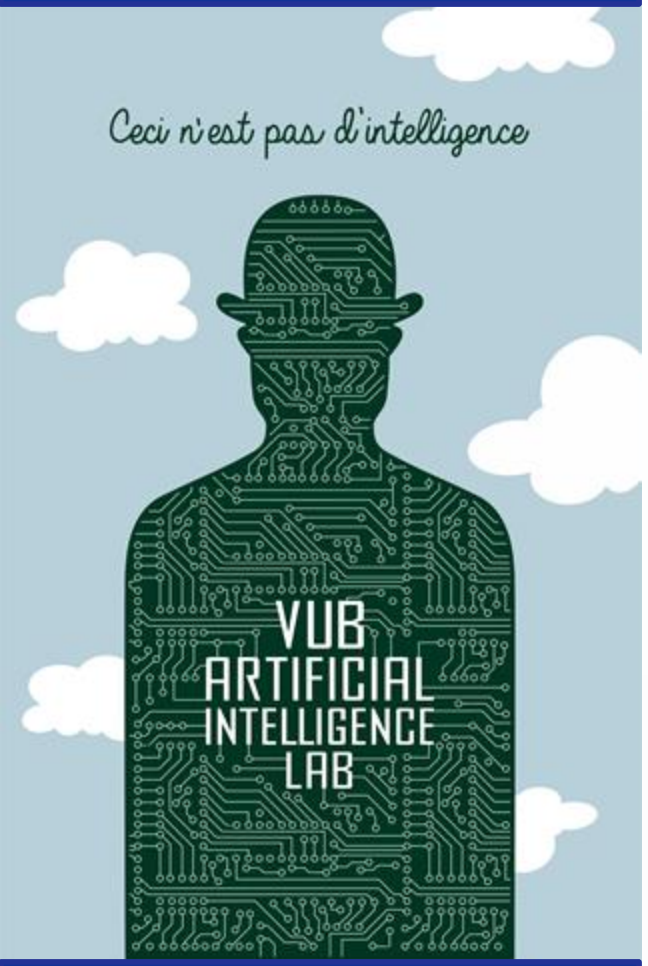


Continuous BDPI

Denis Steckelmacher
denis.steckelmacher@vub.be

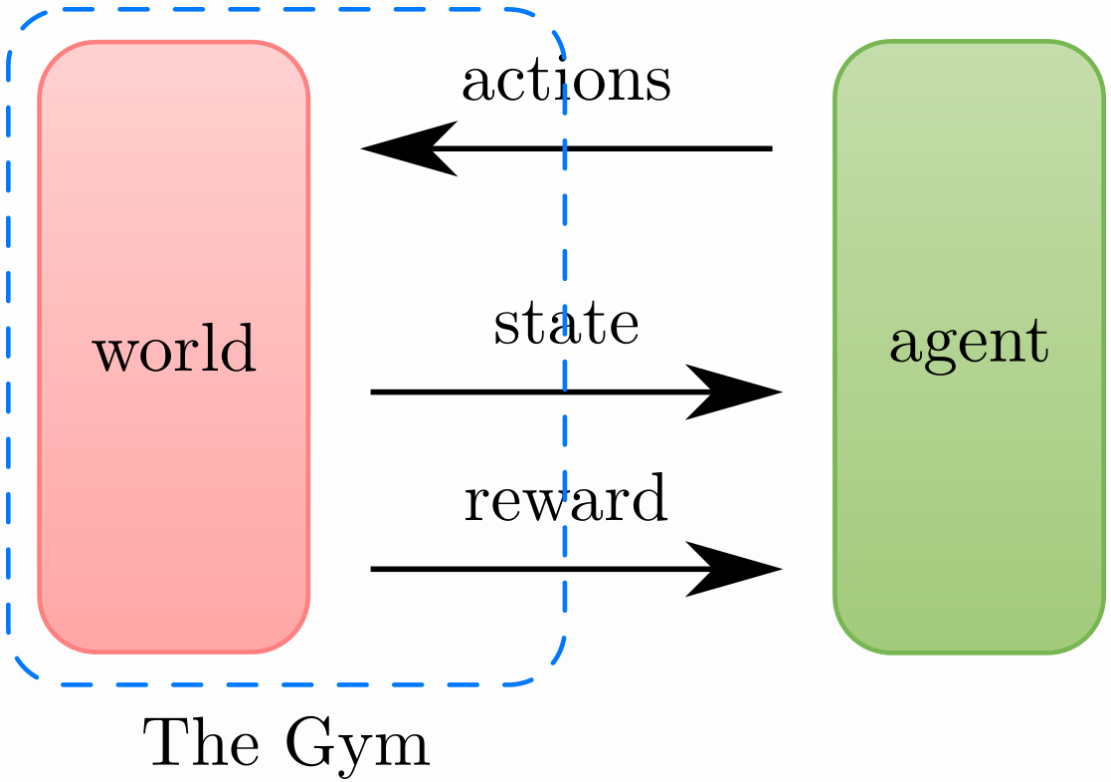
VUB Artificial Intelligence Lab

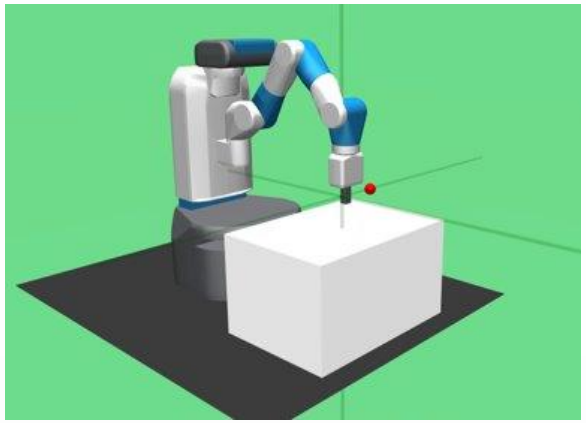
ai.vub.ac.be

@aibrussels

ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Reinforcement Learning





BDPI

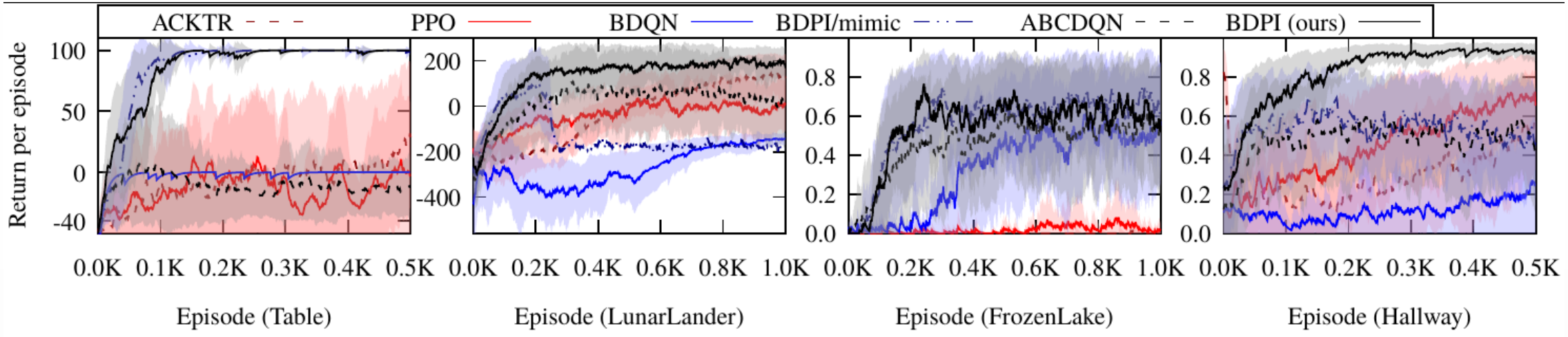
Critic update:

$$Q(s_t, a_t) \mathrel{+}= \alpha * (r_t + \gamma \, \mathbf{max_{a'} \, Q(s_{t+1}, a')} - Q(s_t, a_t))$$

