# Optimization for Learning - FRTN50

Assignment 1

November 1, 2020

**Task 1** Derive $f^*$ and $\iota_S^*$ and write down the Fenchel-dual problem.

The conjugate $f^*$ :

We have : $f(x) = \frac{1}{2}x^T Q x + q^T x\,, \quad Q \succ 0 \quad \forall x \in \mathbb{R}^n.$

By definition of the conjugate of $f$:

$$f^*(s) = \sup_x (s^T x - f(x))\,, \quad \forall s \in \mathbb{R}^n$$

- $f^*(s)$ defines an affine minorizer to $f$ with slope $s$.
- $f$ is $\lambda_{\min(Q)}$-strongly convex.

Then the majorizing point $x_0$ exists and is unique, i.e.: $\exists x_0 \in \mathbb{R}^n$ such that $f^*(s) = s^T x_0 - f(x_0)$.

Due to Fenchel Young's equality:

$$s \in \partial f(x_0)$$

Since $f$ is differentiable : $s = \nabla f(x_0) = Qx_0 + q \implies \mathrm{x}_0 = Q^{-1}(s - q)$ (Q is invertible).

Hence :

$$\boxed{f^*(s) = \frac{1}{2}(s-q)^T Q^{-1}(s-q)\,, \quad \forall s \in \mathbb{R}^n}$$

The conjugate $\iota_S^*$ :

By definition: $\iota_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{else} \end{cases}, \quad x = (x_1, ..., , x_n)^T$

Then: $\iota_S(x) = \sum_{i=1}^{n} \iota_{[a_i,b_i]}(x_i)$, i.e. $i_S$ is separable.

Hence:

$$\iota_S^*(s) = \sum_{i=1}^{n} \iota_{[a_i,b_i]}^*(s_i)\,, \quad s = (s_1, ..., , s_n)^T \in \mathbb{R}^n$$

Now take $i \in \{1, ..., n\}$ and $t \in \mathbb{R}$, we have:

$$\iota^*_{[a_i,b_i]}(t) = \sup_{r\in\mathbb{R}} (rt - \iota_{[a_i,b_i]}(r)) = \sup_{r\in[a_i,b_i]} (rt)$$

- For $t \leq 0$, an optimal $r = a_i$ and $\iota^*_{[a_i,b_i]}(t) = a_i t$.
- For $t \geq 0$, an optimal $r = b_i$ and $\iota^*_{[a_i,b_i]}(t) = b_i t$.

$$i.e. \quad \iota^*_{[a_i,b_i]}(t) = \max(a_i t, b_i t)$$

Therefore,

$$\boxed{\iota^*_S(s) = \sum_{i=1}^{n} \max(s_i a_i, s_i b_i) \quad \forall s \in \mathbb{R}^n}$$

Then the Fenchel-dual problem is as follows:

$$\boxed{\min_{\mu} (f^*(-\mu) + \iota^*_S(\mu)) = \min_{\mu} \left( \frac{1}{2}(-\mu - q)^T Q^{-1}(-\mu - q) + \sum_{i=1}^{n} \max(\mu_i a_i, \mu_i b_i) \right)}$$

**Task 2** Show that $f$ and $f^*$ are $L$-, and $L^*$ -smooth respectively. Find $L$ and $L^*$.

We know that $f$ and $f^*$ are twice differentiable with:

$$\nabla^2 f(x) = Q \quad , \quad \nabla^2 f^*(x) = Q^{-1} \qquad \forall x \in \mathbb{R}^n$$

Since $Q \succ 0$, then:
$\forall i \in \{1, .., n\} \quad \lambda_i > 0 \quad$ where $\lambda_i$ are the eigenvalues of the matrix $Q$.
Hence:

$$\boxed{\forall x \in \mathbb{R}^n \quad 0 \leq \nabla^2 f(x) \leq \lambda_{\max}(Q) I_n} \quad \text{with} \quad \lambda_{\max}(Q) = \max(\lambda_1, ...., \lambda_n)$$

Therefore, $f$ is $\lambda_{\max}(Q)$-smooth convex function. i.e. $L = \lambda_{\max}(Q)$

As a result:
$\forall i \in \{1, .., n\} \quad \dfrac{1}{\lambda_i}$ are the eigenvalues of the matrix $Q^{-1}$.

Then:

$$\boxed{\forall x \in \mathbb{R}^n \quad 0 \leq \nabla^2 f^*(x) \leq \frac{1}{\lambda_{\min}(Q)} I_n} \quad \text{with} \quad \lambda_{\min}(Q) = \min(\lambda_1, ...., \lambda_n)$$

Therefore, $f^*$ is $\dfrac{1}{\lambda_{\min}(Q)}$-smooth convex function. i.e. $L^* = \dfrac{1}{\lambda_{\min}(Q)}$

**Task 3** Derive expressions for $\nabla f$, $\nabla f^*$, $\underset{\gamma\iota_S}{\text{prox}}$ and $\underset{\gamma\iota_S^*}{\text{prox}}$.

