# Grundlagen adaptiver Wissenssysteme

## Übungen zur Vorlesung

Prof. Dr. Thomas Gabel
Frankfurt University of Applied Sciences
Faculty of Computer Science and Engineering
tgabel@fb2.fra-uas.de

# Aufgabenblatt 4

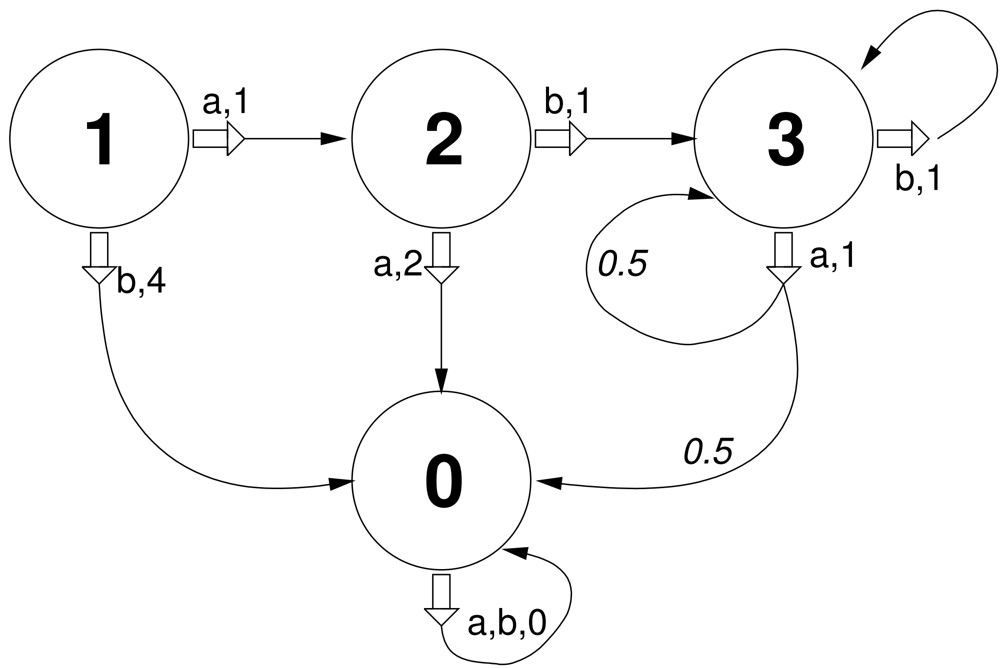
1. **Aufgabenblatt 4 – Übung 11**
2. Aufgabenblatt 4 – Übung 12
3. Aufgabenblatt 4 – Übung 13

# Aufgabe 11: Strategiebewertung

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

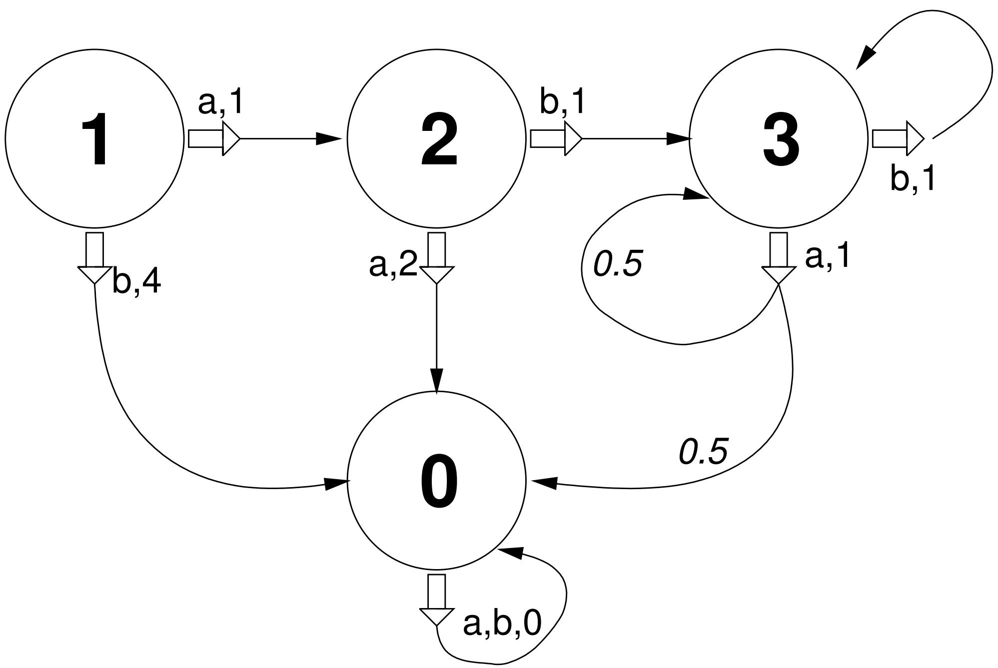
# Aufgabe 11: Strategiebewertung

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



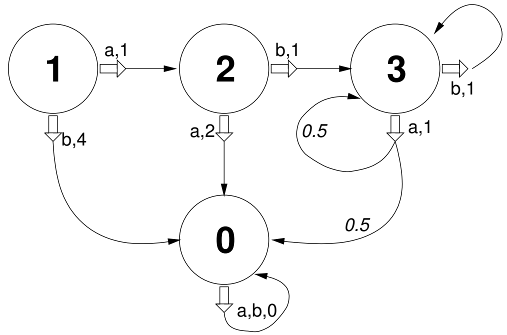
# Aufgabe 11: Strategiebewertung

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



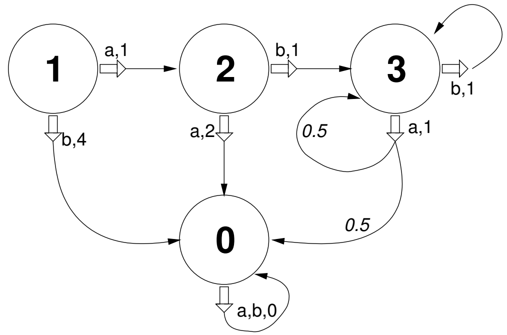

(a) Konvergiert der Wertiterationsalgorithmus, wenn er auf das gegebene Problem ohne Diskontierung angewendet wird?

# Aufgabe 11: Strategiebewertung




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

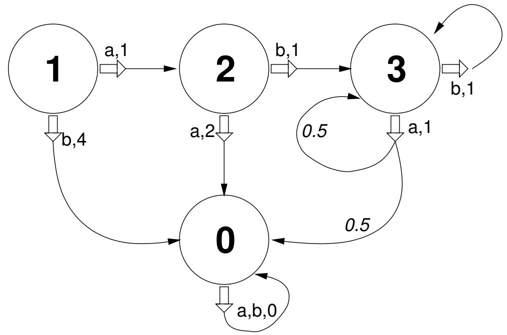
# Aufgabe 11: Strategiebewertung




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

- *Fakt:* Es handelt sich um ein SKP-Problem, da ein absorbierender Zustand (Zustand “0”) existiert, in dem keine Kosten anfallen.
- Aus der Vorlesung ist bekannt, dass für die Konvergenz des Wertiterationsverfahrens die folgenden Voraussetzungen erfüllt sein müssen.

# Aufgabe 11: Strategiebewertung




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

- *Fakt:* Es handelt sich um ein SKP-Problem, da ein absorbierender Zustand (Zustand “0”) existiert, in dem keine Kosten anfallen.
- Aus der Vorlesung ist bekannt, dass für die Konvergenz des Wertiterationsverfahrens die folgenden Voraussetzungen erfüllt sein müssen.
  - Es muss mind. eine erfüllende Strategie existieren (SKP-V1).
  - Für alle nicht erfüllenden Strategien muss es mindestens einen Zustand geben, für den die Pfadkosten unendlich sind (SKP-V2).

# Aufgabe 11: Strategiebewertung

- SKP-V1: Es muss mindestens eine erfüllende Strategie existieren.
- Def.: Eine Strategie $\pi$ ist erfüllend, wenn der Agent aus allen Zuständen nach maximal $n$ Schritten in den Terminalzustand mit positiver Wahrscheinlichkeit übergehen kann.

