Logistic Regression: Complete Matrix Derivation

1 Problem Setup

We observe N labeled samples (x_i, y_i) with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We model the conditional probability of the positive class as

$$P(y_i = 1 \mid \mathbf{x}_i; \theta) = \hat{y}_i = \sigma(z_i), \qquad z_i = \mathbf{w}^T \mathbf{x}_i + w_0, \tag{1}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$. Our goal is to maximize the likelihood—or equivalently, minimize the **binary cross-entropy (BCE)**

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \right]. \tag{2}$$

2 Chain Rule Strategy

Our Goal: Find $\frac{\partial \mathcal{L}}{\partial \mathbf{w}}$ and $\frac{\partial \mathcal{L}}{\partial w_0}$ (or equivalently, $\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}}$ where $\tilde{\mathbf{w}}$ includes the bias).

Chain Rule Decomposition: The loss $\mathcal{L}$ depends on the weights $\mathbf{w}$ through the following chain:

$$\mathbf{w} \to \mathbf{z} \to \hat{\mathbf{y}} \to \mathcal{L} \tag{3}$$

Therefore, by the multivariate chain rule:

$$\boxed{\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}}} \tag{4}$$

Strategy: We will compute each piece systematically:

1. **Step 1:** Compute $\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}}$ — how the loss changes with predictions
2. **Step 2:** Compute $\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}}$ — how predictions change with linear combinations (sigmoid derivative)
3. **Step 3:** Compute $\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}}$ — how linear combinations change with weights (design matrix)
4. **Step 4:** Multiply them together using the chain rule

This systematic approach ensures all partial derivatives are dimensionally consistent and mathematically coherent.

3 Preliminaries

(A) Sigmoid function and its derivative (scalar proof)

The logistic sigmoid is $\sigma(z) = \frac{1}{1+e^{-z}}$.

Differentiate with respect to z using the **quotient rule**: For $f(z) = \frac{g(z)}{h(z)}$, we have $f'(z) = \frac{h(z) \cdot g'(z) - g(z) \cdot h'(z)}{[h(z)]^2}$ ('lo d-hi minus hi d-lo over lo-squared").

Here: $g(z) = 1$ and $h(z) = 1 + e^{-z}$, so $g'(z) = 0$ and $h'(z) = -e^{-z}$.

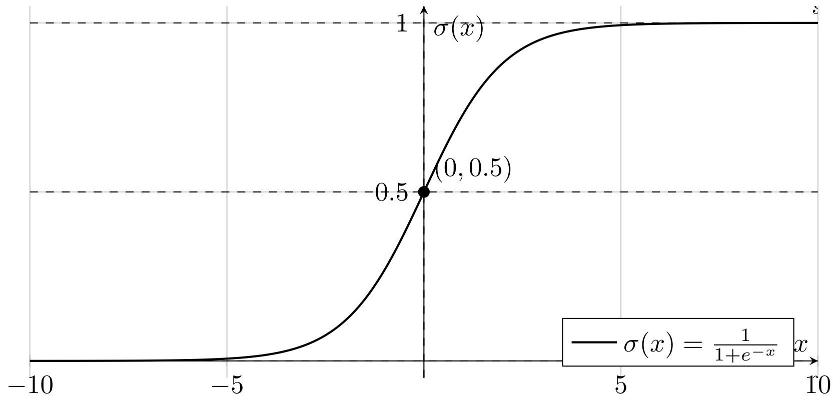

Figure 1: Sigmoid with inflection and asymptotes highlighted.

$$
\begin{aligned}
\frac{d\sigma}{dz} &= \frac{(1+e^{-z}) \cdot 0 - 1 \cdot (-e^{-z})}{(1+e^{-z})^2} \\
&= \frac{0 + e^{-z}}{(1+e^{-z})^2} \\
&= \frac{e^{-z}}{(1+e^{-z})^2}
\end{aligned}
$$

Multiply numerator and denominator by e^z:

$$
\begin{aligned}
&= \frac{e^{-z} \cdot e^z}{(1+e^{-z})^2 \cdot e^z} = \frac{1}{(1+e^{-z})(1+e^{-z}) \cdot e^z} \\
&= \frac{1}{(1+e^{-z})} \cdot \frac{1}{(1+e^{-z}) \cdot e^z} \\
&= \frac{1}{1+e^{-z}} \cdot \frac{1}{e^z + 1} \\
&= \frac{1}{1+e^{-z}} \cdot \frac{1}{1+e^z}
\end{aligned}
$$

