Gedankenprotokoll SS22, Reinforcement Learning, 16.08.2022 (M. Niepert)

9 tasks, 60 points, 120min

1. **True-false** (0.5Pkt for correct, -0.5Pkt for incorrect, 7Pkt)
    a. Does MC have a bias?
    b. REINFORCE with baseline reduces variance and introduces Bias
    c. TD can learn before an episode ends
    d. In epsilon greedy policy, the greedy action is chosen with epsilon and the non-greedy action with 1-epsilon
    e. Q learning is on policy
    f. In off-policy the behavior policy and target policy are the same
    g. Greedy action maximizes the q-value
    h. MCTS uses TD(0) for evaluation of the leaf nodes
    i. In MC learning, if a policy is deterministic, all states are visited with non-zero probability
    j. Episodic tasks can be transformed to non episodic tasks
    k. …

2. **Epsilon greedy policies and bandits (similar to 2. From SS20)**
    a. 2 actions and their probabilities in an epsilon-greedy policy ($p(a_1) = 0.9, p(a_2) = 0.1$) given. Calculate epsilon and give possible q-values.
    b. Can we calculate epsilon and $q(s)$ from an arbitrary probability distribution of 3 actions($p(a_1), p(a_2), p(a_3)$)? Give a proof or counter example.

3. Draw **backup diagrams** for Sarsa, Q-learning, DP (action-values) (6Pkt)

4. **Policy Improvement (same as 6. From SS20)**
    a. Consider a deterministic policy π and a policy μ that is greedy w.rt. $\boldsymbol{v_\pi}$. Prove that the value functions fulfil $v_\mu(s) \geq v_\pi(s) \forall s$.
    b. Assume $v_\mu(s) = v_\pi(s) \forall s$. Prove that this policy must be optimal.

5. **Policy optimization**
    a. Express the value function $v_\pi(s) = E_\pi[G_t | S_t = s]$ in terms of $r(s, a, s') \& \, p(s'|s, a)$ under a deterministic policy.
    b. Express the action value function $q_\pi(s, a)$ in terms of the result from a)
    c. Prove that the Bellman maximization operator is a gamma-contraction (same as exercise 3.1)

6. **MDP**
   Given grid world (S = (1,1) starting state):

		+5
S		**-5**

    a. Give optimal policy for the shown grid world (Not for terminal states)
    b. Calculate 2 iterations of Value iteration for all states
    c. Calculate MC value estimates for some given trajectories
       (1,1)→(1,2)→(1,3)
       (1,1)→(2,1)→(2,2)→(2,3)
       (1,1)→(1,2)→(2,2)→(2,3)

d. Calculate TD updates for the values for the same trajectories (2 iterations)

7. **Temporal difference learning**
    a. Give Sarsa update rule for tabular cases
    b. Calculate Sarsa updates for relevant states according to 3 given trajetories
    c. Some update rule (it was q-learning) given: Name it and say if it is off/on policy

8. **Monte Carlo**
    a. Give the equation to calculate first visit MC value estimates
    b. Calculate state values for First-Visit/Every Visit MC for given trajectories

9. **Policy Gradient**
    a. State Policy gradient theorem
    b. Give update rule for parameters θ
    c. Calculate one update step for parameters with Softmax policy, assuming we started in state s=1 and observed return $G_0 = 5$ after doing action 1. (formula for softmax and score function given). $\gamma = 1, \alpha = 1$, values for $\phi(s, a)$ given. (Similar to 10.b from SS21)

Questions contained basically nothing about n-step Returns and function approximation