# Aufgabe 11: Strategiebewertung

- SKP-V1: Es muss mindestens eine erfüllende Strategie existieren.
- Def.: Eine Strategie $\pi$ ist erfüllend, wenn der Agent aus allen Zuständen nach maximal $n$ Schritten in den Terminalzustand mit positiver Wahrscheinlichkeit übergehen kann.

$$\varrho_\pi := \max_{i=1,\dots,n} P(s_n \neq 0 | \pi, s_0 = i) < 1$$

bzw.

$$\overline{\varrho}_\pi := \min_{i=1,\dots,n} P(s_n = 0 | \pi, s_0 = i) > 0$$

# Aufgabe 11: Strategiebewertung

- SKP-V1: Es muss mindestens eine erfüllende Strategie existieren.
- Def.: Eine Strategie $\pi$ ist erfüllend, wenn der Agent aus allen Zuständen nach maximal $n$ Schritten in den Terminalzustand mit positiver Wahrscheinlichkeit übergehen kann.

$$\varrho_\pi := \max_{i=1,\ldots,n} P(s_n \neq 0 | \pi, s_0 = i) < 1$$

bzw.

$$\overline{\varrho}_\pi := \min_{i=1,\ldots,n} P(s_n = 0 | \pi, s_0 = i) > 0$$

- Beh.: $\pi$ mit $\pi(0) = *$, $\pi(1) = b$, $\pi(2) = a$, $\pi(3) = a$ ist erfüllend.

# Aufgabe 11: Strategiebewertung

- SKP-V1: Es muss mindestens eine erfüllende Strategie existieren.
- Def.: Eine Strategie $\pi$ ist erfüllend, wenn der Agent aus allen Zuständen nach maximal $n$ Schritten in den Terminalzustand mit positiver Wahrscheinlichkeit übergehen kann.

$$\varrho_\pi := \max_{i=1,\dots,n} P(s_n \neq 0 | \pi, s_0 = i) < 1$$

bzw.

$$\overline{\varrho}_\pi := \min_{i=1,\dots,n} P(s_n = 0 | \pi, s_0 = i) > 0$$

- Beh.: $\pi$ mit $\pi(0) = *$, $\pi(1) = b$, $\pi(2) = a$, $\pi(3) = a$ ist erfüllend.
- Es gilt: $P(s_n \neq 0 | \pi, s_0 = i)$ ist 0 für $i = 1$, 0 für $i = 2$ sowie $(\frac{1}{2})^4$ für $i = 3$.

# Aufgabe 11: Strategiebewertung

- SKP-V1: Es muss mindestens eine erfüllende Strategie existieren.
- Def.: Eine Strategie $\pi$ ist erfüllend, wenn der Agent aus allen Zuständen nach maximal $n$ Schritten in den Terminalzustand mit positiver Wahrscheinlichkeit übergehen kann.

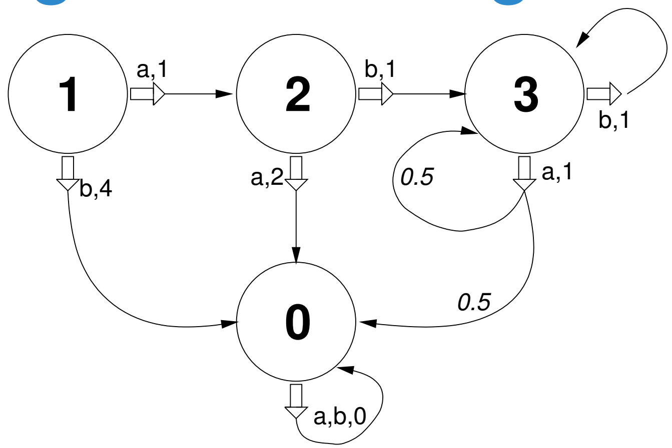
$$\varrho_\pi := \max_{i=1,\dots,n} P(s_n \neq 0 | \pi, s_0 = i) < 1$$

bzw.

$$\overline{\varrho}_\pi := \min_{i=1,\dots,n} P(s_n = 0 | \pi, s_0 = i) > 0$$

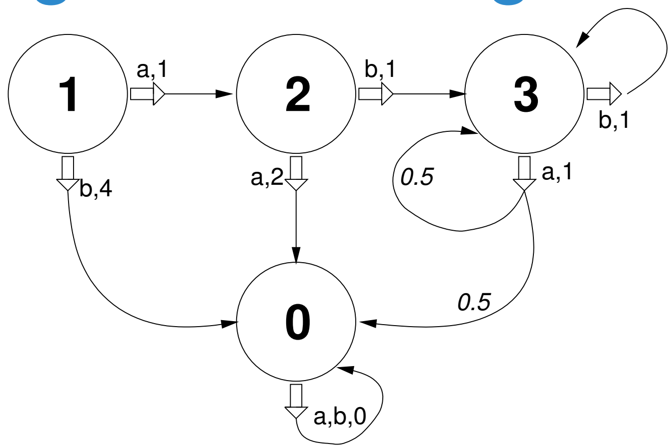
- Beh.: $\pi$ mit $\pi(0) = *$, $\pi(1) = b$, $\pi(2) = a$, $\pi(3) = a$ ist erfüllend.
- Es gilt: $P(s_n \neq 0 | \pi, s_0 = i)$ ist 0 für $i = 1$, 0 für $i = 2$ sowie $(\frac{1}{2})^4$ für $i = 3$.
- Also ist $\varrho_\pi = (\frac{1}{2})^4 < 1$ und damit SKP-V1 erfüllt.

# Aufgabe 11: Strategiebewertung



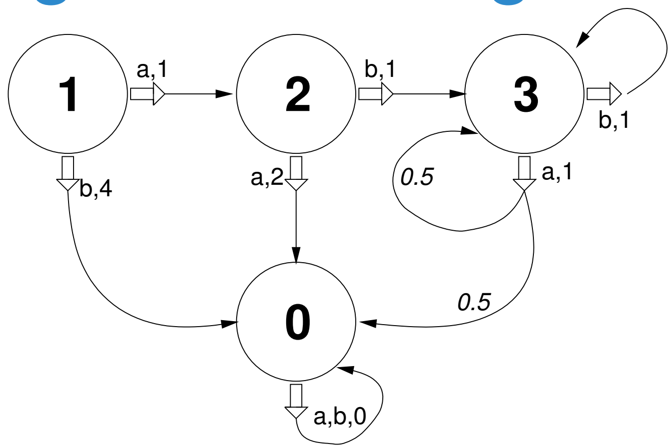

- SKP-V2: Für alle nicht erfüllenden Strategien muss es mindestens einen Zustand geben, für den die Pfadkosten unendlich sind.

# Aufgabe 11: Strategiebewertung



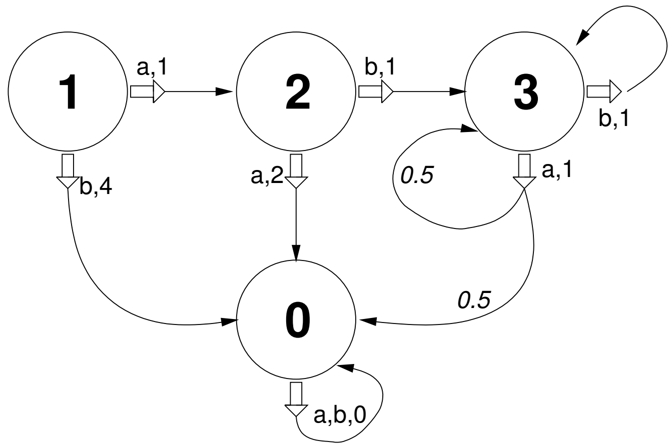

- SKP-V2: Für alle nicht erfüllenden Strategien muss es mindestens einen Zustand geben, für den die Pfadkosten unendlich sind.
- Im gegebenen MDP gibt es nur eine einzige nicht erfüllende Strategie:

