# Numerical Optimization
# Week 3 Assignment

**CHM564**

## 1 Introduction

In this week, I implemented the *SteepestDescent* and *Newtons* algorithm using a backtracking Line-Search. This report will discuss the choice of parameters and the results of these two algrorithm in different functions.

## 2 Theory analysis

### 2.1 Exact decrease

To proof (3.28) on book, we have

$$x_{k+1} = x_k - \alpha \bigtriangledown f \Longrightarrow x_{k+1} - x^* = x_k - x^* - \alpha \bigtriangledown f$$

By $\| \cdot \|_Q^2$, we can have:

$$\begin{aligned}
\|x_{k+1} - x^*\|_Q^2 &= (x_{k+1} - x^*)^T Q(x_{k+1} - x^*) \\
&= (x_k - x^* - \alpha \bigtriangledown f)^T Q(x_k - x^* - \alpha \bigtriangledown f) \\
&= (x_k - x^*)^T Q(x_k - x^*) - 2\alpha \bigtriangledown f^T Q(x_k - x^*) + \alpha^2 \bigtriangledown f^T Q \bigtriangledown f \\
&= \|x_k - x^*\|_Q^2 - 2\alpha \bigtriangledown f^T Q(x_k - x^*) + \alpha^2 \bigtriangledown f^T Q \bigtriangledown f
\end{aligned}$$

Then, by $\bigtriangledown f = Q(x_k - x^*)$ and $\alpha = \frac{\bigtriangledown f^T \bigtriangledown f}{\bigtriangledown f^T Q \bigtriangledown f}$

$$\begin{aligned}
\|x_{k+1} - x^*\|_Q^2 &= \|x_k - x^*\|_Q^2 - 2\alpha \bigtriangledown f^T \bigtriangledown f + \alpha^2 \bigtriangledown f^T Q \bigtriangledown f \\
&= \|x_k - x^*\|_Q^2 - \frac{(\bigtriangledown f^T \bigtriangledown f)^2}{\bigtriangledown f^T Q \bigtriangledown f} \\
&= \|x_k - x^*\|_Q^2 [1 - \frac{(\bigtriangledown f^T \bigtriangledown f)^2}{(\bigtriangledown f^T Q \bigtriangledown f)\|x_k - x^*\|_Q^2}] \\
&= \|x_k - x^*\|_Q^2 [1 - \frac{(\bigtriangledown f^T \bigtriangledown f)^2}{(\bigtriangledown f^T Q \bigtriangledown f)(\bigtriangledown f^T Q^( - 1) \bigtriangledown f))}]
\end{aligned}$$

## 2.2 Invariance properties

If $g(x) = f(Ax)$, then $\bigtriangledown g(x) = A^T \bigtriangledown f(Ax)$
*Proof* :

$$Ax = A_{ij}x_{ij} = \begin{bmatrix} a_{11}x_{11} & ... & a_{j1}x_{1j} \\ ... & & ... \\ a_{1j}x_{j1} & ... & a_{ij}x_{ij} \end{bmatrix}$$

$$\bigtriangledown g(x) = f'(Ax) = \begin{bmatrix} a_{11} & ... & a_{j1} \\ ... & & ... \\ a_{1j} & ... & a_{ij} \end{bmatrix} f(A_{ij}x_{ij}) = A^T \bigtriangledown f(Ax)$$

Then, we assume $\tilde{x_0} = A^T x_0$, for steepest descent with back tracking line search, we can get:$\alpha\rho \to \alpha, p = - \bigtriangledown g(\tilde{x_i})$ until:

$$f(x_k + \alpha\rho_k) \leq f(x_k) + c\alpha \bigtriangledown f_k^T p_k \qquad \rho \in (0,1), c \in (0,1)$$

$$\begin{aligned} \tilde{x_i} &= \tilde{x}_{i-1} + \alpha p \\ &= \tilde{x}_{i-1} - \alpha \bigtriangledown g(\tilde{x}_{i-1}) \\ &= A^T x_{i-1} - \alpha A^T \bigtriangledown f(Ax_{i-1}) \\ &= A^T(x_{i-1} - \alpha \bigtriangledown f(Ax_{i-1})) \end{aligned}$$

$$\begin{aligned} \alpha_{i-1} &= \frac{\bigtriangledown g_{i-1}^T \bigtriangledown g_{i-1}}{\bigtriangledown g_{i-1}^T Q \bigtriangledown g_{i-1}} \\ &= \frac{A \bigtriangledown f_{i-1}^T A^T \bigtriangledown f_{i-1}}{A \bigtriangledown f_{i-1}^T Q A^T \bigtriangledown f_{i-1}} \\ &= \frac{\bigtriangledown f_{i-1}^T \bigtriangledown f_{i-1}}{\bigtriangledown f_{i-1}^T Q \bigtriangledown f_{i-1}} \end{aligned}$$

So,

$$\begin{aligned} \tilde{x_i} &= A^T(x_{i-1} - \alpha \bigtriangledown f(Ax_{i-1})) \\ &= A^T x_i \end{aligned}$$

So, the sequence of iterates satisfies $\tilde{x_i} = A^T$ and $f(x_i) = g(\tilde{x_i})$. For some function, which convex area is small, like $f_3$. We can use invariance property, to increase the convex interval. So that we can aviod overflow.

