universität freiburg

Machine Learning for Stochastics **Summer semester 2025**

Lecturer: Prof. Dr. Thorsten Schmidt
Assistance: Simone Pavarana, MSc.

Tutorial 2

Exercise 1 (3 Points)**.**
Let $N = 2$ be the planning horizon of a Markov Decision Process (MDP) with state space $E = \{0,1\}$ and action space $A = D_n(x) = \{0,1\}$ for all $x \in E$. The transition probability is defined as:

$$Q(x' \mid x,a) = \begin{cases} 1, & \text{if } a = x', \\ 0, & \text{otherwise.} \end{cases}$$

The reward functions are given by $r_n(x,a) = a$ for $n = 0,1$, and the terminal reward is $g_N(x) = x$.

(a) Use dynamic programming to show that the policy $\pi^* = (f_0^*,f_1^*)$, defined by $f_0^*(x) = f_1^*(x) = 1$ for all $x \in E$, is an optimal policy.

(b) Is the policy π^* defined in part (a) the unique optimal policy? If so, prove uniqueness. If not, provide an alternative policy that yields the same value function.

Exercise 2 (3 Points)**.**
Consider the following simple card game: A dealer uncovers, one by one, the cards from a well-shuffled deck that initially contains b_0 black cards and r_0 red cards. At any point, the player may choose to stop the game. If the next card revealed is black (resp. red), the player wins (resp. loses) 1 Euro.

(a) What is the expected gain at the initial time if the player stops the game?

(b) Formulate the problem as a stationary Markov Decision Process (MDP). Clearly define the state space, action space, transition probabilities, and reward function.

(c) Let $g(b,r)$ denote the terminal reward and let E denote the state space. Prove that for all states $(b,r) \in E$ such that $b + r \geq 2$, the value function satisfies:

$$(\mathcal{T}g)(b,r) = g(b,r),$$

where $\mathcal{T}$ is the Bellman operator defined by

$$(\mathcal{T}v)(x) = \max_{a \in A} \left\{ r(x,a) + \beta \int_E v(y) Q(dy \mid x,a) \right\}.$$

(d) Conclude that the value function $J_{b_0+r_0}$ coincides with the expected gain computed in point (a), and thus equals the trivial solution.

Exercise 3 (3 Points)**.**

Consider the following homogeneous Markov Decision Process. The state space and action space are both given by $E = A = \{1,2\}$. Rewards are discounted by a factor $\beta \in (0,1)$, and the one-stage rewards are defined as:

$$r(1,1) = 6, \quad r(2,1) = -3, \quad r(1,2) = 4, \quad r(2,2) = -5.$$

The terminal reward is given by $g(1) = 105$, $g(2) = 100$.
The transition probabilities under actions $a = 1$ and $a = 2$ are represented by the matrices:

$$Q(\cdot \mid \cdot,1) = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}, \quad Q(\cdot \mid \cdot,2) = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}.$$

(a) Let $v : E \to \mathbb{R}$ be any measurable function, and define $\Delta v := v(1) - v(2)$. Show that for all $a \in A$,

$$\mathcal{L}v(1,a) - \mathcal{L}v(2,a) = 9 + (0.1\beta)\Delta v.$$

(b) Let J_n denote the value function at time n, and define $\Delta J_n := J_n(1) - J_n(2)$. Show that for all $n \in \mathbb{N}$,

$$\Delta J_n = 9 \sum_{k=0}^{n-1} (0.1\beta)^k + (0.1\beta)^n \cdot 5.$$

Hint: From part (a), you obtain the recurrence relation

$$\Delta J_{n+1} = 9 + (0.1\beta)\Delta J_n.$$

This is a linear difference equation. Solve it by first finding the general solution to the homogeneous equation, and then adding a particular solution for the non-homogeneous part.