A Unified Scaling Law for Bootstrapped DQNs

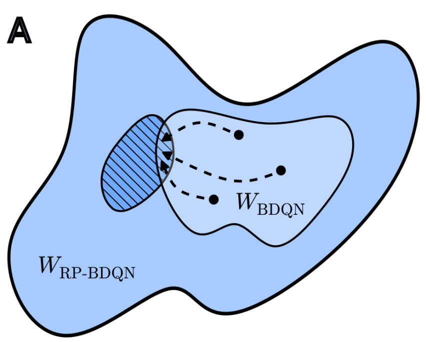
Roman Knyazhitskiy, Pascal van der Vaart, Neil Yorke-Smith

Delft University of Technology

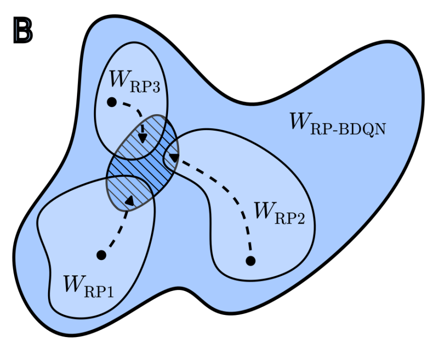
Introduction

Reinforcement Learning (RL) is concerned with finding a policy that maximizes cumulative reward over a sequence of actions. A successful algorithm has to balance (epistemic) exploration with exploitation [1].

Bootstrapped-based methods allow for such balancing by means of an approximate posterior over Q-networks.



Training dynamics of BDQN



Training dynamics of RP-BDQN

We are analyzing and comparing the convergence behavior of Bootstrapped DQN (BDQN) [2] and BDQN with Randomized Priors (RP-BDQN) [3].

We show that the probability of discovering the only rewarding path on DeepSea [4] (PoD) for both methods is governed by a simple K-trials binomial law:

$$P(\text{discovery}) \approx 1 - (1 - \psi^n)^K$$

We also show that Randomized Priors help to prevent posterior collapse, which allows them to have higher ψ.

Research Questions

1. How do BDQN and RP-BDQN scale with the ensemble size and DeepSea [4] environment size?

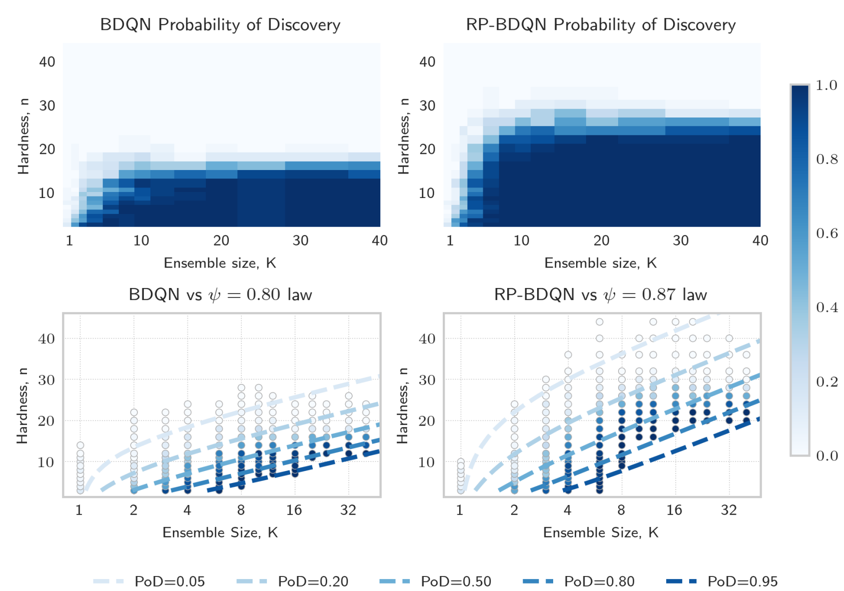
2. Can we describe the probability of these methods discovering a solution using a closed-form scaling law that is robust to changes in hyperparameters?