Note that $\frac{1}{1+e^z} = \frac{e^{-z}}{e^{-z}(1+e^z)} = \frac{e^{-z}}{e^{-z}+1} = 1 - \frac{1}{1+e^{-z}}$:

$$
\begin{aligned}
&= \underbrace{\frac{1}{1+e^{-z}}}_{\sigma(z)} \cdot \left(1 - \underbrace{\frac{1}{1+e^{-z}}}_{\sigma(z)} \right) \\
&= \sigma(z)[1 - \sigma(z)]
\end{aligned}
$$

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

(B) Chain-rule ingredients (scalar)

For one sample, define the loss $L_i = -y_i \log \hat{y}_i - (1 - y_i)\log(1 - \hat{y}_i)$. We will need:

$$\frac{dL_i}{d\hat{y}_i} = -\frac{y_i}{\hat{y}_i} + \frac{1 - y_i}{1 - \hat{y}_i} \tag{5}$$

$$\frac{d\hat{y}_i}{dz_i} = \hat{y}_i(1 - \hat{y}_i) \tag{6}$$

4 Step-by-Step Derivative for Single Sample

Take one sample index i.

$$\begin{aligned}
\frac{dL_i}{dz_i} &= \frac{dL_i}{d\hat{y}_i} \cdot \frac{d\hat{y}_i}{dz_i} \\
&= \left(-\frac{y_i}{\hat{y}_i} + \frac{1 - y_i}{1 - \hat{y}_i}\right) \cdot \hat{y}_i(1 - \hat{y}_i) \\
&= -y_i(1 - \hat{y}_i) + (1 - y_i)\hat{y}_i \\
&= \hat{y}_i - y_i.
\end{aligned}$$

Hence the familiar scalar gradient: $\frac{dL_i}{dz_i} = \hat{y}_i - y_i$.

5 Matrix Notation Setup

(A) Augmented variables

Augment each feature vector with a bias coordinate:

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \in \mathbb{R}^{(d+1)\times 1}, \qquad \tilde{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^{(d+1)\times 1} \tag{7}$$

Design matrix (each row is one augmented sample):

$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ 1 & x_{31} & x_{32} & \cdots & x_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix} \in \mathbb{R}^{N\times(d+1)} \tag{8}$$

Compute the linear combinations:

$$\mathbf{z} = A\tilde{\mathbf{w}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ 1 & x_{31} & x_{32} & \cdots & x_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_N \end{bmatrix} \tag{9}$$

6 Vector Derivatives

(A) Derivative of $\mathbf{z}$ with respect to $\tilde{\mathbf{w}}$

Since $z_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id}$, we have:

$$\frac{\partial z_i}{\partial w_0} = 1, \quad \frac{\partial z_i}{\partial w_1} = x_{i1}, \quad \frac{\partial z_i}{\partial w_2} = x_{i2}, \quad \ldots, \quad \frac{\partial z_i}{\partial w_d} = x_{id} \tag{10}$$

Therefore:

$$\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}} = \frac{\partial}{\partial \tilde{\mathbf{w}}} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial w_0} & \frac{\partial z_1}{\partial w_1} & \frac{\partial z_1}{\partial w_2} & \cdots & \frac{\partial z_1}{\partial w_d} \\ \frac{\partial z_2}{\partial w_0} & \frac{\partial z_2}{\partial w_1} & \frac{\partial z_2}{\partial w_2} & \cdots & \frac{\partial z_2}{\partial w_d} \\ \frac{\partial z_3}{\partial w_0} & \frac{\partial z_3}{\partial w_1} & \frac{\partial z_3}{\partial w_2} & \cdots & \frac{\partial z_3}{\partial w_d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_N}{\partial w_0} & \frac{\partial z_N}{\partial w_1} & \frac{\partial z_N}{\partial w_2} & \cdots & \frac{\partial z_N}{\partial w_d} \end{bmatrix} = A \tag{11}$$

(B) Derivative of $\hat{\mathbf{y}}$ with respect to $\mathbf{z}$

Since $\hat{y}_i = \sigma(z_i)$ and $\frac{d\sigma}{dz} = \sigma(z)[1 - \sigma(z)]$:

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z_1} & \frac{\partial \hat{y}_1}{\partial z_2} & \cdots & \frac{\partial \hat{y}_1}{\partial z_N} \\ \frac{\partial \hat{y}_2}{\partial z_1} & \frac{\partial \hat{y}_2}{\partial z_2} & \cdots & \frac{\partial \hat{y}_2}{\partial z_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_N}{\partial z_1} & \frac{\partial \hat{y}_N}{\partial z_2} & \cdots & \frac{\partial \hat{y}_N}{\partial z_N} \end{bmatrix} \tag{12}$$

$$\hat{y}_i \;=\; \sigma(z_i), \qquad \sigma(t) = \frac{1}{1 + e^{-t}}$$

$$\frac{\partial \hat{y}_i}{\partial z_j} \;=\; 0 \qquad \text{whenever } i \neq j,$$

hence every off–diagonal entry of the Jacobian $\partial \hat{\mathbf{y}} / \partial \mathbf{z}$ is zero.

