UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	IN3310/IN4310/IN5400/IN9400 —
Day of examination:	June 15, 2023
Examination hours:	9:00 – 13:00
This exercise set consists of 16 pages.	
Appendices:	None
Permitted aids:	None

General information:

- Read the entire exercise text before you start solving the exercises. Please check that the exam paper is complete.
- Remember that your exam answers must be anonymous; do not state either your name or that of fellow students.
- Most of your text answers should include a discussion and be brief, typically at most a few sentences.
- When you do calculations to answer an exercise, include the intermediate steps in your answer.
- If you lack information in the exam text or think that some information is missing, you may make your own assumptions, as long as they are justifiable within the context of the exercise. In such a case, you should make it clear what assumptions you have made.
- Plan your time so that you can try to answer as many subtasks as possible. Each subtask is normally weighted equally.

Cheating on school exams

"For school exams, it is considered as cheating to use support materials, unless it is explicitly stated in the exam question that this is permitted. Having access to illegal support materials may also be considered cheating, even if they are not used.

All communication between candidates in the exam room is prohibited. Contact with other candidates or other persons during trips to the lavatory and breaks is also prohibited. Mobile phones and other electronic equipment should be turned off and packed away."

`https://www.uio.no/english/studies/examinations/cheating/index.html`

(Continued on page 2.)

Exercise 1

Consider the logistic regression classifier:

$$f(x) = \sigma(\sum_{d=0}^{D-1} w_d x_d + b) \tag{1}$$

where:

$$\sigma(x) = 1/(1 + e^{-x}) \tag{2}$$

1a

The following question is multi-select question, that is, **more than one choice can be true**. What is true? Briefly explain your selection(s)!

1. The set $\{x : f(x) = const\}$ is parallel to w
2. The set $\{x : f(x) = const\}$ is orthogonal to w
3. If x is such that $\sum_{d=0}^{D-1} w_d x_d = -3b$, and we consider $z = x + 3bw/\|w\|$, then $f(z) = 0.5$
4. If x is such that $\sum_{d=0}^{D-1} w_d x_d = -3b$, and we consider $z = x + 2bw/\|w\|^2$, then $f(z) = 0.5$
5. If x is such that $\sum_{d=0}^{D-1} w_d x_d = -3b$, and we consider $z = x + 3bw/\|w\|^2$, then $f(z) = 0.5$

(Continued on page 3.)

Exercise 2

2a

What is the essential difference between a fully connected layer and a 1-D convolution layer with respect to computing outputs? Answer in 3 sentences at most, but name the difference specific to a 1-D convolution.

2b

A 2-dimensional convolutional layer applies filters with spatial kernel size $(5, 5)$, stride 4, and a padding of 5 to an input of shape (112,224). What will be the spatial output shape?

2c

Assume the 2-dimensional convolutional layer, applying filters with spatial size $(4, 4)$, stride 2, and a padding of 3, has 100 input channels and 30 output channels, and no bias terms. What is the number of trainable parameters in this layer?

(Continued on page 4.)

Exercise 3

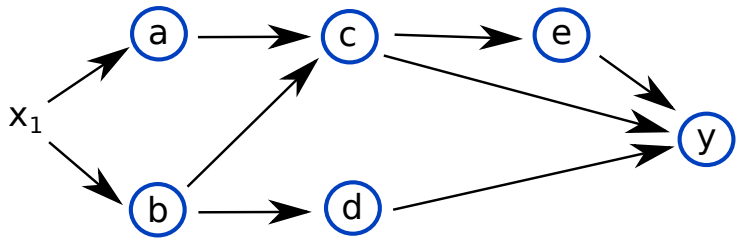
3a

What is true? Briefly explain your selection(s)!

1. Backpropagation is the method of finite differences executed on a graph.
2. Backpropagation computes directional derivatives on a graph.
3. Backpropagation is chainrule executed on a graph.

3b

Consider the following neural network:



The equation for any neuron $k \in \{a, b, c, \ldots\}$ is:

$$n\left(s + \sum_l v_l x_l + \sum_k w_k k\right) \tag{3}$$

where $n(\cdot)$ is some activation function, s is the bias term, $v_l = 0$ for neurons that have no direct inputs x_l, and $w_k = 0$ for neurons that have no inputs from other neurons.

This task asks you to write down expressions for gradients of this network. Please consider the following advices before doing so:

- Write the expression in terms of:
  - $\frac{\partial k_1}{\partial k_2}$, where k_2 is direct input to k_1, and
  - $\frac{\partial k}{\partial x_l}$, where x_l is input to the network.
- This task is about the sequences of partial derivatives along the edges.
- You **do not need** to plug in or compute how $\frac{\partial k_1}{\partial k_2}$ or $\frac{\partial k}{\partial x_l}$ looks like.
- You **do not need** to multiply out terms in parentheses, so $(a + b)c$ or $((a + b)c + (d + e)f)g$ is fine to keep like that.

