{\partial x_2}{\partial u_1}, \frac{\partial x_2}{\partial u_2}] = [1, -1]$
  - $$\frac{\partial \vec{x}}{\partial \vec{u}} = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}$$
- **Ingredient 3: $\frac{\partial f}{\partial \vec{x}}$ (Gradient Row Vector)**
  - Derivative of scalar `f` with respect to vector `x`.
  - $f = x_1 x_2$
  - $\frac{\partial f}{\partial x_1} = x_2$
  - $\frac{\partial f}{\partial x_2} = x_1$
  - $$\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} x_2 & x_1 \end{bmatrix}$$

Time to Assemble!

$$\frac{df}{dt} = \begin{bmatrix} x_2 & x_1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2t \\ 3 \end{bmatrix}$$

Let's calculate from right to left.

- **Step A:** Calculate `(Jacobian Matrix) * (Column Vector)`

$$\begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2t \\ 3 \end{bmatrix} = \begin{bmatrix} (1 \cdot 2t + 2 \cdot 3) \\ (1 \cdot 2t + -1 \cdot 3) \end{bmatrix} = \begin{bmatrix} 2t + 6 \\ 2t - 3 \end{bmatrix}$$

- **Step B:** Calculate `(Row Vector) * (Result from Step A)`

$$\begin{bmatrix} x_2 & x_1 \end{bmatrix} \begin{bmatrix} 2t + 6 \\ 2t - 3 \end{bmatrix} = x_2(2t + 6) + x_1(2t - 3)$$

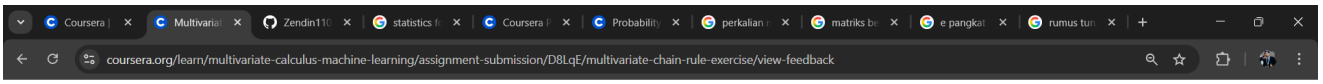
- **Step C:** Return everything to the language of `t`.
  - Our result is still in terms of x_1 and x_2. We need to substitute back.
    - $x_1 = u_1 + 2u_2 = t^2 + 2(3t) = t^2 + 6t$
    - $x_2 = u_1 - u_2 = t^2 - 3t$
  - Substitute into $x_2(2t + 6) + x_1(2t - 3)$:
    $= (t^2 - 3t)(2t + 6) + (t^2 + 6t)(2t - 3)$
  - Now, just plain algebra:
    $= (2t^3 + 6t^2 - 6t^2 - 18t) + (2t^3 - 3t^2 + 12t^2 - 18t)$
    $= (2t^3 - 18t) + (2t^3 + 9t^2 - 18t)$
    $= 4t^3 + 9t^2 - 36t$

Conclusion

Compare the two results:

- **Brute Force Method:** $4t^3 + 9t^2 - 36t$
- **Chain Rule Method:** $4t^3 + 9t^2 - 36t$

They are EXACTLY the same!
This proves that the Multivariate Chain Rule, while it may look complicated with all its matrices and vectors, really works and gives the correct result. For this simple example, the brute force method might seem faster. But for very long and complex chains (like in a Neural Network), the Chain Rule method is far more systematic and forms the basis of the Backpropagation algorithm



2. For the following functions, calculate the expression $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{d\mathbf{x}}{dt}$ in matrix form, where $\mathbf{x} = (x_1, x_2, x_3)$. 1 / 1 point

