Exkurs:

Reinforcement Learning

Dr. Lars Metzger
Fakultät Wirtschaftswissenschaften
TU Dortmund

Reinforcement Learning

Wie kann ein Agent in einem „Setting“ aufeinander folgende Entscheidungen treffen, um seine aufsummierten Auszahlungen zu maximieren?

Das Reinforcement Learning ist Teil des Machine Learning. Es wird als Trainingsmethode für KI benutzt.

Die Spieltheorie erlaubt dynamische Settings, in welchem der Agent Entscheidungen zu treffen hat: Das Setting wird von dem Agenten beeinflusst.

Exploration: Entscheidung unabhängig von den Auszahlungen
$\rightarrow$ Erschließung neuer Möglichkeiten

Exploitation: Entscheidung abhängig von den Auszahlungen
$\rightarrow$ Wählen besserer Möglichkeiten

„A simple adaptive procedure leading to correlated equilibrium“

von Sergiu Hart & Andreu Mas-Colell
erschienen in Econometrica, Vol. 68, No. 5 (2000)

Das Korrelierte Gleichgewicht wird später in Kapitel 3 behandelt.

Wir fokussieren hier auf die „einfache adaptive Prozedur“:

Es wird ein Spiel in Normalform wieder und wieder gespielt.

Zu jedem Zeitpunkt kann ein:e Spieler:in entweder die gleiche Strategie spielen wie in der Vorperiode, oder er/sie wechselt zu einer anderen Strategie mit einer gewissen Wahrscheinlichkeit.

Diese Wahrscheinlichkeit ist proportional zu um wieviel höher die durchschnittlichen Auszahlungen wären, wäre diese andere Strategie schon immer anstelle der Strategie der Vorperiode benutzt worden.

Formale Darstellung der Prozedur

- U: durchschnittliche Auszahlungen bis zum aktuellen Zeitpunkt
- j: Strategie der Vorperiode
- $V(k)$: durchschnittliche Auszahlungen bis zum aktuellen Zeitpunkt, falls bisher anstelle von Strategie j immer Strategie k gewählt worden wäre
- Falls $V(k) > U$: Wechsle von Strategie j zu Strategie k mit Wahrscheinlichkeit $\alpha(V(k) - U)$, wobei α eine Zahl ist, die klein genug ist.
- Spiele weiterhin Strategie j mit Wahrscheinlichkeit $1 - \sum_{k:V(k)>U} \alpha(V(k) - U)$

Beispiel: Mutprobe („Chicken“)

$U(\text{sitzen}, 50/50) = -1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{1}{2}$

$U(\text{springen}, 50/50) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$

		Buzz: sitzen (50%)	Buzz: springen (50%)
Jimmie	sitzen	$-1, -1$	$2, 0$ *
	springen	$0, 2$ *	$1, 1$

Nash Gleichgewichte in reinen Strategien:
(springen, sitzen) & (sitzen, springen)

Nash Gleichgewicht in gemischten Strategien:
(50% springen+50% sitzen,50% springen+50% sitzen)

Beispiel für einen Spielverlauf des Chicken-Spiels

Die Anpassung aus Sicht von Sp. 1

$t = 1$ (sitzen, sitzen) $\rightarrow u_1 = -1$

$t = 2$ (sitzen, sitzen) $\rightarrow u_1 = -1$

$t = 3$ (springen, springen) $\rightarrow u_1 = 1$

$t = 4$ (sitzen, springen) $\rightarrow u_1 = 2$

$t = 5$ (springen, sitzen) $\rightarrow u_1 = 0 \Rightarrow U = 1/5$

Ersetze nun jeweils springen von Sp. 1 durch sitzen und berechne die alternativen Auszahlungen:

$t = 1$ (sitzen, sitzen) $\rightarrow u_1 = -1$

$t = 2$ (sitzen, sitzen) $\rightarrow u_1 = -1$

$t = 3$ (sitzen, springen) $\rightarrow u_1 = 2$

$t = 4$ (sitzen, springen) $\rightarrow u_1 = 2$

$t = 5$ (sitzen, sitzen) $\rightarrow u_1 = -1 \Rightarrow V(sitzen) = 1/5$

Spieler:in 1 bleibt also bei springen.

Beispiel für einen Spielverlauf des Chicken-Spiels

Die Anpassung aus Sicht von Sp. 2

$t = 1$ (sitzen, sitzen) $\rightarrow u_2 = -1$

$t = 2$ (sitzen, sitzen) $\rightarrow u_2 = -1$

$t = 3$ (springen, springen) $\rightarrow u_2 = 1$

$t = 4$ (sitzen, springen) $\rightarrow u_2 = 0$

$t = 5$ (springen, sitzen) $\rightarrow u_2 = 2 \Rightarrow U = 1/5$

Ersetze nun jeweils sitzen von Sp. 2 durch springen und berechne die alternativen Auszahlungen:

$t = 1$ (sitzen, springen) $\rightarrow u_2 = 0$

$t = 2$ (sitzen, springen) $\rightarrow u_2 = 0$

$t = 3$ (springen, springen) $\rightarrow u_2 = 1$

$t = 4$ (sitzen, springen) $\rightarrow u_2 = 0$

$t = 5$ (springen, springen) $\rightarrow u_2 = 1 \Rightarrow V(springen) = 2/5$

$\alpha \cdot \left(\frac{2}{5} - \frac{1}{5}\right) = \alpha \cdot \frac{1}{5}$

Spieler:in 2 wechselt mit positiver Wahrscheinlichkeit zu springen.

