

# Grundlagen der Künstlichen Intelligenz

## **12 Handeln unter Unsicherheit**

---

Maximieren des erwarteten Nutzens

*Volker Steinhage*

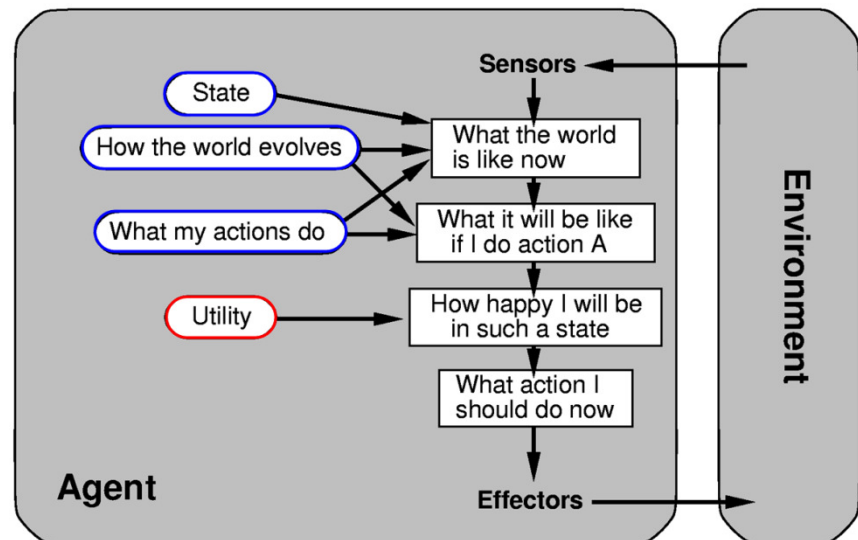
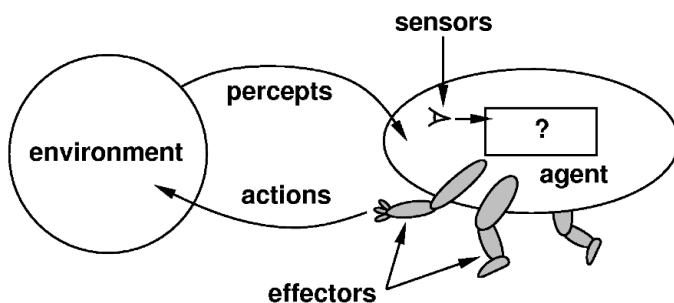
# Inhalt

---

- Einführung in Nutzentheorie
- Auswahl einzelner Aktionen
- Sequentielle Entscheidungsprobleme
- Markov Entscheidungsprozesse
- Value Iteration

# Grundlagen der Nutzentheorie (1)

- Die *Nutzenfunktion* (Utility Function)  $U(S)$ 
  - bewertet Zustände
  - formalisiert so das *Bevorzugen bestimmter Zustände*
  - steht für den Nutzen (Utility) des Zustandes  $S$  für den Agenten



# Grundlagen der Nutzentheorie (2)

Vor.: Eine nichtdeterministische Aktion  $A$  kann zu einer Menge von Folgezuständen  $Result_i(A)$  mit  $i > 1$  führen

Die  $W'$ keiten für das Erreichen der Zustände  $Result_i(A)$  nach Ausführung von  $A$  im aktuellen Zustand unter Evidenz  $E$  sind  $P(Result_i(A) | Do(A), E)$

- Erwarteter Nutzen (*Expected Utility* ( $EU$ )):

Expectation

Utility

$$EU(A | E) = \sum_i P(Result_i(A) | Do(A), E) \cdot U(Result_i(A))$$

- Mit dem Prinzip des *maximalen erwarteten Nutzens* (*Maximum Expected Utility* ( $MEU$ )) sollte ein rationaler Agent die Aktion auswählen, die  $EU(A|E)$  maximiert

# Vorgehen

---

Zunächst Betrachtung des *MEU-Prinzips* für  
Entscheidungen bezogen auf einzelne Situationen:

- 1) Sprache der Nutzentheorie
- 2) Formulierung sinnvoller Bedingungen für das Entscheiden unter Unsicherheit  
→ *Axiome der Nutzentheorie*
- 3) Zeigen, dass aus diesen Bedingungen das *MEU-Prinzip* ableitbar ist

Danach Erweiterung für den effizienten Umgang mit *Folgen* von Aktionen.

# Die Sprache der Nutzentheorie → Lotterien

- komplexe Szenarien werden als *Lotterien*  $L$  bezeichnet
- mögliche Ergebnisse sind dann mögliche *Gewinne*
- das Ergebnis wird vom Zufall bestimmt

Beispiel: Lotterie  $L$  mit 2 mögl. Ergebnissen  $A$  mit  $P(A) = p$  und  $B$  mit  $P(B) = 1 - p$ :

$$L = [p, A ; 1-p, B]$$

Allgemein: Lotterie  $L$  mit  $n$  möglichen Ergebnissen  $A_i$  mit  $P(A_i) = p_i$ , ( $i = 1, \dots, n$ ):

$$L = [p_1, A_1 ; \dots ; p_n, A_n] \text{ mit } \sum_{i \in \{1, \dots, n\}} p_i = 1$$

Sonderfall: Lotterie  $L$  mit nur einem möglichen Ergebnis  $A$ :

$$L = [1, A] \text{ oder kurz } L = A$$

Eine Lotterie ist also eine W'keitsverteilung über einer Menge möglicher Ergebnisse. Jedes einzelne Ergebnis einer Lotterie kann ein atomarer Zustand oder eine weitere Lotterie sein

# Präferenzen in der Nutzentheorie

---

Der Agent soll entscheiden!

→ Ziel ist die Ableitung von *Präferenzen* zwischen verschiedenen Lotterien aufgrund von Präferenzen zwischen den zugrundeliegenden Zuständen

Die hier notwendige *Notation für Präferenz*:\*

$A \succ B$ : Agent bevorzugt A gegenüber B

$A \sim B$ : Agent ist unentschieden zwischen A und B

Welche vernünftigen Bedingungen sollten für die Präferenzrelation gelten?

---

\* Es gibt noch weitere Notationen zu Präferenzen (z.B.  $A \succeq B$ ,  $A \succsim B$ ) in der Nutzentheorie, die aber in diesem Kontext nicht nötig sind.

# Die Axiome 1 und 2 der Nutzentheorie

Sechs Axiome der Nutzentheorie spezifizieren die inhaltlichen Anforderungen für den Umgang mit Lotterien und Prioritäten.

Gegeben seien Zustände  $A, B, C$ .

## 1) Sortierbarkeit:

$$(A \succ B) \oplus (B \succ A) \oplus (A \sim B)$$

Der Agent muss wissen, was er will: entweder einen Zustand bevorzugen oder unentschieden sein ( $\oplus$  für exklusive Disjunktion)

## 2) Transitivität:

$$(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$$

Verletzen der Transitivität verursacht irrationales Verhalten:  $A \succ B \succ C \succ A$ :

Annahme: Agent hat  $A$  und würde es mit Aufpreis für  $C$  eintauschen

~  $C$  würde er wieder für  $A$  und Geldzahlung eintauschen

~ Agent verliert Geld und würde bei laufender Wiederholung alles verlieren



# Die Axiome 3 und 4 der Nutzentheorie

---

## 3) Stetigkeit (Kontinuität):

$$A \succ B \succ C \Rightarrow \exists p [p, A; 1 - p, C] \sim B$$

Wenn  $B$  in einer Folge von Präferenzen zwischen  $A$  und  $C$  liegt, lässt sich eine Lotterie über  $A$  und  $C$  konstruieren, so dass der Agent unentschieden zwischen dieser Lotterie und sicherem  $B$  ist

## 4) Ersetzbarkeit:

$$A \sim B \Rightarrow [p, A; 1 - p, C] \sim [p, B; 1 - p, C]$$

Ist der Agent zwischen zwei einfachen Lotterien unentschlossen, so ist er auch unentschlossen zwischen zwei komplexeren Lotterien, die genau gleich sind, außer dass in einer Lotterie  $A$  durch  $B$  ersetzt ist.

