1.

a

ValueIterationAgents assign values to states, whereas qlearning agents assign values to

combinations of a state and an action.

b

In reinforcement learning, an agent has to experimentally find out

how good actions are. If these methods would be available, it would be possible

to make an offline planning based on the results, and the entire learning would be

unnecessary

c

When a state is left, you get the reward for that state. The living reward is only earned

if you don't earn another reward, which makes sense because it would be a constant reduction

anyway, so you might as well combine it in the reward by math. A terminal state rewards 0

when it is left.

The rewards earned are only dependent on the state that is left by convention.

2.

An online planner makes choices while the actions are being executed, and uses feedback

from previous actions to choose the next action. An offline planner plans ahead, and

does not depend on realtime feedback.

3.

a

A terminal state.

b

Finite problem

c

It means that future rewards must always outweigh the living reward because there is a

possibility to end the getting of rewards if that is more favourable.

4

a

with noise = 0.2:

Discount 0 - failure

discount 0.2 - failure

discount 0.4 - failure

discount 0.6 - failure

discount 0.8 - failure

with discount = 0.8:

noise 0 - yay

noise 0.2 - failure

noise 0.4 - failure

noise 0.6 - failure

noise 0.8 - failure

b

We changed the noise to 0

c

We did this because when the noise is 0, the agent won't accidentally fall off the bridge.

This made the agent cross the bridge, because it didn't accidentally fall off.

5

a

- (1, 0.001, -5)

- (0.5, 0.3, -1)

- (1, 0.001, -0.1)

- (1, 0.3, -1)

- (42, 0, 42)

b

-The first one works because it has relatively low noise, so taking a risky part pays off.

The emmidiate reward causes the AI to try and finish the level as fast as possible,

without falling off the cliff.

-The second one has quite a bit of noise and a high discount rate, the AI takes the safe path

to the closest reward.

-The third: Because of the low noise, no discount and midly negative reward the AI takes the

dangerous path to the furthest away reward.

-The fourth: because of the somewhat present noise the safe path will be taken, the very mellow living

reward together with the discount rate cause the long route to be more rewarding.

-The fifth: 42 is the answer to everything and therefore it is correct.

No noise combined with a positive living reward and a >1 discount causes the AI to never end.

42 is the answer to everything therefore it is correct.

10

Provisional grades

==================

Question q1: 6/6

Question q2: 1/1

Question q3: 5/5

Question q4: 5/5

Question q5: 3/3

Question q6: 1/1

Question q7: 1/1

Question q8: 3/3

------------------

Total: 25/25