Chair of Visual Computing
Department of Informatics
Technical University of Munich

Esolution
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Introduction to Deep Learning

Exam:	IN2346 / endterm	**Date:**	Tuesday 8th February, 2022
Examiner:	Prof. Dr. Matthias Nießner	**Time:**	15:00 – 11:30

- The blackened exam has the same layout as the non-blackened exam with the acutal questions, which is going to be released once the working time starts.
- Only submit your personalized blackened exam. **DO NOT submit the non-blackened/non-personalized exam** (clearly indicated with “DO NOT SCAN/UPLOAD”).
- This final exam consists of **16 pages** with a total of **7 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this simulation is **90 credits**.
- No additional resources are allowed.

Problem 1 Multiple Choice (18 credits)

Mark correct answers with a cross ☒
To undo a cross, completely fill out the answer option ■
To re-mark an option, use a human-readable marking ×■

Please note:

- For all multiple choice questions any number of answers, i.e. either zero (!), one or multiple answers can be correct.
- For each question, you'll receive 2 points if all boxes are answered correctly (i.e. correct answers are checked, wrong answers are not checked) and 0 otherwise.

1.1 You are training a network to classify images of handwritten digits in the range of [0,...,9] on the MNIST dataset. Which of the following data augmentation techniques are suitable to use for this task?

- [x] Add Gaussian noise to the images
- [ ] Vertically flip the images
- [x] Rotation of the images by 10 degrees
- [x] Change the contrast of the images

1.2 What is true about Residual Blocks?

- [ ] Reduce the number of computations in the forward pass
- [x] Act as a highway for gradient flow
- [x] Enable a more stable training of larger networks
- [ ] Act as a regularizer

1.3 For a fully-convolutional 2D CNN, if we double the spatial dimensions of input images, ...

- [ ] ... the number of network parameters doubles
- [x] ... the number of network parameters stays the same
- [ ] ... the receptive field of an arbitrary pixel in an intermediate activation map can decrease
- [ ] ... the dropout coefficient p must be corrected to $\sqrt{p}$ in test time

1.4 What is true about Generative Adversarial Networks?

- [x] The Generator minimizes the probability that the Discriminator is correct
- [ ] They learn the real dataset's distribution directly.
- [x] The Discriminator acts as a classifier
- [ ] The Discriminator samples from a latent space

1.5 Given input x , which of the following statements are always true? Note: For dropout, assume the same set of neurons are chosen.

- [ ] BatchNorm(ReLU(x)) $\equiv$ ReLU(BatchNorm(x))
- [x] Dropout(ReLU(x)) $\equiv$ ReLU(Dropout(x))
- [x] MaxPool(ReLU(x)) $\equiv$ ReLU(MaxPool(x))
- [ ] ReLU(Sigmoid(x)) $\equiv$ Sigmoid(ReLU(x))

1.6 When you are using a deep CNN to train a semantic segmentation model, which of the following can be chosen to help with overfitting issues?

- [ ] Decrease the weight decay parameter
- [x] Increase the probability of switching off neurons in dropout
- [x] Apply random Gaussian noise to the input images
- [ ] Remove parts of the validation set

1.7 In terms of (full-batch) gradient descent (GD) and (mini-batch) stochastic gradient descent (SGD), which of the following statements are true?

- [ ] The computed gradient of the loss w.r.t model parameters in SGD is equal to the computed gradient in GD
- [x] The expected gradient of the loss w.r.t model parameters in SGD is equal to the expected gradient in GD over the same images
- [x] There exists some batch size, for which the gradient of the loss w.r.t model parameters in SGD is equal to the gradient in GD
- [ ] SGD and GD will converge to the same model parameters, but SGD requires less memory at the expense of more iterations

1.8 What is true about batch normalization assuming your train and test set are sampled from the same distribution?

- [ ] Batch normalization cannot be used together with dropout
- [x] Batch normalization makes the gradients more stable, so we can train deeper networks
- [ ] At test time, Batch normalization uses a mean and variance computed on test set samples to normalize the data
- [x] Batch normalization has learnable parameters

1.9 What is true for common architectures like VGG-16 or LeNet? (check all that apply)

- [x] The number of filters tends to increase as we go deeper into the network
- [ ] The width and height of the activation maps tends to increase as we go deeper into the network
- [ ] The input can be an image of any size as long as its width and height are equal
- [x] They follow the paradigm: Conv $\rightarrow$ Pool ... $\rightarrow$ Conv $\rightarrow$ Pool $\rightarrow$ FC ... $\rightarrow$ FC

  (Conv = Conv + activation)

Problem 2 Short Questions (18 credits)

0 1 2

2.1 In k -fold cross validation, choosing a larger value for k increases our confidence in the validation score. What could be a practical disadvantage in doing so? Explain how it arises.

