National Technical University of Athens

Interdisciplinary Master's Programme in

Data Science and Machine Learning

Convex Optimization

Exercise Sheet 1

Written by: Christos Nikou
ID: 03400146

email:christosnikou@mail.ntua.gr

May 8, 2022

Exercise 1

Let $f : \mathbb{R}^n \to \mathbb{R},\ f(x) = \frac{1}{2}x^T Ax - b^T x$, where $A \in \mathbb{R}^{1000\times 1000}$ given by

$$A = \begin{pmatrix} 11 & -1 & 0 & \dots & 0 \\ -1 & 11 & -1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 11 \end{pmatrix}, \text{and } b = \begin{pmatrix} 10 \\ 8 \\ \vdots \\ 8 \\ 10 \end{pmatrix}.$$

(a) Prove that f is strictly convex.

(b) Write a program about the methods of *Steepest Descent* and *Conjugate Gradients* for the quadratic case to determine the minimum value of f. Suppose your initial starting point is $x_0^T = (-1/2,\ 0,\ \dots,\ 0,\ 1/3)^T$. Compare the result of the two methods.

(c) The *Hilbert* matrix $H \in \mathbb{R}^{n\times n}$ is given by $h_{ij} = \frac{1}{i+j-1},\ 1 \le i,j \le n$. Apply the *Steepest Descent* and *Conjugate Gradients* algorithms in order to minimize the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}x^T Hx - b^T x$, where $b = \begin{pmatrix} 10 \\ 2 \\ \vdots \\ 2 \\ 10 \end{pmatrix}$, $x_0 = \begin{pmatrix} -\frac{1}{2} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ and $n = 4^k$ for $k = 1, 2, 3$. Does the algorithms converge?

Note: For your programming results include only the first and last five coordinates of the vectors produced in the first and last iteration of the algorithms.

(a) It is easily seen that A is *positive definite*. Indeed, let $\bar{x}^T = (x_1, x_2, \dots, x_{n-1}, x_n)^T \in \mathbb{R}^n$, then by simple calculations we obtain

$$\bar{x}^T A\bar{x} = (x_1, x_2, \dots, x_{n-1}, x_n) \cdot \begin{pmatrix} 11 & -1 & 0 & \dots & 0 \\ -1 & 11 & -1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 11 \end{pmatrix} \cdot (x_1, x_2, \dots, x_{n-1}, x_n)^T$$

$$\begin{aligned}
&= (11x_1 - x_2, -x_1 + 11x_2 - x_3,\ \dots, -x_{n-2} + 11x_{n-1} - x_n, -x_{n-1} + 11x_n) \cdot (x_1, x_2, \dots, x_{n-1}, x_n)^T \\
&= 11x_1^2 - 2x_1x_2 + 11x_2^2 - 2x_2x_3 + 11x_3^2 + \dots \\
&+ 11x_{n-2}^2 - 2x_{n-2}x_{n-1} + 11x_{n-1}^2 - 2x_nx_{n-1} + 11x_n^2 \\
&= 10x_1^2 + (x_1 - x_2)^2 + 9x_2^2 + (x_2 - x_3)^2 + 9x_3^2 + \dots \\
&+ 9x_{n-2}^2 + (x_{n-2} - x_{n-1})^2 + 9x_{n-1}^2 + (x_{n-1} - x_n)^2 + 10x_n^2.
\end{aligned}$$

Therefore, we obtain the following equation

$$\bar{x}^T A \bar{x} = 10x_1^2 + 10x_n^2 + 9\sum_{k\neq 1,n} x_k^2 + \sum_{k=1}^{n-1}(x_k - x_{k+1})^2,$$

which is certainly positive for all $\bar{x} \neq 0$. In addition, A is also symmetric and it is the *Hessian* matrix of f. From the following general fact it follows that f is strictly convex.

Positive definite Hessian matrix implies strictly convexity: Indeed, suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function with positive definite *Hessian* matrix $\nabla^2 f(x)$ for all $x \in \mathbb{R}^n$. Then f is strictly convex.

Proof: Let $x, y \in \mathbb{R}^2$ such that $x \neq y$. Taylor's theorem implies that

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x + t(y-x))(y-x), \tag{1.1}$$

for some $0 < t < 1$. Since $\nabla^2 f(x)$ is everywhere positive definite, the quadratic term in the equation (1.1) is always strictly positive . Thus we obtain that

$$f(y) > f(x) + \nabla f(x)^T(y-x). \tag{1.2}$$

Therefore, (1.2) holds for every $x, y \in \mathbb{R}^n$ and the inequality holds if and only if $x = y$. Now we can easily prove that f is convex. Indeed, let $z \neq w$ two points in $\mathbb{R}^n$ and $0 < \lambda < 1$. Let $x = \lambda w + (1-\lambda)z$. Then the inequality (1.2) implies that

$$\begin{aligned} f(w) &> f(x) + \nabla f(x)^T(w-x), \\ f(z) &> f(x) + \nabla f(x)^T(z-x). \end{aligned} \tag{1.3}$$

Now, observe that $w - x = (1-\lambda)(w-z)$ and $z - x = \lambda(z-w)$. Thus, if we multiply the first line in (1.3) with λ, the second line with $1-\lambda$, and then add the two inequalities, we obtain

$$\begin{aligned} \lambda f(w) + (1-\lambda)f(z) &> f(x) + \lambda\nabla f(x)^T(1-\lambda)(w-z) + (1-\lambda)\nabla f(x)^T\lambda(z-w) \\ &= f(\lambda w + (1-\lambda)z), \end{aligned}$$

which proves that f is strictly convex.

(b) Starting the algorithms with initial point $x_0^T = (-1/2, 0, \ldots, 0, 1/3)^T$ we get that *Steepest Descent* converges after 13 iterations satisfying the condition $\|r_k\|_2 < 10^{-9}$ with estimated minimum value -3558.4341643046982. The first and last five coordinates of the point in the first iteration are summarized in the following table.

Table 1: Coordinates in the first iteration of *Steepest Descent*

Coordinates	**1st**	**2nd**	**3rd**	**4th**	**5th**
First 5	1.22118262	0.8328303	0.88835232	0.88835232	0.88835232
Last 5	0.88835232	0.88835232	0.88835232	0.925367	1.03661225

Below we see the first and last five coordinates in the last iteration.

Table 2: Coordinates in the last (13th) iteration of *Steepest Descent*

Coordinates	**1st**	**2nd**	**3rd**	**4th**	**5th**
First 5	0.99074787	0.89822662	0.88974491	0.88896736	0.88889608
Last 5	0.88889608	0.88896736	0.88974491	0.89822662	0.99074787

On the other hand, *Conjugate Gradients* converges after 11 iterations satisfying the condition $||r_k||^2 < 10^{-9}$ with estimated minimum value -3558.434164304692. The first and last five coordinates of the point in the first iteration are summarized below.

Table 3: Coordinates in the first iteration of *Conjugate Gradients*

Coordinates	**1st**	**2nd**	**3rd**	**4th**	**5th**
First 5	1.22118262	0.8328303	0.88835232	0.88835232	0.88835232
Last 5	0.88835232	0.88835232	0.88835232	0.925367	1.03661225

and the corresponding coordinates for the last iteration are given below.

