Bayesian Bellman Operators

Matthew Fellows* **Kristian Hartikainen** **Shimon Whiteson**
Department of Computer Science
University of Oxford

Abstract

We introduce a novel perspective on Bayesian reinforcement learning (RL); whereas existing approaches infer a posterior over the transition distribution or Q-function, we characterise the uncertainty in the Bellman operator. Our Bayesian Bellman operator (BBO) framework is motivated by the insight that when bootstrapping is introduced, model-free approaches actually infer a posterior over Bellman operators, not value functions. In this paper, we use BBO to provide a rigorous theoretical analysis of model-free Bayesian RL to better understand its relationship to established frequentist RL methodologies. We prove that Bayesian solutions are consistent with frequentist RL solutions, even when approximate inference is used, and derive conditions for which convergence properties hold. Empirically, we demonstrate that algorithms derived from the BBO framework have sophisticated deep exploration properties that enable them to solve continuous control tasks at which state-of-the-art regularised actor-critic algorithms fail catastrophically.

1 Introduction

A Bayesian approach to reinforcement learning (RL) characterises uncertainty in the Markov decision process (MDP) via a posterior [35, 78]. A great advantage of Bayesian RL is that it offers a natural and elegant solution to the exploration/exploitation problem, allowing the agent to explore to reduce uncertainty in the MDP, but only to the extent that exploratory actions lead to greater expected return; unlike in heuristic strategies such as ε-greedy and Boltzmann sampling, the agent does not waste samples trying actions that it has already established are suboptimal, leading to greater sampling efficiency. Elementary decision theory shows that the only admissible decision rules are Bayesian [22] because a non-Bayesian decision can always be improved upon by a Bayesian agent [24]. In addition, pre-existing domain knowledge can be formally incorporated by specifying priors.

In model-free Bayesian RL, a posterior is inferred over the Q-function by treating samples from the MDP as stationary labels for Bayesian regression. A major theoretical issue with existing model-free Bayesian RL approaches is their reliance on bootstrapping using a Q-function approximator, as samples from the exact Q-function are impractical to obtain. This introduces error as the samples are no long estimates of a Q-function and their dependence on the approximation is not accounted for. It is unclear what posterior, if any, these methods are inferring and how it relates to the RL problem.

In this paper, we introduce Bayesian Bellman Operators (BBO), a novel model-free Bayesian RL framework that addresses this issue and facilitates a theoretical exposition of the relationship between model-free Bayesian and frequentist RL approaches. Using our framework, we demonstrate that, by bootstrapping, model-free Bayesian RL infers a posterior over *Bellman operators*. For our main contribution, we prove that the BBO posterior concentrates on the true Bellman operator (or the closest representation in our function space of Bellman operators). Hence a Bayesian method using the BBO posterior is consistent with the equivalent frequentist solution in the true MDP. We derive convergent gradient-based approaches for Bayesian policy evaluation and uncertainty estimation. Remarkably, our consistency and convergence results still hold when approximate inference is used.

*Correspondence to matthew.fellows@cs.ox.ac.uk

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Our framework is general and can recover empirically successful algorithms such as BootDQNprior+ [57]. We demonstrate that BootDQNprior+'s lagged target parameters, which are essential to its performance, arise from applying approximate inference to the BBO posterior. Lagged target parameters cannot be explained by existing model-free Bayesian RL theory. Using BBO, we extend BootDQNprior+ to continuous domains by developing an equivalent Bayesian actor-critic algorithm. Our algorithm can learn optimal policies in domains where state-of-the-art actor-critic algorithms like soft actor-critic [39] fail catastrophically due to their inability to properly explore.

2 Bayesian Reinforcement Learning

To aid the reader with notation, we provide a mathematical glossary in Appendix A.

2.1 Preliminaries

Formally, an RL problem is modelled as a Markov decision process (MDP) defined by the tuple $\langle \mathcal{S}, \mathcal{A}, r, P, P_0, \gamma \rangle$ [72, 60], where $\mathcal{S}$ is the set of states and $\mathcal{A}$ the set of available actions. At time t, an agent in state $s_t \in \mathcal{S}$ chooses an action $a_t \in \mathcal{A}$ according to the policy $a_t \sim \pi(\cdot|s_t)$. The agent transitions to a new state according to the state transition distribution $s_{t+1} \sim P(\cdot|s_t, a_t)$ which induces a scalar reward $r_t \coloneqq r(s_{t+1}, a_t, s_t) \in \mathbb{R}$ with $\sup_{s',a,s}|r(s', a, s)| < \infty$. The initial state distribution for the agent is $s_0 \sim P_0$ and the state-action transition distribution is denoted as $P^\pi(s', a'|s, a)$, which satisfies $dP^\pi(s', a'|s, a) = d\pi(a'|s')dP(s'|s, a)$.. As the agent interacts with the environment it gathers a trajectory: $(s_0, a_0, r_0, s_1, a_1, r_1, s_2...)$. We seek an optimal policy $\pi^* \in \arg\max_\pi J^\pi$ that maximises the total expected discounted return: $J^\pi \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r_t\right]$ where $\mathbb{E}_\pi$ is the expectation over trajectories induced by π. The Q-function is the total expected reward as a function of a state-action pair: $Q^\pi(s, a) \coloneqq \mathbb{E}_\pi[\sum_{t=0}^\infty r_t|s_0 = s, a_0 = a]$. Any Q-function satisfies the Bellman equation $\mathcal{B}[Q^\pi] = Q^\pi$ where the Bellman operator is defined as:

$$\mathcal{B}[Q^\pi](s, a) \coloneqq \mathbb{E}_{P^\pi(s',a'|s,a)} \left[r(s', a, s) + \gamma Q^\pi(s', a')\right]. \tag{1}$$

2.2 Model-based vs Model-free Bayesian RL

Bayes-adaptive MDPs (BAMDPs) [27] are a framework for *model-based* Bayesian reinforcement learning where a posterior marginalises over the uncertainty in the unknown transition distribution and reward functions to derive a Bayesian MDP. BAMDP optimal policies are the gold standard, optimally balancing exploration with exploitation but require learning a model of the unknown transition distribution which is typically challenging due to its high-dimensionality and multi-modality [67]. Furthermore, planning in BAMDPs requires the calculation of high-dimensional integrals which render the problem intractable. Even with approximation, most existing methods are restricted to small and discrete state-action spaces [6, 38]. One notable exception is VariBAD [82] which exploits a meta learning setting to carry out approximate Bayesian inference. Unfortunately this approximation sacrifices the BAMDP's theoretical properties and there are no convergence guarantees.

Existing model-free Bayesian RL approaches attempt to solve a Bayesian regression problem to infer a posterior predictive over a value function [78, 35]. Whilst foregoing the ability to separately model reward uncertainty and transition dynamics, modelling uncertainty in a value function avoids the difficulty of estimating high dimensional conditional distributions and mimics a Bayesian regression problem, for which there are tractable approximate methods [44, 10, 47, 61, 33, 51]. These methods assume access to a dataset of N samples: $\mathcal{D}^N \coloneqq \{q_i\}_{i=1:N}$ from a distribution over the true Q-function at each state-action pair: $q_i \sim P_Q(\cdot|s_i, a_i)$. Each sample is an estimate of a point of the true Q-function $q_i = Q^\pi(s_i, a_i) + \eta_i$ corrupted by noise η_i. By introducing a probabilistic model of this random process, the posterior over the Q-function $P(Q^\pi|s, a, \mathcal{D}^N)$ can be inferred, which characterises the aleatoric uncertainty in the sample noise and epistemic uncertainty in the model. Modeling aleatoric uncertainty is the goal of distributional RL [11]. In Bayesian RL we are more concerned with epistemic uncertainty, which can be reduced by exploration [57].

2.3 Theoretical Issues with Existing Approaches

Unfortunately for most settings it is impractical to sample directly from the true Q-function. To obtain efficient algorithms the samples q_i are approximated using bootstrapping: here a parametric function approximator $\hat{Q}_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ parametrised by $\omega \in \Omega$ is learnt as an approximation of the Q-function $\hat{Q}_\omega \approx Q^\pi$ and then a TD sample is used in place of q_i. For example a one-step TD estimate approximates the samples as: $q_i \approx r_i + \gamma \hat{Q}_\omega(s_i, a_i)$, introducing an error that is dependent on ω. Existing approaches do not account for this error's dependency on the function approximator.

Samples are no longer noisy estimates of a point $Q^\pi(s_i, a_i)$ and the resulting posterior predictive is not $P(Q^\pi|s, a, \mathcal{D}^N)$ as it has dependence on $\hat{Q}_\omega$ due to the dataset. This problem is made worse when a posterior is inferred over an optimal Q-function as it is impossible to sample from the optimal policy a priori to obtain unbiased samples. This is a major theoretical issue that raises the following questions:

1. Do model-free Bayesian RL approaches that use bootstrapping still infer a posterior?
2. If it exists, how does this posterior relate to solving the RL problem?
3. What effect does approximate inference have on the solution?
4. Do methods that sample from an approximate posterior converge?

Contribution: Our primary contribution is to address these questions by introducing the BBO framework. In answer to Question 1, BBO shows that, by introducing bootstrapping, we actually infer a posterior over Bellman operators. We can use this posterior to marginalise over all Bellman operators to obtain a Bayesian Bellman operator. Our theoretical results provide answers to Questions 2-4, proving that the Bayesian Bellman operator can parametrise a TD fixed point as the number of samples $N \to \infty$ and is analogous to the projection operator used in convergent reinforcement learning. Our results hold even under posterior approximation. Although our contributions are primarily theoretical, many of the benefits afforded by Bayesian methods play a significant role in a wide range of real-world applications of RL where identifying decisions that are being made under high uncertainty is crucial. We discuss the impact of our work further in Appendix B.

3 Bayesian Bellman Operators

Detailed proofs and a discussion of assumptions for all theoretical results are found in Appendix C.

To introduce the BBO framework we consider the Bellman equation using a function approximator: $\mathcal{B}[\hat{Q}_\omega] = \hat{Q}_\omega$. Using Eq. (1), we can write the Bellman operator for $\hat{Q}_\omega$ as an expectation of the empirical Bellman function b_ω:

$$\mathcal{B}[\hat{Q}_\omega](s, a) = \mathbb{E}_{P^\pi(s',a'|a,s)}\left[b_\omega(s', a', s, a)\right], \quad b_\omega(s', a', s, a) \coloneqq r(s', a, s) + \gamma \hat{Q}_\omega(s', a'). \tag{2}$$

When evaluating the Bellman operator to solve the Bellman equation, we can evaluate the function approximator $\hat{Q}_\omega(s, a)$ but we cannot evaluate $\mathcal{B}[\hat{Q}_\omega](s, a)$ due to the uncertainty in reward function and transition distribution. In BBO we capture this uncertainty by treating the empirical Bellman function as a transformation of variables $b_\omega(\cdot, s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each (s, a). The transformed variable $B : \mathbb{R} \to \mathbb{R}$ has a conditional distribution $P_B(b|s, a, \omega)$ which is the *pushforward* of $P^\pi(s', a'|s, a)$ under the transformation $b_\omega(\cdot, s, a)$. For any P_B-integrable function $f : \mathbb{R} \to \mathbb{R}$, the pushforward satisfies:

$$\mathbb{E}_{P_B(b|s,a,\omega)}\left[f(b)\right] = \mathbb{E}_{P^\pi(s',a'|s,a)}\left[f \circ b_\omega(s', a', s, a)\right]. \tag{3}$$

As the pushforward $P_B(b|s, a, \omega)$ is a distribution over empirical Bellman functions, each sample $b \sim P_B(\cdot|s, a, \omega)$ is a noisy sample of the Bellman operator at a point: $b_i = \mathcal{B}[\hat{Q}_\omega](s_i, a_i) + \eta_i$. To prove this, observe that taking expectations of b recovers $\mathcal{B}[\hat{Q}_\omega](s, a)$:

$$\mathbb{E}_{P_B(b|s,a,\omega)}[b] \underbrace{=}_{\text{Eq. (3)}} \mathbb{E}_{P^\pi(s',a'|s,a)}\left[b_\omega(s', a', s, a)\right] \underbrace{=}_{\text{Eq. (2)}} \mathcal{B}[\hat{Q}_\omega](s, a).$$

As the agent interacts with the environment, it obtains samples from the transition distribution $s'_i \sim P(\cdot|s_i, a_i)$ and policy $a'_i \sim \pi(\cdot|s'_i)$. From Eq. (3) a sample from the distribution $b_i \sim P_B(\cdot|s_i, a_i, \omega)$ is obtained from these state-action pairs by applying the empirical Bellman function $b_i = r_i + \gamma \hat{Q}_\omega(s'_i, a'_i)$. As we discussed in Section 2.3, existing model-free Bayesian RL approaches incorrectly treat each b_i as a sample from a distribution over the value function $P_Q(Q^\pi|s, a)$. BBO corrects this by modelling the true conditional distribution: $P_B(b|s, a, \omega)$ that generates the data.

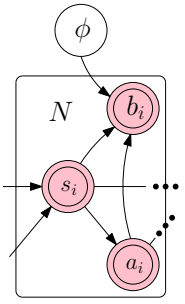


Figure 1: Graphical Model for BBO.

The graphical model for BBO is shown in Fig. 1. To model $P_B(b|s, a, \omega)$ we assume a parametric conditional distribution: $P(b|s, a, \phi)$ with model parameters

$\phi \in \Phi$ and a conditional mean: $\mathbb{E}_{P(b|s,a,\phi)}[b] = \hat{B}_\phi(s,a)$. It is also possible to specify a nonparametric model: $P(b|s,a)$. The conditional mean of the distribution $\hat{B}_\phi$ defines a function space of approximators that represents a space of Bellman operators, each indexed by $\phi \in \Phi$. The choice of $P(b|s,a,\phi)$ should therefore ensure that the space of approximate Bellman operators characterised by $\hat{B}_\phi$ is expressive enough to sufficiently represent the true Bellman operator. As we are not concerned with modelling the transition distribution in our model-free paradigm, we assume states are sampled either from an ergodic Markov chain, or i.i.d. from a buffer. Off-policy samples can be corrected using importance sampling.

Assumption 1 (State Generating Distribution)**.** *Each state s_i is drawn either i) i.i.d. from a distribution $\rho(s)$ with support over S or ii) from an ergodic Markov chain with stationary distribution $\rho(s)$ defined over a σ-algebra that is countably generated from S.*

We represent our preexisting beliefs in the true Bellman operator by specifying a prior $P(\phi)$ with a density $p(\phi)$ which assigns mass over parameterisations of function approximators $\phi \in \Phi$ in accordance with how well we believe they represent $\mathcal{B}[\hat{Q}_\omega]$. Given the prior and a dataset $\mathcal{D}_\omega^N := \{b_i, s_i, a_i\}_{i=1:N}$ of samples from the true distribution P_B, we infer the posterior $P(\phi|\mathcal{D}_\omega^N)$ using Bayes' rule which has the density (see Appendix D.1 for a derivation using both state generating distributions of Assumption 1):

$$p(\phi|\mathcal{D}_\omega^N) = \frac{\prod_{i=1}^N p(b_i|s_i,a_i,\phi)p(\phi)}{\int_\Phi \prod_{i=1}^N p(b_i|s_i,a_i,\phi)dP(\phi)}. \tag{4}$$

To be able to make predictions, we infer the posterior predictive: $p(b|\mathcal{D}_\omega^N, s, a) := \int_\Phi p(b|s,a,\phi)dP(\phi|\mathcal{D}_\omega^N)$. Unlike existing approaches, our posterior density is a function of ω, which correctly accounts for the dependence on $\hat{Q}_\omega$ in our data and the generating distribution $P_B(b|s,a,\omega)$. We highlight that it is possible to define a likelihood and prior that are functions of ω to encode any prior knowledge of how the underlying Bellman operator varies with ω, however this is not strictly necessary as the posterior automatically accounts for this dependence due to its conditioning on $\mathcal{D}_\omega^N$. As we anticipate that most applications of BBO will seek to learn an optimal policy, and hence and optimal Q-function, a prior that incorporates any knowledge available about the optimal Bellman operator will speed learning and give the agent an advantage.

As our data depends on $\hat{Q}_\omega$, we must introduce a method of learning the correct Q-function approximator. As every Bellman operator characterises an MDP, the posterior predictive mean represents a Bayesian estimate of the true MDP by using the posterior to marginalise over all Bellman operators that our model can represent according to our uncertainty in their value:

$$\mathcal{B}^\star_{\omega,N}(s,a) := \mathbb{E}_{P(b|\mathcal{D}_\omega^N,s,a)}[b] = \mathbb{E}_{P(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi(s,a)\right]. \tag{5}$$

For this reason, we refer to the predictive mean $\mathcal{B}^\star_{\omega,N}$ as the *Bayesian Bellman operator* and our Q-function approximator should satisfy a Bellman equation using $\mathcal{B}^\star_{\omega,N}$. Our objective is therefore to find $\omega^\star$ such that $\hat{Q}_{\omega^\star} = \mathcal{B}^\star_{\omega^\star,N}$. A simple approach to learn $\omega^\star$ is to minimise the mean squared Bayesian Bellman error (MSBBE) between the posterior predictive and function approximator:

$$\text{MSBBE}_N(\omega) := \left\|\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi} \tag{6}$$

Here the distribution on the ℓ_2-norm is $\rho(s)\pi(a|s)$ where recall $\rho(s)$ is defined in Assumption 1. Although the MSBBE has a similar form to a mean squared Bellman error with a Bayesian Bellman operator in place of the Bellman operator, our theoretical results in Section 3.1 show its frequentist interpretation is closer to the mean squared projected Bellman operator used by convergent TD algorithms [70]. We derive the MSBBE gradient in Appendix D.3:

$$\begin{aligned}&\nabla_\omega \text{MSBBE}_N(\omega)\\ &\quad= \mathbb{E}_{\rho,\pi}\left[\left(\hat{Q}_\omega - \mathbb{E}_{P(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi\right]\right)\left(\nabla_\omega \hat{Q}_\omega - \mathbb{E}_{P(\phi|\mathcal{D}_\omega^N)}\left[\hat{B}_\phi \nabla_\omega \log p(\phi|\mathcal{D}_\omega^N)\right]\right)\right].\end{aligned} \tag{7}$$

If we can sample from the posterior then unbiased estimates of $\nabla_\omega \text{MSBBE}_N(\omega)$ can be obtained, hence minimising the MSBBE via a stochastic gradient descent algorithm is convergent if the standard Robbins-Munro conditions are satisfied [62]. When existing approaches are used, the posterior has

no dependence on ω and the gradient $\nabla_\omega \log p(\phi|\mathcal{D}_\omega^N)$ is not accounted for, leading to gradient terms being dropped in the update. Stochastic gradient descent using these updates does not optimise any objective and so may not converge to any solution. The focus of our analysis in Section 4.1 is to extend convergent gradient methods for minimising the MSSBE to approximate inference techniques in situations where sampling from the posterior becomes intractable.

Minimising the MSBBE also avoids the double sampling problem encountered in frequentist RL where to minimise the mean squared Bellman error, two independent samples from $P(s'|s,a)$ are required to obtain unbiased gradient estimates [7]. In BBO, this issue is avoided by drawing two independent approximate Bellman operators B_{ϕ_1} and B_{ϕ_2} from the posterior $\phi_1, \phi_2 \sim P(\cdot|\mathcal{D}_\omega^N)$ instead.

3.1 Consistency of the Posterior

To address Question 2, we develop a set of theoretical results to understand the posterior's relationship to the RL problem. We introduce some mild regularity assumptions on our choice of model:

Assumption 2 (Regularity of Model)**.** *i) $\hat{Q}_\omega$ is bounded and (Φ, d_Φ) and (Ω, d_Ω) are compact metric spaces; ii) $\hat{B}_\phi$ is Lipschitz in ϕ, $P(b|s,a,\phi)$ has finite variance and a density $p(b|s,a,\phi)$ which is Lipschitz in ϕ and bounded; and iii) $p(\phi) \propto \exp(-R(\phi))$ where $R(\phi)$ is bounded and Lipschitz.*

Our main result is a Bernstein-von-Mises-type theorem [49] applied to reinforcement learning. We prove that the posterior asymptotically converges to a Dirac delta distribution centered on the set of parameters that minimise the KL divergence between the true and model distributions, which are the frequentist maximum likelihood parameters:

$$\phi_\omega^\star \coloneqq \arg\min_{\phi\in\Phi} \mathrm{KL}(P_B(b,s,a|\omega) \parallel P(b,s,a|\phi)) = \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)}\left[-\log p(b,s,a|\phi)\right], \quad (8)$$

where the expectation is taken with respect to distribution that generates the data: $P_B(b,s,a|\omega)$ that satisfies $dP_B(b,s,a|\omega) = dP_B(b|s,a,\omega)d\pi(a|s)d\rho(s)$. We make a simplifying assumption that there is a single maximum likelihood parameter, which eases analysis and exposition of our results. We discuss the more general case where it does not hold in Appendix C.3.

Assumption 3 (Single Minimiser)**.** *The set of maximum likelihood parameters $\phi_\omega^\star$ defined in Eq.* (8) *exists and is a singleton.*

Theorem 1. *Under Assumptions 1-3, in the limit $N \to \infty$ the posterior concentrates weakly on $\phi_\omega^\star$: i) $P(\phi|\mathcal{D}_\omega^N) \implies \delta(\phi = \phi_\omega^\star)$ a.s.; ii) $\mathcal{B}_{\omega,N}^\star \xrightarrow{a.s.} \hat{B}_{\phi_\omega^\star}$; and iii) MSBBE$_N(\omega) \xrightarrow{a.s.} \|\hat{Q}_\omega - \hat{B}_{\phi_\omega^\star}\|_{\rho,\pi}^2$.*

If our model can sufficiently represent the true conditional distribution then $\mathrm{KL}(P_B(b,s,a|\omega) \parallel P(b,s,a|\phi_\omega^\star)) = 0 \implies P_B(b|s,a,\omega) = P(b|s,a,\phi_\omega^\star)$. Theorem 1 proves that the posterior concentrates on the frequentist solution $\phi_\omega^\star$ and hence the Bayesian Bellman operator converges to the true Bellman operator: $\hat{B}_{\phi_\omega^\star}(s,a) = \mathbb{E}_{P(b|s,a,\phi_\omega^\star)}[b] = \mathbb{E}_{P_B(b|s,a,\omega)}[b] = \mathcal{B}[\hat{Q}_\omega](s,a)$. As every Bellman operator characterises an MDP, any Bayesian RL solution obtained using the BBO posterior such as an optimal policy or value function is consistent with the true RL solution. When the true distribution is not in the model class, $B_{\phi_\omega^\star}$ converges to the closest representation of the true Bellman operator according to the parametrisation that maximises the likelihood $\mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b,s,a|\phi)\right]$. This is analogous to frequentist convergent TD learning where the function approximator converges to a parametrisation that minimises the projection of the Bellman operator into the model class [70, 71, 12]. We now make this relationship precise by considering a Gaussian model.

3.2 Gaussian BBO

To showcase the power of Theorem 1 and to provide a direct comparison to existing frequentist approaches, we consider the nonlinear Gaussian model $P(b|s,a,\phi) = \mathcal{N}(\hat{B}_\phi(s,a), \sigma^2)$ that is commonly used for Bayesian regression [55, 33]. The mean is a nonlinear function approximator that best represents the Bellman operator $B_\phi \approx \mathcal{B}[\hat{Q}_\omega]$ and the model variance $\sigma^2 > 0$ represents the aleatoric uncertainty in our samples. Ignoring the log-normalisation constant c_{norm}, the log-posterior is an empirical mean squared error between the empirical Bellman samples and the model mean $\hat{B}_\phi(s_i,a_i)$ with additional regularisation due to the prior (see Appendix D.2 for a derivation):

$$-\log p(\phi|\mathcal{D}_\omega^N) = c_{\text{norm}} + \sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i,a_i))^2}{2\sigma^2} + R(\phi), \quad \phi_\omega^\star \in \arg\min_{\phi\in\Phi} \|\hat{B}_\phi - \mathcal{B}[\hat{Q}_\omega]\|_{\rho,\pi}^2. \quad (9)$$

Theorem 1 proves that in the limit $N \to \infty$, the effect of the prior diminishes and the Bayesian Bellman operator converges to the parametrisation: $\mathcal{B}^\star_{\omega,N} \xrightarrow{a.s.} \hat{B}_{\phi^\star_\omega}$. As $\phi^\star_\omega$ is the set of parameters that minimise the mean squared error between the true Bellman operator and the approximator, $\hat{B}_{\phi^\star_\omega}$ is a *projection* of the true Bellman operator onto the space of functions represented by $\hat{B}_\phi$:

$$\hat{B}_{\phi^\star_\omega} = \mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega] \coloneqq \{\hat{B}_{\phi'} : \phi' \in \arg\min_{\phi \in \Phi} \|\hat{B}_\phi - \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi}\}. \tag{10}$$

Finally, Theorem 1 proves that the MSBBE converges to the mean squared projected Bellman error $\text{MSBBE}_N(\omega) \xrightarrow{a.s.} \text{MSPBE}(\omega) \coloneqq \|\hat{Q}_\omega - \mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi}$. By the definition of the projection operator in Eq. (10), a solution $\hat{Q}_\omega = \mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega]$ is a TD fixed point; hence any asymptotic MSBBE minimiser parametrises a TD fixed point should it exist. To further highlight the relationship between BBO and convergent TD algorithms that minimise the mean squared projected Bellman operator, we explore the linear Gaussian regression model as a case study in Appendix E, allowing us to derive a regularised Bayesian TDC/GTD2 algorithm [71, 70].

