Optimering för maskinlärning hand in 1

Karl Tengelin & Daniel Jogstad

September 2019

1 Introduction

The goal of this exercise is to solve the following problem:

$$\min_{x \in S} f(x) = \min_{x} f(x) + \iota_S(x) \tag{1}$$

Where the function f(x) is defined as:

$$f(x) = \frac{1}{2} x^T Q x + q^T x \tag{2}$$

And the set S is defined as $S = \{x : \forall i, a_i \leq x_i \leq b_i\}$.

For this we will use the proximal gradient method which solves problem on the form given by 1. The proximal gradient method updates a solution in an iterative manor until the solution converges using the following updating formula:

$$x^{k+1} = \text{prox}_{\gamma \iota_S} \left(x^k - \gamma \nabla f \left(x^k \right) \right) \tag{3}$$

Task 1

Derive f^*

$$f(s)^* = \sup_x (s^T x - f(x)) = \sup_x (s^T x - \frac{1}{2} x^T Q x - q^T x) \tag{4}$$

$$\nabla f(x) = Qx + q \tag{5}$$

Fermats rule (Let f : $\mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then x minimizes f iff $0 \in \partial f(x)$) gives us that $s - Qx - q = 0$

$$
\begin{gathered}
s = Qx + q \\
x = Q^{-1}(s - q) \\
\Downarrow \\
f^*(s) = \sup_x (s^T Q^{-1}(s-q) - \frac{1}{2} Q^{-1T}(s-q)^T Q Q^{-1}(s-q) - q^T Q^{-1}(s-q) \\
= \frac{1}{2}(s-q)^T Q^{-1}(s-q)
\end{gathered} \tag{6}
$$

Derive i_S^*

$$
\iota^*_{[a_i,b_i]}(s) = \sup_x (s^T x - i_{[a_i,b_i]}) = \sup_{x \in [a_i,b_i]} (s^T x) = \sum_i \begin{cases} s_i a_i, s_i < 0 \\ s_i b_i, s_i \geq 0 \end{cases} \tag{7}
$$

Important to note here is that depending on the sign of s and the values of a and b different x maximizes the function.

Fenchel-dual problem

We define the Fenchel dual problem as:

$$
\inf_x (f(Lx) + g(x)) = \max_\mu \left(-f^*(\mu) - g^*\left(-L^T \mu\right)\right) \tag{8}
$$

Here we can identify that in our case L = I and $g^* = \iota_S^*$, hence our dual problem can be written as:

$$
\max_\mu (-f^*(\mu) - \iota_S^*(-I^T \mu)) = \min_\mu f^*(\mu) + \iota_S^*(-I\mu) \tag{9}
$$

Task 2

Showing that f is L-smooth

The definition of Lipschitz smoothness is that a function is L-smooth if its derivatives are Lipschitz continous for constant L. I.e if the following holds:

$$
||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||| \tag{10}
$$

Here we use $\nabla f(x)$ given in task 3 such that:

$$
\begin{aligned}
&||\nabla f(x) - \nabla f(y)||_2 \leq L||x-y|||\\
&\Downarrow\\
&||Qx + q - Qy - q||_2 \leq L||x-y||\\
&\Downarrow\\
&||Q||_2||x-y|| \leq L||x-y||\\
&\Downarrow\\
&||Q||_2 \leq L
\end{aligned}
\tag{11}
$$

Showing that f^* is L^*-smooth

$$
\begin{aligned}
&||\nabla f(x)^* - \nabla f(y)^*||_2 \leq L^*||x-y|||\\
&\Downarrow\\
&||Q^{-1}(x-q) - Q^{-1}(y-q)||_2 \leq L^*||x-y||\\
&\Downarrow\\
&||Q^-1||_2||x-y|| \leq L^*||x-y||\\
&\Downarrow\\
&||Q^{-1}||_2 \leq L^*
\end{aligned}
\tag{12}
$$

