UNIVERSITÄT PADERBORN

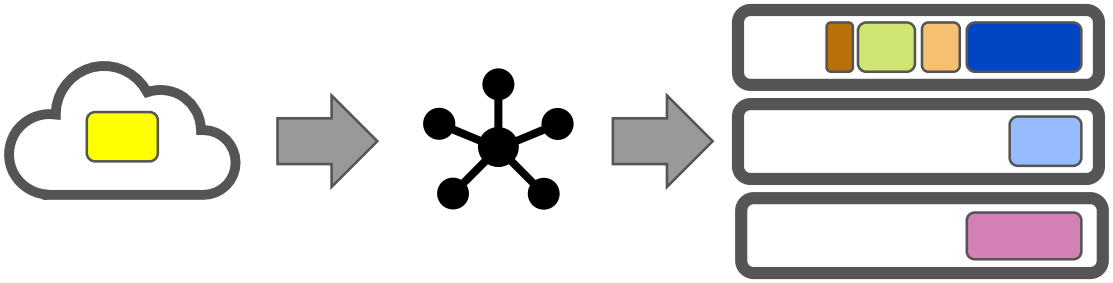
COMPUTER NETWORKS GROUP

VARIANCE REDUCTION FOR REINFORCEMENT LEARNING IN INPUT-DRIVEN ENVIRONMENTS

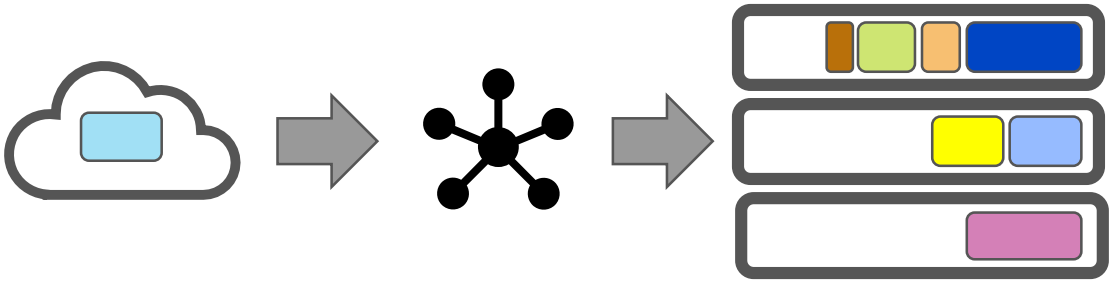
AICON PROJECT GROUP

Stefan Werner

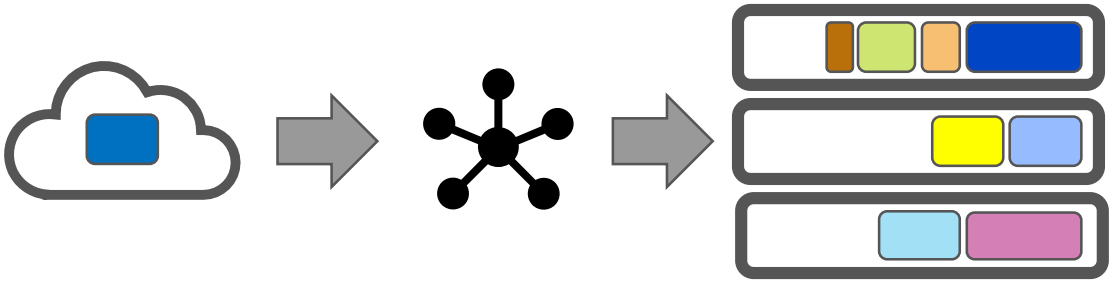
Foundations



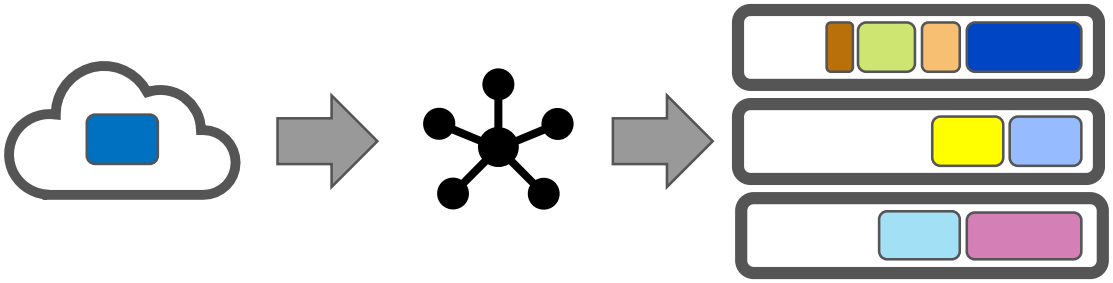
Foundations



Foundations



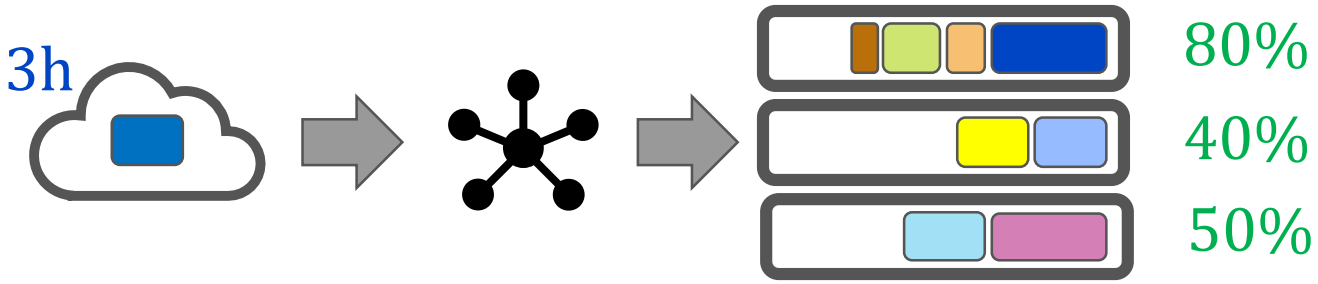
Foundations



MDP

