Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNTI Winter Term 2024/2025

Exercise Sheet 8

Theory of Neural Networks

Deadline: 18.12.2024 23:59

Guidelines: You are expected to work in a group of 2-3 students. While submitting the assignments, please make sure to include the following information for all our teammates in your PDF/python script:

Name:

Student ID (matriculation number):

Email:

Your submissions should be zipped as **Name1_id1_Name2_id2_Name3_id3.zip** when you have multiple files. For assignments where you are submitting a single file, use the **same naming convention** without creating a zip. For any clarification, please reach out to us on the **CMS Forum**. These instructions are mandatory. If you are not following them, tutors can decide not to correct your exercise.

Exercise 8.1 - Universal Approximation Theorem (0.5 + 0.5 + [0.25 + 0.25]=1.5 points)

Read section 6.4.1 in the Deep Learning book, and answer the following questions::

a) What does the Universal Approximation Theorem state? (2 sentences)

b) Suppose we consider only linear neurons, i.e., neurons with the activation function $s(z) = z$. Explain why these networks built on these neurons don't satisfy the Universal Approximation Theorem. (2 sentences)

c) Given the universal approximation properties of feed-forward neural networks, why do we still: (2 reasons suffice for both)

   (i) care about designing neural network architectures different from fully connected feed-forward networks?

   (ii) prefer to use (often many) more than two layers in practice?

Exercise 8.2 - Regularization (0.5+1+0.5+0.5=2.5 points)

a) Why do we penalize only weights but not biases?

b) For a linear model with L_2 regularization (ridge regression), the optimal weights are given by

$$\boldsymbol{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{X}$ is the data matrix and $\boldsymbol{y}$ is the vector of targets. Show that for predictions made using this model, the following holds:

$$\mathbf{X}\boldsymbol{w}^* = \sum_{i=1}^{r} \boldsymbol{u}_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \boldsymbol{u}_i^T \boldsymbol{y},$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the data matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of targets, $\boldsymbol{u}_i \in \mathbb{R}^n$ and $r = rank(\mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the diagonal matrix from the SVD decomposition of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Here $\boldsymbol{u}_i$ are the left singular vectors of X and σ_i^2 are the eigenvalues of $\mathbf{X}^T\mathbf{X}$. This shows that for a linear model, L_2-regularization regularization shrinks contributions of directions corresponding to smaller singular values, reducing the influence of low-variance components in X.

c) Consider the regularized objective $\tilde{J}(w) = J(w) + \frac{\lambda}{2}w^t w$. For SGD, the weight update for $J(w)$ is

$$w_{i+1} = w_i - \eta \nabla_w J(w)$$

where η is the learning rate. Derive the weight update rule for the regularized loss. How is it related to weight decay?

d) Early stopping limits the number of updates that can be made to the parameters. Prove that it is equivalent to using L_2 regularization. Mention explicitly any assumptions you made for this proof.

Hint: Look up Section 7.8, page number 246 in the Deep Learning book.

Exercise 8.3 - Parameter Norm Penalties (3 points)

See attached notebook

Exercise 8.4 - Regularization in Neural Networks (3 points)

See attached notebook