Task 3

Derive ∇f

$f(x)$ is given as:

$$f(x) = \frac{1}{2}x^T Qx + q^T x \tag{13}$$

hence:

$$\nabla f(x) = Qx + q \tag{14}$$

Derive ∇f^*

$f(s)^*$ is calculated as:

$$\frac{1}{2}(s-q)^T Q^{-1}(s-q) \tag{15}$$

Therefore:

$$\nabla f^*(s) = Q^{-1}(s-q) \tag{16}$$

Derive $\text{prox}_{\gamma\iota_S}$

The definition of $prox_{\gamma h}$ is:

$$\text{prox}_{\gamma g}(z) = \underset{x}{\text{argmin}} \left(g(x) + \frac{1}{2\gamma}\|x - z\|_2^2 \right) \tag{17}$$

And in our case $g(x) = \iota_S$. Hence we solve the following problem:

$$\begin{gathered} \text{prox}_{\gamma\iota_S}(z) = \underset{x}{\text{argmin}} \left(\iota_S(x) + \frac{1}{2\gamma}\|x - z\|_2^2 \right) \\ \Updownarrow \\ \text{prox}_{\gamma\iota_S}(z) = \underset{x \in S}{\text{argmin}} \left(\frac{1}{2\gamma}\|x - z\|_2^2 \right) \end{gathered} \tag{18}$$

We can see that in order to minimize the function x has to lie as close as possible to z (or x_i needs to lie close to z_i when looking at multiple dimensions). However we need that $x \in S$ so if z_i is below a_i we choose $x_i = a_i$ as it is as close as we get while still being inside the interval. The same reasoning holds if $b_i < z_i$. So the answer is:

$$\left(prox_{\gamma\iota_S}(z)\right)_i = \begin{cases} a_i, z_i < a_i \\ z_i, a_i \leq z_i \leq b_i \\ b_i, b_i < z_i \end{cases} \tag{19}$$

So this method needs to be performed element-wise in order to evaluate the whole z-vector.

Derive $\text{prox}_{\gamma\iota_S^*}$

Here we use Moreau decomposition:

$$z = \text{prox}_{\gamma f}(z) + \text{prox}_{(\gamma f)^*}(z) = \text{prox}_{\gamma f}(z) + \gamma \, \text{prox}_{\gamma^{-1} f^*}\left(\gamma^{-1} z\right) \tag{20}$$

And since the indicator function ι has the property that its biconjugate is the indicator function itself we can rewrite 20 as:

$$\begin{gathered} z = \text{prox}_{\gamma\iota}(z) + \text{prox}_{(\gamma\iota)^*}(z) = \text{prox}_{\gamma\iota}(z) + \gamma \, \text{prox}_{\gamma^{-1}\iota^*}\left(\gamma^{-1} z\right) \\ \Downarrow \\ z = \text{prox}_{\gamma\iota^*}(z) + \gamma \, \text{prox}_{\gamma^{-1}\iota}(\gamma^{-1} z) \\ \Downarrow \\ \text{prox}_{\gamma\iota^*}(z) = z - \gamma \, \text{prox}_{\gamma^{-1}\iota}(\gamma^{-1} z) \end{gathered} \tag{21}$$

Hence the conjugate function ι^* can be computed using γ to rescale in equation 19:

$$\Big(prox_{\gamma\iota_S^*}(z)\Big)_i = \begin{cases} z_i - \gamma a_i, \frac{z_i}{\gamma} < a_i \\ 0, a_i \leq \frac{z_i}{\gamma} \leq b_i \\ z_i - \gamma b_i, b_i < \frac{z_i}{\gamma} \end{cases} \tag{22}$$

Task 4

$f(s)^*$ and i_S are both closed convex functions and constraint qualification holds. Then it also holds that

$$0 \in \partial f(x) + \partial g(x) \quad \Leftrightarrow \quad \begin{matrix} y \in \partial f(x) & x \in \partial f^*(y) \\ -y \in \partial g(x) & x \in \partial g^*(-y) \end{matrix} \tag{23}$$

since $x \in \partial f^*(y)$ we can create a candidate solution $\hat{x} \in \partial f^*\left(y^\star\right)$
Since f^* is differentiable, $\partial f^*\left(y^*\right)$ will be unique for all y^* and subsequently

$$\hat{x} = \nabla f^*\left(y^*\right) = Q^{-1}(y^* - q), \tag{24}$$

which means that from a solution y^* to the dual problem we can recover a solution $\hat{x}$.

