Lecture 3 – Fitting models & Regularization

Provide short answers (max. 25 words each) to the following questions.

- What does “fitting a model” mean in deep learning?
- What is convexity in optimization?
- Why are most deep learning loss landscapes non-convex?
- What is the main idea behind Stochastic Gradient Descent (SGD)?
- What is an epoch in Stochastic Gradient Descent (SGD)?
- List two advantages of Stochastic Gradient Descent (SGD) over full-batch gradient descent.
- Why does Stochastic Gradient Descent (SGD) not converge in the traditional sense?
- What is momentum in optimization?
- What is Nesterov accelerated momentum?
- What does Adam optimizer combine?
- Name three key hyperparameters in deep learning optimization.
- Why is a validation set needed?
- What is double descent?
- What is inductive bias in model selection?
- What is the goal of regularization?
- What is explicit regularization?
- What is the most common explicit regularization method?
- Why does L2 regularization help?
- What is implicit regularization?
- How does early stopping act as regularization?
- What is Dropout?
- How is Dropout applied during testing?
- What is Ensembling?
- What is the probabilistic interpretation of regularization?
- What does Bayesian learning aim to compute?
- What is Transfer Learning?
- What is Multi-task Learning?
- What is Self-Supervised Learning?
- Why is data augmentation considered a regularization technique?

Multiple-choice questions

What does “fitting a model” mean in deep learning?

A) Choosing activation functions
B) Increasing the number of layers
C) Optimizing parameters to minimize a loss function
D) Normalizing inputs

Which condition indicates a convex function?

A) First derivative is zero everywhere
B) Second derivative is positive everywhere
C) Gradient is negative everywhere
D) Function has multiple local minima

Why are deep learning models' loss landscapes usually non-convex?

A) They use linear activations
B) They use small datasets
C) They lack regularization
D) They have multiple layers and nonlinearities

What is the main idea behind Stochastic Gradient Descend (SGD)?

A) Use full dataset for each update
B) Compute gradients using mini-batches
C) Avoid using gradients
D) Increase learning rate continuously

What is an epoch in Stochastic Gradient Descend (SGD)?

A) One mini-batch update
B) One gradient computation
C) One complete pass through the dataset
D) One validation step

Which is NOT an advantage of Stochastic Gradient Descend (SGD)?

A) Escapes local minima
B) Less computational cost
C) Converges exactly to global minimum
D) Uses all data equally over time

Why does SGD not converge in the traditional sense?

A) It uses too large batches
B) Noise from mini-batches prevents exact convergence
C) It ignores gradients
D) It uses convex functions only

What does momentum do in optimization?

A) Stops updates when gradient is zero
B) Reduces learning rate automatically
C) Accumulates past gradients to accelerate convergence

D) Normalizes gradients

What is Nesterov accelerated momentum?

A) Applies momentum after gradient computation
B) Looks ahead before computing gradient
C) Ignores previous gradients
D) Uses adaptive learning rates only

What does Adam optimizer combine?

A) Momentum and adaptive learning rates
B) SGD and Dropout
C) L1 and L2 regularization
D) Batch normalization and SGD

Which is NOT a hyperparameter in optimization?

A) Learning rate
B) Momentum
C) Batch size
D) Activation output

Why is a validation set used?

A) To compute gradients
B) To tune hyperparameters without overfitting to test set
C) To increase training data size
D) To measure training loss

What is double descent?

A) Error decreases monotonically with model size
B) Test error decreases, then increases, then decreases again
C) Training error increases with capacity
D) Gradient magnitude doubles during training

What is inductive bias?

A) Random initialization of weights
B) Model's use of batch normalization
C) Gradient clipping
D) Model's tendency to prefer certain solutions

What is the main goal of regularization?

A) Increase training accuracy
B) Reduce generalization gap

C) Eliminate activation functions
D) Speed up optimization

What is explicit regularization?

A) Adding noise to inputs
B) Adding penalty terms to the loss function
C) Using dropout
D) Early stopping

Which is the most common explicit regularization method?

A) L1 regularization
B) Dropout
C) L2 regularization
D) Batch normalization

Why does L2 regularization help?

A) Encourages large weights
B) Increases learning rate
C) Removes activation functions
D) Encourages smaller weights and smoothness

What is implicit regularization?

A) Regularization added manually
B) Regularization effects from optimization dynamics
C) Dropout during training
D) Adding noise to labels

How does early stopping act as regularization?

A) Stops training before overfitting occurs
B) Increases model complexity
C) Removes Dropout
D) Adds penalty terms

What is Dropout?

A) Removing layers permanently
B) Randomly dropping units during training
C) Reducing learning rate
D) Adding noise to gradients

How is Dropout applied during testing?

A) Units are still dropped

B) Learning rate is reduced
C) Outputs are scaled by dropout probability
D) Gradients are normalized

What is Ensembling?

A) Combining predictions from multiple models
B) Using dropout during inference
C) Adding noise to inputs
D) Increasing batch size

How does adding noise to inputs help?

A) Makes model robust and prevents overfitting
B) Reduces training data size
C) Increases learning rate
D) Removes activation functions

What is the probabilistic interpretation of regularization?

A) Equivalent to adding a prior over parameters
B) Equivalent to increasing batch size
C) Equivalent to Dropout
D) Equivalent to Early Stopping

What does Bayesian learning compute?

A) A single best parameter set
B) A posterior distribution over parameters
C) A fixed learning rate
D) A deterministic gradient

What is Transfer Learning?

A) Training from scratch
B) Removing regularization
C) Increasing dataset size
D) Using a pre-trained model for a new task

What is Multi-task Learning?

A) Training one model for one task
B) Training a model for multiple related tasks
C) Using dropout for multiple layers
D) Adding noise to multiple inputs

What is Self-Supervised Learning?

A) Learning from labeled data only
B) Using multiple optimizers
C) Learning representations from unlabeled data using pretext tasks
D) Removing activation functions

Why is data augmentation considered regularization?

A) It reduces dataset size
B) It increases data diversity and reduces overfitting
C) It removes noise
D) It eliminates hyperparameters