Bayesian Exploration Networks

Mattie Fellows [*1] **Brandon Kaplowitz** [*2] **Christian Schroeder de Witt** [1] **Shimon Whiteson** [3]

Abstract

Bayesian reinforcement learning (RL) offers a principled and elegant approach for sequential decision making under uncertainty. Most notably, Bayesian agents do not face an exploration/exploitation dilemma, a major pathology of frequentist methods. However theoretical understanding of model-free approaches is lacking. In this paper, we introduce a novel Bayesian model-free formulation and the first analysis showing that model-free approaches can yield Bayes-optimal policies. We show all existing model-free approaches make approximations that yield policies that can be arbitrarily Bayes-suboptimal. As a first step towards model-free Bayes optimality, we introduce the Bayesian exploration network (BEN) which uses normalising flows to model both the aleatoric uncertainty (via density estimation) and epistemic uncertainty (via variational inference) in the Bellman operator. In the limit of complete optimisation, BEN learns true Bayes-optimal policies, but like in variational expectation-maximisation, partial optimisation renders our approach tractable. Empirical results demonstrate that BEN can learn true Bayes-optimal policies in tasks where existing model-free approaches fail.

1. Introduction

In reinforcement learning (RL), an agent is tasked with learning an optimal policy that maximises expected return in a Markov decision process (MDP). In most cases, the agent is in a learning setting and does not know the underlying MDP a priori: typically, the reward and transition distributions are unknown. A Bayesian approach to reinforcement learning characterises the uncertainty in unknown governing variables in the MDP by inferring a posterior over their values conditioned on observed histories of interactions. Using the posterior it is possible to marginalise across unknown variables and derive a belief transition distribution that characterises how the uncertainty will evolve over all future timesteps. The resulting Bayesian RL (BRL) objective transforms a learning problem into a planning problem with a well defined set of optimal policies, known as Bayes-optimal policies, which are a gold standard for exploration (Martin, 1967; Duff, 2002). Moreover, any non-Bayesian policy is suboptimal in terms of optimising the expected returns according to the belief induced by the prior and model of the state and reward transition distributions.

Despite the formal theoretical benefits, learning Bayes-optimal policies remains a significant challenge due to several sources of intractability. Firstly, model-based approaches must maintain a posterior over a model of the state transition dynamics, which is notoriously computationally complex for even low dimensional state spaces (Wasserman, 2006). Secondly, even if it is tractable to calculate and maintain the posterior, the marginalisation needed to find the Bayesian transition and reward distributions requires high dimensional integrals. Finally, given the Bayesian distributions, a planning problem must then be solved in belief space for every history-augmented state to obtain the Bayes-optimal policy.

Alternatively, model-free approaches characterise uncertainty in a Bellman operator. This avoids the issues of modelling uncertainty in high dimensional transition distributions, as Bellman operators require the specification of a one-dimensional conditional distribution. Whilst model-free approaches to BRL exist, little is known about their theoretical properties. Our main contribution is to provide a theoretical analysis to answer three core questions: I) Can model-free approaches learn Bayes-optimal policies? II) What are the relative benefits of model-free approaches? and III) Are existing model-free approaches Bayes-optimal? We answer these questions by introducing a novel model-free formulation that is provably Bayes-optimal whilst still characterising uncertainty in a low-dimensional distribution over Bellman operators. Moreover, we prove that all existing methods inadvertently solve an approximation to the true Bayesian objective that prevents them from learning a

[*]Equal contribution [1]Department of Engineering Science, University of Oxford, Oxford, United Kingdom [2]Department of Economics, New York University, New York, United States of America [3]Department of Computer Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Mattie Fellows <matthew.fellows@eng.ox.ac.uk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

true Bayes-optimal policy.

Motivated by our analysis, we introduce a Bayesian exploration network (BEN) for model-free BRL. BEN reduces the dimensionality of inputs to a one-dimensional variable using a Q-function approximator. The output is then passed through a Bayesian network. Like in an actor-critic approach, BEN can be trained using partial stochastic gradient descent (SGD) methods at each timestep, bypassing computational complexity issues associated with finding a Bayes-optimal policy. This comes at the expense of learning an approximately Bayes-optimal policy but one that converges to the true Bayes-optimal policy in the limit of complete optimisation. To verify our theoretical claims, we evaluate BEN in a search and rescue environment, which is a novel higher dimensional variant of the tiger problem (Kaelbling et al., 1998). We show BEN solves the task while oracles of existing state-of-the-art model-free BRL approaches based on BootDQN+Prior (Osband et al., 2018) and Bayesian Bellman Actor Critic (Fellows et al., 2021) fail due to their inability to learn Bayes-optimal policies. Moreover, our results show that, whilst in the limit of complete optimisation BEN recovers true Bayes-optimal policies, complete optimisation is not necessary as BEN behaves near Bayes-optimally after taking only a few optimisation steps on our objective for every observation.

2. Preliminaries

2.1. Contextual RL

We define a space of infinite-horizon, discounted contextual Markov decision processes (CMDPs) (Hallak et al., 2015) by introducing a context variable $\phi \in \Phi \subseteq \mathbb{R}^d$: $\mathcal{M}(\phi) := \langle \mathcal{S}, \mathcal{A}, P_0, P_S(s,a,\phi), P_R(s,a,\phi), \gamma \rangle$ where each ϕ indexes a specific MDP by parametrising a transition distribution $P_{\mathcal{S}}(s,a,\phi) : \mathcal{S} \times \mathcal{A} \times \Phi \to \mathcal{P}(\mathcal{S})$ and reward distribution $P_R(s,a,\phi) : \mathcal{S} \times \mathcal{A} \times \Phi \to \mathcal{P}(\mathbb{R})$. We denote the corresponding joint conditional state-reward transition distribution as $P_{R,S}(s,a,\phi)$. We assume that the agent has complete knowledge of the set of states $\mathcal{S} \subseteq \mathbb{R}^n$, set of actions $\mathcal{A}$, initial state distribution $P_0 \in \mathcal{P}(\mathcal{S})$ and discount factor γ. An agent follows a policy $\pi : \mathcal{S} \times \Phi \to \mathcal{P}(\mathcal{A})$, taking actions $a_t \sim \pi(s_t, \phi)$. We denote the set of all context-conditioned policies as $\Pi_\Phi := \{\pi : \mathcal{S} \times \Phi \to \mathcal{P}(\mathcal{A})\}$. The agent is assigned an initial state $s_0 \sim P_0$. As the agent interacts with the environment, it observes a history of data $h_t := \{s_0, a_0, r_0, s_1, a_1, r_1, \dots a_{t-1}, r_{t-1}, s_t\} \in \mathcal{H}_t$ where $\mathcal{H}_t$ is the corresponding state-action-reward product space. We denote the context-conditioned distribution over history h_t as: $P_t^\pi(\phi)$ with density $p_t^\pi(h_t|\phi) = p_0(s_0) \prod_{i=0}^t \pi(a_i|s_i,\phi) p(r_i, s_{i+1}|s_i, a_i, \phi)$.

In the infinite-horizon, discounted setting, the goal of an agent in MDP $\mathcal{M}(\phi)$ is to find a policy that optimises the objective: $J^\pi(\phi) = \mathbb{E}_{\tau_\infty \sim P_\infty^\pi(\phi)} [\sum_{t=0}^\infty \gamma^t r_t]$. We denote an optimal policy as $\pi^\star(\cdot, \phi) \in \Pi_\Phi^\star(\phi) := \arg\max_{\pi \in \Pi_\Phi} J^\pi(\phi)$, where $\Pi_\Phi^\star(\phi)$ is the set of all optimal MDP-conditioned policies that are optimal for $\mathcal{M}(\phi)$. For an optimal policy $\pi^\star$, the optimal quality function (Q-function) $Q^\star : \mathcal{S} \times \mathcal{A} \times \Phi \to \mathbb{R}$ satisfies the optimal Bellman equation: $\mathcal{B}^\star [Q^\star] (s_t, a_t, \phi) = Q^\star(s_t, a_t, \phi)$ where

$$\mathcal{B}^\star [Q^\star] (s_t, a_t, \phi) := \mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} [r_t + \max_{a' \in \mathcal{A}} Q^\star(s_{t+1}, a', \phi)],$$

is the optimal Bellman operator.

If the agent has access to the true MDP $\mathcal{M}(\phi^\star)$, computational complexity issues aside, an optimal policy can be obtained by solving a *planning* problem . In the more realistic setting, the agent does not have access to the MDP's transition dynamics and/or reward function. The agent must balance learning these variables through exploration of the MDP at the cost of behaving suboptimally with solving the underlying planning problem by exploiting the information it has observed. This setting is known as a *learning* problem and solving the exploration/exploitation dilemma remains a major challenge for any agent learning to behave optimally.

2.2. Bayesian RL

A Bayesian epistemology characterises the agent's uncertainty in the MDP through distributions over Φ. We start by defining the prior distribution P_Φ which represents the a priori belief in the true value $\phi^\star$ before the agent has observed any transitions. Priors are a powerful aspect of BRL, allowing practitioners to provide the agent with any information about the MDP and transfer knowledge between agents and domains. In the tabula rasa setting, priors can be uninformative, can be used to encode optimism or pessimism in unknown states; or can be a minimax prior representing the worst possible prior distribution over MDPs an agent could face (Buening et al., 2023). Given a history h_t, we aim to reason over future trajectories; thus, Bayesian agents follow policies that condition on histories rather than single states. We denote the space of all histories $\mathcal{H} := \{\mathcal{H}_t | t \geq 0\}$ and the set of all history-conditioned policies as $\Pi_\mathcal{H} := \{\pi : \mathcal{H} \to \mathcal{P}(\mathcal{A})\}$. A Bayesian agent characterises the uncertainty in the MDP by inferring the posterior $P_\Phi(h_t)$ for each $t \geq 0$.

The prior is a special case of the posterior with $h_t = \varnothing$. The posterior $P_\Phi(h_t)$ represents the agent's beliefs in the MDP and can be used to *marginalise* across all CMDPs according to the agent's uncertainty. This yields the Bayesian state-reward transition distribution: $P_{R,S}(h_t, a_t) := \mathbb{E}_{\phi \sim P_\Phi(h_t)} [P_{R,S}(s_t, a_t, \phi)]$. Given this distribution, we can reason over counterfactual future trajectories using the prior predictive distribution over trajectories P_t^π with density: $p_t^\pi(h_t) = p_0(s_0) \prod_{i=0}^t \pi(a_i|h_i) p(r_i, s_{i+1}|h_i, a_i)$. Using the predictive distribution, we define the BRL objective as

$J^{\pi}_{\text{Bayes}} \coloneqq \mathbb{E}_{h_\infty \sim P^{\pi}_{\infty}} \left[\sum_{i=0}^{\infty} \gamma^i r_i\right]$. A corresponding optimal policy is known as a Bayes-optimal policy, which we denote as $\pi^{\star}_{\text{Bayes}}(\cdot) \in \Pi^{\star}_{\text{Bayes}} \coloneqq \arg\max_{\pi \in \Pi_{\mathcal{H}}} J^{\pi}_{\text{Bayes}}$.

Unlike in frequentist RL, Bayesian variables depend on histories obtained through posterior marginalisation; hence the posterior is often known as the *belief state*, which augments each ground state s_t like in a partially observable MDP (POMDP). Analogously to the state-transition distribution in frequentist RL, we can define a belief transition distribution $P_{\mathcal{H}}(h_t, a_t)$ using the Bayesian reward and transition distributions, which has the density: $p_{\mathcal{H}}(h_{t+1}|h_t, a_t) = p(s_{t+1}, r_t|h_t, a_t) \underbrace{p(h_t, a_t|h_t, a_t)}_{=1} = p(s_{t+1}, r_t|h_t, a_t)$.

Using the belief transition, we define the *Bayes-adaptive MDP* (BAMDP) (Duff, 2002):

$$\mathcal{M}_{\text{BAMDP}} \coloneqq \langle \mathcal{H}, \mathcal{A}, P_0, P_{\mathcal{H}}(h, a), \gamma \rangle,$$

which can be solved using planning methods to obtain a Bayes-optimal policy (Martin, 1967).

A Bayes-optimal policy naturally balances exploration with exploitation: after every timestep, the agent's uncertainty is characterised via the posterior conditioned on the history h_t, which includes all future trajectories to marginalise over. The BRL objective, therefore, accounts for how the posterior evolves after each transition, and hence any Bayes-optimal policy $\pi^{\star}_{\text{Bayes}}$ is optimal not only according to the epistemic uncertainty at a single timestep but also to the epistemic uncertainty at every future timestep, decaying according to the discount factor.

Unlike in frequentist RL, if the agent is in a learning problem, finding a Bayes-optimal policy is always possible given sufficient computational resources. This is because any uncertainty in the MDP is marginalised over according to the belief characterised by the posterior. BRL thus does not suffer from the exploration/exploitation dilemma as actions are sampled from optimal policies that only condition on historical observations h_t, rather than the unknown MDP $\phi^{\star}$. More formally, this is a direct consequence of the *conditionality principle*, which all Bayesian methods adhere to, meaning that Bayesian decisions never condition on data that the agent has not observed. From this perspective, the exploration/exploitation dilemma is a pathology that arises because frequentist approaches violate the conditionality principle.

For a Bayes-optimal policy $\pi^{\star}$, we define the optimal Bayesian Q-function as $Q^{\star}(h_t, a_t) \coloneqq Q^{\pi^{\star}_{\text{Bayes}}}(h_t, a_t)$, which satisfies the optimal Bayesian Bellman equation $Q^{\star}(h_t, a_t) = \mathcal{B}^{\star}[Q^{\star}](h_t, a_t)$ where:

$$\mathcal{B}^{\star}[Q^{\star}](h_t, a_t) \coloneqq \mathbb{E}_{h_{t+1} \sim P_{\mathcal{H}}(h_t, a_t)} \left[r_t + \gamma \max_{a'} Q^{\star}(h_{t+1}, a') \right],$$

is the optimal Bayesian Bellman operator. It is possible to construct a Bayes-optimal policy by choosing the action that maximises the optimal Bayesian Q-function $a_t \in \arg\max_{a'} Q^{\star}(h_t, a')$; hence learning $Q^{\star}(h_t, \cdot)$ is sufficient for solving the BAMDP. We take this value-based approach in this paper.

2.3. Related Work

BEN is the first model-free approach to BRL that can learn Bayes-optimal policies. To relate BEN to other approaches, we clarify the distinction between model-free and model-based BRL:

Definition 2.1. Model-based approaches define a prior P_{Φ} over and a model of the MDP's state and reward transition distributions: $P_S(s, a, \phi)$ and $P_R(s, a, \phi)$. Model-free approaches define a prior P_{Φ} over and a model of the MDP's Bellman operators: $P_B(\cdot, \phi)$.

This definition mirrors classical interpretations of model-based and model-free RL, which categorises algorithms according to whether a model of transition dynamics is learnt or the Q-function is estimated directly (Sutton & Barto, 2018). We formulate a novel Bayesian model-free approach in Section 3.1 before proving in Theorem 3.2 that whichever approach is taken, a Bayes-optimal policy may still be learnt.

