Bayesian Ensembles for Exploration in Deep Q-Learning

(Extended Abstract)

Pascal R. van der Vaart
Delft University of Technology
Delft, Netherlands
p.r.vandervaart-1@tudelft.nl

Neil Yorke-Smith
Delft University of Technology
Delft, Netherlands
n.yorke-smith@tudelft.nl

Matthijs T. J. Spaan
Delft University of Technology
Delft, Netherlands
m.t.j.spaan@tudelft.nl

ABSTRACT

Exploration in reinforcement learning remains a difficult challenge. In order to drive exploration, ensembles with randomized prior functions have recently been popularized to quantify uncertainty in the value model. There is no theoretical reason for these ensembles to resemble the actual posterior, however. In this work, we view training ensembles from the perspective of Sequential Monte Carlo, a Monte Carlo method that approximates a sequence of distributions with a set of particles. In particular, we propose an algorithm that exploits both the practical flexibility of ensembles and theory of the Bayesian paradigm. We incorporate this method into a standard Deep Q-learning agent (DQN) and experimentally show qualitatively good uncertainty quantification and improved exploration capabilities over a regular ensemble.

KEYWORDS

Reinforcement Learning, Exploration, Bayesian, Uncertainty

ACM Reference Format:

Pascal R. van der Vaart, Neil Yorke-Smith, and Matthijs T. J. Spaan. 2024. Bayesian Ensembles for Exploration in Deep Q-Learning: (Extended Abstract). In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning (RL) algorithms are still notoriously sample inefficient. One pressing reason is the difficulty of exploring an environment efficiently while assuming little prior knowledge [32]. A promising approach that is currently studied is to quantify uncertainty in the value models learned by the agent, and then either provide intrinsic reward or use Thompson sampling to explore [1–4, 7, 14, 15, 21, 26, 27, 29]. However, quantifying uncertainty for deep neural networks is in itself a difficult task [18, 23].

Ensembles of neural networks have been shown to provide better predictive accuracy over a single model in supervised learning tasks [11, 20], as well as suitable methods for uncertainty quantification for exploration in reinforcement learning [13, 26, 27]. While ensembles with independent models of identical architecture tend to collapse to the same predictive model [17], several techniques have been developed to prevent this, such as adversarial learning [20], bootstrapping the data [27], and adding additive priors [26]. Furthermore, some techniques such as Stein Variational Gradient Descent [8, 22] alleviate this issue by interpreting the ensemble as an approximation to the Bayesian posterior and training it as such. The method that we propose falls into this last category and aims to be closer to the posterior for more accurate uncertainty quantification, while retaining the flexibility of ensembles.

Bayesian neural networks have been applied to RL through Variational Inference [14, 15, 30], as well as MCMC [1, 12, 19]. Due to the complex nature of Neural Networks, however, it is unclear how the model class in Variational inference biases uncertainty quantification. On the other hand, MCMC is in theory unbiased and also shows strong results in large networks in supervised learning [6, 16, 31]. However, MCMC methods such as Hamiltonian Monte Carlo [25] can struggle to find every mode for complex multimodal distributions [10]. This is an important drawback in deep learning, where the posterior distribution is likely very ill behaved, and especially in RL where under-approximation of the posterior complexity might lead to underestimating the uncertainty and therefore failure of exploration. Sequential Monte Carlo (SMC), which uses a set of particles to approximate the posterior, can be a remedy to these issues in non-deep learning applications [10].

In this work, we forego Variational Inference to avoid a decision in model class, and instead alleviate the issues in MCMC by using SMC. Noting the success of ensembles in deep learning, we unify ensembles and MCMC methods by using SMC algorithms to train an ensemble in a Bayesian manner, to benefit from both the practical effectiveness of ensembles and theoretical foundations of MCMC. Specifically, we adapt existing SMC algorithms to a minibatch setting, and show that they are feasible methods to train ensembles as approximations to the Bayesian posterior. Furthermore, as our main contribution, we introduce Sequential Monte Carlo DQN (SMC-DQN), a RL-algorithm which uses SMC to track a posterior over the Q-values and uses this posterior to explore efficiently. We experimentally test our agent's exploration capabilities on several environments, observing significantly stronger performance compared with regular ensembles, and results that are competitive with a strong baseline.