# Aufgabe 11: Strategiebewertung



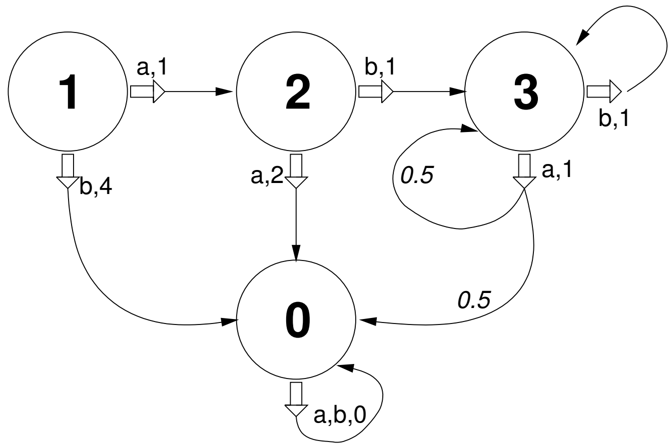

- SKP-V2: Für alle nicht erfüllenden Strategien muss es mindestens einen Zustand geben, für den die Pfadkosten unendlich sind.
- Im gegebenen MDP gibt es nur eine einzige nicht erfüllende Strategie: $\pi(1) = a$, $\pi(2) = b$, $\pi(3) = b$.
- Wir betrachten den Zustand $i = 3$ und ermitteln die Pfadkosten $V^\pi(3)$.
- Es gilt $V^\pi(3) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N} \gamma^t c(s_t, \pi(s_t)) | s_0 = 3\right]$

# Aufgabe 11: Strategiebewertung




- Wegen $\gamma = 1$ und $p_{33}(\pi(3)) = 1.0$ erhalten wir $V^{\pi}(3) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N} 1\right] = \infty$

# Aufgabe 11: Strategiebewertung




- Wegen $\gamma = 1$ und $p_{33}(\pi(3)) = 1.0$ erhalten wir $V^{\pi}(3) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N} 1\right] = \infty$
- Damit ist auch SKP-V2 erfüllt.
  $\Rightarrow$ Value Iteration konvergiert.

# Aufgabe 11: Strategiebewertung

(b) Gegeben sei eine Strategie $\pi : S \rightarrow A$ mit:
$\pi(0) = a$, $\pi(1) = b$, $\pi(2) = b$, $\pi(3) = a$.

Bewerten Sie die Strategie $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- geg.: ABBA-Strategie $\pi$ und $V_0^\pi(i) = 0$ für alle $i$, keine Diskontierung, d.h. $\gamma = 1$

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- geg.: ABBA-Strategie $\pi$ und $V_0^\pi(i) = 0$ für alle $i$, keine Diskontierung, d.h. $\gamma = 1$
- Zur Erinnerung: Aktualisierungsregel

$$V_k(i) := \sum_{j=0}^{n} p_{ij}(\pi(i)) \cdot (c(i, \pi(i)) + \gamma V_{k-1}(j))$$

# Aufgabe 11: Strategiebewertung

(b) Gegeben sei eine Strategie $\pi : S \rightarrow A$ mit:
$\pi(0) = a$, $\pi(1) = b$, $\pi(2) = b$, $\pi(3) = a$.

Bewerten Sie die Strategie $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Erste Iteration ($k = 1$):
  - $V_1(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)

# Aufgabe 11: Strategiebewertung

(b) Gegeben sei eine Strategie $\pi : S \rightarrow A$ mit: $\pi(0) = a$, $\pi(1) = b$, $\pi(2) = b$, $\pi(3) = a$.

Bewerten Sie die Strategie $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Erste Iteration ($k = 1$):
  - $V_1(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)
  - $V_1(1) = 1 \cdot (4 + 1 \cdot 0) = 4$ ($j$ ist hier 0)
  - $V_1(2) = 1 \cdot (1 + 1 \cdot 0) = 1$ ($j$ ist hier 3)
  - $V_1(3) = 0.5 \cdot (1 + 1 \cdot 0) + 0.5 \cdot (1 + 1 \cdot 0) = 0.5 + 0.5 = 1$
    ($j$ ist hier entweder 0 oder 3, jeweils mit einer Wahrscheinlickeit von 0.5)

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Erste Iteration ($k = 1$):
  - $V_1(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)
  - $V_1(1) = 1 \cdot (4 + 1 \cdot 0) = 4$ ($j$ ist hier 0)
  - $V_1(2) = 1 \cdot (1 + 1 \cdot 0) = 1$ ($j$ ist hier 3)
  - $V_1(3) = 0.5 \cdot (1 + 1 \cdot 0) + 0.5 \cdot (1 + 1 \cdot 0) = 0.5 + 0.5 = 1$
    ($j$ ist hier entweder 0 oder 3, jeweils mit einer Wahrscheinlickeit von 0.5)
  - Das Ergebnis der Aktualisierung ist $V_1$.

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Zweite Iteration ($k = 2$):
  - $V_2(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Zweite Iteration ($k = 2$):
  - $V_2(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)
  - $V_2(1) = 1 \cdot (4 + 1 \cdot 0) = 4$ ($j$ ist hier 0)
  - $V_2(2) = 1 \cdot (1 + 1 \cdot 1) = 2$ ($j$ ist hier 3)
  - $V_2(3) = 0.5 \cdot (1 + 1 \cdot 1) + 0.5 \cdot (1 + 1 \cdot 0) = 1 + 0.5 = 1.5$
    ($j$ ist hier entweder 0 oder 3, jeweils mit einer Wahrscheinlickeit von 0.5)
  - Das Ergebnis der Aktualisierung ist $V_2$.

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Dritte Iteration ($k = 3$):
  - $V_3(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)

# Aufgabe 11: Strategiebewertung

(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 $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Dritte Iteration ($k = 3$):
  - $V_3(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)
  - $V_3(1) = 1 \cdot (4 + 1 \cdot 0) = 4$ ($j$ ist hier 0)
  - $V_3(2) = 1 \cdot (1 + 1 \cdot 1.5) = 2.5$ ($j$ ist hier 3)
  - $V_3(3) = 0.5 \cdot (1 + 1 \cdot 1.5) + 0.5 \cdot (1 + 1 \cdot 0) = 1.25 + 0.5 = 1.75$ ($j$ ist hier entweder 0 oder 3, jeweils mit einer Wahrscheinlickeit von 0.5)
  - Das Ergebnis der Aktualisierung ist $V_3$;

# Aufgabe 11: Strategiebewertung

(b) Gegeben sei eine Strategie $\pi : S \rightarrow A$ mit:
$\pi(0) = a$, $\pi(1) = b$, $\pi(2) = b$, $\pi(3) = a$.

Bewerten Sie die Strategie $\pi$, indem Sie den Aktualisierungsschritt im Rahmen der Strategiebewertung drei Mal anwenden (also drei Iterationen durchführen) und mit einer Pfadkostenfunktion $V_0^\pi$ starten, der für alle Zustände Nullkosten annimmt.

- Dritte Iteration ($k = 3$):
  - $V_3(0) = 1 \cdot (0 + 1 \cdot 0) = 0$ ($j$ ist hier 0)
  - $V_3(1) = 1 \cdot (4 + 1 \cdot 0) = 4$ ($j$ ist hier 0)
  - $V_3(2) = 1 \cdot (1 + 1 \cdot 1.5) = 2.5$ ($j$ ist hier 3)
  - $V_3(3) = 0.5 \cdot (1 + 1 \cdot 1.5) + 0.5 \cdot (1 + 1 \cdot 0) = 1.25 + 0.5 = 1.75$ ($j$ ist hier entweder 0 oder 3, jeweils mit einer Wahrscheinlickeit von 0.5)
  - Das Ergebnis der Aktualisierung ist $V_3$; Resultat ist: $V_3(0) = 0$, $V_3(1) = 4$, $V_3(2) = 2.5$, $V_3(3) = 1.75$.