Eigenschaften dieser Prozedur

- Einfachheit: leicht zu erklären und umzusetzen
- Nicht vom beste Antwort-Typ: Spieler:innen wehlen bessere Antworten, nicht beste Antworten.
- Trägheit: Es ist immer möglich, dass bei der alten Strategie verharrt wird.
- Unwissenheit: die Spieler:innen kennen die Auszahlungen der anderen Spieler:innen nicht (ihre eigenen aber schon)
- Kurzsichtigkeit: Zukünftige Reaktionen auf die eigenen Entscheidungen werden ignoriert

Stationarität und Konvergenz

Nash Gleichgewichte in reinen Strategien sind unter dieser Prozedur stationär.

Begründung: Wurde bisher immer das gleiche reine Nash Gleichgewicht gespielt, so wäre eine unilaterale Abweichung in der Vergangenheit niemals besser gewesen. Daher sind die Wechselwahrscheinlichkeiten für alle anderen Strategien gleich null.

Sergiu Hart und Andreu Mas-Collel zeigen zudem, dass die empirische Verteilung über die Menge der Profile von Strategien zu der Menge der korrelierten Gleichgewichte konvergiert.

Jedes Nash Gleichgewicht (egal ob rein oder gemischt) ist ein korreliertes Gleichgewicht. Die Umkehrung gilt aber nicht.

„Social aspiration reinforcement learning in Cournot games“

von Fatas, Morales, Jaramillo-Gutiérrez
erschienen in Economic Theory (2024)

Es wird ein Cournot-Spiel wieder und wieder gespielt.

Zu jedem Zeitpunkt passen alle Spieler:innen ihre Mengen gemäß einer Wahrscheinlichkeitsverteilung an.

Die Wahrscheinlichkeiten passen sich durch die tatsächlich erzielten Auszahlungen an.

Die Autoren zeigen, dass die Spieler:innen irgendwann die Mengen des Wettbewerbgleichgewichts spielen.

Formale Darstellung

n Firmen

Beschränkung der auswählbaren Mengen auf

$$\{0, \epsilon, 2\epsilon, \dots, q^w\}$$

wobei ϵ eine kleine positive Zahl ist und q^w die Wettbewerbsmenge mit $P(nq^w) = c'(q^w)$.

Gewinn von Firma i: $\pi_i(q_1, \dots, q_n) = P(\sum_{j=1}^{n} q_j) \cdot q_i - c(q_i)$

Aspiration level $A_q(t)$:
Neigung, zum Zeitpunkt t die Menge q zu spielen

Wahrscheinlichkeit zum Zeitpunkt t die Menge q zu spielen:

$$p_q(t) = \frac{A_q(t)}{A_0(t) + A_\epsilon(t) + A_{2\epsilon}(t) + \dots + A_{q^w}(t)}$$

Anpassung der Aspirations

Allgemein:

Nachdem in Periode t die Strategie s gewählt wurde und die Auszahlung π realisiert wurde, werden die Aspiration levels $A_s(t)$ wie folgt angepasst:

$$A_s(t+1) = \begin{cases} A_s(t) + \pi & \text{falls } s \text{ in } t \text{ gespielt wurde} \\ A_s(t) & \text{sonst} \end{cases}$$

Speziell in diesem Aufsatz:

Nachdem $i = 1, \ldots, n$ in Periode t die Mengen $q_1, \ldots, q_n$ gewählt haben und die Gewinne $\pi_1, \ldots, \pi_n$ erzielt wurden, werden die Aspiration levels $A_q(t)$ wie folgt angepasst:

$$A_q(t+1) = \begin{cases} A_q(t) + \pi_i - \bar{\pi} & \text{falls } q = q_i \text{ für ein } i = 1, \ldots, n \\ A_q(t) & \text{sonst} \end{cases}$$

Beispiel

↗ Grenzkosten: $c'(q) = 0$

Sei $P(Q) = 120 - Q$, $c(q) = 0$ und $n = 3$.

$q_i = 20$: symmetrische Kartellmengen
$q_i = 30$: Cournotmengen
$q_i = 40$: Wettbewerbsmengen

$\rightarrow$ zulässige Mengen:

$$\{0, 10, 20, 30, 40\}$$

$\frac{1}{3}(120 - Q)(q_1 + q_2 + q_3)$, Q

$$\begin{aligned} \pi_i(q_1, q_2, q_3) &= (120 - Q)\, q_i \\ \bar{\pi} &= \frac{1}{3}(\pi_1 + \pi_2 + \pi_3) = \frac{1}{3}(120 - Q)\, Q \\ \longrightarrow \quad \pi_i - \bar{\pi} &= \frac{1}{3}(120 - Q)(3q_i - Q) \end{aligned}$$

Falls $q_i < \frac{1}{3}Q \Rightarrow \pi_i < \bar{\pi}$ Falls $q_i > \frac{1}{3}Q \Rightarrow \pi_i > \bar{\pi}$

Anpassung im Beispiel

Mit

$$\pi_i - \bar{\pi} = \frac{1}{3}(120 - Q)(3q_i - Q)$$

und

$$A_q(t+1) = \begin{cases} A_q(t) + \pi_i - \bar{\pi} & \text{falls } q = q_i \text{ für ein } i = 1, \ldots, n \\ A_q(t) & \text{sonst} \end{cases}$$

werden überdurchschnittlichen Mengen verstärkt und die unterdurchschnittlichen Mengen geschwächt:

$$q_i > \frac{1}{3}Q \Leftrightarrow \pi_i > \bar{\pi} \Leftrightarrow A_{q_i}(t+1) > A_{q_i}(t)$$

$\Rightarrow$ Die Wahrscheinlichkeit $q_i = 40$ zu spielen konvergiert gegen 1.