4 Approximate BBO

We have demonstrated in Eq. (7) that if it is tractable to sample from the posterior, a simple convergent stochastic gradient descent algorithm can be used to minimise the MSBBE. We derive the gradient update for the linear Gaussian model as part of our case study in Appendix E. Unfortunately, models like linear Gaussians that have analytic posteriors are often too simple to accurately represent the Bellman operator for domains of practical interest in RL. We now extend our analysis to include approximate inference approaches.

4.1 Approximate Inference

To allow for more expressive nonlinear function approximators, for which the posterior normalisation is intractable, we introduce a tractable posterior approximation: $q(\phi|\mathcal{D}^N_\omega) \approx P(\phi|\mathcal{D}^N_\omega)$. In this paper, we use randomised priors (RP) [57] for approximate inference. Randomised priors (PR) inject noise into the maximum a posteriori (MAP) estimate via a noise variable $\epsilon \in \mathcal{E}$ with distribution $P_E(\epsilon)$ where the density $p_E(\epsilon)$ has the same form as the prior. We provide a full exposition of RP for BBO in Appendix F, including derivations of our objectives. RP in practice uses ensembling: L prior randomisations $\mathcal{E}_L \coloneqq \{\epsilon_l\}_{l=1:L}$ are first drawn from P_E. To use RP for BBO, we write the Q-function approximator as an ensemble of L parameters $\Omega_L \coloneqq \{\omega_l\}_{l=1:L}$ where $\hat{Q}_\omega = \frac{1}{L}\sum_{l=1}^L \hat{Q}_{\omega_l}$ and require an assumption about the prior and the function spaces used for approximators:

Assumption 4 (RP Function Spaces)**.** *i) $\hat{Q}_{\omega_l}$ and $\hat{B}_{\omega_l}$ share a function space, that is $\hat{Q}_{\omega'_l} = \hat{B}_{\omega'_l}$ for any $\omega'_l \in \Omega$, where $\Phi = \Omega \subset \mathbb{R}^n$ is compact, convex with a smooth boundary. ii) $\mathcal{E} \subseteq \mathbb{R}^n$ and $R(\phi - \epsilon)$ is defined for any $\phi \in \Phi, \epsilon \in \mathcal{E}$.*

For each $l \in \{1 : L\}$, a set of solutions to the prior-randomised MAP objective are found:

$$\psi^\star_l(\omega_l) \in \arg\min_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}^N_{\omega_l}, \epsilon_l) \coloneqq \arg\min_{\phi \in \Phi} \frac{1}{N}\left(R(\phi - \epsilon_l) - \sum_{i=1}^N \log p(b_i|s_i, a_i, \phi)\right). \tag{11}$$

The RP solution $\psi^\star_l(\omega_l)$ has dependence on ω_l that mirrors the BBO posterior's dependence on ω. To construct the RP approximate posterior $q(\phi|\mathcal{D}^N_\omega)$, we average the set of perturbed MAP estimates over all ensembles: $q(\phi|\mathcal{D}^N_\omega) \coloneqq \frac{1}{L}\sum_{l=1}^L \delta(\phi \in \psi^\star_l(\omega_l))$. To sample from the RP posterior $\phi \sim q(\cdot|\mathcal{D}^N_\omega)$, we sample an ensemble uniformly $l \sim \text{Unif}(\{1 : L\})$ and set $\phi = \psi^\star_l(\omega_l)$. Although BBO is compatible with any approximate inference technique, we justify our choice of RP by proving that it preserves the consistency results developed in Theorem 1:

Corollary 1.1. *Under Assumptions 1-4, results i)-iii) of Theorem 1 hold with $P(\phi|\mathcal{D}^N_\omega)$ replaced by the RP approximate posterior $q(\phi|\mathcal{D}^N_\omega)$ both with or without ensembling.*

In answer to Question 3), Corollary 1.1 shows that the difference between using the RP approximate posterior and the true posterior lies in their characterisation of uncertainty and not their asymptotic behaviour. Existing work shows that RP uncertainty estimates are conservative [59, 21] with strong empirical performance in RL [57, 58] for the Gaussian model that we study in this paper.

The RP approximate posterior $q(\phi|\mathcal{D}^N_\omega)$ depends on the ensemble of Q-function approximators $\hat{Q}_{\omega_l}$ and like in Section 3 we must learn an ensemble of optimal parametrisations $\omega^\star_l$. We substitute

for $q(\phi|\mathcal{D}_\omega^N)$ in place of the true posterior in Eqs. (5) and (6) to derive an ensembled RP MSBBE: $\text{MSBBE}_{\text{RP}}(\omega_l) \coloneqq \|\hat{Q}_{\omega_l} - \hat{B}_{\psi_l^\star(\omega_l)}\|^2_{\rho,\pi}$. When a fixed point $\hat{Q}_{\omega_l} = \hat{B}_{\psi_l^\star(\omega_l)}$ exists, minimising $\text{MSBBE}_{\text{RP}}(\omega_l)$ is equivalent to finding $\omega_l^\star$ such that $\psi_l^\star(\omega_l^\star) = \omega_l^\star$. To learn $\omega_l^\star$ we can instead minimise the simpler parameter objective $\omega_l^\star \in \arg\min_{\omega_l \in \Omega} \mathcal{U}(\omega_l; \psi_l^\star)$:

$$\mathcal{U}(\omega_l; \psi_l^\star) \coloneqq \|\omega_l - \psi_l^\star(\omega_l)\|_2^2 \quad \text{such that} \quad \psi_l^\star(\omega_l) \in \arg\min_{\phi \in \Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l), \tag{12}$$

which has the advantage that deterministic gradient updates can be obtained. $\mathcal{U}(\omega_l; \psi_l^\star)$ can still provide an alternative auxilliary objective when a fixed point does not exist as the convergence of algorithms minimising Eq. (12) does not depend on its existence and has the same solution as minimising $\text{MSBBE}_{\text{RP}}(\omega_l)$ for sufficiently smooth B_ϕ. Solving the bi-level optimisation problem in Eq. (12) is NP-hard [8]. To tackle this problem, we introduce an ensemble of parameters $\Psi_L \coloneqq \{\psi_l\}_{1:L}$ to track $\psi_l^\star(\omega_l)$ and propose a two-timescale gradient update for each $l \in \{1:L\}$ on the objectives in Eq. (12) with per-step complexity of $\mathcal{O}(n)$:

$$\psi_l \leftarrow \mathcal{P}_\Omega\left(\psi_l - \alpha_k \nabla_{\psi_l}\left(R(\psi_l - \epsilon_l) - \log p(b_i|s_i, a_i, \psi_l)\right)\right), \quad \text{(fast)} \tag{13}$$

$$\omega_l \leftarrow \mathcal{P}_\Omega(\omega_l - \beta_k(\omega_l - \psi_l)), \quad \text{(slow)} \tag{14}$$

where α_k and β_k are asymptotically faster and slower stepsizes respectively and $\mathcal{P}_\Omega(\cdot) \coloneqq \arg\min_{\omega \in \Omega} \|\cdot - \omega\|_2^2$ is a projection operator that projects its argument back into Ω if necessary. From a Bayesian perspective, we are concerned with characterising the uncertainty after a *finite* number of samples $N < \infty$ and hence (b_i, s_i, a_i) should be drawn uniformly from the dataset $\mathcal{D}_{\omega_l}^N$ to form estimates of the summation in Eq. (11), which becomes intractable with large N. When compared to existing RL algorithms, sampling from $\mathcal{D}_{\omega_l}^N$ is analogous to sampling from a replay buffer [54]. A frequentist analysis of our updates is also possible by considering samples that are drawn online from the underlying data generating distribution $(b_i, s_i, a_i) \sim P_B$ in the limit $N \to \infty$. We discuss this frequentist interpretation further in Appendix C.5.

To answer Question 4), we prove convergence of updates (13) and (14) using a straightforward application of two-timescale stochastic approximation [15, 14, 42] to BBO. Intuitively, two timescale analysis proves that the faster timescale update (13) converges to an element in Ω using standard martingale arguments, viewing the parameter ω_l as quasi-static as it behaves like a constant. Since the perturbations are relatively small, the separation of timescales then ensures that ψ_l tracks $\psi_l^\star(\omega_l)$ whenever ω_l is updated in the slower timescale update (14), viewing the parameter ψ_l as quasi-equilibrated [14]. We introduce the standard two-timescale regularity assumptions and derive the limiting ODEs of updates (13) and (14) in Appendix C.3:

Assumption 5 (Two-timescale Regularity)**.** *i) $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$ and $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_l)$ are Lipschitz in ψ_l and $(b_i, s_i, a_i) \sim \text{Unif}(\mathcal{D}_{\omega_l}^N)$; ii) $\psi^\circledast(\omega_l)$ and $\omega_l^\circledast$ are local aysmptotically stable attractors of the limiting ODEs of updates (13) and (14) respectively and $\psi_l^\circledast(\omega_l)$ is Lipschitz in ω_l; and iii) The stepsizes satisfy:* $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$, $\sum_{k=1}^\infty \alpha_k = \sum_{k=1}^\infty \beta_k = \infty$, $\sum_{k=1}^\infty \left(\alpha_k^2 + \beta_k^2\right) < \infty$.

Theorem 2. *If Assumptions 1 to 5 hold, ψ_l and ω_l converge to $\psi_l^\circledast(\omega_l^\circledast)$ and $\omega_l^\circledast$ almost surely.*

As ω_l are updated on a slower timescale, they lag the parameters ψ_l. When deriving a Bayesian actor-critic algorithm in Section 4.2, we demonstrate that these parameters share a similar role to a *lagged critic*. There is no Bayesian explanation for these parameters under existing approaches: when applying approximate inference to $P(Q^\pi|s, a, \mathcal{D}^N)$, the RP solution $\psi_l^\star$ has no dependence on ω_l. Hence, minimising $\mathcal{U}(\omega_l; \psi_l^\star)$ and the approximate MSBBE has an exact solution by setting $\omega_l^\star = \psi_l^\star$. In this case, $\hat{Q}_{\omega_l^\star} = \hat{B}_{\psi_l^\star}$ meaning that existing approaches do not distinguish between the Q-function and Bellman operator approximators.

4.2 Bayesian Bellman Actor-Critic

BootDQN+Prior [57, 58] is a state-of-the-art Bayesian model-free algorithm with Thompson sampling [74] where, in principle, an optimal Q-function is drawn from a posterior over optimal Q-functions at the start of each episode. As BootDQN+Prior requires bootstrapping, it actually draws a sample from the Gaussian BBO posterior introduced in Section 3.2 using RP approximate inference with the empirical Bellman

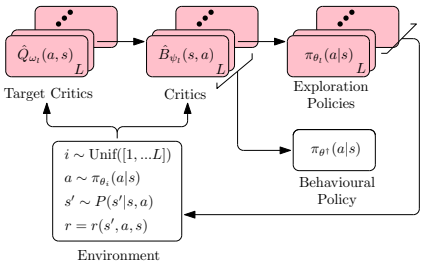


Figure 2: Schematic of RP-BBAC.

function $b_\omega(s', a, s) = r(s', a, s) + \gamma \max_{a'} \hat{Q}_\omega(s', a')$.
This empirical Bellman function results from substituting an optimal policy $\pi(a|s) = \delta(a \in \arg\max_{a'} \hat{Q}_\omega(s, a'))$ in Eq. (3). A variable l is drawn uniformly and the optimal exploration policy $\pi_l^\star(a|s) = \delta(a \in \arg\max_{a'} B_{\phi_l}(s, a'))$ is followed. BootDQN+Prior achieves what Osband et al. [58] call *deep exploration* where exploration not only considers immediate information gain but also the consequences of an exploratory action on future learning. Due its use of the $\arg\max$ operator, BootDQN+Prior is not appropriate for continuous action or large discrete action domains as a nonlinear optimisation problem must be solved every time an action is sampled. We instead develop a randomised priors Bayesian Bellman actor-critic (RP-BBAC) to extend BootDQN+Prior to continuous domains. A schematic of RP-BBAC is shown in Fig. 2 which summarises Algorithm 1. Additional details are in Appendix G.

Comparison to existing actor-critics: Using a Gaussian model also allows a direct comparison to frequentist actor-critic algorithms [50]: as shown in Fig. 2, every ensemble $l \in \{1...L\}$ has its own *exploratory actor* π_{θ_l}, *critic* B_{ψ_l} and *target critic* $\hat{Q}_{\omega_l}$. In BBAC, each critic is the solution to its unique ϵ_l-randomised empirical MSBBE objective from Eq. (12): $\mathcal{L}_{\text{critic}}(\psi_l) \coloneqq -\frac{1}{\sigma^2}\sum_{i=1}^{N}(b_i - \hat{B}_{\psi_l}(s_i, a_i))^2 + R(\psi_l - \epsilon_l)$. The target critic parameters ω_l for each Bellman sample $b_i = r_i + \gamma \hat{Q}_{\omega_l}(s_i', a_i')$ are updated on a slower timescale to the critic parameters, which mimics the updating of target critic parameters after a regular interval in frequentist approaches [54, 39]. We introduce an ensemble of parametric exploration policies $\pi_{\theta_l}(a|s)$ parametrised by a set of parameters $\Theta_L \coloneqq \{\theta_l\}_{l=1:L}$. Each optimal exploration policy $\pi_l^\star(a|s)$ is parametrised by the solution to its own optimisation problem: $\theta_l^\star \in \arg\max_{\theta_l \in \Theta} \mathbb{E}_{\rho(s)\pi_{\theta_l}(a|s)}[B_{\phi_l}(s, a')]$. Unlike frequentist approaches, an exploratory actor is selected at the start of each episode in accordance with our current uncertainty in the MDP characterised by the approximate RP posterior. Exploration is thus both deep and adaptive as actions from an exploration policy are directed towards minimising epistemic uncertainty in the MDP and the posterior variance reduces in accordance with Corollary 1.1 as more data is collected. BBAC's explicit specification of lagged target critics is unique to BBO and, as discussed in Section 4.1, corrects the theoretical issues raised by applying bootstrapping to existing model-free Bayesian RL theory, which does not account for the posterior's dependence on $\hat{Q}_\omega$. Finally, exploration policies may not perform well at test time, so we learn a behaviour policy $\pi_{\theta^\dagger}(a|s)$ parametrised by $\theta^\dagger \in \Theta$ from the data collected by our exploration policies using the ensemble of critics: $\{\hat{B}_{\psi_l}\}_{l=1:L}$. Theoretically, this is the optimal policy for the Bayesian estimate of the true MDP by using the approximate posterior to marginalise over the ensemble of Bellman operators. We augment our behaviour policy objective with entropy regularisation, allowing us to combine the exploratory benefits of Thompson sampling with the faster convergence rates and algorithmic stability of regularised RL [77].

Algorithm 1 RP-BBAC

Initialise $\Theta_L, \Omega_L, \Psi_L, \mathcal{E}_L, \theta^\dagger$ and $\mathcal{D} \leftarrow \varnothing$
Sample initial state $s \sim P_0$
while not converged **do**
  Sample policy $\theta_l \sim \text{Unif}(\Theta_L)$
  for $n \in \{1, ...N_{\text{env}}\}$ **do**
    Sample action $a \sim \pi_{\theta_l}(\cdot|s)$
    Observe next state $s' \sim P(\cdot|s, a)$
    Observe reward $r = r(s', a, s)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, a, r, s'\}$
  end for
  $\Theta_L, \Omega_L, \Psi_L \leftarrow$ UPDATEPOSTERIOR
  $\theta^\dagger \leftarrow$ UPDATEBEHAVIOURALPOLICY
end while

5 Related Work

Existing model-free Bayesian RL approaches assume either a parametric Gaussian [34, 57, 32, 52, 58, 75] or Gaussian process regression model [28, 29]. Value-based approaches use the empirical Bellman function $b_\omega(s', a, s) = r(s', a, s) + \gamma \max_{a'} \hat{Q}_\omega(s', a')$ whereas actor-critic approaches use the empirical Bellman function $b_\omega(s', a', s, a) = r(s', a, s) + \gamma \hat{Q}_\omega(s', a')$. In answering Questions 1-4, we have shown existing methods that use bootstrapping inadvertently approximate the posterior predictive over Q-functions with the BBO posterior predictive $P(Q^\pi|s, a, \mathcal{D}^N) \approx P(b|s, a, \mathcal{D}_\omega^N)$. These methods minimise an approximation of the MSBBE where the Bayesian Bellman operator is treated as a supervised target, ignoring its dependence on ω: gradient descent approaches drop gradient terms and fitted approaches iteratively regress the Q-function approximator onto the Bayesian Bellman operator $\hat{Q}_{\omega_{k+1}} \leftarrow \mathcal{B}^\star_{\omega_k, N}$. In both cases, the updates may not be a contraction mapping for the same reasons as in non-Bayesian TD [76] and so it is not possible to prove general convergence.

The additional Bayesian regularisation introduced from the prior can lead to convergence, but only in specific and restrictive cases [4, 5, 31, 17].

Approximate inference presents an additional problem for existing approaches: many existing methods naïvely apply approximate inference to the Bellman error, treating $\mathcal{B}[Q^\pi](s,a)$ and $Q^\pi(s,a)$ as independent variables [32, 52, 75, 34]. This leads to poor uncertainty estimates as the Bellman error cannot correctly propagate the uncertainty [56, 57]. Osband et al. [58] demonstrate that this can cause uncertainty estimates of $Q^\pi(s,a)$ at some (s,a) to be zero and propose BootDQN+Prior as an alternative to achieve deep exploration. BBO does not suffer this issue as the posterior characterises the uncertainty in the Bellman operator directly. In Section 4.2 we demonstrated that BootDQN+Prior derived from BBO specifies the use of target critics. Despite being essential to performance, there is no Bayesian explanation for target critics under existing model-free Bayesian RL theory, which posits that sampling a critic from $P(Q^\pi|s,a,\mathcal{D}^N)$ is sufficient.

6 Experiments

Convergent Nonlinear Policy Evaluation To confirm our convergence and consistency results under approximation, we evaluate BBO in several nonlinear policy evaluation experiments that are constructed to present a convergence challenge for TD algorithms. We verify the convergence of nonlinear Gaussian BBO in the famous counterexample task of Tsitsiklis and Van Roy [76], in which the TD(0) algorithm is provably divergent. The results are presented in Fig. 3. As expected, TD(0) diverges, while BBO converges to the optimal solution faster than convergent frequentist nonlinear TDC and GTD2 [12]. We also consider three additional policy evaluation tasks commonly used to test convergence of nonlinear TD using neural network function approximators: 20-Link Pendulum [23], Puddle World [16], and Mountain Car [16]. Results are shown in Fig. 11 of Appendix H.3 from which we conclude that i) by ignoring the posterior's dependence on ω, existing model-free Bayesian approaches are less stable and perform poorly in comparison to the gradient based MSBBE minimisation approach in Eq. (7), ii) regularisation from a prior can improve performance of policy evaluation by aiding the optimisation landscape [26], and iii) better solutions in terms of mean squared error can be found using BBO instead of the local linearisation approach of nonlinear TDC/GTD2[12].

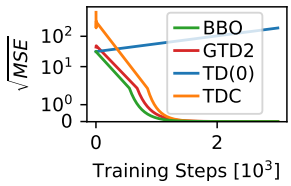
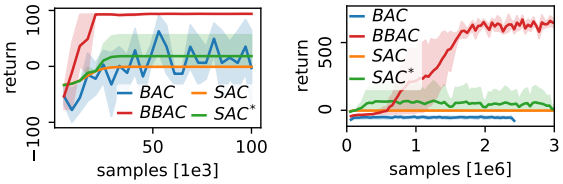


Figure 3: Tsitsiklis counterexample.

Exploration for Continuous Control In many benchmark tasks for continuous RL, such as the locomotion tasks from MuJoCo Gym suite [18], the environment reward is shaped to provide a smooth gradient towards a successful task completion and naïve Boltzmann dithering exploration strategies from regularised RL can provide a strong inductive bias. In practical real-world scenarios, dense rewards are difficult to specify by hand, especially when the task is learned from raw observations like images. Therefore, we consider a set of continuous control tasks with sparse rewards as continuous analogues of the discrete experiments used to test BootDQN+Prior [57]: *MountainCar-Continuous-v0* from Gym benchmark suite and a slightly modified version of the *cartpole-swingup_sparse* from DeepMind Control Suite [73]. Both environments have a sparse reward signal and penalize the agent proportional to the magnitude of executed actions. As the agent is always initialised in the same state, it has to deeply explore costly states in a directed manner for hundreds of steps until it reaches the rewarding region of the state space. We compare RP-BBAC with two variants of the state-of-the-art soft actor-critic: SAC, which is the exact algorithm presented in [40]; and SAC*, a tailored version which uses a single Q-function to avoid *pessimistic underexploration* [20] due to the use of the double-minimum-Q trick (see Appendix I for details). To understand the practical implications of our theoretical results, we also compare against BAC which is a variant of RP-BBAC where $\hat{Q}_{\omega_l^\star} = \hat{B}_{\psi_l^\star}$. As we discussed in Section 4.1, BAC is the Bayesian actor-critic that results from applying RP approximate inference to the posterior over Q-functions used by existing model-free Bayesian approaches with bootstrapping.



(a) MountainCar (b) Cartpole

Figure 4: Continuous control with sparse reward.

The results are shown in Fig. 4. Due to the lack of smooth signal towards the task completion, SAC consistently fails to solve the tasks and converges to always executing the 0-action due to the action cost term, while SAC* achieves the goal in one out of five seeds. RP-BBAC succeeds for all five seeds in both tasks. To understand why, we provide a state support analysis in for

MountainCar-Continuous-v0 Appendix I.1. The final plots are shown in Fig. 5 and confirm that the deep, adaptive exploration carried out by RP-BBAC leads agents to systematically explore regions of the state-action space with high uncertainty. The same analysis for SAC and SAC* confirms the inefficiency of the exploration typical of RL as inference: the agent repeatedly explores actions that lead to poor performance and rarely explores beyond its initial state. The state support analysis for BAC in Appendix I.1 confirms that by using the posterior over Q-functions with bootstrapping, existing model-free Bayesian RL cannot accurately capture the uncertainty in the MDP. Initially, exploration is similar to RP-BBAC but epistemic uncertainty estimates are unstable and cannot concentrate due to the convergence issues highlighted in this paper, preventing adaptive exploration. Our results in Fig. 4 demonstrate that the theoretical issues with existing approaches have negative empirical consequences, verifying that it is essential for Bayesian model-free RL algorithms with bootstrapping to sample from the BBO posterior as BAC fails to solve both tasks where sampling from the correct posterior in RP-BBAC succeeds. In Appendix I.2, we also investigate RP-BBAC's sensitivity to randomized prior hyperparameters. The range of working hyperparameters is wide and easy to tune.

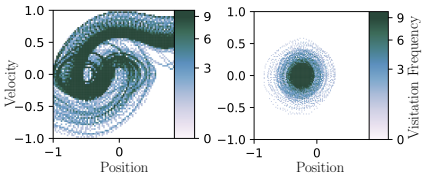


Figure 5: State Support for RP-BBAC (left) and SAC (right) in MountainCar-Continuous-v0.

