Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

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

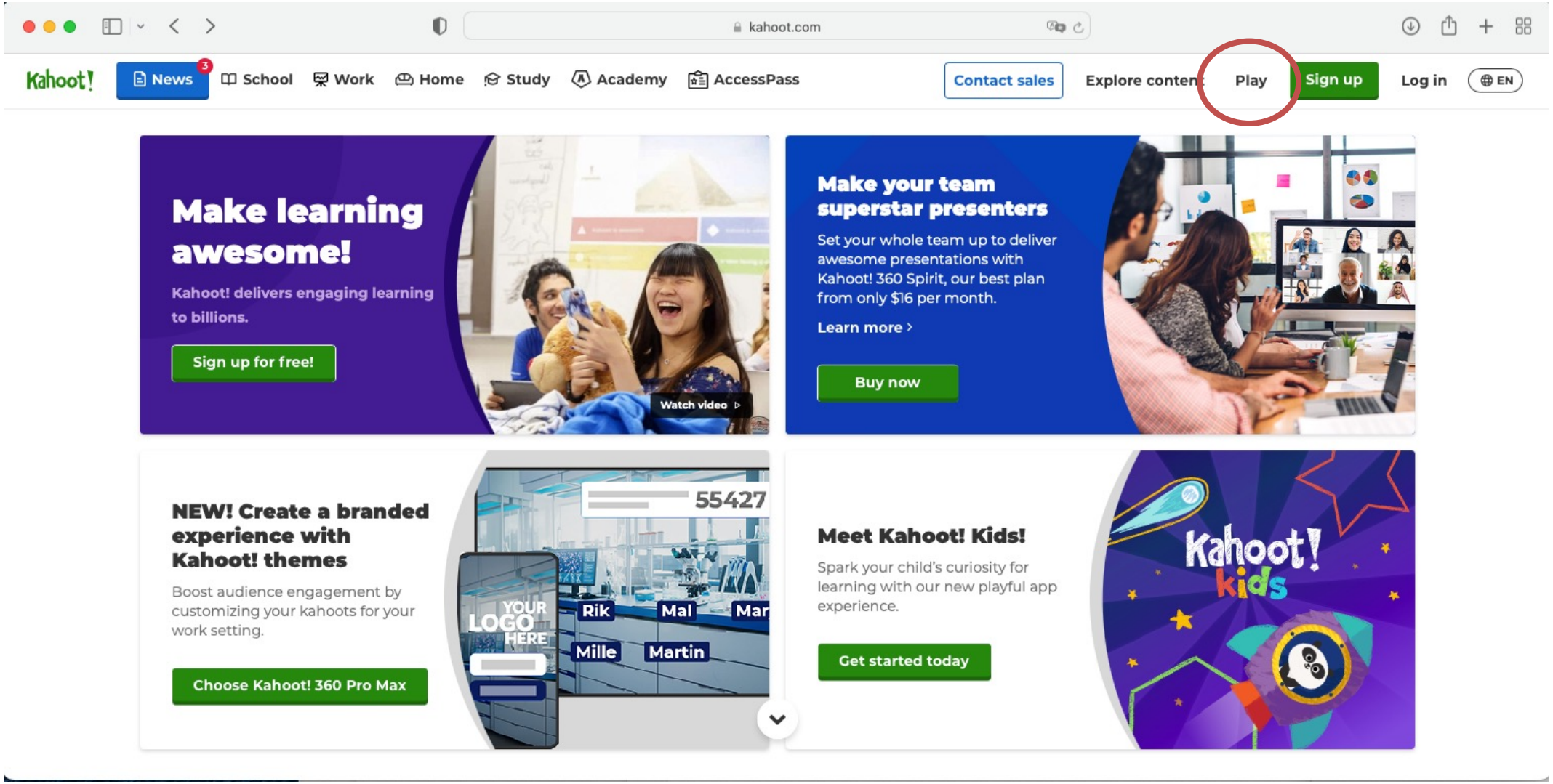
23.06.2022

Christopher Mutschler

Let‘s play Kahoot! again…

Kahoot!

Let‘s play Kahoot!



play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

Tami

Tami

FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

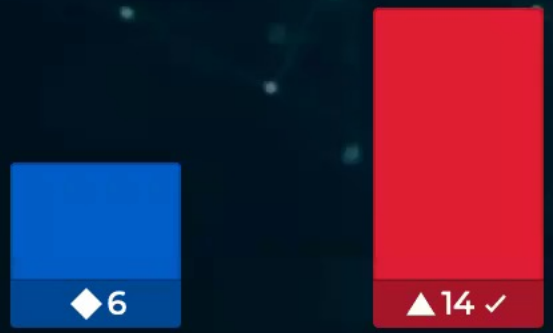
VJ

1/11

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

The estimation of the policy gradient requires transition probabilities!