(1p) Increases training time or more computations. (1p) Making use of more folds, will present the model with more data to train on, but will require way more time as it has to train and validate K separate times. (0p) Overfitting. (0p) High variance. (1p) Less data in validation set.

0 1 2 3 4

2.2 Consider the activation function $f : \mathbb{R} \to \mathbb{R}$ and $f(x) = \ln(1 + e^x)$.
Which one of the following activation functions is most closely approximated by f ? Briefly justify your answer (2 points). What is the benefit of f over the activation function it closely approximates (2 points)?

- Tanh
- ReLU
- Sigmoid

(1p) ReLU.(1p) ReLU is the only function which is unbounded up or any other valid explanation why ReLU or correct drawing of softplus showing similarity to ReLU. (0p) Positive output since it is also valid for Sigmoid.
One of both benefits is sufficient: (2p) Unlike ReLU, this function (softplus) is smooth everywhere in $\mathbb{R}$, so its differentiable everywhere in $\mathbb{R}$. ReLU is not differentiable at 0. (2p) Softplus does not have a dead area for negative inputs or any explanation related to dead ReLU.
(0p) If not ReLU.

0 1 2

2.3 Explain the difference between the validation set and the test set. In your answer, explain the role of each subset and how they are used differently.

Validation set is used for testing **generalization** (0.5p) with different **hyperparameters/ hyperparameter tuning** (0.5p). Test set is only used **at the end/ Not touched during training** (0.5p) to test generalization **on unseen data once** (0.5p). For each missing keyword -0.5p.

0 1 2

2.4 You notice vanishing/exploding gradients in a deep network using the tanh activation function. Suggest two possible changes you can make to the network in order to diminish this issue, without changing the number of trainable parameters. Explain how each of these changes helps.

(0.5p) For naming a correct change. (0.5p) For correct explanation. (1p) Use ReLU activation, does not saturate, large consistent gradients. (1p) Add residual connections, highway for gradient flow, can learn to skip layers. (1p) Xavier initialization, improved weight initialization targets active area of the activation function. (0p) Gradient clipping (does not resolve vanishing gradients). (0p) BatchNorm. (0p) Regularization.

dropping probabilities

2.5 Can two consecutive dropout layers with q and p be replaced with one dropout operation? Explain.

0 1 2

Yes and no. The second Dropout layer drops neurons independently of the first one. Therefore, the probability of a neuron on the input of thise sequence to be dropped is z = 1 - (1-p)(1-q) = p + q - pq. However, it is just the probability of a neuron to be dropped, and one cannot deduct how many neurons will be surely dropped. Thus, the new probability z cannot be used as the hyperparameter to a new Dropout(z) layer, that always drops z of the neurons.

So, actually - no, as two consecutive layers don't drop a fixed amount of neurons.

2.6 Can one encounter overfitting in an unsupervised learning setting? If your answer is *no*, provide a mathematical reasoning. If your answer is *yes*, provide an example.

0 1 2 3

(1p) Yes. (0p) No. (2p) Valid example: clustering N datapoints with N clusters (k-means with $k = N$), autoencoder with large bottleneck/ overfitted on one image, PCA with many components. (0.5p or 1p) For mentioning an unsupervised algorithm with vague explanation.

State the name of

2.7 For each of the following functions, one common problem when choosing them as the activation function for your deep neural network: (a) Sigmoid, (b) ReLU, (c) Identity

0 1 2 3

(1p) Sigmoid (not zero-centered or saturates), (1p) Identity (does not introduce non-linearity), (1p) ReLU (dead RelU or not zero-centered)

Problem 3 **Autoencoder (11 credits)**

Consider a given **unlabeled** image dataset consisting of 10 distinct classes of animals.

