Využitie neurónových sietí pri simulovanej robotike

Bc. Martin Kubovčík

Vedúci: prof. RNDr. Jiří Pospíchal, DrSc.

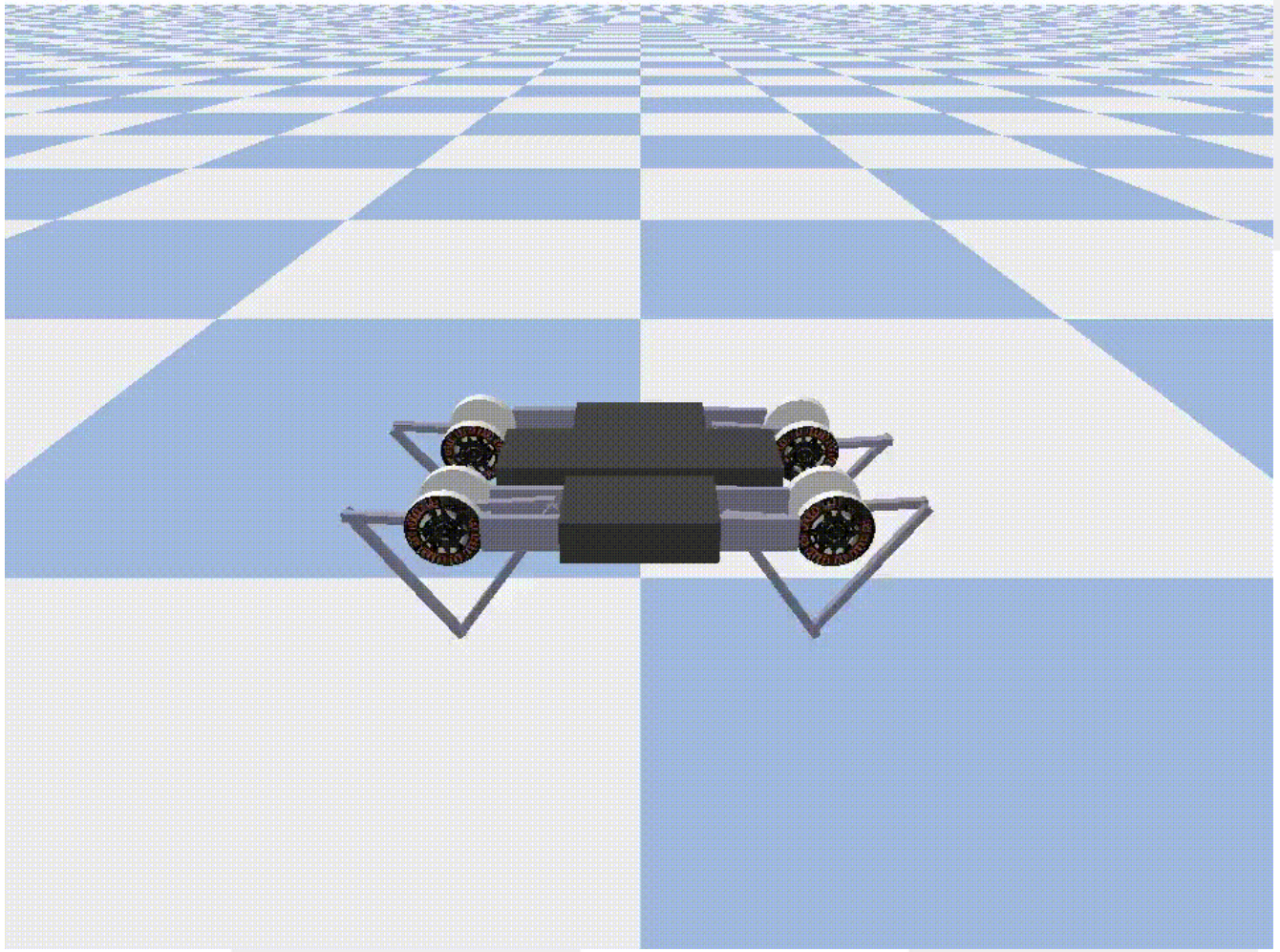
Obsah

- Prostredie
- Soft Actor-Critic
- DeepMind Reverb
- Vylepšenia
- Imitačné učenie
- Sim-to-Real