$\underline{\nabla f,\ \nabla f^* :}$

$f$ and $f^*$ are differentiable then:

$$x \in \mathbb{R}^n \quad \begin{cases} \nabla f(x) = Qx + q \\ \nabla f^*(x) = Q^{-1}(x - q) \end{cases}$$

$\underline{prox_{\gamma\iota_S} :}$

By definition:

$$\underset{\gamma\iota_S}{\text{prox}}(x) = (I + \gamma\partial\iota_S)^{-1}(x)$$

And since $i_S$ is separable i.e. $\iota_S(x) = \sum_{i=1}^{n} \iota_{[a_i,b_i]}(x_i)$.

Then:

$$\boxed{\underset{\gamma\iota_S}{\text{prox}}(x) = \begin{pmatrix} \underset{\gamma\iota_{[a_1,b_1]}}{\text{prox}}(x_1) \\ \underset{\gamma\iota_{[a_2,b_2]}}{\text{prox}}(x_2) \\ \vdots \\ \underset{\gamma\iota_{[an,bn]}}{\text{prox}}(x_n) \end{pmatrix} \in \mathbb{R}^n}$$

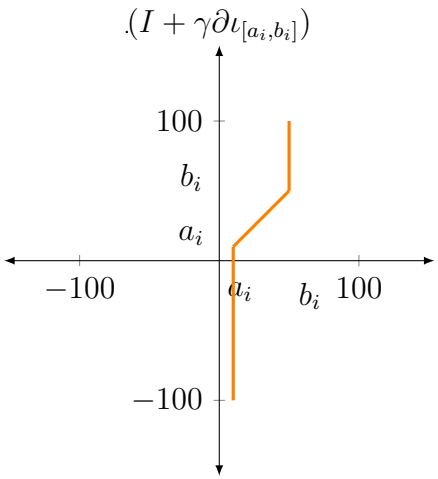
And $\forall i \in \{1, .., n\}$ $\underset{\gamma\iota_{[a_i,b_i]}}{\text{prox}}(x_i) = (I + \gamma\partial\iota_{[a_i,b_i]})^{-1}(x_i)$

We have:

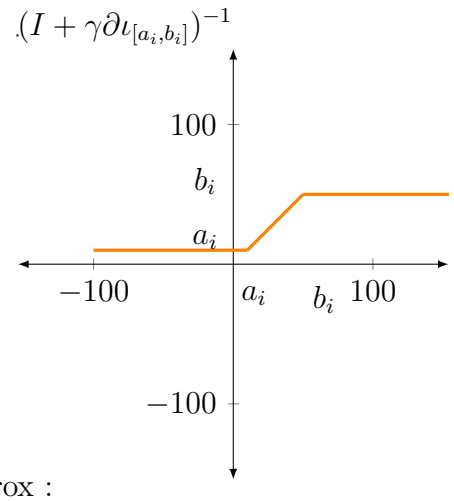
$$\partial_{\iota_{[a_i,b_i]}}(x_i) = \begin{cases} ]-\infty,\ 0] & \text{if } x_i = a_i \\ 0 & \text{if } x_i \in ]a_i,\ b_i[ \\ [0,\ +\infty] & \text{if } x_i = b_i \end{cases}$$

Then:

$$(I+\gamma\partial\iota_{[a_i,b_i]})(x_i) = \begin{cases} ]-\infty,\ a_i] & \text{if } x_i = a_i \\ x_i & \text{if } x_i \in ]a_i,\ b_i[ \\ [b_i,\ +\infty] & \text{if } x_i = b_i \end{cases}$$




We flip the figure and we get:




$$\operatorname{prox}_{\gamma\iota_{[a_i,b_i]}}(x_i) = \begin{cases} a_i & \text{if } x_i \le a_i \\ x_i & \text{if } x_i \in [a_i,\ b_i]\ , \forall i \in \{1,..,n\} \\ b_i & \text{if } x_i \ge b_i \end{cases}$$

prox : $\underline{\gamma\iota_S^*}$

By definition:

$$\operatorname*{prox}_{\gamma\iota_S^*}(x) = (I + \gamma\partial\iota_S^*)^{-1}(x)$$

And since $i_S^*$ is separable i.e. $\iota_S^*(s) = \sum_{i=1}^{n} \max(s_i a_i, s_i b_i) \quad \forall s \in \mathbb{R}^n$.