# Die Axiome 5 und 6 der Nutzentheorie

## 5) Monotonie:

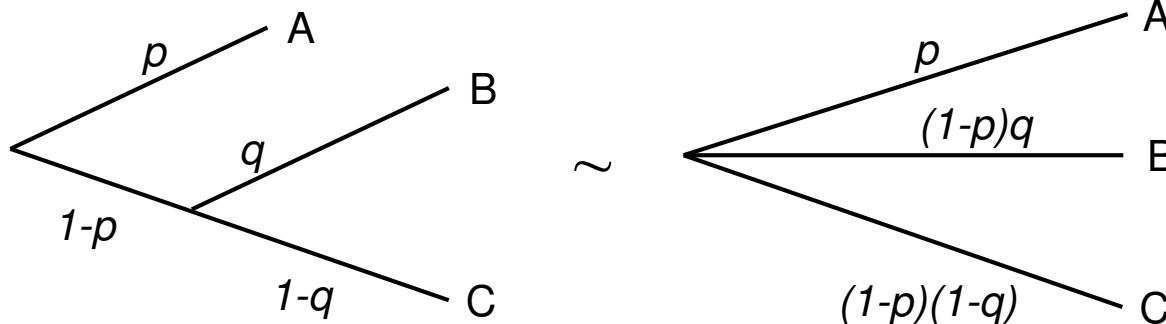
$$A \succ B \Rightarrow (p > q \Leftrightarrow [p, A; 1 - p, B] \succ [q, A; 1 - q, B])$$

Bevorzugt ein Agent das Ereignis  $A$  gegenüber Ereignis  $B$ , dann wird er die Lotterie bevorzugen, die  $A$  mit höherer Wahr'keit als Ergebnis liefert

## 6) Zerlegbarkeit:

$$[p, A; 1 - p, [q, B; 1 - q, C]] \sim [p, A; (1 - p) \cdot q, B; (1 - p) \cdot (1 - q), C]$$

Kombinierte Lotterien sind über die Gesetze der Wahrscheinlichkeitstheorie in einfachere Lotterien zerlegbar



# Nutzenprinzip: Nutzenaxiome $\rightarrow$ Nutzenfunktion

Aus **Präferenzen** gemäß den Nutzenaxiomen folgt die Existenz von **Nutzenfunktionen** (Neumann und Morgenstern, 1944):\*

## Nutzenprinzip:

Beachtet ein Agent in seinen Präferenzen die 6 Axiome der Nutzentheorie, so gibt es eine reellwertige **Nutzenfunktion**  $U: S \rightarrow \mathbf{R}$  auf der Zustandsmenge  $S$  mit

$$U(A) > U(B) \Leftrightarrow A \succ B$$

$$U(A) = U(B) \Leftrightarrow A \sim B$$

\* von Neumann, J. and Morgenstern, O. (1944). Theory of Games and Economic Behavior (first edition). Princeton University Press.

# Nutzenaxiome $\leadsto$ Nutzenfunktion $\leadsto$ MEU

---

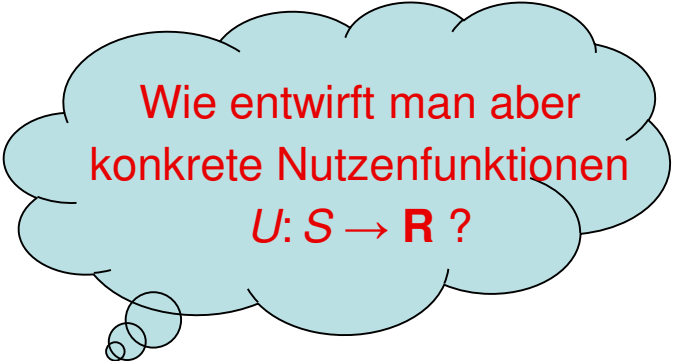
Das Prinzip des maximalen erwarteten Nutzens ist nun in der Sprache der Nutzentheorie formulierbar:

Prinzip des maximalen erwarteten Nutzens (**Max. Expected Utility Principle**):

Der Nutzen einer Lotterie ist die Summe über den Produkten aus den mit ihren Wahrscheinlichkeiten gewichteten Nutzen der möglichen Ergebnisse:

$$U([p_1, S_1; \dots; p_n, S_n]) = \sum_i p_i \cdot U(S_i)$$

Diese Summe  $\sum_i p_i \cdot U(S_i)$  ist zu maximieren.



Wie entwirft man aber  
konkrete Nutzenfunktionen  
 $U: S \rightarrow \mathbf{R}$  ?

# Monetäre Nutzenfunktionen

... auf der Grundlage von Marktmodellen

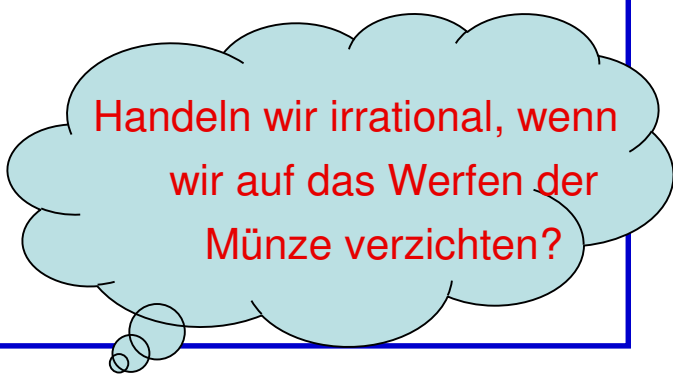
... weil die Nutzentheorie dort ihre Wurzeln hat, da sich Geld als universales Zahlungs- und Bewertungsmittel für alle Güter und Leistungen anbietet

Um zu verstehen, wie man **monetäre Entscheidungen unter Unsicherheit** trifft, untersuchen wir das Verhalten von Agenten bei Entscheidungen mit Lotterien, bei denen Geld im Spiel ist:

Annahme: Wir *haben bereits 1 Mio. Euro* in einem Quiz gewonnen

Angebot: Wir können eine Münze werfen und

- bei Kopf *insgesamt 2,5 Mio. Euro* gewinnen
- bei Zahl *alles verlieren*



Handeln wir irrational, wenn wir auf das Werfen der Münze verzichten?

# Expected Monetary Value

Die Wette also:

Annahme: Wir *haben bereits 1 Mio. Euro* in einem Quiz gewonnen

Angebot: Wir können eine Münze werfen und

- bei Kopf *insgesamt 2,5 Mio. Euro* gewinnen
- bei Zahl *alles verlieren*

## Erster Ansatz:

- Annahme: Die Münze ist fair
- Kriterium für den erwarteten Nutzen sei der „*Expected Monetary Value*“ (EMV):

$$\sum_i p_i \cdot MV(S_i)$$

~  $EU(\text{Accept}) = \frac{1}{2} \cdot (0 \text{ Euro}) + \frac{1}{2} \cdot (2.500.000) = 1.250.000 \text{ Euro}$

~  $EU(\text{Decline}) = 1 \cdot 1.000.000 \text{ Euro}$

~ Der Agent nimmt die Wette an!

Oder wäre ein anderer  
Ansatz sinnvoller?

# Expected Monetary Value

Die Wette also:

Annahme: Wir *haben bereits 1 Mio. Euro* in einem Quiz gewonnen

Angebot: Wir können eine Münze werfen und

- bei Kopf *insgesamt 2,5 Mio. Euro* gewinnen
- bei Zahl *alles verlieren*

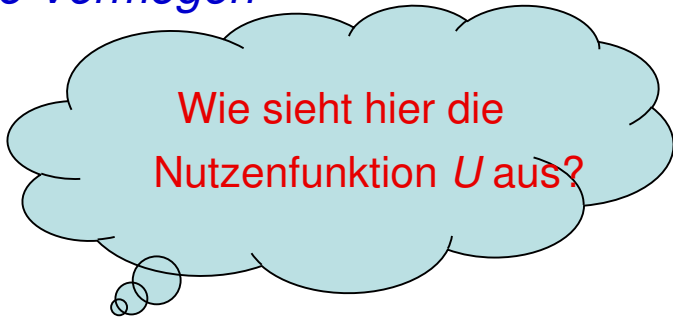
Zweiter Ansatz:

- Annahme: Die Münze ist fair
- Annahme: wir besitzen im Zustand  $S_k$  *bereits  $k$  Euro Vermögen*

$$\leadsto EU(\text{Accept}) = \frac{1}{2} \cdot U(S_k) + \frac{1}{2} \cdot U(S_{k+2.500.000})$$

$$\leadsto EU(\text{Decline}) = 1 \cdot U(S_{k+1.000.000})$$

$\leadsto$  und nun?



Wie sieht hier die  
Nutzenfunktion  $U$  aus?

# Vermögen und Expected Monetary Value

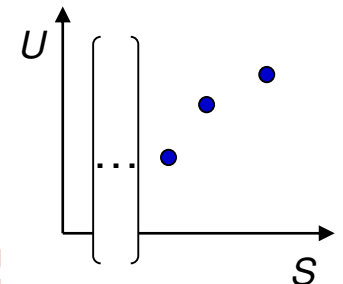
Beobachtung: Der Nutzen von Geld ist ggf. nicht proportional zum Betrag:

- die erste Million Euro mag zu einer anderen Lebensweise führen
- für Milliardäre wird sich wahrscheinlich nur wenig ändern

Übertragen auf unser Beispiel:

- Annahme: Die Münze ist fair
- Annahme: wir besitzen im Zustand  $S_k$  bereits  $k$  Euro *Vermögen*
  - seien  $U(S_k) = 5$ ,  $U(S_{k+1.000.000}) = 8$ ,  $U(S_{k+2.500.000}) = 9$
  - $EU(\text{Accept}) = \frac{1}{2} \cdot U(S_k) + \frac{1}{2} \cdot U(S_{k+2.500.000}) = 2,5 + 4,5 = 7$
  - $EU(\text{Decline}) = U(S_{k+1.000.000}) = 8$

~ Der Agent lehnt die Wette mit dieser Nutzenfunktion  $U$  ab!