- *Trajectories :* $\tau = (s_1, a_1, r_1, s_2, a_2, r_2, s_3, \dots, s_T)$

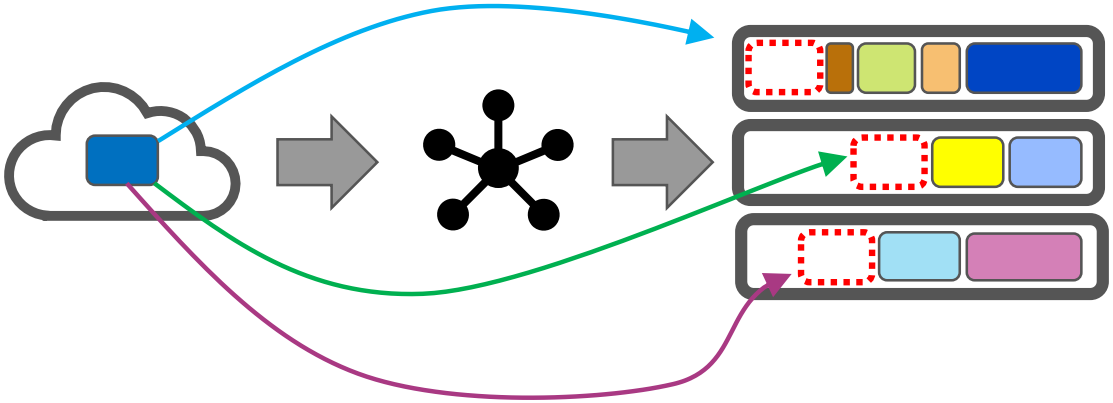
Foundations



MDP

- *Trajectories :* $\tau = (s_1, a_1, r_1, s_2, a_2, r_2, s_3, \dots, s_T)$
  - $s_t \equiv$ *server load and job size*

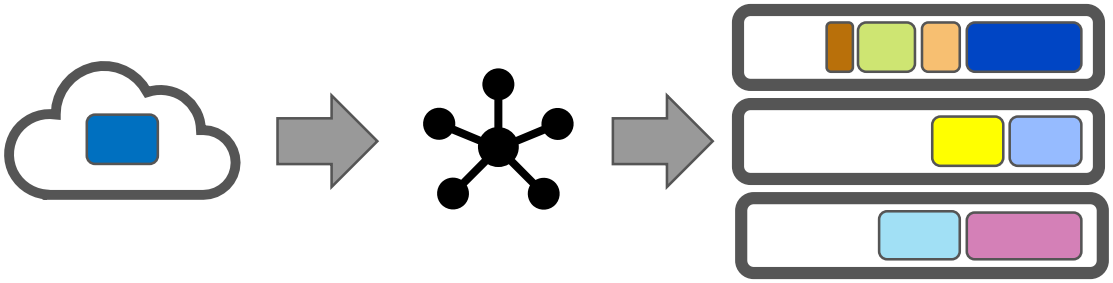
Foundations



MDP

- *Trajectories :* $\tau = (s_1, a_1, r_1, s_2, a_2, r_2, s_3, \dots, s_T)$
  - $s_t \equiv$ *server load and job size*
  - $a_t \equiv$ *schedule job to server* $i \in \{1, 2, 3\}$

