Planning to the Information Horizon of BAMDPs via Epistemic State Abstraction

Dilip Arumugam  
Department of Computer Science  
Stanford University  
dilip@cs.stanford.edu

Satinder Singh  
DeepMind, London  
baveja@deepmind.com

Abstract

The Bayes-Adaptive Markov Decision Process (BAMDP) formalism pursues the Bayes-optimal solution to the exploration-exploitation trade-off in reinforcement learning. As the computation of exact solutions to Bayesian reinforcement-learning problems is intractable, much of the literature has focused on developing suitable approximation algorithms. In this work, before diving into algorithm design, we first define, under mild structural assumptions, a complexity measure for BAMDP planning. As efficient exploration in BAMDPs hinges upon the judicious acquisition of information, our complexity measure highlights the worst-case difficulty of gathering information and exhausting epistemic uncertainty. To illustrate its significance, we establish a computationally-intractable, exact planning algorithm that takes advantage of this measure to show more efficient planning. We then conclude by introducing a specific form of state abstraction with the potential to reduce BAMDP complexity and gives rise to a computationally-tractable, approximate planning algorithm.

1 Introduction

The Bayes-Adaptive Markov Decision Process (BAMDP) [Duff, 2002] is a classic formalism encapsulating the optimal treatment of the exploration-exploitation trade-off by a reinforcement-learning agent with respect to prior beliefs over an uncertain environment. Unfortunately, the standard formulation suffers from an intractably-large hyperstate space (that is, the joint collection of environment states coupled with the agent’s current state of knowledge over the unknown environment) and much of the literature has been dedicated to identifying suitable approximations [Bellman and Kalaba, 1959, Dayan and Sejnowski, 1996, Duff and Barto, 1997, Dearden et al., 1998, Strens, 2000, Duff, 2001, 2003b,a, Wang et al., 2005, Poupart et al., 2006, Castro and Precup, 2007, Kolter and Ng, 2009, Asmuth et al., 2009, Dimitrakakis, 2009, Sorg et al., 2010, Araya-López et al., 2012, Guez et al., 2012, 2013, 2014, Ghavamzadeh et al., 2015, Zintgraf et al., 2019]. In this work, we take steps toward clarifying the hardness of BAMDPs before outlining an algorithmic concept that may help mitigate problem difficulty and facilitate near-optimal solutions.

First, we introduce the notion of information horizon as a complexity measure on BAMDP planning, characterizing when it is truly difficult to identify the underlying uncertain environment. Naturally, the agent’s state of knowledge at each timestep (a component of the overall BAMDP hyperstate) reflects its current epistemic uncertainty and, as the agent accumulates data, this posterior concentrates, exhausting uncertainty and identifying the true environment; after this point, the Bayes-optimal policy naturally coincides with the optimal policy of the underlying Markov Decision Process (MDP). Simply put, the information horizon quantifies the worst-case number of timesteps needed for the agent to reach this point whereupon there is no more information to be gathered about the uncertain environment.
With this complexity measure in hand, we then entertain the idea of epistemic state abstraction as an effective algorithmic tool for trading off between reduced information horizon (complexity) and near-Bayes-optimality of the corresponding planning solution. Intuitively, as the total number of knowledge states an agent may take on drives the intractable size of the hyperstate space, we operationalize state abstraction [Li et al., 2006, Abel et al., 2016] to perform a lossy compression of the epistemic state space, inducing a “smaller” and more tractable BAMDP for planning; our results not only mirror those of analogous work on state aggregation for improved efficiency in traditional MDP planning [Van Roy, 2006] but also parallel similar findings [Hsu et al., 2007, Zhang et al., 2012] on the effectiveness of belief state aggregation in partially-observable MDP (POMDP) planning [Kaelbling et al., 1998].

On the whole, our work provides one possible answer to a question that has already been asked and answered several times in the context of MDPs [Bartlett and Tewari, 2009, Jaksch et al., 2010, Farahmand, 2011, Maillard et al., 2014, Bellemare et al., 2016, Arumugam et al., 2021, Abel et al., 2021]: how hard is my BAMDP? While the remainder of the paper goes on to examine how one particular mechanism for reducing this complexity can translate into a more efficient planning algorithm, we anticipate that this work can serve as a starting point for building a broader taxonomy of BAMDPs, paralleling existing structural classes of MDPs [Jiang et al., 2017, Sun et al., 2019, Agarwal et al., 2020, Jin et al., 2021].

2 Problem Formulation

In this section, we formally define BAMDPs as studied in this paper. As a point of contrast, we begin by presenting the standard MDP formalism used throughout the reinforcement-learning literature [Sutton and Barto, 1998]. We use $\Delta(\mathcal{X})$ to denote the set of all probability distributions with support on an arbitrary set $\mathcal{X}$ and denote, for any natural number $N \in \mathbb{N}$, the index set as $[N] = \{1, 2, \ldots, N\}$. For any two arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the class of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$.

2.1 Markov Decision Processes

We begin with a sequential decision-making problem represented via the traditional finite-horizon Markov Decision Process (MDP) [Bellman, 1957, Puterman, 1994] $\langle S, A, R, T, \beta, H \rangle$ where $S$ is a finite set of states, $A$ is a finite set of actions, $R : S \times A \to [0, 1]$ is a deterministic reward function, $T : S \times A \to \Delta(S)$ is a transition function prescribing next-state transition distributions for all state-action pairs, $\beta \in \Delta(S)$ is an initial state distribution, and $H \in \mathbb{N}$ is the horizon denoting the agent’s total number of steps or interactions with the environment. An agent’s sequential interaction within this environment proceeds in each timestep $h \in [H]$, starting with an initial state $s_1 \sim \beta(\cdot)$, by observing the current state $s_h \in S$, selecting an action $a_h \in A$, and then enjoying a reward $R(s_h, a_h)$ as the environment transitions to $s_{h+1} \sim T(\cdot | s_h, a_h)$. Action selections made by the agent are governed by its non-stationary policy $\pi$: a collection of $H$ stationary, deterministic policies $\pi = (\pi_1, \pi_2, \ldots, \pi_H)$, where $\forall h \in [H], \pi_h : S \to A$. We quantify the performance of policy $\pi$ at timestep $h \in [H]$ by its induced value function $V^\pi_h : S \to \mathbb{R}$ denoting the expected sum of future rewards by deploying policy $\pi$ from a particular state $s \in S$:

$V^\pi_h(s) = \mathbb{E} \left[ \sum_{h'=h}^H R(s_{h'}, a_{h'}) \mid s_h = s \right]$, where the expectation integrates over randomness in the environment transitions. Analogously, we define the action-value function induced by policy $\pi$ at timestep $h$ as $Q^\pi_h : S \times A \to \mathbb{R}$ which denotes the expected future sum of rewards by being in a particular state $s \in S$, executing a particular action $a \in A$, and then following policy $\pi$ thereafter:

$Q^\pi_h(s, a) = \mathbb{E} \left[ \sum_{h'=h}^H R(s_{h'}, a_{h'}) \mid s_h = s, a_h = a \right]$. We are guaranteed the existence of an optimal policy $\pi^*$ that achieves supremal value $V^*_{h} = \sup_{\pi \in \Pi} V^\pi_h(s)$ for all $s \in S, h \in [H]$ where the policy class contains all deterministic policies $\Pi = \{\pi \mid \pi : S \to A\}$. Since rewards are bounded in $[0, 1]$, we have that $0 \leq V^\pi_h(s) \leq V^*_{h} \leq H - h + 1$ for all $s \in S, h \in [H]$, and $\pi$. These value functions obey the Bellman equation and the Bellman optimality equation, respectively:

$V^\pi_h(s) = Q^\pi_h(s, \pi_h(s)) \quad V^\pi_h(s) = \max_{a \in A} Q^\pi_h(s, a), \quad V^\pi_{H+1}(s) = 0 \quad V^*_{H+1}(s) = 0 \quad \forall s \in S$, 

2
\[ Q_h^\pi(s, a) = \mathcal{R}(s, a) + \mathbb{E}_{s' \sim \mathcal{T}(\cdot | s, a)} \left[ V_{h+1}^\pi(s') \right], \quad Q_h^\beta(s, a) = \mathcal{R}(s, a) + \mathbb{E}_{s' \sim \mathcal{T}(\cdot | s, a)} \left[ V_{h+1}^\beta(s') \right]. \]

### 2.2 Bayes-Adaptive Markov Decision Processes

The BAMDP formalism offers a Bayesian treatment of an agent interacting with an uncertain MDP. More specifically, a decision-making agent is faced with a MDP \( M = (S, A, \mathcal{R}, \mathcal{T}_\theta, \beta, H) \) defined around an unknown transition function \( \mathcal{T}_\theta \), for some latent parameter \( \theta \in \Theta \). Prior uncertainty in \( \theta \) is reflected by the distribution \( p(\theta) \). In classic work on BAMDPs with finite state-action spaces, the parameters \( \theta \) denote visitation counts and \( p(\theta) \) is a Dirichlet distribution, so as to leverage the convenience of Dirichlet-multinomial conjugacy for exact posterior updates [Duff, 2002, Poupart et al., 2006]. For our purposes, we will assume an alternative parameterization whose importance will be made clear later when defining our complexity measure.

