The Science of Deep Learning

Iddo Drori

Massachusetts Institute of Technology

Columbia University

Cambridge University Press, 2022.

1 Exercises

1.1 Neural Networks, Forward and Backpropagation

Problem 1. (4 points) **Successive gradient steps.**
In typical gradient descent, we take steps of a constant size, so that:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta f(\theta_t)$$

In the following, assume that f is an arbitrary differentiable function.

Problem 1.1 (1 points)
For very small η, what will generally be true?

a. $f(\theta_t) \geq f(\theta_{t+1})$
b. $f(\theta_t) \leq f(\theta_{t+1})$
c. cannot say

Problem 1.2 (1 points)
For very big η, what will generally be true?

a. $f(\theta_t) \geq f(\theta_{t+1})$
b. $f(\theta_t) \leq f(\theta_{t+1})$
c. cannot say

Problem 1.3 (1 points)
We would like to pick a perfect step size on every step and propose a new update rule that selects η' to be the value step-size η that decreases the objective as much as possible in the direction $\nabla_\theta f(\theta)$ and then uses η' as the step size:

$$\eta' = \arg\min_\eta f(\theta_t - \eta \cdot \nabla_\theta f(\theta_t))$$
$$\theta_{t+1} = \theta_t - \eta' \cdot \nabla_\theta f(\theta_t)$$

What will generally be true?

a. $f(\theta_t) \geq f(\theta_{t+1})$
b. $f(\theta_t) \leq f(\theta_{t+1})$
c. cannot say

Problem 1.4 (1 points)
Assume that we pick η to be the optimal η' from the previous question. The relationship between successive steps taken by consecutive directions must be:

a. Orthogonal, otherwise we could minimize f further by using a larger step.
b. Parallel, since we have found the optimal direction to reach the goal
c. Neither orthogonal nor parallel since the gradient at θ_{t+1} can be anything.

Problem 2. (1 points)
How many weights are in a fully connected neural network with input dimension 5, output dimension 1, and 3 hidden layers (not including the output layer) with 7 activation units each (no bias terms)?

Problem 3. (1 points)
Suppose that the input to the non-linear component of the first activation in the first layer of a neural network is $4x + 3y + 2z - 1$ and to the second activation is $2x - y + 5z + 2$. What are the corresponding matrix weights and inputs?

Problem 4. (2 points) **Computational Graph**

Problem 4.1 (1 points)
Draw the computational graph to compute the function $f(x, y) = y(x + y)$ and use the graph to compute $f(4, 3)$.

Problem 4.2 (1 points)
Draw a backpropagation graph to compute the derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ for $f(x, y) = y(x + y)$. Use the graph to find the derivatives at $x = 4$ and $y = 3$.

Problem 5. (2 points) Repeat (1) and (2) above for $f(x, y) = x^2(x - y)$.

Problem 6. (1 points)
What is the advantage of computing the derivatives of the output with respect to each of the input variables in a backward vs. forward passes?

Problem 7. (2 points)

Show that the $\tanh$ function and the logistic sigmoid function σ are related by

$$\tanh(a) := \frac{e^a - e^{-a}}{e^a + e^{-a}} = 2\sigma(2a) - 1.$$

Problem 8. (2 points) **Logic gates**

Problem 8.1 (2 points)
Draw a network that receives a two dimensional input, with at most two layers, that has at most width two such that:

$$f(x, W_1, b_1, W_2, b_2) > 0 \tag{1.1}$$

iff

$$(x_1 > 0) \text{ OR } (x_2 > 0) \tag{1.2}$$

Problem 8.2 (2 points)
Draw a network that receives a two dimensional input, with at most three layers, that has at most width four such that:

$$f(x, W_1, b_1, W_2, b_2) > 0 \tag{1.3}$$

iff

$$(x_1 < 0 \text{ AND } x_2 > 0) \text{ OR } (x_1 > 0 \text{ AND } x_2 < 0) \tag{1.4}$$

Problem 8.3 (2 points)
Draw a network that implements a NAND gate, which can be used as a building block of any logic function.

Problem 9. (2 points) **Logistic regression**

Problem 9.1 (1 points)
Describe why logistic regression is a linear classifier.

Problem 9.2 (1 points)
Demonstrate how softmax function is a multi-class extension of the logistic regression.

Problem 10. (1 points)
Consider a simple neural network composed of a single layer, containing a sigmoid unit with two inputs. Can this network learn to correctly classify the following four data points?

$$x^1 = (0, 0), \quad y^1 = 0;$$

$$x^2 = (0, 1), \quad y^2 = 1;$$

$$x^3 = (1, 0), \quad y^3 = 1;$$

$$x^4 = (1, 1), \quad y^4 = 0$$

Problem 11. (1 points)
Consider a simple neural network composed of a single layer, containing a sigmoid unit with two inputs. Is this a linear, non-linear or unknown classifier?

a. Liner
b. Non-Linear
c. Unknown

Problem 12. (2 points) **Loss and activation functions**
For each of the following loss functions which activation functions in the last layer is appropriate?