Foundations

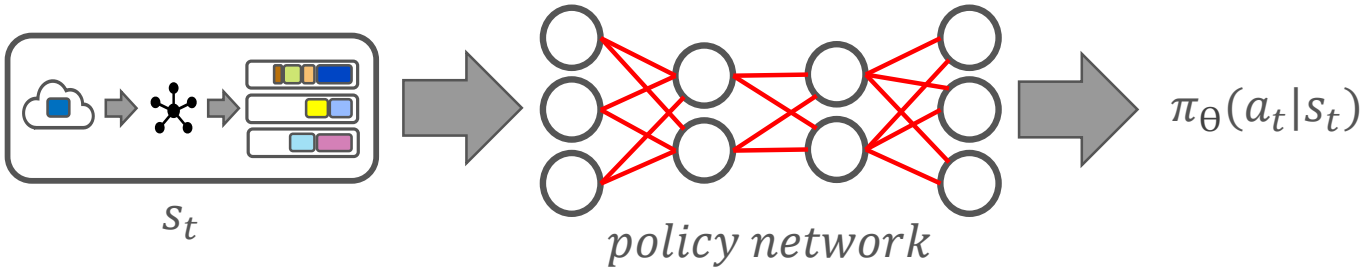


MDP

- *Trajectories :* $\tau = (s_1, a_1, r_1, s_2, a_2, r_2, s_3, \dots, s_T)$
  - $s_t \equiv$ *server load and job size*
  - $a_t \equiv$ *schedule job to server* $i \in \{1, 2, 3\}$
  - $r_t \equiv (-1) \cdot$ *average job completion time*

