236330 - Introduction to Optimization: Homework #2

May 13, 2020

Amit Rotner
123456789
Or Steiner
123456789

Gradient Descent method and Newton's 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_1 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 $\forall x_1, x_2 \in C$, $\forall \beta \in [0,1]$:

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

Let $x_1, x_2 \in L$:

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

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

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

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

Q3:

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

Show that $g\left(x\right) = f\left(Ax\right)$ 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 $\forall x_1, x_2 \in C$, $\forall \beta \in [0,1]$:

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

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

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

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

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

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

Hence, H is PSD.

Q4:

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

Solution:

Jensen's inequality:

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

If $\alpha_1, \alpha_2, \ldots, \alpha_n$ 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 $n = 2$: $a_2 = 1 - \alpha_1$.
  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 $\alpha_1, \alpha_2, \ldots, \alpha_{n+1}$ 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:

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

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:

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

Since $log\left(\cdot\right)$ 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 $f\left(x\right)=\frac{1}{2}x^TQx-b^Tx+c$ 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 β that we found in the tutorial,
   using $\theta\left(Q\right)$ - the condition number of Q.
2. Assume that the step size can be modified at any iteration.
   Find the optimal step size α_k^* .

Solution:

We start by finding x^*:

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

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

From the definition of the gradient decent method,

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

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:

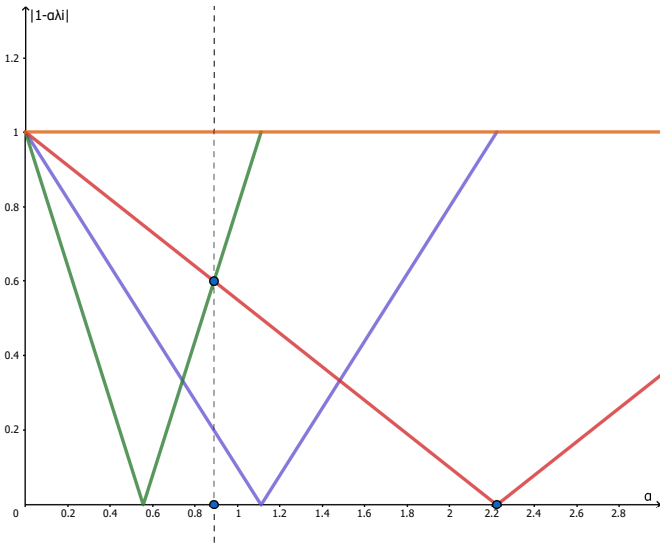
$$\|(I - \alpha Q)\, (x_k - x^*)\|_2 \leq \|(I - \alpha Q)\|_2 \cdot \|(x_k - x^*)\|_2$$

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

Hence, we need to find:

$$\alpha_{opt} = argmin_{\alpha} max_i\, \{|1 - \alpha \lambda_i|\}$$

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,

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

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:

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

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

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

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

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

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

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

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

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 $\forall x \in Dom(f):\ mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$ 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:

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

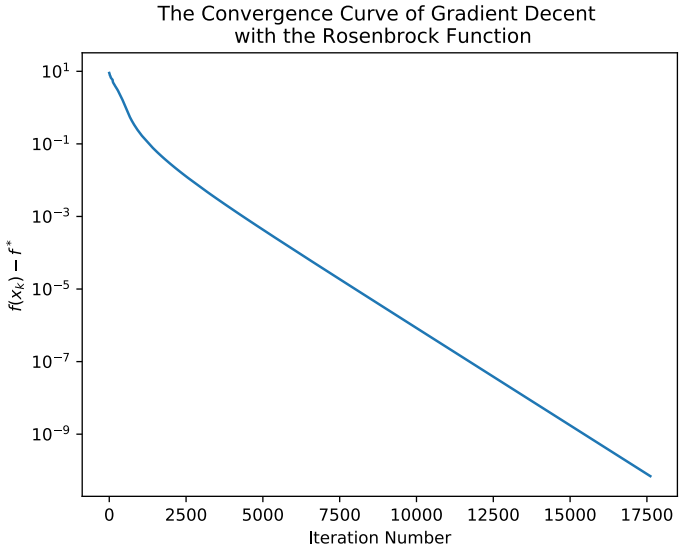
Deriving analytically the gradient, yields:

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

Deriving analytically the Hessian, yields:

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

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

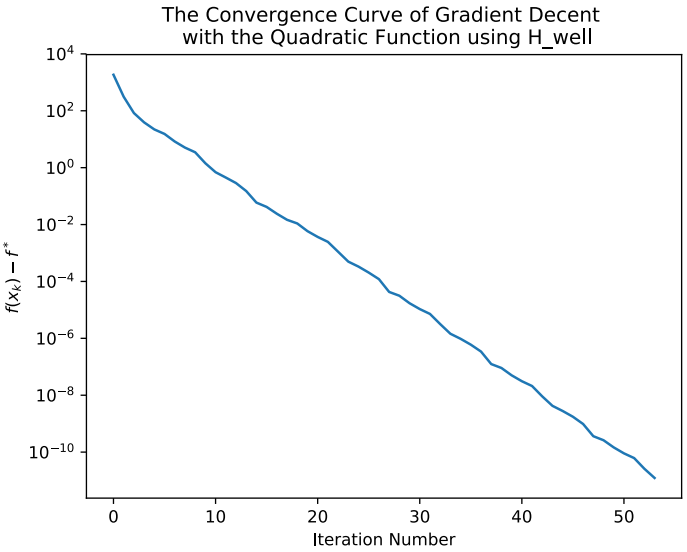


The convergence rate is linear with $\frac{f(x_{k+1})-f^*}{f(x_k)-f^*} \leq 0.9991365848557804 < 1 \ \forall k$ in range.

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

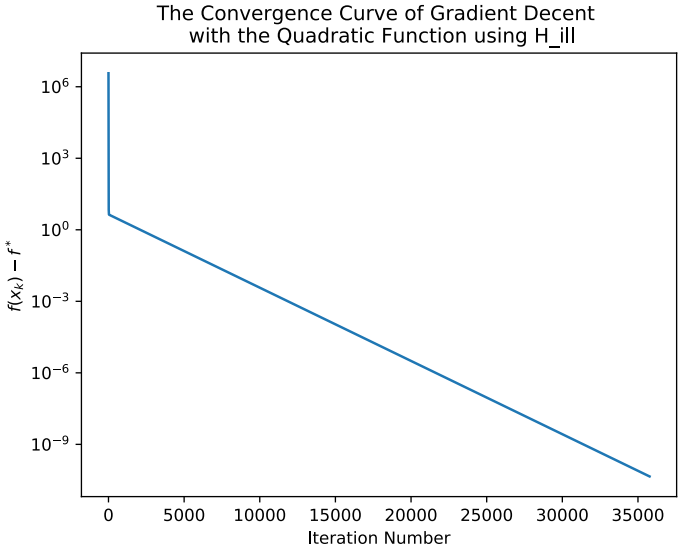
   Using the Gradient Descent method with the starting point x_{0_quad}, we get:

   - With $H = H\_well$:



The convergence rate is linear with $\frac{f(x_{k+1})-f^*}{f(x_k)-f^*} \leq 0.7441284726564737 < 1 \ \forall k$ in range.

- With $H = H\_ill$:



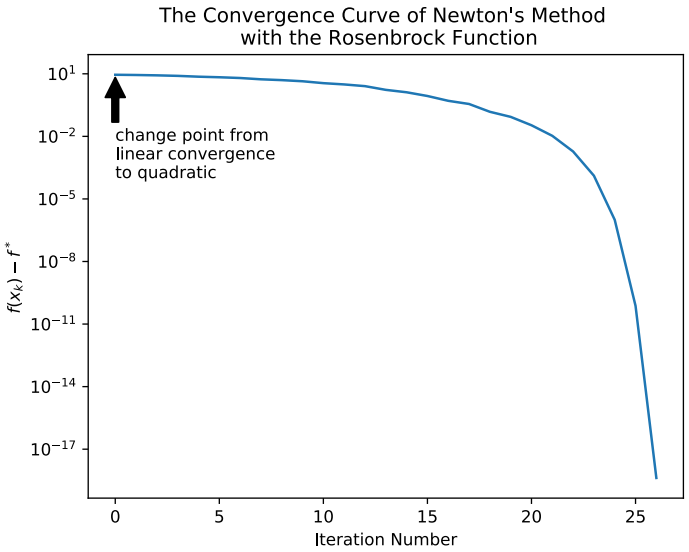
The convergence rate is linear with $\frac{f(x_{k+1})-f^*}{f(x_k)-f^*} \leq 0.9996612219610987 < 1 \ \forall k$ in range.