# Aufgabe 11: Strategiebewertung

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

- Im Grenzfall (d.h. im Fall der Konvergenz der Folge der $V_k$-Funktionen gegen $V^\pi$) gilt:

# Aufgabe 11: Strategiebewertung

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

- Im Grenzfall (d.h. im Fall der Konvergenz der Folge der $V_k$-Funktionen gegen $V^\pi$) gilt:

$$\begin{aligned} V^\pi(3) &= 0.5 \cdot (1 + \gamma V^\pi(3)) + 0.5 \cdot (1 + \gamma \cdot 0) \\ &= 0.5 + 0.5 \cdot V^\pi(3) + 0.5 \\ \Leftrightarrow 0.5 \cdot V^\pi(3) &= 1 \\ \Leftrightarrow V^\pi(3) &= 2 \end{aligned}$$

- Damit ergibt sich für die anderen Zustände:

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

# Aufgabe 11: Strategiebewertung

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

- Im Grenzfall (d.h. im Fall der Konvergenz der Folge der $V_k$-Funktionen gegen $V^\pi$) gilt:

$$\begin{aligned} V^\pi(3) &= 0.5 \cdot (1 + \gamma V^\pi(3)) + 0.5 \cdot (1 + \gamma \cdot 0) \\ &= 0.5 + 0.5 \cdot V^\pi(3) + 0.5 \\ \Leftrightarrow 0.5 \cdot V^\pi(3) &= 1 \\ \Leftrightarrow V^\pi(3) &= 2 \end{aligned}$$

- Damit ergibt sich für die anderen Zustände:
  - $V^\pi(2) = 1 \cdot (1 + \gamma V^\pi(3)) = 1 + 2 = 3$
  - $V^\pi(1) = 1 \cdot (1 + \gamma V^\pi(0)) = 4 + 0 = 4$

# Aufgabe 11: Strategiebewertung

(d) Extrahieren Sie eine gierige Strategie $\pi'$ von dem in Teilaufgabe (c) ermittelten Kostenvektor $V^{\pi}$.

# Aufgabe 11: Strategiebewertung

(d) Extrahieren Sie eine gierige Strategie $\pi'$ von dem in Teilaufgabe (c) ermittelten Kostenvektor $V^{\pi}$.

- Zur Erinnerung:
  Def.: $\pi_{k+1} = \arg\min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a) \, (c(i, a) + \gamma V^{\pi_k}(j))$

# Aufgabe 11: Strategiebewertung

(d) Extrahieren Sie eine gierige Strategie $\pi'$ von dem in Teilaufgabe (c) ermittelten Kostenvektor $V^\pi$.

- Zur Erinnerung:
  Def.: $\pi_{k+1} = \arg\min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a) \left(c(i, a) + \gamma V^{\pi_k}(j)\right)$
- Damit:
  - $\pi'(0) = \arg\min_{u \in \{a,b\}} 1 \cdot (0 + 1 \cdot 0) = \arg\min_{u \in \{a,b\}} 0 = a$ oder $b$

# Aufgabe 11: Strategiebewertung

(d) Extrahieren Sie eine gierige Strategie $\pi'$ von dem in Teilaufgabe (c) ermittelten Kostenvektor $V^{\pi}$.

- Zur Erinnerung:
  Def.: $\pi_{k+1} = \arg\min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)\,(c(i,a) + \gamma V^{\pi_k}(j))$
- Damit:
  - $\pi'(0) = \arg\min_{u \in \{a,b\}} 1 \cdot (0 + 1 \cdot 0) = \arg\min_{u \in \{a,b\}} 0 = a$ oder $b$
  -

$$
\begin{aligned}
\pi'(1) &= \arg\min_{u \in \{a,b\}} \begin{cases} \text{u=a:} & 1 \cdot (1 + \gamma V^{\pi}(2)) \\ \text{u=b:} & 1 \cdot (4 + \gamma V^{\pi}(0)) \end{cases} \\
&= \arg\min_{u \in \{a,b\}} \begin{cases} \text{u=a:} & 4 \\ \text{u=b:} & 4 \end{cases} \\
&= a \text{ bzw. } b
\end{aligned}
$$

# Aufgabe 11: Strategiebewertung

- Damit (Forts.):
  -

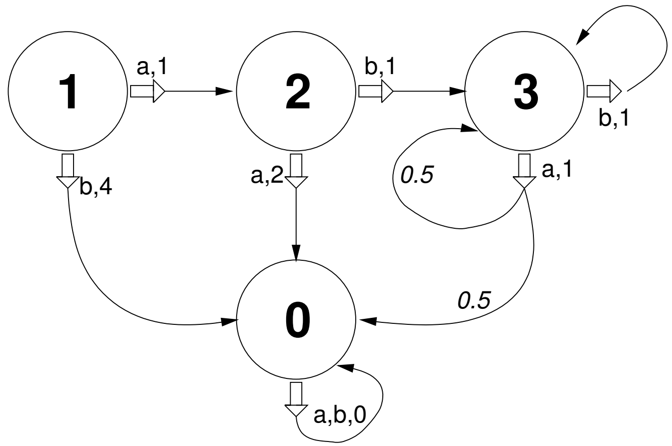
$$
\begin{aligned}
\pi'(2) &= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=b:} & 1 \cdot (1 + \gamma V^{\pi}(3)) \\ \text{u=a:} & 1 \cdot (2 + \gamma V^{\pi}(0)) \end{cases} \\
&= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=b:} & 3 \\ \text{u=a:} & 2 \end{cases} \\
&= a
\end{aligned}
$$

# Aufgabe 11: Strategiebewertung

- Damit (Forts.):
  - $$\begin{aligned}\pi'(2) &= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=b:} & 1\cdot(1+\gamma V^\pi(3)) \\ \text{u=a:} & 1\cdot(2+\gamma V^\pi(0)) \end{cases} \\ &= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=b:} & 3 \\ \text{u=a:} & 2 \end{cases} \\ &= a \end{aligned}$$
  - $$\begin{aligned}\pi'(3) &= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=a:} & 0.5\cdot(1+\gamma V^\pi(3)) \\ & \quad +0.5\cdot(1+\gamma V^\pi(0)) \\ \text{u=b:} & 1\cdot(1+\gamma V^\pi(3)) \end{cases} \\ &= \arg\min_{u\in\{a,b\}} \begin{cases} \text{u=a:} & 1.5+0.5=2 \\ \text{u=b:} & 4 \end{cases} \end{aligned}$$

# Aufgabe 11: Strategiebewertung

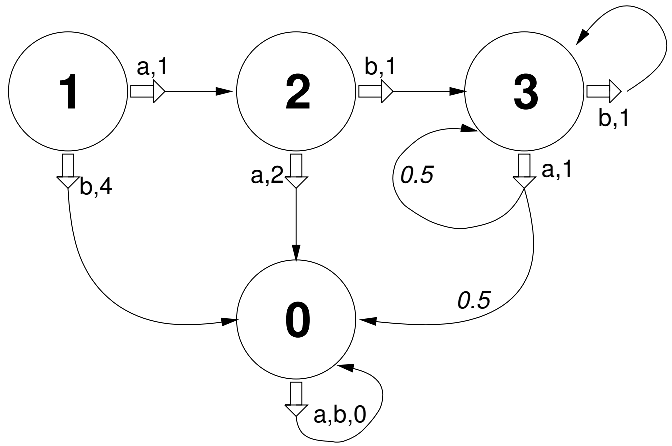
(d) Handelt es sich bei $\pi'$ um die optimale Strategie? Falls nein, wie viele weitere Iterationen des Strategieiterationsverfahrens sind notwendig, um die optimale Strategie zu erhalten?




