QDHUAC – Discovering Diverse Behaviours in a Fraction of the Time

Earlier this month, we were honoured to receive Best Paper Award at the Genetic and Evolutionary Computing Conference (GECCO) for our paper, "Distributional Value Estimation Without Target Networks for Robust Quality-Diversity." We demonstrate a *target-free actor-critic algorithm* for use in QD-RL, a significant advancement combining the latest advancements in Deep Reinforcement Learning (RL) with Quality Diversity (QD), accelerating AI training and dramatically improving results. In the rest of this blog post, we'll explain how.

Across robotics and industry, RL is an increasingly essential technique for creating controllers (in RL, *actors*) capable of complex, evolving tasks – think landing a plane in poor weather, or navigating a deceptive maze. Typically, RL produces only one actor, trained through incremental improvements against a defined goal. However, it has long been known in the field that "Greatness Cannot Be Planned": that uncompromising pursuit of an objective can prevent you from achieving it, and that prioritising a *diversity* of approaches often yields the best result. This is where evolutionary approaches, and particularly Quality Diversity algorithms, excel by discovering rich repertoires of actors with diverse behaviours [1]. It is possible to combine these approaches, and QD-RL [2] marries QD with modern gradient-based Deep RL to achieve state of the art performance on many challenging baseline problems.

However, existing QD-RL algorithms have a major flaw: they require many environment interactions to arrive at high-quality solutions, or in other words, they exhibit very poor *sample efficiency*. In recent years, RL papers have shown that sample efficiency can be improved by simply performing more training steps per environment interaction, known as the *Update-to-Data ratio* (UTD) [3] – but prior attempts to apply this in QD-RL have failed. In our paper, we demonstrate that the culprit is a foundational component common to many RL algorithms: **the target network**.

QD-RL uses a specific form of RL: *actor-critic Deep Q-learning*. In short, each actor is a neural network; and the actors are evaluated by a "critic" known as the Q-network. The Q-network aims to learn how good an action is based on the current state, , which it does by observing actors' experiences in the environment and their associated rewards. However, the definition of the Q function is recursive, and this causes numerical instability in naive training formulations. One very popular solution to this is the *target network*, introduced in the now-famous 2015 Nature paper [4]. which is simply a lagging copy of the Q-network, designated θ^-, which transforms the training error term:

$$L(\theta) = \mathrm{E}\left[\left(r + \gamma \max_{a'} Q(s',\, a';\, \theta) - Q(s,\, a;\, \theta)\right)^2\right]$$

Into this (note the θ^-):

$$L(\theta) = \mathrm{E}\left[\left(r + \gamma \max_{a'} Q(s',\, a';\, \theta^{-}) \;-\; Q(s,\, a;\, \theta)\right)^2\right]$$

By decoupling the target from the Q-network parameters in this way, we can train Q-networks without instability, even when increasing the UTD.

