Part 1

Task 1

First, we shall write the network function explicitly:

$$F(x,W) = W_3^T\phi_2(W_2^T\phi_1(W_1^T x + b_1) + b_2) + b_3$$

Therefore:

$$\psi(r) = (F(x,W) - y)^2 = (W_3^T\phi_2(W_2^T\phi_1(W_1^T x + b_1) + b_2) + b_3 - y)^2$$

Now we can easily derive the loss.

We shall calculate the derivative of tanh as well:

$$\phi'(x) = tanh'(x) = \frac{2\cdot e^{-2x}(1+e^{-2x}) + 2\cdot e^{-2x}(1-e^{-2x})}{(1+e^{-2x})^2} = \frac{4x\cdot e^{-2x}}{(1+e^{-2x})^2}$$

Last thing we want to define before calculating the gradients is the layers' outputs:

$$q_1 = \phi_1(W_1^T x + b_1)$$
$$q_2 = \phi_2(W_2^T\phi_1(W_1^T x + b_1) + b_2)$$

w.r.t b_3:

$$d_{b_3}\psi = d_{b_3}(F-y)^2 = 2(F-y)\cdot d_{b_3}F = 2(F-y)\cdot d_{b_3}(W_3^T q_2 + b_3) = \langle 2(F-y), d_{b_3}b_3\rangle$$

$$\boxed{\nabla_{b_3}\psi = 2(F-y)}$$

w.r.t W_3:

$$d_{W_3}\psi = d_{W_3}(F-y)^2 = 2(F-y)\cdot d_{W_3}F = 2(F-y)\cdot d_{W_3}(W_3^T q_2 + b_3) =$$
$$= \underbrace{2(F-y)}_{scalar}\cdot d_{W_3}(W_3^T)q_2 = \langle 2(F-y)q_2, dW_3\rangle$$

$$\boxed{\nabla_{W_3}\psi = 2(F-y)q_2}$$

w.r.t b_2:

$$d_{b_2}\psi = d_{b_2}(F-y)^2 = 2(F-y)\cdot d_{b_2}F = 2(F-y)\cdot d_{b_2}(W_3^T\phi_2(W_2^T q_1 + b_2) + b_3) =$$
$$= 2(F-y)\cdot W_3^T d_{b_2}\phi_2(W_2^T q_1 + b_2) = 2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{b_2}(W_2^T q_1 + b_2) =$$
$$= 2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{b_2}b_2 = \left\langle 2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right), d_{b_2}b_2\right\rangle$$

$$\boxed{\nabla_{b_2}\psi = 2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right)}$$

w.r.t W_2:

$$d_{W_2}\psi = d_{W_2}(F-y)^2 = 2(F-y)\cdot d_{W_2}F = 2(F-y)\cdot d_{W_2}(W_3^T\phi_2(W_2^T q_1 + b_2) + b_3) =$$
$$= 2(F-y)\cdot W_3^T d_{W_2}\phi_2(W_2^T q_1 + b_2) = 2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{W_2}(W_2^T q_1 + b_2) =$$
$$= Tr\left(2(F-y)\cdot W_3^T diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{W_2}W_2^T q_1\right) =$$
$$= Tr\left(2(F-y)diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{W_2}W_2^T\cdot q_1 W_3^T\right) =$$
$$= Tr\left(2(F-y)q_1 W_3^T\cdot diag\left(\phi_2'(W_2^T q_1 + b_2)\right)\cdot d_{W_2}W_2^T\right) = \langle 2(F-y)q_1 W_3^T\cdot diag\left(\phi_2'(W_2^T q_1 + b_2)\right), d_{W_2}W_2^T\rangle$$

$$\boxed{\nabla_{W_2}\psi = 2(F-y)q_1 W_3^T\cdot diag\left(\phi_2'(W_2^T q_1 + b_2)\right)}$$

