

13. Handeln unter Unsicherheit

Inhalt:

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

Grundlagen der Nutzentheorie

- Die Nutzenfunktion (*Utility function*) bewertet Zustände und formalisiert so das Bevorzugen bestimmter Zustände durch den Agenten.
- $U(S)$ Nutzen des Zustandes S für den Agenten
- Eine nichtdeterministische Aktion A kann zu den Folgezuständen $Result_i(A)$ führen.
- Wie hoch ist die Wahrscheinlichkeit, dass die Folgezustände $Result_i(A)$ erreicht werden, wenn A unter Evidenz E im aktuellen Zustand ausgeführt wird?

→ $P(Result_i(A) \mid Do(A), E)$

Grundlagen der Nutzentheorie

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

$$EU(A \mid E) = \sum_i P(Result_i(A) \mid Do(A), E) \times U(Result_i(A))$$

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

Probleme mit dem *MEU*-Prinzip

- Ansatz erfordert zunächst die Erfassung des initialen Weltzustandes (Perzeption, Wissensrepräsentation, Lernen).
- Die Berechnung der Folgezustände basierend auf $P(Result_i(A) \mid Do(A), E)$ erfordert prinzipiell vollständiges kausales Modell der Welt:
 - permanent an Veränderungen anpassen
 - i.A. NP-vollständig für Bayessche Netze
- Die Berechnung des erwarteten Nutzens $U(Result_i(A))$ erfordert Suche oder Planen. Denn um einen Zustand zu bewerten, muss man prinzipiell alle möglichen Folgezustände kennen („Wirkung des Zustands auf die Zukunft“).

Die Axiome der Nutzentheorie (1)

Die Rechtfertigung des *MEU*-Prinzips:

- Ein Agent verwende eine Nutzenfunktion, welche seine Performanz korrekt bewertet. Der Agent wird dann die maximale Performanzbewertung erzielen, wenn die Maximierung über alle möglichen Szenarien des Agenten führt: Maximierung des Durchschnittsnutzens.
- Dies bedeutet eine Reduktion des umfassenden Anspruchs der Performanzbewertung über umfangreiche Historien der Szenarien auf eine lokale Bewertung durch Maximierung der auf den Folgezustand anzuwendenden Nutzenfunktion.
- Ist *MEU*-Prinzip dennoch die einzig mögliche Wahl?
- Antwort: Notiere vernünftige Bedingungen für die Entscheidungsfindung eines rationalen Agenten und überprüfe, ob das *MEU*-Prinzip aus diesen folgt!

Die Axiome der Nutzentheorie (2)

Zunächst zur Sprache der Nutzentheorie:

- Komplexe Szenarien werden als *Lotterien* L bezeichnet,
- mögliche Ergebnisse sind dann mögliche Gewinne,
- das Ergebnis wird vom Zufall bestimmt.

Beispiele:

- 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 $P(A_i) = p_i$, $(i = 1; \dots ; n)$: $L = [p_1, A_1; \dots ; p_n, A_n]$
- Sonderfall: nur ein mögl. Ergebnis A : $L = [1, A]$ oder kurz $L = A$
- Eine Lotterie ist somit eine Wahrscheinlichkeitsverteilung über einer Menge möglicher Ergebnisse. Jedes einzelne Ergebnis einer Lotterie kann ein atomarer Zustand oder eine weitere Lotterie sein.

Die Axiome der Nutzentheorie (3)

Ziel ist nun die Ableitung von Präferenzen zwischen verschiedenen Lotterien aufgrund von Präferenzen zwischen den zugrunde liegenden Zuständen.

Die Notation für Präferenz bzw. fehlende Möglichkeit zur Präferenz des Agenten:

$A \succ B$ Agent bevorzugt A gegenüber B

$A \sim B$ Agent ist unentschieden bzgl. A und B

$A \succeq B$ Agent bevorzugt A gegenüber B oder ist
unentschieden

Welche „vernünftigen“ Bedingungen sollten nun für die Präferenzrelation gelten?

Die Axiome der Nutzentheorie (4)

Gegeben Zustände A, B, C.

