Reinforcement Learning: Parameterizations for Distributional Continuous Control

Felix Grün, [1,2] *Muhammed Saif-ur-Rehman* [1] *and Ioannis Iossifidis* [1]

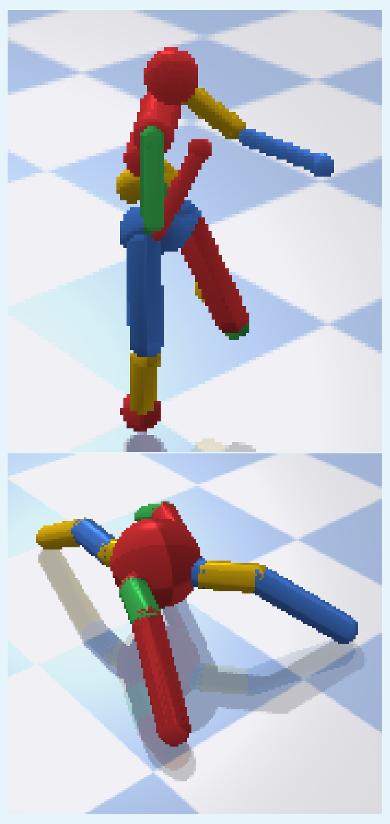
[1]Ruhr West University of Applied Sciences, Robotics and BCI Laboratory, Mülheim, Germany
[2]Ruhr University Bochum, Institute for Neural Computation, Bochum, Germany

Background

- Distributional RL: learn value distribution instead of just the expected value of an action. Evidence for equivalent mechanism in the brain.
- Distributional algorithms have empirically proven to be significant improvements over their non-distributional equivalents, e.g. [1].
- Main areas of variation among DRL algorithms:
  - representation/parameterization of distributions
  - (pseudo-) metric used to measure distance between distributions
- Several significant parameterizations were introduced based on DQN:
  1. Categorical [1]
  2. Quantile Regression [2]
  3. Implicit Quantile Networks [3]
  4. Fully Parameterized Quantile Function [4]
  5. Maximum Mean Discrepancy DQN [5]
- In this work options 2 to 4 are compared in the continuous action setting