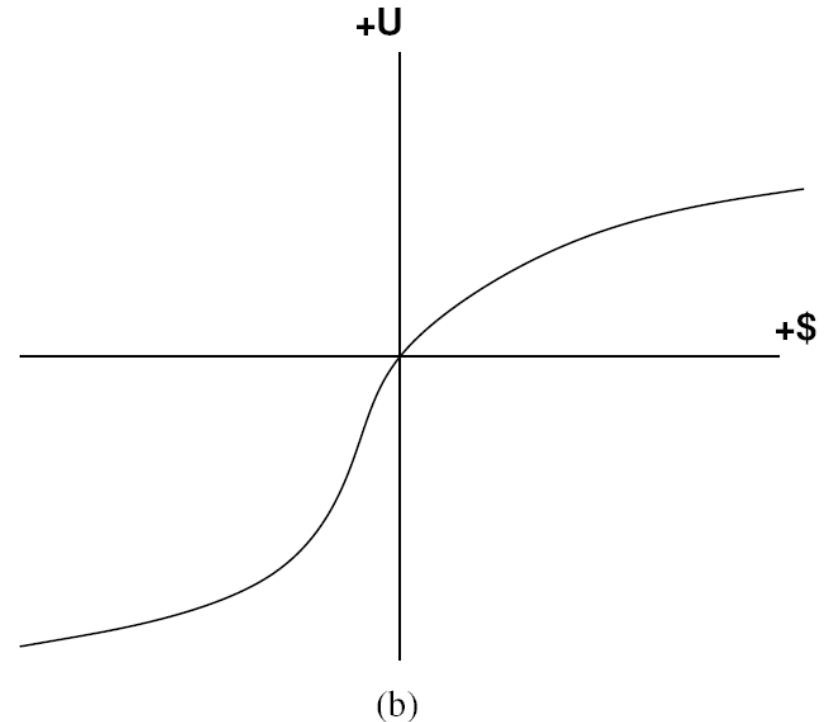
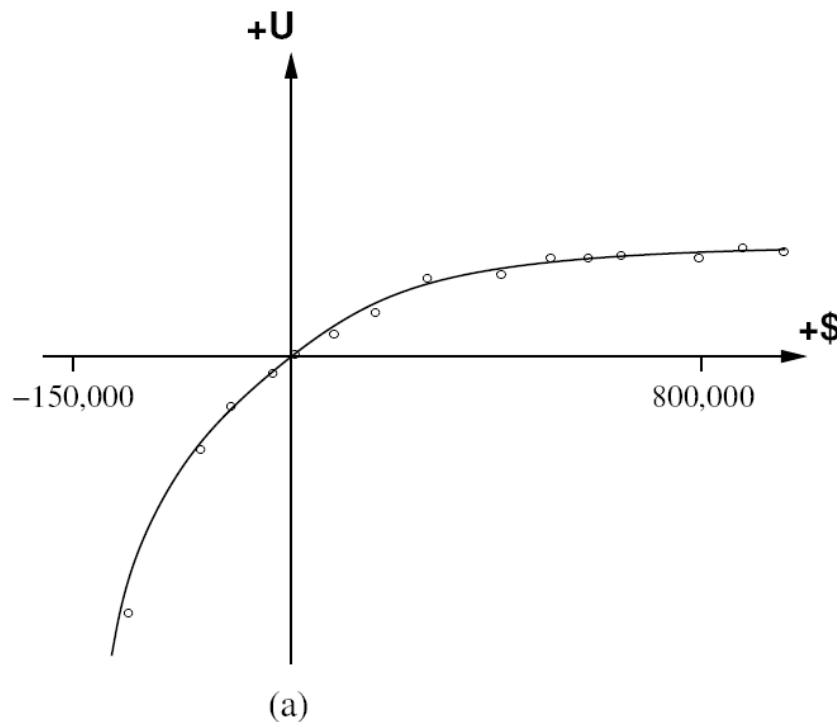


# Verlauf von monetären Nutzenfunktionen

Pionierstudien von Grayson (1960)\* über echte Nutzenfunktionen zeigen:

nahezu exakte Proportionalität zum Logarithmus

~ abnehmender Gradient der Nutzenfunktion

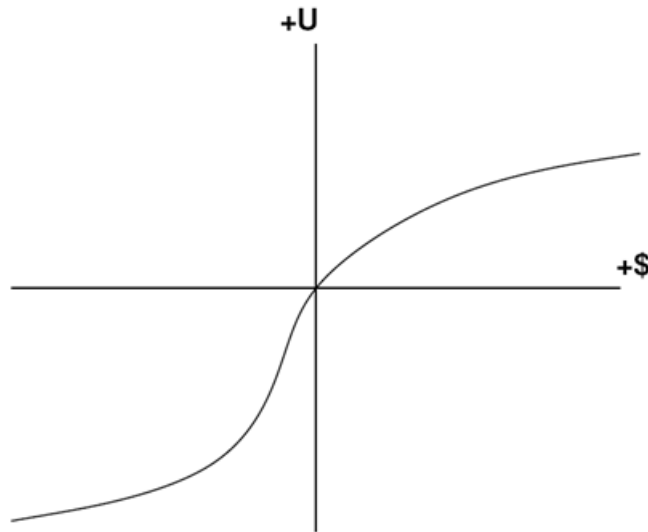


(a) Empirische Kurve eines bestimmten Probanden:  $U(S_{k+n}) = -263.31 + 22.09 \cdot \log(n+150,000)$   
für  $-150,000\$ \leq n \leq 800,000\$$  (b) typische Kurve für den vollständigen Bereich

\* Grayson Jr., C. J. [1960]: Decisions under uncertainty. Boston: Harvard University Press 1960.

# Risikoverhalten

- Im positiven Teil der Nutzenfunktion ist der Agent *risikoscheu*: die Lotterie  $L$  wird geringer bewertet als der sichere Erwartungswert:  $U(L) < U(S_{EMV(L)})$
- Im negativen Teil der Nutzenfunktion ist der Agent *risikofreudig*: die Lotterie  $L$  wird stärker bewertet als der sichere Erwartungswert:  $U(L) > U(S_{EMV(L)})$
- Für kleine Abschnitte der Nutzenfunktion ist der Agent *risikoneutral*: die Funktion ist fast linear:  $U(L) \approx U(S_{EMV(L)})$



# Skalierung und Normalisierung

Die Nutzenaxiome spezifizieren keine eindeutige Nutzenfunktion. Zwei Agenten mit Nutzenfunktion  $U_1(S)$  und  $U_2 = k_1 + k_2 \cdot U_1(S)$  mit Konstante  $k_1$  und positiver Konstante  $k_2$  verhalten sich gleich.

Ein Ansatz zur vergleichenden Beurteilung:

1) **Skalierung** der Nutzen zwischen

- bestem Preis  $U(S^\top) := u_\top$
- schlimmster Katastrophe  $U(S^\perp) := u_\perp$

2) **Normalisierung** der Nutzen:

- bester Preis  $u_\top = 1$
- schlimmste Katastrophe  $u_\perp = 0$

Werte der Zustände  $S \neq S^\top, S^\perp$  erhält man durch Variieren der W'keit  $p$  in einer Standardlotterie  $[p, u_\top; (1-p), u_\perp]$  und Abgleichen mit  $S$  bis der Agent zu  $S$  und Lotterie indifferent ist (s. Axiom 3):

$$[p, u_\top; (1-p), u_\perp] \sim S$$

$\leadsto U(S) = p$  bei normalisierten Nutzen

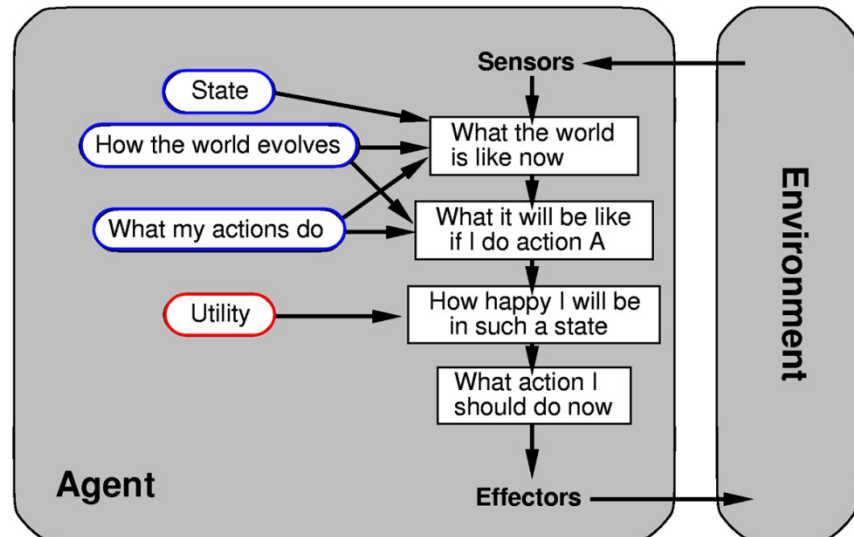
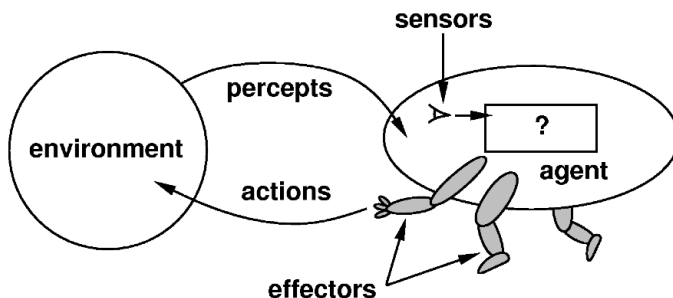
# Sequentielle Entscheidungsprobleme