- Die optimalen Strategie $\pi^*$ lässt den Agenten im Zustand 3 die Aktion

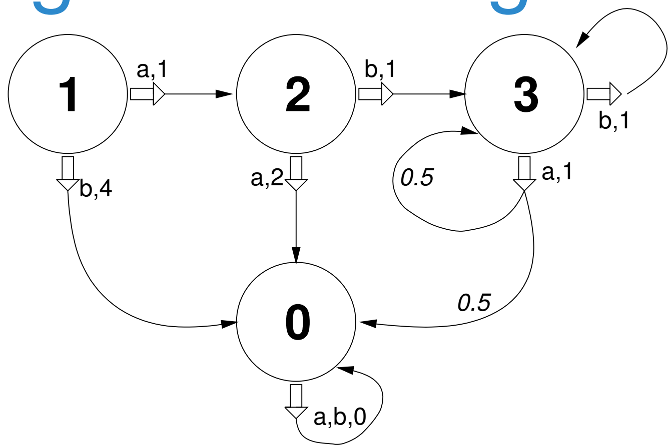
# Aufgabe 11: Strategiebewertung

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



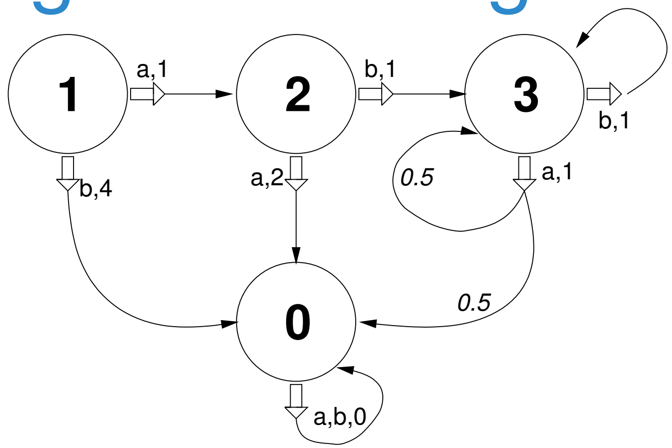

- Die optimalen Strategie $\pi^*$ lässt den Agenten im Zustand 3 die Aktion *a* wählen.
- Dementsprechend (vgl. Teilaufgabe (c)) gilt $V^\star(3) = 2$.

# Aufgabe 11: Strategiebewertung



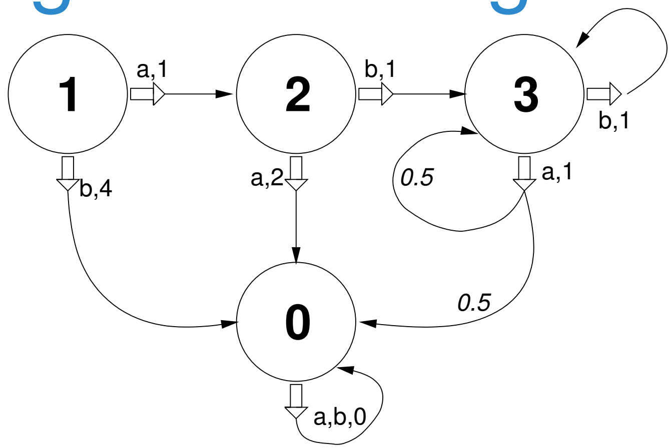

- Damit ist im Zustand 2 die optimale Aktion

# Aufgabe 11: Strategiebewertung




- Damit ist im Zustand 2 die optimale Aktion $\pi^\star(2) = a$, und es gilt dann $V^\star(2) = 2$.
- Schließlich ist die optimale Aktion im Zustand 1 die Aktion

# Aufgabe 11: Strategiebewertung



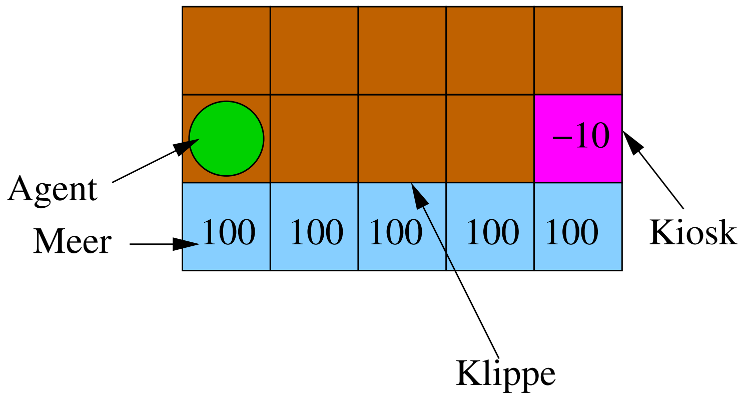

- Damit ist im Zustand 2 die optimale Aktion $\pi^\star(2) = a$, und es gilt dann $V^\star(2) = 2$.
- Schließlich ist die optimale Aktion im Zustand 1 die Aktion $\pi^\star(1) = a$ mit $V^\star(1) = 3$.
- Wenn wir davon ausgehen, dass der arg min-Operator bei gleichem Wert für mehrere $u$ stets die “erste” mögliche Aktion wählt, so erhalten wir $\pi'(0) = a$, $\pi'(1) = a$, $\pi'(2) = a$, $\pi'(3) = a$.
- Dies entspricht einer optimalen Strategie $(\pi^*(0, 1, 2, 3) = (\star, a, a, a))$.

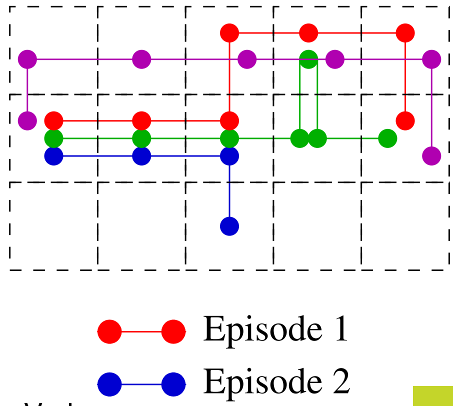
# Aufgabenblatt 4

1. Aufgabenblatt 4 – Übung 11
2. **Aufgabenblatt 4 – Übung 12**
3. Aufgabenblatt 4 – Übung 13

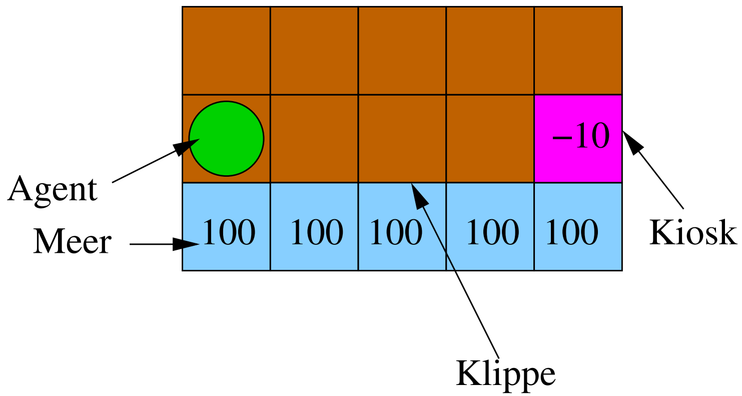
# Aufgabe 12: Monte Carlo und TD($\lambda$)

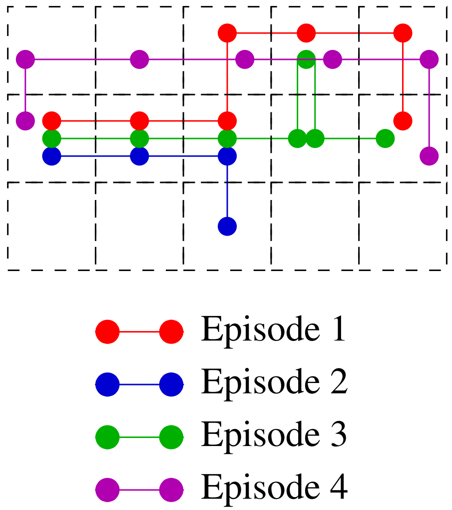
Betrachten Sie den in der Abbildung dargestellten MDP, in dem alle Aktionen (Bewegungen in jeweils vorgegebene Richtung) mit einer Wahrscheinlichkeit von 0.8 ausgeführt werden. Mit der Restwahrscheinlichkeit wird der Agent zufällig in eine der anderen benachbarten Gitterzellen bewegt. Alle Zustandsübergange 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 $s_{2,1}$.