Table 4: Coordinates in the last (11th) iteration of *Conjugate Gradients*

Coordinates	**1st**	**2nd**	**3rd**	**4th**	**5th**
First 5	0.99074787	0.89822662	0.88974491	0.88896736	0.88889608
Last 5	0.88889608	0.88896736	0.88974491	0.89822662	0.99074787

From the previous results and the following figure we see that both methods exhibit the same convergence behavior.

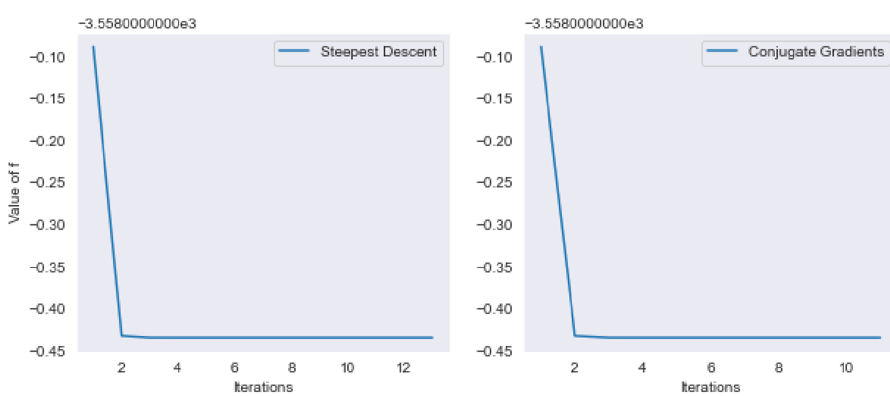

Figure 1: The Convergence of the two methods.

(c) ***case***: $k = 1$ *Steepest Descent* converges after 97825 iterations with estimated minimum value for the function $f(x) = x^T Hx - b^T x$ equal to -83359.99999998607. *Conjugate Gradients* method converges after 4 iterations with -83359.99999998683 estimated minimum value. Below we see the points produced by the two methods in the first iteration.

Table 5: The points in the first iteration

Method	1st	2nd	3rd	4th
Steepest Descent	9.30350104	2.10075022	2.02294466	9.453376
Conjugate Gradients	9.30350104	2.10075022	2.02294466	9.453376

Below we see the points produced in the last iteration of the two methods.

Table 6: The points in the last iteration

Method	1st	2nd	3rd	4th
Steepest Descent	-999.99974513	12599.9971335	-32039.9930981	21559.9955127
Conjugate Gradients	-1000.00000001	12599.9999999	-32040	21559.9999999

Below we see the corresponding graphs for the two methods.

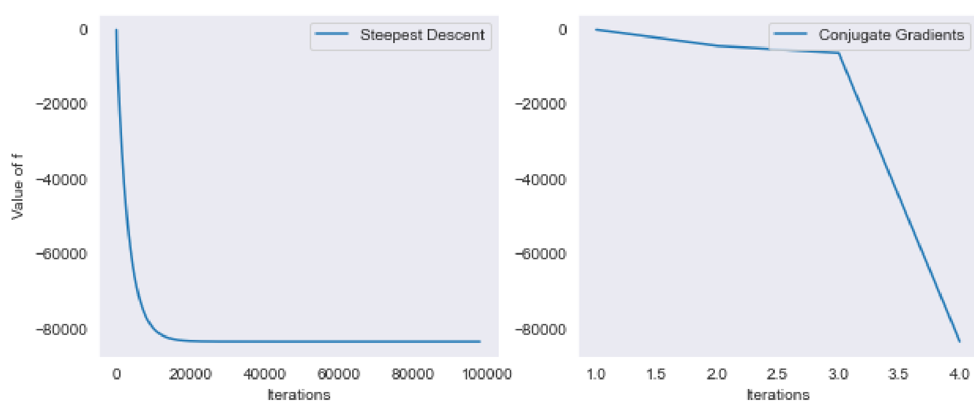

Figure 2: The Convergence of the two methods.

case: $k = 2$ In this case both methods do not convergence. Below we can see the behavior of the two methods.

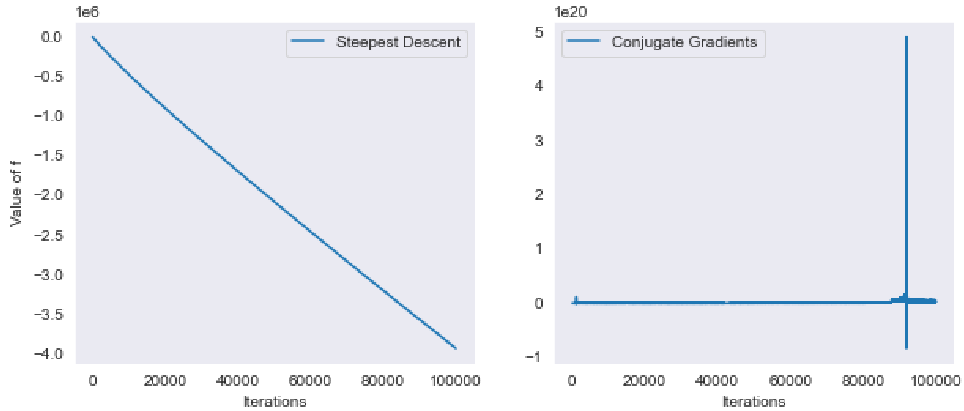

Figure 3: Behavior of the two methods after 100000 iterations.

As we can see, in the first case the values of f slowly decrease but without succeeding to reach the global minimum. In the second case, we see that the values of f bounce rapidly from negative to positive values without any sign of convergence.

case: $k = 3$ Also in this case the methods do not converge to a solution. In Fig 4 see the behavior of the *Conjugate Gradients* algorithm.

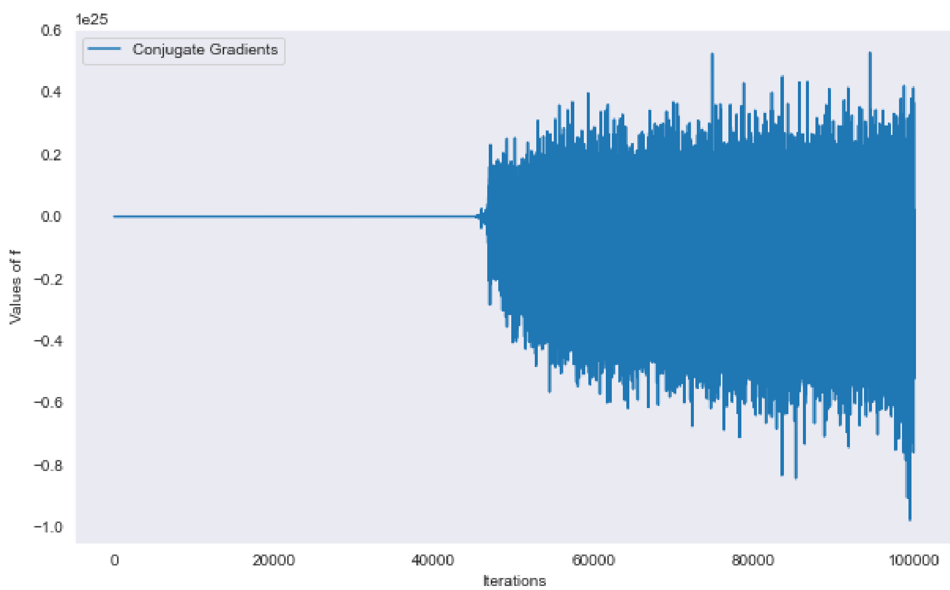

Figure 4: The behavior of Conjugate Gradients method for the case k=3.