**Assumption 1.** We assume that \( \Theta \) is known and \( |\Theta| < \infty \) such that an agent is only ever reasoning about its uncertainty over a finite set of \( |\Theta| \) known MDPs. We further make a realizability assumption that the true parameters reside in this finite set, \( \theta \in \Theta \).

Under Assumption 1, an agent’s prior uncertainty in \( \mathcal{T}_\theta \) is reflected by the distribution \( p(\theta) \in \Delta(\Theta) \) which, with each step of experience encountered by the agent, may be updated via Bayes’ rule to recover a corresponding posterior distribution in light of observed data from the environment. For simplicity, we do not concern ourselves with the computation of the posterior and instead assume access to a deterministic \( B : \Delta(\Theta) \times S \times A \times S \to \Delta(\Theta) \) that performs an exact posterior update to any input distribution \( p \in \Delta(\Theta) \) based on the experience tuple \( (s, a, s') \in S \times A \times S \) in \( O(1) \) time.

The corresponding BAMDP for \( M \) is defined around a so-called *hyperstate* space \( \mathcal{X} = S \times \Delta(\Theta) \) such that any hyperstate \( x = \langle s, p \rangle \in \mathcal{X} \) denotes the agent’s original or physical state \( s \in S \) within the true MDP while \( p \in \Delta(\Theta) \) denotes the agent’s information state or epistemic state [Lu et al., 2021] about the uncertain environment; intuitively, the epistemic state represents the agent’s knowledge of the environment based on all previously observed data. This gives rise to the BAMDP \((\mathcal{X}, A, \mathcal{R}, \mathcal{T}, \beta, H)\) where \( A \) is the same action set as the original MDP \( M, \mathcal{R} : \mathcal{X} \times A \to [0, 1] \) is the same reward function as in \( M \) (that is, \( \mathcal{R}(\langle s, p \rangle) = \mathcal{R}(s, a) \forall (s, p) \in \mathcal{X}, a \in A \)), \( \beta \in \Delta(\mathcal{X}) \) is defined as \( \beta = \beta \times \delta_{p(\theta)} \) where \( \delta_{p(\theta)} \) denotes a Dirac delta centered around the agent’s prior \( p(\theta) \), and \( H \) is the same horizon as MDP \( M \). Due to the determinism of the posterior updates given by \( B \), the BAMDP transition function \( \mathcal{T} : \mathcal{X} \times A \to \Delta(\mathcal{X}) \) is defined as

\[ \mathcal{T}(x' \mid x, a) = \sum_{\theta \in \Theta} \mathcal{T}_\theta(s' \mid s, a)p(\theta)\mathbb{1}(p' = B(p, s, a, s')), \]

where \( x' = \langle s', p' \rangle \in \mathcal{X} \). The associated BAMDP policy \( \pi = (\pi_1, \pi_2, \ldots, \pi_H) \), \( \pi_h : \mathcal{X} \to A, \forall h \in [H] \) selects actions based on the current state of the MDP as well as the agents accumulated knowledge of the environment thus far. With these components, we may define the associated BAMDP value functions with \( x' = \langle s', B(p, s, a, s') \rangle \):

\[ V_h^\pi(x) = Q_h^\pi(x, \pi_h(x)), \quad V_h^x(x) = \max_{a \in A} Q_h^\pi(x, a), \quad V_{H+1}^\pi(x) = 0 \quad \forall x \in \mathcal{X}, \]

\[ Q_h^\pi(x, a) = \mathcal{R}(s, a) + \sum_{s' \neq \theta} \mathcal{T}_\theta(s' \mid s, a)p(\theta)V_{h+1}^\pi(x'), \quad Q_h^\beta(x, a) = \mathcal{R}(s, a) + \sum_{s' \neq \theta} \mathcal{T}_\theta(s' \mid s, a)p(\theta)V_{h+1}^\beta(x'). \]

Based on these optimality equations, we see that the optimal policy of a BAMDP achieving supremal value \( V_h^\pi \) across all timesteps \( h \in [H] \) is the Bayes-optimal policy which appropriately balances the exploration-exploitation trade-off in reinforcement learning. An observation is that this Bayes-optimal policy will tend to achieve lower value than the optimal policy of MDP \( M \) as the agent takes more informative (possibly sub-optimal) actions to identify the true underlying environment.

### 3 Related Work

Bellman and Kalaba [1959] offer the earliest formulation of Bayesian reinforcement learning, whereby the individual actions of a decision-making agent not only provide an update to the physical state\footnote{For ease of exposition, we focus on uncertainty in transition dynamics although one could model uncertainty in either the reward function or the full MDP model (rewards & transitions).}.
of the world but also impact the agent’s internal model of how the world operates. Dayan and Sejnowski [1996] follow this line of thinking to derive implicit exploration bonuses based on how an agent performs posterior updates. Kolter and Ng [2009] make this more explicit and incorporate a specific visitation-based bonus that decays with the concentration of the agent’s Dirichlet posterior. As an alternative, Sorg et al. [2010] incorporate an exploration bonus based on the variance of the agent’s posterior while Araya-López et al. [2012] achieve optimistic exploration by boosting transition probabilities. Duff and Barto [1997] identify multi-armed bandits (that is, MDPs with exactly one state and arbitrarily many actions) as a unique setting where the Bayes-optimal solution is computationally tractable through the use of Gittins indices [Gittins, 1979]. While the vast space of more complicated BAMDPs is computationally intractable, a goal of this paper is to add a bit of nuance and clarify when one might still hope to recover efficient, approximate planning. This is also distinct from the PAC-BAMDP framework introduced by Kolter and Ng [2009], which serves as a characterization of algorithmic efficiency, rather than problem hardness.

Representing uncertainty in the optimal value function rather than environment transition function, Dearden et al. [1998] derive a practical Bayesian Q-learning algorithm by foregoing representation of the epistemic state and instead resampling Q*-values at each timestep. Strens [2000] finds an alternate, tractable solution by lazily updating the epistemic state at the frequency of whole episodes, rather than individual timesteps; a long line of work [Agrawal and Jia, 2017, Osband et al., 2016a,b, Osband and Van Roy, 2017, O'Donoghue et al., 2018, Osband et al., 2019] analyzes this type of approximation to the Bayesian reinforcement-learning problem theoretically and also explores how to scale these solution concepts with deep neural networks.

Duff [2001] finds tractability in representing policies as finite-state stochastic automata, noting structural similarities between BAMDPs and partially-observable MDPs (POMDPs) [Kaelbling et al., 1998]; this type of thinking is further extended by Poupart et al. [2006] who exploit similar structure between the optimal value functions of BAMDPs and POMDPs. Duff [2003a] examine improved memory requirements when applying actor-critic algorithms [Konda and Tsitsiklis, 2000] to BAMDPs while Duff [2003b] consider how to approximately model the stochastic of the evolving epistemic state via diffusion models. Wang et al. [2005] introduce a sparse-sampling approach [Kearns et al., 2002] for balancing computational efficiency against fidelity to Bayes-optimal action selection. An analogous sparse-sampling approach is also developed by Castro and Precup [2007], but with a linear-programming methodology for value-function approximation. A line of work [Guez et al., 2012, 2013, 2014] develops more scalable, sparse-sampling lookahead approaches on the back of Monte-Carlo tree search [Kocsis and Szepesvári, 2006]; these algorithms are somewhat similar in spirit to the approach of Asmuth et al. [2009] who merge multiple posterior samples into a single model while Guez et al. [2014] keep each sample distinct and integrate out the posterior randomness. For a more complete and detailed survey of Bayesian reinforcement learning, we refer readers to Ghavamzadeh et al. [2015]. Crucially, the aforementioned approaches largely revolve around ignoring the epistemic state, lazily updating the epistemic state, or approximating the impact of the epistemic state via random sampling. In contrast, this work offers a new approach and highlights how lossy compression of the epistemic state may naturally reduce BAMDP hardness. Perhaps the most related prior work is by Lee et al. [2018] who introduce a practical approximate-planning approach by quantizing the epistemic state space; this paper clarifies the theoretical ramifications of this quantization step.

