Frankfurt University of Applied Sciences
Fachbereich 2: Informatik und Ingenieurwissenschaften

Grundlagen adaptiver Wissenssysteme (SS2025)

Prof. Dr. Thomas Gabel

Aufgabenblatt 3

Aufgabe 7: SKP-Probleme vs. Diskontierung

Betrachten Sie den in Abbildung 1 dargestellten MDP. Alle Transitionen sind deterministisch; ein Diskontierungsfaktor von $\gamma < 1$ wird verwendet.

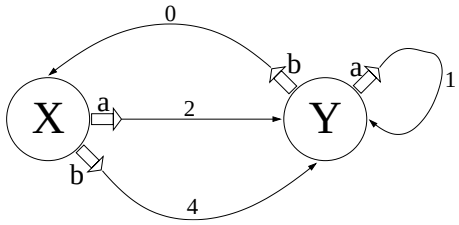


Abbildung 1: Zwei-Zustands-MDP

(a) Begründen Sie, warum es sinnvoll ist, für den dargestellten MDP Diskontierung zu verwenden.

(b) Gibt es eine Klasse von Problemstellungen, für die es problematisch wäre, diese als SKP-Probleme zu modellieren?

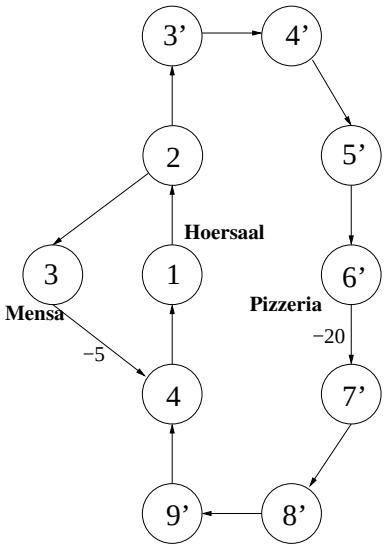
(c) Wie viele Strategien gibt es in diesem MDP? Geben Sie für alle möglichen Strategien π und für alle Zustände i die erwarteten Pfadkosten $V^{\pi}(i)$ als Funktion von γ an und ermitteln Sie die optimale Strategie.

(d) Diskontierte Probleme können als Spezialfall von SKP-Problemen angesehen werden. Konstruieren Sie einen MDP mit Terminalzustand, der zu dem in Abbildung 1 MDP äquivalent ist.

Aufgabe 8: Mensa oder Pizzeria

Betrachten Sie den 11-Zustands-MDP, dessen Übergangsgraph in Abbildung gegeben ist. Alle Transitionen sind deterministisch. Der Agent erfährt Kosten in Höhe von -5, wenn er von die Mensa verlässt und zum Hörsaal zurückkehrt und Kosten von -20, wenn er aus der Pizzeria zurückkehrt. Alle anderen Übergange sind kostenfrei.

(a) Wie viele verschiedene deterministische Strategien gibt es in diesem MDP?

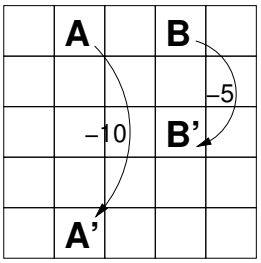


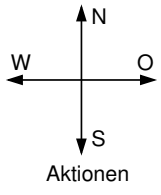
(b) Geben Sie für jede Strategie einen Ausdruck für die im Zustand 1 zu erwartenden Kosten an (Diskontierung mit $\gamma < 1$ wird vorausgesetzt).

(c) Für welche Werte von γ wird die optimale Strategie den Agenten in die Mensa führen? Für welche Werte eher in die Pizzeria?

Aufgabe 9: Bellman-Gleichung

In der in Abbildung dargestellten Gitterwelt sind die Zellen des Gitters die Zustände. In jeder Zelle sind vier Aktionen möglich (Norden, Süden, Osten, Westen), die den Agenten deterministisch in die jeweilige Nachbarzelle des Gitters bewegen. Aktionen, bei denen der Agent das Gitter verlassen würde, führen zu keinem Zellenwechsel, verursachen aber direkte Kosten in Höhe von 1. Alle anderen Aktionen sind mit keinen Kosten verbunden, mit Ausnahme der Zustände A und B. Jede in A ausgeführte Aktion bewegt den Agenten nach A' unter Kosten von -10, jede in B ausgeführte Aktion bewegt den Agenten nach B' unter Kosten von -5.





