Leveraging the Value of Information in POMDP Planning

Zakariya Laouar[1], **Qi Heng Ho**[2], **Zachary Sunberg**[1]

[1]Department of Aerospace Engineering Sciences, University of Colorado Boulder
[2]Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech
zakariya.laouar@colorado.edu, qihengho@vt.edu, zachary.sunberg@colorado.edu

Abstract

Partially observable Markov decision processes (POMDPs) offer a principled formalism for planning under state and transition uncertainty. Despite advances made towards solving large POMDPs, obtaining performant policies under limited planning time remains a major challenge due to the curse of dimensionality and the curse of history. For many POMDP problems, the value of information (VOI) — the expected performance gain from reasoning about observations — varies over the belief space. We introduce a dynamic programming framework that exploits this structure by conditionally processing observations based on the value of information at each belief. Building on this framework, we propose Value of Information Monte Carlo planning (VOIMCP), a Monte Carlo Tree Search algorithm that allocates computational effort more efficiently by selectively disregarding observation information when the VOI is low, avoiding unnecessary branching of observations. We provide theoretical guarantees on the near-optimality of our VOI reasoning framework and derive non-asymptotic convergence bounds for VOIMCP. Simulation evaluations demonstrate that VOIMCP outperforms baselines on several POMDP benchmarks.

Introduction

Decision making in uncertain environments presents a critical challenge. The partially observable Markov decision process (POMDP) framework provides a principled way to plan in such environments (Åström 1965; Smallwood and Sondik 1973). Their applications include robotics (Lauri, Hsu, and Pajarinen 2023), resource management (Papakonstantinou and Shinozuka 2014), human-computer interaction (Chen et al. 2020), and medical diagnosis (Ayer, Alagoz, and Stout 2012). However, finding even approximate solutions for POMDPs is computationally intractable (Madani, Hanks, and Condon 2003) for two main reasons.

Firstly, the computational complexity of the problem scales exponentially with the dimensionality of the state, action, and observation spaces. Secondly, since the state is not directly observable, optimal decisions may depend on the entire history of previous actions and observations, the space of which grows exponentially with the planning horizon. These two fundamental challenges are known as the *curse of dimensionality* and *curse of history*. In this work, we aim to alleviate the curse of history by developing a framework that enables planners to dynamically prune areas of the action-observation histories while maintaining performance guarantees.

To address the challenges of POMDP planning, many practical solvers have been proposed in the past few decades (Shani, Pineau, and Kaplow 2013; Silver and Veness 2010; Kurniawati, Hsu, and Lee 2009; Smith and Simmons 2005; Somani et al. 2013; Sunberg and Kochenderfer 2018). Sampling-based approximate solvers such as Silver and Veness (2010); Sunberg and Kochenderfer (2018); Somani et al. (2013) have made good progress, especially regarding the curse of dimensionality in the state space, by employing particles in search and exploring a subset of histories in the search tree. Nonetheless, problems with large observation spaces still remain relatively difficult for state-of-the-art planners. In such problems, many planners tend to generate shallow trees, which lead to myopic plans that fail to capture long-horizon consequences of such plans.

For many POMDP problems, the value of information (VOI) (Wei 2024; Flaspohler, Roy, and Fisher III 2020), which is the performance gain from fully reasoning about all possible observations from a belief or history, varies significantly over the belief space. This suggests that it may not always be necessary to consider all possible observations at every decision point: when the VOI is small, ignoring observations may incur negligible performance loss while substantially reducing the effective branching on observations.

In this work, we introduce an adaptive VOI framework that formalizes this structure by selectively processing observations based on the VOI at a belief. When VOI is low, this framework disregards uninformative observations. We further derive an upper bound on the suboptimality incurred by this adaptive VOI framework relative to the original POMDP value function. This framework offers a principled mechanism that formalizes the trade-off between open-loop execution (disregarding observations) and closed-loop execution (reasoning over observations). Then, we introduce the VOI-POMDP, a structural transformation that encodes this meta-level choice directly into the problem dynamics.

Leveraging this theoretical framework, we propose value of information Monte Carlo planning (VOIMCP), an MCTS-based POMDP planning algorithm that selectively

branches on observations when useful, enabling more efficient planning for POMDPs with lower value of information by implicitly pruning the history space and reducing the effective branching factor on observations. We prove the non-asymptotic theoretical properties of VOIMCP, and show that we can also recover the optimal POMDP policy. Simulation experiments across a series of benchmarks show that incorporating our VOI framework in online time-constrained planning leads to better performance by enabling long-horizon reasoning.

Related Work

Incorporating the value of information in decision making can be traced back to work on information value theory (Howard 1966). In this line of work, the value of computation is treated such that computations are selected according to the expected improvement in decision quality resulting from their execution. Russell and Wefald (1988) and Russell and Wefald (1991) formulated a meta-level decision problem designed to enable optimal allocation of computation, while other works employ heuristic approximations (Hay et al. 2012). However, these approaches are limited to fully observable sequential decision making.

The use of the VOI for partially observable problems has been relatively less explored. Wei (2024) discusses the value of information based on the difference between purely open-loop and purely closed-loop policies. In contrast, we consider a recursive formulation of VOI that enables adaptive switching between the two modalities.

A recent work by Flaspohler, Roy, and Fisher III (2020) shows how VOI can be used to construct macro-actions in POMDPs. The authors propose a VOI framework similar to the one we introduce in this paper. However, while Flaspohler, Roy, and Fisher III (2020) define a Bellman backup based on an absolute difference threshold between open-loop and closed-loop values, our formulation computes a relative percentage difference threshold of the closed-loop value. This formulation is more general and intuitive since the VOI reasoning does not depend on the specific POMDP instantiation, its reward structure or the magnitude of its optimal value. This enables adaptive reasoning. Additionally, we introduce a novel alternative POMDP representation that exposes the machinery relevant for VOI reasoning in planning. Finally, we leverage our framework in a variant of POMCP for sampling-based planning instead of a method to construct macro-actions.

A core challenge when using tree search methods to solve POMDPs lies in the search strategy: when to simulate empirically valuable actions vs. lesser explored actions. This exploration-exploitation tradeoff has been the subject of many works attempting to address the curse of history. Several approaches have incorporated heuristic information gathering procedures. Do Carmo Alves et al. (2023); Pokharel (2024) use belief and observation entropy heuristics to bias the policy towards entropy-reducing actions as a proxy for actions with high VOI. However, neither belief nor observation entropy alone can sufficiently characterize the true VOI at a belief. Instead, we tackle the exploration-exploitation tradeoff by explicitly reasoning over VOI.

