Simplifying Complex Observation Models in Continuous POMDP Planning with Probabilistic Guarantees and Practice

Idan Lev-Yehudi[1], **Moran Barenboim**[1], **Vadim Indelman**[2]

[1]Technion Autonomous Systems Program (TASP), Technion - Israel Institute of Technology, Haifa 32000, Israel
[2]Department of Aerospace Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel
{idanlev, moranbar}@campus.technion.ac.il, vadim.indelman@technion.ac.il

Abstract

Solving partially observable Markov decision processes (POMDPs) with high dimensional and continuous observations, such as camera images, is required for many real life robotics and planning problems. Recent researches suggested machine learned probabilistic models as observation models, but their use is currently too computationally expensive for online deployment. We deal with the question of what would be the implication of using simplified observation models for planning, while retaining formal guarantees on the quality of the solution. Our main contribution is a novel probabilistic bound based on a statistical total variation distance of the simplified model. We show that it bounds the theoretical POMDP value w.r.t. original model, from the empirical planned value with the simplified model, by generalizing recent results of particle-belief MDP concentration bounds. Our calculations can be separated into offline and online parts, and we arrive at formal guarantees without having to access the costly model at all during planning, which is also a novel result. Finally, we demonstrate in simulation how to integrate the bound into the routine of an existing continuous online POMDP solver.

Introduction

Partially observable Markov decision processes (POMDP) are a flexible mathematical framework for modeling real world decision-making and planning problems with inherent uncertainty. Yet, POMDPs are notoriously hard to solve (Papadimitriou and Tsitsiklis 1987). Online solvers like POMCP (Silver and Veness 2010) and DESPOT (Ye et al. 2017) focus on finding a solution for the current belief only, rather than solving for the entire belief space, hereby restricting the complexity of solution.

These planners are not trivially suitable for problems with continuous observation or action spaces. In recent years there have been several practical approaches for continuous solvers, such as POMCPOW, PFT-DPW (Sunberg and Kochenderfer 2018) and LABECOP (Hoerger and Kurniawati 2021). However the current state of the art is divided between practical algorithms and those with guarantees (Lim, Tomlin, and Sunberg 2021).

Recently were suggested deep neural networks for probabilistic visual observation models in AI. (Eslami et al. 2018) suggested a neural model for scene rendering. Both (Jonschkowski, Rastogi, and Brock 2018) and (Karkus, Hsu, and Lee 2018) suggested DPF and PF-NET respectively, which are models aimed at learning particle filtering. In the context of model free planning, examples include QMDP-net (Karkus, Hsu, and Lee 2017) and Deep Variational Reinforcement Learning (Igl et al. 2018). In the context of model based planning, recent approaches include DualSMC (Wang et al. 2020) and VTS (Deglurkar et al. 2023). All of these works demonstrate the computational challenges arising from incorporating deep neural visual models into practical POMDP planning.

Our research focuses on the implications of using a different observation model, less accurate but computationally favorable, during planning in continuous POMDPs. This could describe for example the case of using a shallower neural

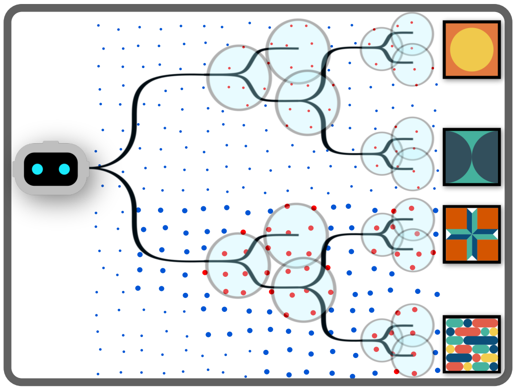


Figure 1: An illustration of a planning session with a simplified observation model. The scattered dots are the presampled states, and the dot size is relative to Δ_Z, the estimated discrepancy between the simplified and original observation models. The simplified observation model is less accurate on the bottom where the surroundings are more visually complex. For the two policies, we compute the bound as a summation over Δ_Z weighted by the transition model. We bound the summation to a truncation distance indicated by the cyan circles, and Δ_Z within it is marked in red. The bottom policy chooses actions that give higher weights to states with greater Δ_Z, resulting in looser bounds.

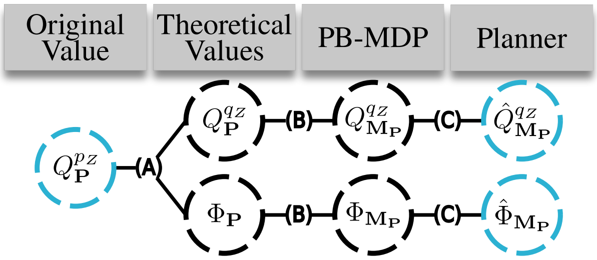


Figure 2: The various relationships between the action value functions in Corollary 3 for probably-approximately bounding $|Q^{pz}_{\mathbf{P}} - \hat{Q}^{qz}_{\mathbf{M_P}}| \leq \hat{\Phi}_{\mathbf{M_P}}$. (A) is given by Theorem 2, connecting theoretical value functions with the theoretical local state bound. (B) is given by Corollary 2, connecting theoretical action value functions with their PB-MDP approximation. (C) is given by any planner with performance guarantees, such as POWSS, approximating the PB-MDP values.

network for planning than for inference. To the best of our knowledge, there still isn't a clear-cut answer as to how accurate a learned observation model needs to be in order to attain a target performance in continuous POMDP planning. Our work provides insight into the complexity-performance trade-off in planning with different observation models.

Contribution

In this paper, we employ a simplified observation model and derive guarantees in the form of concentration inequalities. To the best of our knowledge, this is the first to work to do so. Specifically,

- We derive a novel non-parametric bound on the difference between value functions when planning with different observation models, and we observe that it can be formulated as a local state function.
- Our bound is practically computed by separating calculations to offline and online parts, such that the online estimator of the bounds refrains from accessing the computationally expensive original observation model.
- We derive concentration bounds on the online estimator of the bound. We do so by generalizing previous convergence results of particle-belief MDPs to state rewards under general policies.
- Finally, we demonstrate the practical computation of the bound's estimator in a simple simulated environment.

Related Work

Simplification of Probabilistic Models in POMDP Planning In (Hoerger, Kurniawati, and Elfes 2023) the authors introduce considered several simplification levels of the transition model, whereas we consider only two levels. Yet Hoerger, Kurniawati, and Elfes had to access the most complex model online, and their convergence guarantee is only asymptotic. (Ha and Schmidhuber 2018) consider learning a simplified generative model, for environments in which direct training would be infeasible. On this "world model" they train the policy, yet they too do not have any performance guarantees for the learned policy. A technique for measuring non-linearity based on total variation distance in POMDP planning was considered by (Hoerger et al. 2020). While there are some similarities in the approach to our work, they did not show how their practical estimator is related to the theoretical bound, nor did they give any performance guarantees.

