Finite-state Offline Reinforcement Learning with Moment-based Bayesian Epistemic and Aleatoric Uncertainties

Filippo Valdettaro
Department of Computing
Imperial College London
filippo.valdettaro20@imperial.ac.uk

Aldo A. Faisal
Department of Computing
Imperial College London

Abstract

Reinforcement learning (RL) agents can learn complex sequential decision-making and control strategies, often above human expert performance levels. In real-world deployment, it becomes essential from a risk, safety-critical, and human interaction perspective for agents to communicate the degree of confidence or uncertainty they have in the outcomes of their actions and account for it in their decision-making. We assemble here a complete pipeline for modelling uncertainty in the finite, discrete-state setting of offline RL. First, we use methods from Bayesian RL to capture the posterior uncertainty in environment model parameters given the available data. Next, we determine exact values for the return distribution's standard deviation, taken as the measure of uncertainty, for given samples from the environment posterior (without requiring quantile-based or similar approximations of conventional distributional RL) to more efficiently decompose the agent's uncertainty into epistemic and aleatoric uncertainties compared to previous approaches. This allows us to build an RL agent that quantifies both types of uncertainty and utilises its epistemic uncertainty belief to inform its optimal policy through a novel stochastic gradient-based optimisation process. We illustrate the improved uncertainty quantification and Bayesian value optimisation performance of our agent in simple, interpretable gridworlds and confirm its scalability by applying it to a clinical decision support system (AI Clinician) which makes real-time recommendations for sepsis treatment in intensive care units, and address the limitations that arise with inference for larger-scale MDPs by proposing a sparse, conservative dynamics model.

1 Introduction

In safety-critical machine learning applications, accurately quantifying confidence and uncertainty in decision outcomes becomes imperative for regulatory and trust resons [7, 24]. In general, uncertainties that such systems face can be epistemic, arising from limited data availability, or aleatoric, originating from inherent environmental randomness. Uncertainty quantification is particularly relevant in Reinforcement Learning (RL) systems as uncertainty in decisions compounds in sequential decision-making. The task of disentangling these uncertainties gains importance in real-world decision-making scenarios that are either discrete state in nature or arise where continuous environmental variables are clustered into a finite number of discrete states [26, 13]. In such cases, aleatoric uncertainty encapsulates the uncertainty introduced by this clustering process. While this may lead to information loss and poor scaling with high-dimensional state spaces, it increases the tractability of the resulting environment. In this work, we provide a thorough analysis of uncertainty in these finite-state environments. Utilising Bayesian RL, with distributional elements, we account for epistemic

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

uncertainty via a Bayesian dynamics model with exact inference, assigning posterior probabilities to potential environments [12]. Aleatoric uncertainty is quantified by analytically solving linear equations for higher return distribution moments [35]. We then combine these uncertainties to derive overall aleatoric and epistemic standard deviations. We compare the computational complexity and accuracy of our method with prior work. We propose a novel stochastic gradient-based method for policy optimisation that accounts for model dynamics uncertainty. We empirically demonstrate its superior optimisation performance and scalability over previous methods [11], providing results on gridworlds with varying offline dataset sizes. Our methods find application in clinical decision support systems (CDSS), which leverage vast patient data sets to train RL algorithms for treatment suggestions [17, 30]. We analyse a setup used for sepsis treatment [26], where patients' condition and treatment options were clustered into finite states and actions, originally tackled by applying dynamic programming methods [5]. We enhance this approach with uncertainty quantification and uncertainty-aware control. We investigate the scalability of our methods in such practical environments and address additional challenges, especially in constructing a meaningful dynamics prior, tackled by including domain-specific conservatism in the dynamics model.

2 Related Work

This section reviews uncertainty treatment in offline RL. We focus on epistemic uncertainty in Robust and Adaptive MDP settings, aleatoric uncertainty for risk-averse policy suggestion, and recent work quantifying both types of uncertainty.

Robust and Adaptive MDPs. A simple model-based approach for an MDP uses relative visitation frequencies as the ground truth transition probabilities. This can introduce bias and result in policies that generalise poorly [31, 37, 6]. To address this, a Bayesian approach is often employed to account for uncertainty in ambiguous transition dynamics, a common method in Bayesian RL [16]. Bayesian dynamics models used in Bayes-Adaptive MDPs (BAMDPs) [12, 18] maintain the current belief in transition dynamics and enable optimal 'offline' planning of adaptable 'online' policy rollouts. However, these models may be intractable beyond simple MDPs [33, 29, 40].

In high-risk offline settings, exploration is undesirable. For instance, in the clinical decision support system suggested in [26], novel actions are avoided by only selecting actions above a minimum visitation threshold. Therefore, we focus on optimal *memoryless*, stationary (non-adaptive) policies depending only on the state [10]. Finding such policies that are robust to the worst-case realisation of uncertain dynamics can often lead to overly conservative policies, making average value optimization across a distribution of MDPs a better alternative [32, 21, 38]. The work in [11] provides a method to find such policies, but scalability is not addressed. We propose using stochastic gradient-based policy value function optimization to overcome this limitation.

Risk-averse policies. Accounting for inherent environmental stochasticity is often desirable. Using the distributional RL framework [3], policies are often informed by return distribution properties other than its mean to select risk-averse actions [9, 8]. However, optimal policies for such statistical functionals are generally neither memoryless nor time-consistent [35, 4]. Therefore, we focus on using the mean of the return distribution to guide the agent's policy.

Aleatoric and Epistemic Uncertainty in RL for Healthcare. Several recent efforts have tried to model both types of uncertainties. In healthcare, [23] used a Bayesian dynamics model and Monte Carlo trajectory sampling to model uncertainties and determine when to defer treatment. In contrast, [14] trained an ensemble of distributional deep neural networks (DNNs) to learn the return distribution, effectively learning a 'distribution over distributions' of the return. Our work aims to improve uncertainty representation by replacing DNNs with exact dynamic programming methods and substituting the epistemic uncertainty from DNN parameter disagreement with the epistemic uncertainty due to uncertainty in transition dynamics.

3 Background

Dynamic Programming A Markov Decision Process $\mathcal{M}$ (MDP) [34] is defined by a tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \rho)$, where $\mathcal{S}$ and $\mathcal{A}$ are the (assumed finite) state and action spaces respectively, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ the transition kernel, $\gamma \in [0, 1]$ a discount factor and ρ the distribution over initial states. Given a policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, the *return* of an episode starting from state s is a random variable given by $G^\pi(s) = \sum_{t=0}^{\infty} \gamma^t R_t$, where $R_t = R(s_t, a_t)$, $a_t \sim \pi(\cdot|s_t)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1})$ given that $s_0 = s$. For our purposes we will be taking reward R as deterministic given state, with reward upon departing state s given by $r(s)$ and therefore independent of policy. This is a natural modelling step for MDPs where a certain state is associated with a particular reward, which in practice is common when constructing MDPs. When π is deterministic at state s, we denote the certain action as $\pi(s)$.