Problem 12.1 (0.5 points)
Negative Log-Likelihood Multiclass (NLLM) loss

a. Linear
b. Softmax
c. Sigmoid

Problem 12.2 (0.5 points)
Squared loss

a. Linear
b. Softmax
c. Sigmoid

Problem 12.3 (0.5 points)
Negative Log-Likelihood (NLL) loss

a. Linear
b. Softmax
c. Sigmoid

Problem 12.4 (0.5 points)
Hinge loss

a. Linear
b. Softmax
c. Sigmoid

Problem 13. (2 points) **Applications and activations**
For each of the following applications which activation function is appropraite?

Problem 13.1 (0.5 points)
Map words in a news page to a predicted numerical change in a stock market mean

a. Linear
b. Softmax

Problem 13.2 (0.5 points)
Map a satellite image to the probability it will rain at that location during the next day

a. Linear
b. Softmax

Problem 13.3 (0.5 points)
Map words in an email to which one of a fixed set of folders it should be filed in.

a. Linear
b. Softmax

Problem 14. (2 points) **Applications and loss functions**
Match the following applications to a loss function that make sense for them.

Problem 14.1 (0.5 points)
Map words in a news page to a predicted numerical change in a stock market mean

a. Negative Log-Likelihood (NLL)
b. Squared

Problem 14.2 (0.5 points)
Map a satellite image to the probability it will rain at that location during the next day

a. Negative Log-Likelihood (NLL)
b. Squared

Problem 14.3 (0.5 points)
Map words in an email to which one of a fixed set of folders it should be filed in.

a. Negative Log-Likelihood (NLL)
b. Squared

Problem 15. (3 points)
The function $f(x, y) = x^2 + (x + 6y)^4$ has a minimum $f(0, 0) = 0$.

Problem 15.1 (2 points)
What is the gradient of the function at $(1, 1)$?

Problem 15.2 (1 points)
If we initialize gradient descent to $(1, 1)$ with $\alpha = 0.0001$.
What are the values of (x, y) after the first iteration of gradient descent?

Problem 16. (1 points)
Is the following statement true or false?

During forward propagation for layer $\mathcal{L}$ you need to know the activation function i.e. Sigmoid, ReLU etc. But, during backward propagation, the backward function does not need to know the corresponding activation function for layer $\mathcal{L}$.

Problem 17. (2 points)
You are given a neural network with a sigmoid activation functions in the hidden layer and the following information:

Input:

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Hidden Layer weights:

$$\begin{pmatrix} 0.8 & 0.4 & 0.3 \\ 0.2 & 0.9 & 0.5 \end{pmatrix}$$

Output Layer weights:

$$\begin{pmatrix} 0.3 & 0.5 & 0.9 \end{pmatrix}$$

What is the network output to within 3 digits of precision?

Problem 18. (2 points)
For $a = (0.3, 0.5, 0.2)^T$ and $y = (0, 0, 1)^T$, what is the negative log likelihood (NLL) loss?

1.2 Optmization

Problem 19. (2 points)
A "loss landscape" of your neural network's loss function has a plateau shape, in which there are large areas where the loss is flat. This means the gradient signal will be weak and gradient descent will not update the network parameters very much in each iteration.
What are two activation functions that may cause such a problem, and one activation function that will not?

Problem 20. (2 points)
A "loss landscape" of your neural network's loss function has a saddle shape. Suppose your gradient descent algorithm has converged to the center point of the saddle. Which of the following statements are true?

a. The gradient vector (1st derivative) of the loss function is 0 in all dimensions.
b. The gradient vector (1st derivative) of the loss function has some positive values and some negative values.
c. The Hessian matrix (2nd derivative) of the loss function has some positive eigenvalues and some negative eigenvalues.
d. The Hessian matrix (2nd derivative) of the loss function has all 0 eigenvalues.

Problem 21. (1 points)
Stochastic gradient descent helps accelerate convergence since:

a. A smaller mini-batch reduces noise in each mini-batch, which results in faster convergence.
b. A smaller mini-batch introduces noise approximating the gradient instead of computing the true gradient, which improves generalization.
c. A larger mini-batch size aggregates gradient information, which results in faster optimization.

Problem 22. (2 points)
Which methods help accelerate the optimization of a model that uses batch gradient descent?

a. Using Adam.
b. Fine tuning the learning rate using grid search.
c. Initializing all the weights to zero.
d. Using mini-batch gradient descent.

Problem 23. (2 points)
Calculate the value of the function

$$f(\theta) = (2\theta - 2)^4$$

after updating the θ value in one step of gradient descent. Have $\theta = 4$ and $\eta = 0.01$.

