**HW\_1\_1: Simulate a Function: Cos 2∏x\*x3**

I have used three neural networks of varying depth, but comparable parameter counts:

1. **Model 1 (Shallow):** 1 hidden layer (249 units), 748 parameters.
2. **Model 2 (Deep):** 3 hidden layers (18, 20, 15 units), 747 parameters.
3. **Model 3 (Deeper):** 5 hidden layers (5, 10, 20, 15, 8 units), 742 parameters.

A graph of a training model

AI-generated content may be incorrect.

Figure 1: Training loss vs epochs for shallow, 3-layer, and 5-layer models*.*

A graph with red and green lines

AI-generated content may be incorrect.

Figure 2: Predicted curves of all models compared with the ground-truth function.

**Results**

* **Training Loss:**

The shallow model converged more slowly and required significantly more epochs to reduce the loss, while deeper models achieved faster convergence. Model 3 (5 hidden layers) reached the lowest training loss the fastest.

* **Function Approximation:**

The shallow model (Model 1) and the deepest model (Model 3) captured the target function curve closely, matching oscillations and amplitude. The deep model (Model 2) with 3 hidden layers failed to approximate the function properly (particularly when the input values ranged from -2.2 to +2.2), showing almost flat predictions in this region.

**Comments:**

Depth improves representation power and optimization efficiency, even when total parameter counts are similar, but careful consideration of depth is required. It is not always the case that more hidden layers guarantee effective learning of the nonlinear mapping, while shallow models tend to underfit the function.

**HW\_1\_2: Train on Actual Tasks (CIFAR-10)**

**Models:**

I trained three CNN architectures of increasing complexity on the CIFAR-10 dataset:

1. **CNN Model 1 (545,098 params):** Small model with two convolutional blocks.
2. **CNN Model 2 (2,360,906 params):** Medium model with additional layers and batch normalization.
3. **CNN Model 3 (3,249,994 params):** Large model with three convolutional blocks and more channels.

