# Bayesian Deep Q-Learning via Sequential Monte Carlo

**Pascal van der Vaart  Matthijs T.J. Spaan  Neil Yorke-Smith**
Delft University of Technology
p.r.vandervaart-1@tudelft.nl

## Abstract

Exploration in reinforcement learning remains a difficult challenge. Recently, ensembles with randomized prior functions have been popularized to quantify uncertainty in the value model, in order to drive exploration with success. However these ensembles have no theoretical guarantee to resemble the actual posterior. In this work, we view training ensembles from the perspective of sequential Monte Carlo, and propose an algorithm that exploits both the practical flexibility of ensembles and theory of the Bayesian paradigm. We incorporate this method into a standard DQN agent and experimentally show improved exploration capabilities over a regular ensemble.

## 1 Introduction

Reinforcement learning algorithms are still notoriously sample inefficient. One pressing reason is the difficulty of exploring an environment efficiently while assuming little prior knowledge. 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 [Bellemare et al., 2016, Ostrovski et al., 2017, Burda et al., 2018, Osband et al., 2016, 2018, Gal and Ghahramani, 2016, Fortunato et al., 2017]. However, quantifying uncertainty for deep neural networks is in itself a difficult task [Hüllermeier and Waegeman, 2021, Lockwood and Si, 2022].

Ensembles of neural networks have been shown to provide better predictive accuracy over a single model in supervised learning tasks [Dietterich, 2000, Lakshminarayanan et al., 2017], as well as suitable methods for uncertainty quantification for exploration in reinforcement learning [Osband et al., 2016, 2018, Fellows et al., 2021]. While ensembles with independent models of identical architecture tend to collapse to the same predictive model [Geiger et al., 2020], there are several techniques developed to prevent this, such as regularization, adversarial learning [Lakshminarayanan et al., 2017], bootstrapping the data [Osband et al., 2016], and adding additive priors [Osband et al., 2018]. Furthermore, some techniques such as Stein Variational Gradient Descent [Liu and Wang, 2016, D'Angelo and Fortuin, 2021] 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, while retaining the practicality of ensembles.

Bayesian neural networks can have desirable properties if the posterior can be inferred accurately. They have in theory optimal predictive accuracy given the correct likelihood and prior and also provide accurate uncertainty quantification. Unfortunately, exactly inferring the posterior is intractable already for some simple statistical models, and accurately approximating this posterior is very difficult for neural networks.

Typically, posterior approximation methods fall into one of two categories. Markov Chain Monte Carlo (MCMC) methods, which are theoretically unbiased but have high variance, and Variational Inference methods, which have low variance but are typically biased. Recently both types of methods have been altered specifically for application to neural networks, with MCMC showing strong results in large networks [Chen et al., 2014, Wenzel et al., 2020, Garriga-Alonso and Fortuin, 2021]. However, for complex multimodal distributions, MCMC methods can struggle to find every mode [Del Moral et al., 2006]. This is an important drawback in deep learning, where the posterior distribution is

16th European Workshop on Reinforcement Learning (EWRL 2023).

likely very ill behaved, and especially in reinforcement learning where under-approximation of the posterior complexity might lead underestimating the uncertainty and therefore failure of exploration. Sequential Monte Carlo can be a remedy to these issues in non-deep learning applications [Del Moral et al., 2006], and is also promising for deep learning noting the success of ensembles.

In this work, 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.

Our contributions are two-fold:

1. We introduce existing sequential Monte Carlo algorithms as feasible methods to train ensembles so that they serve as proper approximations to the Bayes posterior.
2. We modify the BootDQN algorithm [Osband et al., 2016] to use sequential Monte Carlo instead, which keeps track of a posterior over the Q-values in a theoretically sound manner.

## 2 Background

### 2.1 Markov Decision Processes

A Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ of a state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and discount factor $0 \leq \gamma < 1$. At each time step $t$, an agent observes the current state $s_t$, chooses an action $a_t \sim \pi(s_t)$ according to its policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and receives reward $r_t = R(s_t, a_t)$. The goal of reinforcement learning is to find a policy $\pi$ that maximizes the discounted cumulative reward $\mathbb{E}_{T,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. Of central importance is the Q-function

$$Q^\pi(s, a) = R(s, a) + \mathbb{E}_{T,\pi}\left[\sum_{t=1}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a\right], \tag{1}$$

denoting the expected discounted future reward if the agent executes action $a$ in state $s$ and then follows the policy $\pi$.

Since the transition function $T$ and reward function $R$ are assumed to be unknown to the agent, computing a strong policy requires exploration of the environment to learn which actions result in optimal return.