7 Conclusion

By introducing the BBO framework, we have addressed a major theoretical issue with model-free Bayesian RL by analysing the posterior that is inferred when bootsrapping is used. Our theoretical results proved consistency with frequentist RL and strong convergence properties, even under posterior approximation. We used BBO to extend BootDQN+Prior to continuous domains. Our experiments in environments where rewards are not hand-crafted to aid exploration demonstrate that sampling from the BBO posterior characterises uncertainty correctly and algorithms derived from BBO can succeed where state-of-the-art algorithms fail catastrophically due to their lack of deep exploration. Future research could experiment with novel inference techniques, more complex priors and likelihoods, especially if they depend on ω, or extend our convergence analysis to the actor-critic algorithm.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement number 637713). Matthew Fellows and Kristian Hartikainen are funded by the EPSRC. The experiments were made possible by a generous equipment grant from NVIDIA. We would like to thank Piotr Miłoś, whose proof for a similar problem inspired our proof of Lemma 3.

References

[1] Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation, and work. In *The Economics of Artificial Intelligence: An Agenda*, pages 197–236. National Bureau of Economic Research, Inc, 2018. URL `https://EconPapers.repec.org/RePEc:nbr:nberch:14027`. B

[2] Daron Acemoglu and Pascual Restrepo. Unpacking skill bias: Automation and new tasks. *AEA Papers and Proceedings*, 110:356–61, May 2020. doi: 10.1257/pandp.20201063. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20201063`. B

[3] Donald Andrews. Generic uniform convergence. *Econometric Theory*, 8(2):241–257, 1992. 1

[4] András Antos, Rémi Munos, and Csaba Szepesvári. Fitted q-iteration in continuous action-space mdps. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 9–16, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520. 5

[5] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008. doi: 10.1007/s10994-007-5038-2. 5

[6] John Asmuth and Michael Littman. Learning is planning: near bayes-optimal reinforcement learning via monte-carlo tree search. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 19–26, 01 2011. 2.2

[7] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. *Machine Learning-International Workshop Then Conference-*, pages 30–37, July 1995. ISSN 00043702. doi: 10.1.1.48.3256. 3

[8] J. F. Bard. Some properties of the bilevel programming problem. *J. Optim. Theory Appl.*, 68(2): 371–378, February 1991. ISSN 0022-3239. 4.1

[9] R.F. Bass. *Real Analysis for Graduate Students*, chapter 21. Createspace Ind Pub, 2013. ISBN 9781481869140. 1

[10] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003. 2.2

[11] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2.2

[12] Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid R. Maei, and Csaba Szepesvári. Convergent temporal-difference learning with arbitrary smooth function approximation. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1204–1212. Curran Associates, Inc., 2009. 3.1, 6, C.3, E.2, H.2.1, H.2.2, H.2.3

[13] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication. 3, 1, 1.1

[14] Vivek Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, 01 2008. ISBN 978-81-85931-85-2. doi: 10.1007/978-93-86279-38-5. 4.1, C.3, 2

[15] Vivek S. Borkar. Stochastic approximation with two time scales. *Syst. Control Lett.*, 29(5): 291–294, February 1997. ISSN 0167-6911. doi: 10.1016/S0167-6911(97)90015-3. 4.1

[16] Justin A. Boyan and Andrew W. Moore. Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 369–376. MIT Press, 1995. 6, H.3.1

[17] David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear td learning. In *ICLR 2020*, 2019. 5

[18] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 6, 2, I.1

[19] Wesley Chung, Somjit Nath, Ajin Joseph, and Martha White. Two-timescale networks for nonlinear value function approximation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=rJleN20qK7`. C.3

[20] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*, 2019. 6, I

[21] Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *Eighth International Conference on Learning Representations*, April 2020. 4.1

[22] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. Chapman and Hall, London, 1974. ISBN 0412124203. 1

[23] Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014. 6, H.1, H.1.1, 1, 2, H.3.1, H.3.2, H.3.2

[24] Bruno de Finetti. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7(1):1–68, 1937. 1

[25] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017. B

[26] Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1049–1058, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 6

[27] Michael O'Gordon Duff and Andrew Barto. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, 2002. AAI3039353. 2.2

[28] Yaakov Engel, Shie Mannor, and Ron Meir. Bayes meets bellman: The gaussian process approach to temporal difference learning. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 154–161, 2003. ISBN 1577351894. 5

[29] Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 201–208, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102377. URL https://doi.org/10.1145/1102351.1102377. 5

[30] Matthew Fellows, Kamil Ciosek, and Shimon Whiteson. Fourier Policy Gradients. In *ICML*, 2018. G

[31] Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 15456–15467. Curran Associates, Inc., 2019. 5

[32] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alexander Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *Proceedings of the International Conference on Representation Learning (ICLR 2018)*, Vancouver (Canada), 2018. 5

[33] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 2.2, 3.2

[34] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org, 2016. 5

[35] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. *Bayesian Reinforcement Learning: A Survey*. now, 2015. ISBN null. 1, 2.2

[36] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*, chapter Introduction to General State-Space Markov Chain Theory. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995. ISBN 9780412055515. 1

[37] Adam Greenfield. *Radical Technologies: The Design of Everyday Life*. Verso, 2018. ISBN 1784780456. B

[38] Arthur Guez, David Silver, and Peter Dayan. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883, 10 2013. doi: 10.1613/jair.4117. 2.2

[39] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 4.2, I

[40] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *CoRR*, abs/1812.05905, 2018. 6, I.1

[41] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2944–2952. Curran Associates, Inc., 2015. G

[42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017. 4.1, C.3, 2, C.5

[43] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. 06 2017. B

[44] Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-60032-3. 2.2

[45] Prasenjit Karmakar and Shalabh Bhatnagar. Two time-scale stochastic approximation with controlled markov noise and off-policy temporal-difference learning. *Math. Oper. Res.*, 43(1): 130–151, February 2018. ISSN 0364-765X. doi: 10.1287/moor.2017.0855. URL `https://doi.org/10.1287/moor.2017.0855`. 2, C.5

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. H.3.4

[47] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014. 2.2, G

[48] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey, 2020. B

[49] B.J.K. Kleijn and A.W. van der Vaart. The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012. doi: 10.1214/12-EJS675. 3.1

[50] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 1008–1014. MIT Press, 2000. 4.2

[51] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017. 2.2

[52] Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and li Deng. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *AAAI*, 11 2018. 5

[53] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 08 2019. B

[54] 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. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. 4.1, 4.2

[55] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 7. The MIT Press, 2012. ISBN 0262018020, 9780262018029. 3.2, E

[56] Brendan O'Donoghue, Ian Osband, Remi Munos, and Vlad Mnih. The uncertainty Bellman equation and exploration. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3839–3848, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 5

[57] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8617–8629. Curran Associates, Inc., 2018. 1, 2.2, 4.1, 4.1, 4.2, 5, 6, F, G

[58] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. 4.1, 4.2, 5, F, I.2

[59] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Uncertainty in neural networks: Bayesian ensembling. *ArXiv Preprint*, abs/1810.05546, 10 2019. 4.1, F

[60] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779. 2.1

[61] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 37:1530–1538, 07–09 Jul 2015. 2.2

[62] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`. 3, C.3

[63] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998. 1.1, 1.1

[64] Alexander Shapiro. On differentiability of metric projections in rn, 1: Boundary case. *Proceedings of the American Mathematical Society*, 99(1):123–128, 1987. ISSN 00029939, 10886826. URL `http://www.jstor.org/stable/2046282`. C.3

[65] Alexander Shapiro. Directional differentiability of metric projections onto moving sets at boundary points. *Journal of Mathematical Analysis and Applications*, 131(2):392–403, 1988. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(88)90213-2. URL `https://www.sciencedirect.com/science/article/pii/0022247X88902132`. 2, C.3

[66] Adam Smith and Janna Anderson. Ai, robotics, and the future of jobs. 2017. B

[67] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013. doi: 10.1109/MSP.2013.2252713. 2.2

[68] Thomas Spooner, Rahul Savani, John Fearnley, and Andreas Koukorinis. Market making via reinforcement learning. In *17th International Conference on Autonomous Agents and Multiagent Systems*, 07 2018. B

[69] Nick Srnicek and Alex Williams. *Inventing the future: postcapitalism and a world without work*. Verso, 2015. ISBN 9781784780968. B

[70] Richard S Sutton, Hamid R. Maei, and Csaba Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1609–1616. Curran Associates, Inc., 2009. 3, 3.1, 3.2, E.2, E.2

[71] Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 993–1000, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553501. 3.1, 3.2, E.2, E.2

[72] Csaba Szepesvári. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010. ISSN 1939-4608. doi: 10.2200/S00268ED1V01Y201005AIM009. 2.1

[73] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. 6, I, I.2

[74] William R Thomson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. 4.2

[75] Ahmed Touati, Harsh Satija, Joshua Romoff, Joelle Pineau, and Pascal Vincent. Randomized value functions via multiplicative normalizing flows. In Amir Globerson and Ricardo Silva, editors, *UAI*, page 156. AUAI Press, 2019. 5

[76] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 2334-3303. doi: 10.1109/9.580874. 5, 6, 10, H.2, H.2.1

[77] Nino Vieillard, Tadashi Kozuno, B. Scherrer, O. Pietquin, Rémi Munos, and M. Geist. Leverage the average: an analysis of regularization in rl. *Advances in Neural Information Processing Systems*, 33, 2020. 4.2

[78] Nikos Vlassis, Mohammad Ghavamzadeh, Shie Mannor, and Pascal Poupart. *Bayesian Reinforcement Learning*, pages 359–386. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_11. 1, 2.2

[79] David Williams. *Probability with Martingales.* Cambridge mathematical textbooks. Cambridge University Press, 1991. ISBN 978-0-521-40605-5. 1, 2

[80] Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: a survey. *arXiv preprint arXiv:1908.08796*, 2019. B

[81] EDUARDO H. ZARANTONELLO. Projections on convex sets in hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In Eduardo H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, pages 237–424. Academic Press, 1971. ISBN 978-0-12-775850-3. doi: https://doi.org/10.1016/B978-0-12-775850-3.50013-3. URL https://www.sciencedirect.com/science/article/pii/B9780127758503500133. C.3

[82] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *8th International Conference on Learning Representations, ICLR 2020, Virtual Conference, Formerly Addis Ababa ETHIOPIA*, 2020. 2.2

Checklist

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In this work we identified a major theoretical issue with existing model-free Bayesian RL approaches that claim to infer a posterior over Q-functions. We answered Questions 1 to 2 of Section 2.3 in Section 3 and Questions 3 to 4 in Section 4.

(b) Did you describe the limitations of your work? [Yes] The purpose of this work is about addressing a major theoretical limitation with existing model-free Bayesian RL. We show that this limitation can have profound empirical consequences in Section 6. As we have discussed, the assumptions of our theories are relatively weak and apply to a wide range of settings and function approximators used for RL including nonlinear neural networks.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 2.3 and the further discussion in Appendix B

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1 to 5 and our extensive discussion in Appendices C.1 and C.3

(b) Did you include complete proofs of all theoretical results? [Yes] We provide a high level intuitive explanation of our theorems in the main text (see Section 3.1 and Section 4.1) with rigorous and detailed proofs in Appendix C.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Provided in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Mathematical Glossary

Term	Definition
$P_0(s_0)$	Initial state distribution
$P(s';\|s,a)$	Transition distribution
$r(s',a,s)$	Reward function
$\pi(a\|s)$	Policy
$P^\pi(s',a'\|s,a)$	State-action Transition Distribution
J^π	Reinforcement learning objective
$\pi_\theta(a\|s)$	Parametric policy
$Q^\pi(s,a)$	Q-function
$\hat{Q}_\omega(s,a)$	Q-function approximator
$\theta \in \Theta$	Policy Parameters
$\omega \in \Omega$	Q-function approximator parameters
$\mathcal{B}[Q^\pi](s,a)$	Bellman Operator (See Eq. (1))
$P_Q(Q^\pi\|s,a)$	Distribution over Q-functions
$P(Q^\pi\|s,a,\mathcal{D}^N_\omega)$	Posterior over Q-functions
$b_\omega(s',a',s,a)$	Empirical Bellman function (see Eq. (2))
$p_b(b\|s,a,\omega)$	Pushforward of $P^\pi(s',a'\|s,a)$ through $b_\omega(s',a',s,a)$ (see Eq. (3))
BBO Model:	
$\rho(s)$	BBO state sampling distribution (see Assumption 1)
$\phi \in \Phi$	Model parameters
$P(b\|s,a,\phi)$	BBO likelihood
$\hat{B}_\phi(s,a)$	Likelihood mean
$\mathcal{D}^N_\omega$	BBO dataset of samples
$p(\phi) \propto \exp(-R(\phi))$	Prior
$p(\phi\|\mathcal{D}^N_\omega)$	Posterior (See Eq. (4))
$p(b\|\mathcal{D}^N_\omega,s,a)$	Posterior predictive
$\mathcal{B}^\star_N(s,a;\omega,\theta)$	Posterior predictive mean i.e. Bayesian Bellman operator (see Eq. (5))
$\text{MSBBE}_N(\omega)$	Mean squared Bayesian Bellman error (see Eq. (6))
$P_B(b,s,a\|\omega)$	Data generating distribution: $dP_B(b,s,a\|\omega) = dP_B(b\|s,a,\omega)d\pi(a\|s)d\rho(s)$
$\phi^\star_\omega$	Maximum likelihood parameters (see Eq. (8))
$\mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega]$	Projection of Bellman operator $\mathcal{B}$ (see Eq. (10))
$\text{MSPBE}(\omega)$	Mean squared projected Bellman error
Approximate Inference:	
$q(\phi\|\mathcal{D}^N_\omega)$	Approximate posterior
$P_E(\epsilon)$	Prior randomisation distribution
$\mathcal{L}(\phi;\mathcal{D}^N_{\omega_l},\epsilon_l)$	Prior randomised MAP objective (see Eq. (11))
$\psi^\star_l(\omega_l)$	RP solutions (see Eq. (11))
$\mathcal{U}(\omega_l;\psi^\star_l)$	Auxiliary objective (see Eq. (12))
$\mathcal{E}_L := \{\epsilon_l\}_{l=1:L}$	Ensemble of prior randomisations
$\Omega_l := \{\omega_l\}_{l=1:L}$	Ensemble of critic parameters
$\Psi_L := \{\psi_l\}_{l=1:L}$	Ensemble of target critic parameters
$\Theta_L := \{\theta_l\}_{l=1:L}$	Ensemble of exploration policy parameters
$\pi_{\theta_l}(a\|s)$	Exploration policy
$\pi_{\theta^\dagger}(a\|s)$	Behavioural policy

B Broader Impact

Many of the benefits afforded by Bayesian RL methods play a significant role in a wide range of real-world applications, for example, financial applications of deep RL [25] have the potential to destabilise economies, and precise uncertainty quantification can enable safer applications of RL for trading [43, 68]. Incorporating prior knowledge can have substantial impact on speed and accuracy, for example in medical applications of RL [80]. Furthermore, safe exploration and convergence guarantees are crucial in many physical-world domains such as robotics or autonomous vehicles [48], where undesirable actions and algorithmic divergence could cause direct human harm or even fatalities.

Stronger theoretical guarantees increases confidence in algorithms and can accelerate the implementation of RL for real-world applications. The societal effects of rapidly increased automation are uncertain and disputed [66]. Positive effects include claims of increased job creation, productivity [1] and even techno-utopianism [69]. A critical approach to these claims is important, especially given the gulf between the ideals and the reality of what automation has achieved so far [37] and the potential automation has for increased inequality [2] and discrimination [53].

C Proofs

C.1 Assumptions and Preliminaries for Theorem 1

Assumption 1 (State Generating Distribution)**.** *Each state s_i is drawn either i) i.i.d. from a distribution $\rho(s)$ with support over S or ii) from an ergodic Markov chain with stationary distribution $\rho(s)$ defined over a σ-algebra that is countably generated from S.*

Observe that the ergodic Markov chain in Assumption 1 does not have to be that followed by $\pi(a|s)$ and hence our algorithms can also be off-policy as long as the underlying Markov chain is ergodic with stationary distribution: $\rho(s)$. In such a case, the expectation under the evaluation policy $\pi(a|s)$ used in the empirical Bellman function may be estimated using importance sampling.

Assumption 2 (Regularity of Model)**.** *i) $\hat{Q}_\omega$ is bounded and (Φ, d_Φ) and (Ω, d_Ω) are compact metric spaces; ii) $\hat{B}_\phi$ is Lipschitz in ϕ, $P(b|s,a,\phi)$ has finite variance and a density $p(b|s,a,\phi)$ which is Lipschitz in ϕ and bounded; and iii) $p(\phi) \propto \exp(-R(\phi))$ where $R(\phi)$ is bounded and Lipschitz.*

Any Q-function is upper and lower bounded as:

$$\frac{r_{\min}}{1-\gamma} \leq Q^\pi \leq \frac{r_{\max}}{1-\gamma},$$

which can be used as a natural bound when designing the Q-function approximator to ensure Assumption 2 i) is satisfied. As the Q-function approximator and reward function are bounded, it follows from Eq. (2) that each $b \sim P_B$ is bounded too, hence $P_B(b|s,a,\omega)$ has finite variance. The assumption that the model $P(b|s,a,\phi)$ has finite variance under Assumption 2 ii) therefore does not affect its capacity to represent $P_B(b|s,a,\omega)$. Finally, the prior being bounded under Assumption 2 iii) avoids pathological cases where the prior places all its mass on a finite number of parametrisations.

Assumption 3 (Single Minimiser)**.** *The set of maximum likelihood parameters $\phi^\star_\omega$ defined in Eq.* (8) *exists and is a singleton.*

Assumption 3 is used to simplify analysis and exposition, allowing us to prove convergence to a single Dirac-delta measure in Lemma 3. In the more realistic situation where the KL divergence may have multiple, disjoint minimisers, our analysis holds, however the posterior converges to a weighted sum of Dirac-delta measures centred on each element of the set of minimisers specific to the exact MDP being studied.

Central to our proofs is the empirical regularised log-likelihood (ERLL):

$$\text{ERLL}_N(\phi, \mathcal{D}^N_\omega) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \log p(b_i|s_i,a_i,\phi) - \frac{\sigma^2}{N} R(\phi).$$

Using this notation, we can write the posterior density as:

$$p(\phi|\mathcal{D}^N_\omega) = \frac{\exp\left(\frac{N}{\sigma^2}\text{ERLL}_N(\phi, \mathcal{D}^N_\omega)\right)}{\int_\Phi \exp\left(\frac{N}{\sigma^2}\text{ERLL}_N(\phi, \mathcal{D}^N_\omega)\right) d\lambda(\phi)} \tag{15}$$

where λ is the Lebesgue measure. Our proofs require three separate notions of convergence which we now make precise.

$\mathbf{P}_{\mathcal{D}}$-almost sure convergence: We denote the distribution of the complete data $\mathcal{D}_\omega := \{b_0, s_0, a_0, b_1, s_1, a_1 ...\}$ as $P_{\mathcal{D}}$. $\mathcal{D}_\omega^N \subset \mathcal{D}_\omega$ for any finite N. A sequence of random variables $Z_N : \mathcal{D}_\omega \to \mathbb{R}$ converges $P_{\mathcal{D}}$-almost surely to Z if $P_{\mathcal{D}}(\mathcal{D} : \lim_{N\to\infty} Z_N(\mathcal{D}) = Z(\mathcal{D})) = 1$ for all $\mathcal{D} \in \mathcal{D}_\omega$. We denote $P_{\mathcal{D}}$-almost sure convergence of Z_N as $Z_N \xrightarrow{P_{\mathcal{D}}-a.s.} Z$.

Weak $\mathbf{P}_{\mathcal{D}}$-almost sure convergence: Our theorems analyse the behaviour of the posterior $P(\phi|\mathcal{D}_\omega^N)$ which depends on the data. We must therefore extend the usual notion of weak convergence to weak $P_{\mathcal{D}}$-almost sure convergence to account for this dependence by characterising the $P_{\mathcal{D}}$-almost sure convergence of the random variable $\int_\Phi f(\phi) dP(\phi|\mathcal{D}_\omega^N)$:

Definition 1 (Weak $P_{\mathcal{D}}$-almost sure convergence:)**.** *A distribution $P_N(\phi|\mathcal{D})$ converges weakly to $P^\star(\phi)$ $P_{\mathcal{D}}$-almost surely if for any continuous, bounded $f : \Phi \to \mathbb{R}$:*

$$P_{\mathcal{D}}\left(\mathcal{D} : \lim_{N\to\infty} \int_\Phi f(\phi) dP_N(\phi|\mathcal{D}) = \int_\Phi f(\phi) dP^\star(\phi)\right) = 1,$$

for all $\mathcal{D} \in \mathcal{D}_\omega$.

We denote weak $P_{\mathcal{D}}$-almost sure convergence of $P_N(\phi|\mathcal{D})$ as $P_N(\phi|\mathcal{D}) \xrightarrow{P_{\mathcal{D}}-a.s.} P^\star(\phi)$

Uniform and Uniform $\mathbf{P}_{\mathcal{D}}$-almost sure convergence Informally, uniform convergence strengthens the notion of pointwise convergence, ensuring that for every scalar $\varepsilon > 0$, a sequence of functions remain uniformly bounded within a margin ε of their limiting value after a fixed number in the sequence. Uniform convergence is formally defined as:

Definition 2 (Uniform Convergence)**.** *Let $f_N : X \to \mathbb{R}$ be a sequence of real-valued functions. The sequence (f_N) converges uniformly to $f : X \to \mathbb{R}$ on X if for every $\varepsilon > 0$ there exists some natural number K such that for all $x \in X$ and $N \geq K$,*

$$|f_N(x) - f(x)| < \varepsilon.$$

An equivalent definition of uniform converge is:

$$\lim_{N\to\infty} \sup_{x\in X} |f_N(x) - f(x)| = 0.$$

We will denote uniform convergence of $f_N(x)$ as $f_N \xrightarrow{\text{unif}} f$.

As we will be proving uniform convergence of the ERLL, which is a sequence of random variables, we must extend the notion of uniform convergence to uniform $P_{\mathcal{D}}$-almost sure convergence by replacing the pointwise convergence condition used to define almost sure convergence with the uniform convergence condition:

Definition 3 (Uniform $P_{\mathcal{D}}$-almost sure convergence)**.** *Let $Z_N : \mathcal{D}_\omega \to \mathbb{R}$ be a sequence of random variables. The sequence (Z_N) converges uniformly to $Z : \mathcal{D}_\omega \to \mathbb{R}$ on $P_{\mathcal{D}}$-almost surely if for every $\varepsilon > 0$ there exists some natural number K such that for all $\mathcal{D} \in \mathcal{D}_\omega$ and $N \geq K$,*

$$|Z_N(\mathcal{D}) - Z(\mathcal{D})| < \varepsilon,$$

except possibly a some subset $\mathcal{D}' \subset \mathcal{D}_\omega$ such that $P_{\mathcal{D}}(\mathcal{D}') = 0$. An equivalent definition of uniform $P_{\mathcal{D}}$-almost sure converge is:

$$P_{\mathcal{D}}\left(\mathcal{D}' : \lim_{N\to\infty} \sup_{\mathcal{D}\in\mathcal{D}'} |Z_N(\mathcal{D}) - Z(\mathcal{D})| = 0\right) = 1.$$

We denote uniform $P_{\mathcal{D}}$-almost sure convergence of the sequence (f_N) as $f_N \xrightarrow{\text{unif}-P_{\mathcal{D}}} f$. We now start with a proposition that establishes a few useful facts about functions that depend on ϕ from our assumptions.