Continuous POMDP Planning With Guarantees (Lim, Tomlin, and Sunberg 2020) proved that POWSS, a Monte Carlo tree search algorithm for continuous POMDPs based on a modification of Sparse Sampling (Kearns, Mansour, and Ng 2002), converges to the optimal policy in high probability. However, is not efficient enough for practical use. Later in (Lim et al. 2023) the convergence results were extended to prove that generally the particle-belief MDP (PB-MDP) accurately approximates the original POMDP with high probability. Recently (Shienman and Indelman 2022; Barenboim, Lev-Yehudi, and Indelman 2023) developed pruning of data association hypotheses in continuous POMDPs with guarantees. The former focuses on a specific reward of entropy of hypotheses weights, while the latter provides looser bounds but for any general state reward. In both works, the hybrid belief setting and type of simplification are different from our work.

Preliminaries

A POMDP is the tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{Z}, p_T, p_Z, r, \gamma, L, b_0 \rangle$. $\mathcal{X}, \mathcal{A}, \mathcal{Z}$ are the state space, action space and observation space respectively. Throughout the paper, $x \in \mathcal{X}$, $a \in \mathcal{A}$, $z \in \mathcal{Z}$ denote individual states, actions and observations respectively. We consider $\mathcal{X}$ and $\mathcal{Z}$ that are continuous, while $\mathcal{A}$ can be either discrete or continuous

The conditional probability distributions p_T and p_Z are the transition and observation models respectively. $p_T(x' \mid x, a)$ models the uncertainty of taking action $a \in \mathcal{A}$ from state $x \in \mathcal{X}$, and $p_Z(z \mid x)$ models the uncertainty of receiving an observation $z \in \mathcal{Z}$ at state x. The reward function $r \colon \mathbb{N} \times \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ gives the immediate reward of applying action a at state x and time t. We denote $r_t \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as the reward function at time t. We assume that the POMDP starts at time 0 and terminates after $L \in \mathbb{N}$ steps. $\gamma \in (0, 1]$ is the discount factor.

Because of the partial observability, the agent has to maintain a probability distribution of the current state given past actions and observations, known as the belief. The initial belief of the agent, denoted as b_0, captures the uncertainty in the initial state. A history at time t is defined as a sequence of the starting belief, followed by actions taken and observations received until t: $H_t \triangleq (b_0, a_0, z_1, \ldots, a_{t-1}, z_t)$. The belief at time t is defined as the conditional distribution of the state given the history $b_t(x_t) \triangleq \mathbb{P}(x_t \mid H_t)$. We denote $H_t^- \triangleq (b_0, a_0, z_1, \ldots, a_{t-1})$ for the same history without the last measurement, and the propagated belief $b_t^-(x_t) \triangleq \mathbb{P}(x_t \mid H_t^-)$. The belief reward is the expected state reward: $r_i(b_i, a) \triangleq \mathbb{E}_{x_i \sim b_i}[r_i(x_i, a)]$.

A policy at time i, denoted as π_i, is a mapping from the space of all histories to the action space $\pi_i \colon \mathcal{H}_i \to \mathcal{A}$. A policy π is a collection of policies from the start time until the

POMDP terminates, i.e. $\pi \triangleq (\pi_t)_{t=0}^{t=L-1}$, and for brevity we denote π_i instead of $\pi_i(H_i)$. It can be shown that optimal decision-making can be made given the belief, instead of considering the entire history (Kaelbling, Littman, and Cassandra 1998). Since the belief and history fulfill the Markov property, a POMDP is a Markov decision process (MDP) on the belief space, commonly referred to as the belief MDP.

The value function of a policy from time t is the expected sum of discounted rewards until the horizon of the POMDP: $V_t^\pi(b_t) \triangleq \mathbb{E}_{z_{t+1:L}}[\sum_{i=t}^{L} \gamma^{i-t} r_i(b_i, \pi_i)]$. Most often the action value function, defined as $Q_t^\pi(b_t, a) \triangleq r(b_t, a) + \gamma \mathbb{E}_{z_{t+1}}[V_{t+1}^\pi(b_{t+1})]$, is used as an intermediate step in the estimation of the value function. The goal of planning is to compute a policy that would maximize the value function. The optimal policy is denoted $\pi^* = \arg\max_\pi V_t^\pi(b_t)$, and the corresponding value and action value functions for the optimal policy are V_t^* and Q_t^*.

Often in real world applications it is impractical to update the belief exactly. Particle filters are used instead to represent a belief non parametrically. The particle belief of C particles at time i is represented by a set state samples and weights $\bar{b}_t = \{(x_t^j, w_t^j)\}_{j=1}^C$, and is defined as the discrete distribution $\mathbb{P}(x \mid \bar{b}_t) = \frac{\sum_{j=1}^C w_t^j \cdot \delta(x - x_t^j)}{\sum_{j=1}^C w_t^j}$. The basic particle filtering algorithm, Sequential Importance Sampling (SIS) (Doucet, de Freitas, and Gordon 2001, 1.3.2), approximates the target posterior distribution by recursively updating the importance weights of the particles with the observation likelihood, and normalizes the weights accordingly. Sequential Importance Resampling (SIR) adds resampling of state particles after the weight updates, to avoid weight degeneracy (Kong, Liu, and Wong 1994).

The Monte Carlo (MC) estimator of an expectation, denoted with $\hat{\mathbb{E}}$, is its approximation with a finite number of samples. This applies to quantities that are defined as expectations, such as $\hat{V}_t^\pi$. For example, the empirical value function based on N samples $\{z_{t+1:L}^j\}_{j=1}^N$ would be $\hat{V}_t^\pi = \frac{1}{N}\sum_{j=1}^N \gamma^{i-t} r_i(b_i^j, \pi_i^j)$, $\pi_i^j = \pi(H_i^j)$, where b_i^j and H_i^j are updated with $z_{t+1:i}^j$, i.e. the sampled sequence of observations until time i.

(Lim et al. 2023) considered the PB-MDP, where the belief space is explicitly the space of particle beliefs $\bar{b}$ of C particles. They proved that with high probability the optimal value in the PB-MDP is bounded from the optimal POMDP value, and this bound converges as $C \to \infty$.

We similarly denote the original POMDP as $\mathbf{P}$ and the PB-MDP as $\mathbf{M_P}$. Quantities that are measured w.r.t. each are denoted accordingly, f.e. the value function starting at time t, given policy π and starting belief b_t in the original POMDP is denoted as $V_{\mathbf{P},t}^\pi(b_t)$. When a result applies to both $\mathbf{P}$ and $\mathbf{M_P}$, we omit this notation.

Methodology

Problem Formulation

We consider cases in which the observation model p_Z is simplified for computational reasons. We denote the simplified model as q_Z, which is only used in planning. In our setting, inference is performed with the original model, such that the initial belief b_t for planning session starting at time t was updated with p_Z. We denote the future value function based on belief updates in planning with either the original or simplified observations models as V_t^{π,p_Z} and V_t^{π,q_Z} respectively, and similarly for Q_t^{π,p_Z} and Q_t^{π,q_Z}.

In the settings we consider, p_Z is computationally more expensive than q_Z, and both are considerably more expensive than p_T. In the example of a ground robot equipped with a camera, q_Z could be a neural network of similar architecture to p_Z but much shallower, whereas p_T might be a Gaussian or some other simple parametric distribution. Note that in part the difference in complexity arises from the difference in dimensions of the state and observation spaces.

Our goal is to derive calculable probabilistic bounds on $|V_t^{\pi,p_Z} - \hat{V}_t^{\pi,q_Z}|$, while restricting access to p_Z to offline computations only. Additionally, we wish to analyze the difference in using the bounds for post guarantees, i.e. after a policy has been extracted, to when they're used during the decision-making, i.e. policy computation.