Problem 24. (2 points)
Which statement is true about the step size in gradient descent?

a. The step size is related to the learning rate.
b. If the step size is too big, gradient descent oscillates and may overshoot.
c. The smaller the step size, the faster we can reach the optimal minima.

Problem 25. (2 points)
Given 256 training examples, what is the number of weight updates that are performed in a single training epoch using

a. Full
b. Stochastic
c. Mini-batch

gradient descent with a batch size of 32?

1.3 Regularization

Problem 26. (2 points)
Which are true about the softmax?

a. Sigmoid is a generalization of the softmax.
b. It can be used for binary classification.
c. The sum of the output is equal to 1.
d. Softmax can be interpreted as a probability of each class.

Problem 27. (2 points)
Which of these are regularization methods?

a. Adding a term of λ-squared magnitude of coefficients.
b. Dropout.
c. Batch norm.
d. Data augmentation.

Problem 28. (1 points)
Which of the following distributions may be regarded as the prior corresponding to $\mathcal{L}_2$ regularization?

a. Gaussian
b. Laplace
c. Cauchy

Problem 29. (1 points)
Which of the following distributions may be regarded as the prior corresponding to $\mathcal{L}_1$ regularization?

a. Gaussian
b. Laplace
c. Cauchy

Problem 30. (3 points)
True or false?

Problem 30.1 (1 points)
$\mathcal{L}_2$ regularization encourages sparse weights.

Problem 30.2 (1 points)
Increasing dropout in a neural network increases the number of computations

to be performed polynomially.

Problem 30.3 (1 points)
You notice while training your neural network that the test loss initially decreases but then starts increasing, while training loss continues to decrease. This means that test loss will only continue to increase if we train any further.

1.4 Convolutional Neural Networks

Problem 31. (2 points)
Determine how many pixels of padding are required for as the input of size 80 by 80 to ensure that convolution with a 31 by 31 filter results in an output of the same size.

Problem 32. (2 points)
Which of the following are true of convolutional neural networks for image analysis?

a. They are trained using backpropagation.
b. They are trained using gradient descent.
c. They are trained using supervised learning .
d. They are trained using unsupervised learning.

Problem 33. (2 points)
Consider an input vector x of size 10 and a filter k of size 4. What is the size of the output vector of the convolution of x with k with padding of size 1?

Problem 34. (2 points)
Which of the following are true of pooling layers in CNNs?

a. They reduce the size of the input to the next layer.
b. They increase the number of parameters.
c. They reduce the number of connections to the next layer.
d. They reduce the number of parameters.

Problem 35. (2 points)
Consider the following CNN architecture:
Input: (10, 10, 3)

- A convolutional layer with 32 5 $\times$ 5 filters, stride 1, and padding 2
- A ReLU activation layer
- A max pooling layer with size 2 and stride 2
- A convolutional layer with 64 5 $\times$ 5 filters, stride 1, and padding 2
- A ReLU activation layer
- A max pooling layer with size 2 and stride 2
- A fully-connected layer with 128 neurons
- A ReLU activation layer
- A dropout layer with drop probability 0.5

- A fully-connected layer with 10 neurons
- A softmax activation layer

What is the total number of parameters in this network?

Problem 36. (2 points)
What is the convolution of filter K with image X?

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 1 & 4 & 4 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 4 & 4 \end{pmatrix}$$

Problem 37. (2 points)
For the flattened image represented of the vector $x = [1, 0, 1, 0, 0, 0, 0, 1, 1, 1]$ and filter $k = [-1, 1, -1]$, what is the convolution of k with x with stride 1 and no padding?

Problem 38. (2 points)
What is the result of performing average pooling on the image X?

$$X = \begin{pmatrix} 1 & 17 & 43 & 4 & 5 \\ 2 & 2 & 6 & 8 & 7 \\ 12 & 9 & 4 & 46 & 5 \\ 3 & 4 & 78 & 9 & 62 \\ 12 & 11 & 14 & 42 & 15 \end{pmatrix}$$

with a 2×2 kernel, with stride of 2, and padding with values of (1, 1) pixel on the right side and bottom of the image.

Problem 39. (2 points)
How many weights are in the max pooling operation of the image X above with a 2x2 kernel, with stride of 2, and padding with a value of 1, 1 pixel on the right side and bottom of the image.

Problem 40. (2 points)
Which of the following is the correct filter for detecting horizontal edges?

a.

$$\begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

b.

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$$

c.

$$\begin{pmatrix} 1 & -1 & 1 \\ 0 & -1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$$

d.

$$\begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

Problem 41. (2 points)
Consider one layer of weights in a convolutional neural network processing grayscale images, connecting one layer of units to the next layer of units. Which of the following types of layers has the least parameters?

a. A fully-connected layer from 20 hidden units to 4 output units.
b. A convolutional layer with eight 5×5 filters.
c. A max-pooling layer that reduces a 10×10 image to a 5×5.
d. A convolutional layer with ten 3×3 filters.

