Homework Assignment 2

In this assignment we continue exploring algorithms for MDPs. This time we consider model free and model based learning algorithms; in particular, SARSA and Dyna-SARSA-RTDP.

The assignment is due by Monday October 6, 1:30pm - in hardcopy (in class) and electronically (via provide).

We evaluate performance on the grid domain from assignments 1 with exactly the same parameters. Recall that we evaluated policies as follows. Your implementation should be such that you can “empirically evaluate” the greedy policy with respect to the current value function after every iteration. To “empirically evaluate” a policy, we run it for 100 episodes where each episode is at most 100 steps long. Each run starts at the bottom left state and the quality of the policy is evaluated as the average of the total discounted reward received in these runs.

The first algorithm is SARSA performing “on-policy TD control” (see review slides). The MDP should automatically reset to the bottom left state every time the goal is reached. You should use exploration parameter $\epsilon = 0.1$ and learning rate $\alpha = 0.1$ at least for initial experiments. For SARSA you should evaluate performance after every 100 steps.

The second algorithm is Dyna-SARSA-RTDP (see review slides) that extends SARSA by learning a model and using it for simulated updates. Simulated updates are done by drawing traces from the model according to the current $\epsilon$-greedy policy and making single state Bellman updates in these. Use $\epsilon$ and $\alpha$ as for SARSA and use $m = 10$ simulated episodes per step. You should evaluate performance after roughly every 100 steps (“real” experience and simulated experience together).

For your report plot the quality of the policies as a function of the “effort” of the algorithm. For SARSA plot quality as a function of the number of steps. For Dyna give three plots. One plot as a function of the number of “real steps” (that is, here we assume simulated experience is very cheap so we do not charge for it). One plot as a function of combined experience and simulated steps (here real and simulated experience are charged the same). The last plot as a function of an approximation of “single action backups” where here TD backups are charged one unit but in RTDP for each action $a$ (note that in the code we update $Q$ for each $a$) the update costs one unit per potential next state $s'$.

Finally, experiment a little with parameters $\epsilon$, $\alpha$, $m$ and study their effect on the results.

It is up to you to decide on language for implementation but please make sure to write in good style and document your code clearly so we can read your code.

1 Submitting Your Assignment

- You should submit the following items both electronically and in hardcopy:
  (1) All your code for the assignment (please write clear code and document it as needed),
(2) Results of the experiments which should include at least the graph explained above (put all 4 plots together) and any observations on this graph as well as results that vary parameters and any observations you have.

- Please submit a hardcopy in class.
- Please submit electronically using provide: Put all the files from the previous item into a zip or tar archive, for example call it myfile.zip. Then submit using provide comp150aml a2 myfile.zip.