Actor update:

$$\Pi \, (s_t) \leftarrow \mathbf{argmax_a \, Q_i(s_t, a)}$$

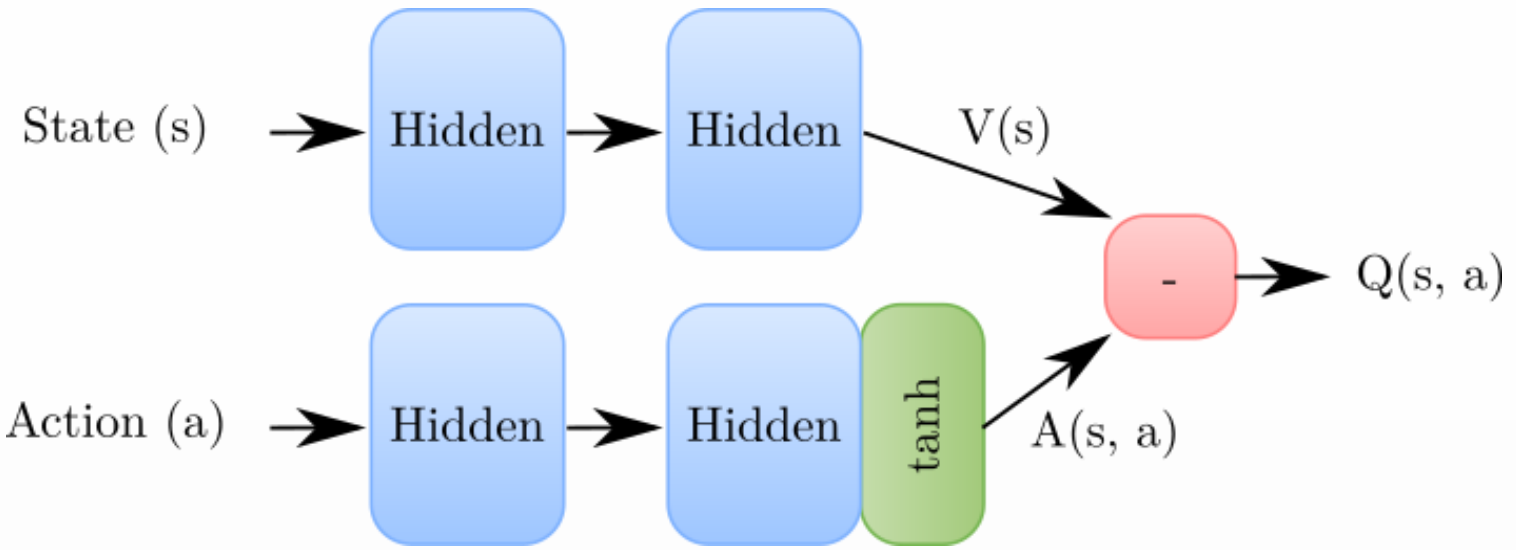
BDPI is promizing



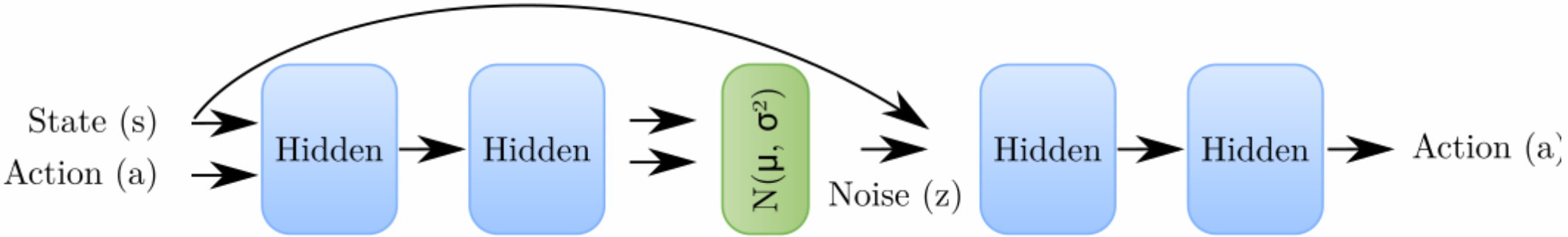
Continuous BDPI

- Keep the off-policy critics
  - But with continuous actions
- Keep the stochastic actor
  - But take care of the argmax