Prostredie

- OpenAI Gym
- PyBullet
- DeepMind Control Suite

Prostredie MinitaurBulletEnv-v0

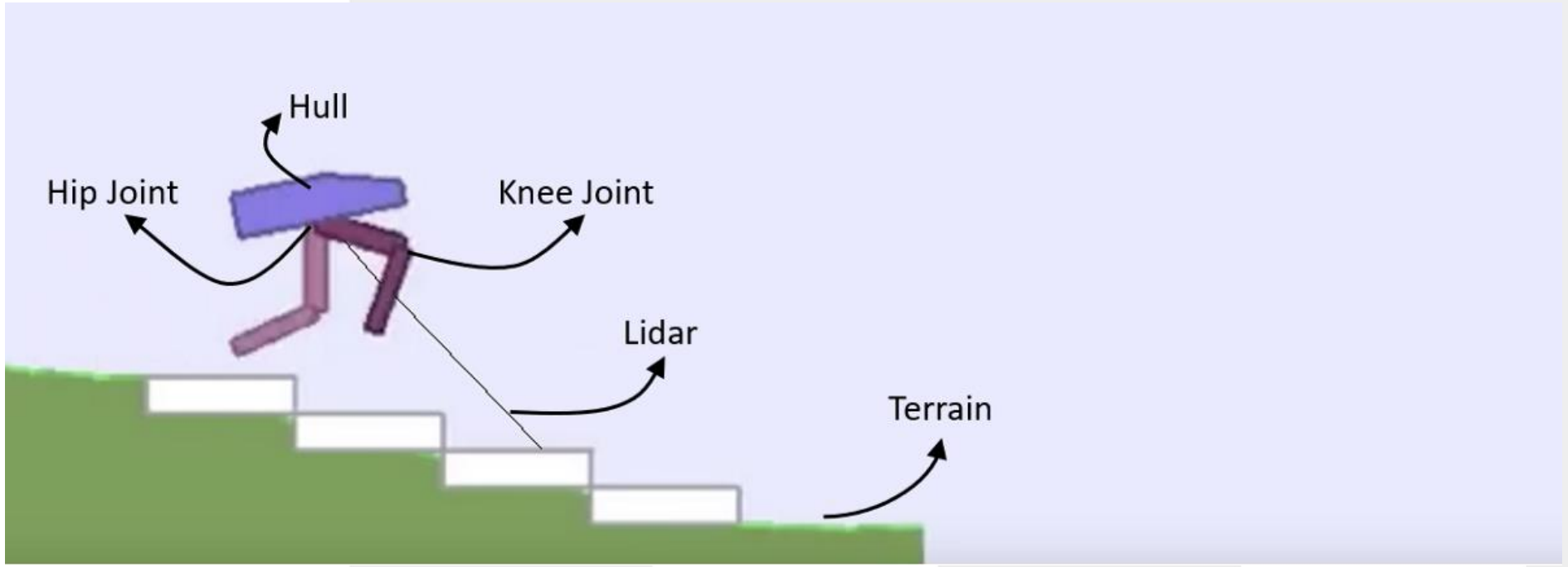


Prostredie MinitaurBulletEnv-v0

- 28 príznakov v stavovom priestore
- 8 akčných členov
- Rozsah akcií [-1.0, 1.0]
- **Vstupy:** uhly motorov, rýchlosti rotácie motorov, krútiace momenty motorov, pozícia robota, orientácia robota
- **Odmena:** pohyb vpred, trest za nepriamočiaru chôdzu, trest za trasenie sa, trest za používanie motorov

ÚCMFPV

Prostredie BipedalWalkerHardcore-v3

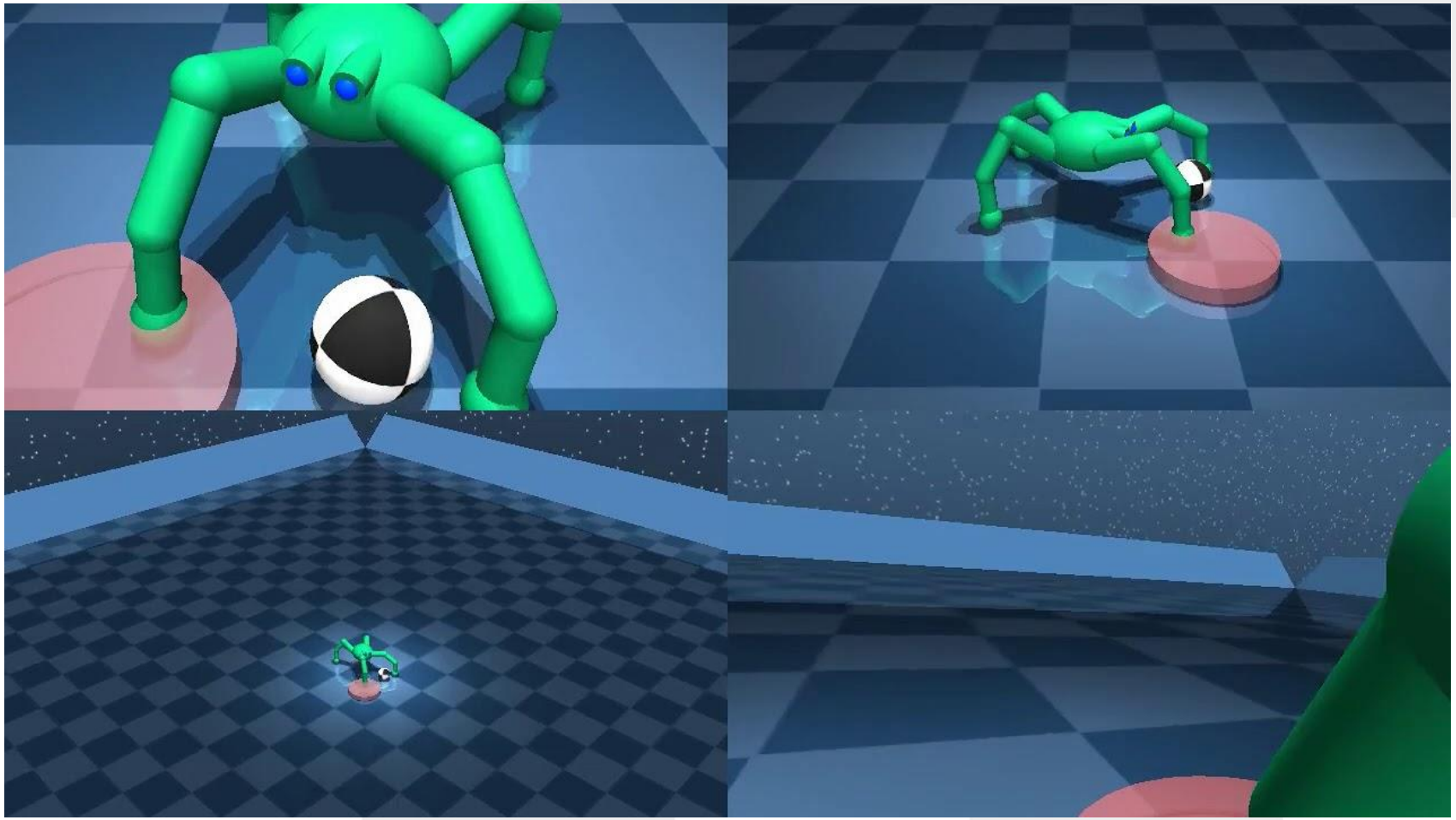


Zdroj: https://pylessons.com/BipedalWalker-v3-PPO#

Prostredie BipedalWalkerHardcore-v3

- 24 príznakov v stavovom priestore
- 4 akčné členy
- Rozsah akcií [-1.0, 1.0]
- **Vstupy:** uhol trupu, uhlová rýchlosť trupu, rýchlosť robota, uhly kĺbov, rýchlosti rotácie kĺbov, kontakty nôh so zemou, LIDAR
- **Odmena:** pohyb vpred (normalizovaný na 300 bodov ak dosiahne cieľ), držanie hlavy rovno, trest za používanie motorov, -100 za pád na zem

Prostredie Quadruped



Zdroj: https://www.deepmind.com/publications/dm-control-software-and-tasks-for-continuous-control

Prostredie Quadruped

- 90 príznakov v stavovom priestore
- 12 akčných členov
- Rozsah akcií [-1.0, 1.0]
- **Vstupy:** uhly kĺbov, rýchlosti rotácie kĺbov, rýchlosť robota, senzor IMU (GYRO+ACC), rýchlosť rotácie, rýchlosť lopty relatívna k rýchlosti robota, pozícia lopty relatívna k pozícii robota, cieľová pozícia pre loptu relatívna k pozícii robota
- **Odmena:** vzdialenosť od lopty, vzdialenosť lopty od cieľa, trup vzpriamený v osi z (+1 ak je v tolerancii, inak <1)