Exercise 2

(a) Let $f(x,y) = 100(y-x^2)^2 + (1-x^2)^2 + 4$. Apply the *Newton - Raphson* method with starting point $(-1.2, 1)$ to solve the minimization problem $f(\bar{x}, \bar{y}) = \min_{(x,y)^T \in \mathbb{R}^2} f(x,y)$.

(b) Modify *Newton - Raphson*'s method by using a positive definite approximation of the *Hessian Matrix* of f (*Quasi Newton method*).

(a) Before we apply *Newton - Raphson* method we can analytically determine the minimum values of f and compare with the result with *Newton - Raphson*'s method. To determine the stationary points we first calculate the partial derivatives of f. By writing,

$$\begin{aligned}\frac{\partial f}{\partial x} &= 200(y-x^2)(-2x) + 2(1-x^2)(-2x)\\ &= -400x(y-x^2) - 4x(1-x^2)\\ &= -400xy + 400x^3 - 4x + 4x^3\\ &= 404x^3 - 400xy - 4x,\end{aligned}$$

and $\partial f/\partial y = 200(y-x^2)$ we obtain $\nabla f(x,y) = (404x^3 - 400xy - 4x, 200(y-x^2))^T$. Now solving for $\nabla f(x,y) = \vec{0}$ we have that

$$\left.\begin{aligned} &404x^3 - 400xy - 4x = 0\\ &y = x^2\end{aligned}\right\} \Longleftrightarrow \left.\begin{aligned} &404x^3 - 400x^3 - 4x = 0\\ &y = x^2\end{aligned}\right\},$$

which is equivalent to $x(x^2-1) = 0$ and $y = x^2$. Solving for x we obtain the following three stationary points $\boxed{\vec{x_1} = (0,0)}$, $\boxed{\vec{x_2} = (1,1)}$ and $\boxed{\vec{x_3} = (-1,1)}$. To determine whether these points are minimums/maximums we proceed by calculating the *Hessian* matrix $\nabla f^2(x,y)$ of f. By elementary calculations it is readily verified that

$$\nabla^2 f(x,y) = \begin{pmatrix} 1212x^2 - 400y - 4 & -400x\\ -400x & 200\end{pmatrix}. \tag{2.1}$$

For $(x,y) = (0,0)$ in (2.1) we obtain

$$\nabla^2 f(0,0) = \begin{pmatrix} -4 & 0\\ 0 & 200\end{pmatrix}.$$

The above matrix has negative determinant and since $\partial^2 f/\partial^2 y > 0$ it follows that $(0,0)$ is a local minimum of f. For $(x,y) = (1,1)$ we obtain $\begin{pmatrix} 808 & -400\\ -400 & 200\end{pmatrix}$ which again has a non-negative determinant and $\partial^2 f/\partial^2 y > 0$, hence $(1,1)$ is also a local minimum. In similar fashion, for $(x,y) = (-1,1)$ we have the matrix $\begin{pmatrix} 808 & 400\\ 400 & 200\end{pmatrix}$ which also has a non-zero determinant and $\partial^2 f/\partial^2 y > 0$. Therefore, we conclude that all stationary points of f are local minimums. In particular, the point $(-1,1)$ is a local minimum of f. Now, the *Newton - Raphson* method can be summarized as follows.

Algorithm 1 Newton's Method

Given a starting point x_0 and an objective function f

Set $x \leftarrow x_0$

for $k = 1, \ldots,$ until convergence or some stopping criterion met **do**

  Compute the Hessian matrix H of f at x

  Set $x \leftarrow x - H^{-1} \cdot \nabla f(x)$

end for

Implementing Algorithm 1 with starting point $[-1.2, 1]^T$ we deduce that the method converges in 6 iterations to the local minimum $[-1, 1]^T$ of f with value 4. In Fig. 5 we see the convergence of this method.

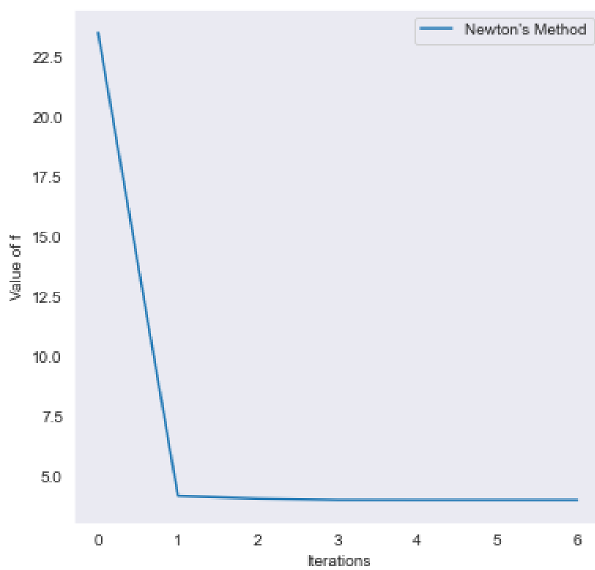

Figure 5: Newton - Raphson algorithm with starting point $(-1.2, 1)^T$.

(b) For the *Quasi Newton method* the *BFGS method* is used (source code taken from this *GitHub* repo). *BFGS* method is one of the most popular *quasi-Newton* algorithms named after its discoverers Broyden, Fletcher, Goldfarb, and Shanno. The method requires a *backtracking line search* with *Wolfe* conditions which we describe in the following code block.

Algorithm 2 Backtracking Line Search - Wolfe Conditions

Given x_k and a direction p_k

Choose $\bar{a} > 0,\ c_1 \in (0, 1),\ c_2 \in (c_1, 1),\ \rho \in (0, 1)$; Set $a \leftarrow \bar{a}$;

while $f(x_k + ap_k) \geq f(x_k) + c_1 a \nabla f(x_k)^T \cdot p_k$ or $\nabla f(x_k + ap_k) \leq c_2 \nabla f(x_k) \cdot p_k$ **do**

  $a \leftarrow \rho a$;

end while

return a

And the *BFGS* method proceeds as follows.