Q9:

Solution:

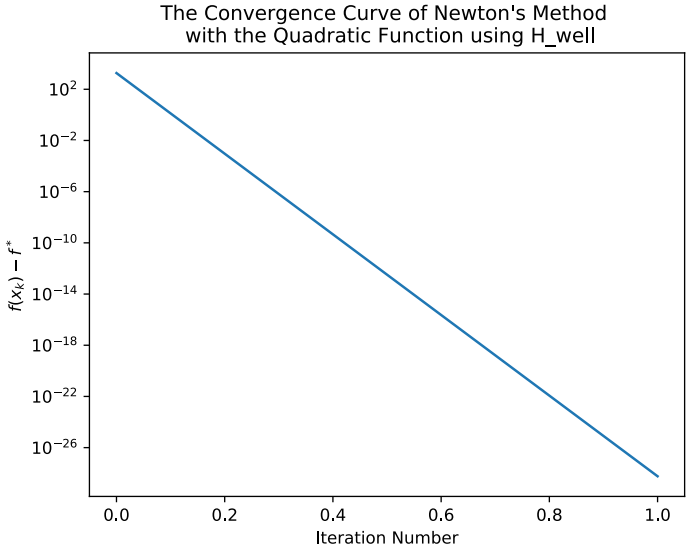
1. Given the Rosenbrock function,

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



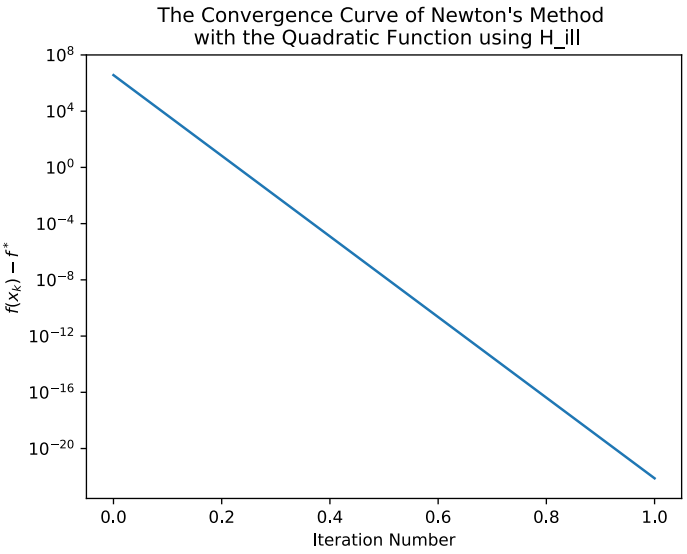
The convergence rate is quadratic with $\frac{f(x_{k+1})-f^*}{(f(x_k)-f^*)^2} \leq C = 75.63226955247731 \; \forall k$ in range.
Therefore, the convergence rate is quadratic from the first iteration.
In addition, while using Newton's method we managed to speed up convergence from 17500 iterations with Gradient Descent to 26 iterations. But, we had to pay $O\left(\frac{n^3}{6}\right)$ for the Cholesky factorization and $O\left(n^2\right)$ for storing the Hessian.

2. Given the quadratic function $f(x) = \frac{1}{2}x^T Hx$,
   Using the Newton's method with the starting point x_{0_quad}, we get:
   - With $H = H\_well$:



The function to minimize using newton method is quadratic, therefore a second order Taylor expansion of the function, is the function itself. Hence, after one iteration of the algorithm we have found the value that minimizes that quadratic function. While, when we used the Gradient Descent method, it took over 50 iterations.

- With $H = H\_ill$:



The function to minimize using newton method is quadratic, therefore a second order Taylor expansion of the function, is the function itself. Hence, after one iteration of the algorithm we have found the value that minimizes that quadratic function. While, when we used the Gradient Descent method, it took over 35000 iterations.