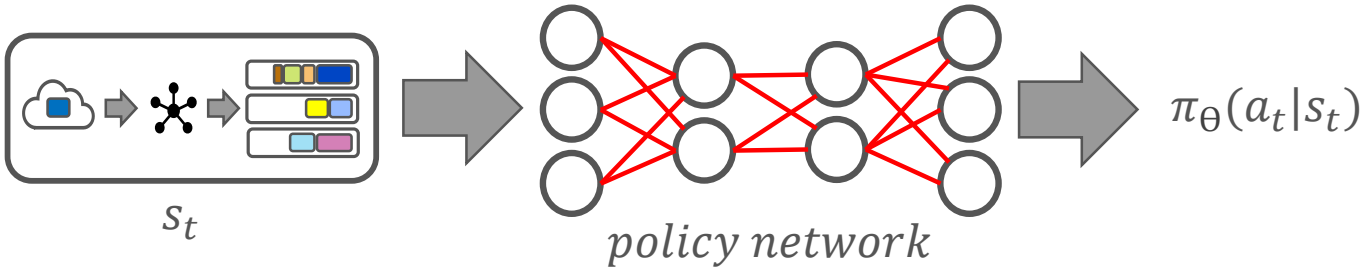
Policy Gradient Methods

- *Acting policy* π_θ *with parameterization* θ

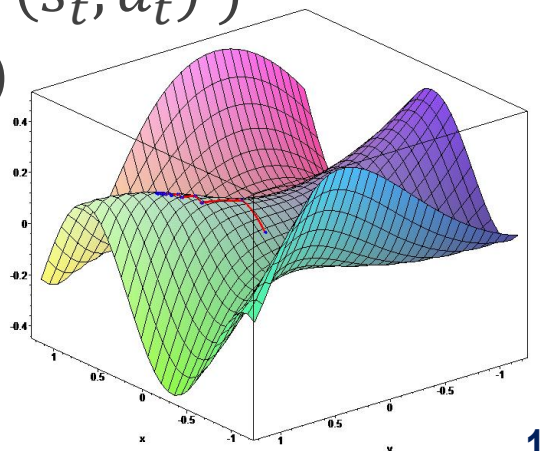


Policy Gradient Methods

- *Acting policy* π_θ *with parameterization* θ



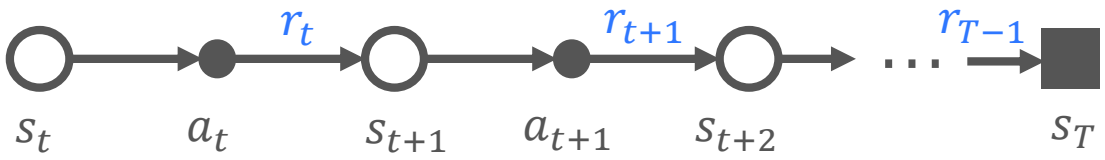
- *Expected return*: $J(\theta) = \mathrm{E}_{\tau\sim p_\theta(\tau)}\left(\sum_t r(s_t, a_t)\right)$
- $\nabla_\theta J(\theta) = \mathrm{E}_\pi\left(Q^\pi(s,a) \cdot \nabla_\theta ln\, \pi_\theta\,(a|s)\right)$



Source: [2]

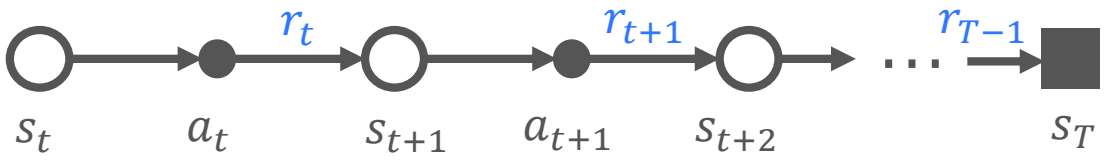
REINFORCE

- *Updates:* $\theta_{t+1} = \theta_t + \alpha \cdot G_t \cdot \nabla_{\theta_t} \ln \pi_\theta (a_t|s_t)$
- $G_t = r_t + \gamma \cdot r_{t+1} + \cdots \gamma^{T-1-t} r_{T-1}$



REINFORCE

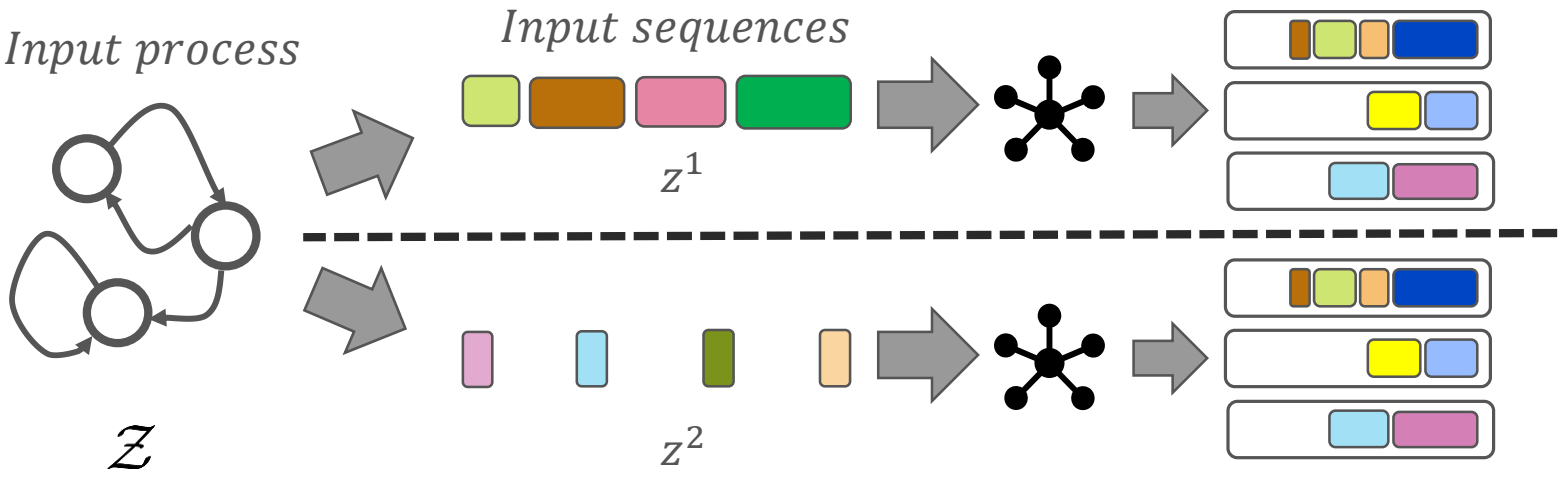
- *Updates:* $\theta_{t+1} = \theta_t + \alpha \cdot G_t \cdot \nabla_{\theta_t} \ln \pi_\theta(a_t|s_t)$
- $G_t = r_t + \gamma \cdot r_{t+1} + \cdots \gamma^{T-1-t} r_{T-1}$



