Gradient of the Negative Log-Likelihood for Logistic Regression

This note derives, step by step, the gradient of the **negative log-likelihood** (binary cross-entropy loss) used in logistic regression.

Setup and Notation

We are given a dataset $\{(x_i, y_i)\}_{i=1}^n$, where:

- $x_i \in \mathbb{R}^d$ is the feature vector for observation i
- $y_i \in \{0, 1\}$ is the corresponding binary label
- $w \in \mathbb{R}^d$ is the weight vector
- $b \in \mathbb{R}$ is the bias (intercept)

Define the linear predictor and predicted probability:

$$z_i = w^\top x_i + b, \qquad p_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}}.$$

Negative Log-Likelihood

The Bernoulli log-likelihood for a single observation is

$$\log p(y_i \mid x_i) = y_i \log p_i + (1 - y_i) \log(1 - p_i).$$

The **negative log-likelihood (NLL)** over the dataset is

$$J(w, b) = -\sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big].$$

We now compute the gradient of $J(w, b)$ with respect to w and b.

Step 1: Derivative of the Sigmoid Function

The sigmoid function satisfies the identity

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z)).$$

Thus,

$$\frac{dp_i}{dz_i} = p_i(1 - p_i).$$

Step 2: Derivative of the Loss w.r.t. z_i

Consider the contribution of a single data point:

$$J_i = -\Big[y_i \log p_i + (1 - y_i) \log(1 - p_i)\Big].$$

First, differentiate with respect to p_i:

$$\frac{dJ_i}{dp_i} = -\left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i}\right).$$

Apply the chain rule:

$$\frac{dJ_i}{dz_i} = \frac{dJ_i}{dp_i} \cdot \frac{dp_i}{dz_i} = -\left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i}\right) p_i(1 - p_i).$$

Now simplify:

$$\frac{dJ_i}{dz_i} = -\big[y_i(1 - p_i) - (1 - y_i)p_i\big] = p_i - y_i.$$

Key result:

$$\boxed{\frac{dJ_i}{dz_i} = p_i - y_i.}$$

This cancellation is what makes logistic regression especially clean to optimize.

Step 3: Gradient with Respect to w

Recall that

$$z_i = w^\top x_i + b.$$

The gradient of z_i with respect to w is

$$\frac{\partial z_i}{\partial w} = x_i.$$

Applying the chain rule:

$$\nabla_w J = \sum_{i=1}^{n} \frac{dJ_i}{dz_i} \frac{\partial z_i}{\partial w} = \sum_{i=1}^{n} (p_i - y_i) x_i.$$

$$\boxed{\nabla_w J(w, b) = \sum_{i=1}^{n} (p_i - y_i) x_i.}$$

Step 4: Gradient with Respect to b

Since

$$\frac{\partial z_i}{\partial b} = 1,$$

we obtain

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{n} \frac{dJ_i}{dz_i} = \sum_{i=1}^{n} (p_i - y_i).$$

$$\boxed{\frac{\partial J(w,b)}{\partial b} = \sum_{i=1}^{n} (p_i - y_i).}$$

Vectorized Form

Let:

- $X \in \mathbb{R}^{n \times d}$ be the design matrix (rows $x_i^\top$)
- $y \in \mathbb{R}^n$ be the label vector
- $p = \sigma(Xw + b\mathbf{1})$

Then the gradients can be written compactly as

$$\boxed{\nabla_w J = X^\top (p - y), \qquad \frac{\partial J}{\partial b} = \mathbf{1}^\top (p - y).}$$

If the **mean** negative log-likelihood $\frac{1}{n} J$ is used instead, both gradients are scaled by $1/n$.

Interpretation

- The residual term $(p_i - y_i)$ measures the prediction error in probability space.
- The gradient is a weighted sum of feature vectors, where errors determine the direction and magnitude of updates.
- This form directly motivates gradient descent and stochastic gradient descent for logistic regression.