Then:

$$\operatorname*{prox}_{\gamma\iota_S^*}(x) = \begin{pmatrix} \operatorname{prox}_{\gamma\iota^*_{[a_1,b_1]}}(x_1) \\ \operatorname{prox}_{\gamma\iota^*_{[a_2,b_2]}}(x_2) \\ \vdots \\ \operatorname{prox}_{\gamma\iota^*_{[a_n,b_n]}}(x_n) \end{pmatrix} \in \mathbb{R}^n$$

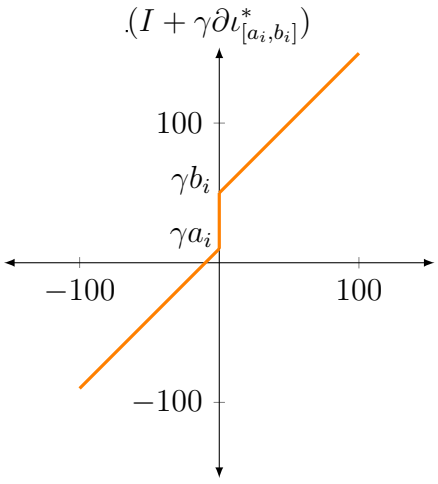
And $\forall i \in \{1,..,n\}$ $\operatorname{prox}_{\gamma\iota^*_{[a_i,b_i]}}(x_i) = (I + \gamma\partial\iota^*_{[a_i,b_i]})^{-1}(x_i)$

We have:

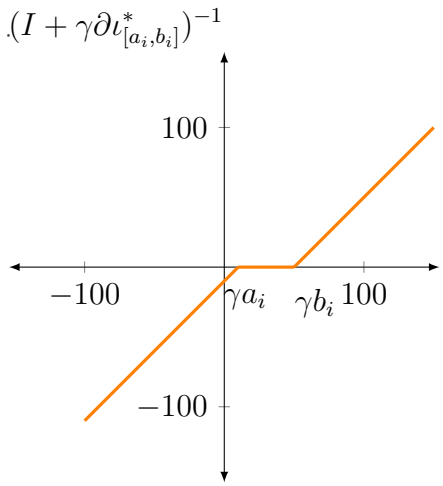
$$\iota^*_{[a_i,b_i]}(x_i) = \max(a_i x_i, b_i x_i) \Longrightarrow \quad \partial_{\iota^*_{[a_i,b_i]}}(x_i) = \begin{cases} a_i & \text{if } x_i < 0 \\ [a_i,\ b_i] & \text{if } x_i = 0 \\ b_i & \text{if } x_i > 0 \end{cases}$$

Then:

$$(I + \gamma\partial\iota^*_{[a_i,b_i]})(x_i) = \begin{cases} x_i + \gamma a_i & \text{if } x_i < 0 \\ [\gamma a_i,\ \gamma b_i] & \text{if } x_i = 0 \\ x_i + \gamma b_i & \text{if } x_i > 0 \end{cases}$$




We flip the figure and we get:




$$\forall i \in \{1, .., n\}$$

$$\operatorname{prox}_{\gamma\iota^*_{[a_i,b_i]}}(x_i) = \begin{cases} x_i - \gamma a_i & \text{if } x_i \leq \gamma a_i \\ 0 & \text{if } x_i \in [\gamma a_i,\ \gamma b_i] \\ x_i - \gamma b_i & \text{if } x_i \geq \gamma b_i \end{cases}$$

**Task 4** Let $y^*$ be a solution to the dual problem, derive an expression that gives a solution to the primal problem given $y^*$.

Since the functions $f$ and $\iota_S$ are convex and the primal and the dual constraint qualifications hold, we can recover a primal solution from the primal-dual optimality condition, that satisfies

$$x^* \in \partial f^*(-y^*) = \{\nabla f^*(-y^*)\}$$
$$x^* \in \partial g^*(y^*)$$

Since $f$ is differentiable and the problem is strongly convex, then the solution $x^*$ is unique.

$$\boxed{x^* = \nabla f^*(-y^*) = Q^{-1}(-y^* - q)}$$

Hence:

$$\boxed{\min_{x \in S} f(x) = f(x^*) = \frac{1}{2}(-y^* + q)^T Q^{-1}(-y^* - q)}$$

## Task 6

### 1 ) Range of different step-sizes:

In order to understand how the condition $\gamma < \frac{2}{L}$ is the best range of step-sizes, we create a function that takes parameters $l$ and $n$ in argument, randomly chooses an initial point $x_0$ and runs the proximal gradient method over $n$ iterations for the following step-sizes:

$$\gamma \in \{\gamma_k = k\frac{\gamma_m}{l} \quad / \gamma_m = \frac{2}{L}, \quad k \in \{1, .., 2l-1\}\}$$

Then plot the Norm of the step-length/residual: $||x^{k+1} - x^k||$ for $k \in \{1, .., n\}$ of each step.

