Optimale Kontrolle von Markovschen Entscheidungsproblemen

Ein Markovsches Entscheidungsproblem wird definiert durch ein Tupel $(S, A, ((P(a))_{a\in A}, c, T)$. Die Räume A und S seien hier endliche Mengen. Eine Strategie bezeichnen wir mit $\pi = (u_0, \ldots, u_{T-1})$ oder $\pi = (u_0, u_1, u_2, \ldots)$ für $T = \infty$.

1 Endlicher Zeithorizont

- Erwartete Kosten unter π bei Start in s: $\quad V^\pi(s) := \mathbb{E}_s^\pi \left[\sum_{t=0}^{T-1} c(X_t, Y_t) + c_T(X_T)\right]$
- optimale Strategie π^* : $\quad V^{\pi^*}(s) \leq V^\pi(s) \quad s \in S \qquad \forall \pi \in \Pi^H$
- Wertefunktion: $\quad V^*(s) := \inf_{\pi\in\Pi^H} V^\pi(s) \quad s \in S$
- erwartete Kosten unter π ab Zeitpunkt t bei Start in s

$$V_t^\pi(h_t) = \mathbb{E}^\pi \left[\sum_{n=t}^{T-1} c(X_n, Y_n) + c_T(X_T) \,\middle|\, Z_t = h_t\right]$$

- Policy Evaluation Iterationsschritt

$$V_t^\pi(h_t) := c(s_t, u_t(h_t)) + \sum_{j\in S} p(j \mid s_t, u_t(h_t)) V_{t+1}^\pi(h_t, u_t(h_t), j)$$

- minimale erwartete Kosten ab t: $\quad V_t^*(h_t) = \min_{\pi\in\Pi^H} V_t^\pi(h_t)$
- Bellman Gleichung Iterationsschritt

$$V_t(h_t) = \min_{a\in A} \left[c(s_t, a) + \sum_{j\in S} p(j \mid s_t, a) V_{t+1}(h_t, a, j)\right] \quad t = T-1, \ldots 0$$

Algorithmus (Backward Induction)**.** *1.* $V_T^*(s_T) = c_T(s_T)$

2. $V_t^* = \min_{a\in A} \left[c(s_t, a) + \sum_{j\in S} p(j \mid s_t, a) V_{t+1}^*(s_{t+1}\right]$

3. $A_{s,t}^* = argmin \left[c(s_t, a) + \sum_{j\in S} p(j \mid s_t, a) V_{t+1}^*(s_{t+1}\right]$

Mit $u_t^*(s_t) \in A_{s,t}^* \quad \forall s_t \in S$ und $\pi^* = (u_0^*(s_0), \ldots, u_{T-1}^*(s_{T-1})$ gilt dann, dass π^* optimale Strategie ist und $\pi^* \in \Pi^M$

2 Unendlicher Zeithorizont

Beim unendlichen Zeithorizont betrachtet man die Folge von Funktionen $V_n : S \to \mathbb{R}_+$, welche definiert ist durch

$$\begin{aligned} V_0(s) &= \inf_{a\in A} c(s,a), \\ V_{n+1}(s) &= \inf_{a\in A}\left\{ c(s,a) \; + \; \sum_{s'\in S} p_{s,s'}(a) V_n(s') \right\}. \end{aligned}$$

Satz (Value Iteration)**.** *Es gelten die folgenden Aussagen:*

(i) $V_n(s) \overset{n\to\infty}{\to} V^*(s)\ \forall s \in S.$

(ii) Gilt für eine stationäre Steuerung u^*

$$c(s,u^*(s)) + \sum_{s'} p_{s,s'}(u^*(s)) V^*(s') = \inf_{a\in A}\left\{ c(s,a) + \sum_{s'} p_{s,s'}(a) V^*(s') \right\}$$

für alle $s \in S$*, dann ist* u^* *optimale, d.h.*

$$V^{u^*}(s) = V^*(s) \;\; \forall s \in S.$$

Satz (Policy improvement)**.** *Für eine vorgegebene Steuerung* $w : S \to A$ *defienieren wir iterativ eine Folge von Steurungen durch*

$$w_0 := w, \quad w_{n+1}(s) := argmin_{a\in A}\left\{ c(s,a) + \sum_{s'\in S} p_{s,s'}(a) V^{w_n}(s') \right\}.$$

Dann bekommen wir, dass

(i) $V^{w_{n+1}} \le V^{w_n}$.

(ii) Durch die Enldichkeit von S *und* A *erreicht der Algorithmus in endlich vielen Schritten die Abbruchbedingung* $w_{n+1} = w_n =: w^*$*. Die Steuerung* w^* *ist dann eine optimale Steuerung.*

Satz. *1. Wir definieren die durch* α *diskontierte Folge* $(V_{n,\alpha})$ *durch* $V_{0,\alpha} := V_0$ *und*

$$V_{n+1,\alpha}(s) = \inf_a \left\{ c(s,a) + \alpha \sum_{s'} p_{s,s'}(a) V_{n,\alpha}(s'). \right\}$$

2. Für eine gegebene stationäre Steuerung u *definieren wir die diskontierte Folge* $w_0^\alpha := u$*,*

$$w_{n+1}^\alpha(s) := argmin_{a\in A}\left\{ c(s,a) + \alpha \sum_{s'} p_{s,s'}(a) V_\alpha^{w_n}(s') \right\}.$$

Mit diesen Definitionen erhält man die analogen Aussagen der vorhergehenden beiden Sätze auch für den diskontierten Fall.