Ausarbeitung zum Vortrag: Reinforcement Learning

0 (Diskreter) bedingter Erwartungswert

Vorab eine Erinnerung an den diskreten bedingten Erwartungswert.

Definition 0.1 (Bedingter Erwartungswert im diskreten Fall)**.** Sei Y integrierbar und X diskret mit $\mathbb{P}(X = x) > 0$. Dann ist

$$\mathbb{E}[Y \mid X = x] := \sum_y y\, \mathbb{P}(Y = y \mid X = x),$$

wobei die Summe über alle Werte y von Y läuft.

Satz 0.1 (Turmregel)**.** *Seien X, Z diskret und Y integrierbar. Dann gilt*

$$\mathbb{E}\big[\mathbb{E}[Y \mid X, Z] \mid X\big] = \mathbb{E}[Y \mid X] \quad \textit{und} \quad \mathbb{E}\big[\mathbb{E}[Y \mid X]\big] = \mathbb{E}[Y].$$

Satz 0.2 (Gesetz der totalen Erwartung)**.** *Sei X diskret und Y integrierbar. Dann gilt*

$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y \mid X = x]\, \mathbb{P}(X = x).$$

(allgemeiner): Für Ereignis $\{Z = z\}$ mit $\mathbb{P}(Z = z) > 0$,

$$\mathbb{E}[Y \mid Z = z] = \sum_x \mathbb{E}[Y \mid X = x, Z = z]\, \mathbb{P}(X = x \mid Z = z).$$

Bemerkung 0.1. In den folgenden Bellman-Beweisen werden explizit die Summenform des Gesetzes der totalen Erwartung angewandt, da diese mehr Intuition ueber das Geschehen gibt.

1 Allgemeine Definitionen und Konstruktion der Markov-Kette (MDPs)

1.1

Wir betrachten diskrete Zeit $t \in \{0, 1, 2, \dots\}$. Sei $T \in \mathbb{N} \cup \{\infty\}$ der terminale Zeitschritt; im finite Horizon Fall ist $T < \infty$, im im infinite Horizon Fall $T = \infty$. Eine *Trajektorie* ist die Folge

$$S_0, A_0, R_0, S_1, A_1, R_1, \dots, S_T \quad (\text{ bei } T = \infty \text{ ist die Folge unendlich}).$$

Dabei sind S_t Zustände, A_t Aktionen und R_t Rewards.

1.2 Return und discounted Return

Definition 1.1 (Discounted Return)**.** Für $\gamma \in [0,1)$ definieren wir den *discounted return* ab Zeit t als

$$G_t := \sum_{k=0}^{\infty} \gamma^k R_{t+k}.$$

Bemerkung 1.1. Angenommen, der Reward ist beschränkt: $|R_t| \leq R_{\max}$ für alle t und ein $R_{\max} < \infty$. Dann konvergiert die Reihe für G_t absolut und es gilt die Schranke.

$$|G_t| \leq \sum_{k=0}^{\infty} \gamma^k |R_{t+k}| \leq R_{\max} \sum_{k=0}^{\infty} \gamma^k = \frac{R_{\max}}{1-\gamma} < \infty.$$

Das motiviert die Definition vom discounted Return und Wahl von γ.

1.3 Policy

Definition 1.2. Eine (stationäre) Policy ist eine Abbildung $\pi : \mathcal{A} \times \mathcal{S} \to [0,1]$ mit

$$\pi(a \mid s) = \mathbb{P}(A_t = a \mid S_t = s), \qquad \sum_{a \in \mathcal{A}} \pi(a \mid s) = 1 \quad \forall s \in \mathcal{S}.$$

Man bezeichne eine Policy als deterministisch, wenn $\pi(a \mid s) \in \{0,1\}$ $\forall (a,s) \in \mathcal{A} \times \mathcal{S}$.

Bemerkung 1.2. Ist π deterministisch, so gilt für jedes $s \in \mathcal{S}$, dass es ein $a \in \mathcal{A}$ mit $\pi(a \mid s) = 1$ gibt. Also definiert man:

$$\pi(s) := a \quad \text{falls} \quad \pi(a \mid s) = 1.$$

1.4 Rekursive Konstruktion des MDP

Wir betrachten endliche Mengen $\mathcal{S}, \mathcal{A}$ und einen vereinfachten Reward, der nur von (s,a) abhängt und deterministisch ist.