Problem 42. (2 points)
Which of the following are true of convolutional neural networks for image analysis?

a. Filters in earlier layers tend to include edge detectors.
b. Pooling layers reduce the spatial resolution of the image.
c. They have more parameters than fully-connected networks with the same number of layers and the same number of neurons in each layer.
d. A CNN can be trained for unsupervised learning tasks, whereas an ordinary neural net cannot.

Problem 43. (2 points)
Given an input image X, and the kernel K, with a stride of 1, and a zero padding of 1 pixel around the image. What is the output of convolution of K with X with

stride 1 and zero padding of size one?

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 4 & 3 \\ 2 & 1 & 2 \\ 4 & 3 & 4 \end{pmatrix}$$

$$K = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$$

1.5 Sequence Models

Problem 44. (2 points) Describe the different ways RNNs are used in the applications of sentiment analysis, named-entity recognition, and text synthesis, and draw an architecture diagram for each application.

Problem 45. (2 points)
Describe the advantages of the Transformer architecture over CNNs and RNNs.

Problem 46. (2 points)
Which of the following is true of the weights in a recurrent neural network?

a. The weights are shared between all time steps.
b. The weights are different for each time step.
c. The weights are different for each input.
d. The weights are different for each output.

Problem 47. (2 points)
What is the output of the following RNN?

$$\begin{cases} x_1 = 1, x_2 = 1, x_3 = 1, w = 1, u = 1, v = 1, s_0 = 0 \\ s_1 = w * s_0 + x_1 \\ v_1 = s_1 \\ s_2 = w * s_1 + x_2 \\ v_2 = s_2 \\ s_3 = w * s_2 + x_3 = \ ? \end{cases}$$

Problem 48. (2 points)
Which of the following is true of the LSTM cell?

a. It is a more general form of the RNN cell
b. It is a more general form of the GRU cell
c. It is a less general form of the GRU cell
d. It is a less general form of the RNN cell

Problem 49. (2 points)
For which of the following sequences can the next character be correctly predicted by a model using only the previous character?

a. $1, 1, 1, -1, 1, 1, 1, -1, ?$
b. $1, -1, 1, -1, 1, -1, ?$
c. $1, 1, -1, 1, 1, -1, 1, 1, -1, ?$

d. $-1, -1, 1, -1, -1, 1, ?$

Problem 50. (2 points)
What is the weight w for an RNN which has input $(x_1, x_2, x_3) = (1, 1, 1)$ and output $(v_1, v_2, v_3) = (1, 2, 3)$ and is defined by $s_t = w * s_{t-1} + x_t$ and $v_t = s_t$?

Problem 51. (2 points)
Given an RNN defined by $s_t = W * s_{t-1} + U * x_t$ with $W = [[-1, 0], [0, -1]]$, $U = [[1], [1]]$, and $s_0 = [[0], [0]]$, what is s_2 for $x = (x_1, x_2) = (1, 0)$?

Problem 52. (2 points)
Suppose you would like to classify restaurant reviews as good or bad. Why is a bag of words not an appropriate solution? Explain.

Problem 53. (2 points)
Select the best type of model for each use case, matching the use case to the options provided:
Use cases:

1. Tumor segmentation analysis or recognizing tumor patches in medical "slides".
2. Sentiment analysis on a large text corpus.
3. Time series forecasting to predict price of automobile in the next month.
4. Predict the price of a house based on square feet, ratio of bedrooms/bathrooms.

Options: CNN, RNN, Fully connected neural network

Problem 54. (2 points)
What is the benefit of using attention in RNN?

a. We can model long dependencies.
b. The inputs are processed in parallel rather than sequentially.
c. Resolves vanishing gradients.
d. All of the above.

1.6 Graph Neural Networks

Problem 55. (2 points)
Citation networks can be treated as graphs where researchers cite fellow researchers' papers. Given a paper "A" we would like to predict if the paper cites another paper "B" or not. Which type of prediction task can you model this to be?

a. Node prediction.
b. Link prediction.
c. Graph prediction.
d. Sub graph prediction.

Problem 56. (2 points)
Select the correct statement among the following:

a. Combine operation gathers information from all nodes and aggregate operation updates the collected information with its self information.
b. Aggregate operation gathers information from all nodes and combine operation updates the collected information with its self information.
c. Combine operation gathers information from it's neighboring nodes and aggregate operation updates the collected information with its self information.
d. Aggregate operation gathers information from it's neighboring nodes and combine operation updates the collected information with its self information.