Our work is also connected to analyses of approximate value iteration [Bellman, 1957] in the MDP setting [Tseng, 1990, Littman et al., 1995], where more recent work has managed to recover improved sample complexity bounds for approximate value iteration [Sidford et al., 2018b,a]. Like Kearns and Singh [1999], our algorithms utilize exact value iteration almost as a black box and it is an open question for future work to see if similar ideas and proof techniques for these approximate variants might be leveraged in the BAMDP setting. Crucially, the variants of value iteration introduced in this work are merely a backdrop for illustrating the utility of our complexity measure and, more generally, a regard for underlying information structure in BAMDPs.

The particular class of epistemic state abstraction introduced and studied in this work revolves around the covering number of the epistemic state space. Curiously, this deepens an existing connection between BAMDPs and POMDPs [Duff, 2001], where a line of work establishes the covering number of the belief state space as a viable complexity measure for the latter both in theory [Hsu et al., 2007] and in practice [Zhang et al., 2012]. In a less related but similar vein, Kakade et al. [2003a] establish a provably-efficient reinforcement-learning algorithm when the MDP state space is a metric space;
their corresponding sample complexity guarantee depends on the covering number of the state space under the associated metric. Our planning complexity result for abstract BAMDPs mirrors those established by these works in its dependence on the covering number of the epistemic state space.

4 The Complexity of BAMDP Planning

In this section, we examine the difficulty of solving BAMDPs through the lens of a classic planning algorithm: value iteration [Bellman, 1957]. Due to space constraints, we relegate pseudocode for all discussed algorithms to Appendix A. We begin with an quick review of the traditional algorithm applied to our setting before introducing the information horizon as a complexity measure for BAMDPs. This quantity gives rise to a more efficient planning algorithm for BAMDPs that waives excessive dependence on the original problem horizon.

Assumption 2. We assume the existence of a suitable, fixed quantization of simplex $\tilde{\Delta}(\Theta) \subset \Delta(\Theta)$ where $|\tilde{\Delta}(\Theta)| < \infty$ such that the BAMDP hyperstate space $\mathcal{X} = \mathcal{S} \times \tilde{\Delta}(\Theta)$ is finite, $|\mathcal{X}| < \infty$.

4.1 Naive Value Iteration

To help build intuitions, we begin by presenting a typical version of value iteration for finite-horizon BAMDPs as Algorithm 1. This algorithm iterates backwards through the $H$ timesteps, computing $(Q^\pi_h)_b$ across every hyperstate-action pair. With the provision of our posterior update oracle $\mathcal{B}$, we avoid a square dependence on the hyperstate space $(|\mathcal{X}|^2)$ and instead only require $O(|\mathcal{S}| |\Theta|)$ to compute next-state value. Consequently, the resulting planning complexity of Algorithm 1 is $O(|\mathcal{X}| |\mathcal{A}| |\mathcal{S}| |\Theta| H)$. Clearly, this represents an onerous burden for two distinct reasons: (1) we are forced to contend with a potentially very large horizon $H$ and (2) we must also search through the entirety of the hyperstate space, $\mathcal{X}$. In the sections that follow, we alleviate the burdens of challenges (1) and (2) in series, using our new notion of information horizon to mitigate the impact of $H$ and leveraging epistemic state abstraction to further reduce the role of $|\mathcal{X}|$, where the latter occurs at the cost of introducing approximation error.

4.2 Information Horizon

As noted in the previous section, our planning complexity suffers from its dependence on the BAMDP horizon $H$. A key observation, however, is that once an agent has completely resolved its uncertainty and identified one of the $|\Theta|$ environments, all that remains is to deploy the optimal policy for that particular MDP. As an exaggerated but illustrative example of this, consider a BAMDP where any action executed at the first timestep completely identifies the true environment $\Theta \in \Theta$. With no residual epistemic uncertainty left, the Bayes-optimal policy would now completely coincide with the optimal policy and take actions without changing the epistemic state since, at this point, the agent has acquired all the requisite information about the previously unknown environment. Even if the problem horizon $H$ is substantially large, a simple BAMDP like the one described should be fairly easy to solve as epistemic uncertainty is so easily diminished and information is quickly exhausted; it is this principle that underlies our hardness measure.

Let $\pi$ be an arbitrary non-stationary policy. For any hyperstate $x \in \mathcal{X}$, we denote by $\mathbb{P}_\pi(x_h = x)$ the probability that policy $\pi$ visits hyperstate $x$ at timestep $h$. With this, we may define the reachable hyperstate space of policy $\pi$ at timestep $h \in [H]$ as $\mathcal{X}_h^\pi = \{x \in \mathcal{X} \mid \mathbb{P}_\pi(x_h = x) > 0\} \subset \mathcal{X}$. In words, the reachable hyperstate space of a policy $\pi$ at a particular timestep is simply the set of all possible hyperstates that may be reached by $\pi$ at that timestep with non-zero probability. Recall that for any hyperstate $x = (s, p) \in \mathcal{X}$, the epistemic state $p \in \Delta(\Theta)$ is a (discrete) probability distribution, for which we may denote its corresponding entropy as $\mathbb{H}(p)$. Given a BAMDP, we define the information horizon of a policy $\pi$ as $\mathcal{I}(\pi) = \inf \{h \in [H] \mid \forall x_h = (s_h, p_h) \in \mathcal{X}_h^\pi, \mathbb{H}(p_h) = 0\}$. The information horizon of a policy, if it exists, identifies the first timestep in $[H]$ where, regardless of precisely which hyperstate is reached by following $\pi$ at this timestep, the agent has fully resolved all of its epistemic uncertainty over the environment $\Theta$. At this point, we call attention back to our structural Assumption 1 for BAMDPs and note that, under the standard parameterization of epistemic
The key insight from the previous section is that, when the information horizon exists and once an
4.3 Informed Value Iteration
number of state-action pair visitations.
approximately-optimal value function.
next section considers one path for reducing this burden at the cost of only being able to realize an
this is likely inescapable for the problem of computing the exact optimal BAMDP value function, the
voids Assumption 2, still renders both Algorithms 1 and 2 as computationally intractable. Since
guarantee bears an unfavorable dependence on the size of the hyperstate space
dramatically reduce the computational complexity of planning. Still, however, the corresponding
In this section, we’ve demonstrated how the information horizon of a BAMDP has the potential to
standard value iteration procedure of Algorithm 1.
I
characteristic complexity of
O
complexity that scales the hyperstate space size by a potentially large problem horizon, we incur a
information horizon, the agent has no uncertainty in
θ
epistemic state
b
that in Line 9, we could also compute the corresponding
′
, however, we use this calculation to make explicit the fact that, by definition of the
information horizon of the BAMDP
I
non-stationary policy class:
I
to compute the
information horizon of the BAMDP
I
non-stationary policy class:
I
to compute the
information horizon of the BAMDP
I
non-stationary policy class:
I
to compute the
information horizon of the BAMDP
I
non-stationary policy class:
I
epistemic uncertainty through differential entropy which, unlike the traditional (Shannon) entropy
H(·), is potentially negative and has no constant lower bound [Cover and Thomas, 2012].
2 Naturally, to compute the information horizon of the BAMDP, we need only take the supremum across the
non-stationary policy class: 
I = \sup_{\pi \in \Pi^H} I(\pi), where \Pi = \{X \rightarrow A\}.
Clearly, when it exists, we have that 1 ≤ I ≤ H; the case where I = 1 corresponds to having a
prior \rho(\theta) that is itself a Dirac delta \delta_\theta centered around the true environment, in which case, \theta is
known completely and the agent may simply compute and deploy the optimal policy for the MDP
(S, A, R, \theta, H). At the other end of the spectrum, an information horizon I = H suggests that,
in the worst case, an agent may need all H steps of behavior in order to fully identify the environment.
In the event that there exists any single non-stationary policy \pi for which the infimum of I(\pi) does
not exist (that is, I(\pi) = \infty), then clearly I = \infty; this represents the most difficult, worst-case
scenario wherein an agent may not always capable of fully resolving its epistemic uncertainty within
the specified problem horizon H. For certain scenarios, the supremum taken over the entire non-
stationary policy class may be exceedingly strict and, certainly, creates a computational intractability
should one wish to operationalize the information horizon algorithmically; in these situations, it
may be more natural to consider smaller or regularized policy classes (for instance, the collection of
expressible policies under a chosen neural network architecture) that yield more actionable notions
of BAMDP complexity. We now go on to show how the information horizon can be used to design
a more efficient BAMDP planning algorithm whose planning complexity bears a more favorable
dependence on H.
4.3 Informed Value Iteration
The key insight from the previous section is that, when the information horizon exists and once an
agent has acted for I timesteps, the Bayes-optimal policy necessarily falls back to the optimal policy
associated with the true environment. Consequently, if the solutions to all |\Theta| possible underlying
MDPs are computed up front, an agent can simply backup their optimal values starting from the Ith
timestep, rather than backing up values beginning at the original horizon H. This high-level idea is
implemented as Algorithm 2 which assumes access to a sub-routine mdp_value_iteration that
consumes a MDP and produces the associated optimal value function for the initial timestep, V^*_1.
Since the underlying unknown MDP is one of |\Theta| possible MDPs, Algorithm 2 proceeds by first
computing the optimal value function associated with each of them in sequence using standard value
iteration, incurring a time complexity of O(|\Theta||S||A|(H - I)). Note that the horizon of each MDP
is reduced to H - I acknowledging that, after identifying the true MDP in I steps, an agent has only
H - I steps of interaction remaining with the environment. With these |\Theta| solutions in hand, the
remainder of the algorithm proceeds with standard value iteration for BAMDPs (as in Algorithm 1),
only now bootstrapping value from the I timestep, rather than the original problem horizon H. Note
that in Line 9, we could also compute the corresponding \bar{\theta} in question by taking the mean of the next
epistemic state \bar{\theta}', however, we use this calculation to make explicit the fact that, by definition of the
information horizon, the agent has no uncertainty in \theta at this point. As a result, instead of planning
complexity that scales the hyperstate space size by a potentially large problem horizon, we incur a
complexity of O(|\Theta||S||A|(|X|I + |S|(H - I))). Naturally, as the gap between the information
horizon I and problem horizon H increases, the more favorably Algorithm 2 performs relative to the
standard value iteration procedure of Algorithm 1.
In this section, we’ve demonstrated how the information horizon of a BAMDP has the potential to
dramatically reduce the computational complexity of planning. Still, however, the corresponding
guarantee bears an unfavorable dependence on the size of the hyperstate space X which, in the reality
that voids Assumption 2, still renders both Algorithms 1 and 2 as computationally intractable. Since
this is likely inescapable for the problem of computing the exact optimal BAMDP value function, the
next section considers one path for reducing this burden at the cost of only being able to realize an
approximately-optimal value function.