The critics (Q-Functions)



The critics (VAE)

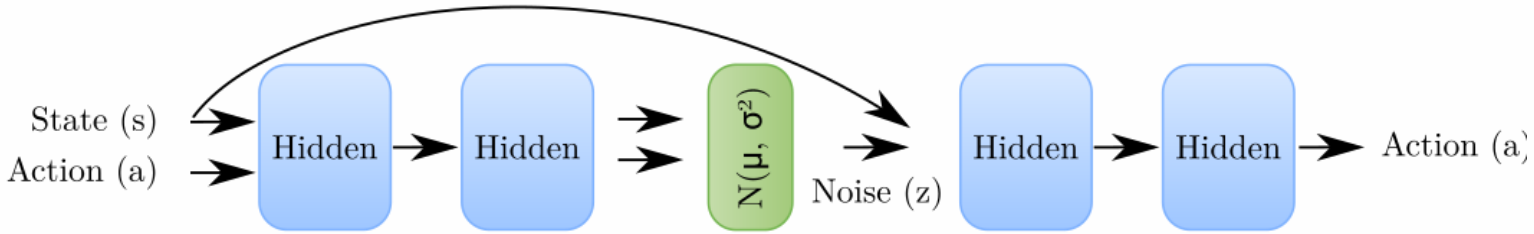


Continuous BDPI

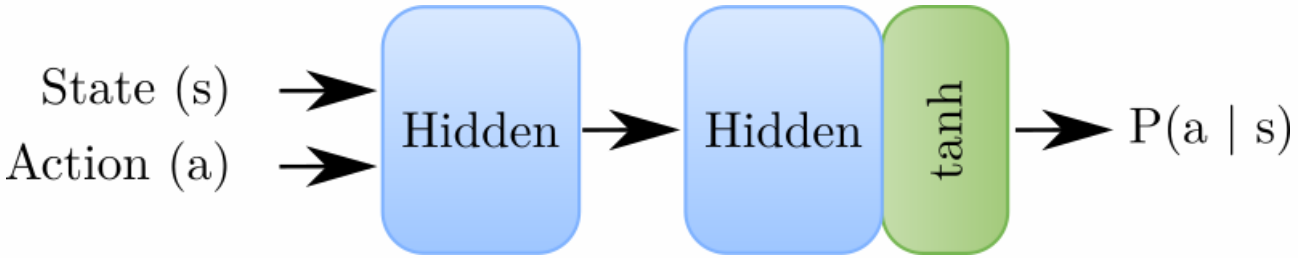
Critic update:

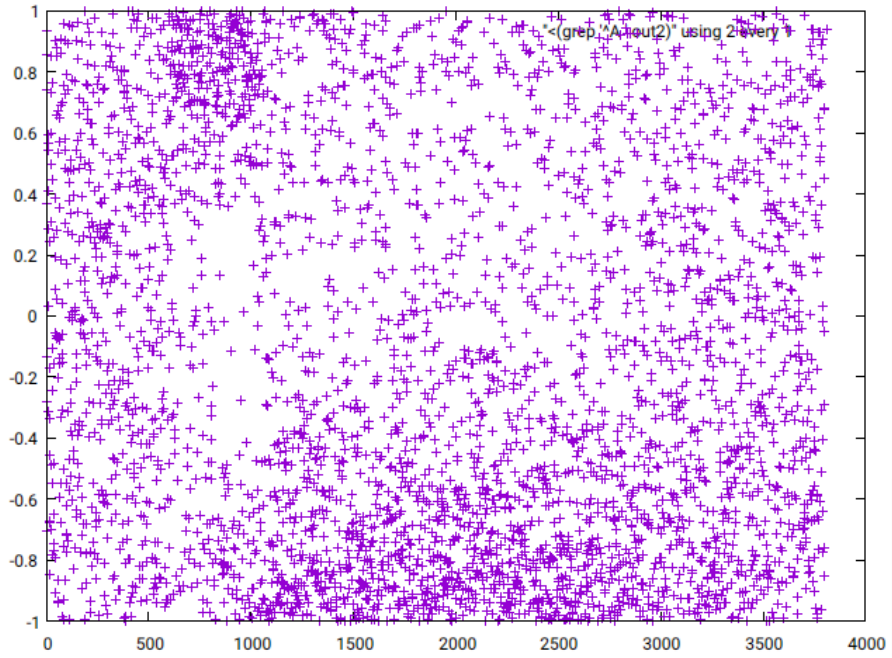
$$Q(s_t, a_t) \mathrel{+}= \alpha * (r_t + \gamma \, \mathbf{max}_{\mathbf{a' \sim VAE(st+1)}} Q(s_{t+1}, a') - Q(s_t, a_t))$$

The actor: a VAE?



The actor: rejection sampling





- Sample actions uniformly
- For each action a:
  - If `random() < P(a | s)`, accept

Continuous BDPI

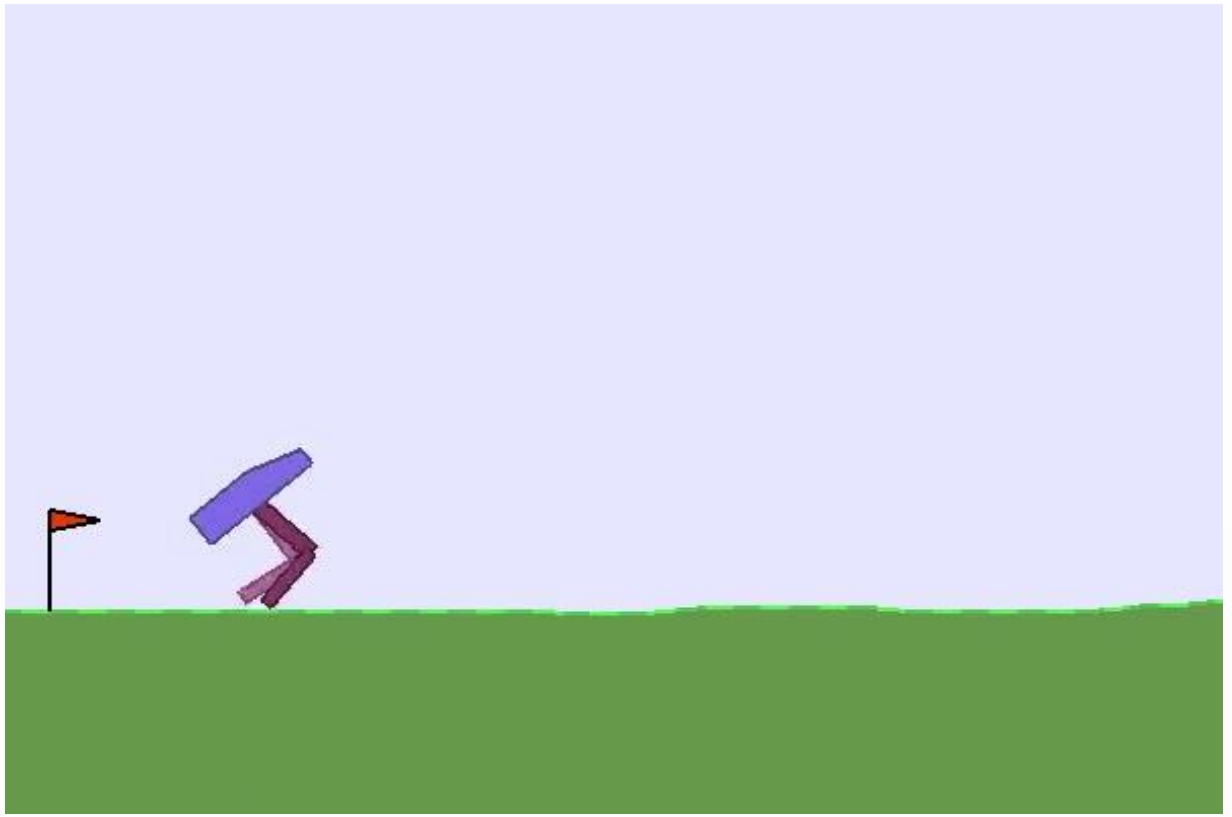
Actor update:

$$\Pi(s_t) \leftarrow \mathbf{argmax}_{a \sim Unif}\ \mathbf{Q_i(s_t, a)}$$

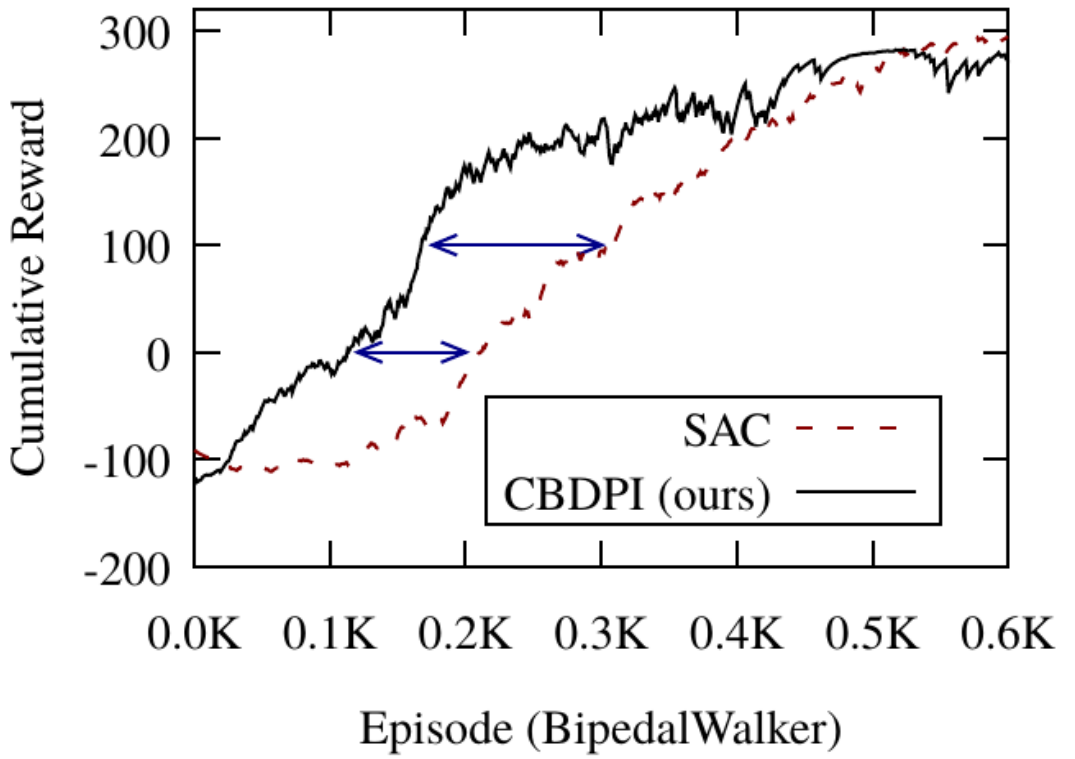
Implemented as:

$$\Pi(s_t, a^*) \mathrel{+}= alr$$
$$\Pi(s_t, a^- \sim \text{Unif}) \mathrel{-}= alr$$

Experiments



Results


300
200
100
0
-100
-200
Cumulative Reward
SAC
CBDPI (ours)
0.0K
0.1K
0.2K
0.3K
0.4K
0.5K
0.6K
Episode (BipedalWalker)