Proposition 1 (Useful Facts About Functions of ϕ)**.** *Under Assumption 2, $i)$ $\hat{B}_\phi$ is bounded; and $ii)$ $\mathbb{E}_{P_B(b,s,a,|\omega)}\left[|\log p(b|s,a,\phi)|\right] < \infty \; \forall \; \phi \in \Phi$.*

Proof. To prove i), we recall that $\hat{B}_\phi$ is defined as the model's mean: $\hat{B}_\phi(s,a) \coloneqq \mathbb{E}_{P(b|s,a,\phi)}[b]$. If $\hat{B}_\phi(s,a)$ is not bounded for all (s,a), the variance of $P(b|s,a,\phi)$ would also not be bounded, which would lead to a contradiction as $P(b|s,a,\phi)$ has finite variance under Assumption 2. To prove ii), we note that as $p(b|s,a,\phi)$ is bounded under Assumption 2, it must be P_B-integrable and hence $\mathbb{E}_{P_B(b,s,a,|\omega)}\left[|\log p(b|s,a,\phi)|\right] < \infty \; \forall \; \phi \in \Phi$. □

C.2 Proof of Theorem 1

Our first proof establishes uniform almost sure convergence of the ERLL (see Appendix C.1 for a detailed definition of the ERLL and uniform almost sure convergence) under the assumptions of Theorem 1.

Lemma 1 (Uniform Almost Sure Convergence of ERLL)**.** *Under Assumptions 1 and 2,* $\text{ERLL}_N(\phi, \mathcal{D}_\omega^N) \xrightarrow{\text{unif}-P_\mathcal{D}} \mathbb{E}_{P_B(b|s,a,\omega)}\left[\log p(b|s,a,\phi)\right]$.

Proof. Applying the triangle inequality to Definition 3 yields:

$$\begin{aligned}
&\sup_{\phi\in\Phi}\left|\text{ERLL}_N(\phi,\mathcal{D}_\omega^N) - \mathbb{E}_{P_B(b|s,a,\omega)}\left[\log p(b|s,a,\phi)\right]\right| \\
&= \sup_{\phi\in\Phi}\left|\frac{1}{N}\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\phi) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right] - \frac{1}{N}R(\phi)\right|, \\
&\leq \sup_{\phi\in\Phi}\left|\frac{1}{N}\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\phi) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right| + \frac{1}{N}\sup_{\phi\in\Phi}\left|R(\phi)\right|.
\end{aligned}$$

As $R(\phi)$ is bounded by Assumption 2, $\lim_{N\to\infty}\frac{1}{N}\sup_{\phi\in\Phi}\left|R(\phi)\right| = 0$, hence:

$$\begin{aligned}
&\lim_{N\to\infty}\sup_{\phi\in\Phi}\left|\text{ERLL}_N(\phi,\mathcal{D}_\omega^N) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right| \\
&\quad\leq \lim_{N\to\infty}\sup_{\phi\in\Phi}\left|\frac{1}{N}\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\phi) - \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]\right|.
\end{aligned}$$

We are therefore left to prove:

$$\frac{1}{N}\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\phi) \xrightarrow{\text{unif}-P_\mathcal{D}} \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right] \tag{16}$$

Theorem 3 of Andrews [3] states that (16) holds if i) Φ is totally bounded, ii) $\log p(b|s,a,\phi)$ is Lipschitz in ϕ and iii) the empirical mean converges pointwise almost surely, that is:

$$\frac{1}{N}\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\phi) \xrightarrow{P_\mathcal{D}-a.s.} \mathbb{E}_{P_B(b,s,a|\omega)}\left[\log p(b|s,a,\phi)\right]$$

Conditions i) and ii) are satisfied by Assumption 2. To prove pointwise strong convergence for condition iii), we use the strong law of large numbers (SLLN) under the two sampling options presented in Assumption 1. As $\mathbb{E}_{P_B(b,s,a|\omega)}[|\log p(b|s,a,\phi)|] < \infty$ from Proposition 1, the pointwise SLLN holds for i.i.d. samples by (for example) Williams [79] Theorem 14.5 or for sampling from an ergodic Markov chain under the conditions of Assumption 1 by Theorem 4.3 of Gilks et al. [36]. Conditions i)-iii) are satisfied, hence (16) holds, completing our proof. □

An important consequence of uniform convergence is that any sequence of supremums and infimums of the sequence of continuous, bounded functions also converges:

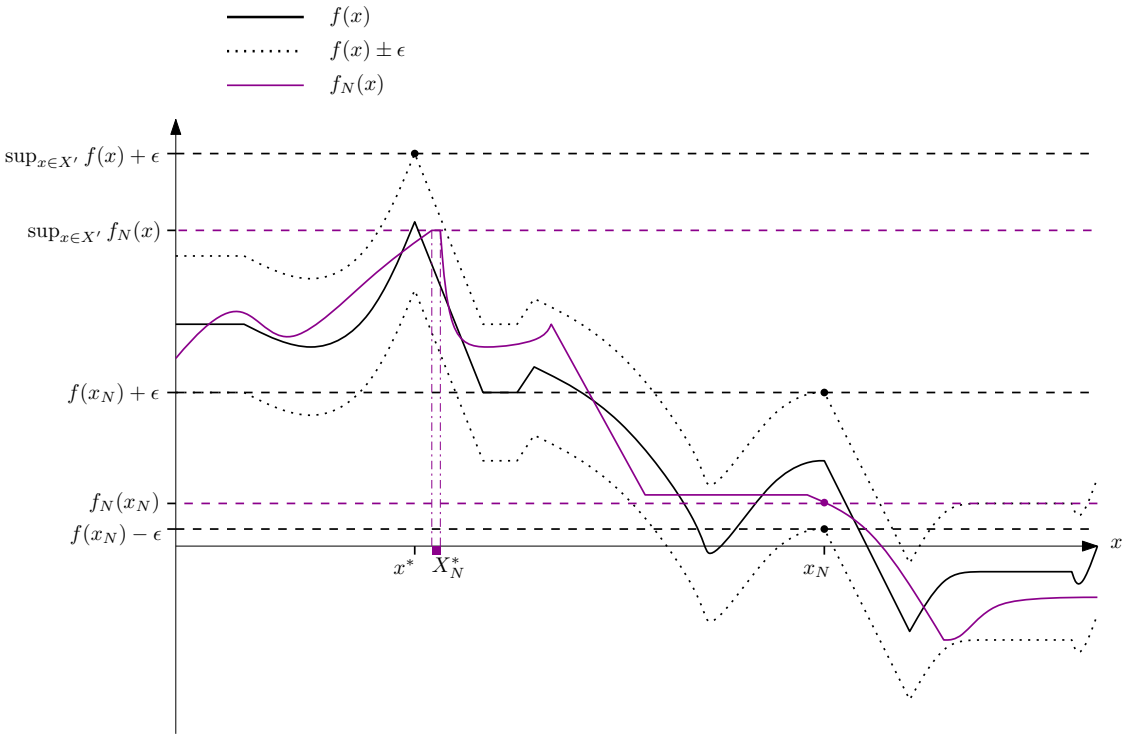


Figure 6: Sketch of Lemma 2.

Lemma 2 (Continuity of sup and inf)**.** *If a sequence of continuous, bounded functions $f_N : X' \to \mathbb{R}$ converges uniformly to f on $X \subseteq X'$, then*

$$i) \lim_{N \to \infty} \sup_{x \in X} f_N(x) = \sup_{x \in X} f(x),$$
$$ii) \lim_{N \to \infty} \inf_{x \in X} f_N(x) = \inf_{x \in X} f(x).$$

Proof. We begin by proving i). To aid the reader's understanding, a sketch of our proof is given in Fig. 6. The bounded sequence $(\sup_{x \in X} f_N(x))$ converges to $\sup_{x \in X} f(x)$ if and only if

$$\limsup_{N \to \infty} \sup_{x \in X} f_N(x) = \liminf_{N \to \infty} \sup_{x \in X} f_N(x) = \sup_{x \in X} f(x). \tag{17}$$

As

$$\limsup_{N \to \infty} \sup_{x \in X} f_N(x) \geq \liminf_{N \to \infty} \sup_{x \in X} f_N(x),$$

it suffices to show:

$$\limsup_{N \to \infty} \sup_{x \in X} f_N(x) \leq \sup_{x \in X} f(x), \tag{18}$$
$$\liminf_{N \to \infty} \sup_{x \in X} f_N(x) \geq \sup_{x \in X} f(x). \tag{19}$$

We begin by proving the lower bound in (18). Denote the preimage of $\sup_{x \in X} f_N(x)$ as:

$$X_N^* \coloneqq \left\{ x : f_N(x) = \sup_{x \in X} f_N(x) \right\}.$$

By the definition of uniform convergence (Definition 2), for every $\varepsilon > 0$ there exists a $K \in \mathbb{N}$ such that for all $N \geq K$ and $x_N^* \in X_N^*$:

$$f_N(x_N^*) < f(x_N^*) + \varepsilon \leq \sup_{x \in X} f(x) + \varepsilon,$$
$$\implies \sup_{x \in X} f_N(x) < \sup_{x \in X} f(x) + \varepsilon. \tag{20}$$

As ε is arbitrary, it follows that:

$$\limsup_{N\to\infty} \sup_{x\in X} f_N(x) \leq \sup_{x\in X} f(x),$$

or else it would be possible to find a $\varepsilon' > 0$ and K such that $\sup_{x\in X} f_N(x) \geq \sup_{x\in X} f(x) + \varepsilon'$ for all $N \geq K$, which contradicts Eq. (20).

Now to show the upper bound in (19) holds using a similar analysis, consider any sequence $x_N \to x^*$ where $f(x^*) = \sup_{x\in X} f(x)$. By the definition of uniform convergence (Definition 2), for every $\varepsilon > 0$ there exists a $K \in \mathbb{N}$ such that for all $N \geq K$:

$$\begin{aligned} &\sup_{x\in X} f_N(x) \geq f_N(x_N) > f(x_N) - \varepsilon, \\ \implies &\sup_{x\in X} f_N(x) > f(x_N) - \varepsilon. \end{aligned} \tag{21}$$

As ε is arbitrary, it follows that:

$$\liminf_{N\to\infty} \sup_{x\in X} f_N(x) \geq \liminf_{N\to\infty} f(x_N),$$

or else it would be possible to find a $\varepsilon' > 0$ and K such that $\sup_{x\in X} f_N(x) \leq f(x_N) - \varepsilon'$ for all $N \geq K$, which contradicts Eq. (21). By the definition of continuity of f:

$$\liminf_{N\to\infty} f(x_N) = \lim_{N\to\infty} f(x_N) = f(x^*) = \sup_{x\in X} f(x).$$

which proves that the upper bound holds:

$$\liminf_{N\to\infty} \sup_{x\in X} f_N(x) \geq \sup_{x\in X} f(x).$$

As (18) and (19) hold then i) follows immediately from Eq. (17).

To prove ii), we note that as f_N and f are bounded, $\inf_{x\in X} f(x) = -\sup_{x\in X}(-f(x))$ and $\inf_{x\in X} f_N(x) = -\sup_{x\in X}(-f_N(x))$. Using $-f$ and $-f_N$ in place of f and f_N, the inequalities (18) and (19) still hold, hence:

$$\begin{aligned} &\limsup_{N\to\infty} \sup_{x\in X}(-f_N(x)) = \liminf_{N\to\infty} \sup_{x\in X}(-f_N(x)) = \sup_{x\in X}(-f(x)), \\ \implies &\limsup_{N\to\infty} \inf_{x\in X} f_N(x) = \liminf_{N\to\infty} \inf_{x\in X} f_N(x) = \inf_{x\in X} f(x), \\ \implies &\lim_{N\to\infty} \inf_{x\in X} f_N(x) = \inf_{x\in X} f(x), \end{aligned}$$

as required. □

In the context of BBO, Lemma 2 implies that any sequence of minimisers/maximisers of the ERLL converges pointwise, which we now use in Lemma 3 to prove that our posterior concentrates on the KL-minimising parameters:

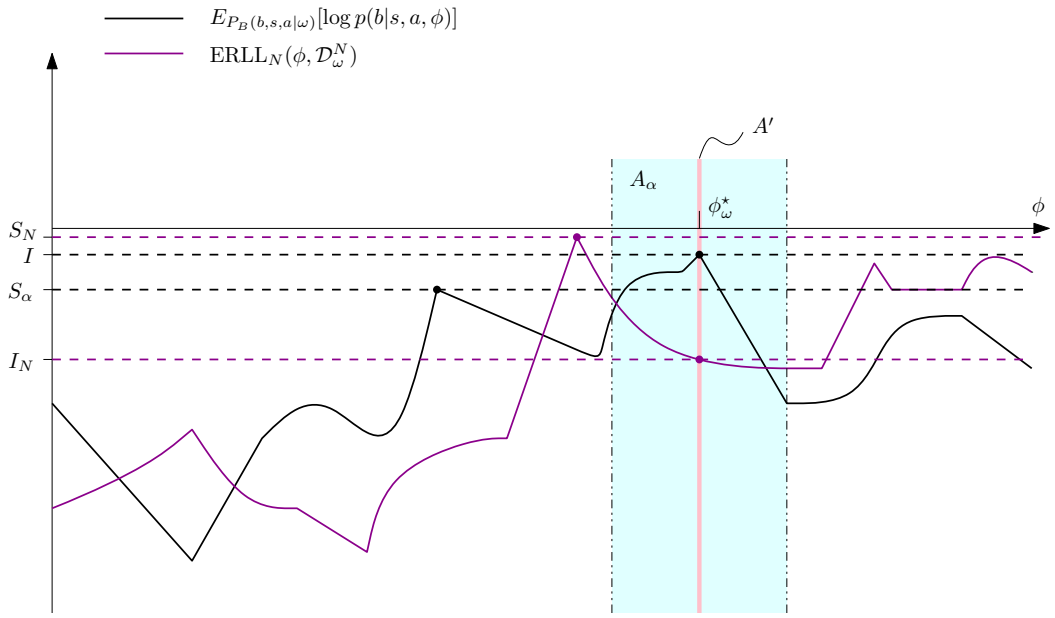


Figure 7: Sketch of Lemma 3.

Lemma 3 (Posterior Concentration)**.** *Under Assumptions 1-3, the posterior converges weak $P_{\mathcal{D}}$-almost surely to a Dirac delta distribution centered on the parameters that minimises the KL divergence:*

$$P(\phi|\mathcal{D}_\omega^N) \xrightarrow{P_{\mathcal{D}}-a.s.} \delta(\phi = \phi^\star_\omega)$$

where

$$\phi^\star_\omega = \arg\max_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$$

Proof. Consider an open ball of radius α centered on $\phi^\star_\omega$:

$$A_\alpha \coloneqq \{\phi : d_\Phi(\phi, \phi^\star_\omega) < \alpha\}$$

for $\alpha > 0$. From the definition of weak convergence of measures [13], it suffices to show that $\lim_{N\to\infty} P(A_\alpha|\mathcal{D}_\omega^N) = 1\ P_{\mathcal{D}} - a.s.\ \forall\ \alpha > 0$ or equivalently,

$$\lim_{N\to\infty} P(\Phi \setminus A_\alpha|\mathcal{D}_\omega^N) = 0 \quad P_{\mathcal{D}} - a.s.\ \forall\ \alpha > 0.$$

From Kolomogrov's first axiom $P(\Phi \setminus A_\alpha|\mathcal{D}_\omega^N) \geq 0\ \forall\ N$, hence we are left to prove:

$$\lim_{N\to\infty} \int_{\Phi\setminus A_\alpha} dP(\phi|\mathcal{D}_\omega^N) \leq 0 \quad P_{\mathcal{D}} - a.s.\ \forall\ \alpha > 0.$$

To aid the reader's understanding of our proof, we provide a sketch in Fig. 7. Let

$$S_\alpha \coloneqq \sup_{\phi\in\Phi\setminus A_\alpha} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right] < \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi^\star_\omega)\right].$$

From Assumption 3, $\phi^\star_\omega$ is a unique maximiser of $\mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right]$, hence by continuity there exists some open subset $A' \subset A_\alpha$ such that

$$I \coloneqq \inf_{\phi\in A'} \mathbb{E}_{P_B(b,s,a|\omega)} \left[\log p(b|s,a,\phi)\right] > S_\alpha\ \forall\ \alpha > 0.$$

Writing the posterior measure as an integral over its density using Eq. (15) yields:

$$P(\Phi \setminus A_\alpha|\mathcal{D}_\omega^N) = \frac{\int_{\Phi\setminus A_\alpha} \exp\left(\frac{N}{\sigma^2}\mathrm{ERLL}_N(\phi, \mathcal{D}_\omega^N)\right) d\lambda(\phi)}{\int_\Phi \exp\left(\frac{N}{\sigma^2}\mathrm{ERLL}_N(\phi, \mathcal{D}_\omega^N)\right) d\lambda(\phi)}, \tag{22}$$

where λ is the Lebesgue measure. Let $I_N := \inf_{\phi \in A'} \text{ERLL}_N(\phi, \mathcal{D}_\omega^N)$ and $S_N := \sup_{\phi \in \Phi \setminus A_\alpha} \text{ERLL}_N(\phi, \mathcal{D}_\omega^N)$. Consider the numerator of Eq. (22), which we can upper bound using S_N:

$$\begin{aligned}\int_{\Phi \setminus A_\alpha} \exp\left(\frac{N}{\sigma^2}\text{ERLL}_N(\phi, \mathcal{D}_\omega^N)\right) d\lambda(\phi) &\leq \int_{\Phi \setminus A_\alpha} \exp\left(\frac{N}{\sigma^2} S_N\right) d\lambda(\phi), \\ &= \lambda(\Phi \setminus A_\alpha) \exp\left(\frac{N}{\sigma^2} S_N\right), \end{aligned} \tag{23}$$

Similarly, we can lower bound the denominator of Eq. (22) using I_N and, since the integrand is positive over Φ, by changing the integration over Φ to over A' :

$$\begin{aligned}\int_{\Phi} \exp\left(\frac{N}{\sigma^2}\text{ERLL}_N(\phi, \mathcal{D}_\omega^N)\right) d\lambda(\phi) &\geq \int_{A'} \exp\left(\frac{N}{\sigma^2}\text{ERLL}_N(\phi, \mathcal{D}_\omega^N)\right) d\lambda(\phi), \\ &\geq \int_{A'} \exp\left(\frac{N}{\sigma^2} I_N\right) d\lambda(\phi), \\ &= \lambda(A') \exp\left(\frac{N}{\sigma^2} I_N\right). \end{aligned} \tag{24}$$

Together, Eqs. (23) and (24) allow us to upper bound the sequence of integrals in Eq. (22):

$$\begin{aligned} P(\Phi \setminus A_\alpha | \mathcal{D}_\omega^N) &\leq \frac{\lambda(\Phi \setminus A_\alpha) \exp\left(\frac{N}{\sigma^2} S_N\right)}{\lambda(A') \exp\left(\frac{N}{\sigma^2} I_N N\right)}, \\ &= \frac{\lambda(\Phi \setminus A_\alpha)}{\lambda(A')} \exp\left((S_N - I_N)\frac{N}{\sigma^2}\right). \end{aligned}$$

Using Lemma 2 and Lemma 1, we can take the limit of $S_N - I_N$ as

$$\lim_{N \to \infty} (S_N - I_N) = S_\alpha - I, \qquad P_\mathcal{D} - a.s.$$

Since $I > S_\alpha$ and hence $S_\alpha - I < 0$, we can take limits to obtain the desired bound:

$$\begin{aligned} \lim_{N \to \infty} P(\Phi \setminus A_\alpha | \mathcal{D}_\omega^N) &\leq \frac{\lambda(\Phi \setminus A_\alpha)}{\lambda(A')} \lim_{N \to \infty} \exp\left((S_N - I_N)\frac{N}{\sigma^2}\right), \\ &= \frac{\lambda(\Phi \setminus A_\alpha)}{\lambda(A')} \exp\left((S_\alpha - I)\frac{\lim_{N \to \infty} N}{\sigma^2}\right) \; P_\mathcal{D} - a.s., \\ &= 0 \quad P_\mathcal{D} - a.s. \; \forall \; \alpha > 0. \end{aligned}$$

□

Theorem 1. *Under Assumptions 1-3, in the limit $N \to \infty$ the posterior concentrates weakly on $\phi^\star$: i) $P(\phi|\mathcal{D}_\omega^N) \xrightarrow{P_\mathcal{D} - a.s.} \delta(\phi = \phi_\omega^\star)$ with ii) $\mathcal{B}_{\omega,N}^\star \xrightarrow{P_\mathcal{D} - a.s.} \hat{B}_{\phi_\omega^\star}$, iii) MSBBE$_N(\omega) \xrightarrow{P_\mathcal{D} - a.s.} \|\hat{Q}_\omega - \hat{B}_{\phi_\omega^\star}\|_{\rho,\pi}^2$.*

Proof.

$$\textbf{Claim } \boldsymbol{i}) : \quad P(\phi|\mathcal{D}_\omega^N) \xrightarrow{P_\mathcal{D} - a.s.} \delta(\phi = \phi_\omega^\star).$$

Claim $i)$ follows immediately from Lemma 3.

$$\textbf{Claim } \boldsymbol{ii}) : \quad \mathcal{B}_{\omega,N}^\star \xrightarrow{P_\mathcal{D} - a.s.} \hat{B}_{\phi_\omega^\star}.$$

To prove $ii)$ we analyse the convergence of the Bayesian Bellman operator, writing it as an expectation:

$$\lim_{N \to \infty} \mathcal{B}_{\omega,N}^\star = \lim_{N \to \infty} \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}_\omega^N).$$

From $i)$, the posterior converges weakly to a Dirac delta distribution centered on $\phi^\star_\omega$ $P_\mathcal{D}$-almost surely. From Proposition 1 $\hat{B}_\phi$ is bounded and from Assumption 2 is Lipschitz in ϕ, so we can apply the portmanteau theorem for weak convergence of measures [13] to find the limit:

$$\lim_{N\to\infty} \mathcal{B}^\star_{\omega,N} = \lim_{N\to\infty} \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}^N_\omega) = \int_\Phi \hat{B}_\phi d\delta(\phi = \phi^\star_\omega) = \hat{B}_{\phi^\star_\omega} \quad P_\mathcal{D} - a.s., \tag{25}$$

as required.

$$\textbf{Claim } \textit{iii)}: \quad \text{MSBBE}_N(\omega) \xrightarrow{P_\mathcal{D}-a.s.} \|\hat{Q}_\omega - \hat{B}_{\phi^\star_\omega}\|^2_{\rho,\pi}.$$

To prove $iii)$, we start with the definition of the MSBBE in the limit $N \to \infty$:

$$\begin{aligned}\lim_{N\to\infty} \text{MSBBE}_N(\omega) &= \lim_{N\to\infty} \left\|\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi} \\ &= \lim_{N\to\infty} \frac{1}{2} \int_{\mathcal{S}\times\mathcal{A}} \left(\hat{Q}_\omega(s,a) - \mathcal{B}^\star_{\omega,N}(s,a)\right)^2 d(\rho(s)\pi(a|s)).\end{aligned} \tag{26}$$

To apply the dominated convergence theorem to Eq. (26), we must show the integrand is dominated and convergences pointwise [9]. Firstly, from Eq. (25), we have $\mathcal{B}^\star_{\omega,N} \xrightarrow{P_\mathcal{D}-a.s.} \hat{B}_{\phi^\star_\omega}$, which is bounded by Proposition 1. Now consider the Bayesian Bellman operator for finite N: $\mathcal{B}^\star_{\omega,N} = \int_\Phi \hat{B}_\phi dP(\phi|\mathcal{D}^N_\omega)$. As $\hat{B}_\phi$ is bounded, Φ is compact and $P(\phi|\mathcal{D}^N_\omega)$ is absolutely continuous with respect to the Lebesgue measure, it follows that $\hat{B}_\phi$ is $P(\phi|\mathcal{D}^N_\omega)$-integrable hence $\mathcal{B}^\star_{\omega,N}$ is bounded for all N by some positive constant $\tilde{C}$.

Consider now the integrand in Eq. (26). Applying the triangle inequality, we have:

$$|\hat{Q}_\omega(s,a) - \mathcal{B}^\star_{\omega,N}(s,a)|^2 \leq (|\hat{Q}_\omega(s,a)| + |\mathcal{B}^\star_{\omega,N}(s,a)|)^2 \leq (|\hat{Q}_\omega(s,a)| + \tilde{C})^2.$$

From Assumption 2 $\hat{Q}_\omega$ is bounded and so the integrand is dominated by some finite constant. The dominated convergence theorem therefore applies for Eq. (26), hence

$$\lim_{N\to\infty} \text{MSBBE}_N(\omega) = \left\|\hat{Q}_\omega - \lim_{N\to\infty} \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi} = \left\|\hat{Q}_\omega - \hat{B}_{\phi^\star_\omega}\right\|^2_{\rho,\pi} \quad P_\mathcal{D} - a.s,$$

as required. □

Finally, we prove our corollary, which establishes the results of Theorem 1 when RP approximate posterior is used in place of the true posterior.