Definition 1.3. Ein MDP ist ein Tupel $(\mathcal{S}, \mathcal{A}, p, r, \nu, \gamma)$ aus

- endlichem Zustandsraum $\mathcal{S}$,
- endlichem Aktionsraum $\mathcal{A}$,
- state transition kernel $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, geschrieben $p(s' \mid s,a)$, wobei

  $$p(s' \mid s,a) := \mathbb{P}(S_{t+1} = s' \mid S_t = s,\ A_t = a) \quad (\forall t \geq 0),$$

  und für alle $s \in \mathcal{S}, a \in \mathcal{A}$ gilt $p(s' \mid s,a) \geq 0$ sowie

  $$\sum_{s' \in \mathcal{S}} p(s' \mid s,a) = 1,$$

- Reward-Funktion $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|r\|_\infty := \max_{s,a} |r(s,a)| < \infty$,
- Initialverteilung ν auf $\mathcal{S}$,
- Discount-Faktor $\gamma \in [0,1)$.

Definition 1.4 (Rekursive Konstruktion)**.** Gegeben sei eine Policy π und ν. Wir definieren dann $(S_t, A_t, R_t)_{t \geq 0}$ rekursiv:

1. $S_0 \sim \nu$.

2. Für $t \geq 0$: Ziehe $A_t \sim \pi(\cdot \mid S_t)$.

3. Ziehe danach $S_{t+1} \sim p(\cdot \mid S_t, A_t)$.

4. Setze $R_t := r(S_t, A_t)$.

Proposition 1.1 (Markoveigenschaft)**.** *Sei* $(S_0, A_0, S_1, A_1, \ldots, S_t)$*.Es gilt für alle* $t \geq 0$*: (i)*

$$\mathbb{P}(A_t = a \mid (S_0, A_0, S_1, A_1, \ldots, S_t),\, S_t = s) = \mathbb{P}(A_t = a \mid S_t = s) = \pi(a \mid s).$$

(ii)

$$\mathbb{P}(S_{t+1} = s' \mid (S_0, A_0, S_1, A_1, \ldots, S_t),\, S_t = s,\, A_t = a) = \mathbb{P}(S_{t+1} = s' \mid S_t = s,\, A_t = a) = p(s' \mid s, a).$$

Insbesondere folgt für die gemeinsame Verteilung von A_t *und* S_t*:*

$$\mathbb{P}(A_t = a,\, S_{t+1} = s' \mid (S_0, A_0, S_1, A_1, \ldots, S_t),\, S_t = s) = \pi(a \mid s)\, p(s' \mid s, a).$$

Beweis. Das folgt direkt aus der rekursiven Konstruktion. □

2 Value Function und Q-Function (action state value Function)

Definition 2.1 (Value Function)**.** Für eine Policy π definieren wir

$$v^\pi(s) := \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \;\middle|\; S_t = s\right].$$

Durch die Markov-Eigenschaft ist v^π unabhängig von t.

Definition 2.2 (Q-Function)**.**

$$q^\pi(s, a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a].$$

Lemma 2.1. $\forall\ s \in \mathcal{S}$ *gilt*

$$v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q^\pi(s, a).$$

Beweis. Nach dem Gesetz der totalen Erwartung gilt:

$$v^\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_{a \in \mathcal{A}} \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]\, \mathbb{P}(A_t = a \mid S_t = s) = \sum_a q^\pi(s, a)\, \pi(a \mid s).$$

□

3 Bellman-Konsistenzgleichung

3.1 Bellman-Konsistenzgleichung für v^π

Satz 3.1 (Bellman-Konsistenzgleichung)**.** *Für alle* $s \in \mathcal{S}$ *gilt*

$$v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\Big(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a)\, v^\pi(s')\Big).$$

Beweis. Aus der Definition des discounted return gilt punktweise die Rekursion

$$G_t = R_t + \gamma G_{t+1}.$$

Wende das Gesetz der totalen Erwartung in Summenform auf (A_t, S_{t+1}) an:

$$v^\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a, S_{t+1} = s'] \, \mathbb{P}(A_t = a, S_{t+1} = s' \mid S_t = s).$$

$R_t = r(S_t, A_t) = r(s, a)$ ist deterministisch, also gilt:

$$\mathbb{E}_\pi[G_t \mid S_t = s, A_t = a, S_{t+1} = s'] = r(s, a) + \gamma \, \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'].$$

