Modern Optimization Techniques – Group 01

Exercise Sheet 04

Submitted by: Muhammad Inaam Ashraf (Matrikel-Nr: 307524)

Semester 2 MSc. Data Analytics

Question 1: Newton Method

1. In your own words, describe what is the intuition of using the Newton step for function minimization. You can use a sketch.

Newton step works by finding the minimum of a quadratic approximation to find the starting value of x and then moves iteratively towards the minimum value of x by repeating the process. In other word, we fit a parabola over our function at each step instead of a line in gradient descent. The intuition behind this is to achieve faster convergence since the minima of a convex function is similar to a parabola and fitting a quadratic approximation helps in much quicker convergence.

2. For the following equations, compute their derivatives and second derivatives, write down the Newton Update Formula and execute 10 iterations of the Newton Method. Discuss what is happening.

(a) $\boldsymbol{f_1: R \to R,\ f_1(x) = x^3 - 2x - 5\ for\ an\ initial\ x = 8\ and\ x = -10}$

Newton Update step is given by:

$$\boldsymbol{x^{(t+1)} = x^{(t)} + \mu^{(t)} \Delta x^{(t)}, where\ \Delta x^{(t)} = -\nabla^2 f(x)^{-1} \nabla f(x)}$$

We have

$$\nabla f_1(x) = 3x^2 - 2 \qquad and \qquad \nabla^2 f_1(x) = 6x$$

And

$$x^{(t+1)} = x^{(t)} - \frac{\mu^{(t)}\left(3x^{(t)^2} - 2\right)}{6x^{(t)}}$$

For $\boldsymbol{x^{(0)} = 8}$ and $\boldsymbol{\mu^{(t)} = 1}$:

$$x^{(1)} = 8 - \frac{1(3(8)^2 - 2)}{6(8)} = 4.04, \qquad x^{(2)} = 4.04 - \frac{1(3(4.04)^2 - 2)}{6(4.04)} = 2.1$$

$$x^{(3)} = 2.1 - \frac{1(3(2.1)^2 - 2)}{6(2.1)} = 1.21, \qquad x^{(4)} = 1.21 - \frac{1(3(1.21)^2 - 2)}{6(1.21)} = 0.88$$

$$x^{(5)} = 0.88 - \frac{1(3(0.88)^2 - 2)}{6(0.88)} = 0.8188, \qquad x^{(6)} = 0.8188 - \frac{1(3(0.8188)^2 - 2)}{6(0.8188)} = 0.8165$$

$$x^{(7)} = 0.8165 - \frac{1(3(0.8165)^2 - 2)}{6(0.8165)} = 0.8165, \qquad x^{(8)} = 0.8165 - \frac{1(3(0.8165)^2 - 2)}{6(0.8165)} = 0.8165$$

$$x^{(9)} = 0.8165 - \frac{1(3(0.8165)^2 - 2)}{6(0.8165)} = 0.8165, \qquad x^{(10)} = 0.8165 - \frac{1(3(0.8165)^2 - 2)}{6(0.8165)} = 0.8165$$

For $x^{(0)} = -10$ and $\mu^{(t)} = 1$:

$$x^{(1)} = -10 - \frac{1(3(-10)^2 - 2)}{6(-10)} = -5.03, \qquad x^{(2)} = -5.03 - \frac{1(3(-5.03)^2 - 2)}{6(-5.03)} = -2.583$$

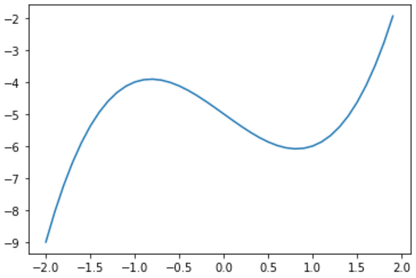
$$x^{(3)} = -2.583 - \frac{1(3(-2.583)^2 - 2)}{6(-2.583)} = -1.42, \qquad x^{(4)} = -1.42 - \frac{1(3(-1.42)^2 - 2)}{6(-1.42)} = -0.945$$

$$x^{(5)} = -0.945 - \frac{1(3(-0.945)^2 - 2)}{6(-0.945)} = -0.825, \qquad x^{(6)} = -0.825 - \frac{1(3(-0.825)^2 - 2)}{6(-0.825)} = -0.8165$$

$$x^{(7)} = -0.8165 - \frac{1(3(-0.8165)^2 - 2)}{6(-0.8165)} = -0.8165, \qquad x^{(8)} = -0.8165 - \frac{1(3(-0.8165)^2 - 2)}{6(-0.8165)} = -0.8165$$

$$x^{(9)} = -0.8165 - \frac{1(3(-0.8165)^2 - 2)}{6(-0.8165)} = -0.8165, \qquad x^{(10)} = -0.8165 - \frac{1(3(-0.8165)^2 - 2)}{6(-0.8165)} = -0.8165$$

