Reinforcement Learning
Gedächtnisprotokoll – 15.08.2023

August 15, 2023

Materials: non-red non-erasable pen
Time: 120 min
Total Points: 60 min
The time was enough, they really do not want you to have protocols of the exam, sheets will be collected afterwards.
All in all nearly all tasks where identical or very similar to the previous two years.

1 True or False (7P)

1/2 Points per Question, there will be no redaction if false crosses are given.

T F

☐ ☐ A softmax policy is effective for a continous action space Reinforcement Learning Problem

☐ ☐ Policy based Reinforcement Learning algorithms can also learn deterministic policies

☐ ☐ In ε greedy policiy selection, the greedy action is taken with probability ε and the random action is taken with probability $1 - \varepsilon$

☐ ☐ A greedy action is the action that maximizes the action-value function at a timestep

☐ ☐ Value iteration needs a model of the MDP dynamics

☐ ☐ In Monte Carlo COntrol some state action pairs may never be visited if the policy is deterministic

☐ ☐ Temporal Difference has typically a lower variance than MC Prediction

☐ ☐ In off-policy learning, the target policy is equivalent to the behaviour policy

☐ ☐ SARSA is an on-policy learning algorithm

☐ ☐ Temporal difference combines sampling of Monte Carlo with bootstrapping of Dynamic Programming

☐ ☐ Importance Sampling is used for off-policy Monte Carlo COntrol

☐ ☐ Temporal Difference has to wait until the episode has terminated

☐ ☐ Reinforce with baseline reduces variance and remains unbiased

☐ ☐ Monte Carlo Tree Search uses TD estimates to evaluate its leaves

2 Bandits with ε-greedy policy - 5P

Consider a two-armed bandit with actions a_1 and a_2. Under an stochastic ε-greedy policy π action a_1 is selected with probability $p = 0.9$ and a_2 with probability $1 - p = 0.1$.

1. Give possible action values $Q(a_1)$ and $Q(a_2)$ and a parameter ε that results in the given action probabilites (3P)

2. Can you derive the unique action values an an ε given an arbitrary probability distribution (p_1, p_2, p_3) over 3 actions ? (Give either a prove or a counterexample) (2P)

3 Backup - 6P

Draw the backup diagrams for SARSA, Q-Learning and Monte Carlo.

4 Value-Functions - 6P

1. Given a deterministic policy π, derive the recursive definition of the value function $v_\pi(s)$ in terms of the reward $r(s, a, s')$ and the probability transition model $p(s'|s, a)$ by starting from the following definition. (3P)
   $v_\pi(s) = E[G_t|S_t = s] = ...$

2. Still assuming a deterministic policy π express q_π in terms pf v_π (3P).

5 Policy improvement - 8P

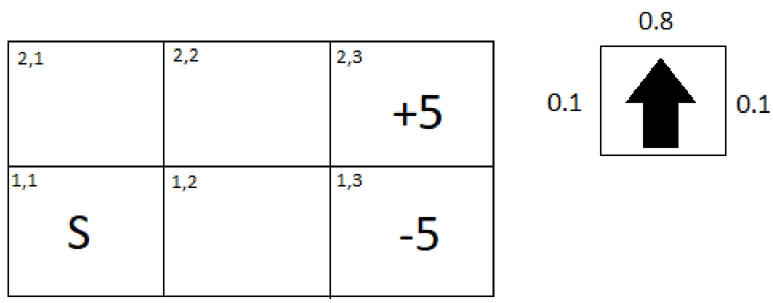
Consider a deterministic policy π and a policy μ that is greedy w.r.t. V_π.

1. Show that μ must be better than or equal to π i.e. $v_\mu(s) \geq v_\pi(s)$.
   Hint: Start with the recursive defnitions of the value function for policies π and μ. (4P)

2. Assume $\forall s : v_\mu(s) = v_\pi(s)$. Show that μ must be the optimal policy. (4P)

6 Markov Decision Processes - 10P

Given the following MDP analogous to the Grid world example from the exercises:



1. What is the optimal policy ? Simply list 'NA' for the terminal states ? (2P)

2. Suppose the agent knows the transition prob. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. Assume V_0 is initally zero for all states and compute V_i for times $i = 1$ and $i = 2$ (3P)

3. The agent start with the policy that all ways chooses to go right and executes the following 3 trajectories:

   (a) (1,1) - (1,2) - (1,3)

   (b) (1,1) - (1,2) - (2,2) - (2,3)

   (c) (1,1) - (2,1) - (2,2) - (2,3)

   What are the MC estimates of the value function for states (1,1) and (2,2) given the sampled trajectories? (2P)

4. Using a discout of $\gamma = 0.9$ and a learning rate of $\alpha = 0.1$ and assuming initial values of zero, what updates does the TD-learning agent kame to the value function given the trajectories 1 and 2 from subtask 3? (3P)

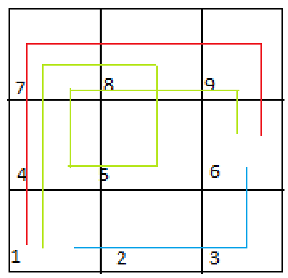
7 MC - Methods - 5P

Assume for the next task that $\gamma = 1.0$

1. State the equation/rule for calculating V_π with first visit MC Prediction (1P)

2. Consider the following episodes. The experience of the agent is given by the tuples (s, a, s', r):
   Calculate the value function v_π for all states with ... (4P)

episode 1	episode 2	episode 3
(1,↑,4,0)	(1,→,2,1)	(1,\uparrow,4,0)
(4,↑,7,0)	(2,→,3,0)	(4,\uparrow,7,1)
(7,→,8,0)	(3,↑,6,3)	(7,\rightarrow,8,0)
(8,→,8,1)		(8,\downarrow,5,0)
(9,↓,6,2)		(5,\leftarrow,4,0)
		(4,\uparrow,7,2)
		(7,\rightarrow,8,0)
		(8,\rightarrow,9,1)
		(9,\downarrow,6,4)



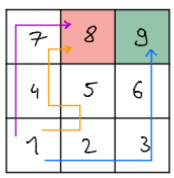
(a) first visit MC

(b) every visit MC

8 TD-Learning

1. State the SARSA Update rule (2P)

2. Some example with given episodes to tell which values changed an by how much.
   Here I used the exercise from a previous semester as they were the same:



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

9 Policy Gradient

1. State the Policy Gradient Theorem $\nabla_\theta J(\theta) = ...$ (2P)

2. Give the update rule for the policy weights θ of the REINFORCE algorith, (without baseline)(2P)

3. Consider a softmax policy with score function:

$$\pi_\theta(s,a) = \frac{e^{\phi(s,a)^T\theta}}{\sum_{k=1}^{N} e^{\phi(s,a_k)^T\theta}} \qquad \nabla_\theta log\pi_\theta(s,a) = \phi(s,a) - \mathrm{E}_{\pi_{theta}}[\phi(s,\cdot)]$$

   We consider an environment with 4 actions. You observed the state $s_0 = 1$ and the action $a_0 = \uparrow$ with a return $G_0 = 5$. Features $\phi(s,a)$ are given by $\phi(1,\uparrow) = (4,2)^T$, $\phi(1,\leftarrow) = (1,3)^T$, $\phi(1,\rightarrow) = (1,2)^T$ and $\phi(1,\downarrow) = (2,5)^T$.
   Compute the update to $\phi \in \mathrm{R}^2$ applied by REINFORCE without baseline for the observed state-action pair (s_0, a_0). Assume θ is intizialized to zero and $\alpha = 1$ and $\gamma = 1$ (3 P)