

# 236330 - Introduction to Optimization: Homework #2

May 11, 2020

Amit Rotner  
123456789  
Or Steiner  
123456789

## Gradient Descent method and Newton method

### Task 1 – Convex sets and functions:

#### Q1:

Show that if  $f_1$  and  $f_2$  are convex functions on a convex domain  $C$ , then  $g(x) = \max_{i=1,2} f_i(x)$  is also a convex function.

#### Solution:

We know that  $f_1$  is a convex function, hence  $\forall x_1, x_2 \in C, \forall \alpha \in [0, 1]$ :

$$f_1(\alpha x_1 + (1 - \alpha) x_2) \leq \alpha f_1(x_1) + (1 - \alpha) f_1(x_2)$$

In addition,  $f_2$  is a convex function, hence  $\forall x_1, x_2 \in C, \forall \alpha \in [0, 1]$ :

$$f_2(\alpha x_1 + (1 - \alpha) x_2) \leq \alpha f_2(x_1) + (1 - \alpha) f_2(x_2)$$

Now, let  $x_1, x_2 \in C, \alpha \in [0, 1]$ :

$$\begin{aligned} g(\alpha x_1 + (1 - \alpha) x_2) &= \max_{i=1,2} f_i(\alpha x_1 + (1 - \alpha) x_2) \\ &\leq \max_{i=1,2} (\alpha f_i(x_1) + (1 - \alpha) f_i(x_2)) \\ &\leq \alpha \max_{i=1,2} f_i(x_1) + (1 - \alpha) \max_{i=1,2} f_i(x_2) \\ &= \alpha g(x_1) + (1 - \alpha) g(x_2) \end{aligned}$$

Therefore,  $g(x)$  is a convex function.

#### Q2:

Let  $f(x)$  be a convex function defined over a convex domain  $C$ .

Show that the level set  $L = \{x \in C : f(x) \leq \alpha\}$  is convex.

**Solution:**

We know that  $f$  is a convex function, hence  $\forall x_1, x_2 \in C, \forall \beta \in [0, 1]$ :

$$f(\beta x_1 + (1 - \beta) x_2) \leq \beta f(x_1) + (1 - \beta) f(x_2)$$

Let  $x_1, x_2 \in L$ :

From the definition of  $L$ ,  $f(x_1) \leq \alpha$  and  $f(x_2) \leq \alpha$ .

Now, let  $\beta \in [0, 1]$ :

$$\begin{aligned} f(\beta x_1 + (1 - \beta) x_2) &\leq \beta f(x_1) + (1 - \beta) f(x_2) \\ &\leq \beta \alpha + (1 - \beta) \alpha \\ &= \alpha \end{aligned}$$

Therefore,  $\beta x_1 + (1 - \beta) x_2 \in L$ . Hence,  $L$  is a convex set.

**Q3:**

Let  $f(x)$  be a smooth and twice differentiable convex function continuously.

Show that  $g(x) = f(Ax)$  is convex, where  $A$  is a matrix of appropriate size.

Check positive semi-definiteness of the Hessian.

**Solution:**

We know that  $f$  is a convex function, hence  $\forall x_1, x_2 \in C, \forall \beta \in [0, 1]$ :

$$f(\beta x_1 + (1 - \beta) x_2) \leq \beta f(x_1) + (1 - \beta) f(x_2)$$

Let  $x_1, x_2 \in L, \beta \in [0, 1]$ :

$$\begin{aligned} g(\beta x_1 + (1 - \beta) x_2) &= f(A\beta x_1 + A(1 - \beta) x_2) \\ &= f(\beta Ax_1 + (1 - \beta) Ax_2) \\ &\leq \beta f(Ax_1) + (1 - \beta) f(Ax_2) \\ &= \beta g(Ax_1) + (1 - \beta) g(Ax_2) \end{aligned}$$

Therefore,  $g(x)$  is a convex function.

As we have seen in HW1, the Hessian of  $g$  equals to:

$$H(x) = A^T \nabla^2 f(Ax) A$$

Given that  $f$  is smooth and twice differentiable convex function continuously,  $\nabla^2 f$  is PSD.

Therefore,  $\forall x \in \mathbb{R}^n : x^T \nabla^2 f x \geq 0$ .

$\forall x \in \mathbb{R}^n$ :

$$x^T H x = x^T A^T \nabla^2 f A x \stackrel{z=Ax}{=} z^T \nabla^2 f z \geq 0$$

Hence,  $H$  is PSD.

