Einführung MDP - Policy Iteration

Arne Huckemann

16.11.2022

Motivation

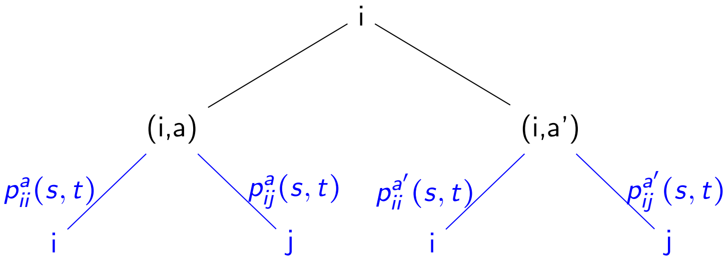
- Markov Decision Processes (MDP) aus 50/60er; Howard, Ronald A. (1960)
- Möglichst gute Entscheidungen finden

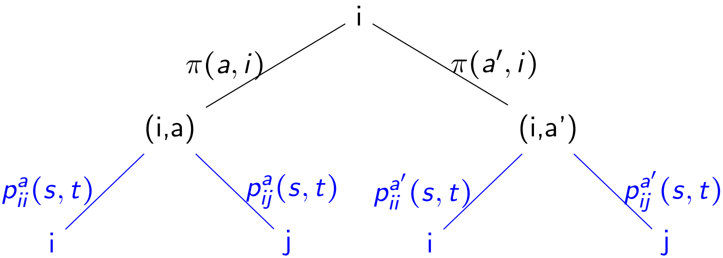
Motivation

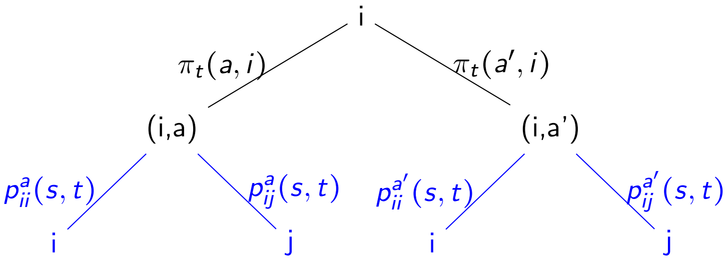
- Markov Decision Processes (MDP) aus 50/60er; Howard, Ronald A. (1960)
- Möglichst gute Entscheidungen finden
- Anwendung
  → Robotik
  → **Epidemiologie**

Motivation

- Markov Decision Processes (MDP) aus 50/60er;
  Howard, Ronald A. (1960)
- Möglichst gute Entscheidungen finden
- Anwendung
  → Robotik
  → **Epidemiologie**
- Wie findet man die optimale Lösung?
- Guo, X.; Hernández-Lerma, O. (2009) Continuous-Time Markov Decision Processes. Stochastic Modelling and Applied Probability, Springer


i
(i,a)
(i,a')
$p_{ii}^{a}(s,t)$
$p_{ij}^{a}(s,t)$
$p_{ii}^{a'}(s,t)$
$p_{ij}^{a'}(s,t)$
i
j
i
j


i
$\pi(a,i)$
$\pi(a',i)$
(i,a)
(i,a')
$p_{ii}^{a}(s,t)$
$p_{ij}^{a}(s,t)$
$p_{ii}^{a'}(s,t)$
$p_{ij}^{a'}(s,t)$
i
j
i
j


i
$\pi_t(a, i)$
$\pi_t(a', i)$
(i,a)
(i,a')
$p^{a}_{ii}(s, t)$
$p^{a}_{ij}(s, t)$
$p^{a'}_{ii}(s, t)$
$p^{a'}_{ij}(s, t)$
i
j
i
j

Annahme 1.1

- Zustands-Menge/Raum: $\mathcal{Z}$ abzählbar, mit $\mathbb{N}$ indiziert; $(\mathcal{Z}, \mathcal{P}(\mathcal{Z}))$

Annahme 1.1

- Zustands-Menge/Raum: $\mathcal{Z}$ abzählbar, mit $\mathbb{N}$ indiziert; $(\mathcal{Z}, \mathcal{P}(\mathcal{Z}))$
- Zeit: Kontinuierliche Zeit: $t \in [0, \infty) =: T$
- W-Raum: $(\Omega, \mathcal{F}, \mathbb{P})$

Annahme 1.1

- Zustands-Menge/Raum: $\mathcal{Z}$ abzählbar, mit $\mathbb{N}$ indiziert; $(\mathcal{Z}, \mathcal{P}(\mathcal{Z}))$
- Zeit: Kontinuierliche Zeit: t $\in [0, \infty) =: T$
- W-Raum: $(\Omega, \mathcal{F}, \mathbb{P})$

Definition 1.2 (Stochastischer Prozess)

- Familie von Zufallsvariablen
- x: $\Omega \times T \to \mathcal{Z}$, $(\omega, t) \mapsto x(\omega, t) = j$