The expected value of G is called the value function $V^\pi(s) = \mathbb{E}G^\pi(s)$, and it can be shown that with this definition, V satisfies the Bellman equation

$$V^\pi(s) = r(s) + \gamma \sum_{a,s'} P(s'|s,a)\pi(a|s)V^\pi(s'). \tag{1}$$

Dynamic programming methods, such as value iteration, can evaluate V and provide the policy that optimises V [36]. It can be shown that the value of any arbitrary policy is

$$\mathbf{v}(\pi) = (\mathbf{I} - \gamma \mathbf{T}(\pi))^{-1}\mathbf{r}, \tag{2}$$

with $\mathbf{v}$ and $\mathbf{r}$ $|\mathcal{S}|$-dimensional vectors with i^{th} element being $V^\pi(s_i)$ and $r(s_i)$ respectively (for s the i^{th} state in $\mathcal{S}$) and $\mathbf{T}(\pi)$ the policy-dependent transition matrix with element i, j given by

$$\mathbf{T}_{i,j} = \sum_a \pi(a|s_i)P(s_j|s_i, a).$$

For clarity we have highlighted here the dependence of $\mathbf{T}$ on π and note that $\mathbf{r}$ does not depend on policy as we are working with state-dependent rewards.

Return Distribution The most common approach to analysing the return distribution, referred to as distributional RL, involves applying distributional Bellman operators [3] which, in the finite-state setting, compute the return distribution arbitrarily accurately for a given MDP (assuming a sufficiently expressive parametrisation) [4]. However, we pursue a different path to usual distributional RL, as for our purposes we only require the first two moments of the return distribution. These can be determined exactly in closed-form for a given MDP without resorting to the full distributional RL framework.

Methods analogous to those developed to evaluate the value of policies by solving the Bellman value equation (Eq. 1) can be extended to determine more general properties of the return distribution $G^\pi(s)$. For example, it can be shown that the variances of the return distribution satisfy an analogous set of linear Bellman equations, with solution given in vector form by [35]:

$$\mathbf{var}(\pi) = (\mathbf{I} - \gamma^2 \mathbf{T}(\pi))^{-1}\mathbf{r}^{(\text{var})}(\pi), \tag{3}$$

where the vector of variances $\mathbf{var}$ has element i corresponding to the variance at state s_i and $\mathbf{r}^{(\text{var})}$ is the vector with element i being

$$\mathbf{r}_i^{(\text{var})}(\pi) = \sum_j P^\pi(s_j|s_i)(r(s_i) + \gamma V^\pi(s_j))^2 - V^\pi(s_i)^2,$$

where $P^\pi(s'|s) = \sum_a \pi(a|s)P(s'|s,a)$.

Bayesian Dynamics Model The dynamics model we employ is standard in Bayesian RL, and is equivalent to the one used in BAMDPs [16, 33] with an unchanging belief and similar to the one proposed in [23], but stationary. By modelling the belief over dynamics parameters of the MDP, this line of work effectively captures the uncertainty due to not being able to narrow down what the true underlying MDP is: with a finite number of transitions, there may be a number of potential MDPs that may have generated the observations, to which we can assign posterior probabilities by using Bayes' rule. For our purposes, we take the reward function of the MDP as known (and deterministic), ultimately because in our applications we will define reward directly as a deterministic function of state, but treat the dynamics of the world as unknown.

Let $\theta_{s,a}^{s'}$ be a parameter representing the probability of transitioning to state s' given action a at state s, and consider a dataset of observed transitions $(s, a, r, s') \in \mathcal{D}$. The likelihood of observing a transition from s, a to s' is thus $p(s'|s,a) = \theta_{s,a}^{s'}$. Next, we specify a conjugate Dirichlet prior on θ, so that for each state-action the resulting posterior probability is also Dirichlet. Assuming a symmetric Dirichlet prior with parameter α_p, the posterior distribution satisfies

$$p(\{\theta_{s,a}^{s_i} | s_i \in \mathcal{S}\} | \mathcal{D}) \propto \prod_j (\theta_{s,a}^{s_j})^{n_j + \alpha_p - 1}, \tag{4}$$

with n_s being the number of times s, a transitioned to state s' and the proportionality constant is given (in closed form) by the multivariate Beta function [27].

When the number of possible outcomes, in this case next states, is large then inference on the Dirichlet parameters can be very data-inefficient when a generic maximum-entropy prior parameter is employed and assigns a disproportionate amount of posterior probability to unobserved outcomes. To mitigate this, one may scale the prior parameter inversely to the number of outcomes, as done in a BAMDP context in [18], or induce sparsity in the possible outcomes by modelling the belief of feasible next states through a hierarchical Bayesian model [15]. The approach we propose is to use a *conservative* dynamics model: in our applications, we have a very clearly defined failure (death) state, so we model the possible next state outcomes for each state-action as being just the observed ones and the failure state. The number of possible outcomes modelled will therefore be much smaller than the full range of outcomes and a maximum-entropy prior can usefully be employed. We compare the effect such a modelling choice has on the posterior values and learned policies to a symmetric prior chosen through Bayesian model selection.

Aleatoric and Epistemic Uncertainty In order to quantify and distinguish between epistemic uncertainty due to ambiguity in MDPs $\mathcal{M}$ given limited data and aleatoric uncertainty in the return G, we use the common decomposition formula that arises after applying the law of total variance [24, 23] to the return G:

$$\operatorname{Var} G(s) = \underbrace{\operatorname{Var}_{\mathcal{M}} \mathbb{E}\, G_{\mathcal{M}}(s)}_{\text{epistemic}} + \underbrace{\mathbb{E}_{\mathcal{M}} \operatorname{Var} G_{\mathcal{M}}(s)}_{\text{aleatoric}}, \tag{5}$$

where we have made clear that the dependence on the return random variable G is conditioned on the MDPs $\mathcal{M}$. The epistemic variance term captures the overall variance in the expected returns due to ambiguity in the MDPs and the aleatoric variance term is an estimate of the intrinsic variance averaged over the posterior MDP distribution. Equations 2 and 3 allow us to determine $\mathbb{E}\, G_{\mathcal{M}}(s) = V(s)$ and $\operatorname{Var} G_{\mathcal{M}}(s)$ exactly, while averages and variances over the MDPs can be approximated through Monte Carlo sampling of the posterior over MDPs. In the limit of infinite data, the epistemic variance should tend to 0 as the probability mass of the posterior focuses in on a specific $\mathcal{M}$ (see Appendix A for a brief discussion), but the aleatoric term won't in general behave similarly.