2 BACKGROUND

We consider standard Markov Decision Process with discounted rewards in the infinite horizon setting. Ensembles of Q-networks $Q_{\theta_1}(s,a), \ldots, Q_{\theta_n}(s,a)$ are a widespread method to improve exploration in an unknown environments. For example, the BootDQN algorithm [27] achieves deep exploration through Thompson sampling, sampling uniformly $i \in \{1, \ldots, n\}$ and acting greedily with respect to the network Q_{θ_i} for a full episode. In BootDQN each

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

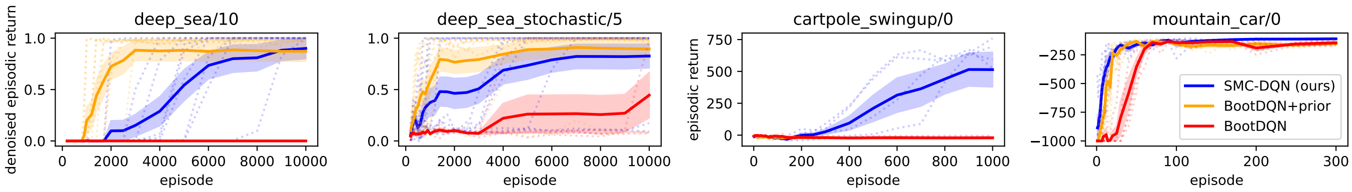

Figure 1: Learning curves over the BSuite environments. The solid line is the mean of 10 seeds for the Deep Sea environments, and 5 seeds for Cartpole Swingup and Mountain Car. The shaded area denotes the standard error of the mean.

network Q_{θ_i} is equipped with its own target network $Q_{\theta'_i}$, and gets updated using transitions (s, a, r, s') from a replay buffer:

$$\theta_i \leftarrow \theta_i - \nabla_{\theta_i} \left[Q_{\theta_i}(s,a) - r - \max_{a'} Q_{\theta'_i}(s', a') \right]^2 . \tag{1}$$

To ensure diversity in the ensemble, independently initialized prior functions are added to each ensemble member's outputs. However, while effective, this technique lacks theoretical motivation when considered as Bayesian priors for neural networks. In problems with well-defined likelihoods and priors, the Bayesian posterior can therefore be expected to outperform prior functions.

Sequential Monte Carlo algorithms sample a sequence of distributions $p_0(\theta), \dots, p_m(\theta)$. Leveraging this fact, we can set the target distributions to a sequence interpolating between the prior and posterior distribution over the parameters of a Q-learner: $p_t(\theta) \propto p(\mathcal{D}|\theta)^{\lambda_t} p(\theta)$, where λ_t is a sequence of temperatures $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_m = 1$, which can be dynamically optimised for [5, 9].

3 SEQUENTIAL MONTE CARLO DQN

To improve exploration, we construct an agent that quantifies uncertainty in its Q-values by approximating the posterior distribution over its parameters using SMC. Specifically, we extend a standard DQN agent [24] by replacing its point-wise estimator $Q_\theta(s,a)$ with an ensemble $Q_{\theta_1}(s,a), \dots, Q_{\theta_n}(s,a)$ and sampling weights $w_1, \dots w_n$ to maintain an approximation of the posterior $p(\theta|\mathcal{D}, \boldsymbol{\theta}')$, conditioned on the replay buffer $\mathcal{D} = ((s_t, a_t, r_t, s_{t+1}))_{t=1\dots N}$ and target parameters $\boldsymbol{\theta}' = (\theta'_1, \dots, \theta'_n)$. In line with Schmitt et al. [30], a normal distribution

$$Q_\theta(s,a) - r(s,a) - \gamma \max_{a'} Q_{\theta'}(s', a') \sim \mathcal{N}(0, \sigma)$$

is used as a probabilistic interpretation of the squared temporal difference error, and to represent the uncertainty in the targets we define the likelihood to be a mixture distribution

$$\log p(s,a,r,s'|\theta,\boldsymbol{\theta}') = \log \sum_{i=1}^{n} \frac{1}{n} \exp\left(-\frac{1}{2\sigma^2} [Q_\theta(s,a) - r(s,a) - \gamma \max_{a'} Q_{\theta'_i}(s', a')]^2 \right), \tag{2}$$

contrasting BootDQN which shares no target values between ensemble members. After collecting a new batch of trajectories $\mathcal{B}$ by acting in the environment, the posterior distribution can be updated efficiently by interpolating between the previous posterior $p(\theta|\boldsymbol{\theta}', \mathcal{D})$ and the new posterior $p(\theta|\boldsymbol{\theta}', \mathcal{D} \cup \mathcal{B})$ with SMC.

Updating the target networks $\boldsymbol{\theta}'$ changes the target distribution, meaning that the sample $(\theta_1, \dots, \theta_n, \; w_1, \dots, w_n)$ is no longer a sample of the posterior with respect to the updated targets, i.e., $p(\theta|\boldsymbol{\theta}'_{\text{new}}, \mathcal{D})$. Therefore, the typical target update $\theta'_i \leftarrow \theta_i$ is now accompanied by another SMC step, which interpolates between $p(\theta|\boldsymbol{\theta}'_{\text{old}}, \mathcal{D})$ and $p(\theta|\boldsymbol{\theta}'_{\text{new}}, \mathcal{D})$.

4 EXPERIMENTAL STUDY

We test our agent in the exploration environments as well as Mountain Car in BSuite [28], against BSuite's baseline BootDQN agent with and without prior. Figure 1 shows the performance of the agents on each task. It can be seen that SMC-DQN outperforms BootDQN without priors on all our benchmarks. On Deep Sea it achieves comparable performance to BootDQN with priors, and significantly outperforms BootDQN with priors on Cartpole Swingup, where BootDQN at this ensemble size fails to learn a meaningful policy even with prior functions. Further, on Mountain Car SMC-DQN learns at the same speed as BootDQN with priors in the beginning, but converges to a slightly better policy.

