Faculty of Applied Sciences
University of Liege
February 20, 2019

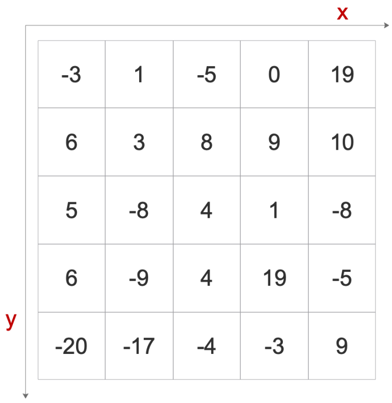
-

Optimal decision making for complex problems:

Assignment 1 - Section 4 to 5

s141770
Derroitte Natan

Continuing this project, the axis system remains a key choice that will influence each of the algorithms implemented. Therefore, it is recalled at the beginning of this report on the second part of the project. 1.



-3	1	-5	0	19
6	3	8	9	10
5	-8	4	1	-8
6	-9	4	19	-5
-20	-17	-4	-3	9

Figure 1: Representation of the axis system

4 Optimal Policy

The purpose of this fourth question was to establish the optimal policy and deduce $J^N_{\mu^*}$. Therefore, a Markov Decision Process (MDP) had to be introduced and its associated transition matrix $p(x'|x,u)$ and reward signal $r(x,u)$ had to be calculated.

Using the theoretical formulas seen during the course, these results, as well as $Q_N(x,u)$ can be directly obtained. Knowing that for a sufficiently high number of iterations (N), $Q_N(x,u)$ tends towards $Q(x,u)$ ($\lim_{N\to\infty} ||Q_N(x,u) - Q(x,u)||_\infty$), it is possible to deduce the optimal policy from these results:

$$\mu^*(x) \in argmaxQ_{u\in U}(x,u)$$

The number of iterations used in the following calculations is 1000. It had indeed been observed during the previous questions that this value of N reduced the error below 0.082 for our grid ($Br = 19$) and our value of γ.
It is now possible to determine the cumulative reward signal using this optimal policy. The results are displayed when the script is executed.
It is also possible to format $J^N_{\mu^*}$ as follows:

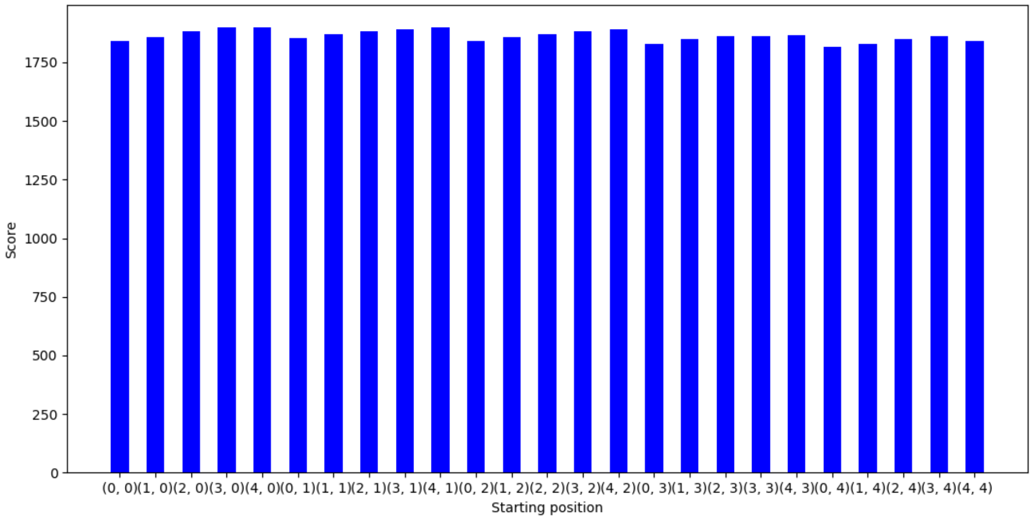


Figure 2: Cumulative reward signal using the computed optimal policy for $= 0$ and $N = 1000$.

These first results correspond to the deterministic case. The influence of the β parameter on $J^N_{\mu^*}$ will be studied in sub-section 4.2.
It is interesting to note that the optimal starting position is (4, 0), *ie* the upper right corner.
As it will be explained in section 4.1, this cell has the highest reward and is located against a wall, allowing the agent to stay there indefinitely. This explains why it is ranked as the best starting position.