Bislang: Einfache Entscheidungsprobleme

~ Nutzen für jedes Ergebnis *einer Aktion* bekannt

Jetzt: Sequentielle Entscheidungsprobleme

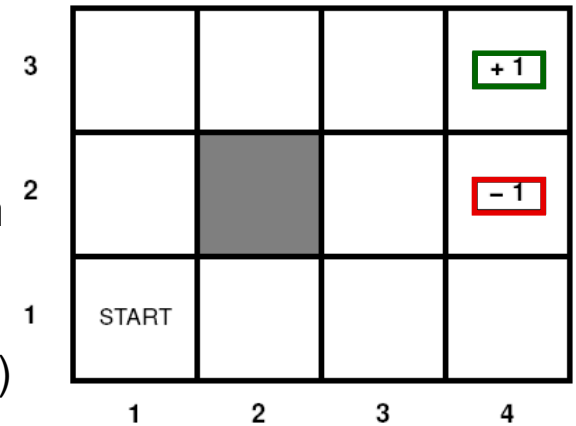
~ Nutzen ist vom Ergebnis einer *Aktionsfolge* abhängig



# Ein sequentielles Entscheidungsproblem

Anwendungsbeispiel:

- Agent soll **Endzustand (4,3)** mit **Belohnung = +1** erreichen  
Agent soll **Endzustand (4,2)** mit **Bestrafung = -1** vermeiden
- **Aktionen:** gehe *ein* Feld nach Nord/Süd/West/Ost  
(Erreichen der Wand → Agentenposition bleibt unverändert)



→ Mit Ausnahme der Endzustände gibt es kein Indiz für den Nutzen der Zustände

→ Die Nutzenfunktion muss auf der Bewertung von Aktionsfolgen gründen:

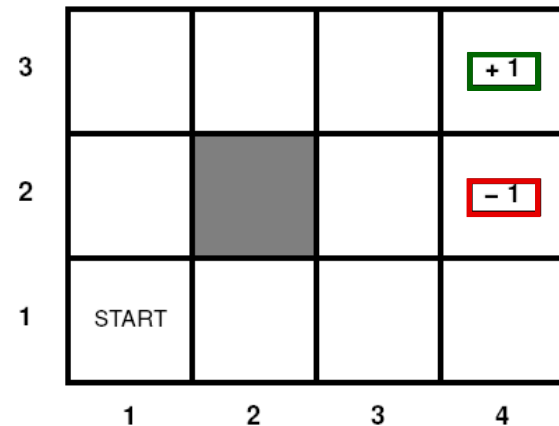
$$\text{Bsp.: } U(\text{Aktionsfolge}) = U(\text{Endzustand}) - 0,04 \cdot L(\text{Aktionsfolge})$$

mit  $L(\text{Aktionsfolge})$  = Länge der Aktionsfolge und Wegekosten pro nichtterminal. Feld = -0,04

~ Aktionsfolge der Länge 10 zum **Zustand (4,3)** hat Gesamtnutzen 0,6

# Deterministische Variante

- *Deterministische Variante*: Alle Aktionen führen immer zum nächsten Feld in der gewählten Richtung
- Der *deterministische Fall* ist durch
  - uninformierte Suchverfahren
  - informierte Suchverfahren
  - Planungsverfahren



lösbar, da der Agent alle Folgezustände genau voraussagen kann

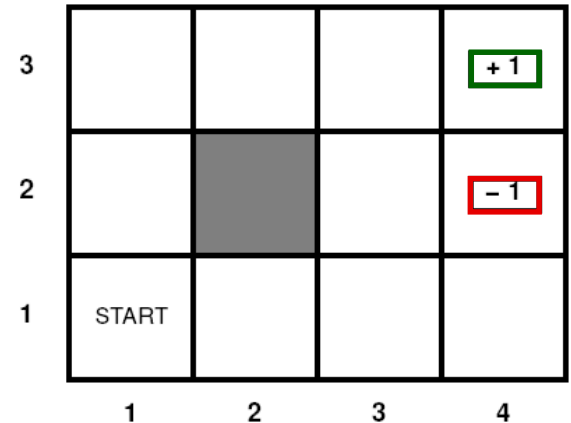
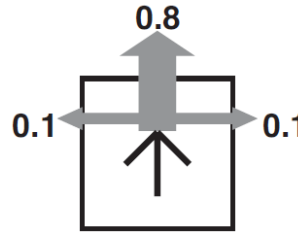
# Stochastische Variante

- *Stochastische Variante*: Intendierter Effekt trete mit 80% ein, mit jeweils 10% bewege sich der Agent rechtwinklig zur gewünschten Richtung:

$$P((1,2) \mid Do(nord, (1,1))) = 0,8,$$

$$P((2,1) \mid Do(nord, (1,1))) = 0,1,$$

$$P((1,1) \mid Do(nord, (1,1))) = 0,1.$$



Im *stochastischen Fall* könnte der Agent das MEU-Prinzip auf Aktionsfolgen anwenden und die erste Aktion der optimalen Aktionsfolge wählen usw.

- ~ Agent muss der *gesamten* Aktionsfolge vor Ausführung *vertrauen*
- ~ unvorteilhaft, wenn etwas falsch läuft, *aber vollständig beobachtbar ist*
- ~ flexiblerer Ansatz möglich, der neue Sensorinformation integriert
- ~ **Markov-Entscheidungsproblem**

○

# Markov-Entscheidungsproblem (1)

---

- Die Spezifikation der Ergebniswahrscheinlichkeiten

- für jede Aktion
- in jedem Zustand

wird als *Übergangsmodell* oder *Transitionsmodell* bezeichnet

- Wir gehen davon aus, dass es sich um *Markov-Übergänge* handelt:

- die Wahrscheinlichkeit, Zustand  $s'$  vom Zustand  $s$  aus zu erreichen, ist nur von  $s$  abhängig
- ... und damit *von der bisherigen Zustandshistorie unabhängig*

- Weil das Entscheidungsproblem sequentiell ist, ist die Nutzenfunktion von einer Zustandsfolge abhängig. Der Nutzen einer Zustandsfolge ergebe sich wie im Beispiel einfach aus der Summe aller erhaltenen Gewinne.



# Markov-Entscheidungsproblem (2)

---

Die Spezifikation eines sequentiellen Entscheidungsproblems

- für eine *vollständig beobachtbare Umgebung*
- mit *Markov-Übergangsmodell*
- mit *additiven Gewinnen*

wird auch als ***Markov Entscheidungsproblem (MDP)*** bezeichnet.\*

Ziel ist die Berechnung einer *optimalen Strategie* (*Policy*):

- eine Strategie spezifiziert für jeden Zustand  $s$ , welche Aktion  $a$  auszuführen ist
- die optimale Strategie maximiert den erwarteten Nutzen

---

\* nach dem russischen Mathematiker Andrei A. Markov: Markovs Arbeiten sind so stark mit der Annahme der Zugänglichkeit bzw. vollständigen Beobachtbarkeit verknüpft, dass Entscheidungsprobleme oft klassifiziert werden als „Markov“ und „Nicht-Markov“.

# Markov-Entscheidungsproblem (3)

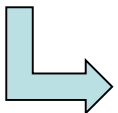
---

## Gegeben:

- Menge von Zuständen in einer zugänglichen, stochastischen Umgebung
- Menge von Zielzuständen
- Menge von Aktionen
- Transitionsmodell  $M(s,a,s')$  bzw.  $M_{ss'}^a$
- Nutzenfunktion

**Transitionsmodell:**  $M(s,a,s')$  spezifiziert die W'keit, mit der Zustand  $s'$  erreicht wird, wenn die Aktion  $a$  in Zustand  $s$  ausgeführt wird

**Strategie:** spezifiziert vollständige Abbildung von allen möglichen Zuständen auf die möglichen Aktionen



**Gesucht:** optimale Strategie, die den erwarteten Nutzen maximiert

# MDP-basierter Agent

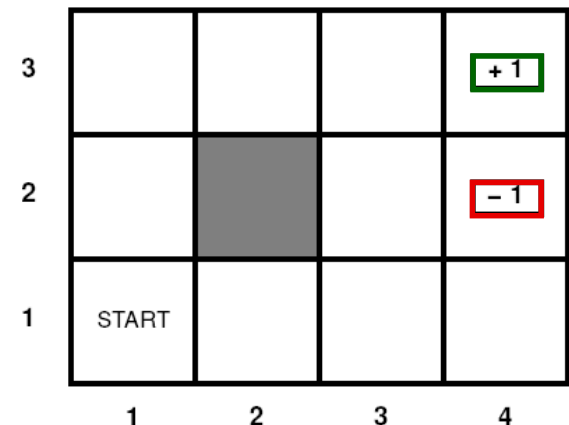
**function** SIMPLE-POLICY-AGENT(*percept*) **returns** an *action*

**static:**  $M$ , a transition model  
 $U$ , a utility function on environment histories  
 $P$ , a policy, initially unknown

**if**  $P$  is unknown **then**  $P \leftarrow$  the optimal policy given  $U, M$   
**return**  $P[\textit{percept}]$

Für das Beispiel sind  $M$  und  $U$ :

- $M$ :  $p(\text{Transition in gewählte Richtung}) = 80\%$   
 $p(\text{Transition in gewählte Richtung} + 90^\circ) = 10\%$   
 $p(\text{Transition in gewählte Richtung} - 90^\circ) = 10\%$
- $U(\text{Aktionsfolge}) = U(\text{Endzustand}) - 0,04 \cdot \# \text{Aktionen}$
- $P$ :  $\leadsto$  Herleitung durch sog. *Value Iteration*



# Value Iteration (1)

*Value Iteration*: ein Algorithmus zur Berechnung einer optimalen Strategie.

