RL | PA -3 : SMDP & Intra-Options Q-learning | Report

By, Chandresh Sutariya (21f3001415)

Colab link : https://colab.research.google.com/drive/1Pu0wWvtX9--MVL5DjSzXuF5D7JmnRgfN?usp=sharing

Comparison between SMDP & IOQL

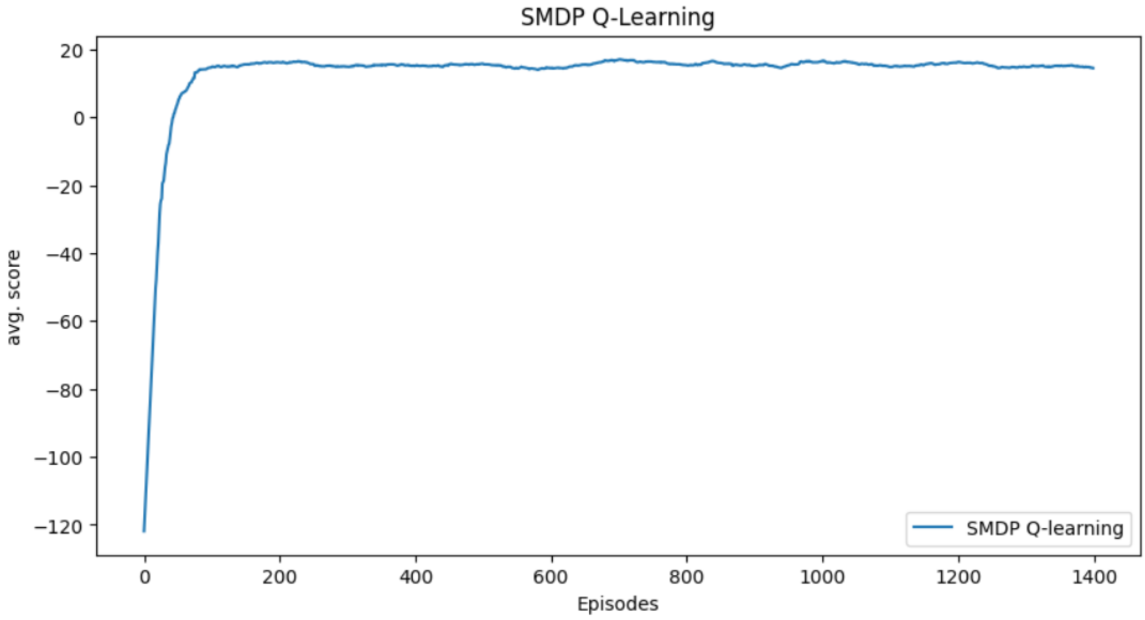
In this particular task/environment, **I observed that IOQL's and SMDP-Q-Learning's rewards <u>do not differ</u> <u>significantly</u>**.

I tried running the whole notebook several times to see if the reward curves change significantly, but they were the same (of course at a high level view).

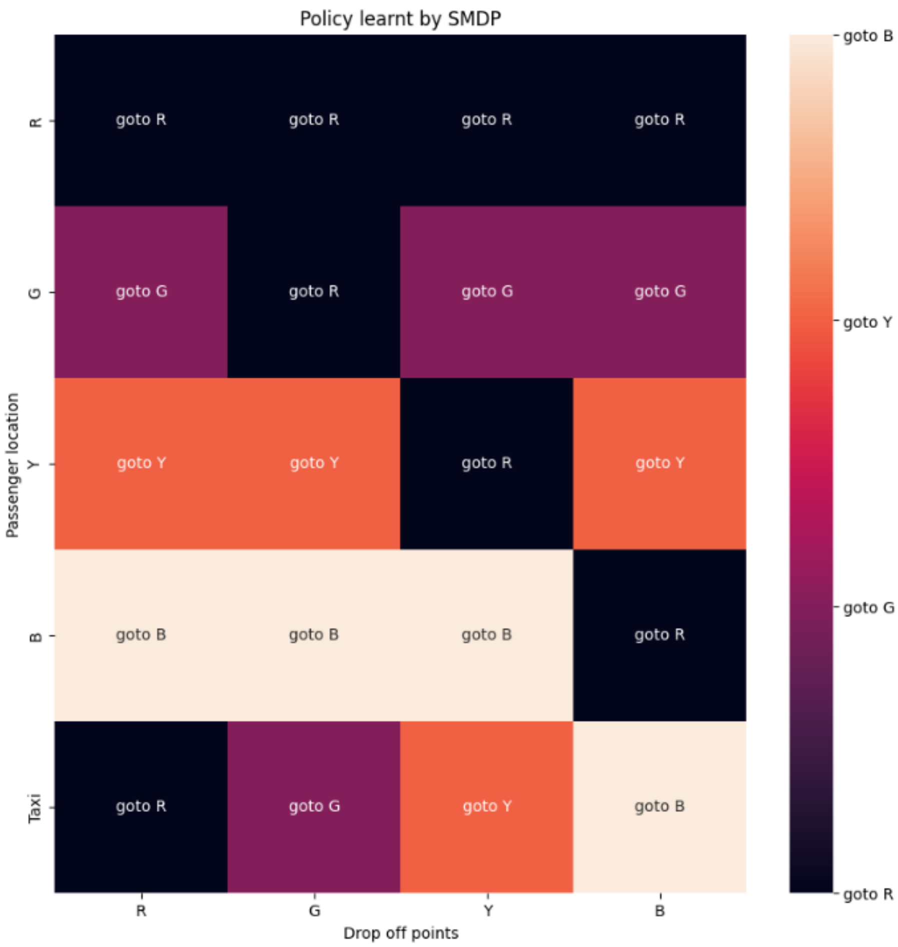
While running the notebook several times and observing, I saw that the-number-of-times the Q-leaning gives slightly better **rewards** were more. This may be because, **in Q-learning, we update the option q-value regardless if the option is ending or not, but in SMDP, that is not the case. In SMDP we only update the q-value of the option when the option is terminating.**

