Stochastische dynamische Optimierung

- Bisher: Neuer Zustand s' auf Stufe $n+1$ ist durch alten Zustand s auf Stufe n und Aktion a eindeutig bestimmt.

$$s' = z_n(s, a)$$

- Jetzt: Neuer Zustand s' ist zusätzlich vom Zufall abhängig.
- Genauer: Zufallsvariable, deren Verteilung von s und a abhängt.
- Übergangwahrscheinlichkeit: $p_n(s, a, s')$

  Wahrscheinlichkeit für s' auf Stufe $n+1$ unter der Bedingung, dass auf Stufe n in Zustand s Aktion a gewählt wurde.

Basismodell

Definition 4.12

Ein endlich-stufiges stochastisches dynamisches Optimierungsproblem (SDO) ist ein Tupel $(N, \mathcal{S}, \mathcal{A}, p, r, V_N)$ mit (i), (ii), (iii), (v), (vi) wie in Definition 4.3 und

(iv) p, das Übergangsgesetz. Für alle $n < N$ und alle $(s, a) \in \mathcal{D}_n$ ist $p_n(s, a, \cdot)$ eine Wahrscheinlichkeitsfunktion (Zähldichte) auf $\mathcal{S}_{n+1}$.

Die Zahl $p_n(s, a, s')$ legt die Wahrscheinlichkeit fest, mit der Zustand $s' \in \mathcal{S}_{n+1}$ angenommen wird, bei Wahl von Aktion a in Zustand $s \in \mathcal{S}_n$.

Bemerkungen

- Nur die Zustansübergänge sind stochastisch, nicht die stufenbezogenen Gewinne und auch nicht die terminalen Kosten.
- Der Gesamtgewinn ist nun keine deterministische Größe mehr sondern eine Zufallsvariable.
- Wir werden daher den Erwartungswert des Gesamtgewinns maximieren.

Generalvoraussetzung

- Wir betrachten im Folgenden nur endliche Zustandsräume $\mathcal{S}$.
- Dies garantiert insbesondere, das die nachfolgend definierten Größen (Extremwerte) existieren.

Entscheidungsfunktion und Politik

Definition 4.13

Eine Funkion $f_n : \mathcal{S}_n \to \mathcal{A}$ mit $f_n(s) \in \mathcal{A}_n(s)$, die auf der Stufe $n < N$ jedem Zustand $s \in \mathcal{S}_n$ eine zulässige Aktion $a = f_n(s)$ zuordnet, heißt Entscheidungsfunktion.

Eine Folge $\delta = (f_0, \ldots, f_{N-1})$ von Entscheidungsfunktionen heißt Politik.

Die Menge aller Politiken bezeichnen wir wieder mit Δ.

Durch den erweiterten Politikbegriff mit Hilfe der Entscheidungsfunktionen ist durch eine Politik für jede eintretende Zustandsfolge auf jeder Stufe eine Aktion vorgegeben.

Optimale Politik

Definition 4.14

Bei Anwendung einer Politik δ und einer Realisation $(s_1, \ldots, s_N)$ der Zustände ergibt sich der Gesamtgewinn

$$R_{s_0\delta}(s_1, \ldots, s_N) := \sum_{n=0}^{N-1} r_n(s_n, f_n(s_n)) + V_N(s_N).$$

Für eine Politik δ ist der erwartete Gesamtgewinn

$$\overline{R}_\delta(s_0) := \sum_{s_1 \in \mathcal{S}_1} \cdots \sum_{s_N \in \mathcal{S}_N} R_{s_0\delta}(s_1, \ldots, s_N) P_{s_0\delta}(I_1 = s_1, \ldots, I_N = s_N)$$

mit

$$P_{s_0\delta}(I_1 = s_1, \ldots, I_N = s_N) = \prod_{n=0}^{N-1} p_n(s_n, f_n(s_n), s_{n+1}).$$

Fortsetzung Definition.

Eine Politik δ^* heißt optimal, wenn

$$\overline{R}_{\delta *}(s_0) \geq \overline{R}_{\delta}(s_0) \text{ für alle } \delta \in \Delta$$

gilt.