Task 5

See appended julia-file.

Task 6

The file problem.jl contains a function for generating Q, q, a, and b that define the quadratic f and the box constraint set S. Use Task 5 to solve the primal problem using the proximal gradient method. Try a range of different step-sizes. What seems to be the best choice? Does the upper bound $\gamma < 2/L$ seem reasonable?

Test different initial points for the algorithm, does this affect the solution the algorithm converge to? Reason about why/why not it affects the solution? Does your solution satisfy the constraint $x^* \in S$? What about the iterates, do they always satisfy the constraint, $x^k \in S$? Why/why not?

As concluded in task 3, f^* was proven to be L-smooth for $L = \|Q\|_2$. Thus the maximum step size to guarantee convergence is $\gamma < \frac{2}{L}$, as given in the task description. For this task the six different step sizes were tested ranging from $\frac{1}{5} \cdot \frac{2}{L}$ to $\frac{6}{5} \cdot \frac{2}{L}$ and table 1 shows how many iterations until the norm $||x^{k+1} - x^k||_2$ was smaller than 10^{-15}.

Step length γ:	$\frac{1}{5} \cdot \frac{2}{L}$	$\frac{2}{5} \cdot \frac{2}{L}$	$\frac{3}{5} \cdot \frac{2}{L}$	$\frac{4}{5} \cdot \frac{2}{L}$	$\frac{5}{5} \cdot \frac{2}{L}$	$\frac{6}{5} \cdot \frac{2}{L}$
Iterations:	7268	3810	2620	2016	1609	$> 10\ 000$

Table 1: For starting point 1: Shows how many iterations needed for each respective step length in order to achieve $||x^{k+1} - x^k||_2 < 10^{-15}$

As one can observe in 1 the optimal step length appears to be somewhere around the limit $\frac{2}{L}$. And as soon as that threshold is breached the solution does not seem to converge. Also when doing the same calculations but with a different starting point the convergence times alter slightly, as can be seen in 2. This seems reasonable since different starting points would imply that each value in each dimension has to "travel" different lengths to reach the solution. A trivial example would be if the solution was 5 for a variable x, then it takes less time to get from 4 to 5 than from 3 to 5.

Step length γ:	$\frac{1}{5} \cdot \frac{2}{L}$	$\frac{2}{5} \cdot \frac{2}{L}$	$\frac{3}{5} \cdot \frac{2}{L}$	$\frac{4}{5} \cdot \frac{2}{L}$	$\frac{5}{5} \cdot \frac{2}{L}$	$\frac{6}{5} \cdot \frac{2}{L}$
Iterations:	8001	4073	2714	1983	1531	$> 10\ 000$

Table 2: For starting point 2: Shows how many iterations needed for each respective step length in order to achieve $||x^{k+1} - x^k||_2 < 10^{-15}$