The one reason I think that both IOQL and SMDP-QL give the same reward curves **is that the task here is very small, there are not many options and the options themselves are also small (in terms of actions that the option has).** AND the policy learned by both algorithms is also the same, **I think that is also due to the same reason.**

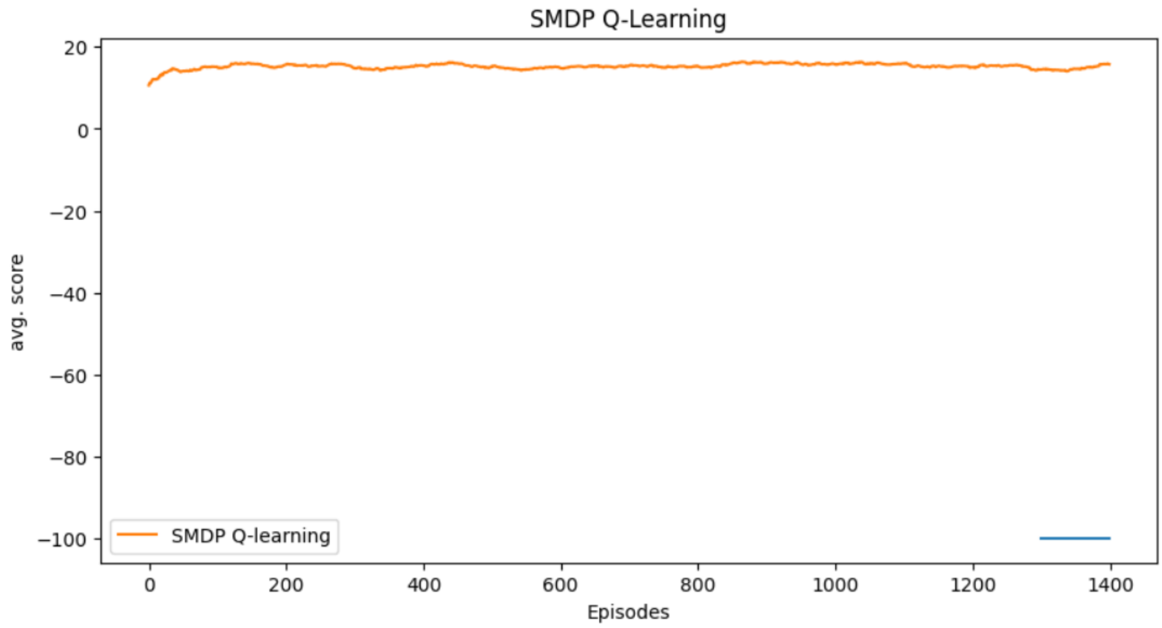
Task 1: SMDP | Reward Curve while learning