**Q4:**

Phrase and prove Jensen's inequality for the discrete case.

**Solution:**

Jensen's inequality:

Let  $f : C \rightarrow \mathbb{R}$  be a convex function and let  $x_1, x_2, \dots, x_n \in C$ .

If  $\alpha_1, \alpha_2, \dots, \alpha_n$  are positive numbers such that  $\sum_{i=1}^n \alpha_i = 1$  then:

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i f(x_i)$$

Proof, by induction:

- For  $n = 2$ :  $\alpha_2 = 1 - \alpha_1$ .

The statement:

$$f(\alpha_1 x_1 + (1 - \alpha_1) x_2) \leq \alpha_1 f(x_1) + (1 - \alpha_1) f(x_2)$$

is true by the convexity of  $f$ .

- Suppose that the statement is true for some  $n$ , we need to prove that it's true for  $n + 1$ :

Let  $\alpha_1, \alpha_2, \dots, \alpha_{n+1}$  be positive numbers such that  $\sum_{i=1}^{n+1} \alpha_i = 1$ .

By the convexity of  $f$ :

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \alpha_i x_i\right) &= f\left(\alpha_1 x_1 + \sum_{i=2}^{n+1} \alpha_i x_i\right) \\ &= f\left(\alpha_1 x_1 + (1 - \alpha_1) \sum_{i=2}^{n+1} \frac{\alpha_i}{1 - \alpha_1} x_i\right) \\ &\leq \alpha_1 f(x_1) + (1 - \alpha_1) f\left(\sum_{i=2}^{n+1} \frac{\alpha_i}{1 - \alpha_1} x_i\right) \end{aligned}$$

Since  $\sum_{i=2}^{n+1} \frac{\alpha_i}{1 - \alpha_1} = 1$ , by the induction hypotheses we get:

$$\begin{aligned} \alpha_1 f(x_1) + (1 - \alpha_1) f\left(\sum_{i=2}^{n+1} \frac{\alpha_i}{1 - \alpha_1} x_i\right) &\leq \alpha_1 f(x_1) + (1 - \alpha_1) \sum_{i=2}^{n+1} \frac{\alpha_i}{1 - \alpha_1} f(x_i) \\ &= \alpha_1 f(x_1) + \sum_{i=2}^{n+1} \alpha_i f(x_i) \\ &= \sum_{i=1}^{n+1} \alpha_i f(x_i) \end{aligned}$$

■

**Q5:**

Using Jensen inequality, prove arithmetic geometric mean inequality:

$$\frac{x_1 + x_2 + \dots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Where  $\forall i, x_i > 0$ .

**Solution:**

We know that  $\log(\cdot)$  is a concave function. Hence,  $-\log(\cdot)$  is a convex function.

Therefore, using Jensen's inequality, where  $\forall i: \alpha_i = \frac{1}{n}$  yields:

$$\begin{aligned}
\log\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) &= \log\left(\sum_{i=1}^n \alpha_i x_i\right) \\
&\geq \sum_{i=1}^n \alpha_i \log(x_i) \\
&= \frac{1}{n} \log(x_1) + \cdots + \frac{1}{n} \log(x_n) \\
&= \frac{1}{n} \log(x_1 \cdot x_2 \cdots x_n) \\
&= \log(\sqrt[n]{x_1 \cdot x_2 \cdots x_n})
\end{aligned}$$

Since  $\log(\cdot)$  is strictly increasing,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

■

## Task 2 – Gradient Descent Analytical Convergence:

### Q6:

Let  $f(x) = \frac{1}{2}x^T Qx - b^T x + c$  be the function to minimize, where  $Q \succ 0$  is a symmetric matrix.

1. We will define the condition number of a positive definite matrix  $A$  as  $\theta \triangleq \frac{\lambda_{max}}{\lambda_{min}}$ .

Write an upper bound on the convergence ratio  $\beta$  that we found in the tutorial, using  $\theta(Q)$  - the condition number of  $Q$ .