4.1 Discussion on the Optimal Policy

The figure 3 shows the optimal policy for $N = 1$ and for $N = 100$.
In a first case, when N =1, the agent has only one action to do. It therefore seems natural that the optimal policy is simply to move to the cell that optimises the immediate reward. In other words, the agent takes the action that has the greatest direct reward.
In the second case, the agent has 100 moves to do. It can therefore afford to take a smaller immediate reward (even negative) if this choice maximises its cumulative reward signal.

It should be noted that when $N = 100$, the policy seems to tend towards moving to the upper right corner. This makes perfect sense! This cell contains the greatest reward and allows the agent to stay there forever since it is next to a wall. It will therefore be favoured by the agent.

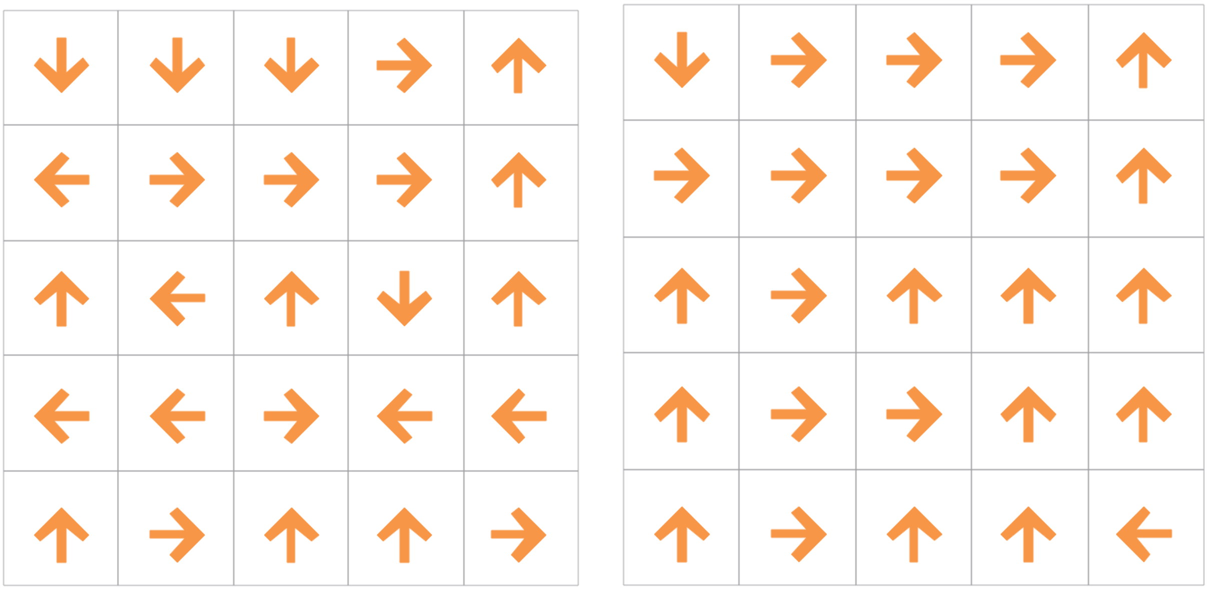


Figure 3: Optimal policy for $\beta = 0$. On the left : $N = 1$ and $N = 100$ on the right.

4.2 Impact of β

In this subsection, the influence of the β parameter on the $J_{\mu^*}^N$ score will be studied. The figure 4 will support these discussions.