4 Methods

4.1 Uncertainty quantification

Existing approaches for estimating aleatoric and epistemic uncertainty in discrete-space MDPs either overlook uncertainty in the transition model [14] or rely on extensive Monte Carlo sampling [23]. As a consequence, the former does not scale consistently with additional data (see Appendix B for empirical evidence for this claim) and we can introduce improvements in the latter for the infinite-horizon MDP case by using closed-form expressions for the first two moments of the return distribution.

We present in Algorithm 1 a way to estimate the value, aleatoric and epistemic variances in Eq. 5. Its computational complexity scales as $O(|\mathcal{S}|^3)$ per dynamics sample due to requiring an $|\mathcal{S}| \times |\mathcal{S}|$ matrix inversion for each of the N_M dynamics samples. In contrast, methods that rely on Monte Carlo return samples to estimate aleatoric and epistemic return will require a larger number of Dirichlet samples and large simulation trajectory lengths to achieve comparable accuracy, but no matrix inversion. We investigate this trade-off quantitatively in Appendix C and conclude that the larger number of samples required for a full Monte Carlo-style evaluation (similar to [23]) is not worth the additional sampling overhead for the MDPs we are considering ($|\mathcal{S}| < 1000$). Note in principle one could also use some iterative policy evaluation scheme [36] to solve for the first and second moments of the return distribution, in so doing sacrificing accuracy to avoid calculating a matrix inverse.

Algorithm 1 Bayesian Value, Epistemic and Aleatoric Uncertainty Evaluation

```
Require: Policy $\pi$, state $s_i$, posterior distribution over transition parameters $p(\mathcal{M}|\mathcal{D})$
  $\theta^{s'}_{sa\{1:N_M\}} \leftarrow N_M$ transition matrix samples from $p(\mathcal{M}|\mathcal{D})$
  $\forall s, s' \in S\, \{\mathbf{T}_{ss'}\}_{\{1:N_M\}} \leftarrow \sum_a \pi(a|s)\theta^{s'}_{sa}$   ▷ $\mathbf{T}_t$ is the $t^{\text{th}}$ action-marginalised transition matrix
  for $t = 1$ to $N_M$ do
    $\mathbf{v}_T \leftarrow (\mathbf{I} - \gamma \mathbf{T}_t(\pi))^{-1}\mathbf{r}$   ▷ Eq. 2 for sampled dynamics
    $\forall s_k \in \mathcal{S}, \mathbf{r}^{(\text{var})}_k(\pi) \leftarrow \sum_j P^{\pi}_t(s_j|s_k)(r(s_k) + \gamma V^{\pi}(s_j))^2 - V^{\pi}(s_k)^2$
    $\mathbf{var}_T \leftarrow (\mathbf{I} - \gamma^2 \mathbf{T}_t(\pi))^{-1}\mathbf{r}^{(\text{var})}(\pi)$   ▷ Equation 3
    $v_t \leftarrow$ element $i$ of $\mathbf{v}_T$
    $var_t \leftarrow$ element $i$ of $\mathbf{var}_T$
  end for
  $bayes\_value \leftarrow \frac{1}{N_M}\sum_{t=1}^{N_M} v_t$
  $aleatoric\_var \leftarrow \frac{1}{N_M}\sum_{t=1}^{N_M} var_t$
  $epistemic\_var \leftarrow \frac{1}{N_M - 1}\sum_{t=1}^{N_M}(v_t - bayes\_value)^2$
      return $bayes\_value, aleatoric\_var, epistemic\_var$
```

4.2 Policy improvement

Beyond evaluating uncertainty, having a belief over the possible range of dynamics that an MDP can exhibit can allow us to account for this uncertain belief when carrying out control. Thus, we seek to find a policy that maximises the value objective under the Bayesian dynamics posterior belief

$$\max_{\pi} \sum_{s} \rho(s)\mathbb{E}_{\mathcal{M}\sim p(\cdot|\mathcal{D})} V^{\pi}_{\mathcal{M}}(s), \tag{6}$$

where the Bayesian value of each state $\mathbb{E}_{\mathcal{M}\sim p(\cdot|\mathcal{D})} V^{\pi}_{\mathcal{M}}(s)$ has been marginalised with respect to the initial state distribution ρ. Algorithm 2 shows the gradient-based approach we suggest to optimise this objective. In contrast to other methods ([26], [11]) this does not introduce bias due to a finite number of transition samples: by re-sampling from the posterior every gradient step, we remove the bias that would occur by picking a smaller finite sample, and standard stochastic gradient optimisation guarantees ensure that we converge to a local optimum.

Algorithm 2 Stochastic Gradient Policy Optimisation

```
Require: Initial deterministic $\pi$, posterior distribution over transition parameters $p(\mathcal{M}|\mathcal{D})$, initial
    policy softness parameter $\eta$, learning rate $\alpha$
  $\forall s \in \mathcal{S}, a \in \mathcal{A}\, z_{sa} \leftarrow \log(\eta/(|\mathcal{A}| - 1))$
  $\forall s \in \mathcal{S}\, z_{s\pi(s)} \leftarrow \log(1 - \eta)$   ▷ Set initial policy parametrisation
  while not converged do
    $\forall s \in \mathcal{S}, a \in \mathcal{A}\, \pi(a|s) \leftarrow \frac{\exp(z_{sa})}{\sum'_a \exp(z_{sa'})}$
    $\theta^{s'}_{sa\{1:n\}} \leftarrow n$ minibatch samples from $p(\mathcal{M}|\mathcal{D})$
    $\forall s, s' \in S\, \{\mathbf{T}_{ss'}\}_{\{1:n\}} \leftarrow \sum_a \pi(a|s)\theta^{s'}_{sa}$   ▷ $\mathbf{T}_t$ is the $t^{\text{th}}$ action-marginalised transition matrix
    $\forall i\, \mathbf{v}_i \leftarrow (\mathbf{I} - \gamma \mathbf{T}_i(\pi))^{-1}\mathbf{r}$   ▷ Eq. 2 for sampled dynamics
    $\mathcal{L} = -\sum_i \rho \cdot \mathbf{v_i}$   ▷ Marginalise over MDP posterior and initial state distribution
    $\forall s \in \mathcal{S}, a \in \mathcal{A}\, z_{sa} \leftarrow z_{sa} - \alpha\frac{\partial \mathcal{L}}{\partial z_{sa}}$   ▷ Policy parameters step towards improving value
  end while
  $\forall s \in \mathcal{S}, a \in \mathcal{A}\, \pi(a|s) \leftarrow \frac{\exp(z_{sa})}{\sum'_a \exp(z_{sa'})}$
      return $\pi$
```