0 1 2

3.1 To train an Autoencoder on images, which type of losses you would use? Name two suitable losses.

(1p each) Image reconstruction losses: L1, L2, SSIM, PSNR, MSE .... (-1p) For L2 together with MSE. (0p) For CE, Hinge, BCE, KL-divergence First two named losses count.

0 1 2

3.2 Explain the effect of choosing a bottleneck dimension which is too small, and the effect of a too large bottleneck dimension in Autoencoders.

(1p) Bottleneck dimension too small leads to poor reconstruction/underfitting, or loss of important information/ too much compression. (1p) Bottleneck dimension too big leads to no compression/overfitting/learning identity.

0 1 2

3.3 Having trained an Autoencoder on this dataset, how would you use the trained Autoencoder (without further training/fine-tuning) to partition the dataset into 10 subsets, where each subset consists only of images of a distinct type of animal?

(1p) Use the trained encoder to get latent embedding for each unlabeled image. (1p) Do clustering (e.g k-means with $k = 10$). Assign each image to it's cluster centre. (0p) Adding FC layers. (0p) Using full autoencoder as feature extractor. (0p) Only mentioning clustering.

3.4 We want to use the same network architecture for de-noising and colorizing old, degraded, gray-scale images of animals. Given the dataset you already have, explain the steps you would take to train your model. In your answer, elaborate on your model's inputs, outputs, and losses.

0 1 2 3

(1p) Augment input image by adding noise. (1p) Transform input images by converting to grayscale. (1p) Use original image as target, use L1/L2 as loss. (0.5p) Name correct loss. (0.5p) Loss between original RGB images and output of the network. (0p) Only loss name. (0p) Proposing another architecture.

3.5 Explain the differences between Autoencoders and Variational Autoencoders. How do they differ during training?

0 1 2

(2p) Variational autoencoders will constrain the bottleneck distribution into a probability distribution but autoencoders don't constrain the latent space. (1p) Having a constraint in the latent space. (1p) Sampling from the latent space. (0.5p) KL - divergence. (-0.5p) Autoencoders generate images.

Problem 4 CNNs (10 credits)

You are given the following network that classifies RGB images into one of 4 classes.
All Conv2d layers use `kernel` = 3 , `padding` = 1 , `stride` = 1 , `bias` = `True` and are defined as $\texttt{Conv2d}(<\texttt{channels}_{\texttt{in}}>, <\texttt{channels}_{\texttt{out}}>)$.
All MaxPool2d layers use `stride` = 2 , `padding` = 0 , and are defined as $\texttt{MaxPool}(<\texttt{kernel}>)$.
The input dimension x of the Linear layer is unknown.
The network's architecture is as follows:

- Conv2d(3, 8) $\rightarrow$ MaxPool2d(2) $\rightarrow$ BatchNorm2d() $\rightarrow$ ReLU() $\rightarrow$
- Conv2d(8, 16) $\rightarrow$ MaxPool2d(2) $\rightarrow$ BatchNorm2d() $\rightarrow$ ReLU() $\rightarrow$
- Conv2d(16, 32) $\rightarrow$ MaxPool2d(2) $\rightarrow$ BatchNorm2d() $\rightarrow$ ReLU() $\rightarrow$
- Flatten() $\rightarrow$
- Linear(x , 4) $\rightarrow$ Softmax()

0 1 2

4.1 In terms of x , what is the total number of trainable parameters of the last linear layer? Include a bias term in your calculation.

$4x + 4 = 4(x + 1)$.
(1p) Matrix is shape $4 \times x$ (1p) plus 4 bias terms.
(-1p) Weight or bias wrong/missing.

0 1 2

4.2 Given RGB input images of size 80×80 pixels, what should the value of x in the Linear layer be? Explain your calculation.

(1p) Each conv2d preserves spatial dimensions. Each maxpool reduces spatial dimensions by 2. Height and width take shape 80 / 2 / 2 / 2 = 10 at linear layer. Depth is 32 as given by final conv2d.
(1p) $x = 10 \times 10 \times 32 = 3200$
(0.5p) Same convolution. (0.5p) Maxpool halves spatial dimension. (0.5p) Correct concept: channels x input x output.

4.3 Explain the main difference between the usage of a BatchNorm layer in a convolutional network in comparison to a fully connected network.

