Tutorial: Gradients, Jacobians & The Trace Trick

Duration: 90 Minutes

Instructor/TA Note

This tutorial is designed to be extremely granular. Students often get lost in the jump from scalar to matrix notation.

- **Part 1 (Scalar to Vector):** Focus on the mechanics of partial differentiation.
- **Part 2 (Vector to Vector):** Focus on **Dimensions**. Always ask: "What is the size of the input? What is the size of the output?"
- **Part 3 (Matrix to Scalar):** Focus on the **Trace Trick**. The goal is to manipulate the differential dL until it looks like $\text{tr}(\mathbf{G}^T d\mathbf{X})$.

---

1 Part 1: Basic Gradient Calculation (Direct Method)

Time Allocation: 20 Minutes

Problem 1: Gradient of a Quadratic Function

Let $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$. Consider the scalar function:

$$f(\mathbf{x}) = 3x_1^2 + 2x_1x_2 + x_2^2$$

1. Calculate the gradient vector $\nabla f(\mathbf{x})$.
2. Evaluate the gradient at the point $\mathbf{p} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

Detailed Solution:

Step 1: Understand the Goal The gradient $\nabla f(\mathbf{x})$ is a vector that collects all the partial derivatives. Since $\mathbf{x}$ has 2 components (x_1, x_2), the gradient will be a vector of size 2.

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$$

Step 2: Compute Partial Derivative w.r.t x_1 *Rule:* When differentiating with respect to x_1, treat x_2 as a constant number (like 5 or π).

$$\begin{aligned}
f(\mathbf{x}) &= \underbrace{3x_1^2}_{\text{Depends on } x_1} + \underbrace{2x_1x_2}_{\text{Linear in } x_1} + \underbrace{x_2^2}_{\text{Constant w.r.t } x_1} \\
\frac{\partial f}{\partial x_1} &= \frac{d}{dx_1}(3x_1^2) + \frac{d}{dx_1}(2x_1x_2) + \frac{d}{dx_1}(x_2^2) \\
&= 6x_1 + 2x_2(1) + 0 \\
&= 6x_1 + 2x_2
\end{aligned}$$

Step 3: Compute Partial Derivative w.r.t x_2 *Rule:* Now treat x_1 as a constant.

$$\begin{aligned}
f(\mathbf{x}) &= \underbrace{3x_1^2}_{\text{Constant}} + \underbrace{2x_1x_2}_{\text{Linear in } x_2} + \underbrace{x_2^2}_{\text{Depends on } x_2} \\
\frac{\partial f}{\partial x_2} &= 0 + 2x_1(1) + 2x_2 \\
&= 2x_1 + 2x_2
\end{aligned}$$

Step 4: Assemble the Vector Stack the results:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 6x_1 + 2x_2 \\ 2x_1 + 2x_2 \end{bmatrix}$$

Step 5: Numerical Evaluation Substitute $x_1 = 1$ and $x_2 = -1$:

$$\nabla f(1, -1) = \begin{bmatrix} 6(1) + 2(-1) \\ 2(1) + 2(-1) \end{bmatrix} = \begin{bmatrix} 6 - 2 \\ 2 - 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

Interpretation: At the point $(1, -1)$, the function increases most rapidly in the direction $(4, 0)$ (purely along the x-axis).

2 Part 2: Jacobians & Chain Rule

Time Allocation: 25 Minutes

Problem 2: The Affine Transformation

Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{m \times n}$, and $\mathbf{b} \in \mathbb{R}^m$. Define the function:

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{b}$$

1. What are the dimensions of the Jacobian matrix $\mathbf{J} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$?
2. Calculate $\mathbf{J}$ explicitly by analyzing the partial derivative $\frac{\partial y_i}{\partial x_j}$.

Detailed Solution:

1. Dimension Analysis (Crucial Step):

- **Input:** $\mathbf{x}$ is a vector of size $n \times 1$.
- **Output:** $\mathbf{y}$ is a vector of size $m \times 1$.
- **Definition:** The Jacobian $\mathbf{J}$ contains the derivative of *every* output component w.r.t *every* input component.
- **Conclusion:** Rows = Output Size (m), Columns = Input Size (n). $\mathbf{J}$ is $m \times n$.

2. Explicit Calculation: Let's look at the equation for just *one* element of the output, say y_i (the i-th row of $\mathbf{y}$).

$$y_i = (\text{Row } i \text{ of } \mathbf{W}) \cdot \mathbf{x} + b_i$$

Written as a sum:

$$y_i = \sum_{k=1}^{n} W_{ik} x_k + b_i$$

Now, calculate the partial derivative of y_i with respect to a specific input x_j:

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j} \left(W_{i1} x_1 + \cdots + W_{ij} x_j + \cdots + W_{in} x_n + b_i \right)$$

Logic Check:

- b_i is constant w.r.t x_j. Derivative is 0.
- For any $k \neq j$, $W_{ik} x_k$ is constant w.r.t x_j. Derivative is 0.
- The only term that survives is $W_{ij} x_j$.