We can also see how the solution converges by looking in figure 1

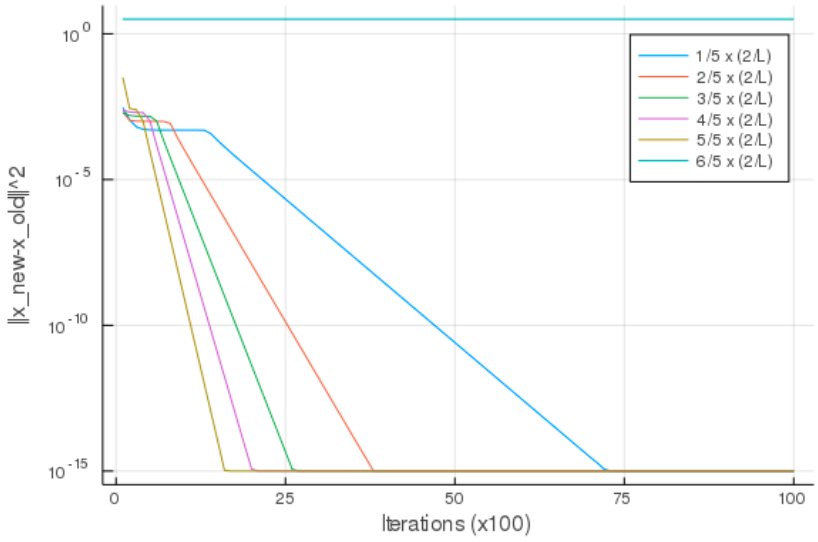

Figure 1: Shows the convergence rate for different step sizes.

In order to check if $x^* \in S$ (for a specific starting point and step size) one can check the end solution x^* element-wise and compare to the a values and b values, see matrix 25. Also one can see that if the starting values change the solution is still the same and that for both starting points the solution is within the interval $[a,b]$. For this illustration the converged solution for step size $\frac{1}{5} \cdot \frac{2}{L}$ was used for each starting point.

$$\begin{bmatrix} a_i & x_i^1 & x_i^2 & b_i \\ -0.128 & 0.336 & 0.336 & 0.733 \\ -0.148 & 0.35 & 0.35 & 0.691 \\ -0.874 & -0.126 & -0.126 & 0.528 \\ -0.486 & -0.022 & -0.022 & 0.683 \\ -0.496 & 0.034 & 0.034 & 0.922 \\ -0.641 & 0.19 & 0.19 & 0.71 \\ -0.375 & -0.05 & -0.05 & 0.751 \\ -0.034 & 0.264 & 0.264 & 0.31 \\ -0.893 & 0.081 & 0.081 & 0.835 \\ -0.886 & -0.423 & -0.423 & 0.378 \\ -0.398 & 0.431 & 0.431 & 0.431 \\ -0.989 & 0.261 & 0.261 & 0.457 \\ -0.274 & -0.127 & -0.127 & 0.689 \\ -0.783 & -0.383 & -0.383 & 0.877 \\ -0.995 & -0.826 & -0.826 & 0.12 \\ -0.661 & -0.194 & -0.194 & 0.612 \\ -0.29 & 0.165 & 0.165 & 0.165 \\ -0.958 & 0.028 & 0.028 & 0.089 \\ -0.55 & -0.103 & -0.103 & 0.598 \\ -0.274 & 0.246 & 0.246 & 0.718 \end{bmatrix} \tag{25}$$

The reason to why the solutions are the same irregardless of the starting point is that we in essence are solving a strong convex problem. We know that the function 2 is a strongly convex function due to its quadratic term and we know that the indicator function is convex. Strongly convex function + convex function = strongly convex function and this means that for the problem we are solving the local minima is the same as the global minima and irregardless of the starting point we always find the same minima, hence the solutions are the same.

If we look at a snapshot of how the solution looks like after 100 iterations and compare to a converged solution it looks like the following matrix, where we use a step size of $\frac{1}{5} \cdot \frac{2}{L}$:

$$\begin{bmatrix} a_i & x_i & x_i - snapshot\ 100\ iterations & b_i \\ -0.128 & 0.336 & 0.315 & 0.733 \\ -0.148 & 0.35 & 0.278 & 0.691 \\ -0.874 & -0.126 & 0.053 & 0.528 \\ -0.486 & -0.022 & 0.128 & 0.683 \\ -0.496 & 0.034 & -0.049 & 0.922 \\ -0.641 & 0.19 & -0.014 & 0.71 \\ -0.375 & -0.05 & -0.102 & 0.751 \\ -0.034 & 0.264 & 0.07 & 0.31 \\ -0.893 & 0.081 & 0.067 & 0.835 \\ -0.886 & -0.423 & -0.192 & 0.378 \\ -0.398 & 0.431 & 0.171 & 0.431 \\ -0.989 & 0.261 & 0.286 & 0.457 \\ -0.274 & -0.127 & -0.207 & 0.689 \\ -0.783 & -0.383 & -0.283 & 0.877 \\ -0.995 & -0.826 & -0.345 & 0.12 \\ -0.661 & -0.194 & -0.238 & 0.612 \\ -0.29 & 0.165 & 0.165 & 0.165 \\ -0.958 & 0.028 & 0.014 & 0.089 \\ -0.55 & -0.103 & 0.036 & 0.598 \\ -0.274 & 0.246 & 0.054 & 0.718 \end{bmatrix} \tag{26}$$

We can see that the snapshot differs from the solution but is still within the interval [a,b]. This is probably due to the nature of our solution method. Each solution-step we use the proximal operator to inch closer to the minima. However the first step our proximal operator does is projecting the value x onto the interval [a,b] (see functions.jl and New_Task_6.jl), this means that even if we only allow out method to do one iteration the "solution" would then lie within the interval because if the value is outside the interval it is projected onto the interval.

Task 7

> Solve the dual problem. Similar to the previous task, find/verify the upper bound on the step-size and find a good step-size choice. Compare the solutions from the primal and the one extracted from the dual, are they the same? Do they satisfy the constraint $x^* \in S$? Let y^k be the iterates of the dual method, using the expression from Task 4, extract the primal iterates $x\hat{}k$ from y^k. Does $x\hat{}k$ always satisfy the constraint $x\hat{}k \in S$? How does the function values, $f(x^k)$, develop over the iterations? What about $f(x\hat{}k) + \iota_S(x\hat{}k)$?

Solving the dual problem is done by solving the problem stated in equation 9. This is the same as solving problem 6, but we now have to use expressions for ∇f^* and $\text{prox}_{\gamma\iota_S^*}$ found in equations 16 and 21. As proved in equation 12, f^* is $\left\|Q^{-1}\right\|_2$-smooth (∇f^* is $\left\|Q^{-1}\right\|_2$-Lipschitz continuous) and as such the maximal step size to guarantee convergence is $\gamma < \frac{2}{L^*}, L^* = \left\|Q^{-1}\right\|_2$. Similarly to assignment 6, different step sizes were tried with results presented in the table below.

Step length γ:	$\frac{1}{5}\cdot\frac{2}{L^*}$	$\frac{2}{5}\cdot\frac{2}{L^*}$	$\frac{3}{5}\cdot\frac{2}{L^*}$	$\frac{4}{5}\cdot\frac{2}{L^*}$	$\frac{5}{5}\cdot\frac{2}{L^*}$	$\frac{6}{5}\cdot\frac{2}{L^*}$
Iterations:	> 150000	102172	68898	52098	41989	> 150000

Table 3: Shows how many iterations needed for each respective step length in order to achieve $||x^{k+1} - x^k||_2 < 10^{-15}$.

As one can observe, the number of iterations needed to convergence is far greater than in assignment 6. With increasing length in step size, the number of iterations gets smaller until the limit of $\frac{2}{L^*}$ is passed, where convergence is not guaranteed (and in this case probably does not happen).

This test was done with a maximum number of iterations being 150000. The fact that the smallest step size did not reach convergence does not mean that it will not converge given enough time, it just means that it is slow to do so.