Corollary 1.1. *Under Assumptions 1-4, results i)-iii) of Theorem 1 hold with $P(\phi|\mathcal{D}^N_\omega)$ replaced by the RP approximate posterior $q(\phi|\mathcal{D}^N_\omega)$ both with or without ensembling.*

Proof. As results $ii)$ and $iii)$ of Theorem 1 follow directly from $i)$ under Assumptions 1-3, we only need to prove $i)$ still holds with the approximate posterior, i.e. that

$$q(\phi|\mathcal{D}^N_\omega) \xrightarrow{P_\mathcal{D}-a.s.} \delta(\phi = \phi^\star_\omega).$$

We consider the case of using the exact RP posterior (without ensembling) $q(\phi|\mathcal{D}^N_\omega) \coloneqq \int_\mathcal{E} \delta(\phi \in \psi(\mathcal{D}^N_\omega, \epsilon))dP_E(\epsilon)$. From the Portmanteau Theorem for the weak convergence of measures [13], it suffices to prove:

$$\int_\Phi f(\phi)dq(\phi|\mathcal{D}^N_\omega) \xrightarrow{P_\mathcal{D}-a.s.} \int_\Phi f(\phi)d\delta(\phi = \phi^\star_\omega) = f(\phi^\star_\omega). \tag{27}$$

for any Lipschitz, bounded function $f : \Phi \to \mathbb{R}$. Substituting for the definition of the RP approximate posterior $q(\phi|\mathcal{D}^N_\omega) \coloneqq \int_\mathcal{E} \delta(\phi \in \psi(\mathcal{D}^N_\omega, \epsilon))dP_E(\epsilon)$, proving (27) holds is equivalent to proving:

$$\int_\mathcal{E} f \circ \psi_N(\epsilon)dP_E(\epsilon) \xrightarrow{P_\mathcal{D}-a.s.} f(\phi^\star_\omega), \tag{28}$$

for any sequence $(\psi_N(\epsilon))$ where $\psi_N(\epsilon) \in \psi(\mathcal{D}_\omega^N, \epsilon)$. Under the definition of RP (see Appendix F), it is implicitly assumed that $f \circ \psi_N : \mathcal{E} \to \mathbb{R}$ is P_E-integrable for any bounded, Lipschitz $f : \Phi \to \mathbb{R}$. Hence, we can apply the dominated convergence theorem to (28) to derive an equivalent condition to prove that (27) holds:

$$f \circ \psi_N(\epsilon) \xrightarrow{P_\mathcal{D}-a.s.} f(\phi_\omega^\star) \quad \forall\, \epsilon \in \mathcal{E},$$

which, from the continuity of f, is equivalent to proving that a sequence of minimisers of $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ converge almost surely to the maximisers of $\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$, that is:

$$\psi_N(\epsilon) \xrightarrow{P_\mathcal{D}-a.s.} \phi_\omega^\star \quad \forall\, \epsilon \in \mathcal{E}. \tag{29}$$

Theorem 7.33 of Rockafellar and Wets [63] states that (29) holds if for all $\epsilon \in \mathcal{E}$: $a)$ $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ and $-\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$ are proper, lower semi-continuous functions of ϕ; $b)$ the sequence $(\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon))$ is eventually level-bounded; and $c)$ $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ epi-converges to $-\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)]$.

Condition $a)$ is trivially satisfied by the assumption of $p(b|s,a,\phi)$ being bounded and Lipschitz in ϕ in Assumption 2, and as a continuous function, is lower semi-continuous as any continuous mapping between any two metric spaces is proper. Recall the definition of $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$:

$$\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) \coloneqq -\frac{1}{N}\sum_{i=1}^N \log p(b_i|s_i,a_i,\phi) + \frac{1}{N}R(\phi - \epsilon).$$

As $p(b|s,a,\phi)$ is bounded and $R(\phi-\epsilon)$ is bounded by some $R_{\text{max}} > 0$ under Assumption 2, there exists a $K > 0$ such that $|\log p(b_i|s_i,a_i,\phi)| \leq K$ which we use to bound $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$:

$$\begin{aligned}
|\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)| &\leq \frac{1}{N}\sum_{i=1}^N |\log p(b_i|s_i,a_i,\phi)| + \frac{1}{N}|R(\phi-\epsilon)|, \\
&\leq \frac{1}{N}\sum_{i=1}^N K + \frac{R_{\text{max}}}{N}, \\
&= K + \frac{R_{\text{max}}}{N}.
\end{aligned}$$

As $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ is bounded for all $N \in \mathbb{N}$ and $\epsilon \in \mathcal{E}$, the sequence is trivially eventually level-bounded, hence $b)$ is satisfied.

To establish epi-convergence, we first establish uniform convergence. As $\lim_{N\to\infty} \frac{1}{N}\sup_{\phi\in\Phi}|R(\phi-\epsilon)| = 0$ for all $\phi \in \Phi$ and $\epsilon \in \mathcal{E}$, we can use Lemma 1 with $\text{ERLL}_N(\phi, \mathcal{D}_\omega^N)$ replaced by $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ to prove uniform almost sure convergence:

$$\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon) \xrightarrow{\text{unif}-P_\mathcal{D}} -\mathbb{E}_{P_B(b|s,a,\omega)}[\log p(b|s,a,\phi)] \quad \forall\, \epsilon \in \mathcal{E}. \tag{30}$$

As we have already proved that each $\mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ is lower semi-continuous, Proposition 7.15 a) of Rockafellar and Wets [63] applies, which strengthens uniform convergence of (30) to epi-convergence $P_\mathcal{D}$-almost surely. Condition $c)$ is satisfied and hence our desired result hold. All arguments above also hold using the ensembled approximation $q(\phi|\mathcal{D}_\omega^N) \approx \frac{1}{L}\sum_{l=1}^N \delta(\phi \in \psi(\mathcal{D}_\omega^N, \epsilon_l))$ if we replace expectations of any function $f(\epsilon)$ under $P_E(\epsilon)$: $\int_\mathcal{E} f(\epsilon) dP_E(\epsilon)$ with expectations over ensembles: $\frac{1}{L}\sum_{l=1}^N f(\epsilon_l)$. □

C.3 Assumptions and Preliminaries for Theorem 2

As we are concerned with finite N in our Bayesian analysis, we analyse the RP objective ignoring the factor of $\frac{1}{N}$ as it leaves the solution unchanged:

$$\mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l) \coloneqq R(\phi - \epsilon_l) - \sum_{i=1}^N \log p(b_i|s_i,a_i,\phi).$$

We repeat the sequence of updates from Section 4 here for convenience:

$$\psi_l \leftarrow \mathcal{P}_\Omega \left(\psi_l - \alpha_k \nabla_{\psi_l} \left(R(\psi_l - \epsilon_l) - \log p(b_i|s_i, a_i, \psi_l)\right)\right), \quad \text{(fast)} \tag{31}$$

$$\omega_l \leftarrow \mathcal{P}_\Omega(\omega_l - \beta_k(\omega_l - \psi_l)). \quad \text{(slow)} \tag{32}$$

To analyse the limiting ODE of this sequence of updates we require an assumption regarding the parameter space:

Assumption 4 (RP Function Spaces)**.** *i) $\hat{Q}_{\omega_l}$ and $\hat{B}_{\omega_l}$ share a function space, that is $\hat{Q}_{\omega_l'} = \hat{B}_{\omega_l'}$ for any $\omega_l' \in \Omega$, where $\Phi = \Omega \subset \mathbb{R}^n$ is compact, convex with a smooth boundary. ii) $\mathcal{E} \subseteq \mathbb{R}^n$ and $R(\phi - \epsilon)$ is defined for any $\phi \in \Phi, \epsilon \in \mathcal{E}$.*

Assumption 4 i) is easily satisfied by defining Ω to be a closed ball of arbitrary radius, which should be applicable to the majority of cases where parametric function approximators are used. It is common in RP to use Gaussian a prior P_Φ, which technically violates the assumption that Φ is compact. We can still satisfy Assumption 4 i) by defining P_Φ to be a truncated Gaussian. This won't affect the algorithm or objectives in practice as P_Φ simply adds a regularisation term and there are no restrictions on $\mathcal{E}$, hence P_E can still be Gaussian.

Provided Ω is large enough, there is diminishing probability that our update ever leaves Ω, especially with regularisation. Hence we do not expect to require projection in practice unless the environment is particularly ill posed. In the unlikely eventuality that projection is required, using a ball makes projection simple as the operator projects back along the line connecting a point to the ball's origin, which defines a normal. In order to derive the limiting ODEs of our updates, we characterise the directional derivative of the projection operator in Proposition 2:

Proposition 2 (Shapiro [65])**.** *The directional derivative of the projection operator at $\omega \in \Omega$ in the direction $y \in \mathbb{R}^n$ given by:*

$$\Gamma_\omega(y) \coloneqq \lim_{\epsilon \downarrow 0} \left(\frac{\mathcal{P}_\Omega(\omega + \epsilon y) - \omega}{\epsilon} \right)$$

which always exists under Assumption 4 and is equivalent to the projection of y onto the tangent cone $T_\Omega(\omega)$:

$$\Gamma_\omega(y) = \mathcal{P}_{T_\Omega(\omega)}(y) \coloneqq \underset{t \in T_\Omega(\omega)}{\arg\min} \|y - t\|_2^2$$

where $T_\Omega(\omega) \coloneqq \text{Closure}\left(\bigcup_{\varepsilon > 0} \bigcup_{\omega' \in \Omega} \frac{1}{\varepsilon}(\omega' - \omega)\right)$

Proof. See Shapiro [65]. □

The directional derivative of the projection operator has been well studied [81, 64, 65, 14] and we provide an informal sketch of its interpretation in Fig. 8 for the reader's intuition. We see that for any point $\omega_1 \in \Omega$ in the interior of Ω, the projection operator is simply the identity function: $\Gamma_{\omega_1}(y) = y$ as $y \in T_\Omega(\omega_1)$. For any point ω_2 on the boundary of Ω, the tangent cone is the closure of the cone formed by all half-lines emanating from ω_2 intersecting Ω in at least one point distinct from ω_2. There are two cases to consider for boundary points. Firstly, when the directional vector y_1 defines a half-line from ω_2 that intersects Ω, the projection operator is the identity $\Gamma_{\omega_2}(y) = y$. In the second case when the directional vector y_2 defines a half-line from ω_2 that leaves Ω, $\Gamma_{\omega_2}(y_2)$ returns the nearest element along the boundary of $T_\Omega(\omega_2)$ (solid purple) according to the projection $\mathcal{P}_{T_\Omega(\omega_2)}(y_2)$.

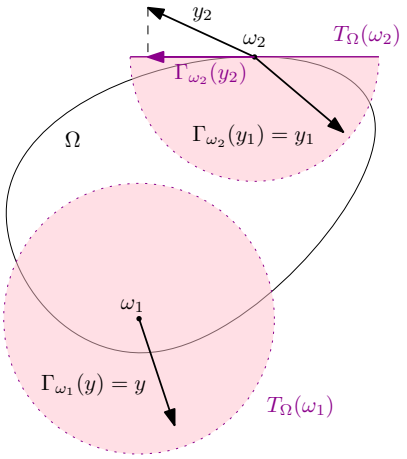


Figure 8: Sketch of Tangent Cones in Ω

Using the directional derivative of the projection operator, we can derive the limiting ODE of update (31):

$$\partial_t \psi_l(t) = -\Gamma_{\psi_l(t)}(\nabla_{\psi_l} \mathcal{L}(\psi_l(t); \mathcal{D}_{\omega_l}^N, \epsilon_l)) \tag{33}$$

Under standard ODE analysis, any equilibria of Eq. (33) satisfy $-\Gamma_{\psi_l(t)}(\nabla_{\psi_l}\mathcal{L}(\psi_l(t);\mathcal{D}^N_{\omega_l},\epsilon_l))=0$ and we denote an asymptotically stable local equilibrium as $\psi^{\circledast}_l(\omega_l)$. Of course, there may be several or even infinite stable local equilibrium for the ODE, but as our assumptions only require the existence of at least one stable attractor within the domain of attraction defined by the initial parametrisations, we lose no generality by considering $\psi^{\circledast}_l(\omega_l)$ in isolation.

As the fast update converges asymptotically quicker than the slower update, we consider ψ_l to be equilibrated at $\psi^{\circledast}_l(\omega_l)$ when analysing the limiting ODE of the slow update. We make this argument rigorously in Theorem 2, from which we derive the limiting ODE for the update (32) as:

$$\partial_t\omega_l(t) = -\Gamma_{\omega_l(t)}(\omega_l(t) - \psi^{\circledast}_l(\omega_l(t))). \tag{34}$$

Crucially, this provides reassurance that our updates and any equilibria satisfying $\Gamma_{\omega_l(t)}(\omega_l(t) - \psi^{\circledast}_l(\omega_l)) = 0$ preserve the dependence of ψ_l on ω_l. We denote an asymptotically stable local equilibrium of Eq. (34) as $\omega^{\circledast}_l$.

Assumption 5 (Two-timescale Regularity)**.** *i)* $\nabla_{\psi_l}R(\psi_l-\epsilon_l)$, $\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)$ *and* $\Gamma_{\psi_l}\left[\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_l},\epsilon)\right]$ *are Lipschitz in* ψ_l, $\Gamma_{\omega_l}\left[-(\omega_l-\psi_l)\right]$ *is Lipschitz in* ω_l *and* $(b_i,s_i,a_i)\sim$ Unif$(\mathcal{D}^N_{\omega_l})$*; ii)* $\psi^{\circledast}(\omega_l)$ *and* $\omega^{\circledast}_l$ *are local aysmptotically stable attractors of the limiting ODEs of updates (13) and (14) respectively and* $\psi^{\circledast}_l(\omega_l)$ *is Lipschitz in* ω_l*; and iii) The stepsizes satisfy:* $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k}=0$, $\sum_{k=1}^{\infty}\alpha_k=\sum_{k=1}^{\infty}\beta_k=\infty$, $\sum_{k=1}^{\infty}\left(\alpha_k^2+\beta_k^2\right)<\infty$.

Compared the presentation in the main body of our paper, we have introduce the additional requirement in $i)$ that $\Gamma_{\psi_l}\left[\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_k},\epsilon)\right]$ and $\Gamma_{\omega_l}\left[-(\omega_l-\psi_l)\right]$ are Lispchtiz. We justify this exclusion from the main body of our text for two reasons: firstly, using the same arguments as Proposition 3 below, it is easy to establish that $\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_l},\epsilon)$ is Lipschitz and by inspection, $(\omega_l-\psi_l)$ is Lipschitz and hence Lipschitzness of $\Gamma_{\psi_k}\left[\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_l},\epsilon)\right]$ and $\Gamma_{\omega_l}\left[-(\omega_l-\psi_l)\right]$ can be established provided Ω is sufficiently large to contain the entire limiting flow of the ODE. For this reason we are essentially encouraged to consider the non-projected form [19]. Secondly, this subtlety is often ignored altogether by existing papers considering two-timescale analysis [12] as it carries a heavy expositional burden but does not effect the algorithm in practice.

As noted by Heusel et al. [42], the assumption of locally asymptotically stable ODEs in Assumption 5 ii) can be ensured by an additional weight decay term in the loss function which increases the eigenvalues of the Hessian. This fits naturally in a Bayesian setting where prior regularisation introduces weight decay into our objectives. Finally, Assumption 5 iii) extends the classic Robbins-Munro stepsize conditions [62] to the two-timescale case. The additional assumption $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k}=0$ ensures that the faster timescale update converges asymptotically faster than the slower update. We now use Assumption 5 to establish the boundedness of several quantities that will be essential for our main proof:

Proposition 3. *Under Assumptions 4 and 5, the quantities i)* $\nabla_{\psi_l}R(\psi_l-\epsilon_l)$ *and* $\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)$*; ii)* $\sum_{j\neq i}^{N}\nabla_{\psi_l}\log p(b_j|s_j,a_j,\psi_l)$*; and iii)* $\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_l},\epsilon_l)$ *are all bounded.*

Proof. We establish $i)$ by noting that from Assumption 5, $\nabla_{\psi_l}R(\psi_l-\epsilon_l)$ and $\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)$ are Lipschitz in ψ_l, and from Assumption 4, Φ is compact, hence as any Lipschitz function defined on a compact set must be bounded, it follows that $\nabla_{\psi_l}R(\psi_l-\epsilon_l)$ and $\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)$ are bounded. To prove $ii)$, we bound $\sum_{j\neq i}^{N}\nabla_{\psi_l}\log p(b_j|s_j,a_j,\psi_l)$ using the triangle inequality:

$$\begin{aligned}\left|\sum_{j\neq i}^{N}\nabla_{\psi_l}\log p(b_j|s_j,a_j,\psi_l)\right| &\leq \sum_{j\neq i}^{N}\left|\nabla_{\psi_l}\log p(b_j|s_j,a_j,\psi_l)\right|,\\ &< N\sup_{j\in\{1:N\}}\left|\nabla_{\psi_l}\log p(b_j|s_j,a_j,\psi_l)\right|,\end{aligned}$$

which is bounded from $i)$ and the fact that N is finite in our Bayesian regime. To prove $iii)$, we establish a similar bound:

$$\begin{aligned}\left|\nabla_{\psi_l}\mathcal{L}(\psi_l;\mathcal{D}^N_{\omega_l},\epsilon_l)\right| &= \left|\nabla_{\psi_l}\left(R(\psi_l-\epsilon_l)-\sum_{i=1}^{N}\log p(b_i|s_i,a_i,\psi_l)\right)\right|,\\ &\leq \left|\nabla_{\psi_l}R(\psi_l-\epsilon_l)\right|+\sum_{i=1}^{N}\left|\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)\right|,\\ &\leq \left|\nabla_{\psi_l}R(\psi_l-\epsilon_l)\right|+N\sup_{i\in\{1:N\}}\left|\nabla_{\psi_l}\log p(b_i|s_i,a_i,\psi_l)\right|,\end{aligned}$$

which is bounded from $i)$ and N being finite. □

C.4 Proof of Theorem 2

To ease the notational burden of our proof, we drop dependence on the ensemble index $l \in \{1:L\}$ as the convergence proof is the same for all ensembles. We formalise updates (31) and (32) by analysing the recursive sequence:

$$\begin{aligned}\psi_{k+1} &= \mathcal{P}_\Omega\left(\psi_k-\alpha_k\nabla_\psi\left(R(\psi_k-\epsilon)-\log p(b_i|s_i,a_i,\psi_k)\right)\right), \quad \text{(fast)}\\ \omega_{k+1} &= \mathcal{P}_\Omega(\omega_k-\beta_k(\omega_k-\psi_k)). \quad \text{(slow)}\end{aligned} \tag{35}$$

where $i \sim \text{Unif}(\{1:N\})$.

Theorem 2. *If Assumptions 1 to 5 hold, ψ_l and ω_l converge to $\psi_l^{\circledast}(\omega_l^{\circledast})$ and $\omega_l^{\circledast}$ almost surely.*

Proof. Firstly, we define the martingale:

$$\mathcal{M}^i_k \coloneqq -\nabla_\psi\sum_{j\neq i}^{N}\log p(b_j|s_j,a_j,\psi_k).$$

Using the identity:

$$\nabla_\psi\left(R(\psi_k-\epsilon)-\log p(b_i|s_i,a_i,\psi_k)\right)=\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)-\mathcal{M}^i_k \tag{36}$$

we re-write the recursive update in Eq. (35) as a martingale difference sequence:

$$\psi_{k+1}=\mathcal{P}_\Omega\left(\psi_k-\alpha_k(\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)-\mathcal{M}^i_k)\right).$$

As $\mathcal{P}_\Omega$ has a well-defined directional derivative under Proposition 2, we can re-write the fast update using a series expansion about ψ_k:

$$\begin{aligned}&\mathcal{P}_\Omega\left(\psi_k-\alpha_k(\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)-\mathcal{M}^i_k)\right)\\ &\quad=\mathcal{P}_\Omega(\psi_k)-\Gamma_{\psi_k}\left[\alpha_k(\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)-\mathcal{M}^i_k)\right]+o(\alpha_n^2),\\ &\quad=\psi_k+\alpha_k\Gamma_{\psi_k}\left[\mathcal{M}^i_k-\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)\right]+o(\alpha_n^2),\\ &\quad=\psi_k+\alpha_k\left(\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right]-\Gamma_{\psi_k}\left[\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)\right]+o(\alpha_n)\right),\end{aligned}$$

where we have used the identity $\mathcal{P}_\Omega(\psi_k)=\psi_k$ in deriving the second equality and the linearity of the derivative for the final equality. Consider now the slow update, which we can write using a similar expansion:

$$\begin{aligned}\mathcal{P}_\Omega(\omega_k-\beta_k(\omega_k-\psi_k)) &= \mathcal{P}_\Omega(\omega_k)-\Gamma_{\psi_k}\left[\beta_k(\omega_k-\psi_k)\right]+o(\beta_k^2),\\ &= \omega_k+\beta_k\left(\Gamma_{\psi_k}\left[-(\omega_k-\psi_k)\right]+o(\beta_k)\right).\end{aligned}$$

We write the two updates together here for clarity:

$$\psi_{n+1}=\psi_k+\alpha_k\left(\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right]-\Gamma_{\psi_k}\left[\nabla_\psi\mathcal{L}(\psi_k;\mathcal{D}^N_{\omega_k},\epsilon)\right]+o(\alpha_n)\right), \tag{37}$$

$$\omega_{n+1}=\omega_k+\beta_k\left(\Gamma_{\psi_k}\left[-(\omega_k-\psi_k)\right]+o(\beta_k)\right). \tag{38}$$

Our proof adapts Theorem 1 of Heusel et al. [42], which is a generalisation of Borkar [14] using the proofs of Karmakar and Bhatnagar [45]. There is one subtle difference between the updates of Heusel et al. [42] and the updates for BBO in Eqs. (37) and (38): in Eqs. (37) and (38), there is an

additional term $o(\alpha_k)$ and $o(\beta_k)$ such that by Assumption 5, $\lim_{k\to\infty} o(\alpha_k) = o(\beta_k) = 0$. $o(\beta_k)$ will be absorbed in the error term ϵ_n in the slow update in Eq. 14 and $o(\alpha_k)$ will be absorbed in the error term ϵ'_n in the fast update in Eq. 17 of Karmakar and Bhatnagar [45]. As $\lim_{n\to\infty} \epsilon_n = \epsilon'_n = 0$, these additional terms $o(\alpha_n)$ and $o(\beta_n)$ leave the proof of Theorem 5 of Karmakar and Bhatnagar [45] unaffected. Theorem 1 of Heusel et al. [42] therefore remains unchanged, which uses Theorem 5 of Karmakar and Bhatnagar [45] to establish that the sequences of Eqs. (37) and (38) converge almost surely to a set of local attractors of the underlying ODE:

$$\{\omega_k, \psi_k\} \xrightarrow{k\to\infty} \{\omega^{\circledast}, \psi^{\circledast}(\omega^{\circledast})\} \; a.s.$$

if

1. $\Gamma_{\psi_k}\left[-\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon)\right]$ is Lipschitz in ψ_k and $\Gamma_{\omega_k}\left[-(\omega_k - \psi_k)\right]$ is Lipschitz in ω_k under Assumption 5,
2. Stepsizes α_k and β_k satisfy Assumption 5
3. For all k, $(\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right])$ is a martingale difference sequence with respect to the filtration of increasing σ-algebras:

   $$\mathscr{F}_k \coloneqq \sigma\left(\psi_j, \Gamma_{\psi_j}\left[\mathcal{M}^i_j\right] : j = 1, ...k\right),$$

   with

   $$\mathbb{E}\left[\left\|\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right]\right\|^2 |\mathscr{F}_k\right] < \infty,$$

4. The limiting ODEs $\partial_t \psi(t) = -\Gamma_{\psi(t)}(\nabla_\psi \mathcal{L}(\psi(t); \mathcal{D}^N_\omega, \epsilon))$ and $\partial_t \omega(t) = -\Gamma_{\omega(t)}(\omega_l(t) - \psi^{\circledast}_l(\omega_l(t)))$ satisfy Assumption 5,
5. $\sup_n \|\psi_n\| < \infty$ and $\sup_n \|\omega_n\| < \infty$ a.s.

