Name

MAP545 : Deep Learning and Optimization - Final Exam
2h - No documents allowed

All questions must be answered with phrases and explanations. Simply providing the right answer won't be enough to obtain the maximum grade. As a result, imprecise or illegible answers may result in a null score at the question. Unless specified otherwise, each question is graded between 0 and 1.

Exercise 1. Deep Learning

1. Describe the operation performed by a neuron of a vintage neural network (MLP).

2. Describe the perceptron algorithm and give the update equation.

3. How can we choose the input layer and the output layer of a neural network?

4. What is backpropagation? Describe a naive approach and explain why backpropagration improves upon it.

5. Can we use the perceptron algorithm to train deep neural networks?

6. What is the main benefit of using a ReLU activation function? What is the main drawback of ReLU? How can you limit this flaw?

7. Explain the vanishing gradient problem.

8. In which setting(s) can we use the 0-1 loss function to train a neural network?

9. Describe the dropout procedure with details (both train and prediction steps).

10. Give four *different* ways of regularizing neural networks (without explaining them).

11. We prefer to use CNN for image classification. Could we use vintage neural networks (MLP) instead? Justify.

12. In a CNN, rank the following layer types from the one that involves the largest number of trainable parameters to the one that involves the smallest number of trainable parameters: Convolutional layer, Pooling layer, Dense layer. Justify.

13. Give two differences between a convolutional layer and a pooling layer.

14. What is the operation performed by a flatten layer? How many parameters does it contain? Where is it used in CNN?
15. Connect each neural network to one of its feature (± 0.2 per correct/incorrect association):

- GoogLeNet
- Xception
- ZFnet
- ResNet
- AlexNet

- Split computations across two GPUs
- Involves shortcut connections
- Decompose a convolution into one convolution that operates on the depth only and one on the spatial dimensions only
- Use 1x1 convolutions to reduce the number of computations
- Use deconvnet to understand feature maps

16. What is the difference between segmentation and object detection?

17. What tasks can be solved with YOLO? What is the main improvement of YOLO compared to previous algorithms?

18. Is it possible to consider more than one hidden layer in a RNN? Give the equation used to compute the hidden state h_t in regular (neither LSTM nor GRU) RNN.

19. A problem called vanishing gradient phenomenon occurs in RNN. What is it exactly?

20. What is truncated backpropagation? What is the interest of using such a procedure? Give a potential drawback.

21. Describe the Stochastic Gradient Descent algorithm.

22. What is the mathematical intuition behind gradient descent update? What is the main mathematical assumption required to make Gradient Descent consistent?

23. Assume that we can prove that for any function f in a class of functions, after k steps of an algorithm, we have $f(x_k) - f(x^*) \leq \rho^k \|x_0 - x_*\|^2$, for x_k the output and x^* the minimum of the function, x_0 the initial point. If I have reduced my error by factor 10^{-3}, compared to the starting point, after 200 iterations, how many more iterations would be necessary to complete a reduction of 10^{-9}?

24. Give an algorithm (other than Gradient Descent) that provides such a rate for a class of function (to be precised). What is the factor ρ?

25. What is the intuition for the choice of the learning rate in SGD? What step-size sequence $(\gamma_t)_{t\geq 0}$ is typically used?

26. What does good/bad conditioning correspond to? Give a mathematical definition of the condition number. What are the 2 main techniques well suited to improve convergence for a poorly conditioned convex function?

27. We consider the case where $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ is convex and smooth but **not** strongly convex. We want to minimize f over $\mathbb{R}^d$. Complete the following comparison of SGD and GD

Name

	GD	SGD
Convergence rate after k iterations	$\frac{1}{k}$	$\frac{1}{\sqrt{k}}$
Complexity per iteration		
Number of iterations to reach precision $\frac{1}{n^{1/4}}$		
Number of iterations to reach precision $\frac{1}{n^3}$		
Total complexity to reach precision $\frac{1}{n^{1/4}}$		
Total complexity to reach precision $\frac{1}{n^3}$		

28. How would you summarize the conclusion of the last two lines of the previous table?

29. What is the main idea behind Nesterov Accelerated Gradient Descent? Is it a first or second-order method?

30. What is the intuition behind variance-reduced algorithms? Give an example of such an algorithm, and describe it.

31. Give one advantage and one drawback of SVRG, compared to SAG.

32. Give the update equation in Newton's method. What happens when this method is applied to quadratic functions?

33. **(3 points)** We consider an optimization problem with the following characteristics: we want to solve

$$\arg\min_{w\in\mathbb{R}^d}\left\{F(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w)\right\}$$

over $\mathbb{R}^d$. 1) What does f_i correspond to in ML? 2) Can you give a precise example of a problem which results in d vary large, n very large, f_i being non-convex, non smooth.

1)

2)

For each of the following methods, explain why it is well suited or not and propose a ranking.

Method	Advantage	Drawback	Ranking
LBFGS			
RMSprop			
SAGA			