We denote shortened notations for the following densities: $\tau_i \triangleq p_T(x_i \mid x_{i-1}, \pi_{i-1})$, $[\zeta/\xi]_i \triangleq [p_Z/q_Z](z_i \mid x_i)$, $[\zeta/\xi]_i^H \triangleq [p_Z/q_Z](z_i \mid H_i^-)$. The values in $[\cdot]$ can be replaced with either respective option. The product of densities is denoted $[\tau/\zeta/\xi]_{i:j} \triangleq \prod_{l=i}^{j} [\tau/\zeta/\xi]_l$. We denote the expectations:

$$\mathbb{E}_{b_i}[\star] \triangleq \int_{x_i} b_i(x_i) \star \mathrm{d}x_i \tag{1}$$

$$\mathbb{E}_{i:j}^{p_T}[\star] \triangleq \int_{x_{i:j}} \tau_{i:j} \star \mathrm{d}x_{i:j} \tag{2}$$

$$\mathbb{E}_{i:j}^{[p_Z/q_Z]}[\star] \triangleq \int_{z_{i:j}} [\zeta/\xi]_{i:j}^H \star \mathrm{d}z_{i:j} \tag{3}$$

$$\mathbb{E}_{i:j}^{p_T,[p_Z/q_Z]}[\star] \triangleq \int_{z_{i:j}} \mathbb{E}_{b_{i-1}}[\mathbb{E}_{i:j}^{p_T}[[\zeta/\xi]_{i:j}\star]] \, \mathrm{d}z_{i:j} \tag{4}$$

$$\mathbb{E}_{i:j+1^-}^{p_T,[p_Z/q_Z]}[\star] \triangleq \int_{z_{i:j}} \mathbb{E}_{b_{i-1}}[\mathbb{E}_{i:j+1}^{p_T}[[\zeta/\xi]_{i:j}\star]] \, \mathrm{d}z_{i:j}, \tag{5}$$

It holds that $[\zeta/\xi]_{i:j}^H = \mathbb{E}_{b_{i-1}}[\mathbb{E}_{i:j}^{p_T}[[\zeta/\xi]_{i:j}]]$ by the law of total probability, and it follows that $\mathbb{E}_{i:j}^{[p_Z/q_Z]}[\star] = \mathbb{E}_{i:j}^{p_T,[p_Z/q_Z]}[\star]$. Additionally, it holds that $\mathbb{E}_{i:j}^{p_T,[p_Z/q_Z]}[\mathbb{E}_{j+1}^{p_T}[\star]] = \mathbb{E}_{i:j+1^-}^{p_T,[p_Z/q_Z]}[\star]$ by Fubini's theorem (Durrett 2019, 1.7).

We denote for a bounded quantity its lower and upper bounds as $\mathcal{LB}$ and $\mathcal{UB}$ respectively. F.e, if $|A| \leq B$, then $\mathcal{LB}(A) = -B$ and $\mathcal{UB}(A) = B$.

For the formulation and practical computation of the bounds, we take the following assumptions:

i. The reward at each time step is bounded: $|r_i(x_i, a)| \leq R_i^{\max}$, for time index i. Hence, follows from the triangle inequality that the value function at a certain time step is bounded by $|V_t^\pi(b_t, a)| \leq \sum_{i=t}^{L} \gamma^{i-t} R_i^{\max} \triangleq V_t^{\max}$, and we define $V_{\max} \triangleq \max_t V_t^{\max}$.

ii. The reachable state space is totally bounded, i.e. for every $\varepsilon > 0$ it can be covered with a finite number of open balls of radius ε.

iii. The models p_T, p_Z, q_Z can be sampled from, and queried for PDF values given samples.

For the result of probabilistic convergence guarantees in Theorem 3, we additionally assume the following:

iv. Weights of particles in particle beliefs are updated via the Sequential Importance Sampling (SIS) algorithm, i.e. $w_t^j \propto [\zeta/\xi]_t \cdot w_{t-1}^j$ for the j'th observation sample z_t^j. Specifically, there is no resampling step.

v. The normalized observation likelihoods of the original and simplified models are bounded almost everywhere w.r.t. the particle filter's proposal distribution: $\sup_{z_{1:t}} \operatorname{ess\,sup}_{x_{0:t} \sim b_0 \cdot \tau_{1:t}} \frac{[\zeta/\xi]_{1:t}}{[\zeta/\xi]^H_{1:t}} \leq d_\infty^{\max}$, where $d_\infty^{\max} \in \mathbb{R}$. See (Lim et al. 2023) for further details.

Bound for Simplified Observation Model

In this section we derive the theoretical bound for the value function with the original observation model and a given policy, w.r.t. the value with the simplified model. We start by stating the following known equivalence relationship between expectation over rewards.

Lemma 1. *The expected belief-dependent reward w.r.t. histories, is equivalent to the expected state-dependent reward w.r.t. the joint distribution of states and observations.*

$$\mathbb{E}_{t+1:i}^{[p_Z/q_Z]}[r_i(b_i, \pi_i)] = \mathbb{E}_{t+1:i}^{p_T,[p_Z/q_Z]}[r_i(x_i, \pi_i)]. \tag{6}$$

Proof. We refer to the supplementary material for all proofs.

We define the following state dependent total variation distance (TV-distance) function,

$$\Delta_Z(x) \triangleq \int_{\mathcal{Z}} |p_Z(z \mid x) - q_Z(z \mid x)| \, \mathrm{d}z. \tag{7}$$

There are several motivations for considering $\Delta_Z(x)$. First, we can estimate it via samples for any density from which we can sample and evaluate for its PDF. It does not require any parametric form of the density, which is useful when considering general learned models. Second, it is state-dependent, hence can be computed locally as we show later for a given belief or trajectory in the belief tree. Different actions might result in different bounds and the locality helps differentiate actions that result in tighter or looser bounds. Third, as given by Pinsker's lemma, the TV-distance is bounded from above by the KL-divergence (Tsybakov 2009). Thus, current approaches that learn probabilistic models with ELBO or directly minimize empirical KL-divergence (Kingma and Welling 2014; Sohn, Lee, and Yan 2015; Rezende and Mohamed 2015; Winkler et al. 2019) also indirectly minimize the TV-distance, therefore it is appropriate to assume that models trained with these objectives will also indirectly minimize Δ_Z.

Theorem 1. *Assume current belief is b_t and a given policy is π. Denote with $b_i^{p_Z}$, $b_i^{q_Z}$ the future belief at time step i updated with either p_Z or q_Z, respectively. Then it holds that*

$$\begin{aligned} &|\mathbb{E}_{t+1:i}^{p_Z}[r_i(b_i^{p_Z}, \pi_i)] - \mathbb{E}_{t+1:i}^{q_Z}[r_i(b_i^{q_Z}, \pi_i)]| \\ &\leq R_i^{\max} \textstyle\sum_{l=t+1}^{i} \mathbb{E}_{t+1:l^-}^{p_T,q_Z}[\Delta_Z(x_l)], \end{aligned} \tag{8}$$

i.e. the difference between the expected reward at future time i for the original and simplified POMDP is bounded by the maximum reward, times the sum of the expected state-dependent TV-distances between the observation models.

