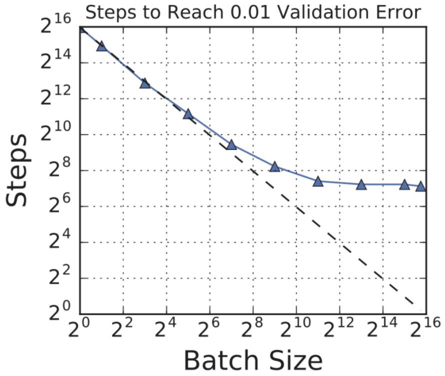
Exam Advanced Machine Learning
11 January 2022, 18.45–21.00

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) The figure below shows the number of stochastic gradient descent (SGD) iterations required to reach a given loss, as a function of the batch size.



For small batch sizes, the number of iterations required to reach the target loss decreases as the batch size increases. Explain this behavior.

(b) You are training a machine learning model with batch gradient descent. Suppose that your training loss increases with the number of iterations. What could be a possible issue with the learning process?

(c) You are training a logistic regression model. You initialize half of the parameters with 0.5, and the other half of the parameters with -0.5. Does this cause any problems? Explain your answer.

(d) You are solving a binary classification task of classifying images as cat versus non-cat. You design a convolutional neural network with a single output neuron. Let the output of this neuron be z. The final output y of your network is given by $y = \sigma(\text{ReLU}(z))$, with σ the sigmoid function. You classify all inputs with final value $y \geq 0.5$ as cat images. What problem are you going to encounter?

(e) In convolutional neural networks, we use weight sharing to deal with image data without using too many parameters. Please explain the effect of weight sharing on the bias and the variance of the model.

(f) We have learned that dense word vectors learned through word2vec or GloVe have many advantages over using sparse one-hot word vectors. Explain if the following statement is an advantage or not. Models using dense word vectors generalize better to rare words than those using sparse vectors.

Question 2: Neural networks

Suppose that you are building a neural network for multi-class classification. The first layer takes the input values $x_1, \dots, x_n$. The second layer transforms the inputs to M hidden units, where unit j is given by $z_j = \sum_{i=1}^{n} w_{ji} x_i + b_j$, where w_{ji} and b_j are the weights and bias terms, respectively. The activation function that is used to generate M outputs is the softmax function, i.e., for output k we have $y_k = \exp(z_k) / \sum_{j=1}^{M} \exp(z_j)$.

(a) Calculate $\partial y_k / \partial z_j$.

(b) Calculate $\partial y_k / \partial w_{ji}$.

(c) The softmax function suffers from numerical instability. One trick may be used when implementing the softmax function. Let $g = \max_{j=1}^{M} z_j$. Then

$$\hat{y}_k = \frac{\exp(z_k - g)}{\sum_{j=1}^{M} \exp(z_j - g)}$$

resolves the problem. State what the numerical problem with the initial softmax computation is, and why the modified formula would help resolving that problem.

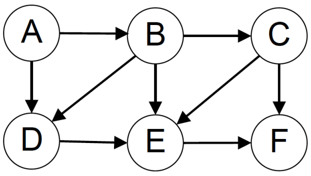
(d) Recall the ReLU(x) $= \max\{x, 0\}$ activation function. Consider an alternative to the ReLU function called the Exponential Linear Unit (ELU) given by

$$\mathrm{ELU}(x) = \begin{cases} x, & x \geq 0, \\ \alpha(e^x - 1), & x < 0. \end{cases}$$

Name one major advantage of using ELU over ReLU.

Question 3: Graphical models

The following figure shows a graphical model over six binary-valued variables $A, \dots, F$. We do not know the parameters of the probability distribution associated with the graph.