−3.3	−8.8	−4.4	−5.3	−1.5
−1.5	−3.0	−2.3	−1.9	−0.5
−0.1		−0.7	−0.4	0.4
1.0	0.4	0.4	0.6	1.2
1.9	1.3	1.2	1.4	

(a) Wir nehmen an, dass der Agent in allen Zuständen alle Aktionen mit gleicher Wahrscheinlichkeit auswählt. Im rechten Teil von Abbildung ist die zugehörige Wertfunk-

tion V^π für einen Diskontierungsfaktor von $\gamma = 0.9$ eingetragen. Ermitteln Sie die fehlenden Einträge.

(b) Zeigen Sie beispielhaft für den Zustand $s_{3,3}$ in der Mitte des Gitters mit $V^\pi(s_{3,3}) = -0.7$, dass die Bellman-Gleichung bezüglich aller seiner Nachbarzustände erfüllt ist.

(c) Erläutern Sie, warum im Zustand B die zu erwartenden Pfadkosten niedriger sind als die direkten Kosten.

(d) Ermitteln Sie die optimalen Pfadkosten der Zustände A und B unter der optimalen Strategie, d.h. $V^\star(A)$ und $V^\star(B)$. Zeichnen Sie ferner die optimale Strategie in das Gitter ein.

(e) Im Beispiel der Gitterwelt werden Kosten größer null vergeben, wenn der Agent in eine Wand hineinläuft, Kosten kleiner null für das Erreichen der Zielpunkte und Nullkosten in allen anderen Situationen. Sind die Vorzeichen der direkten Kosten von Bedeutung oder aber nur die Abstände zwischen ihnen?

Aufgabe 10: Würfeln

Aus der Pro7-Sendung "Schlag den Star" ist das Spiel "Würfeln" bekannt. Zwei Spieler würfeln gegeneinander, die jeweils erzielten Punkte werden summiert. Ein Spieler darf solange würfeln und weitere Punkte sammeln, wie er möchte. Aber erst, wenn er den Würfel an den anderen Spieler abgibt, werden ihm die bis dahin erwürfelten Punkte auf sein Konto gutgeschrieben. Außerdem gilt: Sobald eine 6 fällt, muss der Würfel ebenfalls an den anderen Spieler abgegeben werden, wobei in diesem Fall die bis dahin erwürfelten Punkte verfallen und nicht dem Konto des Spielers gutgeschrieben werden. Der würfelnde Spieler muss also in jedem Zeitschritt entscheiden, ob er den Würfel abgibt und die bislang erzielten Punkte einstreicht, oder ob er weiterwürfelt. Das Spiel gewinnt, wer als Erster 50 oder mehr Punkte auf seinem Konto hat.

In dieser Aufgabe betrachten wir zunächst eine *Vereinfachung* der Aufgabenstellung, bei der nur ein einzelner Spieler beteiligt ist. Das Spiel endet zudem, sobald das erste Mal eine 6 fällt oder sobald der Spieler sich entscheidet aufzuhören und die bislang erwürfelten Punkte seinem Konto gutzuschreiben. Die 50-Punkte-Grenze entfällt also; stattdessen ist es das Ziel, möglichst viele Punkte auf seinem Konto zu haben.

(a) Modellieren Sie die vereinfachte Version des Spieles als SKP-Problem. Begründen Sie Ihre Definition der direkten Kosten.

(b) Zeichnen Sie den Zustandsübergangsgraphen (ausschnittsweise) für das Problem.

(c) Wir bezeichnen die Aktion des Spielers, die ihn den Würfel abgeben (und damit das Spiel beenden) lässt, im Folgenden als GIBAB. Spieler A verfolgt die Strategie, die Aktion GIBAB zu wählen, sobald er mindestens 18 Punkte erwürfelt hat. Spieler B hingegen gibt bereits ab, wenn er 10 Punkte erzielt hat. Bewerten Sie die beiden Strategien, indem Sie V^{π_A} und V^{π_B} ermitteln.

(d) Schließen Sie an Ihre in Teilaufgabe (c) durchgeführten Strategiebewertungen (policy evaluation) jeweils einen Strategieverbesserungsschritt (policy improvement) an. Ermitteln Sie die erwarteten Pfadkosten der so verbesserten Strategie. Sind die nach dem Strategieverbesserungsschritt erhaltenen Strategien optimal?