As 5) is satisfied by virtue of the projection operators, we are left to prove 3).

As $\Gamma_{\psi_k}[\cdot]$ is a linear operator with $\Gamma_{\psi_k}[0] = 0$, $(\Gamma_{\psi_k}\left[\mathcal{M}^i_k\right])$ is a Martingale difference sequence if $(\mathcal{M}^i_k)$ is a Martingale difference sequence. To prove 3) holds, it therefore suffices to show that $(\mathcal{M}^i_k)$ is a Martingale difference sequences with respect to the filtration of increasing σ-algebras:

$$\mathscr{F}'_k \coloneqq \sigma\left(\psi_j, \mathcal{M}^i_j : j = 1, ...k\right),$$

To prove this, we observe that: i) $\mathcal{M}^i_k$ is $\mathscr{F}'_k$-measurable for all k by construction, from Proposition 3 $\mathcal{M}^i_k$ is bounded, hence ii) $\mathbb{E}\left[\mathcal{M}^i_k\right] < \infty$ and from Eq. (36) iii) the conditional expectation of $\mathcal{M}^i_k$ satisfies for all k:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{M}^i_k | \mathscr{F}'_k\right] &= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \mathbb{E}\left[\nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right) \middle| \mathscr{F}'_k\right], \\
&= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \nabla_\psi R(\psi_k - \epsilon) + \mathbb{E}\left[\nabla_\psi \log p(b_i|s_i, a_i, \psi_k) | \mathscr{F}'_k\right], \\
&= \nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \nabla_\psi R(\psi_k - \epsilon) + \sum_{i=1}^{N} \nabla_\psi \log p(b_i|s_i, a_i, \psi_k), \\
&= 0.
\end{aligned}$$

Together i) to iii) form the definition of a martingale difference sequence with respect to the filtration $(\mathscr{F}'_k)$ [79]. As we sample from a discrete uniform distribution and from Proposition 3 $\nabla_{\psi_l} R(\psi_l - \epsilon_l)$, $\nabla_{\psi_l} \log p(b_i|s_i, a_i, \psi_k)$ and $\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon)$ are bounded, the variance of the updates is finite:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{M}^i_k\|^2 | \mathscr{F}'_k\right] \\
&\quad = \mathbb{E}\left[\left(\nabla_\psi \mathcal{L}(\psi_k; \mathcal{D}^N_{\omega_k}, \epsilon) - \nabla_\psi \left(R(\psi_k - \epsilon) - \log p(b_i|s_i, a_i, \psi_k)\right)\right)^2 \middle| \mathscr{F}'_k\right] < \infty,
\end{aligned}$$

and hence 3) is satisfied as required. □

C.5 A Frequentist Analysis of Theorem 2

In the frequentist regime, we are concerned with the objective in the limit $N \to \infty$ and we consider data to be arriving online. Our convergence analysis holds if we replace the assumption of uniform sampling from the dataset $\mathcal{D}_\omega^N$ with (assuming samples have finite variance) sampling i.i.d. from the underlying data distribution $(b_i, s_i, a_i) \sim P_B$. The condition of i.i.d. unbiased estimates leaves our analysis unchanged, however our proof can be extended to cases where the updates are sampled from a Markov chain. In particular, if assumptions (A1), (A5) and (A6)' of Karmakar and Bhatnagar [45] are satisfied, Theorem 2 can be trivially extended to the non-i.i.d. case. We do not discuss these assumptions further as they have been extensively discussed (with a case study) in Heusel et al. [42] which sufficiently covers extending BBO to sampling from a Markov chain in the frequentist regime. Note that the ergodicity of the Markov chain in Assumption 1 greatly simplifies their verification as the set of ergodic occupancy measures is a singleton.

D Derivations

To ease the notational burden of our derivations, we abuse notation slightly by writing the density and the distribution of the policy as $\pi(a|s)$ and the density and distribution of the state sampling distribution as $\rho(s)$. The distinction will be clear from context.

D.1 Posterior Density Derivation

We now derive the posterior density in Eq. (4), which is identical for the two sampling regimes outlined in Assumption 1. We start by deriving the likelihood of the data $\mathcal{D}_\omega^N$ for the i.i.d. case:

$$p(\mathcal{D}_\omega^N|\phi) = \prod_{i=1}^{N} \rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi).$$

Using our prior $p(\phi)$ and Bayes' rule, we can infer the posterior as:

$$\begin{aligned}
p(\phi|\mathcal{D}_\omega^N) &= \frac{p(\mathcal{D}_\omega^N|\phi)p(\phi)}{\int_\Phi p(\mathcal{D}_\omega^N|\phi)dP(\phi)}, \\
&= \frac{\prod_{i=1}^N \left(\rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi)\right) p(\phi)}{\int_\phi \prod_{i=1}^N \left(\rho(s_i)\pi(a_i|s_i)p(b_i|s_i, a_i, \phi)\right) dP(\phi)}, \\
&= \frac{\prod_{j=1}^N(\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi)p(\phi)}{\int_\phi \prod_{j=1}^N(\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)}, \\
&= \frac{\prod_{j=1}^N(\rho(s_j)\pi(a_j|s_j)) \prod_{i=1}^N p(b_i|s_i, a_i, \phi)p(\phi)}{\prod_{j=1}^N(\rho(s_j)\pi(a_j|s_j)) \int_\phi \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)}, \\
&= \frac{\prod_{i=1}^N p(b_i|s_i, a_i, \phi)p(\phi)}{\int_\phi \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)}.
\end{aligned} \tag{39}$$

For sampling from an ergodic Markov chain, we denote the initial state-action density as $p_0(s, a)$ and the transition density as $p(s', a'|s, a)$. For sampling on-policy these distributions are defined as:

$$\begin{aligned}
p_0(s, a) &\coloneqq p_0(s)\pi(a|s) \\
p(s', a'|s, a) &\coloneqq p(s'|s, a)\pi(a'|s').
\end{aligned}$$

We write our likelihood as:

$$\begin{aligned}
p(\mathcal{D}_\omega^N|\phi) &= p_0(s_1, a_1)p(b_1|s_1, a_1, \phi) \prod_{i=2}^{N} (p(s_i\, a_i|s_{i-1}, a_{i-1})p(b_i|s_i, a_i, \phi)), \\
&= p_0(s_1, a_1) \prod_{j=2}^{N} p(s_j\, a_j|s_{j-1}, a_{j-1}) \prod_{i=1}^{N} p(b_i|s_i, a_i, \phi), \\
&= p(S_N, A_N) \prod_{i=1}^{N} p(b_i|s_i, a_i, \phi),
\end{aligned}$$

where $S_N \coloneqq \{s_1, ...s_N\}$, $A_N \coloneqq \{a_1, ...a_N\}$ and

$$p(S_N, A_N) = p_0(s_1, a_1) \prod_{j=2}^{N} p(s_j\, a_j | s_{j-1}, a_{j-1}).$$

We now infer the posterior using Bayes' rule:

$$\begin{aligned} p(\phi|\mathcal{D}_\omega^N) &= \frac{p(\mathcal{D}_\omega^N|\phi)p(\phi)}{\int_\Phi p(\mathcal{D}_\omega^N|\phi)dP(\phi)}, \\ &= \frac{p(S_N, A_N)p(b_i|s_i, a_i, \phi)p(\phi)}{p(S_N, A_N)\int_\phi \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)}, \\ &= \frac{\prod_{i=1}^N p(b_i|s_i, a_i, \phi)p(\phi)}{\int_\phi \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)}, \end{aligned}$$

which has the same form as the i.i.d. case in Eq. (39).

D.2 Gaussian BBO Derivation

Now, using a Gaussian model:

$$P(b|s, a, \phi) \coloneqq \mathcal{N}(\hat{B}_\phi(s, a), \sigma^2),$$

and defining the log-normalisation constant as $c_{\text{norm}} \coloneqq \log \int_\phi \prod_{i=1}^N p(b_i|s_i, a_i, \phi)dP(\phi)$, we can derive the exact form of the log-posterior given in Eq. (9):

$$\begin{aligned} p(\phi|\mathcal{D}_\omega^N) &= \exp(-c_{\text{norm}}) \prod_{i=1}^N p(b_i|s_i, a_i, \phi)p(\phi), \\ &= \exp(-c_{\text{norm}}) \prod_{i=1}^N \exp\left(-\frac{1}{2\sigma^2}(b_i - \hat{B}_\phi(s_i, a_i))^2\right) \exp\left(-R(\phi)\right), \\ &= \exp\left(-c_{\text{norm}} - \frac{1}{2\sigma^2}\sum_{i=1}^N (b_i - \hat{B}_\phi(s_i, a_i))^2 - R(\phi)\right), \\ \implies -\log p(\phi|\mathcal{D}_\omega^N) &= c_{\text{norm}} + \sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i, a_i))^2}{2\sigma^2} + R(\phi), \end{aligned}$$

as required. We now derive the minimising objective for the set of KL minimising parameters in Eq. (9) by substituting for the definition of the Gaussian model and ignoring terms independent of ϕ:

$$\begin{aligned} \phi_\omega^\star &\coloneqq \arg\min_{\phi\in\Phi} \text{KL}(P_B(b, s, a|\omega) \parallel P(b, s, a|\phi)) \\ &= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b, s, a|\phi)\right], \\ &= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b|s, a, \phi)\rho(s)\pi(a|s)\right], \\ &= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[-\log p(b|s, a, \phi)\right], \\ &= \arg\min_{\phi\in\Phi} \mathbb{E}_{P_B(b,s,a|\omega)} \left[(b - \hat{B}_\phi(s, a))^2\right]. \end{aligned} \tag{40}$$

Now, we consider the inner expectation with respect to $P_B(b|\cdot)$ which we denote as $\mathbb{E}_B[\cdot]$ for convenience:

$$\begin{aligned} \mathbb{E}_B\left[(b - \hat{B}_\phi)^2\right] &= \mathbb{E}_B\left[b^2\right] - 2\mathbb{E}_B\left[b\right]\hat{B}_\phi + \hat{B}_\phi^2, \\ &= \mathbb{E}_B\left[b^2\right] - 2\mathcal{B}[\hat{Q}_\omega]\hat{B}_\phi + \hat{B}_\phi^2, \end{aligned}$$

Denoting the conditional variance of b as $\mathbb{V}[b]$, we substitute for $\mathbb{E}_B\left[b^2\right] = \mathbb{V}[b] + \mathcal{B}[\hat{Q}_\omega]^2$:

$$\begin{aligned}\mathbb{E}_B\left[(b - \hat{B}_\phi)^2\right] &= \mathbb{V}[b] + \mathcal{B}[\hat{Q}_\omega]^2 - 2\mathcal{B}[\hat{Q}_\omega]\hat{B}_\phi + \hat{B}_\phi^2, \\ &= \mathbb{V}[b] + (\mathcal{B}[\hat{Q}_\omega] - \hat{B}_\phi)^2.\end{aligned}$$

As $\mathbb{V}[b]$ has no dependence on ϕ and we are finding $\arg\min_{\phi\in\Phi}$, we can ignore it from our derivation. Taking expectations with respect to ρ and π, scaling by $\frac{1}{2}$ and substituting into Eq. (40) yields our desired result:

$$\phi^\star_\omega = \arg\min_{\phi\in\Phi} \|\mathcal{B}[\hat{Q}_\omega] - \hat{B}_\phi\|^2_{\rho,\pi}.$$

D.3 MSBBE Gradient Derivation

We now derive an analytic form for the derivative of the MSBBE in Eq. (7). Starting from the definition of the MSBBE:

$$\begin{aligned}\nabla_\omega \text{MSBBE}_N(\omega) &= \nabla_\omega \left\|\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi}, \\ &= \frac{1}{2}\nabla_\omega \mathbb{E}_{\rho,\pi}\left[(\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N})^2\right], \\ &= \mathbb{E}_{\rho,\pi}\left[(\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N})(\nabla_\omega\hat{Q}_\omega - \nabla_\omega\mathcal{B}^\star_{\omega,N})\right],\end{aligned}$$

Substituting for the definition of the Bayesian Bellman operator from Eq. (5) yields our desired result:

$$\begin{aligned}\nabla_\omega \text{MSBBE}_N(\omega) &= \mathbb{E}_{\rho,\pi}\left[\left(\hat{Q}_\omega - \mathbb{E}_{P(\phi|\mathcal{D}^N_\omega)}\left[\hat{B}_\phi\right]\right)\left(\nabla_\omega\hat{Q}_\omega - \nabla_\omega\mathbb{E}_{P(\phi|\mathcal{D}^N_\omega)}\left[\hat{B}_\phi\right]\right)\right], \\ &= \mathbb{E}_{\rho,\pi}\left[\left(\hat{Q}_\omega - \mathbb{E}_{P(\phi|\mathcal{D}^N_\omega)}\left[\hat{B}_\phi\right]\right)\left(\nabla_\omega\hat{Q}_\omega - \mathbb{E}_{P(\phi|\mathcal{D}^N_\omega)}\left[\hat{B}_\phi\nabla_\omega \log p(\phi|\mathcal{D}^N_\omega)\right]\right)\right].\end{aligned}$$

E Linear BBO

We now consider a simple Gaussian linear regression model as a case study where the Bellman operator and Q-function approximators take the forms $\hat{B}_\phi(s,a) = v(s,a)^\top\phi$ and $\hat{Q}_\omega(s,a) = v(s,a)^\top\omega$. $v(s,a)$ is an n-dimensional feature vector and ϕ and ω are n-dimensional parameter vectors. We first derive the MSBBE and its derivative, showing that an analytic solution exists. For the sake of analysis we use a Gaussian conjugate prior over model parameters ϕ:

$$p(\phi|\phi_0, \Sigma_0) = \mathcal{N}(\phi_0, \Sigma_0).$$

For mathematical convenience, we form an $N \times n$ matrix of features $V_N \coloneqq (v_1 \cdots v_N)^\top$ from our data where $v_i \coloneqq v(s_i, a_i)$ and an N-dimensional vector of datapoints $\beta^N_\omega \coloneqq \{b_1, b_2, ...b_N\}$. Our Bayesian Gaussian linear regression model has been well studied (see Murphy [55]) and the posterior can be shown to be a Gaussian:

$$p(\phi|\mathcal{D}^N_\omega) = \mathcal{N}(\phi^N_\omega, \Sigma_N),$$

where

$$\Sigma_N = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}V_N^\top V_N\right)^{-1}, \tag{41}$$

$$\phi^N_\omega = \Sigma_N\left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}V_N^\top\beta^N_\omega\right). \tag{42}$$

We can also derive the posterior predictive:

$$p(b|s,a,\mathcal{D}^N_\omega) = \mathcal{N}(v(s,a)^\top\phi^N_\omega, \sigma^2 + v(s,a)^\top\Sigma_N v(s,a)).$$

By definition, the predictive mean is the Bayesian Bellman operator $\mathcal{B}^\star_{\omega,N}(s,a) = v(s,a)^\top\phi^N_\omega$ from which we derive the MSBBE for linear BBO:

$$\begin{aligned}\text{MSBBE}_N(\omega) &\coloneqq \left\|\hat{Q}_\omega - \mathcal{B}^\star_{\omega,N}\right\|^2_{\rho,\pi}, \\ &= \left\|v(s,a)^\top\omega - v(s,a)^\top\phi^N_\omega\right\|^2_{\rho,\pi}, \\ &= \left\|v(s,a)^\top(\omega - \phi^N_\omega)\right\|^2_{\rho,\pi}, \\ &= \frac{1}{2}(\omega - \phi^N_\omega)^\top\mathbb{E}_{\rho,\pi}\left[v(s,a)v(s,a)^\top\right](\omega - \phi^N_\omega).\end{aligned} \tag{43}$$

The variance of the posterior predictive is composed of two terms: the first term, σ^2, is the aleatoric uncertainty of the data due to observation noise; the second term, $v(s,a)^\top \Sigma_N v(s,a)$, characterises the epistemic uncertainty, and grows whenever the test point $v(s,a)$ is far from observed data. It is this epistemic uncertainty we are concerned with for exploration as it provides a measure of how well explored regions of the state-action space are. Alongside the predictive mean, an agent can gauge how worthwhile it is to explore a region of high epistemic uncertainty depending upon its potential for high returns. Areas of low epistemic uncertainty will be 'crossed-off' by the agent and are likely not to be explored again, only exploited if their predictive returns are high enough.

We now take derivatives of Eq. (43) directly to derive the MSBBE gradient required by our algorithms:

$$\begin{aligned}\nabla_\omega \text{MSBBE}_N(\omega) &= (\nabla_\omega \omega - \nabla_\omega \phi_\omega^N)^\top \mathbb{E}_{\rho,\pi}\left[v(s,a)v(s,a)^\top\right](\omega - \phi_\omega^N), \\ &= (I - \nabla_\omega (\phi_\omega^N)^\top)\mathbb{E}_{\rho,\pi}\left[v(s,a)v(s,a)^\top\right](\omega - \phi_\omega^N), \end{aligned} \tag{44}$$

To proceed, we must find a expression for the derivative of $(\phi_\omega^N)^\top$:

$$\begin{aligned}\nabla_\omega (\phi_\omega^N)^\top &= \nabla_\omega (\phi_\omega^N)^\top, \\ &= \nabla_\omega \left(\Sigma_N \left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2} V_N^\top \beta_\omega^N\right)\right)^\top, \\ &= \nabla_\omega \frac{1}{\sigma^2}(\beta_\omega^N)^\top V_N \Sigma_N^\top.\end{aligned}$$

Now, for each b_i forming the vector β_ω^N, we have $\nabla_\omega b_i = \nabla_\omega (r + \gamma \omega^\top v_i') = \gamma v_i'$. The derivative of the matrix-vector product $(\beta_\omega^N)^\top V_N$ can thus be found as:

$$\nabla_\omega (\beta_\omega^N)^\top V_N = \gamma \sum_{i=1}^N v_i' v_i^\top.$$

We substitute $\nabla_\omega (\beta_\omega^N)^\top V_N = \gamma \sum_{i=1}^N v_i' v_i^\top$ into Eq. (44) to obtain the gradient:

$$\nabla_\omega \text{MSBBE}_N(\omega) = \left(I - \frac{\gamma}{\sigma^2}\sum_{i=1}^N v_i' v_i^\top\right)\Sigma_N^\top \mathbb{E}_{\rho,\pi}\left[v(s,a)v(s,a)^\top\right](\omega - \phi_\omega^N). \tag{45}$$

E.1 Deriving the LSTD algorithm

Although we have derived the exact gradient of the MSBBE, a simple approach for inferring the posterior predictive mean is to solve the equation $\omega = \phi_\omega^N$: indeed, as a sanity check, we see from Eq. (45) that any $\omega^\star$ such that $\omega^\star = \phi_{\omega^\star}^N$ trivially results in an MSBBE gradient of 0 and parametrises a global minimiser. Expanding our equation using the definition of ϕ_ω^N from Eq. (42), we obtain:

$$\begin{aligned}\omega^\star &= \phi_{\omega^\star}^N, \\ &= \Sigma_N \left(\Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2} V_N^\top \beta_{\omega^\star}^N\right), \\ &= \Sigma_N \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\Sigma_N V_N^\top \beta_{\omega^\star}^N, \\ &= \Sigma_N \Sigma_0^{-1}\phi_0 + \frac{1}{\sigma^2}\Sigma_N \sum_{i=1}^N v_i b_i. \end{aligned} \tag{46}$$

Introducing the shorthand $v_i' \coloneqq v(s_i', a_i')$, we substitute for the definition of the empirical Bellman equation from Eq. (2) using the linear function approximator, $b_i = r_i + \gamma {v_i'}^\top \omega$. This allows us to

factorise Eq. (46):

$$\omega^\star = \Sigma_N \Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \Sigma_N \sum_{i=1}^{N} v_i (r_i + \gamma v_i'^\top \omega^\star),$$

$$\Sigma_N^{-1} \omega^\star = \Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i r_i + \frac{\gamma}{\sigma^2} \left(\sum_{i=1}^{N} v_i v_i'^\top \right) \omega^\star,$$

$$\left(\Sigma_N^{-1} - \frac{\gamma}{\sigma^2} \sum_{i=1}^{N} v_i v_i'^\top \right) \omega^\star = \Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i r_i,$$

Substituting for the definition of the predictive covariance from Eq. (41):

$$\Sigma_N^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} V_N^\top V_N = \Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i v_i^\top,$$

yields:

$$\left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i \left(v_i - \gamma v_i' \right)^\top \right) \omega^\star = \Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i r_i,$$

Define the matrix:

$$D_N \coloneqq \Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i \left(v_i - \gamma v_i' \right)^\top,$$

and the vector:

$$\chi_N \coloneqq \Sigma_0^{-1} \phi_0 + \frac{1}{\sigma^2} \sum_{i=1}^{N} v_i r_i$$

with which we obtain a solution for $\omega^\star$:

$$\omega^\star = D_N^{-1} \chi_N$$

If the problem is tractable enough to store D_N^{-1}, then the exact solution $\omega^\star$ can be obtained using Sherman-Morrison updates as outlined in Algorithm 2. We use an importance weight $w_i \coloneqq \frac{\pi(a_i|s_i)}{\pi_e(a_i|s_i)}$ for the off-policy case when an exploratory policy π_e that is difference from the evaluation policy π is used to gather data. When the on-policy sampling is used, $w_i = 1$. If the prior takes a form that can be inverted with complexity $\mathcal{O}(n^2)$ or less, the overall complexity of Algorithm 2 is $\mathcal{O}(Nn^2)$.

Algorithm 2 Calculate $\omega^\star$

for $i \in \{1, ...N\}$ **do**
  if $i == 1$ **then**
    $D^{-1} \leftarrow \left(\sigma^2 \Sigma_0^{-1}\right)^{-1}$
    $\chi \leftarrow \sigma^2 \Sigma_0^{-1} \phi_0$
  end if
  $\Delta \leftarrow w_i (v_i - \gamma v_i')$
  $D^{-1} \leftarrow D^{-1} - \frac{D^{-1} v_i \Delta^\top D^{-1}}{1 + \Delta^\top D^{-1} v_i}$
  $\chi \leftarrow \chi + w_i v_i r_i$
end for
$\omega^\star \leftarrow D^{-1} \chi$
return $\omega^\star$

To obtain a frequentist equivalent of our algorithm, we use an uninformative prior which can be obtained by choosing $\phi_0 = 0$ and $\Sigma_0 = \frac{1}{\sigma_0^2} I$ with $\sigma_0^2 \to 0$. To obtain an algorithm with complexity $\mathcal{O}(Nn^2)$ that uses this prior, we first choose $(\sigma^2 \Sigma_0)^{-1} = \epsilon I$ for some $\epsilon >> 1$. For each datapoint $i \leq n$, we can then add the outer product $\epsilon 1_i 1_i^\top$ using the Sherman-Morrison formula, where 1_i is a vector of zeros except for the ith entry which is a 1. After n datapoints, we have then 're-moved' the original prior, obtaining exactly $\omega^\star$ inferred using an uniformative prior. We outline this procedure in Algorithm 3 where $D_i^{-1} \coloneqq D^{-1} 1_i 1_i^\top D^{-1}$ is the outer product of the ith column and row of D^{-1} and $[D^{-1}]_{i,i}$ is the ith diagonal element of D^{-1}.