Algorithm 3 BFGS Method

Given starting point x_0, convergence tolerance $\epsilon > 0$,
  inverse Hessian approximation H_0; (the identity)
$k \leftarrow 0$
while $\|\nabla f(x_k)\| > \epsilon$ **do**
  Compute seach direction
    $p_k = -H_k \cdot \nabla f(x_k)$
  Set $x_{k+1} = x_k + a_k p_k$ where a_k is computed by Alg. 2
  Define $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$
  $\rho_k \leftarrow 1/y_k^T s_k$
  Compute H_{k+1} by
    $H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$
  $k \leftarrow k + 1$
end while

Implementing Algorithm 3 in Python we obtain that the method converges in 19 iterations to the local minimum $[-1, 1]^T$ as in the case of *Newton*'s method. In Fig. 6 we can see the path of the Algorithm along with the values of f.

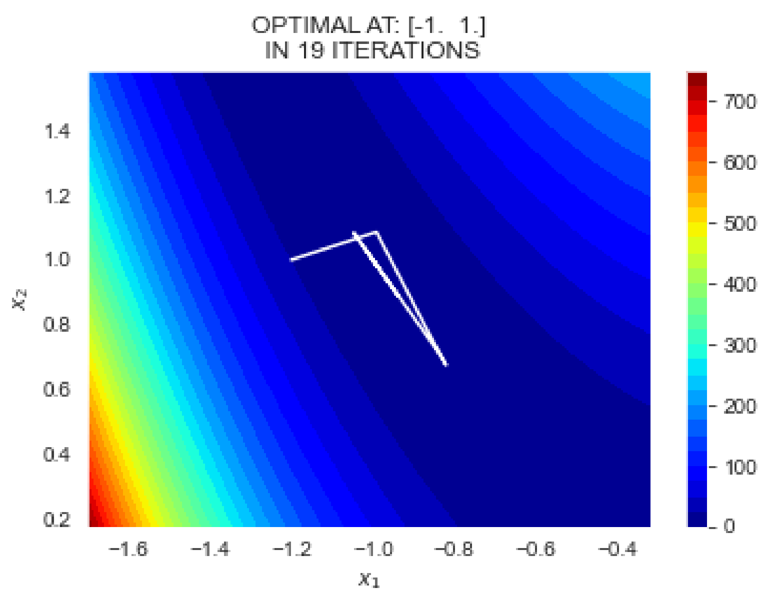

Figure 6: The convergence of *BFGS* method.

Exercise 3

Write a program about *Frank - Wolfe*'s method for the minimization problem $\min_{(x,y)^T \in S} f(x, y) = f(\bar{x}, \bar{y})$, where $S = \{(x, y)^T \in \mathbb{R}^2 : \; x^2 + y^2 \leq 2\}$. Use $(-1/2, 1/2)$ as a starting point and give an estimation (with more than 5 iterations) of the minimum value.

Frank - Wolfe algorithm can be summarized as follows:

Algorithm 4 Frank - Wolfe

Given $x_1 \in S$. Set $y_0 := x_1$
for $k = 1, \ldots,$ until convergence or some stopping criterion met **do**
  Compute $y_k \in S$ such that
    $\min_{y \in S} \nabla f(x_k)^T \cdot (y - x_k) = \nabla f(x_k)^T \cdot (y_k - x_k) := \delta_k$
  if $\delta_k = 0$ **then**
    Stop
  else
    Compute α_k such that
      $\min_{\alpha \in [0,1]} f(x_k + \alpha(y_k - x_k)) = f(x_k + \alpha_k(y_k - x_k))$
  end if
  $x_{k+1} \leftarrow x_k + \alpha_k(y_k - x_k)$
end for

In general, it is not easy to implement an algorithm to analytically solve the problem

$$y_k = \arg\min_{y \in S} \nabla f(x_k)^T \cdot (y - x_k). \tag{3.1}$$

But in some cases, depending on the set S, it is possible to provide an analytical solution. For example, in the case of a circle $S = \{(x, y)^T \in \mathbb{R}^2 : \; x^2 + y^2 \leq R^2\}$ centered at 0 with radius $R > 0$; y_k is given by

$$y_k = -R \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}. \tag{3.2}$$

To this end, suppose that $\nabla f(x_k) = (\alpha, \beta)^T \neq (0, 0)$[1] and observe that the function $y \mapsto \nabla f(x_k)^T \cdot (y - x_k)$, for $y \in S$ is convex. Therefore, any solution of *KTL* equations is also a solution to (3.1). Setting $y = (y_1, y_2)$, $x_k = (u, v)$, $h(y_1, y_2) = \nabla f((u, v))^T \cdot (y_1 - u, y_2 - v)$, $h_1(y_1, y_2) = y_1^2 + y_2^2 - R^2$ and applying *KTL* theorem we obtain the following set of equations.

$$\begin{cases} \nabla h(y_1, y_2) + \lambda_1 \nabla h_1(y_1, y_2) = 0 \\ \lambda_1 h_1(y_1, y_2) = 0 \\ \lambda_1 \geq 0 \end{cases}.$$

[1] If $\nabla f(x_k) = 0$ the algorithm terminates.

Equivalently,

$$\begin{cases} (\alpha,\beta)+\hat{\lambda}_1(2y_1,2y_2)=0 \\ \hat{\lambda}_1(y_1^2+y_2^2)=\hat{\lambda}_1 R^2 \\ \hat{\lambda}_1 \geq 0 \end{cases} . \tag{3.3}$$

Now, if $\hat{\lambda}_1 = 0$ then $(\alpha,\beta) = (0,0)$ which contradicts the fact that $(\alpha,\beta) \neq (0,0)$. Therefore, (3.3) is equivalent to

$$\begin{cases} y_1 = -\alpha/2\hat{\lambda}_1 \\ y_2 = -\beta/2\hat{\lambda}_1 \\ y_1^2+y_2^2 = R^2 \\ \hat{\lambda}_1 > 0 \end{cases} . \tag{3.4}$$

Substituting the expressions for y_1, y_2 we get

$$\begin{aligned} y_1^2+y_2^2 = R^2 &\iff \frac{\alpha^2}{4\hat{\lambda}_1^2}+\frac{\beta^2}{4\hat{\lambda}_1^2} = R^2 \\ &\iff \alpha^2+\beta^2 = 4\hat{\lambda}_1^2 R^2 \\ &\overset{\hat{\lambda}_1>0}{\iff} \hat{\lambda}_1 = \frac{\sqrt{\alpha^2+\beta^2}}{2R}. \end{aligned}$$

Hence, $\boxed{y_1 = -R\frac{\alpha}{\sqrt{\alpha^2+\beta^2}}}$ and $\boxed{y_2 = -R\frac{\beta}{\sqrt{\alpha^2+\beta^2}}}$. Equivalently we can write $(y_1, y_2) = -R\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$ which coincides with (3.2). In our case the set $S = \{(x,y)^T \in \mathbb{R}^2 : x^2+y^2 \leq R^2\}$ is a circle of radius $\sqrt{2}$ centered at 0 which means that the solution to (3.1) is unique and it is given by $y_k = -\sqrt{2}\nabla f(x_k)/\|\nabla f(x_k)\|$. Implementing Algorithm 4 in Python we have the following results for the first 10 iterations.

