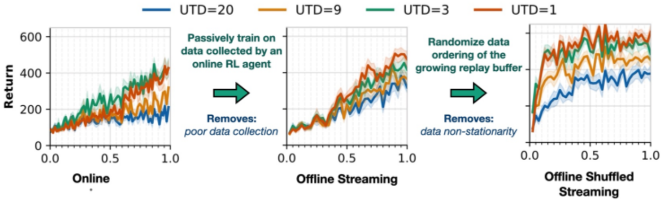
High UTD ratio algorithms

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Towards high UTD ratio

High UTD ratio

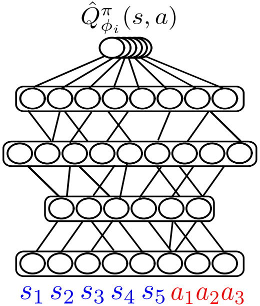


- UTD: Update to data
- High UTD ratio: update the policy (and the critic) further from the same data
- Boost in performance on DQN, PPO and SAC, but with limitations
- An issue is the overestimation bias, another is overfitting to the sampled data

Li, Q., Kumar, A., Kostrikov, I., and Levine, S. (2023) Efficient deep reinforcement learning requires regulating overfitting. *arXiv preprint arXiv:2304.10466*

TQC

TQC: Distributional estimation



- Using a distribution of estimates is more stable than a single estimate
- C51, D4PG, QR-DQN...
- TQC uses N critic heads to estimate a distribution of Q-values
- Taking the Q-value as a random variable rather than a maximum likelihood estimate

Bellemare, M. G., Dabney, W., and Munos, R. (2017) A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*

Truncated Quantile Critics

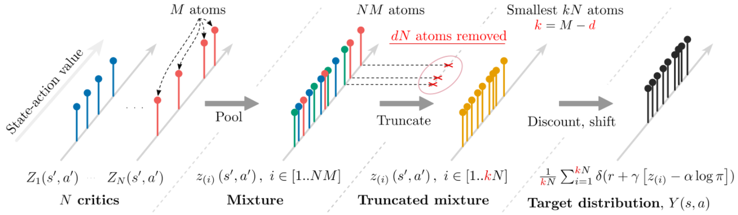
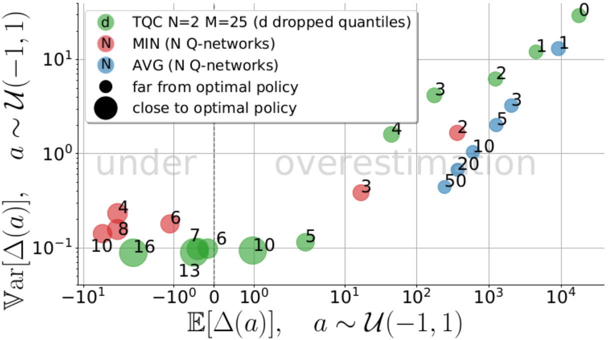


Figure 2. Step-by-step construction of the temporal difference target distribution $Y(s, a)$. First, we compute approximations of the return distribution conditioned on s' and a' by evaluating N separate target critics. Second, we make a mixture out of the N distributions from the previous step. Third, we truncate the right tail of this mixture to obtain atoms $z_{(i)}(s', a')$ from equation 11. Fourthly, we add entropy term, discount and add reward as in soft Bellman equation.

- Each atom is a Q-value estimate
- To fight overestimation bias, TD3 and SAC take the min over two critics
- TQC truncates the higher quantiles

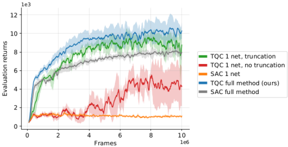
Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pp. 5556–5566. PMLR, 2020

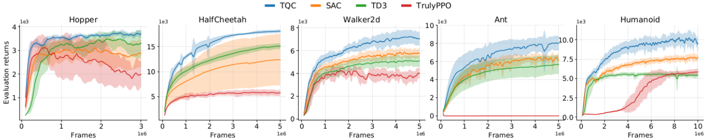
Rationale: bias-variance diagram



- x-axis = bias, y-axis = variance
- Taking the min or the average over N networks is not flexible
- Truncating the higher quantiles results in getting closer to the optimal policy