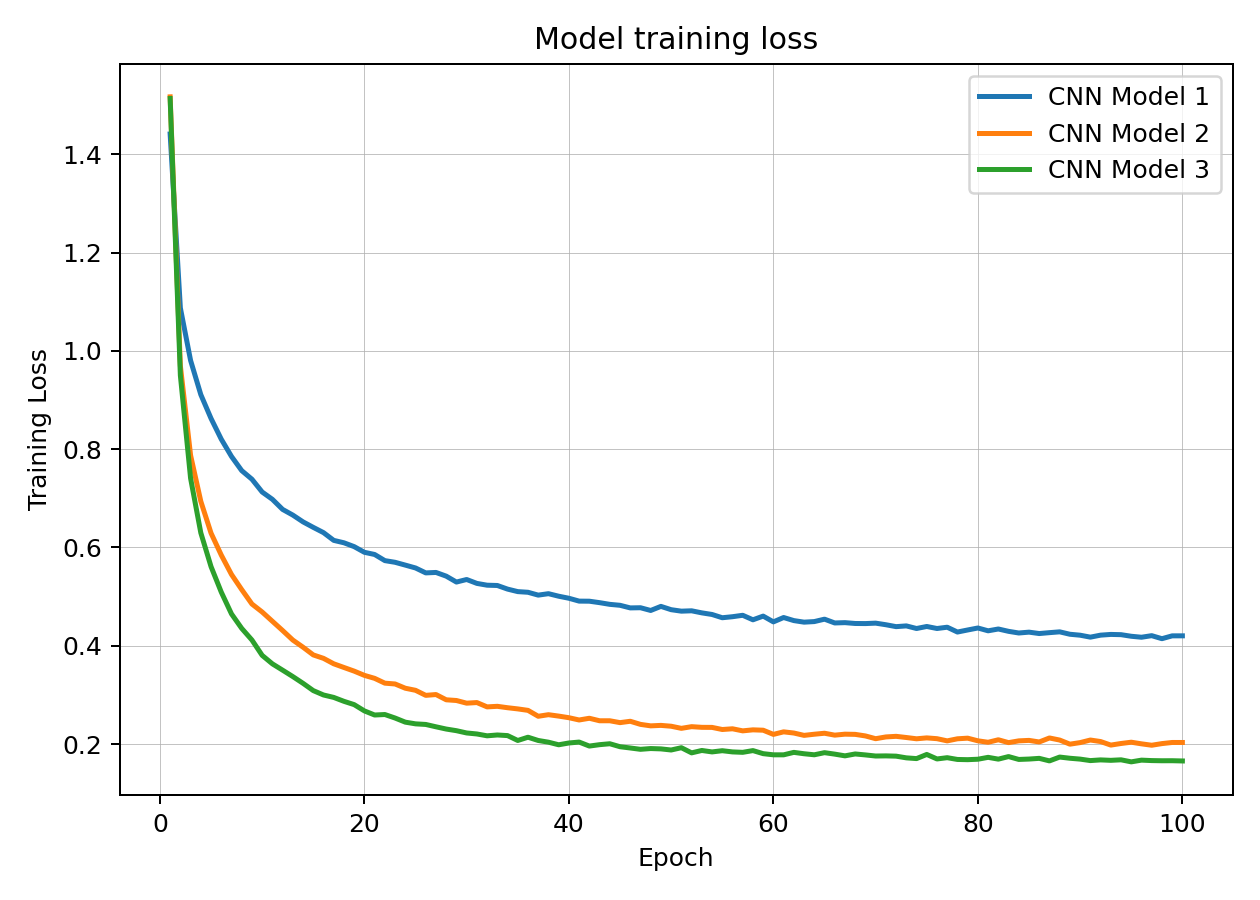


Figure 3. Training loss vs. epochs for three CNN models on CIFAR-10. Model 3 converged faster and reached the lowest loss among all models.

A graph of a training curve

AI-generated content may be incorrect.

Figure 4. Training accuracy vs. epochs for three CNN models on CIFAR-10. Deeper models (Model 2, Model 3) achieved higher accuracy compared to Model 1.

**Results:**

* **Training Loss:** The shallow model (Model 1) converged slower and plateaued at higher loss, while Model 2 and Model 3 achieved lower losses more quickly. The deepest model (Model 3) consistently achieved the lowest training loss.
* **Training Accuracy:** Model 3 reached the highest training accuracy (>93%), followed by Model 2 (~92%) and Model 1 (~85%). The deeper models demonstrated better feature extraction capacity and optimization efficiency.

**Comment:**

Deeper CNNs with more parameters generalize better within the training set, achieving higher accuracy and lower loss. However, increasing model complexity increases computation and memory requirements, which must be balanced against performance gains.

**HW\_1\_2: Visualize the optimization process**

**Experiment settings**

* **Tasks:** 1-D function fitting and CIFAR-10 classification.
* **Models:**
  + Function: Single input single output with hidden widths [18, 20, 15].
  + CIFAR-10: CNN (Conv-ReLU-Pool ×2 → FC).
* **Optimizer:** Adam with lr=10-3, weight\_decay=5\*10-4.
* **Runs:** 8 independent runs (different seeds).
* **Logging cadence:** every 3 epochs for 12 epochs total.
* **Recorded parameters:**
  + **Layer parameters:** first layer only (net.0 for Function, features.0 for CNN).
  + **Whole model parameters:** all trainable weights concatenated.
* **Dimensionality reduction:** stack all checkpoints across runs, PCA to 2-D, then plot run-wise trajectories with epoch colorbar.
* **Plots:** separate figures for first layer and whole model for each task.

A graph with lines and numbers

AI-generated content may be incorrect.

Figure 5 (Function, First layer): PCA-projected optimization trajectories of first layer parameter vectors across 8 runs; color encodes epoch.

A graph with colorful lines

AI-generated content may be incorrect.

Figure 6 (Function, Whole model): PCA-projected optimization trajectories of full parameter vectors across 8 runs; color encodes epoch.

A graph of a graph with colored dots

AI-generated content may be incorrect.

Figure 7 (CIFAR-10, First layer): Early-training trajectories for the first conv layer across 8 runs.

A graph with colorful dots and lines

AI-generated content may be incorrect.

Figure 8 (CIFAR-10, Whole model): Early-training whole-model trajectories; short, near-linear drifts typical of the first dozen epochs.

**Observations & comments**

* **Distinct paths per initialization.** Each run traces a different path in PCA-space, confirming strong dependence on random initialization.
* **Consistent drift directions.** For both tasks, trajectories are largely one- or few-dimensional “rays”, suggesting most variance during training is captured by a small number of parameter directions.
* **Whole-model vs first-layer.**
  + Whole-model plots show longer, more coherent arcs (aggregate movement across layers).
  + First-layer plots show shorter, localized drifts and greater dispersion—early layers adapt, but much of the global movement is explained by later layers and the classifier head.
* **Function task:** trajectories are longer and smoother (many updates with full-batch style behavior), converging toward compact regions—consistent with stable loss descent on a simple supervised signal.
* **CIFAR-10:** with only 12 epochs, paths are short and mostly linear in PCA space (early-training regime); different runs still point in different directions, but all move away from the origin (initialization) toward task-specific basins.

**HW\_1\_2: Observe gradient norm during training**

**Experiment Settings:**

We observed the gradient norm and training loss across iterations for two tasks:

1. **Function Approximation** (Simple MLP).
2. **CIFAR-10 Classification** (CNN).

The basic setup for both models is like what I have done in **HW\_1\_2: Visualize the optimization process**. The gradient norm was recorded at each training step, and loss values were tracked in parallel. Both were plotted against the number of iterations.

