Deep Learning

Theoretical Exercises – Week 11 – Chapter 8

Exercises on the book ”Deep Learning” written by Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Exercises and solutions by T. Méndez and G. Schuster

FS 2024

1 Exercises on Optimization for Training Deep Models

1. Given is the following loss function

$$f(\boldsymbol{w}) = \frac{1}{4}\,w_1^4 + w_1^3 - \frac{17}{4}\,w_1^2 - 6\,w_1 + \frac{1}{5}\,w_2^4 + \frac{6}{5}\,w_2^3 + 89,$$

which has a global minimum at the point $\boldsymbol{c}_0 = [-4.572 \quad -4.5]^T$ and a local minimum at the point $\boldsymbol{c}_1 = [2.175 \quad -4.5]^T$.

(a) Determine the gradient of the loss function.

(b) Search for the global minimum by using gradient descent. In doing so, start at the point $\boldsymbol{w}^{(0)} = [4 \;\; 4]^T$ and use the learning rate $\epsilon = 0.05$. Finish the learning algorithm after 10 iterations and check whether you have found the global minimum or not.

(c) Repeat Exercise (b) with the method of momentum and use $\alpha = 0.5$.

Solution:

(a) The gradient of the loss function is

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = \begin{bmatrix} \dfrac{\partial f(\boldsymbol{w})}{\partial w_1} \\ \dfrac{\partial f(\boldsymbol{w})}{\partial w_2} \end{bmatrix} = \begin{bmatrix} w_1^3 + 3\,w_1^2 - \dfrac{17}{2}\,w_1 - 6 \\ \dfrac{4}{5}\,w_2^3 + \dfrac{18}{5}\,w_2^2 \end{bmatrix}.$$

(b) The update rule for gradient descent is

$$\begin{aligned} \boldsymbol{w}^{(i)} &= \boldsymbol{w}^{(i-1)} - \epsilon \cdot \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \\ &= \begin{bmatrix} w_1^{(i-1)} \\ w_2^{(i-1)} \end{bmatrix} - \epsilon \cdot \begin{bmatrix} \left(w_1^{(i-1)}\right)^3 + 3\left(w_1^{(i-1)}\right)^2 - \dfrac{17}{2}\,w_1^{(i-1)} - 6 \\ \dfrac{4}{5}\left(w_2^{(i-1)}\right)^3 + \dfrac{18}{5}\left(w_2^{(i-1)}\right)^2 \end{bmatrix}. \end{aligned}$$

Starting from point $\boldsymbol{w}^{(0)} = [4 \ \ 4]^T$, using a learning rate of $\epsilon = 0.05$ and performing 10 steps, results in the point $\boldsymbol{w}^{(10)} = [2.1753 \ \ \text{-}4.4885]^T$, which corresponds to the local minimum $\boldsymbol{c}_1$. The steps of this calculation are shown in the following table:

$\boldsymbol{w}^{(0)}$	$\boldsymbol{w}^{(1)}$	$\boldsymbol{w}^{(2)}$	$\boldsymbol{w}^{(3)}$	$\boldsymbol{w}^{(4)}$	$\boldsymbol{w}^{(5)}$
$\begin{bmatrix} 4.0000 \\ 4.0000 \end{bmatrix}$	$\begin{bmatrix} 0.4000 \\ -1.4400 \end{bmatrix}$	$\begin{bmatrix} 0.8428 \\ -1.6938 \end{bmatrix}$	$\begin{bmatrix} 1.3645 \\ -2.0158 \end{bmatrix}$	$\begin{bmatrix} 1.8381 \\ -2.4196 \end{bmatrix}$	$\begin{bmatrix} 2.1020 \\ -2.9068 \end{bmatrix}$
$\boldsymbol{w}^{(6)}$	$\boldsymbol{w}^{(7)}$	$\boldsymbol{w}^{(8)}$	$\boldsymbol{w}^{(9)}$	$\boldsymbol{w}^{(10)}$	
$\begin{bmatrix} 2.1682 \\ -3.4453 \end{bmatrix}$	$\begin{bmatrix} 2.1749 \\ -3.9461 \end{bmatrix}$	$\begin{bmatrix} 2.1753 \\ -4.2911 \end{bmatrix}$	$\begin{bmatrix} 2.1753 \\ -4.4450 \end{bmatrix}$	$\begin{bmatrix} 2.1753 \\ -4.4885 \end{bmatrix}$	

