Assignment 3

STAT 497-H | Reinforcement Learning

Matteo Esposito (40024121), William Ngo (40031586), Spyros Orfanos (40032280)

Concordia University

April 14th, 2019

Question 1

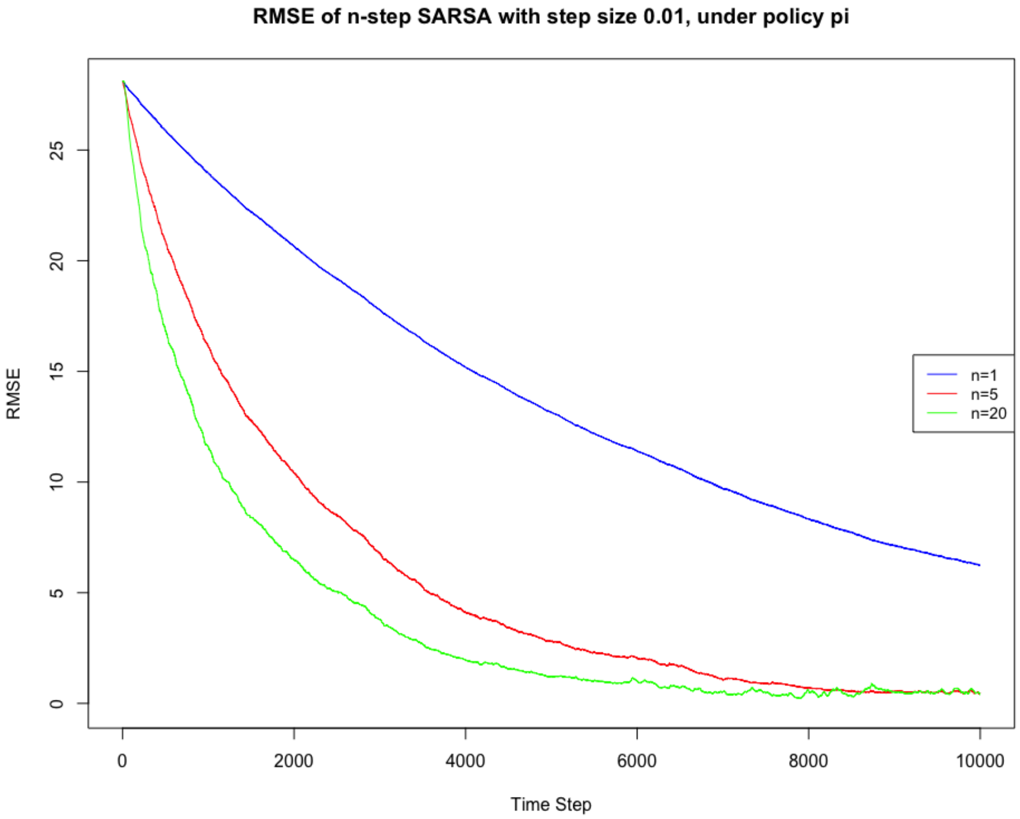
a) Action-value function for policy π.

$$\pi = \begin{bmatrix} \pi(1|1) = 0.7 & \pi(2|1) = 0.2 & \pi(3|1) = 0.1 \\ \pi(1|2) = 0.1 & \pi(2|2) = 0.7 & \pi(3|2) = 0.2 \end{bmatrix}$$

$$\texttt{PolicyAVF} = \begin{bmatrix} 12.57745 & 12.10109 & 11.10109 \\ 11.05382 & 10.79200 & 11.10109 \end{bmatrix}$$

b), **c)** See `main.R` code.

d) For the given problem, $n = 20$ seems to be the best as its RMSE value decays at a quicker rate compared to the $n = 1$ and $n = 5$ curves.



e)

$$\text{Estimated} = \hat{Q}_* = \begin{bmatrix} \hat{q}_*(1,1) & \hat{q}_*(1,2) & \hat{q}_*(1,3) \\ \hat{q}_*(2,1) & \hat{q}_*(2,2) & \hat{q}_*(2,3) \end{bmatrix} = \begin{bmatrix} 14.67647 & 14.18796 & 13.14585 \\ 13.10830 & 12.88070 & 13.19346 \end{bmatrix}$$

