Bayesian Bellman Operators

Anonymous Author(s)
Affiliation
Address
email

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 [34, 77]. 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 [21] because a non-Bayesian decision can always be improved upon by a Bayesian agent [23]. 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.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Our framework is general and can recover empirically successful algorithms such as BootDQNprior+ [56]. 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 [38] fail catastrophically due to their inability to properly explore.

2 Bayesian Reinforcement Learning

2.1 Preliminaries

Formally, an RL problem is modelled as a Markov decision process (MDP) defined by the tuple $\langle \mathcal{S}, \mathcal{A}, r, P, P_0, \gamma \rangle$ [71, 59], 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 $r_t := r(s'_t, a_t, s_t) \in \mathbb{R}$ 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 defined as $P^\pi(s', a'|s, a) := \pi(a'|s')P(s'|s, a)$. 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: $J^\pi := \mathbb{E}_\pi \left[\sum_{i=0}^\infty \gamma^i r_i\right]$ 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: $Q^\pi(s, a) := \mathbb{E}_{\pi_\theta}[\sum_{i=0}^\infty r_i|s_0 = s, a_0 = a]$. Any Q-function satisfies the Bellman equation $\mathcal{B}[Q^\pi] = Q^\pi$ where the Bellman operator is defined as:

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

2.2 Model-based vs Model-free Bayesian RL

Bayes-adaptive MDPs (BAMDPs) [26] 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 [66]. Futhermore, 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, 37]. One notable exception is VariBAD [81] 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 [77, 34]. 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 [43, 10, 46, 60, 32, 50]. These methods assume access to a dataset of N samples: $\mathcal{D}^N := \{q_i\}_{i=1:N}$ 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 predictive $P(Q^\pi|s, a, \mathcal{D}^N)$ 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 [56].

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 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 a positive answer 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 A.

3 Bayesian Bellman Operators

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

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:

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

When solving the Bellman equation, the function approximator $\hat{Q}_\omega$ is known but we are uncertain of its value under the Bellman operator due to the 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 distribution 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 $P(Q^\pi|s,a)$. BBO corrects this by modelling the true conditional distribution: $P_B(b|s,a,\omega)$ that generates the data.

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.

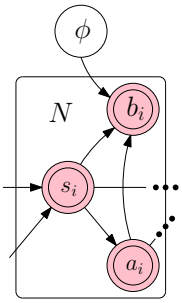


Figure 1: Graphical Model for BBO.

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 density using Bayes' rule (see Appendix C.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 must therefore 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 [69]. We derive the MSBBE gradient in Appendix C.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 [61]. 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 [48] 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:

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

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

Assumption 3 (Single Minimiser)**.** *The set of minimum KL parameters $\phi^\star_\omega$ exists and is a singleton.*

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

If our model can sufficiently represent the true conditional distribution then $\mathrm{KL}(P_B(b,s,a|\omega) \parallel P(b,s,a|\phi^\star_\omega)) = 0 \implies P_B(b|s,a,\omega) = P(b|s,a,\phi^\star_\omega)$. Theorem 1 proves that the posterior concentrates on $\phi^\star_\omega$ and hence the Bayesian Bellman operator converges to the true Bellman operator: $\hat{B}_{\phi^\star_\omega}(s,a) = \mathbb{E}_{P(b|s,a,\phi^\star_\omega)}[b] = \mathbb{E}_{P_B(b|s,a,\omega)}[b] = \mathcal{B}[\hat{Q}_\omega](s,a)$. 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, $B_{\phi^\star_\omega}$ 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 [69, 70, 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 [54, 32]. 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 C.2 for a derivation):

$$-\log p(\phi|\mathcal{D}^N_\omega) = c_{\text{norm}} + \sum_{i=1}^{N} \frac{(b_i - \hat{B}_\phi(s_i,a_i))^2}{2\sigma^2} + R(\phi), \quad \phi^\star_\omega \in \arg\min_{\phi\in\Phi} \|\hat{B}_\phi - \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi}. \quad (8)$$

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 Bellman operator onto the space of functions represented by $\hat{B}_\phi$:

$$\hat{B}_{\phi^\star_\omega} = \{\hat{B}_\phi : \phi \in \arg\min_{\phi\in\Phi} \|\hat{B}_\phi - \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi}\} =: \mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega]. \quad (9)$$

Finally, Theorem 1 proves that the MSBBE converges to the mean squared projected Bellman error $\mathrm{MSBBE}_N(\omega) \xrightarrow{a.s.} \|\hat{Q}_\omega - \mathcal{P}_{\hat{B}_\phi} \circ \mathcal{B}[\hat{Q}_\omega]\|^2_{\rho,\pi} =: \mathrm{MSPBE}(\omega)$. By the definition of the projection operator in Eq. (9), 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 D, allowing us to derive a regularised Bayesian TDC/GTD2 algorithm [70, 69].

