Modern Optimization Techniques – Group 01

Exercise Sheet 03

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

Semester 2 MSc. Data Analytics

Question 1: Gradient Descent

Using the GD algorithm reviewed in class, minimize the function showed below:

$$\boldsymbol{f}: \boldsymbol{R^2} \rightarrow \boldsymbol{R} \quad \boldsymbol{f(x_1, x_2) = {x_1}^2 + 3{x_2}^2 + 2x_1 + 0.5x_2}$$

(a) Compute the optimal solution analytically i.e. the minimum $\boldsymbol{x^* = (x_1{}^*, x_2{}^*)}$ and the associated $\boldsymbol{p^*}$

For finding the minimum, I will take the first derivatives of the function with respect to both x_1 and x_2 and put those equal to zero:

$$\frac{d(f(x_1, x_2))}{dx_1} = 2x_1 + 0 + 2 + 0, \;\; 2x_1 + 2 = 0, Thus\ {x_1}^* = -1$$

$$\frac{d(f(x_1, x_2))}{dx_2} = 0 + 6x_2 + 0 + 0.5, \;\; 6x_2 + 0.5 = 0, Thus\ {x_2}^* = -\frac{1}{12}$$

Putting these values in the function to find p^*

$$f({x_1}^*, {x_2}^*) = (-1)^2 + 3(-\frac{1}{12})^2 + 2(-1) + 0.5\left(-\frac{1}{12}\right) = -\frac{49}{48}\ or - 1.0208$$

So

$$\boldsymbol{({x_1}^*, {x_2}^*) = \left(-1, -\frac{1}{12}\right) \;\; \& \;\; p^* = -\frac{49}{48}\ or - 1.0208}$$

(b) Perform 5 iterations and evaluate the function at the end of each iteration. Use an initial point $\boldsymbol{x_0 = (3, -1)}$ and a step size $\boldsymbol{\mu = 0.2}$. Is the algorithm minimizing it?

We have,

$$\frac{d(f(x_1, x_2))}{d(x_1, x_2)} = \begin{bmatrix} 2x_1 + 2 \\ 6x_2 + 0.5 \end{bmatrix}$$

For gradient descent

$$x^i = x^{i-1} - \mu.\frac{d\big(f(x_1, x_2)\big)}{d(x_1, x_2)}, \qquad or \qquad \begin{bmatrix} {x_1}^i \\ {x_2}^i \end{bmatrix} = \begin{bmatrix} {x_1}^{i-1} \\ {x_2}^{i-1} \end{bmatrix} - \mu.\begin{bmatrix} 2{x_1}^{i-1} + 2 \\ 6{x_2}^{i-1} + 0.5 \end{bmatrix}$$

with $x^0 = (3, -1)$, evaluating the function:

$$f(3, -1) = (3)^2 + 3(-1)^2 + 2(3) + 0.5(-1) = 17.5$$

Now we compute x^1i.e. first iteration

$$\begin{bmatrix} {x_1}^1 \\ {x_2}^1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} - 0.2 \begin{bmatrix} 2(3)+2 \\ 6(-1)+0.5 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 1.6 \\ -1.1 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.1 \end{bmatrix}$$

Evaluating the function:

$$f(1.4, 0.1) = (1.4)^2 + 3(0.1)^2 + 2(1.4) + 0.5(0.1) = 4.84$$

Now x^2i.e. second iteration

$$\begin{bmatrix} {x_1}^2 \\ {x_2}^2 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.1 \end{bmatrix} - 0.2 \begin{bmatrix} 2(1.4)+2 \\ 6(0.1)+0.5 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 0.1 \end{bmatrix} - \begin{bmatrix} 0.96 \\ 0.22 \end{bmatrix} = \begin{bmatrix} 0.44 \\ -0.12 \end{bmatrix}$$