  $\to$ $\mathcal{F}$ - $\mathcal{P}(\mathcal{Z})$ messbar

Annahme 1.1

- Zustands-Menge/Raum: $\mathcal{Z}$ abzählbar, mit $\mathbb{N}$ indiziert; $(\mathcal{Z}, \mathcal{P}(\mathcal{Z}))$
- Zeit: Kontinuierliche Zeit: t $\in [0, \infty) =: T$
- W-Raum: $(\Omega, \mathcal{F}, \mathbb{P})$

Definition 1.2 (Stochastischer Prozess)

- Familie von Zufallsvariablen
- x: $\Omega \times T \to \mathcal{Z}$, $(\omega, t) \mapsto x(\omega, t) = j$
  $\to$ $\mathcal{F}$ - $\mathcal{P}(\mathcal{Z})$ messbar
- x: $T \to \mathcal{Z}, t \mapsto x(t) = j$, (lassen ω weg)

Definition 1.3 (Markov Eigenschaft)

Ein stochastischer Prozess heißt Markov Prozess, falls

$\forall n \in \mathbb{N} : 0 \leq s_1 \leq \ldots \leq s_n \leq s \leq t < \infty;\ i_1, \ldots, i_n, i, j \in \mathcal{Z} :$

$$\mathbb{P}(x(t) = j \mid x(s) = i, x(s_n) = i_n, \ldots, x(s_1) = i_1) = \mathbb{P}(x(t) = j \mid x(s) = i)$$

Definition 1.3 (Markov Eigenschaft)

Ein stochastischer Prozess heißt Markov Prozess, falls

$$\forall n \in \mathbb{N} : 0 \leq s_1 \leq \ldots \leq s_n \leq s \leq t < \infty;\ i_1, \ldots, i_n, i, j \in \mathcal{Z} :$$

$$\mathbb{P}(x(t) = j \mid x(s) = i, x(s_n) = i_n, \ldots, x(s_1) = i_1) = \mathbb{P}(x(t) = j \mid x(s) = i)$$

- Übergangswahrscheinlichkeit:

$$p_{ij}(s,t) := \mathbb{P}(x(t) = j \mid x(s) = i)\ \forall s \leq t \text{ und } i,j \in \mathcal{Z}$$

Definition 1.3 (Markov Eigenschaft)

Ein stochastischer Prozess heißt Markov Prozess, falls

$\forall n \in \mathbb{N} : 0 \leq s_1 \leq \ldots \leq s_n \leq s \leq t < \infty;\ i_1, \ldots, i_n, i, j \in \mathcal{Z}:$

$$\mathbb{P}(x(t) = j \mid x(s) = i, x(s_n) = i_n, \ldots, x(s_1) = i_1) = \mathbb{P}(x(t) = j \mid x(s) = i)$$

- Übergangswahrscheinlichkeit:

$$p_{ij}(s,t) := \mathbb{P}(x(t) = j \mid x(s) = i)\ \forall s \leq t \text{ und } i, j \in \mathcal{Z}$$

Definition 1.4

- Stabil, wenn

$$\lim_{\Delta \to 0^+} p_{ij}(t, t + \Delta) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Bemerkung 1.5

$$\sum_{j \in \mathcal{Z}} p_{ij}(s,t) = 1 \ \forall s \leq t \text{ und } i,j \in \mathcal{Z}$$

Bemerkung 1.5

$$\sum_{j \in \mathcal{Z}} p_{ij}(s,t) = 1 \ \forall s \leq t \text{ und } i,j \in \mathcal{Z}$$

Definition 1.6

Übergangsmatrix: $P(s,t) := (\ p_{ij}(s,t)\ )_{(i,j) \in \mathcal{Z} \times \mathcal{Z}} \ \forall s \leq t < \infty$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t), \ \forall s \leq u \leq t$$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t), \ \forall s \leq u \leq t$$

Beweis:

$$\mathbb{P}(x(t) = j \mid x(s) = i) = \mathbb{P}\Big(\dot{\bigcup}_{k \in \mathcal{Z}} \ \{x(t) = j, x(u) = k\} \mid x(s) = i\Big)$$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t),\ \forall s \le u \le t$$

Beweis:

$$\mathbb{P}(x(t) = j \mid x(s) = i) = \mathbb{P}\Big(\dot\bigcup_{k \in \mathcal{Z}} \{x(t) = j, x(u) = k\} \mid x(s) = i\Big)$$

$$\overset{\sigma Add.}{=} \sum_{k \in \mathcal{Z}} \mathbb{P}(x(t) = j, x(u) = k \mid x(s) = i) \quad \mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid B, C)\mathbb{P}(B \mid C)$$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t), \ \forall s \leq u \leq t$$

Beweis:

$$\mathbb{P}(x(t) = j \mid x(s) = i) = \mathbb{P}\left(\dot{\bigcup}_{k \in \mathcal{Z}} \{x(t) = j, x(u) = k\} \mid x(s) = i\right)$$

$$\overset{\sigma Add.}{=} \sum_{k \in \mathcal{Z}} \mathbb{P}(x(t) = j, x(u) = k \mid x(s) = i) \quad \mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid B, C)\mathbb{P}(B \mid C)$$

$$\overset{Bed.WK}{=} \sum_{k \in \mathcal{Z}} \mathbb{P}(x(t) = j \mid x(u) = k, x(s) = i)\ \mathbb{P}(x(u) = k \mid x(s) = i)$$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t),\ \forall s \le u \le t$$