Kim, Karunanayake, and Kurniawati (2023); Kim and Kurniawati (2025) have explored sampling actions based on a reference policy instead of fully expanding them at each node, showing promise in addressing challenges with long-horizon planning. While Kim, Karunanayake, and Kurniawati (2023) focus on reducing the effective action branching by embedding a reference policy in the objective, our work can be seen as reducing the effective observation branching factor when the value gained by branching is low.

Our work is also related to abstraction (Ho et al. 2019) and selective perception (McCallum 1996). In particular, taking an open-loop execution mode in our framework is equivalent to abstracting all observations into a single observation. In online POMDP planning, Wu et al. (2021) propose a belief packing procedure that merges similar beliefs, effectively creating a branch that corresponds to an abstracted observation. On the other hand, our methodology adaptively ignores observations based on the VOI rather than belief similarity. This distinction is critical because observations may generate statistically dissimilar beliefs that nonetheless support the same optimal policy. Whereas similarity-based metrics would force branching in these cases, our value-based approach identifies that the information is decision-irrelevant, enabling more aggressive pruning of the search space.

Partially Observable Markov Decision Processes

A finite-horizon *Partially Observable Markov Decision Process* (POMDP) can be formally represented as a tuple $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \mathcal{Z}, D, b_0, \gamma)$, where: $\mathcal{S}, \mathcal{A}$, and $\mathcal{O}$ are finite sets of states, actions, and observations, respectively. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [-R_{max}, R_{max}]$, for some $R_{max} > 0$, is the immediate reward function, $\mathcal{Z} : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$ is the probabilistic observation function, D is the horizon of the problem, $b_0 \in \Delta(\mathcal{S})$ is the initial belief or probability distribution over states, and $\gamma \in [0, 1]$ is the discount factor.

A defining characteristic of POMDPs is that the state $s \in \mathcal{S}$ cannot be observed directly. Instead, after taking an action $a \in \mathcal{A}$, an observation $o \in \mathcal{O}$ is received which provides incomplete information about the state. Unlike in MDPs, where optimal actions depend on states alone, in POMDPs, the optimal action must be computed from a history $h_t \in \mathcal{H}$ at time t, which defines a sequence of action-observation pairs: $h_t \equiv \{b_0, a_1, o_1, ..., a_t, o_t\}$. Action-terminated histories are defined as $ha_{t+1} \equiv \{b_0, a_1, o_1, ..., a_t, o_t, a_{t+1}\}$. It is sufficient to summarize the information contained in h_t as the belief $b_t(s) \equiv \mathbb{P}(s_t = s|h_t) \in \Delta(\mathcal{S})$. With a slight abuse of notation, we sometimes drop the subscript t and express histories as h and action-terminated histories as ha.

A general finite-horizon policy is a sequence $\pi = (\pi_D, \ldots, \pi_1)$ where each $\pi_d : \Delta(\mathcal{S}) \to \mathcal{A}$ describes how an agent behaves at any belief.

The objective of the agent is to find a policy that maximizes the discounted sum of future rewards for each belief starting from an initial belief b_0. Formally, the optimal pol-

icy is defined as:

$$\pi^* \in \arg\max_{\pi} \mathbb{E}\Big[\sum_{t=0}^{D-1} \gamma^t r(b_t, \pi_{D-t}(b_t)) \;\Big|\; b_0\Big], \quad (1)$$

where b_t is the updated belief at time t.

Given an action and observation, a belief update can be performed using Bayes' rule:

$$b'(s') = \tau(b,a,o)(s') = \frac{\mathcal{Z}(s',o,a)\ \sum_{s\in\mathcal{S}} \mathcal{T}(s,a,s')\, b(s)}{\mathbb{P}(o \mid b,a)}, \quad (2)$$

where the probability of observing o is

$$\mathbb{P}(o \mid b,a) = \sum_{s'\in\mathcal{S}} \mathcal{Z}(s',o,a) \sum_{s\in\mathcal{S}} \mathcal{T}(s,a,s') b(s). \quad (3)$$

The optimal value function, denoted $V_d^*(b)$, represents the maximum discounted expected reward obtainable from belief b with d steps remaining. It is defined recursively by the Bellman optimality equation:

$$V_d^*(b) = \max_{a\in\mathcal{A}} \Big\{ r(b,a) + \gamma \mathbb{E}[V_{d-1}^*(\tau(b,a,o))] \Big\}, \quad (4)$$

with $V_0^*(b) = 0$. Here, the expected immediate reward is $r(b,a) = \sum_{s\in\mathcal{S}} b(s)\, \mathcal{R}(s,a)$, and the expectation over future values is computed with respect to the observation probabilities:

$$\mathbb{E}[\cdot] = \sum_{o\in\mathcal{O}} \mathbb{P}(o \mid b,a)\, V_{d-1}^*(\tau(b,a,o)). \quad (5)$$

It is often useful to express the optimal value for a particular action at a belief using the Q-value function:

$$Q_d^*(b,a) = r(b,a) + \gamma \mathbb{E}[V_{d-1}^*(\tau(b,a,o))]. \quad (6)$$

Consequently, the optimal decision rule for the horizon step d can be retrieved directly from the Q-values:

$$\pi_d^*(b) \in \arg\max_{a\in\mathcal{A}}\ Q_d^*(b,a). \quad (7)$$

Value of Information for POMDPs

The *curse of history* renders exact planning intractable for non-trivial horizons and problem sizes. However, the standard formulation implicitly assumes that every observation is critical for decision-making, necessitating reasoning over all possible reachable histories. In practice, the utility of observation information is rarely uniform across the belief space; there are often regions where an agent can act effectively without immediate observation feedback. In this section, we introduce our adaptive value of information (VOI) reasoning framework to exploit this non-uniformity. By formalizing the trade-off between open-loop execution (disregarding observations) and closed-loop execution (branching on observations), we expose a principled mechanism that enables an agent to selectively prune the history space while maintaining bounds on optimality. This lays the fundamental groundwork for our algorithm that exploits this tradeoff for improved search efficiency.