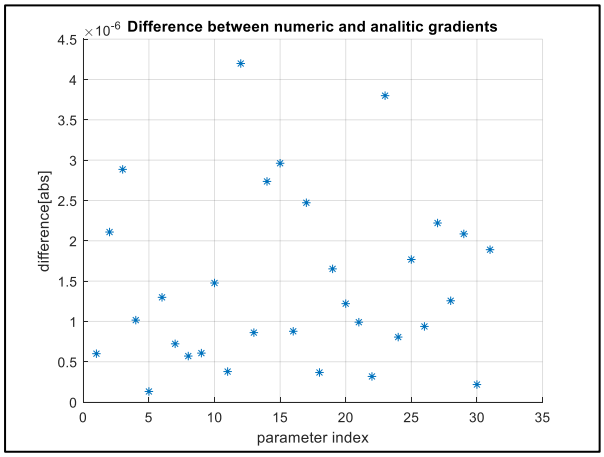
w.r.t b_1:

$$d_{b_1}\psi = d_{b_1}(F-y)^2 = 2(F-y)\cdot d_{b_1}F =$$
$$= 2(F-y)\cdot d_{b_1}(W_3^T\phi_2(W_2^T\phi_1(W_1^Tx+b_1)+b_2)+b_3) =$$
$$= 2(F-y)\cdot W_3^T d_{b_1}\phi_2(W_2^T\phi_1(W_1^Tx+b_1)+b_2) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot d_{b_1}(W_2^T\phi_1(W_1^Tx+b_1)+b_2) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T d_{b_1}\phi_1(W_1^Tx+b_1) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{b_1}(W_1^Tx+b_1) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{b_1}b_1 =$$
$$= \Big\langle 2(F-y)\cdot diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big), d_{b_1}b_1\Big\rangle$$
$$\boxed{\nabla_{b_1}\psi = 2(F-y)\cdot diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)}$$

w.r.t W_1:

$$d_{W_1}\psi = d_{W_1}(F-y)^2 = 2(F-y)\cdot d_{W_1}F =$$
$$= 2(F-y)\cdot d_{W_1}(W_3^T\phi_2(W_2^T\phi_1(W_1^Tx+b_1)+b_2)+b_3) =$$
$$= 2(F-y)\cdot W_3^T d_{W_1}\phi_2(W_2^T\phi_1(W_1^Tx+b_1)+b_2) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot d_{W_1}(W_2^T\phi_1(W_1^Tx+b_1)+b_2) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T d_{W_1}\phi_1(W_1^Tx+b_1) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{W_1}(W_1^Tx+b_1) =$$
$$= 2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{W_1}W_1^Tx =$$
$$= Tr\Big(2(F-y)\cdot W_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{W_1}W_1^Tx\Big) =$$
$$= Tr\Big(2(F-y)\cdot diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{W_1}W_1^TxW_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T\Big) =$$
$$= Tr\Big(2(F-y)\cdot xW_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T\cdot diag\Big(\phi_1'(W_1^Tx+b_1)\Big)d_{W_1}W_1^T\Big) =$$
$$= \langle 2(F-y)\cdot xW_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T\cdot diag\Big(\phi_1'(W_1^Tx+b_1)\Big), d_{W_1}W_1^T\rangle$$
$$\boxed{\nabla_{W_2}\psi = 2(F-y)\cdot xW_3^T diag\Big(\phi_2'(W_2^T\phi_1(W_1^Tx+b_1)+b_2)\Big)\cdot W_2^T\cdot diag\Big(\phi_1'(W_1^Tx+b_1)\Big)}$$