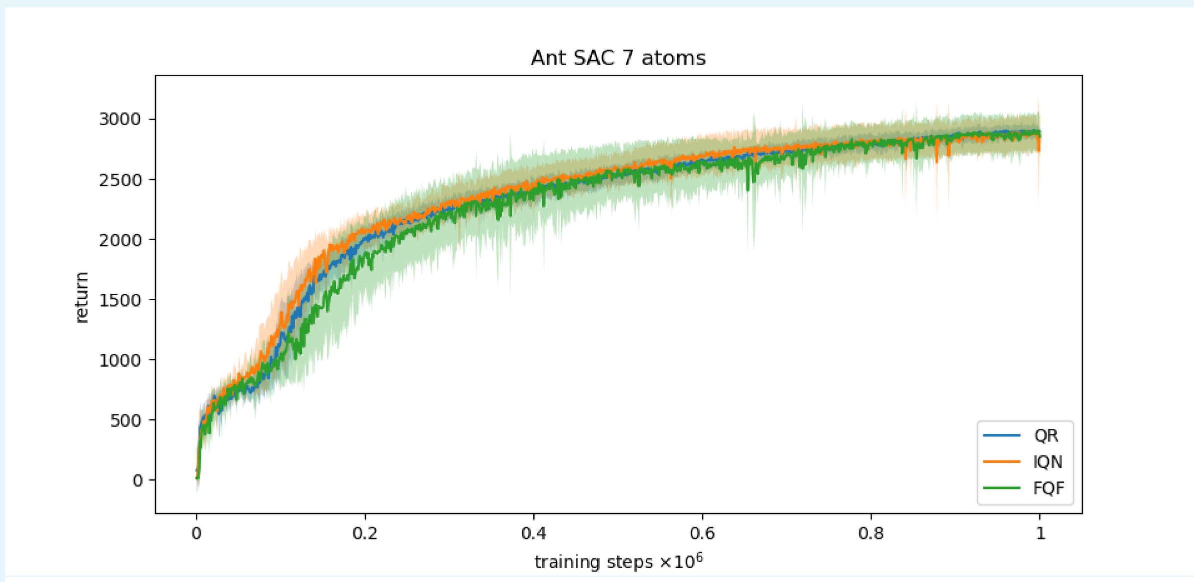
Methods

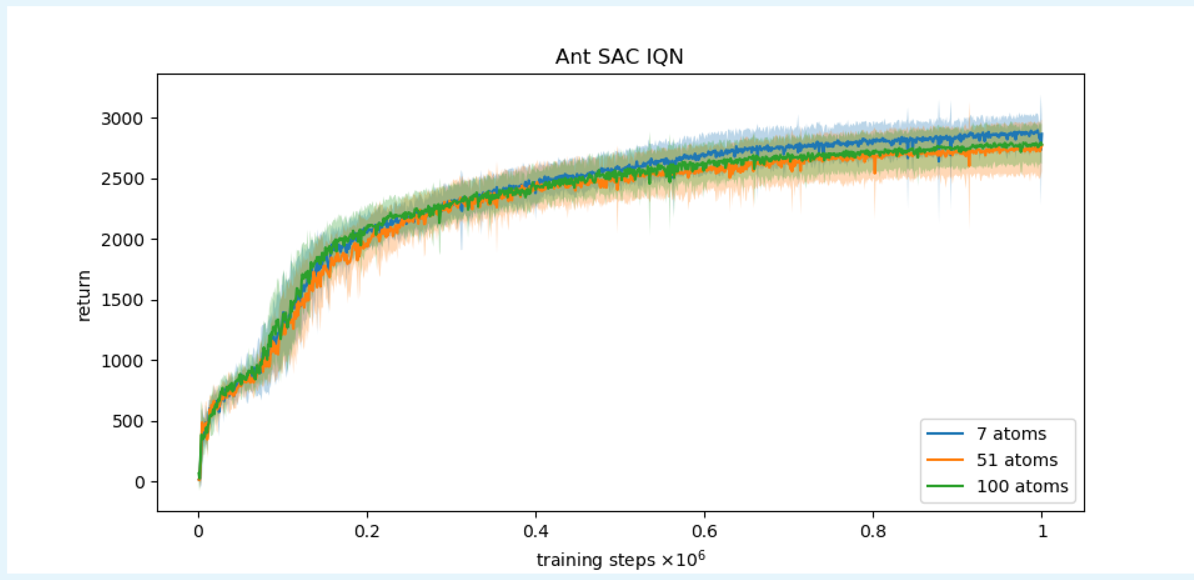
- **Comparison:** Three distributional types of critics were implemented on top of SAC and TD3 in the popular sb3 [6] library. We test each critic with both base algorithms using 7, 51 and 100 atoms.
- **Hyper Parameter Search:** Hyper parameters were tuned separately for each algorithm, parameterization and number of atoms using optuna [7] on the hardest task from the set we selected (humanoid). In order to isolate the effects of varying the number of atoms the same hyperparameters were used across resolutions.
- **Evaluation:** each algorithm was trained **10** times in each setting. Each evaluation was done determi- nistically averaged over **5** episodes. Plots show the mean and std of those values over the **10** trainings.

