Erinnerung Folie 196

Verwendung unsicherer Info. beim Planen

Unterschiedliche Aktionen eines Agenten erzeugen mit unterschiedlichen W‘keiten Effekte (Sicht zur Planungszeit).

Beispiel Flughafenfahrt-Aktion Folie 195:

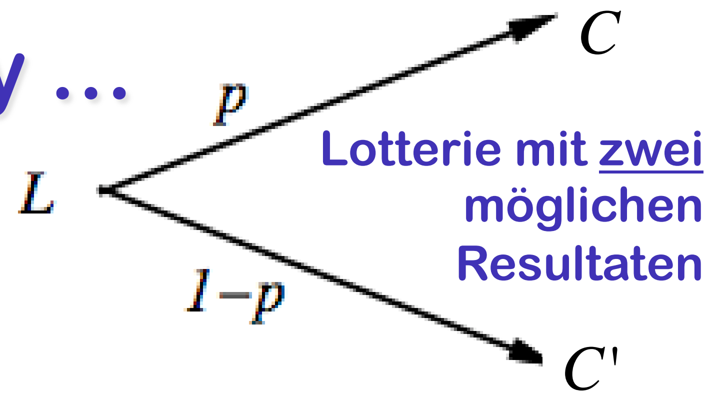
$P(A(25)$ bringt mich rechtzeitig zum Flug $| \ldots) = 0{,}04$
$P(A(90)$ bringt mich rechtzeitig zum Flug $| \ldots) = 0{,}8$
$P(A(120)$ bringt mich rechtzeitig zum Flug $| \ldots) = 0{,}95$
$P(A(1440)$ bringt mich rechtzeitig zum Flug $| \ldots) = 0{,}9999999999999999$

Und nun?

Handle so, dass Du den maximalen erwarteten Nutzen erzielst!
(s. Kapitel 1 (Folie 17): Rationaler Agent)

Entscheidung ≈ Effektw‘keit x Nutzen

Life is a Lottery …



Modelliere Aktionen mit nicht-deterministischen/unbekannten Effekten als **Lotterie** (W'keitsverteilung über den möglichen Konsequenzen/Effekte d. Aktion):

Notation: $L = [p_1, C_1;\ p_2, C_2;\ \ldots;\ p_n, C_n]$

für alternative Effekte (*consequences*) oder Folge-Lotterien C_i und ihre W'keiten p_i, *wobei* $\sum_i p_i = 1$

Zustände entsprechen Lotterien der Form $[1, C]$

Gegeben seien **numerische Nutzen-Werte** U für die Effekte C_i von A (allgemeiner: Präferenzrelation auf den C_i)

Definiere den **Nutzen U (*utility*) der Lotterie/Aktion A**:

$U(A) = U([p_1, C_1;\ \ldots;\ p_n, C_n]) = \sum_i p_i U(C_i)$ (oft normiert als $0 \leq U(A) \leq 1$)

Erwarteter Nutzen unter Evidenz

Zur Modellierung von Planen unter Unsicherheit fehlt noch:

- Berücksichtigung von vorhandener **Evidenz** für die Aktionseffekte
  (z.B. W'keit von Zuspätkommen mit und ohne Wissen der Staumeldungen)
- Einbeziehung unabhängiger Ursachen für Aktionseffekte
  (z.B. W'keit, dass ich einen Flieger erreiche, hängt ab von meiner Ankunft, aber auch davon, ob der Flieger da ist)

Erwarteter Nutzen EU der Lotterie/Aktion A unter Evidenz E:

$$EU(A|E) = \sum_i \underbrace{P(C_i|E)}_{p_i} \cdot U(C_i)$$

Entscheidung ≈ Effektw'keit x Nutzen (Folie 196) wird zu:

Für Aktionen $\mathcal{A}$, Evidenz E wähle $\operatorname{argmax}_{A\in\mathcal{A}} EU(A|E)$, wobei

$$EU(A|E) := \textstyle\sum_i \left[P(C_i(A) \mid Do(A), E) \cdot U(C_i(A))\right]$$