3. Where does the identified scaling law break down, and are there any properties of the ensemble that are related?

Method & Results

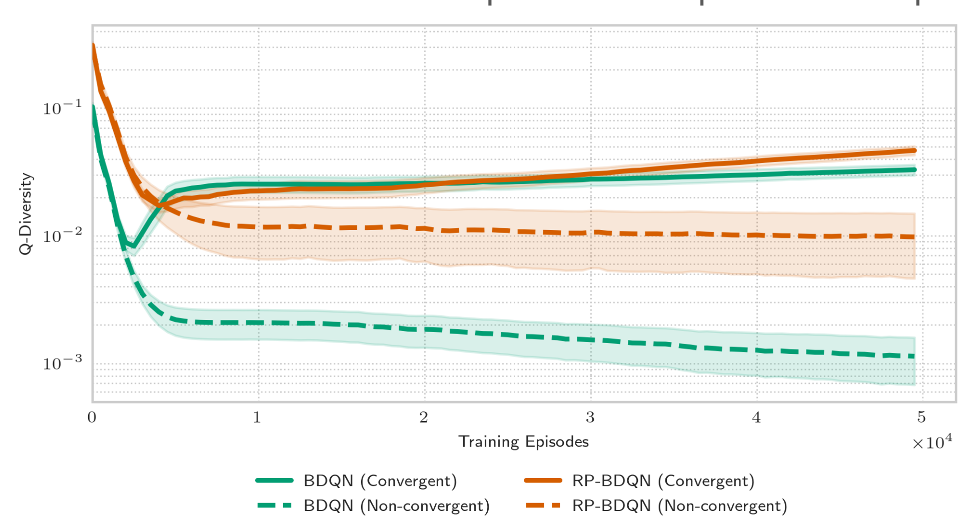
We ran >40,000 experiments on DeepSea [4] environment; each one for 50k per-ensemble episodes.
We show that the probability of discovering the rewarding path is fairly well approximated by our simple model:

Algorithm	Parameter (ψ)	Goodness-of-fit (R^2)	Dispersion	MSE
BDQN	0.80 ± 0.02	0.84	4.1	0.024
RP-BDQN	0.87 ± 0.01	0.69	8.1	0.049



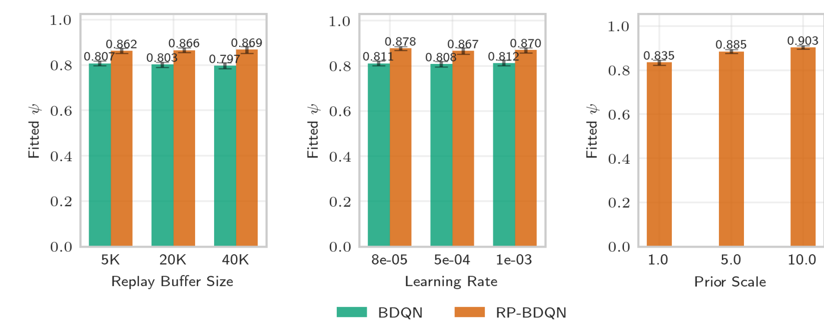
Posterior Collapse

The introduced Q-Diversity metric – standard deviation of Q-values over different ensemble members on hold-out set of states – shows that randomized priors reduce posterior collapse.



Hyperparameter Sensitivity

We find no effect on the law parameter from changing learning rate and replay buffer size, while a significant effect from changing the prior scale.



Limitations

1. The law works well in medium-K regime, but fails in small- and large- K regimes. We attribute this to the compute budget limit and our member-independent model. The dispersion indicates an overall poor fit.
2. No use of more complex environments, because of no clearly defined "hardness" metric, which limits applicability.

Future work

1. Optimizing for ψ: Moving from a descriptive to a prescriptive use of our scaling law by designing algorithms that explicitly aim to maximize ψ.
2. Refining the Scaling Law: Developing a more nuanced model that accounts for the cooperative effects to better capture ensemble dynamics.

GitHub

References

[1] - "Reinforcement Learning: An Introduction", Sutton R. and Barto A.
[2] - "Deep exploration via bootstrapped dqn", Osband I. et al
[3] - "Randomized prior functions for deep reinforcement learning", Osband I. et al
[4] - "Behaviour suite for reinforcement learning", Osband I. et al