### 2.2 Bootstrapped DQN

A common strategy to find an optimal policy is Deep Q-learning (DQN), where a parameterized neural network $Q_\theta$ is learned and the corresponding policy picks $\operatorname{argmax}_a Q_\theta(s, a)$ in every state [Mnih et al., 2015]. To improve exploration, the BootDQN algorithm [Osband et al., 2016] leverages the uncertainty estimation abilities of ensembles. By learning an ensemble of $Q$-networks

$$Q_{\theta_1}(s, a), \ldots, Q_{\theta_n}(s, a)$$

the agent achieves deep exploration through Thompson sampling, sampling uniformly $i \in \{1, \ldots, n\}$ and acting greedily with respect to the network $Q_{\theta_i}$ for a full episode by picking $\operatorname{argmax}_a Q_{\theta_i}(s, a)$ in every state $s$. To update its predictions, each network $Q_{\theta_i}$ is equipped with their own target network $Q_{\theta'_i}$, and gets updated with their own targets:

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

where $s, a, r, s'$ are transitions sampled uniformly from a replay buffer.

Crucially, the ensemble $Q_{\theta_1}(s, a), \ldots, Q_{\theta_n}(s, a)$ has to stay diverse in underexplored states in order to keep exploration going. This can be achieved by bootstrapping the data for each ensemble member, or by using randomized prior functions. Randomized prior functions have been shown to be effective at keeping ensemble diversity [Osband et al., 2018].

A randomized prior function is a fixed prior function $Q_{\vartheta_i}(s, a)$ that is sampled independently for each ensemble member $Q_{\theta_i}(s, a)$, at the start of training. It remains unmodified during training and is added to the model outputs:

$$(Q_{\theta_i} + Q_{\vartheta_i})(s, a) = Q_{\theta_i}(s, a) + Q_{\vartheta_i}(s, a). \tag{3}$$

**Algorithm 1:** Base Sequential Monte Carlo

**Input:** sequence of target distributions $p_0(\theta), \dots, p_m(\theta)$ and MCMC kernels $P_1, \dots P_m$
**Result:** a sample $\theta_1, \dots, \theta_n \sim p_m(\theta)$
$\theta_1, \dots, \theta_n \sim p_0$ ; `/* Draw initial particles */`
$w_1, \dots, w_n \leftarrow 1$;
**for** $t = 1, \dots, m$ **do**
  **for** $j = 1, \dots n$ **do**
    $w_j \leftarrow \frac{p_t(\theta_j)}{p_{t-1}(\theta_j)}$ ; `/* Reweighting */`
  **end**
  **for** $j = 1, \dots n$ **do**
    $\theta_j \leftarrow P_t(\theta_j)$ ; `/* Apply Markov chain kernel */`
  **end**
  $\theta_1, \dots, \theta_n \sim$ `resample`$(\theta_1, \dots, \theta_n | w_1, \dots, w_n)$ ; `/* Resample with replacement */`
**end**

During action selection, the $Q$-values of ensemble member $i$ are given by $(Q_{\theta_i} + Q_{\vartheta_i})(s, a)$, and the BootDQN update rule is modified:

$$\theta_i \leftarrow \theta_i - \nabla_{\theta_i} \left[ (Q_{\theta_i} + Q_{\vartheta_i})(s, a) - r - \max_{a'} (Q_{\theta_i'} + Q_{\vartheta_i})(s', a') \right]^2. \tag{4}$$

Each ensemble member having a unique prior function causes unique generalization behaviour on unobserved data, which keeps the ensemble outputs diverse in underexplored states. Note that randomized prior functions are not unique to reinforcement learning and can be used in supervised learning settings as well.

However, randomized prior functions lack 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 methods that rely on randomized prior functions.

### 2.3 Bayesian Neural networks

A Bayesian Neural Network (BNN) is any neural network $f_\theta$ parameterized by $\theta \in \Theta$, with some prior distribution $p(\theta)$ over $\Theta$.

Given training data $(x_1, \dots, x_n)$ and labels $(y_1, \dots, y_n)$ i.i.d. from some likelihood $\mathcal{L}(y|f_\theta(x))$, the goal is to compute or sample from the posterior distribution over the parameter $\theta$.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} = \frac{\prod_{i=1}^n \mathcal{L}(y_i|f_\theta(x_i)p(\theta)}{\int \prod_{i=1}^n \mathcal{L}(y_i|f_\theta(x_i)p(\theta)d\theta} \tag{5}$$

Unfortunately, especially in the case of large neural networks, the posterior is intractable to compute or sample from exactly. Therefore it is necessary to resort to approximation methods such as variational inference and Markov Chain Monte Carlo.