By applying to the value function with the triangle inequality, and changing the order of summation, we arrive at the following result.

Corollary 1. *The difference between the original and simplified value functions can be bounded by the following sum of scaled expected TV-distance terms:*

$$|V_t^{\pi,p_Z}(b_t) - V_t^{\pi,q_Z}(b_t)|$$

$$\leq \textstyle\sum_{i=t+1}^{L} \mathbb{E}_{t+1:i^-}^{p_T,q_Z}[\Delta_Z(x_i)] \sum_{l=i}^{L} \gamma^{l-t} R_l^{\max} \tag{9}$$

$$= \textstyle\sum_{i=t+1}^{L} V_i^{\max} \cdot \mathbb{E}_{t+1:i^-}^{p_T,q_Z}[\Delta_Z(x_i)]. \tag{10}$$

Equivalence to State-Action Local Bound

Corollary 1 requires computing the theoretical expectations $\mathbb{E}_{t+1:i^-}^{p_T,q_Z}[\Delta_Z(x_i)]$, which is not trivial with continuous states and observations. Our key insight is that the bound can be viewed as the expected sum of a state-action function. This serves two purposes. The first is that the bounds become more easily estimated, as we can integrate our calculations in existing POMDP planners by adding a secondary reward-like function to compute. The second is that we can extend convergence results of POMDP algorithms to the empirical estimate of the bound.

First we define the following time dependent state-action function, and its extension to a belief function,

$$m_i(x_i, \pi_i) \triangleq V_{i+1}^{\max} \cdot \mathbb{E}_{i+1}^{p_T}[\Delta_Z(x_{i+1})], \tag{11}$$

$$m_i(b_i, \pi_i) \triangleq \mathbb{E}_{b_i}[m(x_i, \pi_i)]. \tag{12}$$

Intuitively speaking, m_i bounds the loss of the value function at time i when considering the action of π_i from state x_i or belief b_i, based on Δ_Z. It is then natural to extend this definition with the following:

1. Cumulative bound for a given policy and initial belief, $M_t^\pi(b_t) \triangleq \mathbb{E}_{t+1:L-1}^{q_Z}[\sum_{i=t}^{L-1} m_i(b_i, \pi_i)]$, analogous to $V_t^\pi(b_t)$. Note that the time horizon has decreased by 1, and there is no discount factor.
2. Action cumulative bound for a given policy and initial belief, $\Phi_t^\pi(b_t, a) \triangleq m_t(b_t, a) + \mathbb{E}_{t+1}^{q_Z}[M_{t+1}^\pi(b_{t+1})]$, analogous to $Q_t^\pi(b_t, a)$.

Theorem 2. *Under the conditions of Theorem 1, the difference between the original and simplified value function can be bounded by*

$$|V_t^{\pi,p_Z}(b_t) - V_t^{\pi,q_Z}(b_t)| \leq M_t^\pi(b_t). \tag{13}$$

In addition, the respective difference in action value function can be bounded by the action cumulative bound,

$$|Q_t^{\pi,p_Z}(b_t, a) - Q_t^{\pi,q_Z}(b_t, a)| \leq \Phi_t^\pi(b_t, a). \tag{14}$$

Computing $M_t^\pi(b_t)$ would be useful when trying to arrive at a bound for a specific policy, usually as post-guarantees after a planning session. On the other hand, the computation of $\Phi_t^\pi(b_t, a)$ can be defined recursively over an entire belief tree, meaning it extends the original definition of the bound to all extractable policies from a given tree. Hence, the action cumulative bound can be used during planning, when computing a policy or pruning low value branches like in (Sztyglic and Indelman 2022).

Online Estimator of Local Bound

We now show how we estimate m_i during online planning without requiring access to p_Z. We do this by offline pre-sampling state samples $\{x_n^{\Delta}\}_{n=1}^{N_\Delta} \sim Q_0(x)$, named delta states, and evaluating $\Delta_Z(x_n^{\Delta})$ for each. During online planning, we simply reweight Δ_Z using the importance sampling formalism for state sample x_i^j. We use this to redefine m_i and also to define the empirical estimate $\tilde{m}_i$.

$$m_i(x_i^j, a) = V_{i+1}^{\max} \cdot \mathbb{E}_{x' \sim Q_0}[\tfrac{p_T(x'|x_i^j, a)}{Q_0(x')} \Delta_Z(x')], \tag{15}$$

$$\tilde{m}_i(x_i^j, a) \triangleq V_{i+1}^{\max} \cdot \tfrac{1}{N_\Delta} \textstyle\sum_{i=1}^{N_\Delta} \tfrac{p_T(x_n^{\Delta}|x_i^j, a)}{Q_0(x_n^{\Delta})} \Delta_Z(x_n^{\Delta}). \tag{16}$$

Based on this, we define $\tilde{\Phi}_t^{\pi}(b_t, a)$ when computed with $\tilde{m}_i$ instead of m_i. We explicitly define this notation to differentiate $\tilde{\Phi}$ from $\hat{\tilde{\Phi}}$; The former is defined with a theoretical expectation over the observation space, while the latter is an empirical estimate with sampled belief trajectories. For notational brevity, in the rest of the paper we denote $\hat{\Phi} \triangleq \hat{\tilde{\Phi}}$.

The support of Q_0 must cover all of $\mathcal{X}$ for the estimator in (16) to be consistent (Doucet, de Freitas, and Gordon 2001, 1.3.2). Although it is possible to construct such densities over unbounded state spaces, we find that in practice it is often not required as $\mathcal{X}$ is either naturally bounded, such as the configuration space of a manipulator, or can be defined to be large enough but bounded to cover the relevant domain of the planning problem. Hence, it could be sufficient to choose Q_0 with a finite support. To the end of simplifying the discussion, we took assumption ii.

Under this condition, we can establish with Hoeffding's inequality the following simple probabilistic bound:

$$\mathbb{P}(|m_i(x, a) - \tilde{m}_i(x, a)| \geq \nu) \leq 2\exp(-2N_\Delta \tfrac{\nu^2}{B_i^2}), \tag{17}$$

where $B_i = 2 \cdot V_i^{\max} \cdot \max_{x, x', a} \frac{p_T(x'|x, a)}{Q_0(x')}$ since $0 \leq \Delta_Z(x) \leq 2$. We assume B_i to be computable offline. This is sufficient for obtaining the concentration bounds of Theorem 3, we approximate this even further N_Δ may be very large, as we describe in section .

Convergence Guarantees

Our major convergence result is based on an adaptation of Lemma 1 of (Lim et al. 2023). Lemma 1 shows that the theoretical optimal POMDP action value $Q_{\mathbf{P}}^*$ is bounded with high probability from the theoretical optimal PB-MDP action value $Q_{\mathbf{M_P}}^*$. We extend their result in two ways: First, proving the bound for general policies, rather than only the optimal one, allows for proving convergence of Φ_t^{π} w.r.t. a policy that is chosen to maximize $V_t^{\pi, qz}$, or any other objective. Second, by introducing an additional assumption of a probabilistically bounded reward, denoted with $\tilde{r}$, we are able to apply Theorem 3 to $\tilde{m}_i$.

