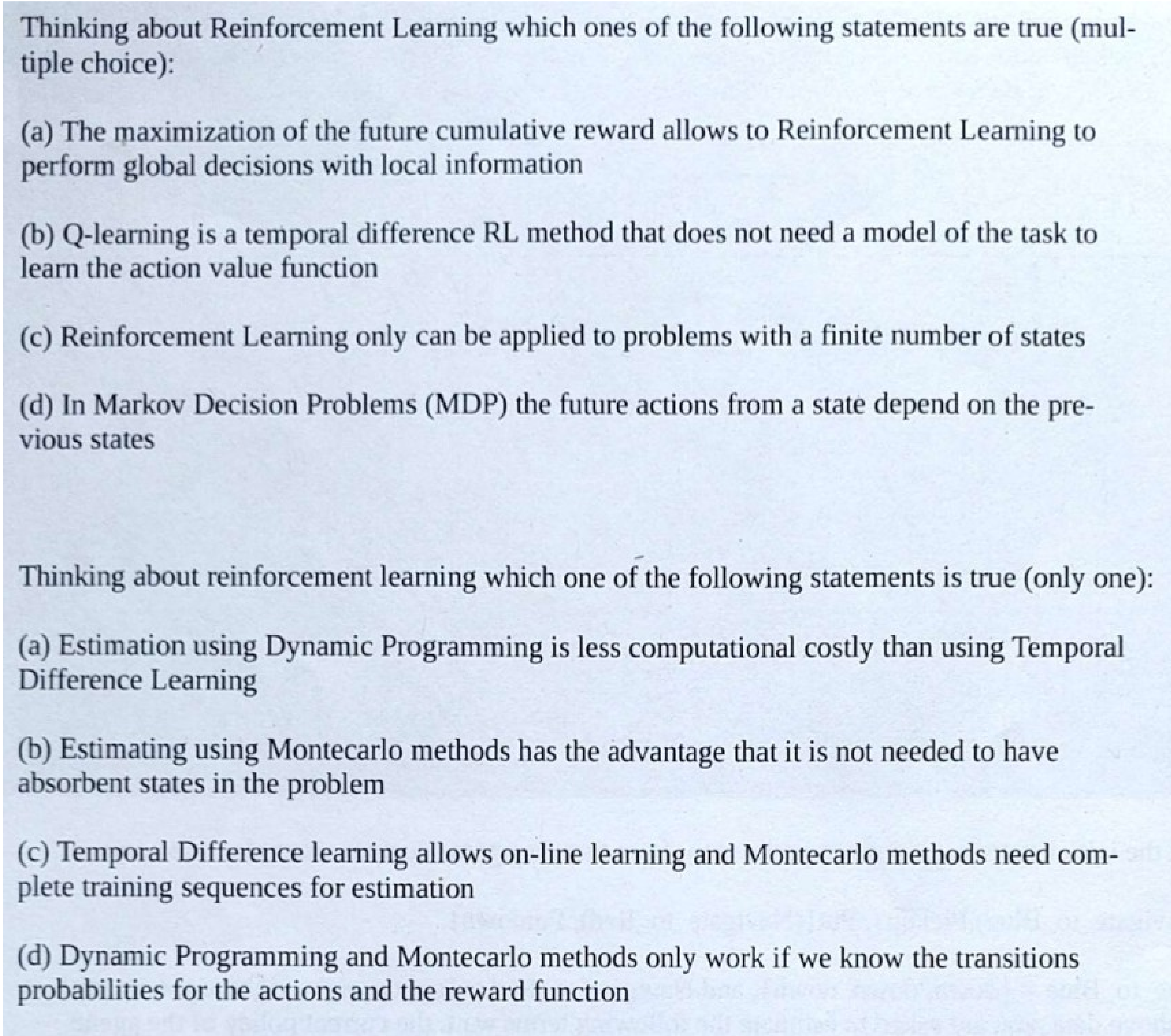
Reinforcement Learning SS17

63 points all together. I could not take any pictures so it is more or less a collection of the exam from SS16 and the questions that I could remember.