For the diagonal case $i = j$:

$$\frac{\partial \hat{y}_i}{\partial z_i} = \sigma'(z_i) = \sigma(z_i)\bigl(1 - \sigma(z_i)\bigr) = \hat{y}_i\bigl(1 - \hat{y}_i\bigr).$$

Since $\hat{y}_i$ only depends on z_i, this becomes a diagonal matrix:

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & 0 & 0 & \cdots & 0 \\ 0 & \hat{y}_2(1-\hat{y}_2) & 0 & \cdots & 0 \\ 0 & 0 & \hat{y}_3(1-\hat{y}_3) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{y}_N(1-\hat{y}_N) \end{bmatrix} \tag{13}$$

(C) Derivative of $\mathcal{L}$ with respect to $\hat{\mathbf{y}}$

The loss function is $\mathcal{L} = -\sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$.

$$\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \hat{y}_1} \\ \frac{\partial \mathcal{L}}{\partial \hat{y}_2} \\ \frac{\partial \mathcal{L}}{\partial \hat{y}_3} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \hat{y}_N} \end{bmatrix} = \begin{bmatrix} -\frac{y_1}{\hat{y}_1} + \frac{1-y_1}{1-\hat{y}_1} \\ -\frac{y_2}{\hat{y}_2} + \frac{1-y_2}{1-\hat{y}_2} \\ -\frac{y_3}{\hat{y}_3} + \frac{1-y_3}{1-\hat{y}_3} \\ \vdots \\ -\frac{y_N}{\hat{y}_N} + \frac{1-y_N}{1-\hat{y}_N} \end{bmatrix} \tag{14}$$

7 Chain Rule Application (Step 4)

Bringing it all together: Now we apply the chain rule formula with our three computed components.

Dimensional Analysis

Before multiplying, let's verify the dimensions work out:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}} \tag{15}$$

$$\mathbb{R}^{(d+1)\times 1} = \mathbb{R}^{1\times N} \cdot \mathbb{R}^{N\times N} \cdot \mathbb{R}^{N\times(d+1)} \tag{16}$$

Note: We need to transpose $\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}}$ and $\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}}$ to get the dimensions to work out correctly.

Step 4a: Intermediate step — $\frac{\partial \mathcal{L}}{\partial \mathbf{z}}$

First, let's compute how the loss changes with respect to the linear combinations:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}} = \left(\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}}\right)^T \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} \tag{17}$$

$$= \begin{bmatrix} -\frac{y_1}{\hat{y}_1} + \frac{1-y_1}{1-\hat{y}_1} & -\frac{y_2}{\hat{y}_2} + \frac{1-y_2}{1-\hat{y}_2} & \cdots & -\frac{y_N}{\hat{y}_N} + \frac{1-y_N}{1-\hat{y}_N} \end{bmatrix} \tag{18}$$

$$\cdot \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & 0 & \cdots & 0 \\ 0 & \hat{y}_2(1-\hat{y}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{y}_N(1-\hat{y}_N) \end{bmatrix} \tag{19}$$

Multiplying the i-th component (this is where the magic happens!):

$$\left(\frac{\partial \mathcal{L}}{\partial \mathbf{z}}\right)_i = \left(-\frac{y_i}{\hat{y}_i} + \frac{1-y_i}{1-\hat{y}_i}\right) \cdot \hat{y}_i(1-\hat{y}_i) \tag{20}$$

$$= -y_i(1-\hat{y}_i) + (1-y_i)\hat{y}_i \tag{21}$$

$$= -y_i + y_i\hat{y}_i + \hat{y}_i - y_i\hat{y}_i \tag{22}$$

$$= \hat{y}_i - y_i \tag{23}$$

Therefore:

$$\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{z}} = \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \hat{y}_3 - y_3 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix} = \hat{\mathbf{y}} - \mathbf{y}} \tag{24}$$

Step 4b: Final result — $\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}}$