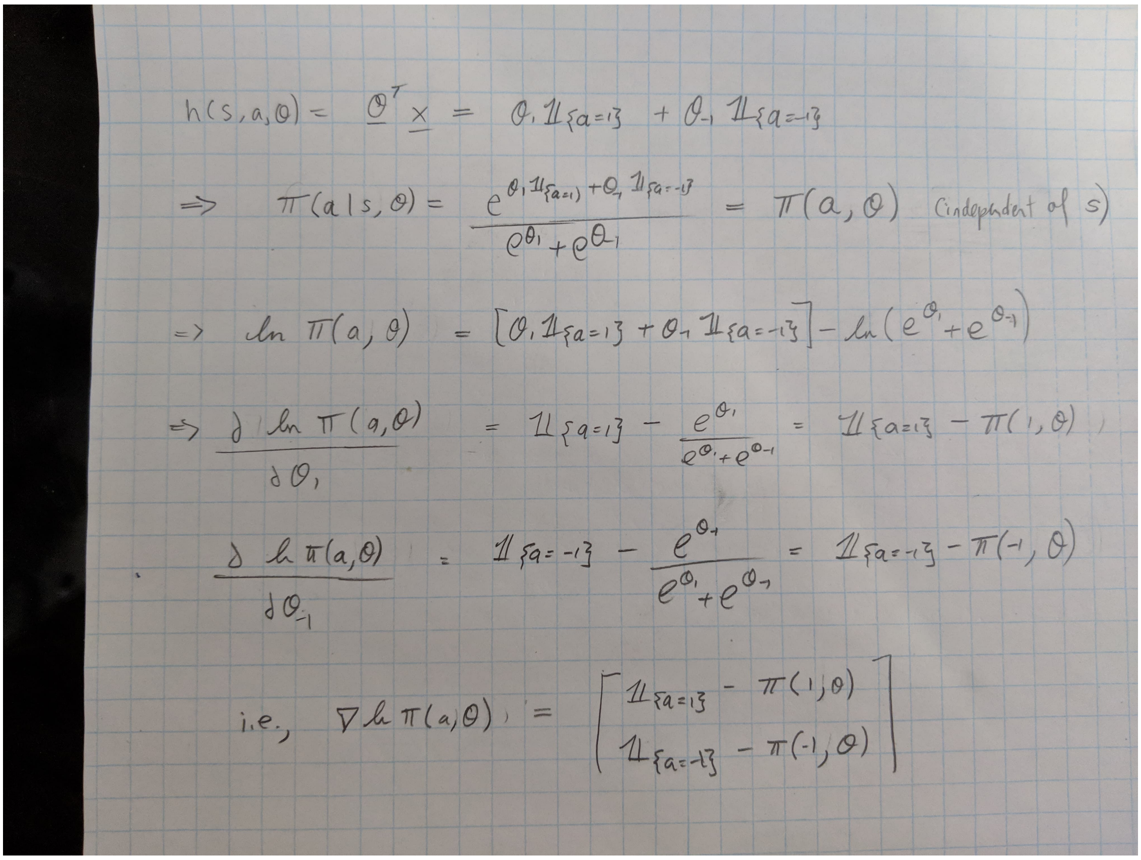
$$\text{True} = Q_* = \begin{bmatrix} q_*(1,1) & q_*(1,2) & q_*(1,3) \\ q_*(2,1) & q_*(2,2) & q_*(2,3) \end{bmatrix} = \begin{bmatrix} 14.70588 & 14.23529 & 13.23529 \\ 13.17647 & 12.91176 & 13.23529 \end{bmatrix}$$

Question 2

a)

Corridor Problem		
Probability of Selecting Right	0.05	0.95
Expected Return	-44.11255	-81.8691
True State Values	-44	-82

b)



$$h(s,a,\theta) = \underline{\theta}^T \underline{x} = \theta_1 \mathbb{1}\{a=1\} + \theta_{-1} \mathbb{1}\{a=-1\}$$

$$\Rightarrow \quad \pi(a \mid s,\theta) = \frac{e^{\theta_1 \mathbb{1}\{a=1\} + \theta_{-1} \mathbb{1}\{a=-1\}}}{e^{\theta_1} + e^{\theta_{-1}}} = \pi(a,\theta) \quad \text{(independent of } s\text{)}$$

$$\Rightarrow \quad \ln \pi(a,\theta) = \left[\theta_1 \mathbb{1}\{a=1\} + \theta_{-1} \mathbb{1}\{a=-1\}\right] - \ln\left(e^{\theta_1} + e^{\theta_{-1}}\right)$$

$$\Rightarrow \quad \frac{\partial \ln \pi(a,\theta)}{\partial \theta_1} = \mathbb{1}\{a=1\} - \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_{-1}}} = \mathbb{1}\{a=1\} - \pi(1,\theta)$$

$$\frac{\partial \ln \pi(a,\theta)}{\partial \theta_{-1}} = \mathbb{1}\{a=-1\} - \frac{e^{\theta_{-1}}}{e^{\theta_1} + e^{\theta_{-1}}} = \mathbb{1}\{a=-1\} - \pi(-1,\theta)$$

$$\text{i.e.,} \quad \nabla \ln \pi(a,\theta) = \begin{bmatrix} \mathbb{1}\{a=1\} - \pi(1,\theta) \\ \mathbb{1}\{a=-1\} - \pi(-1,\theta) \end{bmatrix}$$

c) See `main.R` code.

d) The curves corresponding to step sizes $\alpha_\theta = 2^{-13}$and 2^{-14}, (red and green curves respectively), are very similar to those presented in the textbook. However, our return curve for step size $\alpha_\theta = 2^{-12}$ (blue curve) is not reflective of the one in the textbook.