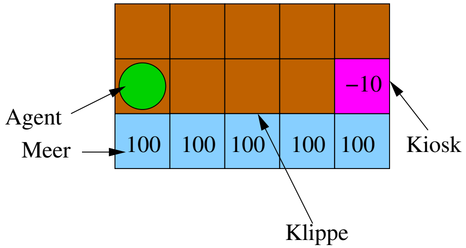
# Aufgabe 12: Monte Carlo und TD($\lambda$)

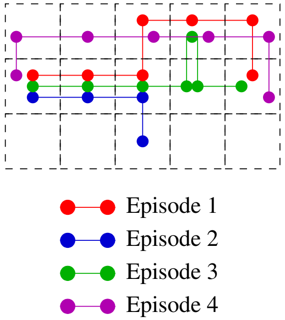



(a) Berechnen Sie mittels Monte-Carlo-Strategieevaluation auf Basis der dargestellten Episoden 1 bis 3 die Funktion $V_3(i)$ für alle Zustände $i$. Setzen Sie $\alpha = \frac{1}{m}$ sowie $V_0(i) = 0 \forall i$.

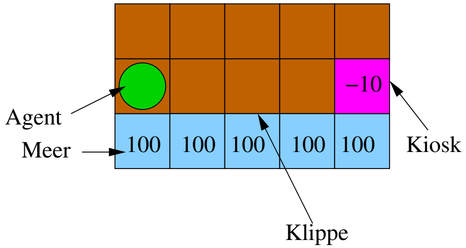
# Aufgabe 12: Monte Carlo und TD($\lambda$)

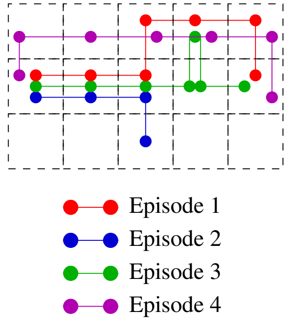



- Beginn: $V_0(i) = 0$ für alle $i \in S$
- 1.Episode, d.h. $t = 1$:

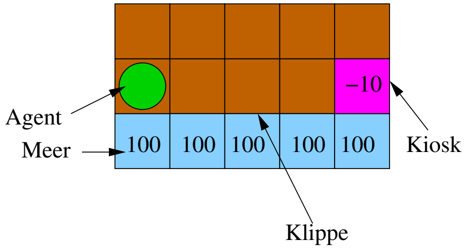
# Aufgabe 12: Monte Carlo und TD($\lambda$)

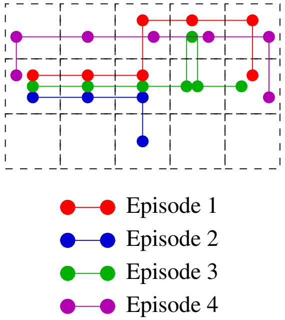



- Beginn: $V_0(i) = 0$ für alle $i \in S$
- 1.Episode, d.h. $t = 1$:
  - $g(i, t) = c(s_0) + c(s_1) + c(s_2) + \cdots + c(s_{N-1})$ mit $s_0 = i = s_{2,1}$

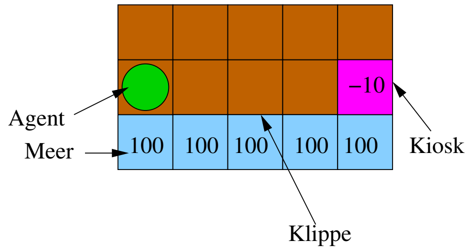
# Aufgabe 12: Monte Carlo und TD($\lambda$)

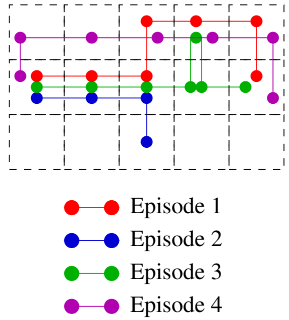



- Beginn: $V_0(i) = 0$ für alle $i \in S$
- 1.Episode, d.h. $t = 1$:
  - $g(i, t) = c(s_0) + c(s_1) + c(s_2) + \cdots + c(s_{N-1})$ mit $s_0 = i = s_{2,1}$
  - Damit ergeben sich für die Pfadkosten in der ersten Trajektorie:
    - $g(s_0, 1) = g(s_{2,1}, 1) = c(s_{2,1}, s_{2,2}) + c(s_{2,2}, s_{2,3}) + c(s_{2,3}, s_{1,3}) + c(s_{1,3}, s_{1,4}) + c(s_{1,4}, s_{1,5}) + c(s_{1,5}, s_{2,5})$
    - $g(s_0, 1) =$

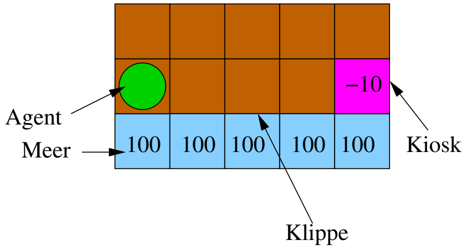
# Aufgabe 12: Monte Carlo und TD($\lambda$)

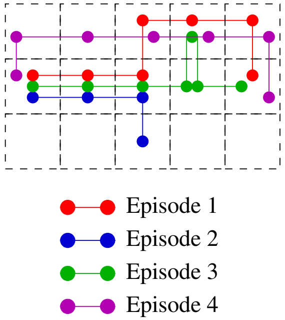



- Beginn: $V_0(i) = 0$ für alle $i \in S$
- 1.Episode, d.h. $t = 1$:
  - $g(i,t) = c(s_0) + c(s_1) + c(s_2) + \cdots + c(s_{N-1})$ mit $s_0 = i = s_{2,1}$
  - Damit ergeben sich für die Pfadkosten in der ersten Trajektorie:
    - $g(s_0,1) = g(s_{2,1},1) = c(s_{2,1},s_{2,2}) + c(s_{2,2},s_{2,3}) + c(s_{2,3},s_{1,3}) + c(s_{1,3},s_{1,4}) + c(s_{1,4},s_{1,5}) + c(s_{1,5},s_{2,5})$
    - $g(s_0,1) = 1 + 1 + 1 + 1 + 1 + (-10) = -5$

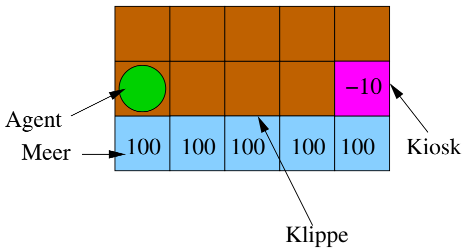
# Aufgabe 12: Monte Carlo und TD($\lambda$)

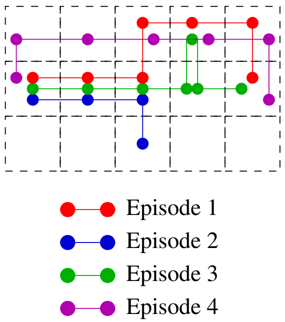



- Damit ergeben sich für die Pfadkosten in der ersten Trajektorie:
  - $g(s_1, 1) = g(s_{2,2}, 1) =$
    $c(s_{2,2}, s_{2,3}) + c(s_{2,3}, s_{1,3}) + c(s_{1,3}, s_{1,4}) + c(s_{1,4}, s_{1,5}) + c(s_{1,5}, s_{2,5})$
  - $g(s_1, 1) = 1 + 1 + 1 + 1 + (-10) = -6$

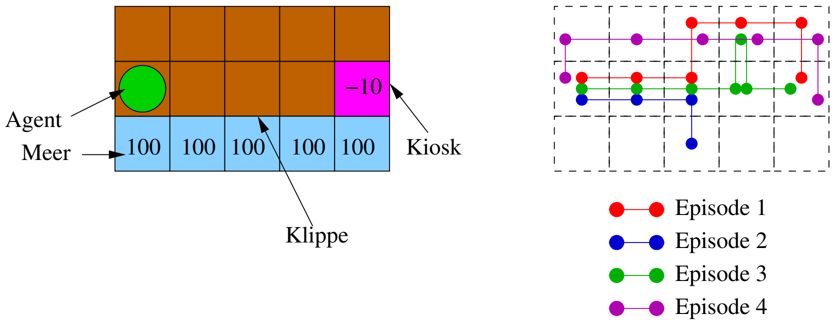
# Aufgabe 12: Monte Carlo und TD($\lambda$)



