

Maschinelles Lernen I - Grundverfahren

V09 Erweiterte Verfahren des Reinforcement Learning



Wintersemester 19/20

Prof. Dr. J.M. Zöllner, Karl Kurzer & Karam Daaboul

INSTITUT FÜR ANGEWANDTE INFORMATIK UND FORMALE BESCHREIBUNGSVERFAHREN



Reinforcement Learning (RL)

- Reinforcement Learning
Wiederholung und Beispiele
- Strategiebasiertes Lernen
Policy Gradient
- Kombiniertes Lernen
Actor Critic Verfahren

Imitation Learning und Reinforcement Learning

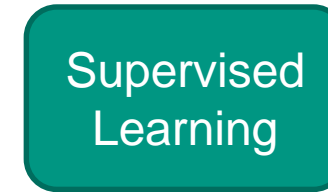



Was ist der Beobachtungsraum?

Was ist der Aktionsraum?

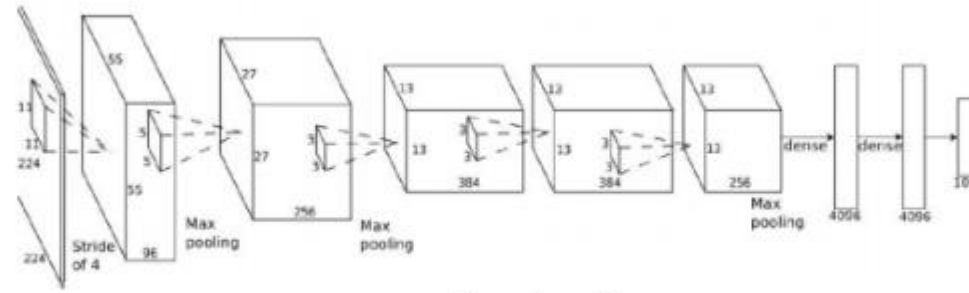
Was ist das Ziel?

Imitation Learning und Reinforcement Learning


 o_t
 a_t



 $\pi_{\theta}(a_t|o_t)$

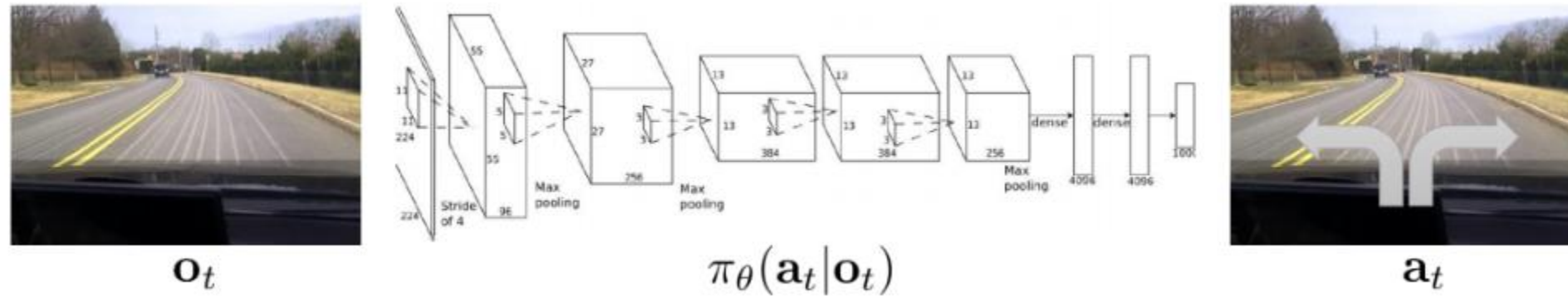
DriveNet, Hubschneider et al. 2017


 o_t

 $\pi_{\theta}(a_t|o_t)$

 a_t

Deep Reinforcement Learning, Berkeley, 2019

Imitation Learning und Reinforcement Learning



Welche Aktion ist besser?

Belohnungsfunktion $r(s_t, a_t)$

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{(s,a) \sim \pi_{\theta}} [r(s, a)]$$

Unendlicher Horizont

Markov Decision Process

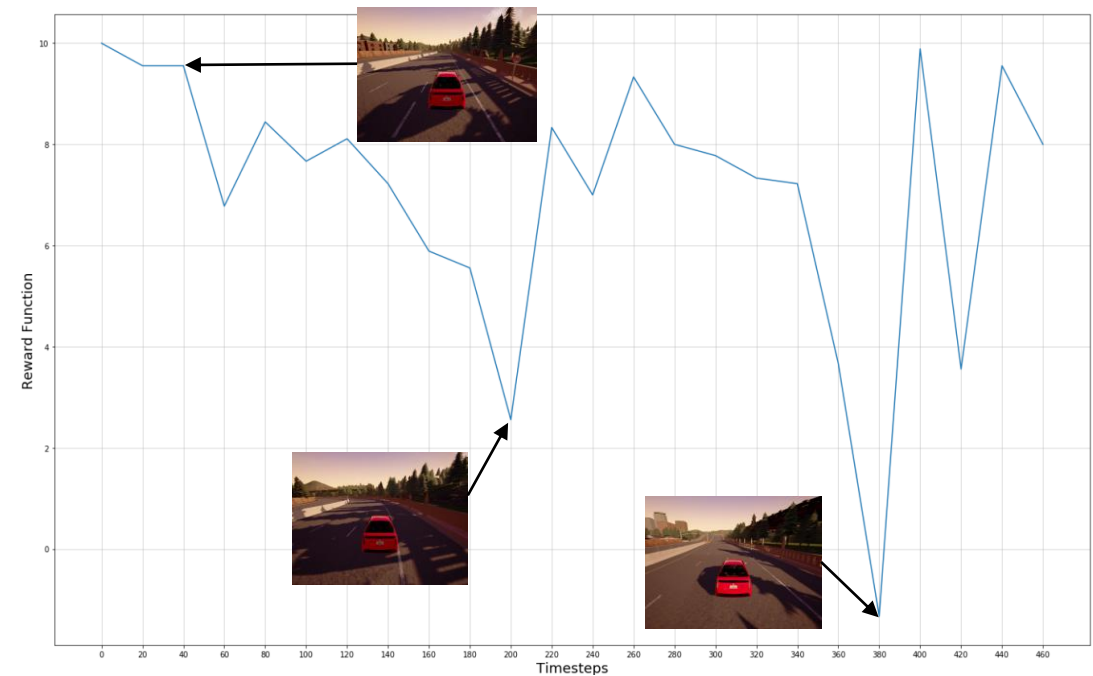
$s, a_t, r(s_t, a_t)$ und $r(s_{t+1} | s_t, a_t)$

$$\theta^* = \arg \max_{\theta} \sum_{t=1}^T \mathbb{E}_{(s_t, a_t) \sim \pi_{\theta}} [r(s_t, a_t)]$$