Figure 1: Performance of REINFORCE on the short-corridor gridworld (Total reward on episode averaged over 100 runs)

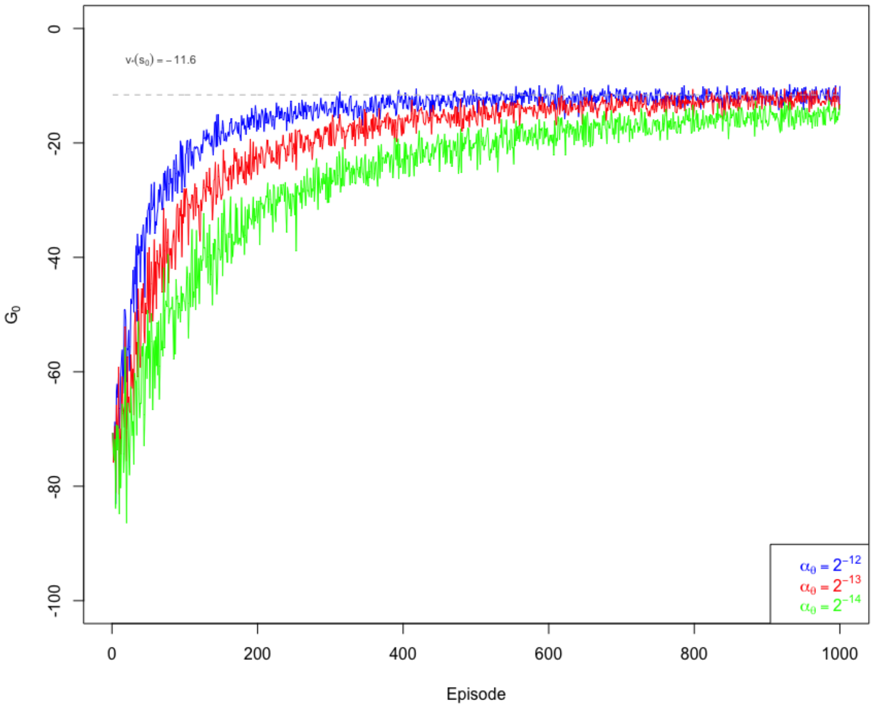
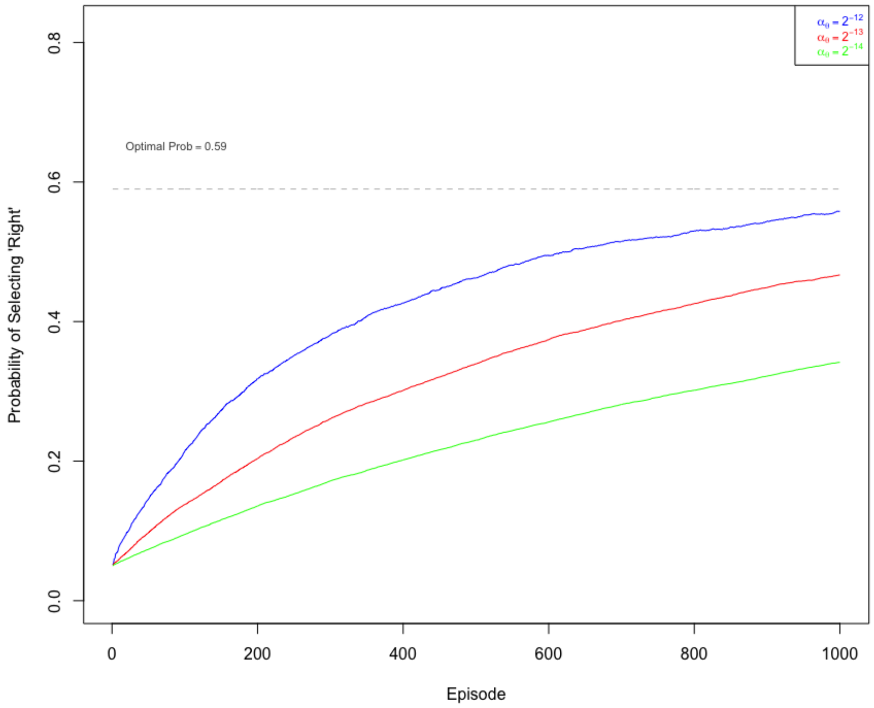