Problem 57. (2 points)
What are the two key operations used for updating a node representation in a GNN?

a. Aggregate and Combine.
b. Aggregate and Message.
c. Combine and Update.
d. Aggregate and Max Pooling.

Problem 58. (4 points)
Can a GNN compute the number of nodes in a graph and the number of triangles in a graph? In both cases, either explain why not or write the aggregation function that provides a solution.

Problem 59. (2 points)

What is the difference between the following two GNN architectures: Graph Sage and GAT?

a. Equal contribution of information from neighborhood nodes in graph sage versus controlled information from neighborhood nodes in GAT.
b. Graph Sage has a single matrix for neighborhood and self embedding whereas GAT has two.
c. GAT concatenates self embedding and neighbordhood embedding matrices whereas GraphSage does not.

Problem 60. (2 points)
What is the difference between a link prediction and a node classification task?

a. Link prediction predicts the relationship between nodes, whereas node prediction predicts features or categories of a node.
b. Node prediction predicts the relationship between nodes, whereas link prediction predicts features or categories of a node.
c. Node prediction predicts characteristics of a subgraph, whereas link prediction predicts characteristics of the edges of the graph.

Problem 61. (2 points)
What is the difference between a node embedding and a node representation?

a. A node embedding is a special case of node representation.
b. A node representation is a special case of node embedding.
c. There is no difference between the two.

Problem 62. (2 points)
What is the difference between a dynamic and static graph?

a. Static graphs are more flexible compared to dynamic graphs.
b. Dynamic graphs are faster compared to static graphs.
c. Dynamic graphs change over time and are more flexible.

Problem 63. (2 points)
Among the following, which is the most suitable similarity metric to obtain information about mutual friends or mutual connections in a social network?

a. Adjacency similarity.
b. Overlap similarity.
c. Edit distance.
d. Multi-hop similarity.

Problem 64. (2 points) Consider a tree graph where node 1 is connected individually to nodes 0, 2 and 3. Let $s(i, j)$ denote the similarity between two nodes i and j in a graph. If we use random walk embeddings, which relationships are correct?

a. $s(0, 1) = s(0, 2) = s(0, 3)$
b. $s(0, 1) > s(0, 2) > s(0, 3)$
c. $s(0, 1) > s(0, 2) = s(0, 3)$
d. $s(0, 1) < s(0, 2) < s(0, 3)$

Problem 65. (2 points) Consider a tree graph where node 1 is connected individually to nodes 0, 2 and 3. Let $s(i, j)$ denote the similarity between two nodes i and j in a graph. If we use a one-hot encoded vector where each node is of dimension $|V| \times 1$ where $|V|$ represents the number of nodes in a graph, which relationships are correct?

a. $s(0, 1) > s(0, 2) > s(0, 3)$
b. $s(0, 1) = s(0, 2) = s(0, 3)$
c. $s(0, 1) < s(0, 2) < s(0, 3)$
d. $s(0, 1) > s(0, 2) = s(0, 3)$

Problem 66. (2 points) What is the relationship between the adjacency matrix and the degree matrix?

a. Diagonal values of degree matrix are equal to the row sum of adjacency matrix.
b. There is no relationship between the two
c. Diagonal values of degree matrix are equal to the column sum of adjacency matrix.
d. Diagonal values of degree matrix are equal to the transpose of adjacency matrix.

Problem 67. (2 points) What is the Laplacian matrix for a graph with nodes $\{1, 2, 3, 4, 5\}$ and edges $\{(1.5), (1, 3), (2, 3), (2, 5), (3, 4)\}$?

1.7 Generative Adversarial Networks (GANs)

Problem 68. (4 points) Most GANs use different loss functions for the discriminator and generator. Write the equations for the discriminator and generator losses of the original GAN, Least Squares GAN (LS-GAN), and Wasserstein GAN (WGAN).

Problem 69. (4 points) GANs are trained by alternating between training the discriminator and training the generator. In this case, describe how to identify that GAN training has converged.

Problem 70. (4 points) Two common problems in training GANs are mode collapse and vanishing gradients. Explain each problem and describe their solutions.

Problem 71. (4 points) Explain how the Cycle GAN loss function and architecture avoid learning the identity function.

Problem 72. (4 points) Select all true statements.

a. GAN input is random vector.
b. Discriminator goal is to generate random noise.
c. Discriminator goal is to maximize the loss.
d. Generator receives input from the discriminator.

Problem 73. (4 points) Consider a GAN which successfully produces images of oranges. Which of the following are true?

a. The generator's goal is to learn the distribution of orange images.
b. After GAN training, the discriminator loss reaches a constant value.
c. The generator may produce unseen images of oranges.
d. The discriminator may be used to classify images as oranges or not.

Problem 74. (4 points) What is the accuracy of the discriminator at a global optimum for a trained GAN model given an ideally trained generator?

a. 0.5
b. $p_{\text{data}}/(p_g + p_{\text{data}})$
c. 1
d. None of the above.