Evaluating the function:

$$f(0.44, -0.12) = (0.44)^2 + 3(-0.12)^2 + 2(0.44) + 0.5(-0.12) = 1.0568$$

Now x^3i.e. third iteration

$$\begin{bmatrix} {x_1}^3 \\ {x_2}^3 \end{bmatrix} = \begin{bmatrix} 0.44 \\ -0.12 \end{bmatrix} - 0.2 \begin{bmatrix} 2(0.44)+2 \\ 6(-0.12)+0.5 \end{bmatrix} = \begin{bmatrix} 0.44 \\ -0.12 \end{bmatrix} - \begin{bmatrix} 0.576 \\ -0.044 \end{bmatrix} = \begin{bmatrix} -0.136 \\ -0.076 \end{bmatrix}$$

Evaluating the function:

$$f(-0.136, -0.076) = (-0.136)^2 + 3(-0.076)^2 + 2(-0.136) + 0.5(-0.076) = -0.274$$

Now x^4i.e. fourth iteration

$$\begin{bmatrix} {x_1}^4 \\ {x_2}^4 \end{bmatrix} = \begin{bmatrix} -0.136 \\ -0.076 \end{bmatrix} - 0.2 \begin{bmatrix} 2(-0.136)+2 \\ 6(-0.076)+0.5 \end{bmatrix} = \begin{bmatrix} -0.136 \\ -0.076 \end{bmatrix} - \begin{bmatrix} 0.3456 \\ 0.0088 \end{bmatrix} = \begin{bmatrix} -0.4816 \\ -0.0848 \end{bmatrix}$$

Evaluating the function:

$$f(-0.4816, -0.0848) = (-0.4816)^2 + 3(-0.0848)^2 + 2(-0.4816) + 0.5(-0.0848) = -0.752$$

Now x^5i.e. fifth iteration

$$\begin{bmatrix} {x_1}^5 \\ {x_2}^5 \end{bmatrix} = \begin{bmatrix} -0.4816 \\ -0.0848 \end{bmatrix} - 0.2 \begin{bmatrix} 2(-0.4816)+2 \\ 6(-0.0848)+0.5 \end{bmatrix} = \begin{bmatrix} -0.4816 \\ -0.0848 \end{bmatrix} - \begin{bmatrix} 0.20736 \\ -0.00176 \end{bmatrix} = \begin{bmatrix} -0.689 \\ -0.083 \end{bmatrix}$$

Evaluating the function:

$$f(-0.689, -0.083) = (-0.689)^2 + 3(-0.083)^2 + 2(-0.689) + 0.5(-0.083) = -0.92$$

The algorithm is definitely minimizing the function as the function value after each iteration is going towards the function minimum i.e.

$$f(x_1, x_2) \rightarrow [17.5 > 4.84 > 1.0568 > -0.274 > -0.752 > -0.92] \rightarrow -1.0208$$

(c) What happen if you change the step size to 0:5? Perform 3 iterations and evaluate the function at the end of each iteration as well.

We have,

$$\frac{d(f(x_1, x_2))}{d(x_1, x_2)} = \begin{bmatrix} 2x_1 + 2 \\ 6x_2 + 0.5 \end{bmatrix}$$

For gradient descent

$$x^i = x^{i-1} - \mu . \frac{d\big(f(x_1, x_2)\big)}{d(x_1, x_2)}, \qquad or \qquad \begin{bmatrix} {x_1}^i \\ {x_2}^i \end{bmatrix} = \begin{bmatrix} {x_1}^{i-1} \\ {x_2}^{i-1} \end{bmatrix} - \mu . \begin{bmatrix} 2{x_1}^{i-1} + 2 \\ 6{x_2}^{i-1} + 0.5 \end{bmatrix}$$

with $x^0 = (3, -1)$, evaluating the function:

$$f(3, -1) = (3)^2 + 3(-1)^2 + 2(3) + 0.5(-1) = 17.5$$

Now we compute x^1i.e. first iteration

$$\begin{bmatrix} {x_1}^1 \\ {x_2}^1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} - 0.5 \begin{bmatrix} 2(3) + 2 \\ 6(-1) + 0.5 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 4 \\ -2.75 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.75 \end{bmatrix}$$