We provide a review of several model-based approaches and their approximations in Appendix A, focusing instead on model-free approaches here. The majority of existing model-free approaches purportedly infer a posterior over Q-functions $P^{\pi}_{Q}(h_t)$ given a history of samples h_t, thus requiring a model of the aleatoric uncertainty in Q-function samples $q \sim P^{\pi}_{Q}(s, a, \phi)$. $P^{\pi}_{Q}(s, a, \phi) : \mathcal{S} \times \mathcal{A} \times \Phi \to \mathcal{P}(\mathbb{R})$ is typically a parametric Gaussian, which is a conditional distribution over a one-dimensional space, allowing for standard techniques from Bayesian regression to be applied. As inferring a posterior over Q-functions requires samples from complete returns, some degree of bootstrapping using function approximation is required for algorithms to be practical (Kuss & Rasmussen, 2003; Engel et al., 2005; Gal & Ghahramani, 2016; Osband et al., 2018; Fortunato et al., 2018; Lipton et al., 2018; Osband et al., 2019; Touati et al., 2019). By introducing bootstrapping, model-free approaches actually infer a posterior over *Bellman operators*, which concentrates on the true Bellman operator with increasing samples under appropriate regularity assumptions (Fellows et al., 2021). Instead of attempting to solve the BAMDP exactly, existing model-free approaches employ posterior sampling where a single MDP is drawn from the posterior at the start of each episode (Thomson, 1933; Strens, 2000; Osband et al., 2013), or optimism in the face of uncertainty (OFU) (Lai & Robbins, 1985; Kearns & Singh, 2002) where exploration is increased or decreased by a heuristic to reflect the uncertainty characterised by

the posterior variance (Jin et al., 2018; Ciosek et al., 2019; Luis et al., 2023). Unfortunately, both posterior sampling and OFU exploration can be highly inefficient and far from Bayes-optimal (Zintgraf et al., 2020; Buening et al., 2023). Exploration strategies aside, a deeper issue with existing model-free Bayesian approaches is that an optimal policy under their formulations is not Bayes-optimal, but instead solves either a *myopic* or *QBRL* approximation to the BRL objective. We explore these approximations theoretically in Sections 3.4 and 3.5 respectively.

3. Analysis of Model-Free BRL

Detailed proofs for theorems are provided in Appendix B. We now provide an analysis of model-free BRL under function approximation to answer the following questions:

I) Can model-free BRL approaches yield Bayes-optimal solutions and is there an analytic relationship between model-based and model-free approaches?

II) What are the benefits of model-free approaches over model-based approaches?

III) Are existing model-free BRL approaches Bayes-optimal and, if not, what solutions do these approaches learn instead?

We introduce a parametric function approximator $Q_\omega : \mathcal{H} \times \mathcal{A} \to \mathbb{R}$ parametrised by $\omega \in \Omega$ to approximate the optimal Bayesian Q-function. The function approximator satisfies the following regularity assumption:

Assumption 3.1. Assume that $Q_\omega(h_t, a_t)$ is a Lebesgue measurable mapping $Q_\omega(h_t \setminus s_t, a_t, \cdot) : \mathcal{S} \to \mathbb{R}$ for all $\omega \in \Omega, h_t \setminus s_t \in \mathcal{H}$ and $a_t \in \mathcal{A}$ and the set $\arg\max_{a'} Q_\omega(h_t, a')$ always exists.

Assumption 3.1 should automatically be satisfied for most function approximators of interest in RL (for example bounded Lipschitz functions defined on closed sets).

3.1. Model-Free BRL Formulation

In model-free BRL, our goal is to characterise uncertainty in the optimal Bayesian Bellman operator instead of the reward-state transition distribution. Given samples from the true reward-state distribution $r_t, s_{t+1} \sim P^\star_{R,S}(s_t, a_t)$ we use *bootstrapping* to estimate the optimal Bayesian Bellman operator

$$b_t = \beta_\omega(h_{t+1}) \coloneqq r_t + \gamma \max_{a'} Q_\omega(h_{t+1}, a').$$

We refer to $\beta_\omega(h_{t+1})$ as the bootstrap function. Similarly to Fellows et al. (2021), we interpret bootstrapping as making a change of variables under the mapping $\beta_\omega(\cdot, h_t, a_t) : \mathbb{R} \times \mathcal{S} \to \mathbb{R}$. The bootstrapped samples b_t have distribution $P^\star_B(h_t, a_t; \omega)$ which is the *pushforward* distribution over next period's possible updated Q-values satisfying: $\mathbb{E}_{b_t \sim P^\star_B(h_t, a_t; \omega)}[f(b_t)] = \mathbb{E}_{r_t, s_{t+1} \sim P^\star_{R,S}(s_t, a_t)}[f(r_t + \gamma \max_{a'} Q_\omega(h_{t+1}, a'))]$ for any measurable function $f : \mathbb{R} \to \mathbb{R}$. We refer to $P^\star_B(h_t, a_t; \omega)$ as the Bellman distribution.

When predicting b_t given an observation h_t, a, there are two sources of uncertainty to take into account: firstly, even if $P^\star_B(h_t, a_t; \omega)$ is known, there is natural stochasticity due to the environment's reward-state transition dynamics that prevents b_t being determined. This type of uncertainty is known as *aleatoric uncertainty*. Secondly, in a learning problem, the Bellman distribution $P^\star_B(h_t, a_t; \omega)$ cannot be determined a priori and must be inferred from observations of b_t. This type of uncertainty is known as *epistemic uncertainty*. Unlike aleatoric uncertainty, epistemic uncertainty can always be reduced with more data as the agent explores.

Our first step is to introduce a model of the process $b_t \sim P^\star_B(h_t, a_t; \omega)$, which characterises the aleatoric uncertainty in the optimal Bellman operator. In this paper, we choose a parametric model $P_B(h_t, a_t, \phi; \omega)$ with density $p(b_t | h_t, a_t, \phi; \omega)$ parametrised by $\phi \in \Phi$; however a non-parametric model may also be specified. We specify a prior P_Φ over the parameter space which represents the agent's initial belief over the space of models. The space of models $P_B(h_t, a_t, \phi; \omega)$ can be interpreted as a hypothesis space over the true Bellman distribution $P^\star_B(h_t, a_t; \omega)$, with each hypothesis indexed by a parameter $\phi \in \Phi$.

Let $\mathcal{D}_\omega(h_t) \coloneqq \{(b_i, h_i, a_i)\}_{i=0}^{t-1}$ denote the dataset of bootstrapped samples. Once the agent has observed $\mathcal{D}_\omega(h_t)$, it updates its belief in ϕ by inferring a posterior $P_\Phi(\mathcal{D}_\omega(h_t))$, whose density can be derived using Bayes' rule:

$$p(\phi | \mathcal{D}_\omega(h_t)) = \frac{\prod_{i=0}^{t-1} p(b_i | h_i, a_i, \phi; \omega) p(\phi)}{\mathbb{E}_{\phi \sim P_\Phi}\left[\prod_{i=0}^{t-1} p(b_i | h_i, a_i, \phi; \omega)\right]}.$$

The posterior $P_\Phi(\mathcal{D}_\omega(h_t))$ characterises the epistemic uncertainty over the hypothesis space, which we use to obtain the predictive optimal Bellman distribution: $P_B(h_t, a_t; \omega) = \mathbb{E}_{\phi \sim P_\Phi(\mathcal{D}_\omega(h_t))}[P_B(h_t, a_t, \phi; \omega)]$.

The predictive optimal Bellman distribution is analogous to the predictive reward-state transition distribution introduced in Section 2.2 as it is derived by marginalising across the hypothesis space according to the epistemic uncertainty in each model under the posterior. Taking expectations over the variable b_t using $P_B(h_t, a_t; \omega)$, we derive the predictive optimal Bellman operator: $B^+[Q_\omega](h_t, a_t) \coloneqq \mathbb{E}_{b_t \sim P_B(h_t, a_t; \omega)}[b_t]$, which integrates both the aleatoric and epistemic uncertainty in b_t to make a Bayesian prediction of the optimal Bellman operator at each timestep t. Intuitively, we expect $B^+[Q_\omega](h_t, a_t)$ to play the same role as the optimal Bayesian Bellman operator introduced in Section 2.2, which we now formally confirm:

Theorem 3.2. *Let Assumption 3.1 hold, then* $B^+[Q_\omega](h_t, a_t) = B^\star[Q_\omega](h_t, a_t)$.

In answer to Question I, Theorem 3.2 proves that model-free approaches can be Bayes-optimal: if $Q_{\omega^\star}$ satisfies $Q_{\omega^\star}(\cdot) = B^+[Q_{\omega^\star}](\cdot)$, it also satisfies an optimal Bayesian Bellman equation: $Q_{\omega^\star}(\cdot) = \mathcal{B}^\star[Q_{\omega^\star}](\cdot)$ hence any agent taking action $a_t \in \arg\max_{a'} Q_{\omega^\star}(h_t, a')$ is thus Bayes-optimal with respect to the prior P_Φ and likelihood defined by the model $P_B(h_t, a_t, \phi; \omega^\star)$. Theorem 3.2 is a consequence of the *sufficiency principle*: this result confirms that it does not matter whether we characterise uncertainty in the reward-state transition distributions or the optimal Bellman operator, a Bayes-optimal policy may still be learned.

3.2. Aleatoric Uncertainty Matters

By making a Lipschitz assumption, we answer the final part of Question I to find an exact relationship between a model-based and the equivalent model-free approach:

Assumption 3.3. In addition to Assumption 3.1, assume $Q_\omega(s, a)$ is Lipschitz in s for all $\omega \in \Omega$, $a \in \mathcal{A}$ and $|J_\beta(r_t, s_{t+1})| \coloneqq \|\nabla_{r_t, s_{t+1}} \beta_\omega(h_{t+1})\|_2 > 0$ a.e.

To determine the relationship between model-free and model-based approaches, we must study the *pre-image* of the bootstrap function defined in Section 3.1:

$$\beta_\omega^{-1}(b_t, h_t, a_t) \coloneqq \{r_t, s_{t+1} | \beta_\omega(h_{t+1}) = b_t\}.$$

Intuitively, $\beta_\omega^{-1}(b_t, h_t, a_t)$ returns all of the reward-next state pairs that produce an equivalent b_t, thereby mapping from Bellman operator space (whose uncertainty is characterised by model-free methods) to reward-state space (whose uncertainty is characterised by model-based methods). We sketch this mapping Figure 1. We now derive the exact Bellman distribution model used by model-free approaches starting from a given reward-state transition model as specified by a model-based approach:

Corollary 3.4. *Let Assumption 3.3 hold. If there exists a $\omega^\star$ satisfying $\mathcal{B}^\star[Q_{\omega^\star}](h_t, a_t) = Q_{\omega^\star}(h_t, a_t)$, then the Bayes-optimal policy under a parametric reward-state transition model with density $p(r_t, s_{t+1} | s_t, a_t, \phi)$ and prior P_Φ is equivalent to a Bayes-optimal policy using the same prior under a optimal Bellman model with density:*

$$p(b_t | h_t, a_t, \phi; \omega) = \mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} \left[\frac{\delta(r_t, s_{t+1} \in \beta_\omega^{-1}(b_t, h_t, a_t))}{|J_\beta(r_t, s_{t+1})|} \right].$$

Corollary 3.4 relies on the coarea formula (Federer, 1969, Theorem 3.2.12), which generalises the change of variables formula to non-injective mappings between dimensions. Like in SurVAE flows (Nielsen et al., 2020) the change of variables for non-injective mappings must be stochastic in the reverse direction $b_t \to r_t, s_{t+1}$ to account for the fact that several variables can map to a single b_t.

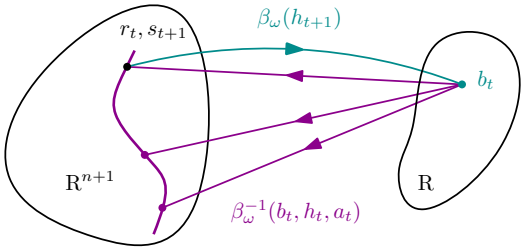


Figure 1. Sketch of transformation of variables β_ω.

$P_{R,S}(b_t, h_t, a_t; \omega)$ is thus a mapping to the set of probability distributions over the pre-image $\beta_\omega^{-1}(b_t, h_t, a_t)$ (see Figure 1 for a sketch). Due to Assumption 3.3, $\beta_\omega^{-1}(b_t, h_t, a_t)$ is an n-dimensional manifold in $\mathbb{R}^{n+1}$ space, for example a 1-dimensional curved line in $\mathbb{R}^2$ as shown in our sketch.

Accurately representing the aleatoric uncertainty through the model $P_B(h_t, a_t; \omega)$ is the focus of distributional RL (Bellemare et al., 2017) and has been ignored by the model-free BRL community. As discussed in Section 2.3, most existing parametric model-free BRL approaches have focused on representing the epistemic uncertainty in the posterior under a parametric Gaussian model (Osband et al., 2018). One notable exception is model-based Q-variance estimation (Luis et al., 2023). However, this approach is derived from BootDQN+Prior (Osband et al., 2018) which, as we prove in Section 3.5, forfeits Bayes-optimality.

Using Corollary 3.4, we investigate whether a Gaussian distribution is appropriate. If the function $\beta_\omega(h_{t+1})$ is bijective with inverse $r_t, s_{t+1} = \beta_\omega^{-1}(b_t, h_t, a_t)$ then the density is:

$$p(b_t | h_t, a_t, \phi; \omega) = \frac{p(\beta_\omega^{-1}(b_t, h_t, a_t) | s_t, a_t, \phi)}{|J_\beta(\beta_\omega^{-1}(b_t, h_t, a_t))|}.$$

Section 3.2 demonstrates that even in simple MDPs with bijective transformation of variables, the equivalent space of reward-state transition distributions cannot be modelled well by Gaussian models over Bellman operators as the density $p(b_t | h_t, a_t, \phi; \omega)$ can be arbitrarily far from Gaussian depending upon the choice of function approximator. This issue has been investigated empirically when modelling uncertainty in Q-functions (Janz et al., 2019), where improving the representative capacity of a Gaussian model using successor features reduces the learning time from $\mathcal{O}(L^3)$ to $\mathcal{O}(L^{2.5})$ in the L-episode length chain task (Osband et al., 2018) under a posterior sampling exploration regime. This issue is particularly pertinent if we are concerned with finding polices that approach Bayes-optimality. Epistemic uncertainty estimates are rendered useless if the space of MDPs that the agent is uncertain over does not reflect the agent's environment. The key insight is that accurately representing both aleatoric and epistemic uncertainty is crucial for learning Bayesian policies with successful exploration strategies as epistemic uncertainty cannot be considered in isolation from aleatoric uncertainty.

3.3. Benefits of Model-Free over Model-Based BRL

As discussed in Section 1, obtaining exact Bayes-optimal policies is hopelessly intractable for all but the simplest domains. Models with many parameters $\Phi \subseteq \mathbb{R}^d$ exacerbate this problem as inferring a posterior and carrying out posterior marginalisation becomes exponentially more computationally complex in d (Bellman, 1961). With this in mind, we answer Question II by showing that model-free approaches need fewer parameters than equivalent model-based approaches.

As many real-world problems of interest have high dimensional state spaces $\mathcal{S} \subseteq \mathbb{R}^n$, representing the state transition distribution in model-based approaches accurately requires a model $P_S(s, a, \phi)$ with a large number of parameters. By contrast, $b_t \in \mathbb{R}$ and hence representing the distribution $P_B^\star(h_t, a_t; \omega)$ accurately requires a model $P_B(h_t, a_t, \phi; \omega)$ with fewer parameters. Formally, the sample efficiency for density estimation of conditional distributions scales poorly with increasing dimensionality (Grünewälder et al., 2012): Wasserman (2006) show that when using a nonparametric frequentist kernel approach to density estimation, even with optimal bandwidth, the mean squared error scales as $\mathcal{O}(N^{\frac{-4}{n+4}})$ where N is the number of samples from the true density. A mean squared error of less than 0.1 when the target density is a multivariate Gaussian of dimension 10 requires 842,000 samples compared to 19 for a two-dimensional problem.