Medien anzeigen

◆ Wahr ✖

▲ Falsch ✓

Tami

Tami

FAU084414

FAU084414

Pelin Genc

Pelin Genc

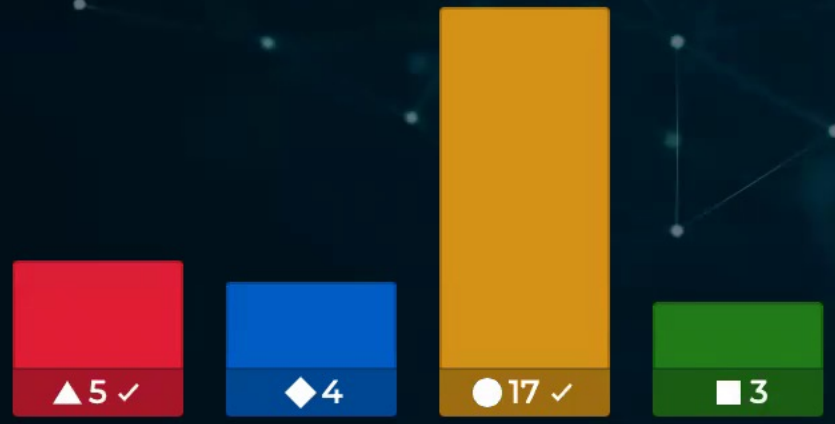
VJ

VJ

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

3/11

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

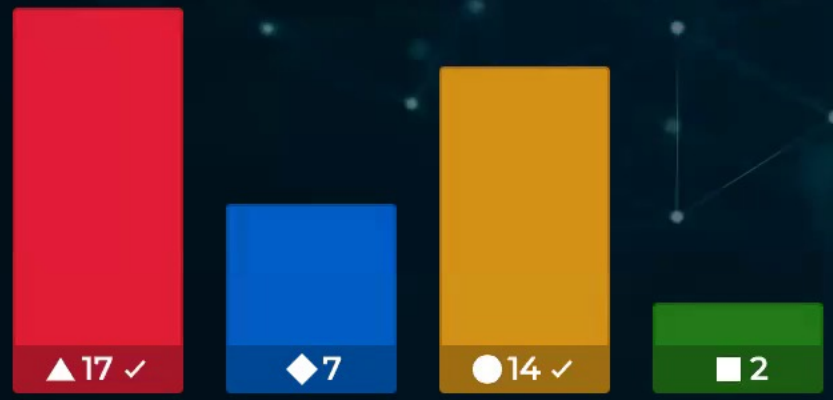
VJ

4/11

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

5/11

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

- ◆ REINFORCE with G(τ)
- ● REINFORCE (with reward-to-go)
- ▲ REINFORCE (with rewards-to-go & baselines)
- ■ Advantage Actor-Critic (with GAE and λ>0) ✓

Tami

Tami

FAU084414

FAU084414

Pelin Genc

Pelin Genc

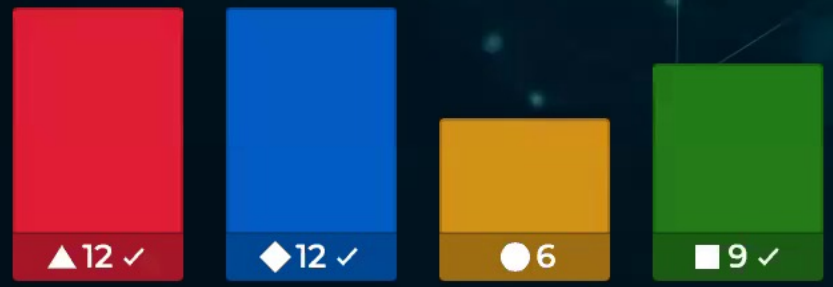
VJ

VJ

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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



▲ a low bias ✓

◆ a low variance ✓

● a low bias and low variance ✗

■ some bias and some variance ✓


Christopher Mutschler

Tami

Tami


Georg Rabenstein

FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

VJ

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

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

◆ regularlize the entropy of the action distribution ✔

Tami

Tami

FAU084414

FAU084414

Pelin Genc

Pelin Genc

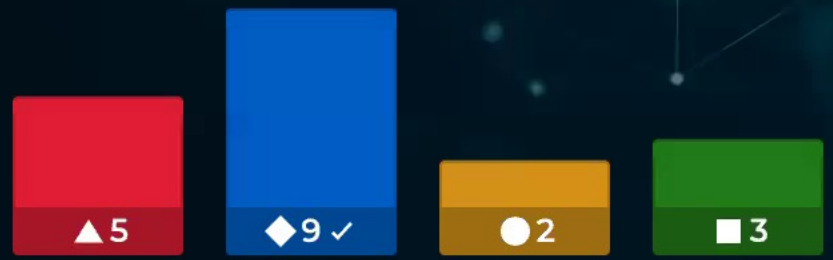
VJ

VJ

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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 ✖

Christopher Mutschler

Tami

Tami

Georg Rabenstein

FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

VJ

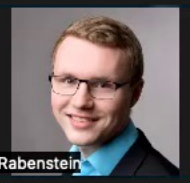
kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

Tami

Tami



FAU084414

FAU084414

Pelin Genc

Pelin Genc

VJ

VJ

◆18 ✓

▲2

◆ Wahr ✓

▲ Falsch ✗

10/11

kahoot.it Spiel-PIN: 8955279

play.kahoot.it/v2/gameblock?quizId=b131697c-1091-4a7e-8fb1-0961dc0a5ab0

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

kahoot.it Spiel-PIN: 8955279