Evaluating the function:

$$f(-1, 1.75) = (-1)^2 + 3(1.75)^2 + 2(-1) + 0.5(1.75) = 9.0625$$

Now x^2i.e. second iteration

$$\begin{bmatrix} {x_1}^2 \\ {x_2}^2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.75 \end{bmatrix} - 0.5 \begin{bmatrix} 2(-1) + 2 \\ 6(1.75) + 0.5 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.75 \end{bmatrix} - \begin{bmatrix} 0 \\ 5.5 \end{bmatrix} = \begin{bmatrix} -1 \\ -3.75 \end{bmatrix}$$

Evaluating the function:

$$f(-1, -3.75) = (-1)^2 + 3(-3.75)^2 + 2(-1) + 0.5(-3.75) = 39.3125$$

Now x^3i.e. third iteration

$$\begin{bmatrix} {x_1}^3 \\ {x_2}^3 \end{bmatrix} = \begin{bmatrix} -1 \\ -3.75 \end{bmatrix} - 0.5 \begin{bmatrix} 2(-1) + 2 \\ 6(-3.75) + 0.5 \end{bmatrix} = \begin{bmatrix} -1 \\ -3.75 \end{bmatrix} - \begin{bmatrix} 0 \\ -11 \end{bmatrix} = \begin{bmatrix} -1 \\ 7.25 \end{bmatrix}$$

Evaluating the function:

$$f(-1, 7.25) = (-1)^2 + 3(7.25)^2 + 2(-1) + 0.5(7.25) = 160.3$$

Here we have

$$f(x_1, x_2) \rightarrow [17.5 > 9.0625 < 39.3125 < 160.3] \nrightarrow -1.0208$$

Clearly, the step size is too large and it missed the minima between iteration 1 and 2 as highlighted above. The function will never converge no matter how many more iterations we perform. Therefore, 0.5 is too big a step size.

(d) What is the function being minimized in a non-regularized least squares linear regression? Show how to derive its closed form solution..

The function being minimized in a non-regularized least squares is the loss function given by:

$$RSS(\beta) = \sum_{n=1}^{N} (y_n - \widehat{y_n})^2$$

From the lecture slides, for multiple linear regression with several predictors, we can write the predicted $\hat{y}(x_n)$ as:

$$\widehat{y_n} = \widehat{\beta_0} + \widehat{\beta_1} x_{n,1} + \widehat{\beta_2} x_{n,2} + \cdots + \widehat{\beta_M} x_{n,M}$$

$$= \widehat{\beta_0} + \sum_{m=1}^{M} \widehat{\beta_m} x_{n,m}$$

Which we can write as

$$\widehat{y_n} = \hat{\beta}^T x_n$$

Where

$$\hat{\beta} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \cdots \\ \widehat{\beta_M} \end{bmatrix}, \qquad and \qquad x_n = \begin{bmatrix} 1 \\ x_{n,1} \\ \cdots \\ x_{n,M} \end{bmatrix}$$

Or

$$\hat{y} = X\hat{\beta}$$

And loss function becomes

$$RSS(\hat{\beta}) = \sum_{n=1}^{N} (y_n - \widehat{y_n})^2 = \|y - \hat{y}\|^2 = \left\|y - X\hat{\beta}\right\|^2$$

To find its closed form, we do

$$\hat{\beta} := \underset{\hat{\beta} \in R^M}{argmin} \left\|y - X\hat{\beta}\right\|^2$$

So, we take the derivative of $RSS(\beta)$ and put it = 0

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}} = \frac{\partial \left\|y - X\hat{\beta}\right\|^2}{\partial \hat{\beta}} = \frac{\partial (y - X\hat{\beta})(y - X\hat{\beta})}{\partial \hat{\beta}}$$