Endlicher Horizont

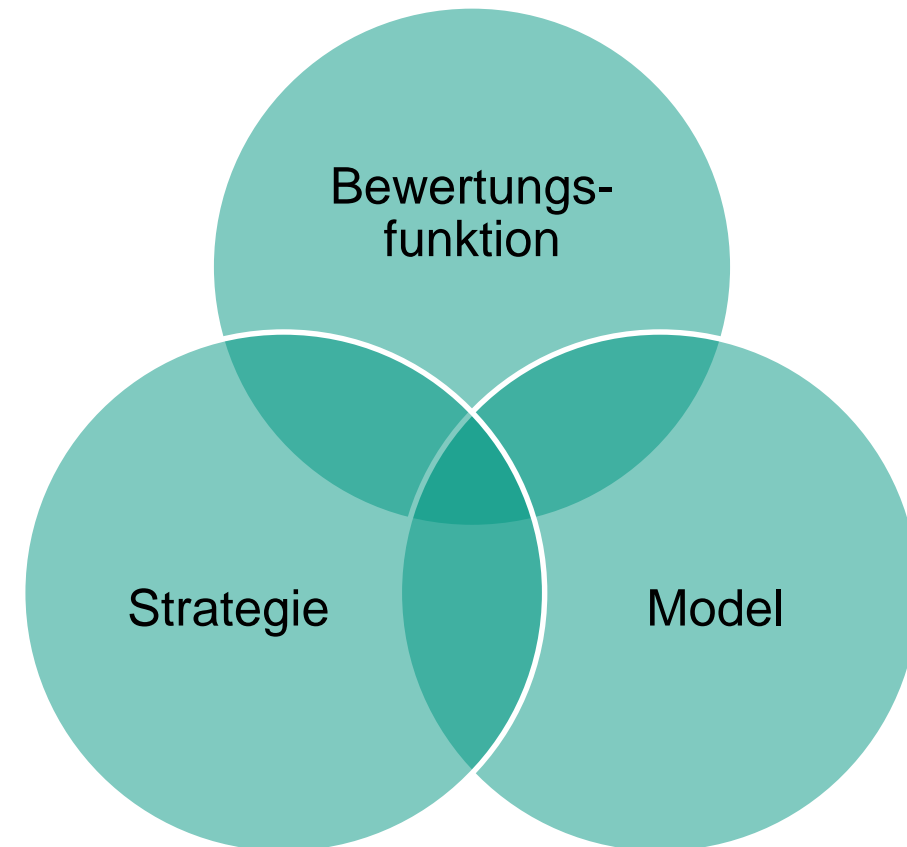
Ziele können als Maximierung der kummulierten Belohnung beschrieben werden.

Imitation Learning und Reinforcement Learning



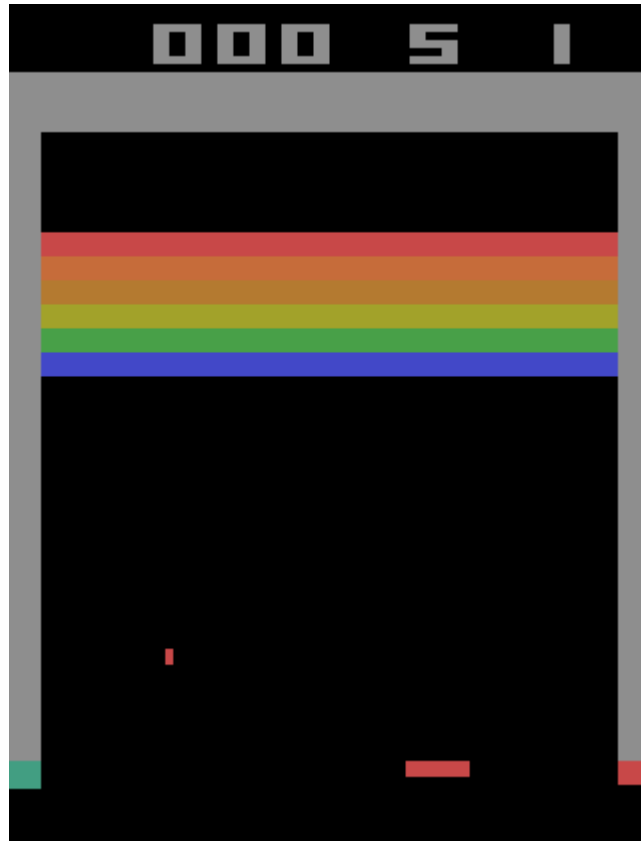
Taxonomie des Reinforcement Learning

- **Bewertungsbasiert**
 - Keine Strategie
 - Bewertungsfunktion
- **Strategiebasiert**
 - Strategie
 - Keine Bewertungsfunktion
- **Actor Critic**
 - Strategie
 - Bewertungsfunktion



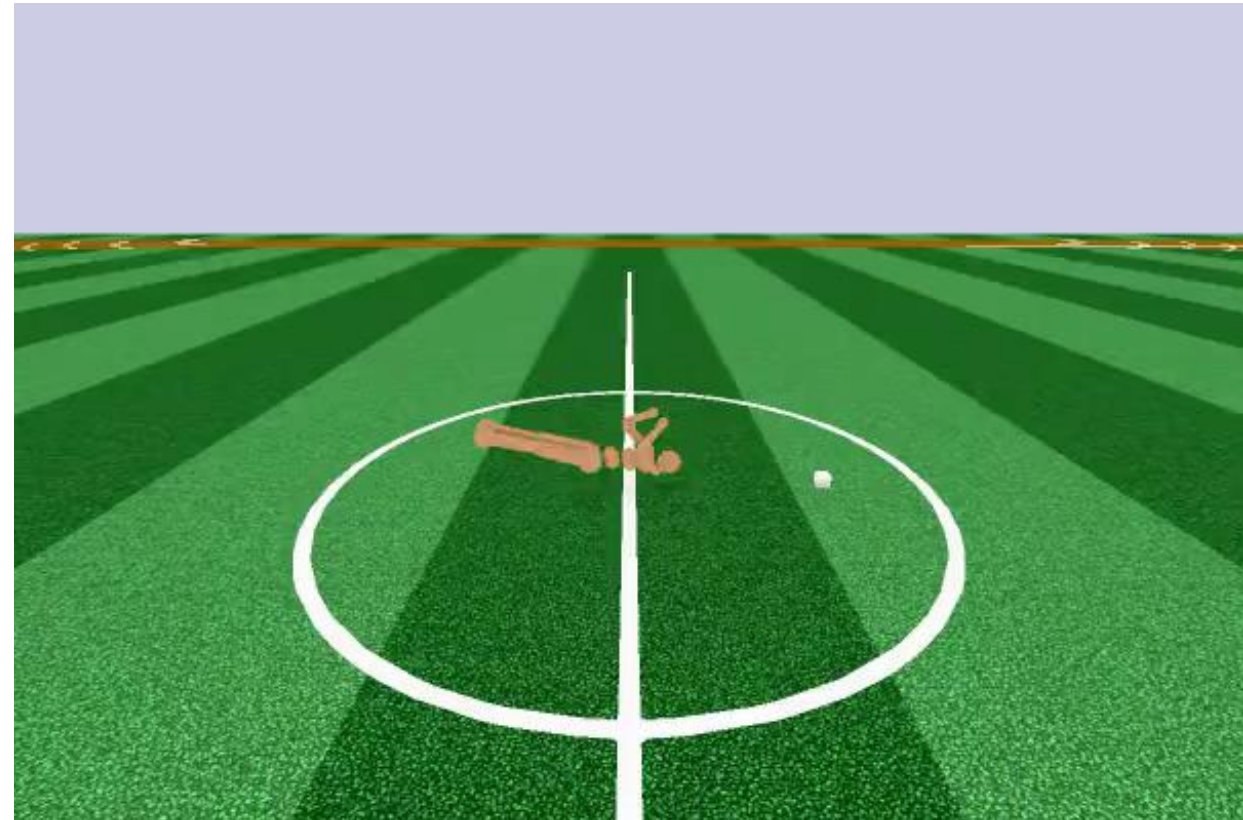
Reinforcement Learning: Beispiel (1)