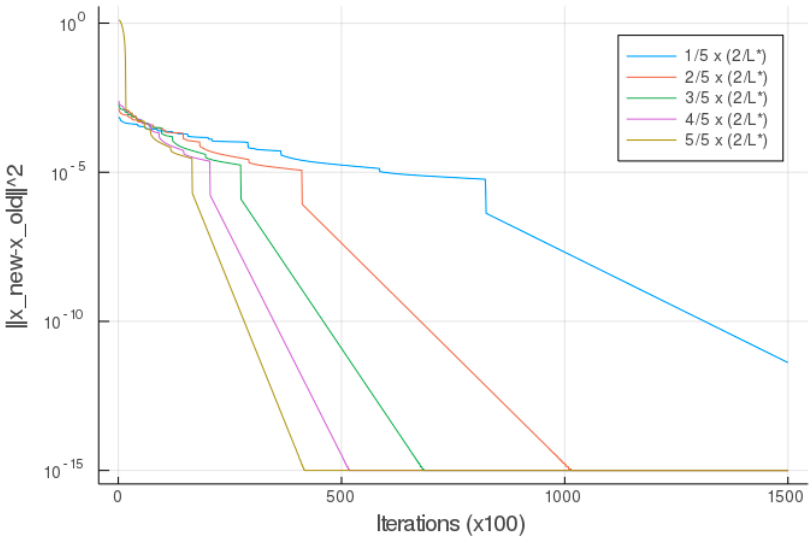

Figure 2: Shows the convergence rate for different step sizes ranging from $\frac{1}{5} \cdot \frac{2}{L^*}$ to $\frac{5}{5} \cdot \frac{2}{L^*}$

As seen in the figure, the shape of step size $\frac{1}{5} \cdot \frac{2}{L^*}$ is very similar to the other ones, which suggests convergence for that step size as well, only that it converge at a slower rate.

If we look at the convergence rate for a step size of $\frac{6}{5} \cdot \frac{2}{L^*}$ we can see that the solution skyrockets quite quickly and never converges, see figure 3.

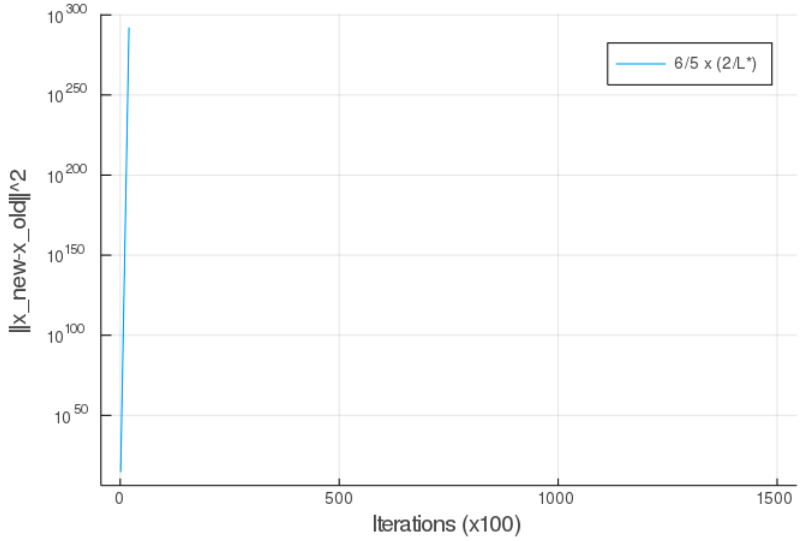

Figure 3: Shows the convergence rate for step size $\frac{6}{5} \cdot \frac{2}{L^*}$

We know from equation 24 that from the dual solution y^*, we can recover a solution $\hat{x}$. We are interested if this $\hat{x}$ is within the interval [a,b]. For a step size of $\frac{2}{L^*}$ and a fixed number of iterations 10^6 this was calculated and is shown in the following matrix with limits a and b:

$$\begin{bmatrix} a_i & \hat{x}_i & b_i \\ -0.128 & 0.336 & 0.733 \\ -0.148 & 0.35 & 0.691 \\ -0.874 & -0.126 & 0.528 \\ -0.486 & -0.022 & 0.683 \\ -0.496 & 0.034 & 0.922 \\ -0.641 & 0.19 & 0.71 \\ -0.375 & -0.05 & 0.751 \\ -0.034 & 0.264 & 0.31 \\ -0.893 & 0.081 & 0.835 \\ -0.886 & -0.423 & 0.378 \\ -0.398 & 0.431 & 0.431 \\ -0.989 & 0.261 & 0.457 \\ -0.274 & -0.127 & 0.689 \\ -0.783 & -0.383 & 0.877 \\ -0.995 & -0.826 & 0.12 \\ -0.661 & -0.194 & 0.612 \\ -0.29 & 0.165 & 0.165 \\ -0.958 & 0.028 & 0.089 \\ -0.55 & -0.103 & 0.598 \\ -0.274 & 0.246 & 0.718 \end{bmatrix}$$

Here, the digits are rounded to three decimal places. We can see that all values in $\hat{x}$ lies within the interval $[a,b]$. That being said, there is a possibility that some values will only be very close to the interval. This is because we are using $\left\|y^{k+1} - y^k\right\|_2 < 10^{-15}$ as an approximation of convergence. Thus, we could have a solution y that is almost optimal so that when recovering $\hat{x}$ we might end up with something that only almost lies in the interval. This is proven when looking closely on $\hat{x}_{17}$ with 13 decimal places:

$$\begin{bmatrix} a_{17} & \hat{x}_{17} & b_{17} \\ -0.290004394019 & 0.1654385857491 & 0.1654385857490 \end{bmatrix}$$

We can see that $\hat{x}_{17}$ lies slightly above the interval.

We can also look at $\hat{x}$ recovered from a snapshot of y after 100 iterations:

a_i	$\hat{x}$-snapshot: 100 iterations	b_i
−0.128	−0.955	0.733
−0.148	2.988	0.691
−0.874	−1.037	0.528
−0.486	−0.879	0.683
−0.496	−1.657	0.922
−0.641	0.249	0.71
−0.375	−0.651	0.751
−0.034	−1.462	0.31
−0.893	−0.159	0.835
−0.886	−0.678	0.378
−0.398	0.72	0.431
−0.989	−0.26	0.457
−0.274	−0.944	0.689
−0.783	1.336	0.877
−0.995	0.103	0.12
−0.661	−1.173	0.612
−0.29	1.431	0.165
−0.958	−0.061	0.089
−0.55	−0.571	0.598
−0.274	0.251	0.718

Clearly some points are outside the interval [a,b]. As we concluded in assignment 6, any snapshot from solving the primal problem will always be inside the interval. By definition, they will not be the same as these ones recovered from the dual problem.

With some of these values being large, especially in the first iterations, the function values [$f(\hat{x}^k)$] will be large as well (f is quadratic, will give large positive values for large $\hat{x}$-values).

