Empirical Evaluation of Random Network Distillation for DQN Agents

Alberto Moreno Sánchez[1] supervised by Neil Yorke-Smith[1], Pascal van der Vaart[1]

[1]EEMCS, Delft University of Technology, The Netherlands

Abstract

This paper investigates how Random Network Distillation (RND), coupled with Boltzmann exploration, influences exploration behaviour and learning dynamics in value-based agents such as Deep Q-Network (DQN) across a range of environments, from classic control tasks to behaviour suite benchmarks and contextual bandits. The study addresses the sensitivity of RND to key hyperparameters, the impact of exploration strategy design, and the transferability of settings across tasks. The results reveal that RND remains benefitial within DQN in both sequential and non-sequential tasks, but requires careful tuning of reward scaling, temperature, and network capacity to be effective. No universal hyperparameter configuration generalises across environments, and inappropriate tuning can lead to unstable learning or suboptimal outcomes. These findings provide practical insights into the strengths and limitations of applying RND within value-based reinforcement learning frameworks.

Background

Reinforcement Learning (RL) [1] trains agents to maximize cumulative rewards through trial-and-error interaction with an environment. **Deep Q-Network (DQN)** [2] learns optimal behaviour by maximizing expected rewards, but struggles with exploration [3].

Random Network Distillation (RND) [4]: A method to improve exploration in environments with sparse or deceptive rewards by providing intrinsic motivation based on novelty. Originally proposed in combination with recurrent and convolutional policy networks, but not integrated into DQN.

Intrinsic Reward Mechanism: RND uses a fixed, randomly initialized target network f_{target} and a trainable predictor network $f_{\text{predictor}}$. The intrinsic reward is computed as the squared error between their outputs, the intrinsic and extrinsic rewards are combine, scaled by coefficients :

$$r_t^{\text{int}} = \|f_{\text{target}}(\hat{s}_t) - f_{\text{predictor}}(\hat{s}_t)\|^2$$

$$r_t = \beta_{\text{ext}} \cdot r_t^{\text{ext}} + \beta_{\text{int}} \cdot r_t^{\text{int}}$$

RND in Deep Q-Network: Integrating RND into DQN requires design choices to ensure compatibility with value-based learning. Boltzmann exploration [5] replaces traditional ϵ-greedy to better leverage intrinsic rewards, and intrinsic rewards are stored in the replay buffer to maintain stable learning.

Results and Discussion

1. Are Robust Settings Transferable Across Environments? Hyperparameter sweeps in `CartPole`, `Acrobot`, and `DeepSea` show no universally optimal configuration. Strong performance requires environment-specific tuning of key hyperparameters such as reward scaling and exploration strategy.

2. Which Hyperparameters Are Most Influential? Temperature (τ) and reward scaling critically affect both exploration stability and task performance. Poorly tuned τ leads to unstable learning or insufficient exploration, while imbalanced reward scaling hinders task completion.

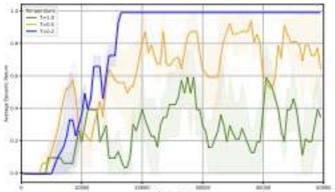

Figure 1. High temperatures cause instability in `DeepSea`.

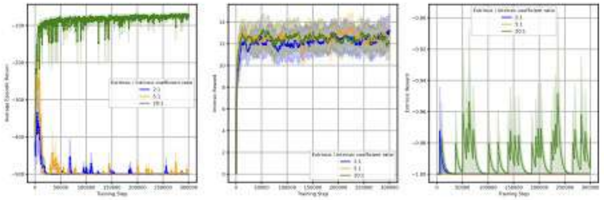

Figure 2. Correct reward scaling enables task completion in `Acrobot`.

3. How Well Does RND Integrate with Deep Q-Network? RND, combined with Boltzmann exploration, enhances directed exploration in value-based agents, even in sparse-reward settings like `DeepSea` and `Acrobot`. However, integration requires careful architectural and hyperparameter design to avoid destabilizing learning.

4. Does RND Improve Exploration in Contextual Bandits? In the `MNISTBandit` contextual badit, RND-augmented agents outperform ϵ-greedy baselines in classification accuracy and learning speed. This suggests that novelty-driven exploration provides benefits even in non-sequential, simplified environments.

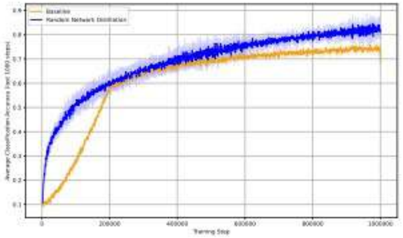

Figure 3. RND agents achieve faster learning in `MNISTBandit`.

Conclusions

Key Findings:

- RND improves exploration in value-based agents like DQN, enhancing learning in sparse-reward and deceptive environments.
- Strong performance requires **environment-specific tuning**; no single hyperparameter configuration generalizes across tasks.
- Boltzmann exploration outperforms ϵ-greedy by providing smoother, graded action selection that better complements intrinsic rewards.
- RND benefits extend to single-step tasks, as shown in the `MNISTBandit` contextual bandit setting.

Limitations & Future Work:

- Due to limited computational resources, large-scale experiments on textttMontezumaRevenge remain an open direction for future work.
- Fixed temperature and reward coefficient settings limit adaptability; future work should explore **temperature annealing** and **dynamic reward scaling**.
- The **relationship between network capacity, observation space, and intrinsic reward variance** warrants deeper investigation.
- More flexible **replay buffer strategies** could mitigate outdated exploration incentives.

References

[1] A. Sukovic and G. Radanovic, "Reward design for justifiable sequential decision-making," in *International Conference on Learning Representations (ICLR)*, 2024.

[2] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[3] K. Tan, "Exploration of obstacle tower using random network distillation.", 2019, Preprint.

[4] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] L. Pan, Q. Cai, Q. Meng, W. Chen, and L. Huang, "Reinforcement learning with dynamic boltzmann softmax updates," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 2785–2791.