<u>Beweis:</u>

$$\mathbb{P}(x(t)=j \mid x(s)=i) = \mathbb{P}\left(\dot\bigcup_{k\in\mathcal{Z}} \{x(t)=j, x(u)=k\} \mid x(s)=i\right)$$

$$\overset{\sigma Add.}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j, x(u)=k \mid x(s)=i) \quad \mathbb{P}(A,B \mid C) = \mathbb{P}(A \mid B,C)\mathbb{P}(B \mid C)$$

$$\overset{Bed.WK}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j \mid x(u)=k, x(s)=i)\ \mathbb{P}(x(u)=k \mid x(s)=i)$$

$$\overset{Markov}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j \mid x(u)=k)\ \mathbb{P}(x(u)=k \mid x(s)=i)$$

Satz 1.7 (Chapman-Kolmogorov Equation)

$$P(s,t) = P(s,u)P(u,t),\ \forall s \le u \le t$$

<u>Beweis:</u>

$$\mathbb{P}(x(t)=j \mid x(s)=i) = \mathbb{P}\Big(\dot\bigcup_{k\in\mathcal{Z}} \{x(t)=j, x(u)=k\} \mid x(s)=i\Big)$$

$$\overset{\sigma Add.}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j, x(u)=k \mid x(s)=i) \quad \mathbb{P}(A,B \mid C) = \mathbb{P}(A \mid B,C)\mathbb{P}(B \mid C)$$

$$\overset{Bed.WK}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j \mid x(u)=k, x(s)=i)\ \mathbb{P}(x(u)=k \mid x(s)=i)$$

$$\overset{Markov}{=} \sum_{k\in\mathcal{Z}} \mathbb{P}(x(t)=j \mid x(u)=k)\ \mathbb{P}(x(u)=k \mid x(s)=i)$$

$$\overset{Def.}{=} \sum_{k\in\mathcal{Z}} p_{kj}(u,t)p_{ik}(s,u)$$

Satz und Definition 1.8

Für einen stabilen Markov Prozess gilt $\forall\ 0 \leq s:$

$$\infty > \lim_{t \to s^+} \frac{p_{ij}(s,t) - \delta_{ij}}{t - s} =: q_{ij}(s)$$

Satz und Definition 1.8

Für einen stabilen Markov Prozess gilt $\forall\ 0 \leq s$:

$$\infty > \lim_{t \to s^+} \frac{p_{ij}(s,t) - \delta_{ij}}{t - s} =: q_{ij}(s)$$

Beweis: Nur für Fall i = j:

Annahme: $p_{ii}(s,t) \neq 1$

Satz und Definition 1.8

Für einen stabilen Markov Prozess gilt $\forall\ 0 \leq s:$

$$\infty > \lim_{t \to s^+} \frac{p_{ij}(s,t) - \delta_{ij}}{t - s} =: q_{ij}(s)$$

<u>Beweis:</u> Nur für Fall i = j:

Annahme: $p_{ii}(s,t) \neq 1$

1) $f(s,t) := -log(p_{ii}(s,t))$

Satz und Definition 1.8

Für einen stabilen Markov Prozess gilt $\forall \ 0 \leq s$:

$$\infty > \lim_{t \to s^+} \frac{p_{ij}(s,t) - \delta_{ij}}{t - s} =: q_{ij}(s)$$

<u>Beweis:</u> Nur für Fall i = j:

Annahme: $p_{ii}(s,t) \neq 1$

1) $f(s,t) := -log(p_{ii}(s,t))$

$p_{ii}(s,t) = \sum_{k \in \mathcal{Z}} p_{ik}(s,u) p_{ki}(u,t) \geq p_{ii}(s,u) p_{ii}(u,t)$

Satz und Definition 1.8

Für einen stabilen Markov Prozess gilt $\forall\ 0 \leq s$:

$$\infty > \lim_{t \to s^+} \frac{p_{ij}(s,t) - \delta_{ij}}{t - s} =: q_{ij}(s)$$

Beweis: Nur für Fall i = j:

Annahme: $p_{ii}(s,t) \neq 1$

1) $f(s,t) := -log(p_{ii}(s,t))$

$p_{ii}(s,t) = \sum_{k \in \mathcal{Z}} p_{ik}(s,u) p_{ki}(u,t) \geq p_{ii}(s,u) p_{ii}(u,t)$

$\Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

1) $f(s,t) := -log(p_{ii}(s,t))$

2) $p_{ii}(s,t) \geq p_{ii}(s,u)p_{ii}(u,t) \Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

1) $f(s,t) := -log(p_{ii}(s,t))$

2) $p_{ii}(s,t) \geq p_{ii}(s,u)p_{ii}(u,t) \Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

$\Rightarrow^{[2]} \sup_{t \to s^+} \frac{f(s,t)}{t-s} = \lim_{t \to s^+} \frac{f(s,t)}{t-s}$

1) $f(s,t) := -log(p_{ii}(s,t))$

2) $p_{ii}(s,t) \geq p_{ii}(s,u)p_{ii}(u,t) \Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

