

Exam 2023/24 - first take

1. Consider two consecutive layers x, y of size n in an MLP. For simplicity we assume no biases and no activations, so these layers are related by $y = Ax$, where A is an $n \times n$ matrix. When computing the gradient of loss function L over x by backpropagation, which formula should you use:
 - a. $dL/dx = A dL/dy$
 - b. $dL/dx = A^{-1} dL/dy$
 - c. $dL/dx = A^T dL/dy$ (V)
 - d. $dL/dx = (A^{-1})^T dL/dy$
2. Which of the following optimizers include a gradient momentum term in updating weights:
 - a. SGD (N)
 - b. Nesterov's momentum SGD (Y)
 - c. Adagrad (N)
 - d. Adam (Y)
3. Consider two neural networks N_1, N_2 with equivalent architectures, i.e. for any setting of weights of N_1 there exists a setting of weights of N_2 that computes the same function, and vice versa. We train these two networks with the same loss, same dataset, and starting points corresponding to the same function. We use SGD using entire dataset each step, with the same learning rate for both networks. Will we obtain equivalent points in N_1 and N_2 in after each step?
 - a. Yes, even if we use L_2 -regularization (N)
 - b. Yes, but we might not if we use L_2 -regularization (N)
 - c. Not necessarily, even without any regularization (Y)
4. Consider a simple convnet with three subsequent convolutional layers with filters of size $3 \times 3, 5 \times 5$ and 3×3 respectively (each with stride 1). What is the size of the effective receptive field of each neuron (square area).
 - a. 81
5. Consider a GAN training setup. Which of the following are true:
 - a. In a discriminator training step, we need to compute the gradient of the generator. (N)
 - b. In a generator training step, we need to compute the gradient of the discriminator. (Y)
 - c. In a generator training step we need access to a fully differentiable random number generator (N).
 - d. None of the above (N).
6. Consider a feed-forward layer in a standard transformer architecture (with two MLPs). If we double the size of the internal representation called d_{ff} or $inner_dim$ (keeping other aspects of the architecture constant), the number of parameters of this layers (ignoring biases) will increase by a factor of :
 - a. 2, (Y)

- b. 4, (N)
 - c. 8, (N)
7. Consider a layer with dropout applied to it with rate p during training, i.e. every neuron is “removed” with probability p . In such a scenario, at test time, each neuron output for this layer is multiplied by a constant. This constant is:
- a. p
 - b. $1-p$ (Y)
 - c. $1/p$
 - d. $1/(1-p)$
8. Consider the following 4x6 grid. The agent starts in the lower left corner and can move only up and to the right. The game ends when the agent reaches the top right corner. When stepping on a field with A a reward +3 is generated, stepping on B generates reward -1, stepping on C generates -100. Which of the following sentences are true:
- a. All the states have the same V-value (N).
 - b. All the policies generate the same expected reward (Y).
 - c. There are more states with positive V-value than with a negative V-value (Y).
 - d. None of the above (N).

A	B	B	B	B	A
C	A	B	B	B	B
A	C	A	B	B	B
	A	C	A	B	B

Exam 2023/24 - retake

1. Consider a convolutional layer without biases which has 25 filters of size 4x4, with 4 input channels. What is the number of parameters of this layer?
 - a. 400 (n)
 - b. 1600 (y)
 - c. 64 (n)
2. Consider a fully convolutional neural network. If we swap the order of layers, then the size of the effective receptive field is:
 - a. always going to stay the same (Y),
 - b. always change (N),
 - c. potentially change, depending on the details of the architecture (N).

3. Does it make sense to use L2-regularization in a neural network that uses Batch Normalization (BN) after each layer?
 - a. Yes, both L2-regularization and BN will work fine in this setting. (N)
 - b. No, BN makes actual values of weights meaningless, so using L2-regularization is pointless. (N)
 - c. It does, however instead of preventing overfitting, in presence of BN, L2-regularization controls effective learning rate. (Y)*
4. Which of the equalities hold (* denotes values for an optimal strategy)?
 - a. $v^*(s) = \max_a q^*(s,a)$
 - b. $v^*(s) = \min_a q^*(s,a)$
 - c. $v^*(s) = \sum_a q^*(s,a)$
5. Consider a regular MLP (multi-layer perceptron) architecture with 10 fully connected layers with ReLU activation function. The input to the network is a vector of size 100, where each dimension has zero mean and standard deviation equal to 1 across the dataset.