REINFORCE with Baseline

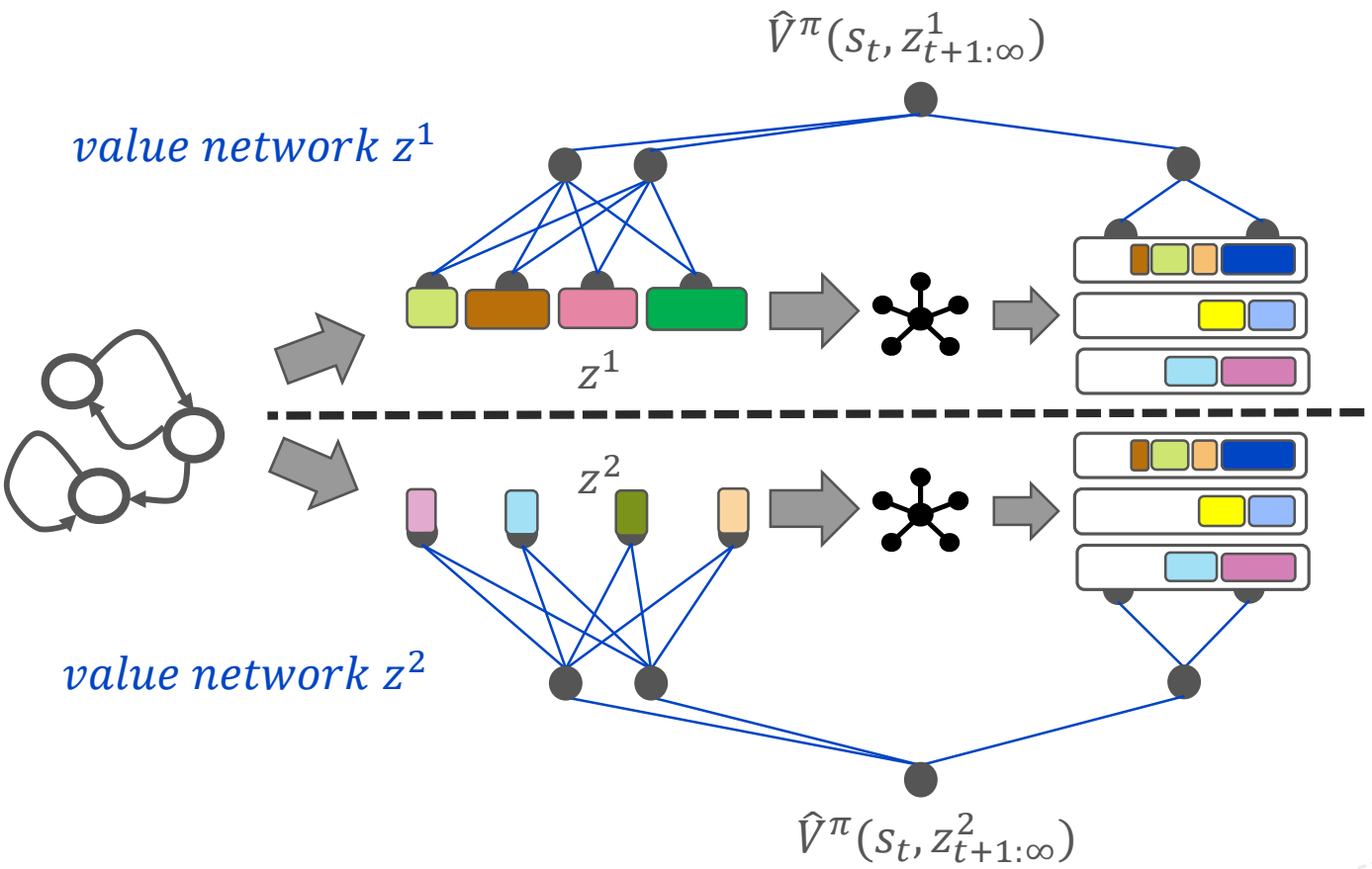
- $\theta_{t+1} = \theta_t + \alpha \cdot (G_t - \hat{V}^\pi(s_t)) \cdot \nabla_{\theta_t} \ln \pi_\theta(a_t|s_t)$