Question 1



Thinking about Reinforcement Learning which ones of the following statements are true (multiple choice):

(a) The maximization of the future cumulative reward allows to Reinforcement Learning to perform global decisions with local information

(b) Q-learning is a temporal difference RL method that does not need a model of the task to learn the action value function

(c) Reinforcement Learning only can be applied to problems with a finite number of states

(d) In Markov Decision Problems (MDP) the future actions from a state depend on the previous states

Thinking about reinforcement learning which one of the following statements is true (only one):

(a) Estimation using Dynamic Programming is less computational costly than using Temporal Difference Learning

(b) Estimating using Montecarlo methods has the advantage that it is not needed to have absorbent states in the problem

(c) Temporal Difference learning allows on-line learning and Montecarlo methods need complete training sequences for estimation

(d) Dynamic Programming and Montecarlo methods only work if we know the transitions probabilities for the actions and the reward function

-first part a)+b)

-The second part was slightly different → there was no correct answer. Here it is c)

Also there where 3 open questions:

1. Describe a Marcov Process

→ see slides for definition

2. Do VI and PI converge to the same policy?

→ not in general they but they converge to the same Value function

3. …

Reinforcement Learning SS17

Question 2

I can not remember the exact formulation of the task.

Basically there was a dice with the number as reward, so when the dice shows one the reward you get is 1 and so on.

There are two actions, either you roll the dice once more, then you will recieve no reward. Or you can choose to stop and then you get the reward of the dice rolled.

At time step h you stop to roll.

a) what is the Q-Value at time step h-1

→ calculate $Q_{h}(roll,s) = 1/6\sum_{i=1}^6 s_i = 3,5$ and $Q_{h}(stop,s)=r$

→ when you stop you get the reward of the dice rolled in state h-1 so $r = s_{h-1}$

→ when you roll again you get the transition probability of the state s_h times the value of the dice, since you know that there will be no action afterwards

b) calculate the value function for every state

→ for every state it is the max over the possible actions so $V(s_i) = \max_a (Q(stop,s_i),Q(roll,s_i))$

→ $V(s_1) =3,5$, $V(s_2) =3,5$, $V(s_3) =3,5$, $V(s_4) =4$, $V(s_5) =5$, $V(s_6) =6$