Bewertungsfunktion



DQN, Mnih et al. 2013

Strategie



PPO, Schulman et al. 2017

<https://openai.com/blog/openai-baselines-ppo/>

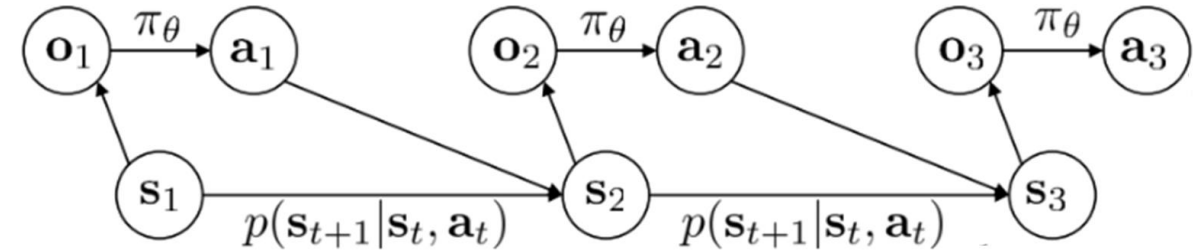
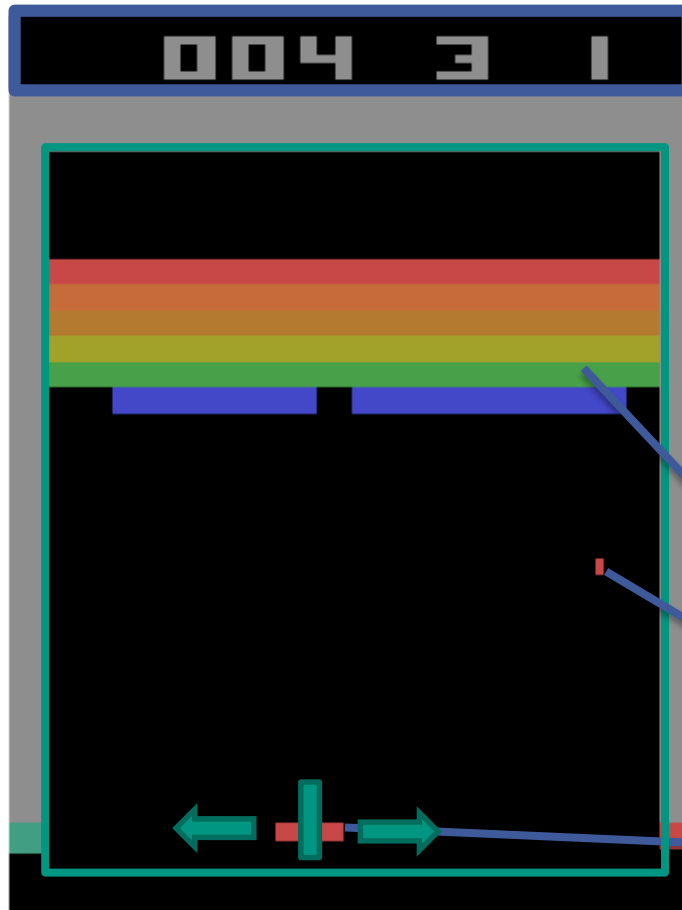
Reinforcement Learning: Beispiel (2)

Belohnung

Messung

Aktionsraum:

- rechts
- Stehen bleiben
- links



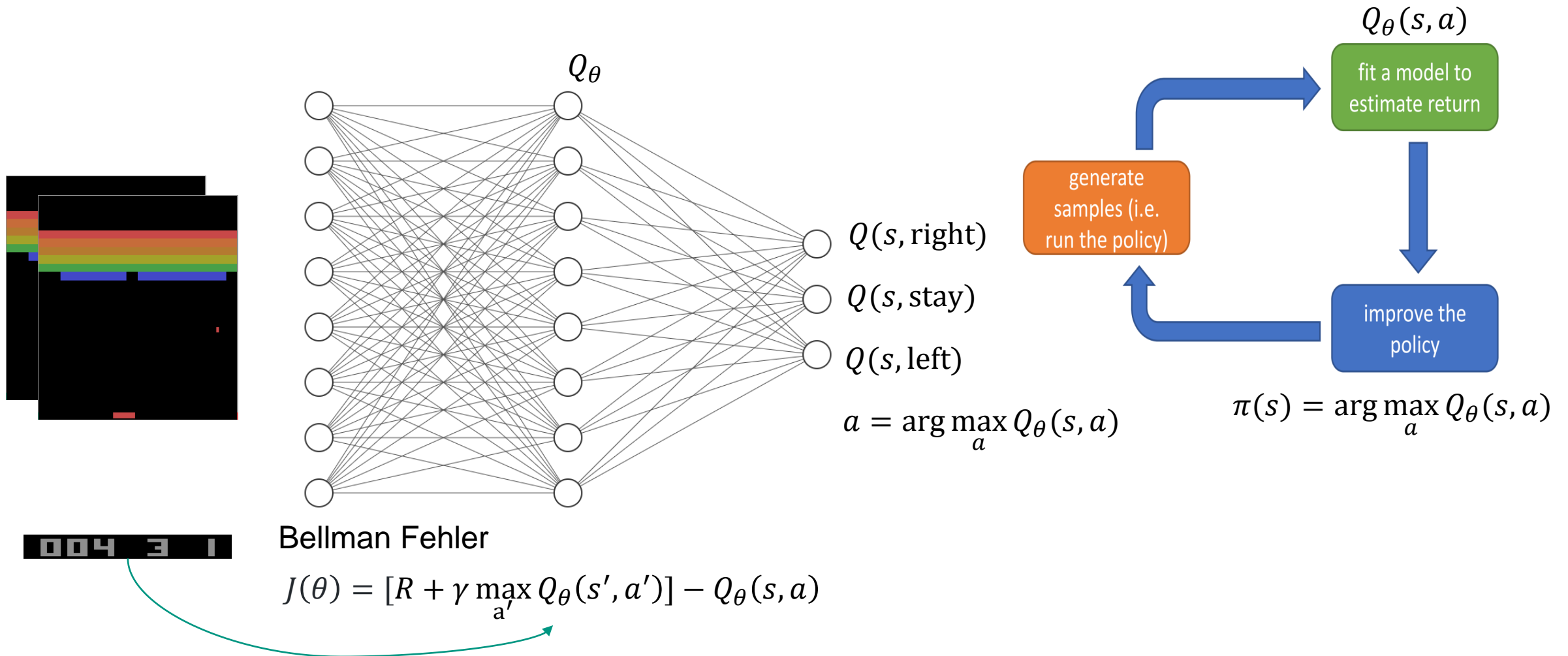
Ein Zustand erfüllt die Markov Eigenschaft wenn und nur wenn:

$$p[S_{t+1} | S_t] = p[S_{t+1} | S_1, \dots, S_t]$$

Zustandsraum:

- Die Position der Blöcke
- Die Position und die Richtung des Balls
- Die Position des Agenten