(c) The update rule for the method of momentum is

$$\begin{aligned} \boldsymbol{w}^{(i)} &= \boldsymbol{w}^{(i-1)} + \boldsymbol{v}^{(i)} \\ &= \begin{bmatrix} w_1^{(i-1)} \\ w_2^{(i-1)} \end{bmatrix} + \begin{bmatrix} v_1^{(i)} \\ v_2^{(i)} \end{bmatrix} \end{aligned} \tag{1.1}$$

with

$$\begin{aligned} \boldsymbol{v}^{(i)} &= \alpha \cdot \boldsymbol{v}^{(i-1)} - \epsilon \cdot \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \\ &= \alpha \cdot \begin{bmatrix} v_1^{(i-1)} \\ v_2^{(i-1)} \end{bmatrix} - \epsilon \cdot \begin{bmatrix} \left(w_1^{(i-1)}\right)^3 + 3\left(w_1^{(i-1)}\right)^2 - \frac{17}{2} w_1^{(i-1)} - 6 \\ \frac{4}{5}\left(w_2^{(i-1)}\right)^3 + \frac{18}{5}\left(w_2^{(i-1)}\right)^2 \end{bmatrix} \end{aligned}$$

and $\boldsymbol{v}^{(0)} = [0 \ \ 0]^T$. Starting from point $\boldsymbol{w}^{(0)} = [4 \ \ 4]^T$, using a learning rate of $\epsilon = 0.05$ and $\alpha = 0.5$ and performing 10 steps, results in the point $\boldsymbol{w}^{(10)} = [\text{-}4.7790 \ \ \text{-}4.5724]^T$, which corresponds to the global minimum $\boldsymbol{c}_0$. The steps of this calculation are shown in the following table:

$\boldsymbol{w}^{(0)}$	$\boldsymbol{w}^{(1)}$	$\boldsymbol{w}^{(2)}$	$\boldsymbol{w}^{(3)}$	$\boldsymbol{w}^{(4)}$	$\boldsymbol{w}^{(5)}$
$\begin{bmatrix} 4.0000 \\ 4.0000 \end{bmatrix}$	$\begin{bmatrix} 0.4000 \\ -1.4400 \end{bmatrix}$	$\begin{bmatrix} -0.9572 \\ -4.4138 \end{bmatrix}$	$\begin{bmatrix} -1.8362 \\ -5.9679 \end{bmatrix}$	$\begin{bmatrix} -2.9523 \\ -4.6537 \end{bmatrix}$	$\begin{bmatrix} -4.4858 \\ -3.8635 \end{bmatrix}$
$\boldsymbol{w}^{(6)}$	$\boldsymbol{w}^{(7)}$	$\boldsymbol{w}^{(8)}$	$\boldsymbol{w}^{(9)}$	$\boldsymbol{w}^{(10)}$	
$\begin{bmatrix} -5.3641 \\ -3.8484 \end{bmatrix}$	$\begin{bmatrix} -4.3818 \\ -4.2269 \end{bmatrix}$	$\begin{bmatrix} -4.1263 \\ -4.6113 \end{bmatrix}$	$\begin{bmatrix} -4.4934 \\ -4.7088 \end{bmatrix}$	$\begin{bmatrix} -4.7790 \\ -4.5724 \end{bmatrix}$	

Thus, in this case it was possible to overcome the local minimum through the use of the method of momentum.