### 2.4 Sequential Monte Carlo

Sequential Monte Carlo algorithms model a sequence of distributions $p_0(\theta), \dots, p_m(\theta)$. An initial state of particles is drawn from $p_0$, and by repeatedly applying importance sampling, a Markov chain Monte Carlo (MCMC) algorithm and resampling steps, the initial samples are transformed to be a sample of $p_n(\theta)$. A basic outline is given in Algorithnm 1. Under certain conditions, notably that sample distributions have to be invariant under the MCMC steps, this Monte Carlo scheme converges to the correct target [Del Moral et al., 2006].

Leveraging this fact, we can set $p_m(\theta) = p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ and pick a sequence of temperatures $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_n = 1$ and use SMC to sample from the sequence $p_0(\theta), \dots, p_m(\theta)$, where distribution is given by

$$p_t(\theta) \propto p(\mathcal{D}|\theta)^{\lambda_t} p(\theta). \tag{6}$$

This effectively interpolates between the prior and the posterior. Setting the temperature sequence correctly is important, because a too course interpolation can cause the importance sampling weights

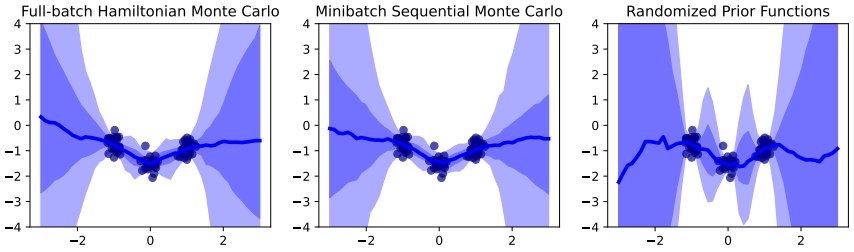


Figure 1: Sequential Monte Carlo and randomized prior functions in a supervised learning setting, compared against the gold standard Hamiltonian Monte Carlo without noise. Sequential Monte Carlo more closely resembles the posterior distribution as approximated by Hamiltonian Monte Carlo.

to be unstable and a too fine interpolation wastes computation. Fortunately, automatic on the fly tuning methods exist that choose the next temperature based on the effective sample size of the current sample [Cai et al., 2021, Dau and Chopin, 2022].

## 3 Sequential Monte Carlo for Bayesian Neural networks

Applying SMC to model the posterior of a Bayesian neural network is in practice similar to an ensemble. The particles $\theta_1, \ldots, \theta_n$ are individual models, and equipped with importance sampling weights $w_1, \ldots, w_n$ model an unbiased approximation of the predictive posterior distribution.

The model is initialized by sampling initial parameters $\theta_1, \ldots, \theta_n$ from the prior $p(\theta)$, and trained by running Adaptive SMC with target distributions $p(\mathcal{D}|\theta)^{\lambda_t} p(\theta)$. Unfortunately, typically in deep learning the data set $\mathcal{D}$ is so large that mini batches are required to tractably compute the likelihood and gradients. This poses two potential problems:

1. The reweighting step is now noisy, which lowers the quality of importance sampling.
2. The MCMC step, which typically is gradient based, is now noisy. This means we may have to rely on MCMC methods that only approximately leave the target distribution invariant.

However, these problems can largely be alleviated, as the theoretical results by Llorente et al. [2022] show that noisy importance sampling has higher variance but remains unbiased. Furthermore, Wenzel et al. [2020] and Garriga-Alonso and Fortuin [2021] show that accurate noisy MCMC schemes that leave the target distribution (approximately) invariant do exist.

Specifically, in our experiments we estimate the reweighting steps by sampling a single batch independently for every particle every iteration. As MCMC kernel we use the Symplectic Euler Langevin scheme with hyper-parameters as suggested by Wenzel et al. [2020]. At each iteration $t$, the temperature in the next step $\lambda_{t+1} = \lambda_t + \delta$ is picked by keeping the effective sample size (ESS) at a desired level $d$:

$$\max_{\delta} \delta, \text{ such that: } \delta < 1 - \lambda_t, \text{ and}$$
$$\text{ESS} = \frac{\left(\sum_{j=1}^n w_j\right)^2}{\sum_{j=1}^n w_j^2} = \frac{\left(\sum_{j=1}^n p(\mathcal{D}|\theta_j)^\delta\right)^2}{\sum_{j=1}^n p(\mathcal{D}|\theta_j)^{2\delta}} > d. \tag{7}$$

Our initial experimental findings in a supervised learning setting, shown in Figures 1 and 2, show that even with very small mini batches, SMC with noisy likelihoods and gradients results in performance on par with the gold standard noise-free Hamiltonian Monte Carlo [Neal et al., 2011], albeit at the cost of requiring a finer grained temperature schedule for lower batch sizes. This is to be expected, as the temperature schedule depends on maintaining a high enough effective sample size, which is known to be lower when using noisy reweighting [Llorente et al., 2022].