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

Final Assembly: The entry at row i, column j of the Jacobian corresponds to W_{ij}. Therefore, the Jacobian matrix is exactly the weight matrix.

$$\mathbf{J} = \mathbf{W}$$

Problem 3: The Element-wise Activation

Let $\mathbf{h} \in \mathbb{R}^k$. Let $\mathbf{z} = \sigma(\mathbf{h})$, where σ is the sigmoid function applied **element-wise** (i.e., $z_i = \sigma(h_i)$).

Compute the Jacobian matrix $\frac{\partial \mathbf{z}}{\partial \mathbf{h}}$. Explain why this matrix is diagonal.

Detailed Solution:

1. Dimensions: Input size k, Output size k. Jacobian is $k \times k$.

2. The "Cross-Talk" Check: We need to calculate $\frac{\partial z_i}{\partial h_j}$. Ask yourself: "Does changing input h_j affect output z_i?"

- **Case A (Off-Diagonal, $i \neq j$):** Since the function is element-wise, z_1 depends ONLY on h_1. z_1 does NOT depend on h_2. Therefore, $\frac{\partial z_i}{\partial h_j} = 0$ for all $i \neq j$.
- **Case B (Diagonal, $i = j$):** Here, $z_i = \sigma(h_i)$. This is just a standard scalar derivative. $\frac{\partial z_i}{\partial h_i} = \sigma'(h_i)$.

3. Constructing the Matrix:

$$\mathbf{J} = \begin{bmatrix} \sigma'(h_1) & 0 & \dots \\ 0 & \sigma'(h_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

This is a diagonal matrix. In vector notation:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{h}} = \text{diag}(\sigma'(\mathbf{h}))$$

3 Part 3: Matrix Calculus using the Trace Trick

Time Allocation: 45 Minutes

The Trace Identification Theorem: If you can manipulate the differential of a scalar L into the form:

$$dL = \text{tr}(\mathbf{G}^T d\mathbf{X})$$

Then the gradient is:

$$\nabla_{\mathbf{X}} L = \mathbf{G}$$

Problem 4: Trace of a Linear Product

Let $L = \text{tr}(\mathbf{AX})$, where $\mathbf{A}$ and $\mathbf{X}$ are square matrices. Find $\nabla_{\mathbf{X}} L$.

Detailed Solution:

Step 1: Take the Differential Apply the operator d to the equation. The trace is a linear operator, so d moves inside.

$$dL = d(\text{tr}(\mathbf{AX})) = \text{tr}(d(\mathbf{AX}))$$

Step 2: Apply Matrix Rules $\mathbf{A}$ is a constant matrix, so it does not change ($d\mathbf{A} = 0$). $\mathbf{X}$ is the variable.

$$d(\mathbf{AX}) = \mathbf{A}(d\mathbf{X})$$

Substitute this back:

$$dL = \text{tr}(\mathbf{A}d\mathbf{X})$$

Step 3: Match the Identification Form We need the form $\text{tr}(\mathbf{G}^T d\mathbf{X})$. Currently we have $\text{tr}(\mathbf{A}d\mathbf{X})$. Set them equal to find $\mathbf{G}$:

$$\mathbf{G}^T = \mathbf{A}$$

Take the transpose of both sides:

$$\mathbf{G} = \mathbf{A}^T$$

Answer: $\nabla_{\mathbf{X}} L = \mathbf{A}^T$.

Problem 5: Trace of a Quadratic Product

Let $L = \text{tr}(\mathbf{X}^T\mathbf{AX})$, where $\mathbf{A}$ is a constant square matrix. Find $\nabla_{\mathbf{X}} L$.

Detailed Solution:

Step 1: Product Rule for Differentials Treat $\mathbf{X}^T$, $\mathbf{A}$, and $\mathbf{X}$ as three separate terms being multiplied. Rule: $d(UVW) = (dU)VW + U(dV)W + UV(dW)$. Since $\mathbf{A}$ is constant ($d\mathbf{A} = 0$), the middle term vanishes.

$$dL = \text{tr}\Big(\underbrace{(d\mathbf{X}^T)}_{\text{Diff first term}} \mathbf{AX} + \mathbf{X}^T\mathbf{A} \underbrace{(d\mathbf{X})}_{\text{Diff last term}} \Big)$$

Note that $d(\mathbf{X}^T) = (d\mathbf{X})^T$.