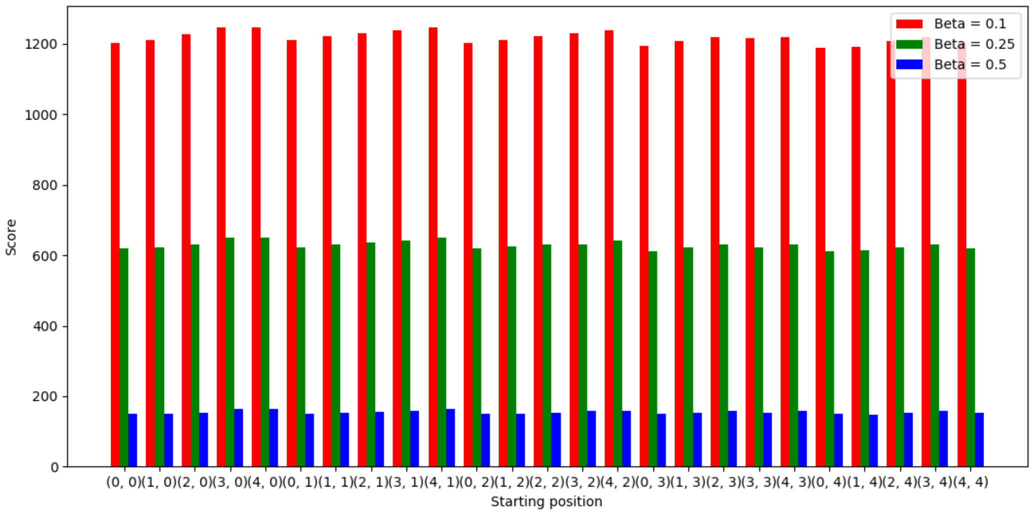


Figure 4: Comparison of the cumulative reward signal $J_{\mu^*}^N$ for different value of β. The number of iterations is 1000.

It appears that the closer the β parameter is to 1, the more the results decrease. Indeed, $J_{\mu^*}^N$ decreases with the value of β, quickly taking negative values for $\beta > 0.5$. It can be concluded that, in this environment, the optimal β is thus 0.
This is explained by the presence of a negative reward in (0, 0). In almost all cases, all squares have a

action with a better reward.

The minimum value of the cumulative reward signal corresponds to the limit of -300 for $\beta = 1$.
This value of -300 is not insignificant: it corresponds to the iteration where $0.99J^{N-1} = -297$. The value of the reward being -3, it appears that $J_{\mu^*}^N(x) = -3 + 0.99J_{\mu^*}^{N-1}(x) = -300 \; \forall x \in X$.

Using the same logic, it is possible to calculate the maximum value of the cumulative reward signal of this grid.
In the case where $\beta = 0$, it was previously concluded that the agent would remain indefinitely in (4, 0), where the reward is 19. One just has to solve:

$$J_{\mu^*}^N\left((4,0)\right) = 19 + \gamma J_{\mu^*}^{N-1}\left((4,0)\right)$$

with $J_{\mu^*}^N\left((4,0)\right) = J_{\mu^*}^{N-1}\left((4,0)\right)$, it appears that $J_{\mu^*}^N\left((4,0)\right) = 1900$.

5 System Identification

For this question, it was asked to estimate the $p(x'|x,u)$ transition matrix and the $r(x,u)$ reward function of a MDP from a $h_t = x_0u_0r_0x_1...x_t$ trajectory.

This trajectory was randomly generated from random starting position and using legal actions. From the outset, it cannot be missed that the size of this trajectory has a crucial influence on the results and will be detailed in the subsection 5.2.
This trajectory allows to compute $\hat{p}(x'|x,u)$ and $\hat{r}(x,u)$, the estimations of $p(x'|x,u)$ and $r(x,u)$.

It is interesting to specify how the rewards and the transitions not previously encountered in the history are initiated.
Regarding the reward, a neutral reward of 0 is then considered.
For the transition probabilities, it is possible to base ourselves on our a priori knowledge of the agent, omitting the environment. Since the agent can only move in the cells around him, or in (0, 0) in the stochastic case, it can be deduced that the majority of transitions will therefore be impossible. The unknown transitions were therefore set to 0.
It should be noted that other values to instantiate these unknown transitions could have been considered, such as $\frac{1}{nb\_cells}$, considering that each box is equiprobable.

Once $\hat{r}(x,u)$ and $\hat{p}(x'|x,u)$ have been calculated, determining $\hat{Q}(x,u)$, $\hat{\mu}^*$ and $J_{\hat{\mu}^*}^N$ simply consists of performing the same manipulations as for the previous question. The procedures will therefore not be recalled.

The results are displayed when the script corresponding to the question is executed and present in figure 5. This is done for a history size equal to 1000 ($t = 1000$).
Again, the deterministic case is first studied and the influence of β will be discussed in a future subsection

Although the results seem to correspond perfectly to the unestimated value presented in the previous section, it is interesting to introduce more accurate error measurements.