Table 7: Frank - Wolfe Algorithm

Iteration	**Point x**	**Value of f**	δ_k
0	$[-0.5, 0.5]^T$	10.812	1
1	$[-0.563, 0.315]^T$	4.466	-100.76
2	$[-0.570, 0.317]^T$	4.462	-0.910
3	$[-0.568, 0.319]^T$	4.459	-1.995
4	$[-0.659, 0.411]^T$	4.375	-0.796
5	$[-0.648, 0.415]^T$	4.339	-10.652
6	$[-0.647, 0.418]^T$	4.337	-0.779
7	$[-0.650, 0.417]^T$	4.336	-0.857
8	$[-0.649, 0.419]^T$	4.335	-0.893
9	$[-0.653, 0.420]^T$	4.333	-0.707
10	$[-0.651, 0.421]^T$	4.332	-1.37

As we observe from Table 7 the point slowly approaches the point $[-1, 1]^T$ which is a local minimum of f as proved in Exercise 2 (a). Our stopping criterion for the *Frank - Wolfe* implementation is with regard to the absolute value of δ_k. In Fig. 7 we see the variations on the values of f with respect to the sequence of points generated by the algorithm.

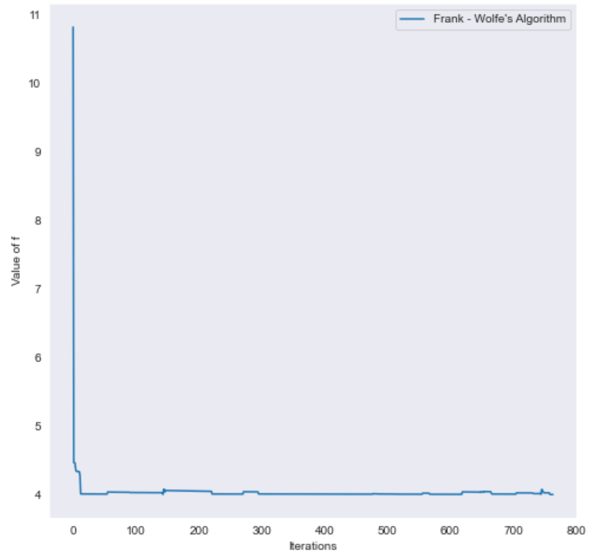

Figure 7: Frank - Wolfe algorithm with starting point $(1/2, 1/2)^T$.

The above implementation terminated in 763 iterations with value $|\delta_k| < 10^{-4}$. In Fig. 8 we see the sequence of points generated by the algorithm. Observe that every point of the sequence lies in S as the algorithm suggests.

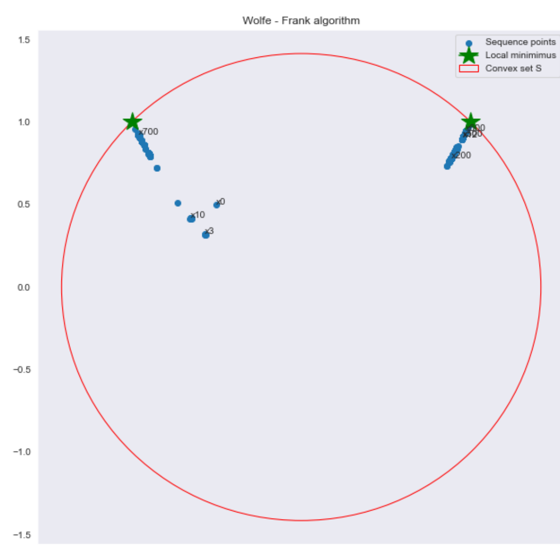

Figure 8: The sequence of points generated by Frank - Wolfe algorithm.

Exercise 4

Examine the existence and uniqueness of the minimization problem $\min\limits_{(x,y,z)^T \in S} f(x,y,z) = f(\bar{x},\bar{y},\bar{z})$ in each of the following cases.

1. $f(x,y,z) = e^{x^2+y^2+z^2} - (x^2+y^2+z^2)^4,\ \ S = \{(x,y,z)^T \in \mathbb{R}^3 : x^2 + 4y^2 \leq 1\}$.
2. $f(x,y,z) = e^{-(x^2+y^2+z^2)} - (x^2+y^2+z^2)^2,\ \ S = \{(x,y,z)^T \in \mathbb{R}^3 : (x-3)^2 + (y-1)^2 \leq 1,\ |z| \leq 1\}$.
3. $f(x,y,z) = e^{x^2+y^2+z^2} - (x^2+y^2+z^2)^3 - (x+y+z),\ \ S = \{(x,y,z)^T \in \mathbb{R}^3 : (x-1)^2 + y^2 = 1\}$.

For the first two problems state the necessary condition. Is the necessary condition sufficient?

1. In this case, since the set S has no restrictions with respect to the variable z it follows that S is unbounded. To examine if a solution of $\min\limits_{(x,y,z)^T \in S} f(x,y,z) = f(\bar{x},\bar{y},\bar{z})$ exists we proceed as follows: First we observe that $\lim\limits_{x\to\infty} e^x - x^4 = +\infty$. Therefore, it follows that

$$\lim_{\substack{\|(x,y,z)\|_2 \to \infty \\ (x,y,z)^T \in S}} f(x,y,z) = +\infty. \tag{4.1}$$

The above implies that there exists a radius $R > 0$ such that $f(x,y,z) > 1$ for every $(x,y,z)^T \in S \cap (\mathbb{R}^3 \setminus \hat{B}(0,R))$, where $\hat{B}(0,R)$ denotes the closed ball of radius R centered at 0, i.e

$$\hat{B}(0,R) = \{(x,y,z) \in \mathbb{R}^3 : \|(x,y,z)\|_2 \leq R\}.$$

Now, since $\hat{B}(0,R)$ is bounded, it follows that $S \cap \hat{B}(0,R)$ is closed and bounded. In particular, $S \cap \hat{B}(0,R)$ is compact. By the continuity of f it follows that there exists $x^*, x^{**} \in S \cap \hat{B}(0,R)$ such that $f(x^*) \leq f(\vec{x}) \leq f(x^{**})$ for every $\vec{x} \in S \cap \hat{B}(0,R)$. Now since $(0,0,0) \in S \cap \hat{B}(0,R)$ and $f(0,0,0) = 1 > 0$ it follows that $f(x^{**}) \geq 1 > 0$. Using again (4.1) we may find an $R' > R$ such that $f(x,y,z) > f(x^{**})$ for every $(x,y,z)^T \in S \cap (\mathbb{R}^3 \setminus \hat{B}(0,R'))$. Again by the compactness of $S \cap \hat{B}(0,R')$ it follows that f has a minimum value on $S \cap \hat{B}(0,R')$, say $(\bar{x},\bar{y},\bar{z})$. We claim that $(\bar{x},\bar{y},\bar{z})$ is a solution of $\min\limits_{(x,y,z)^T \in S} f(x,y,z)$. Indeed, since $(\bar{x},\bar{y},\bar{z})$ is the minimum value of f on $S \cap \hat{B}(0,R')$ it follows that

$$f(\bar{x},\bar{y},\bar{z}) \leq f(x,y,z) \text{ for every } (x,y,z) \in S \cap \hat{B}(0,R'). \tag{4.2}$$