Adaptive Value of Information

As discussed in the previous section, the optimal solution to a POMDP satisfies the Bellman equation given in (4). Here, the agent performs a Bayesian belief update (2) after every action and observation pair. An agent that integrates feedback at every step is said to act in a *closed-loop* manner: by acting, observing, and updating its belief, it closes the feedback loop prior to every subsequent decision. The corresponding closed-loop value function is

$$V_d^{CL}(b) = \max_{a\in\mathcal{A}} \Big\{ r(b,a) + \gamma \mathbb{E}_o\big[V_{d-1}^{CL}(\tau(b,a,o))\big] \Big\}. \quad (8)$$

From a computational complexity perspective, the number of distinct histories at depth d is

$$|\mathcal{H}_d|^{CL} = (|\mathcal{A}| \cdot |\mathcal{O}|)^d. \quad (9)$$

Since the branching factor depends on the size of the observation space, the number of histories grows exponentially with the planning horizon. This relationship directly highlights the *curse of history*.

Alternatively, an agent may choose to ignore observations, planning based purely on expected beliefs. This is referred to as *open-loop* reasoning. Formally, an agent employs an *open-loop belief update* by marginalizing over the observation space in (2), such that

$$\tau(b,a)(s') = \mathbb{E}_o[\tau(b,a,o)] = \sum_s \mathcal{T}(s,a,s')\, b(s). \quad (10)$$

The resulting *open-loop value function* is defined as

$$V_d^{OL}(b) = \max_{a\in\mathcal{A}} \Big\{ r(b,a) + \gamma V_{d-1}^{OL}(\tau(b,a)) \Big\}, \quad (11)$$

with $V_0^{OL}(b) = 0$. The computational advantage is immediately apparent: the number of distinct histories at depth d when acting open-loop drops by a factor of $|\mathcal{O}|$ to

$$|\mathcal{H}_d|^{OL} = |\mathcal{A}|^d \ll |\mathcal{H}_d|^{CL}. \quad (12)$$

However, acting open-loop is generally suboptimal. Since the value function is convex with respect to the belief state (Smallwood and Sondik 1973), Jensen's inequality implies that the value of the expected belief is upper bounded by the expected value of the posterior beliefs:

$$\mathbb{E}_o[V_d^*(\tau(b,a,o))] \geq V_d^*(\mathbb{E}_o[\tau(b,a,o)]), \quad (13)$$

$$V_d^{CL} \geq V_d^{OL}. \quad (14)$$

We can quantify this suboptimality as the simple *value of information* (VOI), defined as the difference between the closed-loop and open-loop values (Wei 2024):

$$VOI_d(b) \quad = \quad V_d^{CL}(b) - V_d^{OL}(b). \quad (15)$$

The simple VOI presents a static dichotomy: it compares a policy that *always* observes against one that *never* observes over the horizon d. However, this fails to capture the potential for adaptive reasoning over the value of information. In many POMDPs, the performance gain from reasoning about observations varies over the belief space. For instance, an

agent may justifiably act in an open-loop manner while observations are uninformative, but must switch to closed-loop planning as uncertainty accumulates. Therefore, a more nuanced treatment of VOI requires evaluating the value of information locally at each belief state, allowing the agent to dynamically switch between open-loop and closed-loop modalities.

In what follows, we outline a framework designed to enable this selective reasoning. We begin by defining a new value function, denoted as $\hat{V}_d^*$, which is the value of a policy that selects between open-loop and closed-loop reasoning strategies based on the criterion defined below.

The core of this framework is our proposed *adaptive value of information* criterion. At each decision step, the agent computes the value difference between the two strategies and acts closed-loop only if the value gain exceeds a threshold parameter κ. We define the adaptive VOI as:

$$\hat{VOI}_d^{\kappa}(b) \quad = \quad \hat{V}_d^{CL}(b) - \hat{V}_d^{OL}(b), \tag{16}$$

where

$$\hat{V}_d^{OL}(b) = \max_{a \in \mathcal{A}} \Big\{ r(b,a) + \gamma \hat{V}_{d-1}^*(\tau(b,a)) \Big\}, \tag{17}$$

$$\hat{V}_d^{CL}(b) = \max_{a \in \mathcal{A}} \Big\{ r(b,a) + \gamma \mathbb{E}_o \big[\hat{V}_{d-1}^*(\tau(b,a,o)) \big] \Big\}. \tag{18}$$

Crucially, unlike the simple VOI, where open-loop and closed-loop values are computed independently, our components $\hat{V}_d^{OL}$ and $\hat{V}_d^{CL}$ are coupled recursively through the optimal κ-adaptive value function $\hat{V}_{d-1}$, defined as

$$\hat{V}_d^*(b) = \begin{cases} \hat{V}_d^{OL}(b), & \text{if } \hat{V}_d^{OL}(b) \geq \hat{V}_d^{CL}(b) - \kappa |\hat{V}_d^{CL}(b)|, \\ \hat{V}_d^{CL}(b), & \text{otherwise,} \end{cases} \tag{19}$$

for $\kappa \in [0,1]$.

With $\hat{V}_0^* = 0$, it is clear by induction on d that $\hat{V}_d^*$ is bounded and exists for any $d \in \{0, \ldots, D\}$. (19) represents a Bellman-like update under the operator $\mathbb{B}_\kappa$, where

$$\hat{V}_d^*(b) \; = \; \mathbb{B}_\kappa \hat{V}_{d-1}^*(b). \tag{20}$$

$\hat{V}^*$ defines a decision rule at each belief b based on κ. If the open-loop backup value at b is within κ of the closed-loop value, the $\hat{VOI}_d^{\kappa}(b)$ is sufficiently *low*. Therefore, the agent performs the open-loop backup, which ignores redundant information. On the other hand, if the $VOI_d^{\kappa}(b)$ is relatively *high*, there is too much value to be lost by disregarding observations; therefore, the agent performs the full closed-loop backup. This adaptive reasoning is applied at every decision step, enabling the agent to *selectively* incur the cost of closed-loop reasoning only when the adaptive VOI justifies it. As a result, the policy interpolates between purely open-loop behavior (when κ is large and VOI is generally low) and fully closed-loop behavior (when κ is small and VOI is high), reducing unnecessary reasoning over future observations in low-VOI regions while preserving near-optimal performance where observations are valuable.

Bounded Suboptimality

Here, we analyze the suboptimality of the adaptive VOI value. We define the regret ρ of the adaptive VOI value for d steps at a belief b as

$$\rho_d(b) = \left| V_d^*(b) - \hat{V}_d^*(b) \right|. \tag{21}$$

In the following theorem, we prove that the worst-case regret is bounded by a function of κ.