Take an example of $l = 6$ and $n = 1400$:

```
julia> range_step_sizes(6, 1400)
```

- 5 steps w.r.t the upper bound ( $\forall i \in \{1, .., 5\} \quad \lambda_i < \frac{2}{L}$ )
- $\lambda_6 = \frac{2}{L}$
- 5 steps ( $\forall i \in \{7, .., 11\} \quad \lambda_i > \frac{2}{L}$ )

Of course we use the same initial point $x_0$ for each step-size, then we plot the results using the log scale on the y-axis.

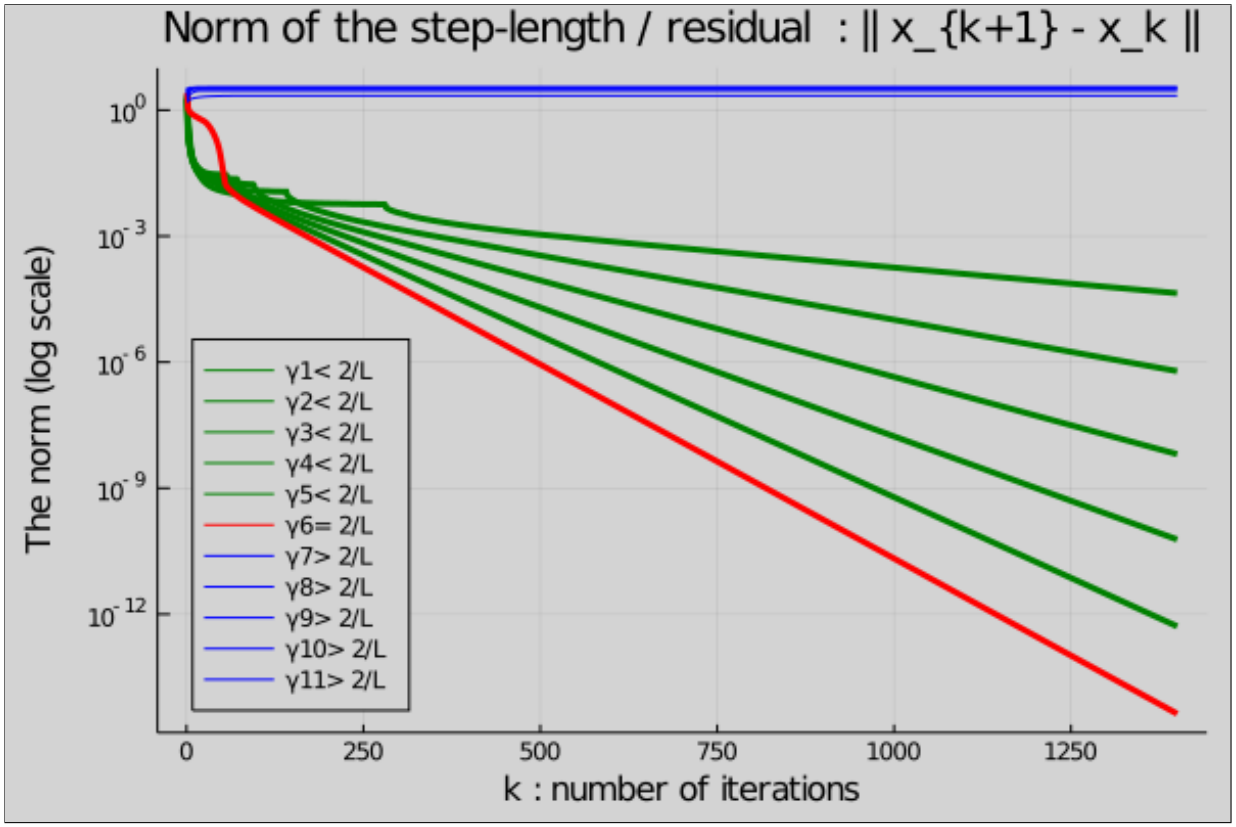



Figure 1: **The norm of the step-length/residual using a log scale on the y-axis**

**Result:**

As we can see, the convergence of the norm has the same behaviour over iterations, that means the norm of all step sizes $\gamma < \frac{2}{L}$ (The green labels) still decreasing over iterations to very low values, and the norm of all step sizes $\gamma > \frac{2}{L}$ (The blue labels) converges to different non zero values $\geq 1$. It is clear that using the step-size $\gamma \leq 2/L$ for large number of iterations will push the norm to low values around $10^{-15}$, which is low enough to say that the norm converges to zero.