Theorem 3 (Generalized PB-MDP Convergence)**.** *Assume that the immediate state reward estimate is probabilistically bounded such that $\mathbb{P}(|r_i^j - \tilde{r}_i^j| \geq \nu) \leq \delta_r(\nu, N_r)$, for a number of reward samples N_r and state sample x_i^j. Assume that $\delta_r(\nu, N_r) \to 0$ as $N_r \to \infty$. For all policies π, $t = 0, \ldots, L$ and $a \in \mathcal{A}$, the following bounds hold with probability of at least $1 - 5(4C)^{L+1}(\exp(-C \cdot \acute{k}^2) + \delta_r(\nu, N_r))$:*

$$|Q_{\mathbf{P},t}^{\pi,[pz/qz]}(b_t, a) - Q_{\mathbf{M_P},t}^{\pi,[pz/qz]}(\bar{b}_t, a)| \leq \alpha_t + \beta_t, \tag{18}$$

where,

$$\alpha_t = (1+\gamma)\lambda + \gamma\alpha_{t+1},\ \alpha_L = \lambda \geq 0, \tag{19}$$

$$\beta_t = 2\nu + \gamma\beta_{t+1},\ \beta_L = 2\nu \geq 0, \tag{20}$$

$$k_{\max}(\lambda, C) = \frac{\lambda}{4V_{\max} d_{\infty}^{\max}} - \frac{1}{\sqrt{C}} > 0, \tag{21}$$

$$\acute{k} = \min\{k_{\max}, \lambda/4\sqrt{2}V_{\max}\}. \tag{22}$$

If we require the bound to hold for all possible policies that can be extracted from a given belief tree simultaneously, then under the assumption of a finite action space, the probability is at least $1 - 5(4|\mathcal{A}|C)^{L+1}(\exp(-C \cdot \acute{k}^2) + \delta_r(\nu, N_r))$.

Corollary 2. *For arbitrary precision ε and accuracy δ we can choose constants λ, ν, C, N_r such that the following holds with probability of at least $1 - \delta$:*

$$|Q_{\mathbf{P},t}^{\pi,[pz/qz]}(\bar{b}_t, a) - Q_{\mathbf{M_P},t}^{\pi,[pz/qz]}(\bar{b}_t, a)| \leq \varepsilon. \tag{23}$$

By applying Corollary 2 twice, once to the original reward function then once to the action cumulative bound, we can conclude that the estimated simplified PB-MDP value is probabilistically bounded from the original theoretical POMDP value.

We now arrive at our key theoretical result.

Corollary 3. *Assuming that $\mathcal{P}$ is an MDP planner that can approximate Q-values with arbitrary precision $\varepsilon^{\mathcal{P}}$ at an accuracy $1 - \delta^{\mathcal{P}}$, we denote the precision and accuracy of the action value and action cumulative bound functions:*

$$\mathbb{P}(|Q_{\mathbf{M_P},t}^{\pi,qz}(\bar{b}_t, a) - \hat{Q}_{\mathbf{M_P},t}^{\pi,qz}(\bar{b}_t, a)| \leq \varepsilon_Q^{\mathcal{P}}) \geq 1 - \delta_Q^{\mathcal{P}} \tag{24}$$

$$\mathbb{P}(|\Phi_{\mathbf{M_P},t}^{\pi}(\bar{b}_t, a) - \hat{\Phi}_{\mathbf{M_P},t}^{\pi}(\bar{b}_t, a)| \leq \varepsilon_\Phi^{\mathcal{P}}) \geq 1 - \delta_\Phi^{\mathcal{P}} \tag{25}$$

From Corollary 2 it holds that we can choose constants λ, ν, C, N_r such that the following holds,

$$\mathbb{P}(|Q_{\mathbf{P},t}^{\pi,qz}(b_t, a) - Q_{\mathbf{M_P},t}^{\pi,qz}(\bar{b}_t, a)| \leq \varepsilon_Q) \geq 1 - \delta_Q, \tag{26}$$

$$\mathbb{P}(|\Phi_{\mathbf{P},t}^{\pi}(b_t, a) - \tilde{\Phi}_{\mathbf{M_P},d}^{\pi}(\bar{b}_t, a)| \leq \varepsilon_\Phi) \geq 1 - \delta_\Phi. \tag{27}$$

Then with probability of at least $1 - (\delta_Q + \delta_Q^{\mathcal{P}} + \delta_\Phi + \delta_\Phi^{\mathcal{P}})$

$$|Q_{\mathbf{P},t}^{\pi,pz}(b_t, a) - \hat{Q}_{\mathbf{M_P},t}^{\pi,qz}(\bar{b}_t, a)| \leq \tag{28}$$

$$\hat{\Phi}_{\mathbf{M_P},t}^{\pi}(\bar{b}_t, a) + \varepsilon_Q + \varepsilon_Q^{\mathcal{P}} + \varepsilon_\Phi + \varepsilon_\Phi^{\mathcal{P}}. \tag{29}$$

In summary, any planner that can approximate the PB-MDP values $Q_{\mathbf{M_P},t}^{\pi,qz}(\bar{b}_t, a)$ with high probability can also approximate the cumulative bound $\Phi_{\mathbf{M_P},t}^{\pi}(\bar{b}_t, a)$ with high probability, since it is mathematically formulated like a state reward. Therefore, we can construct a bound for the theoretical value $Q_{\mathbf{P},t}^{\pi,pz}(b_t, a)$ with high probability from online calculations that do not involve access to p_Z.

An important observation is that the probabilistic bound obtained by Theorem 3 is independent of the chosen policy, i.e. constant w.r.t. the actions. Therefore, when using the bound in Corollary 3 for decision-making, if our goal is to distinguish between actions that result in maximal lower or upper bound, i.e. $\max_a [\mathcal{LB}/\mathcal{UB}](Q^{\pi,p_Z}_{\mathbf{P},t}(b_t, a))$, we will obtain the same action choice by computation of $\max_a \hat{Q}^{\pi,q_Z}_{\mathbf{M_P},t}(\bar{b}_t, a)[-/+]\hat{\Phi}^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a)$.

Implementation

We verify the computability of our approach in a 2D beacons POMDP. [1] We integrate the computation of $\tilde{m}_i$ and $\hat{\Phi}_{\mathbf{M_P},t}$ into PFT-DPW, and showcase an example of where the bounds could affect a policy's decision-making.

Simulative Setting

Our experimental setting is a 2D light-dark inspired simulation, shown in figure 3.

An agent is in a wall-surrounded arena, with a gate at the bottom to a goal region. The agent's starting location is either to the left or to the right of the goal region. The agent's task is to enter the goal region without colliding with a wall. The observations are noisy measurements of the agent's location, being more certain when in the "light" region, and less when in the "dark" region.

The light region $\mathcal{X}_{\text{light}}$ is defined by circles centered at each of the 6 beacons located at the top of the arena, and $\mathcal{X}_{\text{dark}} = \mathcal{X} \setminus \mathcal{X}_{\text{light}}$. The observation model in the light region $p_Z(z \mid x \in \mathcal{X}_{\text{light}})$ is defined as a Gaussian mixture model (GMM) with 1126 components arranged such that they approximately form a truncated Gaussian distribution centered at x. The simplified observation model differs only in the light region, and it approximates p_Z with a single Gaussian: $q_Z(z \mid x \in \mathcal{X}_{\text{light}}) \sim \mathcal{N}(x, \Sigma^{q_Z}_{\text{light}})$. We refer readers to (Bar-Shalom, Li, and Kirubarajan 2004, 1.4.16) on approximating GMMs with a single Gaussian. In the dark region, $p_Z(z \mid x \in \mathcal{X}_{\text{dark}}) = q_Z(z \mid x \in \mathcal{X}_{\text{dark}}) = \mathcal{N}(x, \Sigma_{dark})$

This setting demonstrates a case where the original observation model is an overly parameterized model, like a complex neural network, and one would like to replace it with a less parameterized albeit similar model.