Nach der Markov-Eigenschaft gilt:

$$\mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'] = \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s'] = v^\pi(s').$$

und

$$\mathbb{P}(A_t = a, S_{t+1} = s' \mid S_t = s) = \mathbb{P}(A_t = a \mid S_t = s) \, \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) = \pi(a \mid s) \, p(s' \mid s, a).$$

Zusammen gibt das

$$v^\pi(s) = \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \big(r(s, a) + \gamma v^\pi(s')\big) \, \pi(a \mid s) \, p(s' \mid s, a).$$

Da $r(s, a)$ unabhaengig von s' sind und $\sum_{s'} p(s' \mid s, a) = 1$, gilt

$$v^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \Big(r(s, a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s, a) \, v^\pi(s')\Big).$$

□

Bemerkung 3.1. Analog kann man für alle $s \in \mathcal{S}, a \in \mathcal{A}$ die Bellman-Gleichung

$$q^\pi(s, a) = r(s, a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s, a) \, v^\pi(s').$$

beweisen.

4 Optimalität und Bellman-Optimalitätsgleichungen

Definition 4.1 (Optimale Value Functions)**.** Die optimale Value Function ist

$$v^*(s) := \sup_\pi v^\pi(s) \qquad (s \in \mathcal{S}),$$

und die optimale Q-Function ist

$$q^*(s, a) := \sup_\pi q^\pi(s, a) \qquad (s \in \mathcal{S}, \ a \in \mathcal{A}).$$

Bemerkung 4.1. In endlichen MDPs wird das Supremum als Maximum erreicht.

4.1 Halbordnung auf Policies

Definition 4.2. Für zwei Policies π, π' definieren wir eine Halbordnung:

$$\pi \preceq \pi' \quad :\Longleftrightarrow \quad v^{\pi}(s) \leq v^{\pi'}(s) \;\; \forall s \in \mathcal{S}.$$

Bemerkung 4.2. Somit existiert ein maximales Element (Policy), da der Aktionsraum und Zustandsraum endlich sind, aber kein Maximum. Die "maximale"Policy muss also nicht eindeutig sein.

Satz 4.1 (Bellman-Optimalitätsgleichungen)**.** *Für alle $s \in \mathcal{S}$ gilt*

$$v^*(s) = \max_{a \in \mathcal{A}} q^*(s,a),$$

und für alle $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\, v^*(s').$$

Da $R_t = r(S_t, A_t)$ laesst sich die letztere Gleichung auch wie folgt schreiben:

$$q^*(s,a) = \mathbb{E}\big[R_t + \gamma v^*(S_{t+1}) \;\big|\; S_t = s,\; A_t = a\big],$$

Beweis. Gleichung 2: Aus der Bemerkung zu q^π gilt für jede Policy π:

$$q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\, v^{\pi}(s').$$

Nach Maximieren über π gilt:

$$q^*(s,a) = \sup_{\pi} q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) \sup_{\pi} v^{\pi}(s') = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a)\, v^*(s').$$

Gleichung 1: Für jede Policy π gilt

$$v^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, q^{\pi}(s,a) \leq \max_{a \in \mathcal{A}} q^{\pi}(s,a) \leq \max_{a \in \mathcal{A}} q^*(s,a).$$

Somit gilt auch: $v^*(s) \leq \max_a q^*(s,a)$. Für die Umkehrung beachte, dass in endlichen MDP eine (nicht unbedingt eindeutige) optimale Policy π^* existiert und $v^*(s) = v^{\pi^*}(s)$. Somit gilt:

$$v^*(s) = v^{\pi^*}(s) = \sum_{a \in \mathcal{A}} \pi^*(a \mid s)\, q^{\pi^*}(s,a) \leq \max_{a \in \mathcal{A}} q^{\pi^*}(s,a) \leq \max_{a \in \mathcal{A}} q^*(s,a).$$

□

5 Policy Evaluation als lineares Gleichungssystem

Sei $\mathcal{S} = \{s_1, \ldots, s_n\}$, $n = |\mathcal{S}|$. Definiere:

$$v_\pi \in \mathbb{R}^n, \quad (v_\pi)_i = v^{\pi}(s_i), \qquad (R_\pi)_i = \sum_{a \in \mathcal{A}} \pi(a \mid s_i)\, r(s_i, a),$$

$$(P_\pi)_{ij} = \sum_{a \in \mathcal{A}} \pi(a \mid s_i)\, p(s_j \mid s_i, a).$$

Dann ist die Bellman-Konsistenzgleichung äquivalent zu dem LGS:

$$v_\pi = R_\pi + \gamma P_\pi v_\pi.$$

Bemerkung 5.1. Die direkte Lösung des LGS laeuft in $\mathcal{O}(|S|^3)$. Deswegen kann ein iterativer Ansatz, wie in Abschnitt 6 beschrieben, sinnvoller sein.