0 1 2

(2p) Normalization acts on channel dimension instead of per feature/different channels normalization/statistics. (1p) Only CNN or only FC. (0p) Over batch/all samples. (0p) Normalize weights. (0p) Normalize each pixel. (0p) Normalize input data.

4.4 Compute the total number of trainable parameters of the first convolutional layer, Conv2d(3,8).

0 1 2

(2p) $3 \times 3 \times 3 \times 8 + 8 = 216 + 8 = 224$
$k \times k \times channels_{in} \times N_{filters} + bias$
(0.5p) Weights wrong, bias correct. (-1p) Bias missing/wrong. (-0.5p) Correct answer, additionally specified batchnorm.

4.5 Compute the total number of trainable parameters in all of the BatchNorm layers.

0 1 2

(2p) Each BatchNorm2d layer has two weights per channel. The number of channels it has is given by the output of the preceding conv2d layer. Therefore # trainable BatchNorm weights = $2 * 8 + 2 * 16 + 2 * 32 = 2 * (56) = 112$ weights. (0.5p) Only $2 + 2 + 2$ without channels. (1.5p) Correct expression, final answer wrong.

Problem 5 Optimization and Gradients (16 credits)

You are training a large fully-connected neural network and select as an initial choice an SGD optimizer. In order to overcome the limitations of SGD, your colleague suggests adding momentum.

0 1 2 3

5.1 Name two limitations of SGD that momentum can potentially solve. Explain how momentum solves them.

1) limitations: slow learning / small steps, SGD is noisy, SGD only has one lr for all dimensions (1p each, 2p max)
2) explanation: speeds up learning if gradient keeps pointing in the **same direction**, adjusts lr down if oscillating over local minimum,

0 1 2

5.2 One can apply momentum, as shown in the formula:

$$\nu^{k+1} = \beta \cdot \nu^k - \alpha \cdot \nabla_\theta L(\theta^k)$$

What do the hyperparameters α and β represent?

1in alpha = learning rate (1pt) beta = accumulation rate of velocity/friction (1pt) momentum (0.5pt), only accumulation rate (0.5pt)

0 1 2

5.3 How does Nesterov Momentum differ from standard momentum? Explain.

Demonstrates understanding of Nestrov momentum but no further insights (only 0.5pt). A step in direction of previous momentum/accumulated gradient (only gradient is not enough) (1pt). Gradient term computed from position calculated with previous gradient i.e. look ahead step (1pt). Gradient corrects potential overshooting of momentum already in the same step (1pt).
Common mistakes: formulas without explaining them, not mentioning that the "jump" is calculated using accumulated gradients / previous momentum

0 1 2 3

5.4 Is RMSProp considered a first or second order method (1p)? What is the main difference between RMSProp and SGD+Momentum?

Second order (1pt) **Explanation:** RMSProp dampens oscillation /exponentially decaying average of variance/ uses second moment (1pt) SGD + Momentum accumulates gradient / uses first moment (1pt)

For the following questions, consider the convex optimization objective:

$$\min_{x \in \mathbb{R}} x^2$$

5.5 What is the optimal solution of this optimization problem?

0
1

$x^* = 0$ (1p)

5.6 You are working with an initialization of $x_0 = 5$ and a learning rate of $\texttt{lr} = 1$. How many iterations would gradient descent (without momentum) need in order to converge to the optimal solution? Explain.

0
1

Won't convergence/ infinite iterations (0.5pt). Explanation as overshoot/ oscillate (0.5pt)

5.7 deleted

5.8 What is the main advantage of using a second order method such as Newton's Method? Why are second order methods not used often in practice for training deep neural networks?

0
1
2

Advantages: less iterations (1pt) if only mentioned "converge faster" without specifying in terms of iterations (0.5pt), only 1 step (0.5p), no need to choose learning rate.
Drawbacks: Hessian costly to compute, Second order methods don't work well with mini-batches (1pt each; 1pt max)

5.9 How many iterations would Newton's method need to converge (using the same initialization $x_0 = 5, \texttt{lr} = 1$)? Explain.