Theorem 1 (Bounded Regret)**.** *Let $\kappa \in [0,1]$. Then, for any b and $d \geq 1$,*

$$\rho_d(b) \leq \kappa \frac{R_{max}}{1-\gamma} \left(\frac{1-\gamma^d}{1-\gamma} \right).$$

Proof. Let $\rho_d(b) = \left| V_d^*(b) - \hat{V}_d^*(b) \right|$. Let $\mathbb{B}$ be the Bellman operator for (4). Then, by the triangle inequality,

$$\begin{aligned} \rho_d(b) &= \left| \mathbb{B} V_{d-1}^*(b) - \mathbb{B}_\kappa \hat{V}_{d-1}^*(b) \right| \\ &\leq \left| \mathbb{B} V_{d-1}^*(b) - \mathbb{B} \hat{V}_{d-1}^*(b) \right| + \left| \mathbb{B} \hat{V}_{d-1}^*(b) - \mathbb{B}_\kappa \hat{V}_{d-1}^*(b) \right| \end{aligned}$$

The first term is a γ-contraction. For the second term, (19) implies the single-step suboptimality is at most $\kappa \left| \hat{V}_d^{CL}(b) \right| \leq \kappa \frac{R_{max}}{1-\gamma}$, and since $V_0^* = \hat{V}_0^*$, we have that

$$\begin{aligned} \rho_d(b) \leq \sum_{t=1}^{d} \gamma^{d-t} (\kappa |\hat{V}_d^{CL}(b)|) &\leq \sum_{t=1}^{d} \gamma^{d-t} (\kappa \frac{R_{max}}{1-\gamma}) \\ &= \kappa \frac{R_{max}}{1-\gamma} \left(\frac{1-\gamma^d}{1-\gamma} \right) \end{aligned}$$

□

Theorem 1 shows that the adaptive VOI reasoning provides a sound and bounded approximation of V_d^*. In particular, the adaptive VOI regret scales linearly with the parameter κ, ensuring that the approximation error remains controlled. This result motivates the use of the adaptive VOI to efficiently allocate planning computation in the pursuit of mitigating the curse of history, especially in settings where the intrinsic VOI is low and the resulting regret is small.

POMDP Planning using Value of Information

This section demonstrates how the adaptive VOI reasoning framework can be integrated into a planning algorithm to effectively allocate computational effort. Optimizing the adaptive value function $\hat{V}^*$ directly would traditionally require evaluating and comparing the open-loop (17) and closed-loop (18) value functions at every step. Rather than performing this comparison as an external step, we introduce the *Value of Information POMDP* (VOI-POMDP). This meta-level representation explicitly encodes the choice between open-loop and closed-loop execution modes as distinct actions within the problem structure. By unifying these modalities into a single decision space, we enable standard solvers to automatically perform VOI reasoning during the search process. Subsequently, we propose Value of Information Monte Carlo Planning (VOIMCP), a solver based on Monte Carlo Tree Search (MCTS) that exploits the VOI-POMDP structure for efficient planning.

VOI-POMDP Representation

The adaptive VOI framework establishes a mathematical basis for disregarding observations. Consequently, computing $\hat{V}^*$ can be viewed as a meta-level choice between open-loop and closed-loop execution modalities at each belief. The following POMDP transformation encodes this meta-level choice as distinct actions directly into the problem structure.

Given a POMDP $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \mathcal{Z}, D, b_0, \gamma)$, we define the VOI-POMDP as the tuple $\mathcal{P}' = (\mathcal{S}, \mathcal{A}', \mathcal{O}', \mathcal{T}', \mathcal{R}', \mathcal{Z}', D, b_0, \gamma)$. The state space $\mathcal{S}$, initial belief b_0, horizon D, and discount factor γ remain identical to the original problem. We construct the augmented components as follows. Let $\mathcal{A}$ be the original action space. We duplicate each action and define an augmented action space with an open-loop action set and a closed-loop action set:

$$\mathcal{A}_{\mathrm{OL}} = \{\, a_{\mathrm{OL}} \mid a \in \mathcal{A}\}, \quad \mathcal{A}_{\mathrm{CL}} = \{\, a_{\mathrm{CL}} \mid a \in \mathcal{A}\},$$
$$\mathcal{A}' = \mathcal{A}_{\mathrm{OL}} \cup \mathcal{A}_{\mathrm{CL}} = \{(a, m) \mid a \in \mathcal{A},\, m \in \{\mathrm{OL}, \mathrm{CL}\}\}.$$

Further, we augment the observation space with a *null* (or uninformative) observation

$$\mathcal{O}' := \mathcal{O} \cup \{o_{\mathrm{null}}\},$$

where o_{null} is received with probability 1 under open-loop actions while observations under closed-loop actions behave as in the original POMDP:

$$\mathcal{Z}'(s', o, a_m) = \begin{cases} \mathbf{1}\{o = o_{\mathrm{null}}\}, & m = \mathrm{OL}, \\ Z(s', o, a), & m = \mathrm{CL}. \end{cases} \tag{22}$$

$\mathcal{T}'$ and $\mathcal{R}'$ map both open-loop and closed-loop action variants to the same values:

$$\mathcal{T}'(s, a_m, s') = \mathcal{T}(s, a, s'), \tag{23}$$
$$\mathcal{R}'(s, a_m) = \mathcal{R}(s, a). \tag{24}$$

The belief update for a VOI-POMDP depends on the action taken. For open-loop actions, the computed belief update marginalizes over observations, as in (10). Closed-loop actions allow for the full belief update as in (2). Thus,

$$b'(s') = \begin{cases} \tau(b, a_m, o_{\mathrm{null}}) = \sum_s \mathcal{T}'(s, a, s')\, b(s), & m = \mathrm{OL}, \\ \tau(b, a_m, o) = \tau(b, a, o), & m = \mathrm{CL}. \end{cases} \tag{25}$$

The adaptive VOI value function $\hat{V}'_d$ for $\mathcal{P}'$ is computed recursively for $d \in \{0, \ldots, D\}$, starting from the base case $\hat{V}'_0 \equiv \hat{V}_0 = 0$. To simplify notation, we denote the optimal value function for $\mathcal{P}'$ as $\hat{V}'_d$, dropping the superscript $*$ (i.e., $\hat{V}'_d = \hat{V}'^*_d$). The value backup is defined as:

$$\hat{V}'_d(b) = Q'_d(b, \hat{\pi}^*_d(b)), \quad \forall d \in \{0, \ldots, D\}, \tag{26}$$

where Q'_d is the standard action-value for $\mathcal{P}'$, computed as:

$$Q'_d(b, a) = r(b, a) + \gamma \mathbb{E}\left[\hat{V}'_{d-1}(\tau(b, a, o))\right]. \tag{27}$$

and $\hat{\pi}^*_d$ denotes the optimal adaptive VOI-based policy at depth d that satisfies for every belief b

$$\hat{\pi}^*_d(b) = \arg\max_{a \in \mathcal{A}'} \begin{cases} Q'_d(b, a), & a \in \mathcal{A}_{OL}, \\ Q'_d(b, a) - \kappa |Q'_d(b, a)|, & a \in \mathcal{A}_{CL}. \end{cases} \tag{28}$$

We see that the VOI-POMDP representation preserves the same (optimal) adaptive VOI value function as (19).