■ **Anordenbarkeit**

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

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

■ **Transitivität**

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

Verletzen der Transitivität verursacht irrationales Verhalten:

$$A \succ B \succ C \succ A.$$

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 der Nutzentheorie (5)

- 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 B „for sure“ ist.

- Substituierbarkeit

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

Einfachere Lotterien können gegen kompliziertere Lotterien ersetzt werden, ohne dass sich die Unentschiedenheit ändert.

Die Axiome der Nutzentheorie (6)

- Monotonie

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

Bevorzugt ein Agent das Ergebnis A, dann muss er auch die Lotterie bevorzugen, die A mit höherer Wahrscheinlichkeit als Ergebnis liefert.

- Dekomponierbarkeit

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

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

Nutzenfunktion und Axiome

Aus Axiomen der Präferenzrelation folgt Existenz einer Nutzenfunktion!

1. *Utility Principle:*

Beachtet ein Agent in seinen Präferenzen Axiome, dann existiert eine reellwertige Funktion $U : S \rightarrow \mathbb{R}$ auf der Zustandsmenge S mit

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

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

2. *Maximum Expected Utility Principle:*

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

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

→ MEU-Prinzip: Maximiere $\sum_i p_i \times U(S_i)$

Wie entwirft man nun aber konkrete Nutzenfunktionen?

Monetäre Nutzenfunktionen

- . . . aus Marktmodellen! Dort hat die Nutzentheorie ihre Wurzeln, da sich Geld als universales Zahlungs- und Bewertungsmittel für alle Güter und Leistungen anbietet.**
- Um nun zu verstehen, wie man monetäre Entscheidungen unter Unsicherheit trifft, können wir das Verhalten von Agenten bei Entscheidungen mit Lotterien, bei denen Geld im Spiel ist, untersuchen.**
- Annahme: Wir haben bereits 1 Mio. Euro in einem Quiz gewonnen.**
- Angebot: Wir können eine Münze werfen und**
 - bei Kopf insgesamt 3 Mio. Euro gewinnen,**
 - bei Zahl alles verlieren.**
- Handeln wir irrational, wenn wir auf das Werfen der Münze verzichten?**

Expected Monetary Value

Betrachten wir den folgenden Ansatz:

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

$$\leadsto EU(Accept) = \frac{1}{2}(0 \text{ Euro}) + \frac{1}{2}(3.000.000 \text{ Euro}) = 1.500.000 \text{ Euro.}$$

$$\leadsto EU(Decline) = 1.000.000 \text{ Euro.}$$

Oder wäre ein anderer Ansatz sinnvoller?

Nutzenfunktion und Expected Monetary Value

Betrachten wir den folgenden zweiten Ansatz:

- Annahme: Die Münze ist fair.
- Annahme: Wir sind im Zustand S_k , in dem wir bereits k Millionen Euro besitzen.

$$\leadsto EU(Accept) = \frac{1}{2}U(S_k) + \frac{1}{2}U(S_{k+3.000.000}).$$

$$\leadsto EU(Decline) = U(S_{k+1.000.000}).$$

Wie sieht die Nutzenfunktion U aus?

Vermögen und Expected Monetary Value

Beobachtung: Der Nutzen, den wir aus Geld ziehen, ist ggf. nicht proportional zum Betrag des Geldes:

- Die erste Million Euro mag uns zu einer anderen Lebensweise führen.
- Mit der 503. Million wird sich unser Leben ggf. nur wenig ändern.

Übertragen auf unser Beispiel:

- Annahme: Die Münze ist fair.
- Annahme: Wir sind im Zustand S_k , in dem wir bereit k Millionen Euro besitzen.
- $U(S_k) = 5$, $U(S_{k+1.000.000}) = 8$, $U(S_{k+3.000.000}) = 10$.

$$\leadsto EU(\text{Accept}) = \frac{1}{2}U(S_k) + \frac{1}{2}U(S_{k+3.000.000}) = 2,5 + 5 = 7,5.$$

$$\leadsto EU(\text{Decline}) = U(S_{k+1.000.000}) = 8.$$

→ Der Agent lehnt die Wette mit dieser Nutzenfunktion U ab!.