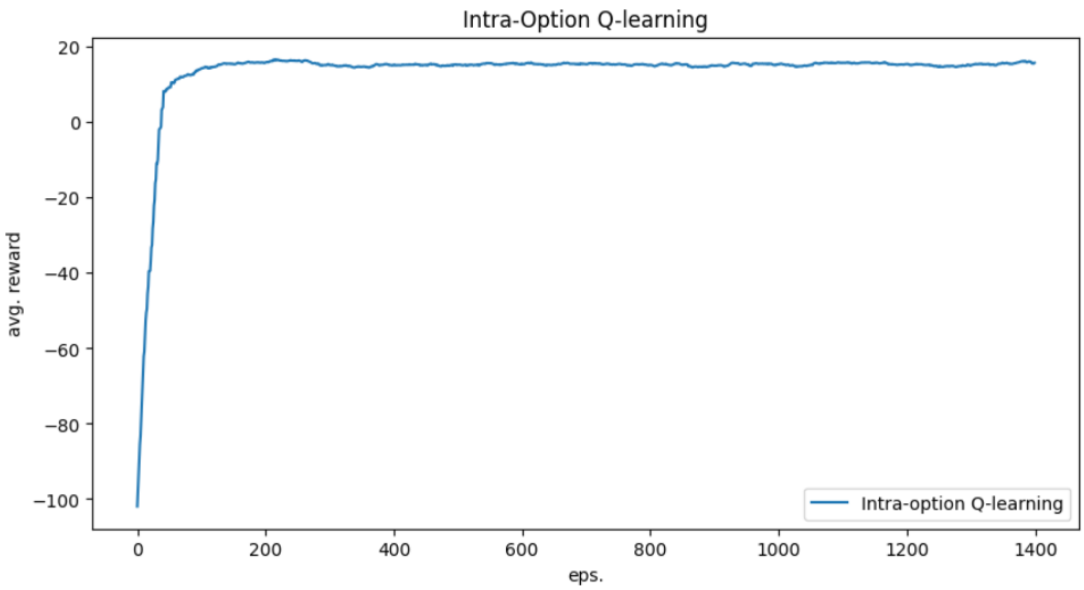
Task 1 : SMDP | Learned Option policy



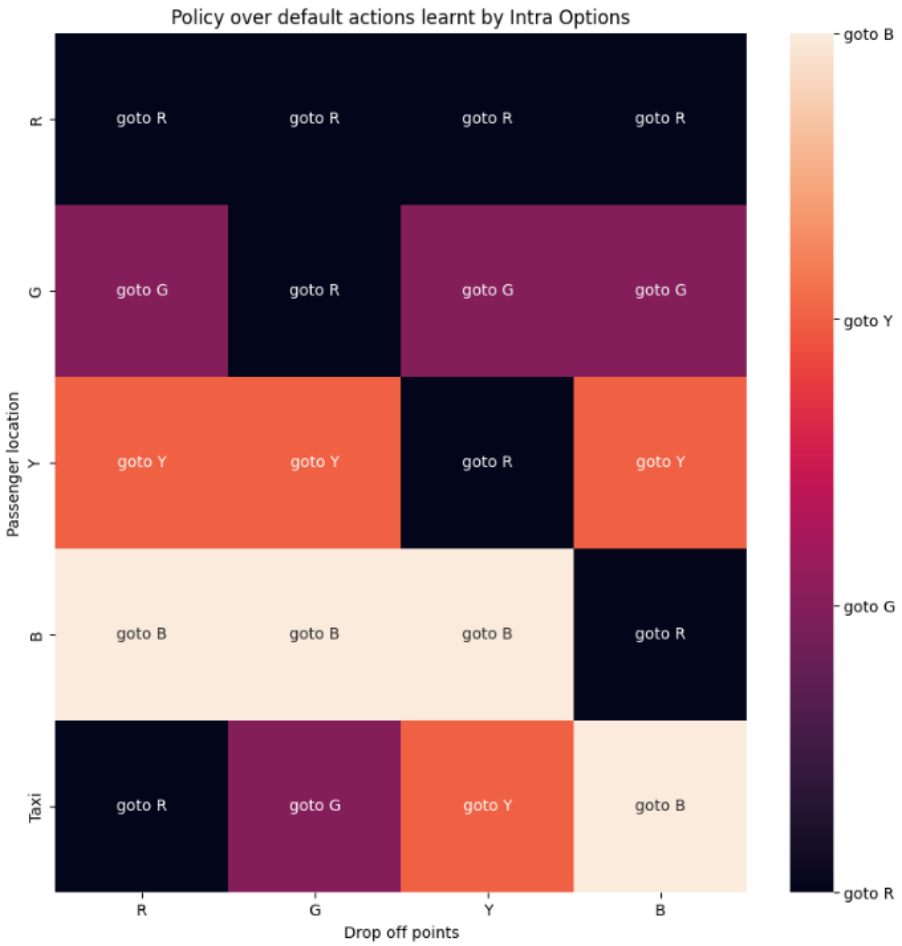
Task 1 : SMDP | Reward Curve after learning (using learned Q-values) (not updating SMDP Q-values)



Task 2: SMDP | Reward Curve while learning



Task 2 : SMDP | Learned Option policy



Task 2 : SMDP | Reward Curve after learning (using learned Q-values) (not updating SMDP Q-values)