Q-Learning

- Off-policy algoritmus
- Markov decision process
- s_0,a_0,r_1,s_1,a
- $Q(s, a) = r + \gamma * max_{a'} Q(s', a')$

kvalita Q vykonanej akcie a_t

discount factor

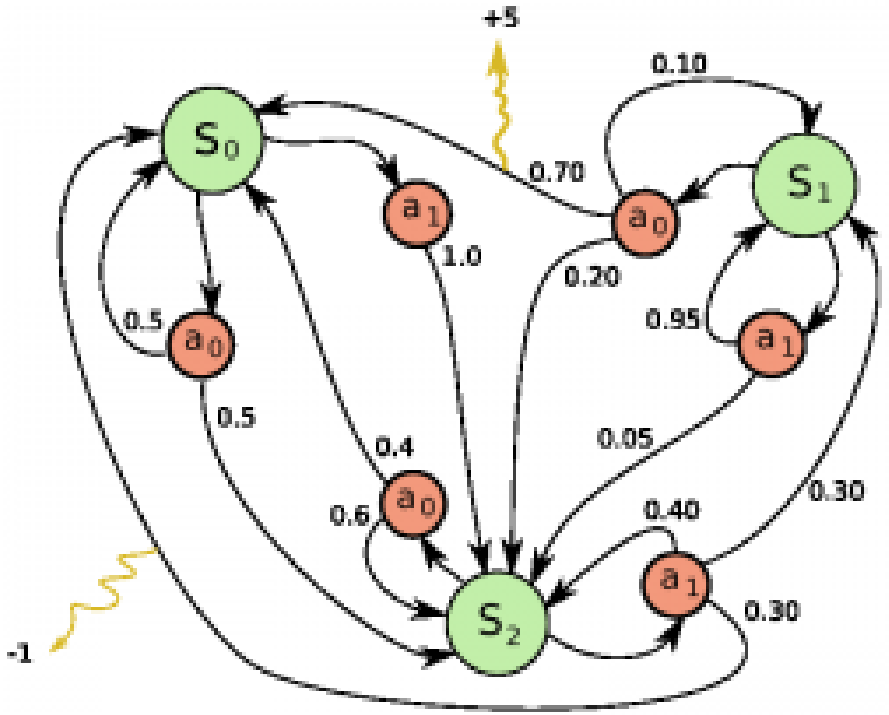
maximálna kvalita Q akcie a_{t+1} v stave s_{t+1}