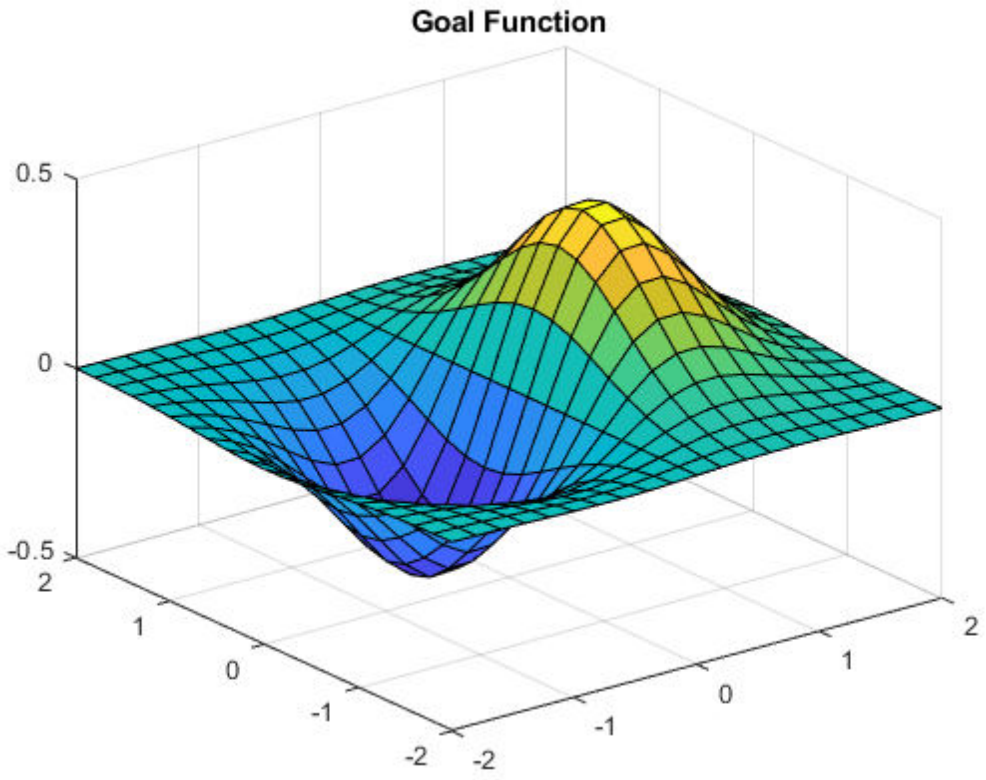
We randomly picked weights and input values for the model and calculated numerically and analytically so we can verify the above expressions.
The results of $|analitical - numerical|$ are as follow:



We may see that the difference for each one of the 31 weights is bounded by $\sim 4.5e^{-6}$ meaning that the gradients were calculated correctly.

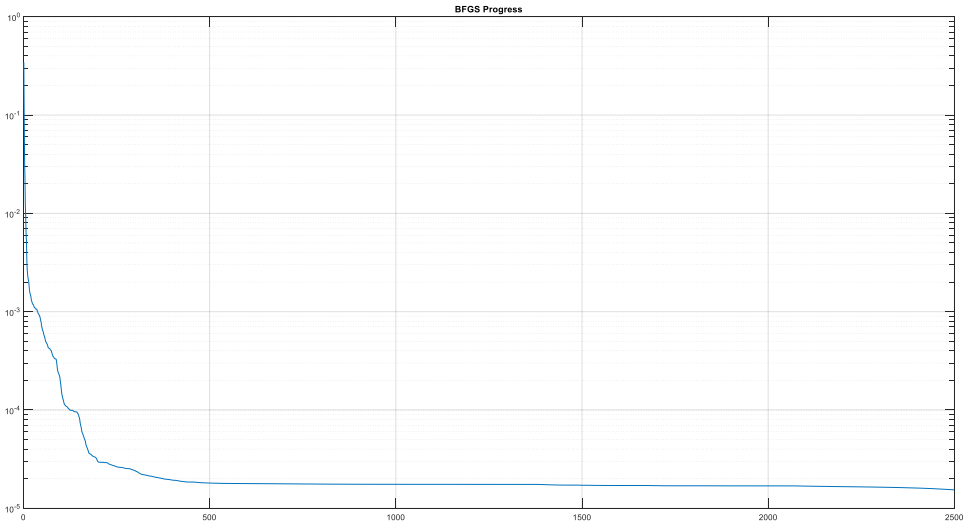
Task 2:

The function we'd like to estimate look like:

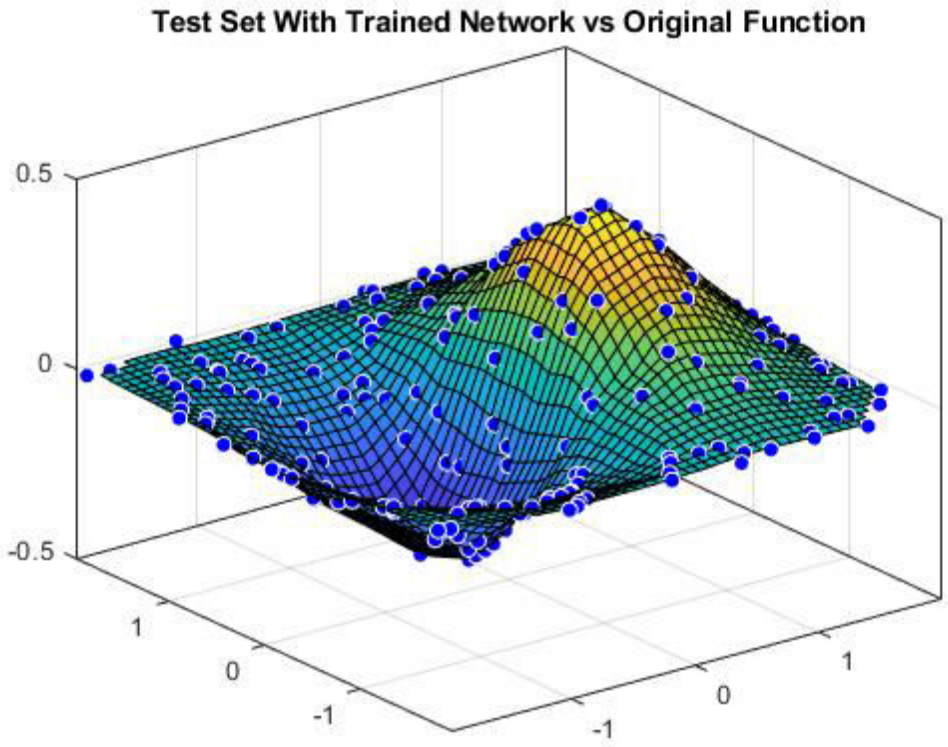


Task 3

We have trained our model using BFGS, we may see that the model has converged very fast:



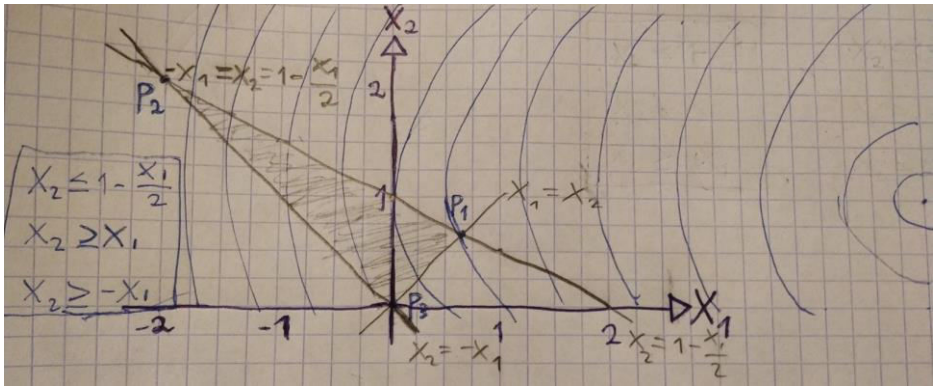
Afterwards we inserted the test set to the model and plotted the results on the target function:



Part 2 – Augmented Lagrangian:

Task 1 – Quadratic Programming:

$$f(x) = 2(x_1 - 5)^2 + (x_2 - 1)^2$$

1. 