$$\text{Then the upper bound } \gamma < \frac{2}{L} \text{ seems reasonable.}$$

## 2 ) Different initial points

Our goal in this part is to show that for any initial point the solution will be the same. So we create a function that takes three parameters $k$, $n$ and $\epsilon$ and:

- chooses the step-size $\gamma = \frac{2}{L} - \epsilon$ which verifies the reasonable condition $\gamma < \frac{2}{L}$.
- computes a landmark solution $x_f$ over $n$ iterations.
- for each $i \in \{1, .., k\}$:
  - randomly creates an initial point $x_{0_i}$
  - computes its final solution $x_i^*$ over $n$ iterations.
  - calculates the distance between this solution and the landmark solution: $d_i = ||x_i^* - x_f||$

Take an example of $k = 15$, $n = 3000$ and $\epsilon = 0.005$:

```
julia> initial_points(15, 3000, 0.005)
```

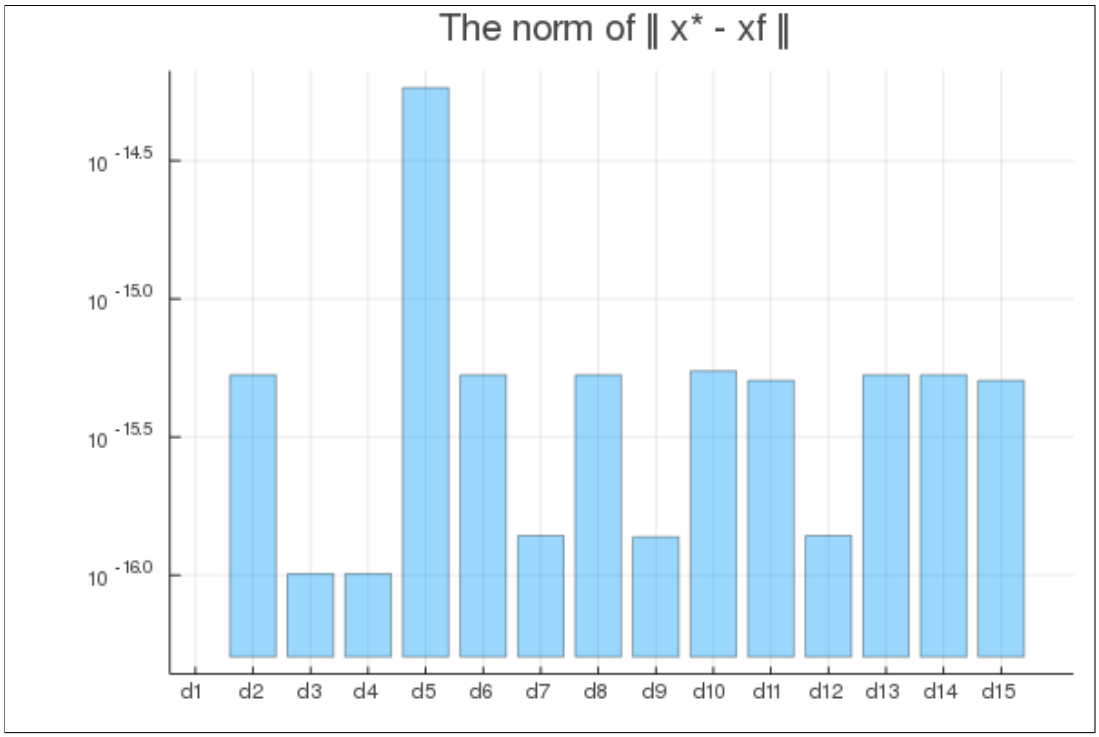



Figure 2: **Comparing the distances between solutions of 15 random initial points**

As you can see, we use the barplot in order to compare the distances, and we find that the max distance is around 15 orders of magnitude, corresponding to machine precision.

**Explanation:**
The fact that the initial points does not affect the final solution is due to the unique solution of the primal problem, since the function is strongly convex.

**3 ) $x^* \in S$ ? $x^k \in S$ ? and why?**

We have $\forall k \in \mathbb{N}$:

$$x^{k+1} = \underset{\gamma\iota_S}{\text{prox}}(x^k - \gamma\nabla\phi(x^k)) \in S \quad \text{Due to the definition of the } \underset{\gamma\iota_S}{\text{prox}} \text{ (see result Task 3).}$$

Then:

$$x^* \in S$$

## Task 7

Recall : the dual problem is:

$$\boxed{\min_{\mu} \left(f^*(-\mu) + \iota_S^*(\mu)\right)}$$

We know that:

- $f^*$ is differentiable and strongly convex.
- $\iota_S^*$ is proximable.