0
1

Only takes 1 iteration(0.5pt). Jumps to minimum right away / convex problem / 2nd order taylor approximation exactly approximate quadratic problem / calculation that it converges after one step (0.5pt).
Common mistakes: uses second derivative, uses hessian instead of lr

Problem 6 **Derivatives (9 credits)**

Consider the formula of the Sigmoid function $\sigma(x) : \mathbb{R} \to \mathbb{R}$:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

0 1 6.1 Compute the derivative $\frac{d\sigma(x)}{dx}$ in terms of x .

$$\frac{d\sigma}{dx} = \frac{0 \cdot (1 + e^{-x}) - 1 \cdot (-e^{-x})}{(1 + e^{-x})^2} = \frac{e^{-x}}{(1 + e^{-x})^2}$$

correct intermediate step (0.5p) & correct final answer (1p)

0 1 6.2 A special property of this function is that its derivative can be expressed in terms of the Sigmoid function itself. Denote $y = \sigma(x)$, and show how the derivative you computed can be re-written in terms of y , the output of the Sigmoid function. *Hint:* Your answer should only depend on y .

$$\frac{dy}{dx} = y(1 - y)$$

$$y = \frac{1}{(1 + e^{-x})}$$

$$1 - y = \frac{1 + e^{-x}}{(1 + e^{-x})} - \frac{1}{(1 + e^{-x})} = \frac{e^{-x}}{(1 + e^{-x})}$$

Final correct answer (1p). Wrong answer with some correct derivation (0.5p).

An affine Layer is described by $\mathbf{z} = XW + b$.
Consider the following affine layer, which has 2 input neurons and 1 output neuron:

$$W = \begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times 1}$$

$$b = 2 \in R^1$$

and input:

$$X = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}_{2\times 2}$$

The forward pass of the network would be:

$$\sigma(\mathbf{z}) = \sigma(XW + b) = \sigma(\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 2) = \sigma(\begin{bmatrix} 3 \\ -2 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \end{bmatrix}) = \sigma(\begin{bmatrix} 5 \\ 0 \end{bmatrix}) = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \text{ (rounded up).}$$

Let's compute the backward pass of the network. Assume L(z) = sum(z)

6.3 If $\mathbf{y} = \sigma(\mathbf{z}) = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$, calculate the gradient of the output after the Sigmoid activation function w.r.t $\mathbf{z}$, $\frac{d\mathbf{y}}{d\mathbf{z}}$:

0 1 2 3

dL/dz = dL/dy * dy/dz = $\frac{\partial L}{\partial z} = \mathbf{y} \circ (1 - \mathbf{y})$ x ^{1}

^{1} x ($\begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \circ \left(1 - \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 0 \\ 0.25 \end{bmatrix}$

^{1}

writing the derivative correctly (element-wise multiplication and NOT matrix multiplication) (1p), correct intermediate calculation (dimensions are correct) (1p), correct final answer (1p)

6.4 We will use the computed gradient to perform back-propagation through the affine layer to the network's parameters.
Let *dout* be the upstream derivative of the Sigmoid that you have calculated in question 3. Calculate the derivatives $\frac{d\mathbf{y}}{dW}$ and $\frac{d\mathbf{y}}{db}$.
Hint: Pay attention to the shapes of the results; they should be compatible for a gradient update.
Note: In case you skipped the previous question, you can get partial points by writing the correct formulas using *dout* symbolically.

0 1 2 3 4

$$dW = X^T \cdot dout = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.25 \end{bmatrix}$$

$$db \rightarrow sum(dout, axis = 0) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0.25 \end{bmatrix} = 0.25$$

dW:(2p) *db*:(2p). For case, chain rule (0.5p) writing the matrices correctly (e.g $X_T * dout$) (1p), correct answer (0.5p). if missed the correct answer by 1/n (-0.5p)

Problem 7 Model Evaluation (8 credits)

Two students, *Erika* and *Max* train a neural network for the task of image classification. They use a dataset which is divided into train and validation sets. They each train their own network for 25 epochs.

0 1 7.1 Erika selects a model and obtains the following curves. Interpret the model's behaviour from the curves. Then, suggest what could Erika do in order to improve its performance?