Algorithm 3 Calculate $\omega^\star$ with Frequentist prior

for $i \in \{1, ...N\}$ **do**
  if i==0 **then**
    $D^{-1} \leftarrow \epsilon I$
    $\chi \leftarrow 0$
  end if
  $\Delta \leftarrow w_i (v_i - \gamma v_i')$
  $D^{-1} \leftarrow D^{-1} - \frac{D^{-1} v_i \Delta^\top D^{-1}}{1 + \Delta^\top D^{-1} v_i}$
  $\chi \leftarrow \chi + w_i v_i r_i$
  if $i \leq n$ **then**
    $D^{-1} \leftarrow D^{-1} \frac{\epsilon D_i^{-1}}{1 + \epsilon [D^{-1}]_{i,i}}$
  end if
end for
$\omega^\star \leftarrow -D^{-1} \chi$
return $\omega^\star$

E.2 Recovering TDC/GTD2

To derive TDC/GTD2 from our Gaussian linear regression model, we observe that the Bayesian Bellman operator can be interpreted as an N-sample Monte-Carlo estimate of the projection operator with additional bias due to the prior:

$$\begin{aligned}\mathcal{B}^{\star}_{\omega,N} &= v^{\top}\left(\sigma^2\Sigma_0^{-1} + V_N^{\top}V_N\right)^{-1}\left(\sigma^2\Sigma_0^{-1}\phi_0 + V_N^{\top}\beta_{\omega}^{N}\right), \\ &= v^{\top}\left(\frac{1}{N}\left(\sigma^2\Sigma_0^{-1} + \sum_{i=1}^{N} v_i v_i^{\top}\right)\right)^{-1}\frac{1}{N}\left(\sigma^2\Sigma_0^{-1}\phi_0 + \sum_{i=1}^{N} v_i b_i\right). \end{aligned} \tag{47}$$

As the TDC/GTD2 is a frequentist algorithm, we must derive the Bayesian Bellman operator in the limit $N \to \infty$. Taking the limit $N \to \infty$ of Eq. (47) using the strong law of large numbers, we see that the effect of the prior will diminish (observe that it does not scale with increasing N) and the Bayesian Bellman operator converges to the projection operator:

$$\begin{aligned}\lim_{N\to\infty}\mathcal{B}^{\star}_{\omega,N} &= \lim_{N\to\infty} v^{\top}\left(\frac{1}{N}\left(\sigma^2\Sigma_0^{-1} + \sum_{i=1}^{N} v_i v_i^{\top}\right)\right)^{-1}\frac{1}{N}\left(\sigma^2\Sigma_0^{-1}\phi_0 + \sum_{i=1}^{N} v_i b_i\right), \\ &= \lim_{N\to\infty} v^{\top}\left(\frac{1}{N}\sum_{i=1}^{N} v_i v_i^{\top}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} v_i b_i, \\ &= v^{\top}\mathbb{E}_{\rho,\pi}\left[vv^{\top}\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v\mathcal{B}[\hat{Q}_{\omega}]\right], \\ &= \mathcal{P}_{\hat{B}_{\phi}} \circ \mathcal{B}[\hat{Q}_{\omega}]. \end{aligned}$$

This confirms the consistency results that we established between the mean squared projected Bellman error (MSPBE) and MSBBE in the limit $N \to \infty$ under Theorem 1 in Section 3.2 for Gaussian models. The MSPBE has been well studied for linear function approximators [70, 71, 12] and can be shown to take the form:

$$\text{MSPBE}(\omega) = \mathbb{E}_{\rho,\pi}\left[v^{\top}(\mathcal{B}[\hat{Q}_{\omega}] - \hat{Q}_{\omega})\right]\mathbb{E}_{\rho,\pi}\left[vv^{\top}\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v(\mathcal{B}[\hat{Q}_{\omega}] - \hat{Q}_{\omega})\right]. \tag{48}$$

Taking gradients to minimise the MSPBE via stochastic gradient descent leads to the TDC/GTD2 algorithms, which are derived in full in Sutton et al. [70, 71] from the above objective. To avoid the costly matrix inversion in Eq. (48), a set of weights to approximate the term $\zeta \approx \mathbb{E}_{\rho,\pi}\left[vv^{\top}\right]^{-1}\mathbb{E}_{\rho,\pi}\left[v(\mathcal{B}[\hat{Q}_{\omega}] - \hat{Q}_{\omega})\right]$ is learnt, which are updated on a slower timescale:

$$\zeta_{k+1} \leftarrow \zeta_k + \alpha_k^{\zeta}(r + \gamma\hat{Q}'_{\omega_k} - \hat{Q}_{\omega_k} - v_k^{\top}\zeta_k)v_k.$$

This approximation is then used to update the function approximator. When a TD estimate is used, this results in two algorithms :

$$\begin{aligned}\omega_{k+1} &\leftarrow \omega_k + \alpha_k^{\omega}\left(v_k - \gamma v'_k\right)v_k^{\top}\zeta_k \quad \text{(GTD2)}, \\ \omega_{k+1} &\leftarrow \omega_k + \alpha_k^{\omega}v_k(r + \gamma\hat{Q}'_{\omega_k} - \hat{Q}_{\omega_k}) - \alpha_k^{\omega}\gamma v'_k v_k^{\top}\zeta_k \quad \text{(TDC)}.\end{aligned}$$

As the TDC/GTD2 algorithm is derived from a frequentist perspective, it does not characterise uncertainty in the MDP. Conversely, our framework allows us to estimate the predictive variance due to state visitation after N updates of GDT2/TDC, which is essential for achieving deep exploration.

F Randomised Priors for BBO

Randomised priors is a method for obtaining an approximation of an intractable posterior by 'randomising' the maximum a posteriori (MAP) point estimate. Here, a noise variable $\epsilon \in \mathcal{E}$ with distribution $P_E(\epsilon)$ where the density $p_E(\epsilon)$ has the same form as the prior is introduced, which defines a continuum of ϵ-randomised MAP estimates:

$$\psi^{\star}(\epsilon; \mathcal{D}_{\omega}^{N}) = \arg\max_{\phi\in\Phi}\mathcal{L}(\phi; \mathcal{D}_{\omega}^{N}, \epsilon), \quad \mathcal{L}(\phi; \mathcal{D}_{\omega}^{N}, \epsilon) \coloneqq \frac{1}{N}\left(R(\phi - \epsilon) - \sum_{i=1}^{N}\log p(b_i|s_i, a_i, \phi)\right).$$

For simplicity, it is implicitly assumed when using RP that the prior-randomised MAP estimate $\psi^\star(\epsilon; \mathcal{D}_\omega^N)$ is P_E-integrable and that each $\arg\max_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$ is a singleton. We also assumed in Assumption 4 that $R(\phi - \epsilon)$ is well-defined for any $\epsilon \in \mathcal{E}$ and $\phi \in \Phi$. The RP approximate posterior can then be constructed by averaging over all MAP estimates using P_E:

$$q(\phi|\mathcal{D}_\omega^N) = \int_{\mathcal{E}} \delta(\phi = \psi^*(\epsilon; \mathcal{D}_\omega^N)) dP_E(\epsilon).$$

The RP approximate posterior is motivated by the fact that sampling from $q(\phi|\mathcal{D}_\omega^N)$ is equivalent to sampling from the true posterior when a linear Gaussian model is used [57, 58]. When a nonlinear model is used, we still expect the RP posterior to provide a good approximation to the true posterior. Pearce et al. [59] confirm that this holds where nonlinear neural networks are used as function approximations.

A naïve way to sample from the approximate posterior is to first sample $\epsilon \sim P_E$ and solve the optimisation problem $\psi^\star(\epsilon; \mathcal{D}_\omega^N) = \arg\max_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_\omega^N, \epsilon)$. This approach is not tractable as solving the optimisation problem is NP-hard, so instead an ensembling approach is used. As outlined in Section 4, L prior randomisations $\mathcal{E}_L \coloneqq \{\epsilon_l\}_{l=1:L}$ are drawn from P_E. For each $l \in \{1 : L\}$, a set of solutions to the prior-randomised MAP objective are found:

$$\psi_l^\star(\omega_l) \in \arg\min_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l) \coloneqq \arg\min_{\phi\in\Phi} \frac{1}{N} \left(R(\phi - \epsilon_l) - \sum_{i=1}^N \log p(b_i|s_i, a_i, \phi) \right).$$

When using RP with ensembling for reinforcement learning, we write the Q-function approximator as an ensemble of L parameters $\Omega_L \coloneqq \{\omega_l\}_{l=1:L}$ where $\hat{Q}_\omega = \frac{1}{L}\sum_{l=1}^L \hat{Q}_{\omega_l}$ and treat learning each ω_l as a separate problem for each ensemble, leading to the randomised priors ensembled MSBBE:

$$\text{MSBBE}_{\text{RP}}(\omega_l) \coloneqq \|\hat{Q}_{\omega_l} - \hat{B}_{\psi_l^\star(\omega_l)}\|_{\rho,\pi}^2$$

G BBAC Algorithm

For BBAC, we use the Gaussian BBO model introduced in Section 3.2 with a Gaussian prior: $R(\phi) = \frac{1}{\sigma_0^2}\|\phi - \phi_0\|_2^2$. The log-posterior is thus:

$$-\log p(\phi|\mathcal{D}_\omega^N) = c_{\text{norm}} + \sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i, a_i))^2}{2\sigma^2} + \frac{1}{\sigma_0^2}\|\phi - \phi_0\|_2^2$$

We choose a prior parameterisation of $\phi_0 = 0$, which implies a corresponding Gaussian noise distribution: $P_E = \mathcal{N}(0, \sigma_0^2 I)$ [57]. The RP (critic) objective can be derived from the log-posterior as:

$$\mathcal{L}(\psi_l; \mathcal{D}_{\omega_l}^N, \epsilon_l) = \sum_{i=1}^N \frac{(b_i - \hat{B}_\phi(s_i, a_i))^2}{2\sigma^2} + \frac{1}{\sigma_0^2}\|\psi_l - \epsilon_l\|_2^2,$$

where we have ignored $\frac{1}{N}$ as N is finite and so does not contribute to the objective's solution. The two-timescale updates in (13) and (14) are thus:

$$\psi_l \leftarrow \mathcal{P}_\Omega \left(\psi_l - \alpha_k \nabla_{\psi_l} \hat{\mathcal{L}}_{\text{BBAC}}^i(\psi_l) \right), \quad \text{(critic)}$$
$$\omega_l \leftarrow \mathcal{P}_\Omega(\omega_l - \beta_k(\omega_l - \psi_l)), \quad \text{(target critic)}$$

where:

$$\hat{\mathcal{L}}_{\text{BBAC}}^i(\psi_l) \coloneqq \frac{1}{2\sigma^2} \left(r_i + \gamma \hat{Q}_{\omega_l}(s_i', a_i') - \hat{B}_{\psi_l}(s_i, a_i) \right)^2 + \frac{1}{\sigma_0^2}\|\psi_l - \epsilon_l\|_2^2.$$

For our ensembled actors, we choose a Gaussian policy: $\pi_{\theta_l} = \mathcal{N}(\mu_{\theta_l}, \Sigma_{\theta_l})$. Using the reparametrisation trick for actor-critic [47, 41, 30], we can derive low variance policy gradient updates by introducing a variable $\upsilon \sim P_\upsilon(\cdot)$ where $P_\upsilon = \mathcal{N}(0, I)$. Defining the transformation of variables $t_{\theta_l}^{-1}(\upsilon, s) = \Sigma_{\theta_l}^{\frac{1}{2}}(s)(\mu_{\theta_l}(s) - \upsilon)$, we can write our actor objective as an expectation under P_υ:

$$\mathbb{E}_{\rho(s)\pi_{\theta_l}(a|s)}[B_{\phi_l}(s, a)] = \mathbb{E}_{\rho(s)P_\upsilon(\upsilon)}[B_{\phi_l}(s, a = t_{\theta_l}^{-1}(\upsilon, s))]. \tag{49}$$

Algorithm 4 UPDATEPOSTERIOR$(\Theta_L, \Omega_L, \Psi_L, \mathcal{E}_L, \mathcal{H})$

$k \leftarrow 0$
while not converged **do**
  Sample mini-batch of transitions $T \sim \mathcal{D}$
  for $l \in \{1, ...L\}$ **do**
    for $\{s_i, a_i, r_i, s'_i\} \in T$ **do**
      $a'_i \sim \pi_{\theta_l}(\cdot|s'_i)$
      $\psi_l \leftarrow \mathcal{P}_\Omega\left(\psi_l - \frac{\alpha_k}{|T|}\nabla_{\psi_l}\hat{\mathcal{L}}^i_{\text{BBAC}}(\psi_l)\right)$
    end for
    $\omega_l \leftarrow \mathcal{P}_\Omega(\omega_l - \beta_k(\omega_l - \psi_l))$
    Sample batch B_l of states from $s \sim d(\cdot)$
    for $s \in B_l$ **do**
      $\upsilon \sim P_\upsilon(\cdot)$
      $\theta_l \leftarrow \theta_l + \frac{\zeta_k}{|B_l|}\nabla_{\theta_l} B_{\phi_l}(s, a = t^{-1}_{\theta_l}(\upsilon, s))$
    end for
  end for
  $k \leftarrow k + 1$
end while

To minimising the objective in Eq. (49), we propose a stochastic gradient descent update:

$$\theta_l \leftarrow \theta_l + \zeta_k \nabla_{\theta_l} B_{\phi_l}(s, a = t^{-1}_{\theta_l}(\upsilon, s)), \quad \text{(actor)}$$

where ζ_k is the actor learning rate. The pseudocode for updating the posterior using these objectives is shown in Algorithm 4 using batch updates.

Proving full convergence of our actor-critic is beyond the scope of this paper, however we leverage insights from our two-timescale analysis in Section 4.1 and ensure that $\alpha_k > \beta_k > \zeta_k$ to stabilise learning. Finally, we include the pseudocode used to learn our Gaussian behavioural policy in Algorithm 5. Like in soft actor-critic, the algorithm randomly samples two critics from the ensemble and updates the actor using an entropy regularised objective with the minimum of the samples.

Algorithm 5 UPDATEBEHAVIOURALPOLICY$(\theta^\dagger, \Psi_L, \mathcal{H})$

$k \leftarrow 0$
while not converged **do**
  Sample mini-batch of states $B \sim \mathcal{H}$
  Sample $\psi_1, \psi_2 \sim \text{Unif}(\Psi_L)$
  for $s \in B$: **do**
    $\upsilon \sim p_\upsilon(\cdot)$
    $\theta^\dagger \leftarrow \theta^\dagger + \frac{\alpha_k}{|B|}\nabla_{\theta^\dagger}\hat{J}(\theta^\dagger; s, \upsilon, \psi_1, \psi_2)$
  end for
  $k \leftarrow k + 1$
end while

where

$$\hat{J}(\theta^\dagger; s, \upsilon, \psi_1, \psi_2) \coloneqq \min_{i \in \{1,2\}} \left(\hat{B}_{\psi_i}(s, a = t^{-1}_{\theta^\dagger}(\upsilon, s))\right) - \alpha \log \pi_{\theta^\dagger}(a = t^{-1}_{\theta^\dagger}(\upsilon, s)|s).$$

H Policy Evaluation Experiments

In this section, we empirically study the properties of BBO in the policy evaluation regime. We start by presenting linear BBO, showing that it performs on-par with well-studied existing linear policy evaluation algorithms. We then consider nonlinear regime, first in a simplistic counter example, which shows the convergence of nonlinear BBO in an easily-understandable practical setting, and further in more complex domains with neural network function approximators, which demonstrate convergence and consistency of BBO.

H.1 Linear Policy Evaluation

We first evaluate the linear BBO from Algorithm 2 in a suite of 9 policy evaluation tasks, comparing it with 6 other methods: TD, TDC, GTD2, BRM, and two LSTD variants, which we refer to as LSTD and LSTD+. LSTD corresponds to vanilla version of the algorithm and thus can be seen as the non-regularized frequentist version of BBO, whereas LSTD+ corresponds to LSTD with additional improvements, such as improved off-policy reweighting, eligibility traces, and regularization, as presented in Dann et al. [23] (Section 3.4). We note that LSTD+ is not directly comparable with BBO, but is rather presented here as a strong baseline. All the environments, evaluated policies, and other training configurations follow exactly those presented by Dann et al. [23] (Section 3.1). We omit three tasks included in Dann et al. [23]: the Baird counterexample due to its solution corresponding to 0 weights and thus being unfairly trivial for least-squares methods like BBO and LSTD to solve, and the on- and off-policy Cart-Pole Swingup tasks due to our lack of access to the software required to run the policies.

H.1.1 Hyperparameters

We use the hyperparameters provided by [23] (Section 3.1.7) for all the baseline algorithms (LSTD, TD, TDC, GTD2, BRM). For BBO, we test the following values for the prior variance: $\{1e{-}1, 3e{-}1, 1e0, 3e0, 1e1\}$. In the results reported, we use $1e{-}1$ for Cart-Pole with impoverished features and $1e1$ for the rest of the tasks.

H.1.2 Results

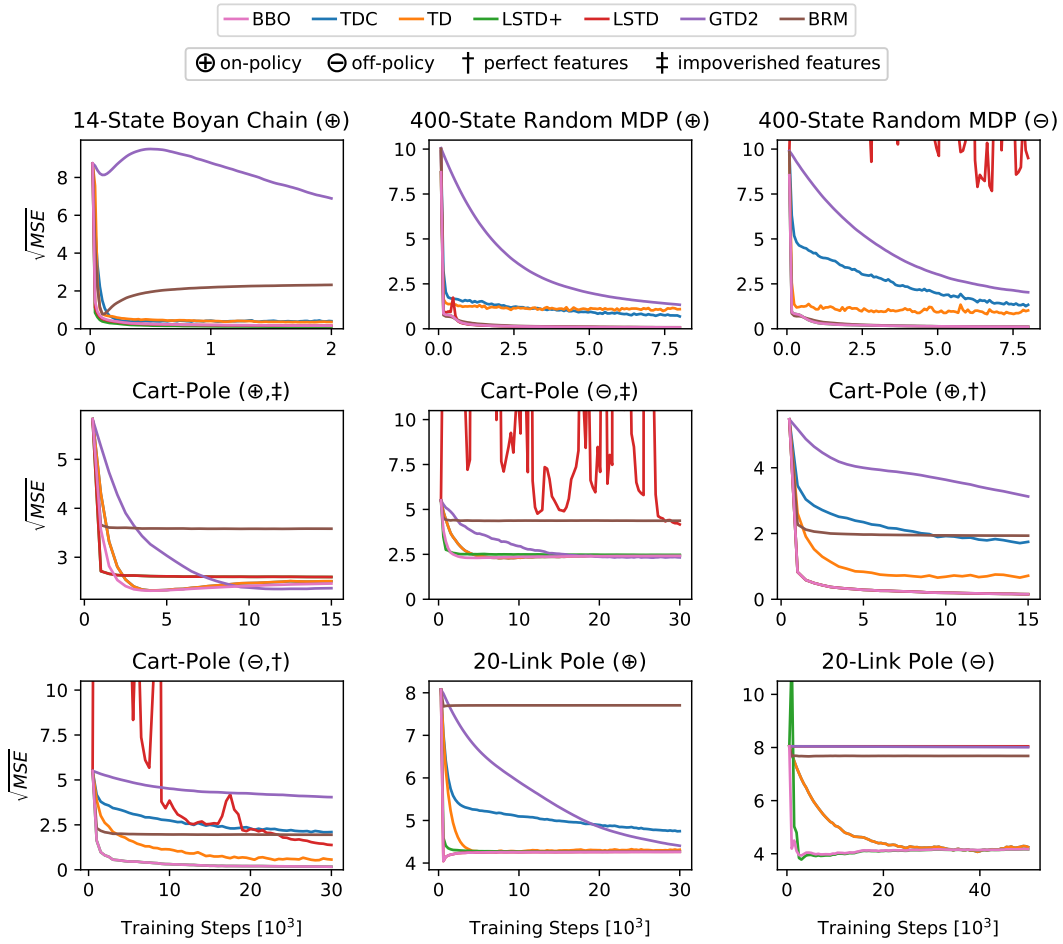


Figure 9: Linear policy evaluation results.

The full training curves are presented in Fig. 9 and summarized in tables Table 1 and Table 2. Table 1 presents the final mean squared error for each algorithm in each task. BBO achieves top final mean squared error in 6 out of 9 tasks, and performs close to the best method in all the rest. In terms of

Table 1: MSE of final predictions. The values for all methods except for BBO are obtained with code provided by [23]. ⊕=on-policy, ⊖=off-policy, †=perfect features, ‡=impoverished features.

	BBO	GTD2	TD	TDC	LSTD+	LSTD	BRM
14-State Boyan Chain (⊕)	0.16	6.89	0.36	0.40	**0.10**	0.16	2.32
400-State Random MDP (⊕)	**0.07**	1.34	1.09	0.69	**0.07**	**0.07**	0.08
400-State Random MDP (⊖)	**0.11**	2.03	1.02	1.33	**0.11**	9.50	**0.11**
Cart-Pole (⊕,‡)	2.46	**2.37**	2.51	2.51	2.60	2.60	3.58
Cart-Pole (⊖,‡)	2.42	**2.33**	2.44	2.44	2.47	4.17	4.37
Cart-Pole (⊕, †)	**0.15**	3.13	0.72	1.75	**0.15**	**0.15**	1.93
Cart-Pole (⊖, †)	**0.17**	4.04	0.57	2.10	**0.17**	1.38	1.95
20-Link Pole (⊕)	**4.26**	4.41	4.31	4.75	4.27	**4.26**	7.71
20-Link Pole (⊖)	**4.17**	8.01	4.25	4.25	**4.17**	8.04	7.68

Table 2: Sum of square root MSE over all timesteps. The values for all methods except for BBO are obtained with code provided by [23]. ⊕=on-policy, ⊖=off-policy, †=perfect features, ‡=impoverished features.

	BBO	GTD2	TD	TDC	LSTD+	LSTD	BRM
14-State Boyan Chain (⊕)	33.46	841.45	61.33	61.55	**25.06**	32.21	214.35
400-State Random MDP (⊕)	**24.74**	342.56	122.51	119.84	**24.74**	27.75	27.87
400-State Random MDP (⊖)	**29.65**	442.96	113.30	266.36	**29.65**	$> 10^3$	32.13
Cart-Pole (⊕,‡)	**76.99**	89.96	79.65	79.61	81.56	81.56	109.88
Cart-Pole (⊖,‡)	**243.86**	291.02	253.96	253.88	253.53	$> 10^3$	438.94
Cart-Pole (⊕, †)	**13.51**	116.98	31.85	68.75	**13.51**	**13.51**	62.78
Cart-Pole (⊖, †)	24.58	267.92	67.91	159.23	**24.57**	681.19	122.12
20-Link Pole (⊕)	428.98	555.56	441.97	508.58	431.56	**428.97**	770.76
20-Link Pole (⊖)	**415.37**	802.78	470.12	470.20	421.74	804.45	768.59

cumulative mean squared error, as shown in Table 2, BBO outperforms other methods in 4 out of 9 tasks, again performing similarly to top-performing methods across all tasks. This demonstrates the benefits of Bayesian methods and incorporating priors in policy evaluation models.

H.2 Tsitsiklis' Triangle Counterexample

We then consider the three-state Tsitsiklis' Triangle MDP [76] designed to prove divergence of TD methods with nonlinear function approximators. The purpose of this experiment is to empirically validate the convergence properties of nonlinear BBO algorithms.

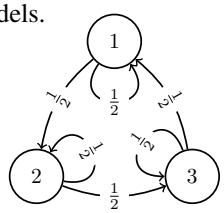


Figure 10: Tsitsiklis' Triangle MDP [76].

H.2.1 Environment and Value Function

The environment, illustrated in Fig. 10, consists of the state space $S = \{1, 2, 3\}$ and the action-independent transition kernel

$$P \coloneqq [p(s' = j|\cdot, s = i)]_{i,j} = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Let $\hat{V}(\omega) \coloneqq [\hat{V}_\omega(s = 1), \hat{V}_\omega(s = 2), \hat{V}_\omega(s = 3)]^\top$ be the value function vector. It can be shown that any $\hat{V}(\omega)$ parametrised by $\omega \in \mathbb{R}$ and satisfying the linear dynamical system

$$\frac{d\hat{V}(\omega)}{d\omega} = (Q + \epsilon I)\,\hat{V}(\omega), \tag{50}$$

with the condition that $\hat{V}(0)^\top 1 = 0$ where $\epsilon > 0$ is a small constant and:

$$Q \coloneqq \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{2} \\ \frac{3}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} & 1 \end{bmatrix},$$

diverges when updated using the TD algorithm [76]. To be consistent with Bhatnagar et al. [12] we choose $\hat{V}(\omega) = \exp(\epsilon\omega)(a\cos(\lambda\omega) - b\sin(\lambda\omega))$ with $a = (-14.9996, -35.0002, 50.0004)$, $b = (-49.0753, 37.5278, 11.5469)$, $\lambda = \sqrt{3}/2$, and $\epsilon = 10^{-2}$ as a solution to Eq. (50) (using the values that we received from the authors' implementation), we use the discount $\gamma = 0.9$ and normalize the gradient steps to stabilize the updates. Each update batch includes all 6 environment transitions.