Grundidee:

- für jeden Zustand wird dessen Nutzen berechnet
- ausgehend von den Nutzen der möglichen *Nachfolgezustände* kann eine optimale Aktion für jeden Zustand ausgewählt werden

Dafür nötig:  
Separierbarkeit

*Aktionsfolge, Zustandsfolgen & Separierbarkeit:*

- eine *Aktionsfolge* generiert einen Baum möglicher *Zustandsfolgen* (*Historien*)
- eine Nutzenfunktion  $U_h$  auf Zustandsfolgen  $s_0, s_1, \dots, s_n$  heißt *separierbar* gdw. es eine Funktion  $f$  derart gibt, dass

$$U_h([s_0, s_1, \dots, s_n]) = f(s_0, U_h([s_1, \dots, s_n])).$$

Die einfachste Form ist eine *additive Gewinnfunktion*  $R$  (für *Reward*):

$$U_h([s_0, s_1, \dots, s_n]) = R(s_0) + U_h([s_1, \dots, s_n]).$$

Rewards im Beispiel:  $R((4,3)) = +1$ ,  $R((4,2)) = -1$ ,  $R(\text{sonstige}) = -0,04 = -1/25$

## Value Iteration (2)

Wegen der Additivität der Nutzenfunktion kann man den Nutzen  $U(s)$  eines Zustands  $s$  auf den maximalen erwarteten Nutzen seiner Nachfolger reduzieren:

$$U(s) = R(s) + \max_a \sum_{s'} M_{ss'}^a U(s')$$

Aus dieser sogenannten *Bellmann-Gleichung* folgt die optimale Strategie *policy*<sup>\*</sup>:

$$policy^*(s) = \arg \max_a \sum_{s'} M_{ss'}^a U(s')$$

**Operationalisierung:** Approximative Berechnung durch iterative Anwendung der sogenannten *Bellmann-Aktualisierung*:

$$U_{t+1}(s) \leftarrow R(s) + \max_a \sum_{s'} M_{ss'}^a U_t(s'),$$

wobei  $U_t(s)$  die Utility des Zustands  $s$  nach  $t$  Iterationen ist

**Man kann zeigen:** Mit  $t \rightarrow \infty$  konvergieren die Utilities aller Zustände

# Value Iteration (3)

```
function VALUE-ITERATION( $M, R$ ) returns a utility function
  inputs:  $M$ , a transition model
            $R$ , a reward function on states
  local variables:  $U$ , a utility function, initially identical to  $R$  *
                     $U'$ , a utility function, initially identical to  $R$ 
  repeat
     $U \leftarrow U'$ 
    for each state  $s$  do
       $U'[s] \leftarrow R[s] + \max_a \sum_{s'} M_{ss'}^a U[s']$ 
    end
  until CLOSE-ENOUGH( $U, U'$ )
  return  $U$ 
```

Zeitkomplexität: Pro Iteration quadratisch in der Zahl der Zustände und linear in der Zahl der Aktionen. Die Transitionsmatrix  $M$  ist i.A. dünn besetzt und eine mittlere Zahl von Nachfolgezuständen pro Zustand gegeben: dann Zeitkomplexität auch linear in der Zahl der Zustände.

---

\*  $U[s]$  und  $U'[s]$  könnten auch alle mit 0 initialisiert werden. Dann würde die erste Iteration diese auf die Rewards  $R(s)$  setzen.

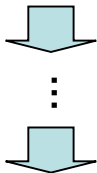
# Anwendung der Value Iteration (1)

Am Beispiel:

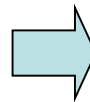
← **Start:** initiale Nutzen =  $R(s)$   
für alle Zustände  $s$

**Ergebnis** für die **Nutzen** der einzelnen Zustände nach Konvergenz und daraus resultierende **Strategie**:

3	-0.04	-0.04	-0.04	<span style="border: 1px solid green; padding: 2px;">+ 1</span>
2	-0.04		-0.04	<span style="border: 1px solid red; padding: 2px;">- 1</span>
1	-0.04	-0.04	-0.04	-0.04
	1	2	3	4

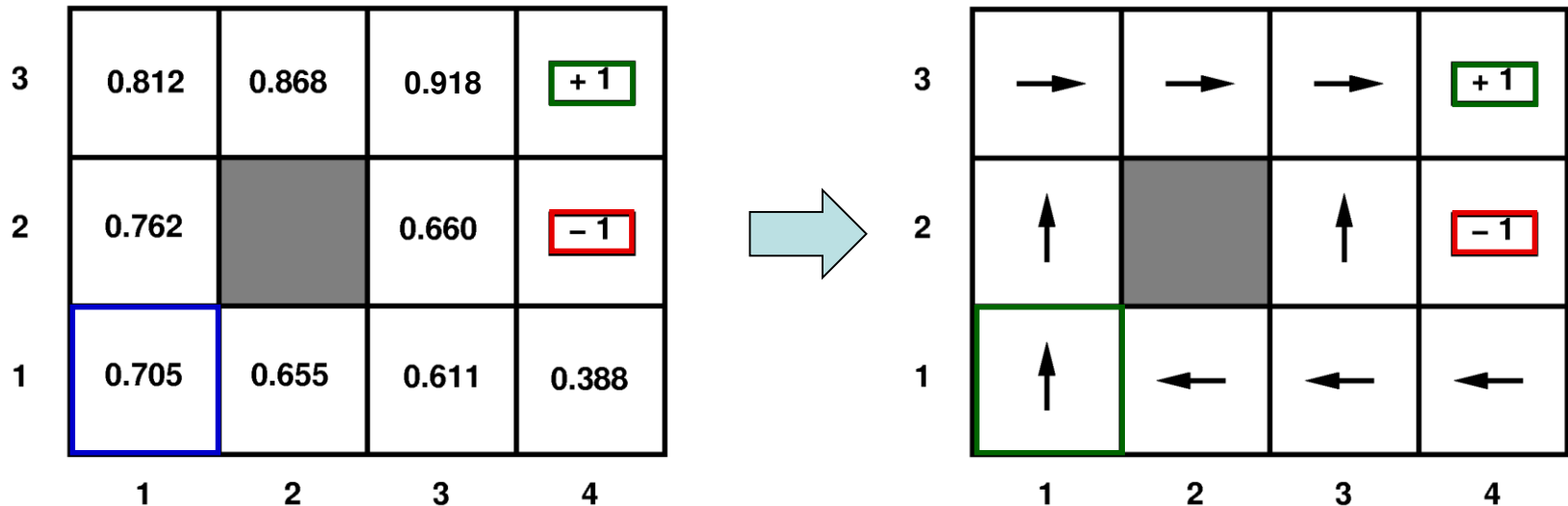


3	0.812	0.868	0.918	<span style="border: 1px solid green; padding: 2px;">+ 1</span>
2	0.762		0.660	<span style="border: 1px solid red; padding: 2px;">- 1</span>
1	0.705	0.655	0.611	0.388
	1	2	3	4



3	→	→	→	<span style="border: 1px solid green; padding: 2px;">+ 1</span>
2	↑		↑	<span style="border: 1px solid red; padding: 2px;">- 1</span>
1	↑	←	←	←
	1	2	3	4

# Anwendung der Value Iteration (2)



Ableitung der Strategie aus der errechneten Nutzenfunktion:

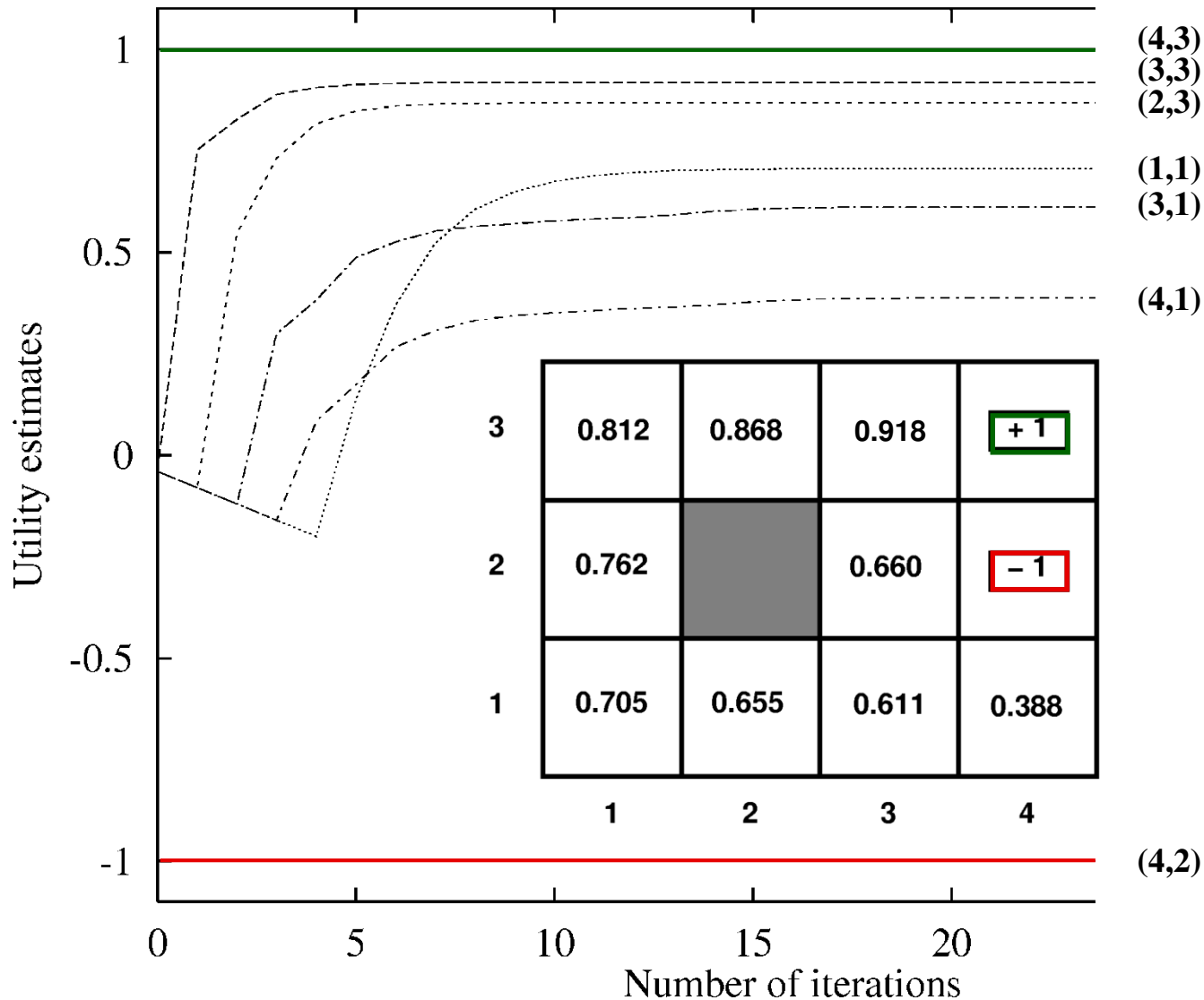
Geg.:  $U(1,1) = -0.04 + \max \{ 0.8 \cdot U(1,2) + 0.1 \cdot U(2,1) + 0.1 \cdot U(1,1), \quad (\text{Aktion } \textit{nord})$   
 $0.9 \cdot U(1,1) + 0.1 \cdot U(1,2), \quad (\text{Aktion } \textit{west})$   
 $0.9 \cdot U(1,1) + 0.1 \cdot U(2,1), \quad (\text{Aktion } \textit{süd})$   
 $0.8 \cdot U(2,1) + 0.1 \cdot U(1,2) + 0.1 \cdot U(1,1) \} \quad (\text{Aktion } \textit{ost})$   
 $= -0.04 + \max\{0.75 \text{ (nord)}, 0.71 \text{ (west)}, 0.70 \text{ (süd)}, 0.67 \text{ (ost)}\}.$

→ also wird in Feld [1,1] die **Aktion nord** gewählt.



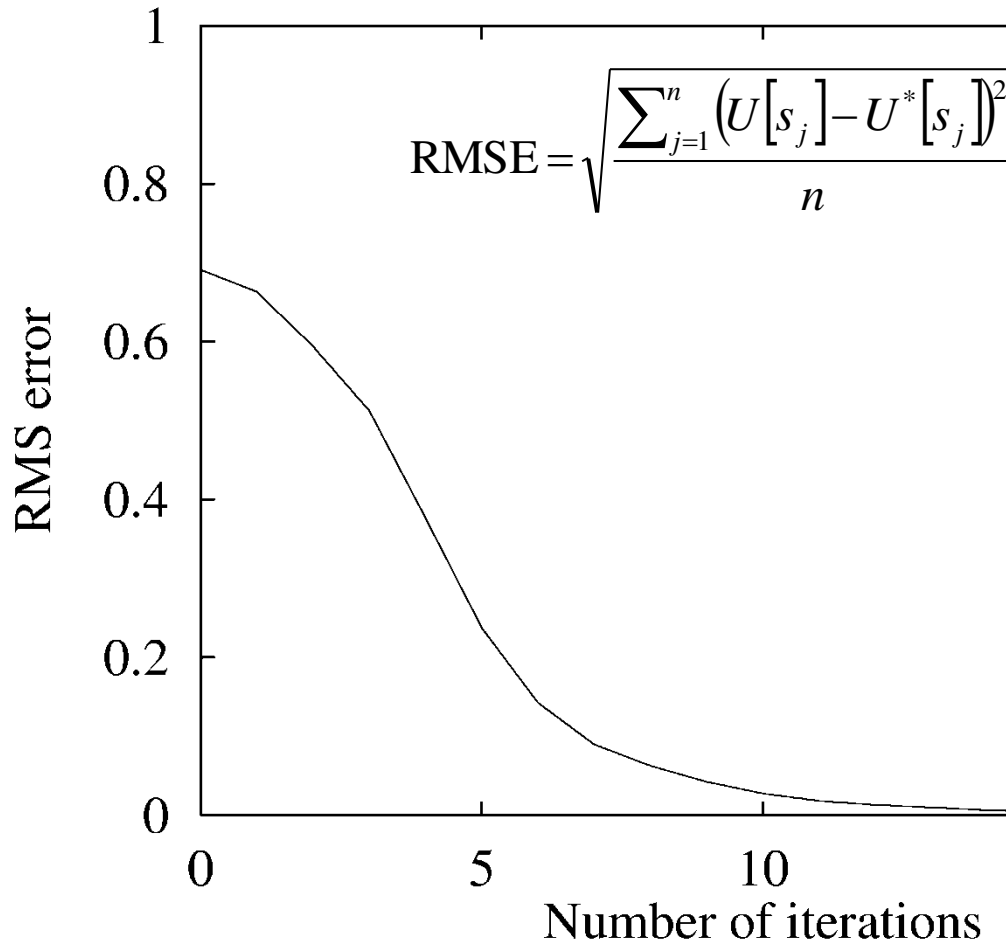
# Anwendung der Value Iteration (3)