St. Petersburg-Paradoxon (Bernoulli, 1738)

- Annahme: Wiederholtes Werfen einer fairen Münze.
- Chance: Gewinn von 2^n Euro, wenn Kopf beim n -ten Wurf das 1. Mal fällt.
- Frage: Welcher Einsatz wäre angemessen?

$$\leadsto EMV(St.P) = \sum_i P(Heads_i)MV(Heads_i) = \sum_i \frac{1}{2^i} 2^i = \frac{2}{2} + \frac{4}{4} + \frac{8}{8} + \dots = \infty.$$

→ Ein Agent, der EMV maximieren will, böte jede endliche Summe für die Chance!

Bernoulli löste das Paradoxon durch eine logarithmische Gewinnbewertung: $U(S_{k+n}) = \log_2 n$.

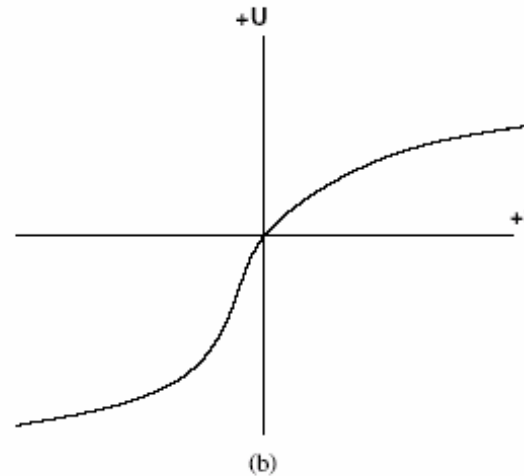
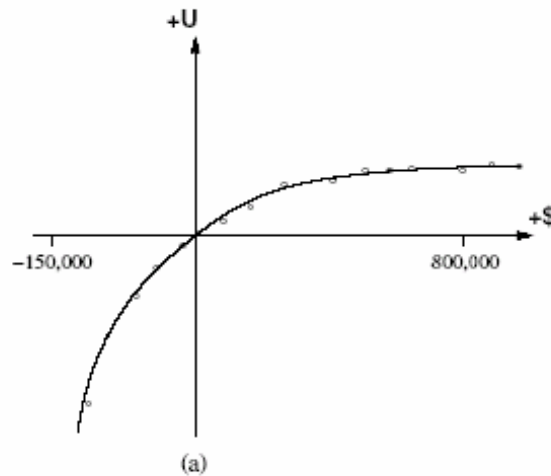
$$\leadsto EU(St.P) = \sum_i P(Heads_i)U(Heads_i) = \sum_i \frac{1}{2^i} \log_2 2^i = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \dots = 2.$$

→ Ein Agent mit Nutzenfunktion U bezahlt max. 4 Euro, da $U(S_{k+4}) = \log_2 4 = 2$.

(Alternativer Ansatz: Beschränkung des Gewinns!)

Verlauf von monetären Nutzenfunktionen

- Auch heute zeigen Studien zu aktuellen Nutzenfunktionen von Grayson (s. Abb. (a))
 - nahezu exakte Proportionalität zum Logarithmus,
→ abnehmenden Gradienten der Nutzenfunktion.
- Für den Schuldenfall kann eine analoge Differenzierung zwischen unterschiedlichen Schulden zu der S-förmigen Gesamtkurve führen (s. Abb. (b)).
- Mit dieser Nutzenfunktion kann ein Agent mit 10 Mio. \$ Schulden eine Wette mit fairer Münze und 10 Mio. \$ Gewinn für Kopf und 20 Mio. \$ Verlust für Zahl akzentuieren.