Proposition 1. *Let $\mathcal{P}$ be the original POMDP and let $\mathcal{P}'$ be the corresponding transformed VOI-POMDP defined above. We have that for every depth $d \in \{0, \ldots, D\}$ and belief $b \in \Delta(\mathcal{S})$,*

$$\hat{V}'_d(b) = \hat{V}^*_d(b). \tag{29}$$

The VOI-POMDP representation has distinct benefits in planning. By mapping the meta-level choice to the action and value backup space, we allow standard solvers to optimize for $\hat{V}^*$ using minor modifications to existing search heuristics. The representation structurally simplifies the search process. Although duplicating the action space doubles the nominal action branching ($|\mathcal{A}'| = 2|\mathcal{A}|$) and raises the worst-case branching factor from $O((|\mathcal{A}||\mathcal{O}|)^d)$ to $O((|\mathcal{A}||\mathcal{O}| + |\mathcal{A}|)^d)$, this cost is offset by the collapse of the observation space under open-loop actions. Under open-loop actions, the action-observation branching factor drops to 1. In practice, this prevents expanding dense observation branches of a search tree whenever VOI is low, significantly reducing the effective branching factor and enabling deeper planning under the same computational budget.

Value of Information Monte Carlo Planning

To leverage the advantages of VOI reasoning for POMDP planning, we build on PO-UCT (Silver and Veness 2010), a widely used Monte Carlo tree search (MCTS) algorithm for large POMDPs. PO-UCT extends the upper confidence trees (UCT) algorithm to partially observable problems by building a search tree over the space of histories instead of states, relying solely on a generative model $\mathcal{G}$ of the problem.

Our algorithm, called Value of Information Monte Carlo Planning (VOIMCP), models adaptive VOI reasoning by solving for $\hat{V}^*_d$ on $\mathcal{P}'$. The procedure for VOIMCP is outlined in Algorithm 1. The global parameters include the number of tree queries n, the exploration bonus β, the problem horizon D, the discount factor γ, and κ. The SEARCH procedure serves as the entry point, taking as input a belief b and repeatedly calling the SIMULATE procedure to build out the tree T. At each history-action node $T(ha)$ in the tree, the visitation count $N(ha)$, and the estimated value $\bar{Q}(ha)$ are stored. These statistics are all initialized to 0. While our theoretical VOIMCP formulation has every simulation iteration reaching the full depth D, our practical implementation follows the standard PO-UCT approach for efficiency (Laouar, Ho, and Sunberg 2026). Specifically, when the SIMULATE procedure expands a new history node $h \notin T$, it terminates the iteration and invokes a VALUEESTIMATE procedure to estimate the value of initialized leaf nodes. A common evaluation approach is to simulate with a heuristic rollout policy.

We adopt the polynomial upper confidence bound (UCB) from Shah, Xie, and Xu (2022). Our primary contribution is a modified polynomial UCB, which implements the

adaptive VOI backup in (26). We approximate the adaptive VOI backup operator by replacing the exact values Q'_d with Monte Carlo estimates $\bar{Q}$ and applying a penalty κ directly to the closed-loop exploitation term. For open-loop actions, the selection metric follows the standard optimistic principle:

$$UCB_{VOI}(ha_{OL}) = \bar{Q}(ha_{OL}) + B_N(ha_{OL}), \tag{30}$$

where $\bar{Q}(\cdot)$ denotes the mean value estimate of simulations that visit a history-action pair, and $B_N(\cdot)$ denotes the polynomial bonus term proposed by Shah, Xie, and Xu (2022):

$$B_N(ha_m) = \beta^{1/\xi} \cdot \frac{N(h)^{\alpha/\xi}}{N(ha_m)^{1-\eta}}, \quad \forall m \in \{OL, CL\}, \tag{31}$$

parameterized by β, ξ, α, and η.

For closed-loop actions, we leverage the adaptive VOI objective to enforce a κ-dependent bias against unnecessary observation branching. This is achieved by deflating the exploitation term in the UCB calculation:

$$\begin{aligned} UCB_{VOI}(ha_{CL}) = \bar{Q}(ha_{CL}) \; - \; & \kappa|\bar{Q}(ha_{CL})| \\ & + B_N(ha_{CL}). \end{aligned} \tag{32}$$

This deflation biases the search towards open-loop actions whenever the potential VOI is insufficient to meet relative tolerance κ, effectively steering the search toward the computationally cheaper open-loop branches. By integrating this VOI-guided reasoning into our selection policy, VOIMCP focuses computation on actions and observations that are likely to yield significant value, enabling deeper tree search while maintaining near-optimal policy quality.

Annealing for Global Convergence While a fixed κ prunes unnecessary observation branching based on a constant threshold (converging to $\hat{V}^*_D$), recovering the optimal POMDP value V^*_D requires this bias to vanish over time. To achieve this, we employ an annealing schedule where κ_N decays as a function of the visitation counts. Specifically, we require κ to decay as a function of the confidence width, i.e., $0 \leq \kappa_i(t, s_i) \leq \frac{1}{c_\kappa R_{max}} B_{t,s_i}$ where $c_\kappa > 1$. This ensures that the selection bias is always dominated by the confidence width; as $B_N \to 0$, the bias vanishes and the algorithm recovers the optimal value function in the limit.