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

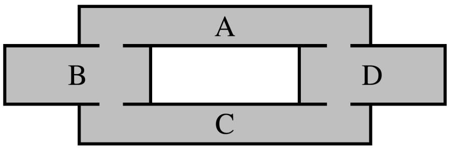
(b) Which of the following conditional independence assertions are true?
i) $A \perp\!\!\!\perp F \mid D$
ii) $C \perp\!\!\!\perp D \mid E$
iii) $A \perp\!\!\!\perp E$

(c) For this network, we want the following relations to hold:
i) $A \perp\!\!\!\perp D \mid B$
ii) $A \perp\!\!\!\perp F \mid C$
iii) $C \perp\!\!\!\perp D \mid B$

What is the minimum set of edges that must be removed such that the relations hold simultaneously. Explicitly state the edges that need to be removed.

Question 4: Hidden Markov Models (HMMs)

Imagine you have a smart house that wants to track your location within itself so it can turn on the lights in the room you are in and make your food in your kitchen. Your house has 4 rooms (A, B, C, D) in the floor plan below (A is connected to B and D, B is connected to A and C, C is connected to B and D, and D is connected to A and C).



At the beginning of the day ($t = 0$), your probabilities of being in each room are p_A, p_B, p_C, and p_D for rooms A, B, C, and D, respectively. At each time t, your position (following a Markovian process) is given by X_t. At each time, your probability of staying in the same room is q_0, your probability of moving clockwise to the next room is q_1, and your probability of moving counterclockwise to the next room is $q_{-1} = 1 - q_0 - q_1$.

(a) Initially, assume your house has no way of sensing where you are. What is the probability that you will be in room D at time $t = 1$?

Now assume that your house contains a sensor M_A that detects motion ($M_A = 1$) or no motion ($M_A = 0$) in room A. This sensor has probability $1 - 2\gamma$ for detecting motion if you are in room A, for γ small. The sensor is a bit noisy and can be tricked in adjacent rooms with probability γ, resulting in the conditional distributions for the sensor given as follows.

$$\begin{aligned}\mathbb{P}(M_A = 1\,|X = A) &= 1 - 2\gamma = 1 - \mathbb{P}(M_A = 0\,|X = A),\\ \mathbb{P}(M_A = 1\,|X = B) &= \gamma = 1 - \mathbb{P}(M_A = 0\,|X = B),\\ \mathbb{P}(M_A = 1\,|X = C) &= 0 = 1 - \mathbb{P}(M_A = 0\,|X = C),\\ \mathbb{P}(M_A = 1\,|X = D) &= \gamma = 1 - \mathbb{P}(M_A = 0\,|X = D)\end{aligned}$$

(b) Model this process as a Hidden Markov Model, where the observations are the sensor readings and the latent variables are the locations. Specify all the parameters (π, A, φ).

(c) Suppose that you start in room A with $q_0 = 0$ and $q_1 = 1/2 = q_{-1}$. Explain which observation is the most likely observation: $(1, 0, 0)$, $(1, 1, 0)$, or $(1, 1, 1)$, where the numbers represent the readings of M_A at $t = 0, 1$, and 2, respectively.

Question 5: Optimization in machine learning

You are given a training set $\{(t_i, \mathbf{x}_i)\}_{i=1,\ldots,N}$ with N data points. You decide to use a machine learning algorithm with weights $\mathbf{w}$ that transforms an input $\mathbf{x}$ to an output t. You suppose that the relationship is given by $t = y(\mathbf{x}, \mathbf{w}) + \varepsilon$, with ε having a Gaussian distribution with variance σ^2. More specifically, $\mathbb{P}(\varepsilon\,|\,\sigma^2) = \mathcal{N}(\varepsilon\,|\,0, \sigma^2)$ with $\mathcal{N}(x\,|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/(2\sigma^2)\}$.

(a) Write down the likelihood function $\mathbb{P}(\mathbf{t}\,|\,\mathbf{X}, \mathbf{w}, \sigma^2)$ using all data.

(b) Show that the log likelihood function is given by

$$-\frac{N\ln(\sigma^2)}{2} - \frac{N\ln(2\pi)}{2} - \frac{\sum_{i=1}^{N}(t_i - y(\mathbf{x}_i, \mathbf{w}))^2}{2\sigma^2}.$$

(c) Show that maximizing the log likelihood function is equivalent to minimizing the root of the mean-squared error.

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

Final grade is: (sum of partial grades) / 3.0 + 1.0