5 Results

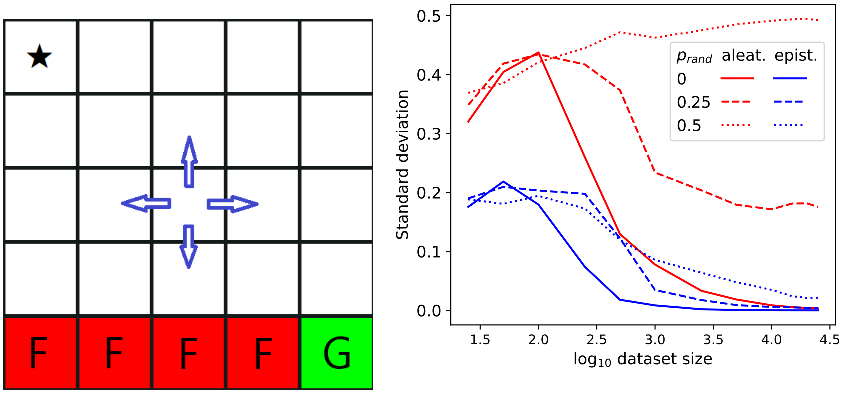
Here we present some illustrative results on gridworld environments as well as on a clinical dataset. The gridworld experiments demonstrate the salient features of our methods in the case where a ground-truth MDP can be easily investigated and modified, while the application to clinical data confirms its applicability to MDPs with practical use. We first examine uncertainty evaluation for a specific policy and then consider policy improvement. We then apply the same methods to the MIMIC-III dataset [22], and present results on the impact that carrying out Bayesian policy improvement has on this dataset.

5.1 Gridworld

We consider a gridworld with stochastic transitions: at each step there is a probability p_{rand} of being pushed down regardless of action taken. Otherwise, the agent moves up, down, left or right by one square determined by the action. The observed transitions dataset $\mathcal{D}$ is generated by repeatedly spawning an agent in a non-terminal random state and carrying out a random action. Experiments are ran on the gridworld visualised in Fig. 1a. The results presented here are for datasets that are proper subsets of any one of the larger datasets to ensure that the latter are strictly more informative.

Uncertainty Quantification We first highlight the main differences compared to recent methods that have been suggested to quantify aleatoric and epistemic uncertainty. To focus on this particular feature, we consider the policy *evaluation* problem, comparing how results from our Bayesian approach differ from others when evaluating the uncertainty for the policy that is optimal under the MLE dynamics parameter estimates. We see that our uncertainty quantification in Algorithm 1 scales consistently with varying dataset size (epistemic uncertainty always becomes small) and intrinsic stochasticity (higher p_{rand} corresponds to higher aleatoric uncertainty).

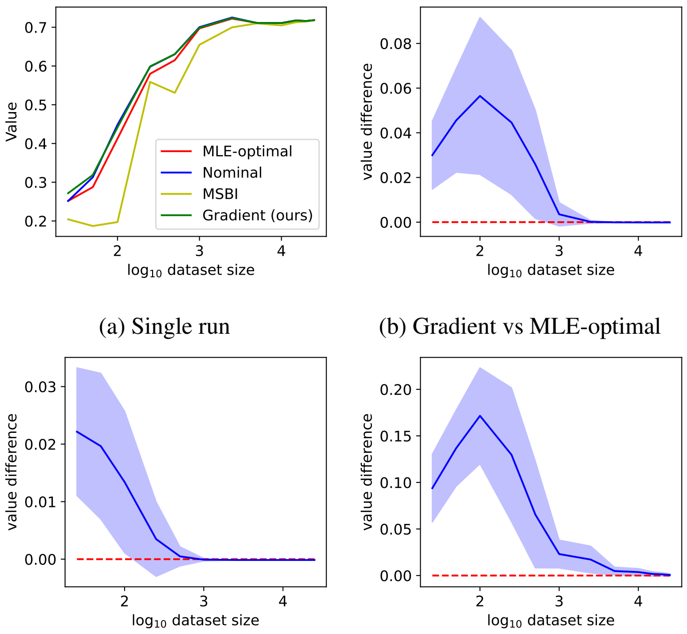
In contrast, we find that the approach in [14] always leads to low epistemic uncertainty at the end of training, as the lack of knowledge of the underlying MDP is not modelled, and thus does not scale consistently with data. In Appendix B we visualise how this quantity evolves during training with different datasets after adapting the algorithm to carry out SARSA policy evaluation on the same, fixed policy and observe that it always tends to be small regardless of how informative the dataset is by the end of training. Additionally, as discussed in section 4.1, the computation of aleatoric and epistemic uncertainty through closed-form moments as in Algorithm 1 achieves better accuracy for similar computation compared to previous methods that carry out aleatoric and epistemic uncertainty quantification in discrete MDPs.



(a) Gridworld

(b) Aleatoric and epistemic uncertainty

Figure 1: Fig. 1a shows the gridworld used in the experiments. The terminal states are the failure states (cliff) marked as **F** in red, and the goal state marked as **G** in green. The agent can move up, down, left, or right (or remain stationary if it hits the boundary of the grid). The transition dynamics have intrinsic stochasticity controlled by the probability p_{rand}, which is the probability of pushing the agent down regardless of action taken. Offline training datasets were created by randomly sampling actions at random non-terminal states. State ★ is chosen as an exemplar state to plot state-dependent uncertainties. In Fig. 1b, the epistemic (blue) and aleatoric (red) standard deviations (taken as the square root of the variances in Eq. 5) are shown as a function of training dataset size, with different levels of intrinsic stochasticity indicated by solid, dashed, and dotted lines.



(a) Single run

(b) Gradient vs MLE-optimal

(c) Gradient vs Nominal-optimal

(d) Gradient vs MSBI

Figure 2: Fig. 2a shows the average return ('Value') as a function of dataset size, averaged across the Bayesian posterior, as in Eq. 6. We compare the performance on this objective of four methods: (i) MLE-optimal policy with naive transition probabilities, (ii) the optimal policy for the expected (nominal) MDP, (iii) MSBI policy from [11], and (iv) our proposed gradient-optimized policy. Higher values indicate better performance at equal dataset sizes. The example gridworld has $p_{\text{rand}} = 0.25$. As value will be dataset-dependent, we show the average and standard deviation between the difference in Bayesian state value at state ★ for the same dataset in Fig. 2b and 2d, where values above the red dashed line signify an improvement. These plots report the average and standard deviation across 50 generated datasets for each dataset size.