na za vykonanú akciu

vzťah

- Čím viac ideme do budúcna tým nižšiu váhu

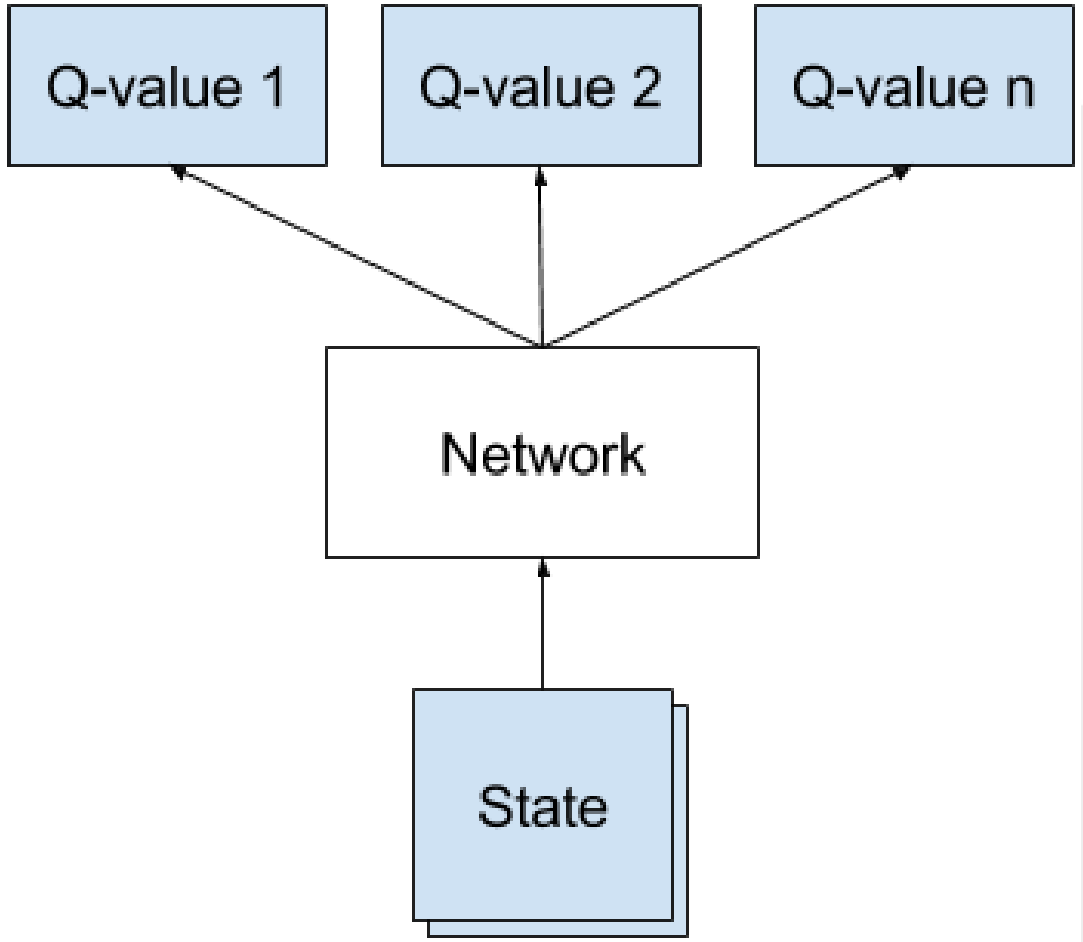
  γ tomu prikladáme



Zdroj: https://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/

Deep Q Network

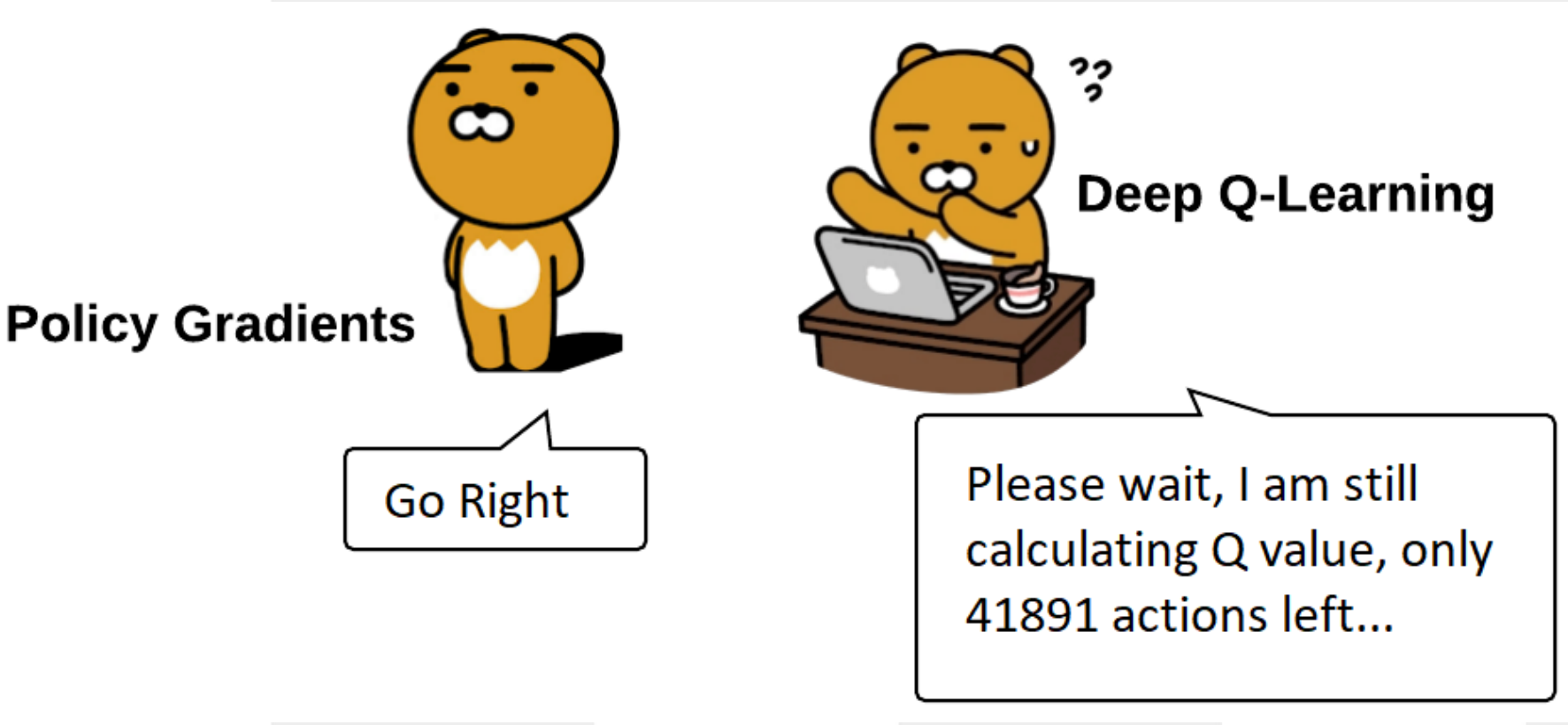
- Univerzálny aproximátor nelineárnych funkcií
- Transformácia vysoko rozmerných dát (2D obraz) do Q-hodnôt
- Problémy s nestabilitou počas učenia (local minimum) – množstvo trikov – target network, experience replay buffer



Zdroj: https://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/

ÚCMFPV

Soft Actor-Critic



Zdroj: https://pylessons.com/Beyond-DQN#

Soft Actor-Critic

- Maximalizácia entropie
- Maximalizácia Q funkcie
- Dvojica nezávislých modelov Actor (policy) a Critic (Q value)

entropia náhodnej premennej X

vyjadrenie entropie pre Gaussovu distribúciu

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) = -\mathbb{E}\left[log\mathcal{N}(\mu, \sigma^2)\right]$$

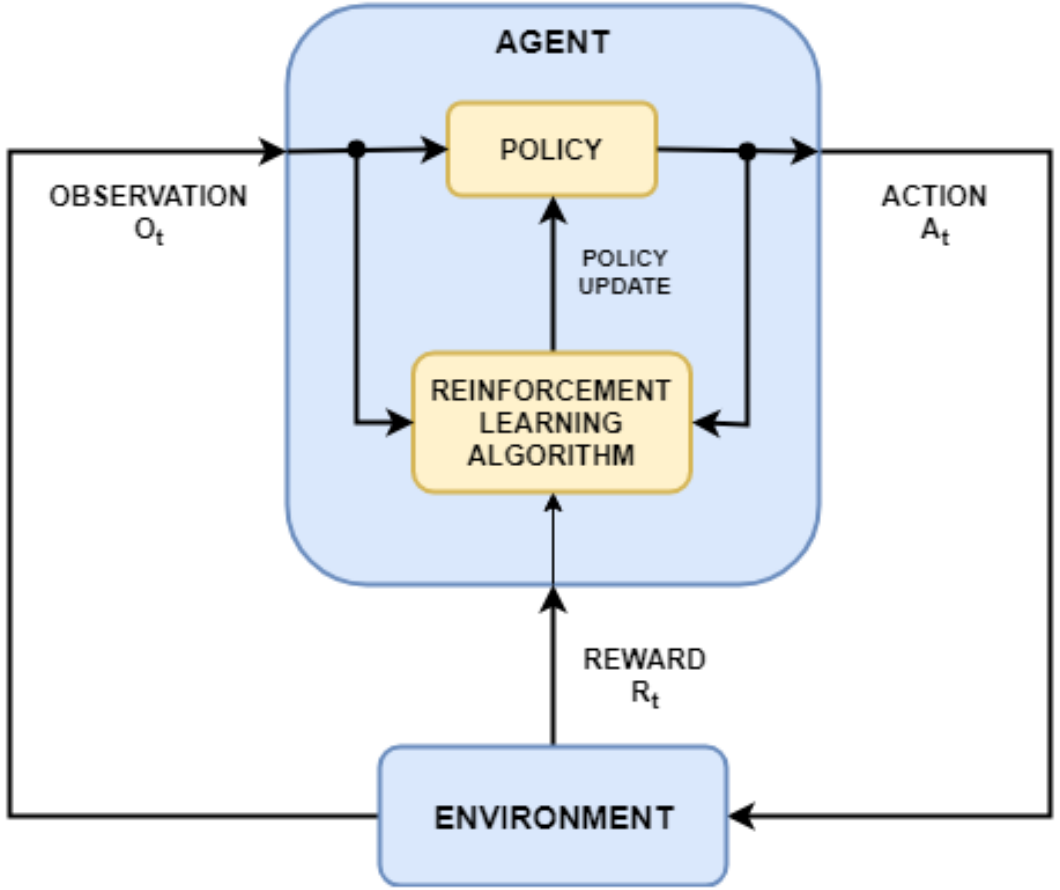
pravdepodobnosť vygenerovania konkrétnej hodnoty x_i

