

[RL22] Q&A Session on Policy-Based RL #2

23.06.2022

Christopher Mutschler

Let's play Kahoot! again...

Kahoot!

Let's play Kahoot!

The image shows a screenshot of the Kahoot! website homepage. The browser address bar shows 'kahoot.com'. The navigation menu includes 'Kahoot!', 'News', 'School', 'Work', 'Home', 'Study', 'Academy', 'AccessPass', 'Contact sales', 'Explore content', 'Play', 'Sign up', 'Log in', and 'EN'. The 'Play' button is circled in red. Below the navigation are four promotional cards:

- Make learning awesome!**
Kahoot! delivers engaging learning to billions.
[Sign up for free!](#)
- Make your team superstar presenters**
Set your whole team up to deliver awesome presentations with Kahoot! 360 Spirit, our best plan from only \$16 per month.
[Learn more >](#)
[Buy now](#)
- NEW! Create a branded experience with Kahoot! themes**
Boost audience engagement by customizing your kahoots for your work setting.
[Choose Kahoot! 360 Pro Max](#)
- Meet Kahoot! Kids!**
Spark your child's curiosity for learning with our new playful app experience.
[Get started today](#)

what is the latest material you **have** studied (i.e., where are you with the lecture content as today)

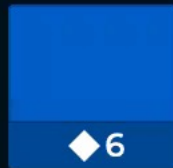


Medien anzeigen

▲ Lecture 2: Dynamic Programming 25%	◆ Lecture 3: Model-Free Prediction
● Lecture 4: Model-Free Control 13%	■ Lecture 5: Value Function Approximation 0%
🏠 Lecture 6: Policy-based RL #2 0%	▼ Lecture 7: Policy-based RL #2 56%

Christopher Mutschler
Tami
Tami
Georg Rabenstein
FAU084414
FAU084414
Pelin Genc
Pelin Genc
VJ

The estimation of the policy gradient requires transition probabilities!



Medien anzeigen

Wahr

Falsch



Tami

Tami



FAU084414

FAU084414

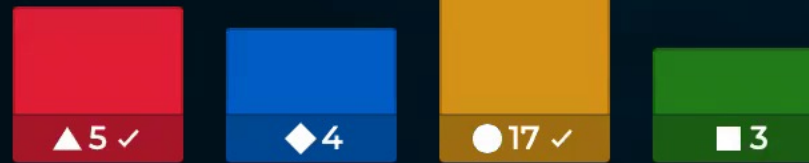
Pelin Genc

Pelin Genc

VJ

VJ

what are suitable baselines to subtract from the return to reduce the variance of the estimation of the p gradient?



▲ reward-to-go ✓

◆ return of the trajectory ✗

● state-value function ✓

■ discount factor ✗

Christopher Mutschler

Tami

Tami

Georg Rabenstein

FAU084414

FAU084414

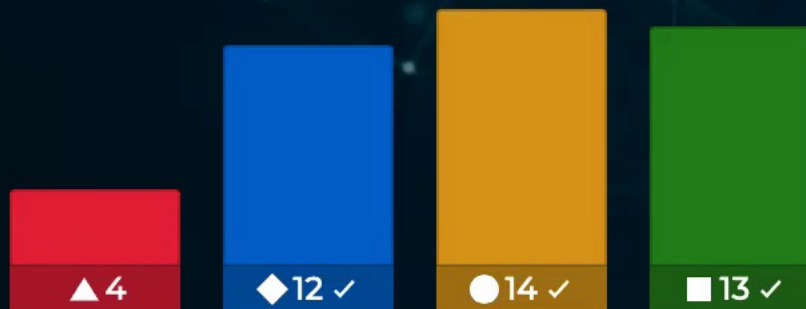
Pelin Genc

Pelin Genc

VJ

VJ

The policy gradient ...



Medien anzeigen

▲ might have a large bias. ✗

◆ usually has a high variance. ✓

● can be estimated through sampling. ✓

■ must be re-estimated when the policy is updated. ✓



Christopher Mutschler

Tami

Tami



Georg Rabenstein

FAU084414

FAU084414

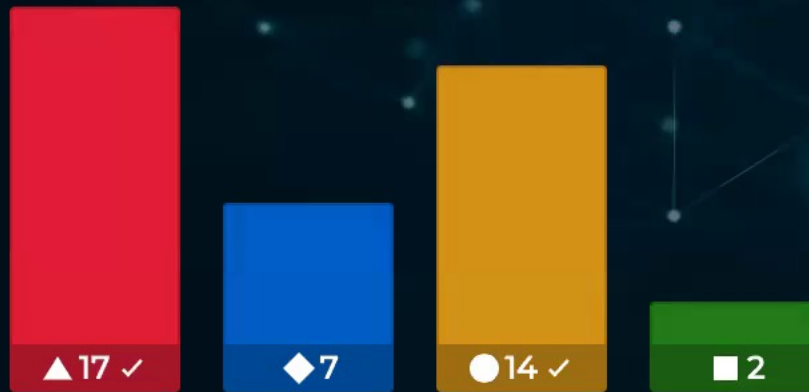
Pelin Genc

Pelin Genc

VJ

VJ

What is true for RL using policy gradients?