What is happening here is that we have converged to the local minima and local maxima for x = 8 and x = -10 respectively. This is because the second derivative of the function $\nabla^2 f_1(x) = 6x$ can take any sign as x can have any sign.



(b) $\boldsymbol{f_2 : R \rightarrow\ R,\ f_2(x) = 3x^{\frac{1}{3}}\ \mathit{for\ an\ initial}\ x = -0.5\ \mathit{and}\ x = 1}$

Newton Update step is given by:

$$\boldsymbol{x^{(t+1)} = x^{(t)} + \mu^{(t)}\Delta x^{(t)}, \mathit{where}\ \Delta x^{(t)} = -\nabla^2 f(x)^{-1}\nabla f(x)}$$

We have

$$\nabla f_2(x) = \frac{1}{x^{\frac{2}{3}}} \qquad \mathit{and} \qquad \nabla^2 f_1(x) = -\frac{2}{3x^{\frac{5}{3}}}$$

And

$$x^{(t+1)} = x^{(t)} - \mu^{(t)}\left(\frac{1}{x^{(t)\frac{2}{3}}} \times -\frac{3x^{(t)\frac{5}{3}}}{2}\right) = x^{(t)} + \frac{3}{2}\mu^{(t)}x^{(t)} = x^{(t)}(1 + \frac{3}{2}\mu^{(t)})$$

For $\boldsymbol{x^{(0)} = -0.5}$ and $\boldsymbol{\mu^{(t)} = 1}$:

$$x^{(1)} = -0.5\left(1+\frac{3}{2}(1)\right) = -1.25, \qquad x^{(2)} = -1.25\left(1+\frac{3}{2}(1)\right) = -3.125$$

$$x^{(3)} = -3.125\left(1+\frac{3}{2}(1)\right) = -7.8125, \qquad x^{(4)} = -7.8125\left(1+\frac{3}{2}(1)\right) = -19.53$$

$$x^{(5)} = -19.53\left(1+\frac{3}{2}(1)\right) = -48.828, \qquad x^{(6)} = -48.828\left(1+\frac{3}{2}(1)\right) = -122.07$$

$$x^{(7)} = -122.07\left(1+\frac{3}{2}(1)\right) = -305.8, \qquad x^{(8)} = -305.8\left(1+\frac{3}{2}(1)\right) = -762.94$$

$$x^{(9)} = -762.94\left(1+\frac{3}{2}(1)\right) = -1907.35, \qquad x^{(10)} = -1907.35\left(1+\frac{3}{2}(1)\right) = -4768.37$$

For $\boldsymbol{x^{(0)} = 1}$ and $\boldsymbol{\mu^{(t)} = 1}$:

$$x^{(1)} = 1\left(1+\frac{3}{2}(1)\right) = 2.5, \qquad x^{(2)} = 2.5\left(1+\frac{3}{2}(1)\right) = 6.25$$

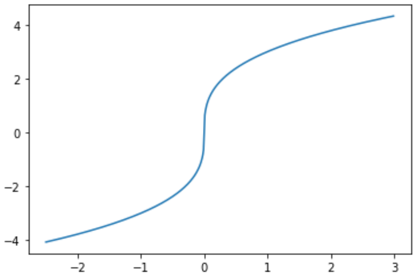
$$x^{(3)} = 6.25\left(1+\frac{3}{2}(1)\right) = 15.625, \qquad x^{(4)} = 15.625\left(1+\frac{3}{2}(1)\right) = 39.0625$$

$$x^{(5)} = 39.0625\left(1+\frac{3}{2}(1)\right) = 97.66, \qquad x^{(6)} = 97.66\left(1+\frac{3}{2}(1)\right) = 244.14$$

$$x^{(7)} = 244.14\left(1+\frac{3}{2}(1)\right) = 610.35, \qquad x^{(8)} = 610.35\left(1+\frac{3}{2}(1)\right) = 1525.88$$

$$x^{(9)} = 1525.88\left(1+\frac{3}{2}(1)\right) = 3814.7, \qquad x^{(10)} = 3814.7\left(1+\frac{3}{2}(1)\right) = 9536.74$$