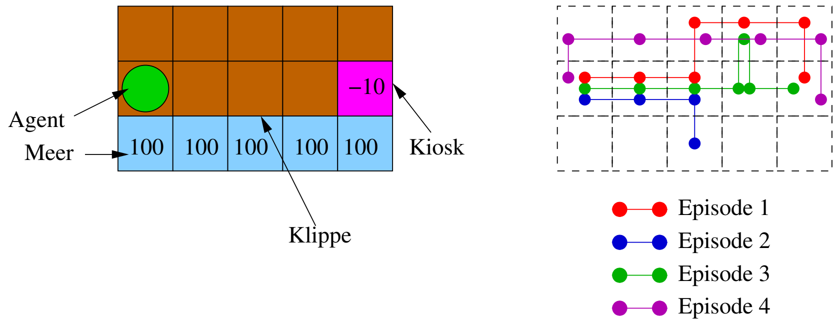

- Damit ergeben sich für die Pfadkosten in der ersten Trajektorie:
  - $g(s_1, 1) = g(s_{2,2}, 1) =$
    $c(s_{2,2}, s_{2,3}) + c(s_{2,3}, s_{1,3}) + c(s_{1,3}, s_{1,4}) + c(s_{1,4}, s_{1,5}) + c(s_{1,5}, s_{2,5})$
  - $g(s_1, 1) = 1 + 1 + 1 + 1 + (-10) = -6$
  - Und weiter in Analogie:
    - $g(s_2, 1) = g(s_{2,3}, 1) = 1 + 1 + 1 + (-10) = -7$

# Aufgabe 12: Monte Carlo und TD($\lambda$)



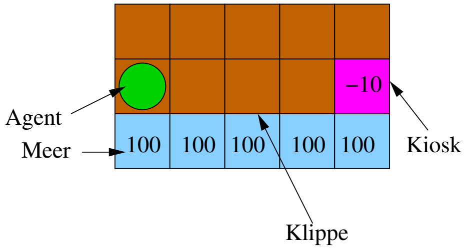

- Damit ergeben sich für die Pfadkosten in der ersten Trajektorie:
  - $g(s_1, 1) = g(s_{2,2}, 1) =$
    $c(s_{2,2}, s_{2,3}) + c(s_{2,3}, s_{1,3}) + c(s_{1,3}, s_{1,4}) + c(s_{1,4}, s_{1,5}) + c(s_{1,5}, s_{2,5})$
  - $g(s_1, 1) = 1 + 1 + 1 + 1 + (-10) = -6$
  - Und weiter in Analogie:
    - $g(s_2, 1) = g(s_{2,3}, 1) = 1 + 1 + 1 + (-10) = -7$
    - $g(s_3, 1) = g(s_{1,3}, 1) = 1 + 1 + (-10) = -8$
    - $g(s_4, 1) = g(s_{1,4}, 1) = 1 + (-10) = -9$
    - $g(s_5, 1) = g(s_{1,5}, 1) = -10$
    - $g(s_6, 1) = g(s_{2,5}, 1) = 0 \quad$ (Terminalzustand)

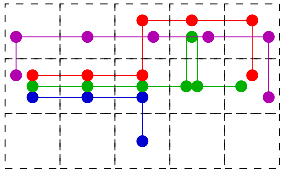
# Aufgabe 12: Monte Carlo und TD($\lambda$)




- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:

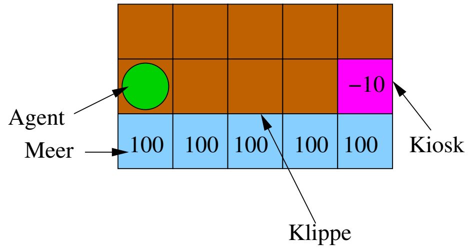
# Aufgabe 12: Monte Carlo und TD($\lambda$)

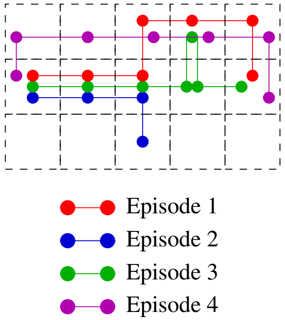



- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s) = V_0(s)$ für alle nicht in dieser Trajektorie besuchten Zustände, d.h. für $s \in \{s_{1,1}, s_{1,2}, s_{2,4}, s_{3,k}\}$ mit $k = 1 \ldots 5$

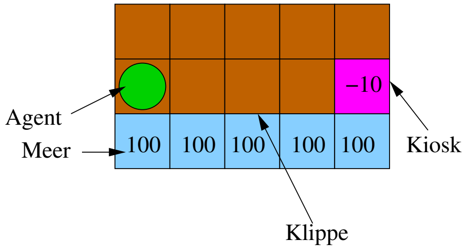
# Aufgabe 12: Monte Carlo und TD($\lambda$)

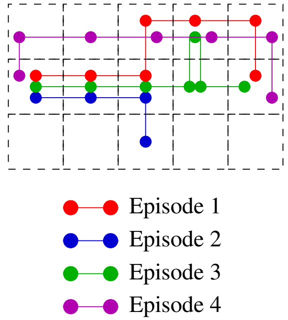



- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s) = V_0(s)$ für alle nicht in dieser Trajektorie besuchten Zustände, d.h. für $s \in \{s_{1,1}, s_{1,2}, s_{2,4}, s_{3,k}\}$ mit $k = 1 \ldots 5$
  - Weiter für $V_1(s_{2,1})$:

# Aufgabe 12: Monte Carlo und TD($\lambda$)




- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s) = V_0(s)$ für alle nicht in dieser Trajektorie besuchten Zustände, d.h. für $s \in \{s_{1,1}, s_{1,2}, s_{2,4}, s_{3,k}\}$ mit $k = 1 \ldots 5$
  - Weiter für $V_1(s_{2,1})$:

$$\begin{aligned} V_1(s_{2,1}) &= V_0(s_{2,1}) + \alpha_t(g(s_{2,1}, t) - V_0(s_{2,1})) \\ &= 0 + 1(-5 - 0) \\ &= -5 \end{aligned}$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

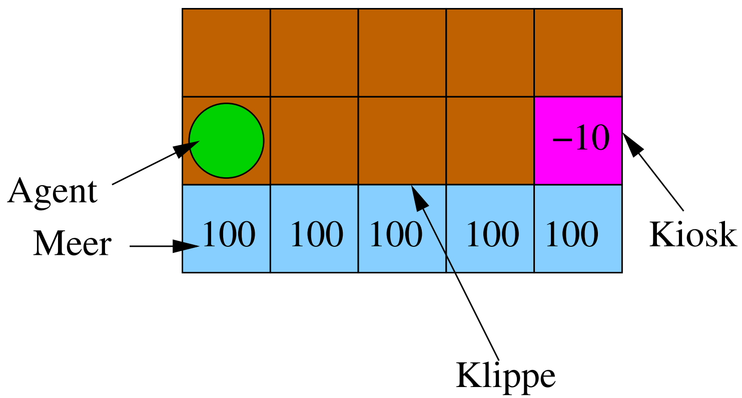
- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s_{2,2}) = 0 + 1(-6 - 0) = -6$
  - $V_1(s_{2,3}) = 0 + 1(-7 - 0) = -7$
  - $V_1(s_{1,3}) = 0 + 1(-8 - 0) = -8$
  - $V_1(s_{1,4}) = 0 + 1(-9 - 0) = -9$
  - $V_1(s_{1,5}) = 0 + 1(-10 - 0) = -10$

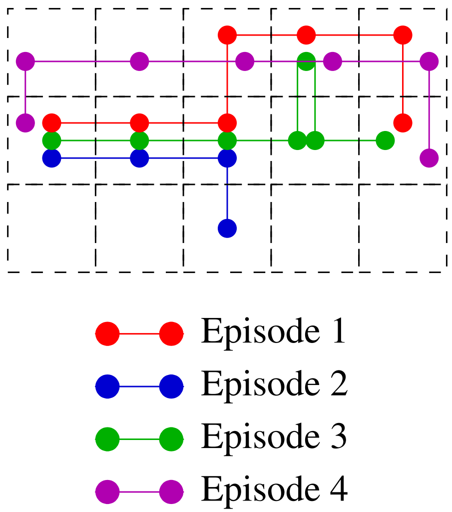
# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s_{2,2}) = 0 + 1(-6 - 0) = -6$
  - $V_1(s_{2,3}) = 0 + 1(-7 - 0) = -7$
  - $V_1(s_{1,3}) = 0 + 1(-8 - 0) = -8$
  - $V_1(s_{1,4}) = 0 + 1(-9 - 0) = -9$
  - $V_1(s_{1,5}) = 0 + 1(-10 - 0) = -10$
- Darstellung und folgende Episoden ($t = 2$ und $t = 3$) $\rightarrow$ Tafel