$\Rightarrow^{[2]} \sup_{t \to s^+} \frac{f(s,t)}{t-s} = \lim_{t \to s^+} \frac{f(s,t)}{t-s}$

Damit Existenz gezeigt. Nun Eindeutigkeit:

$\lim_{t \to s^+} \frac{p_{ii}(s,t)-1}{t-s} = \lim_{t \to s^+} \frac{e^{-f(s,t)}-1}{t-s} = \lim_{t \to s^+} \frac{e^{-f(s,t)}-1}{f(s,t)} \frac{f(s,t)}{t-s}$

1) $f(s,t) := -log(p_{ii}(s,t))$

2) $p_{ii}(s,t) \geq p_{ii}(s,u)p_{ii}(u,t) \Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

$\Rightarrow^{[2]} \sup_{t \to s^+} \frac{f(s,t)}{t-s} = \lim_{t \to s^+} \frac{f(s,t)}{t-s}$

Damit Existenz gezeigt. Nun Eindeutigkeit:

$\lim_{t \to s^+} \frac{p_{ii}(s,t)-1}{t-s} = \lim_{t \to s^+} \frac{e^{-f(s,t)}-1}{t-s} = \lim_{t \to s^+} \frac{e^{-f(s,t)}-1}{f(s,t)} \frac{f(s,t)}{t-s}$

Da $\frac{e^{-f(s,t)}-1}{f(s,t)} = \frac{(1-f(s,t)+f(s,t)^2/2-\ldots)-1}{f(s,t)} = \frac{f(s,t)}{f(s,t)}(-1 + \underbrace{f(s,t)/2 + \ldots}_{\to 0, t \to s^+})$

1) $f(s,t) := -log(p_{ii}(s,t))$

2) $p_{ii}(s,t) \geq p_{ii}(s,u)p_{ii}(u,t) \Rightarrow f(s,t) \leq f(s,u) + f(u,t)$

$\Rightarrow^{[2]} \sup_{t\to s^+} \frac{f(s,t)}{t-s} = \lim_{t\to s^+} \frac{f(s,t)}{t-s}$

Damit Existenz gezeigt. Nun Eindeutigkeit:

$\lim_{t\to s^+} \frac{p_{ii}(s,t)-1}{t-s} = \lim_{t\to s^+} \frac{e^{-f(s,t)}-1}{t-s} = \lim_{t\to s^+} \frac{e^{-f(s,t)}-1}{f(s,t)} \frac{f(s,t)}{t-s}$

Da $\frac{e^{-f(s,t)}-1}{f(s,t)} = \frac{(1-f(s,t)+f(s,t)^2/2-\dots)-1}{f(s,t)} = \frac{f(s,t)}{f(s,t)}(-1 + \underbrace{f(s,t)/2 + \dots}_{\to 0, t\to s^+})$

$\Rightarrow -\lim_{t\to s^+} \frac{f(s,t)}{t-s} = -\sup_{t\to s^+} \frac{f(s,t)}{t-s} =: q_{ii}(s)$

Korollar 1.9

Außerdem gilt für die Transition Rates:

i) Nicht-Homogen: $q_{ii}(s) \leq 0$ und $q_{ij}(s) \geq 0$

ii) Konservativ: $\sum_{j \in \mathcal{Z}} q_{ij}(s) = 0$

Korollar 1.9

Außerdem gilt für die Transition Rates:

i) Nicht-Homogen: $q_{ii}(s) \leq 0$ und $q_{ij}(s) \geq 0$

ii) Konservativ: $\sum_{j \in \mathcal{Z}} q_{ij}(s) = 0$

Beweis:

i) $p_{ii}(s,t) - 1 \leq 0$ und $p_{ij}(s,t) \geq 0$

Korollar 1.9

Außerdem gilt für die Transition Rates:

i) Nicht-Homogen: $q_{ii}(s) \leq 0$ und $q_{ij}(s) \geq 0$

ii) Konservativ: $\sum_{j \in \mathcal{Z}} q_{ij}(s) = 0$

Beweis:

i) $p_{ii}(s,t) - 1 \leq 0$ und $p_{ij}(s,t) \geq 0$

ii) $p_{ii}(s,t) = 1 - \sum_{i \neq j} p_{ij}(s,t)$

$$\Rightarrow q_{ii}(s) := \lim_{t \to s^+} \frac{p_{ii}(s,t) - 1}{t - s} = \lim_{t \to s^+} \frac{1 - \sum_{i \neq j} p_{ij}(s,t) - 1}{t - s}$$

$$= -\sum_{i \neq j} \lim_{t \to s^+} \frac{p_{ij}(s,t)}{t - s} = -\sum_{i \neq j} q_{ij}(s,t)$$

Satz 1.10 (Umkehrung)

$t \geq 0: Q(t) := (q_{ij}(t))_{i,j \in \mathcal{Z}}$ mit nicht-homogenen, konservativen, messbaren und integrierbaren Einträgen auf $\mathbb{R}$.
Falls für $L_1, L_2 > 0$ und $w \in [1, \infty)^{\mathcal{Z}}$:

$$w^T Q(t) \leq L_1 w^T \text{ und } -diag(Q(t)) \leq L_2 w$$

dann gibt es ein eindeutiges $P(s,t)$ das durch:

- $q_{ij}(t) := \lim_{\Delta \to 0^+} \frac{p_{ij}(t, t+\Delta) - \delta_{ij}}{\Delta}$ bestimmt ist und es gilt

Satz 1.10 (Umkehrung)

$t \geq 0 : Q(t) := (q_{ij}(t))_{i,j \in \mathcal{Z}})$ mit nicht-homogenen, konservativen, messbaren und integrierbaren Einträgen auf $\mathbb{R}$.
Falls für $L_1, L_2 > 0$ und $w \in [1, \infty)^{\mathcal{Z}}$:

$$w^T Q(t) \leq L_1 w^T \text{ und } -diag(Q(t)) \leq L_2 w$$

dann gibt es ein eindeutiges $P(s, t)$ das durch:

- $q_{ij}(t) := \lim_{\Delta \to 0^+} \frac{p_{ij}(t,t+\Delta)-\delta_{ij}}{\Delta}$ bestimmt ist und es gilt
- Kolmogorov's Backward und Forward:
  - $\frac{\partial}{\partial t} P(s, t) = P(s, t) Q(t)$
  - $\frac{\partial}{\partial s} P(s, t) = -Q(s) P(s, t)$

Beispiel Epidemiologie (1/4)

- $\mathcal{Z} := \mathbb{N}_0$
- Population $= N \in \mathbb{N}$

Beispiel Epidemiologie (1/4)

- $\mathcal{Z} := \mathbb{N}_0$
- Population $= N \in \mathbb{N}$

$$q_{ij}(s) = \begin{cases} \lambda_i + \hat{\lambda}_i & j = i+1 \\ -(\lambda_i + \hat{\lambda}_i + \mu_i) & j = i \\ \mu_i & j = i-1 \\ 0 & sonst \end{cases}$$

- λ_i Ansteckungsrate innerhalb der Population
- $\hat{\lambda}_i$ Ansteckungsrate außerhalb der Population
- μ_i Genesungsrate

Beispiel Epidemiologie 2/4 (Überprüfen der Annahmen)

$a_i := \lambda_i + \hat{\lambda}_i$ und $b_i := \mu_i$

$$w^T Q(t) \leq L_1 w^T$$

$$\mathrm{Q(t)} = \begin{pmatrix} -(a_1+b_1) & a_1 & 0 & 0 & \cdots \\ b_2 & -(a_2+b_2) & a_2 & 0 & \cdots \\ 0 & b_3 & -(a_3+b_3) & a_3 & \cdots \\ 0 & 0 & b_4 & -(a_4+b_4) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Beispiel Epidemiologie 2/4 (Überprüfen der Annahmen)

$a_i := \lambda_i + \hat{\lambda}_i$ und $b_i := \mu_i$

$$w^T Q(t) \leq L_1 w^T$$

$$\mathrm{Q(t)} = \begin{pmatrix} -(a_1 + b_1) & a_1 & 0 & 0 & \cdots \\ b_2 & -(a_2 + b_2) & a_2 & 0 & \cdots \\ 0 & b_3 & -(a_3 + b_3) & a_3 & \cdots \\ 0 & 0 & b_4 & -(a_4 + b_4) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

$w = (1, 1, \ldots)^T$
$L_1 := ||w^T Q(t)||_\infty$

Beispiel Epidemiologie 3/4 (Überprüfen der Annahmen)

$$-diag(Q(t)) \leq L_2 w$$

$$\text{-diag(Q(t))} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \dots \end{pmatrix} \leq L_2 \begin{pmatrix} \mathrm{w}_1 \\ \mathrm{w}_2 \\ \dots \end{pmatrix}$$

Beispiel Epidemiologie 3/4 (Überprüfen der Annahmen)

$$-diag(Q(t)) \leq L_2 w$$

$$\text{-diag(Q(t))} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \dots \end{pmatrix} \leq L_2 \begin{pmatrix} \mathrm{w}_1 \\ \mathrm{w}_2 \\ \dots \end{pmatrix}$$

Und $L_2 = ||diag(Q(t))||_\infty$

Definition 2.1

- $(\mathcal{A}, \sigma(\mathcal{A}))$ Action Space
- $\mathcal{A}(i) \in \sigma(\mathcal{A})$ mögliche Actions in i $\in \mathcal{Z}$

Definition 2.1

- $(\mathcal{A}, \sigma(\mathcal{A}))$ Action Space
- $\mathcal{A}(i) \in \sigma(\mathcal{A})$ mögliche Actions in i $\in \mathcal{Z}$

Definition 2.2 (Markov Kern)

Für jedes i ist $\pi_\cdot(\cdot, i)$ Markovkern:

- $\pi_\cdot(\cdot, i) : [0, \infty) \times \sigma(\mathcal{A}(i)) \to [0, 1]$

Definition 2.1

- $(\mathcal{A}, \sigma(\mathcal{A}))$ Action Space
- $\mathcal{A}(i) \in \sigma(\mathcal{A})$ mögliche Actions in i $\in \mathcal{Z}$

Definition 2.2 (Markov Kern)