Empirical results verify theoretical analysis that density estimation using normalising flows scales poorly with increasing dimension of the target distribution (Papamakarios et al., 2021). From a Bayesian perspective, we also expect the posterior to concentrate at a slower rate with increasing dimensionality as the agent requires more data to decrease its uncertainty in the transition model parameters. As we discuss in Appendix A, tractable model-based approaches based on VariBAD (Zintgraf et al., 2020) circumvent this issue by only inferring a posterior over a small subset of model parameters $m \subset \phi$ at the expense of Bayes-optimality.

An additional benefit of modelling uncertainty in the Bellman operator is that, as we sketch in Figure 1, the transformation of variables $\beta_\omega(h_{t+1})$ is surjective. This reflects the fact that several reward-state transition distributions can yield the same optimal Bellman operator. As such, a single hypothesis over Bellman operator models may account for several equivalent hypotheses over reward-state transition models. As we proved in Theorem 3.2 that it does not matter which variable we model uncertainty in, there may be unnecessary information in reward-state transition models that becomes redundant when finding a Bayes-optimal policy.

We now answer Question III by analysing the two families of existing model-free approaches, myopic BRL and QBRL, showing that both methods make approximations that prevent them from learning Bayes-optimal policies.

3.4. Myopic BRL

The most common approximation to exact BRL, whether intentional or not, is to solve a variation of the true BAMDP where the epistemic uncertainty update is *myopic*. Here the distribution: $P_{R,S}(h_t, s_{t'}, a_{t'}) = \mathbb{E}_{\phi \sim P_\Phi(h_t)}\left[P_{R,S}(s_{t'}, a_{t'}, \phi)\right]$ is used to characterise the epistemic uncertainty over all future timesteps $t' \geq t$ and does not account for how the posterior evolves after each transition through the Bayesian Bellman operator. The corresponding distribution over a trajectory $h_{t:t'}$ from timestep t to t' having observed h_t is: $p^\pi_{\text{Myopic}}(h_{t:t'}|h_t) = \prod_{i=t}^{t'-1} \pi(a_i|s_i, h_t) \cdot p_{R,S}(r_i, s_{i+1}|s_i, a_i, h_t)$. Here, only $P_\Phi(h_t)$ is used to marginalise over uncertainty at each timestep $t' \geq t$ and information in $h_{t:t'}$ is not used to update the posterior.

Several existing model-free approaches (Kuss & Rasmussen, 2003; Engel et al., 2005; Gal & Ghahramani, 2016; Fortunato et al., 2018; Lipton et al., 2018; Touati et al., 2019) naively introduce a Q-function approximator $Q_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ whose parameters minimise the mean-squared Bayesian Bellman error: $\omega^\star \in \arg\min_{\omega \in \Omega} \|Q_\omega(s, a) - \mathcal{B}^\star_{\text{Myopic}}[Q_\omega](h_t, s, a)\|^2_\rho$ where $\mathcal{B}^\star_{\text{Myopic}}[Q_\omega]$ is the myopic Bellman operator: $\mathcal{B}^\star_{\text{Myopic}}[Q_\omega](h_t, s_{t'}, a_{t'}) = \mathbb{E}_{r_{t'}, s_{t'+1} \sim P_{R,S}(h_t, s_{t'}, a_{t'})}\left[r_{t'} + \gamma \max_{a'} Q_\omega(s_{t'+1}, a')\right]$.

This Bellman operator does not propagate epistemic uncertainty across timesteps; hence myopic BRL is optimal with respect to myopic beliefs and is not Bayes-optimal. Recent notable exceptions are BootDQN+Prior (Osband et al., 2018; 2019), its actor-critic analogue BBAC (Fellows et al., 2021) and model-based Q-variance estimation (Luis et al., 2023); whilst these approaches satisfy an uncertainty Bellman equation (O'Donoghue et al., 2018), they can still be far from Bayes optimal, as we now show.

3.5. QBRL

BootDQN+Prior and BBAC assume that Bayesian value functions can be decomposed using an expectation over a contextual value function using the posterior. For example, for the optimal Q-function: $Q^\star_{\text{QBRL}}(h_t, a_t) = \mathbb{E}_{\phi \sim P_\Phi(h_t)}\left[Q^\star(s_t, a_t, \phi)\right]$ where $Q^\star(s_t, a_t, \phi)$ is the contextual optimal Q-function introduced in Section 2.1. An optimal QBRL policy can be constructed using $a^\star \in \arg\max_{a'} Q^\star_{\text{QBRL}}(h_t, a')$ and we denote the set of all optimal QBRL policies as Π_{QBRL}. This approach is analogous to the QMDP approximation used to solve POMDPs (Littman et al., 1995), which amounts to assuming that uncertainty in the agent's current belief over ϕ disappears after the next action. We refer to these methods as QBRL.

In QBRL, practical algorithms introduce a function ap-

proximator $Q_\omega(s_t, a_t, \phi)$ to estimate the contextual Q-function. A dataset $\mathcal{D}(h_t)$ of bootstrapped samples $b_t := r_t + \gamma \max_{a'} Q_\omega(s_{t+1}, a_{t+1}, \phi)$ and a predictive distribution over b_t is inferred following the formulation in Section 3.1. Whilst both QBRL methods use posterior sampling to avoid costly posterior marginalisations, ignoring this approximation, it has not been established whether QBRL methods have the potential to learn Bayes-optimal policies, that is whether $\Pi^\star_{\text{QBRL}} = \Pi^\star_{\text{Bayes}}$. To answer this question, we first define the set of QBRL policies: $\Pi_{\text{QBRL}} := \left\{ \mathbb{E}_{\phi \sim P_\Phi(\mathcal{H}_t)} \left[\pi(\cdot, \phi) \right] | \pi \in \Pi_\Phi \right\} \subset \Pi_{\mathcal{H}}$. Using the following theorem, we show that QBRL approaches maximise the Bayesian RL objective but the set of policies they optimise over is restricted to Π_{QBRL}:

Theorem 3.5. *Under Assumption 3.3,*

$$\begin{aligned} \Pi^\star_{\text{QBRL}} &= \arg\max_{\pi \in \Pi_{QBRL}} J^\pi_{Bayes}, \\ &= \left\{ \mathbb{E}_{\phi \sim P_\Phi(h_t)} \left[\pi^\star(\cdot, \phi) \right] | \pi^\star(\cdot, \phi) \in \Pi^\star_\Phi \right\}. \end{aligned}$$

Theorem 3.5 proves that the set of contextual optimal policies $\Pi^\star_{\text{Contextual}}$ can only be formed from a mixture of optimal policies conditioned on specific MDPs using the posterior. We thus prove that QBRL optimal policies can be arbitrarily Bayes-suboptimal in Corollary 3.6, using the tiger problem (Kaelbling et al., 1998) as a counterexample:

Corollary 3.6. *There exist MDPs with priors such that* $\Pi^\star_{QMDP} \cap \Pi^\star_{Bayes} = \varnothing$.

Finally, we note that a recent algorithm EVE (Schmitt et al., 2023) uses a combination of both myopic and QBRL approximation together.

4. Bayesian Exploration Network (BEN)

Using our formulation in Section 3.4, we develop a model-free BRL algorithm capable of learning Bayes-optimal policies. As we are taking a value-based approach in this paper, we focus on solving the optimal Bayesian Bellman equation. We now introduce the Bayesian Exploration network (BEN), which is comprised of three individual networks: a Q-network to reduce the dimensionality of inputs to a one-dimensional variable and then two normalising flows to characterise both the aleatoric and epistemic uncertainty over that variable as it passes through the Bellman operator.

4.1. Network Specification

Recurrent Q-Network For our choice of function approximator, we encode history using a recurrent neural network (RNN). Similar approximators are used in POMDP solvers (Hausknecht & Stone, 2015; Schlegel et al., 2023). The Q-function approximator is a mapping from history-action pairs, allowing uncertainty to propagate properly through the Bayesian Bellman equation. By contrast, encoding of history is missing from state-of-the-art model-free approaches based on QBRL as these function approximators condition on a context variable instead. We denote the output of the function approximator at time t as $q_t = Q_\omega(h_t, a_t)$.

Aleatoric Network We now specify our proposed model $P_B(h_t, a_t, \phi; \omega)$. Recall from Section 3.2 that the performance of model-free BRL depends on the capacity for representing aleatoric uncertainty. Whilst the model in Corollary 3.4 could be used in principle, there are two practical issues: firstly, we cannot assume knowledge of the preimage $b_\omega^{-1}(b_t, h_t, a_t)$ for arbitrary function approximators; and, secondly, the transformation $b_\omega(h_{t+1})$ is a surjective mapping $\mathbb{R}^{n+1} \to \mathbb{R}$, meaning we would not be taking advantage of the projection down to a lower dimensional space by first specifying a model $P_{R,S}(s_t, a_t, \phi)$.

Instead, we specify $P_B(h_t, a_t, \phi; \omega)$ using a normalising flow for density estimation (Rezende & Mohamed, 2015; Kobyzev et al., 2021), making a transformation of variables $b_t = B(z_{\text{al}}, q_t, \phi)$ where $z_{\text{al}} \in \mathbb{R}$ is a base variable with a zero-mean, unit variance Gaussian P_{al}. Sampling $b_t \sim P_B(h_t, a_t, \phi; \omega)$ is equivalent to sampling $z_t \sim P_{\text{al}}$ and then applying the transformation $b_t = B(z_{\text{al}}, q_t, \phi)$. Details can be found in Appendix C.1. We refer to this density estimation flow as the *aleatoric* network as it characterises the aleatoric uncertainty in the Bellman operator. Unlike in model-based approaches where the hypothesis space must be specified a-priori, in BEN the hypothesis space is determined by the representability of the aleatoric network, which can be tuned to the specific set of problems. Under mild regularity assumptions (Huang et al., 2018), an autoregressive flow as a choice for the aleatoric network can represent any target distribution $P_B(h_t, a_t, \phi; \omega)$ to arbitrary accuracy given sufficient data (Kobyzev et al., 2021).

A key advantage of our approach is that we have preprocessed the input to our aleatoric network through $q_t = Q_\omega(h_t, a_t)$ to extract features that reduce the dimensionality of the state-action space. This architecture hard-codes the prior information that a Bellman operator is a functional of the Q-function approximator, meaning that we only need to characterise aleatoric uncertainty in a lower dimensional input q_t. Unlike in VariBAD, we do not need to introduce frequentist heuristics to learn function approximator parameters ω. Instead these are learnt automatically by solving the optimal Bayesian Bellman equation, which we detail in Section 4.2.

Epistemic Network Inferring the posterior and carrying out marginalisation exactly is intractable for all but the simplest aleatoric networks, which would not have sufficient capacity to represent a complex target distribution $P_B(h_t, a_t, \phi; \omega)$. Instead we use variational inference via normalising flows to learn a tractable approximation P_ψ parametrised by $\psi \in \Psi$ which we learn by minimising the KL-divergence between the two distributions

$\mathrm{KL}(P_\psi \parallel P_\Phi(\mathcal{D}_\omega(h_t)))$. This is equivalent to minimising the tractable evidence lower bound $\mathrm{ELBO}(\psi; h, \omega)$. We provide details in Appendix C.2. We refer to this flow as the epistemic network as it characterises the epistemic uncertainty in ϕ. To our knowledge, BEN is the first time flows have been used for combined density estimation and variational inference.

4.2. Mean Squared Bayesian Bellman Error (MSBBE)

Finally, we learn a parametrisation $\omega^\star$ that satisfies the optimal Bayesian Bellman equation for our Q-function approximator. For BEN, this is equivalent to minimising the Mean Squared Bayesian Bellman Error (MSBBE) between the predictive optimal Bellman operator $B^+[Q_\omega]$ and Q_ω:

$$\mathrm{MSBBE}(\omega; h_t, \psi) \coloneqq \left\| B^+[Q_\omega](h_t, a_t) - Q_\omega(h_t, a_t) \right\|_\rho^2,$$

where ρ is an arbitrary sampling distribution with support over $\mathcal{A}$. Given sufficient compute, at each timestep t it is possible in principle to solve the nested optimisation problem for BEN: $\omega^\star \in \arg\min_{\omega\in\Omega} \mathrm{MSBBE}(\omega; h_t, \psi^\star(\omega))$, such that $\psi^\star(h_t, \omega) \in \arg\min_{\psi\in\Psi} \mathrm{ELBO}(\psi; h, \omega)$. Nested optimisation problems are commonplace in model-free RL and can be solved using two-timescale stochastic approximation: we update the epistemic network parameters ψ using gradient descent on an asymptotically faster timescale than the function approximator parameters ω to ensure convergence to a fixed point (Borkar, 2008; Heusel et al., 2017; Fellows et al., 2021), with ω playing a similar role as target network parameters used to stabilise TD. Solving the MSBBE exactly for every observable history h_t is computationally intractable. Instead, we propose partial minimisation of our objectives as outlined in Algorithm 1: after observing a new tuple $\{a, r, s\}$, the agent takes N_{Update} MSBBE update steps using the new data. This is equivalent to partially minimising the empirical expectation $\mathbb{E}_{h\sim h_t}\left[\mathrm{MSBBE}(\omega; h, \psi^\star(\omega))\right]$, where each $h \sim h_t$ is a sequence drawn from the observed history, analogously to how state-action pairs are drawn from the replay buffer in DQN (Mnih et al., 2016). To ensure a separation of timescales between parameter updates, the agent carries out $N_{\text{Posterior}}$ steps of SGD on the ELBO for every MSBBE update. Our algorithmic is shown in Algorithm 1.

Algorithm 1 APPROXBRL$(P_\Phi, \mathcal{M}(\phi))$

```
Initialise ω, ψ, h = s
Sample initial state s ~ P_0
Take N_Pretrain SGD Steps on MSBBE(ω)
while posterior not converged do
    Take action a ∈ arg max_a' Q_ω(h, a')
    Observe reward r ~ P_R(s, a, φ*)
    Transition to new state s ~ P_S(s, a, φ*)
    h ← {h, a, r, s}
    for N_Update Steps: do
        Take N_Posterior SGD steps on ELBO(ψ; h, ω)
        Take a SGD step on MSBBE(ω; h, ψ)
    end for
end while
```

Additionally, we exploit the fact that the MSBBE can be minimised prior to learning using samples of state-action pairs and so carry out N_{Pretrain} pretraining steps of SGD on the loss using the prior in place of the approximate posterior. If no prior knowledge exists, then the agent can be deployed. If there exists additional domain-specific knowledge, such as transitions shared across all MDPs or demonstrations at key goal states, this can also be used to train the agent using the model-based form of the Bellman operator. Full algorithmic details can be found in Appendix C.3.

5. Experiments

We introduce a novel search and rescue gridworld MDP as a more challenging, higher-dimensional extension to the tiger problem (which we show BEN can solve in Appendix D). An agent is tasked with rescuing N_{victims} victims from a dangerous situation whilst avoiding any one of N_{hazards} hazards. Details can be found in Appendix D.4. We evaluate BEN using a 7×7 grid size with 8 hazards and 4 victims.

Episodic Setting, Tabula Rasa In the episodic setting, the environment is reset after 245 timesteps (5 times the grid size to allow time to visit the edges of larger grids), and a new environment is uniformly sampled from the space of MDPs. After resetting, the epistemic parameters ψ are also reset, representing the belief in the new MDP returning to the prior. Q-network parameters ω are retained so the agent can exploit information that is shared across MDPs. We initialise the agent with a zero-mean Gaussian prior of diagonal variance equal to 0.1. We assume no prior environment knowledge of any kind in this experiment. We compare to PPO using an RNN to encode history as a strong frequentist baseline. We also compare to BootDQN+prior, which is the state-of-the-art Bayesian model-free method. Results are shown in Figure 2 where we plot the cumulative return for the three methods. Whilst BootDQN+prior can outperform RNN PPO due to its more sophisticated deep exploration, it cannot successfully learn to listen to solve the task. This confirms our key theoretical analysis that methods based on QBRL learn contextual policies that can be arbitrarily suboptimal (Corollary 3.6). Unlike BEN, BootDQN+prior cannot be used to approximate Bayes-optimal policies.