Soft Actor-Critic

- Gaussova distribúcia – predikcia priemerov a smerodajných odchýlok
- Efektívnejšie prehľadávanie na rozdiel od Twin Delayed DDPG (TD3) alebo Deep Deterministic Policy Gradient (DDPG)
- Menej citlivé na ladenie hyperparametrov
- Adaptívny podiel entropie na chybových funkciách

Zdroj: https://hbr.org/2007/11/a-leaders-framework-for-decision-making

Soft Actor-Critic



Zdroj: https://medium.com/intro-to-artificial-intelligence/soft-actor-critic-reinforcement-learning-algorithm-1934a2c3087f

Soft Actor-Critic

- Odmeny sú v rozsahu [-1.0, 1.0] – funkcia tanh
- Chybové funkcie:

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D, \epsilon_t \sim \mathcal{N}} \left[\alpha \log \pi_\phi\left(f_\phi(\epsilon_t, s_t) \middle| s_t\right) - Q_\theta(s_t, f_\phi(\epsilon_t, s_t)) \right]$$

entropia distribúcie

podiel entropie na chybovej fu

predikovaná Q-hodnota

$$J_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim D} \left[\frac{1}{2} \left(Q_\theta(s_t, a_t) - \left(r(s_t, a_t) + \gamma \left(Q_\theta(s_{t+1}, a_{t+1}) - \alpha \log \pi_\phi\left(a_{t+1} \middle| s_{t+1}\right) \right) \right) \right)^2 \right]$$