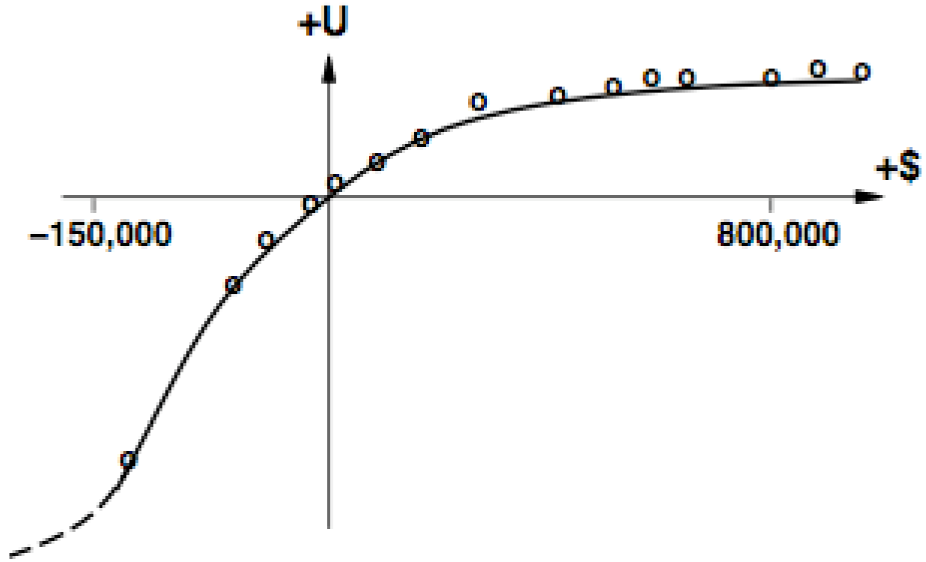
Exkurs: Ist Kontostand eine Nutzenfunktion?

Für die meisten Menschen **nicht**!

z.B.: [1,„Gewinne 1 Mio €"] > [0.5,„Gewinne 0 €"; 0.5,„Gewinne 3 Mio €"],

… was der Definition Folie 318 widerspricht!

Empirisch ermitteltes U (qualitativ):



normative vs. **deskriptive** Entscheidungstheorie

Nutzen im Flughafentransferbeispiel

Beispiel Flughafenfahrt-Aktion Folie 196:

P(rechtzeitig | Do(A(25), $Stau$=kein_Stau) = 0,04

P(rechtzeitig | Do(A(90), $Stau$=kein_Stau) = 0,7

P(rechtzeitig | Do(A(120), $Stau$=kein_Stau) = 0,95

P(rechtzeitig | Do(A(1440), $Stau$=kein_Stau) = 0,99999999

P(rechtzeitig)=1–P(verspätet)

Nutzenwerte für die Effekte

$$EU(A|E) = \sum_i \underbrace{P(C_i|E)}_{p_i} \cdot U(C_i)$$

U(verspätet) = –100

U(rechtzeitig) = 10 – *Frühstrafe*

Frühstrafe = (2^n) für n Stunden erwartete Wartezeit ≥1 Stunde, sonst 0

$$EU\left(A(90)\middle|Stau = \text{kein\_Stau}\right) = 0{,}3 \cdot -100 + 0{,}7 \cdot 10 = -23$$

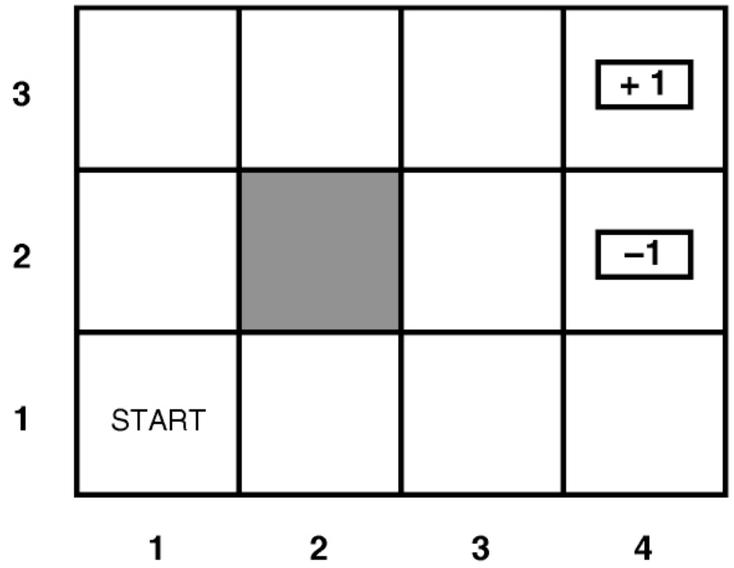
$$EU\left(A(120)\middle|Stau = \text{kein\_Stau}\right) = 0{,}05 \cdot -100 + 0{,}95 \cdot \left(10 - 2^1\right) = 2{,}6$$