2 Prior work (see, for example, Theorem 1 of Kolter and Ng [2009]) operating with the Dirichlet parameteri-
ation will make an alternative assumption for similar effect where epistemic state updates cease after a certain
number of state-action pair visitations.
5 Epistemic State Abstraction

5.1 State Abstraction in MDPs

As numerous sample-efficiency guarantees in reinforcement learning [Kearns and Singh, 2002, Kakade et al., 2003b, Strehl et al., 2009] bear a dependence on the size of the MDP state space, $|S|$, a large body of work has entertained state abstraction as a tool for improving the dependence on state space size without compromising performance [Whitt, 1978, Bertsekas et al., 1988, Singh et al., 1995, Gordon, 1995, Tsitsiklis and Van Roy, 1996, Dean and Givan, 1997, Ferns et al., 2004, Jong and Stone, 2005, Li et al., 2006, Van Roy, 2006, Ferns et al., 2012, Jiang et al., 2015a, Abel et al., 2016, 2018, 2019, Dong et al., 2019, Du et al., 2019, Misra et al., 2020, Arumugam and Van Roy, 2020, Abel, 2020]. Broadly speaking, a state abstraction $\phi: S \rightarrow S_\phi$ maps original or ground states of the MDP into abstract states in $S_\phi$. Typically, one takes $\phi$ to be defined with respect to an abstract state space $S_\phi$ with smaller complexity (in some sense) than $S$; in the case of state aggregation where all spaces in question are finite, this desideratum often takes the very simple form of $|S_\phi| < |S|$. Various works have identified conditions under which specific classes of state abstractions $\phi$ yield no approximation error and perfectly preserve the optimal policy of the original MDP [Li et al., 2006], as well as conditions under which near-optimal behavior is preserved [Van Roy, 2006, Abel et al., 2016]. As its name suggests, our proposed notion of epistemic state abstraction aims to lift these kinds of guarantees for MDPs over to BAMDPs and contend with the intractable hyperstate space.

Before examining BAMDPs, we provide a brief overview of how state abstraction impacts the traditional MDP, as a point of comparison with the BAMDP setting. Given a MDP $(S, A, R, T, \beta, H)$, a state abstraction $\phi: S \rightarrow S_\phi$ induces a new abstract MDP $M_\phi = (S_\phi, A, R_\phi, T_\phi, H)$ where the abstract reward function $R_\phi: S_\phi \times A \rightarrow [0, 1]$ and transition function $T_\phi: S_\phi \times A \rightarrow \Delta(S_\phi)$ are both defined with respect to a fixed, arbitrary weighting function $\omega: S \rightarrow [0, 1]$ that, intuitively, measures the contribution of each individual MDP state $s \in S$ to its allocated abstract state $\phi(s)$. More specifically, $\omega$ is required to induce a probability distribution on the constituent MDP states of each abstract state: $\forall s_\phi \in S_\phi$, $\sum_{s \in \phi^{-1}(s_\phi)} \omega(s) = 1$. This fact allows for well-defined rewards and transition probabilities as given by

\[ R_\phi(s_\phi, a) = \sum_{s \in \phi^{-1}(s_\phi)} R(s, a)\omega(s), \quad T_\phi(s'_\phi \mid s_\phi, a) = \sum_{s \in \phi^{-1}(s_\phi)} \sum_{s' \in \phi^{-1}(s'_\phi)} T(s' \mid s, a)\omega(s). \]

As studied by Van Roy [2006], the weighting function $\omega$ does bear implications on the efficiency of learning and planning. Naturally, one may go on to apply various planning or reinforcement-learning algorithms to $M_\phi$ and induce behavior in the original MDP $M$ by first applying $\phi$ to the current state $s \in S$ and then leveraging the optimal abstract policy or abstract value function of $M_\phi$. Conditions under which $\phi$ will induce a MDP $M_\phi$ that preserves optimal or near-optimal behavior are studied by Li et al. [2006], Van Roy [2006], Abel et al. [2016].

5.2 Compressing the Epistemic State Space

In this section, we introduce epistemic state abstraction for BAMDPs with the goal of paralleling the benefits of state abstraction in MDPs. In particular, we leverage the fact that our epistemic state space $\Delta(\Theta) = \Delta^{(|\Theta| - 1)}$ is the $(|\Theta| - 1)$-dimensional probability simplex. Recall that for any set $Z$; any threshold parameter $\delta > 0$; and any metric $\rho: Z \times Z \rightarrow \mathbb{R}^+$ on $Z$, a set $\{z_1, z_2, \ldots, z_K\}$ is a $\delta$-cover of $Z$ if $\forall z \in Z$, $\exists i \in [K]$ such that $\rho(z, z_i) \leq \delta$. In this work, we will consider $\delta$-covers with arbitrary parameter $\delta > 0$ defined on the simplex $\Delta^{(\Theta)}$ with respect to the total variation distance metric on probability distributions, denoted $\| \cdot \|_{TV}$. Let $e_i \in \Delta(\Theta)$ be the $i$th standard basis vector such that $\|e_i\| = 0, \forall i \in [|\Theta|]$. We define an epistemic state abstraction with parameter $\delta > 0$ as the projection from $\Delta(\Theta)$ onto the smallest $\delta$-cover of $\Delta(\Theta)$ with respect to $\| \cdot \|_{TV}$ that contains all standard basis vectors $\{e_1, e_2, \ldots, e_{|\Theta|}\}$; paralleling notation for the $\delta$-covering number, we use $\mathcal{N}(|\Theta|, \delta, \| \cdot \|_{TV})$ to denote the size of this minimal cover and, for consistency with the state-abstraction literature in MDPs, use $\phi: \Delta(\Theta) \rightarrow \Delta_{\phi}(\Theta)$ to denote the epistemic state abstraction. Briefly, we note that while computing exact $\delta$-covers is a NP-hard problem, approximation algorithms do exist [Hochbaum, 1996, Zhang et al., 2012]; our work here is exclusively concerned with establishing theoretical guarantees that warrant further investigation of such approximation techniques to help solve BAMDPs in practice.
It is important to note that while there are numerous statistical results expressed in terms of covering numbers (for instance, Dudley’s Theorem [Dudley, 1967]), our definition of covering number differs slightly in its inclusion of the standard basis vectors. The simple reason for this constraint is that it ensures we may still count on the existence of abstract epistemic states for which an agent has fully exhausted all epistemic uncertainty in the underlying environment. Consequently, we are guaranteed that the information horizon is still a well-defined quantity under this abstraction\(^1\). As \( \delta \) increases, larger portions of the epistemic state space where the agent has residual, but still non-zero, epistemic uncertainty will be immediately mapped to the nearest standard basis vector under \( \phi \). If such a lossy compression is done too aggressively, the agent’s beliefs over the uncertain environment may prematurely and erroneously converge. On the other hand, if done judiciously with a prudent setting of \( \delta \), one has the potential to dramatically reduce the complexity of planning across a much smaller, finite hyperstate space and recover an approximately-optimal BAMDP value function.