H.2.2 Hyperparameters

We search over the learning rates for all the algorithms. For TDC, and GTD2, we begin with a coarse grid search with values of $\{1\text{e}{-3}, 1\text{e}{-2}, 1\text{e}{-1}, 1\text{e}0\}$, followed by finer search of $\{1\text{e}{-1}, 2\text{e}{-1}, ..., 9\text{e}{-1}, 1\text{e}0\}$. For TD(0), we manually search around the learning rate used in Bhatnagar et al. [12]. We do not search over BBO's prior loss weight and set to 1.0. The final hyperparameters are presented in Table 3.

Table 3: Hyperparameters for reported Tsitsiklis Triangle experiments.

Method	Parameter	Value
BBO	Lower-level learning rate	8e-1
	Upper-level learning rate	1e-1
TD(0)	Learning rate	2e-3
TDC	Fast timescale learning rate	1e0
	Slow timescale learning rate	1e-1
GTD2	Fast timescale learning rate	8e-1
	Slow timescale learning rate	1e-1

H.2.3 Results

As can be seen in Fig. 3, BBO converges to optimal solution similarly to prior convergent nonlinear methods, TDC and GTD2 [12], while TD(0), as expected, diverges. These results verify the convergence properties of the proposed nonlinear BBO algorithms.

H.3 Neural Network Function Approximators

Most interesting real-world tasks demand use of expressive function approximators such as neural networks. Despite their lack of theoretical convergence guarantees, neural networks have been successfully used in practice for estimating value functions in a wide range of recent reinforcement learning applications. Our proposed gradient BBO algorithms provide provably convergent method for policy evaluation with runtime complexity that is linear in the dimension of the parameter space. In this experiment, we evaluate these properties empirically by applying BBO to a nonlinear regime with neural network function approximators. We use a MAP approximate posterior $p(\phi|\mathcal{D}_\omega^N) \approx \delta(\phi = \phi^*_{\mathcal{D}_\omega^N})$ where:

$$\phi^*_{\mathcal{D}_\omega^N} \in \arg\min_{\phi\in\Phi} \left(\sum_{i=1}^{N} \frac{(b_i - \hat{B}_\phi(s_i, a_i))^2}{2\sigma^2} + \frac{1}{\sigma_0^2}\|\phi - \phi_0\|_2^2 \right),$$

The MAP objective is recommended for policy evaluation as there is no need for the agent to explore, hence learning the posterior uncertainty is inappropriate when a point estimate will suffice. Under a similar derivation as in Appendix F, we can minimise the MSBBE with the MAP posterior estimate using the two timescale updates:

$$\phi \leftarrow \mathcal{P}_\Omega \left(\phi - \alpha_k \nabla_\phi \hat{\mathcal{L}}^i_{\text{MAP}}(\phi) \right), \quad \text{(fast)}$$
$$\omega \leftarrow \mathcal{P}_\Omega(\omega - \beta_k(\omega - \phi)), \quad \text{(slow)}$$

where:

$$\hat{\mathcal{L}}^i_{\text{MAP}}(\phi) \coloneqq \frac{1}{2}\left(r_i + \gamma \hat{Q}_\omega(s'_i, a'_i) - \hat{B}_\phi(s_i, a_i)\right)^2 + \frac{1}{\sigma_0^2}\|\phi - \phi_0\|_2^2.$$

We refer to algorithms that minimise the MSBBE in this way as *gradient* BBO as they take the posterior's dependence of $\hat{Q}_\omega$ into account. We also test a version where we ignore this dependence that we call *direct* BBO, ignoring the slow update and setting $\hat{B}_\phi = \hat{Q}_\omega$.

We investigate the performance between gradient vs. direct and Bayesian vs. frequentist variants of Gaussian BBO and compare them to prior nonlinear TD(0) and TDC algorithms. We set ϕ_0 to our initial estimate of the value function parameters, which is initialised using a Glorot uniform initialisation. Our experiments are designed to (1) verify BBO's convergence and consistency properties in nonlinear regime, especially in cases where not all theoretical assumptions are fulfilled exactly (2) investigate the effect of bi-level optimization (i.e. gradient vs. direct methods), and (3) investigate the effect of additional regularization in BBO due to the prior.

H.3.1 Environments

We consider three environments commonly used in policy evaluation literature: 20-Link Pendulum [23] with 40D continuous observation space, Puddle World [16] with 2D continuous observation space, and the continuous variant of Mountain Car [16][2] with 2D continuous observation space.

H.3.2 Datasets

The datasets used for evaluating the policies consist of 20000, 20000, and 30000 on-policy transitions for Puddle World, Mountain Car, and 20-Link Pendulum, respectively. For Puddle World and Mountain Car, each transition is sampled independently by resetting the state uniformly at random in the state space after each transition. Puddle World and Mountain Car observations are normalized to range $[-1, 1]$.

Policies. For Puddle World and Mountain Car experiments, we run the policy evaluation using a simple, non-optimal policies. In Puddle World, the policy selects either up or down action uniformly at random. In Mountain Car, the right action is chosen when velocity > 0, and otherwise the left action. The 20-Link Pendulum experiments, on the other hand, use an optimal policy obtained with dynamic programming, similar to the policy used in the corresponding linear experiments. See [23] for details.

Ground-truth Value Functions. For Puddle World and Mountain Car, the ground-truth value function, for which the mean squared errors are computed, is obtained for a set of 625 evenly-spaced states (25×25 grid) in the 2D state space. We reset the agent to each of the 625 states 1000 times, rolling it out and computing the cumulative sum of rewards for up to 1000 steps, finally averaging over the 1000 resets. The 20-Link Pendulum environment is a linear-quadratic MDP, for which we obtain the exact values for 5000 states, as is done by Dann et al. [23]. All the value functions are computed with the same discount factor that is used for training the approximate value functions.

H.3.3 Network Structure

All the experiments use a single-layer feedforward network with hidden layer size of 256. TDC uses a $tanh$ activation whereas other algorithms use $relu$.

H.3.4 Training Details

The discount factor is set to 0.98 for all tasks. The optimization is carried out with Adam optimizer [46] using randomly sampled mini-batches of size 512. The hyperparameter searches are done using 3 datasets and the final results are reported are averaged over 24 separate datasets and seeds. Each full trial (100k steps) run takes about 15 minutes for BBO variants, 7 minutes for TD(0), and 85 minutes for TDC on standard desktop machine.

H.3.5 Hyperparameters

We use the same grid search for the hyperparameters across all environments. The hyperparameters with their evaluated values for each algorithm are presented in Table 4 for BBO, Table 5 for TD(0), and Table 6 for TDC. For BBO, instead of searching over the full grid at once, we first perform a

[2]We use the MountainCar-Continuous-v0 implementation from the OpenAI Gym suite [18]

coarse grid search for learning rates without using a prior, then do another denser search around the best values (with values still given in the Table 4), and finally perform a grid search over the lower-level gradient steps per training steps and prior values (when applicable). For TDC and TD(0), we perform a full grid search over the listed values.

Table 4: BBO hyperparameter grid

Hyperparameter	Grid values
Learning rates	$\{$ 1e$-$6, 3e$-$6, 1e$-$5, $\ldots$, 1e$-$1, 3e$-$1, 1e0 $\}$
Weight for the prior loss $(1/\sigma_0^2)^*$	$\{$ 1e$-$4, 3e$-$4, 1e$-$3, $\ldots$, 1e$-$1, 3e$-$1, 1e0 $\}$
Lower-level steps / training step	$\{1, 5, 10, 20\}$

* For 20-Link Pendulum, we also include $\{$ 2.5e$-$1, 5e$-$1, 7.5e$-$1 $\}$.

Table 5: TD(0) hyperparameter grid

Hyperparameter	Grid values
Learning rate*	$\{$ 1e$-$6, 3e$-$6, 1e$-$5, $\ldots$, 1e$-$1, 3e$-$1, 1e0 $\}$

* For 20-Link Pendulum, we also include $\{$ 1e$-$7, 3e$-$7 $\}$.

Table 6: TDC hyperparameter grid

Hyperparameter	Grid values
Fast timescale learning rate	$\{$ 1e$-$6, 3e$-$6, 1e$-$5, $\ldots$, 1e$-$1, 3e$-$1, 1e0 $\}$
Slow timescale learning rate	$\{$ 1e$-$6, 3e$-$6, 1e$-$5, $\ldots$, 1e$-$1, 3e$-$1, 1e0 $\}$

The final hyperparameters for each method and task are presented in Table 7. The values are chosen by manually picking the best-performing set from the hyperparameter search based on the final MSE, and the final results are reported for 24 holdout datasets.

H.3.6 Results

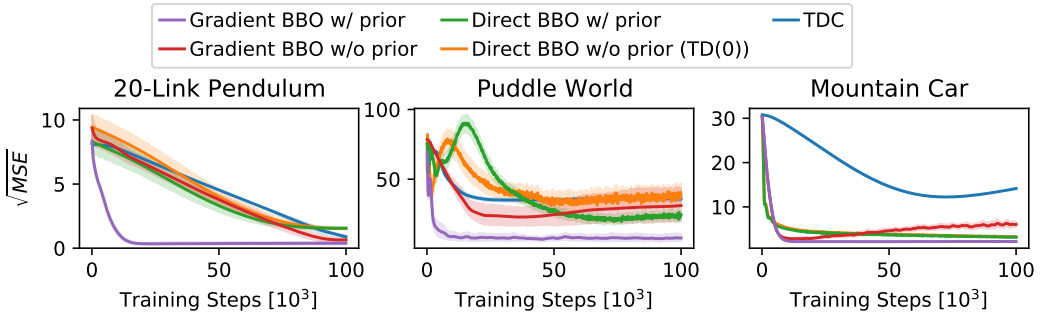


Figure 11: Nonlinear policy evaluation results.

The results are presented in Fig. 11. Gradient BBO with a prior quickly converges to a good solution in all tasks, outperforming the non-Bayesian and direct BBO variants as well as other nonlinear methods (note that direct BBO without prior corresponds to TD(0)) across all tasks. Given that both the bi-level optimization and the prior alone perform worse than Bayesian gradient BBO, the speed and quality of the solution can be attributed to their combination. Furthermore, while not a direct measure, the convergence to near-zero MSE provides empirical evidence of the algorithm's consistency.

Table 7: Hyperparameters for reported nonlinear policy evaluation results.

Method	Parameter	Environment		
		20-Link Pendulum	Puddle World	Mountain Car
Gradient BBO w/ prior	Upper-level learning rate	1e-2	3e-2	1e-2
	Lower-level learning rate	1e-3	1e-2	3e-3
	Prior loss weight	5e-1	3e-4	1e-1
	Lower-level steps / training step	20	10	10
Gradient BBO w/o prior	Upper-level learning rate	3e-4	1e-3	1e-2
	Lower-level learning rate	1e-3	1e-2	3e-3
	Lower-level steps / training step	20	10	10
Direct BBO w/ prior	Learning rate	3e-7	1e-3	3e-4
	Prior loss weight	1.0	3e-4	1e-1
TD(0)	Learning rate	3e-7	3e-3	3e-4
TDC	Fast timescale learning rate	3e-3	1e-3	1e-2
	Slow timescale learning rate	1e-4	1e-3	1e-5

I Continuous Control Experiments

This section extends the continuous control experiments in Section 6 and provides further details and analysis of the BBAC algorithm. We refer to $\lambda \coloneqq \frac{1}{\sigma_0^2}$ as the regularisation weight and σ^2 as the prior scale. We first investigate BBAC's behavior in the *MountainCar-Continuous-v0* task, where the environment's simplicity and low-dimensionality allows us to visually analyze the behavior of the randomized prior ensemble. We then analyze BBAC's sensitivity to randomized prior hyperaparameters in a slightly modified version of DeepMind Control Suite's [73] *cartpole-swingup_sparse* environment.

To improve the exploration capability of SAC in our challenging domains, we also introduce a variant of SAC called SAC* which uses a single Q-function to avoid *pessimistic underexploration* [20]. Here, instead of learning two soft Q-function approximators independently and choosing the minimum for the actor and critic gradient updates as specified by SAC, we train a single soft Q-function and use updates (6) and (13) of Haarnoja et al. [39] directly. This also reduces the inductive bias that SAC has for solving tasks with dense reward structures, making the comparison against BBAC fairer.

All the experiments in this section use the same hyperparameters from Table 8. The policies and function approximators are parameterized as fully-connected neural networks.

Table 8: Common Hyperparameters for BBAC and SAC.

Hyperparameter	Value
Optimizer	Adam
Learning rates	$3\mathrm{e}{-4}$
Discount	0.99
Replay buffer size	$1\mathrm{e}6$
Nonlinearity	ReLU
Hidden layers	2
Hidden units/layer	256
Batch size	256
Target smoothing coefficient	$5\mathrm{e}{-3}$
Training steps per environment step	1

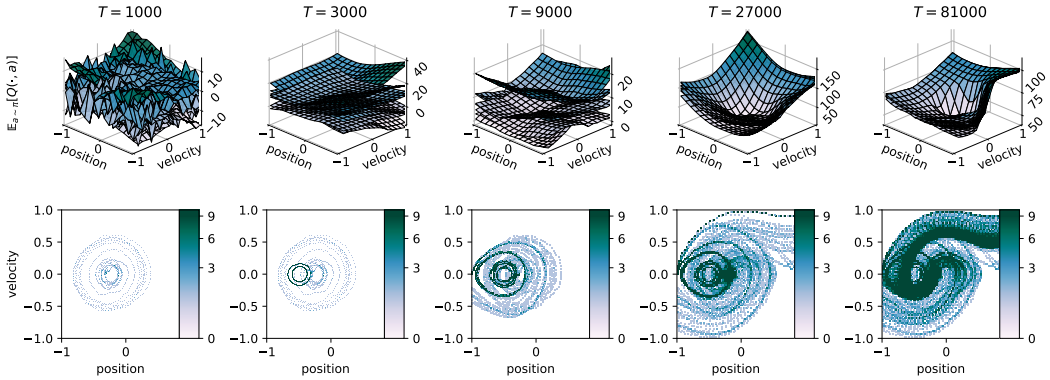


Figure 12: Diagnostics throughout learning of BBAC in MountainCar-Continuous-v0 environment. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. At $T = 9000$, the agent has not yet discovered any reward from the environment, but the disagreement of ensemble members drives the algorithm to deeply explore the state space. At around $T = 14000$ (not shown in the plot), the agent achieves the goal for the first time and the value functions start to shape towards optimal solution. (Bottom) State visitation plots show how the covered states evolve during learning. The deep, adaptive exploration carried out by BBAC leads to the agent systematically exploring regions of the state-action space eventually leading to successful task completion.

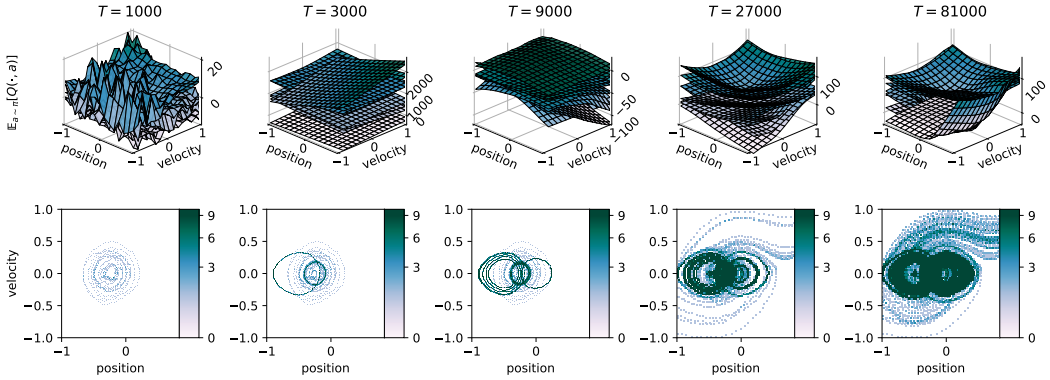


Figure 13: Diagnostics throughout learning of BAC in MountainCar-Continuous-v0 environment. (Top) Expected Q-values for each ensemble member ($L = 8$) over environment states $(position, velocity)$. At $T = 9000$, the agent has not yet discovered any reward from the environment and the disagreement of ensemble members drives the algorithm to explore the state space similarly as with BBAC (see Fig. 12). As corroborated by Fig. 4a, the Q-function learning is relatively unstable without target networks and the agent spends a lot of time re-exploring the states around origin. (Bottom) State visitation plots show how the covered states evolve during learning. The agent is able to reach the goal but, due to unstable learning, spends more time exploring around the origin than BBAC.

I.1 BBAC MountainCar-Continuous-v0

In order to qualitatively measure how the agents explore the state space, we analyze the evolution of critics and coverage of state space throughout learning. We use the standard MountainCar-Continuous-v0 implementation from the OpenAI Gym suite [18], and run each of the algorithms for $1e5$ timesteps, while checkpointing the replay pool and the critics on pre-specified intervals.

For the final SAC runs, we set the target entropy using the heuristic provided in [40] which is the negative number of action dimensions, i.e. -1 in this case. We also tested target entropy values $\{-8, -4, -2, 0, \frac{1}{2}, 1\}$, all of which behaved similarly as the default target. From the 35 runs with these 7 different entropies, a total of 3 seeds were able to achieve the goal. As our experiments are carried out in a tabula rasa setting where the prior functions are drawn as randomly initialized neural

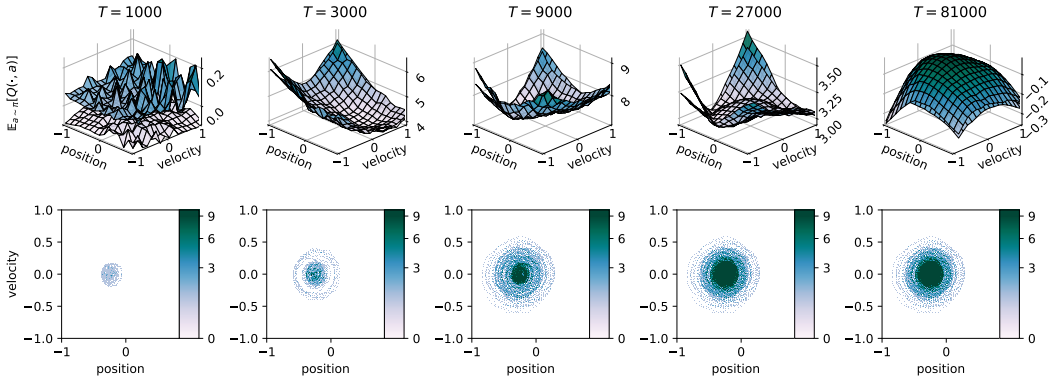
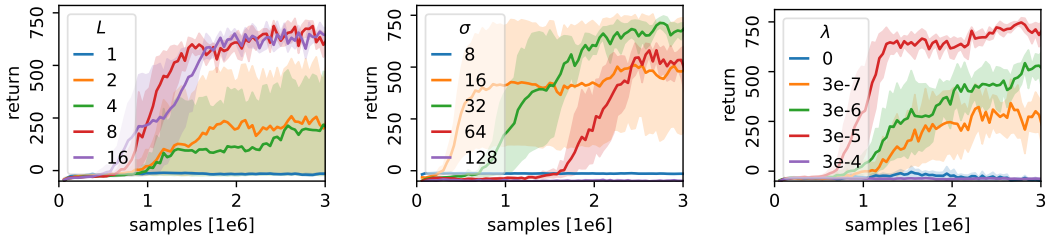


Figure 14: Diagnostics throughout learning of SAC in MountainCar-Continuous-v0 environment. (Top) Ensemble values over environment state $(position, velocity)$. Notice the local maximum in the Q functions around the initial state at $T = 27000$. Similar, but more subtle maximum exists at $T = 9000$. (Bottom) State visitation plots show how the covered states evolve during learning. Due to the naïve exploration, SAC repeatedly explores actions that lead to poor performance and rarely explores beyond the initial state.



(a) Ensemble size L. λ=3e−5, σ=32.

(b) Prior scale σ. L=8, λ=3e−5.

(c) Regularization weight λ. L=8, σ=32.

Figure 15: Evaluation of BBAC's sensitivity to randomized prior hyperparameters. We vary (a) ensemble size L, (b) prior scale σ, and (c) regularization weight λ, while keeping other hyperparameters fixed.

networks, we find that the choice of prior parameters affects the speed with which BBAC solves the tasks. BBAC is relatively insensitive to the randomized prior hyperparameters in this environment, as long as the ensemble size $L \geq 4$, and for the final results we use ensemble size $L = 8$, prior scale $\sigma = 100$, and regularization weight $\lambda = 3\text{e}{-5}$.

The state support analysis, shown in Fig. 14, confirms the inefficiency of the exploration typical of RL-as-inference algorithms like SAC: the agent eschews costly actions that would ultimately lead to rewarding states, thus rarely exploring beyond its initial state. This can also be seen in the value functions, which are prematurely driven to sub-optimal solution.

The same analysis for BBAC, in Fig. 12, shows how the deep, adaptive exploration leads to the agent systematically exploring regions of the state-action space with high uncertainty. The value function plots illustrate how the ensemble uncertainty drives the exploration. In the beginning, even when the actions are costly and no positive rewards are encountered, there still exists ensemble members whose value are optimistic under uncertainty. When at least one such optimistic value function exists for a given state, then that state will eventually be explored by the agent, and the agent will only stop exploring states whose uncertainty is driven down by visiting them.

The analysis for BAC in Fig. 13 confirms that BAC initially explores similarly to BBAC, but due to the convergence issues that stem from ignoring the posterior's dependence on ω, the ensembles never concentrate with increasing number of samples. This means that the approximate posterior cannot characterise the epistemic uncertainty well and the residual stochasticity continues to drive exploration when the agent should have learnt to ignore actions that can't lead to increased returns.

I.2 BBAC cartpole

We further investigate BBAC's sensitivity to the choice of randomized prior hyperparameters – ensemble size (L), prior scale (σ), and regularization weight (λ) – in a slightly modified version of DeepMind Control Suite's [73] *cartpole-swingup_sparse* domain, as a continuous analog to the one presented in [58].

The original *cartpole-swingup_sparse* is modified such that the reward function includes a control cost of $0.1a_t$ for action a_t on each timestep t, and the agent receives a positive reward of 1 only when both the cart is controllably centered and the pole is controllably upright. That is, the reward function $r(s_t, a_t)$ for action a_t at state $s_t = (\cos(\theta_t), \sin(\theta_t), \dot{\theta}_t, x_t, \dot{x}_t)$ is given by:

$$r(s_t, a_t) = -0.1|a_t| + (\mathbb{1}(|x_t| < 0.1) \cdot \mathbb{1}(0.95 < \cos(\theta_t)) \cdot \mathbb{1}(|\dot{x}_t| < 1) \cdot \mathbb{1}(|\dot{\theta}_t| < 1)).$$

The pole is initialized to a stationary downright position, and the motor is not strong enough to turn the pole upright on a single pass, meaning that in order to reach the goal, the agent has to build momentum by moving the cart and swinging the pole back and forth. This requires executing costly actions for more than a hundred steps, making the exploration problem non-trivial.

SAC's performance, shown in Fig. 4b, confirms that naïve exploration strategies, such as noisy actions incorporated by maximum-entropy reinforcement learning, eschew costly exploration actions needed for task completion and converge to sub-optimal strategy. BBAC on the other hand is able to consistently solve the task.

The results for BBAC's hyperparameter sensitivity are presented in Fig. 15. Increasing the ensemble size (L; Fig. 15a) improves the likelihood of solving the task and, as expected, this effect plateaus and the task can be consistently solved with $L \geq 8$. In the case of both prior scale (σ; Fig. 15b) and regularization weight (λ; Fig. 15c), too small values limit the exploration and there is a sweet spot where task performance and exploration are well-balanced. While higher values of prior scale are unable to solve the task within the $3e6$ time steps shown here, we expect them to eventually converge to the optimal solution, whereas higher values of regularization weight are likely to constrain the learning too much to allow convergence.

All our experiments are carried out in a tabula rasa setting where the prior functions are drawn as randomly-initialized neural networks, which is why the effect of the hyperparameter choice to the speed of learning is expected. The range of working hyperparameters is relatively wide and easy to tune, however, and in a real-world scenario, we might have access to prior knowledge on the task, for example through transfer learning, which would further simplify the choice of BBAC's hyperparameters.