Each hidden layer has 10000 neurons and has weights initialized from a normal distribution with zero mean and variance 0.01. Which of the following options is the most likely?

- * The gradients will be exploding (y)
- * The gradients will be vanishing (n)
- * Neither. (n)

Exam 2022/23 - retake

1. Let N be a neural network with weights that minimize a loss function L , and let N_1 be a network with identical architecture and weights that minimize $L + l_1$ penalty on weights. Then
 - * value of L for N_1 is at least as high as for N (YES)
 - * every weight in N_1 has absolute value no bigger than in N (NO)
2. Consider a neural network with 2000 weights and 200 biases. For each of the algorithms below, how many additional numbers (i.e. ignoring the actual values of weights and biases) need to be stored between the epochs.

- * Minibatch Gradient Descent with batch size 10: (0)
- * Minibatch Gradient Descent with momentum and batch size 10: (2200)
- * Adam with batch size 10: (4400)

3. Consider a convolutional layer with biases which has 10 filters of size 5x5, with 4 input channels. What is the number of parameters of this layer?

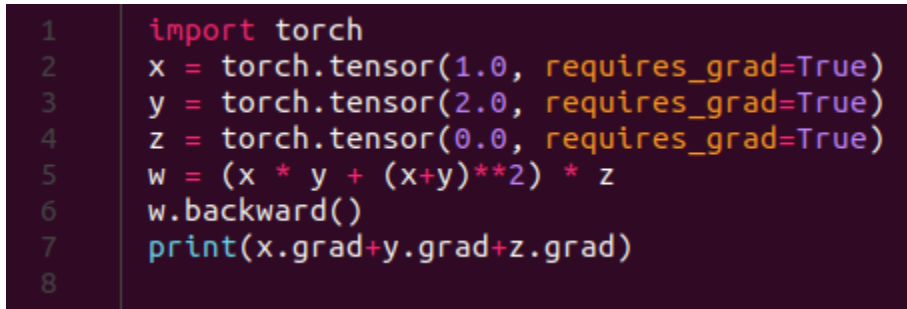
Answer: 1010

4. A convolutional neural network consists of a sequence of 10 convolutions with kernels of size 5x5 each (padding same). What is the number of pixels (area) of the effective receptive field of a single output pixel. You can ignore the boundaries and focus on the middle of an image.

Answer: $(5 + 9 * 4)^2 = 1681$

5. What the following piece of code will produce?

```
import torch
x = torch.tensor(1.0, requires_grad=True)
y = torch.tensor(2.0, requires_grad=True)
z = torch.tensor(0.0, requires_grad=True)
w = (x * y + (x+y)**2) * z
w.backward()
print(x.grad+y.grad+z.grad)
```



Answer: 11 (?)

6. Consider the following 4x6 grid. The agent starts in the lower left corner and can move only up and to the right. The game ends when the agent reaches the top right corner. When stepping on a field with A a reward +1 is generated, stepping on B generates reward -2. Consider a discount factor 0.5 (reward generated after the first move is worth 1.0, after the second move is worth 0.5, then 0.25, etc). Which state has the highest value of the V-function? (indices are 0-based, top left corner has coordinates (0,0))

A	B	B	B	B	A
B	B	B	B	B	B
A	B	B	B	B	B
	A	B	B	B	B

- a) (3,0)
- b) (0,0)
- c) (0,4)
- d) (2,0)

Answer: (0,4)

7. Which of the following are key components of model based reinforcement learning methods that differentiate them from model-free reinforcement learning.

- Learning a transition model
- Learning a q-function
- Learning a reward function

Answer: T, F, T

8 - Consider a regular MLP (multi-layer perceptron) architecture with 15 fully connected layers with ReLU activation function. The input to the network is a vector of size 1000, where each dimension has zero mean and standard deviation equal to 1 across the dataset.

Each hidden layer has 1000 neurons and has weights initialized from a normal distribution with zero mean and standard deviation 0.01. Which of the following options is the most likely?

- * The gradients will be exploding
- * The gradients will be vanishing
- * Neither.

Answer: the gradients will be exploding.

Exam 2022/23 - first take

1. Consider the following architectures:

- Input - linear layer with 100 neurons - relu - output
- Input - linear layer with 1000 neurons - linear layer with 100 neurons - relu - output

Which of the following scenarios are possible when the model is trained to convergence using SGD(answer true or false):

- a) The training error of the first model is lower than the second model.
- b) The training error of the second model is lower than the first model.