To make this intuition more precise, consider an initial BAMDP \( \langle \mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \mathcal{R}, H \rangle \) and, given an epistemic state abstraction \( \phi : \Delta(\Theta) \to \Delta(\Theta) \) with fixed parameter \( \delta > 0 \), we recover an induced abstract BAMDP \( \langle \mathcal{X}_\phi, \mathcal{A}, \mathcal{R}, \mathcal{T}_\phi, \mathcal{R}_\phi, H \rangle \) where, most importantly, \( \mathcal{X}_\phi = S \times \Delta(\Theta) \). Just as in the MDP setting, the model of the abstract BAMDP depends on a fixed, arbitrary weighting function of the original epistemic states \( \omega : \Delta(\Theta) \to [0, 1] \) that adheres to the constraint: \( \forall p_\phi \in \Delta(\Theta), \int_{\phi^{-1}(p_\phi)} \omega(p)dp = 1 \), which means abstract rewards and transition probabilities for a current and next abstract epistates, \( x_\phi = \langle s, p_\phi \rangle \) and \( x'_\phi = \langle s', p'_\phi \rangle \), are given by

\[
\mathcal{R}_\phi(x_\phi, a) = \int_{\phi^{-1}(p_\phi)} \mathcal{R}(x, a)\omega(p)dp = \int_{\phi^{-1}(p_\phi)} \mathcal{R}(s, a)\omega(p)dp = \mathcal{R}(s, a),
\]

\[
\mathcal{T}_\phi(x'_\phi \mid x_\phi, a) = \int_{\phi^{-1}(p_\phi)} \mathcal{T}(x' \mid x, a)dp,
\]

where \( x = \langle s, p \rangle \) and \( x' = \langle s', p' \rangle \).

The initial state abstraction distribution is defined as \( \Gamma_\phi = \beta \times \delta(\rho(\theta)) \) where \( \beta \in \Delta(S) \) denotes the initial state distribution of the underlying MDP while \( \delta(\rho(\theta)) \) is a Dirac delta centered around the agent’s original prior, \( \rho(\theta) \), projected by \( \phi \) into the abstract epistemic state space. Observe that the abstract BAMDP transition function is stochastic with respect to the next abstract epistemic state \( p'_\phi \), unlike the original BAMDP transition function whose next epistemic states are deterministic. This is, perhaps, not a surprising observation as it also occurs in standard state aggregation of deterministic MDPs as well. Nevertheless, it is important to note the corresponding abstract BAMDP value functions must now acknowledge this stochasticity for any abstract policy \( \pi = (\pi_{\phi,1}, \pi_{\phi,2}, \ldots, \pi_{\phi,H}) \), \( \pi_{\phi,h} : \mathcal{X}_\phi \to \mathcal{A}, \forall h \in [H] \):

\[
V^\pi_{\phi,h}(x_\phi) = Q^\pi_{\phi,h}(x_\phi, \pi_{\phi,h}(x_\phi)) \quad V^*_\phi,h(x_\phi) = \max_{a \in \mathcal{A}} Q^*_\phi,h(a, x_\phi, a), \quad V^\pi_{\phi,H+1}(x_\phi) = 0 \quad V^*_\phi,H+1(x_\phi) = 0 \quad \forall x_\phi \in \mathcal{X}_\phi,
\]

\[
Q^\pi_{\phi,h}(x_\phi, a) = \mathcal{R}(s, a) + \sum_{s', \phi'} \mathcal{T}_\phi(x'_\phi \mid x_\phi, a)V^\pi_{\phi,h+1}(x'_\phi), \quad Q^*_\phi,h(x_\phi, a) = \mathcal{R}(s, a) + \sum_{s', \phi'} \mathcal{T}(x'_\phi \mid x_\phi, a)V^*_\phi,h+1(x'_\phi).
\]

Beyond the fact that this abstract BAMDP enjoys a reduced hyperstate space, we further observe that the information horizon of this new BAMDP, \( \mathcal{I}_\phi \), has the potential to be smaller than that of the original BAMDP. That is, if \( \mathcal{I} \) steps are needed to fully resolve epistemic uncertainty in the original BAMDP then, by compressing the epistemic state space via \( \phi \), we may find epistemic uncertainty exhausted in fewer than \( \mathcal{I} \) timesteps within the abstract BAMDP. Furthermore, for a suitably large setting of the \( \delta \) parameter, we also have cases where the original BAMDP has \( \mathcal{I} = \infty \) while \( \mathcal{I}_\phi < \infty \); in words, whereas it may not have been possible to resolve all epistemic uncertainty within \( \mathcal{H} \) timesteps, compression of the epistemic state space reduces this difficulty in the abstract problem as knowledge states near (in the total-variation sense) each vertex of the probability simplex \( \varepsilon_i \) are immediately aggregated. Due to space constraints, we defer further discussion of the abstract information horizon and its relationship with the original information horizon to Appendix C.

\(^1\)Note that an alternative would be to introduce an additional constant \( \gamma \in \mathbb{R}^+ \) and define the information horizon based on \( \mathcal{H}(p) = \gamma \); our construction avoids carrying this cumbersome additional parameter dependence in the results.

\(^2\)One could also imagine abstracting over the original MDP state space \( S \) which, for clarity, we do not consider in this work.
As a toy illustration of last scenario, consider a $\phi$ with $\delta$ sufficiently large such that any step from the agent’s prior distribution immediately maps to a next abstract hyperstate with no epistemic uncertainty. Clearly, regardless of $I$, we have an abstract BAMDP where $I_\phi = 2$. Of course, under such an aggressive abstraction, we should expect to garner an unfavorable degree of approximation error between the solutions of the abstract and original BAMDPs. The next section makes this error analysis and performance loss precise alongside an approximate planning algorithm that leverages the reduced complexity of abstract BAMDPs to recover a near-optimal solution to the original BAMDP.

5.3 Informed Abstract Value Iteration

Observe that if, after inducing the abstract BAMDP according to a given epistemic state abstraction $\phi$, the resulting information horizon is finite $I_\phi < \infty$, then we are in a position to run Algorithm 2 on the abstract BAMDP. Moreover, we no longer need the crutch of Assumption 2 as, by definition of $\phi$, we are guaranteed a finite abstract hyperstate space of size $|\mathcal{X}_\phi| = |\mathcal{S}| \cdot \mathcal{N}(\Delta(\Theta), \delta, || \cdot ||_{TV})$. With the solution to the abstract BAMDP in hand, we can supply values to any input hyperstate of the original BAMDP $x = (s, p) \in \mathcal{X}$ by simply applying $\phi$ to the agent’s current epistemic state $p$ and querying the value of the resulting abstract hyperstate $(s, \phi(p)) \in \mathcal{X}_\phi$. We present this approximate BAMDP planning procedure as Algorithm 3.

By construction, this algorithm inherits the planning complexity guarantee of Algorithm 2, specialized to the abstract BAMDP input, yielding $O ([|\Theta||S|^2|A|] (\mathcal{N}(\Delta(\Theta), \delta, || \cdot ||_{TV})^2 I_\phi + (H - I_\phi)))$. A key feature of this result is that we entirely forego a (direct) dependence on the hyperstate space of the original BAMDP and, instead, take on dependencies with the size of the abstract hyperstate space, $|\mathcal{X}_\phi|^2 = |\mathcal{S}|^2 \mathcal{N}(\Delta(\Theta), \delta, || \cdot ||_{TV})^2$, and the abstract information horizon $I_\phi$. While both terms decrease as $\delta \rightarrow 1$, there is a delicate balance to be maintained between the ease with which one may solve the abstract BAMDP and the quality of the resulting solution when deployed in the original BAMDP of interest. We dedicate the remainder of this section to making this balance mathematically precise. Due to space constraints, all proofs are relegated to Appendix B. A natural first step in our analysis is to establish an approximation error bound:

**Proposition 1.** Let $V^*_h$ and $V^*_{\phi,h}$ denote the optimal original and abstract BAMDP value functions, respectively, for any timestep $h \in [H]$. Let $\phi$ be an epistemic state abstraction as defined above. Then, $\max_{x \in \mathcal{X}} |V^*_h(x) - V^*_{\phi,h}(\phi(x))| \leq 2\delta(H - h)(H - h + 1)$.

