

# Soft Actor Critic

Oliver, Leon Büttinghaus, Thilo Röthemeyer

15. April 2021

# Contents

1 SAC Grundprinzip

2 SAC Update Regeln

3 SAC Algorithmus

# Kontinuierlicher Aktionsraum

- kontinuierliche Aktionsräume benötigen
  - ⇒ Approximation für Q-Funktion
  - ⇒ Approximation für Strategie
- Schritt von Tabellen zu DNNs
- Optimierung mittels gradient descent

# Funktionen und deren Netzwerke

- State Value Funktion:

$V_\psi(s_t)$  → Skalar als Ausgabe

- Q-Funktion:

$Q_\theta(s_t, a_t)$  → Skalar als Ausgabe

- Strategie:

$\pi_\phi(s_t | a_t)$  → Mittelwert und Kovarianz als Ausgabe ⇒ Gauss

Mit Parametervektoren  $\psi$ ,  $\theta$  und  $\phi$

# State Value Funktion

- eigenes Netzwerk nicht notwendig, aber
  - stabilisiert Training
  - macht simultanes Training aller Netzwerke möglich

# Optimierung State Value Funktion

# Q-Funktion

# Optimierung Q-Funktion

# Optimierung der Strategie

# Algorithmus (1/2)

# Algorithmus (2/2)

# Part 2

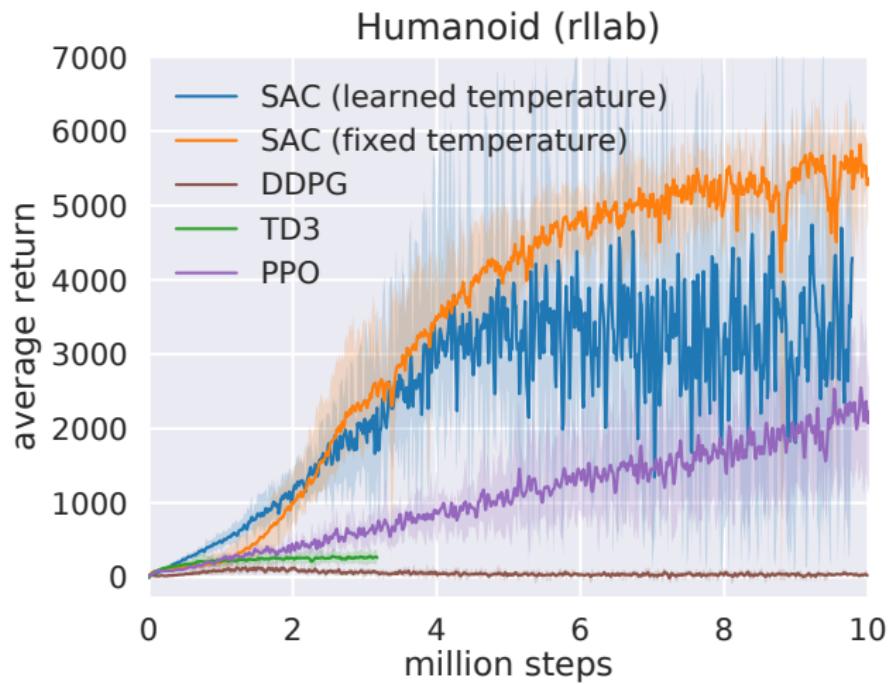
# Ziel der Experimente

- Stabilität und Sample Komplexität im Vergleich zu anderen Algorithmen
  - kontinuierliche Aufgaben
  - verschiedene Schwierigkeitgrade
- OpenAI gym und rllab