entropia distribúcie

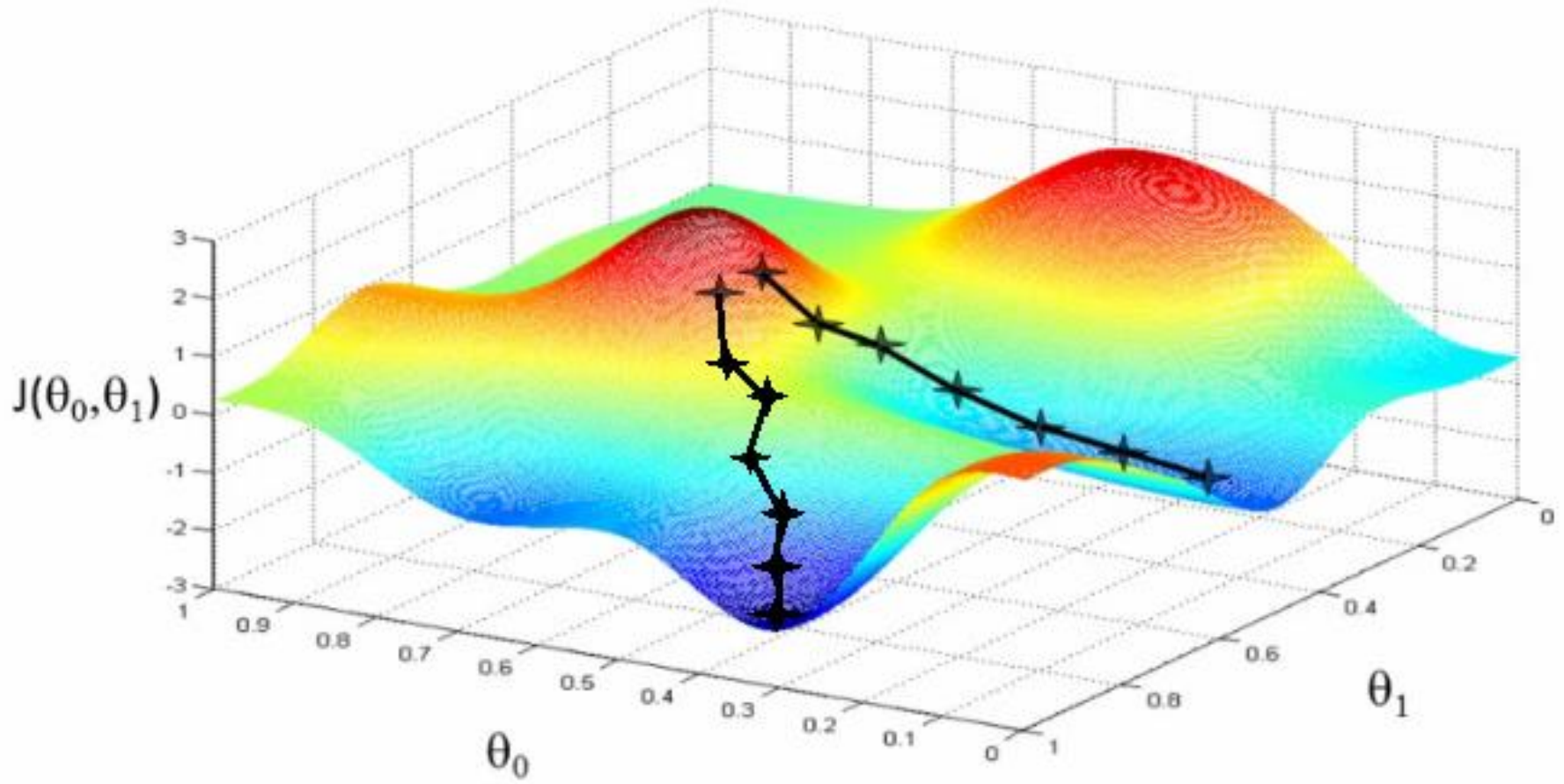
funkcia odmer

faktor

Q-hodnota pre nasledujúci stav

Q-hodnota pre aktuálny stav

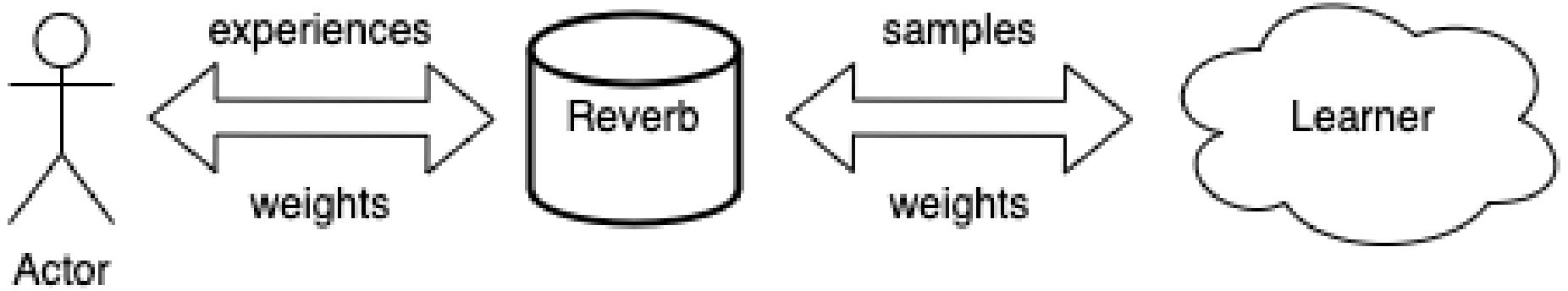
Soft Actor-Critic



Zdroj: https://medium.com/common-notes/gradient-ascent-e23738464a19

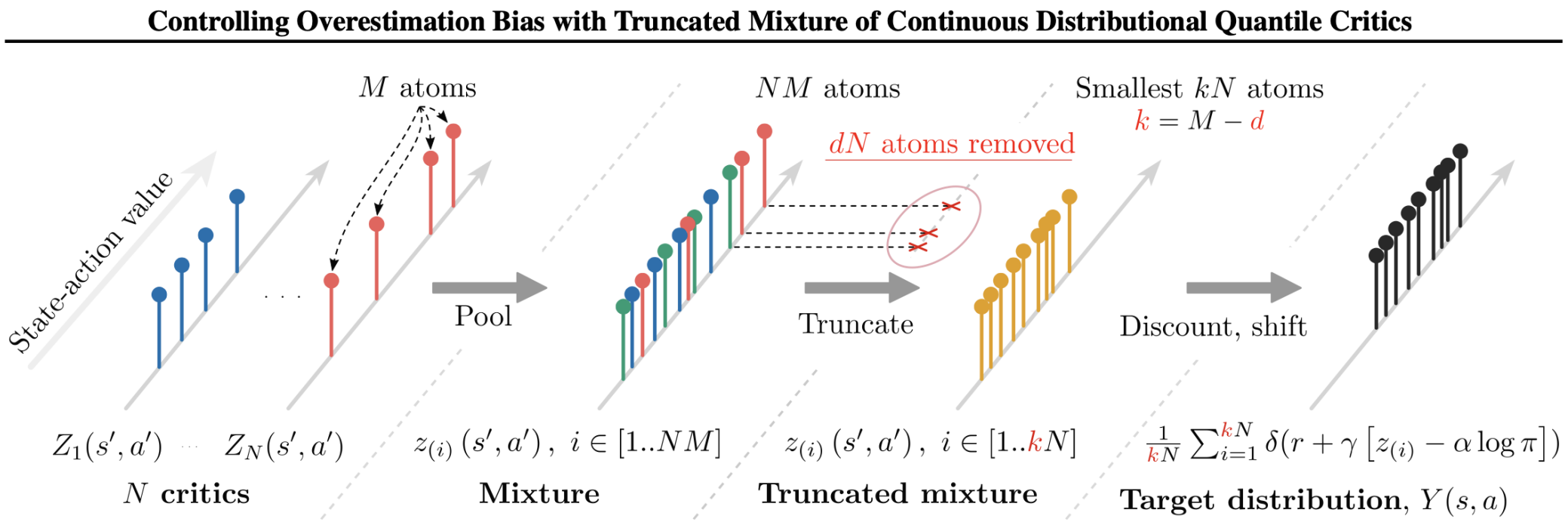
DeepMind Reverb

- Klient-server architektúra
- Actor – zbiera skúsenosti v hernom prostredí
- Learner – aktualizuje parametre neurónových sietí
- Databáza – distribúcia dát po počítačovej sieti (Experience Replay)



Vylepšenia

Controlling Overestimation Bias with Truncated Mixture of Continuous Distributional Quantile Critics



Zdroj: http://proceedings.mlr.press/v119/kuznetsov20a/kuznetsov20a.pdf

Vylepšenia
Generalized State-Dependent Exploration (gSDE)

- Parametre šumu θ_ε sú generované len každých n krokov počas epizódy
- Namiesto stavu s_t

$a_t = \mu(s_t; \theta_\mu) + \varepsilon(s_t; \theta_\varepsilon)$

deterministická akcia predikc

$\varepsilon(s; \theta_\varepsilon) = \theta_\varepsilon z_\mu(s)$

$\theta_\varepsilon \sim N(0, \sigma^2)$

náhodná premenná z Gaussovej distribúcie s nulovým priemerom a smerodajnou odchýlkou danou trénovanými parametrami

latentný priestor modelu

šumová zložka závisla od stavu

Vylepšenia
Generalized State-Dependent Exploration (gSDE)

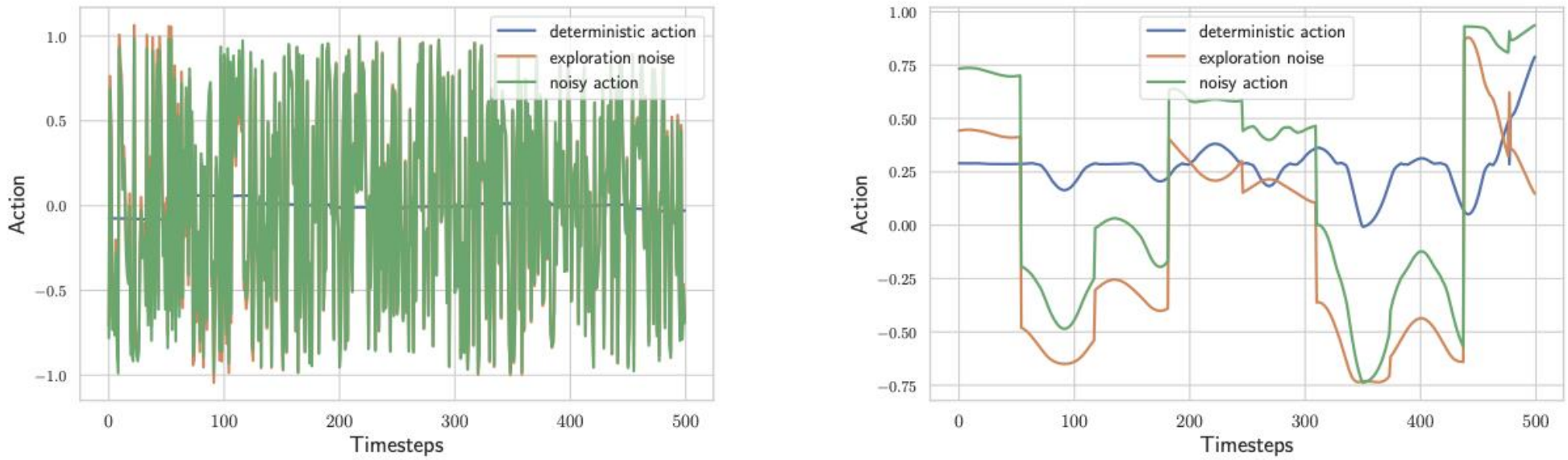


Figure 1: Left: unstructured exploration, as typically used in simulated RL. Right: *g*SDE provides smooth and consistent exploration.