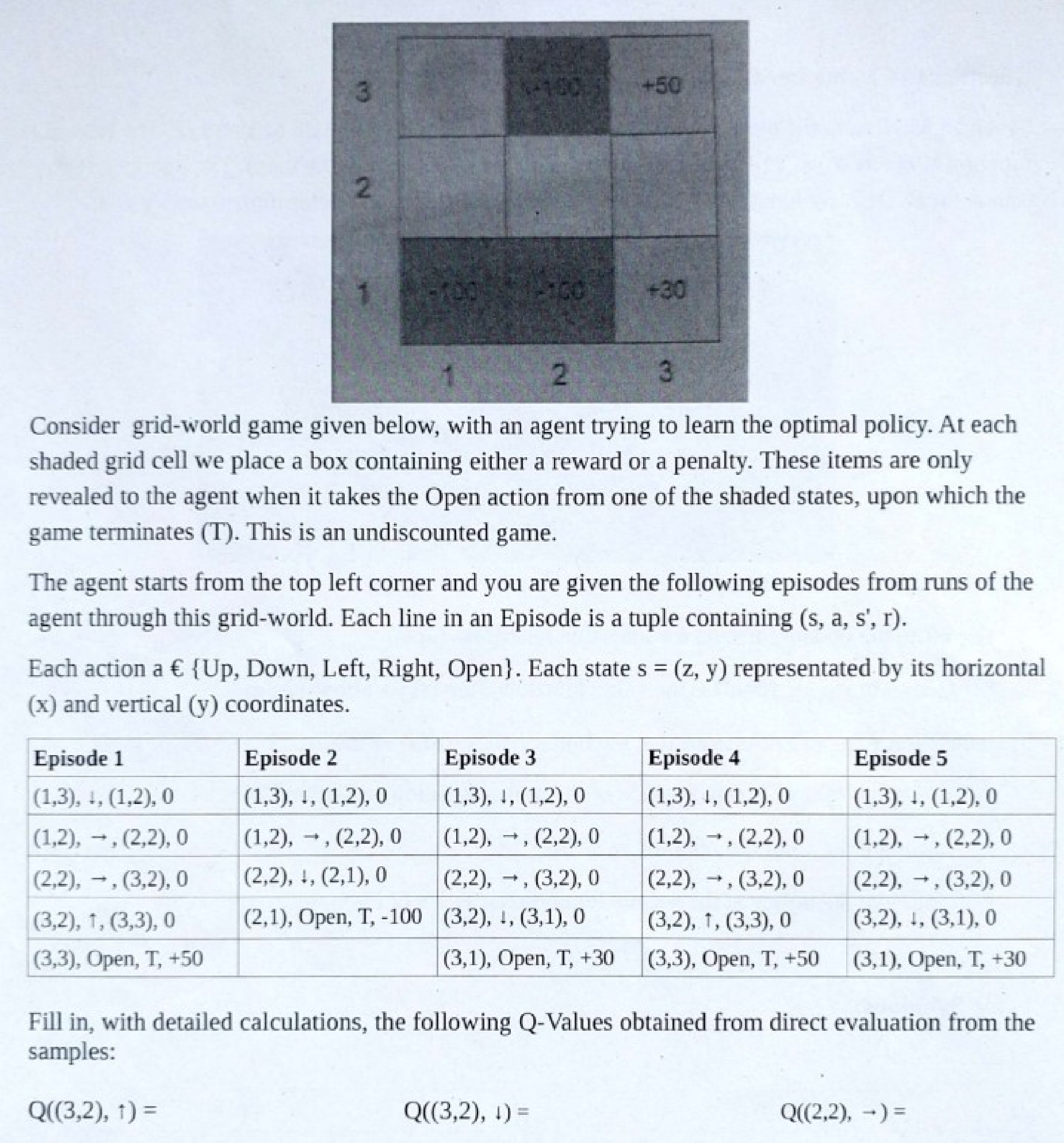
c) optimal policy?

→ roll when you are belwo 3,5, stop when you are above

d) ….

Reinforcement Learning SS17

Question 3



3		-100	+50
2			
1	-100	-100	+30
	1	2	3

Consider grid-world game given below, with an agent trying to learn the optimal policy. At each shaded grid cell we place a box containing either a reward or a penalty. These items are only revealed to the agent when it takes the Open action from one of the shaded states, upon which the game terminates (T). This is an undiscounted game.

The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r).

Each action a € {Up, Down, Left, Right, Open}. Each state s = (z, y) representated by its horizontal (x) and vertical (y) coordinates.

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), ↓, (1,2), 0	(1,3), ↓, (1,2), 0	(1,3), ↓, (1,2), 0	(1,3), ↓, (1,2), 0	(1,3), ↓, (1,2), 0
(1,2), →, (2,2), 0	(1,2), →, (2,2), 0	(1,2), →, (2,2), 0	(1,2), →, (2,2), 0	(1,2), →, (2,2), 0
(2,2), →, (3,2), 0	(2,2), ↓, (2,1), 0	(2,2), →, (3,2), 0	(2,2), →, (3,2), 0	(2,2), →, (3,2), 0
(3,2), ↑, (3,3), 0	(2,1), Open, T, -100	(3,2), ↓, (3,1), 0	(3,2), ↑, (3,3), 0	(3,2), ↓, (3,1), 0
(3,3), Open, T, +50		(3,1), Open, T, +30	(3,3), Open, T, +50	(3,1), Open, T, +30

Fill in, with detailed calculations, the following Q-Values obtained from direct evaluation from the samples:

Q((3,2), ↑) =   Q((3,2), ↓) =   Q((2,2), →) =

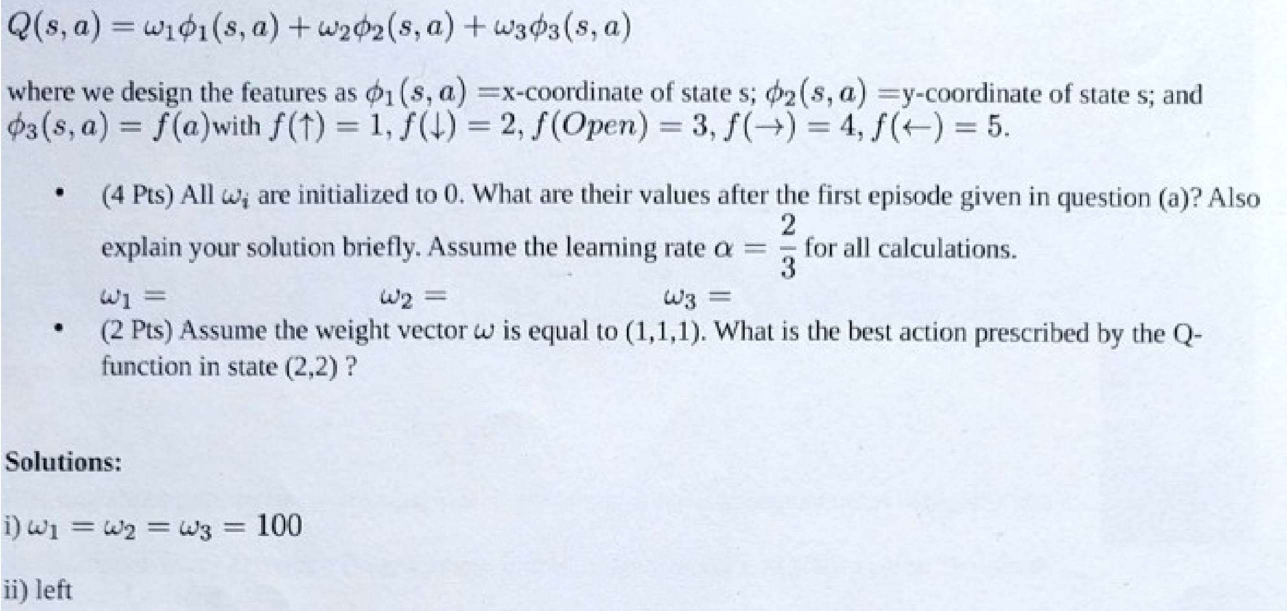
one more question was to write down the equation for computing the update of Q-Values for MC

Reinforcement Learning SS17

Question 4

(using the problem from Qustion 3)

consider Q-Learning algorithm with linear function approximation



$Q(s,a) = \omega_1\phi_1(s,a) + \omega_2\phi_2(s,a) + \omega_3\phi_3(s,a)$

where we design the features as $\phi_1(s,a)$ =x-coordinate of state s; $\phi_2(s,a)$ =y-coordinate of state s; and $\phi_3(s,a) = f(a)$with $f(\uparrow) = 1, f(\downarrow) = 2, f(Open) = 3, f(\rightarrow) = 4, f(\leftarrow) = 5.$

- (4 Pts) All ω_i are initialized to 0. What are their values after the first episode given in question (a)? Also explain your solution briefly. Assume the learning rate $\alpha = \frac{2}{3}$ for all calculations.

  $\omega_1 =$ $\omega_2 =$ $\omega_3 =$
- (2 Pts) Assume the weight vector ω is equal to (1,1,1). What is the best action prescribed by the Q-function in state (2,2) ?