$f(\mathbf{x}) = f(x_1, x_2, x_3) = x_1^3 cos(x_2) e^{x_3}$

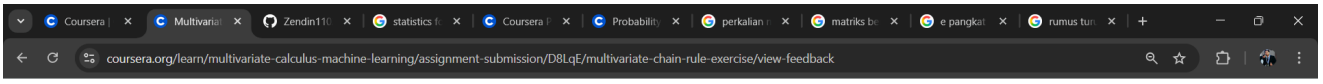
$x_1(t) = 2t$

$x_2(t) = 1 - t^2$

$x_3(t) = e^t$

- (●) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{d\mathbf{x}}{dt} = [3x_1^2 cos(x_2)e^{x_3}, -x_1^3 sin(x_2)e^{x_3}, x_1^3 cos(x_2)e^{x_3}] \begin{bmatrix} 2 \\ -2t \\ e^t \end{bmatrix}$
- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{d\mathbf{x}}{dt} = [3x_1^2 cos(x_2)e^{x_3}, -x_1^3 cos(x_2)e^{x_3}, x_1^3 cos(x_2)e^{x_3}] \begin{bmatrix} 2 \\ 2t \\ e^t \end{bmatrix}$
- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{d\mathbf{x}}{dt} = [3x_1^2 cos(x_2)e^{x_3}, x_1^3 cos(x_2)e^{x_3}, x_1^3 sin(x_2)e^{x_3}] \begin{bmatrix} 2 \\ 2t \\ -e^t \end{bmatrix}$
- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{d\mathbf{x}}{dt} = [3x_1^2 cos(x_2)e^{x_3}, -x_1^3 sin(x_2)e^{x_3}, x_1^3 sin(x_2)e^{x_3}] \begin{bmatrix} 2 \\ 2t \\ e^t \end{bmatrix}$

Correct
Well done!



← Back **Multivariate chain rule exercise**
Practice Assignment • 20 min

5. For the following functions, calculate the expression $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{u}} \frac{d\mathbf{u}}{dt}$ in matrix form, where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{u} = (u_1, u_2)$. 1 / 1 point

$f(\mathbf{x}) = f(x_1, x_2, x_3) = sin(x_1) cos(x_2) e^{x_3}$

$x_1(u_1, u_2) = sin(u_1) + cos(u_2)$

$x_2(u_1, u_2) = cos(u_1) - sin(u_2)$

$x_3(u_1, u_2) = e^{u_1 + u_2}$

$u_1(t) = 1 + t/2$

$u_2(t) = 1 - t/2$

- (●) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{u}} \frac{d\mathbf{u}}{dt} =$

$$[cos(x_1)cos(x_2)e^{x_3}, -sin(x_1)sin(x_2)e^{x_3}, sin(x_1)cos(x_2)e^{x_3}] \begin{bmatrix} cos(u_1) & -sin(u_2) \\ -sin(u_1) & -cos(u_2) \\ e^{u_1+u_2} & e^{u_1+u_2} \end{bmatrix} \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix}$$

- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{u}} \frac{d\mathbf{u}}{dt} =$

$$[cos(x_1)cos(x_2)e^{x_3}, -sin(x_1)cos(x_2)e^{x_3}, sin(x_1)cos(x_2)e^{x_3}] \begin{bmatrix} cos(u_1) & sin(u_2) \\ -sin(u_1) & -cos(u_2) \\ e^{u_1+u_2} & -e^{u_1+u_2} \end{bmatrix} \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix}$$

- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{u}} \frac{d\mathbf{u}}{dt} =$

$$[cos(x_1)cos(x_2)e^{x_3}, -sin(x_1)^2 sin(x_2)e^{x_3}, sin(x_1)cos(x_2)e^{x_3}] \begin{bmatrix} sin(u_1) & -sin(u_2) \\ -sin(u_1) & -cos(u_2) \\ 3e^{u_1+u_2} & e^{u_1+u_2} \end{bmatrix} \begin{bmatrix} -1/2 \\ -1/2 \end{bmatrix}$$

- ( ) $\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{u}} \frac{d\mathbf{u}}{dt} =$

$$[cos(x_1)cos(x_2)e^{x_3}, sin(x_1)sin(x_2)e^{x_3}, sin(x_1)cos(x_2)e^{x_3}] \begin{bmatrix} -cos(u_1) & -sin(u_2) \\ -sin(u_1) & -cos(u_2) \\ e^{u_1+u_2} & 2e^{u_1+u_2} \end{bmatrix} \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

Correct

Tags: #mml-specialization #multivariate-calculus #chain-rule #jacobian #gradient #backpropagation