Now the dual iterates follow the same process:

$$\mu^{k+1} = \underset{\gamma\iota_S^*}{\text{prox}}(\mu^k + \gamma\nabla f^*(-\mu^k))$$

**1 ) range of different step-size**

We create a function simalar to which in Task 6) when we try the following range of step-sizes:

$$\gamma \in \{\gamma_k = k\frac{\gamma_m}{l} \quad / \gamma_m = \frac{2}{L*}, \quad k \in \{1, .., 2l-1\}\}$$

Then plot the Norm of the step-length/residual: $||\mu^{k+1} - \mu^k||$ for $k \in \{1, .., n\}$ of each step.

Take an example of $l = 6$ and $n = 1400$:

```
julia> range_dual_step_size(6, 20000)
```

**Result:**
In order to visualize the norm of all step sizes, we plot separately:

the norm of each $\gamma \leq \frac{2}{L*}$ which is obviously decreasing to low values ( around $10^{-15}$ ), but it requires a high number of iterations ($n \geq 20000$ ).

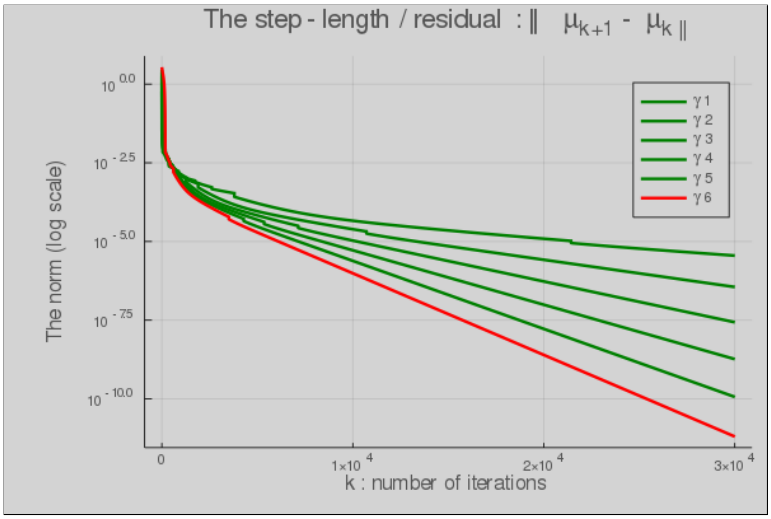
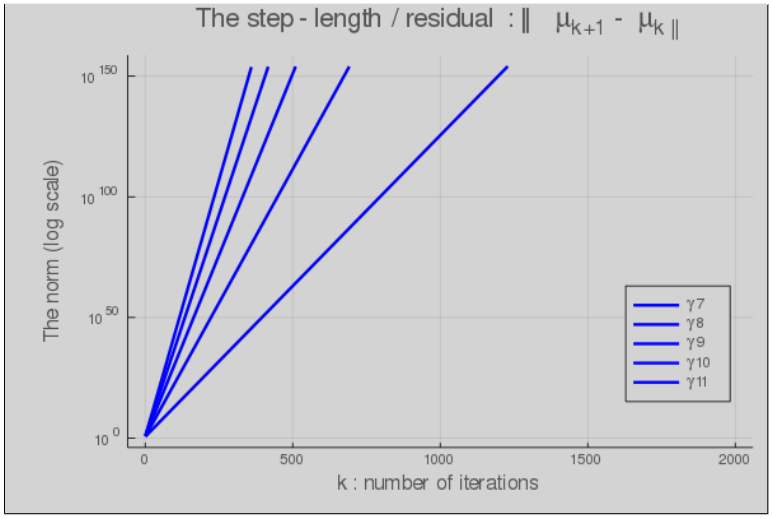



Figure 3: **step sizes** $\gamma \leq \frac{2}{L*}$




the norm of all step sizes $\gamma > \frac{2}{L*}$ which converges to higher values ( around $10^{150}$ ) just after hundreds of iterations.

Figure 4: **step sizes** $\gamma > \frac{2}{L*}$

As a result, we respect this upper bound and we choose as the best step-size $\gamma = \frac{2}{L} - \epsilon$ with $\epsilon$ a small quantity.