Bewertungsbasiert



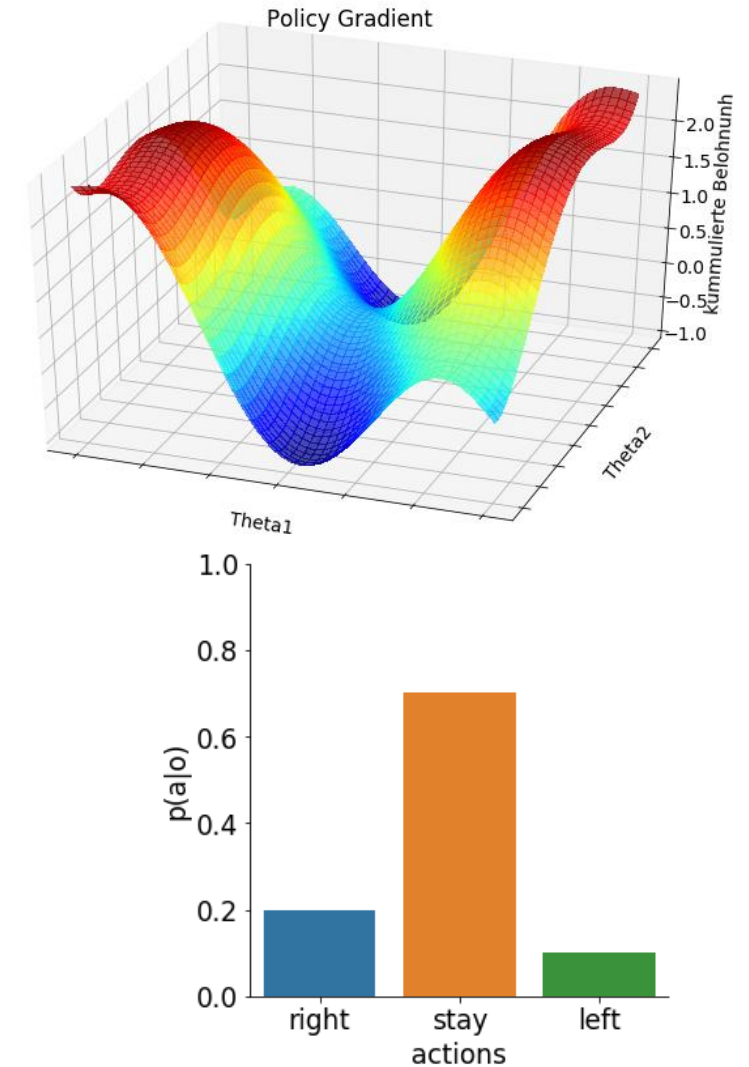
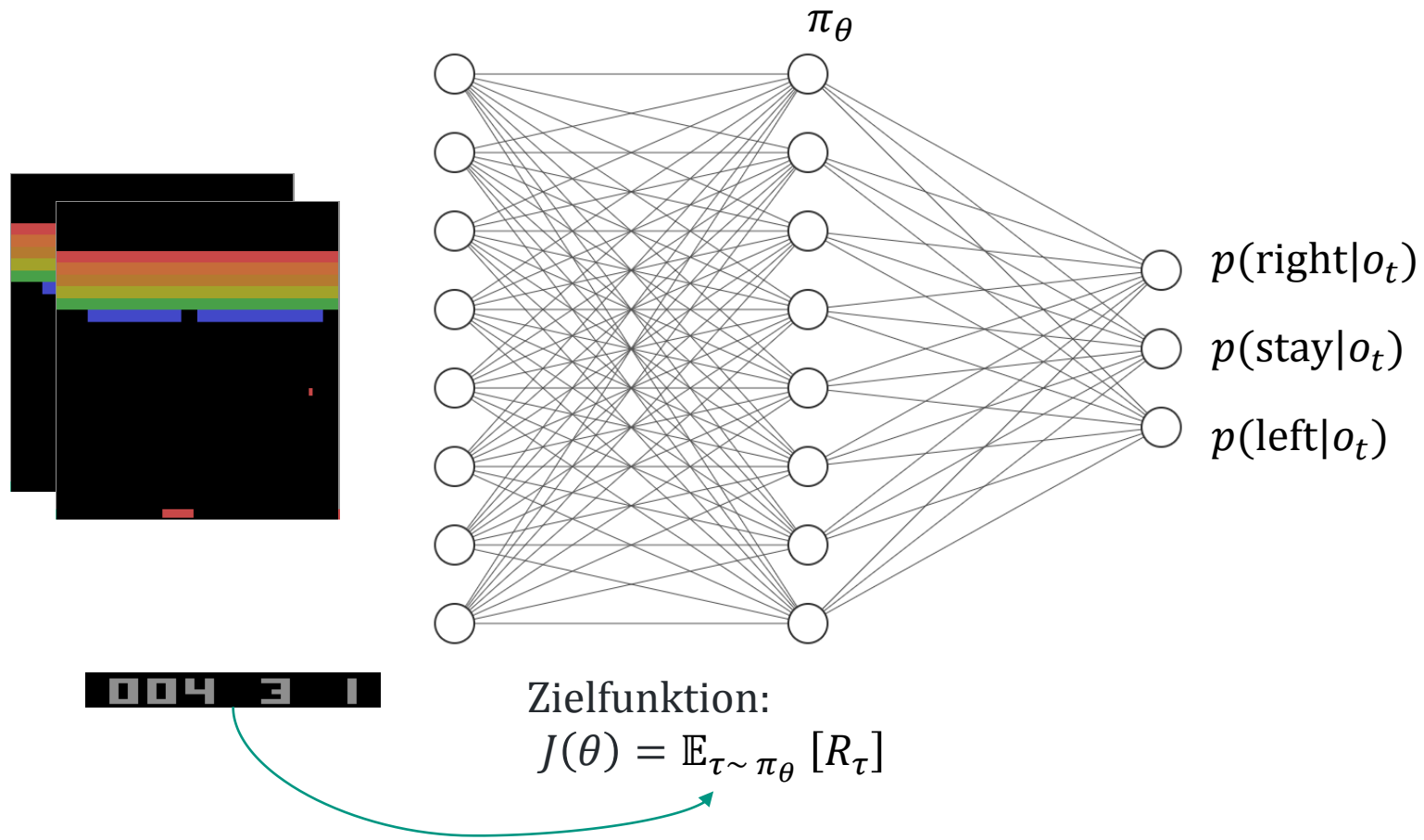
Reinforcement Learning (RL)

- Reinforcement Learning
Wiederholung und Beispiele

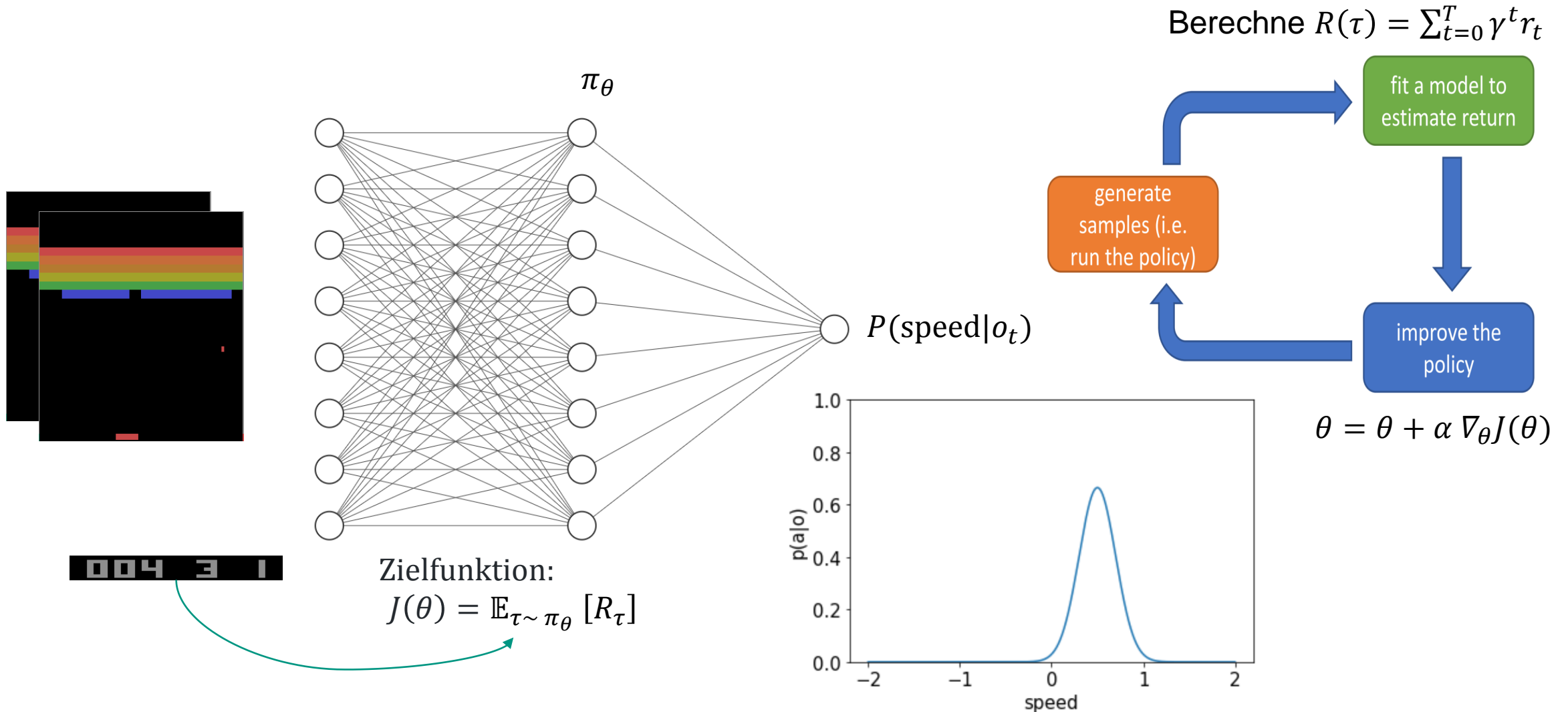
- Strategiebasiertes Lernen
Policy Gradient