Answer: true + true

2. Consider running the batch Gradient Descent (GD) algorithm with loss function L . Let x denote the current solution.

Then (answer true or false):

- in each step of GD, the value of $L(x)$ does not increase
- in each step of GD, the distance of x to some local optimum of L does not increase
- it is possible that the gradient of L is zero with respect to all the variables.

Answer: F, F, T

3. Which of the following strategies were designed to preventing overfitting of a neural network

- Dropout.
- Batch-norm.
- Reparametrization trick.
- Using ReLU instead of sigmoid.

Answer: T, T, F, F.

4. Consider a convolutional layer without biases which has 50 filters of size 1×1 , with 8 input channels. What is the number of parameters of this layer?

Answer: 400

5. Consider a batchnorm layer that follows a linear layer with 100 neurons, how many trainable parameters (in the default setting) does this layer have?

Answer: 200

6. A convolutional neural network consists of a sequence of 3 convolutions with kernels of size 3×3 , 5×5 and 7×7 respectively (padding same). What is the number of pixels (area) of the effective receptive field of a single output pixel. You can ignore the boundaries and focus on the middle of an image.

Answer: $13 * 13 = 169$

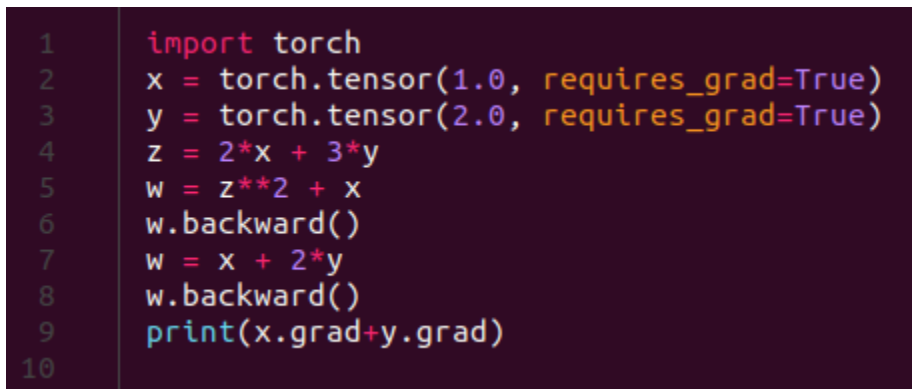
7. Which of the following mechanisms were created to avoid permutation invariance in sequence processing with self-attention?

- Positional encoding.
- Batch normalization.
- 1×1 convolution.

Answer: T, F, F

8. What the following piece of code will produce?

```
import torch
x = torch.tensor(1.0, requires_grad=True)
y = torch.tensor(2.0, requires_grad=True)
z = 2*x + 3*y
w = z**2 + x
w.backward()
w = x + 2*y
w.backward()
print(x.grad+y.grad)
```



```
1  import torch
2  x = torch.tensor(1.0, requires_grad=True)
3  y = torch.tensor(2.0, requires_grad=True)
4  z = 2*x + 3*y
5  w = z**2 + x
6  w.backward()
7  w = x + 2*y
8  w.backward()
9  print(x.grad+y.grad)
10
```

Answer: 84

Exam 2021/22

1. Every gradient descent algorithm has a certain number of hyperparameters, e.g. the batch size, etc. How many hyperparameters do these algorithms have (ignore parameters not directly related to the learning process, like regularization, dropout, etc.): Gradient Descent, Momentum Gradient Descent, Adagrad, RMSProp (the minibatch version in all cases)?
 - a. 2, 3, 3, 3
 - b. 2, 3, 2, 3**
 - c. 2, 3, 3, 4
2. Consider a convolutional layer with biases which has 16 filters of size 3x3, with 8 input channels. What is the number of parameters of this layer?
 - a. 1153
 - b. 1168**
 - c. 1280

3. Consider a batch norm in a convnet, the input of which has 8 channels, each of size 16x16. How many parameters this layer has?
 - a. 8
 - b. 16**
 - c. 256
4. Consider a layer of with dropout applied to it with rate p , i.e. every neuron is “removed” with probability p . In typical implementations, at test time, each neuron output for this layer is multiplied by a constant. This constant is:
 - a. p
 - b. $1-p$**
 - c. $1/p$
 - d. $1/(1-p)$
5. Consider a regular MLP (multi-layer perceptron) architecture with 10 fully connected layers with ReLU activation function. The input to the network is a vector of size 100, where each dimension has zero mean and standard deviation equal to 1 across the dataset.