Solutions:

i) $\omega_1 = \omega_2 = \omega_3 = 100$

ii) left

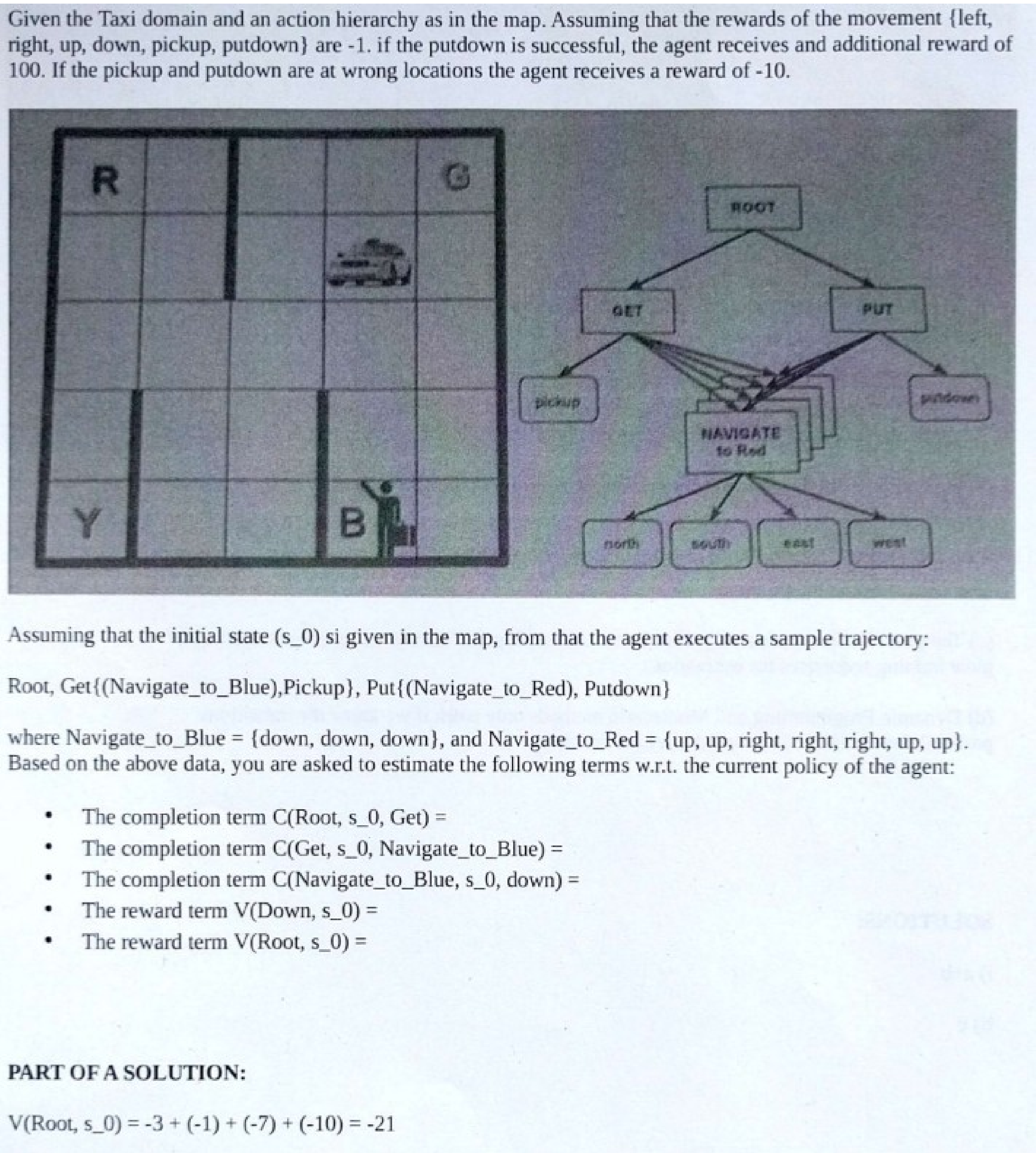
→ there was one more qustion about alpha. What properties must alpha fulfill that the algorithm converges, here you were also asked to show that this is the case.