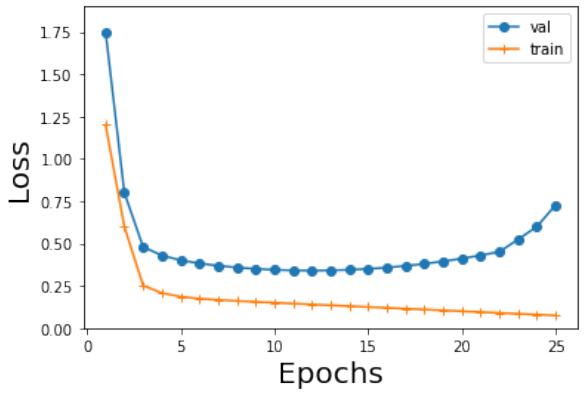

Figure 7.1: Training curves for Erika's model.

(0.5p) Overfitting , (0.5p) regularization (Dropout, l1, l2 weight decay, data augmentation), early stopping and reducing capacity.
Common mistakes: Stop training early. No mention of stopping training early based on validation error.

0 1 2 7.2 *Max* selects a different model and obtains the following curves. Interpret the model's behaviour from the curves. Then, suggest what change could Max make to his model in order to improve its performance?

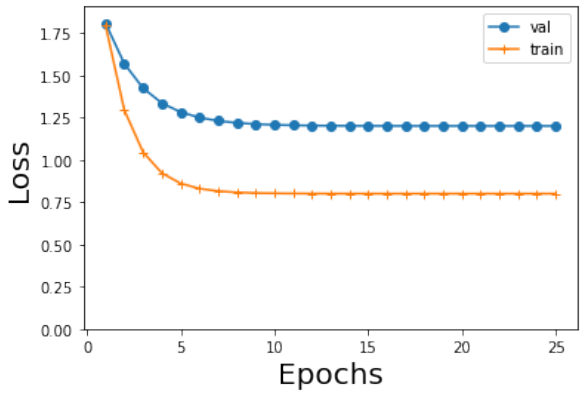

Figure 7.2: Training curves for Max's model.

Underfitting (0.5p), Increase model capacity (1.5p). **OR** Optimization is not optimal (0.5p), Decrease learning rate / learning rate decay / use optimizer that corrects a bad learning rate choice (e.g Adam) / BN (1.5p)
Common mistakes: Just describing what the graphs do. Generalization gap, add regularization

7.3 Both Max and Erika are able to agree on a model architecture and obtain the following curves. However, when deployed in real world, their model seems to perform poorly. What is a possible reason for such an observation and what should they do?

0
1
2

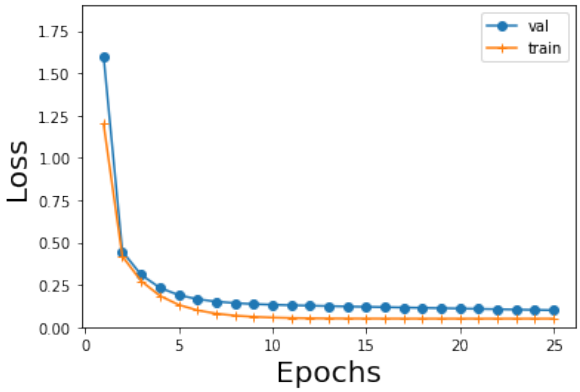

Figure 7.3: Training curves for the new model.

Another option- due to data leakage (or some bug) from the val to the train, we couldn't catch overfitting.
Possible reasoning: test and train/val data is sampled from different distributions / domain gap.(1p) Fix by trying to make test and train data more similar or from same distribution / augmentation / add another dataset to train / (1p).
Common mistakes: change the test set. Shuffle train and val dataset to train again. If train and val come from the same distribution (given in the question) shuffling will not help. Overfitting to "val" data.

After adapting the new network architecture, Max and Erika are training their own model, using the same architecture, with identical initial weights, using exactly the same hyperparameters. They also use the same SGD optimizer (no momentum), batch size, and learning rates. The only difference is that Max normalizes the loss by $1/N$ (where N is the number of training samples in the dataset) while Erika does not.

7.4 How does this affect the optimal model weights that minimize this optimization objective? (1p) After 10 optimizer steps, will they arrive at the same model parameters? Explain.(2p)

0
1
2
3

It doesn't affect the optimal model's optima (1p) The weights will be different (0.5p) after 10 steps (0.5p). A good explanation why (lr is scaled) (1p) Contradictory / Unclear explanation (-0.5p)
Common mistakes: 1/ N would make it independent of the size of the dataset. Irrelevant.

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.