The error made when approximating $r(x,u)$ can simply be computed by calculating an absolute mean deviation between $\hat{r}(x,u)$ and $r(x,u) \forall x \in X, u \in U$.
The same strategy was used to estimate the error between $J_{\hat{\mu}^*}^N(x)$ and $J_{\mu^*}^N(x) \forall x \in X$.
Regarding $p(x'|x,u)$, this simple error measurement is not optimal. Having initiated the unknown transitions at 0 (as explained above), many transitions will be correct and the average error will be very low. An error measure punishing more discrepancies is the root mean square error which will therefore be used to estimate the error on $p(x'|x,u) \forall x, x' \in X, u \in U$.

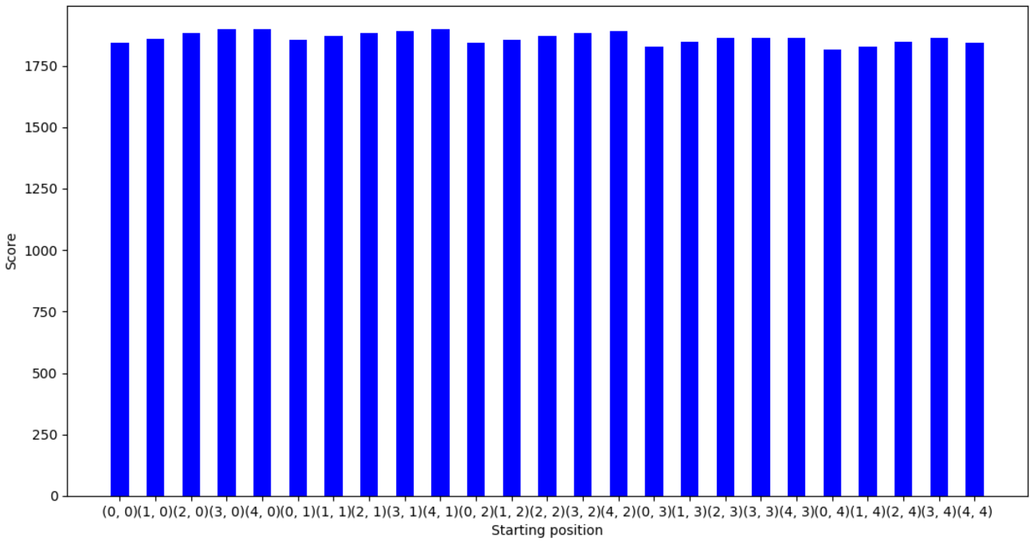


Figure 5: Cumulative reward signal using the estimated optimal policy for $\beta = 0$ and $N = 1000$.

The quality of the results presented in Figure 5 can now be attested:

- Mean absolute deviation on the reward signal : 0.0.
- Root mean square deviation on the transition probabilities : 0.0.
- Absolute mean deviation on $J^N_{\mu^*}$: 0.0.

It is now clear that the estimate made for $t = 1000$ corresponds perfectly to the unestimated results.

5.1 Impact of β

For a high value of t, it has been seen that the estimations made in this question are similar to the results of the previous question. The influence of the β parameter in these cases is therefore similar to what was developed in section 4.2.

When the trajectory is small, however, β finds interest through exploration. Indeed, when h_t is small, the agent, with only a few actions, often finds himself staying in the same area of the grid in the deterministic case. The possibility of going to (0, 0) therefore allows the agent to find himself in an area that may be far from what it already knows.
However, it is important to note that this interest only exists for low values of β. Otherwise, the agent will be blocked in the (0, 0) zone, which will hinder the agent's exploration.

5.2 Impact of the size of the trajectory

In this section, the convergence of estimations made with respect to the size of the trajectory will be studied.

A first comparison can be made using the figure 6. In this one, it is possible to observe the evolution of $J^N_{\hat{\mu}^*}(x) \forall x \in X$ for different values of t.
An expected result immediately appears: the smaller the size of the trajectory, the more the values of the cumulative reward signal seem to deviate from the true value.

