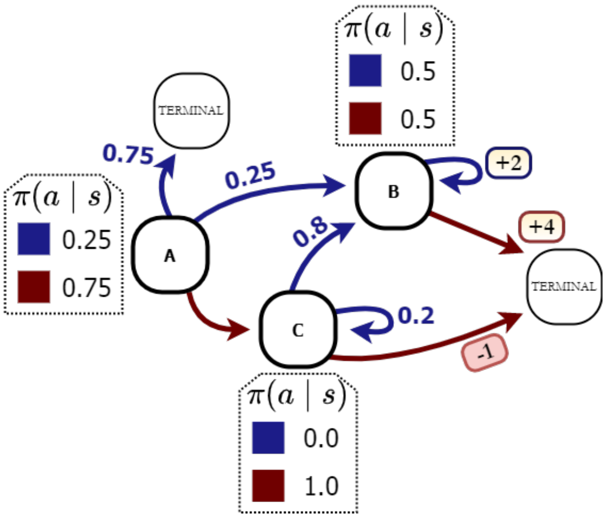
CUB Reinforcement Learning, Fall 2024

Test Exam

During exam no materials can be used. For each task you may get up to 1 point. The total grade for the exam is computed as sum of points for all tasks divided by 10.

Task 1. For given MDP and policy find better policy using policy improvement for $\gamma = 0.5$:

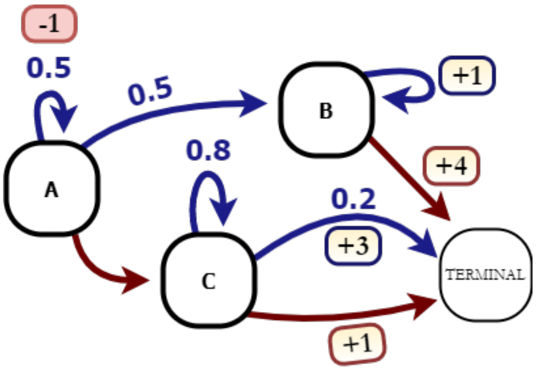


Task 2. For tabular MDP:

- ☐ Optimal policy may not exist;
- ☐ There exists exactly one optimal (and deterministic) policy;
- ☐ There exists at least one optimal policy, but all of them are deterministic;
- ☐ There exists at least one optimal policy and among these policies at least one is deterministic;
- ☐ None of the above.

Task 3. Write down Bellman equations for optimal V and Q functions (in total four equations: 1) Q using Q, 2) V using V, 3) Q using V and 4) V using Q).

Task 4. Find all optimal policies for given MDP and $\gamma = 0.5$:



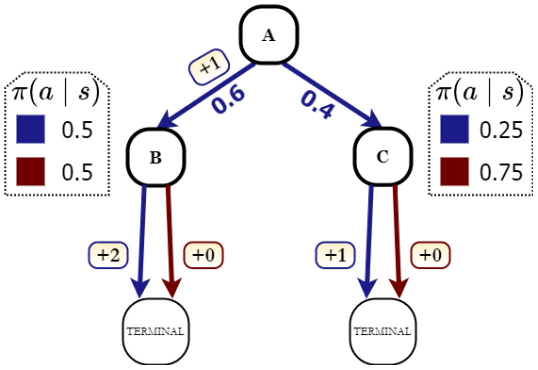
Task 5. Point out off-policy RL algorithms:

- □ CEM (Cross-Entropy Method)
- □ DQN (Deep Q-network)
- □ Rainbow DQN
- □ QR-DQN (Quantile Regression DQN)
- □ REINFORCE
- □ A2C (Advantage Actor-Critic)
- □ PPO (Proximal Policy Optimization)
- □ DDPG (Deep Deterministic Policy Gradient)
- □ TD3 (Twin Delayed DDPG)
- □ SAC (Soft Actor-Critic)

Task 6. Choose correct statements about GAE estimate:

- □ In this estimate all available N-step estimates are aggregated;
- □ In this estimate user should tune a hyperparameter λ, that is reponsible for bias-variance trade-off;
- □ It is required to play full episode in order to be able to compute this estimate;
- □ If in GAE estimate N-step estimates only for large N are used, then GAE value may degrade and for all values of λ may have too large variance;
- □ If in GAE estimate N-step estimates only for small N are used, then GAE value may degrade and for all values of λ may have too large bias.

Task 7. Compute distribution of returns (like in distributional RL) for state A and the blue action. MDP and policy are the following, $\gamma = 0.5$:



Task 8. For the rollout $s_0, a_0, r_0 = +1, s_1, a_1, r_1 = +0, s_2, a_2, r_2 = +1, s_3$, finished at the terminal state s_3, find GAE estimate $\text{GAE}(s_0, a_0)$ for $\lambda = 0.5$ with discounting factor $\gamma = 1$, if the current approximation of V function is:

$$V^\pi(s_0) = +1$$

$$V^\pi(s_1) = +1$$

$$V^\pi(s_2) = 0$$

Task 9. PPO and TRPO advantages compared to simple policy gradient methods (A2C) are in the following:

□ They can learn using data collected by arbitrary behaviour policy;

□ They better address exploration-exploitation dilemma using comparison between two policies;

□ Optimization of lower bound instead of initial criterion function allows to get unbiased gradient estimate even with non-exact critic;

□ They can use ensembles of multi-step estimates;

□ None of the above.

Task 10. Adding policy entropy term to cumulative reward (like in SAC method) leads to:

□ Set of optimal policies does not change;

□ Optimal policy becomes unique;

□ Optimal policy becomes stochastic;

□ Value functions coincide with the ones from original RL formulation;

□ New value function $U(s, a)$ is introduced;

□ None of the above.

Task 11. In which RL algorithms for actor's training a gradient of Q function w.r.t. actions is used?

□ DQN

□ Policy Gradient algorithms (A2C, PPO)

□ DDPG

□ TD3

□ SAC (if reparameterization trick is applicable)

Task 12. What is computed by different heads of neural network in AlphaZero?

□ policy for given state

□ V function value

□ Q function value

□ Exploration Bonus;

□ probabilities for next states;