Risikoverhalten

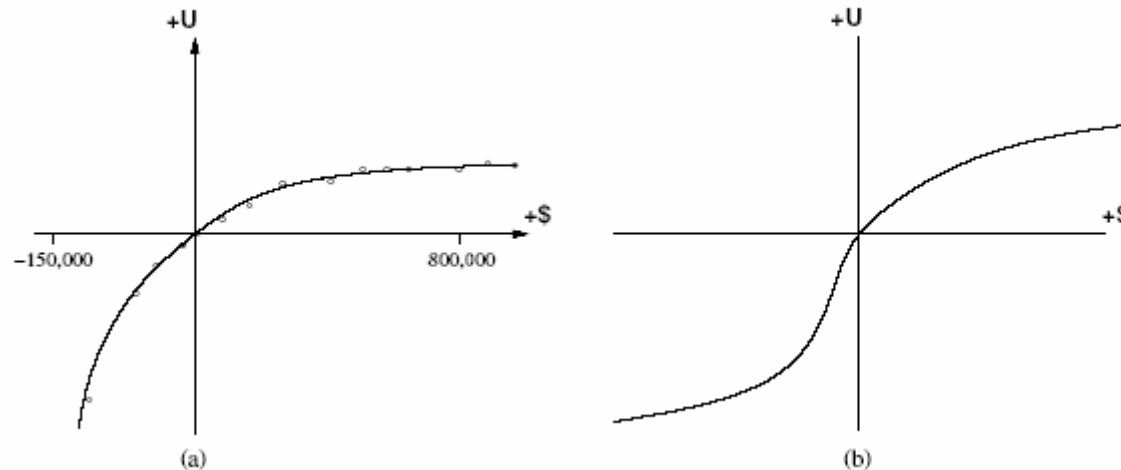
- Im positiven Teil der Nutzenfunktion ist der Agent *risikomeidend* (*Risc-Averse*), d.h.: der Agent bewertet die Lotterie immer geringer als den Erwartungswert:

$$U(S_L) < U(S_{EMV(L)}).$$

- Im negativen Teil der Nutzenfunktion ist der Agent *risikosuchend* (*Risc-Seeking*), d.h.: der Agent bewertet die Lotterie immer stärker als den Erwartungswert:

$$U(S_L) > U(S_{EMV(L)}).$$

- Für kleine Abschnitte der Nutzenfunktion ist der Agent *risikoneutral* (*Risc-Neutral*), d.h.: die Funktion ist fast linear: $U(S_L) \approx U(S_{EMV(L)})$.



Skalierung und Normierung

- Zwei Agenten mit Nutzenfunktionen U bzw. U' mit $U'(S) = k_1 + k_2 U(S)$ mit positiven Konstanten k_1, k_2 verhalten sich gleich.

→ Skalierung der Nutzen zwischen

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

→ Normierung der Nutzen:

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

- Zwischenwerte erhält man durch Variieren der Wahr-scheinlichkeit p in einer Standardlotterie

und Abgleichen mit möglichem Ergebnis $u_\perp = 0$ S , bis der Agent zu S und Lotterie indifferent ist:

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

→ $U(S) = p$ bei normierten Nutzen.

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

Sequentielle Entscheidungsprobleme

Bislang: Einfache Entscheidungsprobleme

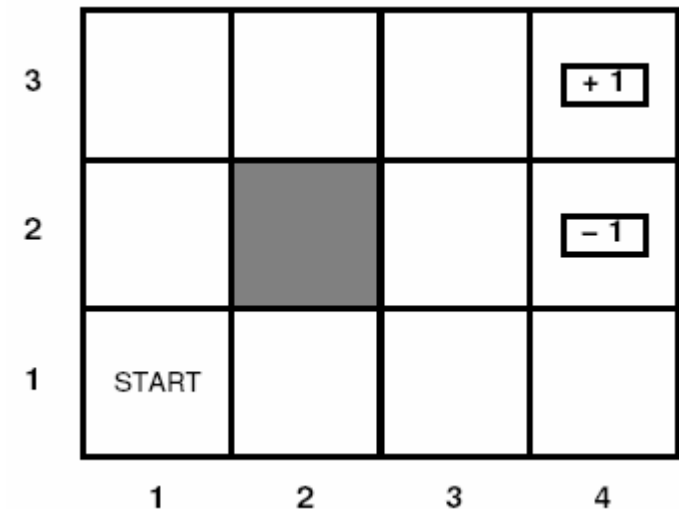
→ Nutzen für jedes Ergebnis einer Aktion bekannt

Jetzt: Sequentielle Entscheidungsprobleme

→ Nutzen ist vom Ergebnis einer Aktionsfolge abhängig.