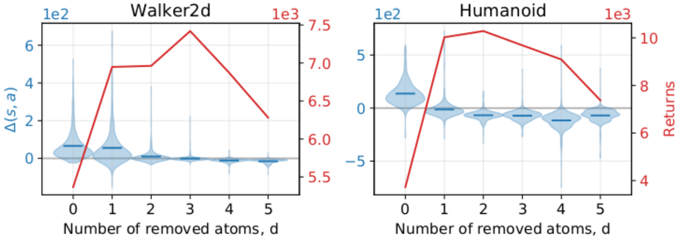
Performance





- Top figure: Humanoid-v2
- From 5 to a single critic
- Outperforms SAC, easier to use

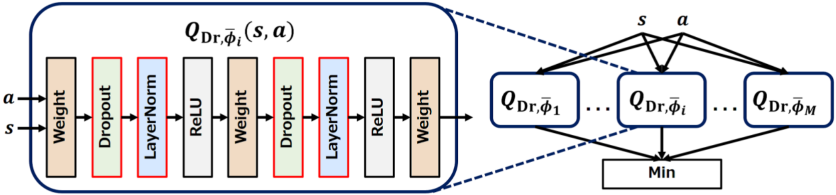
Impact of truncation



- red = performance
- blue = distribution of error
- The optimal number of truncated quantiles is not always the same

DroQ

DroQ: Dropout and ensembling

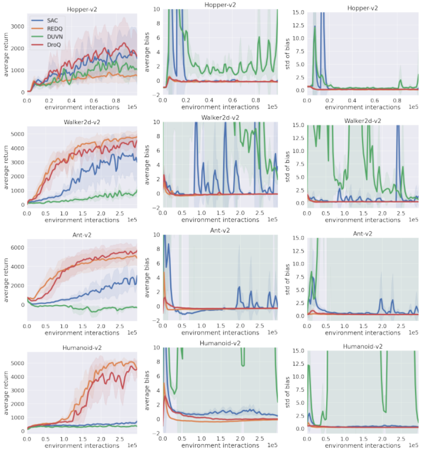


- REDQ: Ensembling from random networks
- DroQ: Dropout, Layer Normalization and ensembling

Chen, X., Wang, C., Zhou, Z., and Ross, K. (2021) Randomized ensembled double Q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., and Tsuruoka, Y. (2021) Dropout Q-functions for doubly efficient reinforcement learning. *arXiv preprint arXiv:2110.02034*

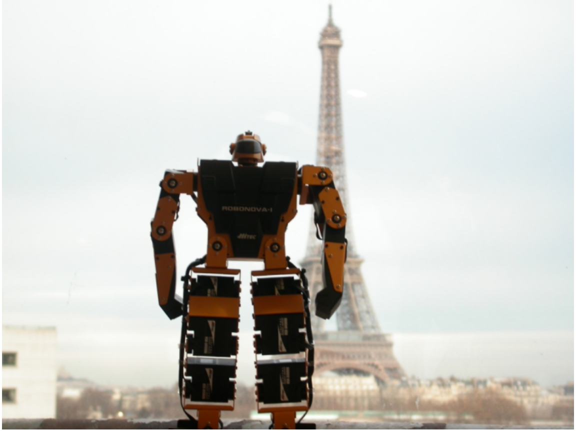
DroQ: Performance



- Outperforms SAC, REDQ and DUVN
- No comparison to TQC

Moerland, T. M., Broekens, J., and Jonker, C. M. (2017) Efficient exploration with double uncertain value networks. *arXiv preprint arXiv:1711.10789*

Any question?



Send mail to: `Olivier.Sigaud@upmc.fr`

Bellemare, M. G., Dabney, W., and Munos, R. (2017).
A distributional perspective on reinforcement learning.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR.

Chen, X., Wang, C., Zhou, Z., and Ross, K. (2021).
Randomized ensembled double Q-learning: Learning fast without a model.
arXiv preprint arXiv:2101.05982.

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., and Tsuruoka, Y. (2021).
Dropout Q-functions for doubly efficient reinforcement learning.
arXiv preprint arXiv:2110.02034.

Kuznetsov, A., Shvechikov, P., Grishin, A., and Vetrov, D. P. (2020).
Controlling overestimation bias with truncated mixture of continuous distributional quantile critics.
In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5556–5566. PMLR.

Li, Q., Kumar, A., Kostrikov, I., and Levine, S. (2023).
Efficient deep reinforcement learning requires regulating overfitting.
arXiv preprint arXiv:2304.10466.

Moerland, T. M., Broekens, J., and Jonker, C. M. (2017).
Efficient exploration with double uncertain value networks.
arXiv preprint arXiv:1711.10789.