Episodic Setting, Weak Prior We repeat the experiment in the episodic setting, this time showing the agent examples of deterministic movement and the average reward for opening a door at random as a weak prior. The results for our implementation are shown in Figure 3. We plot the return at the end of each 245 timestep episode. As expected,

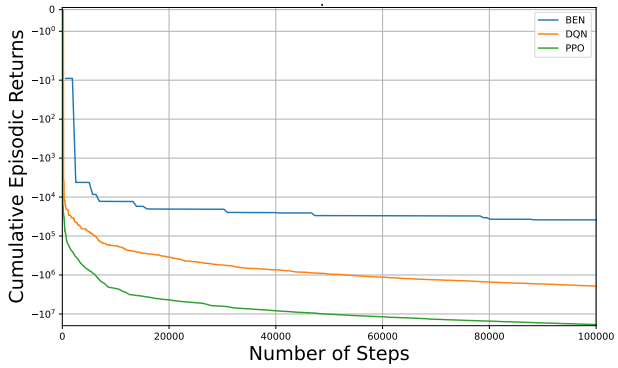


Figure 2. Evaluation in search and rescue episodic problem with no prior knowledge, showing cumulative return of BEN vs RNN PPO and BootDQN+prior.

BEN can solve this challenging problem, exploring in initial episodes to learn about how the listening states correlate to victim and hazard positions, then exploiting this knowledge in later episodes, finding all victims immediately. Our results demonstrate that BEN can scale to higher dimensional domains without forfeiting BRL's strong theoretical properties through approximation.

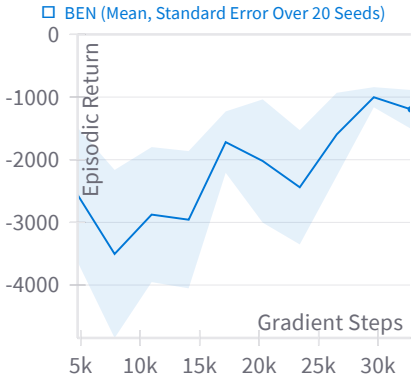


Figure 3. Evaluation in search and rescue episodic problem with weak prior knowledge, showing return of BEN after each episode.

Zero-shot Setting, Strong Prior In this setting, our goal is to investigate how effectively BEN can exploit strong prior knowledge to solve the search and rescue environment in a single episode. We prior-train BEN using simulations in related (but not identical) environments drawn from a uniform prior, showing the agent the effect of listening. This experiment is designed to mimic a real-life application where simulations can be provided by demonstrators in a generic training setting, followed by deployment in a novel environment where the robot has only one chance to succeed. Details can be found in Appendix D.5. To investigate effect of history dependency on the solution, we also compare to a variant of BEN using a Q-function that only conditions on state-actions, representing an ideal QBRL solution. We plot the cumulative return as a function of number of gradient steps over the course of the episode in Figure 4 for both BEN and the QBRL variant. Our results demonstrate that by exploiting prior knowledge, BEN can successfully rescue all victims and avoid all hazards, even when encountering a novel environment. In contrast, the oracle policy for existing state-of-the-art model-free methods, which learn a QBRL Bayes policy, cannot solve this problem because, as our analysis in Section 3.5 shows, $\Pi^{\star}_{\text{QBRL}}$ is limited to mixtures of optimal policies conditioned on ϕ, causing contextual agents to repeatedly hit hazards. This challenging setting showcases the high sample efficiency with low computational complexity of BEN.

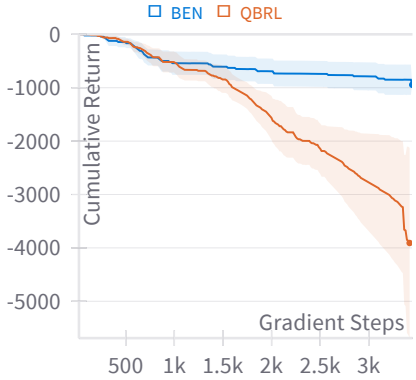


Figure 4. Evaluation in zero-shot search and rescue showing cumulative return for BEN vs. QBRL methods.

In addition, we perform two ablations in Appendix D.7. First, we demonstrate that performance depends on the capacity of the aleatoric network, verifying our claim in Section 4.1 that there is a balancing act specifying a rich enough hypothesis space to represent the true model accurately but that is small enough to generalise effectively. Secondly, we investigate how pre-training affects returns. As we decrease the number of prior pre-training MSBBE minimisation steps, we see that performance degrades in the zero-shot settling, as expected. Moreover, this ablation shows that relatively few pre-training steps are needed to achieve impressive performance once the agent is deployed in an unknown MDP.

6. Conclusions

We carried out theoretical analyses of model-free BRL and formulated the first model-free approach that can learn Bayes-optimal policies. We proved that existing model-free approaches for BRL are limited to optimising over a set of QBRL policies or optimise a myopic approximation of the true BRL objective. In both cases, the corresponding optimal policies can be arbitrarily Bayes-suboptimal. We introduced BEN, a model-free BRL algorithm that can successfully learn true Bayes-optimal policies. Our experimental evaluation confirms our analysis, demonstrating that BEN can behave near Bayes-optimally even under partial minimisation, paving the way for a new generation of model-free BRL approaches with the desirable theoretical properties of model-based approaches.

Acknowledgements

Mattie Fellows is funded by a generous grant from Waymo. Christian Schroeder de Witt is funded by UKRI grant: Turing AI Fellowship EP/W002981/1, Armasuisse Science+Technology, and an EPSRC IAA Impact Fund award. The experiments were made possible by a generous equipment grant from NVIDIA.

Impact Statement

This paper presents work whose goal is to advance the field of Bayesian reinforcement learning. Our primary contribution is theoretical and there are no specific dangers of our work. Any advancement in RL should be seen in the context of general advancements to machine learning. Whilst machine learning has the potential to develop useful tools to benefit humanity, it must be carefully integrated into underlying political and social systems to avoid negative consequences for people living within them. A discussion of this complex topic lies beyond the scope of this work.

References

Asmuth, J. and Littman, M. Learning is planning: near bayes-optimal reinforcement learning via monte-carlo tree search. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 19–26, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Beck, J., Vuorio, R., Liu, E., Xiong, Z., Zintgraf, L., Finn, C., and Whiteson, S. A survey of meta-reinforcement learning. 01 2023. doi: 10.48550/arXiv.2301.08028.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Bellman, R. E. *Adaptive Control Processes*. Princeton University Press, Princeton, 1961. ISBN 9781400874668. doi: doi:10.1515/9781400874668. URL https://doi.org/10.1515/9781400874668.

Bogachev, V. I. *Measure Theory*. Measure Theory ; 1. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 2007. edition, 2007. ISBN 1-280-74570-3.

Borkar, V. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency Gurgaon, 01 2008. ISBN 978-81-85931-85-2. doi: 10.1007/978-93-86279-38-5.

Buening, T. K., Dimitrakakis, C., Eriksson, H., Grover, D., and Jorge, E. Minimax-bayes reinforcement learning. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7511–7527. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/buening23a.html.

Ciosek, K., Vuong, Q., Loftin, R., and Hofmann, K. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*, 2019.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *ICLR*, 2017. URL http://arxiv.org/abs/1605.08803.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. RL^2: Fast reinforcement learning via slow reinforcement learning, 2017. URL https://openreview.net/forum?id=HkLXCE9lx.

Duff, M. O. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, Department of Computer Science, University of Massachusetts Amherst, 2002. AAI3039353.

Engel, Y., Mannor, S., and Meir, R. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 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.

Federer, H. *Geometric Measure Theory*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1969. ISBN 9783540045052. URL https://books.google.co.uk/books?id=QslkQgAACAAJ.

Fellows, M., Hartikainen, K., and Whiteson, S. Bayesian bellman operators. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2021.

Fellows, M., Smith, M., and Whiteson, S. Why target networks stabilise temporal difference methods. In *ICML*, 2023.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., and Legg, S. Noisy networks for exploration. In *Proceedings of the International Conference on Representation Learning (ICLR 2018)*, Vancouver (Canada), 2018.

Gal, Y. and Ghahramani, Z. 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, pp. 1050–1059. JMLR.org, 2016.

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in mdps with rkhs embeddings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1603–1610, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Guez, A., Silver, D., and Dayan, P. 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.

Hallak, A., Castro, D. D., and Mannor, S. Contextual markov decision processes. *ArXiv*, abs/1502.02259, 2015. URL https://api.semanticscholar.org/CorpusID:14616648.

Hausknecht, M. J. and Stone, P. Deep recurrent q-learning for partially observable mdps. *AAAI*, abs/1507.06527, 2015. URL https://aaai.org/papers/11673-deep-recurrent-q-learning-for-partially-observable-mdps/.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6626–6637. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2078–2087. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/huang18d.html.

Janz, D., Hron, J., Mazur, P., Hofmann, K., Hernández-Lobato, J. M., and Tschiatschek, S. Successor uncertainties: Exploration and uncertainty in temporal difference learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1–2):99–134, may 1998. ISSN 0004-3702.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002. doi: 10.1023/A:1017984413808. URL https://doi.org/10.1023/A:1017984413808.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Paper.pdf.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021. doi: 10.1109/TPAMI.2020.2992934.

Kuss, M. and Rasmussen, C. Gaussian processes in reinforcement learning. In Thrun, S., Saul, L., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/7993e11204b215b27694b6f139e34ce8-Paper.pdf.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6: 4–22, 1985.

Lipton, Z., Li, X., Gao, J., Li, L., Ahmed, F., and Deng, l. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *Association for the Advancement of Artificial Intelligence*, 11 2018.

Littman, M. L., Cassandra, A. R., and Kaelbling, L. P. Learning policies for partially observable environments: Scaling up. In Prieditis, A. and Russell, S. (eds.), *Machine Learning Proceedings 1995*, pp. 362–370. Morgan Kaufmann, San Francisco (CA), 1995. ISBN 978-1-55860-377-6. doi: https://doi.org/10.1016/B978-1-55860-377-6.50052-9. URL https://www.sciencedirect.com/science/article/pii/B9781558603776500529.

Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F., and Peters, J. Model-based uncertainty in value functions. *AISTATS 2023*, abs/2302.12526, 2023.

Martin, J. J. *Bayesian decision problems and Markov chains [by] J. J. Martin*. Wiley New York, 1967.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/mniha16.html.

Nielsen, D., Jaini, P., Hoogeboom, E., Winther, O., and Welling, M. Survae flows: Surjections to bridge the gap between vaes and flows. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9578a63fbe545bd82cc5bbe749636af1-Abstract.html.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty Bellman equation and exploration. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3839–3848, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/odonoghue18a.html.

Osband, I., Van Roy, B., and Russo, D. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3003–3011, Red Hook, NY, USA, 2013. Curran Associates Inc.

Osband, I., Aslanides, J., and Cassirer, A. Randomized prior functions for deep reinforcement learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8617–8629. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8080-randomized-prior-functions-for-deep-reinforcement-learning.pdf.

Osband, I., Roy, B. V., Russo, D. J., and Wen, Z. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1530–1538. JMLR, 2015.

Schlegel, M. K., Tkachuk, V., White, A. M., and White, M. Investigating action encodings in recurrent neural networks in reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=K6g4MbAC1r.

Schmitt, S., Shawe-Taylor, J., and van Hasselt, H. Exploration via epistemic value estimation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26164. URL https://doi.org/10.1609/aaai.v37i8.26164.

Strens, M. A bayesian framework for reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 943–950. ICML, 2000.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Thomson, W. R. 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. URL https://doi.org/10.1093/biomet/25.3-4.285.

Touati, A., Satija, H., Romoff, J., Pineau, J., and Vincent, P. Randomized value functions via multiplicative normalizing flows. In Globerson, A. and Silva, R. (eds.), *UAI*, pp. 156. AUAI Press, 2019.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.co.uk/books?id=hV8o5R7_5tkC.

Wasserman, L. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387251456.

Yao, H., Huang, L.-K., Zhang, L., Wei, Y., Tian, L., Zou, J., Huang, J., and Li, Z. . Improving generalization in meta-learning via task augmentation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11887–11897. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yao21b.html.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=Hkl9JlBYvr.

Zintgraf, L. M., Feng, L., Lu, C., Igl, M., Hartikainen, K., Hofmann, K., and Whiteson, S. Exploration in approximate hyper-state space for meta reinforcement learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12991–13001. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zintgraf21a.html.

A. Model-Based Approaches and their Shortcomings

Whilst POMDP solvers such as RL^2 (Duan et al., 2017) or meta-learning approaches (Beck et al., 2023) can be naively applied to the BAMDP, solving the BAMDP exactly is hopelessly intractable for all but the simplest of problems for several reasons: firstly, unless a conjugate model is used, the posterior over model parameters cannot be evaluated analytically as the posterior normalisation constant is intractable to evaluate and conjugate models are often too simple to be of practical value in RL; secondly, even if it is possible to obtain a posterior over the model parameters, solving the BAMDP requires evaluating high dimensional integrals by marginalising over model parameters; finally, finding a Bayes-optimal policy in the MDP requires solving a planning problem at each timestep for all possible histories. Few model-based BRL algorithms scale beyond small and discrete state-action spaces, even under approximation (Asmuth & Littman, 2011; Guez et al., 2013).

One notable exception is the VariBAD framework (Zintgraf et al., 2020), and subsequent related approaches (Yao et al., 2021; Zintgraf et al., 2021), which avoids the problem of intractability by being Bayesian over a small subset of model parameters. These methods sacrifice Bayes-optimally, relying on a frequentist heuristic to learn the non-Bayesian parameters; here, parametric models of the reward and state and transition distributions are introduced: $P_\theta(r|s,a,m)$ and $P_\theta(r|s,a,m)$ which are parametrised by $\theta \in \Theta$ and condition on the random variable m. A posterior over m is then inferred and used to obtain a marginal likelihood over trajectories:

$$p(\tau|\theta, \mathcal{D}) = \mathbb{E}_{m\sim p(m|\mathcal{D})}\left[p_\theta(\tau|m)\right],$$

which is optimised for θ. In this way, VariBAD is a partially Bayesian approach, inferring a posterior over m, but not parameters θ. The dimensionality of m can be kept relatively small to ensure tractability. As the VariBAD optimal policy is not obtained by marginalising over a space of MDPs nor is the uncertainty accounted for in all of the model's parameters, it is not Bayes-optimal in the limit of approximation.

Another issue with model-based approaches, especially those trained in a meta-learning context such as RL^2, is the assumption that the agent has access to a generative hypothesis space where it is possible to sample MDPs from a prior and collect rollouts of transitions from each sampled MDP. In real-world settings, knowing the exact hypothesis space is not a feasible assumption as it is generally not possible to specify transition dynamics a priori for all environments the agent can encounter.

B. Proofs

Theorem 3.2. *Let Assumption 3.1 hold, then* $B^+[Q_\omega](h_t, a_t) = B^\star[Q_\omega](h_t, a_t)$.