This function doesn't have a minima or a maxima that can be seen from the derivatives and the plot.



$$\nabla f_2(x) = \frac{1}{x^{\frac{2}{3}}} \neq 0 \qquad and \qquad \nabla^2 f_1(x) = -\frac{2}{3x^{\frac{5}{3}}} \neq 0$$

3. In which cases the Newton step can overshoot?.

Newton method can overshoot in cases where

- There is no minima of the function as in case of Q 1.2b

- There is a minima, a maxima and an inflection point as in case of Q1.2a. The reason the Newton Method didn’t overshoot in Q1.2a is because we started away from the inflection point in both cases ($x^{(0)} = 8\ and - 10$). Had we started between the minima and maxima, the algorithm would have overshot.

Question 2: Newton Method for Machine Learning

1. The loss function of the linear Regression has the following form:

$$\boldsymbol{\mathcal{L}(X, \beta, Y) = \sum_{i=1}^{m} (x_i\beta - y_i)^2}$$

(a). Compute the Hessian $\boldsymbol{\nabla_\beta}^{\mathbf{2}}\boldsymbol{\mathcal{L}(X, \beta, Y)}$

Computing first derivative

$$\nabla_\beta \mathcal{L}(X, \beta, Y) = \nabla_\beta\left(\sum_{i=1}^{m}\left(x_i^{\,2}\beta^2 - 2\beta x_i y_i + y_i^{\,2}\right)\right)$$

$$= 2\beta\sum_{i=1}^{m} x_i^{\,2} - 2\sum_{i=1}^{m} x_i y_i + 0$$

$$\boldsymbol{\nabla_\beta \mathcal{L}(X, \beta, Y) = 2(X^T X\beta - X^T Y)}$$

Computing Hessian

$$\boldsymbol{\nabla_\beta}^{\mathbf{2}}\boldsymbol{\mathcal{L}(X, \beta, Y)} = \nabla_\beta(2(X^T X\beta - X^T Y) = 2\boldsymbol{X^T X}$$

(b). Does it make sense to use the Newton minimization algorithm here? Discuss your answer.

Newton step for linear regression is given by:

$$\beta^{(t+1)} = \beta^{(t)} - \mu^{(t)}\nabla^2\mathcal{L}(X, \beta, Y)^{-1}\nabla\mathcal{L}(X, \beta, Y)$$

$$\beta^{(t+1)} = \beta^{(t)} - \mu^{(t)}(X^T X)^{-1}(X^T X\beta^{(t)} - X^T Y)$$

For $\mu^{(t)} = 1$

$$\beta^{(t+1)} = \beta^{(t)} - \beta^{(t)} + (X^T X)^{-1}X^T Y)$$

$$\beta^{(t+1)} = (X^T X)^{-1}X^T Y)$$

This is the same as the closed form solution for linear regression. Hence, using Newton method here will not give any benefit as we will be solving the same system of linear equations.

2. The loss function of the logistic regression has the following form:

$$\boldsymbol{\mathcal{L}(X, \beta, Y) = -\sum_{i=1}^{m} y_i log\big(\sigma(x_i\beta)\big) + (1 - y_i) log(1 - \sigma(x_i\beta))}$$

(a). Compute the Hessian $\boldsymbol{\nabla_{\beta}}^{\mathbf{2}}\boldsymbol{\mathcal{L}(X,\beta,Y)}$

Computing first derivative

$$\nabla_{\beta}\mathcal{L}(X,\beta,Y) = \nabla_{\beta}\left(-\sum_{i=1}^{m} y_i log\big(\sigma(x_i\beta)\big) + (1-y_i)log(1-\sigma(x_i\beta))\right)$$

$$\nabla_{\beta}\mathcal{L}(X,\beta,Y) = \left(-\sum_{i=1}^{m}\frac{y_i}{\sigma(x_i\beta)}\nabla_{\beta}\big(\sigma(x_i\beta)\big)\nabla_{\beta}(x_i\beta) + \frac{1-y_i}{1-\sigma(x_i\beta)}\nabla_{\beta}(1-\sigma(x_i\beta))\,\nabla_{\beta}(1-x_i\beta)\right)$$