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 D. 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: $q(\phi|\mathcal{D}_\omega^N) \approx P(\phi|\mathcal{D}_\omega^N)$. In this paper, we use randomised priors (RP) [56] for approximate inference. Randomised priors (PR) inject noise into the maximum a posteriori (MAP) estimate via a random variable $E : \mathcal{E} \to \mathbb{R}$ 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 E, including derivations of our objectives. RP in practice uses ensembling: L prior randomisations $\mathcal{E}_L := \{\epsilon_l\}_{l=1:L}$ are first drawn from P_E. To use RP for BBO, we write the Q-function approximator as an ensemble of L parameters $\Omega_L := \{\omega_l\}_{l=1:L}$ 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 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:

$$\psi_l^\star(\omega_l) \in \arg\min_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l) := \arg\min_{\phi\in\Phi} \frac{1}{N}\left(R(\phi - \epsilon_l) - \sum_{i=1}^{N} \log p(b_i|s_i, a_i, \phi)\right). \tag{10}$$

The RP solution $\psi_l^\star(\omega_l)$ has dependence on ω_l that mirrors the BBO posterior's dependence on ω. To construct the RP approximate posterior $q(\phi|\mathcal{D}_\omega^N)$, we average the set of perturbed MAP estimates over all ensembles: $q(\phi|\mathcal{D}_\omega^N) := \frac{1}{L}\sum_{l=1}^{L} \delta(\phi \in \psi_l^\star(\omega_l))$. To sample from the RP posterior $\phi \sim q(\cdot|\mathcal{D}_\omega^N)$, we sample an ensemble uniformly $l \sim \text{Unif}(\{1 : L\})$ and set $\phi = \psi_l^\star(\omega_l)$. 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 $P(\phi|\mathcal{D}_\omega^N)$ replaced by the RP approximate posterior $q(\phi|\mathcal{D}_\omega^N)$ 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 [58, 20] with strong empirical performance in RL [56, 57] for the Gaussian model that we study in this paper.

The RP approximate posterior $q(\phi|\mathcal{D}_\omega^N)$ 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 $\omega_l^\star$. 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: $\text{MSBBE}_{\text{RP}}(\omega_l) := \|\hat{Q}_{\omega_l} - \hat{B}_{\psi_l^\star(\omega_l)}\|_{\rho,\pi}^2$. 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 minimising the simpler parameter objective $\omega_l^\star \in \arg\min_{\omega_l\in\Omega} \mathcal{U}(\omega_l; \psi_l^\star)$:

$$\mathcal{U}(\omega_l; \psi_l^\star) := \|\omega_l - \psi_l^\star(\omega_l)\|_2^2 \quad \text{such that} \quad \psi_l^\star(\omega_l) \in \arg\min_{\phi\in\Phi} \mathcal{L}(\phi; \mathcal{D}_{\omega_l}^N, \epsilon_l), \tag{11}$$

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. (11) 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. (11) is NP-hard [8]. To tackle this problem we propose a two-timescale gradient update for each $l \in \{1 : L\}$ on the objectives in Eq. (11) with per-step complexity of $\mathcal{O}(n)$:

$$\psi_l \leftarrow \mathcal{P}_\Omega\left(\psi_l - \alpha_k \nabla_{\psi_l}\left(R(\psi_l - \epsilon_l) - \log p(b_i|s_i, a_i, \psi_l)\right)\right), \quad \text{(fast)} \tag{12}$$

$$\omega_l \leftarrow \mathcal{P}_\Omega(\omega_l - \beta_k(\omega_l - \psi_l)), \quad \text{(slow)} \tag{13}$$

where α_k and β_k are asymptotically faster and slower stepsizes respectively and $\mathcal{P}_\Omega(\cdot) := \arg\min_{\omega\in\Omega}\|\cdot - \omega\|_2^2$ 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. (10), 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 [53]. 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 B.5.

To answer Question 4), we prove convergence of updates (12) and (13) using a straightforward application of two-timescale stochastic approximation [15, 14, 41] to BBO. Intuitively, two timescale analysis proves that the faster timescale update (12) 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 $\psi^\star(\omega_l)$ whenever ω_l is updated in the slower timescale update (13), viewing the parameter ψ_l as quasi-equilibrated [14]. We introduce the standard two-timescale regularity assumptions and derive the limiting ODEs of updates (12) and (13) in Appendix B.3.

Assumption 5 (Two-timescale Regularity)**.** *i)* $\nabla_{\psi_l}\left(R(\psi_l - \epsilon_l) - \log p(b_i|s_i, a_i, \psi_l)\right)$ *is Lipschitz in* ψ_l *and* $(b_i, s_i, a_i) \sim \text{Unif}(\mathcal{D}^N_\omega)$*; ii)* $\psi^\circledast(\omega_l)$ *and* $\omega_l^\circledast$ *are local aysmptotically stable attractors of the limiting ODEs of updates (12) and (13) 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 [56, 57] is a state-of-the-art Bayesian model-free algorithm with Thompson sampling [73] 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 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. [57] 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 F.

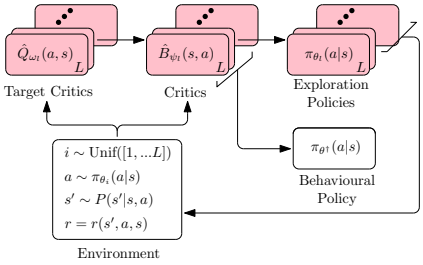


Figure 2: Schematic of RP-BBAC.