## 4 Sequential Monte Carlo DQN

The ensemble of a BootDQN agent can be replaced with an ensemble trained by sequential Monte Carlo with minor changes to the architecture. Most notably, along with the parameters of the ensemble $\theta_1, \ldots, \theta_n$, the agent now also stores the importance sampling weights $w_1, \ldots, w_n$.

**Algorithm 2:** Sequential Monte Carlo for BNNs

**Input:** Prior $p_0(\theta)$ and target $p_0(\theta)p(\mathcal{D}|\theta)$, MCMC algorithm
**Result:** A sample $\theta_1, \ldots, \theta_n \sim p_0(\theta)p(\theta|\mathcal{D})$
$\theta_1, \ldots, \theta_n \sim p_0$ ; `/* Draw initial particles */`
$w_1, \ldots, w_n \leftarrow 1$;
$t \leftarrow 0$;
$\lambda_0 \leftarrow 0$;
**while** $\lambda_t < 1$ **do**
  Pick $\lambda_t > \lambda_{t-1}$ ; `/* Equation 7 */`
  $\log p_t(\theta) \leftarrow \log p_0(\theta) + \lambda_t \log p(\mathcal{D}|\theta)$;
  **for** $j = 1, \ldots n$ **do**
    $\log w_j \leftarrow \lambda_t \log p(\theta_j|\mathcal{D})$ ; `/* Reweighting, evaluating likelihood */`
  **end**
  $w_1, \ldots, w_n = \texttt{normalize}(w_1, \ldots, w_n)$;
  **for** $j = 1, \ldots n$ **do**
    $\theta_j \leftarrow \texttt{MCMC}(\theta_j, \log p_t(\theta))$ ; `/* Apply Markov chain kernel */`
  **end**
  $\theta_1, \ldots, \theta_n \sim \texttt{resample}(\theta_1, \ldots, \theta_n | w_1, \ldots, w_n)$ ; `/* Resample */`
  $t \leftarrow t + 1$;
**end**

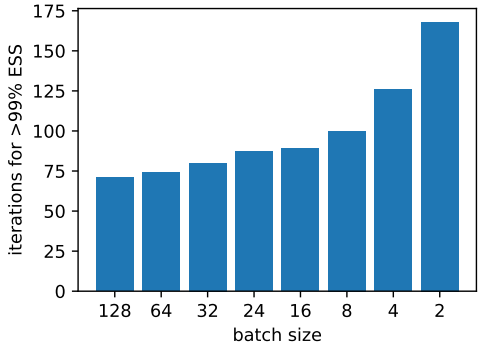


Figure 2: Number of interpolation iterations required in Sequential Monte Carlo to achieve 99% effective sample size for decreasing batch size, in the same experimental setup as Figure 1.

Equipped with a likelihood and prior, the agent attempts to maintain a posterior over the parameters of a Q-network, conditioned on the current replay buffer $\mathcal{D} = ((s_t, a_t, r_t, s_{t+1}))_{t=1\ldots N}$ and current target parameters $\theta'_1, \ldots, \theta'_n$. This is achieved by defining a joint likelihood over $(\theta, \theta')$. In line with [Schmitt et al., 2023], a normal distribution $Q_\theta(s,a) - r(s,a) - \operatorname{argmax}_{a'} Q_{\theta'}(s',a') \sim \mathcal{N}(0,\sigma)$ is used as a probabilistic interpretation of the squared temporal difference error, resulting in the log-likelihood

$$\log \mathcal{L}(\theta, \theta', \mathcal{D}) = -\frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[ Q_\theta(s_t, a_t) - r_t - \operatorname*{argmax}_{a'} Q_{\theta'}(s_{t+1}, a') \right]^2. \tag{8}$$

The log posterior distribution is defined then as

$$\log p(\theta|\theta', \mathcal{D}) \propto \log p(\theta) + \log \mathcal{L}(\theta, \theta', \mathcal{D}). \tag{9}$$