$$dL = \text{tr}((d\mathbf{X})^T\mathbf{AX} + \mathbf{X}^T\mathbf{A}d\mathbf{X})$$

Step 2: Linearity of Trace Split the trace of a sum into a sum of traces:

$$dL = \underbrace{\text{tr}((d\mathbf{X})^T\mathbf{AX})}_{\text{Term 1}} + \underbrace{\text{tr}(\mathbf{X}^T\mathbf{A}d\mathbf{X})}_{\text{Term 2}}$$

Step 3: The Transpose Trick (Crucial Step) We want both terms to have $d\mathbf{X}$ on the right side. Term 2 is already good. Term 1 has $(d\mathbf{X})^T$. *Identity:* $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{M}^T)$. Apply this to Term 1. Let $\mathbf{M} = (d\mathbf{X})^T\mathbf{A}\mathbf{X}$.

$$\mathbf{M}^T = (\mathbf{A}\mathbf{X})^T((d\mathbf{X})^T)^T = \mathbf{X}^T\mathbf{A}^T d\mathbf{X}$$

So, Term 1 becomes: $\text{tr}(\mathbf{X}^T\mathbf{A}^T d\mathbf{X})$.

Step 4: Combine Terms Now substitute the transformed Term 1 back into the equation:

$$dL = \text{tr}(\mathbf{X}^T\mathbf{A}^T d\mathbf{X}) + \text{tr}(\mathbf{X}^T\mathbf{A}d\mathbf{X})$$

Factor out the common parts ($\mathbf{X}^T$ at start, $d\mathbf{X}$ at end):

$$dL = \text{tr}(\mathbf{X}^T(\mathbf{A}^T + \mathbf{A})d\mathbf{X})$$

Step 5: Identify the Gradient Compare with $dL = \text{tr}(\mathbf{G}^T d\mathbf{X})$.

$$\mathbf{G}^T = \mathbf{X}^T(\mathbf{A}^T + \mathbf{A})$$

Take transpose of both sides (remember $(XY)^T = Y^T X^T$):

$$\mathbf{G} = (\mathbf{A}^T + \mathbf{A})^T(\mathbf{X}^T)^T$$

$$\mathbf{G} = (\mathbf{A} + \mathbf{A}^T)\mathbf{X}$$

Problem 6: Linear Regression (Normal Equation)

Let $L = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. Derive $\nabla_{\mathbf{x}} L$ and set to 0.

Detailed Solution:

Step 1: Convert Norm to Trace The squared Euclidean norm $\|\mathbf{v}\|^2$ is equivalent to dot product $\mathbf{v}^T\mathbf{v}$, which is equivalent to $\text{tr}(\mathbf{v}^T\mathbf{v})$. Let $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ (the residual vector).

$$L = \text{tr}(\mathbf{r}^T\mathbf{r})$$

Step 2: Differentiate w.r.t the Residual r From Problem 5 (with $\mathbf{A} = \mathbf{I}$), we know $d(\text{tr}(\mathbf{r}^T\mathbf{r})) = \text{tr}(2\mathbf{r}^T d\mathbf{r})$.

$$dL = \text{tr}(2\mathbf{r}^T d\mathbf{r})$$

Step 3: Find $d\mathbf{r}$ We need to relate changes in $\mathbf{r}$ to changes in $\mathbf{x}$.

$$\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$$

Apply differential:

$$d\mathbf{r} = d(\mathbf{A}\mathbf{x}) - d(\mathbf{b})$$

Since $\mathbf{b}$ is constant, $d\mathbf{b} = 0$.

$$d\mathbf{r} = \mathbf{A}d\mathbf{x}$$

Step 4: Substitute back into dL Replace $d\mathbf{r}$ in the equation from Step 2:

$$dL = \text{tr}(2\mathbf{r}^T(\mathbf{A}d\mathbf{x}))$$

Step 5: Rotate/Associate to isolate $d\mathbf{x}$ Use matrix associativity. We want $d\mathbf{x}$ isolated at the end.

$$dL = \text{tr}((2\mathbf{r}^T\mathbf{A})d\mathbf{x})$$

Step 6: Identify Gradient Match with $\mathrm{tr}(\mathbf{G}^T d\mathbf{x})$. Note that since $\mathbf{x}$ is a vector, $\mathbf{G}$ is a vector, so $\mathbf{G}^T$ is a row vector.

$$\mathbf{G}^T = 2\mathbf{r}^T \mathbf{A}$$

Take transpose:

$$\mathbf{G} = (2\mathbf{r}^T \mathbf{A})^T = 2\mathbf{A}^T \mathbf{r}$$

Step 7: Expand r Substitute $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ back in:

$$\nabla_{\mathbf{x}} L = 2\mathbf{A}^T (\mathbf{A}\mathbf{x} - \mathbf{b})$$