Convergence Analysis of VOIMCP

Here, we analyze the non-asymptotic properties of VOIMCP. First, we prove that for a fixed κ, VOIMCP converges polynomially to the optimal VOI-value function $\hat{V}^*_D$. Let $\bar{V}_n(b_0) = \frac{1}{n}\sum_{a\in\mathcal{A}'} N(h_0a)\bar{Q}(h_0a)$ denote the value estimate at the root after n simulations, where $h_0 = \{b_0\}$. Then we have the following result.

Theorem 2. *Consider a POMDP $\mathcal{P}$ with $D \geq 1$ and a fixed $\kappa \in [0, 1]$. There exists a valid configuration of the algorithm's depth dependent parameters (specifically, depth dependent ξ, α, β detailed in the Technical Appendix of Laouar, Ho, and Sunberg (2026)) such that for any $\eta \in [0.5, 1)$ and any initial belief b_0, the following claim holds for the output $\bar{V}_n(b_0)$ of VOIMCP with n simulations:*

$$|E[\bar{V}_n(b_0)] - \hat{V}^*_D(b_0)| \leq O(n^{\eta-1}). \tag{33}$$

Algorithm 1: VOIMCP

```
procedure SEARCH(b)
    for all i = 1, ..., n do
        s ~ b
        h_0 = {b}
        SIMULATE(s, h_0, 0)
    end for
    return arg max_{a∈A'} Q̄(h_0 a)
end procedure
procedure SIMULATE(s, h, depth)
    if depth = D then
        return 0
    end if
    if h ∉ T then
        for all a ∈ A' do
            T(ha) ← (0, 0)
        end for
    end if
    a* ∈ arg max_{a∈A'} UCB_VOI(ha)
    if a* ∈ A_CL then
        (s', o, r) ~ G(s, a*)
    else
        (s', r) ~ G(s, a*)
        o ← o_null
    end if
    N(h) ← N(h) + 1; N(ha*) ← N(ha*) + 1
    R ← r + γ SIMULATE(s', ha*o, depth+1)
    Q̄(ha*) ← Q̄(ha*) + (R − Q̄(ha*)) / N(ha*)
    return R
end procedure
```

Proof Sketch. We defer the detailed derivation to the Technical Appendix in Laouar, Ho, and Sunberg (2026). The proof structure adapts the analysis of (Shah, Xie, and Xu 2022) and models the search tree as a hierarchy of non-stationary κ-biased Multi-Armed Bandit (MAB) problems. Using backward induction from the leaf nodes to the root, we prove polynomial concentration and convergence to the VOI-optimal value. Finally, we transfer this result to VOIMCP via a history-MDP equivalence. □

We now demonstrate that by annealing κ as a function of the confidence bonus, the selection bias vanishes sufficiently fast to guarantee convergence to V^*_D.

Theorem 3. *Consider a POMDP $\mathcal{P}$ with $D \geq 1$ and $0 \leq \kappa_i(t, s_i) \leq \frac{1}{c_\kappa R_{max}} B_{t,s_i}$ for some $c_\kappa > 1$. There exists a valid configuration of the algorithm's depth dependent parameters (specifically, depth dependent ξ, α, β detailed in the full version (Laouar, Ho, and Sunberg 2026)) such that for any $\eta \in [0.5, 1)$ and any initial belief b_0, the following claim holds for the output $\bar{V}_n(b_0)$ of VOIMCP with n simulations:*

$$|E[\bar{V}_n(b_0)] - V^*_D(b_0)| \leq O(n^{\eta-1}). \tag{34}$$

Proof Sketch. We defer the detailed derivation to the Technical Appendix in Laouar, Ho, and Sunberg (2026). The proof follows the same backward induction structure es-

tablished in Theorem 2. The distinction lies in the selection bias for closed-loop actions: by enforcing $\kappa_i(t, s_i) \leq \frac{1}{c_\kappa R_{max}} B_{t,s_i}$, the VOI penalty effectively acts as a vanishing exploration noise rather than a fixed bias. For every node, as visitation count $\to \infty$, this bias term shrinks to zero and is asymptotically dominated by the suboptimality gap of the actions. Propagating this result from leaves to the root, we demonstrate that the value estimates at each depth concentrate around the true optimal value function V^*. □

Empirical Evaluation

In this section, we evaluate our adaptive VOI approach.

Domains We evaluate VOIMCP with a fixed[1] κ on three POMDP benchmark domains with large observation spaces:

1. **Target Tracking** (Flaspohler, Roy, and Fisher III 2020): An agent aims to track a moving target on a 10×10 grid. The agent's own state is fully observable, while the target is only partially observable through noisy measurements.
2. **FieldVision RockSample** (Ross et al. 2008): A variant of RockSample with a much larger observation space. A robot explores a grid to collect valuable rocks while avoiding bad rocks. Rock quality is observed noisily through sensor readings after each action.
3. **Laser Tag** (Somani et al. 2013): A robot aims to tag a moving target on a 7×11 grid with eight randomly placed obstacles. The robot receives noisy distance measurements to the nearest obstacle in each direction, yielding an observation space of roughly 1.5×10^6 observations.

Comparison Algorithms We evaluate VOIMCP against the following baselines:

1. **PO-UCT** (Silver and Veness 2010): The standard MCTS solver for POMDPs, which uses the original UCB1 selection policy. This also serves as an ablation study, as VOIMCP reduces to PO-UCT if the meta-level action space is removed and standard selection policy is used.
2. **I-UCB POMCP** (Do Carmo Alves et al. 2023): A variant of POMCP that augments the UCB term with an observation entropy heuristic.
3. **Open-Loop**: A VOIMCP variant that acts purely in an open-loop manner by only expanding null observations. This baseline essentially plans a sequence of actions without reasoning about future observations.

To isolate the impact of the search component in the planning algorithms, we standardize the belief update mechanism across all methods by using a Sequential Importance Resampling (SIR) particle filter (Gordon, Salmond, and Smith 1993). Similarly, we estimate leaf node values using a random rollout policy for all approaches.

[1]Additional results for annealing experiments are available in the full version (Laouar, Ho, and Sunberg 2026).

Implementation Details We tuned the key parameters of each approach and report the best performance. The horizon D for each algorithm was optimized from the set $\{20, 40, 60, 80\}$, and the exploration constant c for POMCP and VOIMCP was optimized from the set $\{1, 10, 100, 1000\}$. For VOIMCP, $\beta^{1/\xi} = c$ and $\eta = \frac{1}{2}$ for all nodes in the tree, resulting in a bonus term B_N that scales as $N(h)^{\frac{1}{4}}/\sqrt{N(ha)}$, consistent with Shah, Xie, and Xu (2022). For I-UCB POMCP, we adopt the parameters recommended by Do Carmo Alves et al. (2023). Detailed parameters used for the experiments are outlined in the supplementary material. We evaluate all approaches under computational budgets ranging from 100 to 1000 tree queries, averaged over 1000 trials. All algorithms were implemented in Julia and executed single-threaded on a computer equipped with an Intel Xeon(R) W-1370P processor.

