1. $\frac{1}{2}$ point for right answer, $-\frac{1}{2}$ point for wrong answer

a) every non-greedy action is chosen with probability $\frac{\varepsilon}{n}$, where n is number of actions

b) each action is at least picked with probability $\frac{\varepsilon}{n}$

c) Q-Learning has a maximization bias problem

d) Q-Learning is on-policy learning

e) episodic tasks can be transformed to non episodic tasks

f) TD-learning can learn before episode terminates

g) Inverse RL tries to learn the transition function

h) Value function of Sarsa converges to optimal value function if all state-action pairs are infinitely often visited

i) Value iteration converges after one step for any MDP if $\gamma = 1$

j) Monte Carlo is non-applicable on non-episodic tasks

2. Bandits

a) 3 actions, $p(a_1) = 0.6$, $p(a_2) = p(a_3) = 0.2$

Find Action Function for $Q(a_1)$, $Q(a_2)$ and for ε

b) For arbitrary distribution p_1, p_2, p_3, can ε and Q-Functions be represented?

Give a proof or a counter example

3. Backup diagrams

Give the backup diagrams for Q-Learning, Sarsa and Monte Carlo

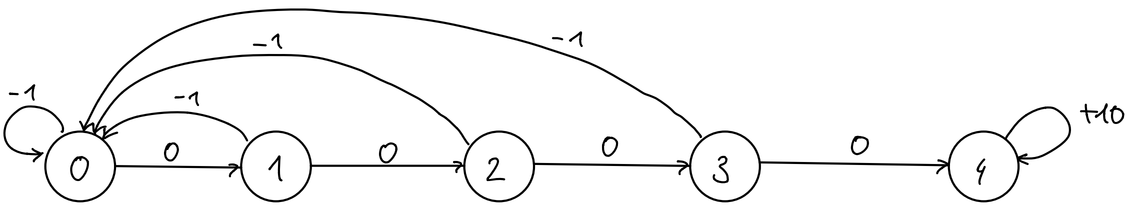
4. Value iteration

a) Give the recursive relationship of v_π

b) Express q_π in terms of v_π

c) Prove the Contraction Theorem (Bellman Operator given)

5. Markov decision process



a) What is the optimal policy

b) Calculate value iteration for V_1, V_2 for all states

c) Calculate optimal value function V^* for all states

Hint: $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$, $|a| < 1$

6. Policy improvement

a) μ is probabilistic policy, that is greedy wrt v_π

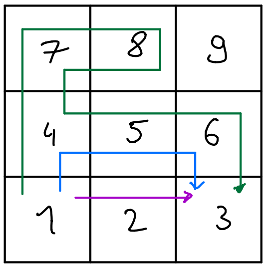
π is a deterministic policy

Prove that $v_\mu(s) \geq v_\pi(s)$ for all s

b) If $v_\mu(s) = v_\pi(s)$ for all s, show that $v_\mu(s)$ is the optimal policy

7. Monte Carlo

a) Give the update rule for first visit MC



3 episodes given as table with rewards

- ep 1 : $s=3$, $r=2$
- ep 2 : $s=3$, $r=2$
- ep 3 : $s=4$, $r=1$, $s=3$, $r=3$

b) Calculate the value function for all states for first visit and for every visit

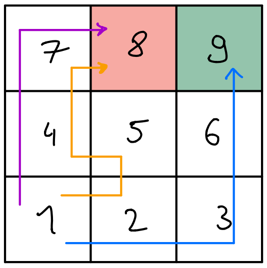
c) Episodes are done with a policy that chooses each action with probability 0.25.

Now we want to learn a target policy with $\pi(\downarrow|s)=0.5=\pi(\rightarrow|s)$ $\forall s$

Perform ordinary importance sampling and provide a value function.

8. Sarsa

a) Provide the update rule for Sarsa



b) Calculate the relevant Q-Functions for states and actions after each episode

c) $Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\left(R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1},a) - Q(S_t,A_t)\right)$

Explain if this update rule is on-policy or off-policy

d) Give the value function for Linear function approximation and explain

e) $w^T = [1, -1, 2, 1]$

features consist of x-position and y-position of field, i.e. 1 has $(1,1)$, 8 has $(2,3)$

and actions encoded as: $\rightarrow$ is $[1,0]$, $\downarrow$ is $[0,-1]$, $\uparrow$ is $[0,1]$ and $\leftarrow$ is $[-1,0]$

features: [x-position, y-position, action]

Calculate $q(1,\uparrow)$, $q(5,\uparrow)$, $q(5,\rightarrow)$, $q(6,\downarrow)$

f) Discuss if this choice of features is a good representation.

9. Policy gradient

Trajectory $\tau = (S_0, A_0, R_1, S_1, A_1, R_2 \dots S_T)$

a) Give $\nabla_\theta \log p(\tau|\theta)$. Start with the likelihood of the probability

b) Give a formula for $\nabla_\theta \mathbb{E}[G_0]$

c) Two actions are given a_0 and a_1. Give a simple parametrization for $\pi(a_0|s;\theta)$ and $\pi(a_1|s;\theta)$ with features $\phi(s)$