Our results show a gap between Deep Sea and the continuous environments in the performance relative to the baselines. We hypothesize that this is due to the fact that the likelihood does not explain the one-hot encoded environment Deep Sea very well. In the continuous environments, agents can exploit the generalization capabilities of neural networks, allowing the posterior to model sensible generalization behaviours. However, this generalization can lead to errors in one-hot encoded environments where unconnected states are independent.

5 CONCLUSION

We introduced the novel idea of using SMC to train an ensemble in order to approximate the Bayesian posterior distribution. Specifically, we modified the BootDQN algorithm to use SMC, thus keeping track of a posterior over the Q-values in a theoretically sound manner. We found that such an approach is able to maintain a diverse set of models that can drive exploration in difficult-to-explore environments such as Deep Sea and Cartpole Swingup. Especially in continuous state environments, the uncertainty quantification provided by the posterior distribution leads to better exploration compared to our baselines. In the future, we intend to investigate the influence of the choice of likelihood and derive methods to synthesize meaningful likelihoods.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreements 964505 (E-pi) and 952215 (TAILOR).

REFERENCES

[1] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. 2018. Efficient exploration through Bayesian Deep Q-Networks. In *2018 Information Theory and Applications Workshop (ITA)*. IEEE.

[2] Chenjia Bai, Lingxiao Wang, Lei Han, Jianye Hao, Animesh Garg, Peng Liu, and Zhaoran Wang. 2021. Principled Exploration via Optimistic Bootstrapping and Backward Induction. In *International Conference on Machine Learning*.

[3] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, Vol. 29.

[4] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random Network Distillation. In *International Conference on Learning Representations*.

[5] Michael Cai, Marco Del Negro, Edward Herbst, Ethan Matlin, Reca Sarfati, and Frank Schorfheide. 2021. Online Estimation of DSGE Models. *The Econometrics Journal* 24, 1 (2021), C33–C58.

[6] Tianqi Chen, Emily Fox, and Carlos Guestrin. 2014. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*.

[7] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. 2019. Better exploration with optimistic actor critic. In *Advances in Neural Information Processing Systems*, Vol. 32.

[8] Francesco D'Angelo and Vincent Fortuin. 2021. Repulsive Deep Ensembles are Bayesian. In *Advances in Neural Information Processing Systems*, Vol. 34.

[9] Hai-Dang Dau and Nicolas Chopin. 2022. Waste-free Sequential Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84, 1 (2022), 114–148.

[10] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. 2006. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 3 (2006), 411–436.

[11] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*. Springer Berlin Heidelberg, 1–15.

[12] Vikranth Dwaracherla and Benjamin Van Roy. 2021. Langevin DQN. arXiv:2002.07282 [cs.LG]

[13] Matthew Fellows, Kristian Hartikainen, and Shimon Whiteson. 2021. Bayesian Bellman Operators. In *Advances in Neural Information Processing Systems*, Vol. 34.

[14] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. 2019. Noisy Networks for Exploration. arXiv:1706.10295 [cs.LG]

[15] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in deep Learning. In *International Conference on Machine Learning*.

[16] Adrià Garriga-Alonso and Vincent Fortuin. 2021. Exact Langevin Dynamics with Stochastic Gradients. *arXiv:2102.01691* (2021).

[17] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. 2020. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment* 2020, 2 (2020).

[18] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* 110, 3 (2021), 457–506.

[19] Haque Ishfaq, Qingfeng Lan, Pan Xu, A. Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. 2023. Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo. arXiv:2305.18246 [cs.LG]

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, Vol. 30.

[21] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2021. Sunrise: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. In *International Conference on Machine Learning*.

[22] Qiang Liu and Dilin Wang. 2016. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, Vol. 29.

[23] Owen Lockwood and Mei Si. 2022. A Review of Uncertainty for Deep Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533.

[25] Radford M. Neal et al. 2011. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

[26] Ian Osband, John Aslanides, and Albin Cassirer. 2018. Randomized Prior Functions for Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 31.

[27] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems*, Vol. 29.

[28] Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado Van Hasselt. 2020. Behaviour Suite for Reinforcement Learning. In *International Conference on Learning Representations*.

[29] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. 2017. Count-Based Exploration with Neural Density Models. In *International Conference on Machine Learning*.

[30] Simon Schmitt, John Shawe-Taylor, and Hado van Hasselt. 2023. Exploration via Epistemic Value Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37.

[31] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. 2020. How Good is the Bayes Posterior in Deep Neural Networks Really?. In *International Conference on Machine Learning*.

[32] Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, Peng Liu, and Zhen Wang. 2021. Exploration in Deep Reinforcement Learning: From Single-Agent to Multiagent Domain. *IEEE Transactions on Neural Networks and Learning Systems* (2021).