The action space is $\mathcal{A} = \{(\pm 1, 0), (0, \pm 1)\}$. The transition model $p_T(x' \mid x, a)$ is a Gaussian centered at the agent's location plus the action. The horizon is $L = 15$, and the POMDP will terminate early if the agent enters the goal region or collides with a wall.

The reward is only state and time dependent, and is a sum of three indicators: $r_t(x) = R_{hit} \cdot \mathbf{1}_{x \in \mathcal{X}_{goal}} + R_{miss} \cdot \mathbf{1}_{x \notin \mathcal{X}_{goal}} + R_{collide} \cdot \mathbf{1}_{x \in \mathcal{X}_{collision}}$. In all time steps, $R_{hit} = 100$, $R_{collide} = -50$. The miss reward is $R_{miss} = -50$ if $t = L$ and is -1 otherwise. The discount factor is $\gamma = 1$.

Further details of the experimental setup can be found in the supplementary material.

Implementation of Bounds

When sampling delta states $\{x^{\Delta}_n\}^{N_{\Delta}}_{n=1}$ offline, without prior knowledge of which states are more likely, we choose a uniform Q_0. In order to assure even coverage of the state space, we chose to sample them from the quasi-random sequence R_2, which has been shown empirically to minimize the discrepancy (Roberts 2018). Hence, we perform in practice quasi Monte Carlo method (QMCM) for the computation of $\tilde{m}_i$ (Caflisch 1998). It has been shown in many empirical examples that QMCM obtains faster convergence in practice than regular MC with an equivalent number of samples. In theory, it is possible to provide a deterministic upper bound to the QMCM integration error via the Koksma-Hlwaka inequality (Lemieux 2009, 5.6). However, it is generally hard to compute, and infinite for several very simple functions.

For each sampled x^{Δ}_n we estimate Δ_Z via observation samples, based on assumption iii. We perform the following importance sampling estimation w.r.t. $(p_Z + q_Z)/2$:

$$\hat{\Delta}_Z(x^{\Delta}_n) = \sum_{j=1}^{N_Z} 2 \cdot \frac{|p_Z(z^n_j \mid x^{\Delta}_n) - q_Z(z^n_j \mid x^{\Delta}_n)|}{p_Z(z^n_j \mid x^{\Delta}_n) + q_Z(z^n_j \mid x^{\Delta}_n)} \tag{30}$$

where $\{z^n_j\}^{N_Z}_{n=1} \overset{i.i.d.}{\sim} (p_Z + q_Z)/2$. It is possible to quantify the MC estimation error from this step, however we assume that with enough offline compute power it could be made negligible, such that $\Delta_Z \approx \hat{\Delta}_Z$.

We did several optimizations in order to compute $\tilde{m}_i$ in a time efficient manner. The first is by pre-filtering to only keep x^{Δ}_n for which $\Delta_Z(x^{\Delta}_n) > \Delta_{Thresh}$. The second optimization is to only consider sampled states within a truncation distance of d_T from state particle x^j_i in (16). We implemented this by keeping all x^{Δ}_n in a KD-tree for efficient radius queries (Maneewongvatana and Mount 1999). We chose Δ_{Thresh} and d_T such that the error in computing $\tilde{m}_i$ is at most $V^{\max} \cdot 10^{-4}$. Lastly, since the runtime complexity of $\tilde{m}_i$ grows linearly with C, the number of particles in the belief $\bar{b}_i$, we limit the number of particles used to N_x by performing MC estimation w.r.t. the particle belief: $\hat{\tilde{m}}_i(\bar{b}_i, a) = \frac{1}{N_x} \sum_{j=1}^{N_X} \tilde{m}_i(x^j_i, a)$ where $\{x^j_i\}^{N_X}_{j=1} \overset{i.i.d.}{\sim} \bar{b}_i$.

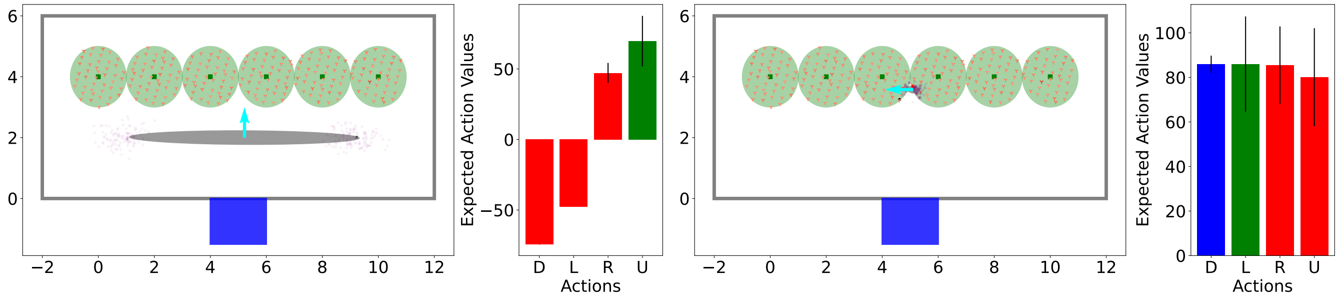
Empirical Bound Evaluation

In all scenarios, a planning session is fixed at 500 simulations of PFT-DPW with the same parameters, as described in the supplementary material.

We record for each time step the estimated expected value $\hat{Q}^{[p_Z/q_Z]}_t$ and the plan session duration. In simplified planning, we also record the expected action cumulative bound $\hat{\Phi}_{\mathbf{M_P},t}$. The policy is maximization of the estimated action value function based on the original or simplified model, i.e. $\pi^{[p_Z/q_Z]}_t \triangleq \max_a \hat{Q}^{[p_Z/q_Z]}_t(\bar{b}_t, a)$, and we respectively name them as the original or simplified value policy. After each planning session, we apply the selected action, update the particle belief according to the observation received, and continue the scenario until the POMDP terminates.

In our first test, we run 100 scenarios of each planning scheme - once with the original observation model, and once with the simplified with the additional computation of $\hat{\Phi}_{\mathbf{M_P},t}$. As shown in figure 4, the simplified planning time is greatly reduced for the same number of simulations compared to planning with the original model. These results were expected because of increased overhead for sampling

[1] Our code is publicly available at https://github.com/IdanLevYehudi/SimplifyingObsPOMDP



(a) Start of scenario, $t = 0$.

(b) Middle of scenario, $t = 4$.