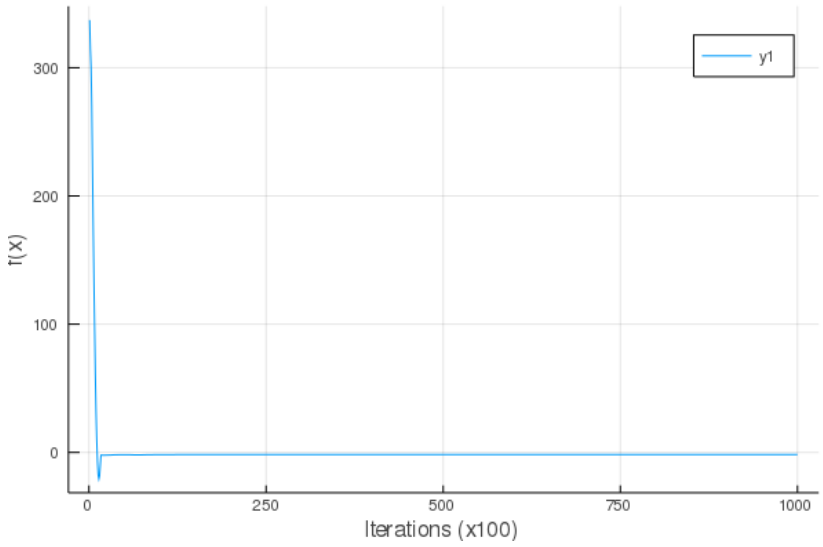

Figure 4: Shows the convergence of f(x) for the dual problem

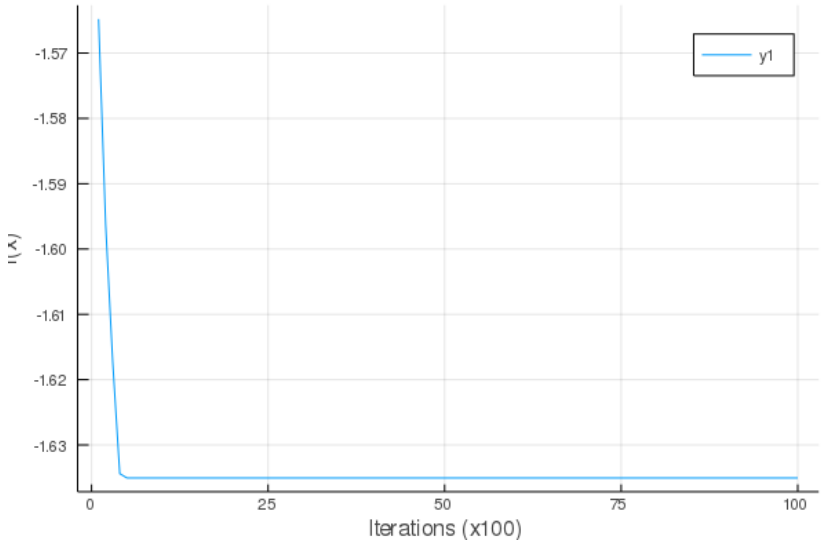

Figure 5: Shows the convergence of f(x) for the primal problem

$f(x)$ for the iterates of the dual problem in figure 4 shows the convergence of f. The slightly negative values that the function converges to are explained by the term $q^T x$ in equation 13. For the original problem, $\hat{x}$-values are sometimes outside the interval meaning that $[f(\hat{x}^k)] + \iota_S(\hat{x}^k)$ will have values at infinity.

We can also compare our recovered primal solution $\hat{x}$ with the primal solution x from assignment 6:

$$\begin{bmatrix} x & \hat{x} \\ 0.336 & 0.336 \\ 0.35 & 0.35 \\ -0.126 & -0.126 \\ -0.022 & -0.022 \\ 0.034 & 0.034 \\ -0.05 & -0.05 \\ 0.264 & 0.264 \\ 0.081 & 0.081 \\ -0.423 & -0.423 \\ 0.431 & 0.431 \\ 0.261 & 0.261 \\ -0.127 & -0.127 \\ -0.383 & -0.383 \\ -0.826 & -0.826 \\ -0.194 & -0.194 \\ 0.165 & 0.165 \\ 0.028 & 0.028 \\ -0.103 & -0.103 \\ 0.246 & 0.246 \end{bmatrix}.$$

Clearly, they appear identical. However, as already stated in assignment 6, we know that all points of x will lie in the interval, while we have proven $\hat{x}$ to only be very close. They are equal in theory but because of the limitations of numerical precision, they are only very similar in practice. We can show this by once again looking at the 17th decimal point:

$$\begin{bmatrix} x_{17} & \hat{x}_{17} \\ 0.1654385857490 & 0.1654385857491 \end{bmatrix}.$$

But for all intents and purposes the solutions are identical and this has to do with the same thing that was mentioned in task 6, namely that the problem we are solving is strongly convex. By going to the dual of the problem we are changing the structure of the problem but since the original problem only has one unique solution it would be illogical for the dual problem to produce a different solution. In such case there would be no point in solving the dual.