**Algorithm 3:** Sequential DQN

**Input:** MDP, batch size $B$, ensemble size $n$
**Result:** Posterior over $Q_\theta(s,a)$

$\theta_1,\dots,\theta_n \sim p(\theta)$ ; `/* Initialize networks */`
$w_1,\dots,w_n \leftarrow 1$;
**while** *training* **do**
  $i \sim \texttt{uniform}(1,\dots,n)$;
  $s_0 \sim \texttt{MDP}$;
  $t \leftarrow 0$;
  **while** *episode not done* **do**
    $a_t = \operatorname{argmax}_a Q_{\theta_i}(s_t)$ ; `/* TS sample for full episode */`
    $r_t, s_{t+1} \sim \texttt{MDP}(s_t, a_t)$;
    $\mathcal{B} = \mathcal{B} \cup \{(s_t, a_t, r_t, s_{t+1})\}$
    **if** $|\mathcal{B}| = B$ **then**
      $\theta_1,\dots,\theta_n, w_1,\dots,w_n \leftarrow \texttt{SMC}(p(\theta|\theta',\mathcal{D}), p(\theta|\theta',\mathcal{D}\cup\mathcal{B}))$ ; `/* main update */`
      $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{B}$;
      $\mathcal{B} \leftarrow \emptyset$
    **end**
    **if** *time for target update* **then**
      $\theta'_{i,\texttt{new}} \leftarrow \theta_i$;
      $\theta_1,\dots,\theta_n, w_1,\dots,w_n \leftarrow \texttt{SMC}(p(\theta|\theta',\mathcal{D}), p(\theta|\theta'_{\texttt{new}},\mathcal{D}))$ ; `/* target update */`
      $\theta'_i \leftarrow \theta'_{i,\texttt{new}}$
    **end**
  **end**
**end**

After collecting a batch of trajectories $\mathcal{B} = ((s_t, a_t, r_t, s_{t+1}))_{t=1,\dots,B}$, the posterior distribution can be updated efficiently by noting that

$$\begin{aligned}
\log p(\theta|\theta',\mathcal{D}\cup\mathcal{B}) &= \log p(\theta) + \log \mathcal{L}(\theta,\theta',\mathcal{D}\cup\mathcal{B}) \\
&= \log p(\theta) + \sum_{t=1}^{N+B} -\frac{1}{2\sigma^2}\left[Q_\theta(s_t,a_t) - r_t - \operatorname*{argmax}_{a'} Q_{\theta'}(s_{t+1},a')\right]^2 \\
&= \log p(\theta) + \log \mathcal{L}(\theta,\theta',\mathcal{D}) \\
&\quad + \sum_{t=N}^{N+B} -\frac{1}{2\sigma^2}\left[Q_\theta(s_t,a_t) - r_t - \operatorname*{argmax}_{a'} Q_{\theta'}(s_{t+1},a')\right]^2 .
\end{aligned} \tag{10}$$

Therefore, since our agent is currently holding a sample of $p(\theta|\theta',\mathcal{D})$, the posterior can be updated by running SMC on the sequence

$$p(\theta|\theta',\mathcal{D}),\ p(\theta|\theta',\mathcal{D})\mathcal{L}(\theta,\theta',\mathcal{B})^{\lambda_1},\dots,p(\theta|\theta',\mathcal{D})\mathcal{L}(\theta,\theta',\mathcal{B})^{\lambda_k},\ p(\theta|\theta',\mathcal{D})\mathcal{L}(\theta,\theta',\mathcal{B}), \tag{11}$$

interpolating between the posterior given what was already known, and the new posterior including the new batch. The temperatures $\lambda_1,\dots\lambda_k$ can be tuned on the fly, depending on how novel the new batch is.

Evaluating the likelihood $\mathcal{L}(\theta,\theta',\mathcal{B})$ only requires the latest batch, and can easily be computed exactly. However, we choose to approximate $p(\theta|\theta',\mathcal{D})$ by sampling mini batches uniformly from the replay buffer, since computing it exactly would require summing over the entire replay buffer.

Updating the target networks naively changes the target distribution, meaning that the weights $w_1,\dots w_n$ are incorrect, and the sample

$$\theta_1,\dots,\theta_n,\quad w_1,\dots,w_n$$

is no longer a sample of the posterior with respect to the updated targets, i.e., $p(\theta|\theta'_{\texttt{new}},\mathcal{D})$. Therefore, the typical target update $\theta'_i \leftarrow \theta_i$ is now accompanied by another sequential Monte Carlo step, which