Each hidden layer has 100 neurons and has weights initialized from a normal distribution with zero mean and standard deviation 0.001. Which of the following options is the most likely?

- * The gradients will be exploding
- * The gradients will be vanishing**
- * Neither.

6. Consider a regular RNN layer **without biases**, with hidden state of size 64, its input elements being vectors of size 8, output elements being vectors of size 8. How many parameters this layer has?
 - a. 1024
 - b. 4096
 - c. 5120**
7. Consider a simple game and a parameterized policy that governs the behaviour of an agent:
 In each turn the agent can pick action A or action B. There are four turns total. The reward is given at the end of the episode and it equals the number of times the action A was chosen.
 The policy we are considering has a single parameter x and this policy in each turn takes action A with probability x and action B otherwise.

What is the gradient of the expected reward with respect to the parameter x evaluated at $x=0.5$?

Correct answer: 4

8. Consider the following game. There are two bandits (slot machines) A and B. If you play A then you get 10 credits with probability 0.5 and 0 credits otherwise. If you play B then you get 20 credits with probability 0.2 and 0 credits otherwise. There are 3 turns total and in each turn you can decide which bandit to use. At the end of the episode you get a reward of 1 if and only if you have collected at least 20 credits.
- What is the q-value $q^*(A)$ in the initial state?
- Assume there is no discount factor ($\gamma=1$).

Correct answer: 0.555

Exam 2020/21

6. Let x be a vector of real numbers. Which of the following operations are equivalent to $\text{softmax}(x)$:
- $\text{softmax}(x+(1,\dots,1))$
 - $\text{softmax}(2 * x)$
 - $\text{softmax}(\text{sigmoid}(x))$
7. Consider a convolutional layer with biases which has 10 filters of size 5×5 , with 4 input channels. What is the number of parameters of this layer?
- 100
 - 110
 - 140
8. Consider a batch norm in a convnet, the input of which has 10 channels, each of size 100×100 . How many parameters this layer has?
- 10
 - 20
 - 100
9. Consider a regular MLP (multi-layer perceptron) architecture with 10 fully connected layers with ReLU activation function. The input to the network is a vector of size 1000, where each dimension has zero mean and standard deviation equal to 1 across the dataset.

Each hidden layer has 1000 neurons and is initialized from a normal distribution with zero mean and standard deviation 0.01. Which of the following options is the most likely?

- * the gradients will be exploding
- * The gradients will be vanishing
- * Neither.

10. Consider a regular RNN layer **without biases**, with hidden state of size 100, its input elements being vectors of size 10, output elements being vectors of size 10. How many parameters this layer has?

- a. 10200
- b. 11000
- c. 12000

11. Consider a reinforcement learning setting, where the state is represented as two features x_1, x_2 , there are two possible actions a_1, a_2 , and the agent picks the action a_1 with probability $\max(0, \min(1, f_1 * x_1 + f_2 * x_2))$ and action a_2 otherwise. Assume only one episode with one turn where the starting state is $(x_1=2, x_2=-1)$, the reward given to the agent after performing action a_1 is 10, while after performing action a_2 is 0. Consider the gradient of the expected reward of a policy with respect to the parameters f_1, f_2 for two cases (i) $f_1=1, f_2=1.5$ (ii) $f_1=1, f_2=1.25$. In which of the cases (i) or (ii) the norm of the gradient is larger?

- a. (i)
- b. (ii)
- c. They are equal.

12. Consider the following 4 datapoints for a classification problem into 2 classes:

(input=(0,0), label=0)

(input=(1,1), label=0)

(input=(0,1), label=1)

(input=(1,0), label=1)

Which of the following MLP architectures can achieve training error of value 0?

- a. Input -> dense layer with 2 neurons -> softmax
- b. Input -> dense layer with 1 neuron -> sigmoid -> dense layer with 2 neurons -> softmax
- c. Input -> dense layer with 2 neurons -> sigmoid -> dense layer with 2 neurons -> softmax

8. Consider a game with 25 coins on the table. In each turn the player can either:

- Take r coins, where $0 \leq r \leq \min(10, \text{\#coins left})$ and receive a reward of $r-2$.
- End the game - no additional rewards.

What is the q-value $q^*(a_4)$ (where a_4 represents the action of taking 4 coins in the first turn)?

Assume there is no discount factor ($\gamma=1$).

- a. 16
- b. 17
- c. 18
- d. 19