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}} = (y - X\hat{\beta})\frac{\partial (y - X\hat{\beta})}{\partial \hat{\beta}} + (y - X\hat{\beta})\frac{\partial (y - X\hat{\beta})}{\partial \hat{\beta}} = (y - X\hat{\beta})(-X) + (y - X\hat{\beta})(-X)$$

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}} = -2X(y - X\hat{\beta}) = -2X^T y + 2X^T X\hat{\beta}$$

Putting equal to zero

$$-2X^T y + 2X^T X\hat{\beta} = 0$$

$$2X^T X\hat{\beta} = 2X^T y$$

$$\boldsymbol{X^T X\hat{\beta} = X^T y \quad or \quad \hat{\beta} = (X^T X)^{-1} X^T y}$$

Which is the required closed form solution.

Question 2: Backtracking Line Search

Let us define a function:

$$\boldsymbol{f: R^2 \to R \quad f(x_1, x_2) = {x_1}^2 + {x_2}^2}$$

(d) Suppose you want to do a backtracking line search using the negative gradient $\Delta x = -\nabla f(x)$ as descent direction. Suppose you are in a current point $x' = (x_1', x_2')$ write down the backtracking condition

$$\boldsymbol{f(x + \mu \Delta x) > f(x) + a\mu \nabla f(x) \Delta x}$$

for these special settings.

Computing left hand side:

Since $\Delta x = -\nabla f(x)$ and $x' = (x_1', x_2')$, we have

$$f(x' + \mu \Delta x') = f(x' - \mu \nabla f(x'))$$

And

$$f(x_1' - \mu \nabla f(x_1'), x_2' - \mu \nabla f(x_2')) = (x_1' - \mu \nabla f(x_1'))^2 + (x_2' - \mu \nabla f(x_2'))^2$$

Now

$$\nabla f(x_1') = 2x_1', \qquad and \qquad \nabla f(x_2') = 2x_2'$$

So

$$f(x_1' - \mu \nabla f(x_1'), x_2' - \mu \nabla f(x_2')) = (x_1' - \mu 2x_1')^2 + (x_2' - \mu 2x_2')^2$$

$$= {x_1'}^2 - 4\mu {x_1'}^2 + 4\mu^2 {x_1'}^2 + {x_2'}^2 - 4\mu {x_2'}^2 + 4\mu^2 {x_2'}^2$$

$$= {x_1'}^2 (1 - 4\mu + 4\mu^{2'}) + {x_2'}^2 (1 - 4\mu + 4\mu^{2'})$$

$$= ({x_1'}^2 + {x_2'}^2)(1 - 4\mu + 4\mu^{2'})$$

Computing right hand side:

Since $\Delta x = -\nabla f(x)$ and $x' = (x_1', x_2')$, we have

$$f(x) = f(x'), and\ a\mu\nabla f(x)\Delta x = -a\mu\nabla f(x')^T\nabla f(x')$$

$$f(x) + a\mu\nabla f(x)\Delta x = f(x') - a\mu\nabla f(x')^T\nabla f(x')$$

$$= x_1'^2 + x_2'^2 - a\mu[2x_1' \quad 2x_2'] \begin{bmatrix} 2x_1' \\ 2x_2' \end{bmatrix}$$

$$= x_1'^2 + x_2'^2 - 4a\mu x_1'^2 - 4a\mu x_2'^2$$

$$= (x_1'^2 + x_2'^2)(1 - 4a\mu)$$

Therefore,

$$f(x' + \mu\Delta x') > f(x') + a\mu\nabla f(x')\Delta x'$$

Becomes

$$(x_1'^2 + x_2'^2)(1 - 4\mu + 4\mu^{2'}) > (x_1'^2 + x_2'^2)(1 - 4a\mu)$$

$$(1 - 4\mu + 4\mu^{2'}) > (1 - 4a\mu)$$

$$1 - 4\mu + 4\mu^2 - 1 + 4a\mu > 0$$

$$4a\mu - 4\mu + 4\mu^2 > 0$$

$$4\mu(a - 1 + \mu) > 0$$

$$a - 1 + \mu > 0$$

$$\boldsymbol{\mu > 1 - a}$$

b) We pick $a = 0.5, b = 0.1$ and start with a rather high initial step size $\mu = 10$. How small does μ have to become for the backtracking condition to be false? How many backtracking iterations will be done until this happens?
[NOTE:] you can use $x = (0.5, 1)$ to show your working for this question.