Bayesian Policy Improvement An optimal memoryless policy that accounts for the model uncertainty will be maximising the average value across the MDP posterior given in Eq. 6. We compare the performance on this objective of our Algorithm 2 (Gradient policy), the optimal policy for naive visitation frequencies (MLE-optimal policy), the optimal policy for the *expected* (marginalised), referred to as Nominal [11, 10], MDP (Nominal-optimal policy) and the Multi-Sample Backwards Induction (MSBI policy) algorithm suggested in [11]. The latter comes with theoretical guarantees of near-optimality; however, the number of samples from the transition matrix posterior required for such guarantees for our setup are of the order of magnitude $(\epsilon(1-\gamma))^{-3} \approx 10^{14}$ using $\gamma = 0.999$ and an error tolerance on the value of $\epsilon = 0.01$, which is a computationally intractable number of samples to store and process for transition matrices. Thus, we use a number of samples ($N_M = 32768$) that roughly matches the computation time of the gradient-optimised policy (30-60s depending on dataset without GPU acceleration for the gridworld experiments). In Algorithm 2, we choose our initial policy to be a softened version of the Nominal-optimal policy. The initial MSBI policy is also taken to be the Nominal-optimal policy for a fairer comparison.

We empirically find that the gradient-optimised policy consistently outperforms this version of MSBI, for the particular MDPs and dynamics parameter distributions we are considering, as well as the MLE- and Nominal-optimal policies especially in lower data regimes when optimising the Bayesian posterior value objective. Results for a sample state and dataset are presented in Fig. 2a for different dataset sizes. The corresponding relative performances of our method against both MLE-optimal and MSBI on the Bayesian objective over 50 sets of generated datasets are presented in Fig. 2b and 2d with error bars (standard deviations), confirming that our method consistently outperforms the other two over a larger number of randomly-generated datasets.

5.2 Clinical Data

We apply Algorithm 2 to the MIMIC-III dataset, as in [26] and [14], using the same state clustering of 752 states and 25 actions. Two terminal states represent patient recovery and death. As in [26], actions at any state with fewer than 5 visits in the dataset are excluded. A patient's recovery gives a reward of 1, death gives 0, and no intermediate rewards are used. This ensures that state value corresponds approximately to probability of survival when $\gamma \approx 1$ ($\gamma = 0.999$).

Bayesian inference with Dirichlet distributions with a large number of possible outcomes (next states) is problematic, as mentioned in section 3 [15], and careful thought must be given to what prior to employ. First we consider a Bayesian model selection approach: we assume all possible states are reachable and symmetric. This allows us to optimise the model evidence with respect to the unique parameter α_p of the prior, in the hope that specifying a prior which is more in line with the observations will lead to better inference (see Appendix D for details). As expected, the optimal α_p is found to be much smaller than 1, $\alpha_p = 0.072$, giving less weight after inference to the prior than the maximum-entropy $\alpha_p = 1$ prior does. However, this approach still fails to accurately model our belief, which can be seen by considering the following scenario: suppose the patient is in a bad state and has two options, namely (a) try a treatment that has been attempted many times with rare success or (b) try a treatment that has always gone wrong, but has been tried a small number of times so has high uncertainty in the outcome. Option (b) is clearly not appealing, but the agent's posterior will still place significant probability mass on unobserved states in the presence of a small number of transitions, thus highly encouraging the agent to take the less visited action and assigning it a disproportionately high value. Upon inspection, this is exactly what is happening in the outlier state in Fig. 3a (at approximate coordinates $(0.6, 0.8)$), and the value given by this Bayesian posterior is likely unreasonable.

To address this, we introduce conservatism by considering only observed states and the death state as next possible states, thus ensuring a more conservative prior. Inducing conservatism in offline RL with datasets that don't adequately cover the full state-action space is in line with literature [1, 28], and conservative MDP models have found success in continuous offline RL by modulating reward [39, 25] or dynamics [19], somewhat analogously to what is being proposed here. By only including observed or negative outcomes, the agent is unable to place probability mass on unsupported next-states and therefore use high uncertainty to inflate the value of poorly visited actions in bad states. The scarcity of outcomes allows for meaningful inference using a maximum-entropy prior with $\alpha_p = 1$, and a high-entropy prior is favorable from a conservatism standpoint. It encourages the agent to select actions that have sufficient support to offset the high prior probability mass assigned to the death state. The Bayesian values inferred with this setup are presented in Fig. 3b. Fig. 3 shows the possible improvement, according to the Bayesian posterior value, of employing the Bayesian gradient-optimised policy compared to the MLE-optimal policy used in [26], resulting in higher probability of survival (according to the dynamics model). In particular, we note that employing the gradient-optimised policy improves the value, and therefore corresponding approximate probability of survival, by about 2.1% when averaged across states, with a maximum improvement on a particular state of 17.8%, according to the conservative Bayesian dynamics model.

6 Limitations

Our methods apply to a specific category of Markov Decision Processes (MDPs) with finite states and known reward structures. We have shown these are capable of handling moderately-sized MDPs that carry practical real-world application possibilities in section 5.2), yet it is unclear exactly how large the MDPs tackled can be before these approaches become computationally intractable. One key limitation of our proposed methods is its sensitivity of the resulting policy and inferred values on the dynamics model prior used, especially when data is inadequate for effective inference across all dynamics priors. For example, we observe that the effects of having a sparse or evidence-optimised model can be significant on both the inferred policy and the associated values (see Fig. 3) and exactly how to best include or combine these elements to select a prior that achieves consistently good performance on the ground-truth MDPs is an important question and one that we defer to future work.

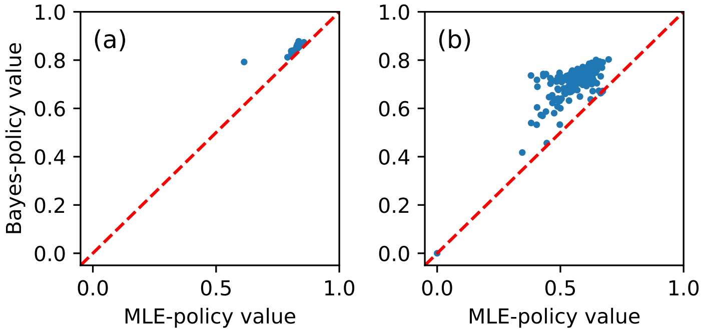


Figure 3: Values of each state under the Bayesian policy and the MLE-optimal policy in the clinical MDP. Each state is represented by its corresponding Bayesian and MLE values, and points above the diagonal indicate superior performance of the Bayesian policy on the Bayesian objective. The left plot (a) demonstrates the impact of different dynamics model priors on performance when employing Bayesian model selection with an optimal parameter of $\alpha_p = 0.072$. The right plot (b) shows the results when using a prior selected through a conservative sparse dynamics model.