Bsp.: Der Agent soll den Zustand (4; 3) erreichen
(Belohnung +1) und den Zustand (4; 2)
vermeiden (Bestrafung)

- Aktionen sind: nord, süd, west, ost.
- Mit Ausnahme der beiden Endzu-stände (+1/-1) gibt es kein Indiz für den Nutzen der Zustände!



- Nutzenfunktion muss auf Bewertung von Aktionsfolgen gründen!

Bsp.: $U(\text{Aktionsfolge}) = U(\text{Endzustand}) - 1/25 L(\text{Aktionsfolge})$

mit $L(\text{Aktionsfolge}) = \text{Länge der Aktionsfolge}$

Aktionsfolge der Länge 6 zum +1-Zustand hat Nutzen 0,76.

Deterministische und stochastische Variante

- ***Deterministische Variante:*** Alle Aktionen führen immer zum nächsten Feld in gewählter Richtung, beim Erreichen der Wand bleibt die Agentenposition unverändert.
- ***Stochastische Variante:*** Intendierter Effekt tritt mit 0.8 ein, mit 0.2 bewegt sich Agent rechtwinklig zur gewünschten Richtung:
 - $P((1, 2) \mid \text{Do}(\text{nord}, (1, 1))) = 0,8$
 - $P((2, 1) \mid \text{Do}(\text{nord}, (1, 1))) = P((1, 1) \mid \text{Do}(\text{nord}, (1, 1))) = 0,1$
- **Der deterministische Fall ist durch uninformierte oder informierte Suchverfahren lösbar, da der Agent alle Folgezustände genau voraussagen kann.**
- **Im stochastischen Fall könnte Agent das *MEU*-Prinzip auf Aktionsfolgen anwenden und erste Aktion der optimalen Aktionsfolge wählen usw., aber**
 - Agent muss gesamter Aktionsfolge vor Ausführung vertrauen!
 - Unvorteilhaft, wenn etwas falsch läuft, aber beobachtbar ist!
 - Flexiblerer Ansatz nötig, der neue Sensorinformation integriert!

Markov Entscheidungsproblem (MDP)

- Die Berechnung einer optimalen Strategie (Policy) in einer zugänglichen, stochastischen Umgebung mit bekanntem Modell der Zustandsübergänge wird als Markov Entscheidungsproblem (MDP) bezeichnet (nach dem russischen Mathematiker Andrei A. Markov).
- Markovs Arbeiten sind so stark mit der Annahme der Zugänglichkeit verknüpft, dass Entscheidungsprobleme oft klassifiziert werden als „Markov“ und „Nicht-Markov“.
- Die Markov-Eigenschaft gilt, wenn die Zustandsübergänge für alle Zustände nur vom aktuellen Ausgangszustand selbst, nicht aber von der bisherigen Historie abhängen.

Markov Entscheidungsproblem (MDP)

Gegeben:

- Menge von Zuständen in zugänglicher, stochastischer Umgebung
- Menge von Zielzuständen
- Menge von Aktionen
- Transitionsmodell M_{ij}^a
- Nutzenfunktion

Transitionsmodell: M_{ij}^a ist die Wahrscheinlichkeit, dass Zustand j erreicht wird, wenn Aktion a in Zustand i ausgeführt wird.

Policy: Vollständige Abbildung von allen möglichen Zuständen auf die möglichen Aktionen.