Figure 3: The results of two planning sessions in 2D beacons. The goal is indicated by the blue rectangle, the beacons and their radii by the green squares and circles, and the outer walls by the grey outer rectangle. The filtered delta states $\{x_n^\Delta\}_{n=1}^{N_\Delta^{kept}}$ are indicated by the tri-downs, with color relative to estimated TV-distance of the simplified observation model $7.06 \leq \hat{\Delta}_Z \cdot 10^2 \leq 12.08$. The colored dots indicate: the true state in black, the observation in red and the particle belief in purple. The belief empirical mean and covariance are the grey ellipse. The bars on the right depict $\hat{Q}_t^{q_Z}$ for all actions, and $\hat{\Phi}_t$ as symmetric error bars. The action chosen by the simplified value policy $\pi_t^{q_Z}$ is colored in green, and the lower bound policy $\pi_t^{\mathcal{LB}}$ in blue if different. At $t = 4$ we can see an inconsistency of action order between $\pi_t^{q_Z}$ that chooses left, whereas $\pi_t^{\mathcal{LB}}$ chooses down.

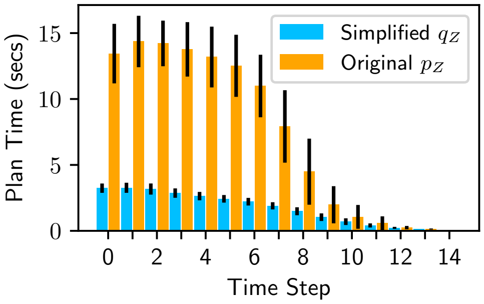


Figure 4: Mean and standard deviation of planning duration over 100 scenarios vs. scenario time step, with the original observation model p_Z or simplified model q_Z.

and evaluating the PDF of the original observation model, leading to longer planning times even when additionally computing $\hat{\Phi}$ in the simplified planning.

In our second test, we quantified how often the simplified policy is different from the lower or upper bound policies. We denote the lower bound policy as $\pi_t^{\mathcal{LB}} \triangleq \arg\max_a\{\hat{Q}_t^{q_Z}(\bar{b}_t, a) - \hat{\Phi}_t(\bar{b}_t, a)\}$ and the upper bound policy as $\pi_t^{\mathcal{UB}} \triangleq \arg\max_a\{\hat{Q}_t^{q_Z}(\bar{b}_t, a) + \hat{\Phi}_t(\bar{b}_t, a)\}$. In our results in figure 5, we can see that there is a great difference between the policies, in particular in time steps 1-7, which is when the agent is mostly around the light region. $\pi^{\mathcal{LB}}$ tends to steer away from the light region, hence it differs mostly in the earlier time steps, whereas $\pi^{\mathcal{UB}}$ prefers the light region, hence it chooses to stay there during the descent of the agent towards the goal. These results indicate that the bounds are non-trivial, and corresponding policies do point towards different objectives as the scenario progresses.

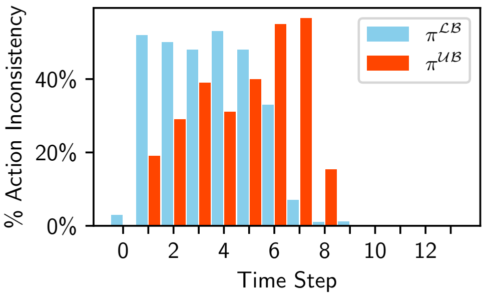


Figure 5: Percentage of scenarios in each time step in which the lower or upper bound policies, $\pi^{\mathcal{LB}}$ or $\pi^{\mathcal{UB}}$, chose an action different from the simplified value policy π^{q_Z}.

Conclusion

This paper builds upon the paradigm of solving POMDP planning problems with simplification for adhering to computational limitations. We suggest a planning framework with a simplified observation model, to the end of practically solving POMDPs with complex high dimensional observations like visual observations, with finite time performance guarantees. We formulate a novel bound based on local reweighting of pre-calculated TV-distances at pre-sampled states, and show that its estimator bounds with high probability the theoretical value function of the original problem. Finally, an example showcasing how the bounds can influence the decision-making process is presented for both the lower and upper bound policies. In future research, we envision the use of our bounds during the planning process itself, for pruning of action branches to the end of runtime improvement, or for certifying performance when they're computed explicitly.

Acknowledgements

This work was supported by the Israel Science Foundation (ISF) and by US NSF/US-Israel BSF.

References

Bar-Shalom, Y.; Li, X. R.; and Kirubarajan, T. 2004. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons.

Barenboim, M.; Lev-Yehudi, I.; and Indelman, V. 2023. Data Association Aware POMDP Planning with Hypothesis Pruning Performance Guarantees. *IEEE Robotics and Automation Letters (RA-L)*.

Caflisch, R. E. 1998. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7: 1–49.

Deglurkar, S.; Lim, M. H.; Tucker, J.; Sunberg, Z. N.; Faust, A.; and Tomlin, C. 2023. Compositional Learning-based Planning for Vision POMDPs. In *Learning for Dynamics and Control Conference*, 469–482. PMLR.

Doucet, A.; de Freitas, N.; and Gordon, N., eds. 2001. *Sequential Monte Carlo Methods In Practice*. New York: Springer-Verlag.

Durrett, R. 2019. *Probability: theory and examples*, volume 49. Cambridge university press.

Eslami, S. A.; Rezende, D. J.; Besse, F.; Viola, F.; Morcos, A. S.; Garnelo, M.; Ruderman, A.; Rusu, A. A.; Danihelka, I.; Gregor, K.; et al. 2018. Neural scene representation and rendering. *Science*, 360(6394): 1204–1210.

Ha, D.; and Schmidhuber, J. 2018. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems (NIPS)*, 31.

Hoerger, M.; and Kurniawati, H. 2021. An On-Line POMDP Solver for Continuous Observation Spaces. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 7643–7649. IEEE.

Hoerger, M.; Kurniawati, H.; Bandyopadhyay, T.; and Elfes, A. 2020. Linearization in motion planning under uncertainty. In *Intl. Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 272–287. Springer.

Hoerger, M.; Kurniawati, H.; and Elfes, A. 2023. Multilevel Monte Carlo for solving POMDPs on-line. In *Intl. J. of Robotics Research*, volume 42, 196–213. Sage Publications Sage UK: London, England.

Igl, M.; Zintgraf, L.; Le, T. A.; Wood, F.; and Whiteson, S. 2018. Deep variational reinforcement learning for POMDPs. In *International Conference on Machine Learning*, 2117–2126. PMLR.

Jonschkowski, R.; Rastogi, D.; and Brock, O. 2018. Differentiable Particle Filters: End-to-End Learning with Algorithmic Priors. In *Robotics: Science and Systems (RSS)*.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1): 99–134.

Karkus, P.; Hsu, D.; and Lee, W. S. 2017. Qmdp-net: Deep learning for planning under partial observability. In *Advances in Neural Information Processing Systems (NIPS)*, 4694–4704.

Karkus, P.; Hsu, D.; and Lee, W. S. 2018. Particle Filter Networks with Application to Visual Localization. In *Conference on Robot Learning*.

Kearns, M.; Mansour, Y.; and Ng, A. Y. 2002. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2): 193–208.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Kong, A.; Liu, J. S.; and Wong, W. H. 1994. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425): 278–288.

Lemieux, C. 2009. *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer.

Lim, M. H.; Becker, T. J.; Kochenderfer, M. J.; Tomlin, C. J.; and Sunberg, Z. N. 2022. Generalized optimality guarantees for solving continuous observation POMDPs through particle belief MDP approximation. *arXiv preprint arXiv:2210.05015*.

