HW3 046203 RL

Submitters:

Submitter 1: David Valensi 342439643

Submitter 2: Yaniv Galron 206765646

Question 1: Worst Case Reward

1. Let’s define

Our goal is to find

Basis:

For

:

:

Finally, we return .

1. For a given MDP, let’s define the following policy , which takes each action with uniform probability. This policy enables every single transition (which exist in the MDP) with some positive probability:

This way, . This is right since the supremum works on every possible stationary stochastic policy and we defined some stationary stochastic policy which enables all transitions in the given MDP. It means that our will cover every possible sequence in the given MDP.

Thus, we can use the previous DP algorithm with maximum instead of minimum to find: by using the following backward for example:

Or similarly:

1. Yes, there exists a deterministic policy that attains .

Let’s denote

We define the deterministic policy:

Let’s prove that ,by induction for all t.

Basis:

Step:

Then

From the definition of :

.

1. i. The supremum is taken over all policies, so the actions are maximizing the expression, but by definition of , we consider the worst-case transition in the MDP

Thus, the DP equations are

ii. It results the following DP operator:

We must prove that

As in class, let’s take

For all state s, it holds:

(\*) As we learned in class, the difference between two function minima is smaller or equal to the maximal difference between two functions.

Similarly, with , we can prove that

Thus, is a contraction.

iii. (Bonus) is a contracting. We remember that from Banach-fixed-point Theorem, there exists a unique solution to the equation .

Thus, we can define the following stationary deterministic policy

Such that by definition we attain the optimal value :

We can write in operator notations.

.

Question 4 – Stochastic Shortest Path

1. We define for this MDP the value function as we learned in class:

\*Since terminal state is cost free, once the state 0 is reached, the best way to behave is to stay at goal state since

The Bellman Equations are

For the optimal policy, it holds:

1. For a fixed stationary policy , the Bellman operator is:

The Bellman Operator is:

1. The cost function is positive (given).

For a non-proper policy, it holds that

I.e. the optimal cost function is not finite.

1. First, we consider the SSP problem with same transitions but with costs

.

For any policy we have since .

, the optimal value from state s, holds that

Thus, defining .

Let’s write the Bellman Optimality Equation for , the optimal value from state s.

Joining a+b, we have:

Now,

The previous inequality holds for all states s. Thus, we can write

We proved that is a contraction operator with the weighted maximum norm that’s defined, and the contraction coefficient is .

Question 5:

As we learned in class, the Value Iteration algorithm produces Value functions and thus greedy policies which are not necessarily better than the previous steps policies. However, there is a convergence guarantee to the optimal .

We will show a counter example to the claim “VI algorithm produces a sequence of and greedy policy w.r.t. s.t. ”

For the following MDP:

A picture containing text, whiteboard

Description automatically generated

With

Of course, for this MDP we have

We use the VI updates:

.

We observe for this sequence that

In other words, we see twice that as needed.