Input-Driven Environments

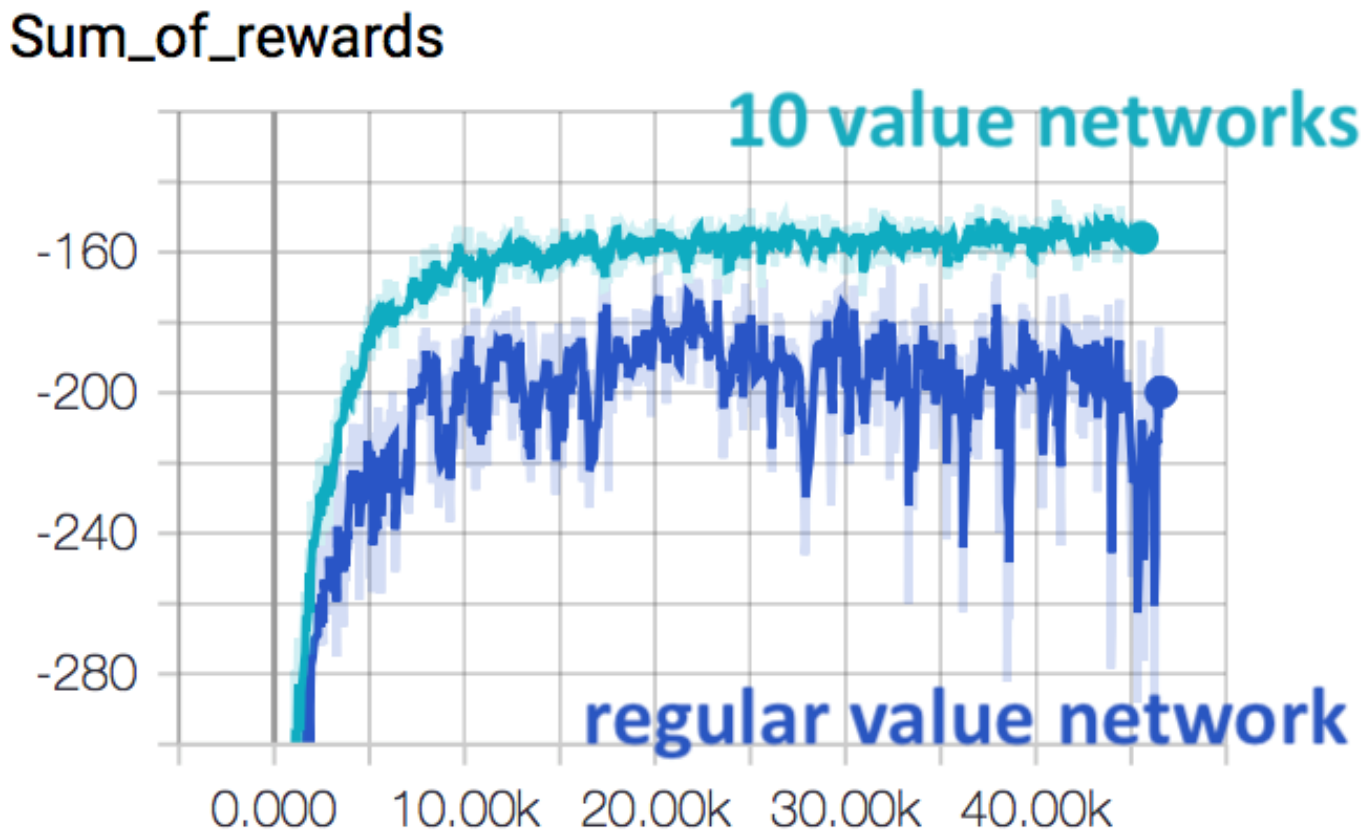


- *Rewards may be the result of favorable inputs*
- $\hat{V}^{\pi}(s_t, z^1_{t+1:\infty}) \neq \hat{V}^{\pi}(s_t, z^2_{t+1:\infty})$
- $\hat{V}^{\pi}(s_t)$ *doest not account for input differences*

Multi-Value Network Baselines



Experimental Results



Source: [3]

References

[1] H. Mao, S. B. Venkatakrishnan, M.Schwarzkopf, and M. Alizadeh. “Vari-ance reduction for reinforcement learningin input-driven environments”. In:arXivpreprint arXiv:1807.02264 (2018).

[2] Image source: https://upload.wikimedia.org/wikipedia/commons/6/68/Gradient_ascent_%28surface%29.png

[3] Image source: https://github.com/hongzimao/input_driven_rl_example/blob/master/figures/training.png