Für jedes i ist $\pi.(\cdot, i)$ Markovkern:

- $\pi.(\cdot, i) : [0, \infty) \times \sigma(\mathcal{A}(i)) \to [0, 1]$
  1. $\pi.(A, i)$ ist $\mathcal{B}([0, \infty))$ - $\mathcal{B}([0, 1])$ messbar $\forall\ A \in \sigma(\mathcal{A}(i)), i \in \mathcal{Z}$

Definition 2.1

- $(\mathcal{A}, \sigma(\mathcal{A}))$ Action Space
- $\mathcal{A}(i) \in \sigma(\mathcal{A})$ mögliche Actions in i $\in \mathcal{Z}$

Definition 2.2 (Markov Kern)

Für jedes i ist $\pi_{\cdot}(\cdot, i)$ Markovkern:

- $\pi_{\cdot}(\cdot, i) : [0, \infty) \times \sigma(\mathcal{A}(i)) \to [0, 1]$
  1. $\pi_{\cdot}(A, i)$ ist $\mathcal{B}([0, \infty))$ - $\mathcal{B}([0, 1])$ messbar $\forall\ A \in \sigma(\mathcal{A}(i)), i \in \mathcal{Z}$
  2. $\pi_t(\cdot, i)$ ist W-Maß auf $\sigma(\mathcal{A}(i))\ \forall\ t \in [0, \infty)$ und $i \in \mathcal{Z}$

Definition 2.1

- $(\mathcal{A}, \sigma(\mathcal{A}))$ Action Space
- $\mathcal{A}(i) \in \sigma(\mathcal{A})$ mögliche Actions in i $\in \mathcal{Z}$

Definition 2.2 (Markov Kern)

Für jedes i ist $\pi_\cdot(\cdot, i)$ Markovkern:

- $\pi_\cdot(\cdot, i) : [0, \infty) \times \sigma(\mathcal{A}(i)) \to [0, 1]$
  1. $\pi_\cdot(A, i)$ ist $\mathcal{B}([0, \infty))$ - $\mathcal{B}([0, 1])$ messbar $\forall\ A \in \sigma(\mathcal{A}(i)), i \in \mathcal{Z}$
  2. $\pi_t(\cdot, i)$ ist W-Maß auf $\sigma(\mathcal{A}(i))$ $\forall\ t \in [0, \infty)$ und $i \in \mathcal{Z}$

Policy: $\pi := (\pi_t)_{t \geq 0} \in \Pi$

Definition 2.3

- Gegebene Transition Rates: $q(j \mid i, a)$ messbar und integrierbar in $a \in \mathcal{A}(i)$
- Durchschnittliche Transition Rate von i nach j der Policy π:

$$q_{ij}^{\pi}(t) := \int_{\mathcal{A}(i)} q(j \mid i, a) \; d\pi_t(a, i)$$

Definition 2.3

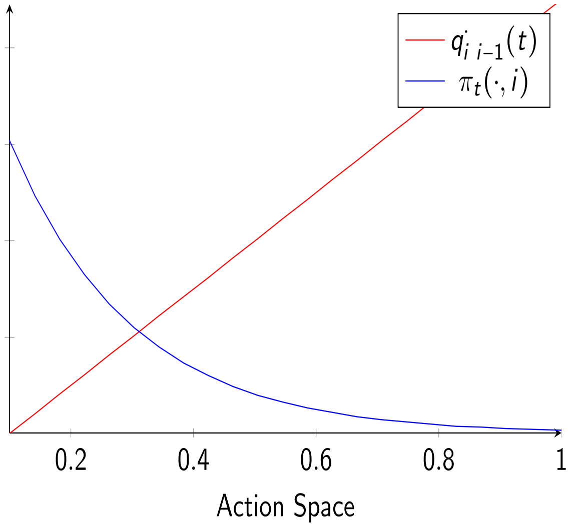
- Gegebene Transition Rates: $q(j \mid i, a)$ messbar und integrierbar in $a \in \mathcal{A}(i)$
- Durchschnittliche Transition Rate von i nach j der Policy π:

$$q_{ij}^{\pi}(t) := \int_{\mathcal{A}(i)} q(j \mid i, a) \; d\pi_t(a, i)$$

Definition 2.4

- Kosten: $c(i \mid a) \geq 0$ messbar und integrierbar
- Durchschnittliche Kosten der Policy π im Zustand i:

$$c_i^{\pi}(t) := \int_{\mathcal{A}(i)} c(i \mid a) \; d\pi_t(a, i)$$


$q_{i\,i-1}^{\cdot}(t)$
$\pi_t(\cdot, i)$
0.2
0.4
0.6
0.8
1
Action Space

Beispiel Epidemiologie (4/4)

- $\mathrm{A} := [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] = A(i)\ \forall i \in \mathcal{Z}$
- $\mathrm{a} = (a_1, a_2) \in A$

Beispiel Epidemiologie (4/4)

- A $:= [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] = A(i) \ \forall i \in \mathcal{Z}$
- a $= (a_1, a_2) \in A$

$\rightarrow a_1 \in [\underline{a}, \overline{a}]$ Level der Quarantäne $\Rightarrow \hat{\lambda}_i := \hat{\lambda}_i(a_1)$