In order to establish a complimentary performance-loss bound, we require an intermediate result characterizing performance shortfall of a BAMDP value function induced by a greedy policy with respect to another near-optimal BAMDP value function. The analogue of this result for MDPs is proven by Singh and Yee [1994], and the proof for BAMDPs follows similarly.

**Proposition 2.** Let $V = \{V_1, V_2, \ldots, V_H\}$ be an arbitrary BAMDP value function. We denote by $\pi_{h,V}$ the greedy policy with respect to $V$ defined $\forall x = (s, p) \in \mathcal{X}$ as

$$\pi_{h,V}(x) = \arg \max_{a \in \mathcal{A}} \mathcal{R}(x, a) + \sum_{\delta, s'} \mathcal{T}_\theta(s' | s, a)p(\theta)V_{h+1}(x'),$$

where $x' = (s', B(p, s, a, q)) \in \mathcal{X}$. Recall that $V^*_h$ denotes the optimal BAMDP value function at timestep $h + 1$ and $\pi^*_h$ denote the Bayes-optimal policy. If for all $h \in [H]$, for all $s \in \mathcal{S}$, and for any $p, q \in \Delta(\Theta)$ $|V^*_h((s, p)) - V_h((s, q))| \leq \varepsilon$, then $||V^*_h - V_{h}^{\pi^*_h,V}||_\infty \leq 2\varepsilon(H - h + 1)$.

Combining Propositions 1 and 2 immediately yields a corresponding performance-loss bound as desired, paralleling the analogous result for state aggregation in MDPs (see Theorem 4.1 of Van Roy [2006]):

**Proposition 3.** Let $\pi^*_{\phi,h}$ denote the greedy policy with respect to $V^*_{\phi,h}$. Then, $||V^*_h - V_{h}^{\pi^*_{\phi,h}}||_\infty \leq 4\delta(H - h)(H - h + 1)^2$.

6 Discussion & Conclusion

In this work, we began by characterizing the complexity of a BAMDP via an upper bound on the total number of interactions needed by an agent to exhaust information and fully resolve its
epistemic uncertainty over the true environment. Under an assumption on the exact form of the agent’s uncertainty, we showed how this information horizon facilitates more efficient planning when smaller than the original problem horizon. We recognize that Assumption 1 deviates from the traditional parameterization of uncertainty in the MDP transition function via the Dirichlet distribution [Poupart et al., 2006, Kolter and Ng, 2009] (sometimes also known as the flat Dirichlet-Multinomial or FDM model [Asmuth, 2013]). The driving force behind this choice is to avoid dealing in differential entropy when engaging with the (residual) uncertainty contained in any epistemic state. Should one aspire to depart from Assumption 1 altogether in a rigorous way that manifests within the analysis, we suspect that it may be fruitful to consider a lossy compression of each epistemic state into a discrete, $|\Theta|$-valued random variable. Under such a formulation, the appropriate tool from information theory for the analysis would be rate-distortion theory [Cover and Thomas, 2012, Csiszár, 1974]. This would, in a theoretically-sound way, allow for an arbitrary BAMDP parameterization and, for the purposes of continuing the use (discrete) Shannon entropy in the definition of the information horizon, induce a lossy compression of each epistemic state whose approximation error relative to the true epistemic state could be accounted for via the associated rate-distortion function.

Recognizing the persistence of the intractable BAMDP hyperstate space, we then proceeded to outline epistemic state abstraction as a mechanism that not only induces a finite, tractable hyperstate space but also has the potential to incur a reduced information horizon within the abstract problem. Through our analysis of approximation error and performance loss, we observe an immediate consequence of Proposition 3: if one wishes to compute an $\varepsilon$-optimal BAMDP value function for an original BAMDP of interest, one need only find the $\frac{4(H−h)(H−h+1)}{H}$-cover of the simplex, $\Delta(\Theta)$, and then apply the corresponding epistemic state abstraction through Algorithm 3, whose planning complexity bears no dependence on the hyperstate space of the original BAMDP and has reduced dependence on the problem horizon. One might observe that the right-hand side of the value-loss bound is maximized at timestep $h = 1$, making this first step the limiting factor when determining what value of $\delta$ to employ for computing the epistemic state abstraction. As this cover could become quite large and detract from the efficiency of utilizing an epistemic state abstraction in subsequent time periods, future work might benefit from considering abstractions formed by a sequence of exactly $H \delta_h$-covers, where the indexing of $\delta_h$ in time $h \in [H]$ affords better preservation of value (via a straightforward extension of Proposition 3) across all timesteps simultaneously.

One caveat and limitation of our planning algorithms (both exact and approximate) is the provision of the information horizon as an input. An agent designer may seldom have the prescience of knowing the underlying BAMDP information structure or, even with suitable regularity assumptions on the policy class, be able to compute it. An observation is that many sampling-based algorithms for approximately solving BAMDPs, like BAMCP [Guez et al., 2012], implicitly hypothesize a small information horizon (typically, a value of 1) through their use of posterior sampling and choice of rollout policy. Meanwhile, recent work has demonstrated strong performance guarantees for a reinforcement-learning agent acting in an arbitrary environment [Dong et al., 2022] through the use of an incrementally increasing discount factor [Jiang et al., 2015b, Arumugam et al., 2018], gradually expanding the effective range over which the agent is expected to demonstrate competent behavior. Taking inspiration from this idea, future work might consider designing more-efficient planning algorithms that, while ignorant of the true information horizon, instead hypothesize a sequence of increasing information horizons, eventually building up to the complexity of the full BAMDP. Of course, prior to development of novel algorithms, the notion that existing BAMDP planners may already make implicit use of the information horizon is in and of itself a task for future work to tease apart and make mathematically rigorous.

Moreover, similar to how the simulation lemma [Kearns and Singh, 2002] provides a principled foundation for model-based reinforcement learning, our analysis might also be seen as offering theoretical underpinnings to the Bayes-optimal exploration strategies learned by meta reinforcement-learning agents [Ortega et al., 2019, Mikulik et al., 2020] whose practical instantiations already rely upon approximate representations of epistemic state [Zintgraf et al., 2019, 2021].

Acknowledgements

The authors gratefully acknowledge the anonymous reviewers for their insightful comments, questions, and discussions.
References


Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning. 1998. 2


Zongzhang Zhang, Michael Littman, and Xiaoping Chen. Covering number as a complexity measure for POMDP planning and learning. In Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012. 2, 4, 7


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
Here we present all algorithms discussed in the main paper.

**Algorithm 1** Value Iteration for BAMDPs

1: **Input:** BAMDP $\langle X, A, R, T, \beta, H \rangle$
2: $V_{H+1}^\star(\langle s, p \rangle) = 0, \forall (s, p) \in X$
3: for $h = H, H - 1, \ldots, 1$ do
4: for $(s, p) \in X$ do
5: for $a \in A$ do
6: $Q_h^\star(\langle s, p \rangle, a) = R(s, a) + \sum_{s', \theta} T_\theta(s' | s, a)p(\theta)V_{h+1}^\star(\langle s', B(p, s, a, s') \rangle)$
7: end for
8: $V_h^\star(\langle s, p \rangle) = \max_{a \in A} Q_h^\star(\langle s, p \rangle, a)$
9: end for
10: end for

**Algorithm 2** Informed Value Iteration for BAMDPs

1: **Input:** BAMDP $\langle X, A, R, T, \beta, H \rangle$, Information horizon $I < \infty$
2: for $\theta \in \Theta$ do
3: $V_\theta^\star = \text{ndp-value-iteration}(\langle S, A, R, T_\theta, \beta, H - I \rangle)$
4: end for
5: for $h = I - 1, I - 2, \ldots, 1$ do
6: for $(s, p) \in X$ do
7: for $a \in A$ do
8: if $h + 1 == I$ then
9: for $s' \in S$ do
10: $p' = B(p, s, a, s')$
11: $\bar{\theta} = \sum_{\theta \in \Theta} \theta 1(p'(\theta) = 1)$
12: $V_{h+1}(\langle s', p' \rangle) = V_{\bar{\theta}}^\star$
13: end for
14: end if
15: $Q_h^\star(\langle s, p \rangle, a) = R(s, a) + \sum_{s', \theta} T_\theta(s' | s, a)p(\theta)V_{h+1}^\star(\langle s', B(p, s, a, s') \rangle)$
16: end for
17: $V_h^\star(\langle s, p \rangle) = \max_{a \in A} Q_h^\star(\langle s, p \rangle, a)$
18: end for
19: end for
Then, we now leverage the standard trick of adding “zero” by subtracting and adding the following between the first term in isolation, we first observe that, by definition of the weighting function, let \( \Delta \) denote its corresponding abstract hyperstate where \( H \) is the abstract information horizon.

