Reinforcement Learning Klausur SS21 (59 Pts)

t

1. True or False:

- In the e-greedy case, all actions are chosen with nonzero probability
- TD can learn before episode terminates
- …

2. Backup diagrams:

- Q-policy evaluation
- Q-value-iteration
- Q-learning
- Sarsa

3. Value functions:

- Recursive definition of v_π = E[Gt|...]
- Bellman equation for q*
- Proof of $||Tv-Tv'||_\infty$

4. MDP:

- What is optimal policy
- 2 Value iterations of given MDP ($V_1(s)$, $V_2(s)$)
- Optimal values with $\sum a_i = 1/(1-a)$

5. GPI

  - Update rule for policy evaluation → Can you compute optimal value func.?
  - Explain GPI with sketch
  - How differ value and policy iterations

- **6. Monte-Carlo:**
  - Update rule for MC
  - First and every visit
  - Importance sampling with $\pi(a=right \mid s) = \pi(a=up \mid s) = 0.5$ and $b(a \mid s) = 0.25$
- **7. TD:**
  - Tabular Q learning update rule
  - 3 episodes and calculate relevant state action pairs (probabilistic env not deterministic)
  - Sarsa whas given you had to guess whoch update rule it is
- **8. Function approximation:**
  - Linear function approximation of q(s,a)
  - With L(w) (value-error, least square) get to update rule with derivative
  - Q-learning linear function approximation → same as from previous task with L(w)
- **9. Softmax**
  - give log of softmax and its derivative

- show that derivative of log-softmax is expected value

- **10. Policy gradients**
  - a) State the policy gradient equation
  - b) Give the policy

$$\pi(a|s) = \frac{e^{\theta(s,a)^{t} w}}{\Sigma_b \quad e^{\theta(s,b)^{T} w}}$$

  -

state	$\theta(s, up)_1$	$\theta(s, right)_1$	$\theta(s, right)_1$	$\theta(s, right)_1$
s = 1	3	2	3	3

  - Given a table of episodes, with rewards and steps such as (current state = 1, action = up, next_state = 4, reward = 0). There are three episodes, use the REINFORCE algorithm to calculate the weights w for state 1 of each episode.