   The active constrains are (g1 and g2):

   $g1(x) = x_2 + \frac{x_1}{2} - 1 \leq 0$

   $g2(x) = x_1 - x_2 \leq 0$

2. Calculation of the intersections of the active constraints:

   P1:

   $$\begin{cases} \boldsymbol{x2 = x1} \\ \boldsymbol{x2 = 1 - \frac{x1}{2}} \end{cases}$$

   $$\Rightarrow x1 = 1 - \frac{x1}{2} \Rightarrow \frac{3x1}{2} = 1 \Rightarrow x1 = \frac{2}{3} = x2$$

   $$f\left(\frac{2}{3}, \frac{2}{3}\right) = 37\frac{2}{3}$$

   In purpose to verify that those are the active constraints, we need to get higher value in the other intersections:

   P2:

   $$\begin{cases} \boldsymbol{x2 = 1 - \frac{x1}{2}} \\ \boldsymbol{x2 = -x1} \end{cases}$$

   $$\Rightarrow -x1 = 1 - \frac{x1}{2} \Rightarrow x1 = -2, x2 = 2$$

   $f(-2,2) = 99$ V

   P3:

   $$\begin{cases} \boldsymbol{x2 = x1} \\ \boldsymbol{x2 = -x1} \end{cases}$$

   $$\Rightarrow x1 = 0 = x2$$

   $f(0,0) = 51$ V

3. The primal problem is:

$$\min_x f(x) \ s.t.$$
$$g1(x) = x_2 + \frac{x_1}{2} - 1 \le 0$$
$$g2(x) = x_1 - x_2 \le 0$$
$$g3(x) = -x_1 - x_2 \le 0$$

The lagrangian is:

$$L(x_1, x_2, \lambda) = f(x) + \sum_{i=1}^{3} \lambda_i g_i(x)$$
$$= 2(x_1 - 5)^2 + (x_2 - 1)^2 + \lambda_1 \left(x_2 + \frac{x_1}{2} - 1\right) + \lambda_2 (x_1 - x_2) + \lambda_3 (-x_1 - x_2)$$

- $\nabla_{x_1} L = 4(x_1 - 5) + \frac{\lambda_1}{2} + \lambda_2 - \lambda_3 = 4x_1 + 0x_2 + \frac{1}{2}\lambda_1 + \lambda_2 - \lambda_3 - 20 = 0$

  $\Downarrow$

  $$4x_1 + 0x_2 + \frac{1}{2}\lambda_1 + \lambda_2 - \lambda_3 = 20$$

- $\nabla_{x_2} L = 2(x_2 - 1) + \lambda_1 - \lambda_2 - \lambda_3 = 0$

  $$0x_1 + 2x_2 + \lambda_1 - \lambda_2 - \lambda_3 - 2 = 0$$

  $\Downarrow$

  $$0x_1 + 2x_2 + \lambda_1 - \lambda_2 - \lambda_3 = 2$$

- $\nabla_{\lambda_1} L = x_2 + \frac{x_1}{2} - 1 = 0$

  $$\frac{1}{2}x_1 + x_2 + 0\lambda_1 + 0\lambda_2 + 0\lambda_3 = 1$$

- $\nabla_{\lambda_2} L = x_1 - x_2 = 0$
- $\nabla_{\lambda_3} L = -x_1 - x_2 = 0$
- Since g3(x) is inactive we get $\lambda_3 = 0$.

  $\Downarrow$

$$\begin{pmatrix} 4 & 0 & \frac{1}{2} & 1 \\ 0 & 2 & 1 & -1 \\ \frac{1}{2} & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 20 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{2}{3} \\ 12 \\ 11\frac{1}{3} \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{2}{3} \\ 12 \\ 11\frac{1}{3} \\ 0 \end{pmatrix}$$

We have $\forall i \in [3]: \lambda_i \geq 0$, and since we got P1 as the solution we know that the constrains are met, so we meet the KKT conditions.