Proof. We start by analysing the transformation of variables $b_t = \beta_\omega(h_{t+1}) \coloneqq r_t + \gamma \max_{a'} Q_\omega(h_{t+1}, a')$. To ease notation, for a variable $x \in \mathbb{R}^d$ we write an integral with respect to the Lebesgue measure $\mathcal{L}^d$ as $\int f(x)d\mathcal{L}^d(x) = \int f(x)dx$. Under Assumption 3.1, the mapping $\beta_\omega(\cdot, h_t, a_t) : \mathbb{R} \times \mathcal{S} \to \mathbb{R}$ is also measurable, hence from the change of variables theorem under measurable mappings (see Bogachev (2007, Theorem 3.6.1)):

$$\begin{aligned}
\mathcal{B}^\star[Q_\omega](h_t, a_t) &= \mathbb{E}_{r_t, s_{t+1}\sim P_{R,S}(h_t,a_t)}\left[\beta_\omega(h_{t+1})\right] \\
&= \int \beta_\omega(h_{t+1}) p(r_t, s_{t+1}|h_t, a_t) d(r_t, s_{t+1}), \\
&= \int b_t p(b_t|h_t, a_t; \omega) db_t.
\end{aligned}$$

Marginalising over $b_0, b_1, ..b_{t-1}$ yields:

$$\begin{aligned}
\mathcal{B}^\star[Q_\omega](h_t, a_t) &= \int b_t p(b_0, b_1, \cdots b_t|h_t, a_t; \omega) d(b_0, b_1, \cdots b_t), \\
&= \int b_t p(b_t|b_0, b_1, \cdots b_{t-1}, h_t, a_t; \omega) dP(b_0, b_1, \cdots b_{t-1}|h_t; \omega) db_t,
\end{aligned}$$

Now, because each b_i is a deterministic transformation of variables $b_i = \beta_\omega(h_{i+1}) = r_i + \gamma \max_{a'} Q_\omega(h_{i+1}, a')$, the

distribution $dP(b_0, b_1, \cdots b_{t-1}|h_t; \omega) = d\delta(b_0 = \beta_\omega(h_1), b_1 = \beta_\omega(h_2), \cdots b_{t-1} = \beta_\omega(h_t))$, hence:

$$\begin{aligned}
\mathcal{B}^\star[Q_\omega](h_t, a_t) \int b_t p(b_t|b_0, b_1, \cdots b_{t-1}, h_t, a_t; \omega) d\delta(b_0 = \beta_\omega(h_1), b_1 = \beta_\omega(h_2), \cdots b_{t-1} = \beta_\omega(h_t)) db_t, \\
&= \int b_t p(b_t|b_0 = \beta_\omega(h_1) \cdots b_{t-1} = \beta_\omega(h_t), h_t, a_t; \omega) db_t, \\
&= \int b_t \int p(b_t|h_t, a_t, \phi; \omega) p(\phi|b_0 = \beta_\omega(h_1) \cdots b_{t-1} = \beta_\omega(h_t), h_t, a_t) d\phi db_t, \\
&= \int b_t \int p(b_t|h_t, a_t, \phi; \omega) p(\phi|\mathcal{D}_\omega(h_t)) d\phi db_t, \\
&= B^+[Q_\omega](h_t, a_t),
\end{aligned}$$

as required. □

Corollary 3.4. *Let Assumption 3.3 hold. Assume there exists a parametrisation $\omega^\star$ such that $\mathcal{B}^\star[Q_{\omega^\star}](h_t, a_t) = Q_{\omega^\star}(h_t, a_t)$. A Bayes-optimal policy under a parametric reward-state transition model with density $p(r_t, s_{t+1}|s_t, a_t, \phi)$ and prior P_Φ is equivalent to a Bayes-optimal policy under a optimal Bellman model with density:*

$$p(b_t|h_t, a_t, \phi; \omega) = \mathbb{E}_{r_t, s_{t+1} \sim p(r_t, s_{t+1}|s_t, a_t, \phi)} \left[\frac{\delta(r_t, s_{t+1} \in \beta_\omega^{-1}(b_t, h_t, a_t))}{|J_\beta(r_t, s_{t+1})|} \right],$$

with the same prior.

Proof. We verify our result by proving $\mathcal{B}^\star[Q_\omega](h_t, a_t) = B^+[Q_\omega](h_t, a_t)$ for $p(b_t|h_t, a_t, \phi; \omega)$ defined above, hence any agent following policy $a^\star \in \arg\max_{a'} B^+[Q_{\omega^\star}](h_t, a')$ will be taking actions $a^\star \in \arg\max_{a'} Q_{\omega^\star}(h_t, a_t) = Q^\star(h_t, a_t)$ and thus following a Bayes-optimal policy. We first prove:

$$\mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} \left[\beta_\omega(h_{t+1}) \right] = \mathbb{E}_{b_t \sim P_B(h_t, a_t, \phi; \omega)} \left[b_t \right].$$

Starting from the left hand side, as the $n+1$ dimensional Hausdorff and Lebesgue measures agree over $\mathbb{R}^{n+1}$, we write the expectation as a Lebesgue integral under the Hausdorff measure $\mathcal{H}^{n+1}$:

$$\begin{aligned}
&\mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} \left[\beta_\omega(h_{t+1}) \right] \\
&\qquad = \int_{\mathbb{R} \times \mathcal{S}} \beta_\omega(h_{t+1}) p(r_t, s_{t+1}|s_t, a_t, \phi) d\mathcal{H}^{n+1}(r_t, s_{t+1}), \\
&\qquad = \int_{\mathbb{R} \times \mathcal{S}} \frac{\beta_\omega(h_{t+1}) p(r_t, s_{t+1}|s_t, a_t, \phi)}{|J_\beta(r_t, s_{t+1})|} |J_\beta(r_t, s_{t+1})| d\mathcal{H}^{n+1}(r_t, s_{t+1}).
\end{aligned}$$

where $\mathcal{H}^n$ is the n-dimensional Hausdorff measure. Because of the Lipschitz assumption under Assumption 3.3, $|J_\beta(r_t, s_{t+1})|$ exists almost everywhere due to Rademacher's theorem (Villani, 2008, Theorem 14.25), hence under Assumption 3.3, we can apply the coarea formula to the transformation of variables $b_t = \beta_\omega(h_t, a_t, r_t, s_{t+1})$ (Federer, 1969, Theorem 3.2.12):

$$\begin{aligned}
&\mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} \left[\beta_\omega(h_{t+1}) \right] \\
&= \int_{\mathbb{R}} \int_{\beta_\omega^{-1}(b_t, h_t, a_t)} b_t \frac{p(r_t, s_{t+1}|s_t, a_t, \phi)}{|J_\beta(r_t, s_{t+1})|} d\mathcal{H}^n(r_t, s_{t+1}) d\mathcal{H}^1(b_t), \\
&= \int_{\mathbb{R}} b_t \int_{\beta_\omega^{-1}(b_t, h_t, a_t)} \frac{p(r_t, s_{t+1}|s_t, a_t, \phi)}{|J_\beta(r_t, s_{t+1})|} \delta(r_t, s_{t+1} \in \beta_\omega^{-1}(b_t, h_t, a_t)) d\mathcal{H}^n(r_t, s_{t+1}) d\mathcal{H}^1(b_t), \\
&= \int_{\mathbb{R}} b_t \mathbb{E}_{r_t, s_{t+1} \sim p(r_t, s_{t+1}|s_t, a_t, \phi)} \left[\frac{\delta(r_t, s_{t+1} \in \beta_\omega^{-1}(b_t, h_t, a_t))}{|J_\beta(r_t, s_{t+1})|} \right] d\mathcal{H}^1(b_t), \\
&= \mathbb{E}_{b_t \sim P_B(h_t, a_t, \phi; \omega)} \left[b_t \right],
\end{aligned}$$

where $P_B(h_t, a_t, \phi; \omega)$ has density:

$$p(b_t|h_t, a_t, \phi; \omega) = \mathbb{E}_{r_t, s_{t+1} \sim p(r_t, s_{t+1}|s_t, a_t, \phi)} \left[\frac{\delta(r_t, s_{t+1} \in \beta_\omega^{-1}(b_t, h_t, a_t))}{|J_\beta(r_t, s_{t+1})|} \right],$$

as required. Using Theorem 3.2 our result follows:

$$\begin{aligned}
\mathcal{B}^\star[Q_\omega](h_t, a_t) &= \mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(h_t, a_t)} \left[\beta_\omega(h_{t+1})\right] \\
&= \mathbb{E}_{\phi \sim P_\Phi(h_t)} \left[\mathbb{E}_{r_t, s_{t+1} \sim P_{R,S}(s_t, a_t, \phi)} \left[\beta_\omega(h_{t+1})\right]\right], \\
&= \mathbb{E}_{\phi \sim P_\Phi(h_t)} \left[\mathbb{E}_{b_t \sim P_B(h_t, a_t, \phi; \omega)} \left[b_t\right]\right], \\
&= \mathbb{E}_{\phi \sim P_\Phi(\mathcal{D}(h_t))} \left[\mathbb{E}_{b_t \sim P_B(h_t, a_t, \phi; \omega)} \left[b_t\right]\right], \\
&= B^+[Q_\omega](h_t, a_t).
\end{aligned}$$

□

Theorem 3.5. *Under Assumption 3.3,*

$$\Pi^\star_{\text{QBRL}} = \arg\max_{\pi \in \Pi_{QBRL}} J^\pi_{Bayes} = \left\{\mathbb{E}_{\phi \sim P_\Phi(h_t)} \left[\pi^\star(\cdot, \phi)\right] | \pi^\star(\cdot, \phi) \in \Pi^\star_\Phi\right\}.$$

Proof. We first show $\Pi^\star_{\text{QBRL}} = \arg\max_{\pi \in \Pi_{\text{QBRL}}} J^\pi_{\text{Bayes}}$. We use the QBRL assumption to derive the expected return:

$$\begin{aligned}
J^{\pi^\star}_{\text{QBRL}} &= \mathbb{E}_{s_0 \sim P_0} \left[V^\star_{\text{QBRL}}(s_0)\right], \\
&= \mathbb{E}_{s_0 \sim P_0} \left[\mathbb{E}_{\phi \sim P_\Phi} \left[V^\star(s_0, \phi)\right]\right], \\
&= \mathbb{E}_{s_0 \sim P_0, \phi \sim P_\Phi} \left[\mathbb{E}_{a_0 \sim \pi^\star(s_0, \phi)} \left[Q^\star(s_0, a', \phi)\right]\right].
\end{aligned}$$

Substituting for the definition of the optimal contextual Q-function:

$$\begin{aligned}
J^{\pi^\star}_{\text{QBRL}} &= \mathbb{E}_{\phi \sim P_\Phi, s_0 \sim P_0, a_0 \sim \pi^\star(s_0, \phi)} \left[\mathbb{E}_{\tau_\infty \sim P^{\pi^\star}_\infty(s_0, a_0, \phi)} \left[\sum_{t=0}^{\infty} \gamma^t r_t\right]\right], \\
&= \mathbb{E}_{\phi \sim P_\Phi} \left[\mathbb{E}_{\tau_\infty \sim P^{\pi^\star}_\infty(\phi)} \left[\sum_{t=0}^{\infty} \gamma^t r_t\right]\right], \\
&= \mathbb{E}_{\phi \sim P_\Phi} \left[J^{\pi^\star}(\phi)\right], \\
&= \mathbb{E}_{\phi \sim P_\Phi} \left[\max_{\pi \in \Pi_\Phi} J^\pi(\phi)\right].
\end{aligned}$$

We now prove $\Pi^\star_\Phi = \arg\max_{\pi \in \Pi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right]$ P_Φ-almost everywhere by contradiction. We first show:

$$\Pi^\star_\Phi \subseteq \arg\max_{\pi \in \Pi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right].$$

Assume that there exists some $\pi^\dagger(\cdot, \phi) \in \Pi^\star_\Phi \notin \arg\max_{\pi \in \Phi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right]$. There then exists some set $\Phi' \subseteq \Phi$ with non-zero measure according to P_Φ such that:

$$J^{\pi^\dagger}(\phi) < \max_\pi J^\pi(\phi),$$

for $\phi \in \Phi'$, which is a contradiction as $\pi^\dagger(\cdot, \phi) \in \arg\max_\pi J^\pi(\phi)$. We are left to show:

$$\Pi^\star_\Phi \supseteq \arg\max_{\pi \in \Pi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right].$$

Assume that there exists some $\pi^\dagger(\cdot, \phi) \in \arg\max_{\pi \in \Phi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right] \notin \Pi^\star_\Phi$. There then exists some set $\Phi' \subseteq \Phi$ with non-zero measure according to P_Φ such that:

$$J^{\pi^\dagger}(\phi) < J^{\pi^\star}(\phi),$$

for any $\phi \in \Phi'$ and $\pi^\star \in \Pi^\star_\Phi$. Taking expectations it follows:

$$\mathbb{E}_{\phi \sim P_\Phi} \left[J^{\pi^\dagger}(\phi)\right] < \mathbb{E}_{\phi \sim P_\Phi} \left[J^{\pi^\star}(\phi)\right] \leq \max_{\pi \in \Pi_\Phi} \mathbb{E}_{\phi \sim P_\Phi} \left[J^\pi(\phi)\right],$$

implying $\pi^\dagger(\cdot,\phi) \notin \arg\max_{\pi\in\Phi_\Phi} \mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)]$, which is a contradiction. As we have proved $\Pi^\star_\Phi = \arg\max_{\pi\in\Pi_\Phi} \mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)]$, we are left to show $\arg\max_{\pi\in\Pi_\Phi} \mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)] = \arg\max_{\pi\in\Pi_{\text{QBRL}}} J^\pi_{\text{Bayes}}$.

$$\begin{aligned}\mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)] =& \mathbb{E}_{\phi\sim P_\Phi}\left[\mathbb{E}_{\tau_\infty\sim P^\pi_\infty(\phi)}\left[\sum_{t=0}^\infty r_t\right]\right],\\ =& \mathbb{E}_{\tau_\infty\sim P^\pi_{\text{QBRL}}}\left[\sum_{t=0}^\infty r_t\right],\end{aligned}$$

where $P^\pi_{\text{QBRL}} = \mathbb{E}_{\phi\sim P_\Phi}[P^\pi_\infty(\phi)]$. As $\pi(\phi, s_t)$ is context-conditioned, we can marginalise over all contexts using the posterior to obtain the predictive history-conditioned QBRL policy: $\pi_{\text{QBRL}}(h_t) = \mathbb{E}_{\phi\sim P_\Phi(h_t)}[\pi(\phi, s_t)]$. It follows that $P^\pi_{\text{QBRL}} = P^{\pi_{\text{QBRL}}}_\infty$ where $P^{\pi_{\text{QBRL}}}_\infty$ is the Bayesian predictive distribution over h_∞ using context-conditioned policies with density:

$$p^{\pi_{\text{QBRL}}}(h_\infty) = p_0(s_0)\prod_{i=0}^\infty \pi_{\text{QBRL}}(a_t|h_t)p(r_t, s_{t+1}|h_t),$$

hence:

$$\begin{aligned}\mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)] = \mathbb{E}_{\tau_\infty\sim P^{\pi_{\text{QBRL}}}_\infty}\left[\sum_{t=0}^\infty r_t\right] = J^{\pi_{\text{QBRL}}}_{\text{Bayes}},\\ \implies \arg\max_{\pi\in\Pi_\Phi} \mathbb{E}_{\phi\sim P_\Phi}[J^\pi(\phi)] = \arg\max_{\pi\in\Pi_\Phi} J^{\pi_{\text{QBRL}}}_{\text{Bayes}}.\end{aligned}$$