Lemma 5.1. *P_π ist stochastisch*

Beweis. Jeder Eintrag ist nichtnegativ, da $\pi(a \mid s) \geq 0$ und $p(\cdot \mid s, a) \geq 0$. Dazu gilt:

$$\sum_{j=1}^{n}(P_\pi)_{ij} = \sum_{j=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid s_i)p(s_j \mid s_i, a) = \sum_{a\in\mathcal{A}} \pi(a \mid s_i)\sum_{j=1}^{n} p(s_j \mid s_i, a) = \sum_{a\in\mathcal{A}} \pi(a \mid s_i)\cdot 1 = 1.$$

□

6 Konvergenz der iterativen Bellman-Gleichung

Lemma 6.1. *Wir definieren die (induzierte) Operatornorm*

$$\|M\|_\infty := \sup_{\|x\|_\infty=1} \|Mx\|_\infty.$$

Für (zeilen)stochastisches P_π gilt $\|P_\pi\|_\infty \leq 1$.

Beweis. Sei $x \in \mathbb{R}^n$. Für jede Zeile i gilt wegen $(P_\pi)_{ij} \geq 0$:

$$|(P_\pi x)_i| = \left|\sum_{j=1}^{n}(P_\pi)_{ij}x_j\right| \leq \sum_{j=1}^{n}(P_\pi)_{ij}|x_j| \leq \left(\sum_{j=1}^{n}(P_\pi)_{ij}\right)\|x\|_\infty = 1\cdot\|x\|_\infty.$$

Also $\|P_\pi x\|_\infty \leq \|x\|_\infty$ für alle x, und damit $\|P_\pi\|_\infty \leq 1$. □

Satz 6.1. *Definiere $T_\pi : \mathbb{R}^n \to \mathbb{R}^n$ durch*

$$T_\pi(v) := R_\pi + \gamma P_\pi v.$$

Dann ist T_π eine Kontraktion und besitzt genau einen Fixpunkt v_π. Die Iteration

$$v^{(k+1)} = T_\pi(v^{(k)})$$

konvergiert gegen v_π und es gilt die Fehlerabschätzung

$$\left\|v^{(k)} - v_\pi\right\|_\infty \leq \gamma^k \left\|v^{(0)} - v_\pi\right\|_\infty.$$

Beweis. Für $v, w \in \mathbb{R}^n$ gilt

$$\|T_\pi(v) - T_\pi(w)\|_\infty = \|\gamma P_\pi(v-w)\|_\infty \leq \gamma\|P_\pi\|_\infty \|v-w\|_\infty \leq \gamma\|v-w\|_\infty,$$

weil $\|P_\pi\|_\infty \leq 1$. Also ist T_π eine Kontraktion. Da $(\mathbb{R}^n, \|\cdot\|_\infty)$ vollständig ist, folgt aus dem Banachschen Fixpunktsatz die Eindeutigkeit des Fixpunktes v_π und die Konvergenz der Iteration. Für den Fehler gilt (Picard-Iteration):

$$\left\|v^{(k)} - v_\pi\right\|_\infty = \left\|T_\pi^k(v^{(0)}) - T_\pi^k(v_\pi)\right\|_\infty \leq \gamma^k \left\|v^{(0)} - v_\pi\right\|_\infty.$$

Bemerkung 6.1. Ein kleines γ sorgt fuer schnellere Konvergenz, jedoch sinkt dadurch die Ausdruckskraft des Return.

□

7 Policy Improvement und Policy Iteration

Satz 7.1 (Policy Improvement Theorem)**.** *Seien π und π' deterministische Policies. Gilt*

$$q^\pi(s, \pi'(s)) \geq v^\pi(s) \quad \forall s \in \mathcal{S},$$

dann folgt

$$v^{\pi'}(s) \geq v^\pi(s) \quad \forall s \in \mathcal{S}.$$

Beweis. Sei $s \in \mathcal{S}$. Dann gilt

$$v^\pi(s) \leq q^\pi(s, \pi'(s)).$$

Nach Definition von q^π und $G_t = R_t + \gamma G_{t+1}$ gilt

$$q^\pi(s, \pi'(s)) = \mathbb{E}\big[R_t + \gamma G_{t+1} \mid S_t = s,\ A_t = \pi'(s)\big].$$

