

CS3317 Artificial Intelligence

Coursework 2: Logistic Regression

[Example Report — For Reference Only]

Name: Zhang San Student ID: 522030XXXXX

April 15, 2026

Note: This is an example report showing the expected structure and depth of analysis. All numerical results shown as `--` are placeholders—replace them with your own experimental results. Do not copy the analysis verbatim; write your own observations based on your experiments.

1 Implementation Overview

Softmax Function

I implemented the softmax function with numerical stability by subtracting the maximum logit from each row before computing the exponential:

$$\text{softmax}(z_i) = \frac{e^{z_i - \max(\mathbf{z})}}{\sum_j e^{z_j - \max(\mathbf{z})}} \quad (1)$$

This prevents overflow when logit values are large while producing mathematically equivalent results.

[Briefly explain why numerical stability matters and how your implementation handles it.]

Loss Functions

Cross-Entropy Loss. The standard loss for multi-class classification:

$$L = -\frac{1}{N} \sum_{i=1}^N \log p_{i,y_i} \quad (2)$$

where p_{i,y_i} is the softmax probability of the correct class.

[For each loss function, provide: (1) the mathematical definition, (2) the gradient formula, and (3) any implementation considerations such as numerical stability tricks.]

Hinge Loss. *[Write the Crammer–Singer formulation and explain the sparse gradient.]*

Exponential Loss. *[Write the formula and explain how you handled numerical issues.]*

Squared Loss. *[Write the formula and note the softmax Jacobian in the gradient.]*

Gradient Descent Optimizer

The optimizer computes parameter gradients via the chain rule:

$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^\top \frac{\partial L}{\partial \mathbf{z}}, \quad \frac{\partial L}{\partial \mathbf{b}} = \sum_i \frac{\partial L}{\partial \mathbf{z}_i} \quad (3)$$

Then updates: $\mathbf{W} \leftarrow \mathbf{W} - \eta \cdot \frac{\partial L}{\partial \mathbf{W}}$.

[Mention whether your gradient checks passed and the relative error magnitude.]

2 Loss Function Comparison

2.1 Training Curves

Figure ?? shows the training loss and validation accuracy for all four loss functions.

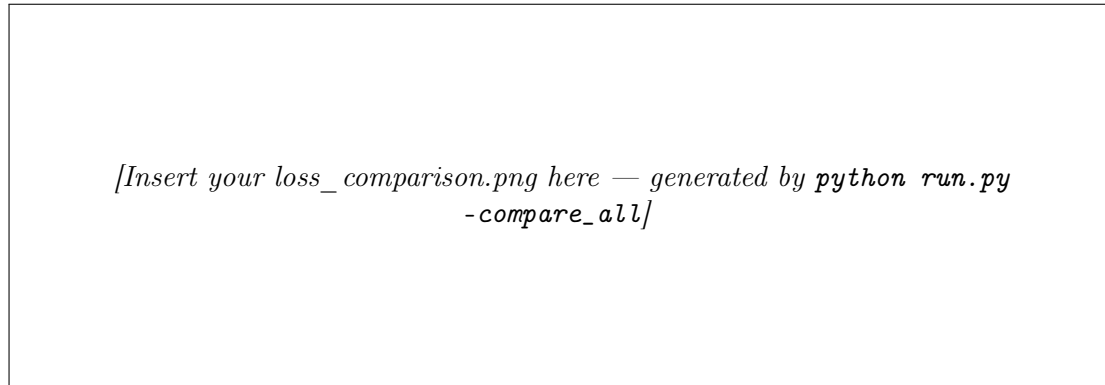


Figure 1: Training loss (left) and validation accuracy (right) for all four loss functions on FashionMNIST.

2.2 Results Summary

Table 1: Test accuracy and convergence comparison across loss functions.