In particular, since $x^{**} \in S \cap \hat{B}(0,R) \subseteq S \cap \hat{B}(0,R')$ it follows that

$$f(\bar{x},\bar{y},\bar{z}) \leq f(x^{**}). \tag{4.3}$$

Since $f(x,y,z) > f(x^{**})$ for every $(x,y,z) \notin S \cap \hat{B}(0,R')$ it follows that

$$f(\bar{x},\bar{y},\bar{z}) < f(x,y,z) \text{ for every } (x,y,z) \notin S \cap \hat{B}(0,R') \tag{4.4}$$

Combining (4.2) and (4.3) we obtain that $f(\bar{x},\bar{y},\bar{z}) \leq f(x,y,z)$ for every $(x,y,z) \in S$. Therefore, $(\bar{x},\bar{y},\bar{z})$ is a solution to the problem $\min\limits_{(x,y,z)^T \in S} f(x,y,z) = f(\bar{x},\bar{y},\bar{z})$. Now, to examine whether the constrained minimization problem has a unique solution or not we reason as follows: First we observe that the point $(0,0,\sqrt{\ln 10})$ lies in S and that $f(0,0,\sqrt{\ln 10}) = 10 - (\sqrt{\ln 10})^8 = 10 - (ln10)^4 \approx 10 - (2.3)^4 < 0$. Therefore, a solution $(x,y,z) \in S$ to the constrained minimization problem satisfies $(x,y,z) \neq (0,0,0)$. Now since f is symmetric, i.e $f(x,y,z) = f(-x,-y,-z)$ it follows that f has at least points $\vec{x_1}, \vec{x_2} \in S$ such that $\vec{x_1} \neq \vec{x_2}$ and $f(\vec{x_1}) = f(\vec{x_2}) = \min_{(x,y,z)^T \in S} f(x,y,z)$. Therefore, the constrained minimization problem has more than one solutions. Now, if $f_1(x,y,z) = x^2 + 4y^2 - 1$ then by *KTL* the necessary condition that $(x,y,z)^T \in S$ must satisfy in order to be a solution is described by the following equations

$$\begin{cases} \nabla f(x,y,z) + \lambda_1 \nabla f_1(x,y,z) = 0 \\ \lambda_1 f_1(x,y,z) = 0 \\ \lambda_1 \geq 0 \end{cases} . \tag{4.5}$$

Calculating the gradients, (4.5) becomes

$$\begin{cases} \left[2e^{x^2+y^2+z^2} - 8(x^2+y^2+z^2)^3\right](x,y,z) + \lambda_1(2x, 8y, 0) = (0,0,0) \\ \lambda_1 = \lambda_1 x^2 + 4\lambda_1 y^2 \\ \lambda_1 \geq 0 \end{cases} .$$

2. In this case, as opposed to the previous one, the set S is closed and bounded and hence compact. Therefore, the continuity of f on S implies that f attains a minimum value on S. In other words, the minimization problem $\min_{(x,y,z)^T \in S} f(x,y,z)$ has at least one solution in this case. To prove that the solution is not unique we only have to prove that the point $(x_0, y_0, z_0) \in S$ with $f(x_0,y_0,z_0) = \min_{(x,y,z)^T \in S} f(x,y,z)$ has $z_0 \neq 0$. In such case, the point $(x_0, y_0, -z_0) \neq (x_0, y_0, z_0)$ and

$$f(x_0, y_0, -z_0) = f(x_0, y_0, z_0) = \min_{(x,y,z)^T \in S} f(x,y,z).$$

Therefore, if $z_0 \neq 0$ we have more than one solution. To prove $z_0 \neq 0$ let $h(z) = f(x_0, y_0, z)$ for $|z| \leq 1$. Then,

$$\begin{aligned} h'(z) &= \frac{\partial}{\partial z}\left(e^{-(x_0^2+y_0^2+z^2)} - (x_0^2+y_0^2+z^2)^2\right) \\ &= -2ze^{-(x_0^2+y_0^2+z^2)} - 4z(x_0^2+y_0^2+z^2). \end{aligned}$$

Now,

$$\begin{aligned} h'(z) = 0 &\Longleftrightarrow -2ze^{-(x_0^2+y_0^2+z^2)} - 4z(x_0^2+y_0^2+z^2) = 0 \\ &\Longleftrightarrow z\left(e^{-(x_0^2+y_0^2+z^2)} + 2(x_0^2+y_0^2+z^2)\right) = 0 \\ &\Longleftrightarrow z = 0, \end{aligned}$$

since $e^{-(x_0^2+y_0^2+z^2)} + 2(x_0^2+y_0^2+z^2) > 0$ for every $|z| \leq 1$. Now, $h'(1) < 0$ and since $h'(z) = 0$ has unique solution $z = 0$ we deduce that $h'(z) < 0$ for every $0 < z \leq 1$. Therefore, h is strictly decreasing in $(0, 1]$. Hence,

$$f(x_0, y_0, z_0) \leq f(x_0, y_0, 1) = h(1) < h(0) = f(x_0, y_0, 0),$$

which in turn implies that $z_0 \neq 0$. Therefore, any solution (x_0, y_0, z_0) to the minimization problem must satisfy $z_0 \neq 0$ and hence the point $(x_0, y_0, -z_0)$ would be a another solution. So, the minimization problem has more than one solution. Now, any solution (x, y, z) must satisfy the *KTL* conditions described by the following system of equations

$$\begin{cases} \nabla f(x,y,z) + \lambda_1 \nabla f_1(x,y,z) + \lambda_2 \nabla f_2(x,y,z) + \lambda_3 \nabla f_3(x,y,z) = 0 \\ \lambda_i f_i(x,y,z) = 0, \ \ i = 1,2,3 \\ \lambda_1, \lambda_2, \lambda_3 \geq 0 \end{cases}, \tag{4.6}$$

where $f_1(x,y,z) = (x-3)^2 + (y-1)^2 - 1$, $f_2(x,y,z) = z - 1$, $f_3(x,y,z) = -z - 1$. Calculating the gradients (4.6) becomes

$$\begin{cases} \left(-2e^{-(x^2+y^2+z^2)} - 4(x^2+y^2+z^2)\right)(x,y,z) + \lambda_1(2(x-3), 2(y-1), 0) + \lambda_2(0,0,1) + \lambda_3(0,0,-1) = (0,0,0) \\ \lambda_1(x-3)^2 + \lambda_1(y-1)^2 = \lambda_1 \\ \lambda_2 z = \lambda_2 \\ \lambda_3 z = -\lambda_3 \\ \lambda_1, \lambda_2, \lambda_3 \geq 0 \end{cases}.$$

or equivalently,

$$\begin{cases} \left(-2e^{-(x^2+y^2+z^2)} - 4(x^2+y^2+z^2)\right)x + 2\lambda_1 x - 6\lambda_1 = 0 \\ \left(-2e^{-(x^2+y^2+z^2)} - 4(x^2+y^2+z^2)\right)y + 2\lambda_1 y - 2\lambda_1 = 0 \\ \left(-2e^{-(x^2+y^2+z^2)} - 4(x^2+y^2+z^2)\right)z + \lambda_2 - \lambda_3 = 0 \\ \lambda_1(x-3)^2 + \lambda_1(y-1)^2 = \lambda_1 \\ \lambda_2 z = \lambda_2 \\ \lambda_3 z = -\lambda_3 \\ \lambda_1, \lambda_2, \lambda_3 \geq 0 \end{cases}.$$

3. Now, as in case 1, the set $S = \{(x,y,z)^T \in \mathbb{R}^3 : (x-1)^2 + y^2 =\}$ has no restrictions with respect to the variable z. Therefore, it follows that S is unbounded. To prove the existence we proceed in similar fashion as in case 1. Once we observe that

$$\lim_{\substack{\|(x,y,z)\|_2 \to \infty \\ (x,y,z)^T \in S}} f(x,y,z) = +\infty,$$

the same argument as in case 1 proves that the constrained minimization problem has at least one solution.

