1 Question

Objective: Train a linear classifier on MNIST using softmax loss and L2 regularization.

Best Performance: Achieved 92.22% test accuracy with optimal hyperparameters (LR = 0.001, $\lambda = 0.001$, 50 epochs).

Key Findings:

- **Learning Rate:** Critical for stability; lower rates (0.001) performed best, while higher rates (0.1) caused divergence.
- **Regularization:** Moderate λ values (0.001–0.01) improved generalization; $\lambda = 1$ led to underfitting ($\sim$ 88% accuracy).

Model Comparison Summary

Model	Training Accuracy	Testing Accuracy	Validation Accuracy
Logistic Regression	94%	92%	92%
Single-Layer Network	93.93% (epoch 20)	91.76%	92.93%

Maximum Accuracy Class: 1, Accuracy: 0.9806
Minimum Accuracy Class: 8, Accuracy: 0.7156

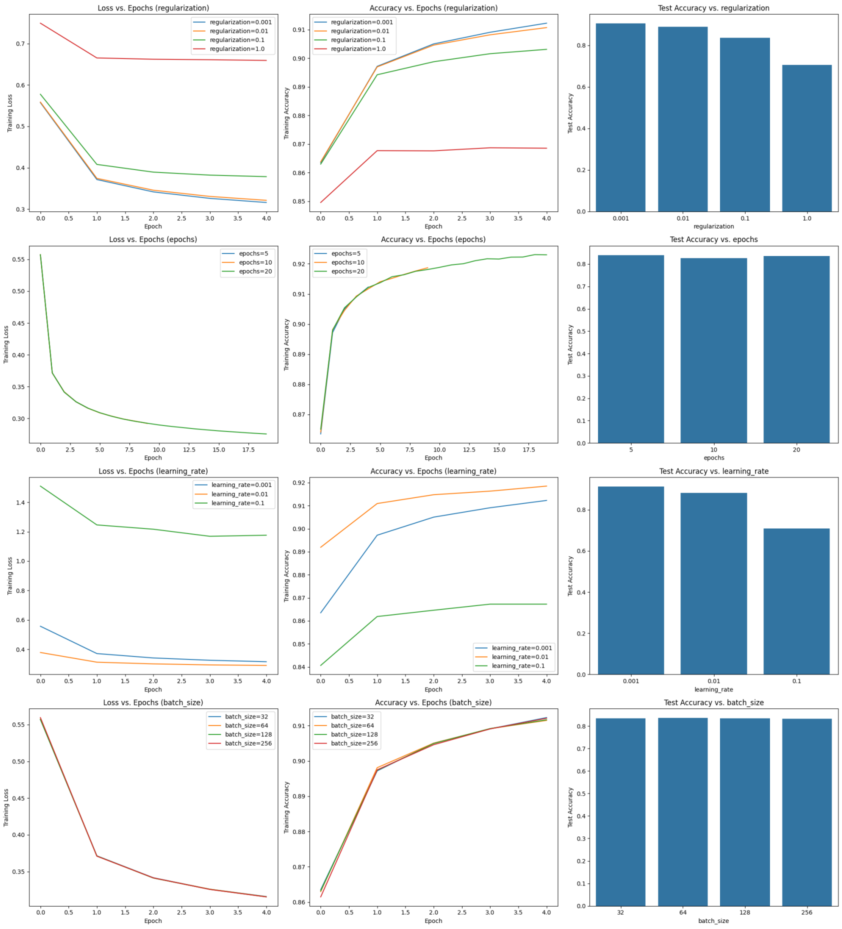

Figure 1: Parameter Tuning

2 Question

x	y	z	$\frac{\partial}{\partial x}$ (dx)	$\frac{\partial}{\partial y}$ (dy)	$\frac{\partial}{\partial z}$ (dz)
2	4	1	$2.1417102890343505 \times 10^{-16}$	0.43720979194276516	-0.3069722756588883
9	14	3	$-2.5313688314302287 \times 10^{-15}$	-1.148344174369598	$2.672161875217988 \times 10^{-7}$
128	42	666	$1.3244281103803983 \times 10^{-14}$	-0.42245219681098717	-0.0
52	14	28	$-6.6263243875773434 \times 10^{-15}$	-0.42245219681098645	-0.0

Table 1: Gradients at different points.