Results and Discussion

In Figure 1, we report the discounted cumulative reward against the number of tree queries. We also report the mean maximum tree depth and effective branching factor in Figure 2. We define the maximum tree depth as the greatest depth reached by any history node in the search tree, and the effective branching factor as the average number of action-observation branches expanded per visited node.

VOIMCP outperforms all baselines in mean discounted cumulative reward across all the benchmark problems. The statistics in Figure 2 reveal the mechanism behind this success: with the exception of the Open-Loop policy which only has 1 observation branch per action, VOIMCP consistently achieves the highest maximum tree depth while maintaining the lowest effective branching factor. These results confirm that the adaptive VOI framework enables VOIMCP to "take shortcuts" via selectively avoiding expanding dense observation branches (closed-loop actions) whenever the value of information is low. By selectively reducing the computational overhead of observation processing, VOIMCP allocates limited planning resources to enable deeper search. This alleviates the curse of history and facilitates a more efficient exploration, leading to superior performance.

In domains with large observation spaces, the performance of PO-UCT degrades. This occurs because the original UCB heuristic distributes exploration across the observations, leading to shallower search trees than VOIMCP. Interestingly, the Open-Loop policy performs better than I-UCB POMCP in both the Target Tracking and FieldVision RockSample problems, and better than PO-UCT in the latter. This counter-intuitive result provides an insight: when observation spaces are large and noisy, the computational cost of fully processing observations may outweigh the benefit. By always ignoring observations, the Open-Loop policy eliminates observation branching and searches deeply, but cannot reason about observation information. VOIMCP bridges this gap through the adaptive VOI, acting open-loop when observations are noisy or irrelevant, while retaining the reactivity of closed-loop planning when necessary.

While the I-POMCP algorithm has demonstrated superior performance on some domains (Do Carmo Alves et al.

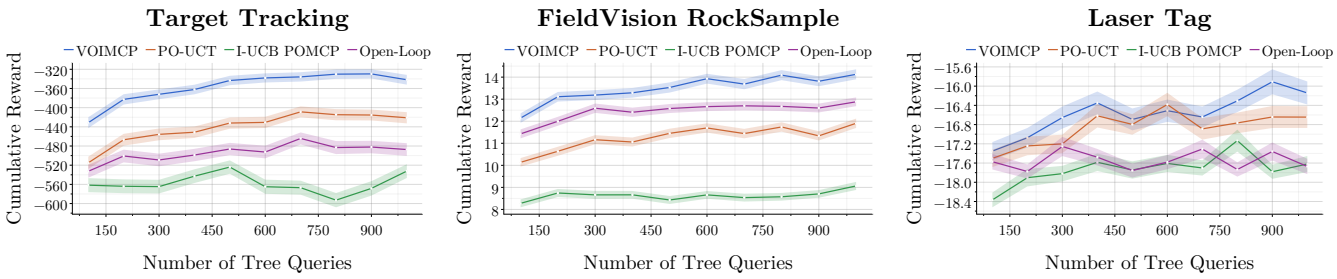

Figure 1: Comparative benchmark results presenting discounted cumulative reward against the number of tree queries over 1000 trials. The lighter colored ribbons around the Monte Carlo mean display the 95% confidence interval.

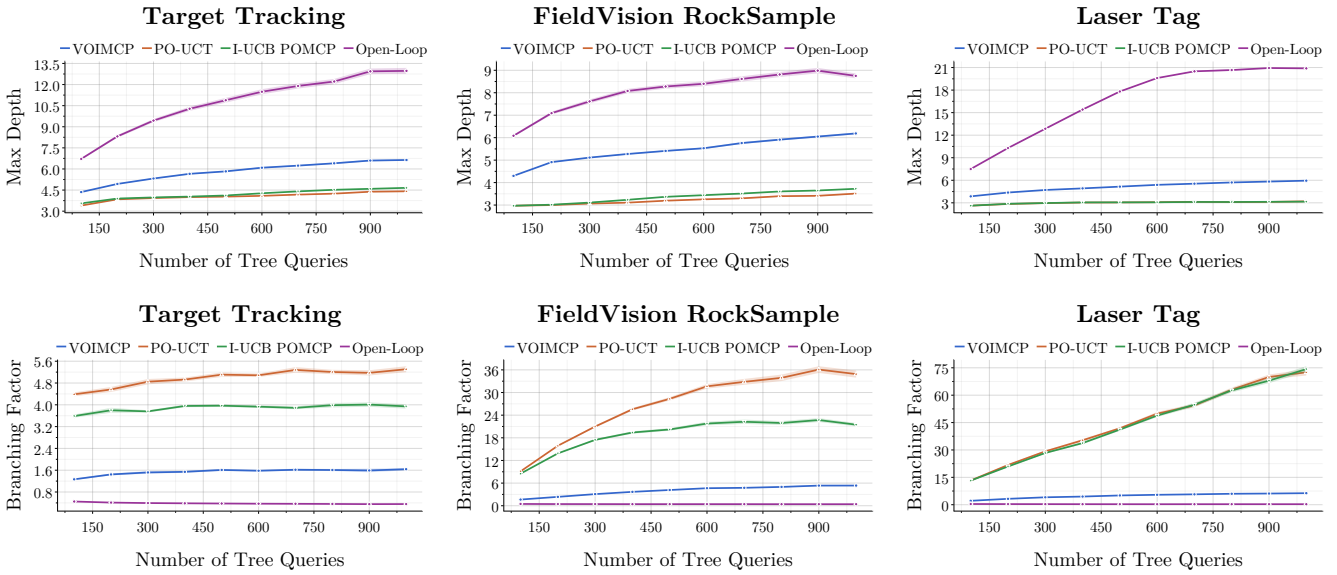

Figure 2: Tree growth statistics. (Top Row) Maximum tree depth vs. number of tree queries. (Bottom Row) Effective action-observation branching factor vs. number of tree queries. Statistics are computed over 100 trials.

