DISCRETE AND CONTINUOUS ACTION REPRESENTATION FOR PRACTICAL RL IN VIDEO GAMES

Olivier Delalleau*, Maxim Peter*, Eloi Alonso, Adrien Logut
Ubisoft Montréal
*Equal contribution, alphabetical order

UBISOFT LA FORGE

Abstract

- We focus on applications of Reinforcement Learning in **AAA video games** under associated performance constraints. We thus use off-policy algorithms with small feedforward networks.
- We propose an extension of Soft Actor-Critic [1] called **Hybrid SAC** to handle either discrete, continuous or parameterized action spaces.
- We apply this algorithm for a high speed driving task in **Watch Dogs 2**©.
- We study the impact of using **Normalizing Flows** on algorithm performance.

Hybrid Soft Actor-Critic

Soft Actor-Critic is part of the maximum entropy framework in which agents maximize the sum of rewards and entropy bonus:

$$E_\pi\left[\sum_t \gamma^t (r_t + \alpha\, H(\pi(\cdot\,|s_t)))\right]$$

When there are both discrete and continuous actions, we propose to separate the entropy bonus between discrete and continuous actions:

$$H\left(\pi(a^d, a^c|s)\right) = H(\pi(a^d|s)) + \sum_{a^d} \pi(a^d|s) H(\pi(a^c|s, a^d))$$

We separate the contributions of the continuous and discrete entropies in the final entropy bonus:

$$\alpha^d H(\pi(a^d|s)) + \alpha^c \sum_{a^d} \pi(a^d|s) H(\pi(a^c|s, a^d))$$

Main references

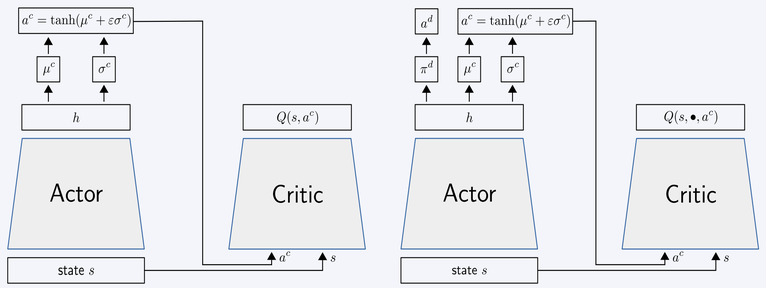
[1] Haarnoja et al. (2018). **Soft actor-critic algorithms and applications**

[2] Bester et al. (2019). **Multipass q-networks for deep reinforcement learning with parameterised action spaces**

[3] Christodoulou (2019). **Soft actor-critic for discrete action settings**

[4] Mazoure et al. (2019). **Leveraging exploration in off policy algorithms via normalizing flows**

Architecture



Standard SAC | *Hybrid SAC (one discrete component and one continuous)*

Results on parameterized action spaces

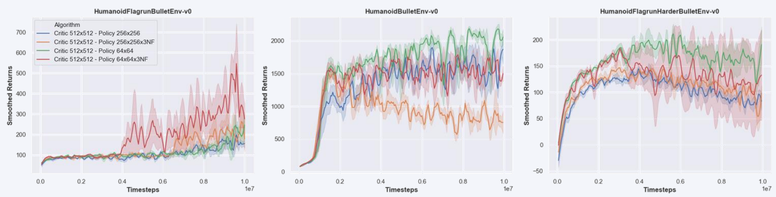
Algorithm	Platform Return	Goal P(goal)	HFO P(goal)
MP-DQN	0.987 ± 0.039	0.789 ± 0.070	0.509 ± 0.110
MP-DQN (with MC)	-	-	0.913 ± 0.070
Hybrid SAC	*0.981 ± 0.013*	*0.728 ± .047*	*0.639 ± 0.141*

Hybrid SAC is competitive with the recently proposed MP-DQN [2].
Results averaged over 30 seeds with 95% confidence interval.
"With MC" means that Monte Carlo rollouts are used.
HFO: Half Field Offense.

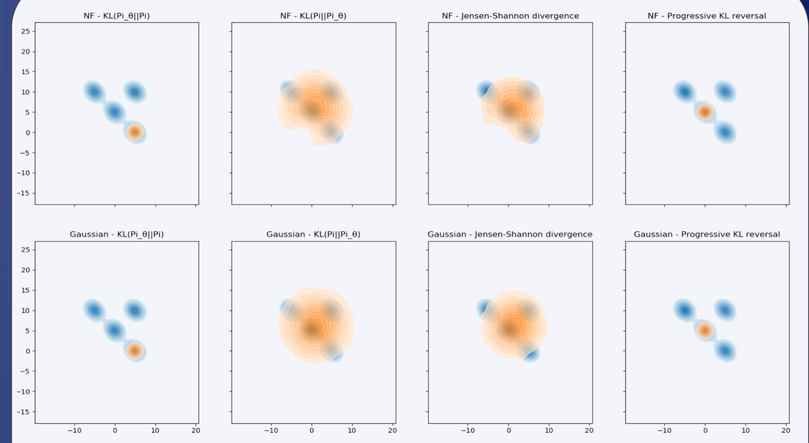
Normalizing Flows

We investigate the claim that radial Normalizing Flows are a cheap way to extend SAC to get more expressive policies.

In our experiments, it is not clear that the policies obtained improve over the baseline. We suspect this is due to the mode matching behavior of the KL divergence in SAC.



SAC with radial normalizing flows does not convincingly outperform regular Gaussian SAC *on three Roboschool PyBullet environments.*
Curves are averaged on 5 random seeds, and smoothed.



Toy experiment: final shapes of policy distributions (in orange) after trying to match a fixed Gaussian mixture (in blue) for 10,000 steps. Top row uses normalizing flows, while the bottom row is using a Gaussian policy. Various divergence losses are evaluated from left to right. The leftmost column shows that ***the loss used in SAC makes normalizing flows collapse to a single mode just like a regular Gaussian policy****.*

Conclusions

- We introduced Hybrid SAC, an extension to the SAC algorithm to handle either discrete, continuous or parameterized action spaces. It exhibits competitive performance with the state-of-the-art on parameterized actions benchmarks.
- Our results suggest that while Normalizing Flows do not seem to improve SAC performance, they might still be leveraged by using other losses.

Acknowledgements

We would like to thank the authors of [4] (Mazoure et al. 2019) for insightful conversations and providing us with their implementation. We also thank Paul Barde for his valuable feedback.

Contact

laforge@ubisoft.com, we're hiring !

QR link to the paper