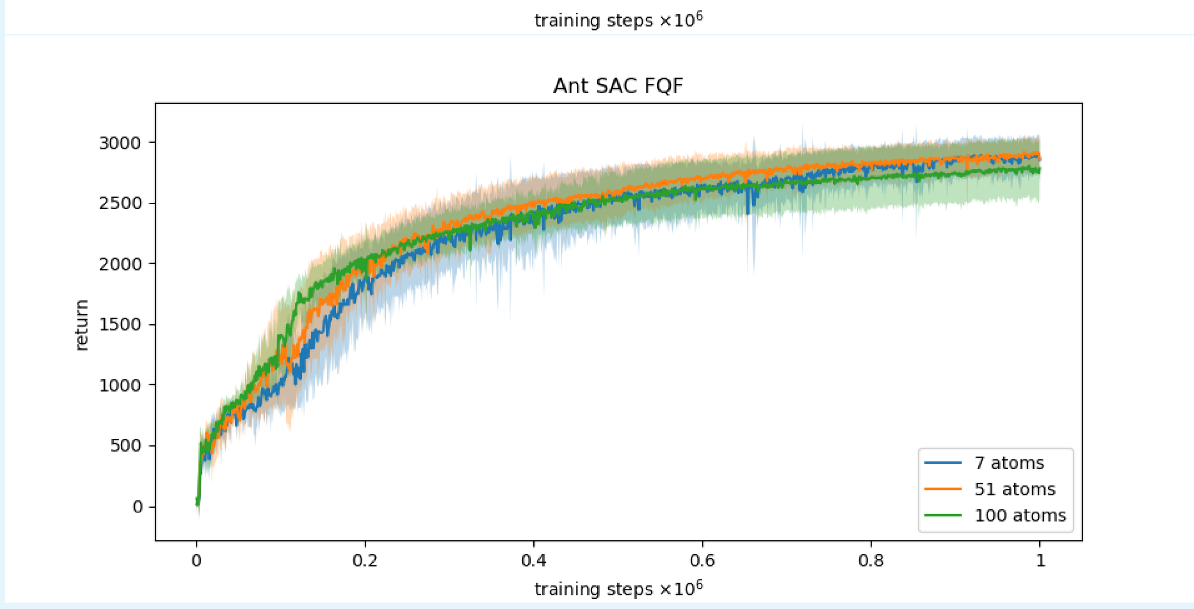


Humanoid (top) and Ant environments (bottom).

Results - Preliminary!







	Mean	Median	>Human	>DQN
DQN	221%	79%	24	0
C51	701%	178%	40	50
QR-DQN	902%	193%	41	54
IQN	1112%	218%	39	54
FQF	1426%	272%	44	54

Qualitatively different results from the discrete action domain, specifically 57 (55) Atari games, as reported by [4].

The results we have so far suggest that the advantages of IQN and FQF do not automatically translate to the continuous action domain: neither the number of atoms nor the parameterization have a significant impact on the learning performance.

Observations Regarding Hyperparameters

Having extensively tuned hyperparameters for the different setting, we can make some observations:

- tuned learning rates are highest for quantile regression with fixed quantiles and lowest for learned quantiles (fully parameterized (FQF))
- tuned learning rates for SAC based algorithms tend to be more than double the equivalent TD3 based optimum.
- SAC based algorithms appear to be less sensitive to choice of hyperparameters than the TD3 based
- original IQN allows the use of
  1. different number of target atoms than predicted atoms
  2. distortion risk measures to derive risk-sensitive policies

1. was ignored in favor of comparability with the other methods. **2.** is not applicable in the same way in an actor-critic setting because the policy is not implicitly defined by the value distribution.

Literature

[1] M. G. Bellemare *et al.*, "A Distributional Perspective on Reinforcement Learning,", 2017.

[2] W. Dabney *et al.*, "Distributional Reinforcement Learning with Quantile Regression,", 2018.

[3] W. Dabney *et al.*, "Implicit Quantile Networks for Distributional Reinforcement Learning," *arXiv:1806.06923 [cs, stat]*, Jun. 2018.

[4] D. Yang *et al.*, "Fully Parameterized Quantile Function for Distributional Reinforcement Learning,", 2019.

[5] T. Nguyen-Tang *et al.*, "Distributional Reinforcement Learning via Moment Matching,", vol. 35, no. 10, May 2021, ISSN: 2374-3468.

[6] A. Raffin *et al.*, *Stable Baselines3*, 2019.

[7] T. Akiba *et al.*, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Acknowledgement

This work is supported by the Ministry of Economics, Innovation, Digitization and Energy of the State of North Rhine-Westphalia and the European Union, grants GE-2-2-023A (REXO) and IT-2-2-023 (VAFES)

iossifidis Lab: Ruhr West University of Applied Sciences, Mülheim, Germany Mail: `felix.gruen@hs-ruhrwest.de` WWW: `http://lab.iossifidis.net`