Der maximale erwartete Gesamtgewinn ist gegeben durch

$$V_0(s_0) := \max\{\overline{R}_{\delta}(s_0) | \delta \in \Delta\}.$$

Herleitung der Optimalitätsgleichung (1)

$$\overline{R}_{n\delta}(s_n) := \sum_{s_{n+1}\in\mathcal{S}_{n+1}} \cdots \sum_{s_n\in\mathcal{S}_N} R_{s_n\delta}(s_{n+1},\ldots,s_N) P_{s_n\delta}(s_{n+1},\ldots,s_N)$$

beschreibt der erwarteten Gesamtgewinn auf den Stufen n bis N, wenn auf Stufe n Zustand s_n vorlieht und Politikk δ angewendet wird. Hierbei sei

$$R_{s_n\delta} := \sum_{t=n}^{N-1} r_t(s_t, f_t(s_t)) + V_N(s_N)$$

und

$$P_{s_n\delta}(s_{n+1},\ldots,s_N) := \prod_{t=n}^{N-1} p_t(s_t, f_t(s_t), s_{t+1}).$$

Herleitung der Optimalitätsgleichung (2)

Der maximale erwartete Gesamtgewinn auf den Stufen n bis N ist dann gegeben durch

$$V_n(s) := \max\{\overline{R}_{n\delta}(s) | \delta \in \Delta\}, \quad s \in \mathcal{S}_n.$$

Definition 4.15

Für $n \leq N$ und $s \in \mathcal{S}_n$ heißt die Funktion $V_n(s)$ Wertfunktion.

Aus den Rechenregeln für Erwartungswerte ergibt sich

$$\overline{R}_{n\delta}(s) = r_n(s, f_n(s)) + \sum_{s' \in \mathcal{S}_{n+1}} p_n(s, f_n(s), s') \overline{R}_{n+1,\delta}(s')$$

mit $\overline{R}_{N\delta}(s) := V_N(s)$, $s \in \mathcal{S}_N$.

Optimalitätsgleichung

Satz 4.16

(i) *Für* $n = 0, \ldots, N-1$ *und alle* $s \in \mathcal{S}_n$ *gilt*

$$V_n(s) = \max_{a \in \mathcal{A}_n(s)} \{r_n(s,a) + \sum_{s' \in \mathcal{S}_{n+1}} p_n(s,a,s') V_{n+1}(s')\}.$$

(ii) *Jede aus den die rechte Seite von (ii) maximierende Aktionen* $f_n^*(s)$ *gebildete Politik* $\delta^* = (f_0^*, \ldots, f_{N-1}^*)$ *ist optimal.*

Beweis.

Tafel ✎.

Wertiteration für stochastische dynamische Optimierung

Algorithmus 4.17

(1) Für alle $s \in \mathcal{S}_N : v'(s) := V_N(s)$
(2) **for** $n = N-1$ **downto** 0 **do**
(3) Für alle $s \in \mathcal{S}_{n+1} : v(s) = v'(s)$
(4) Für alle $s \in \mathcal{S}_n$ berechne
(5) $v'(s) = \max_{a \in \mathcal{A}_n(s)} \{r_n(s,a) + \sum_{s' \in \mathcal{S}_{n+1}} p_n(s,a,s') v(s'))\}$
(6) und bestimme eine Entscheidungsfunktion f_n^*
aus den maximierenden Aktionen $a^* =: f_n^*(s) \in \mathcal{A}_n(s)$
(7) **end**
(8) $V_0(s_0) := v'(s_0)$
(9) $\delta^* := (f_0^*, \ldots, f_{N-1}^*)$

Beispiel 4.18

Wir erweitern Beispiel 4.10 wie folgt:

- Die Lagerkapazität sei auf 2 ME beschränkt.
- In jeder Periode n tritt Bedarf $b_n = 1$ mit Wahrscheinlichkeit 0.6 und Bedarf $b_n = 0$ mit Wahrscheinlichkeit 0.4 auf.
- Wenn der Bedarf in einer Periode nicht gedeckt ist, treten Fehlmengenkosten in Höhe von 20 GE auf.
- Mit Ausnahme der letzten Periode müssen Fehlmengen in der nächsten Periode ausgeglichen werden.
- Restbestände am Ende der Laufzeit können nicht verwertet werden, verursachen aber auch keine Kosten.
- Wie sieht eine Politik aus, die die erwarteten Beschaffungs- und Fehlmengenkosten minimiert?

Modellerweiterungen gegenüber Beispiel 4.10:

Fortsetzung Beispiel.