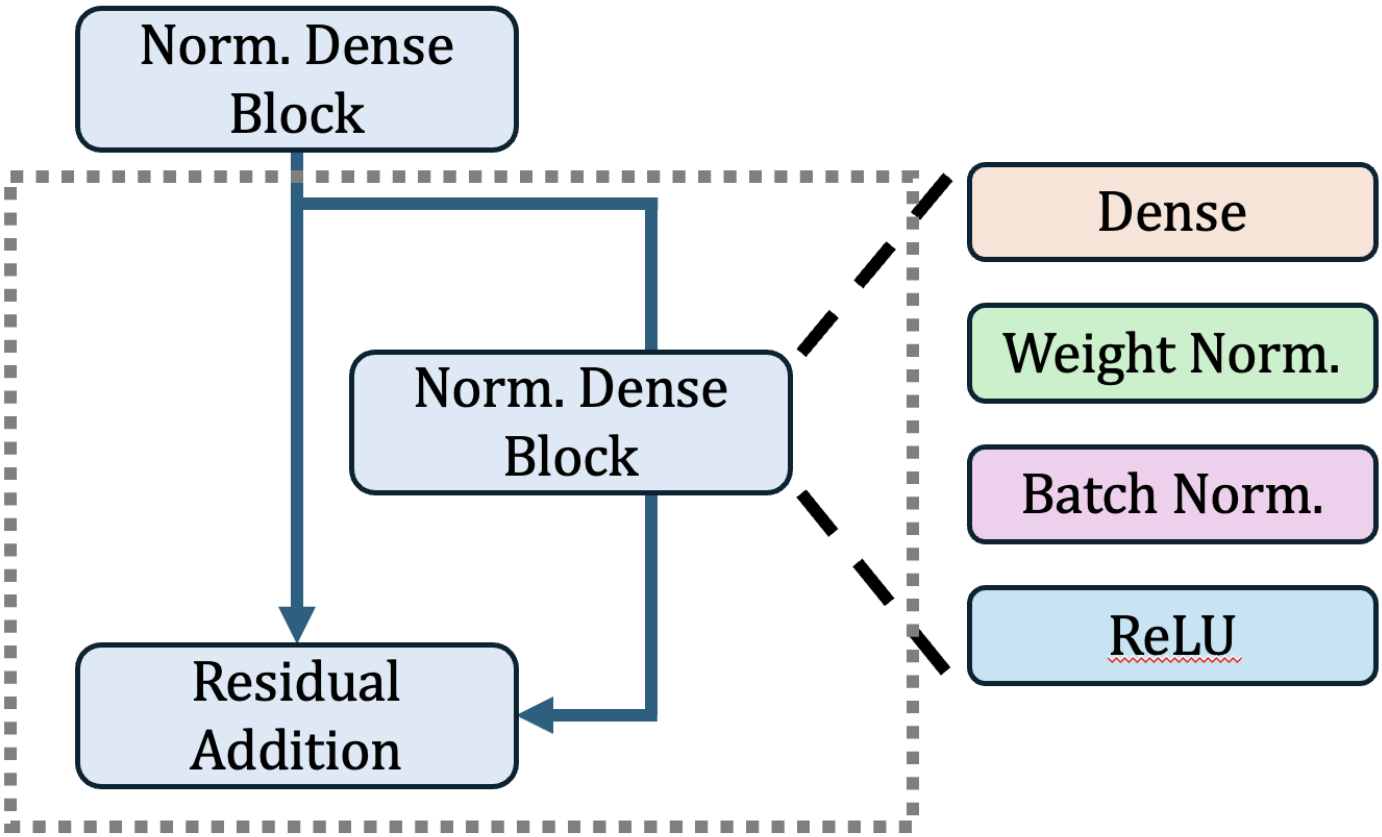
As mentioned, in QD-RL approaches, naively increasing UTD doesn't work, and the target network is to blame. To understand this, remember that in standard RL, we have one actor incrementally improving its behaviour, which rarely changes dramatically. The Q-network learns from broadly similar and related experiences, and the delayed information of the target network causes little harm. In the QD-RL regime, however, we have many diverse actors contributing novel experiences for the Q-network to learn from, increasing the importance of up-to-date evaluations, and making convergence more challenging. We call this phenomenon the **Target Tracking Lag**.

Because the target network is updated slowly, it tends to limit the discovery of new behaviours. Imagine one of our actors suddenly discovers a brilliant new way to run. When the training process tries to evaluate this new, unseen *state-action trajectory*, the outdated target network evaluates it based on old data. It looks at this novel, high-performing solution and suggests that this new behaviour is terrible and should not be attempted. This creates a **Pessimism Gap:** the target network systematically undervalues rapid evolutionary breakthroughs, and by feeding the primary network these pessimistic "false negatives", the target network actively suppresses the discovery of new niches. It has previously been shown that QD is very sensitive to this sort of misevaluation, which stops our algorithm from exploring the very behaviours that QD-RL aims to discover.

Enter *QDHUAC* (Quality-Diversity High Update Actor-Critic), an actor-critic algorithm which resolves these problems by removing the target network entirely. This unlocks high UTD training, but re-introduces the divergence the target network was there to solve. QDHUAC then utilises two key architectural components to stabilise training, tuned for the QD-RL setting.