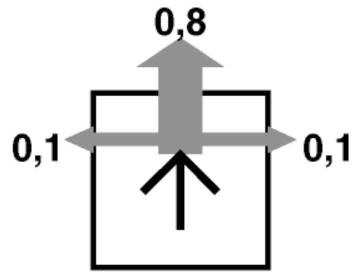
$$EU\left(A(1440)\middle|Stau = \text{kein\_Stau}\right) \approx 1 \cdot \left(10 - 2^{23}\right) = -8.388.598$$

Sequenzielle Entscheidungsprobleme

Problem bislang: Finde optimale Einzelaktion. **Jetzt**:
Finde: Optimale Sequenzen von Aktionen unter Unsicherheit

Beispiel





Aktion/Lotterie
in jedem Zug:

- mit P=0,8 lande im beabsichtigten Feld
- mit je P=0,1 lande rechts/links daneben oder bleibe auf Feld hängen (bei Wand)

Up(1,1) = [0,8,⟨1,2⟩; 0,1,⟨2,1⟩; 0,1,⟨1,1⟩]
Right(,)=[…] **Left**(,)=[…] **Down**(,)=[…]

Up, **Up**, **Right**, **Right**, **Right** erreicht ⟨4,3⟩ mit
$P = 0{,}8^5 + 0{,}1^4 \cdot 0{,}8 = 0{,}32768 + 0{,}00008 = 0{,}32776$

Markowsche Entscheidungsprozesse (MDPs)

Voraussetzungen

- Vollständige Beobachtbarkeit der Umgebung/des Zustands (Agent ermittelt sicher, auf welcher Position er ist)
- **Markow-Eigenschaft**: Folgezustand hängt nur ab von letztem Zustand und Aktion (nicht von früheren Aktionen)

MDP (S_0, T, R)

- **Startzustand** S_0 (**Beispiel**: $\langle 1,1 \rangle$)
- **Transitionsmodell** $T(s,a,s')$: W'keit, von Zustand s mit a in s' zu kommen („auseinandergenommene" Notation der Lotterie) (**Beispiel**: $T(\langle 1,1 \rangle, \mathbf{Up}, \langle 1,2 \rangle)=0{,}8$; $T(\langle 1,1 \rangle, \mathbf{Up}, \langle 1,1 \rangle)=0{,}1$; …)
- ***Reward*-Funktion** $R(s)$: Belohnung (positiv oder negativ), e. Zustand zu erreichen (**Beispiel**: $-0{,}04$ außer für $\langle 4,2 \rangle$, $\langle 4,3 \rangle$) (Nutzen<u>anteil</u> (+/–) für <u>jeden einzelnen</u> Zwischenzustand)

Der Nutzen des Agierens

Voraussetzungen:

- Nutzen einer Handlungssequenz ergibt sich aus der Summe von *Rewards* der besuchten Zustände
- *Rewards* in naher Zukunft sind möglicherweise anders zu gewichten als in ferner Zukunft: Faktor $0 \leq \gamma \leq 1$ (Abschlag, *discount factor*)
- Länge von Aktionssequenzen ist a priori nicht beschränkt

$$U([s_0, s_1, s_2, \ldots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t)$$

Für $\gamma < 1$ und $R_{\max}$ ist der Nutzen jeder Aktionssequenz endlich:

$$U([s_0, s_1, s_2, \ldots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma}$$

