Exam Advanced Machine Learning
05 January 2021, 18.30–21.15

This exam consists of 5 problems, each consisting of several questions. All answers should be motivated, including calculations, formulas used, etc. The use of a calculator is not allowed.

Question 1: Short questions
Please provide an argument for your answer on the following questions.

(a) Label the training loss curves A, B, and C with whether they were likely generated with stochastic gradient descent, mini-batch gradient descent, or batch gradient descent.

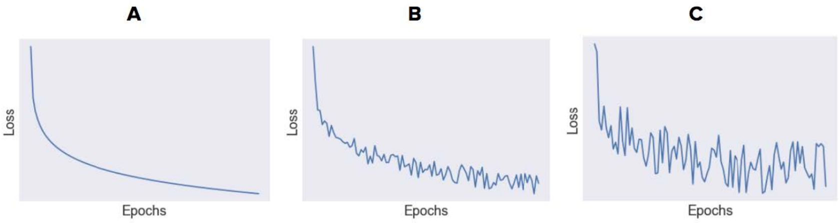

Figure 1: Loss functions.

Batch (A), mini-batch (B), stochastic (C).

(b) Is the following statement true or false? For convex loss functions both stochastic gradient descent and batch gradient descent will eventually converge to the global optimum.

False.

(c) Weight sharing allows convolutional neural networks to deal with image data without using too many parameters. What is the effect of weight sharing on the bias and on the variance of a model?

Weight sharing increases the bias.

(d) You are given a dataset of 10×10 grayscale images. Your goal is to build a 5-class classifier. You have to adopt one of the following two options:

(i) the input is flattened into a 100-dimensional vector, followed by a fully-connected layer with 5 neurons

(ii) the input is directly given to a convolutional layer with five 10×10 filters

Explain which one you would choose.

The two approaches are the same. However, the second one seems better in terms of computational costs (no need to flatten the input). We accept the answer “the two approaches are the same”.

Question 2: Neural networks
You design a fully-connected neural network with 3 hidden layers as depicted in the following figure.

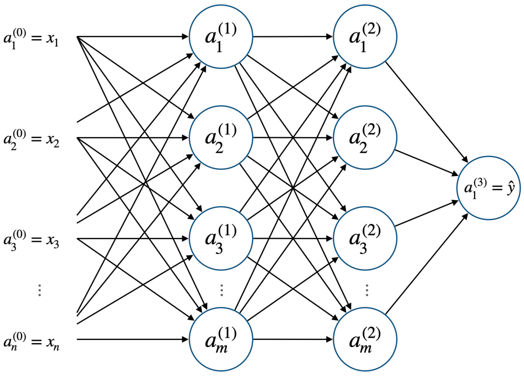

Figure 2: neural network.

(a) You choose all activations functions to be sigmoid functions, and your optimizer is stochastic gradient descent. You initialize all the weights and biases to zero and forward propagate an input $x \in \mathbb{R}^n$ to an output $\hat{y} \in \mathbb{R}$. What is the value of the output $\hat{y}$?

$\hat{y} = 0.5$.

(b) Suppose that the input is 20-dimensional ($n = 20$), and that the hidden layers have 10 hidden units each ($m = 10$). What is the total number of trainable parameters in your network?

The number of parameters is $(20 + 1) \times 10 + (10 + 1) \times 10 + (10 + 1) \times 1 = 331$.

(c) Consider the model defined in question (a) with all parameters initialized with zeros. Let $W^{[1]}$ denote the weight matrix of the first hidden layer. You forward propagate a batch of examples, and then backpropagate the gradients and update the parameters. Which of the following statements is true?

(i) The entries of $W^{[1]}$ may be positive or negative
(ii) The entries of $W^{[1]}$ are all negative
(iii) The entries of $W^{[1]}$ are all positive
(iv) The entries of $W^{[1]}$ are all zeros

Answer (i).

(d) Consider the model defined in question (a) with all activation functions to be sigmoid functions. You initialize the weights in this neural network with large positive numbers. Is this a good idea?

Large W causes Wx to be large. When Wx is large, the gradient is small for sigmoid activation function. Hence, we will encounter the vanishing gradient problem.

(e) Consider the model defined in question (b). Let $W^{[1]} \in \mathbb{R}^{10\times 20}$, $W^{[2]} \in \mathbb{R}^{10\times 10}$, and $W^{[3] \in \mathbb{R}^{1\times 10}}$ denote the weights of the hidden layers. The neural network operates as $z^{(1)} = W^{[1]}x$, $a^{(1)} = f(z^{(1)})$, $z^{(2)} = W^{[2]}a^{(1)}$, $a^{(2)} = g(z^{(2)})$, $z^{(3)} = W^{[3]}a^{(2)}$, and $\hat{y} = h(z^{(3)})$. Choose for f, g, and h the $\tanh$, the sigmoid, and the ReLu function, respectively.

(i) Compute $\partial\hat{y}/\partial w^{[3]}_{1j}$

(ii) Compute $\partial\hat{y}/\partial w^{[2]}_{ji}$

$\partial\hat{y}/\partial w^{[3]}_{1j} = a^{(2)}_j \mathbb{1}(\hat{y} > 0)$.
$\partial\hat{y}/\partial w^{[2]}_{ji} = \sigma(z^{(2)})(1 - \sigma(z^{(2)}))a^{(1)}_i a^{(2)}_j \mathbb{1}(\hat{y} > 0)$.

