Advanced AI Techniques (WS04)

Exercise sheet 11
Deadline: Thursday, 27 Jan 05

Exercise 1 (6 points)

Remember the skiing trip example from last week. [1] *Use Dynamic Programming Techniques to find the optimal policy:*

a. Implement the search for the optimal policy with the policy iteration algorithm (alternating policy evaluation and policy improvement steps). Choose the threshold $\theta = 0.001$.

b.Implement the search for the optimal policy with value iteration algorithm. Again, use $\theta = 0.001$.

c. Compare the convergence behavior of a. and b.: After how much runtime do we get the result? Which is faster and why?

d. The idea of the policy improvement (and policy determination) is to change the policy in a way that for each state, a deterministic action which yields the highest expected reward is chosen. Does this local optimization lead to a global optimization of your policy? Give a short explanation why.

Exercise 2 (4 points)

Implement the five state random walk example from the lecture with the policy that chooses one of the two actions of each state with equal probability, like shown in the figure below. Rewards are written on the arrows, and all transitions are deterministic.

A B C D E

0

[1] You plan to go skiing in a skiing region with three skiing slopes A (red), B (red) and C (blue), and three lifts X (to the top of slope A), Y (to slope B), Z (to slope C). Going down a red slope, you get twice as much pleasure as going down the blue one. From A, you can either run down to lift Y or Z. From B, you can either go to lift X or to lift Y. From C, you can go down to lift X or Z. There is always the option of waiting on top of a slope (without getting any pleasure from it). Assume that you are a bad skier: The transition probabilities are 0.6 for slope B, and 0.75 for slopes A and C. In all other cases, you end in hospital for the rest of the time (each time step in hospital rewards the negative of the pleasure of running down the blue slope).

a. Use the Every-Visit Monte Carlo algorithm with $\alpha = 1, \gamma = 1$. Simulate 1000 episodes.

b. Use the TD(0) (Temporal Difference) algorithm with $\alpha = 1, \gamma = 1$. Simulate 2000 episodes.

**. If you want, implement in another version of you program the First-Visit MC algorithm, and/or experiment with implementations of $TD(\lambda)$ for different values for λ. What is the difference? You can also choose smaller values for α. What is the effect? (This part of the exercise is voluntary and does not count for points.)*

Exercise 3 (2 points)

What are the differences (advantages, disadvantages, prerequisites) of the Monte Carlo algorithm and the TD algorithm? Give an example for a situation when you would rather use Monte Carlo, and when you would prefer the $TD(0)$ algorithm?