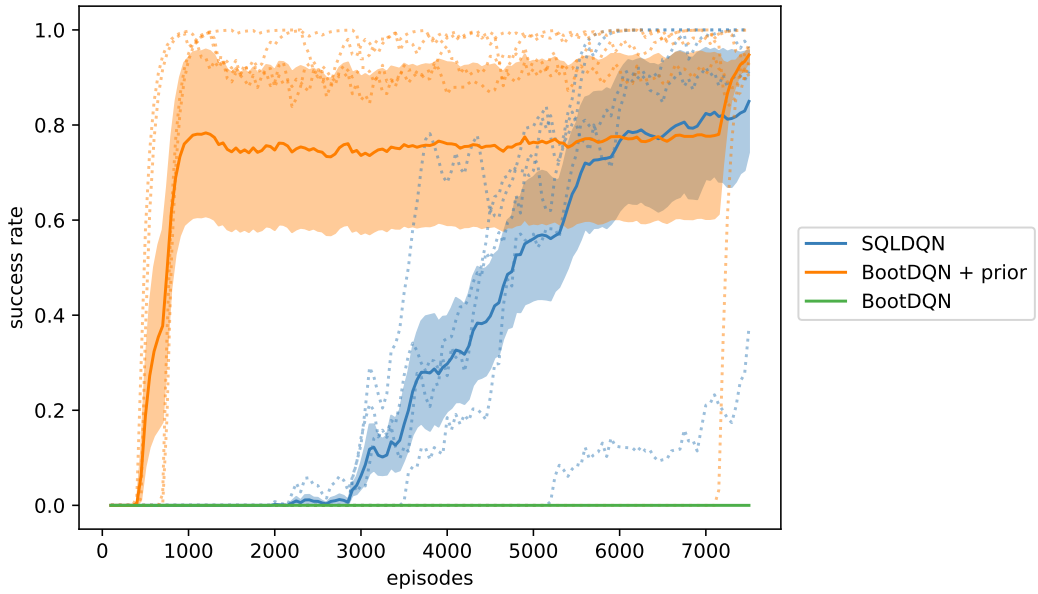


Figure 3: The fraction of runs where the agent reaches the goal state at each episode, averaged over 50 episodes. The shaded area shows the standard error of the mean over 5 seeds. Dotted lines show individual seeds. The agent with an ensemble of DQNs can not solve this environment.

smoothly interpolates between the distribution conditioned on the old targets and the distribution conditioned on the new targets via the sequence

$$p(\theta)\mathcal{L}(\theta, \theta'_{\texttt{old}}, \mathcal{D})^{(1-\lambda_t)}\mathcal{L}(\theta, \theta'_{\texttt{new}}, \mathcal{D})^{\lambda_t} \to p(\theta)\mathcal{L}(\theta, \theta'_{\texttt{new}}, \mathcal{D}), \tag{12}$$

as $\lambda_t \to 1$. This transforms a sample of the posterior with respect to the previous targets to a sample with respect to the new targets. Intuitively speaking, this trains the main networks $\theta_1, \ldots, \theta_n$ to immediately match the new targets.

Apart from the novel parameter updates and target network updates, Sequential Monte Carlo DQN (SQL DQN) makes no further alterations to BootDQN. Algorithm 3 shows the general structure with both updates, where `SMC` refers to Algorithm 2. The symplectic euler Langevin scheme [Wenzel et al., 2020] is used as MCMC step. While the structure is similar to a typical DQN implementation, a major difference is that the SMC steps takes significantly more time than typical Q-value updates. For each batch from the environment, the agent fully adapts its model to match the posterior given the new replay buffer, and also after each target update the agent trains its networks to match the posterior given the new targets. Furthermore, there is a very clear split between samples that are already in the data set $\mathcal{D}$ and samples that have yet to be incorporated. It is assumed that the replay buffer $\mathcal{D}$ is large enough to store every experienced transition.

## 5 Experiments

We evaluate our agent on the $30 \times 30$ Deep Sea environment [Osband et al., 2016, 2020]. This environment is a grid, where the agent starts in the top left corner and at each time step, the agent drops down one row and either moves left or right. The agent receives small negative reward when moving to the right, except when reaching the bottom right state, where the agent obtains reward 1. The optimal policy is to move right along the diagonal to reach this goal state, and a single mistake in an episode will cause the agent to not receive any reward. This environment requires a dedicated exploration method, as it would take a randomly exploring agent on average $2^{30}$ episodes to observe a positive reward.

Because updating the models and targets is more computationally expensive than in regular BootDQN, the agent uses a relatively large batch size of 1024, meaning that the models are only updated approximately every 34 episodes. Furthermore, because the target update step also updates the main models alongside the targets, the target networks can be updated after every main update without