Exercise 5

(a) Examine the existence and uniqueness of the minimization problem $\min\limits_{(x,y,z)^T \in S} f(x,y,z) = f(\bar{x},\bar{y},\bar{z})$ in each of the following cases.

1. $f(x,y,z) = (x-1)^2+(y-1)^2+(z-1)^2,\ S = \{(x,y,z)^T \in \mathbb{R}^3 : x^2+y^2+z^2 \leq 1,\ x+y+z = 0\}$.
2. $f(x,y,z) = (x-1)^2+(y-1)^2+(z-1)^2,\ S = \{(x,y,z)^T \in \mathbb{R}^3 : x^2+2y^2 \leq 3,\ y+z = 0\}$.

Determine the possible points of local minimums by solving analytically *KTL*'s conditions.

(b) Apply *Newton - Raphson*'s method to (1.a) with starting point $(0, 1.2, -1.2)^T$. Does the algorithm converge?

(a) **1.** Let $f_1(x,y,z) = x^2+y^2+z^2-1$ and $f_2(x,y,z) = x+y+z$. Then, *KTL* conditions are described by the following system of equations.

$$\begin{cases} \nabla f(x,y,z) + \lambda_1 \nabla f_1(x,y,z) + \lambda_2 f_2(x,y,z) = 0 \\ \lambda_1 f_1(x,y,z) = 0 \\ f_2(x,y,z) = 0 \\ \lambda_1 \geq 0,\ \lambda_2 \in \mathbb{R}. \end{cases} \tag{5.1}$$

Calculating the gradients we get that (5.1) is equivalent to

$$\begin{cases} \Big(2(x-1), 2(y-1), 2(z-1)\Big) + \lambda_1(2x, 2y, 2z) + \lambda_2(1,1,1) = 0 \\ \lambda_1 = \lambda_1(x^2+y^2+z^2) \\ x+y+z = 0 \\ \lambda_1 \geq 0,\ \lambda_2 \in \mathbb{R}. \end{cases} \tag{5.2}$$

Expanding the equations in the first line in (5.2) we obtain

$$\begin{cases} 2x - 2 + 2x\lambda_1 + \lambda_2 = 0 \\ 2y - 2 + 2\lambda_1 y + \lambda_2 = 0 \\ 2z - 2 + 2\lambda_1 z + \lambda_2 = 0 \\ x+y+z = 0 \\ \lambda_1 = \lambda_1(x^2+y^2+z^2) \\ \lambda_1 \geq 0,\ \lambda_2 \in \mathbb{R}. \end{cases} \tag{5.3}$$

Subtracting the second equation from the first in (5.3) we obtain

$$\begin{aligned} 2x - 2y + 2\lambda_1 x - 2\lambda_1 y = 0 &\iff 2x(1+\lambda_1) = 2y(1+\lambda_1) \\ &\iff x(1+\lambda_1) = y(1+\lambda_1) \\ &\overset{1+\lambda_1>0}{\iff} \boxed{\text{x=y}}. \end{aligned} \tag{5.4}$$

Therefore, substituting $x = y$ the fourth equation in (5.3) we get that $\boxed{z = -2x}$. Hence, by the fifth equation in (5.3) we get that

$$\begin{aligned}\lambda_1 = \lambda_1(x^2 + y^2 + z^2) &\iff \lambda_1 = \lambda_1(2x^2 + 4x^2)\\ &\iff \lambda_1 = 6\lambda_1 x^2.\end{aligned} \tag{5.5}$$

Case 1: $\lambda_1 = 0$. Then using that $\lambda_1 = 0$ in the first two equations in (5.3) we get that

$$\left.\begin{aligned}2x - 2 + \lambda_2 = 0\\ 2z - 2 + \lambda_2 = 0\end{aligned}\right\} \overset{z=-2x}{\iff} \left.\begin{aligned}2x - 2 + \lambda_2 = 0\\ -4x - 2 + \lambda_2 = 0\end{aligned}\right\}$$

Subtracting the last two equations we end up with $6x = 0 \iff \boxed{x = 0}$. And since, $x = y$, $z = -2x$ we get that $\boxed{x = y = z = 0}$. Therefore, the first point satisfying *KTL* conditions is $\boxed{x_1 = (0, 0, 0)}$.

Case 2: $\lambda_1 \neq 0$. In this case, by (5.5) we get that $\boxed{x \pm \frac{\sqrt{6}}{6}}$. If $x = \frac{\sqrt{6}}{6}$ then since $x = y$ and $z = -2x$ we obtain the second point $\boxed{x_2 = (\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{3})}$. On the other hand, if $x = \frac{\sqrt{6}}{6}$ we get $\boxed{x_3 = (-\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3})}$.

Now, since S is convex and $f(x, y, z) = (x - 1)^2 + (y - 1)^2 + (z - 1)^2$ is also convex, f_1 is C^∞ and f_2 is an affine transformation it follows that the *KTL* conditions in (5.2) are sufficient, which means that the points $x_1 = (0, 0, 0)$, $x_2 = (\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{3})$ and $x_3 = (-\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3})$ are indeed local minimums of f on S.