Finally, as each $\pi(\cdot,\phi)\in\Pi_\phi$ indexes a $\pi_{\text{QBRL}}(\cdot,\phi)\in\Pi_{\text{QBRL}}$,

$$\max_{\pi\in\Pi_\Phi} J^{\pi_{\text{QBRL}}}_{\text{Bayes}} = \max_{\pi\in\Pi_{\text{QBRL}}} J^\pi_{\text{Bayes}} \implies \Pi^\star_{\text{QBRL}} = \arg\max_{\pi\in\Pi_{\text{QBRL}}} J^\pi_{\text{Bayes}},$$

where $\Pi^\star_{\text{QBRL}} = \left\{\mathbb{E}_{\phi\sim P_\Phi(h_t)}[\pi^\star(\cdot,\phi)]\,|\pi^\star(\cdot,\phi)\in\Pi^\star_\Phi\right\}$. □

Corollary 3.6. *There exist MDPs with priors such that $\Pi^\star_{QBRL}\cap\Pi^\star_{Bayes} = \varnothing$.*

Proof. We consider the tiger problem as a counter example (Kaelbling et al., 1998) with $\gamma = 0.9$, $r_{\text{tiger}} = -500$, $r_{\text{gold}} = 10$ and $r_{\text{listen}} = -1$. Details of the space of MDPs can be found in Appendix D.1. We index the MDP with the tiger in the left door as $\phi =$ tiger left and the tiger in the right door as $\phi =$ tiger right. Consider the uniform prior over MDPs $P(\phi = \text{tiger left}) = P(\phi = \text{tiger right}) = 0.5$. As agents always start in state s_0, it suffices to find the optimal MDP conditioned policies in s_0:

$$\pi^\star(s_0, \phi = \text{tiger left}) = \delta(a = \text{open right}), \quad \pi^\star(s_0, \phi = \text{tiger right}) = \delta(a = \text{open left})$$

From Theorem 3.5, it follows that the optimal Bayesian contextual policy is a mixture of these two policies using the prior:

$$\pi^\star(s_0) = 0.5(\delta(a = \text{open right}) + \delta(a = \text{open left})).$$

The optimal Q-function for the optimal MDP-conditioned policies are

$$\begin{aligned}Q^\star(s_0, a = \text{open right}, \phi = \text{tiger left}) = Q^\star(s_0, a = \text{open left}, \phi = \text{tiger right}) = \frac{r_{\text{gold}}}{1-\gamma} = 100,\\ Q^\star(s_0, a = \text{open right}, \phi = \text{tiger right}) = Q^\star(s_0, a = \text{open left}, \phi = \text{tiger left})\\ = r_{\text{tiger}} + \frac{\gamma r_{\text{gold}}}{1-\gamma} = -155.\end{aligned}$$

Using Theorem 3.5, we can find the Bayesian value function for the optimal contextual policy:

$$\begin{aligned}&Q^\star_{\text{QBRL}}(s_0, a = \text{open right})\\ &\quad= 0.5\left(Q^\star(s_0, a = \text{open right}, \phi = \text{tiger left}) + Q^\star(s_0, a = \text{open right}, \phi = \text{tiger right})\right),\\ &\quad= -27.5,\\ &Q^\star_{\text{QBRL}}(s_0, a = \text{open left})\\ &\quad= 0.5\left(Q^\star(s_0, a = \text{open left}, \phi = \text{tiger left}) + Q^\star(s_0, a = \text{open left}, \phi = \text{tiger right})\right),\\ &\quad= -27.5,\end{aligned}$$

from which the Bayesian return for the optimal contextual policy follows:

$$\begin{aligned} J^{\pi^\star_{\text{QBRL}}}_{\text{Bayes}} &= \mathbb{E}_{s\sim\delta(s_0)}\left[\mathbb{E}_{a\sim\pi^\star_{\text{QBRL}}(s)}\left[Q^\star_{\text{QBRL}}(s,a)\right]\right], \\ &= 0.5\left(Q^\star_{\text{QBRL}}(s_0, a = \text{open left}) + Q^\star_{\text{QBRL}}(s_0, a = \text{open right})\right), \\ &= -27.5. \end{aligned}$$

Now consider the policy that always listens $\pi^\dagger = \delta(a = \text{listen})$. The Bayesian return for this policy is:

$$J^{\pi^\dagger}_{\text{Bayes}} = \frac{r_{\text{listen}}}{1-\gamma} = -10,$$

hence:

$$\begin{aligned} J^{\pi^\star_{\text{QBRL}}}_{\text{Bayes}} < J^{\pi^\dagger}_{\text{Bayes}} \leq \max_{\pi\in\Pi_{\mathcal{H}}} J^{\pi}_{\text{Bayes}} = J^{\pi^\star}_{\text{Bayes}}, \\ \implies \Pi^\star_{\text{QBRL}} \cap \Pi^\star_{\text{Bayes}} = \varnothing, \end{aligned}$$

as required. □

C. Network Details

Our recurrent Q network architecture is shown Figure 5. We encode the prior history-action pair via a recurrent variable $\hat{h}_t \coloneqq \{h_t, a_t\}$. At each timestep our network inputs $\hat{h}_{t-1}$ and a tuple of observations $o_t \coloneqq \{r_{t-1}, s_t, a_t\}$ and outputs the value variable $q_t = Q_\omega(h_t, a_t) = Q_\omega(\hat{h}_{t-1}, o_t)$ and the new recurrent encoding $\hat{h}_t$.

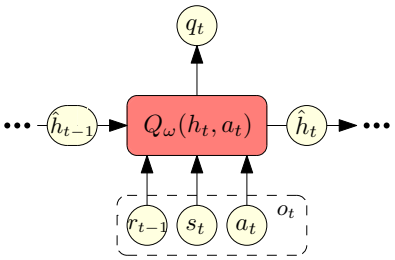


Figure 5. Recurrent Q Network

C.1. Aleatoric Network

As discussed in Section 4.1, to model the distribution over optimal Bellman operators $P_B(h_t, a_t, \phi; \omega)$ we introduce a base variable $z_{\text{al}} \in \mathbb{R}$ with a tractable distribution P_{al}; in this paper we use a zero-mean, unit variance Gaussian $\mathcal{N}(0, 1)$. We then generate b_t using a change of variables $b_t = B(z_{\text{al}}, q_t, \phi)$ parameterised by $\phi \in \Phi$, where $B(z_{\text{al}}, q_t, \phi)$ is a mapping in z_{al} with inverse $z_{\text{al}} = B^{-1}(b_t, q_t, \phi)$. Under this change of variables, $\mathbb{E}_{b_t \sim P_B(h_t, a_t, \phi; \omega)}[f(b_t)] = \mathbb{E}_{z_{\text{al}} \sim P_{\text{al}}}[f \circ B(z_{\text{al}}, q_t, \phi)]$ for any integrable $f : \mathbb{R} \to \mathbb{R}$. We refer to $B(z_{\text{al}}, q_t, \phi)$ as the *aleatoric* network as it characterises the aleatoric uncertainty in the MDP and its expressiveness implicitly determines the space of MDPs that our model can represent.

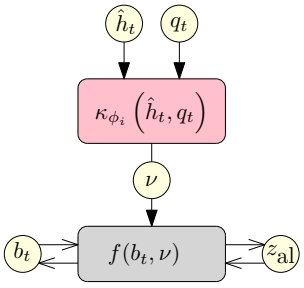


Figure 6. Details of Aleatoric Flow

We define our aleatoric flow by adapting the autoregressive flow (Kingma et al., 2016). We take inputs from the RNN Q-function approximator outputs $\hat{h}_t, q_t$ (including the history encoding) and pass them through a conditioner $\kappa_{\phi_i}(\hat{h}_t, q_t)$, which is a feed forward neural network parametrised by ϕ_i where $\phi_i \subset \phi$. The output of the conditioner is a vector that defines the parameters for the coupling function. We use an inverse autoregressive flow, with $z_{\text{al}} \in \mathbb{R}^2$ followed with a dimensionality reduction layer to reduce the dimension of the output to 1 and abs layers, as detailed in Nielsen et al. (2020). Since only $b_t = B(q_t, z_{\text{al}}, \phi)$ needs to be bijective in z_{al}, there are also no restrictions on $Q_\omega(h_t, a_t)$, allowing us to use any arbitrary RNN. The aleatoric network then consists of L coupling functions in composition:

$$B(z_{\text{al}}, q_t, \phi) = f_L \circ f_{L-1} \circ \cdots f_2 \circ f_1(z_{\text{al}}, \hat{h}_t, q_t, \phi).$$

C.2. Epistemic Network

Given the aleatoric network $B(z_{\text{al}}, q_t, \phi)$, dataset of bootstrapped samples $\mathcal{D}_\omega(h_t) \coloneqq \{b_i\}_{i=0}^{t-1}$ and prior over parameters P_Φ, our goal is to infer the posterior $P_\Phi(\mathcal{D}_\omega(h_t))$ to obtain the predictive mean:

$$\hat{B}[Q_\omega](h_t, a_t) \coloneqq \mathbb{E}_{\phi \sim P_\Phi(\mathcal{D}_\omega(h_t))}\left[\mathbb{E}_{z_{\text{al}} \sim P_{\text{al}}}[B(q_t, z_{\text{al}}, \phi)]\right].$$

Unfortunately, inferring the posterior and carrying out marginalisation exactly is intractable for all but the simplest aleatoric networks, which would not have sufficient capacity to represent a complex target distribution $P_B(h_t, a_t, \phi; \omega)$. We instead look to variational inference using a normalising flow to learn a tractable approximation.

Like in Appendix C.1, we introduce a base variable $z_{\text{ep}} \in \mathbb{R}^d$ with a tractable distribution P_{ep}; again we use a zero-mean Gaussian $\mathcal{N}(0, I^d)$. We then make a transformation of variables $\phi = t_\psi(z_{\text{ep}})$ where $t_\psi : \mathbb{R}^d \to \mathbb{R}^d$ is a bijective mapping parametrised by $\psi \in \Psi$ with inverse $z_{\text{ep}} = t_\psi^{-1}(\phi)$. We refer to $t_\psi(z_{\text{ep}})$ as the epistemic network as it characterises the epistemic uncertainty in ϕ. From the change of variables formula, it follows that the resulting variational distribution P_ψ has a density $p_\psi(\phi) = |\det(J_\psi(\phi))|\, p_{\text{ep}} \circ t_\psi^{-1}(\phi)$ where $J_\psi(\phi) \coloneqq \nabla_\phi t_\psi^{-1}(\phi)$ is the Jacobian of the inverse mapping. Using variational inference, we treat P_ψ as an approximation of the true posterior $P_\Phi(h_t)$, which we learn by minimising the KL-divergence between the two distributions $\text{KL}(P_\psi \parallel P_\Phi(h_t))$. This is equivalent to minimising the following negative evidence lower-bound (ELBO) objective with respect to ψ:

$$\text{ELBO}(\psi; h_t, \omega) \coloneqq \mathbb{E}_{z_{\text{ep}} \sim P_{\text{ep}}}\left[\left(\sum_{i=0}^{t-1}\left(B^{-1}(b_i, q_i, \phi)^2 - \log\left|\partial_b B^{-1}(b_i, q_i, \phi)\right|\right) - \log p_\Phi(\phi)\right)\Bigg|_{\phi = t_\psi(z_{\text{ep}})}\right].$$

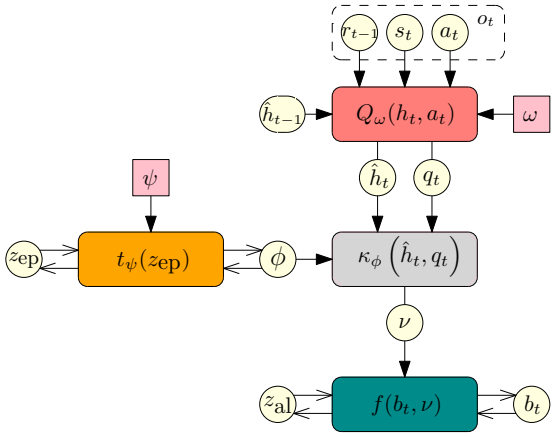


Figure 7. Schematic of BEN

We provide a fully connected schematic of BEN with flow details in Figure 7.

To derive this result, we start with the definition of the KL-divergence $\mathrm{KL}(P_\psi \parallel P_\Phi(h_t))$, using Bayes' rule to re-write the log-posterior:

$$\begin{aligned}\mathrm{KL}(P_\psi \parallel P_\Phi(h_t)) \coloneqq& \mathbb{E}_{\phi\sim P_\psi(\phi)}\left[\log p_\psi(\phi) - \log p_\Phi(\phi|h_t)\right], \\ =& \mathbb{E}_{\phi\sim P_\psi(\phi)}\left[\log p_\psi(\phi) - \log p_\Phi(h_t|\phi) - \log p_\Phi(\phi) + \log p(h_t)\right], \\ =& \mathbb{E}_{\phi\sim P_\psi(\phi)}\left[\log p_\psi(\phi) - \log p_\Phi(h_t|\phi) - \log p_\Phi(\phi)\right].\end{aligned}$$

As the log-evidence $\log p(h_t)$ has no dependence on ψ, we can omit it from our objective, instead maximising the ELBO:

$$\mathrm{ELBO}(\psi; h_t, \omega) = \mathbb{E}_{\phi\sim P_\psi(\phi)}\left[\log p_\Phi(h_t|\phi) + \log p_\Phi(\phi) - \log p_\psi(\phi)\right].$$

We write the expectation with respect to P_ψ using the transformation of variables:

$$\begin{aligned}\mathrm{ELBO}(\psi; h_t, \omega) &= \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\log p_\Phi(h_t|t_\psi(z_{\text{ep}}) + \log p_\Phi(t_\psi(z_{\text{ep}}) - \log p_\psi(t_\psi(z_{\text{ep}})\right], \\ &= \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\log p_\Phi(h_t|t_\psi(z_{\text{ep}})\right] + \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\log p_\Phi(t_\psi(z_{\text{ep}}) - \log p_\psi(t_\psi(z_{\text{ep}})\right].\end{aligned}$$

Now, by the definition of $p_\psi(\phi)$, it follows:

$$\begin{aligned}p_\psi \circ t_\psi(z_{\text{ep}}) &= \left|\det\left(J_\psi \circ t_\psi(z_{\text{ep}})\right)\right| p_{\text{ep}} \circ t_\psi^{-1} \circ t_\psi(z_{\text{ep}}), \\ &= \left|\det\left(\nabla_\phi t_\psi^{-1} \circ t_\psi(z_{\text{ep}})\right)\right| p_{\text{ep}}(z_{\text{ep}}), \\ &= p_{\text{ep}}(z_{\text{ep}}),\end{aligned}$$

hence as $p_{\text{ep}}(z_{\text{ep}})$ has no dependence on ψ, we can omit it from the objective, yielding:

$$\begin{aligned}\mathrm{ELBO}(\psi; h_t, \omega) &= \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\left(\log p_\Phi(h_t|\phi) + \log p_\Phi(\phi)\right)|_{\phi=t_\psi(z_{\text{ep}})}\right], \\ &= \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\left(\log\left(\prod_{i=0}^{t-1} p_B(b_i|h_i, a_i, \phi; \omega)\right) + \log p_\Phi(\phi)\right)\Big|_{\phi=t_\psi(z_{\text{ep}})}\right], \\ &= \mathbb{E}_{z_{\text{ep}}\sim P_{\text{ep}}(z_{\text{ep}})}\left[\left(\sum_{i=0}^{t-1} \log p_B(b_i|h_i, a_i, \phi; \omega) + \log p_\Phi(\phi)\right)\Big|_{\phi=t_\psi(z_{\text{ep}})}\right].\end{aligned}$$

Finally we can derive the exact form of the log-density $p_B(b_i|h_i, a_i, \phi; \omega)$ using the change of variables formula under $b_t = B(z_{\text{al}}, q_t, \phi)$:

$$\begin{aligned}\log p_B(b_i|h_i, a_i, \phi; \omega) &= \log\left(\exp(-B^{-1}(b_i, q_i, \phi)^2)\left|\partial_b B^{-1}(b_i, q_i, \phi)\right|\right), \\ &= -B^{-1}(b_i, q_i, \phi)^2 + \log\left|\partial_b B^{-1}(b_i, q_i, \phi)\right|.\end{aligned}$$