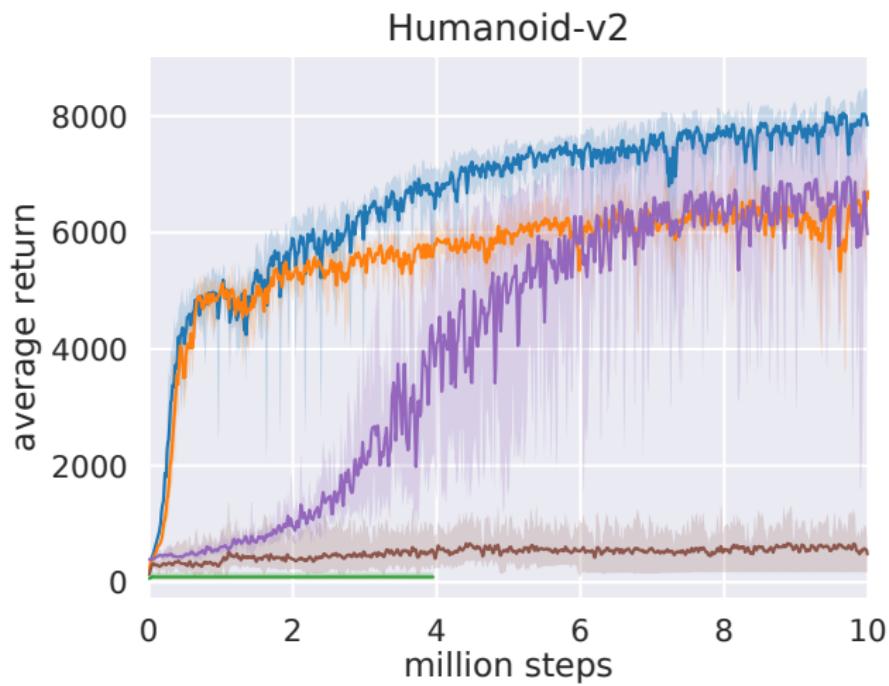
# Vergleich zu anderen Algorithmen

- SAC
  - mean action
  - feste und variable Temperatur
- PPO
- DDPG
- TD3
- SQL mit zwei Q Funktionen
  - evaluated with exploration noise
- 5 Instanzen mit einer Evaluation alle 1000 Schritte
- Total average return shown in the following
- schattierter Verlauf sind alle fünf Durchläufe