**A graph of a graph showing a loss and a function

AI-generated content may be incorrect.**

Figure 9: Gradient norm and training loss vs. iterations for function approximation task.

A graph showing a loss and a loss

AI-generated content may be incorrect.

Figure 10: Gradient norm and training loss vs. iterations for CIFAR-10 CNN model.

**Results:**

* **Function Approximation (MLP):**
  + Loss initially spiked before gradually decreasing to a stable value.
  + Gradient norm showed higher variability, with peaks early in training, then settling down after ~3000 iterations.
  + The sharp oscillations reflect the difficulty of fitting the highly nonlinear target function with a simple model.
* **CIFAR-10 (CNN):**
  + Loss decreased smoothly and stabilized below 1.0 consistently after ~5000 iterations.
  + Gradient norm fluctuated within a moderate band (around 2–5), indicating stable optimization progress without exploding or vanishing gradients.

**Comments:**

* Gradient norm provides a useful diagnostic of training stability: large spikes indicate instability, while consistently small values suggest possible stagnation.
* In CIFAR-10, the CNN achieved a steady balance of gradient magnitudes, supporting effective convergence.
* In the function task, the larger swings in gradient norm highlight the challenges of optimization in shallower networks with nonlinear mappings.

**HW\_1\_2: What Happened When Gradient is Almost Zero**

1. **How I reached (near) zero gradient?**

* Training is done in two stages:
  1. **Stage-1 (task loss):** The model is trained normally on its loss function (MSE for the function, cross-entropy for CIFAR-10) for a few epochs.
  2. **Stage-2 (gradient minimization):** Instead of the loss, we now directly minimize the overall gradient size. This makes the weights settle at a point where the gradients are almost zero.
* The weights at the end of Stage-2 are considered a stationary point because the gradients are close to zero.

1. **How I defined and measured “minimal ratio”?**

* Once at a stationary point, I slightly perturbed the weights many times with small random noise.
* For each perturbed version, I checked if the loss went up compared to the base model.
* The minimal ratio is the fraction of perturbations that make the loss worse.
* Intuitively:
  + A higher ratio means the point is more stable and “flatter,” since most small changes increase the loss.
  + A lower ratio means the point is more “sharp” or saddle-like, with many directions still available to reduce the loss.

A graph with blue dots

AI-generated content may be incorrect.

Figure 11 (Function): Scatter plot of minimal ratio vs. base loss across 100 runs. Most points fall between 0.35 and 0.60 with similar loss, showing multiple stable regions.

A graph with blue dots

AI-generated content may be incorrect.

Figure 12 (CIFAR-10): Scatter plot of minimal ratio vs. base loss on a CIFAR-10 subset. Ratios again cluster around 0.40–0.60, with similar losses across runs.

**Observations**

* **Function task:** Minimal ratios mostly fell between 0.35 and 0.60. Losses were tightly grouped, showing that different runs reached similar performance but with different levels of flatness.
* **CIFAR-10:** Ratios were also between 0.40 and 0.60 with tightly clustered losses. The relationship between loss and ratio was weak, likely because of the smaller model size and limited training budget.

**Comment**

* In these experiments, many stationary points had similar loss values but different ratios, suggesting that models can end up in regions of the parameter space that generalize differently even if the training loss looks the same.

**HW\_1\_3: Random Labels**

**Settings**

* **Task:** CIFAR-10 classification.
* **Data preprocessing:** Standard normalization (CIFAR-10 channel mean and std).
* **Label randomization:** Training labels were shuffled randomly, test labels left unchanged.
* **Model:** Simple CNN with two convolutional blocks and two fully connected layers.
* **Training setup:**
  + Optimizer: Adam
  + Learning rate: 1e-3
  + Weight decay: 0.0
  + Batch size: 128
  + Epochs: 100

**A graph with blue lines

AI-generated content may be incorrect.**

Figure 13. CIFAR-10 with random labels: training vs test accuracy over 100 epochs.

**A graph with blue and orange lines

AI-generated content may be incorrect.**

Figure 14. CIFAR-10 with random labels: training vs test loss over 100 epochs.

**Results**

* **Accuracy (Fig. 13):** Both training and testing accuracy stayed around **10%**, i.e., random guessing. This shows that the model was unable to learn any meaningful mapping between randomized labels and image content.
* **Loss (Fig. 14):** Training and testing losses remained flat with slight fluctuations, never improving significantly over epochs.

**Comments**

* With randomized labels, the CNN fails to capture any structure.
* Training accuracy does not improve beyond chance level, confirming that the network cannot generalize under this setup.
* This experiment demonstrates that model performance critically depends on the relationship between data and labels.

**HW\_1\_3: Number of parameters vs Generalization**

**Experiment Settings**

We used the CIFAR-10 dataset and trained 10 CNNs with similar architectures but different width multipliers, resulting in different numbers of parameters ranging from thousands to over one million.