\[
\max_{p} |V^{*}_h(x) - V^{*}_{\phi,h}(\phi(x))| \leq 2\delta(H - h)(H - h + 1).
\]

**Proof.** With a slight abuse of notation, for any hyperstate \( x \in X \), let \( \phi(x) = \langle s, p \rangle \in X_{\phi} \) denote its corresponding abstract hyperstate where \( p = \phi(p) \in \Delta_{\phi}(\Theta) \). For brevity, we define \( p' \triangleq B(p, s, a, s') \). We have

\[
\max_{x \in X} |V^{*}_h(x) - V^{*}_{\phi,h}(\phi(x))| = \max_{(s, p) \in X} | \max_{a \in A} Q^{*}_h((s, p), a) - \max_{a \in A} Q^{*}_{\phi,h}((s, p), a)| \\
\leq \max_{(s, p), a \in X \times A} |Q^{*}_h((s, p), a) - Q^{*}_{\phi,h}((s, p), a)| \\
= \max_{(s, p), a} \left| \sum_{\theta \in \Theta} \theta(s' \mid s, a)p(\theta)\sum_{\phi, p'} \mathcal{T}_{\phi}(\langle s', p' \rangle \mid \langle s, p \rangle, a)V^{*}_{\phi,h+1}((s', p')) \right|
\]

We now leverage the standard trick of adding “zero” by subtracting and adding the following between our two terms before applying the triangle inequality to separate them:

\[
\sum_{s', p'} \mathcal{T}_{\phi}(\langle s', p' \rangle \mid \langle s, p \rangle, a)V^{*}_{\phi,h+1}((s', p')).
\]

Examining the first term in isolation, we first observe that, by definition of the weighting function, \( \int_{\omega^{-1}(p)} \omega(p)dp = 1 \) and so we have

\[
\sum_{\theta, s'} T_{\theta}(s' \mid s, a)p(\theta)V^{*}_{h+1}((s', p')) = \int_{\phi^{-1}(p)} \omega(p)\sum_{\theta, s'} T_{\theta}(s' \mid s, a)p(\theta)V^{*}_{h+1}((s', p'))dp.
\]
Expanding with the definition of the abstract BMDP transition function, we have

$$
\sum_{s',p'_{\phi}} T_{\phi}(\langle s', p'_{\phi} \rangle | \langle s, p_{\phi} \rangle, a) V^*_{h+1}(\langle s', p' \rangle) = \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) \sum_{\theta, s'} T_{\theta}(s' | s, a) \bar{p}(\theta) V^*_{h+1}(\langle s', p' \rangle) \sum_{p'_{\phi}} 1 (B(\bar{p}, s, a, s') = p'_{\phi}) d\bar{p}
$$

$$
= \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) \sum_{\theta, s'} T_{\theta}(s' | s, a) \bar{p}(\theta) V^*_{h+1}(\langle s', p' \rangle) d\bar{p}
$$

since $\phi(B(\bar{p}, s, a, s')) \leq H - h$ and simplifying, we have

$$
\left| \sum_{\theta, s'} T_{\theta}(s' | s, a) p(\theta) V^*_{h+1}(\langle s', p' \rangle) - \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) \sum_{\theta} p(\theta) - \bar{p}(\theta) d\bar{p} \right|
$$

$$
\leq (H - h) \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) \sum_{\theta} \frac{1}{2} |p(\theta) - \bar{p}(\theta)| d\bar{p}
$$

$$
= (H - h) \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) 2 \cdot \frac{1}{2} \sum_{\theta} |p(\theta) - \bar{p}(\theta)| d\bar{p}
$$

$$
= (H - h) \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) 2 \cdot \|p(\theta) - \bar{p}(\theta)\|_{TV} d\bar{p}
$$

$$
\leq 4\delta(H - h) \int_{\phi^{-1}(p_{\phi})} \omega(\bar{p}) d\bar{p}
$$

$$
= 4\delta(H - h),
$$

where the last upper bound follows from the definition of a $\delta$-cover since $\phi(p) = \phi(\bar{p}) = p_{\phi}, \forall \bar{p} \in \phi^{-1}(p_{\phi})$.

Moving on to the second term and applying Jensen’s inequality, we have

$$
\left| \sum_{s',p'_{\phi}} T_{\phi}(\langle s', p'_{\phi} \rangle | \langle s, p_{\phi} \rangle, a) V^*_{h+1}(\langle s', p' \rangle) - \sum_{s',p'_{\phi}} T_{\phi}(\langle s', p'_{\phi} \rangle | \langle s, p_{\phi} \rangle, a) V^*_{h+1}(\langle s', p'_{\phi} \rangle) \right|
$$

$$
\leq \sum_{s',p'_{\phi}} T_{\phi}(\langle s', p'_{\phi} \rangle | \langle s, p_{\phi} \rangle, a) \left| V^*_{h+1}(\langle s', p' \rangle) - V^*_{h+1}(\langle s', p'_{\phi} \rangle) \right|
$$

$$
\leq \max_{x \in X} |V^*_{h+1}(x) - V^*_{\phi, h+1}(\phi(x))|.
$$

Thus, putting everything together, we have established that

$$
\max_{x \in X} |V^*_{h}(x) - V^*_{\phi, h}(\phi(x))| \leq 4\delta(H - h) + \max_{x \in X} |V^*_{h+1}(x) - V^*_{\phi, h+1}(\phi(x))|
$$

Iterating the same sequence of steps for the latter term on the right-hand side $H - h$ more times, we arrive at a final bound

$$
\max_{x \in X} |V^*_{h}(x) - V^*_{\phi, h}(\phi(x))| \leq \sum_{h=1}^{H} 4\delta(H - h) = 4\delta \sum_{h=1}^{H} (H - h) = 4\delta \frac{(H - h)(H - h + 1)}{2} = 2\delta(H - h)(H - h + 1).
$$

$\square$
B.2 Proof of Proposition 2

Proposition 2. Let $V = \{V_1, V_2, \ldots, V_H\}$ be an arbitrary BAMDP value function. We denote by $\pi_{h,V}$ the greedy policy with respect to $V$ defined as

$$\pi_{h,V}(x) = \arg \max_{a \in \mathcal{A}} \left[ R(x, a) + \sum_{\theta, s'} T_0(s' \mid s, a)p(\theta)V_{h+1}(x') \right] \quad \forall x \in \mathcal{X},$$

where $x' = (s', B(p, s, a, s')) \in \mathcal{X}$. Recall that $V^*_h$ denotes the optimal BAMDP value function at timestep $h + 1$ and $\pi^*_h$ denote the Bayes-optimal policy. If for all $h \in [H]$, for all $s \in \mathcal{S}$, and for any $p, q \in \Delta(\Theta)$

$$|V^*_h(s, p) - V^*_h(s, q)| \leq \varepsilon, \text{ then } \|V^*_h - V^{\pi_{h,V}}_h\|_\infty \leq 2\varepsilon(H - h + 1).$$

Proof. Fix an arbitrary timestep $h \in [H]$. For any $x \in \mathcal{X}$, define $a, \bar{a} \in \mathcal{A}$ such that $a = \pi^*_h(x)$ and $\bar{a} = \pi^*_{h,V}(x)$. Similarly, let $p' = B(p, s, a', s')$ and $\bar{p}' = B(p, s, \bar{a}, s')$. Since, by definition, $\pi_{h,V}$ is greedy with respect to $V_{h+1}$, we have that

$$R(x, a) + \sum_{\theta, s'} T_0(s' \mid s, a)p(\theta)V_{h+1}(x') \leq R(x, \bar{a}) + \sum_{\theta, s'} T_0(s' \mid s, \bar{a})p(\theta)V_{h+1}(x').$$

By assumption, we have that

$$V^*_h((s', p')) - \varepsilon \leq V^*_h((s', p')) \leq V^*_h((s', p')) + \varepsilon.$$

Applying both bounds to the above yields

$$R(x, a) + \sum_{\theta, s'} T_0(s' \mid s, a)p(\theta)V^*_h((s', p')) - \varepsilon \leq R(x, \bar{a}) + \sum_{\theta, s'} T_0(s' \mid s, \bar{a})p(\theta)V^*_h((s', p')) + \varepsilon.$$