Problem 75. (4 points) The goal of the generator is not:

a. Minimize classification error of the discriminator.
b. Maximize classification error of the discriminator.
c. Minimize $\log(1 - D(G(z)))$
d. Maximize $\log(D(G(z)))$

Problem 76. (4 points) The generator is represented as $G(z; \theta_g)$. Which of the following statements are true?

a. $p_z(z)$ represents the probability distribution of real data.
b. θ_g represents the parameters of the discriminator.
c. Input noise is sampled from z.
d. θ_g represents the parameters of the generator.

Problem 77. (4 points) Which of the following describes a correct process of training a GAN discriminator and generator?

a. k steps for the generator followed by one step for the discriminator.
b. One steps for the generator followed by one step for the discriminator.
c. k steps for the discriminator followed by one step for the generator.
d. One steps for discriminator followed by k step for the generator.

Problem 78. (4 points) Generative models learn a:

a. Posterior probability.
b. Joint probability.
c. Prior probability.
d. All of the above.

Problem 79. (4 points) Is the following statement true? The most common problem when training a GAN is a failure to converge which refers to not finding an equilibrium between the discriminator and the generator.

1.8 Variational Autoencoders

Problem 80. (4 points) In principle component analysis (PCA), given data $x_1, x_2 ... x_n \in \mathbb{R}^d$, the goal is to find a linear transformation φ: $\mathbb{R}^d \to \mathbb{R}^k$ ($k \leqslant d$) that best maintains the reconstruction accuracy or equivalently, minimizing the $\mathcal{L}_2$ error. The principal components may be derived in terms of a best-approximating linear subspace. Consider the optimization function:

$$\text{minimize}_{A \in \mathbb{R}^{d \times k}, A^T * A = I_k} \sum_{i=1}^{n} ||x_i - AA^T x_i||_2^2$$

The subspace is defined by the column space of A (with orthonormal columns), and for each point x_i we would like to locate its best approximation in the subspace (in terms of euclidean distance) to find the matrix A such that the goal is satisfied. A solution is given by $\hat{A} = V$, whose columns are the leading k eigenvectors of the matrix $X * X^T$. A single layer autoencoder is a nonlinear version of this problem:

$$\text{minimize}_{W \in \mathbb{R}^{k \times d}} \sum_{i=1}^{n} ||x_i - W' g(W.x_i)||_2^2$$

for a non-linear activation function g. The autoencoder first encodes the input to a latent representation and then decodes the latent representation to reconstruct the input. If the number of units in the hidden layer, latent representation, is less than the input layer then the Autoencoder is undercomplete; otherwise it is overcomplete. In which of the following cases is an Autoencoder reduced to PCA?

a. When the autoencoder is overcomplete and uses nonlinear encoder and decoder networks.
b. When the autoencoder is undercomplete, and uses a linear decoder with mean squared error loss.
c. When the autoencoder is overcomplete, and uses a linear decoder with mean squared error loss.
d. Autoencoders cannot be reduced to PCA.

Problem 81. (4 points) Consider an autoencoder with h hidden units, and input to the autoencoder is a set of d-dimensional unlabeled data, ${x^{(i)}}_{i=1}^{N}$, where W^1 denotes the $d \times h$ weight matrix between the input layer and the hidden layer and W^2 denotes the $h \times d$ weight matrix between the hidden layer and the output layer. f^1 denotes the activation function for hidden layer, and f^2 is the activation

function for the output layer.

$$z_j^{(i)} = \sum_{k=1}^{d} W_{k,j}^1 \cdot x_k^{(i)}$$

$$a_j^{(i)} = f^1(\sum_{k=1}^{d} W_{k,j}^1 \cdot x_k^{(i)})$$

$$t_j^{(i)} = \sum_{k=1}^{H} W_{k,j}^2 \cdot a_k^{(i)}$$

$$\hat{x}_j^{(i)} = f^2(\sum_{k=1}^{H} W_{k,j}^2 \cdot a_k^{(i)})$$

$$J(W^1, W^2)^{(i)} = ||x^{(i)} - \hat{x}^{(i)}||^2 = \sum_{j=1}^{d} (x_j^{(i)} - \hat{x}_j^{(i)})^2$$

is the reconstruction error for example $x^{(i)}$. $J(W^1, W^2) = \sum_{i=1}^{N} J(W^1, W^2)^{(i)}$ is the total reconstruction error. We add 1 to the input and hidden layer so that no bias term is required.
Compute $\frac{\delta J^{(i)}}{\delta W_{k,i}^2}$, $\frac{\delta J^{(i)}}{\delta z_j^{(i)}}$, and $\frac{\delta J^{(i)}}{\delta W_{k,i}^1}$.