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Mit MC-Strategieevaluation und $\alpha_t = \frac{1}{t} = \alpha_1 = \frac{1}{1} = 1$ erhalten wir für $V_1$:
  - $V_1(s_{2,2}) = 0 + 1(-6 - 0) = -6$
  - $V_1(s_{2,3}) = 0 + 1(-7 - 0) = -7$
  - $V_1(s_{1,3}) = 0 + 1(-8 - 0) = -8$
  - $V_1(s_{1,4}) = 0 + 1(-9 - 0) = -9$
  - $V_1(s_{1,5}) = 0 + 1(-10 - 0) = -10$
- Darstellung und folgende Episoden ($t = 2$ und $t = 3$) $\rightarrow$ Tafel



# Aufgabe 12: Monte Carlo und TD($\lambda$)

(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.

- Episode 4: $s_0 = s_{2,1} \rightarrow s_{1,1} \rightarrow s_{1,2} \rightarrow s_{1,3} \rightarrow s_{1,4} \rightarrow s_{1,5} \rightarrow s_{2,5}$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

(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.

- Episode 4: $s_0 = s_{2,1} \rightarrow s_{1,1} \rightarrow s_{1,2} \rightarrow s_{1,3} \rightarrow s_{1,4} \rightarrow s_{1,5} \rightarrow s_{2,5}$
- zeitlicher Differenzfehler:

$$d_k := c(s_k) + V_{t-1}(s_{k+1}) - V_{t-1}(s_k)$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

(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.

- Episode 4: $s_0 = s_{2,1} \rightarrow s_{1,1} \rightarrow s_{1,2} \rightarrow s_{1,3} \rightarrow s_{1,4} \rightarrow s_{1,5} \rightarrow s_{2,5}$
- zeitlicher Differenzfehler:

$$d_k := c(s_k) + V_{t-1}(s_{k+1}) - V_{t-1}(s_k)$$

- hier (Abhängigkeit der Kosten vom Folgezustand):

$$d_k := c(s_k, s_{k+1}) + V_{t-1}(s_{k+1}) - V_{t-1}(s_k)$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

(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.

- Episode 4: $s_0 = s_{2,1} \to s_{1,1} \to s_{1,2} \to s_{1,3} \to s_{1,4} \to s_{1,5} \to s_{2,5}$
- zeitlicher Differenzfehler:

$$d_k := c(s_k) + V_{t-1}(s_{k+1}) - V_{t-1}(s_k)$$

- hier (Abhängigkeit der Kosten vom Folgezustand):

$$d_k := c(s_k, s_{k+1}) + V_{t-1}(s_{k+1}) - V_{t-1}(s_k)$$

- Sei $s_0 = s_{2,1}$, $s_1 = s_{1,1}$, $s_2 = s_{1,2}$, $s_3 = s_{1,3}$, $s_4 = s_{1,4}$, $s_5 = s_{1,5}$, $s_6 = s_{2,5}$.

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$\begin{aligned} d_0 &= c(s_0, s_1) + V_3(s_1) - V_3(s_0) \\ &= c(s_{2,1}, s_{1,1}) + V_3(s_{1,1}) - V_3(s_{2,1}) \\ &= 1 + 0 - 30\frac{2}{3} \\ &= -29\frac{2}{3} \end{aligned}$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$\begin{aligned} d_0 &= c(s_0, s_1) + V_3(s_1) - V_3(s_0) \\ &= c(s_{2,1}, s_{1,1}) + V_3(s_{1,1}) - V_3(s_{2,1}) \\ &= 1 + 0 - 30\frac{2}{3} \\ &= -29\frac{2}{3} \end{aligned}$$

$$\begin{aligned} d_1 &= c(s_1, s_2) + V_3(s_2) - V_3(s_1) \\ &= c(s_{1,1}, s_{1,2}) + V_3(s_{1,2}) - V_3(s_{1,1}) \\ &= 1 + 0 - 0 \\ &= 1 \end{aligned}$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$
\begin{aligned}
d_2 &= c(s_2, s_3) + V_3(s_3) - V_3(s_2) \\
&= c(s_{1,2}, s_{1,3}) + V_3(s_{1,3}) - V_3(s_{1,2}) \\
&= 1 + (-8) - 0 \\
&= -7
\end{aligned}
$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$
\begin{aligned}
d_2 &= c(s_2, s_3) + V_3(s_3) - V_3(s_2) \\
&= c(s_{1,2}, s_{1,3}) + V_3(s_{1,3}) - V_3(s_{1,2}) \\
&= 1 + (-8) - 0 \\
&= -7
\end{aligned}
$$

$$
\begin{aligned}
d_3 &= c(s_3, s_4) + V_3(s_4) - V_3(s_3) \\
&= c(s_{1,3}, s_{1,4}) + V_3(s_{1,4}) - V_3(s_{1,3}) \\
&= 1 + (-9) - (-8) \\
&= 0
\end{aligned}
$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$
\begin{aligned}
d_4 &= c(s_4, s_5) + V_3(s_5) - V_3(s_4) \\
&= c(s_{1,4}, s_{1,5}) + V_3(s_{1,5}) - V_3(s_{1,4}) \\
&= 1 + (-10) - (-9) \\
&= 0
\end{aligned}
$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

- Wir erhalten:

$$\begin{aligned} d_4 &= c(s_4, s_5) + V_3(s_5) - V_3(s_4) \\ &= c(s_{1,4}, s_{1,5}) + V_3(s_{1,5}) - V_3(s_{1,4}) \\ &= 1 + (-10) - (-9) \\ &= 0 \end{aligned}$$

$$\begin{aligned} d_5 &= c(s_5, s_6) + V_3(s_6) - V_3(s_5) \\ &= c(s_{1,5}, s_{2,5}) + V_3(s_{2,5}) - V_3(s_{1,5}) \\ &= -10 + 0 - (-10) \\ &= 0 \end{aligned}$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

(c) Ermitteln Sie nun mit Hilfe des TD($\lambda$)-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 12: Monte Carlo und TD($\lambda$)

(c) Ermitteln Sie nun mit Hilfe des TD($\lambda$)-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.

- Zur Erinnerung: Exponentielle Gewichtung der Pfadkosten-Datenpunkte

$$V_t(s_k) = (1 - \lambda) \cdot \mathbb{E}\left[\sum_{l=0}^{\infty} \lambda^l \cdot \left( V_{t-1}(s_{k+1+l}) + \sum_{m=0}^{l} c(s_{k+m}) \right)\right]$$

# Aufgabe 12: Monte Carlo und TD($\lambda$)

(c) Ermitteln Sie nun mit Hilfe des TD($\lambda$)-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.

- Zur Erinnerung: Exponentielle Gewichtung der Pfadkosten-Datenpunkte

$$V_t(s_k) = (1-\lambda) \cdot \mathbb{E}\left[\sum_{l=0}^{\infty} \lambda^l \cdot \left( V_{t-1}(s_{k+1+l}) + \sum_{m=0}^{l} c(s_{k+m}) \right)\right]$$

- In TD-Form: TD($\lambda$)-Algorithmus

$$V_t(s_k) = V_{t-1}(s_k) + \alpha_t \cdot \sum_{m=0}^{\infty} \lambda^m \cdot d_{k+m}$$

mit $d_m = c(s_m) + V_{t-1}(s_{m+1}) - V_{t-1}(s_m)$ und $s_0, s_1, \ldots$ aus aktueller Episode sowie $\alpha_t$ angemessen fallend mit $t$ (hier: $\alpha_t = \frac{1}{t}$)