## 2 ) Compare the solutions from the primal and the one extracted from the dual

In this part we will run mutiple realizations and compute the norm of the distance $||x^* - \hat{x}^*||$ between the primal solution $x^*$ and the one extracted from the dual $\hat{x}^*$.
The function **dual_primal(k, n,** $\epsilon$**)** does this task for $k$ realizations over $n$ iterations, and of course $\epsilon$ is chosen such that $\gamma = \frac{2}{L^*} - \epsilon$. e.g. **dual_primal(10, 90000, 0.005)**:

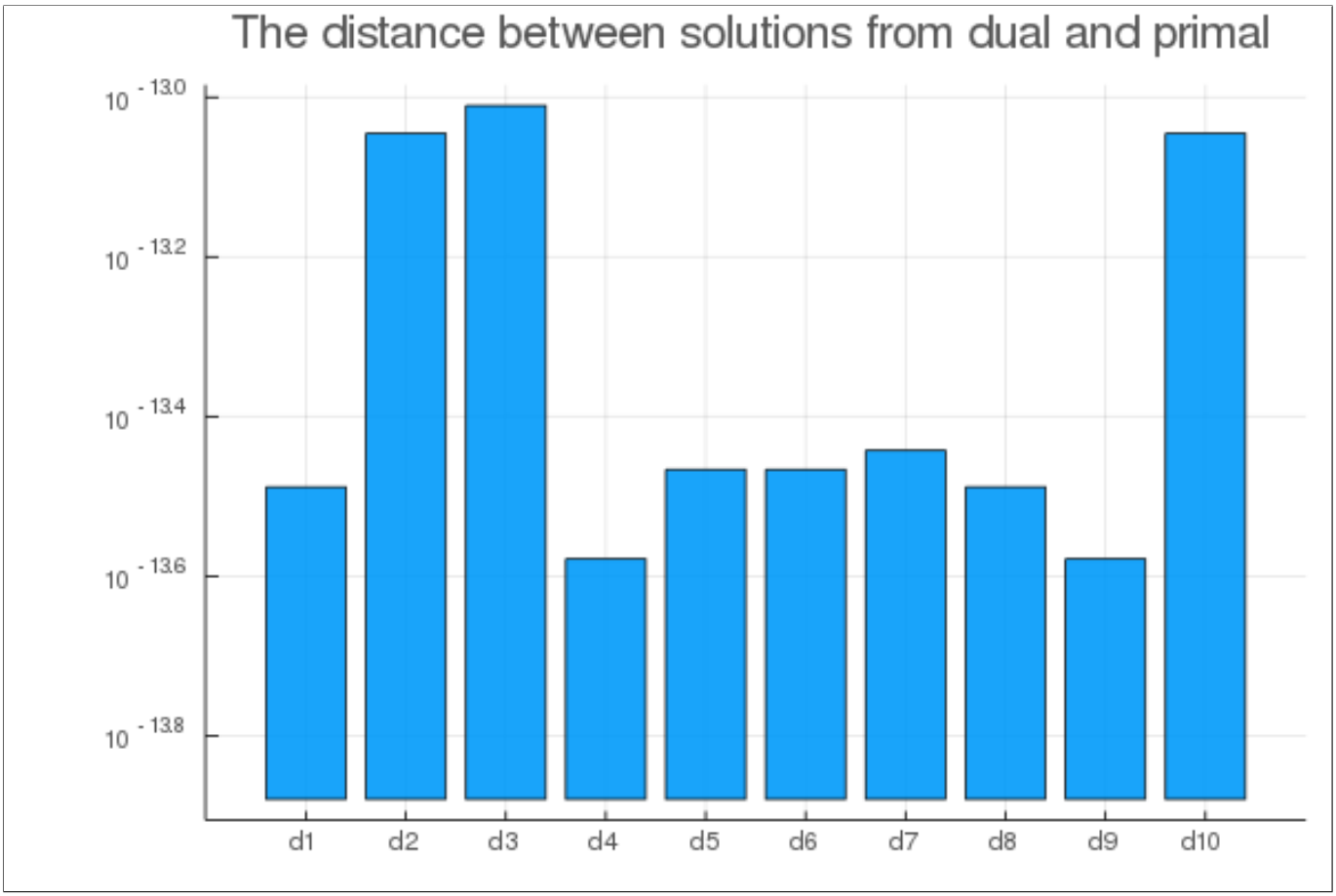



Figure 5: **BarPlot of distances for 10 random realizations**

From the **Figure 5**, we can say that the primary solution is so close to the one extracted from the dual problem since the difference is around $10^{-13}$. But this cannot prove that the extracted solution belongs to $S$ i.e. We can guarantee that it is arbitrary close, but not that it reaches the optimum or that it is in the set $S$.

**2 ) $x^k \in S$? How does the function values develop over the iterations?**

In order to answer this question we will plot over iterations the function values of:

- $f(\hat{x}^k)$ of the extracted iterates.
- $f(x^k)$ of the primal iterates.
- $f(\hat{x}^k) + \iota_S(\hat{x}^k)$ of the extracted iterates.

  And here we replace inf by another number so as to visualize if $x^k \in S$ or not.