Gesucht: **Optimale Strategie (Policy)**, 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

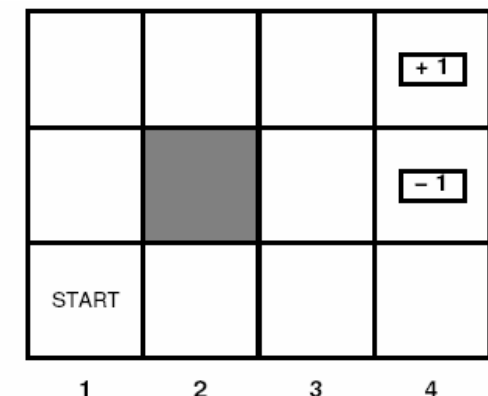
P , a policy, initially unknown

if P is unknown **then** $P \leftarrow$ the optimal policy given U, M

return $P[\textit{percept}]$

Beispiel:

- M : $p(\text{Transition in gewählte Richtung}) = 0.8$,
 $p(\text{Transition in gewählte Richtung} + 90^\circ) = 0.1$
 $p(\text{Transition in gewählte Richtung} - 90^\circ) = 0.1$.
- $U(\text{Aktionsfolge}) = U(\text{Endzustand}) - \frac{\# \text{Aktionen}}{25}$.



Value-Iteration (1)

- Ein Algorithmus zur Berechnung einer optimalen Strategie.
- Grundidee: für jeden Zustand wird sein Nutzen berechnet. Ausgehend vom Nutzen kann eine optimale Aktion für jeden Zustand ausgewählt werden.
- Eine Aktionsfolge generiert einen Baum möglicher Zustände (Histories). Eine Nutzenfunktion U_h auf Histories heißt separierbar gdw. es eine Funktion f derart gibt, dass

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

Einfachste Form ist additive Belohnungsfunktion R (*Reward*):

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

Im Beispiel:

$$R((4, 3)) = +1, \quad R((4, 2)) = -1, \quad R(\text{sonstige}) = -\frac{1}{25}.$$

Value-Iteration (2)

Der Nutzen eines Zustandes i ist durch den erwarteten Nutzen der optimalen Strategie unter Transitionsmodell M in i bestimmt:

$$\begin{aligned} U(i) &= EU(H(i, policy^*) \mid M) \\ &= \sum P(H(i, policy^*) \mid M) \cdot U_h(H(i, policy^*)) \end{aligned}$$

mit optimaler Strategie $policy^*$, dem Historienbaum $H(state, policy)$ aufgrund der Anwendung von $policy$ auf $state$ und dem Nutzen U_h einer Historie.

Wegen der Additivität der Nutzenfunktion kann man die Utility eines Zustands auf den maximalen erwarteten Nutzen seiner Nachfolger reduzieren:

$$\begin{aligned} policy^*(i) &= \operatorname{argmax}_a \sum_j M_{ij}^a U(j) \\ U(i) &= R(i) + \max_a \sum_j M_{ij}^a U(j) \end{aligned}$$

Value-Iteration (3)

- Sind die Nutzen der Endzustände bekannt, dann lässt sich in bestimmten Fällen ein **n-Schritt-Entscheidungsproblem** auf das Berechnen der Nutzen der Endzustände des **(n-1)-Schritt-Entscheidungsproblems** reduzieren.
- Iteratives und effizientes Verfahren durch dynamische Programmierung
 - **n-Schritt-Entscheidungsproblem**
 - **|A| mögliche Aktionen für jeden Schritt**
 - **|S| mögliche Zustände**
 - **Rechne Nutzen für alle Zustände des (n-1)-ten Schrittes mit Hilfe der bekannten Nutzen der Endzustände**
 - **Rechne Nutzen für alle Zustände des (n-2)-ten Schrittes usw.**
- **$O(n |A| |S|)$ statt $O(|A|^n)$ bei Aufzählung**

Value-Iteration (4)

- **Problem:** Typische Entscheidungsprobleme der KI enthalten Zyklen, so dass Histories eine potentiell unendliche Länge haben!
- **Lösung:** Approximative Berechnung durch iterative Anwendung von

$$U_{t+1}(i) \leftarrow R(i) + \max_a \sum_j M_{ij}^a U_t(j),$$

- wobei $U_t(i)$ die Utility des i -ten Zustands nach t Iterationen ist.
- **Beobachtung:** Mit $t \rightarrow \infty$ konvergieren die Utilities aller Zustände.