## 2.3 Convergence of Non-convex function

For function $f$, there exists a constant $L$, such that $p^T \bigtriangledown^2 f(x)p \leq L\|h\|^2$,

$$\begin{aligned} f(x + \alpha p) - f(x) &= \int \bigtriangledown f^T(x + \tau p)p \, d\tau \\ &= \alpha \bigtriangledown f^T(x)p + \int (\bigtriangledown f^T(x + \tau p) - \bigtriangledown f^T(x))p \, d\tau \\ &\leq \alpha \bigtriangledown f^T(\alpha)p + L\|h\|^2 \int \tau \, d\tau \end{aligned}$$

If $p = - \bigtriangledown f$

$$f(x+\alpha p) - f(x) \leq -\alpha \| \bigtriangledown f(x)\|^2 + \frac{\alpha^2 L}{2} \| \bigtriangledown f(x)\|^2 = (\frac{\alpha^2 L}{2} - \alpha) \| \bigtriangledown f(x)\|^2$$

By differentiating the function with respect to $\alpha$ and setting the derivative to zero, we obtain:

$$\alpha * = \frac{1}{L}$$
$$f(x+\alpha p) - f(x) \leq -\frac{1}{2L} \| \bigtriangledown f\|^2$$

For non-convex function,

$$\sum_{k=0}^{M} (f(x^{k+1}) - f(x^k)) = f(x^M) - f(x^0) \leq -\frac{1}{2L} \sum_{0}^{M} \| \bigtriangledown f(x^k)\|^2$$
$$\Rightarrow \frac{1}{2L} \sum_{0}^{M} \| \bigtriangledown f(x^k)\|^2 \leq f(x^0) - f(x^M) \leq C$$

Thus, if $k \to \infty$,gradient$\| \bigtriangledown f(x^k)\| \to 0$,convergence speed in linear.

### 2.4 Steepest descent

Gradient descent method is a special case of the Steepest descent. In the steepest descent, for a $\|\epsilon\|$:

$$\triangle \epsilon_{nsd} = argmin_v(\bigtriangledown f(x)^T \epsilon | \|\epsilon\| = 1)$$

When we use Euclidean norm, the steepest descent is gradient descent. Therefore, the following discussions are based on gradient descent.

## 3 Implement the algorithm

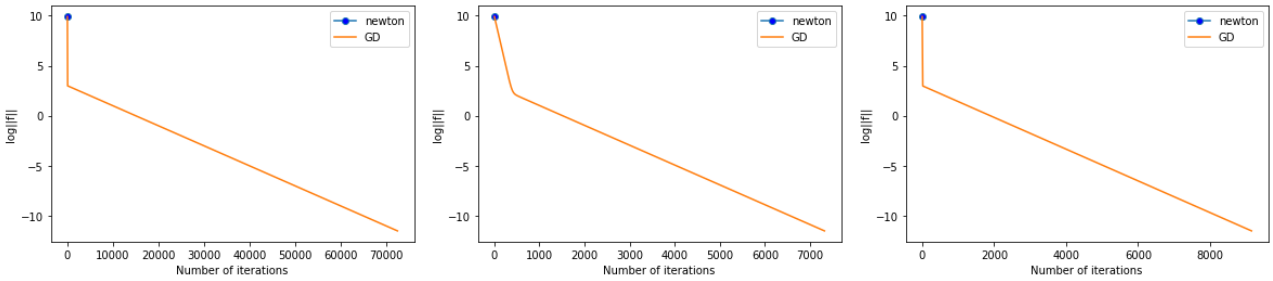
### 3.1 Description of algorithm

We use backtracking line search to choose the step size. If the step size of each gradient descent is fixed, convergence maynot be achieved. If each step is very small, the descent rate will be very slow, requiring many rounds of iterations. So the choice of step size and convergence speed is a trade-off relationship. We use backtracking to find the best situations.

The main problem with Newton's method is that when *Hessian* is not positive definite, its difficult to find the direction.
For steepest algorithm, when *Hessian* is not positive, we use fastest descent direction as next step's direction. For (3.4) on book, it introduces a correction matrix and $B_k = G_k + E_k$ is sufficiently positive definite. However, we find that if we want to it be convergent, $E_k$ must satisfy:
There exists a parameter $C \geq 0$, for any $k$:

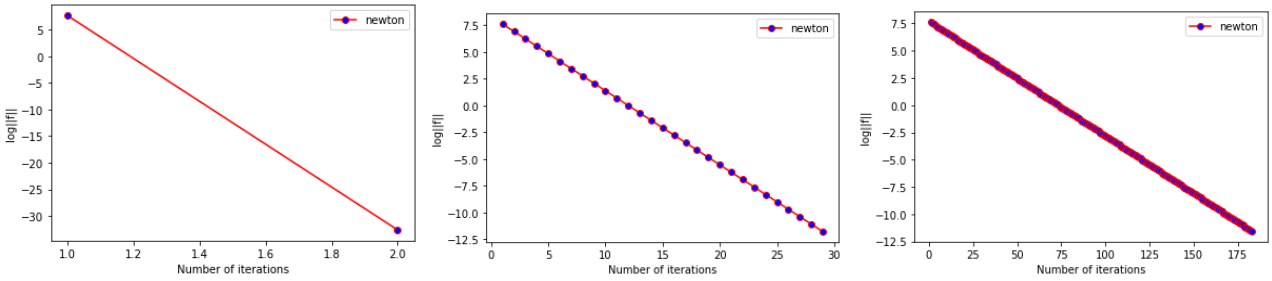
$$cond(B_k) = \|B_k\| \|B_k^{-1}\| \leq C$$

## 3.2 Paremeter selection




(a) c=$10^{-4}$ $\rho = 0.1$ (b) c=$10^{-4}$ $\rho = 0.8$ (c) c=0.1 $\rho = 0.8$

Figure 1: Test on Function1




(a) $\alpha = 1$ (b) $\alpha = 0.5$ (c) $\alpha = 0.1$

Figure 2: Function1 in d=5

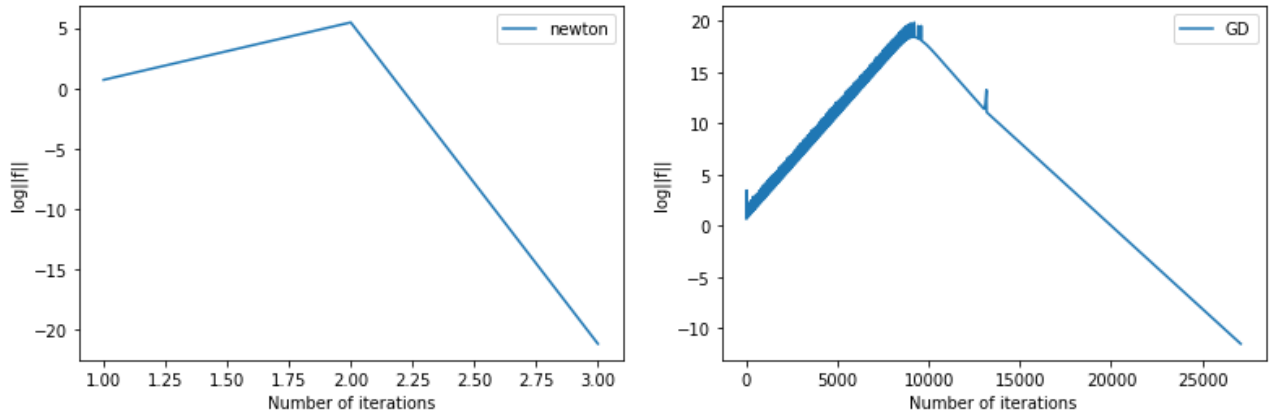
As we know, for backtracking, we need choose $\rho \in (0,1), c \in (0,1)$ and repeat $\rho\alpha \to \alpha$ until

$$f(x_k + \alpha\rho_k) \leq f(x_k) + c\alpha \bigtriangledown f_k^T p_k$$

For inital step, *Newton* can use $\alpha = 1$ to start iterate. But for gradient descent, we will use $\alpha_1 = \frac{\bigtriangledown f_1^T \bigtriangledown f_1}{\bigtriangledown f_1^T Q \bigtriangledown f_1}$.
Firstly, we can easily get if $\rho$ small, the step length $\alpha$ will be short. And the number of iterations will be large. We need $\alpha$ achieves sufficient decrease but isn't too short. Then, $c$ is always chosen to be quite small, it determines the number of $\alpha$ changes. So, in this test, we choose $\rho = 0.8$

From Figure 1, we can see that the results by gradient descent always show different with parameters' change. However, no matter what parameters we choose, *Newton* always iterate 2 times and the second time $\log \|f\| \to \infty$. For function1, *Newton* is not sensitive.
I use $\| \bigtriangledown f(x)\|^2 \leq 10^{-10}$ and iteration$\leq$ 10000 as stopping criterias. For the first item is to make sure the accuary. The second item is because some functions may lose into cycle.
For Figure 2, we want to know the results on ellipsoid with different choices of $\alpha$ in $d = 5$. We can see clearly that, with $\alpha$ became smaller, the number of iterations became more. Also its easy to understand. x drops long in each iteration, so the number of all iterations will be small.

### 3.3 Test on non-convex function




(a) Newton (b) Gradient descent

Figure 3: Test on non-convex area

From previous assignments, for function3, when $\|x\| \leq \sqrt{\epsilon}$, $f_3$ is convex. So, we choose initial points out of this area, to test the algorithms. We can see that, the function can converage. The algorithms can be used in non-convex functions.

## 4 Summary

In this report, we mainly described the algorithms and disscussed tha choice of parameters and stopping criterias. *Gradientdescent* is sensitive by different parameters. In addition, *Newton* always shows better than *GD*.