Projet de Renforcement Learning

Jed Moutahir

December 2023

1 Introduction

The parameters chosen for the following are :
$\lambda = 0.3$, $\mu_1 = 0.2$, $\mu_2 = 0.4$, $\gamma = 0.99$, $Q_1 = 20$, $Q_2 = 20$
The code is available at this Github Folder

2 MDP

2.1 Policy Evaluation

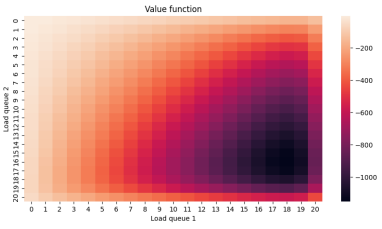
Assume the random policy that dispatches every job with probability 0.5 to either queue 1 and 2.

2.1.1 Bellman's equation

Here is Bellman's equation that characterizes the value function for this policy :

$$V[i,j] = -(i+j)+\gamma*(\frac{\lambda}{2}*(V[i+1,j]+V[i,j+1])+\mu_1*V[i-1,j]+\mu_2*V[i,j-1])+(1-\lambda-\mu_1-\mu_2)*V[i,j]$$

2.1.2 Value function



2.2 Optimal control

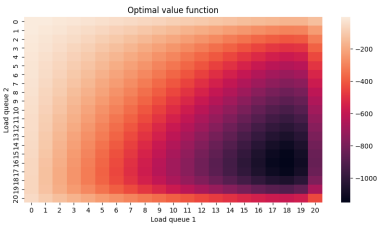
In this part you are asked to find the optimal policy to dispatch incoming jobs.

2.2.1 Bellman's equation

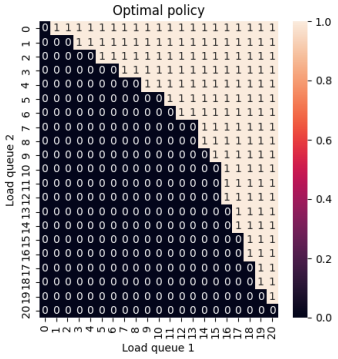
Here is Bellman's equation that characterizes the value function for this policy :

$$V[i,j] = -(i+j)+\gamma*(\lambda*max(V[i+1,j],V[i,j+1])+\mu_1*V[i-1,j]+\mu_2*V[i,j-1])+(1-\lambda-\mu_1-\mu_2)*V[i,j]$$

2.2.2 Value function

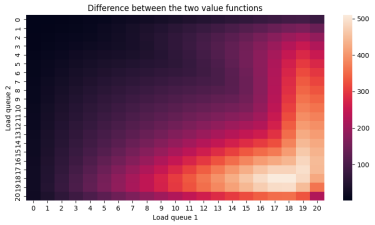


2.2.3 Optimal Policy



2.2.4 Comparing performances

As we can see in the following plot, the optimal policy is always better for every state possible.

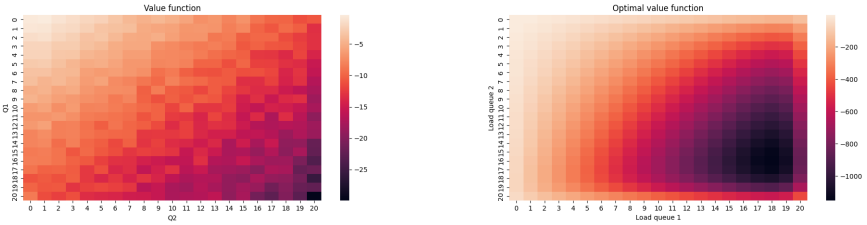


3 Tabular Model-Free control

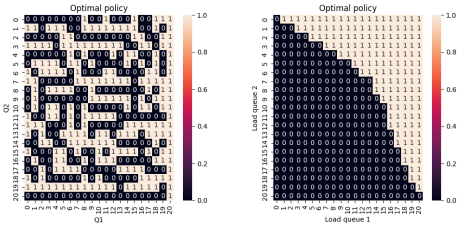
3.1 Policy Evaluation

3.1.1 Comparing TD(0) and MDP

Here are the value functions for TD(0) and MDP :

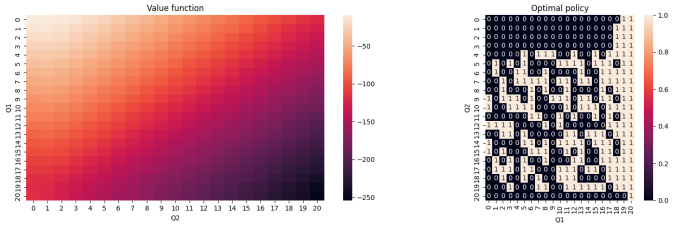


Unfortunately, the value function calculated using TD(0) and $\alpha_n = \frac{1}{n}$ does not correspond with the optimal one. Hence, the optimal policy is also quite different :

