UNIVERSITÄT DES SAARLANDES
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNTI Winter Term 2020/2021

Exercise Sheet 5

Deep Feedforward Networks and Gradient Descent

Deadline: 15.12.2020, 23:59

Exercise 5.1 - Gradient Descent (1 + 0.5 + 1 points)

The loss L of the unsupervised *k-means* clustering algorithm for k clusters, sample points $x_1, x_2, ..., x_n$, and centers $\mu_1, \mu_2, ..., \mu_n$ is denoted by:

$$L = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$$

where S_j refers to the set of data points which are closer to μ_j than to any other cluster mean.

a) μ_j is updated by computing the mean. Instead, let us try to minimize L by **batch** gradient descent while keeping S_j fixed. Derive the update formula for μ_1 with the learning rate ϵ.

b) Derive the update formula for μ_1 with **stochastic** gradient descent on a single sample point x_i. Use learning rate ϵ.

c) Let us now try to connect the batch gradient descent update equation from part (a) with the standard *k-means* algorithm. In the update step of *k-means*, we assign each cluster center to be the mean of the data points closest to that center. Calculate the value of ϵ such that both the equations perform the same update for the first cluster whose center is μ_1.

Exercise 5.2 - Weight Space Symmetry (1 + 0.5 points)

Consider a neural network with a single hidden layer consisting of M neurons and *tanh* activation function. For any neuron in the hidden layer, simultaneous change of sign of input and output weights from the neuron leads to no change in the output layer therefore producing an equivalent transformation. Similarly, for any pair of neurons interchange of input weights between the neurons and simultaneous interchange of output weights produces an equivalent transformation.

a) Find the total number of equivalent transformation for the hidden layer.

b) Consider a deep neural network with N hidden layers. Each hidden layer consists of M_i neurons where $i \in 1, 2, ..., N$ and *tanh* activation function. Find the total number of equivalent transformations for the network.

Exercise 5.3 - Hessian and Optimization (1 + 1 points)

Consider a quadratic form function given by,

$$f(X) = \frac{1}{2}X^T AX + b^T X + c$$

where, $X = [x_1, x_2]$ and A is symmetric. Hessian for the function is given by A. Each eigenvector of the hessian represents a direction where the curvature of f is independent of the other directions. The curvature in the direction of the eigenvector is determined by the eigenvalue.

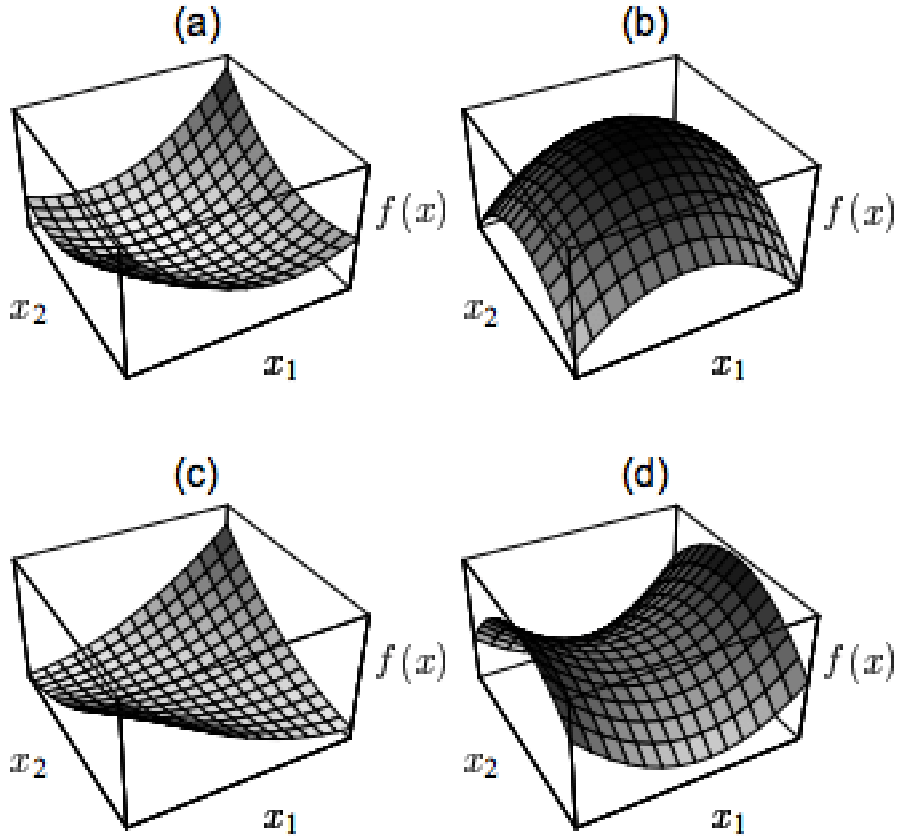

Figure 1

a) For each curve in figure 1, what can you say about the sign of the eigenvalues of the Hessian matrix.

b) For each curve comment whether the Hessian is positive definite, negative definite or neither. Explain your answer.

Exercise 5.4 - Gradient Descent and Newton's Method (1.0 + 1.0 + 0.5 = 2.5 points)

In the optimization setting, Gradient Descent and Newton's Method are two commonly used iterative methods for finding a solution. Let's say we want to minimize a function $f : \mathbb{R}^2 \to \mathbb{R}$ which is defined as $f(\mathbf{x}) = x_1^2 - 3x_1 + x_2^2 - x_1x_2$. The starting point is $\mathbf{x}^{(0)} = [1, 1]^T$.

a) Use Gradient Descent with step size (also known as learning rate) $\epsilon = 0.5$ to minimize the function f. The iteration should stop if the L2-norm of the gradient at the current point is less than 0.2. Show your intermediate steps. [Hint: The iteration should finish within 4 steps].

b) Starting from the same point $\mathbf{x}$, use Newton's Method to find a solution. Is the solution you get a global minimum? Argue why or why not.

c) Is Newton's Method always applicable if the function f is twice continuously differentiable? Argue with the help of the function $f(x) = 2x^3 - 5x$ at $x = 0$.

Exercise 5.5 - Convex Optimization (0.5 + 0.5 + 0.5 = 1.5 points)

a) Show that convexity is invariant under affine maps. That is if f is convex then so is $g(x) = f(Ax + b)$, where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

b) Show that every local minimum of a convex function is a global minimum.

c) Is the cross entropy loss function convex with respect to the weights of a multi layer perceptron model?

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You can and are encouraged to submit the assignment as a team of two students. Submitting as a team will be mandatory for the next assignment.
- Hand in zip file containing the PDF with your solutions and the completed iPython notebook, or you could also hand in just the iPython notebook containing everything.
- Therefore make sure to write the Microsoft Teams user name, student id and the name of each member of your team on your submission.
- Your assignment solution must be uploaded by only **one** of your team members to the 'Assignments' tab of the tutorial team (in **Microsoft Teams**).
- If you have any trouble with the submission, contact your tutor **before** the deadline.