Substituting and multiplying by -1, thus changing to an objective to minimise rather than maximise, yields our desired result:

$$\text{ELBO}(\psi; h_t, \omega) \coloneqq \mathbb{E}_{z_{\text{ep}} \sim P_{\text{ep}}} \left[\left(\sum_{i=0}^{t-1} \left(B^{-1}(b_i, q_i, \phi)^2 - \log \left| \partial_b B^{-1}(b_i, q_i, \phi) \right| \right) - \log p_\Phi(\phi) \right) \Big|_{\phi = t_\psi(z_{\text{ep}})} \right].$$

C.3. Network Training

Prior Initialisation Before any actions have been taken, we can minimise the MSBBE using the initial state s_0 and the prior P_ϕ, which we assume is tractable to obtain samples from. This has the advantage of initialising the Q-function approximator to incorporate any prior domain knowledge we have about the MDP, in addition to ensuring that an optimal Bayesian Bellman equation is approximately satisfied before training starts. Once the posterior is updated using a new observation, we shouldn't expect the solution to the MSBBE to change significantly to reflect the updated belief. Finally, there may be prior knowledge about state and reward transitions that are available to us a priori that we would like to encode in the Q-function approximator. If an agent in state s taking action a always transitions according to a known conditional distribution $P_{R,S}(s, a, \phi)$, then we can use this information to solve the Bayesian Bellman equation conditioned on s, a. We combine all such state-action pairs into a dataset $\mathcal{D}_{\text{prior}} \coloneqq \{s_i, a_i\}_{i=1}^{K_{\text{prior}}}$, for which we minimise the MSSBE:

Algorithm 2 PRIORINITIALISATION($P_\Phi, s_0, \mathcal{D}_{\text{prior}}$)

Initialise ω
for N_{Pretrain} steps **do**
  Sample action $a \sim \rho$
  Sample two MDPs $\phi, \phi' \sim P_\Phi$
  Sample aleatoric variables $z_{\text{al}}, z'_{\text{al}} \sim P_{\text{al}}$
  $q_0 = Q_\omega(s_0, a)$
  $\omega \leftarrow \omega - \alpha(B(q_0, z_{\text{al}}, \phi) - q_0)\nabla_\omega(B(q_0, z'_{\text{al}}, \phi') - q_0)$
  Sample from prior data $s, a \sim \mathcal{D}_{\text{prior}}$
  Sample two MDPs $\phi, \phi' \sim P_\Phi$
  Sample rewards-state transition $r, s_+ \sim P_{R,S}(s, a, \phi)$
  Sample rewards-state transition $r', s'_+ \sim P_{R,S}(s, a, \phi')$
  $\omega \leftarrow \omega - \alpha(r + \gamma \max_{a' \in \mathcal{A}} Q_\omega(s, a, r, s_+, a') - Q_\omega(s, a))\nabla_\omega(\gamma \max_{a' \in \mathcal{A}} Q_\omega(s, a, r', s'_+, a') - Q_\omega(s, a))$
end for

$$\mathcal{L}(\omega; \mathcal{D}_{\text{prior}}) = \sum_{i=1}^{K_{\text{prior}}} \left(\mathbb{E}_{\phi \sim P_\Phi} \left[\mathbb{E}_{r_i, s_{i+1} \sim P_{R,S}(s_i, a_i, \phi)} \left[r + \gamma \max_{a' \in \mathcal{A}} Q_\omega(s_i, a_i, r_i, s_{i+1}, a') \right] \right] - Q_\omega(s_i, a_i) \right)^2.$$

We give specific details of $\mathcal{D}_{\text{prior}}$ in the context of our search and rescue environment in Appendix D.4. Both MSBBE objectives can be minimised using stochastic gradient descent with two independent samples from the prior to avoid bias in our updates, as outlined in Algorithm 2. Note that for domains where we don't have such knowledge, we can take $\mathcal{D}_{\text{prior}} = \varnothing$ and ignore the minimisation steps on $\mathcal{L}(\omega; \mathcal{D}_{\text{prior}})$. BEN's incorporation of prior knowledge does not require a full generative model of the environment dynamics and demonstrations can be from simulated or related MDPs that do not exactly match the set of tasks the agent is in at test time.

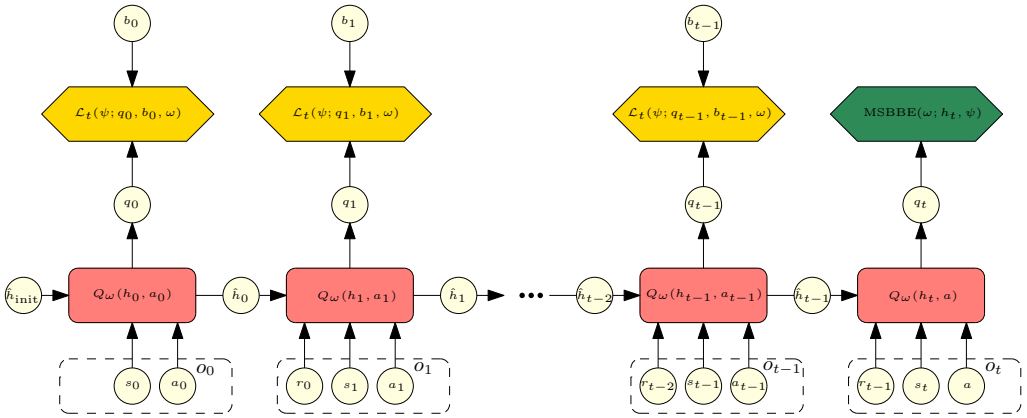


Figure 8. Schematic of BEN Training Regime. Losses are shown as hexagons.

Posterior Updating To obtain an efficient algorithm, we note that the ELBO objective can be written as a summation: $\text{ELBO}(\psi; h_t, \omega) = \sum_{i=0}^{t-1} \mathcal{L}_t(\psi; q_i, b_i, \omega)$, where each sub-objective is:

$$\mathcal{L}_t(\psi; q_i, b_i, \omega) \coloneqq \mathbb{E}_{z_{\text{ep}} \sim P_{\text{ep}}} \left[\left(B^{-1}(b_i, q_i, \phi) \right)^2 - \log \left| \partial_b B^{-1}(b_i, q_i, \phi) \right| - \frac{1}{t} \log p_\Phi(\phi) \Big|_{\phi = t_\psi(z_{\text{ep}})} \right].$$

As shown in Figure 8, we can minimise $\text{ELBO}(\psi; h_t, \omega)$ by unrolling the RNN, starting at $i = 0$. After each timestep, we obtain q_i, which can be used to minimise the loss $\mathcal{L}_t(\psi; q_i, b_i, \omega)$ with the observation b_i whilst keeping ω fixed. Once the network has been unrolled to the timestep t, we can use the output to minimise the MSBBE. Like in for our prior initialisation algorithm in Algorithm 2, it is important that we sample two independent samples $\phi, \phi' \sim P_\phi(h_t)$ from our approximate posterior when minimising the MSBBE to avoid biased gradient estimates. Once t becomes too large, we can truncate the sequences to length t', starting at state $s_{t-t'}$ instead of s_0. Like when target parameters used to stabilise frequentist TD methods (Fellows et al., 2023), this updating ensures that the Q-network is updated on an asymptotically slower timescale to the posterior parameters, and we tune the length of truncation t' for the sequence and stepsizes α_ψ and α_ω to ensure stability.

D. Experiments

D.1. Tiger Problem

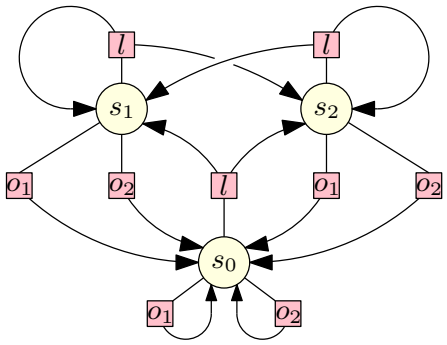


Figure 9. Tiger Problem MDP

Algorithm 3 POSTERIORUPDATING(h_t, ψ, ω)

for N_{Update} steps **do**
  $\hat{h} \leftarrow \hat{h}_{\text{init}}$
  $o \leftarrow \{s_0, a_0\}$
  for $i \in [0 : t-1]$ **do**
    $\hat{h}, q_i \leftarrow Q_\omega(\hat{h}, o)$
    $b_i \leftarrow r_i + \gamma \max_{a'} Q_\omega(\hat{h}, o, r_i, s_{i+1}, a')$
    $g_\psi \sim \nabla_\psi \mathcal{L}_t(\psi; q_i, b_i, \omega)$
    $\psi \leftarrow \psi - \alpha_\psi g_\psi$
    $o \leftarrow r_i, s_{i+1}, a_{i+1}$
  end for
  $q_t \leftarrow Q_\omega(\hat{h}, o)$
  $\beta_\omega \sim \nabla_\omega \text{MSBBE}(\omega; q_t, \psi)$
  $\omega \leftarrow \omega - \alpha_\omega \beta_\omega$
end for

The aim of this empirical evaluation is to verify our claim that BEN can learn a Bayes-optimal policy and compare BEN to existing model-free approaches. We evaluate BEN in the counterexample tiger problem domain from Corollary 3.6, which allows for comparison against a true Bayes-optimal policy. We show our tiger problem MDP in Figure 9. The agent is always initialised in state s_0 and can chose to open door 1 (o_1), open door 2 (o_2) or to listen (l): $\mathcal{A} \coloneqq \{o_1, o_2, l\}$. There are two possible MDPs the agent can be in, with the tiger assigned to either door 1 or door 2 randomly and the gold to the other door. If the agent chooses o_1, door 1 is opened and the agent receives a reward of $r_{\text{tiger}} = -500$ if the tiger is behind the door or $r_{\text{gold}} = 10$ if the gold is behind the door. The agent always transitions to state s_0 after selecting o_1 or o_2. If the agent chooses to listen, it receives a small negative reward of $r_{\text{listen}} = -1$ and if the tiger is behind door 1 transitions to state s_1 with probability 0.85 and state s_2 with probability 0.1, or if the tiger is behind door 2, the agent transitions to state s_2 with probability 0.85 and state s_1 with probability 0.1.

D.2. Tiger Problem Implementation Details

We initialise the agent with a uniform prior over the two MDPs. The posterior for this problem is tractable so we use that in place of the epistemic network:

$$P_\Phi(\phi = \text{tiger in 1}|h_t) = \frac{0.85^{N_1} \cdot 0.1^{N_2}}{0.85^{N_1} \cdot 0.1^{N_2} + 0.1^{N_1} \cdot 0.85^{N_2}},$$
$$P_\Phi(\phi = \text{tiger in 2}|h_t) = 1 - P_\Phi(\phi = \text{tiger in 1})$$

where N_1 is the number of visitations to state s_1 and N_2 is the number of visitations to state s_2. If the agent opens the door, the posterior trivially becomes:

$$P_\Phi(\phi|h_t) = \begin{cases} \delta(\phi = \text{tiger in 1}), & a_{t-1} = o_1, r_{t-1} = r_{\text{tiger}}, \\ \delta(\phi = \text{tiger in 1}), & a_{t-1} = o_2, r_{t-1} = r_{\text{gold}}, \\ \delta(\phi = \text{tiger in 2}), & a_{t-1} = o_2, r_{t-1} = r_{\text{tiger}}, \\ \delta(\phi = \text{tiger in 2}), & a_{t-1} = o_1, r_{t-1} = r_{\text{gold}}. \end{cases}$$

The aleatoric network can be handcoded as the pushforward of known transition distributions. We vary the number of steps for the MSBBE minimisation with a learning rate of 0.02 using ADAM for the stochastic gradient descent. For the for Q-function approximator, we use a fully connected linear layer with ReLU activations, a gated recurrent unit and a final fully connected linear layer with ReLU activations. All hidden dimensions are 32. The dimension of $\hat{h}_0$ is 2. The input dimension is 1 + 2 =3 and the network output is 3 dimensional to reflect the three possible actions the agent can take.

D.3. Tiger Problem Results

We initialise all agents in a tabula rasa setting with a uniform prior over MDPs and plot the median returns after each timestep in Figure 10 for 11 timesteps, averaged over MDPs, each drawn uniformly. We plot the performance of BEN for a varying number of SGD minimisation steps on our MSBBE objective. Figure 10 shows that by increasing the number of SGD minimisation steps, BEN's performance approaches that of the Bayes-optimal oracle and the variance in the policies decreases, with near Bayes-optimal performance attainted using 20 minimisation steps. We also compare BEN to an oracle that is optimal over the space contextual policies, $\Pi^\star_{\text{QBRL}}$, which is an optimal policy for existing model-free approaches. As expected, the contextual optimal policy is limited to a mixture of optimal policies conditioned on ϕ, hence the performance is comparatively poor: median returns are significantly lower than BEN as contextual policies sample an initial action uniformly before acting optimally once the true MDP is revealed.

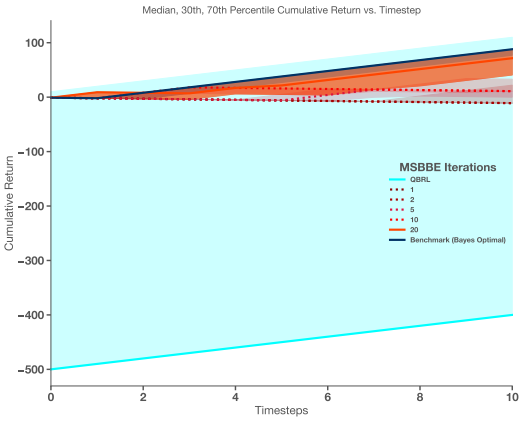


Figure 10. Results of evaluation in tiger problem showing BEN with increasing minimisation steps on MSBBE vs Bayes-optimal and contextual oracles

D.4. Search and Rescue Problem

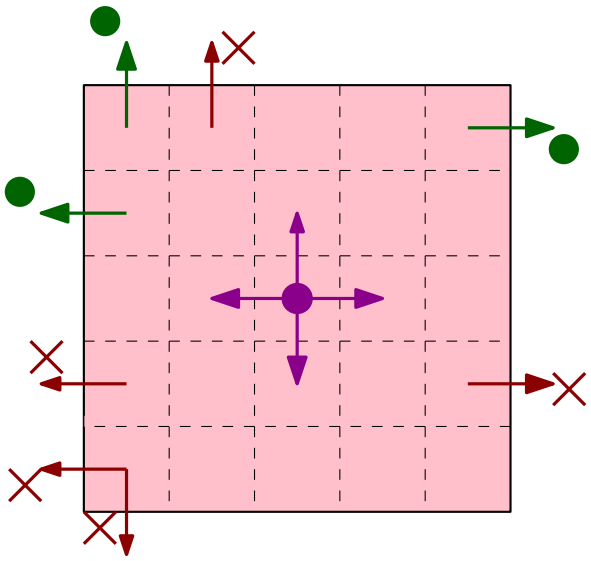


Figure 11. 5×5 Search and Rescue Problem MDP with 5 Hazards (red crossed) and 3 Victims (green circles). Agent (purple circle) is shown in s_0. Green actions yield reward r_{rescue} and red actions yield reward r_{hazard}.