2. Now in the case where $S = \{(x, y, z)^T \in \mathbb{R}^3 : x^2 + 2y^2 \leq 3,\ y + z = 0\}$ we let $f_1(x, y, z) = x^2 + 2y^2 - 3$ and $f_2(x, y, z) = y + z$. Now *KTL* conditions have again the form in (5.1). Calculating the gradients and substituting in (5.1) we obtain

$$\begin{cases}(2(x-1), 2(y-1), 2(z-1)) + \lambda_1(2x, 4y, 0) + \lambda_2(0, 1, 1) = 0\\ 3\lambda_1 = \lambda_1(x^2 + 2y^2)\\ y + z = 0\\ \lambda_1 \geq 0,\ \lambda_2 \in \mathbb{R}\end{cases}. \tag{5.6}$$

Expanding the first equation in (5.6) we obtain the following system of equations

$$\begin{cases}2x - 2 + 2\lambda_1 x = 0\\ 2y - 2 + 4\lambda_1 y + \lambda_2 = 0\\ 2z - 2 + \lambda_2 = 0\\ 3\lambda_1 = \lambda_1(x^2 + 2y^2)\\ y + z = 0\\ \lambda_1 \geq 0,\ \lambda_2 \in \mathbb{R}\end{cases}. \tag{5.7}$$

Now, by the the fifth equation we have that $\boxed{y=-z}$. By the third equation we get $\boxed{\lambda_2 = 2 - 2z}$. Therefore, substituting $\lambda_2 = 2 - 2z$ in the second equation and using $y = -z$ we obtain

$$\begin{aligned} 2y - 2 + 4\lambda_1 y + \lambda 2 = 0 &\iff 2y - \cancel{2} + 4\lambda_1 y + \cancel{2} + 2y = 0 \\ &\iff 4y(1 + \lambda_1) = 0. \end{aligned}$$

Now, since $1 + \lambda_1 \geq 0$ we get that $\boxed{y = 0}$, which in turn implies that $\boxed{z=0}$. Now, by the fourth equation we get $3\lambda_1 = \lambda_1 x^2$.

Case 1: $\lambda_1 = 0$. In this case by the first equation we obtain $\boxed{x = 1}$. Therefore, the first point satisfying the equations in (5.6) is $\boxed{x_1 = (1, 0, 0)}$.

Case 2: $\lambda_1 \neq 0$. In this case by $3\lambda_1 = \lambda_1 x^2$ we obtain $\boxed{x \pm \sqrt{3}}$. Therefore, we obtain the following points $\boxed{x_2 = (\sqrt{3}, 0, 0)}$ and $\boxed{x_3 = (-\sqrt{3}, 0, 0)}$.

Now it easily verified that all the points x_1, x_2, x_3 belong to S.

(b) Applying *Newton - Raphson* (Alg. 1) with starting point $(0, 1.2, -1.2)^T$ we obtain that the algorithm converges after 2 iterations to the global minimum $(1, 1, 1)$ of f which of course does not lie in the set S. In Fig. 9 we see the results.

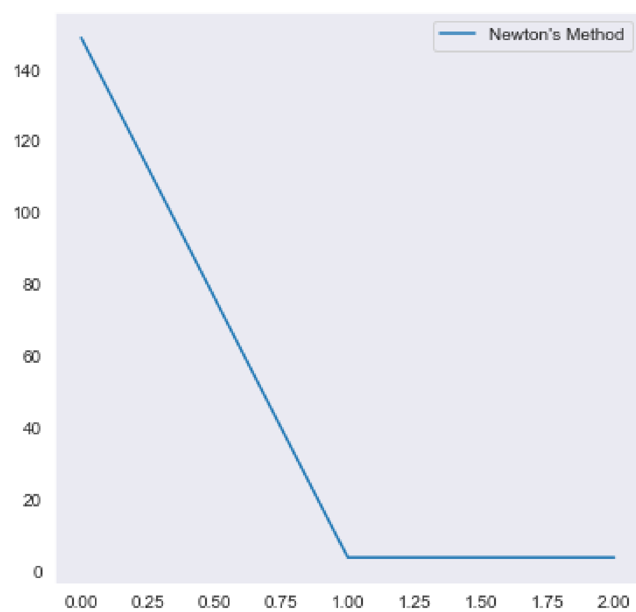

Figure 9: Newton - Raphson Algorithm.

Exercise 6

Let $A \in \mathbb{R}^{n\times n}$ be a positive definite and symmetric matrix. Furthermore, let $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = x^T Ax + b^T x$. Suppose that $m = \min_{\|x\|_2=1} x^T Ax > 0$. Prove that

$$f(x_n) - f(x_{n+1}) \leq \frac{\|r_n\|_2^2}{m}, \tag{6.1}$$

where $(x_n)_{n=0}^{\infty}$ is the sequence generated by *Steepest's Descent* method with optimized step choice.

Proof. The update rule is given by $x_{n+1} = x_n + a_n r_n$, where $r_n = b - Ax_n$. Since the choice of the step a_n is optimal we know that $a_n = r_n^T r_n / r_n^T A r_n$. Now, we write

$$\begin{aligned} f(x_n) - f(x_{n+1}) &= \frac{1}{2}x_n^T Ax_n - b^T x_n - \frac{1}{2}(x_n + a_n r_n)^T A(x_n + a_n r_n) + b^T(x_n + a_n r_n) \\ &= \cancel{\frac{1}{2}x_n^T Ax_n} - \cancel{b^t x_n} - \cancel{\frac{1}{2}x_n^T Ax_n} - \frac{a_n}{2}x_n^T Ar_n - \frac{a_n}{2}r_n^T Ax_n - \frac{a_n^2}{2}r_n^T Ar_n \\ &+ \cancel{b^T x_n} + a_n b^T r_n. \end{aligned}$$

Now using the fact that A is symmetric we have that $r_n^T Ax_n = x_n^T Ar_n$. Therefore,

$$\begin{aligned} f(x_n) - f(x_{n+1}) &= -a_n x_n^T Ar_n - \frac{a_n^2}{2}r_n^T Ar_n + a_n b^T r_n \\ &= -a_n(b^T - r_n^T)r_n - \frac{a_n^2}{2}r_n^T Ar_n + a_n b^T r_n \\ &= -\cancel{a_n b^T r_n} + a_n r_n^T r_n - \frac{a_n^2}{2}r_n^T Ar_n + \cancel{a_n b^T r_n}. \end{aligned}$$

Now substituting $a_n = r_n^T r_n / r_n^T A r_n$ we obtain

$$\begin{aligned} f(x_n) - f(x_{n+1}) &= \frac{r_n^T r_n}{r_n^T Ar_n} r_n^T r_n - \frac{1}{2}\frac{(r_n^T r_n)^2}{(r_n^T Ar_n)^2} r_n^T Ar_n \\ &= \frac{\|r_n\|_2^4}{\|r_n\|_A^2} - \frac{1}{2}\frac{\|r_n\|_2^4}{\|r_n\|_A^2} \\ &= \frac{1}{2}\frac{\|r_n\|_2^4}{\|r_n\|_A^2} = \frac{1}{2}\frac{\|r_n\|_2^2\,\|r_n\|_2^2}{\|r_n\|_A^2} \\ &= \frac{1}{2}\frac{\|r_n\|_2^2}{\left\|\frac{r_n}{\|r_n\|_2}\right\|_A^2}, \end{aligned}$$

but now, since for $x = r_n / \|r_n\|_2$ we have that $\|x\|_2 = 1$ then $m \leq x^T Ax = \|x\|_A^2$. Therefore,

$$f(x_n) - f(x_{n+1}) = \frac{1}{2}\frac{\|r_n\|_2^2}{\left\|\frac{r_n}{\|r_n\|_2}\right\|_A^2} \leq \frac{1}{2}\frac{\|r_n\|_2^2}{m},$$

which completes the proof. $\square$

References

[1] Markus Grasmair. *Basic properties of convex functions.* 2016.

[2] Stephen Wright, Jorge Nocedal, et al. “Numerical optimization”. In: *Springer Science* 35.67-68 (1999).