4. We found that:

$$\nabla_{x_1} L = 4x_1 + 0x_2 + \frac{1}{2}\lambda_1 + \lambda_2 - \lambda_3 - 20 = 0$$

$$\nabla_{x_2} L = 2(x_2 - 1) + \lambda_1 - \lambda_2 - \lambda_3 = 0$$

$$\Downarrow$$

$$x_1^* = \frac{-\frac{1}{2}\lambda_1 - \lambda_2 + \lambda_3 + 20}{4} = \frac{-\lambda_1 - 2\lambda_2 + 2\lambda_3}{8} + 5$$

$$2x_2 - 2 + \lambda_1 - \lambda_2 - \lambda_3 = 0$$

$$x_2^* = \frac{-\lambda_1 + \lambda_2 + \lambda_3}{2} + 1$$

Therefore:

$$g(\lambda) = \inf_{x} \mathcal{L}(x, \lambda) = L(\lambda, x^*) = 2\left(\frac{-\lambda_1 - 2\lambda_2 + 2\lambda_3}{8}\right)^2 + \left(\frac{-\lambda_1 + \lambda_2 + \lambda_3}{2}\right)^2 +$$

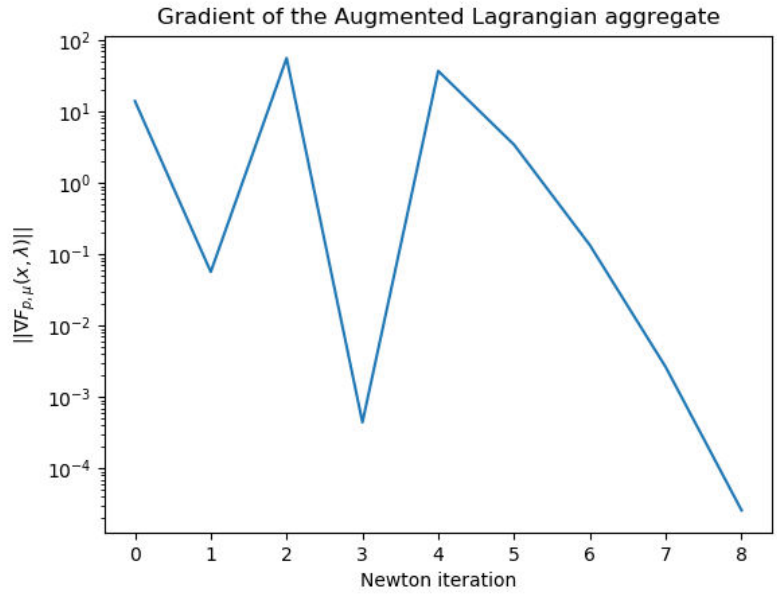
$$\lambda_1\left(-\frac{9}{16}\lambda_1 + \frac{3}{8}\lambda_2 + \frac{5}{8}\lambda_3 + 2.5\right) + \lambda_2\left(\frac{3}{8}\lambda_1 - \frac{3}{4}\lambda_2 - \frac{1}{4}\lambda_3 + 4\right) + \lambda_3\left(\frac{5}{8}\lambda_1 - \frac{1}{4}\lambda_2 - \frac{3}{4}\lambda_3 - 6\right)$$