Comparison to existing actor-critics: Using a Gaussian model also allows a direct comparison to frequentist actor-critic algorithms [49]: 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. (11): $\mathcal{L}_{\text{critic}}(\psi_l) := -\frac{1}{\sigma^2}\sum_{i=1}^N (b_i - \hat{B}_{\psi_l}(s_i, a_i))^2 + R(\psi_l - \epsilon_l)$. 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 [53, 38]. We introduce an ensemble of parametric exploration policies $\pi_{\theta_l}(a|s)$ parametrised by a set of parameters $\Theta_L := \{\theta_l\}_{l=1:L}$. 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 [76].

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 [33, 56, 31, 51, 57, 74] or Gaussian process regression model [27, 28]. 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 $P(Q^\pi|s, a, \mathcal{D}^N) \approx P(b|s, a, \mathcal{D}^N_\omega)$. 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 [75] 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, 30, 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 [31, 51, 74, 33]. This leads to poor uncertainty estimates as the Bellman error cannot correctly propagate the uncertainty [55, 56]. Osband et al. [57] 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 [75], 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 [22], Puddle World [16], and Mountain Car [16]. Results are shown in Fig. 11 of Appendix G.3 from which we conclude that i) by ignoring the posterior's dependence on ω, existing model-free Bayesian approaches are less

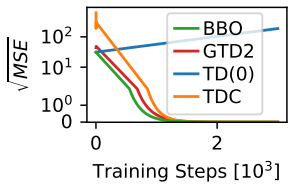
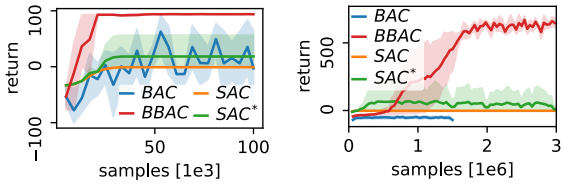


Figure 3: Tsitsiklis counterexample.

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 [25], 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].

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.



(a) MountainCar (b) Cartpole

Figure 4: Continuous control with sparse reward.

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 [56]: *MountainCar-Continuous-v0* from Gym benchmark suite and a slightly modified version of the *cartpole-swingup_sparse* from DeepMind Control Suite [72]. 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 [39]; and SAC*, a tailored version which uses a single Q-function to avoid *pessimistic underexploration* [19] due to the use of the double-minimum-Q trick (see Appendix H for details). To understand the practical implications of our theoretical results, we also compare against BAC which is a variant of BBAC where $\hat{Q}_{\omega_l^\star} = \hat{B}_{\psi_l^\star}$. As we discussed in Section 4, BAC is the Bayesian actor-critic that results from applying RP approximate inference to the posterior used by existing model-free Bayesian approaches with bootstrapping, which introduces error into the posterior.

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 H.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 H.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 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 ignoring this theoretical issue leads to negative empirical consequences on performance as BAC fails to solve both tasks where sampling from the correct posterior in BBAC succeeds, verifying empirically that it is essential for Bayesian model-free RL algorithms with bootstrapping sample from the BBO posterior. In Appendix H.2, we also investigate RP-BBAC's sensitivity to randomized prior hyperparameters. The range of working hyperparameters is wide and easy to tune.

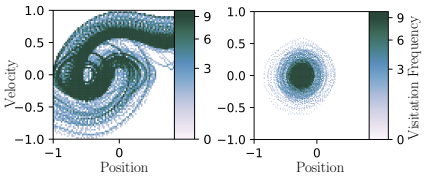


Figure 5: State Support for RP-BBAC (left) and SAC (right) in MountainCar-Continuous-v0.

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.

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. A

[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. A

[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, G.2.1, G.2.2, G.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, B.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, G.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, 1, H.1

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

[20] 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, E.1

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

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

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

[24] 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. A

[25] 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

[26] 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

[27] 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

[28] 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

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

[30] 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

[31] 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

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

[33] 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

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

[35] 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

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

[37] 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

[38] 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

[39] 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, H.1

[40] 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. F

[41] 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, B.3, 2, B.5

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

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

[44] 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, B.5

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

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

[47] 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. A

[48] 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

[49] 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

[50] 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

[51] 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

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

[53] 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

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

[55] 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

[56] 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, E.1

[57] 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, E.1, H.2

[58] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Uncertainty in neural networks: Bayesian ensembling. *ArXiv Preprint*, abs/1810.05546, 10 2019. 4.1, E.1

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

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

[61] 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, B.3

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

[63] 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. B.3

[64] 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. 3, B.3

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

[66] 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

[67] 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. A

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

[69] 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, D.2

[70] 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, D.2

[71] 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

[72] 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, H, H.2

[73] 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

[74] 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

[75] 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, G.2, G.2.1

[76] 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

[77] 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

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

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

[80] 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. B.3

[81] 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 A
   (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 B.1 and B.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 B.
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]