We have

$$\boldsymbol{f(x - \mu\nabla f(x)) > f(x) - a\mu\nabla f(x)^T\nabla f(x)}$$

First iteration:
Computing LHS

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = \begin{bmatrix} 2(0.5) \\ 2(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$x - \mu\nabla f(x) = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} - 10\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -9.5 \\ -19 \end{bmatrix}$$

$$f(x - \mu\nabla f(x)) = f\left(\begin{bmatrix} -9.5 \\ -19 \end{bmatrix}\right) = -9.5^2 + -19^2 = 451.25$$

Computing RHS

$$f(x) = f\left(\begin{bmatrix}0.5\\1\end{bmatrix}\right) = {x_1}^2 + {x_2}^2 = 0.5^2 + 1^2 = 1.25$$

$$a\mu\nabla f(x)^T\nabla f(x) = (0.5)(10)[1 \quad 2]\begin{bmatrix}1\\2\end{bmatrix} = (0.5)(10)(5) = 25$$

$$f(x) - a\mu\nabla f(x)^T\nabla f(x) = 1.25 - 25 = -23.75$$

Checking condition:

$$f\big(x - \mu\nabla f(x)\big) > f(x) - a\mu\nabla f(x)^T\nabla f(x) \Rightarrow 451.25 > -23.75 \Rightarrow True$$

Updating μ

$$\mu = b\mu = 0.1(10) = 1$$

Second iteration:

Computing LHS

$$x - \mu\nabla f(x) = \begin{bmatrix}0.5\\1\end{bmatrix} - 1\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}-0.5\\-1\end{bmatrix}$$

$$f\big(x - \mu\nabla f(x)\big) = f\left(\begin{bmatrix}-0.5\\-1\end{bmatrix}\right) = -0.5^2 + -1^2 = 1.25$$

Computing RHS

$$a\mu\nabla f(x)^T\nabla f(x) = (0.5)(1)[1 \quad 2]\begin{bmatrix}1\\2\end{bmatrix} = (0.5)(1)(5) = 2.5$$

$$f(x) - a\mu\nabla f(x)^T\nabla f(x) = 1.25 - 2.5 = -1.25$$

Checking condition:

$$f\big(x - \mu\nabla f(x)\big) > f(x) - a\mu\nabla f(x)^T\nabla f(x) \Rightarrow 1.25 > -1.25 \Rightarrow True$$

Updating μ

$$\mu = b\mu = 0.1(1) = 0.1$$

Third iteration:

Computing LHS

$$x - \mu\nabla f(x) = \begin{bmatrix}0.5\\1\end{bmatrix} - 0.1\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}0.4\\0.8\end{bmatrix}$$

$$f\big(x - \mu\nabla f(x)\big) = f\left(\begin{bmatrix}0.4\\0.8\end{bmatrix}\right) = 0.4^2 + 0.8^2 = 0.8$$

Computing RHS

$$a\mu\nabla f(x)^T\nabla f(x) = (0.5)(0.1)[1 \quad 2]\begin{bmatrix}1\\2\end{bmatrix} = (0.5)(0.1)(5) = 0.25$$

$$f(x) - a\mu\nabla f(x)^T\nabla f(x) = 1.25 - 0.25 = 1$$

Checking condition:

$$f\big(x - \mu\nabla f(x)\big) > f(x) - a\mu\nabla f(x)^T\nabla f(x) \Rightarrow 0.8 \not> 1 \Rightarrow False$$

Since the condition is violated in the third iteration, optimum value of μ has been achieved in 2 iterations. Therefore,

$\mu = 0.1$ has made the backtracking condition false and it took 2 backtracking iterations