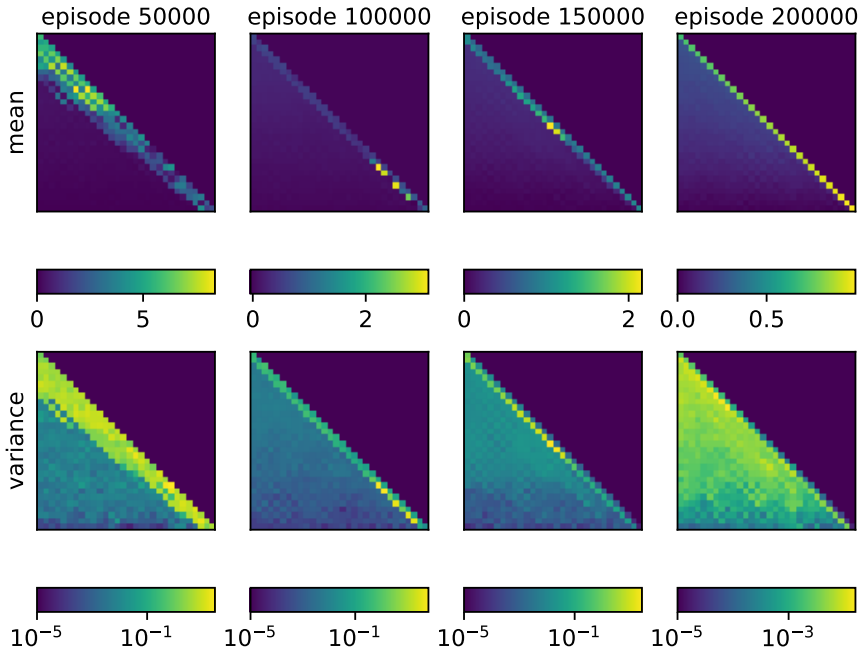


Figure 4: Graphical view the mean and variance of the values during training on Deep sea. Each pixel shows the mean (top row) or variance (bottom row) of that state as predicted by the ensemble. The top left state is the initial state, and the bottom right state is the goal state. The diagonal is the only sequence of states that results in reward. The states in the upper triangle are unreachable. Note the different scale of the color bars across columns.

causing instability. To further increase the diversity of collected trajectories, the active $Q$-network is sampled uniformly from the particles instead of from the distribution. Furthermore, no resampling is performed during SMC in order to maintain more diversity in the target networks.

Figure 3 shows the average success rate, i.e., the number of times the agent reaches the goal state, during training. Crucially, this is the success rate of the Thompson sampling policy which still explores up until all ensemble members agree on the correct path. It can be seen that it takes a long time for the entire ensemble to agree on the correct action sequence. Even after observing the goal, it takes several thousands more episodes to fully converge to the optimal policy. This could be improved upon by switching to a more optimistic evaluation, likelihood or prior, but not without assuming further information about the environment. For example, in this experiment the agent is not aware that the rewards and trajectories are in fact deterministic given the action sequence. In figure 4 this problem is very apparent: even after convergence at 200000 episodes, the value is still reasonably high for early but non-optimal states. This is because the aleatoric uncertainty in the likelihood is compounded from the end to the start of the trajectory, and the argmax in the targets biases the value upwards. BootDQN with prior functions does not suffer from this as much as it simply minimizes the mean squared error without taking a probabilistic point of view. It can be seen clearly in Figure 3 that a well-tuned variant of BootDQN with prior functions outperforms SQL DQN significantly in this scenario. Taking this probabilistic point of view as in SQL DQN may be an advantage in stochastic scenarios, but we leave further comparisons on these environments for future work.

Qualitatively speaking, Figure 4 shows what should be expected of a posterior. The variance decreases as states are explored more often, and the variance decreases faster for states that are closer to the end of the environment. Furthermore, the values are learned bottom up, and the ensemble stays optimistic about states that have not been visited often. This allows the agent to eventually solve the

environment, albeit at more steps than BootDQN with prior functions. At episode 200 000, the mean value prediction is close to $\gamma^n$ for $n = 29, \ldots, 0$ along the diagonal, which is the true value, and the variance is low for every state.

## 6 Conclusion

This paper studied the novel idea of how sequential Monte Carlo can be used to train an ensemble to approximate the Bayesian posterior distribution. We modified the BootDQN algorithm to use sequential Monte Carlo, 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. The algorithm developed here is motivated from a theoretical standpoint, but our findings are that deviating from the theoretically-sound algorithm by not performing resampling and uniformly sampling particles for action selection *improves* performance in reinforcement learning. This could be caused by the likelihood or prior not describing the environment well, especially in deterministic environments such as Deep Sea. In the future, we will extend our experiments to stochastic and continuous environments,

## 7 Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 964505 (E-pi). We thank the reviewers for useful feedback and thank Yaniv Oren and Moritz Zanger for insightful discussions.

## References

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. *Advances in Neural Information Processing Systems*, 2016.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

M. Cai, M. Del Negro, E. Herbst, E. Matlin, R. Sarfati, and F. Schorfheide. Online estimation of dsge models. *The Econometrics Journal*, 24(1):C33–C58, 2021.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, 2014.