7 Conclusion

We have proposed a framework for estimating aleatoric and epistemic uncertainty in the outcome of discretised state space policies and use it for control, including an example to application in the domain of clinical decision support systems. Specifically, the setup analysed here is relevant to the setup presented in one of the key exemplars of offline RL [26] in clinical decision support systems. In comparison to previous frameworks estimating such uncertainties in RL with a similar setup there are two main improvements. 1. we do not require function approximators and therefore entirely bypass complications, numerical inaccuracies or uncertainties that may be introduced during the training of these. 2. by employing a Bayesian dynamics model, the quantification for epistemic uncertainty meaningfully scales with additional data, which is not a feature of ensemble methods [8, 14] as training uncertainty is the only modelled uncertainty. Additionally, the stationarity of the dynamics model employed enables us to compute standard deviations of the return distribution analytically without requiring Monte Carlo trajectory sampling, as done in [23], resulting in more accurate and computationally efficient evaluations.

On the control side, we can account for epistemic uncertainty in the optimisation of a policy and, as highlighted by earlier work [35], address aleatoric uncertainty by suggesting it should be handled by reshaping the reward rather than doing non-expectation-based optimisation. While previous methods to carry out memoryless Bayesian policy optimisation exist [11], the computational overhead to attain the theoretical guarantees in these is intractable for our setup. Therefore, we propose a computationally scalable approach that outperforms its feasible counterpart based on empirical evaluations. Our approach has relevance to the analysis of MDPs with more general uncertainty in dynamics parameters [38, 10] particularly when practical computational considerations take precedence over theoretical guarantees. We have introduced pessimism in the face of uncertainty, a common and necessary ingredient in offline RL [25, 39, 2] especially when the dataset does not adequately span the full state-action space [1], in the form of a conservative dynamics model. This draws an analogy to conservatism in the face of uncertainty commonly used in continuous control offline RL.

On the application side, the methods presented here could be employed for a variety of purposes, such as enhancing possible treatment strategies' interpretability to decision-makers through uncertainty quantification, for example by splitting states and treatment options into varying groups of outcome uncertainty (see Appendix E for an example visualisation). In large MDPs, we caution against using naive symmetrical Dirichlet priors for dynamics modelling, as used for example in [12, 18], when data is finite and limited. Instead, we suggest exploring better prior modelling techniques such as sparse or hierarchical [15] Bayesian models, which could be combined with evidence-based Bayesian model selection, to improve epistemic uncertainty quantification and control robustness.

Acknowledgments and Disclosure of Funding

FV was supported by the Department of Computing, Imperial College London. AF was supported by a UKRI Turing AI Fellowship (EP/V025499/1).

References

[1] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020.

[2] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[4] Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. `http://www.distributional-rl.org`.

[5] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[6] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[7] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G. Lee, Vikram Deshpande, Joseph Schwab, Michael H. Lev, Ramon G. Gonzalez, Michael S. Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, Dec 2022.

[8] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[9] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[10] Erick Delage and Shie Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

[11] Christos Dimitrakakis. Robust Bayesian reinforcement learning through tight lower bounds. In *European Workshop on Reinforcement Learning*, pages 177–188. Springer, 2011.

[12] Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

[13] François Dufour and Tomás Prieto-Rumeau. Approximation of Markov decision processes with general state space. *Journal of Mathematical Analysis and applications*, 388(2):1254–1267, 2012.

[14] Paul Festor, Giulia Luise, Matthieu Komorowski, and A Aldo Faisal. Enabling risk-aware reinforcement learning for medical interventions through uncertainty decomposition. In *Workshop in Interpretable Machine Learning for Healthcare*, 2021.

[15] Nir Friedman and Yoram Singer. Efficient Bayesian parameter estimation in large discrete domains. *Advances in neural information processing systems*, 11, 1998.

[16] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

[17] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

[18] Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. *Advances in neural information processing systems*, 25, 2012.

[19] Kaiyang Guo, Yunfeng Shao, and Yanhui Geng. Model-based offline reinforcement learning with pessimism-modulated dynamics belief. *arXiv preprint arXiv:2210.06692*, 2022.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[21] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[22] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[23] Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*, 2021.

[24] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[25] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

[26] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[27] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

[28] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Gilwoo Lee, Brian Hou, Aditya Mandalika, Jeongseok Lee, Sanjiban Choudhury, and Siddhartha S Srinivasa. Bayesian policy optimization for model uncertainty. *arXiv preprint arXiv:1810.01014*, 2018.

[30] Luchen Li, Ignacio Albert-Smet, and Aldo A Faisal. Optimizing medical treatment for sepsis in intensive care: from reinforcement learning to pre-trial evaluation. *arXiv preprint arXiv:2003.06474*, 2020.

[31] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[32] Arnab Nilim and Laurent Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

[33] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704, 2006.

[34] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 2014.

[35] Matthew J Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

[36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[37] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[38] Huan Xu and Shie Mannor. The robustness-performance tradeoff in Markov decision processes. *Advances in Neural Information Processing Systems*, 19, 2006.

[39] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[40] Luisa Zintgraf, Sebastian Schulze, Cong Lu, Leo Feng, Maximilian Igl, Kyriacos Shiarlis, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: variational bayes-adaptive deep rl via meta-learning. *The Journal of Machine Learning Research*, 22(1):13198–13236, 2021.

A Dynamics posterior with infinite data

Recall that the transition parameters corresponding to the transition probabilities for state-action s, a have posterior probability density of the form given in Eq. 4, which is a Dirichlet distribution with parameters $\alpha_i = n_i + \alpha_p$. A standard property of the Dirichlet distribution is that the variances of its random variables are given by

$$\text{Var}\,\theta_{sa}^i = \frac{\hat{\alpha}_i(1-\hat{\alpha}_i)}{\alpha_0+1},$$

where $\alpha_0 = \sum_j \alpha_j$ and $\hat{\alpha}_i = \alpha_i/\alpha_0$. Since $0 \leq \hat{\alpha}_i \leq 1$, in the limit of infinite data, corresponding to arbitrarily large α_0, the variance of the posterior vanishes. Hence, the posterior probability mass entirely concentrates arbitrarily close to the MDP corresponding to the transition parameters with $\hat{\alpha}_i$ which, in the limit, corresponds to the observed visitation frequency.