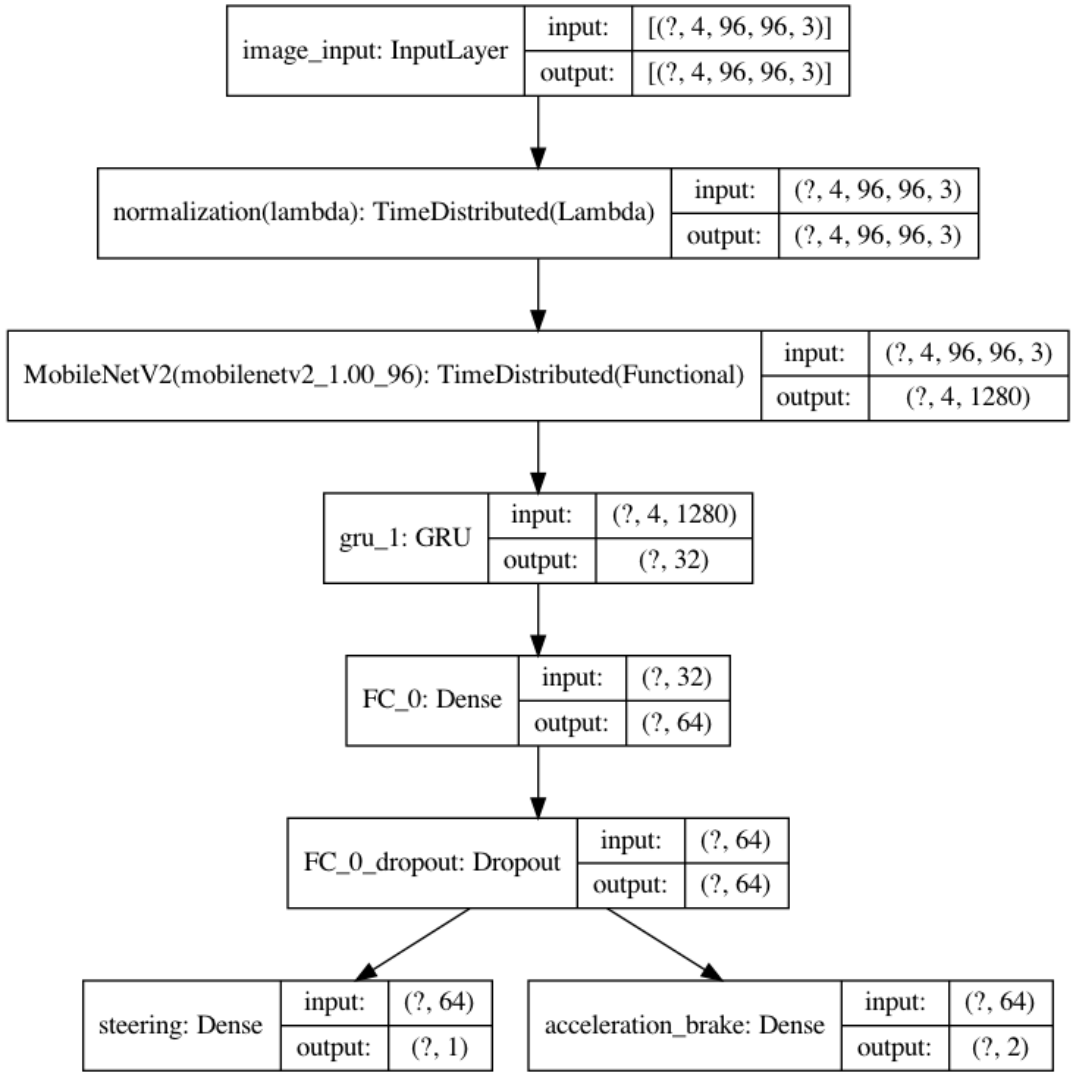
ÜCMFPV

Zdroj: https://proceedings.mlr.press/v164/raffin22a/raffin22a.pdf

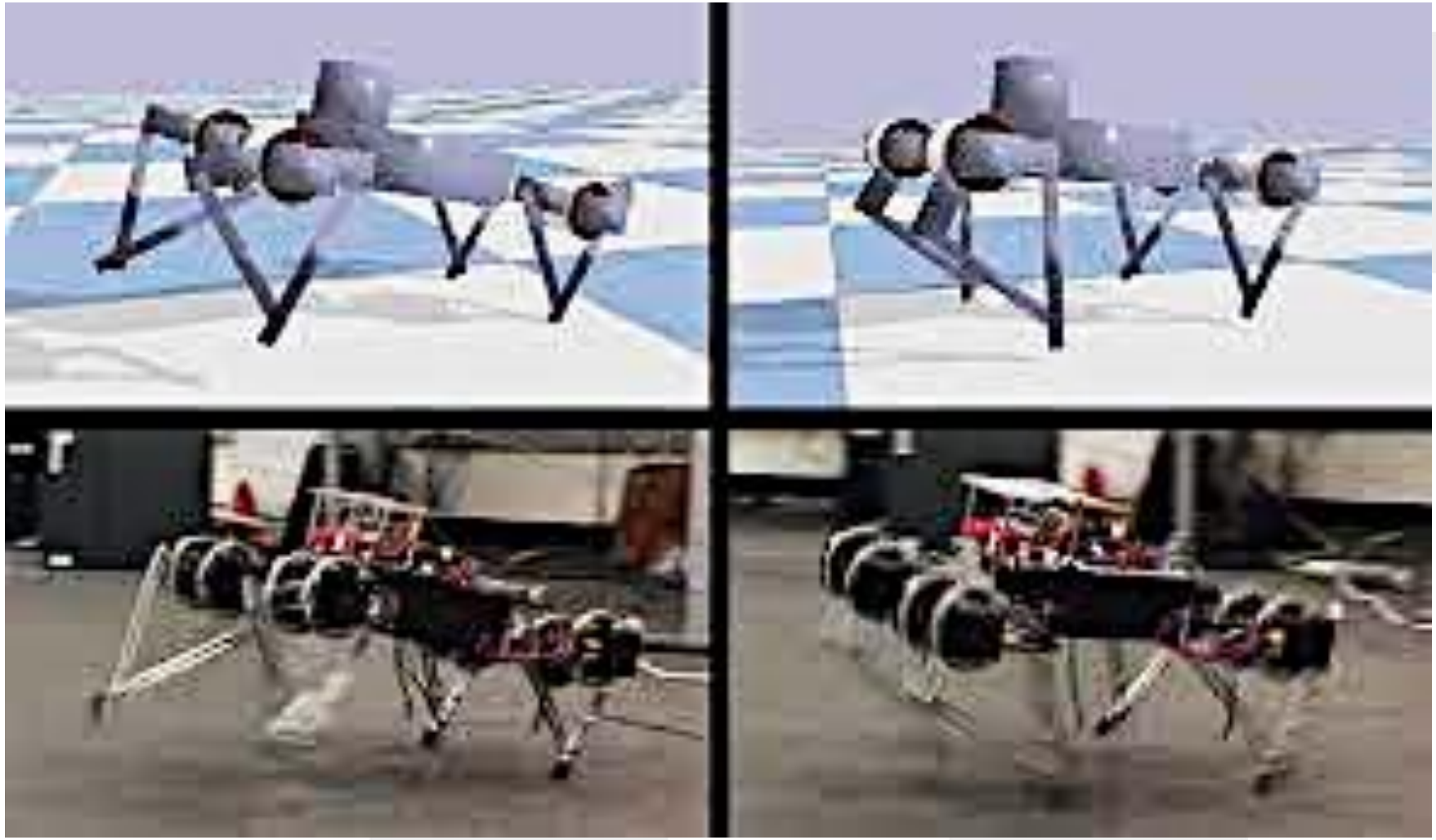
Imitačné učenie

Imitačné učenie

- Vstup: RGB obraz 96x96 pixelov, normalizované do rozsahu (-1, 1)
- Rekurentná jednotka GRU kvôli časovej sérii obrazov
- MobileNetV2
- Detekcia smeru pohybu auta voči okoliu
- Výstup: volant (-1, 1), plynový pedál (0, 1), brzdový pedál (0, 1)
- Využiteľnosť v reálnej robotike – kamera, GPS