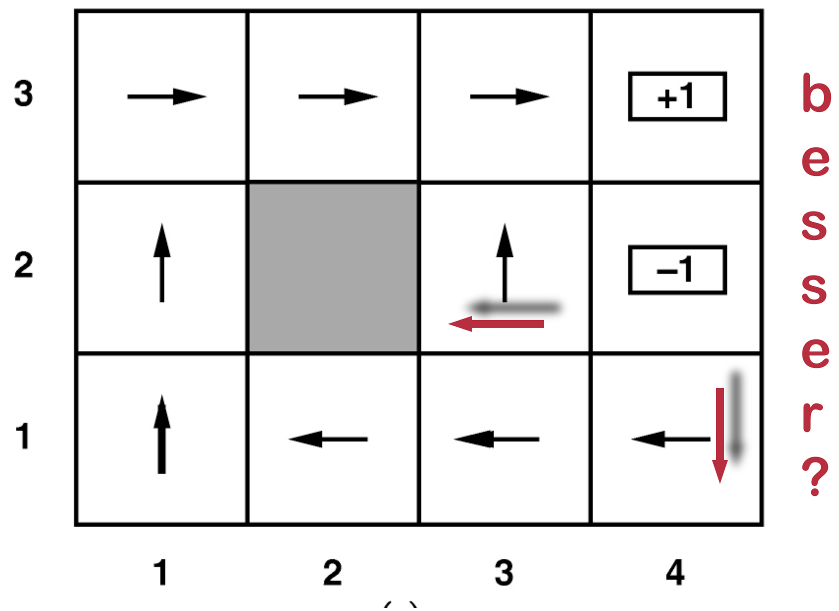
MDP-Pläne (Strategien, Taktiken, *policies*)

Bisher:

Nutzen von Aktionssequenzen bewertet gegebenes Verhalten, hilft aber nicht entscheiden, was der Agent in Zustand s_i tun sollte!

Policy (MDP-Plan) ist e. Abbildung $\pi : S \rightarrow A$

Beispiel



Eine **optimale *Policy*** ist eine *Policy* mit maximalem erwartetem Nutzen:

$$\pi^* := \operatorname*{argmax}_{\pi} EU\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| \pi\right]$$

Wie bewertet man den Nutzen zukünftiger Zustände nach *policy* π?

Der Nutzen eines Zustands…

$$\pi^* := \operatorname*{argmax}_{\pi} EU\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| \pi\right]$$

… gegeben *Policy* π ist: $U^\pi(s) := EU\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| \pi, s_0 = s\right]$

Jetzt wollen wir sagen:

Der (a priori-)Nutzen eines Zustands ist sein *Reward* plus der (erwartete, diskontierte) Nutzen des Nachfolgezustands unter der jeweils optimalen Folgeaktion

Die Bellmann-Gleichung

$$U(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s')U(s')$$

…definiert ein Gleichungssystem über ein MDP, das $U^{\pi^*}(s)$ lokal charakterisiert,

…das aber i.a. nicht praktisch lösbar ist!

Achtung Henne & Ei!
Zustandsnutzen &
optimale Policy

Bester Nachbarzustand statt optimaler *policy*

Zwischenfazit:

- Wir wissen, wie für MDPs eine optimale *policy* definiert ist

… aber was ist ein Algorithmus, sie zu berechnen?

- Wir wissen, wie der Nutzen aller Zustände im MDP statisch definiert ist – dabei optimale *policy* implizit vorausgesetzt

… aber wie löst man das entsprechende Gleichungssystem?

Ausweg:

- Nimm die Bellmann-Gleichung $U(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s')U(s')$
- approximiere optimalen Zustandsnutzen durch Iteration
- „hoffe“, dass die Iteration auf korrekten Wert konvergiert

Value Iteration (Ertel: Wert-Iteration)

MDP ist bekannt und gegeben!

function VALUE-ITERATION(mdp, ϵ) **returns** a utility function
inputs: mdp, an MDP with states S, transition model T, reward function R, discount γ
ϵ, the maximum error allowed in the utility of any state
local variables: U, U', vectors of utilities for states in S, initially zero
δ, the maximum change in the utility of any state in an iteration

repeat
$U \leftarrow U'; \delta \leftarrow 0$
for each state s **in** S **do**
$$U'[s] \leftarrow R[s] + \gamma \max_a \sum_{s'} T(s,a,s')\, U[s']$$
if $|U'[s] - U[s]| > \delta$ **then** $\delta \leftarrow |U'[s] - U[s]|$
until $\delta < \epsilon(1-\gamma)/\gamma$
return U

„Bellmann-Gleichung“,
aber in U und U' !

Hier Abbruch zusätzlich abhängig von γ.
Kann man machen, muss man aber nicht.