B Policy uncertainty evaluation

The policy we present and compare results for is the policy that optimises the maximum likelihood estimate (MLE) of the transition dynamics MDP, where transition probability is taken to be the relative frequency of observed transitions, which we refer to as the MLE-optimal policy.

Running SARSA policy evaluation on the methods proposed in [14] explicitly shows that the epistemic uncertainty in the dynamics transition is not captured by the ensemble method used. Fig. 4 shows that with this setup, epistemic uncertainty correlates with loss but is independent of amount of data observed. This is visible as the curves collapse to small epistemic uncertainty values irrespective of data set size even though the amount of data in the smallest data set size (25) is smaller than the total number of transitions of the MDP (80). This is because it captures information on parametric training uncertainty but not of the dynamics model uncertainty.

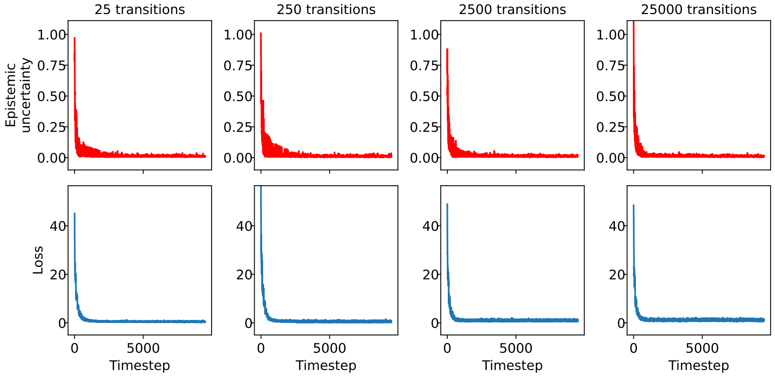


Figure 4: Plot of the epistemic uncertainty and loss as a function of training timestep demonstrating that epistemic is not accurately tracked by previous methods. Epistemic standard deviation (top row, red data) is quantified here over 10k time steps, corresponding to the agent carrying out transitions over many episodes. The corresponding ensemble quantile regression loss (bottom row, blue data) at each training timestep is shown below. Here we show as examplar the results for fixed policy using ensemble methods with a MLE-dynamics model for different number of observed transitions in the dataset generated by the gridworld with $p_{\text{rand}} = 0.5$. The value that the epistemic standard deviation converges to is always small for all visited states and independent of dataset size as the only notion of uncertainty captured in this setup is one of parametric uncertainty and not MDP uncertainty.

C Probabilistic evaluation bounds

Here we provide a quantitative investigation into the choice of method to evaluate the quantities of interest for a given policy, including a comparison of the probabilistic bounds on the errors due to finite numbers of samples. We compare the efficiency required to achieve an evaluation within a certain accuracy ε with a minimum probability $1-\delta$ for methods that (i) (ours, Algorithm 1) carry out an exact calculation of the return distribution moments and then Monte Carlo evaluation with samples from the dynamics posterior or (ii) carry out Monte Carlo sampling for every evaluation step. The quantity we investigate in detail is the Bayesian value at a given state for a given policy (appearing in Eq. 6 for a given policy and state), and since aleatoric and epistemic uncertainties are calculated in very similar fashion, the conclusions regarding Bayesian value estimation will also carry through to the uncertainty quantification case.

C.1 Exact moments

The quantity of interest we wish to approximate is

$$\hat{V} = \mathbb{E}_{\mathcal{M}}(V_{\mathcal{M}}(s)),$$

where the expectation is taken over the Dirichlet posterior of MDP dynamics parameters. For a given set of dynamics parameters $\mathcal{M}$, we have access to the closed form expression for the first moment of the return distribution $V_{\mathcal{M}}(s)$ (in terms of policy, dynamics and reward) as presented in Eq. 2.

We assume a bounded reward $|r| \leq r_{\text{max}}$ and employ the well-known form of the Hoeffding inequality [20] valid for the random variable $S_n = \sum_{i=1}^{n} X_i$ with X_i bounded and i.i.d. such that $\mathbb{E}(S_n) = \mu$:

$$\mathbb{P}(|S_n - \mu| \leq \epsilon) \geq 1 - 2\exp\left(-\frac{2\epsilon^2}{n\Delta^2}\right)$$

with Δ being the size of the interval on which X can take values.

In context, we take $X_i = \frac{1}{N_M} V_i$ as the closed-form expression for the value of the i^{th} of the N_M dynamics samples, so $\mu = \hat{V}$. From the boundedness assumption on the reward, we can also bound $|V_i| \leq \frac{r_{\text{max}}}{1-\gamma} = V_{\text{max}}$ and $\Delta \leq 2V_{\text{max}}/N_M$. We require enough samples so that with probability at least $1-\delta$ the error in our approximation of $\hat{V}$ is within ϵ of the true value. By the Hoeffding inequality, we can ensure this is the case by choosing N_M such that

$$\delta \leq 2\exp\left(-\frac{N_M\epsilon^2}{2V_{\text{max}}^2}\right),$$

which corresponds to the smallest integer N_M such that

$$N_M \geq \log\left(\frac{2}{\delta}\right)\left(\frac{2V_{\text{max}}^2}{\epsilon^2}\right).$$

C.2 Monte-Carlo sampling

The alternative method to using closed-form expressions for the moments of the return distribution given an MDP sample would be to in turn approximate these through Monte Carlo samples, as done in [23]. To do so, given the infinite horizon nature of the MDPs we are considering, we would have to accumulate rewards over a roll-out with a finite number of steps T, thus incurring in some error, which can be bounded above by $\gamma^T V_{\text{max}}$. Note that the tightness of this bound will depend entirely on the reward structure of the MDP, and that this is not a source of error that can be reduced by repeatedly sampling transitions. For the purposes of the analysis presented, we will be generous in mostly ignoring the computational cost associated with sampling trajectories for a given MDP. In practice, sampling from a categorical distribution (i.e. sampling the trajectories for a given MDP) is significantly faster than sampling from a Dirichlet distribution (i.e. sampling the transition matrix), so we incorporate the overall computational cost of trajectory sampling into the modest condition that T cannot be arbitrarily large, but assume infinite trajectory sampling capability otherwise. This assumption allows us to determine the value for the i^{th} given MDP arbitrarily accurately up to this error, so that the distance between the true value V_i to the accumulated finite sum of rewards V_i' will be bounded by $|V_i - V_i'| \leq \gamma^T V_{\text{max}}$.

