Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Oct 17, 2016
(Deep Learning, Exercise series 6 - solutions)

Solution 1 (Fundamentals of Unconstrained Optimization):

(1) The gradient of the Rosenbrock function is

$$\nabla f(x) = \begin{pmatrix} 2(200w_1^3 - 200w_1w_2 + w_1 - 1) \\ 200(w_2 - w_1^2) \end{pmatrix}. \tag{1}$$

The Hessian matrix is

$$H(x) = \begin{pmatrix} -400(w_2 - 3w_1^2) + 2 & -400w_1 \\ -400w_1 & 200 \end{pmatrix}. \tag{2}$$

The gradient at $w^* = (1,1)^\top$ is $\nabla f(w^*) = (0,0)^\top$ while the Hessian is

$$H(w^*) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \tag{3}$$

which is a positive definite matrix. Since H is a $d \times d$ matrix this can be seen, for example by Sylvester's criterion since $802 > 0$ and $det(H(w^*)) = 400 > 0$[1].

(2) The gradient of $f(w)$ is

$$\nabla f(w) = \begin{pmatrix} 2(w_1 + 4) \\ -4(w_2 - 3) \end{pmatrix}. \tag{4}$$

Therefore the gradient vanishes at $w^* = (-4,3)^\top$.

The Hessian matrix is

$$H = \begin{pmatrix} 2 & 0 \\ 0 & -4 \end{pmatrix}$$

which means there is a direction of negative curvature at w^*. The contour plot of the function is shown in figure 1. Minimizing this function is actually *infeasible* since for example for the sequence $(w^\nu)_{\nu\in\mathbb{N}} = (0,\nu)$ we have $\lim_{\nu\to\infty} f(w^\nu) = -\infty$.

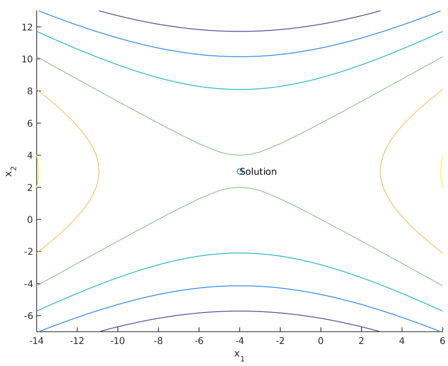

Figure 1: Contour plot

[1] Note that strict second order stationarity is a local criterion and does (by itself) not imply that w^* is a *global* minimizer. You could show global optimality e.g. via the coercivity of f or by showing that $f(w^*) = \inf_x f(w)$, but this is not subject of this lecture.

Solution 2 (Approximate Hessian for feed-forward networks):

First note that $\|\boldsymbol{s}\|_2^2 = \sum_i^d s_i^2 = \boldsymbol{s}^\intercal \boldsymbol{s}$. Then it is easy to see that

$$\nabla m_t(\boldsymbol{s}) = \nabla f(\boldsymbol{w}_t) + L\boldsymbol{s} \tag{5}$$

and

$$\nabla^2 m_t(\boldsymbol{s}) = L\mathbf{I}. \tag{6}$$

Since $L > 0$ we can conclude that the objective $m_t(\boldsymbol{s})$ is strongly convex. Hence any point that satisfies $\nabla m_t(\boldsymbol{s}) = 0$ is a global minimizer (as a matter of fact this point is unique due to the strong convexity of $m_t(\boldsymbol{s})$). Thus, by setting Eq.(5) to zero and solving for $\boldsymbol{s}$ we have

$$\boldsymbol{s}_t^* = -\frac{1}{L}\nabla f(\boldsymbol{w}_t). \tag{7}$$

Solution 3 (Programming exercise: Optimization methods in tensorflow):

The solution to the programming exercise is provided in the jupyter notebook *Optimization-solution.ipynb*.