2. Assume that the step size can be modified at any iteration.  
Find the optimal step size  $\alpha_k^*$ .

### Solution:

We start by finding  $x^*$ :

$$\begin{aligned}
df &= Qx - b = 0 \\
Qx^* &= b \\
x^* &= Q^{-1}b
\end{aligned}$$

$\nabla f(x^*) = 0$  Hence,  $\nabla f(x^*) = Qx^* - b$

From the definition of the gradient decent method,

$$x_{k+1} = x_k - \alpha \nabla f(x_k) = x_k - \alpha (Qx_k - b) = (I - \alpha Q)x_k + \alpha b$$

1. From the above, we get:

$$x_{k+1} = (I - \alpha Q) x_k + \alpha b = (I - \alpha Q) x_k + \alpha Q x^*$$

And therefore,

$$x_{k+1} - x^* = (I - \alpha Q) x_k + \alpha Q x^* - x^* = (I - \alpha Q) x_k + (\alpha Q - I) x^* = (I - \alpha Q) (x_k - x^*)$$

As we have learned in Numerical Algorithms,  $\|Ax\| \leq \|A\| \cdot \|x\|$  for every norm.

Hence:

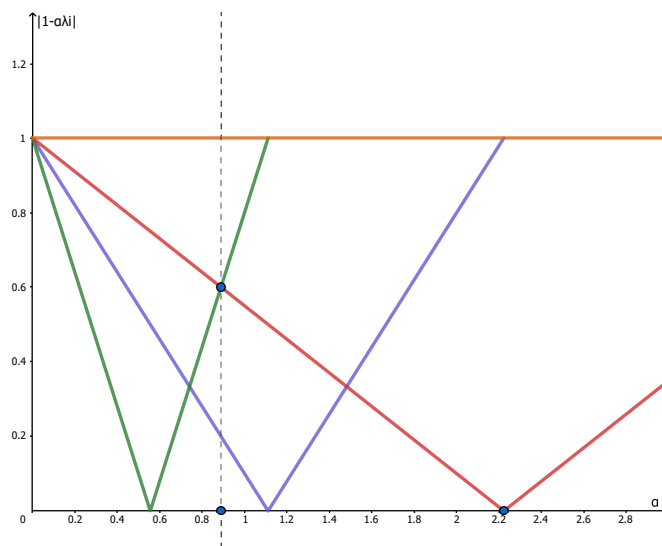
$$\|(I - \alpha Q) (x_k - x^*)\|_2 \leq \|(I - \alpha Q)\|_2 \cdot \|x_k - x^*\|_2$$

$$\begin{aligned} \beta &\triangleq \frac{\|x_{k+1} - x^*\|_2}{\|x_k - x^*\|_2} \\ &= \frac{\|(I - \alpha Q) (x_k - x^*)\|_2}{\|x_k - x^*\|_2} \\ &\leq \frac{\|(I - \alpha Q)\|_2 \cdot \|x_k - x^*\|_2}{\|x_k - x^*\|_2} \\ &= \|(I - \alpha Q)\|_2 \\ &= \sigma_{\max}(I - \alpha Q) \\ &= \max_i \{ |1 - \alpha \lambda_i| \} \end{aligned}$$

Hence, we need to find:

$$\alpha_{\text{opt}} = \operatorname{argmin}_{\alpha} \max_i \{ |1 - \alpha \lambda_i| \}$$

We can solve this problem graphically:



Looking on the graph we see that the optimal solution satisfies:

$$1 - \alpha_{opt}\lambda_{min} = \alpha_{opt}\lambda_{max} - 1$$

$$\alpha_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}}$$

Therefore,

$$\begin{aligned}\beta &\leq \max_i \{|1 - \alpha_{opt}\lambda_i|\} \\ &= \max \{1 - \alpha_{opt}\lambda_{min}, 1 - \alpha_{opt}\lambda_{max}\} \\ &= 1 - \alpha_{opt}\lambda_{min} \\ &= 1 - \frac{2}{\lambda_{min} + \lambda_{max}}\lambda_{min} \\ &= \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} \\ &= \frac{\theta(Q) - 1}{\theta(Q) + 1}\end{aligned}$$

2. We know from above that:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) = x_k + \alpha_k d_k$$

The optimal step is given by:

$$\alpha_k^* = \operatorname{argmin}_{\alpha_k} f(x_{k+1}) = \operatorname{argmin}_{\alpha_k} f(x_k - \alpha_k \nabla f(x_k)) = \operatorname{argmin}_{\alpha_k} f(x_k + \alpha_k d_k)$$

Since  $f(x) = \frac{1}{2}x^T Q x - b^T x + c$ :

$$\begin{aligned}\alpha_k^* &= \operatorname{argmin}_{\alpha_k} \frac{1}{2} (x_k + \alpha_k d_k)^T Q (x_k + \alpha_k d_k) - b^T (x_k + \alpha_k d_k) + c \\ &= \operatorname{argmin}_{\alpha_k} \frac{1}{2} \|A(x_k + \alpha_k d_k)\|^2 - b^T (x_k + \alpha_k d_k) + c\end{aligned}$$

$$\frac{df(x_k - \alpha_k d_k)}{\alpha_k} = d_k^T Q (x_k + \alpha_k d_k) - d_k^T b = 0$$

$$d_k^T (Qx_k - b) + \alpha_k d_k^T Q d_k = 0$$

Because  $\nabla f(x_k) = Qx_k - b$ , we get that  $d_k = -\nabla f(x_k) = b - Qx_k$ .