- Kombiniertes Lernen
Actor Critic Verfahren

Strategiebasiert: Diskrete stochastische Policy



Strategiebasiert: Kontinuierliche stochastische Policy

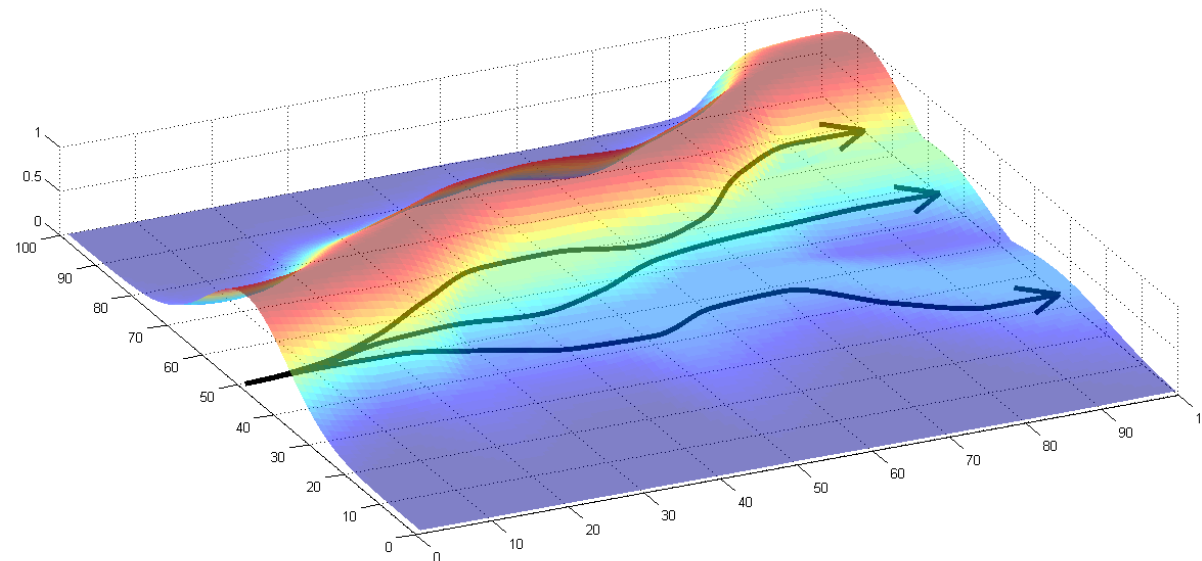


Trajektorien

- Eine Trajektorie τ ist eine Abfolge von Zuständen und Aktionen.

$$\tau_{\pi_{\theta}} = (s_0, a_0, s_1, a_1, \dots)$$

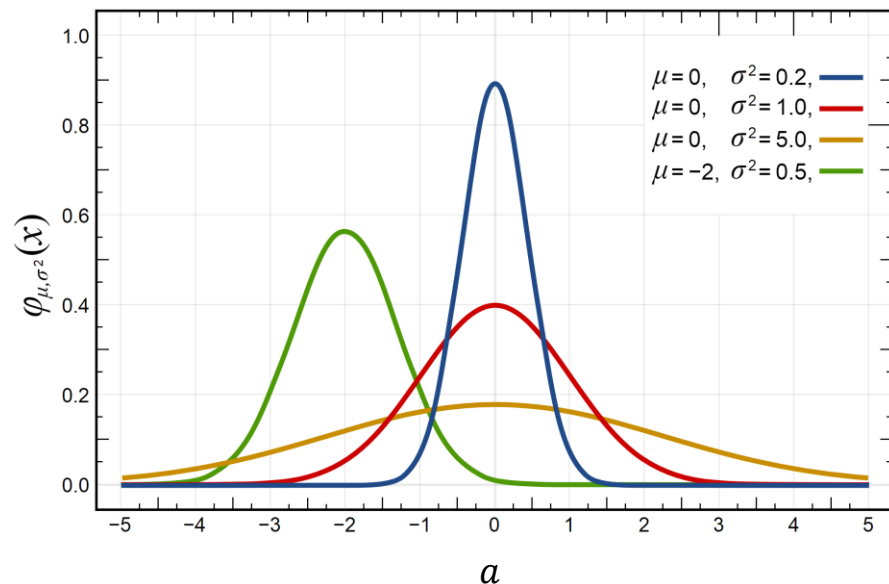
- $R(\tau) = \sum_{t=0}^T R(s_t, a_t)$ ist die Belohnung der Trajektorie



Strategiebasiertes Lernen

- Parametrisiert die Strategie explizit
- Erlaubt stochastische Strategien
- Erlaubt hochdimensionale und kontinuierliche Aktionsräume

Stochastische Strategie



Zielfunktion

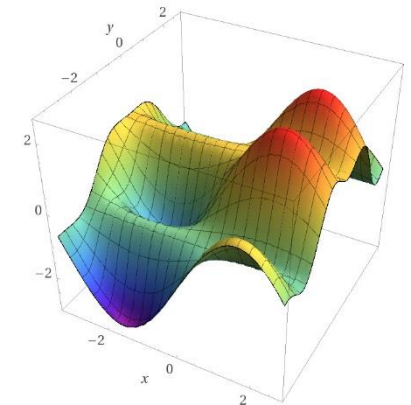
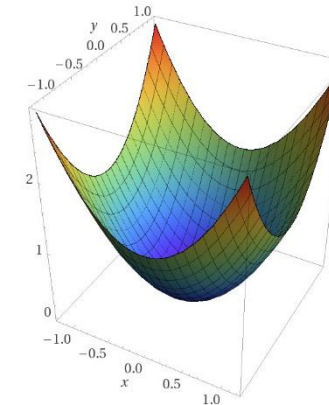
- Ziel: optimale Strategie zu lernen (maximale Belohnung)

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [R_{\tau}]$$

$$\theta^* = \arg \max_{\theta} J(\theta) = \arg \max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R_{\tau}]$$