Now we complete the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}} = \left(\frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}}\right)^T \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{z}} \tag{25}$$

$$= A^T \cdot (\hat{\mathbf{y}} - \mathbf{y}) \tag{26}$$

Explicitly writing out the matrix multiplication:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{N1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{N2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & x_{3d} & \cdots & x_{Nd} \end{bmatrix} \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \hat{y}_3 - y_3 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{N} (\hat{y}_i - y_i) \\ \sum_{i=1}^{N} (\hat{y}_i - y_i) x_{i1} \\ \sum_{i=1}^{N} (\hat{y}_i - y_i) x_{i2} \\ \vdots \\ \sum_{i=1}^{N} (\hat{y}_i - y_i) x_{id} \end{bmatrix}$$

This gives us the gradient components:

$$\frac{\partial \mathcal{L}}{\partial w_0} = \sum_{i=1}^{N} (\hat{y}_i - y_i) \tag{27}$$

$$\frac{\partial \mathcal{L}}{\partial w_j} = \sum_{i=1}^{N} (\hat{y}_i - y_i) x_{ij} \quad \text{for } j = 1, 2, \ldots, d \tag{28}$$

8 Hessian (Convexity Check)

Differentiate the gradient again:

$$H = \frac{\partial^2 \mathcal{L}}{\partial \tilde{\mathbf{w}} \partial \tilde{\mathbf{w}}^T} = A^T \mathrm{diag}[\hat{y}_1(1-\hat{y}_1), \hat{y}_2(1-\hat{y}_2), \ldots, \hat{y}_N(1-\hat{y}_N)] A \tag{29}$$

Explicitly:

$$H = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{Nd} \end{bmatrix} \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & 0 & \cdots & 0 \\ 0 & \hat{y}_2(1-\hat{y}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{y}_N(1-\hat{y}_N) \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix} \tag{30}$$

If A has full column rank, the Hessian is positive definite, proving strict convexity, so gradient-based solvers converge to the global optimum.

9 Why Gradient Descent? (No Closed-Form Solution)

Comparison with Linear Regression

Linear Regression: For ordinary least squares, we minimize:

$$\mathcal{L}_{\text{linear}}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 = \frac{1}{2} ||\mathbf{y} - A\mathbf{w}||^2 \tag{31}$$

This is a **quadratic function** in $\mathbf{w}$. Setting $\nabla_{\mathbf{w}} \mathcal{L}_{\text{linear}} = 0$ gives:

$$A^T A \mathbf{w} = A^T \mathbf{y} \quad \Rightarrow \quad \mathbf{w}^* = (A^T A)^{-1} A^T \mathbf{y} \tag{32}$$

This is the **normal equation** — a closed-form solution!

Why Logistic Regression is Different

Logistic Regression: Our loss function is:

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{N} \left[y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \right] \tag{33}$$

Problem: The sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ is a **transcendental function**. This means:

1. **No Polynomial Form:** Unlike linear regression where the loss is quadratic in $\mathbf{w}$, logistic regression involves exponentials and logarithms.
2. **No Closed-Form Solution:** Setting $\nabla_{\mathbf{w}} \mathcal{L} = 0$ gives us:

   $$A^T(\hat{\mathbf{y}} - \mathbf{y}) = \mathbf{0} \tag{34}$$

   But $\hat{\mathbf{y}} = \sigma(A\mathbf{w})$ involves the sigmoid function! This creates a **transcendental equation** that cannot be solved algebraically.
3. **Implicit Dependence:** The predictions $\hat{\mathbf{y}}$ depend on $\mathbf{w}$ through the sigmoid, making the equation $A^T(\sigma(A\mathbf{w}) - \mathbf{y}) = \mathbf{0}$ impossible to solve directly for $\mathbf{w}$.

Enter Gradient Descent

Since we can't solve directly, we use **iterative optimization**:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^{(t)}) \tag{35}$$

where α is the **learning rate** and our gradient is:

$$\nabla_{\mathbf{w}} \mathcal{L} = A^T(\hat{\mathbf{y}} - \mathbf{y}) \tag{36}$$

10 The Role of the Learning Rate α

What α Controls

The learning rate α determines **how big steps** we take in the direction of the negative gradient:

- **Large α:** Fast convergence, but risk of overshooting the minimum
- **Small α:** Stable convergence, but slow progress
- **Just right α:** Efficient convergence to the global optimum

Why We Need α (Step Size Control)

1. **Gradient gives direction:** $\nabla_{\mathbf{w}} \mathcal{L}$ tells us which way to move
2. **But not how far:** The magnitude of the gradient depends on the scale of our problem
3. **α provides scale:** It converts the gradient direction into an appropriate step size