Hence,

$$0 = d_k^T (Qx_k - b) + \alpha_k d_k^T Q d_k = -d_k^T d_k + \alpha_k d_k^T Q d_k$$

$$\alpha_k d_k^T Q d_k = d_k^T d_k$$



Assuming that  $d_k \neq 0$  (otherwise,  $x_k = x^*$ ), we have that  $d_k^T Q d_k > 0$  because  $Q \succ 0$ .

Therefore:

$$\alpha_k^* = \frac{d_k^T Q d_k}{d_k^T d_k}$$

**Q7:**

Let there be a strongly convex function  $f(x)$ .

Prove that if  $\forall x \in \text{Dom}(f) : mI \preceq \nabla^2 f(x) \preceq MI$  then:

$$\frac{1}{2m} \|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2M} \|\nabla f(x)\|_2^2$$

**Solution:**

From Taylor's multivariate theorem we know that:

$$\forall x, y \in \mathbb{R}^n, \exists z \in [x, y] : f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(z) (y - x)$$

Strong convexity implies that:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T M (y - x) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} M \|y - x\|_2^2$$

The above function minimized at  $y^* = x - \frac{1}{M} \nabla f(x)$ .

Therefore:

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T (y^* - x) + \frac{1}{2} M \|y^* - x\|_2^2 \\ &= f(x) + \nabla f(x)^T \left( x - \frac{1}{M} \nabla f(x) - x \right) + \frac{1}{2} M \left\| x - \frac{1}{M} \nabla f(x) - x \right\|_2^2 \\ &= f(x) - \frac{1}{M} \nabla f(x)^T \nabla f(x) + \frac{M}{2} \frac{1}{M^2} \|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{M} \|\nabla f(x)\|_2^2 + \frac{1}{2M} \|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2 \end{aligned}$$

On the other side of the inequality, we get:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T m (y - x) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} m \|y - x\|_2^2$$

The above function minimized at  $y^* = x - \frac{1}{m} \nabla f(x)$ .

Therefore:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (y^* - x) + \frac{1}{2} \|y^* - x\|_2^2 \\ &= f(x) + \nabla f(x)^T \left( x - \frac{1}{m} \nabla f(x) - x \right) + \frac{1}{2} m \left\| x - \frac{1}{m} \nabla f(x) - x \right\|_2^2 \\ &= f(x) - \frac{1}{m} \nabla f(x)^T \nabla f(x) + \frac{m}{2} \frac{1}{m^2} \|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{m} \|\nabla f(x)\|_2^2 + \frac{1}{2m} \|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \end{aligned}$$

By choosing  $y = x^*$  we get:

$$f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 \geq f(x^*) \geq f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2$$

$$-\frac{1}{2m} \|\nabla f(x)\|_2^2 \geq f(x^*) - f(x) \geq -\frac{1}{2M} \|\nabla f(x)\|_2^2$$

$$\frac{1}{2m} \|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2M} \|\nabla f(x)\|_2^2$$

■

**Q8:**

**Solution:**

1. Given the Rosenbrock function:

$$f((x_1, x_2, \dots, x_N)) = \sum_{i=1}^{N-1} \left[ (1 - x_i)^2 + 100 (x_{i+1} - x_i^2)^2 \right]$$

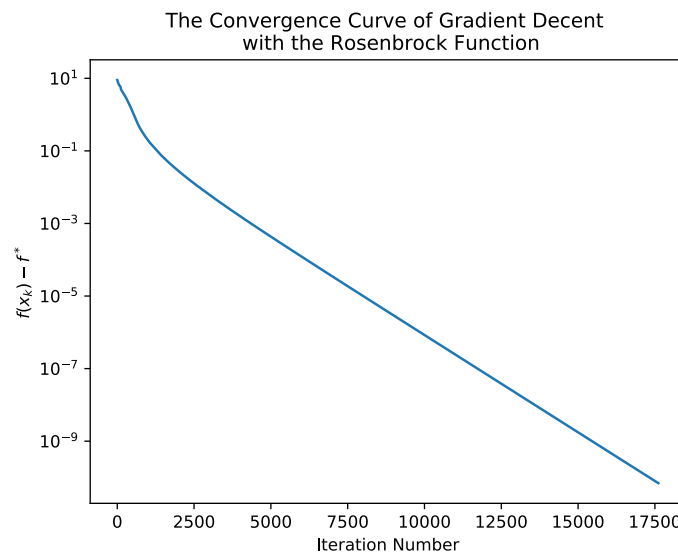
Deriving analytically the gradient, yields:

$$\nabla f(x) = \begin{pmatrix} -2(1-x_1) - 400x_1(x_2 - x_1^2) \\ -2(1-x_2) - 400x_2(x_3 - x_2^2) + 200(x_2 - x_1^2) \\ \vdots \\ -2(1-x_{N-1}) + 400x_{N-1}(x_N - x_{N-1}^2) + 200(x_N - x_{N-1}^2) \\ 200(x_N - x_{N-1}^2) \end{pmatrix}$$

Deriving analytically the Hessian, yields:

$$\nabla^2 f(x) = \begin{pmatrix} 2 - 400x_2 + 1200x_1^2 & -400x_1 & & & \\ -400x_1 & 202 - 400x_3 + 1200x_2^2 & & & \\ & & \ddots & & \\ & & & -400x_{N-1} & \\ & & & 202 - 400x_N + 1200x_{N-1}^2 & -400x_N \\ & & & & -400x_{N-1} & 200 \end{pmatrix}$$

2. Using the Gradient Descent method with the starting point  $x_0 = (0, 0, \dots, 0)$  and  $N = 10$ , we get:

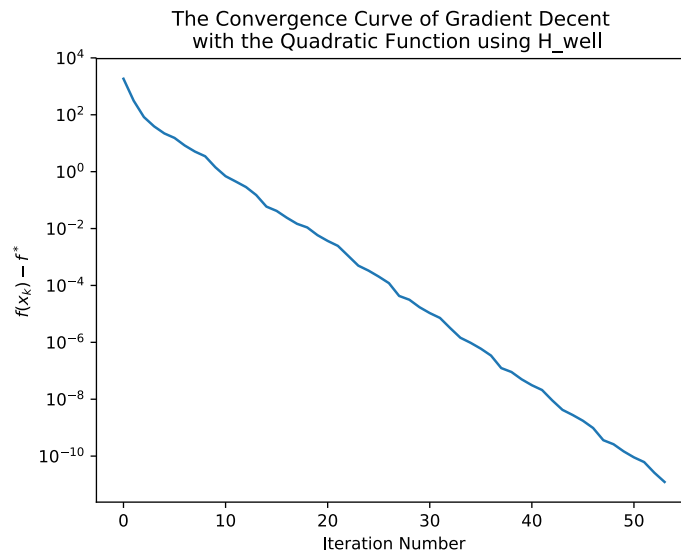


TODO: Add a short explanation about the graph

3. Given the quadratic function  $f(x) = \frac{1}{2}x^T Hx$ .

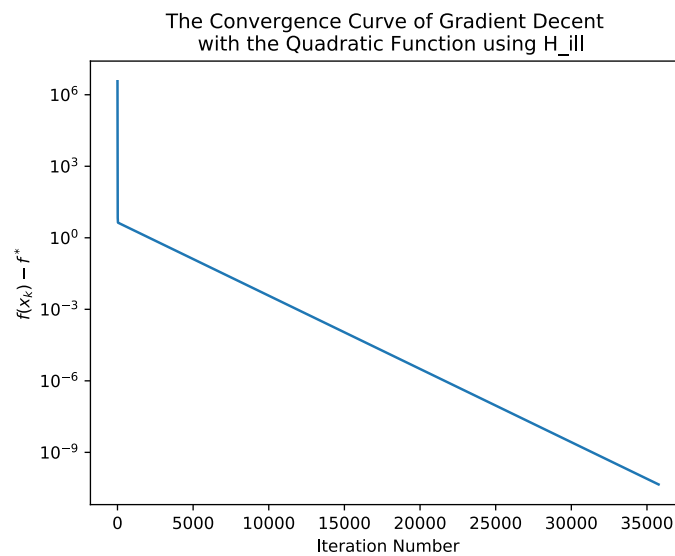
Using the Gradient Descent method with the starting point  $x_{0\_quad}$ , we get:

- With  $H = H\_well$ :