→ then there was a forumla of a expexted SARSA algorithm give. Here you had to decide if the algorithm was on or off policy.

Since it took the same policy for the expected value as for the behaviour it was on policy

Reinforcement Learning SS17

Question 6



Given the Taxi domain and an action hierarchy as in the map. Assuming that the rewards of the movement {left, right, up, down, pickup, putdown} are -1. if the putdown is successful, the agent receives and additional reward of 100. If the pickup and putdown are at wrong locations the agent receives a reward of -10.

R
G
Y
B
ROOT
GET
PUT
pickup
putdown
NAVIGATE to Red
north
south
east
west

Assuming that the initial state (s_0) si given in the map, from that the agent executes a sample trajectory:

Root, Get{(Navigate_to_Blue),Pickup}, Put{(Navigate_to_Red), Putdown}

where Navigate_to_Blue = {down, down, down}, and Navigate_to_Red = {up, up, right, right, right, up, up}. Based on the above data, you are asked to estimate the following terms w.r.t. the current policy of the agent:

- The completion term C(Root, s_0, Get) =
- The completion term C(Get, s_0, Navigate_to_Blue) =
- The completion term C(Navigate_to_Blue, s_0, down) =
- The reward term V(Down, s_0) =
- The reward term V(Root, s_0) =

PART OF A SOLUTION:

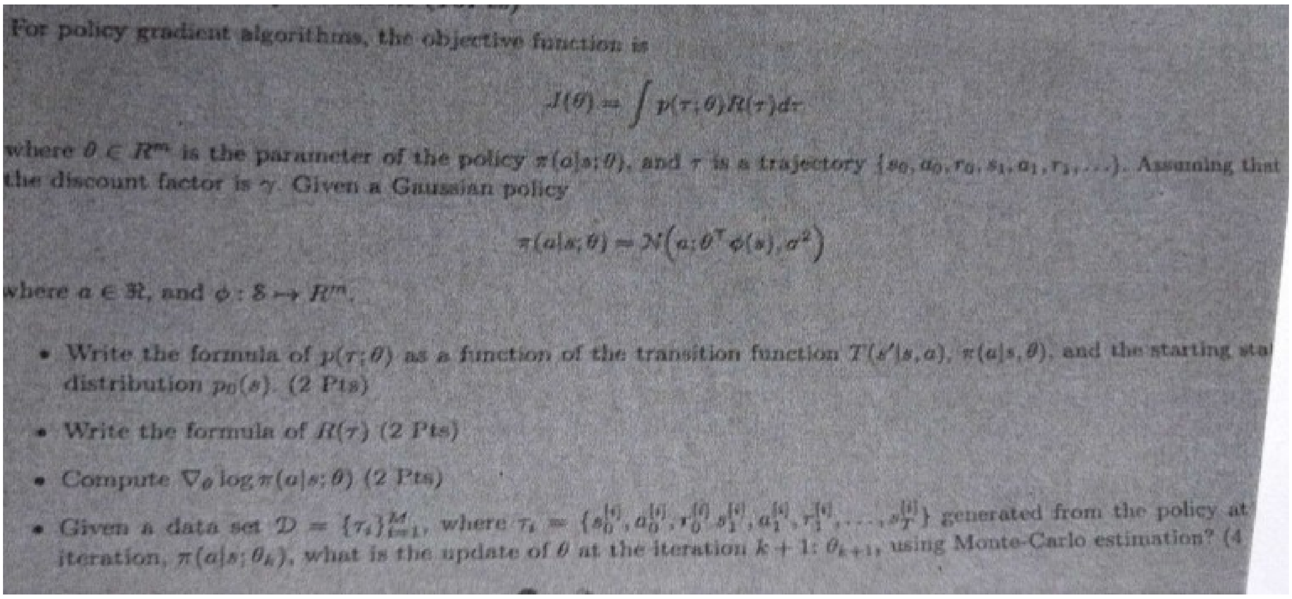
V(Root, s_0) = -3 + (-1) + (-7) + (-10) = -21