- Übergangsgesetz:

$$\begin{aligned} p(s,a,s+a-1) &= 0.6 \\ p(s,a,s+a) &= 0.4 \end{aligned}$$

- Zur Repräsentation der Fehlmengen führen wir den Zustand -1 ein. Für jede Aktion in diesem Zustand fallen stets zusätzliche Kosten von 20 GE an.

$$r_n(s,a) = \begin{cases} -q_n \cdot a & \text{für } s \geq 0 \\ -20 - q_n \cdot a & \text{für } s = -1 \end{cases}$$

- Tritt am Ende eine Fehlmenge auf, verursacht dies terminale Kosten von 20 GE.

$$V_4(s) = \begin{cases} 0 & \text{für } s \geq 0 \\ -20 & \text{für } s = -1 \end{cases}$$

Fortsetzung Beispiel.

- Fehlmengen müssen ausgeglichen und Lagerkapazitäten beachtet werden:

$$\mathcal{A}_n(s) = \begin{cases} \{0\} & \text{für } s = 2 \\ \{0,1\} & \text{für } s = 1 \\ \{0,1,2\} & \text{für } s = 0 \\ \{1,2\} & \text{für } s = -1 \end{cases}$$

Wertiteration:

$$\begin{aligned}
V_3(-1) &= \max\{-20-10+0.6(-20), -20-20\} = -40 \\
V_3(0) &= \max\{0.6(-20), -10, -20\} = -10 \\
V_3(1) &= \max\{0, -10\} = 0 \\
V_3(2) &= 0 \\
\hline
V_2(-1) &= \max\{-20-12+0.6(-40)+0.4(-10), \\
&\qquad -20-24+0.6(-10)\} = -50 \\
V_2(0) &= \max\{0.6(-40)+0.4(-10), -12+0.6(-10), -24\} = -18 \\
V_2(1) &= \max\{0.6(-10), -12\} = -6 \\
V_2(2) &= 0
\end{aligned}$$

Fortsetzung Beispiel.

$$
\begin{aligned}
V_1(-1) &= \max\{-20-9+0.6(-50)+0.4(-18), \\
&\qquad -20-18+0.6(-18)+0.4(-6)\} = -51.2 \\
V_1(0) &= \max\{0.6(-50)+0.4(-18), \\
&\qquad -9+0.6(-18)+0.4(-6), -18+0.6(-6)\} = -21.6 \\
V_1(1) &= \max\{0.6(-18)+0.4(-6), -9+0.6(-6)\} = -12.6 \\
V_1(2) &= 0.6(-6) = -3.6 \\
\hline
V_0(0) &= \max\{0.6(-51.2)+0.4(-21.6), \\
&\qquad -7+0.6(-21.6)+0.4(-12.6), \\
&\qquad -14+0.6(-12.6)+0.4(-3.6)\} = -23
\end{aligned}
$$

Optimale Politik: Tafel ✎.

Graphentheoretisches Modell (1)

vgl. Folie 172

- Zwei Arten von Knoten:
  - Zustandsknoten ◯ wie im deterministischen Fall
  - Zufallsknoten ♢ für die Modellierung der stochastischen Zustandsübergänge
- Für jeden Zustandsknoten (n, s) und jede Aktion $a \in \mathcal{A}_n(s)$ gibt es einen Zufallsknoten (n, s, a) und eine entsprechende gerichtete Kante von (n, s) nach (n, s, a).
- Für jeden Zufallsknoten (n, s, a) und jeden Zustand $s' \in \mathcal{S}_{n+1}$ mit $p_n(s, a, s') > 0$ gibt es eine Kante von (n, s, a) zu $(n + 1, s')$.

Graphentheoretisches Modell (2)

Es sei

$$V_n(s,a) := \sum_{s' \in \mathcal{S}_{n+1}} p_n(s,a,s') V_{n+1}(s')$$

der maximale erwartete Gesamtgewinn ab Stufe $n+1$ wenn in Zustand $s \in \mathcal{S}_n$ Aktion a gewählt wurde.