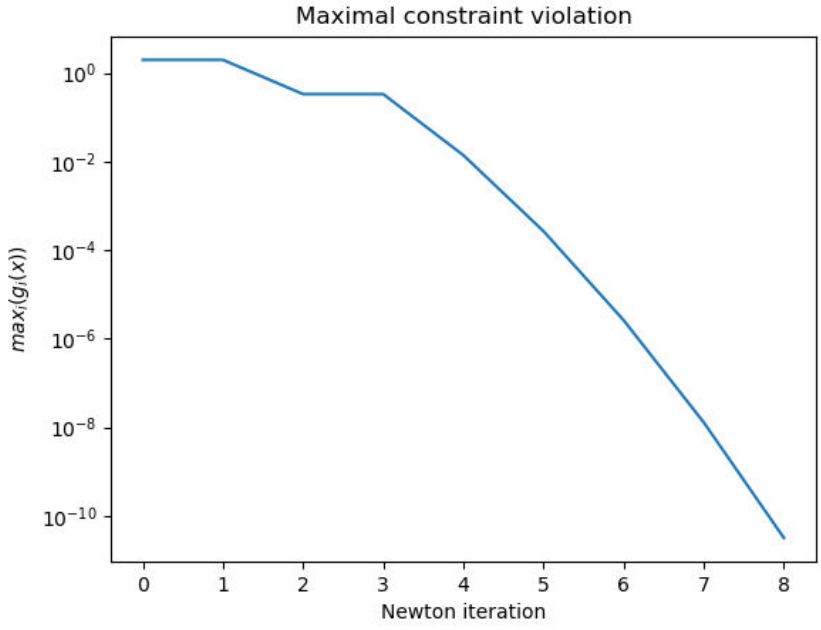
The dual problem is: $\sup_{\lambda_i \geq 0} g(\lambda)$

5. Substituting $(\lambda_1, \lambda_2, \lambda_3) = \left(12, 11\frac{1}{3}, 0\right)$ to $g(\lambda)$, we get: $g(\lambda), = 37\frac{2}{3}$

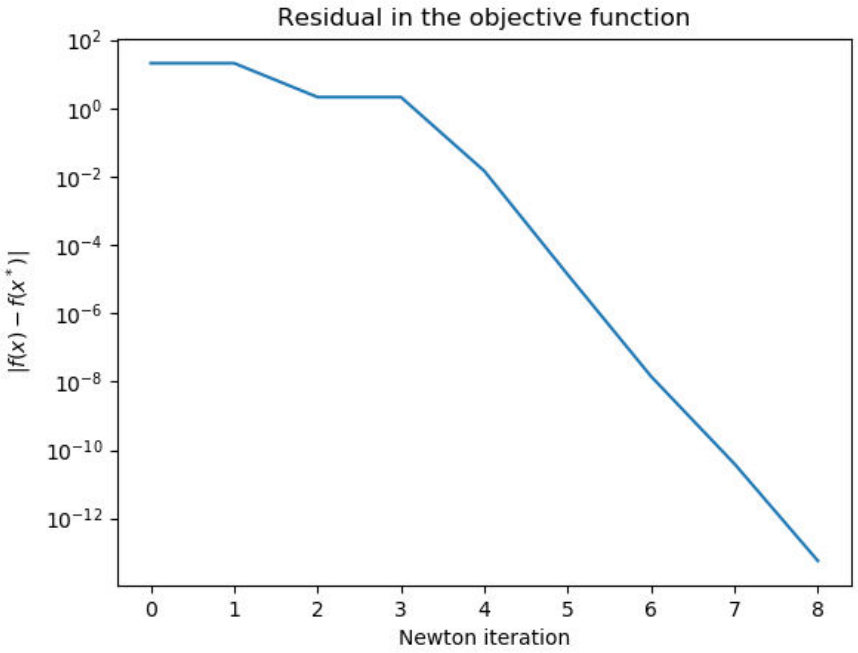
   As we can see $f(x)$ is convex and All the constraints are linear, hence convex. Hence, we can assume strong duality. Thus, the optimum of the dual function is the same as the optimum of the objective function, and we got the same value, so the optimum of the dual function is achieved.