During the exam they corrected the Navigate_to_Red part to {up,up,left,left,left,up,up}

so the soulution is not negative anymore

Reinforcement Learning SS17

Question 7



For policy gradient algorithms, the objective function is

$$J(\theta) = \int p(\tau;\theta)R(\tau)d\tau$$

where $\theta \in R^m$ is the parameter of the policy $\pi(a|s;\theta)$, and τ is a trajectory $\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots\}$. Assuming that the discount factor is γ. Given a Gaussian policy

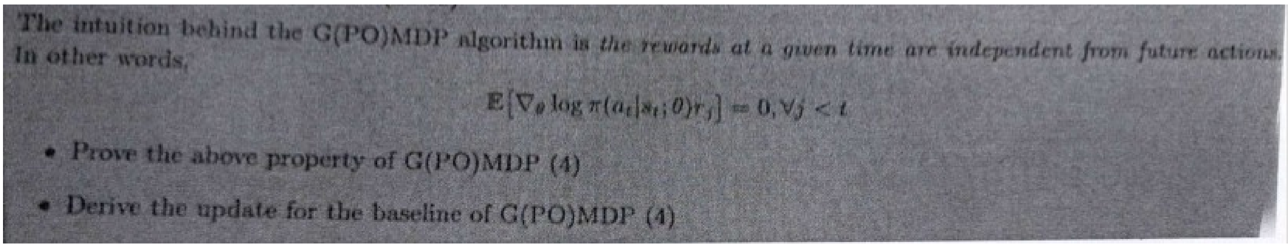
$$\pi(a|s;\theta) = \mathcal{N}\Big(a;\theta^\top\phi(s),\sigma^2\Big)$$

where $a \in \Re$, and $\phi : \mathcal{S} \mapsto R^m$.

- Write the formula of $p(\tau;\theta)$ as a function of the transition function $T(s'|s,a)$, $\pi(a|s,\theta)$, and the starting sta distribution $p_0(s)$. (2 Pts)
- Write the formula of $R(\tau)$ (2 Pts)
- Compute $\nabla_\theta \log \pi(a|s;\theta)$ (2 Pts)
- Given a data set $\mathcal{D} = \{\tau_i\}_{i=1}^M$, where $\tau_i = \{s_0^{[i]}, a_0^{[i]}, r_0^{[i]} s_1^{[i]}, a_1^{[i]}, r_1^{[i]}, \ldots, s_T^{[i]}\}$ generated from the policy at iteration, $\pi(a|s;\theta_k)$, what is the update of θ at the iteration $k+1$: θ_{k+1}, using Monte-Carlo estimation? (4

Reinforcement Learning SS17

Question 8



The intuition behind the G(PO)MDP algorithm is *the rewards at a given time are independent from future actions.* In other words,

$$\mathbb{E}\left[\nabla_\theta \log \pi(a_t|s_t;\theta) r_j\right] = 0, \forall j < t$$

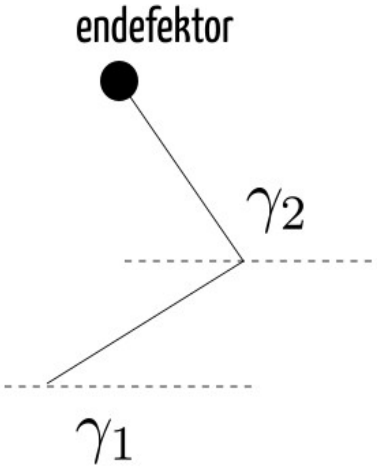
- Prove the above property of G(PO)MDP (4)
- Derive the update for the baseline of G(PO)MDP (4)

Reinforcement Learning SS17

Question 9

It was an open design Task.

The setup was like this



a) design an RL agent that can reach a goal in the 2D state space. There where motors at both joints.

b) how can you change the algorithm that for example 10 goals are reached?

c) what would you do if you are asked for a new goal every time?

→ instead of learning Q-Values, learn a model