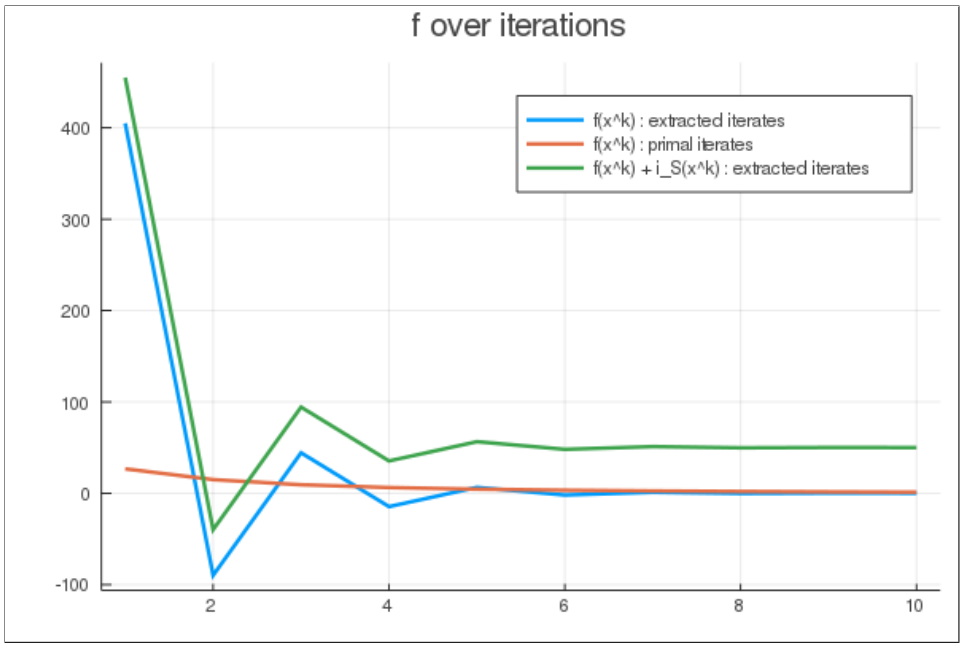



Figure 6: **Values over 10 iterations**

It is obvious that the extracted iterates $\hat{x}^k$ don't all belong to $S$, since the green graph is the translated graph of the blue one.

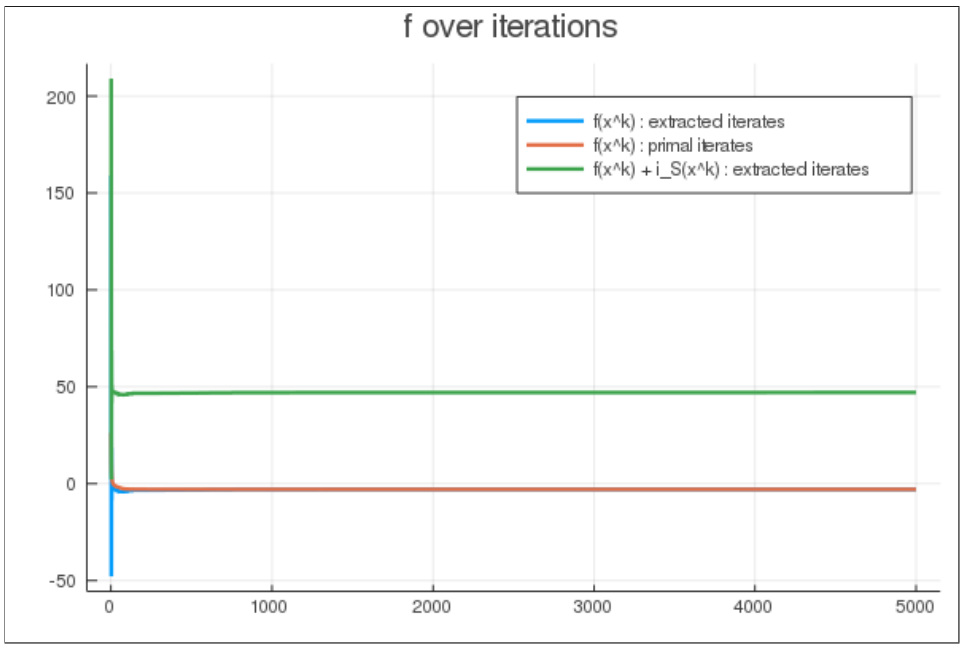



Figure 7: **Values over 5000 iterations**

As we can notice the values of $f(\hat{x}^k) + \iota(\hat{x}^k)$ are equal translated by 50, which proves that $f(\hat{x}^k)$ converges to a fixed value (by having an oscillating behavior).
**Remark:** We notice smaller function values at the beginning, but these values are not the minimum of the problem since their $\hat{x}^k$ don't belong to the set $S$.