Problem 82. (4 points) Given $a = \text{ReLU}((W^1)^T x)$ where W^1 is a 4x3 matrix, and $\tilde{x} = \text{ReLU}((W^2)^T a)$, W^2 is a 3×4 matrix. What values of W^1 and W^2 reconstruct the inputs $x^{(1)} = (0, 1, 0, 1)^T$ and $x^{(2)} = (0, 0, 1, 1)^T$ and $x^{(3)} = (1, 0, 0, 0)^T$?

Problem 83. (4 points) What does an auto-encoder learn about the data?

a. Low dimensional representation of the data.
b. High dimensional representation of the data.
c. No representation of the data is learned.

Problem 84. (4 points) Autoencoders are able to compress the input data in its hidden representation if the input features are:

a. Independent.
b. Unrelated.
c. Correlated.

Problem 85. (4 points) Which of the following is true regarding the MSE reconstruction loss of an autoencoder?

a. It forms distinct clusters in latent space.
b. It is not differentiable and cannot be used for backpropagation.
c. It cannot be optimized with gradient descent.
d. None of the above.

Problem 86. (4 points) For an autoencoder given the input signal x and the reconstructed signal y, which of the following objective functions can we minimize to train its parameters with a gradient descent optimizer?

a. $\mathcal{L}(x, y) = \exp^{-(|x-y|)}$
b. $\mathcal{L}(x, y) = \exp^{(|x-y|)}$
c. $\mathcal{L}(x, y) = -\log(|x - y|)$
d. $\mathcal{L}(x, y) = (x + y)^2$

Problem 87. (4 points) When an autoencoder is used for dimensionality reduction, under what conditions can it learn a more powerful generalization than PCA?

a. When the autoencoder is undercomplete and uses nonlinear encoder and decoder functions.
b. When the autoencoder is undercomplete and uses a linear decoder with an MSE loss function.
c. When the autoencoder is overcomplete and uses a linear decoder with an MSE loss function.
d. When the autoencoder is overcomplete and uses nonlinear encoder/decoder functions.

Problem 88. (4 points) A zero-bias autoencoder has 3 input neurons, 1 hidden neuron and 3 output neurons. If the network is perfectly trained using an input $[2\ 3\ 5]^T$, what would be the values of encoder weights? (Answer in the format $[a\ b\ c]$)

Problem 89. (4 points) A zero-bias autoencoder has 3 input neurons, 1 hidden neuron, and 3 output neurons. If the network is perfectly trained using an input $[2\ 3\ 5]^T$, what would be the values of decoder weights? (Answer in the format $[a\ b\ c]$, transpose assumed)

Problem 90. (4 points) Which is of the following methods may be used for training autoencoders:

a. Training one layer at a time
b. Training encoder first and then decoder
c. End-to-end training

1.9 Transformers

Problem 91. (4 points) For the vectors x_i consider the weighted average $y_i = \sum_j \alpha_{i,j} \cdot x_j$ where $w_{i,j} = x_i^T x_j$ and $\alpha_{i,j} = \text{softmax}(w_{i,j})$. What is $\sum_j \alpha_{i,j}$ for any i?

Problem 92. (4 points) In Seq2Seq models with n words, let $h_1, h_2, ..., h_n \in \mathbb{R}^h$ be the encoder hidden states (h is the embedding dimensionality), $s_t \in \mathbb{R}^h$ be the decoder hidden state at step t, then the attention scores for step t are: $e_t = [(s_t)^T h_1, ..., (s_t)^T h_n] \in \mathbb{R}^n$
Taking the softmax to get the attention distribution α_t for this step: $\alpha_t = \text{softmax}(e_t) \in \mathbb{R}^n$. What is the formula for the attention output a_t?

Problem 93. (4 points) Which of the following architectures cannot be parallelized?

a. CNNs
b. RNNs
c. Transformers

Problem 94. (4 points) What is the optimization objective of BERT?

a. Predicting a masked word.
b. Predicting whether two sentences follow each other.
c. Both a. and b.

Problem 95. (4 points) What is the time complexity of Transformers as a function of sequence length n?

a. $O(n)$
b. $O(n^2)$
c. $O(n^3)$
d. None of the above.

Problem 96. (4 points) Why do we use positional encoding in transformers?

a. Because it helps locate the most important word in the sentence.
b. Because attention encoder outputs depend on the order of the inputs.
c. Because we replaced recurrent connections with attention modules.
d. Because it decreases overfitting in RNN and transformers.

Problem 97. (4 points) What is the time complexity of Transformers as a function of the number of heads h?

a. $O(\log h)$
b. $O(h)$
c. $O(h^2)$
d. None of the above.

Problem 98. (4 points) What is the main difference between the Transformer and the encoder-decoder architecture?

a. Unlike encoder-decoder, transformer uses attention only in encoders.
b. Unlike encoder-decoder, transformer uses attention only in decoders.
c. Unlike encoder-decoder, transformers have multiple encoder-decoder structures layered up together.
d. Unlike encoder-decoder, transformer uses attention in both encoders and decoders.