Task 2 (code):

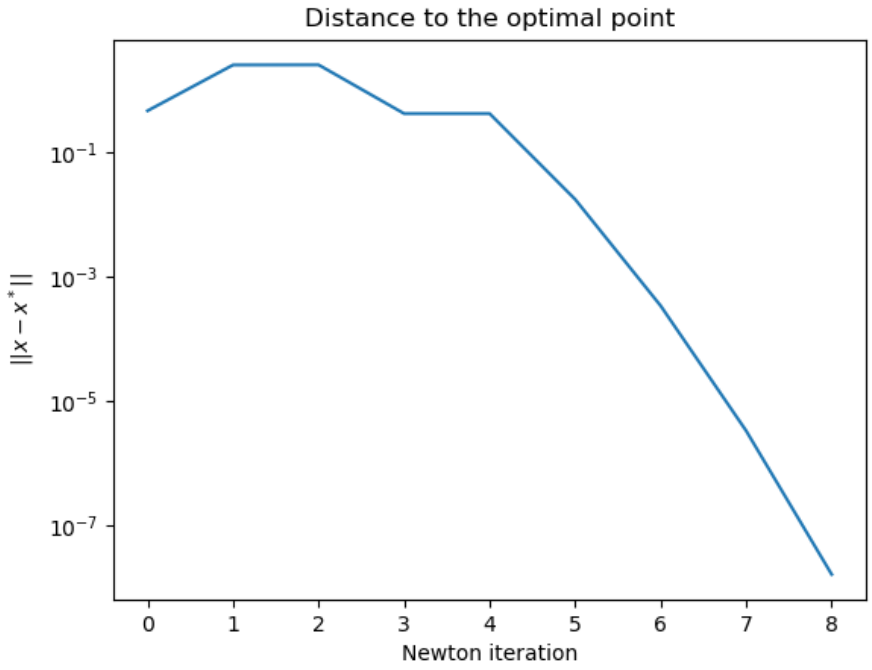
1. 

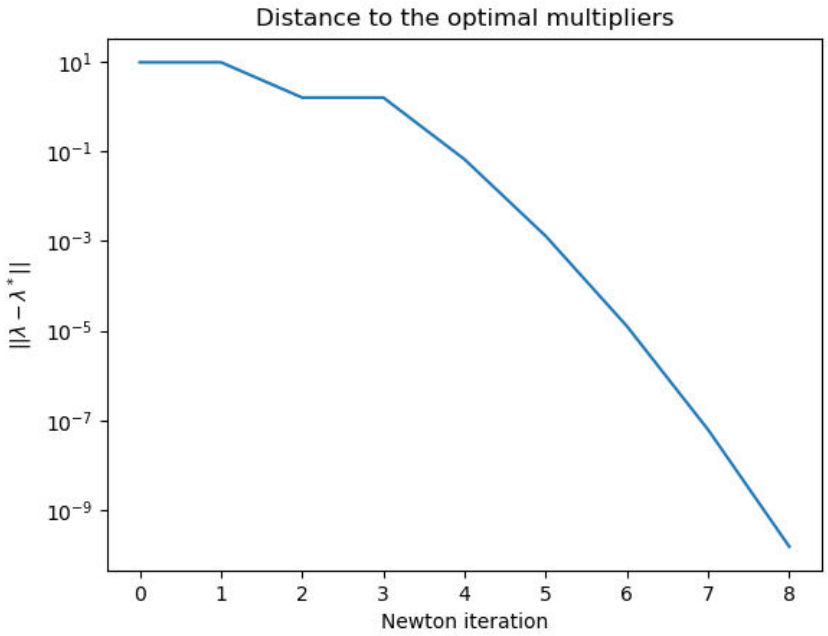
2. 

3.


Residual in the objective function
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
$|f(x) - f(x^*)|$
0
1
2
3
4
5
6
7
8
Newton iteration

4.


Distance to the optimal point
10^{-1}
10^{-3}
10^{-5}
10^{-7}
$||x - x^*||$
0
1
2
3
4
5
6
7
8
Newton iteration


Distance to the optimal multipliers
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
$||\lambda - \lambda^*||$
0
1
2
3
4
5
6
7
8
Newton iteration