Staying Grounded with Hybrid Normalisation

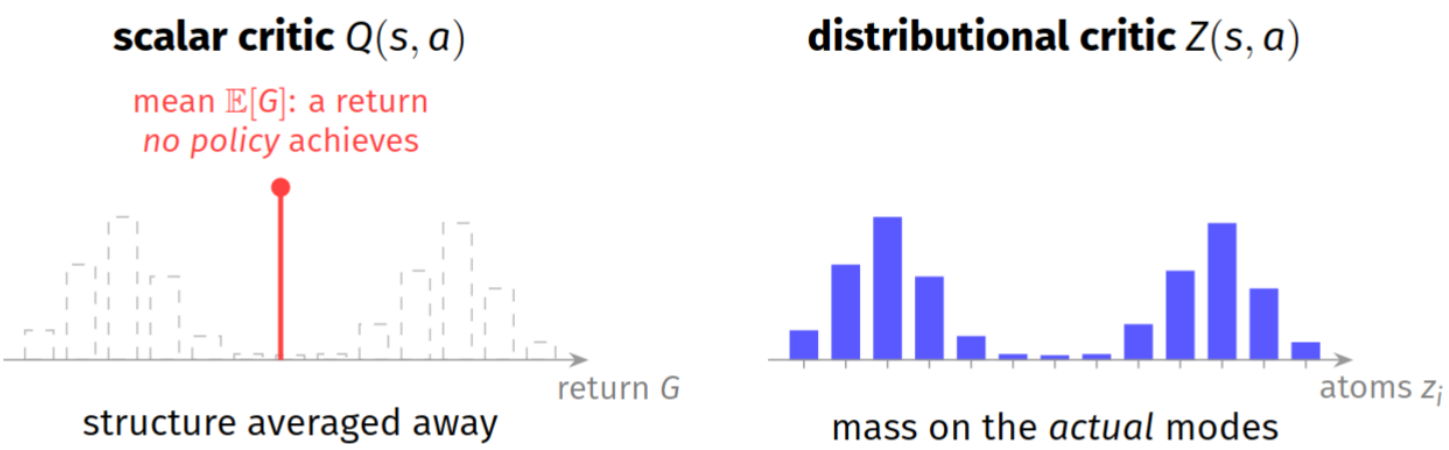
Following work in other RL contexts, removing the target network can be compensated through careful use of *batch normalisation* (BN) [5, 6]. BN smooths out the optimisation problem, which is particularly important when learning from the highly diverse trajectories collected in QD-RL. However, in the QD-RL setting, this isn't sufficient. Even with the other adjustments discussed later, training signals are still too erratic, and performance is poor. In practice, we found that applying *weight normalisation* (WN) [7] was the solution. WN places a hard mathematical limit on how fast the network can alter its weights and smooths the optimisation further, and so we apply both WN+BN in a hybrid normalisation scheme to every layer.



Seeing the Whole Picture with a Distributional Critic

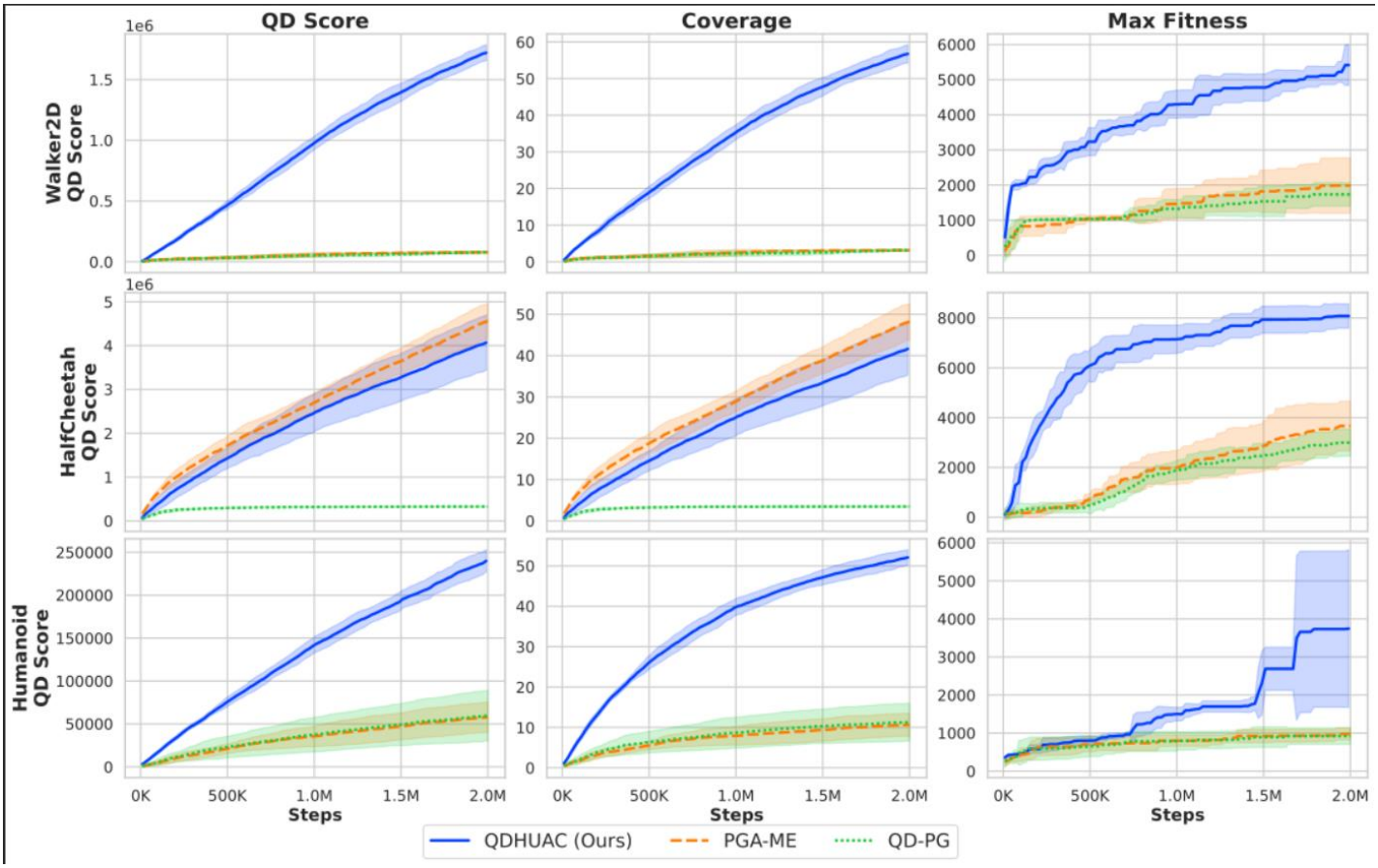
In standard RL, the critic calculates the "average" best total reward. However, a QD-RL critic learns from thousands of wildly different actors. Averaging all diverse experiences into a single value washes out the rich underlying structure.