2023), its planning component I-UCB POMCP was the worst-performing baseline across all three benchmarks. I-UCB augments the traditional UCB with an observation entropy heuristic. This term is designed to induce information-gathering behavior by rewarding actions that heuristically reduce uncertainty. However, our results demonstrate that the observation entropy is an insufficient proxy for the value of information. I-UCB may prioritize high-entropy observations that are irrelevant to the task rewards. In contrast, VOIMCP directly searches based on the value of information, ensuring that the agent only seeks observations that contribute to reward maximization.

The effectiveness of VOIMCP relies on the search-depth gains from avoiding observation branching outweighing the increase of the augmented action space. For many POMDPs, such as our case studies, the existence of open-loop shortcuts enables more efficient search. However, for problems in which the value of information is high but poorly captured during search, the selection policy may persistently select open-loop actions that are counterproductive for high-value policies. We posit that VOI-guided pruning is most effective in domains where the computational curse of history is the primary bottleneck to discovering high-value policies.

Conclusion

This paper presents a recursive framework on reasoning about the value of information in POMDP planning. We propose VOIMCP, an algorithm that selectively disregards observation information when the value to be gained by reasoning over such information is low. We prove the theoretical properties of both the framework and our algorithm, showing bounded regret and non-asymptotic convergence. This work shows that value of information reasoning can reduce the effective branching factor, alleviating the curse of history in POMDPs. Future work includes investigating fundamental structures of POMDPs that are most amenable to VOI reasoning, designing better schemes for tuning and annealing κ, and integrating adaptive VOI reasoning into other POMDP solution techniques.

Acknowledgments

This work was partially supported by the National Science Foundation, Grants 2340958 and 2137269 and the members of the Center for Autonomous Air Mobility and Sensing (CAAMS) IUCRC.

References

Ayer, T.; Alagoz, O.; and Stout, N. K. 2012. OR Forum—A POMDP Approach to Personalize Mammography Screening Decisions. *Operations Research*, 60(5): 1019–1034.

Chen, M.; Nikolaidis, S.; Soh, H.; Hsu, D.; and Srinivasa, S. 2020. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2): 1–23.

Do Carmo Alves, M. A.; Varma, A.; Elkhatib, Y.; and Soriano Marcolino, L. 2023. Information-guided Planning: An Online Approach for Partially Observable Problems. In *Advances in Neural Information Processing Systems*, volume 36, 69157–69177.

Flaspohler, G.; Roy, N. A.; and Fisher III, J. W. 2020. Belief-dependent macro-action discovery in POMDPs using the value of information. *Advances in Neural Information Processing Systems*, 33: 11108–11118.

Gordon, N.; Salmond, D.; and Smith, A. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140: 107–113.

Hay, N.; Russell, S.; Tolpin, D.; and Shimony, S. E. 2012. Selecting computations: theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 346–355.

Ho, M. K.; Abel, D.; Griffiths, T. L.; and Littman, M. L. 2019. The value of abstraction. *Current opinion in behavioral sciences*, 29: 111–116.

Howard, R. A. 1966. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1): 22–26.

Kim, E.; Karunanayake, Y.; and Kurniawati, H. 2023. Reference-Based POMDPs. In *Advances in Neural Information Processing Systems*, volume 36, 40659–40675.

Kim, E.; and Kurniawati, H. 2025. Partially Observable Reference Policy Programming. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 8536–8543.

Kurniawati, H.; Hsu, D.; and Lee, W. S. 2009. SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. In *Robotics: Science and Systems IV*, 65–72.

Laouar, Z.; Ho, Q. H.; and Sunberg, Z. 2026. Leveraging the Value of Information in POMDP Planning. *arXiv preprint arXiv:2604.01434*.

Lauri, M.; Hsu, D.; and Pajarinen, J. 2023. Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Transactions on Robotics*, 39(1): 21–40.

Madani, O.; Hanks, S.; and Condon, A. 2003. On the Undecidability of Probabilistic Planning and Related Stochastic Optimization Problems. *Artificial Intelligence*, 147(1-2): 5–34.

McCallum, A. K. 1996. *Reinforcement Learning with Selective Perception and Hidden State*. University of Rochester.

Papakonstantinou, K.; and Shinozuka, M. 2014. Planning structural inspection and maintenance policies via dynamic programming and Markov processes. *Reliability Engineering & System Safety*, 130: 214–224.

Pokharel, G. 2024. Increasing the Value of Information During Planning in Uncertain Environments. *arXiv preprint arXiv:2409.13754*.

Ross, S.; Pineau, J.; Paquet, S.; and Chaib-Draa, B. 2008. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32: 663–704.

Russell, S.; and Wefald, E. 1988. Decision-Theoretic Control of Reasoning: General Theory and an Application to Game-Playing. Technical Report UCB/CSD-88-435, University of California Berkeley.

Russell, S.; and Wefald, E. 1991. Principles of metareasoning. *Artificial Intelligence*, 49(1): 361–395.

Shah, D.; Xie, Q.; and Xu, Z. 2022. Nonasymptotic Analysis of Monte Carlo Tree Search. *Operations Research*, 70(6): 3234–3260.

Shani, G.; Pineau, J.; and Kaplow, R. 2013. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1): 1–51.

Silver, D.; and Veness, J. 2010. Monte-Carlo Planning in Large POMDPs. In *Advances in Neural Information Processing Systems*, volume 23.

Smallwood, R. D.; and Sondik, E. J. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations research*, 21(5): 1071–1088.

Smith, T.; and Simmons, R. 2005. Point-based POMDP algorithms: improved analysis and implementation. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 542–549. AUAI Press.

Somani, A.; Ye, N.; Hsu, D.; and Lee, W. S. 2013. DESPOT: Online POMDP Planning with Regularization. In *Advances in Neural Information Processing Systems*, volume 26.

Sunberg, Z.; and Kochenderfer, M. 2018. Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces. *Proceedings of the International Conference on Automated Planning and Scheduling*, 28(1): 259–263.

Wei, R. 2024. Value of Information and Reward Specification in Active Inference and POMDPs. In *The First Workshop on NeuroAI @ NeurIPS2024*.

Wu, C.; Yang, G.; Zhang, Z.; Yu, Y.; Li, D.; Liu, W.; and Hao, J. 2021. Adaptive Online Packing-guided Search for POMDPs. In *Advances in Neural Information Processing Systems*, volume 34, 28419–28430. Curran Associates, Inc.

Åström, K. 1965. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1): 174–205.