Slide 1: Title Slide

- **Title:** Deriving Gradient Descent for Logistic Regression
- **Subtitle:** A Step-by-Step Mathematical Explanation
- **Presenter:** [Your Name]
- **Date:** [Date]

Slide 2: Introduction to the Problem

- **Heading:** What are we trying to do?
- **Content:**
  - Explain the goal of logistic regression: to find a function that predicts a probability between 0 and 1.
  - Introduce the key components:
    - **Linear Equation (z):** $z = w^T x + b$
    - **Sigmoid Function ($\hat{y}$):** $\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$
    - **Cost Function (J):** Log Loss, which measures the error.
  - State the ultimate goal: Use gradient descent to find the parameters (w and b) that minimize the cost function J.

Slide 3: The Cost Function (Log Loss)

- **Heading:** The Cost Function: Measuring Error
- **Content:**
  - Present the formula for the Log Loss cost function for a single example:
    $$Cost(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1 - \hat{y}) & \text{if } y = 0 \end{cases}$$
  - Explain why this function is used instead of Mean Squared Error.
  - Show the combined formula for the total cost over all m training examples:
    $J(w, b) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})]$
  -

Slide 4: The General Gradient Descent Rule

- **Heading:** How We Update Parameters
- **Content:**
  - Explain the general concept of gradient descent: iteratively taking steps in the direction opposite to the gradient.
  - Show the general update rule: $parameter := parameter - \alpha \times \frac{\partial J}{\partial parameter}$
  - Define the terms:
    - **Parameter:** w_j or b.
    - **Learning Rate (α):** The step size.
    - **Gradient ($\frac{\partial J}{\partial parameter}$):** The slope of the cost function with respect to the parameter.
  - State the task: We need to derive the gradient for both w_j and b.

Slide 5: Derivation - Step 1 (The Chain Rule)

- **Heading:** The Chain Rule
- **Content:**
  - Explain that we will use the chain rule to break down the complex derivative.
  - Show the breakdown for a single weight (w_j):
    $$\frac{\partial L}{\partial w_j} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_j}$$
  - Explain what each term represents:
    - $\frac{\partial L}{\partial \hat{y}}$: How the cost changes with the prediction.
    - $\frac{\partial \hat{y}}{\partial z}$: How the prediction changes with the linear input.
    - $\frac{\partial z}{\partial w_j}$: How the linear input changes with the weight.

Slide 6: Derivation - Step 2 (Calculating the Derivatives)

- **Heading:** Calculating Each Piece
- **Content:**
  - **Derivative 1:** $\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$
    - Briefly show the steps or state the simplified result.
  - **Derivative 2:** $\frac{\partial \hat{y}}{\partial z} = \hat{y}(1-\hat{y})$
    - Explain that this is a key property of the sigmoid function.
  - **Derivative 3:** $\frac{\partial z}{\partial w_j} = x_j$ and $\frac{\partial z}{\partial b} = 1$
    - This is the simplest part of the derivation, based on the linear equation.

Slide 7: Derivation - Step 3 (Putting it all together)

- **Heading:** Combining the Pieces
- **Content:**
  - Show the substitution for the weight derivative:
    $\frac{\partial L}{\partial w_j} = \left(\frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \right) \times (\hat{y}(1 - \hat{y})) \times x_j$
  - Show the simplification:
    $\frac{\partial L}{\partial w_j} = (\hat{y} - y)x_j$
  - Show the substitution for the bias derivative:
    $\frac{\partial L}{\partial b} = \left(\frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \right) \times (\hat{y}(1 - \hat{y})) \times 1$
  - Show the simplification:
    $\frac{\partial L}{\partial b} = \hat{y} - y$

Slide 8: The Final Update Rules

- **Heading:** The Final Gradient Descent Rules
- **Content:**
  - Present the final update rules, averaged over all m examples.
  - **Weight Update Rule:**
    $w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}^{(i)} - y^{(i)}) x_j^{(i)}]$
  - **Bias Update Rule:**
    $b := b - \alpha \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}^{(i)} - y^{(i)})]$
  - Explain that these are the equations used in every iteration of training to improve the model.