Loss Function	Test Accuracy	Convergence (epoch)	Stability
Cross-Entropy	__._ %	~ __	_____
Hinge	__._ %	~ __	_____
Exponential	__._ %	~ __	_____
Squared	__._ %	~ __	_____

2.3 Analysis

[For each loss function, discuss the following based on your experimental results:]

Cross-Entropy. [Discuss its accuracy ranking and convergence speed. Why does its gradient $(\mathbf{p} - \mathbf{y})$ lead to smooth, well-scaled updates?]

Hinge Loss. [Discuss the effect of sparse gradients on convergence. How does the margin-based formulation differ from probability-based losses?]

Exponential Loss. [Discuss training stability. What happens when p_y is near zero? How does this affect gradient magnitude?]

Squared Loss. [Discuss convergence behavior. What is the “plateau effect” caused by softmax saturation, and how does it affect early training?]

3 Visualization

3.1 Learned Weight Templates

[Describe the visual patterns you observe in the weight templates. What do the positive and negative weights correspond to for specific classes? Why do the templates function as “matched filters”?]

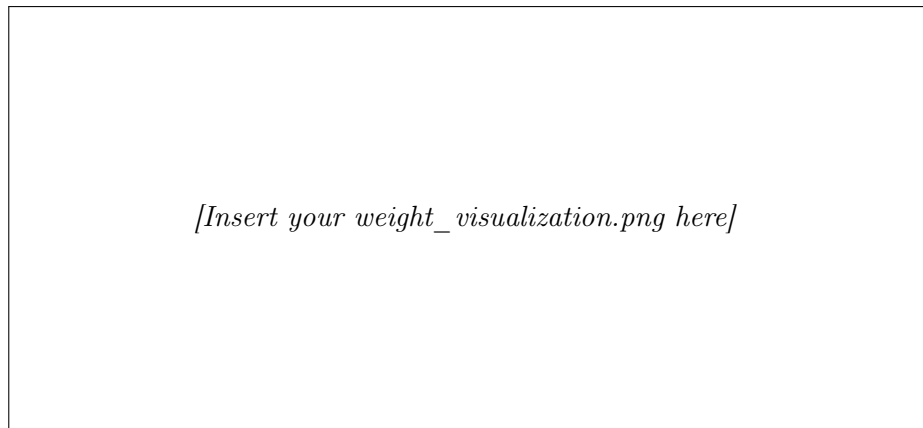


Figure 2: Learned weight templates for each of the 10 classes (cross-entropy loss).

3.2 Confusion Matrix

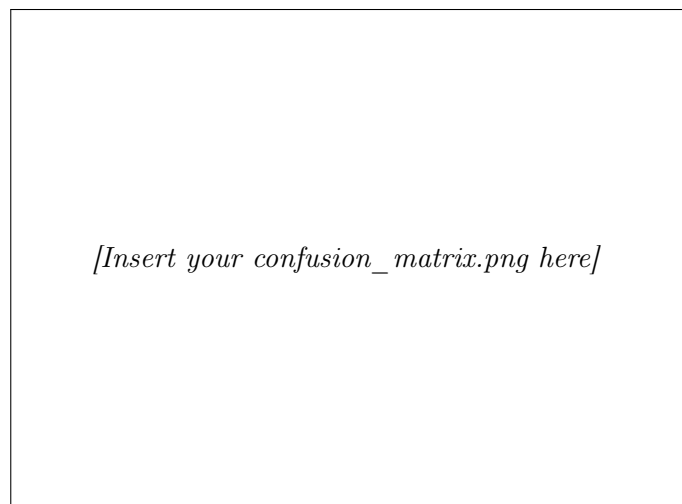


Figure 3: Confusion matrix on the test set (cross-entropy loss).

[Identify the most commonly confused class pairs. Explain why certain classes are hard to distinguish using only flattened pixel vectors. Which classes achieve the highest individual accuracy and why?]

4 Conclusion

[Summarize 2–3 key findings from your experiments. For example:]

- 1. [Which loss function works best for classification and why?]*
- 2. [How do different loss functions exhibit different convergence and stability behaviors?]*
- 3. [What are the fundamental limitations of logistic regression on image data, and what might address them?]*