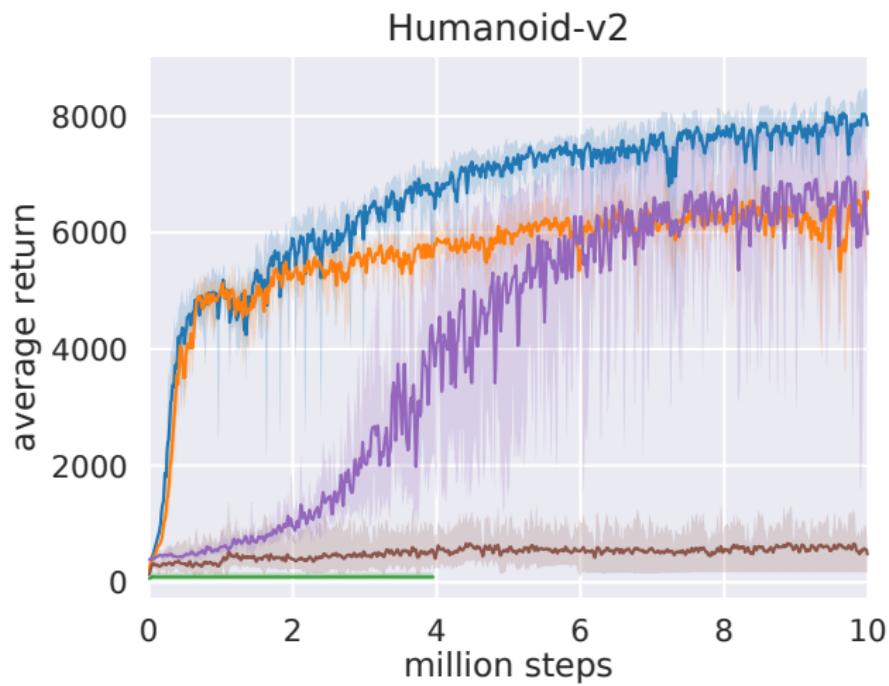
# Ergebnisse



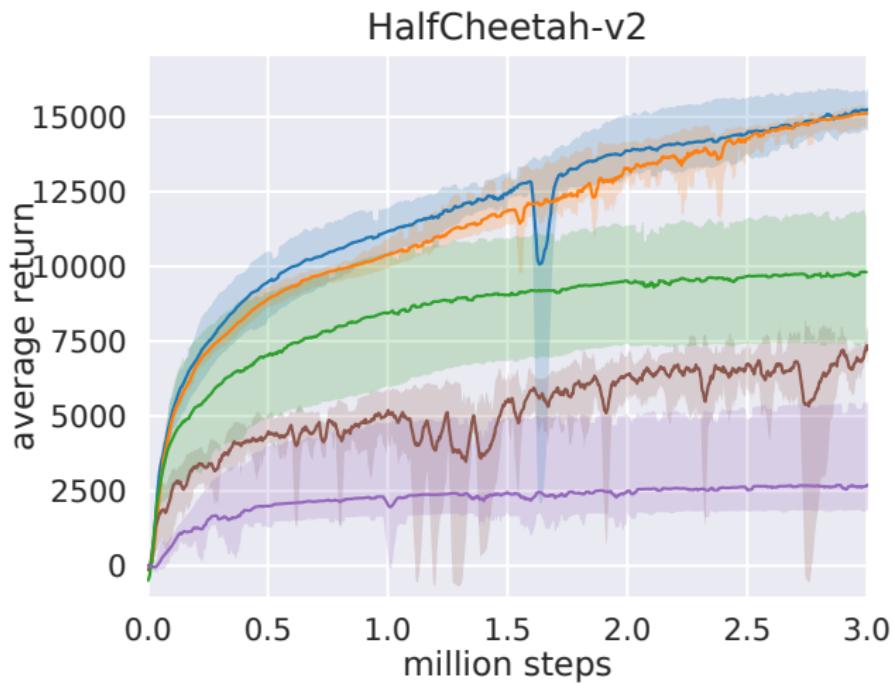
# Ergebnisse



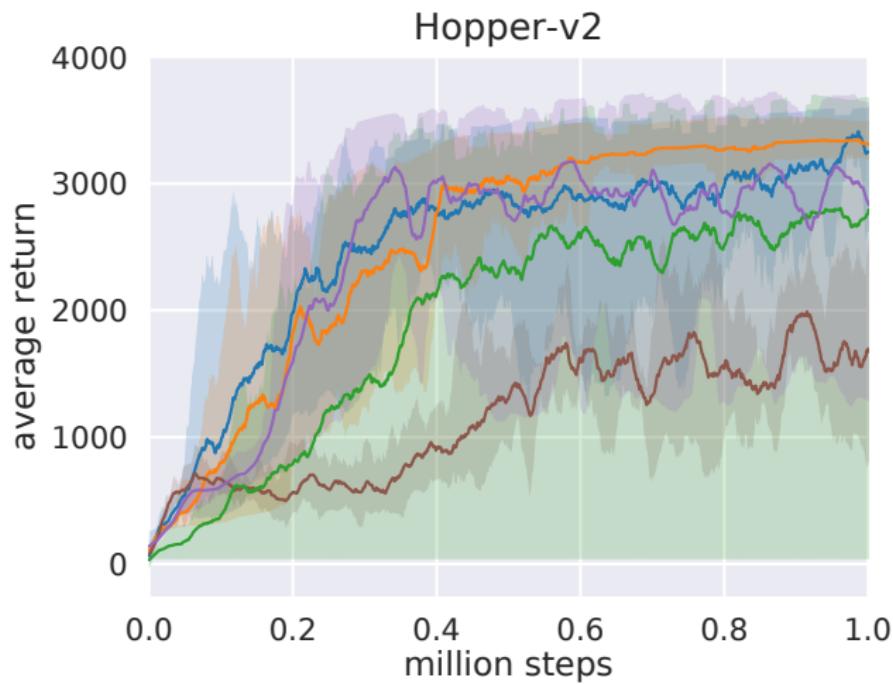
# Ergebnisse



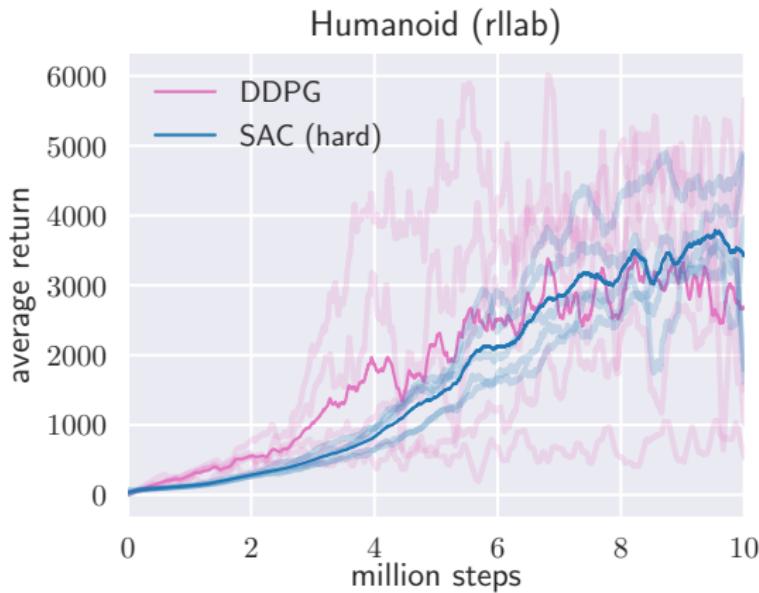
# Ergebnisse



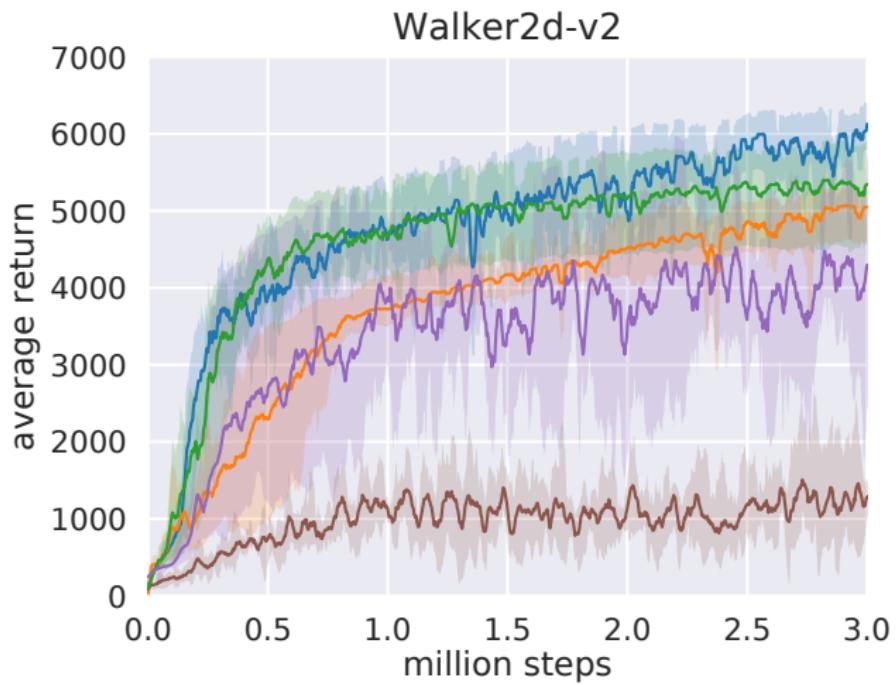