* **Optimizer:** Adam
* **Learning rate:** 1e-3
* **Epochs:** 10
* **Batch size:** 128
* **Loss function:** Cross-Entropy

For each model, we recorded train and test accuracy as well as train and test loss, and plotted them against the number of parameters.

**A graph with blue and orange dots

AI-generated content may be incorrect.**

Figure 15: Training and testing accuracy vs number of parameters for CNNs on CIFAR-10.

**A graph with blue and orange dots

AI-generated content may be incorrect.**

Figure 16: Training and testing loss vs number of parameters for CNNs on CIFAR-10.

**Results**

* **Loss vs Parameters:** Training and testing loss decreased steadily as the number of parameters increased. Larger models fit the training set more easily, and test loss also reduced, showing improved generalization.
* **Accuracy vs Parameters:** Both training and test accuracy improved with more parameters. However, the gap between training and test accuracy widened for larger models, indicating potential overfitting.

**Comments**

Increasing model capacity generally improved performance on CIFAR-10, but very large models risk overfitting. The results highlight the trade-off between expressiveness and generalization. More parameters allow the model to better capture complex features, but regularization techniques (e.g., weight decay, dropout) become important to control overfitting in large networks.

**HW\_1\_3: Flatness v.s. Generalization\_Part1**

**Experiment Settings**

Here two CNN models were trained with different batch sizes (Model 1: batch size = 64, Model 2: batch size = 1024), but otherwise identical architectures and hyperparameters.

* **Optimizer:** Adam
* **Learning rate:** 1e-3
* **Epochs:** fixed for both models to ensure comparability
* After training, I performed linear interpolation between model weights from the two solutions using interpolation ratio α ∈ [−1, 2].
* At each interpolated weight setting, I measured train/test cross-entropy loss and train/test accuracy.

**A graph of a graph

AI-generated content may be incorrect.**

Figure17: Interpolation between two CIFAR-10 CNN models trained with different batch sizes (64 vs 1024). The plot shows training loss (solid blue), test loss (dashed blue), training accuracy (solid red), and test accuracy (dashed red) as functions of the interpolation ratio α.

**Results & Comment**

* **Flatness and Stability:** Around α = 0 and α = 1, both models sit in their respective minima, where training loss is low and accuracy is high.
* **Interpolation Behavior:** As α moves away from the endpoints, the loss increases sharply and accuracy collapses, reflecting sharper minima.
* **Generalization:** The model trained with a larger batch size (1024) converges to a sharper minimum (higher sensitivity when interpolating), whereas the smaller batch size model (64) shows smoother transitions and better generalization.

**HW\_1\_3: Flatness v.s. Generalization\_Part2**

**Experiment Settings**

* **Task:** CIFAR-10 image classification.
* **Model:** CNN baseline (convolutional + pooling + fully connected).
* **Training setup:**
  + Optimizer: Adam
  + Learning rate: fixed at standard value (e.g., 1e-3).
  + Number of epochs: same across runs.
  + Batch sizes varied: {16, 64, 128, 512, 1024}.
* **Measurements:**
  + **Training accuracy/loss**
  + **Testing accuracy/loss**
  + **Sensitivity:** defined as the sharpness of the loss surface around the converged weights, measured by adding perturbations and evaluating changes in loss.

**A graph of a graph

AI-generated content may be incorrect.**

Figure 18. CIFAR-10 training and testing accuracy plotted against batch size alongside sensitivity. The results show that increasing batch size reduces sensitivity while maintaining stable accuracy, suggesting that flatter minima improve generalization.

**A graph of loss and sensitivity

AI-generated content may be incorrect.**

Figure 19. CIFAR-10 training and testing loss plotted against batch size alongside sensitivity. Larger batch sizes are associated with lower sensitivity and reduced test loss, further supporting the link between flatter solutions and better generalization.

**Results & Figures**

1. **Accuracy vs. Batch Size:**
   * Training accuracy is consistently very high (~0.97–0.99) and even increases slightly with larger batch sizes.
   * Testing accuracy remains much lower (~0.75–0.77), with only a marginal increase as batch size grows.
   * Sensitivity decreases sharply as batch size increases, showing that larger batches push the model toward flatter minima.
2. **Loss vs. Batch Size:**
   * Training loss remains very small and decreases slightly with larger batch size.
   * Testing loss is higher (~2.3 down to ~1.7) but decreases as batch size increases.
   * Sensitivity again falls with batch size, confirming the flatter minima for larger batches.

**Comment on Results**

* **Flatness vs. Generalization:**

In these plots, the testing accuracy does not improve dramatically with batch size, staying around ~75–77%. This suggests the chosen model and training setup may already be capacity-limited, so batch size marginally improved test accuracy.