Value-Iteration (5)

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

inputs: M , a transition model

R , a reward function on states

local variables: U , utility function, initially identical to R

U' , utility function, initially identical to R

repeat

$U \leftarrow U'$

for each state i **do**

$U'[i] \leftarrow R[i] + \max_a \sum_j M_{ij}^a U[j]$

end

until CLOSE-ENOUGH(U, U')

return U

Anwendung der Value Iteration

Nutzen der Zustände nach Konvergenz und resultierende Policy:

3	0.812	0.868	0.912	$\boxed{+1}$	3	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\boxed{+1}$
2	0.762		0.660	$\boxed{-1}$	2	$\uparrow$		$\uparrow$	$\boxed{-1}$
1	0.705	0.655	0.611	0.388	1	$\uparrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
	1	2	3	4		1	2	3	4

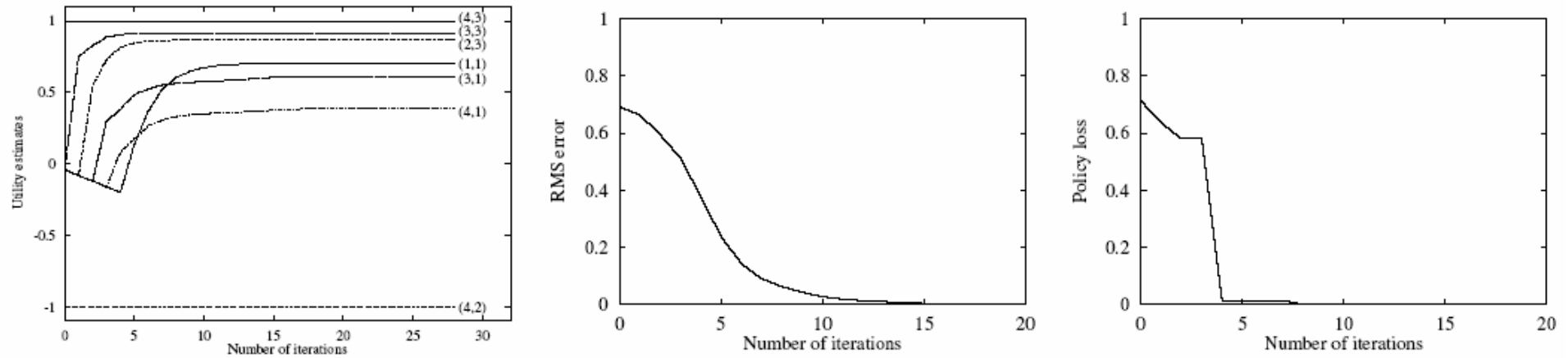
Ableitung der Strategie aus der Nutzenfunktion:

$$\begin{aligned}
 U(1,1) = -0.04 + \max \{ & 0.8 U(1,2) + 0.1 U(2,1) + 0.1 U(1,1), & (\text{nord}) \\
 & 0.9 U(1,1) + 0.1 U(1,2), & (\text{west}) \\
 & 0.9 U(1,1) + 0.1 U(2,1), & (\text{süd}) \\
 & 0.8 U(2,1) + 0.1 U(1,2) + 0.1 U(1,1) \} & (\text{ost})
 \end{aligned}$$

nord gewählt.

Anwendung der Value Iteration

Konvergenz von Nutzen, Nutzenfehlern und Policy-Fehler:



RMS error steht für Root Mean Square Error zwischen den aktuellen Nutzenwerten und den korrekten Nutzenwerten.

Policy loss bezeichnet die Differenz zwischen den erwarteten Nutzen aufgrund der aktuellen Policy und den erwarteten Nutzen aufgrund der korrekten Policy.

Zusammenfassung

- **Rationale Agenten können auf der Basis einer Wahrscheinlichkeitstheorie und einer Nutzentheorie entwickelt werden.**
- **Agenten, die Ihre Entscheidungen entsprechend den Axiomen der Nutzentheorie fällen, besitzen eine Nutzenfunktion.**
- **Sequentielle Probleme in unsicheren Umgebungen (MDP's) können durch Berechnen einer Policy gelöst werden.**
- **Value Iteration ist ein Verfahren zur Berechnung optimaler Policies.**