Am Beispiel: Konvergenz der Nutzenwerte



# Anwendung der Value Iteration (4)

Am Beispiel: Konvergenz der Nutzenfehler



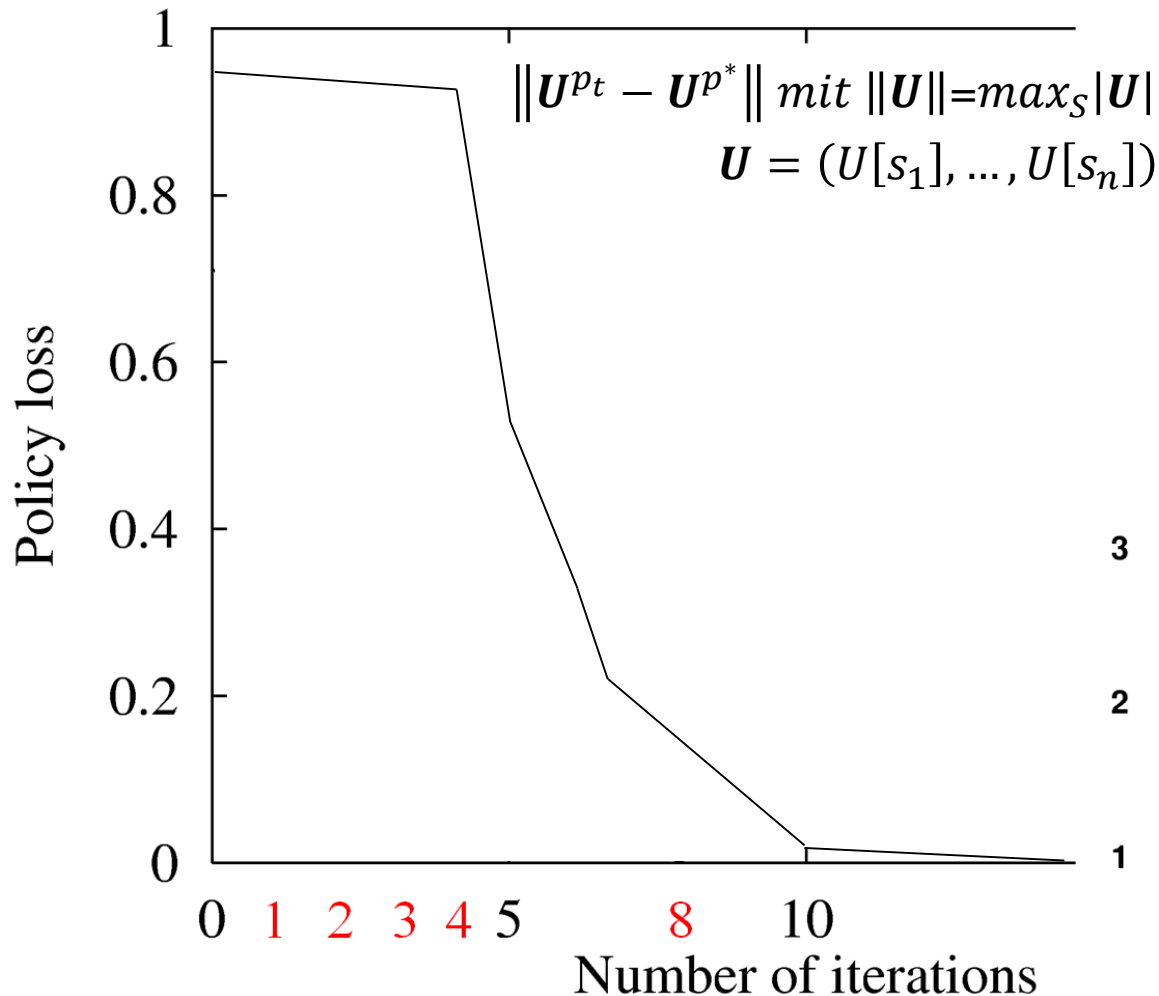
*Root Mean Square Error:*

Wurzel aus dem mittleren quadrat. Fehler der Werte der aktuellen Nutzenfunktion  $U$  im Vergleich zur korrekten Nutzenfunktion  $U^*$ .

3	0.812	0.868	0.918	+ 1
2	0.762		0.660	- 1
1	0.705	0.655	0.611	0.388
	1	2	3	4

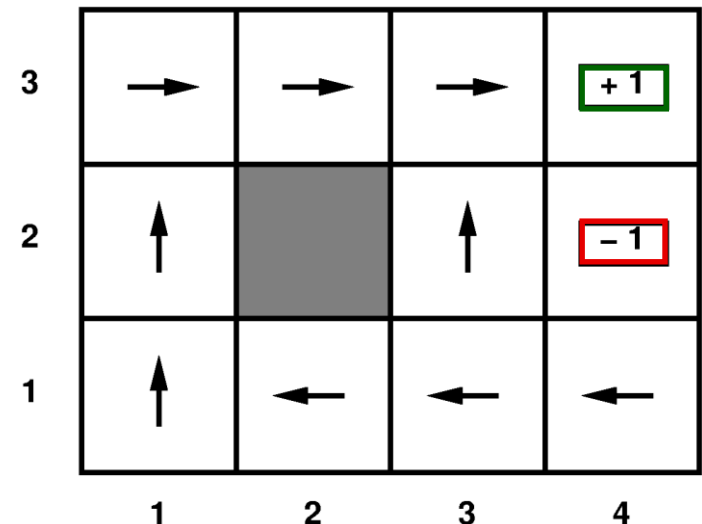
# Anwendung der Value Iteration (5)

Am Beispiel: Konvergenz der Strategie



Policy loss:

Unterschied zwischen den Nutzen der aktuellen Strategie  $p_t$  im Vergleich zur optimalen Strategie  $p^*$ .

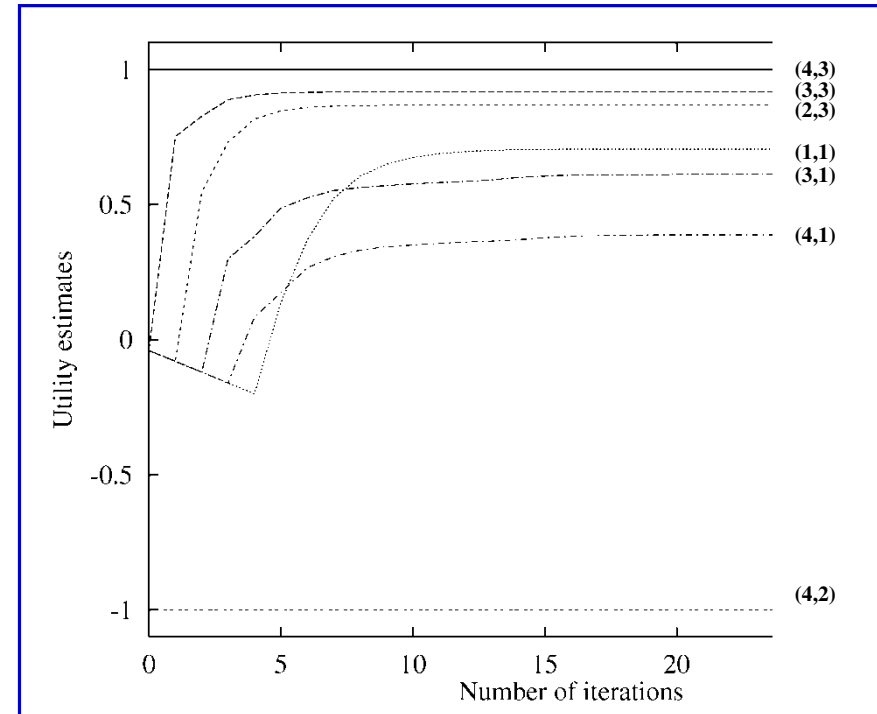


# Anwendung der Value Iteration (6)

Am Beispiel:

Entwicklung der Nutzenwerte

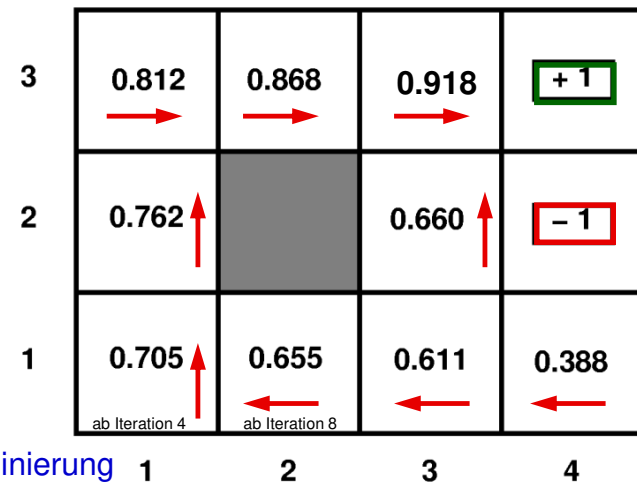
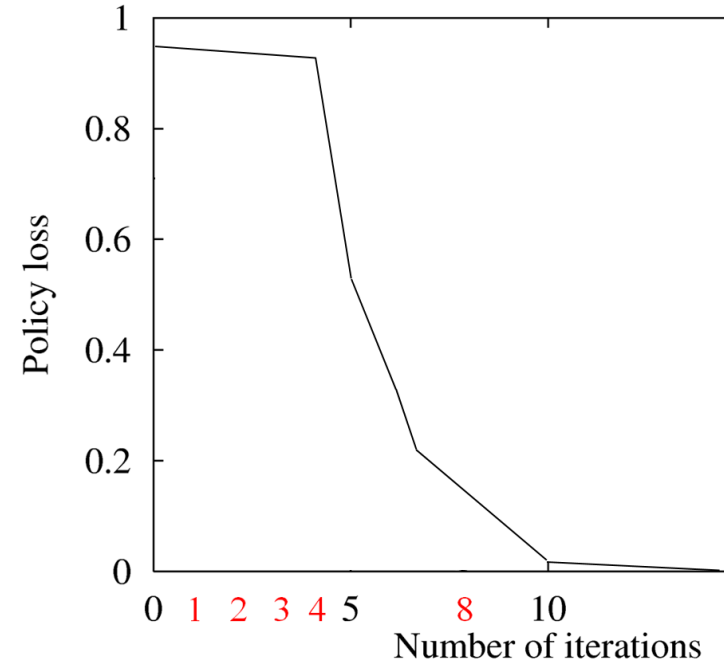
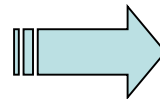
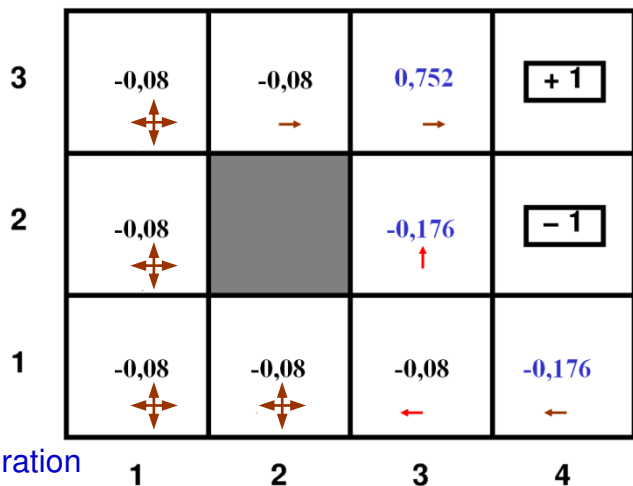
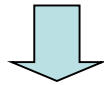
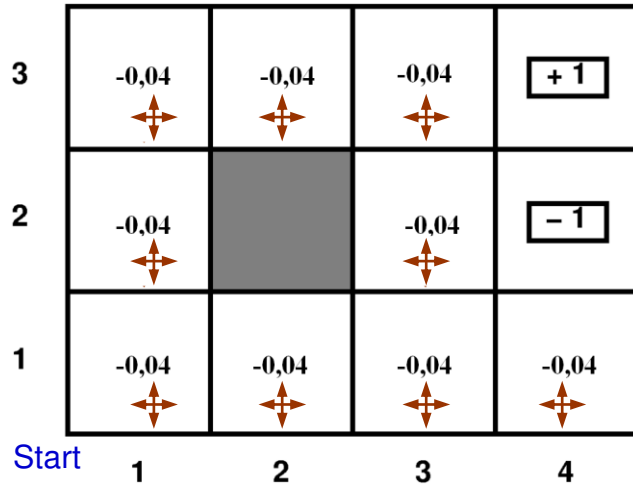
3	0.812	0.868	0.918	+ 1
2	0.762		0.660	- 1
1	0.705	0.655	0.611	0.388
	1	2	3	4



