MNIST MLP: Detailed Math, Design Choices, and Alternatives

Model Summary

Input is a flattened MNIST image $x \in \mathbb{R}^{784\times 1}$. The network is a 3-layer MLP with two hidden layers of width 16 and a residual connection:

$$784 \to 16 \to 16 \to 10$$

Notation

We denote the input column vector by x, the weight matrices by W_0, W_1, W_2, and the bias vectors by b_0, b_1, b_2. The pre-activation values are written as z_0, z_1, z_2, while the post-activation hidden outputs are a_0, a_1. The softmax probabilities are p, and the one-hot target label is y.

Forward Pass (Exact Math)

Input

The input image $X \in \mathbb{R}^{28\times 28}$ is flattened:

$$x = \mathrm{vec}(X) \in \mathbb{R}^{784\times 1}$$

Hidden Layer 1

$$W_0 \in \mathbb{R}^{16\times 784}, \quad b_0 \in \mathbb{R}^{16\times 1}$$

$$z_0 = W_0 x + b_0$$

$$a_0 = \mathrm{ReLU}(z_0)$$

where

$$\mathrm{ReLU}(t) = \max(0, t) \quad \text{(element-wise)}$$

Hidden Layer 2 With Residual

$$W_1 \in \mathbb{R}^{16\times 16}, \quad b_1 \in \mathbb{R}^{16\times 1}$$

$$z_1 = W_1 a_0 + b_1$$

$$h_1 = \mathrm{ReLU}(z_1)$$

$$a_1 = a_0 + h_1$$

The residual term a_0 helps preserve early features and stabilizes optimization.

Output Layer

$$W_2 \in \mathbb{R}^{10\times 16}, \quad b_2 \in \mathbb{R}^{10\times 1}$$

$$z_2 = W_2 a_1 + b_2$$

$$p_i = \frac{e^{z_{2,i}}}{\sum_{j=1}^{10} e^{z_{2,j}}}$$

The vector $p \in \mathbb{R}^{10\times 1}$ is a valid probability distribution:

$$\sum_{i=1}^{10} p_i = 1$$

Loss (Cross Entropy)

Let $y \in \mathbb{R}^{10\times 1}$ be a one-hot label. The element-wise loss is:

$$L_i = -y_i \log(p_i)$$

Total loss is the sum:

$$\mathcal{L} = \sum_{i=1}^{10} L_i$$

Backpropagation (High-Level)

The model computes gradients for each parameter:

$$\nabla W_0, \nabla b_0, \nabla W_1, \nabla b_1, \nabla W_2, \nabla b_2$$

Then applies SGD updates:

$$W \leftarrow W - \eta \nabla W, \quad b \leftarrow b - \eta \nabla b$$

Initialization Used (And Why)

This model uses Xavier/Glorot uniform initialization for each weight matrix:

$$W \sim \mathcal{U}(-\alpha, \alpha), \quad \alpha = \sqrt{\frac{6}{\text{fan\_in} + \text{fan\_out}}}$$

This keeps the variance of activations roughly stable for symmetric activations.

Alternative For ReLU

For ReLU, a common alternative is He/Kaiming initialization:

$$W \sim \mathcal{U}\Big(-\sqrt{\frac{6}{\text{fan\_in}}}, \sqrt{\frac{6}{\text{fan\_in}}}\Big)$$

Or normal:

$$W \sim \mathcal{N}\Big(0, \frac{2}{\text{fan\_in}}\Big)$$

He initialization accounts for ReLU zeroing negative activations, preserving variance.

Why This Architecture?

Why 16 Hidden Units?

This width is small enough to train quickly on CPU while still large enough to learn basic digit features. It also makes the model easy to inspect and reason about during experimentation.

Why Two Hidden Layers?

Two layers can represent more complex feature hierarchies than a single layer. The first layer learns basic edges and strokes, while the second layer combines them into digit-like patterns.

Why Residual Add?

The residual connection $a_1 = a_0 + h_1$ preserves earlier features, improves gradient flow, and reduces the chance of vanishing gradients in deeper stacks.

Design Alternatives

Wider layers such as 32, 64, or 128 units increase capacity at the cost of more computation. Deeper MLPs with batch normalization can improve stability but add complexity. Convolutional networks exploit spatial structure and often outperform MLPs on images. Alternative activations like GELU, Leaky ReLU, or ELU can change optimization behavior. Optimizers such as Adam or RMSProp often converge faster than plain SGD.

What The Parameters Represent

Each row of W_0 is a learned template over 784 pixels. Over training, those rows often resemble stroke detectors. Biases b_0 shift each neuron's activation threshold.

End-to-End Transformation

$$x \in \mathbb{R}^{784\times1} \xrightarrow{W_0,b_0,\text{ReLU}} a_0 \in \mathbb{R}^{16\times1} \xrightarrow{W_1,b_1,\text{ReLU},\text{residual}} a_1 \in \mathbb{R}^{16\times1} \xrightarrow{W_2,b_2,\text{softmax}} p \in \mathbb{R}^{10\times1}$$