Stufenweise Berechnung von Stufe $N-1$ zu Stufe 0:

- An Zustandsknoten (n,s) bilden wir wie üblich das Maximum:

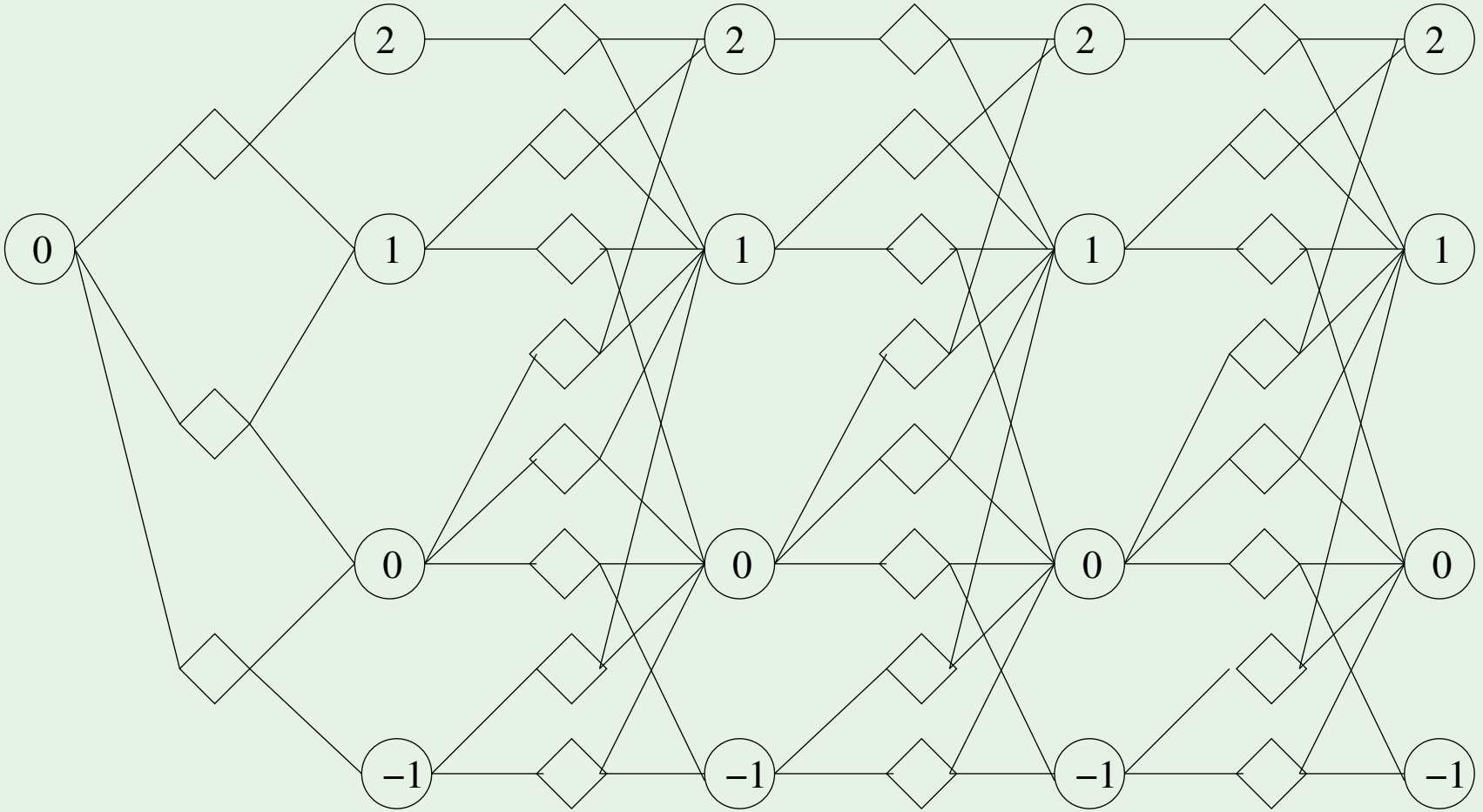
$$V_n(s) = \max_{a \in \mathcal{A}_n(s)} \{r_n(s,a) + V_n(s,a)\}$$

- An Aktionsknoten (n,s,a) bilden wir den Erwartungswert:

$$V_n(s,a) = \sum_{s' \in \mathcal{S}_{n+1}} p_n(s,a,s') V_{n+1}(s')$$

Beispiel 4.19

Graph zu Beispiel 4.18:



Berechnung: Tafel ✎.

Ausblick und verwandte Themen

- unendlicher Planungshorizont mit Diskontierungsfaktor $0 \leq \gamma < 1$
  ☞ Markow-Entscheidungsproblem, Markov Decision Process
- stetige Planung statt zu diskreten Zeitpunkten
  ☞ Kontrolltheorie
- Existenz eines "Gegners", mit abwechselnd zu treffenden Entscheidungen und Zustandsübergängen
  ☞ Spieltheorie
- unbekanntes Übergangsgesetz, optimale Politik lernen
  ☞ Q-Learning als Spezialfall des Reinforcement Learning