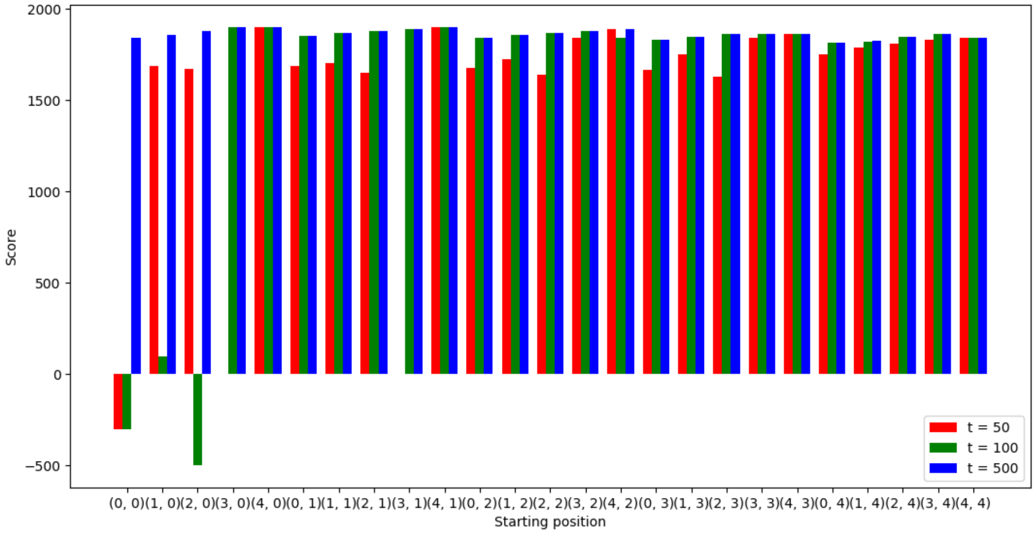
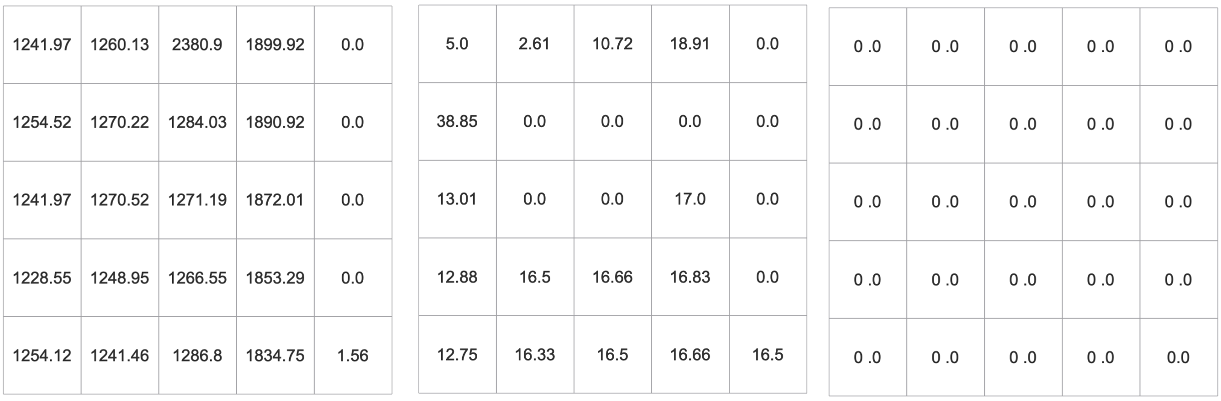


Figure 6: Comparison of the cumulative reward signal $J^N_{\hat{\mu}^*}$ for different size of trajectories(t). The number of iterations is 1000 and $\beta = 0$.

A different way to highlight this conclusion is shown in the figure 7. In this one, it can be observed the absolute mean error made on $J^N_{\hat{\mu}^*}$.[1]
The error is then revealed to be less and less large as a function of t. It can also be observed that for $t = 50$, the estimate was perfect for the right column. This is simply implied: the agent had to do most of his 50 actions in this column, thus getting a good knowledge of it.



1241.97	1260.13	2380.9	1899.92	0.0
1254.52	1270.22	1284.03	1890.92	0.0
1241.97	1270.52	1271.19	1872.01	0.0
1228.55	1248.95	1266.55	1853.29	0.0
1254.12	1241.46	1286.8	1834.75	1.56

5.0	2.61	10.72	18.91	0.0
38.85	0.0	0.0	0.0	0.0
13.01	0.0	0.0	17.0	0.0
12.88	16.5	16.66	16.83	0.0
12.75	16.33	16.5	16.66	16.5

0 .0	0 .0	0 .0	0 .0	0 .0
0 .0	0 .0	0 .0	0 .0	0 .0
0 .0	0 .0	0 .0	0 .0	0 .0
0 .0	0 .0	0 .0	0 .0	0 .0
0 .0	0 .0	0 .0	0 .0	0.0

Figure 7: Absolute mean error on $J^N_{\hat{\mu}^*}$ for different size of trajectories(t). The number of iterations is 1000 and $\beta = 0$. From left to right : $t = 50$, $t = 100$, $t = 500$

In order to further deepen the analysis on the influence of t, it is possible to study the convergence of estimators with respect to t. To do this, the figures 8 show the evolution of the error parameters previously defined as a function of t.