- Finden der optimalen Strategieparameter durch Gradientenaufstieg

$$\theta_{k+1} = \theta_k + \alpha \nabla_{\theta} J(\theta)$$

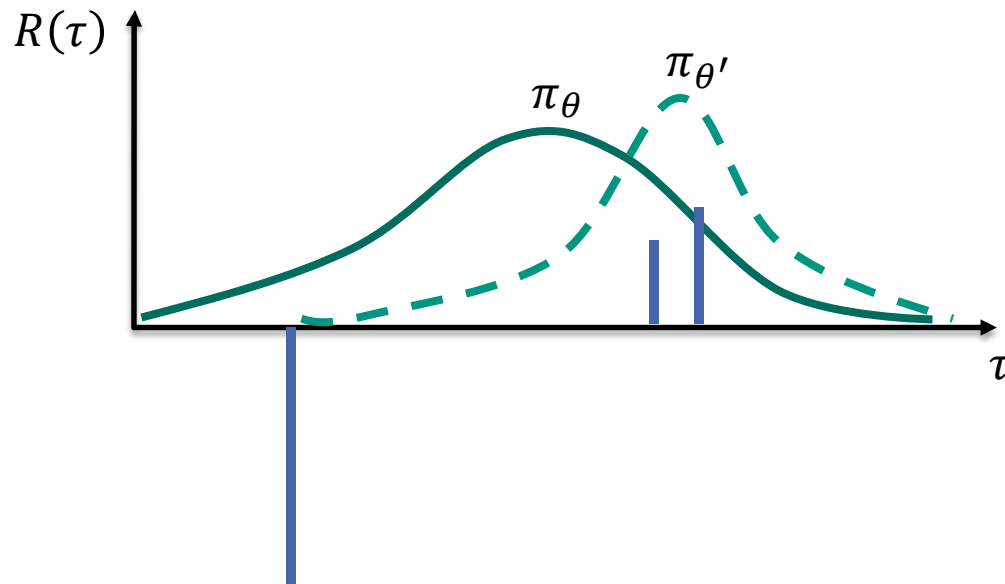


Definition

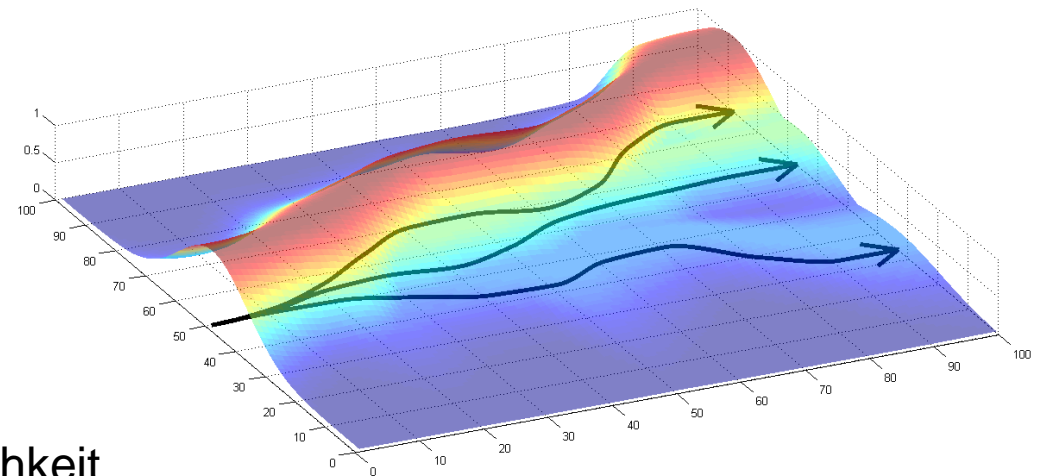
Die Ableitung der Zielfunktion nach Strategieparametern θ

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]$$

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]$$



Der Gradient erhöht die Wahrscheinlichkeit von Trajektorien mit positiven Belohnungen; umgekehrt werden Trajektorien mit negativen Belohnungen unwahrscheinlicher.



Ableiten des Strategiegradienten (Policy Gradient)

- Partielle Ableitung der Zielfunktion nach θ

$$\nabla_{\theta} J(\theta) = \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(\tau)]$$

- Expandieren der Erwartung

$$\nabla_{\theta} J(\theta) = \nabla_{\theta} \int P(\tau|\theta) R(\tau)$$

$$\nabla_{\theta} J(\theta) = \int \nabla_{\theta} P(\tau|\theta) R(\tau)$$

- Log Trick: $\frac{\nabla x}{x} = \nabla \log x$

$$\nabla_{\theta} P(\tau|\theta) = P(\tau|\theta) \frac{\nabla_{\theta} P(\tau|\theta)}{P(\tau|\theta)} = P(\tau|\theta) \nabla_{\theta} \log P(\tau|\theta)$$

$$\nabla_{\theta} J(\theta) = \int P(\tau|\theta) \nabla_{\theta} \log P(\tau|\theta) R(\tau)$$

- Erwartungswert

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau|\theta) R(\tau)]$$

Ableiten von $\nabla_{\theta} \log P(\tau|\theta)$

- $\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau|\theta) R(\tau)]$
- $R(\tau) = \sum_{t=0}^T R(s_t, a_t)$ ist die Belohnung der Trajektorie
- $P(\tau|\theta)$ ist die Wahrscheinlichkeit von Trajektorie τ unter Verwendung der Strategie π_{θ}

$$P(\tau|\theta) = \overbrace{\rho_0(s_0)}^{\text{Startverteilung}} \prod_{t=0}^T P(s_{t+1}|s_t, a_t) \pi_{\theta}(a_t | s_t)$$

- Log-Wahrscheinlichkeit einer Trajektorie

$$\log P(\tau|\theta) = \log \rho_0(s_0) + \sum_{t=0}^T (\log P(s_{t+1}|s_t, a_t) + \log \pi_{\theta}(a_t | s_t))$$

Ableiten von $\nabla_{\theta} \log P(\tau|\theta)$

$$\log P(\tau|\theta) = \underbrace{\log \rho_0(s_0)}_{\text{Unabhängig von } \theta} + \sum_{t=0}^T (\underbrace{\log P(s_{t+1}|s_t, a_t)}_{\text{Unabhängig von } \theta} + \log \pi_{\theta}(a_t | s_t))$$

$$\nabla_{\theta} \log P(\tau|\theta) = \underbrace{\nabla_{\theta} \log \rho_0(s_0)}_{\text{Unabhängig von } \theta} + \sum_{t=0}^T (\underbrace{\nabla_{\theta} \log P(s_{t+1}|s_t, a_t)}_{\text{Unabhängig von } \theta} + \nabla_{\theta} \log \pi_{\theta}(a_t | s_t))$$

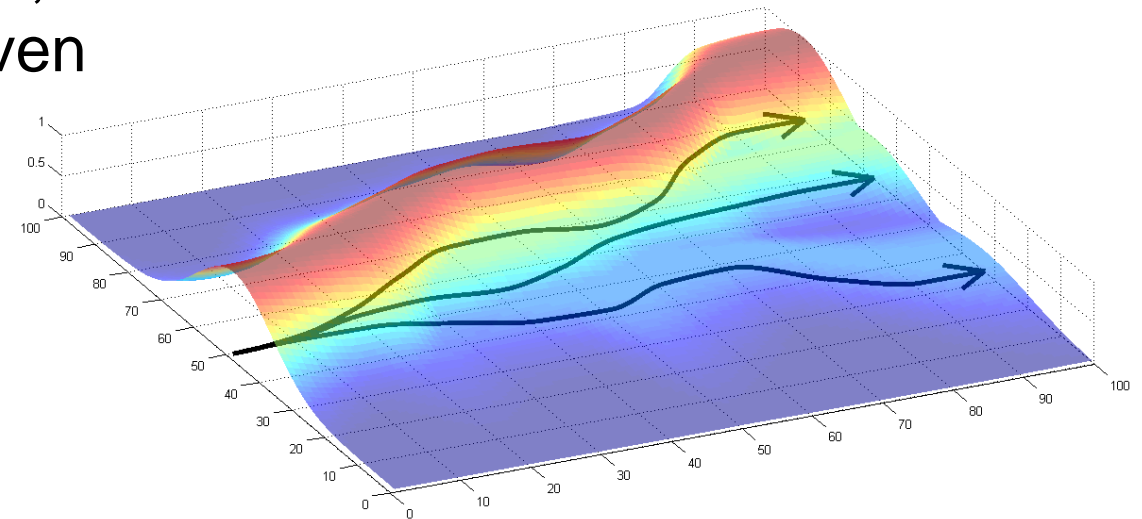
$$\nabla_{\theta} \log P(\tau|\theta) = \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau|\theta) R(\tau)] = \nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]$$

Policy Gradient

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right] \approx \frac{1}{|D|} \sum_{\tau \in D} \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau)$$

Der Gradient erhöht die Wahrscheinlichkeit von Trajektorien mit positiven Belohnungen; umgekehrt werden Trajektorien mit negativen Belohnungen unwahrscheinlicher.