Da $R_t = r(S_t, A_t) = r(s, \pi'(s))$ deterministisch und nach der Markov-Eigenschaft gilt

$$q^\pi(s, \pi'(s)) = \mathbb{E}\big[R_t + \gamma v^\pi(S_{t+1}) \mid S_t = s,\ A_t = \pi'(s)\big] = \mathbb{E}_{\pi'}\big[R_t + \gamma v^\pi(S_{t+1}) \mid S_t = s\big],$$

wobei man hier nutzt, dass π' deterministisch ist. Nun setzen wir die Voraussetzung ein:

$$v^\pi(s) \leq \mathbb{E}_{\pi'}\big[R_t + \gamma q^\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\big].$$

Beachte, dass $q^\pi(S_{t+1}, \pi'(S_{t+1})) = \mathbb{E}_\pi[G_{t+1} \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})]$ setze $G_{t+1} = R_{t+1} + \gamma G_{t+2}$. Ersetzte den Term $v^\pi(S_{t+2})$ wieder durch $q^\pi(S_{t+2}, \pi'(S_{t+2}))$. Nach sukszessivem Einsetzen folgt:

$$v^\pi(s) \leq \mathbb{E}_{\pi'}\Big[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \;\Big|\; S_t = s\Big] = v^{\pi'}(s).$$

□

Algorithm 1 Policy Iteration

```
Input: Modell p, r, γ
Initialisiere (deterministische) Policy π
repeat
    Berechne v^π mit der iterativen Bellman-Gleichung
    fixpunktErreicht ← true
    for all s ∈ S do
        a_alt ← π(s)
        π(s) ← arg max_a (r(s, a) + γ Σ_{s'} p(s'|s, a) v^π(s'))
        if π(s) ≠ a_alt then
            fixpunktErreicht ← false
        end if
    end for
until fixpunktErreicht
```

Bemerkung 7.1. Wir betrachten einen endlichen Zustandsraum und Aktionsraum, also terminiert der Algorithmus nach endlich vielen Schritten. Wenn der Algorithmus terminiert,dann gilt $\pi^{(k+1)} = \pi^{(k)}$. Dann gilt auch $v_{\pi^{(k+1)}} = v_{\pi^{(k)}}$. Dann gilt also $v_{\pi^{(k)}}(s) = \max_a q_{\pi^{(k)}}(a, s) \quad \forall s \in \mathcal{S}$, was genau der Bellman-Optimalitätsgleichung entspricht,somit muss $v_{\pi^{(k)}} = v_*$ gelten, insbesondere ist dann Policy $\pi^{(k)}$*optimal*.

8 Value Iteration

Algorithm 2 Value Iteration

Input: Modell p, r, γ
Initialisiere $v(s) \leftarrow 0 \; \forall s \in \mathcal{S}$
repeat
    Fehler $\leftarrow 0$
    for all $s \in \mathcal{S}$ **do**
        $v_{\text{alt}} \leftarrow v(s)$
        $v(s) \leftarrow \max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')\right)$
        Fehler $\leftarrow \max(\text{Fehler}, |v_{\text{alt}} - v(s)|)$
    end for
until Fehler ≈ 0
Leite aus $v(s)$ Policy her: $\pi(s) \leftarrow \arg\max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')\right)$
Output: $v \approx v^*$, Policy π

Bemerkung 8.1. Hier approximieren wir v^* direkt und leiten dann aus v^* die Policy her, was potentiell eine Faktor $\mathcal{O}\|S|)$ bessere Laufzeit haben wird. Der Beweis fuer die Konvergenz des Algorithmus ist analog zum Beweis der Konvergenz von der iterativen Bellman-Gleichung. Da $T(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')\right)$ eine Kontraktion ist etc.

9 Model-based und Model-free Reinforcement Learning

Bemerkung 9.1. In diesen Algorithmen haben wir die Dynamik des Environments (also den state transition kernel p und reward Function r) als bekannt angenommen. Wenn das nicht der Fall ist, dann approximiert man die Unbekannten mit Monte-Carlo Methoden, was aber nur langsame Konvergenz liefert ($\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$).

10 Literatur

1. The Mathematics of Machine Learning: Lectures on Supervised Methods and Beyond. Maria Han Veiga and François Gaston Ged. De Gruyter, 2024.

2. Reinforcement Learning: An Introduction. Second Edition. Richard S. Sutton and Andrew G. Barto. MIT Press, 2018.