Optimization and Uncertainty

Summer term 2023

Prof. Dr. Martin Hoefer
Conrad Schecker, Lisa Wilhelmi

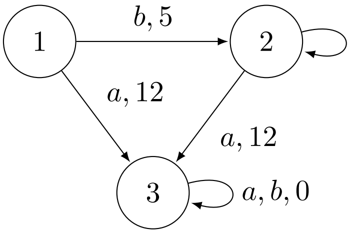
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Institute of Computer Science
Algorithms and Complexity

Assignment 9

Issued: 27.06.2023
Due: 04.07.2023, **10:00h**

Exercise 9.1 *Value Iteration versus Policy Iteration* (2 + 2 + 2 points)

Consider a Markov decision process with states $\mathcal{S} = \{1, 2, 3\}$ and actions $\mathcal{A} = \{a, b\}$ which is depicted below. The state transitions are deterministic. The numbers in the edge labels are the respective rewards. Assume an infinite time horizon with discount factor $\gamma = \frac{1}{2}$.



a) Derive an optimal Markovian policy π^* and $V^*(s)$ for all $s \in \mathcal{S}$.

b) Perform the first six steps of value iteration starting with initial vector $v^{(0)} = (0, 0, 0)$.

c) Starting from the policy that always performs action a, apply policy iteration until convergence.

Exercise 9.2 *Value Iteration with Caution* (5 points)

Consider a more cautious version of value iteration for MDPs with infinite time horizon with state set $\mathcal{S}$ and action set $\mathcal{A}$. It uses the operator t' which is defined by $t'(v)_s = \mu \cdot t(v)_s + (1 - \mu) \cdot v_s$, for all states $s \in \mathcal{S}$, where t is the value iteration defined in the lecture and $\mu \in (0, 1)$ is an arbitrary parameter.

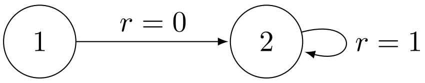
Show that t' converges to the unique fixed point of t.

Exercise 9.3 *Gittins Index* (2 + 2 points)

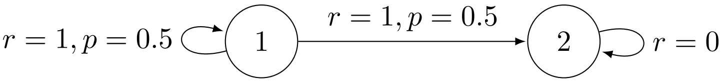
Consider the following instances for the Markovian Single-Armed Bandit problem with charges $\lambda \geq 0$. Let r denote the reward of a transition when action `play` is chosen, and p denotes the probability that the respective transition occurs ($p = 1$ unless stated otherwise). If `pause` is chosen, no transition occurs and the reward is zero in this round. At each iteration step, the probability that the process terminates is $\gamma \in (0, 1)$.

For each of the single-armed bandits, derive the Gittins indices of all states.

a)



b)



Exercise 9.4 *Upper bound for UCB1* (5 points)

When the distributions of the n arms have similar means, the upper bound for UCB1 shown in lecture is very poor. Show that an upper bound on the expected regret, which is independent of the means, is given by $5\sqrt{nT \ln T} + 4n$, where T is the total number of rounds.
Hint: Divide the arms according to whether their expected value deviates more or less than $\varepsilon = \sqrt{\frac{n \ln T}{T}}$ *from the expected value of the best arm* μ_{i^*}.