Write down an expression for the following gradients of this network:

(Continued on page 5.)

- $\frac{dc}{dx_1}$
- $\frac{dy}{dx_1}$

(Continued on page 6.)

Exercise 4

4a

Let $C(x)$ be the output of a stack of convolution layers C, computed from an input feature map x. What does a residual connection compute? Briefly explain your selection(s)!

1. $\sigma(C(x)) * C(x)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$ is element-wise sigmoid weighting
2. $C(x) + x$
3. `avgpool( torch.cat( (C(x),x), dim=(1) ) )`
4. $C(x) - x$
5. $C(x) * x$

4b IN9400 students only

Answers from students in other courses will be ignored :-)

Name one advantage of using residual connections with respect to gradients. 3 sentences max.

(Continued on page 7.)

Exercise 5

Consider the following optimizers:

1. SGD
2. SGD with momentum
3. RMSProp
4. AdamW

5a

Which of these are using normalization of gradients? Briefly explain your selection(s)!

5b

Which of these are using moving averages of gradients which are not applied as normalization? Briefly explain your selection(s)!

(Continued on page 8.)

Exercise 6

6a

Consider the following code:

```
def train_epoch(model,  trainloader,  losscriterion, device, optimizer):

        model.train()
        model = model.to(device)

        losses = []
        for batch_idx, data in enumerate(trainloader):

                inputs=data['image'].to(device)
                labels=data['label'].to(device)

                output = model(inputs)
                loss = losscriterion(output, labels)

                loss.backward()
                optimizer.step()

                losses.append(loss.item())
        return losses
```

Which one line of code is missing in the code above?

(Continued on page 9.)

Exercise 7

7a

- How is a test subset obtained?
- How does a test subset differ from an external dataset?

7b

Consider the following scenario: You train a model and get a good enough accuracy on the validation subset, but the accuracy on the test subset is lower than expected. So you change your training settings and retrain the model. You keep changing settings and retraining until the accuracy is good enough on the test subset. Do you expect that the accuracy is similarly better when the model is applied on new data as you observed for the test subset? Argue in at most 3 sentences.

7c IN5400 and IN9400 students only

Answers from students in other courses will be ignored :-)

One can facilitate a neural network to generalize better by controlling its capacity. Mention any two methods to control a neural network's capacity.

(Continued on page 10.)

Exercise 8

Figure 1: Various Image Augmentations

Figure 1 illustrates the following image transformations:

- Mixup.
- Solarization.
- Affine transformation.
- Crop.

8a

Match the names of the transformations with the corresponding image.

8b

Classify each transformation as either **geometric**, **photometric**, or **other**. Provide a brief justification for your classification.

(Continued on page 11.)

Exercise 9

9a

Describe two methods used to overcome the exploding gradients problem.

9b

Describe any two input-output structures of RNNs. Give an example/application (one for each input-output structure) of where they can be used.

(Continued on page 12.)

Exercise 10

Adversarial attacks can broadly be classified as:

- white box vs. black box attacks,
- targeted vs. untargeted attacks.

10a

Describe what characterizes these attack types.

10b

Briefly explain the Projected Gradient Descent (PGD) method for constructing adverserial attacks, and discuss the applicability of PGD for the aforementioned types of attacks.

(Continued on page 13.)

Exercise 11

11a

Why do CNN architectures like single-shot multibox detector (SSD) and Feature Pyramid Network (FPN) use many features maps from many different layers of the neural network to detect object of different sizes?

11b

Briefly explain how non-maximum suppression (NMS) tries to solve the problem of an object detector predicting multiple overlapping boxes for an object. (There is no need to write the NMS algorithm in detail.)

(Continued on page 14.)

Exercise 12

12a

Given an input matrix of size 3x2 (3 rows and 2 columns), what would be the dimensions of the output when applying transposed convolution with kernel size 2x2, stride=2 and no padding?

12b

What is the difference between semantic segmentation and instance segmentation?

(Continued on page 15.)

Exercise 13

13a IN3310 and IN4310 students only

Answers from students in other courses will be ignored :-)

How does a transformer decoder block use the transformer encoder's output?

13b IN4310 students only

Answers from students in other courses will be ignored :-)

What problem would we face if we were to use Transformers directly on image pixels by treating each pixel as a separate token, i.e., by treating an image as a sequence of pixels? How does the Vision Transformer (ViT) overcome that problem?

(Continued on page 16.)

Exercise 14

14a IN5400 and IN9400 students only

Answers from students in other courses will be ignored :-)

GANs can overfit to data especially when the dataset is small enough. One way to solve that problem is to use data augmentation. But data augmentation can lead to a problem. What is that problem and how can we overcome that problem?

14b IN9400 students only

Answers from students in other courses will be ignored :-)

In progressive growing of GANs, how is each new convolution layer incorporated (phased in) slowly?