▲ trains stochastic policies in an on-policy way ✓

◆ uses a set of old policies to enable a more diverse sampling from the env ✗

● Exploration implicitly becomes less over the course of training ✓

■ likely find the global optimum ✗



Christopher Mutschler

Tami

Tami



Georg Rabenstein

FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

VJ

Sort from high (top) to low (bottom) variance



◆ REINFORCE with $G(\tau)$

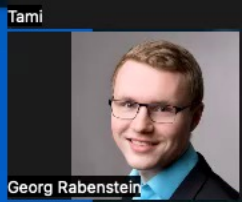
● REINFORCE (with reward-to-go)

▲ REINFORCE (with rewards-to-go & baselines)

■ Advantage Actor-Critic (with GAE and $\lambda > 0$) ✓



Christopher Mutschler



Georg Rabenstein

FAU084414

FAU084414

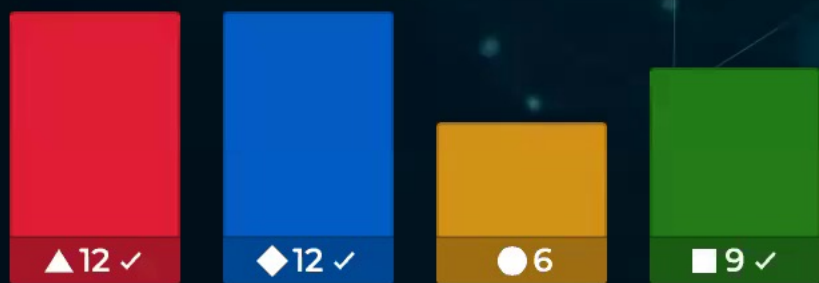
Pelin Genc

Pelin Genc

VJ

VJ

Using the Generalized Advantage Estimation (GAE) we may estimate the policy gradient with



▲ a low bias ✓

◆ a low variance ✓

● a low bias and low variance ✗

■ some bias and some variance ✓



Tami

Tami



Georg Rabenstein

FAU084414

FAU084414

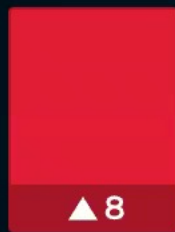
Pelin Genc

Pelin Genc

VJ

VJ

What is the better way to do exploration in policy-based RL?



▲ use a non-greedy action selection such as epsilon-greedy



◆ regularize the entropy of the action distribution



Tami

Tami



Georg Rabenstein

FAU084414

FAU084414

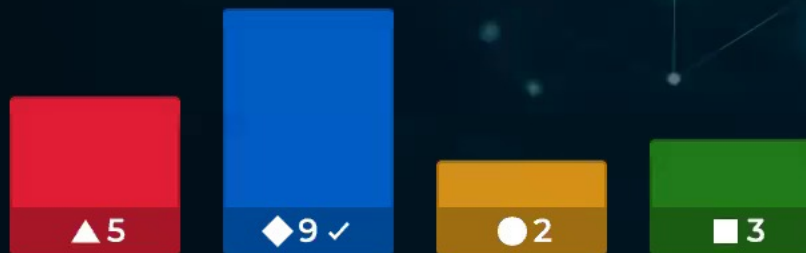
Pelin Genc

Pelin Genc

VJ

VJ

What is the main motivation of TRPO?



▲ Reduce variance of the policy gradient



◆ Not move the policy to far between updates (in policy space)



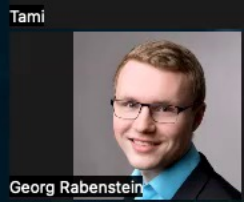
● Non-flat optimization landscapes should be tackled with first-order methods



■ Making use of transition dynamics to better estimate of the policy gradient



Tami



FAU084414

Pelin Genc

VJ

VJ

PPO-clip simply clips the updates to the policy, making it computationally much more efficient than TRPO.

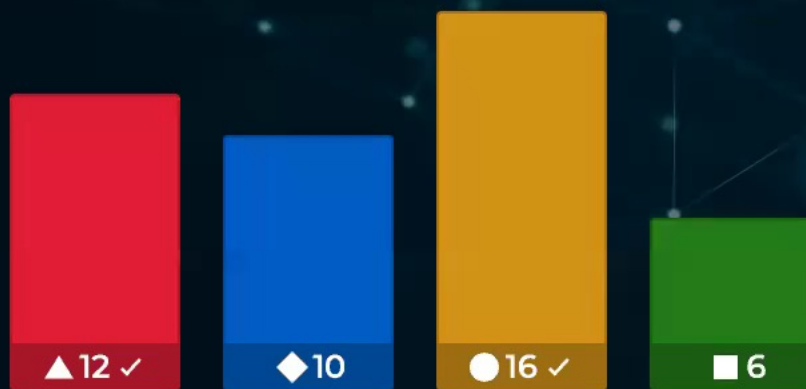


◆ Wahr ✓

▲ Falsch ✗

A vertical list of participant names and profile pictures. From top to bottom: Christopher Mutschler (with video feed), Tami, Tami, Georg Rabenstein (with profile picture), FAU084414, FAU084414, Pelin Genc, Pelin Genc, VJ, and VJ.

Deep Deterministic Policy Gradient



▲ trains off-policy ✓

◆ trains a stochastic policy ✗

● is an actor-critic algorithm ✓

■ is easy to tune ✗



Christopher Mutschler

Tami

Tami



Georg Rabenstein

FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

VJ