Question 3: Graphical models
The following figure shows a graphical model over eight binary-valued variables $A, \ldots, H$. We do not know the parameters of the probability distribution associated with the graph.

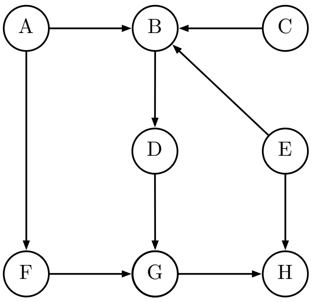

Figure 3: Graphical model.

(a) Write the expression for the joint probability $\mathbb{P}(A, B, C, D, E, F, G, H)$ of the network in its *reduced* factored form.

$\mathbb{P}(A, B, C, D, E, F, G, H) = \mathbb{P}(A)\,\mathbb{P}(B|A, C, E)\,\mathbb{P}(C)\,\mathbb{P}(D|B)\,\mathbb{P}(E)\,\mathbb{P}(F|A)\,\mathbb{P}(G|D, F)$ $\mathbb{P}(H|E, G)$.

(b) Which of the following conditional independence assertions are true?

i) $A \perp\!\!\!\perp B$
ii) $A \perp\!\!\!\perp C$
iii) $A \perp\!\!\!\perp D \,|\, B, H$
iv) $A \perp\!\!\!\perp E \,|\, F$

(i) False,
(ii) True,
(iii) False,
(iv) True.

Question 4: Hidden Markov Models (HMMs)
Bob lives a simple life. Some days he is *Angry* and some days he is *Happy*. However, Bob

hides his true emotional state, and so all you can observe is whether he smiles, frowns, laughs, or yells. We start on day 1 in the *Happy* state, and there is one transition per day.

Bob's behavior can be modelled by a hidden Markov model specified by (π, A, φ) that can output 4 possible values. Thus, the hidden states $z_i \in \{\text{Happy, Angry}\}$, and the output values $x_i \in \{\text{smile, frown, laugh, yell}\}$. The further specification of the hidden Markov model is given as follows:

$$\pi = (1, 0), \quad A = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, \quad \varphi = \begin{pmatrix} 0.5 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.5 & 0.2 & 0.2 \end{pmatrix}.$$

Thus, $\mathbb{P}(z_{t+1} = \text{Angry}|z_t = \text{Happy}) = 0.2$, and $\mathbb{P}(z_{t+1} = \text{Happy}|z_t = \text{Happy}) = 0.8$. But also, $\mathbb{P}(x_t = \text{smile}|z_t = \text{Happy}) = 0.5$, and $\mathbb{P}(x_t = \text{smile}|z_t = \text{Angry}) = 0.1$.

(a) Calculate $\mathbb{P}(z_2 = \text{Happy})$.

$\mathbb{P}(z_2 = \text{Happy}) = 0.8$.

(b) Calculate $\mathbb{P}(x_2 = \text{frown})$.

$\mathbb{P}(x_2 = \text{frown}) = 0.8 \times 0.1 + 0.2 \times 0.5 = 0.18$.

(c) Calculate $\mathbb{P}(z_2 = \text{Happy}|\, x_2 = \text{frown})$.

$\mathbb{P}(z_2 = \text{Happy}|\, x_2 = \text{frown}) = 0.8 \times 0.1/0.18 = 8/18 = 4/9$.

(d) Determine $\mathbb{P}(x_{100} = \text{yell})$.

$\mathbb{P}(x_{100} = \text{yell}) = 0.2$.

(e) Assume that $x_1 = \text{frown}$, $x_2 = \text{frown}$, $x_3 = \text{frown}$, $x_4 = \text{frown}$, and $x_5 = \text{frown}$. What is the most likely sequence of latent states $z_1, \ldots, z_5$?

Happy, Angry, Angry, Angry, Angry.

Question 5: Bayesian linear regression

Suppose that you are doing machine learning with linear regression models using basis functions $\varphi(\cdot)$. Thus, the prediction is given by $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \varphi(\mathbf{x})$. We assume that the data points are drawn independently from the distribution $p(t\,|\,\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t\,|\,y(\mathbf{x}, \mathbf{w}), \sigma^2)$, where σ^2 is the variance. We are applying a Bayesian framework to learn the parameters $\mathbf{w}$. Suppose that we have already observed N data points, then the prior distribution that we start with is given by $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{m}_N, \mathbf{S}_N)$, where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^\top \mathbf{t}/\sigma^2),$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \mathbf{\Phi}^\top \mathbf{\Phi}/\sigma^2,$$

having mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$.

(a) Suppose that you receive an additional data point $(\mathbf{x}_{N+1}, t_{N+1})$. Show by 'completing the squares' that the posterior distribution is again of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{m}_{N+1}, \mathbf{S}_{N+1})$.

(b) Derive the expressions for $\mathbf{m}_{N+1}$ and $\mathbf{S}_{N+1}$.

partial grade	1	2	3	4	5	6
(a)	1	1	1	1	1	1
(b)	2	2	3	1	2	2
(c)	2	2		1	2	2
(d)	1	1		2	1	
(e)	2			1		
(f)	1					

Final grade is: (sum of partial grades) / 4.0 + 1.0