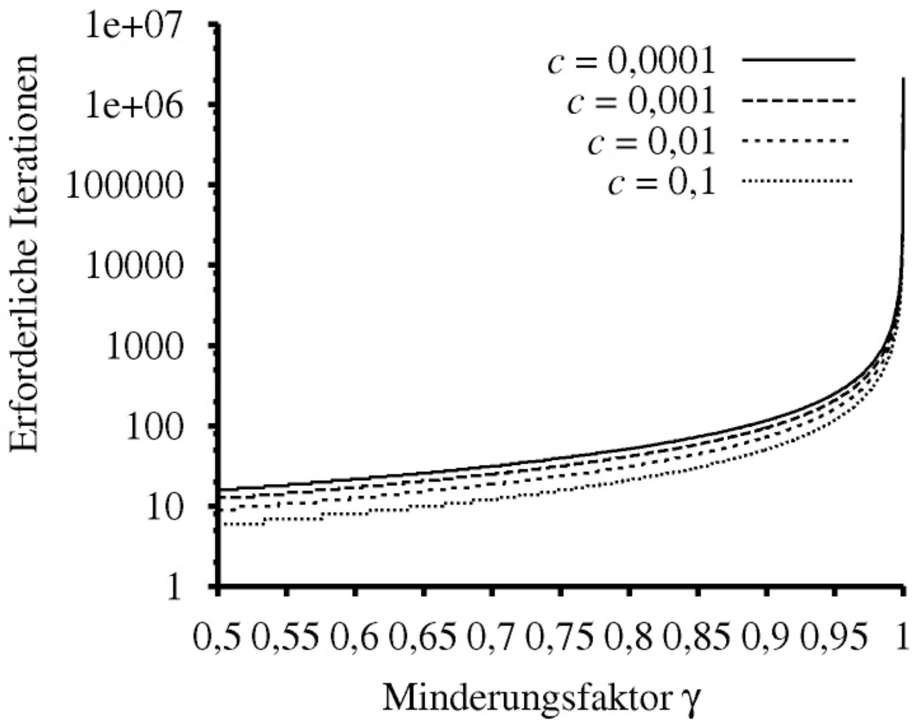
Konvergenz der *Value Iteration*

Satz

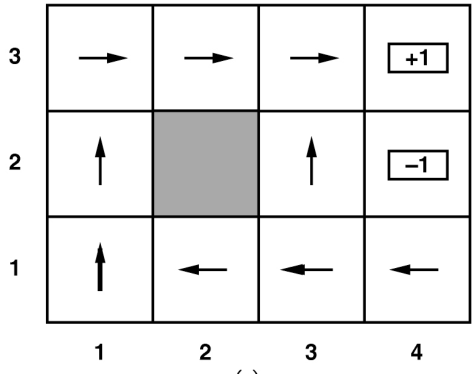
Der *Value Iteration*-Algorithmus konvergiert für alle Zustände s des MDP auf den Nutzen $U^{\pi^*}(s)$, welcher der optimalen *Policy* π^* entspricht

Beweisskizze: s. Russell/Norvig 17.2

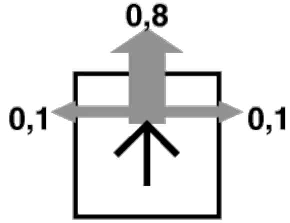
Zahl der Iterationen bis Konvergenz in Abhängigkeit von Abschlagsfaktor γ und erlaubtem Fehler $\varepsilon = c \cdot R_{max}$ (<u>qualitativer</u>(!) Zusammenhang)



Ergebnis für Beispiel-MDP lt. Russell/Norvig



T s. Folie 322



	1	2	3	4
3	0,812	0,868	0,918	+ 1
2	0,762		0,660	–1
1	0,705	0,655	0,611	0,388

angeblich:

- $\gamma=1$,
- $R(s)=–0{,}04$ f. nichtterminale s

Probe:

$$U(\langle 3,3\rangle) = R(\langle 3,3\rangle) + \gamma \times \sum T(\langle 3,3\rangle,\mathbf{Right},s')U(s')$$
$$= –0{,}04 + 0{,}8\text{x}1+ 0{,}1\text{x}0{,}660 + 0{,}1\text{x}0{,}918$$
$$= –0{,}04 + 0{,}8 + 0{,}066 + 0{,}0918 = 0{,}9178$$

Achtung: Probe klappt nicht für andere Zustände!! (z.B. $\langle 3,1\rangle$)