Consequently, we have that

$$R(x, a) - R(x, \bar{a}) \leq 2\varepsilon + \max_{(s, p)} \left[ T_0(s' \mid s, a)p(\theta)V^*_h((s', p')) - T_0(s' \mid s, \bar{a})p(\theta)V^*_h((s', p')) \right].$$

From this, it follows that

$$\|V^*_h - V^{\pi_{h,V}}_h\|_\infty = \max_{x \in \mathcal{X}} |V^*_h(x) - V^{\pi_{h,V}}_h(x)|$$

$$= \max_{x \in \mathcal{X}} |Q^*_h(x, a) - Q^{\pi_{h,V}}_h(x, \bar{a})|$$

$$= \max_{x \in \mathcal{X}} \left[ R(x, a) - R(x, \bar{a}) + \sum_{\theta} p(\theta) \sum_{s'} T_0(s' \mid s, a)V^*_h((s', p')) - T_0(s' \mid s, \bar{a})V^{\pi_{h,V}}_{h+1}(s', \bar{p}') \right]$$

$$\leq 2\varepsilon + \max_{(s, p)} \left[ \sum_{\theta} p(\theta) \sum_{s'} T_0(s' \mid s, \bar{a})V^{\pi_{h,V}}_{h+1}(s', \bar{p}') \right]$$

$$\leq 2\varepsilon + \|V^*_h - V^{\pi_{h+1,V}}_{h+1}\|_\infty$$

$$\vdots$$

$$\leq 2\varepsilon(H - h + 1).$$

where the last inequality follows by iterating the same procedure for the second term in the penultimate inequality across the remaining $H - h$ timesteps. □

B.3 Proof of Proposition 3

Proposition 3. Let $\pi^*_{\phi,h}$ denote the greedy policy with respect to $V^*_{\phi,h+1}$. Then,

$$\|V^*_h - V^{\pi^*_{\phi,h}}_h\|_\infty \leq 4\delta(H - h)(H - h + 1)^2.$$

Proof. Since, for any $x \in \mathcal{X}$, $\phi(x)$ differs only in the epistemic state, the proof follows by realizing that the $\varepsilon$ term of Proposition 2 is established by Proposition 1. Namely,

$$\|V^*_h - V^{\pi^*_{\phi,h}}_h\|_\infty \leq 2(H - h + 1)\max_{x \in \mathcal{X}} |V^*_h(x) - V^*_h(\phi(x))| \leq 4\delta(H - h)(H - h + 1)^2$$

□
C On the Reduction of the Abstract Information Horizon

In this section, we offer two simple yet illustrative examples of BAMDPs where the use of epistemic state abstraction can either decrease or increase the information horizon of the resulting abstract BAMDP. Taken together with the main results of the paper, these examples underscore how epistemic state abstraction, similar to traditional state abstraction in MDPs, is not a panacea to sample-efficient BAMDP planning. Further investigation is needed to clarify the conditions under which an epistemic state abstraction may actually deliver upon the theoretical benefits outlined in this work.

C.1 Decreased Information Horizon under Epistemic State Abstraction

Consider a MDP whose state space is defined on the non-negative integers \( S = \mathbb{Z}^+ = \{0, 1, 2, \ldots\} \) with two actions \( \mathcal{A} = \{+, o\} \). For the purposes of this example, we ignore the reward function and focus on the transition function where, for any timestep \( h \in [H] \), there is a fixed parameter \( q > \frac{1}{2} \) such that \( s_{h+1} = s_h + \Delta \) with \( \Delta \sim \text{Bernoulli}(q) \) if \( a_h = + \) and \( \Delta \sim \text{Bernoulli}(1 - q) \) if \( a_h = o \). In words, action + has a higher probability of incrementing the agent’s current state by one whereas the \( o \) action is more likely to leave the agent’s state unchanged.

For the corresponding BAMDP, we take \( \Theta = \{\theta_1, \theta_2\} \) where \( \theta_1 \) corresponds to the true MDP transition function as described above. Meanwhile, \( \theta_2 \) simply corresponds to the transition function of \( \theta_1 \) with the actions flipped such that \( \Delta \sim \text{Bernoulli}(q) \) if \( a_h = o \) and \( \Delta \sim \text{Bernoulli}(1 - q) \) if \( a_h = + \). Clearly, when \( q = 1 \), epistemic uncertainty in this BAMDP is resolved immediately by the first transition whereas, with \( q \downarrow \frac{1}{2} \), the two hypotheses become increasingly harder to distinguish, requiring more observed transitions from the environment and potentially exceeding the finite problem horizon \( H \).

Now consider the epistemic state abstraction of \( \Delta(\Theta) \) with parameter \( \delta > 0 \). Increasing \( \delta \uparrow 1 \) is commensurate with setting a \( \gamma \in \mathbb{R}_{>0} \) such that once the entropy of the current epistemic state falls below this threshold \( \mathbb{E}(p_{h}) \leq \gamma \), we immediately have that \( \phi(p_{h}) \in \{e_1, e_2\} \), where \( e_i \) denotes the \( i \)th standard basis vector in \( \Delta(\Theta) \). By construction, any action taken in this MDP necessarily reveals information to help reduce epistemic uncertainty. Consequently, with enough observed transitions from the environment, we can use \( \phi \) to collapse the agent’s beliefs around \( \theta_1 \) with far fewer samples than what would be needed to fully exhaust epistemic uncertainty. Moreover, depending on the exact problem horizon \( H \), this could be used to recover a finite abstract information horizon from what was an infinite information horizon in the original BAMDP.

More concretely, suppose \( H = 3 \) and \( q = \frac{1}{2} \). While any policy will quickly stumble upon \( \theta_1 \) as the most likely outcome, such a short horizon \( H \) likely does not allow the entropy of all epistemic states to fall to zero, potentially yielding an infinite information horizon \( I = \infty \). However, there certainly exists a value of \( \delta \) such that the first two steps of behavior under any policy is sufficient for reaching a vertex state in \( \Delta_{\phi}(\Theta) \) and identifying the underlying MDP, ultimately yielding \( I_{\phi} = 2 \).

C.2 Increased Information Horizon under Epistemic State Abstraction

To show how epistemic state abstraction can work unfavorably and increase the information horizon, we consider a BAMDP where all policies rapidly resolve epistemic uncertainty, but the incorporation of an epistemic state abstraction may compromise the agent’s ability to reach a vertex of the simplex \( \Delta(\Theta) \).

For any \( N \in \mathbb{N} \), consider a single initial state connected to two \( N \)-state chains (an upper chain and a lower chain) with two possible actions \( \mathcal{A} = \{a_1, a_2\} \). For each possible \( \theta \in \Theta \), the corresponding transition function \( T_{\theta} \) will either have \( a_1 \) deterministically transition to the next state in the upper chain and have \( a_2 \) send the agent to the next state in the lower chain, or vice versa. In other words, each transition function \( T_{\theta} \) can be concisely encoded as a length \( N \) binary string \( \in \{0, 1\}^N \) where, for each \( n \in [N] \), the \( n \)th bit equal to 1 implies action \( a_1 \) transitions to the upper chain (immediately implying \( a_2 \) transitions to the lower chain) while a value of 0 signifies the opposite transition structure. Further suppose that, for each \( n \in [N] \), the \( n \)th bit of all but one of the transition functions is identical; said differently, this structural assumptions says that, in the \( n \)th stage of the chain, there is a single informative transition that would uniquely identify the underlying MDP. A consequence of this is that, in the worst case, a BAMDP policy will only see one of the uninformative transitions in each
stage and, therefore, can only eliminate exactly one hypothesis from $\Theta$ with each timestep. Thus, for any problem horizon $H \geq N$, we are guaranteed an information horizon of $N$.

To see how epistemic state abstraction might inhibit planning, consider the policy that misses the informative transition in each stage. For an epistemic state abstraction with $\delta$ sufficiently large, the agent will remain stuck (via a self-looping transition) in the initial abstract epistemic state as no single uninformative transition yields sufficient information gain to move the agent through the abstract epistemic state space. This phenomenon of state abstraction ameliorating generalization while drastically worsening the challenge of exploration has already been observed in the standard MDP setting [Abel et al., 2017, 2020]; here, we see that epistemic state abstraction is also vulnerable to the same weakness. As a result, the agent will never converge to one of the vertices of the simplex and never see its epistemic uncertainty in the underlying environment completely diminished, resulting in $I_o = \infty$. 