Beispiel Epidemiologie (4/4)

- A $:= [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] = A(i) \ \forall i \in \mathcal{Z}$
- a $= (a_1, a_2) \in A$

$\rightarrow a_1 \in [\underline{a}, \overline{a}]$ Level der Quarantäne $\Rightarrow \hat{\lambda}_i := \hat{\lambda}_i(a_1)$
$\rightarrow a_2 \in [\underline{b}, \overline{b}]$ Level der medizinischen Behandlung $\Rightarrow \lambda_i := \lambda_i(a_2)$

Beispiel Epidemiologie (4/4)

- A $:= [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] = A(i) \; \forall i \in \mathcal{Z}$
- a $= (a_1, a_2) \in A$

$\rightarrow a_1 \in [\underline{a}, \overline{a}]$ Level der Quarantäne $\Rightarrow \hat{\lambda}_i := \hat{\lambda}_i(a_1)$
$\rightarrow a_2 \in [\underline{b}, \overline{b}]$ Level der medizinischen Behandlung $\Rightarrow \lambda_i := \lambda_i(a_2)$

$$c(i \mid a) := h_0(i) + h_1(a_1) + h_2(i, a_2); \;\; \forall i \in \mathcal{Z}, a = (a_1, a_2) \in A$$

Definition 3.1

Wert der Policy mit Start in i (x(0) = i) und $\alpha > 0$:

$$J_i^\pi := \mathbb{E}_\pi^i \left[\int_0^\infty e^{-\alpha t} \; c_{x(t)}^\pi(t) \; dt \right]$$
$$= \int_0^\infty e^{-\alpha t} \; \mathbb{E}_\pi^i \left[c_{x(t)}^\pi(t) \right] \; dt$$

- $\mathbb{E}_\pi^i [\; c_{x(t)}^\pi(t) \;] = \sum_{j \in \mathcal{Z}} c_j^\pi(t) p_{ij}(0,t)$

Definition 3.2

π^* heißt optimale Policy, falls

$$\pi^* \in argmin_{\pi \in \Pi} J_i^{\pi}$$

Definition 4.1

Sei f: $\mathcal{Z} \to \mathcal{A}$, i $\mapsto f(i) \in \mathcal{A}(i)$
Neue stationäre deterministische Policy:

$$\pi(A, i) = \delta_{f(i)}(A),\ A \in \sigma(\mathcal{A}(i))$$

Definition 4.1

Sei f: $\mathcal{Z} \to \mathcal{A}$, i $\mapsto f(i) \in \mathcal{A}(i)$
Neue stationäre deterministische Policy:

$$\pi(A,i) = \delta_{f(i)}(A),\ A \in \sigma(\mathcal{A}(i))$$

Definition 4.2

$\forall a \sim \pi(\cdot, i):$

$$q_{ij}^{\pi}(t) = \int_{\mathcal{A}(i)} q(j \mid i, a) d\pi(a,i) =: q_{ij}^{f(i)}$$

Definition 4.1

Sei f: $\mathcal{Z} \to \mathcal{A}$, i $\mapsto f(i) \in \mathcal{A}(i)$

Neue stationäre deterministische Policy:

$$\pi(A,i) = \delta_{f(i)}(A),\ A \in \sigma(\mathcal{A}(i))$$

Definition 4.2

$\forall a \sim \pi(\cdot, i):$

$$q_{ij}^{\pi}(t) = \int_{\mathcal{A}(i)} q(j \mid i,a) d\pi(a,i) =: q_{ij}^{f(i)}$$

$$c_{i}^{\pi}(t) = \int_{\mathcal{A}(i)} c(i \mid a) d\pi_t(a,i) =: c_{i}^{f(i)}$$

Definition 4.1

Sei f: $\mathcal{Z} \to \mathcal{A}$, i $\mapsto f(i) \in \mathcal{A}(i)$
Neue stationäre deterministische Policy:

$$\pi(A,i) = \delta_{f(i)}(A),\ A \in \sigma(\mathcal{A}(i))$$

Definition 4.2

$\forall a \sim \pi(\cdot, i):$

$$q_{ij}^{\pi}(t) = \int_{\mathcal{A}(i)} q(j \mid i,a) d\pi(a,i) =: q_{ij}^{f(i)}$$

$$c_{i}^{\pi}(t) = \int_{\mathcal{A}(i)} c(i \mid a) d\pi_t(a,i) =: c_{i}^{f(i)}$$

Transition Matrix P(s,t) existiert und bestimmt x(t) auf $\mathcal{Z}$

Satz 4.3

$$J_i^f = \int_0^\infty e^{-\alpha t}\mathbb{E}_\pi^i\left[c_{x(t)}^{f(i)}\right]dt = \int_0^\infty e^{-\alpha t}\sum_{j\in\mathcal{Z}} c_j^{f(i)} p_{ij}(0,t)dt$$

$$= \frac{1}{\alpha - q_{ii}^{f(i)}}(c_i^{f(i)} + \sum_{i\neq k\in\mathcal{Z}} J_k^f q_{ik}^{f(i)})$$

