RL e09

Mengnan Wang

June 26, 2017

1 SMDP learning in four-room domain

1.1 Computing policies for options

I only compute 4 policy for moving to up/left/right/down door. These policies can be used in any one of four rooms with correct coordinate mapping.

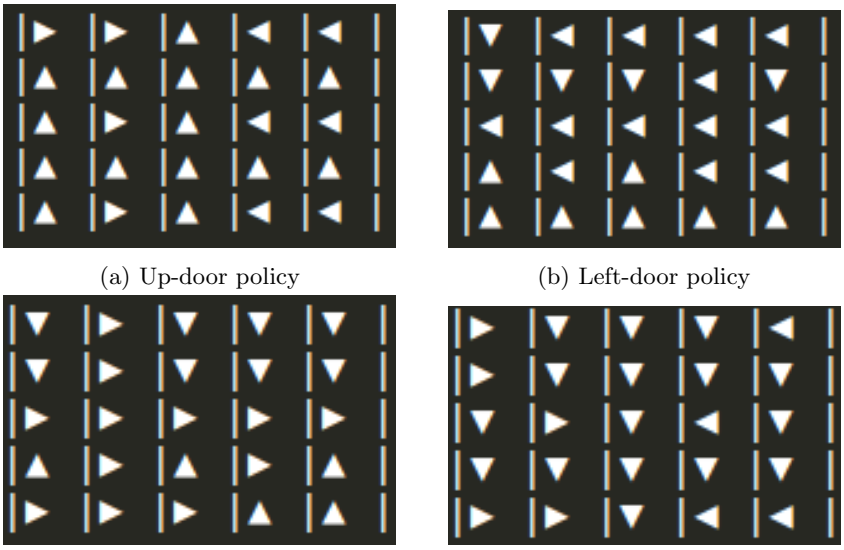
1.2 SMDP planning

'U' stands for moving to up door. 'L' stands for moving to left door. 'D' stands for moving to down door. 'R' stands for moving to right door.

Compared to basic Q-Learning without options. SMDP explores less states, which can be seen in Q value plot. Update type 1 explores less than naive update. But for update type 2, it does not converge after 1000 iteration. As i increase the iteration, it converges at 10000 iteration and the final policy and value plot is similar to basic Q-Learning without options.

1.3 Compute P

To estimate $P(s^{'}, k|s, o)$, we can record the transition $\{s, o, r, s^{'}, k\}$ and compute the expectation.



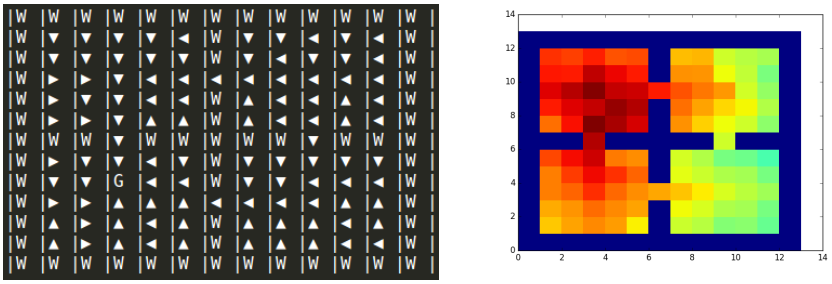
(a) Up-door policy

(b) Left-door policy

(c) Right-door policy

(d) Down-door policy

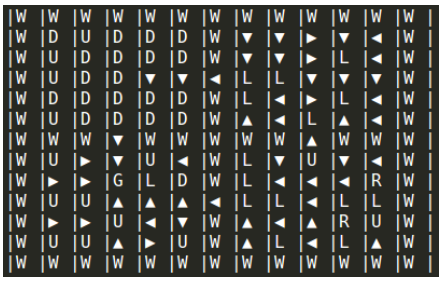
Figure 1: Option policy



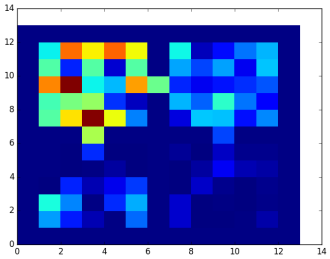
(a) Policy of Q-Learning

(b) Q value of Q-Learning

Figure 2: Basic Q-Learning with 1000 iterations

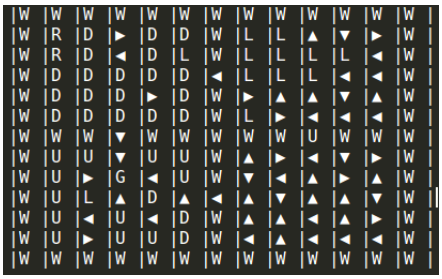


(a) Policy of Q-Learning

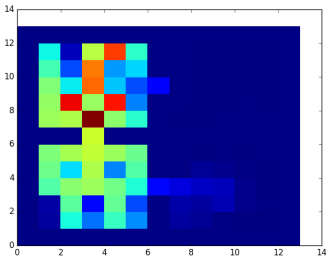


(b) Q value of Q-Learning

Figure 3: SMDP with naive update after 1000 iterations



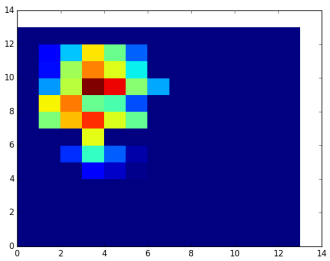
(a) Policy of Q-Learning



(b) Q value of Q-Learning

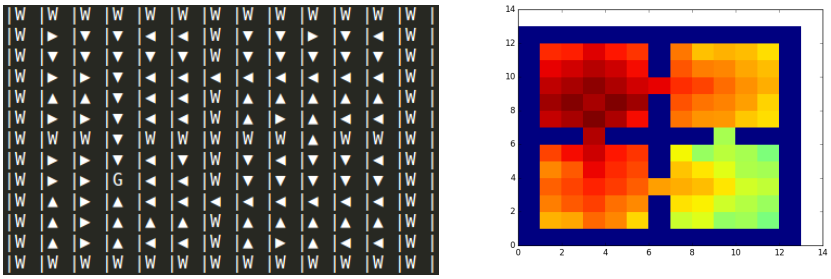
Figure 4: SMDP with update type 1 after 1000 iterations

(a) Policy of Q-Learning



(b) Q value of Q-Learning

Figure 5: SMDP with update type 2 after 1000 iterations



(a) Policy of Q-Learning

(b) Q value of Q-Learning

Figure 6: SMDP with pdate type 2 after 10000 iterations