Instead, QDHUAC uses a **distributional critic** [8]. Rather than guessing a single mean value, it models the full probability distribution of possible returns, which provides a richer and denser gradient signal. By initialising the critic with a uniform distribution, we bake in *optimism* towards new, unexplored actions. Finally, by distributing probability mass across a spectrum of fixed atoms, we prevent the massive, single value spikes that typically destabilise target-free learning.



Learning to Walk, One (Big) Step at a Time

To test whether eliminating the Target Tracking Lag genuinely unlocks faster learning, we evaluated QDHUAC across five challenging, high-dimensional continuous mechanical control tasks from the brax benchmark suite, including the notoriously high-dimensional Ant and Humanoid environments. We benchmarked our approach against state-of-the-art QD-RL baselines, PGA-ME and QD-PG.



QDHUAC achieved competitive coverage and elite maximum fitness scores using an **order of magnitude fewer environment steps** than baselines. In fact, the power of target-free learning becomes even more obvious in environments with difficult exploration dynamics. For example, in the humanoid task, standard baseline critics often destabilise here because the actor constantly falls over early in training; constrained by low update frequencies, they struggle to find a policy capable of surviving the episode. In contrast, QDHUAC achieves approximately 3.5 x higher fitness. Because it isn't held back by

a lagging critic, it rapidly constructs a diverse repertoire of distinct, surviving gaits that standard algorithms struggle to find.

We also observed a fascinating behavioural shift in environments like HalfCheetah and Ant. In HalfCheetah, QDHUAC's maximum fitness soared to nearly 8000, completely dwarfing the baseline's peak of 3500. We hypothesise that this is due to the effectiveness of the high-UTD critic, which can provide a much stronger learning signal earlier in the training process. While this intense exploitation meant the algorithm spent slightly less time mapping out low-performing, divergent regions, it drastically prioritised populating the archive with high-quality solutions.

Conclusion

For nearly a decade, target networks have been viewed as a fundamental requirement for value estimation in off-policy Reinforcement Learning. However, our paper adds to a growing body of work suggesting they are not an absolute necessity, but an architectural crutch. When we step out of the world of standard RL and into the more diverse, chaotic world of Quality-Diversity, we find that the latency introduced by target networks creates a pessimistic bias that actively suppresses the discovery of novel behaviours. By combining a Target-Free Distributional Critic with Hybrid Normalization, QDHUAC demonstrates that we can resolve the structural conflict between high-speed learning and diverse exploration.

This dramatically reduces the computational bottlenecks for evolutionary reinforcement learning and opens the door for a new generation of QD-RL-based learning in resource intensive domains.

To see our full paper: LINK TO ARXIV

[1] Eysenbach, B., Gupta, A., Ibarz, J. and Levine, S., 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

[2] Lim, Bryan, Manon Flageat, and Antoine Cully. "Understanding the synergies between quality-diversity and deep reinforcement learning." In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1212-1220. 2023.

[3] Hiraoka, Takuya, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, and Yoshimasa Tsuruoka. "Dropout q-functions for doubly efficient reinforcement learning." *arXiv preprint arXiv:2110.02034* (2021).

[4] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. and Petersen, S., 2015. Human-level control through deep reinforcement learning. *nature*, *518*(7540), pp.529-533.

[5] Bhatt, Aditya, Daniel Palenicek, Boris Belousov, Max Argus, Artemij Amiranashvili, Thomas Brox, and Jan Peters. "Crossq: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity." In *International conference on learning representations*, vol. 2024, pp. 55293-55311. 2024.

[6] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *International conference on machine learning*, pp. 448-456. pmlr, 2015.

[7] Salimans, Tim, and Durk P. Kingma. "Weight normalization: A simple reparameterization to accelerate training of deep neural networks." *Advances in neural information processing systems* 29 (2016).

[8] Bellemare, M.G., Dabney, W. and Munos, R., 2017, July. A distributional perspective on reinforcement learning. In *International conference on machine learning* (pp. 449-458). Pmlr.