Frankfurt University of Applied Sciences
Fachbereich 2: Informatik und Ingenieurwissenschaften

Grundlagen adaptiver Wissenssysteme (SS2025)

Prof. Dr. Thomas Gabel

Aufgabenblatt 4

Aufgabe 11: Strategiebewertung

Betrachten Sie den in Abbildung 1 dargestellten MDP, in dem alle Transitionen, mit Ausnahme der auf Aktion a im Zustand 3 folgenden Zustandsübergänge, deterministisch sind.

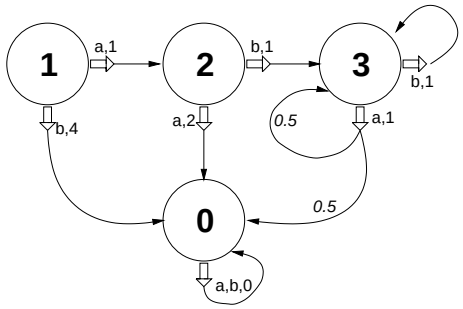

Abbildung 1: Vier-Zustands-MDP

(a) Konvergiert der Wertiterationsalgorithmus, wenn er auf das gegebene Problem ohne Diskontierung angewendet wird?

(b) Gegeben sei eine Strategie $\pi : S \to A$ mit:
$\pi(0) = a$, $\pi(1) = b$, $\pi(2) = b$, $\pi(3) = a$. Bewerten Sie die Strategie π, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion V_0^π starten, der für alle Zustände Nullkosten annimmt.

(c) Nehmen Sie an, Sie würden die Aktualisierung im Rahmen der Strategiebewertung unendlich oft fortführen. Schätzen Sie die resultierende Pfadkostenfunktion V_k^π für $k \to \infty$ ab.

(d) Extrahieren Sie eine gierige Strategie π' von der in Teilaufgabe (c) ermittelten Pfadkostenfunktion V^π.

(e) Handelt es sich bei π' um die optimale Strategie? Falls nein, wie viele weitere Iterationen des Strategieiterationsverfahrens sind notwendig, um die optimale Strategie zu erhalten?

Aufgabe 12: Monte Carlo und TD(λ): Anwendung

Betrachten Sie den in Abbildung 2 dargestellten MDP, in dem alle Aktionen (Bewegungen in jeweils vorgegebene Richtung) mit einer Wahrscheinlichkeit von 0.8 ausgeführt werden. Mit der Restwahrscheinlichkeit wird der Agent zufällig in eine der anderen benachbarten Gitterzellen bewegt. Alle Zustandsübergange verursachen Kosten von 1, lediglich bei Erreichen des Kiosk (Terminalzustand $s_{2,4}$) erfährt der Agent negative Kosten in Höhe von -10, und bei einem Absturz von der Klippe (Terminalzustände $s_{3,1} \dots s_{3,5}$) Kosten in Höhe von 100. Der Agent startet, wie in der Abbildung angedeutet, jeweils im Startzustand $s_{2,1}$.

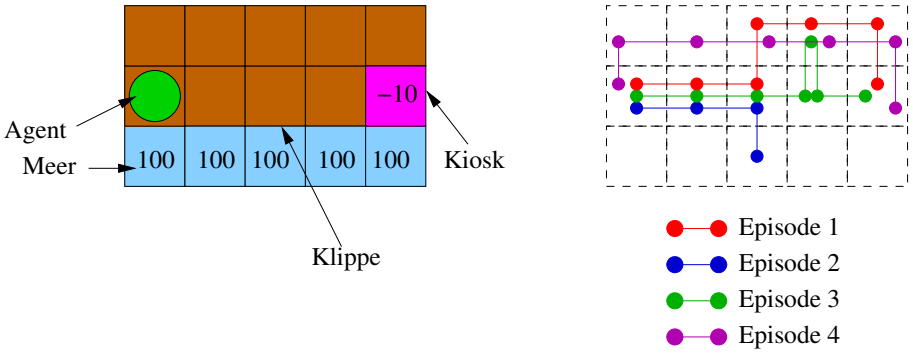

Abbildung 2: Klippen-MDP

(a) Berechnen Sie mittels Monte-Carlo-Strategieevaluation auf Basis der dargestellten Episoden 1 bis 3 die Funktion $V_3(i)$ für alle Zustände i. Setzen Sie $\alpha = \frac{1}{m}$ sowie $V_0(i) = 0 \forall i$.

(b) Betrachten Sie nun die in magenta dargestellte Episode 4. Geben Sie für alle Zustände dieser Episode den temporalen Differenzfehler auf Basis ihrer in Teilaufgabe (a) ermittelten Kostenfunktion V_3 an.

(c) Ermitteln Sie nun mit Hilfe des TD(λ)-Algorithmus, für $\lambda = 0$, $\lambda = 0.5$ und $\lambda = 1.0$, die erwarteten Pfadkosten $V^\pi(s_{2,1})$ auf Basis der ersten drei Episoden.

Aufgabe 13: Q-Lernen

Betrachten Sie die in Abbildung 3 dargestellte deterministische Gitterwelt, in der ein absorbierender Terminalzustand G existiert. Für Übergänge ins Ziel betragen die direkten Kosten -10, für alle anderen Transitionen 0; ein Diskontierungsfaktor von $\gamma = 0.8$ kommt zur Anwendung.

(a) Ermitteln Sie die optimale Kostenfunktion V^* für diesen MDP sowie eine optimale Strategie $\pi^* : S \to A$.

(b) Geben Sie die Q-Werte $Q(s, a)$ für alle Zustands-Aktions-Paare an.

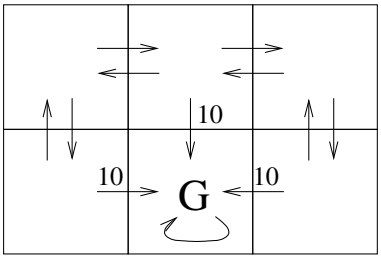

Abbildung 3: Gitterwelt

Startend mit einer zu null initialisierten $\hat{Q}$-Funktion, soll im Folgenden der Q-Lernalgorithmus für den Gitterwelt-MDP angewendet werden. Der Agent startet in der Gitterzelle unten links und bewegt sich im Uhrzeigersinn durch das Gitter, bis er nach 5 Zustandsübergängen im Zielzustand G anlangt, womit die aktuelle Episode beendet wird.

(c) Erläutern Sie, welche $\hat{Q}$-Werte im Laufe jener Episode aktualisiert worden sind und geben Sie die veränderten Q-Werte an (Lernrate 1.0).

(d) Ermitteln Sie die Werte der $\hat{Q}$-Funktion, nachdem der Agent zwei weitere (identische) Episoden durchlaufen hat.