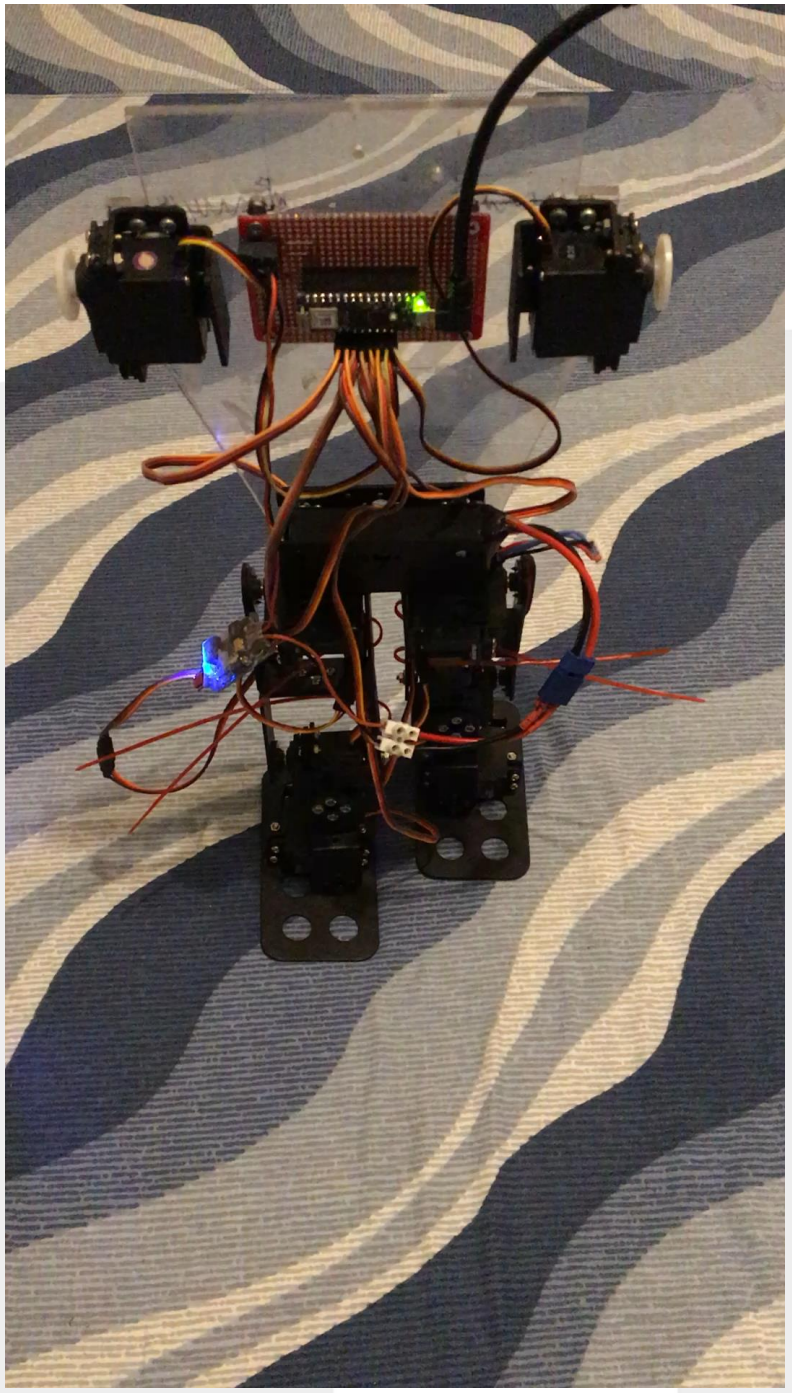
Sim-to-Real



Zdroj: http://www.roboticsproceedings.org/rss14/p10.pdf

Sim-to-Real

- Actor – predstavuje robota v simulovanom/reálnom svete
- Learner – server alebo cloud – trénuje robota
- Databáza – server alebo cloud – uchováva skúsenosti robota



Ďakujem za pozornosť

GitHub: https://github.com/markub3327
Email: kubovcik1@ucm.sk
Instagram: https://www.instagram.com/martin.kubovcik/
Twitter: https://twitter.com/markub3327

ÚCM FPV