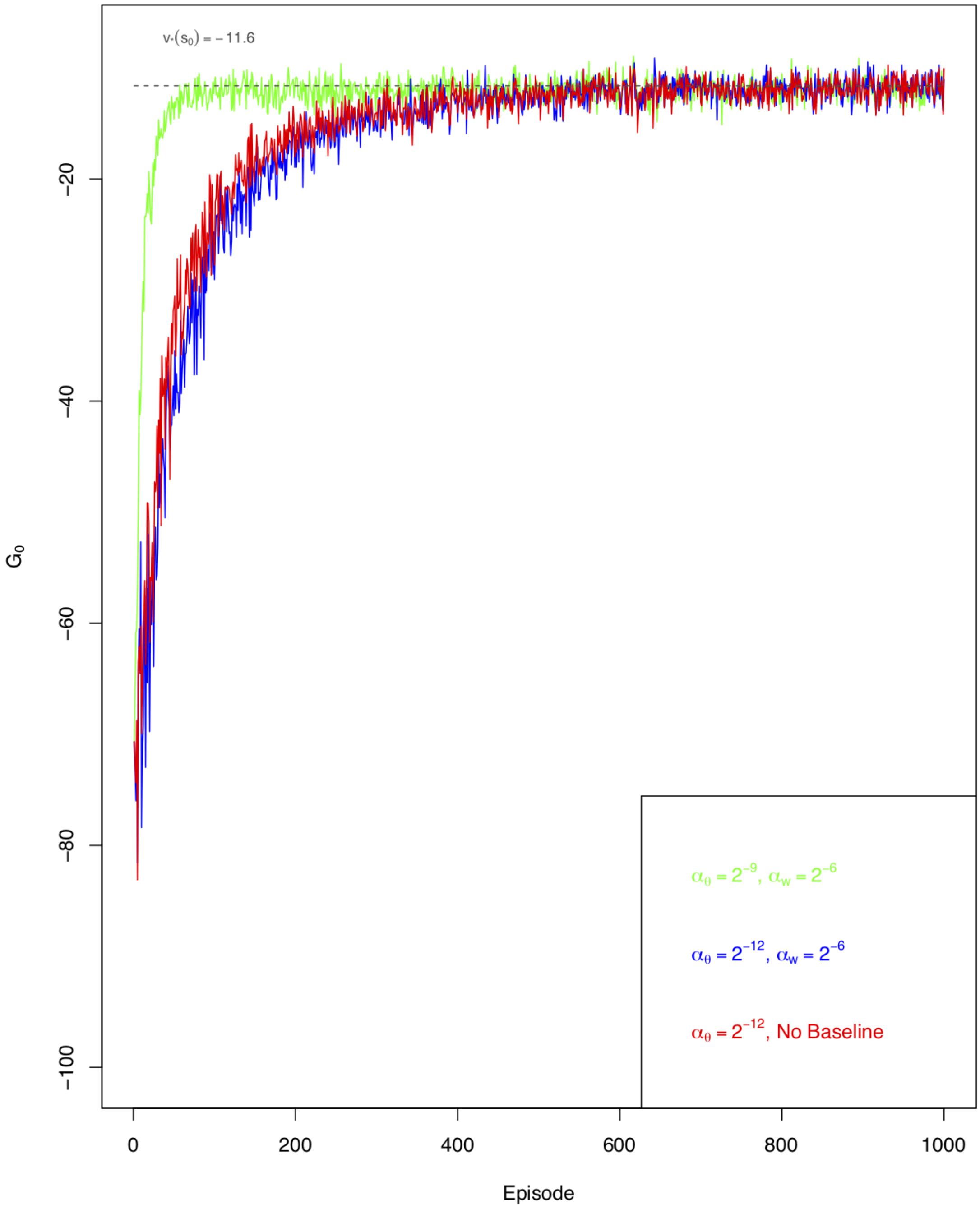
Figure 2: Performance of REINFORCE on the short-corridor gridworld (Probability of selecting "Right" compared to the optimal probability)



e) See `main.R` code.

f) The curve corresponding to step size $\alpha_\theta = 2^{-9}$, (green) curve is very similar to the one presented in the textbook. Regarding the red curve, the one presented in the textbook is of step size $\alpha_\theta = 2^{-13}$ without baseline, whereas ours is of step size $\alpha_\theta = 2^{-12}$ without baseline, therefore there is a discrepancy between the two red curves. However, for $\alpha_\theta = 2^{-12}$ the results are similar with and without baseline (red and blue).

Figure 3: Performance of REINFORCE with Baseline on the short-corridor gridworld (Total reward on episode averaged over 100 runs)


$v_*(s_0) = -11.6$
G_0
Episode
$\alpha_\theta = 2^{-9}$, $\alpha_w = 2^{-6}$
$\alpha_\theta = 2^{-12}$, $\alpha_w = 2^{-6}$
$\alpha_\theta = 2^{-12}$, No Baseline