Gradient Descent Algorithm for Logistic Regression

Initialize: $\mathbf{w}^{(0)}$ randomly
for $t = 0, 1, 2, \ldots$ until convergence **do**
  $\mathbf{z}^{(t)} \leftarrow A\mathbf{w}^{(t)}$
  $\hat{\mathbf{y}}^{(t)} \leftarrow \sigma(\mathbf{z}^{(t)})$ ▷ Apply sigmoid element-wise
  $\nabla\mathcal{L}^{(t)} \leftarrow A^T(\hat{\mathbf{y}}^{(t)} - \mathbf{y})$ ▷ Compute gradient
  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \alpha\nabla\mathcal{L}^{(t)}$ ▷ Update weights
end for

Convexity Saves Us

Good News: Even though we can't solve analytically, the logistic regression loss is **strictly convex** (as shown by our positive definite Hessian). This guarantees:

- There is exactly one global minimum
- Gradient descent will find it (with appropriate α)
- No local minima to get stuck in

$$\tilde{\mathbf{w}}^{(t+1)} = \tilde{\mathbf{w}}^{(t)} - H^{-1}g \tag{37}$$

Because H has the weighted least squares form:

$$H = A^T \operatorname{diag}[\hat{y}_1(1-\hat{y}_1), \hat{y}_2(1-\hat{y}_2), \ldots, \hat{y}_N(1-\hat{y}_N)]A \tag{38}$$

Each Newton iteration solves a linear system akin to ordinary least squares but with weights $\hat{y}_i(1-\hat{y}_i)$. IRLS typically converges faster than gradient descent but requires computing the Hessian inverse at each step.

11 Dimension Table

Quantity	Dimension
A	$\mathbb{R}^{N\times(d+1)}$
$\hat{\mathbf{y}}, \mathbf{y}, \mathbf{z}$	$\mathbb{R}^{N\times 1}$
$\hat{\mathbf{y}} - \mathbf{y}$	$\mathbb{R}^{N\times 1}$
$A^T(\hat{\mathbf{y}} - \mathbf{y})$	$\mathbb{R}^{(d+1)\times 1}$
H	$\mathbb{R}^{(d+1)\times(d+1)}$

12 Key Takeaways

1. **Chain Rule Success:** The systematic application of $\frac{\partial\mathcal{L}}{\partial\tilde{\mathbf{w}}} = \frac{\partial\mathcal{L}}{\partial\hat{\mathbf{y}}} \cdot \frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{z}} \cdot \frac{\partial\mathbf{z}}{\partial\tilde{\mathbf{w}}}$ gives us the clean result $\nabla_{\tilde{\mathbf{w}}}\mathcal{L} = A^T(\hat{\mathbf{y}} - \mathbf{y})$.
2. **Dimensional Coherence:** Each step preserves dimensional consistency:
$$\frac{\partial\mathcal{L}}{\partial\hat{\mathbf{y}}} \in \mathbb{R}^{N\times 1}, \quad \frac{\partial\hat{\mathbf{y}}}{\partial\mathbf{z}} \in \mathbb{R}^{N\times N}, \quad \frac{\partial\mathbf{z}}{\partial\tilde{\mathbf{w}}} \in \mathbb{R}^{N\times(d+1)}$$
$$\Rightarrow \frac{\partial\mathcal{L}}{\partial\tilde{\mathbf{w}}} \in \mathbb{R}^{(d+1)\times 1}$$
3. **Sigmoid Magic:** The sigmoid derivative $\sigma'(z) = \sigma(z)[1-\sigma(z)]$ creates perfect cancellation, reducing the complex chain rule to simply $\hat{y}_i - y_i$ for each sample.

4. **Matrix Structure:** The diagonal structure of $\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}}$ reflects the independence of predictions—each $\hat{y}_i$ depends only on its corresponding z_i.

5. **No Closed-Form Solution:** Unlike linear regression, the transcendental nature of the sigmoid function prevents an analytical solution, necessitating iterative optimization methods.

6. **Gradient Descent Necessity:** The learning rate α controls step size in the optimization, balancing convergence speed with stability.

7. **Convexity Guarantee:** BCE plus sigmoid gives a strictly convex objective, ensuring gradient descent converges to the unique global optimum.

8. **Geometric Interpretation:** The gradient $A^T(\hat{\mathbf{y}}-\mathbf{y})$ shows that we move in the direction that reduces the prediction errors, weighted by the feature values.

13 Summary of Chain Rule Application

$$\boxed{\begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{w}}} &= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \cdot \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \tilde{\mathbf{w}}} \\ &= A^T(\hat{\mathbf{y}} - \mathbf{y}) \end{aligned}} \tag{39}$$