We now present a novel search and rescue MDP designed to present a challenging extension to the toy tiger problem domain. An agent is tasked with rescuing N_{victims} victims from a dangerous situation whilst avoiding any one of N_{hazards} hazards. The agent's action space is $\mathcal{A} = \{\text{up}, \text{down}, \text{left}, \text{right}, \text{listen}\}$. The agent can move in an $N_{\text{grid}} \times N_{\text{grid}}$ gridworld where N_{grid} is an odd number and transitions one square deterministically in the direction of the action taken. If the agent selects an action that would take it off the grid, it remains put and opens the door adjacent to its square in the direction of the action. If the agent opens a door with a victim behind, it receives a reward of $r_{\text{victim}} = 10$ and the victim is removed from the MDP. If the agent . If the agent opens a door with a hazard behind, it receives a reward of $r_{\text{hazard}} = -100$ and the hazard remains in the MDP. We show an example MDP in Figure 11.

The agent is initialised in position $(0, 0)$, which is the central square of the grid. The agent observes state $s \in \mathcal{S} \subset \mathbb{R}^{2+N_{\text{victims}}+N_{\text{hazards}}}$ where $l_{\text{agent}} \coloneqq (s_0, s_1)$ is the agent's location relative to $(0, 0)$. The agent does not directly observe which doors have hazards or victims behind. If the agent chooses the action listen, their location remains put and they transition to a new state s' where s'_i for each $i \in \{2 : 1 + N_{\text{victims}} + N_{\text{hazards}}\}$ is given by:

$$s'_i = \exp\left(-\frac{\|l_{\text{agent}} - l_i\|^2}{N_{\text{grid}}} + \eta\right), \quad \eta \sim \mathcal{N}(0; \sigma^2_{\text{noise}})$$

which a noisy variable correlated to the distance between the agent and each victim/hazard l_i. The victim locations are $\{l_i\}_{i \in \{2:N_{\text{victims}}+1\}}$ and the hazard locations are $\{l_i\}_{i \in \{N_{\text{victims}}+2:1+N_{\text{victims}}+N_{\text{hazard}}\}}$. For each MDP, the victims and hazards are randomly assigned a square each adjacent to the grid and the initialised uniformly across that square. If an agent opens a door with a victim, their location becomes $(N_{\text{grid}} \cdot 1000, N_{\text{grid}} \cdot 1000)$ and no further reward can be obtained for that victim. Agents receive a small negative reward for listening $r_{\text{listen}} = -1$ and no reward for traversing the grid.

D.5. Exploiting Prior Knowledge

For the search and rescue environment, there is domain knowledge that we can use to form $\mathcal{D}_{\text{prior}}$ that is common to all MDPs. The first example of this knowledge is that movement transitions are deterministic and yield no reward when the agent is traversing the grid. To make this precise, we define the set of states in the interior of the grid:

$$\mathcal{S}_{\text{interior}} \coloneqq \left\{ s \middle| |s_0| < \frac{N_{\text{grid}} - 1}{2}, |s_1| < \frac{N_{\text{grid}} - 1}{2}, \{s_i\}_{i \geq 2} = 0 \right\},$$

that is their location is not adjacent to the grid's boundary. All other values $s_2 : s_{1+N_{\text{victims}}+N_{\text{hazards}}}$ are set to 0. We define the set of movement actions to be $\mathcal{A}_{\text{movement}} \coloneqq \{up, down, left, right\}$. Taking an action $a \in \mathcal{A}_{\text{movement}}$ when in state $s \in \mathcal{S}_{\text{interior}}$ always moves the agent in the direction of the action selected without changing the other states and receives a reward of 0, that is:

$$\begin{aligned}
P_S(s \in \mathcal{S}_{\text{interior}}, a = up) &= \delta(s'_1 = s_1 + 1, s'_{\neq 1} = s_{\neq 1}),\\
P_S(s \in \mathcal{S}_{\text{interior}}, a = down) &= \delta(s'_1 = s_1 - 1, s'_{\neq 1} = s_{\neq 1}),\\
P_S(s \in \mathcal{S}_{\text{interior}}, a = right) &= \delta(s'_0 = s_0 + 1, s'_{\neq 0} = s_{\neq 0}),\\
P_S(s \in \mathcal{S}_{\text{interior}}, a = left) &= \delta(s'_0 = s_0 - 1, s'_{\neq 0} = s_{\neq 0}),\\
P_R(s \in \mathcal{S}_{\text{interior}}, a \in \mathcal{A}_{\text{movement}}) &= \delta(r = 0),
\end{aligned}$$

allowing us to sample from $\mathcal{A}_{\text{movement}} \times \mathcal{S}_{\text{interior}} \subset \mathcal{D}_{\text{prior}}$ and apply the above transformation.

In addition to the deterministic transitions, we can also include prior reward information. Firstly we define the boundary states where the agent is adjacent to the edge of the grid

$$\mathcal{S}_{\text{boundary}} \coloneqq \left\{ s || s_0 | = \frac{N_{\text{grid}} - 1}{2}, \{s_i\}_{i \geq 2} = 0 \right\} \cup \left\{ s || | s_1 | = \frac{N_{\text{grid}} - 1}{2}, \{s_i\}_{i \geq 2} = 0 \right\}$$

We note that again all non-locations states $s_2 : s_{1+N_{\text{victims}}+N_{\text{hazards}}}$ are set to 0 because other values are specific to each MDP. As agents and hazards are initialised uniformly in the squares adjacent to the grid, if an agent is in $\mathcal{S}_{\text{boundary}}$ and takes an action to move out of the grid (i.e. open a door), then the expected reward will be:

$$r_{\text{prior}} \coloneqq \frac{N_{\text{victims}}}{4 \times N_{\text{grid}}} \cdot r_{\text{victim}} + \frac{N_{\text{hazards}}}{4 \times N_{\text{grid}}} \cdot r_{\text{hazard}}.$$

Listening Information Key to solving the search and rescue environment is learning to listen before acting.

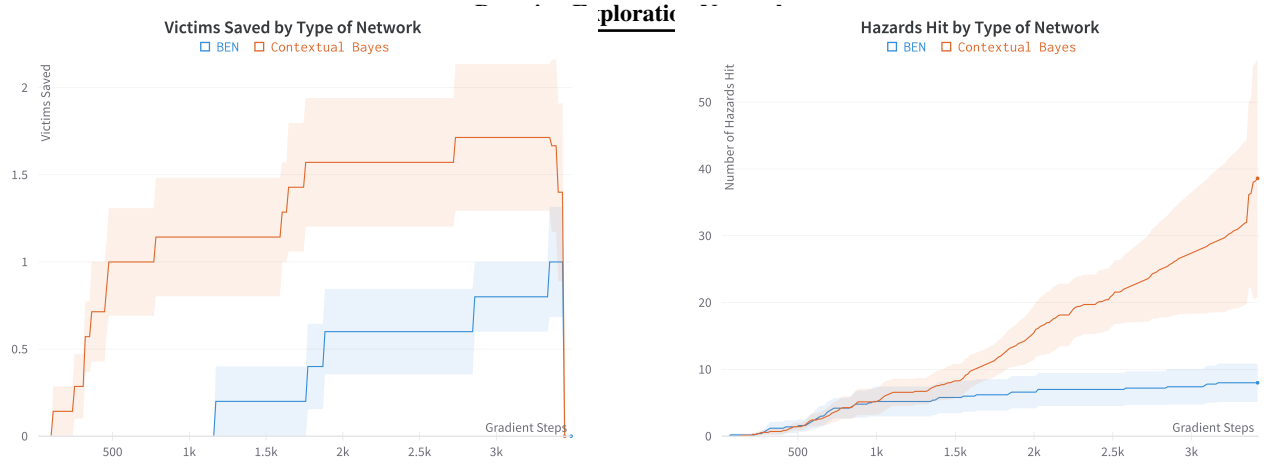
D.6. Search and Rescue Implementation Details

The epistemic network consists of two layers of ActNorm, a Masked Autoregressive Flow with two blocks and a LU linear decomposition and permutation as in Dinh et al. (2017). The base distribution is a unit Gaussian. This takes the number of parameters in the Aleatoric Network is a projection to 2d space from 1d, an Inverse Autoregressive Flow, a LU Linear decomposition and Permutation, a projection back to 1d with the Slice Flow, and an Abs Flow. The base distribution is a standard 1d Gaussian. The AbsFlow consists of 6 applications of the conditioner network, (K=6), and two layers. We vary the number of steps for the MSBBE minimisation with a learning rate of 1e-4 using ADAM for the stochastic gradient descent and use a separate ADAM optimiser, with a learning rate of 1e-4 for the Epistemic Network training on the ELBO. For the for Q-function approximation, we use a fully connected linear layer with ReLU activations, a gated recurrent unit and a final fully connected linear layer with ReLU activations. All hidden dimensions are 64. The dimension of the hidden state $\hat{h}_0$ is 64. The input size is the state space size 14 (4 number of victims + 8 number of hazards + 1 + 1 for x and y dims) The input dimension is state space size + 1 for reward + 1 for action = 16. The network output is 5 dimensional to reflect the five possible actions the agent can take. The input to the conditioner network is number of aleatoric parameters+ hidden dim size + 1 for q value, and the hidden layers and output size are the number of aleatoric parameters. Only a subset of these aleatoric parameters are used as needed in each layer and the rest are dropped.

D.7. Ablations

We carry out the following ablations in the zero-shot setting for the search and rescue environment, averaged over 7 seeds in this zero-shot test and plot the sample standard errors:

QBRL Approaches We repeat the ablation carried out in the Tiger Problem for this new domain, demonstrating the existing approaches that learn a contextual optimal policy (i.e. state of the art model-free approaches such as BBAC (Fellows et al., 2021) and BootDQN+Prior (Osband et al., 2019)) cannot succeed in this challenging setting. This corresponds to using a function approximator with no capacity to represent history. BEN provides a clear improvement over these existing methods in terms of cumulative return in Figure 4. To understand why, for each approach we plot the number of victims



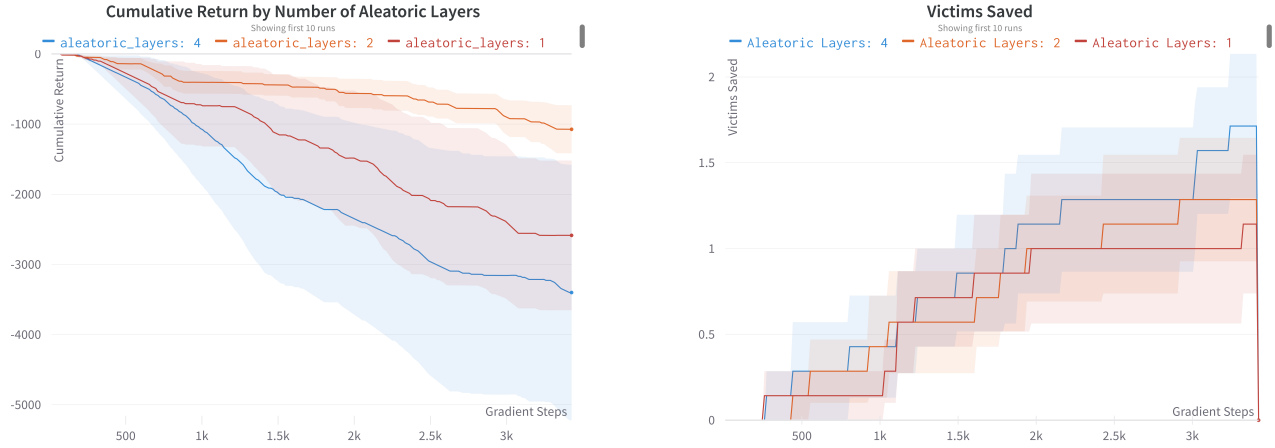
(a) Victims Saved Contextual Ablation

(b) Hazards Hit Contextual Ablation

Figure 12. Contextual vs BEN Ablation

rescued in Figure 12a and hazards hit in Figure 12b. Although both approaches save a similar number of victims, the contextual approach hits an order of 10 times more hazards than BEN. These results demonstrate that QBRL approaches struggle to solve this problem whereas BEN is slightly more conservative, yet does not hit nearly as many hazards, as we would expect given the disproportionally greater negative reward for hitting a hazard than rescuing the victim in our environment.

Capacity for Representing Aleatoric Uncertainty We now investigate how reducing increasing the capacity of our aleatoric network affects the performance in this domain. We increase the number of aleatoric flow layers from 1 to 4 and plot the returns in Figure 13a and the number of victims rescued in Figure 13b. We see that for this environment, 2 flow

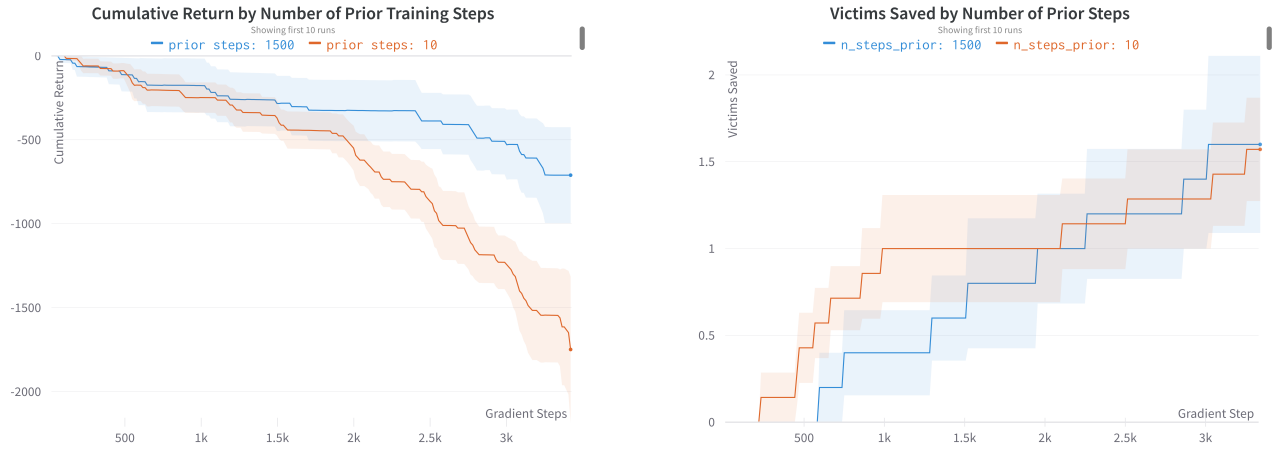


(a) Return Aleatoric Ablation

(b) Victims Saved Aleatoric Ablation

Figure 13. Aleatoric Network Ablation

layers yields the best returns. As the number of aleatoric flow layers is determines the hypothesis space, our results provide evidence that there exists a trade-off between specifying a rich enough hypothesis space and a hypothesis space that is too general for the problem setting. For 4 layers, the hypothesis space is too general to learn how to behave optimally given the number of minimisation steps whereas for 1 layer, the agent cannot represent aleatoric uncertainty sufficiently to learn a policy that is useful for the environment. This ablation also supports our central claim that aleatoric uncertainty cannot be neglected in model-free Bayesian approaches.



(a) Return Prior Ablation

(b) Victims Saved Prior Ablation

Figure 14. Aleatoric Network Ablation

Incorporation of Prior Knowledge Finally, we investigate how our prior training regime affects the performance of BEN, varying the number of prior gradient training steps according to Algorithm 2. Results are plotted in Appendix D.7. A key motivation for taking a Bayesian approach to RL is the ability to formally exploit prior knowledge. We use this ablation to demonstrate how knowledge provided by simple simulations can be incorporated into BEN's pre-training regime. As we decrease the number of prior pretraining MSBBE minimisation steps, we see that performance degrades in the zero-shot settling as expected. Moreover, this ablation shows that a relatively few number of pre-training steps are needed to achieve impressive performance once the agent is deployed in an unknown MDP, supporting our central claim that BEN is computationally efficient.