Satz 4.3

$$J_i^f = \int_0^\infty e^{-\alpha t}\mathbb{E}_\pi^i\left[c_{x(t)}^{f(i)}\right]dt = \int_0^\infty e^{-\alpha t}\sum_{j\in\mathcal{Z}} c_j^{f(i)} p_{ij}(0,t)dt$$

$$= \frac{1}{\alpha - q_{ii}^{f(i)}}(c_i^{f(i)} + \sum_{i\neq k\in\mathcal{Z}} J_k^f q_{ik}^{f(i)})$$

$$<=> J_i^f = \frac{1}{\alpha - q_{ii}^{f(i)}}(\overbrace{\textstyle\sum_{i\neq k\in\mathcal{Z}} J_k^f q_{ik}^{f(i)} + J_i^f q_{ii}^{f(i)}}^{=\sum_{k\in\mathcal{Z}} J_k^f q_{ik}^{f(i)}} - J_i^f q_{ii}^{f(i)})$$

Satz 4.3

$$J_i^f = \int_0^\infty e^{-\alpha t} \mathbb{E}_\pi^i \left[c_{x(t)}^{f(i)} \right] dt = \int_0^\infty e^{-\alpha t} \sum_{j \in \mathcal{Z}} c_j^{f(i)} p_{ij}(0,t) dt$$

$$= \frac{1}{\alpha - q_{ii}^{f(i)}} \left(c_i^{f(i)} + \sum_{i \neq k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)} \right)$$

$$<=> J_i^f = \tfrac{1}{\alpha - q_{ii}^{f(i)}} \Big(\overbrace{\textstyle\sum_{i \neq k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)} + J_i^f q_{ii}^{f(i)}}^{=\sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}} - J_i^f q_{ii}^{f(i)} \Big)$$

$$<=> (\alpha - q_{ii}^{f(i)}) J_i^f = c_i^{f(i)} + \textstyle\sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)} - J_i^f q_{ii}^{f(i)}$$

$$\alpha J_i^f = c_i^{f(i)} + \sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}$$

$$\alpha J_i^f = c_i^{f(i)} + \sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}$$

Satz 4.4 (Policy Improvement)

- $D_f(i, a) := c_i^a + \sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}$ für $i \in \mathcal{Z}$

$$\alpha J_i^f = c_i^{f(i)} + \sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}$$

Satz 4.4 (Policy Improvement)

- $D_f(i,a) := c_i^a + \sum_{k \in \mathcal{Z}} J_k^f q_{ik}^{f(i)}$ für $i \in \mathcal{Z}$
- f' Policy Improvement zu f, falls

$$\rightarrow \; f'(i) := \begin{cases} f(i) & , D_f(i,a) \geq \alpha J_i^f \;\; \forall a \in A(i) \\ a' & ,\text{irgend ein } a' \in A(i) \text{ mit } D_f(i,a') < \alpha J_i^f \end{cases}$$

$$\Rightarrow \; J_i^f \geq J_i^{f'}$$

<u>Beweis:</u> Direkt per Konstruktion

Policy Iteration Algorithmus

1. Wähle k=0 und zufällige/beliebige Policy $f^{(k=0)}$ und Genauigkeitsmaß $\epsilon > 0$

Policy Iteration Algorithmus

1. Wähle k=0 und zufällige/beliebige Policy $f^{(k=0)}$ und Genauigkeitsmaß $\epsilon > 0$

2. Iteriere:
   i) Wert $J_i^{f^{(k)}}$ ermitteln

   ii) Policy Improvement: $f^{(k+1)} = f'^{(k)}$

   iii) Falls $\mid J_i^{f^{(k+1)}} - J_i^{f^{(k)}} \mid < \epsilon \; \forall i \in \mathcal{Z} \Rightarrow$ fertig.
   Sonst zurück zu i) und setze $k = k + 1$

Policy Iteration Algorithmus

1. Wähle k=0 und zufällige/beliebige Policy $f^{(k=0)}$ und Genauigkeitsmaß $\epsilon > 0$
2. Iteriere:
   i) Wert $J_i^{f^{(k)}}$ ermitteln
   ii) Policy Improvement: $f^{(k+1)} = f'^{(k)}$
   iii) Falls $| J_i^{f^{(k+1)}} - J_i^{f^{(k)}} | < \epsilon \ \forall i \in \mathcal{Z} \Rightarrow$ fertig.
   Sonst zurück zu i) und setze $k = k + 1$

Satz 4.5

Dieser Algorithmus konviergiert zu einem lokalen Minimum

Literatur

[1] Guo, X.; Hern andez-Lerma, O. (2009) Continuous-Time Markov Decision Processes. Stochastic Modelling and Applied Probability, Springer

[2] Liuer Ye und Xianping Guo und Onésimo Hernández-Lerma, (2008) Existence and Regularity of a Nonhomogeneous Transition Matrix under Measurability Conditions, J Theor Probab

[3] Kolmogoroff, A. (1930) Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Math. Ann.

[4] Howard, Ronald A. (1960) Dynamic Programming and Markov Processes, The M.I.T. Press

[5] Marek Kuczma, (2009) An Introduction to the Theory of Functional Equations and Inequalities, Second Edition