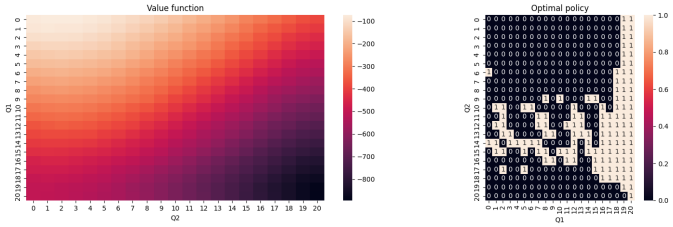


3.1.2 Experimenting with α_n

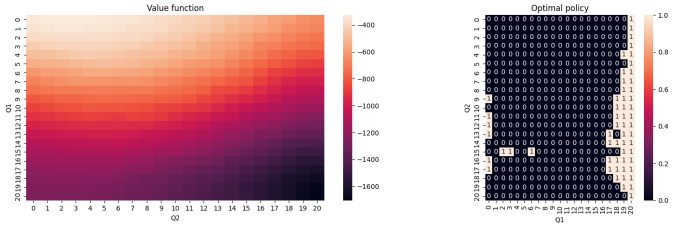
Here are multiple experiments on how $alpha_n$ could help the convergence of TD(0) :
$alpha = \frac{1}{\sqrt{n}}$



$alpha = \frac{1}{ln(n)}$



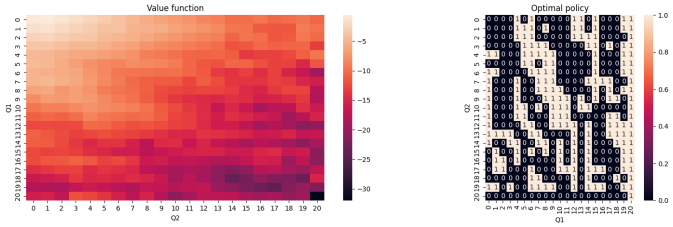
$alpha = \frac{1}{ln(ln(n))}$



3.2 Optimal control

3.2.1 Q-Learning

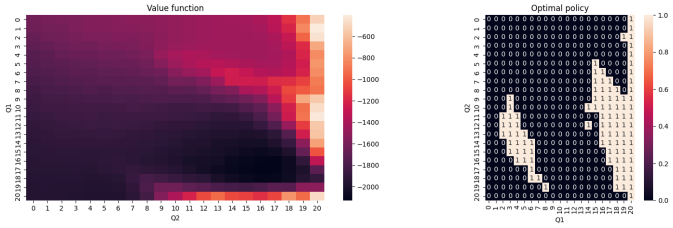
Here are the value function and the optimal policy found using Q-Learning and $\alpha_n = \frac{1}{n}$



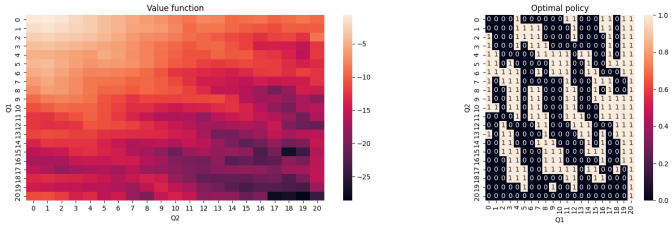
Q-Learning doesn't seem to be converging for this problem.

3.2.2 Experimenting with α_n

$\alpha_n = 0.1$



$\alpha_n = \frac{1}{n^\gamma}$

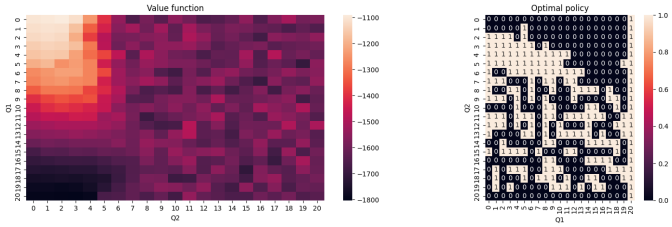


Even after experimenting with α_n we are not able to make the Q-Learning method converge. This method does not seem to fit this problem.

4 Model-free control with Value Function/Policy approximation

The subject chosen is Policy approximation with softmax parametrization

Here are the results obtained :



Unfortunaltely, even after a lot of tuning, the algorithm did not converge.

5 Conclusion

This problem is quite hard to solve using Reinforcement Learning methods other than the Iterative Policy Evaluation.