

Lecture 2 – Loss functions

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

- What is the role of a loss function in training neural networks?
- Why do we minimize the negative log-likelihood instead of maximizing likelihood directly?
- What does the universal approximation theorem state?
- What is depth efficiency in neural networks?
- What is the maximum likelihood criterion?
- Why do we use the log function in likelihood calculations?
- What distribution is commonly assumed in univariate regression?
- What loss function arises from Gaussian likelihood in regression?
- Why does variance disappear in the least squares regression criterion?
- What challenge arises in multivariate regression with outputs of different scales?
- What distribution models binary classification? Provide both the name and the formula.
- What function maps neural network outputs to $[0,1]$ for binary classification?
- What is the binary cross-entropy loss formula?
- What distribution models multiclass classification? Provide both the name and the formula.
- What function ensures outputs sum to one in multiclass classification?
- What is the multiclass cross-entropy loss formula?
- How is the predicted class chosen in multiclass classification? Include in your response, a short description of the model output.
- What is cross-entropy in machine learning? Provide its formula, and a description of what it measures.
- How is KL divergence related to cross-entropy?
- Why is cross-entropy preferred over MSE for classification?
- Why do we treat the outputs of multi-output models as being independent?
- Why do deep networks need tricks for training beyond 20 layers?
- What is the role of activation functions in loss computation?
- Why is log-likelihood maximization equivalent to negative log-likelihood minimization?
- What is the main advantage of probabilistic modeling in loss design?

Multiple-choice questions

What is the primary purpose of a loss function in neural networks?

- A) To increase the number of layers
- B) To measure prediction error
- C) To normalize inputs
- D) To compute activation values

Which of the following is minimized during training?

- A) Likelihood
- B) Activation function
- C) Negative log-likelihood
- D) Gradient norm

Why do we use the log function in likelihood calculations?

- A) To increase complexity
- B) To avoid underflow and simplify multiplication into addition
- C) To normalize outputs
- D) To compute gradients faster

Which theorem states that a single hidden layer network can approximate any continuous function?

- A) Depth Efficiency Theorem
- B) Maximum Likelihood Theorem
- C) Cross-Entropy Principle
- D) Universal Approximation Theorem

Why are deep networks often preferred over shallow networks?

- A) They handle structured data efficiently
- B) They eliminate the need for training
- C) They avoid activation functions
- D) They eliminate the need for loss functions

What is depth efficiency?

- A) Ability to compute gradients faster
- B) Deep networks approximate functions with fewer units than shallow networks
- C) Shallow networks outperform deep networks
- D) Efficiency in activation computation

Which distribution is commonly assumed in univariate regression?

- A) Bernoulli
- B) Categorical
- C) Gaussian
- D) Poisson

Which loss function corresponds to Gaussian likelihood?

- A) Cross-Entropy
- B) KL Divergence

- C) Hinge Loss
- D) Mean Squared Error

Why does variance disappear in least squares regression?

- A) It is learned separately
- B) It becomes a constant term
- C) It is normalized
- D) It is ignored intentionally

Which distribution models binary classification?

- A) Gaussian
- B) Bernoulli
- C) Categorical
- D) Exponential

Which function maps outputs to $[0,1]$ in binary classification?

- A) Softmax
- B) ReLU
- C) Sigmoid
- D) Tanh

Which distribution models multiclass classification?

- A) Bernoulli
- B) Gaussian
- C) Categorical
- D) Poisson

Which function ensures outputs sum to one in multiclass classification?

- A) Sigmoid
- B) Softmax
- C) ReLU
- D) Tanh

How is the predicted class chosen in multiclass classification?

- A) Random selection
- B) Class with smallest probability
- C) Class with largest probability
- D) Average of all probabilities

What does cross-entropy measure?

- A) Distance between two probability distributions

- B) Difference between two activation functions
- C) Gradient magnitude
- D) Variance of outputs

How is KL divergence related to cross-entropy?

- A) $\text{KL divergence} = \text{Cross-Entropy} + \text{Entropy}$
- B) $\text{Cross-Entropy} = \text{KL divergence} + \text{Entropy}$
- C) They are unrelated
- D) $\text{KL divergence} = \text{Cross-Entropy} - \text{Entropy}$

Why is cross-entropy preferred over MSE for classification?

- A) It is easier to compute
- B) It penalizes incorrect confident predictions
- C) It avoids activation functions
- D) It requires fewer parameters

What happens if outputs are treated as independent in multi-output models?

- A) Loss becomes a product of terms
- B) Loss becomes a sum of terms
- C) Loss disappears
- D) Loss becomes zero

Why do deep networks need tricks for training beyond 20 layers?

- A) They lack activation functions
- B) They suffer from vanishing/exploding gradients
- C) They have too few parameters
- D) They cannot compute loss

What is the role of activation functions in loss computation?

- A) To compute gradients
- B) To ensure outputs are in valid ranges for probability distributions
- C) To normalize inputs
- D) To reduce variance

Why is log-likelihood maximization equivalent to negative log-likelihood minimization?

- A) Because log is monotonic
- B) Because they are unrelated
- C) Because of normalization
- D) Because of activation functions

What is the main advantage of probabilistic modeling in loss design?

- A) It avoids optimization
- B) It provides a principled way to handle uncertainty
- C) It eliminates activation functions
- D) It reduces computation

Which loss is commonly used for regression tasks?

- A) Cross-Entropy
- B) Mean Squared Error
- C) Hinge Loss
- D) KL Divergence

Which loss is commonly used for binary classification?

- A) Mean Squared Error
- B) Binary Cross-Entropy
- C) Hinge Loss
- D) KL Divergence

Which loss is commonly used for multiclass classification?

- A) Mean Squared Error
- B) Binary Cross-Entropy
- C) Multiclass Cross-Entropy
- D) Hinge Loss

What is the effect of large-scale outputs in multivariate regression?

- A) They dominate the loss
- B) They reduce variance
- C) They improve accuracy
- D) They normalize automatically

Which technique can handle outputs of different scales?

- A) Ignoring variance
- B) Rescaling inputs and outputs
- C) Removing activation functions
- D) Using softmax

Which concept connects cross-entropy and negative log-likelihood?

- A) They are equivalent in classification tasks
- B) They are unrelated
- C) Cross-Entropy is larger than NLL
- D) NLL is larger than Cross-Entropy