3 Question

Best Model

- **Parameters:** `batch_size=32`, `learning_rate=0.01`
- **Test Accuracy: 96.92%** (near SOTA for simple MLPs)
- **Validation Accuracy:** Peaked at **97.88%** (Epoch 5)

Hyperparameter Insights

- **Optimal LR:** 0.01 (fast convergence)
- **High LR (0.1):** Failed (approximately 10% accuracy, divergence)
- **Batch Sizes:** Smaller batches (32) outperformed larger ones

Comparison

Model	Test Accuracy
Logistic Regression	92%
Single-Layer	91%
Two-Layer MLP	**96.92%**

Table 2: Model Comparison on MNIST Dataset

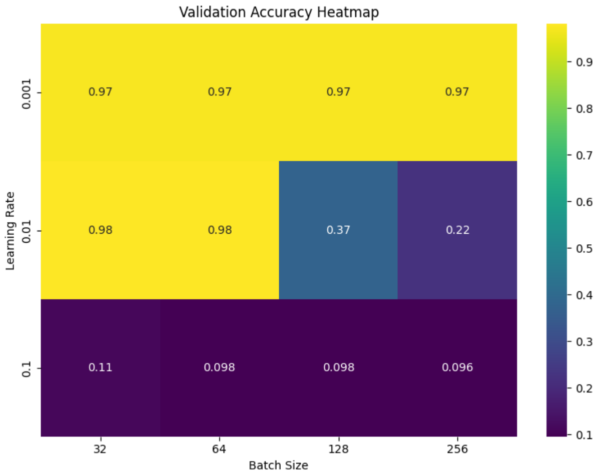

Figure 2: Parameter Tuning

4 Question

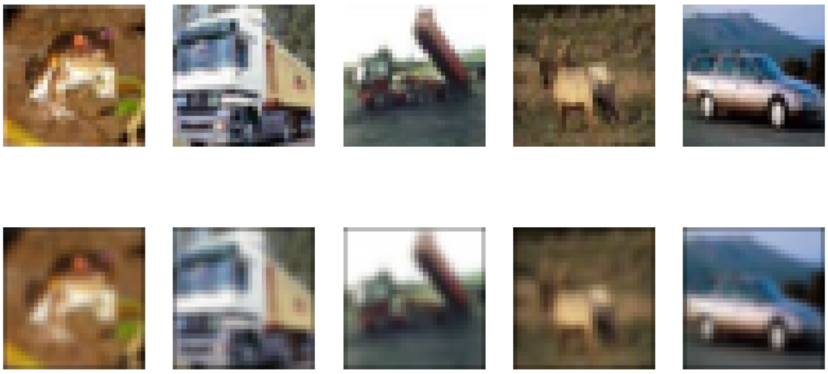

Figure 3: Before and After convolution

5 Question

5.1 1. Effect of Increasing Stride (padding=0, kernel=5)

- Stride = 1 → L2 = 2.9391
- Stride = 2 → L2 = 2.9967
- Stride = 3 → L2 = 3.0114

As the stride increases from 1 to 3, the L2 distance slightly increases. A larger stride results in fewer sampled positions on the feature map (i.e., more downsampling), causing the outputs to deviate more from the baseline C_0.

5.2 2. Effect of Padding=2 vs. Padding=0 (stride=1, kernel=5)

- Padding = 2 → L2 = 2.8642
- Padding = 0 → L2 = 2.9391

When more padding is applied (padding = 2), the L2 distance decreases. Additional padding helps preserve spatial resolution at the edges, making the filtered outputs closer to C_0 in the L2 sense. In contrast, setting padding = 0 cuts off border regions, slightly altering the receptive field and increasing the distance from the baseline.