1.10 Reinforcement Learning

Problem 99. (2 points)
After applying Q-Learning to $q = 6$, what is its value? Let $t = 8$ and $a = 0.2$.

Problem 100. (2 points)
What is the updated Q-value of a tuple (s, a) if $q = 0$, $a = 0.2$, and $t = 2$?

Problem 101. (2 points)
What is a difference between value iteration and Q-Learning?

a. Value iteration is given the transition function and reward function; whereas Q-Learning doesn't.
b. Value iteration is given the transition function and reward function; whereas Q-Learning is only given the reward function
c. Value iteration computes a policy; whereas Q-Learning computes a value function
d. Value iteration computes a value function; whereas Q-Learning computes a policy.

Problem 102. (2 points)
What is a difference between temporal difference (TD) and monte carlo (MC) sampling updates?

a. TD updates the value by a biased estimate looking one step ahead; whereas MC updates the value by a rollout of episodes until termination.
b. MC updates the value by a biased estimate looking one step ahead; whereas TD updates the value by a rollout of episodes until termination.
c. TD performs dynamic programming; whereas MC performs a random walk.
d. TD performs BFS; whereas MC performs DFS.

Problem 103. (2 points)
The Bellman optimality equation for finding Q^*is :

a. Linear and used for evaluating a given policy.
b. Linear and used for finding the best policy.
c. Non-linear and used for evaluating a given policy.
d. Non-linear and used for finding the best policy.

Problem 104. (2 points)
Once we compute Q* we can compute the best policy by:

a. $\arg\max_a Q^*(s,a)$
b. $\arg\min_a Q^*(s,a)$
c. $\arg\max_a Q^*(s,a) - V^*(s)$

Problem 105. (2 points)
A policy maps:

a. Action to state
b. State to action
c. State to reward
d. Reward to action
e. Action to reward

Problem 106. (2 points)
Given 4 states (s_0, s_1, s_2, s_3) and 2 actions (a_1, a_2) in Q-Learning for $t = (s, a, s', r)$ with $\alpha = 0.5$ and $\gamma = 0.9$, where Q is initialized to zeros, we observe $(s_0, a_1, s_2, 0)$. What is $Q(s_0, a_1)$?

Problem 107. (2 points)
In Q-Learning with ε-Greedy, which value of ε may result in not finding the optimal value out of 0, 0.5 and 1?

1.11 Deep Reinforcement Learning

Problem 108. (2 points) **Value Function Approximation**
Neural networks are used for value function approximation in reinforcement learning:

a. Since storing a table of all action-state values is not feasible.
b. Since we would like to generalize to unseen states.
c. Both a and b.
d. To increase the probability of actions that worked well.

Problem 109. (2 points) Prioritized replay samples important transitions from a stored experience buffer more frequently which results in more efficient learning. What criteria determines which transitions are important?

a. TD error
b. MC error
c. DQN error

Problem 110. (2 points) **Actor-Critic**
In actor-critic methods, the following operations are performed in which order:

a. Fit value function using MC or TD learning
b. Update critic parameters
c. Sample trajectory following actor policy
d. Update policy parameters
e. Approximate policy gradient
f. Compute action advantage function

Problem 111. (2 points) **MCTS**
Monte-Carlo tree search selects action based on:

a. Greedy approach.
b. Lower confidence bound.
c. Upper confidence bound.

Problem 112. (2 points) **AlphaZero**
In AlphaZero:

a. The neural network policy converges to a Monte Carlo tree search policy that would not use a neural network.
b. The Monte Carlo tree search uses predictions made by a neural network for guiding the search.

c. The neural network predicts the probabilities of actions given a state and value of a state.
d. All of the above.

Problem 113. (2 points) In AlphaZero self play is between:

a. The network and itself, for example playing both white and black pieces in Chess.
b. A new network and an old network.
c. Both a and b.

Problem 114. (2 points) **Policy Gradient**
Which of the following statements are true?

a. Policy gradient cannot handle discontinuous cost functions.
b. Policy gradient treats environment rewards and dynamics as black boxes.
c. Policy gradient is an example of model-free reinforcement learning, since the agent doesn't try to fit a model of the environment

Problem 115. (2 points) **TRPO**
The Trust region policy optimization (TRPO) optimization objective:

a. Is unconstrained.
b. Constrains the $\mathcal{L}_2$ norm of the policy parameters.
c. Constrains the KL divergence between distributions over policy trajectories.

Problem 116. (2 points) **World Models**
A world model uses the following deep learning architectures:

a. CNN and GAN.
b. RNN and VAE.
c. GNN and Transformer.
d. Autoencoder.