Zellen	Start	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7	Iteration 8	Iteration 9	Iteration 10
1,1	-0,04	-0,08	-0,12	-0,16	-0,2	0,1360704	0,36423424	0,51826067	0,60266986	0,64757771	0,67325386
1,2	-0,04	-0,08	-0,12	-0,16	0,225984	0,4530432	0,60215552	0,68145152	0,72270633	0,74306671	0,75290301
1,3	-0,04	-0,08	-0,12	0,37248	0,559808	0,6894336	0,75127552	0,78302003	0,7981568	0,80536209	0,808717
2,1	-0,04	-0,08	-0,12	-0,16	0,152832	0,2819264	0,40112832	0,45682579	0,49145656	0,5404272	0,5861476
x											
2,3	-0,04	-0,08	0,5456	0,7232	0,813568	0,8462848	0,85959616	0,86463706	0,86659472	0,86734265	0,86762998
3,1	-0,04	-0,08	-0,1296	0,28104	0,3642	0,4809288	0,52075016	0,55011425	0,56251699	0,56977208	0,57632569
3,2	-0,04	-0,176	0,444	0,55848	0,624776	0,6460488	0,65494408	0,65821223	0,65948853	0,65997256	0,66015871
3,3	-0,04	0,752	0,8176	0,88616	0,904464	0,912924	0,91589728	0,91708414	0,91752964	0,91770182	0,91776744
4,1	-0,04	-0,176	-0,2216	-0,26584	0,058248	0,1571848	0,26046152	0,30264628	0,33035603	0,3430492	0,35012259
4,2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
4,3	1	1	1	1	1	1	1	1	1	1	1

# Anwendung der Value Iteration (7)

Am Beispiel: Konvergenz der Strategie



# Stationäre und nicht-stationäre Strategien

Für die Entscheidungsfindung kann es einen **endlichen Horizont** oder ein **unendlichen Horizont** geben.

**Endlicher Horizont:** es gibt eine feste Zeit bzw. Schrittzahl, nach der nichts mehr geht (*Game over*).

Bei einem „engen“ zeitl. Horizont von z.B. 3 Schritten muss der Agent ggf. anders entscheiden als bei einem weiten Horizont von z.B. 100 Schritten. Z.B. hat der Agent bei nur 3 Schritten ggf. keine Zeit zum vorsichtigen „Probieren“, während dies bei 100 Schritten noch der Fall sein mag.

Bei einem endlichen Horizont kann sich die optimale Strategie für einen Zustand also mit der Zeit ändern und ist somit **nicht stationär**\*.

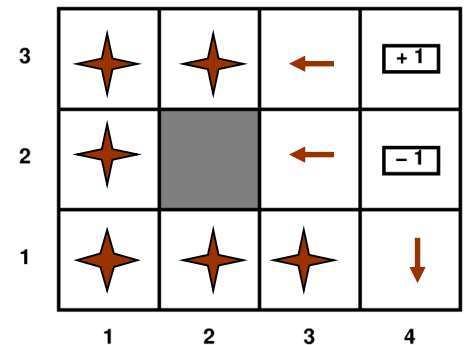
**Unendlicher Horizont:** Es liegen keine Zeitbegrenzungen vor. Die optimale Strategie ist dann stationär. *In dieser Vorlesung* werden nur Aufgaben mit unendlichem Horizont und damit **stationären** Strategien betrachtet.

\* Gilt nur für vollständig zugängliche Umgebungen.

# Unendliche Horizonte

Die Entscheidung für stationäre Strategien bei unendl. Horizonten kann Probleme verursachen, wenn die Umgebung keinen Zielzustand aufweist oder der Agent nie einen solchen findet!

**Bspl.:** mit  $R(s) = -0,04$  für alle Nichtterminalzustände  $s$  wurde eine richtige Strategie abgeleitet. Mit  $R(s) > 0$  würde der Agent mit jedem Schritt reicher und wird die Nähe der Felder (4,2) und (4,3) vermeiden, um einen unendlichen Gesamtgewinn zu erzielen (s. Abb.).



Ohne Zielzustände oder das Nichterreichen von Zielzuständen kommt es also zu unendlichen Umgebungsverläufen. Wie aber sind dann zwei Zustandsfolgen mit Nutzen  $+\infty$  vergleichbar?

Antwort: wir können den Agenten veranlassen, aktuelle Gewinne gegenüber fernen Gewinnen zu bevorzugen!

# Additive und verminderte Gewinne

---

Unter Stationarität gibt es zwei Möglichkeiten, Zustandsfolgen Nutzen zuzuweisen:

1. **Additive Gewinne**: der Nutzen einer Zustandsfolge ist

$$U_h([s_0, s_1, s_2, \dots]) = R(s_0) + R(s_1) + R(s_2) + \dots$$

2. **Verminderte Gewinne**: der Nutzen einer Zustandsfolge ist

$$U_h([s_0, s_1, s_2, \dots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots$$

Der **Verminderungsfaktor**  $\gamma$

- ist eine Zahl zwischen 0 und 1,
- beschreibt die Priorität des Agenten für aktuelle Gewinne gegenüber fernen

Gewinnen:

- kleines  $\gamma$  entspricht hoher Priorität für aktuelle Gewinne
- der Grenzfall  $\gamma = 1$  den additiven Gewinnen



# Value-Iteration mit verminderten Gewinnen

Bei verminderten Gewinnen sind die Nutzen unendlicher Folgen endlich. Wenn die Rewards durch  $R_{\max}$  begrenzt sind und  $\gamma < 1$ , gilt mit der Standardformel für unendliche geometrische Folgen:

$$U_h([s_0, s_1, s_2, \dots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = R_{\max} / (1 - \gamma).$$

**function** VALUE-ITERATION( $M, R$ ) returns a utility function

**inputs:**  $M$ , a transition model,  $R$ , a reward function on states

$\varepsilon$ , the maximum error allowed in the utility of any state

**local variables:**  $U, U'$ , utility vectors for all states, initially identical to  $R$

$\delta$ , the maximum change in utility of any state in an iteration

**repeat**

$U \leftarrow U', \delta \leftarrow 0$

**for each** state  $s$  **do**

$U'[s] \leftarrow R[s] + \gamma \cdot \max_a \sum_{s'} M_{ss'}^a U[s']$

**if**  $|U'[s] - U[s]| > \delta$  **then**  $\delta \leftarrow |U'[s] - U[s]|$

**end**

**until**  $\delta < \varepsilon \cdot (1 - \gamma) / \gamma$

**return**  $U$

Es gilt: wenn  $|U'[s] - U[s]| < \varepsilon \cdot (1 - \gamma) / \gamma$ ,  
dann  $|U'[s] - U_{\text{true}}[s]| < \varepsilon$

# Zusammenfassung

---

- Rationale Agenten können auf der Basis einer **Wahrscheinlichkeitstheorie** und einer **Nutzentheorie** entwickelt werden.
- Agenten, die ihre Entscheidungen entsprechend den Axiomen der **Nutzen-theorie** fällen, besitzen eine **Nutzenfunktion**.
- Sequentielle Probleme in zugänglichen, unsicheren Umgebungen (MDPs) können durch Berechnen einer **Strategie** gelöst werden.
- **Value Iteration** ist ein Verfahren zur Berechnung optimaler Strategien.