[1] The figures 6 and 7 were generated for different trajectories.

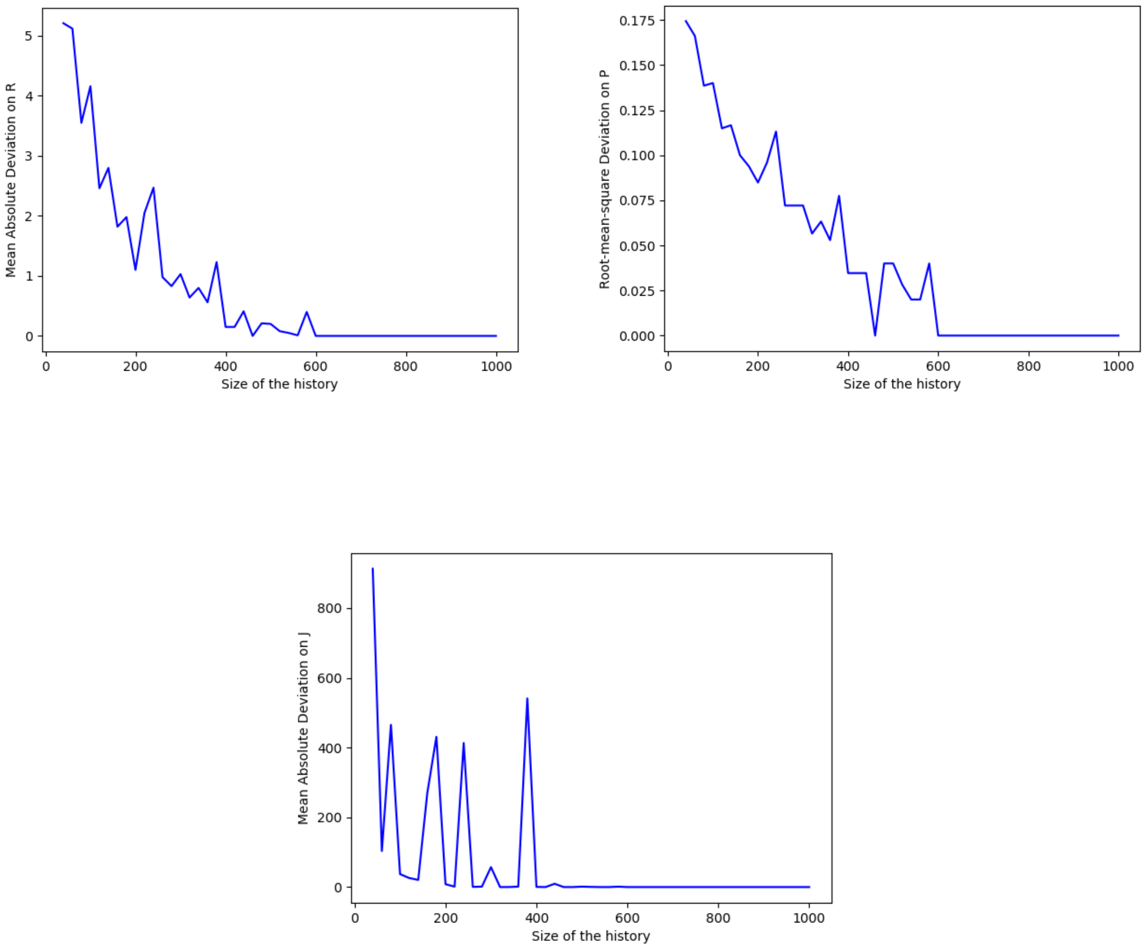


Figure 8: Evolution of the error on $r(x, u)$, $p(x'|x, u)$ and $J^N_{\mu*}$ with respect to t.

The convergence speed is given by the gradient of the curves.
From these results, it appears that the error measurements tend towards 0 when the size of the trajectory exceeds 600. However, this value remains strongly linked to our execution.