# Ergebnisse



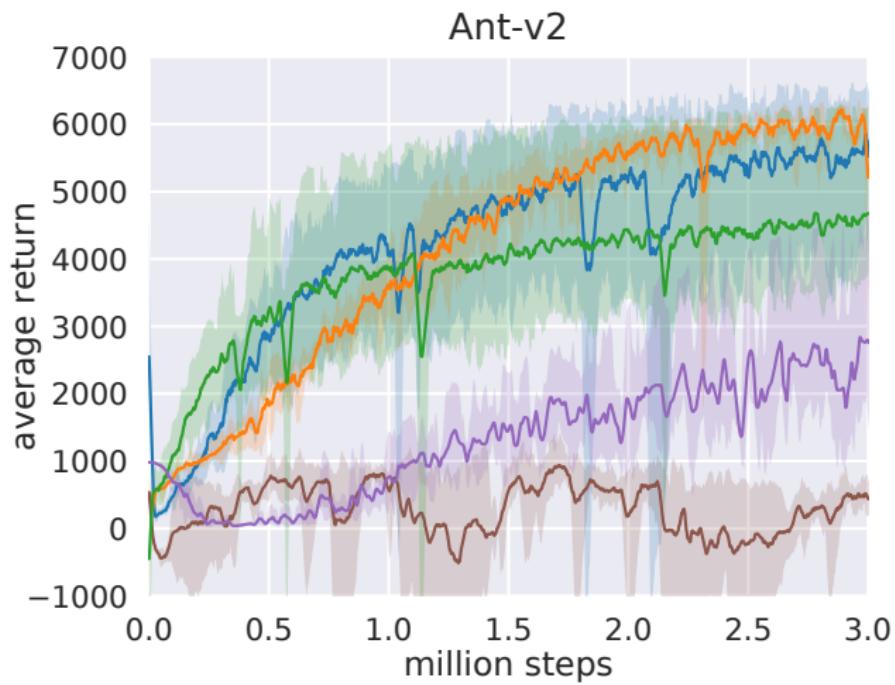
# Ergebnisse



# Ergebnisse



# Ergebnisse



# Zusammenfassung

- soft actor critic vorgestellt
  - off policy
  - Entropiemaximierung verbessert Stabilität
  - Besser als state of the art Algorithmen