Hochschule Düsseldorf
University of Applied Sciences

# Learning to Cooperate: Reinforcement Learning in the Iterated Prisoners Dilemma

Jonah Gräfe

## Das iterierte Gefangenendilemma

Das wiederholte Gefangenendilemma (Iterated Prisoner's Dilemma, IPD) ist ein klassisches Problem der Spieltheorie, das die Spannung zwischen Kooperation und Wettbewerb in wiederholten Interaktionen modelliert. In jeder Runde entscheiden sich zwei Spieler unabhängig voneinander entweder für die Kooperation (C) oder für die Defektion (D). Ihre Entscheidungen bestimmen ihre Auszahlungen auf der Grundlage einer Auszahlungsmatrix:

| Pleayer A/B | Cooperate (C) | Defect (D) |
|---|---|---|
| Cooperate (C) | (3, 3) | (0, 5) |
| Defect (D) | (5, 0) | (1, 1) |

Table 1. Auszahlungsmatrix

## *Deep* Q-Learning

Q-Learning-Agenten

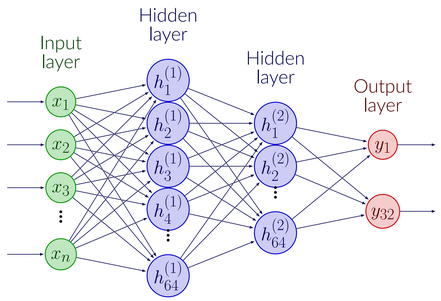
1. **Q-Werte:** Der Q-Wert stellt die erwartete kumulative Belohnung für das Ausführen einer Aktion in einem bestimmten Zustand und das anschließende Befolgen der optimalen Strategie dar.
2. **Aktualisierung:**

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a} Q(s',a) - Q(s,a)]$$

- $s, a, s'$: Aktueller Zustand, durchgeführte Aktion und der nächste Zustand.
- $r$: Erhaltene Belohnung
- $\alpha$: Lernrate
- $\gamma$ Diskontierungsfaktor

Deep Q-Learning-Agenten

1. **Neuronales Netz:** Bildet Zustände auf Q-Werte für alle möglichen Aktionen ab. Dies ersetzt die Q-Tabelle.
2. **Erfahrungswiedergabe:** Der Agent speichert vergangene Erfahrungen in einem Wiederholungspuffer.
3. **Training**: Minimierung des MSE zwischen vorhergesagten Q-Werten und Ziel Q-Werten mit Hilfe Adam-Optimierers




## Hypothese & Null-Hypothese

Hypothese ($H_1$):
Die Q-Learning Agenten erzielen signifikant höhere Belohnungen als zufällig gewählte Strategien (klassische Agenten) im Iterierten Gefangenendilemma (Iterated Prisoner's Dilemma). Dies legt nahe, dass die Q-Learning Agenten effektiv lernt und ihre Strategien anpassen, um etablierte Agenten wie Tit-for-Tat, Always Cooperate und Always Defect zu übertreffen.

Nullhypothese ($H_0$):
Die Leistung der Q-Learning Agent unterscheideen sich nicht signifikant von der Leistung zufällig gewählter Strategien. Alle beobachteten Unterschiede bei den Belohnungen sind auf Zufall oder Rauschen in den Daten zurückzuführen und nicht auf die Lernfähigkeit der Agenten.

Um diese Hypothesen zu bewerten, führen wir strenge statistische Tests durch, um die Leistung der Q-Learning Agenten über mehrere Spiele und Gegner hinweg zu analysieren. Die Ergebnisse geben Aufschluss darüber, ob der lernbasierte Ansatz einen echten strategischen Vorteil bietet.