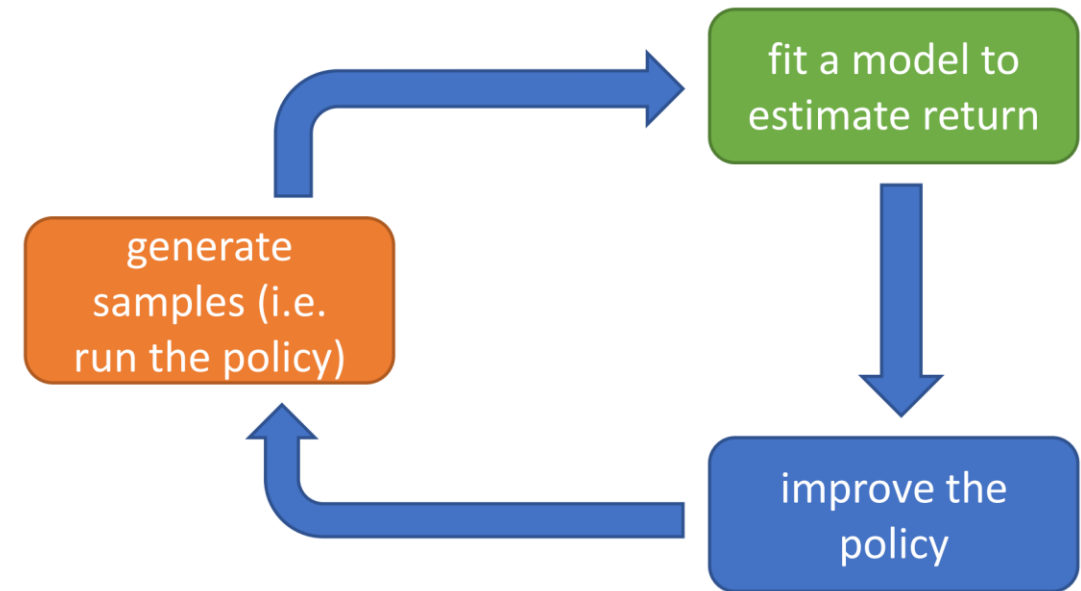


Policy Gradient – REINFORCE

Initialisiere Strategieparameter $\theta \in R^d$ zufällig $\rightarrow \pi(a|s, \theta)$

Wiederhole:

1. Generiere D Trajektorien/Episoden $(s_0, a_0, r_1, \dots, s_{T-1}, a_{T-1}, r_T)$ mit $\pi(\cdot | \cdot, \theta)$
2. $\nabla_{\theta} J(\theta) \leftarrow \frac{1}{|D|} \sum_{\tau \in D} \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau)$
3. $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$



Charakteristika – Policy Gradient

- Strategie wird explizit gelernt/verbessert
- Erlaubt stochastische Strategien
- Generell stabiler (smooth updates)
- Dateneffizient (On-Policy)
- Erlaubt hochdimensionale und kontinuierliche Aktionsräume

Probleme bei Strategiegradientenmethoden

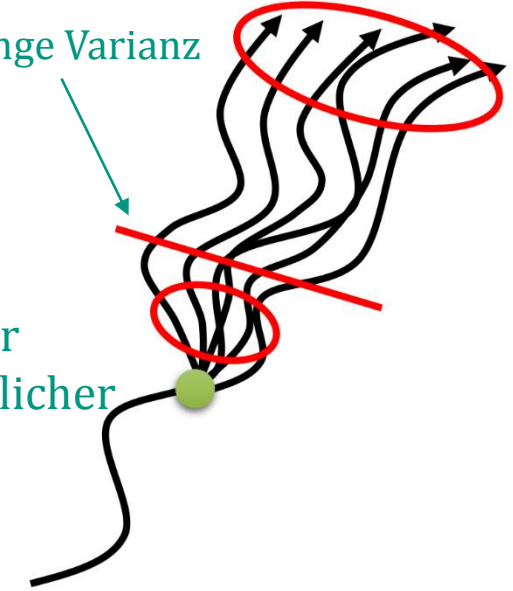
■ Varianz

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau) \right]$$

hohe Varianz

geringe Varianz

jede Trajektorie mit positiver Belohnung wird wahrscheinlicher



■ Lösung: Baselines

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) [R(\tau) - b] \right]$$

$$b = \frac{1}{N} \sum_{i=1}^{|D|} R(\tau) \quad \text{Ist der Gradient weiterhin korrekt?}$$

Baselines

- Die Ableitung der Zielfunktion nach Strategieparametern θ

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(\tau) R(\tau)]$$

- wobei:

$$\nabla_{\theta} \log \pi_{\theta}(\tau) = \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)$$

- Hinzufügen eines Baselines zu der Gleichung

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(\tau) [R(\tau) - b]]$$

$$\mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(\tau) b] = \int \underbrace{\pi_{\theta}(\tau) \nabla_{\theta} \log \pi_{\theta}(\tau)}_{\nabla_{\theta} \pi_{\theta}(\tau)} b d\tau = \int \nabla_{\theta} \pi_{\theta}(\tau) b d\tau = b \nabla_{\theta} \int \pi_{\theta}(\tau) d\tau = b \nabla_{\theta} 1 = 0$$

- Das Subtrahieren einer Baseline bringt kein Bias in den Erwartungswert ein.
- Die durchschnittliche Belohnung ist nicht die optimale Baseline, jedoch hinreichend gut.

Reinforcement Learning (RL)

- Bestandteile des RL Problems
 - Markov'scher Entscheidungsprozess
- Wertbasiertes Lernen
 - Q-Learning
- Strategiebasiertes Lernen
 - Policy Gradient
- Kombiniertes Lernen
 - Actor Critic Verfahren

Actor-Critic Verfahren

- Statt zufällige $R(\tau)$ zu sampeln, wird ein **Critic** hinzugefügt, um eine Q-Funktion zu approximieren:

$$Q_w(s, a) \approx Q^{\pi_\theta}(s, a)$$

- approximierter Policy Gradient:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)]$$

- Actor wird mit diesem Gradient aktualisiert
- Critic durch Policy Evaluation Verfahren

- Erweiterung durch Advantage-Funktion

$$A(s, a) = Q(s, a) - V(s)$$

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) A(s, a)]$$

Probleme mit RL

- Dateneffizienz
 - Performanz
 - Belohnungsfunktion
 - Lokale Optima
 - Generalisierungsprobleme
- Keine Standardverfahren