Computing $\nabla_{\beta}\big(\sigma(x_i\beta)\big)$

$$\nabla_{\beta}\big(\sigma(x_i\beta)\big) = \nabla_{\beta}\left(\frac{1}{1+e^{-x_i\beta}}\right)$$

$$\nabla_{\beta}\big(\sigma(x_i\beta)\big) = -\frac{\nabla_{\beta}(1+e^{-x_i\beta})}{(1+e^{-x_i\beta})^2} = \frac{e^{-x_i\beta}}{(1+e^{-x_i\beta})^2} = \frac{1-1+e^{-x_i\beta}}{(1+e^{-x_i\beta})^2}$$

$$\nabla_{\beta}\big(\sigma(x_i\beta)\big) = \frac{1}{1+e^{-x_i\beta}}\left(1-\frac{1}{1+e^{-x_i\beta}}\right)$$

$$\boldsymbol{\nabla_{\beta}\big(\sigma(x_i\beta)\big) = \sigma(x_i\beta)\big(1-\sigma(x_i\beta)\big)}$$

So, now

$$\nabla_{\beta}\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m}\frac{y_i}{\sigma(x_i\beta)}\sigma(x_i\beta)\big(1-\sigma(x_i\beta)\big)x_i + \frac{1-y_i}{1-\sigma(x_i\beta)}(\sigma(x_i\beta)\big(1-\sigma(x_i\beta)\big))x_i$$

$$\nabla_{\beta}\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m} x_i y_i\big(1-\sigma(x_i\beta)\big) + x_i(1-y_i)(-\sigma(x_i\beta))$$

$$\nabla_{\beta}\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m} x_i y_i - x_i y_i\sigma(x_i\beta) - x_i\sigma(x_i\beta) + x_i y_i\sigma(x_i\beta)$$

$$\boldsymbol{\nabla_{\beta}\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m} x_i(y_i - \sigma(x_i\beta))}$$

Or in matrix form

$$\boldsymbol{\nabla_{\beta}\mathcal{L}(X,\beta,Y) = -X^{T}(Y-\widehat{Y})}$$

Computing Hessian

$$\nabla_{\beta}{}^{2}\mathcal{L}(X,\beta,Y) = \nabla_{\beta}(-\sum_{i=1}^{m} x_i(y_i - \sigma(x_i\beta)))$$

$$\nabla_\beta{}^2\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m}\nabla_\beta(x_i y_i) - x_i\nabla_\beta(\sigma(x_i\beta))$$

$$\nabla_\beta{}^2\mathcal{L}(X,\beta,Y) = -\sum_{i=1}^{m} 0 - x_i\sigma(x_i\beta)(1-\sigma(x_i\beta))x_i$$

$$\boldsymbol{\nabla_\beta{}^2\mathcal{L}(X,\beta,Y)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{x_i{}^2\sigma(x_i\beta)(1-\sigma(x_i\beta)})$$

Or in matrix form

$$\boldsymbol{\nabla_\beta{}^2\mathcal{L}(X,\beta,Y) = X^T\hat{Y}(1-\hat{Y})X}$$

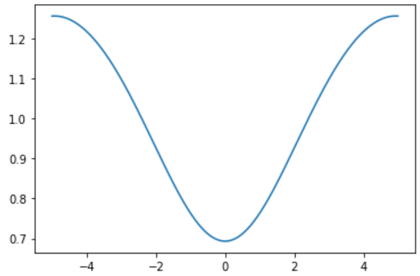
(b). Does it make sense to use the Newton minimization algorithm here? Discuss your answer.

Newton step for logistic regression is given by:

$$\beta^{(t+1)} = \beta^{(t)} - \mu^{(t)}\nabla^2\mathcal{L}(X,\beta,Y)^{-1}\nabla\mathcal{L}(X,\beta,Y)$$

$$\beta^{(t+1)} = \beta^{(t)} - \mu^{(t)}(X^T\hat{Y}(1-\hat{Y})X)^{-1}(-X^T(Y-\hat{Y}))$$

Although computation of the Newton step is costly due to the inverse calculation, the algorithm converges very fast compared to gradient ascent due to its quadratic nature. The reason for fast convergence is that our objective function (logloss function) is convex and twice differentiable and does have a minima as shown in the plot. Therefore, it makes perfect sense to use Newton's algorithm for logistic regression.


1.2
1.1
1.0
0.9
0.8
0.7
−4
−2
0
2
4