F. D'Angelo and V. Fortuin. Repulsive deep ensembles are bayesian, 2021.

H.-D. Dau and N. Chopin. Waste-free sequential monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):114–148, 2022.

P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer Berlin Heidelberg, 2000.

M. Fellows, K. Hartikainen, and S. Whiteson. Bayesian bellman operators. In *Advances in Neural Information Processing Systems*, 2021.

M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, et al. Noisy networks for exploration. *arXiv:1706.10295*, 2017.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

A. Garriga-Alonso and V. Fortuin. Exact langevin dynamics with stochastic gradients. *arXiv:2102.01691*, 2021.

M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

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

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm, 2016.

F. Llorente, L. Martino, J. Read, and D. Delgado-Gómez. Optimality in noisy importance sampling. *Signal Processing*, 194:108455, 2022.

O. Lockwood and M. Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2022.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb. 2015. ISSN 00280836.

R. M. Neal et al. Mcmc using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

I. Osband, C. Blundell, A. Pritzel, and B. V. Roy. Deep exploration via bootstrapped dqn, 2016.

I. Osband, J. Aslanides, and A. Cassirer. Randomized prior functions for deep reinforcement learning, 2018.

I. Osband, Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepesvari, S. Singh, B. V. Roy, R. Sutton, D. Silver, and H. V. Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020.

G. Ostrovski, M. G. Bellemare, A. Oord, and R. Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning*, 2017.

S. Schmitt, J. Shawe-Taylor, and H. van Hasselt. Exploration via epistemic value estimation, 2023.

F. Wenzel, K. Roth, B. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, 2020.

# Appendices

## A Implementation details

### A.1 Supervised learning

The neural network is a fully connected network with two hidden layers of size $100$ and $10$. The data is generated by sampling $x$ i.i.d. from a uniform mixture of three normal distributions with means $-1$, $0$, and $1$ and standard deviation $0.25$. The labels $y$ are generated by randomly sampling a neural network $\theta_{\texttt{prior}}$ from the prior and setting $y_i = f_{\theta_{\texttt{prior}}}(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.25)$.

The hyperparameters for SMC are shown in Table 1, where the batch size is varied for the results shown in Figure 2. For randomized prior functions, the prior function was drawn from the same prior as used for SMC and HMC.

| Sequential Monte Carlo hyperparameters | |
|---|---|
| $p(\theta)$ (prior) | i.i.d. $\mathcal{N}(0,1)$ |
| $\sigma$ (likelihood std) | 0.1 |
| $B$ (batch size) | 1024 |
| MCMC steps (main) | 100 |
| MCMC steps (target) | 100 |
| **Reinforcement Learning hyperparameters** | |
| Total episodes | 7500 |
| Buffer size | $\infty$ |
| Main update frequency | $B$ |
| Target update frequency | $B$ |
| **Symplectic Euler Langevin Scheme hyperparameters** | |
| $\ell$ (learning rate) | $10^{-3}$ |
| cycle length | 20 |
| $\beta$ | 0.98 |
| $h$ (step size) | $\sqrt{\frac{\ell}{\mathtt{n_{data}}}}$ |
| $\gamma$ | $\frac{1}{h}$ |

Table 2: Hyper-parameters of the Deep Sea environment experiments.

| Sequential Monte Carlo hyperparameters | |
|---|---|
| $p(\theta)$ (prior) | i.i.d. $\mathcal{N}(0,1)$ |
| $\sigma$ (likelihood std) | 0.1 |
| $B$ (batch size) | 32 |
| MCMC steps (main) | 10 |
| MCMC steps (target) | 10 |
| **Symplectic Euler Langevin Scheme hyperparameters** | |
| $\ell$ (learning rate) | $10^{-3}$ |
| cycle length | 20 |
| $\beta$ | 0.98 |
| $h$ (step size) | $\sqrt{\frac{\ell}{\mathtt{n_{data}}}}$ |
| $\gamma$ | $\frac{1}{h}$ |

Table 1: Hyperparameters for SMC in the supervised learning experiments shown in Figures 1 and 2.

### A.2 Deep Sea

The environment is the `deepsea/10` environment as implemented in BSuite [Osband et al., 2020], which is $30 \times 30$ and uses a random action mapping.

The Q-value network is the same as the baseline BootDQN agent in BSuite, which is fully connected with two hidden layers of size $50$. Action selection is also unchanged and done by picking one ensemble member uniformly at random, and acting greedily with respect to that network for the rest of the episode.

The hyperparameters used are shown in Table 2.

The hyperparameters for BootDQN + prior and BootDQN are the default hyperparameters in BSuite.