TODO: Add a short explanation about the graph

- With  $H = H_{ill}$ :



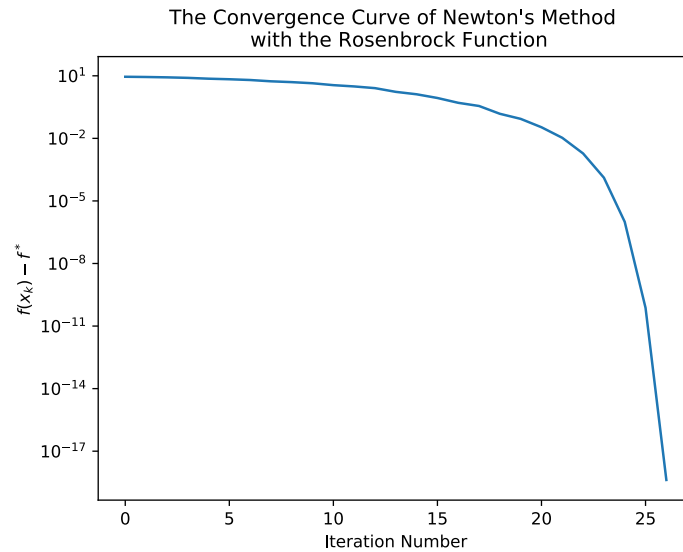
TODO: Add a short explanation about the graph

**Q9:**

**Solution:**

1. Given the Rosenbrock function,

Using the Newton's method with the starting point  $x_0 = (0, 0, \dots, 0)$  and  $N = 10$ , we get:

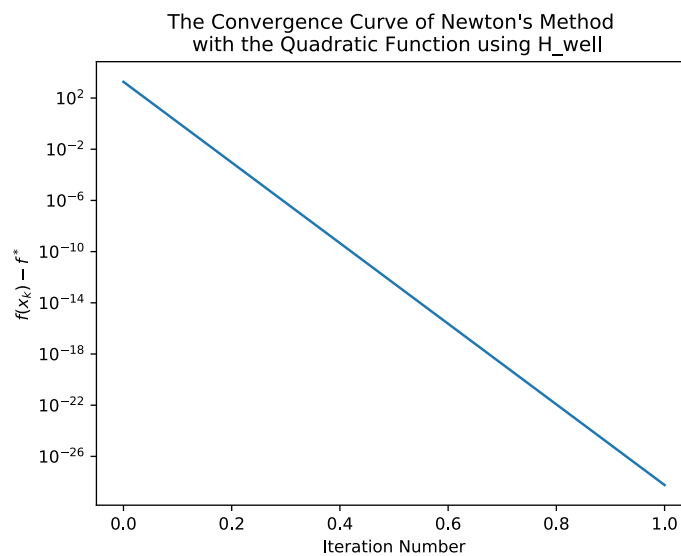


TODO: Add a short explanation about the graph

2. Given the quadratic function  $f(x) = \frac{1}{2}x^T H x$ ,

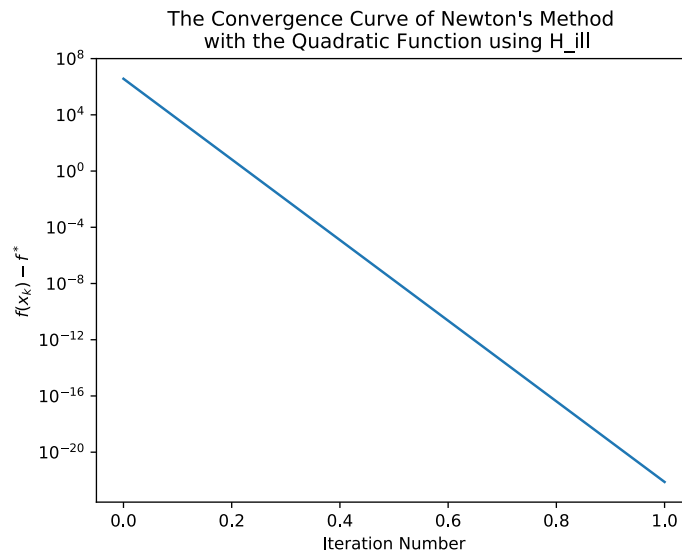
Using the Newton's method with the starting point  $x_{0\_quad}$ , we get:

- With  $H = H\_well$ :



TODO: Add a short explanation about the graph

- With  $H = H\_ill$ :



TODO: Add a short explanation about the graph