Thus, we can consider the distance

$$\begin{aligned}\left|\hat{V} - \frac{1}{N_M}\sum_i V_i'\right| &\leq \left|\hat{V} - \frac{1}{N_M}\sum_i V_i\right| + \left|\frac{1}{N_M}\sum_i V_i - \frac{1}{N_M}\sum_i V_i'\right| \\ &\leq \left|\hat{V} - \frac{1}{N_M}\sum_i V_i\right| + \gamma^T V_{\text{max}},\end{aligned}$$

so that if

$$\left|\hat{V} - \frac{1}{N_M}\sum_i V_i\right| + \gamma^T V_{\text{max}} \leq \epsilon,$$

with probability at least $1 - \delta$, then the distance to the original estimate also satisfies

$$\left|\hat{V} - \frac{1}{N_M}\sum_i V_i'\right| \leq \epsilon.$$

with at least probability $1 - \delta$.

As such, we apply the Hoeffding inequality in the form

$$\mathbb{P}\left(\left|\hat{V} - \frac{1}{N_M}\sum_i V_i\right| \leq \epsilon - \gamma^T V_{\text{max}}\right) \geq 1 - 2\exp\left(-\frac{N_M(\epsilon - \gamma^T V_{\text{max}})^2}{2V_{\text{max}}^2}\right).$$

Note that this also imposes a minimum horizon truncation of $T > \log(\epsilon/V_{\text{max}})/\log\gamma$. Explicitly including the probability threshold δ now corresponds to finding an N_M such that

$$\delta \leq 2\exp\left(-\frac{N_M(\epsilon - \gamma^T V_{\text{max}})^2}{2V_{\text{max}}^2}\right),$$

so

$$N_M \geq \log\left(\frac{2}{\delta}\right)\frac{2V_{\text{max}}^2}{(\epsilon - \gamma^T V_{\text{max}})^2}.$$

This bound corresponds to a worsening by a factor of $(1 - \gamma^T V_{\text{max}}/\varepsilon)^{-2}$ in the number of samples required to get comparable accuracy to the method that uses exact moments. For example, for the gridworld setup considered ($\gamma = 0.999, r_{\text{max}} = 1$ and positing $\epsilon = 0.001$) would require an order of magnitude of $T \approx 10^5$ for every rolled out trajectory, (of which we are assuming to be able to carry out an arbitrarily large number to obtain this bound) at which point the contribution of the trajectory sampling to the bottleneck would be severe and require a completely different bound to take it into account. Thus, for the regime we consider, choosing to compute exact moments does save computation towards the computational bottleneck of taking samples from a Dirichlet posterior.

Note that aleatoric and epistemic uncertainty will behave similarly: aleatoric variance is an analogous expectation over the second instead of first moment (which we again can have in closed-form or can estimate through Monte Carlo samples) and the bound will be analogous. Similarly, for epistemic variance the error in return due to truncated trajectories will compound when calculating the variance over expected returns, and again we expect a similarly greater number of samples for N_M.

D Bayesian model selection

To determine the prior that for the dynamics model with results presented in Fig. 3a, we carry out Bayesian model selection by minimising the negative log-marginal likelihood of the data with respect to the parameter α_p. To remain consistent with the limitation that only actions observed at least 5 times in the data should be employed at each state, we only use the data for such state-action transitions when determining the optimal α_p.

For each state-action, the full form of the Dirichlet prior in terms of α_p is [15]

$$p(\{\theta_{s,a}^{s_j} | s_i \in \mathcal{S}\}) = \frac{\Gamma(|\mathcal{S}|\alpha_p)}{\Gamma(\alpha_p)^{|\mathcal{S}|}}\prod_j (\theta_{s,a}^{s_j})^{\alpha_p - 1},$$

where Γ is the gamma function. The likelihood is

$$p(\mathcal{D}|\theta) = \prod_j (\theta_{s,a}^{s_j})^{n_j},$$

with n_j being the number of observed transitions from state-action s, a to state s_j. Hence, the model evidence is

$$\begin{aligned} p(\mathcal{D}) &= \int d\theta p(\mathcal{D}|\theta)p(\theta) && (7) \\ &= \frac{\Gamma(|\mathcal{S}|\alpha_p)}{\Gamma(\alpha_p)^{|\mathcal{S}|}}\frac{\prod_j \Gamma(\alpha_p + n_j)^{|\mathcal{S}|}}{\Gamma(|\mathcal{S}|\alpha_p + N_{s,a})}, && (8) \end{aligned}$$

with $N_{s,a}$ being the number of observed transitions from state-action s, a. Since transitions are independent across state-actions, taking the negative logarithm of this quantity and summing across all state-actions results in the overall negative log-marginal likelihood for the dataset in terms of α_p. The resulting function of α_p is visualised in Fig.5 and attains a minimum value at approximately $\alpha_p = 0.072$.

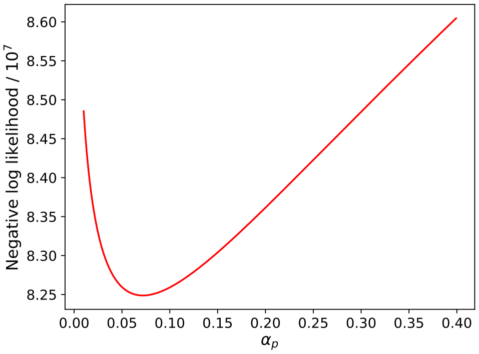


Figure 5: Negative log-marginal likelihood for clinical data dynamics model against parameter α_p of the prior.

E State uncertainty visualisation

In Fig. 6 we show how the MIMIC-III states aleatoric and epistemic uncertainties are related. The values are computed using the same conservative dynamics model of Fig. 3b.

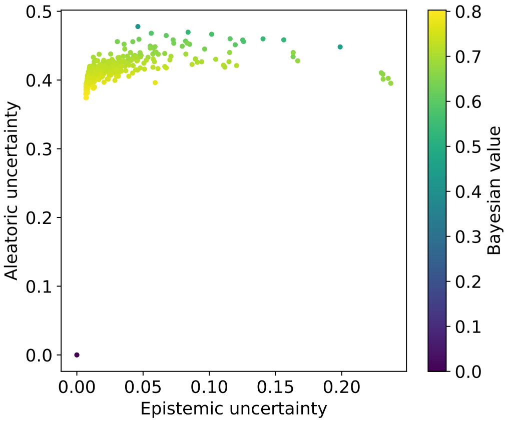


Figure 6: States plotted according to their epistemic and aleatoric standard deviations. Each dot represents a state, with its colour corresponding to its average value according to the Bayesian posterior.

As expected for the particular reward structure of the MDP considered, aleatoric uncertainty and average Bayesian value are strongly related: since the return variable is approximately binomial (approximately 1 for success and 0 for failure) its mean and variance are related straightforwardly. Note this will not be true for MDPs with more general return structures.