Datenineffizienz

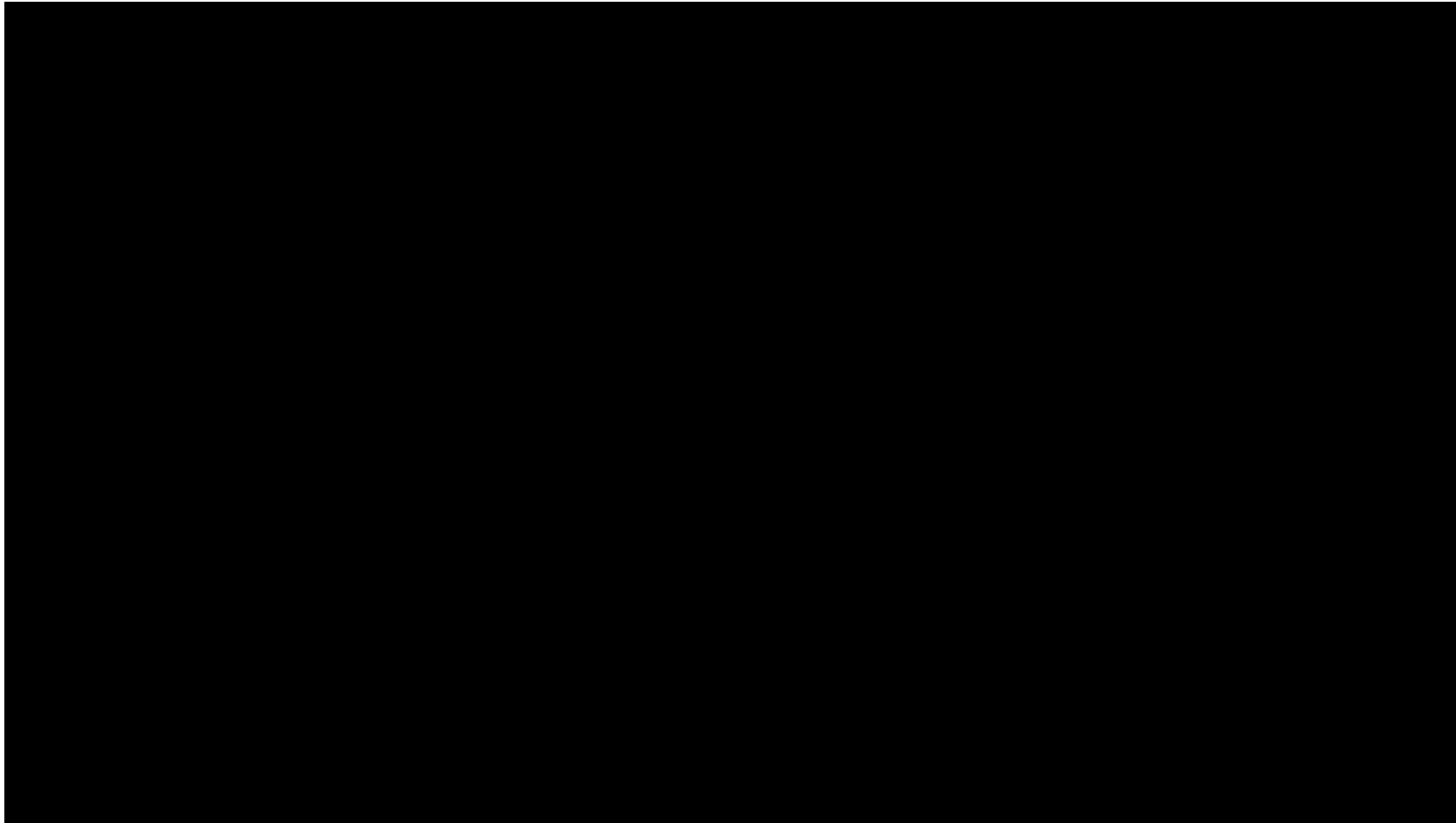


Können Erfahrungen genutzt werden um
Explorationsstrategien abzuleiten?



Generalisierungsprobleme
Levine et al. 2016

Multi-Agent Reinforcement Learning

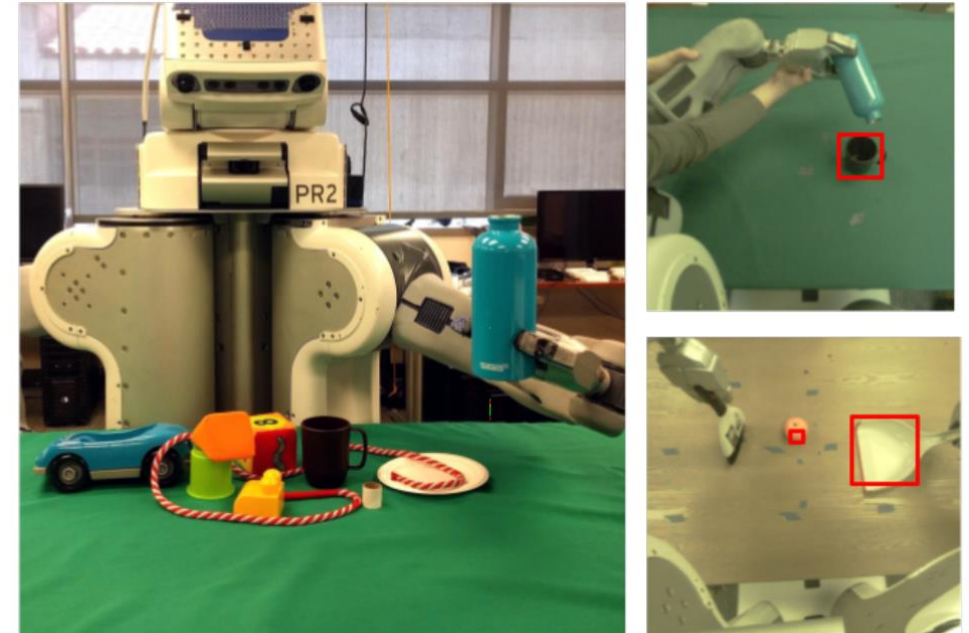
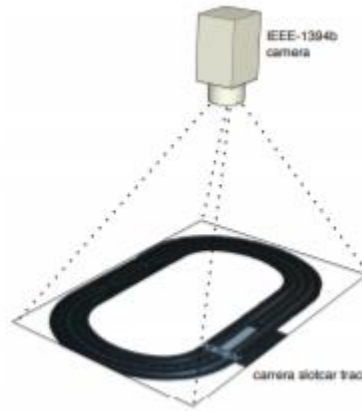


<https://openai.com/blog/emergent-tool-use/>

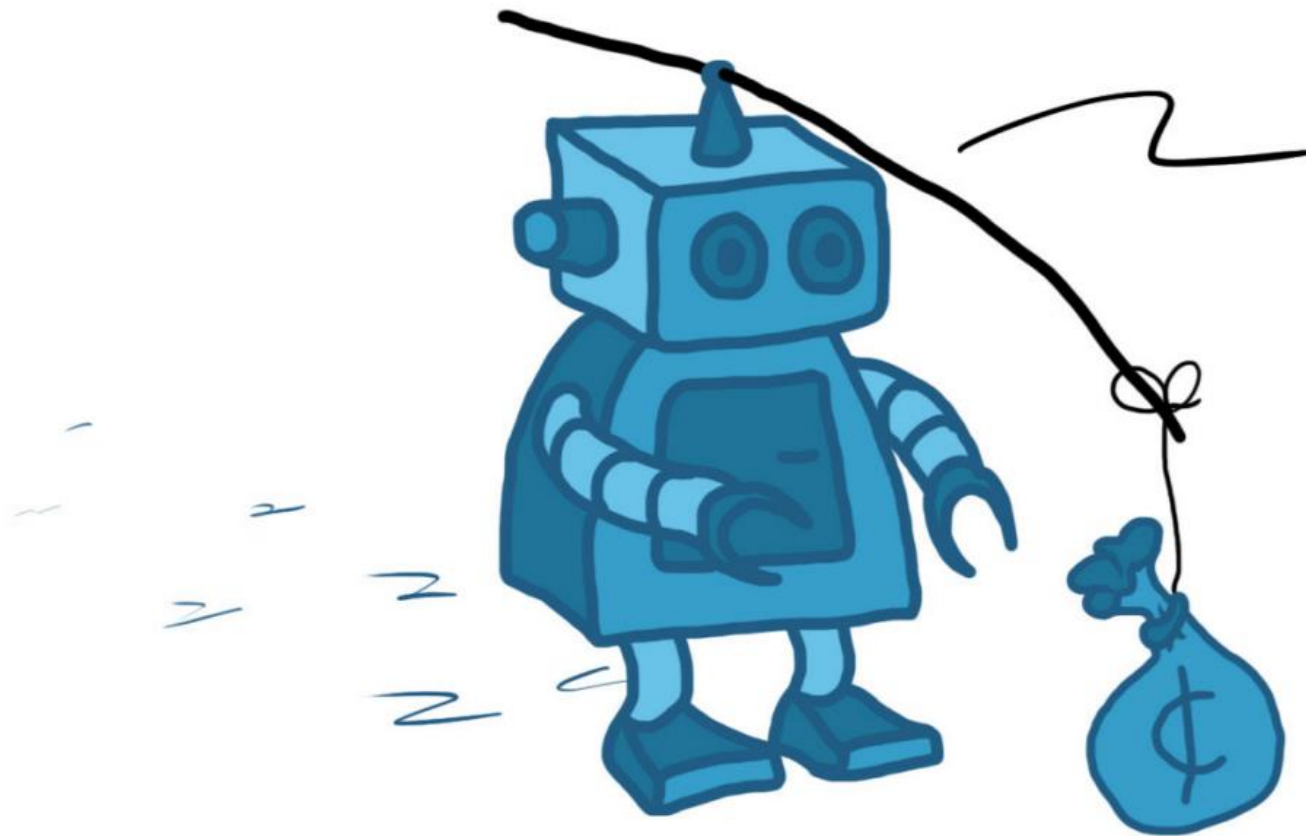
Pixel to Control



Lange et al. 2016



Devin et al. 2017



this is the way you
want me to go?

Literatur

- R. Sutton 2018 – “Reinforcement Learning: An Introduction”
- S. Levine 2018 – “Deep Reinforcement Learning” (Berkley Course on RL)
- P. Abbeel 2017 – “Deep RL Bootcamp” (Berkley Course on RL)
- D. Silver 2015 – “Reinforcement Learning” (UCL Course on RL)
- OpenAi 2018 – “SpinningUp”