Lim, M. H.; Becker, T. J.; Kochenderfer, M. J.; Tomlin, C. J.; and Sunberg, Z. N. 2023. Optimality guarantees for particle belief approximation of POMDPs. *Journal of Artificial Intelligence Research*, 77: 1591–1636.

Lim, M. H.; Tomlin, C.; and Sunberg, Z. N. 2020. Sparse Tree Search Optimality Guarantees in POMDPs with Continuous Observation Spaces. In *Intl. Joint Conf. on AI (IJCAI)*, 4135–4142.

Lim, M. H.; Tomlin, C. J.; and Sunberg, Z. N. 2021. Voronoi progressive widening: efficient online solvers for continuous state, action, and observation POMDPs. In *2021 60th IEEE conference on decision and control (CDC)*, 4493–4500. IEEE.

Maneewongvatana, S.; and Mount, D. M. 1999. It's okay to be skinny, if your friends are fat. In *Center for geometric computing 4th annual workshop on computational geometry*, volume 2, 1–8. Citeseer.

Papadimitriou, C.; and Tsitsiklis, J. 1987. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3): 441–450.

Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *Intl. Conf. on Machine Learning (ICML)*, 1530–1538. PMLR.

Roberts, M. 2018. The Unreasonable Effectiveness of Quasirandom Sequences. https://extremelearning.com.au/unreasonable-effectiveness-of-quasirandom-sequences/. Accessed: 2023-08-12.

Shienman, M.; and Indelman, V. 2022. D2A-BSP: Distilled Data Association Belief Space Planning with Performance Guarantees Under Budget Constraints. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*.

Silver, D.; and Veness, J. 2010. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems (NIPS)*, 2164–2172.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, 3483–3491.

Sunberg, Z.; and Kochenderfer, M. 2018. Online algorithms for POMDPs with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28.

Sztyglic, O.; and Indelman, V. 2022. Speeding up Online POMDP Planning via Simplification. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*.

Tsybakov, A. B. 2009. *Introduction to nonparametric estimation*. New York ; London : Springer.

Wang, Y.; Liu, B.; Wu, J.; Zhu, Y.; Du, S. S.; Fei-Fei, L.; and Tenenbaum, J. B. 2020. DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 4190–4198. International Joint Conferences on Artificial Intelligence Organization. Main track.

Winkler, C.; Worrall, D.; Hoogeboom, E.; and Welling, M. 2019. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Ye, N.; Somani, A.; Hsu, D.; and Lee, W. S. 2017. DESPOT: Online POMDP planning with regularization. *JAIR*, 58: 231–266.

Proofs

Lemma 1

Lemma 1. *The expected belief-dependent reward w.r.t. histories, is equivalent to the expected state-dependent reward w.r.t. the joint distribution of states and observations.*

$$\mathbb{E}^{[p_Z/q_Z]}_{t+1:i}[r_i(b_i,\pi_i)] = \mathbb{E}^{p_T,[p_Z/q_Z]}_{t+1:i}[r_i(x_i,\pi_i)]. \tag{31}$$

Proof. Without loss of generality we prove for p_Z.

From the definition of the expectation, we write it as the following explicit integral:

$$\mathbb{E}^{p_Z}_{t+1:i}[r_i(b_i,\pi_i)] \tag{32}$$

$$= \int\limits_{z_{t+1:i}} p_Z\left(z_{t+1} \mid H^-_{t+1}\right) \cdots p_Z\left(z_i \mid H^-_i\right)$$
$$r_i\left(b_i,\pi_i(H_i)\right) \mathrm{d}z_{t+1:i} \tag{33}$$

From the assumption of a state reward and Fubini's theorem, we replace the belief reward with the expected state reward:

$$\mathbb{E}^{p_Z}_{t+1:i}[r_i(b_i,\pi_i)]$$

$$= \int\limits_{z_{t+1:i}} \int\limits_{x_i} p_Z\left(z_{t+1} \mid H^-_{t+1}\right) \cdots p_Z\left(z_i \mid H^-_i\right)$$
$$b_i\left(x_i\right) r_i\left(x_i,\pi_i(H_i)\right) \mathrm{d}x_i\, \mathrm{d}z_{t+1:i} \tag{34}$$

We apply the following steps of Bayes' rule, and marginalization and chain rule repetitively from time i until time t.

1. We apply Bayes' rule to the belief $b_i(x_i)$. By definition, $b_i(x_i) \triangleq \mathbb{P}(x_i \mid H_{i-1},\pi_{i-1}(H_{i-1}),z_i)$, and therefore $b_i(x_i) = \frac{p_Z(z_i|x_i)}{p_Z\left(z_i|H^-_i\right)} b^-_i\left(x_i\right)$.

$$\mathbb{E}^{p_Z}_{t+1:i}[r_i(b_i,\pi_i)]$$

$$= \int\limits_{z_{t+1:i}} \int\limits_{x_i} p_Z\left(z_{t+1} \mid H^-_{t+1}\right) \cdots \cancel{p_Z\left(z_i \mid H^-_i\right)}$$
$$\frac{p_Z\left(z_i \mid x_i\right) b^-_i\left(x_i\right)}{\cancel{p_Z\left(z_i \mid H^-_i\right)}} r_i\left(x_i,\pi_i(H_i)\right) \mathrm{d}x_i\, \mathrm{d}z_{t+1:i} \tag{35}$$

2. We marginalize over the state x_{i-1} and then apply the chain rule to the propagated belief $b^-_i(x_i)$. By definition the propagated belief satisfies $b^-_i(x_i) \triangleq \mathbb{P}(x_i \mid H_{i-1},\pi_{i-1}(H_{i-1}))$, and therefore it holds that $b^-_i(x_i) = \int_{x_{i-1}} b_{i-1}\left(x_{i-1}\right) p_T\left(x_i \mid x_{i-1},\pi_{i-1}\left(H_{i-1}\right)\right)$.

$$\mathbb{E}^{p_Z}_{t+1:i}[r_i(b_i,\pi_i)]$$

$$= \int\limits_{z_{t+1:i}} \int\limits_{x_{i-1:i}} p_Z\left(z_{t+1} \mid H^-_{t+1}\right) \cdots p_Z\left(z_{i-1} \mid H^-_{i-1}\right)$$
$$b_{i-1}\left(x_{i-1}\right) p_T\left(x_i \mid x_{i-1},\pi_{i-1}\left(H_{i-1}\right)\right) p_Z\left(z_i \mid x_i\right)$$
$$r_i\left(x_i,\pi_i(H_i)\right) \mathrm{d}x_{i-1:i}\, \mathrm{d}z_{t+1:i} \tag{36}$$

By repeating steps 1 and 2 until marginalizing over the state x_t, we conclude that:

$$\mathbb{E}^{p_Z}_{t+1:i}[r_i(b_i,\pi_i)]$$

$$= \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t\left(x_t\right) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1},\pi_{j-1}\left(H_{j-1}\right)\right)$$
$$\cdot \prod_{k=t+1}^{i} p_Z\left(z_k \mid x_k\right) \cdot r_i\left(x_i,\pi_i(H_i)\right) \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i}$$
$$= \mathbb{E}^{p_T,p_Z}_{t+1:i}\left[r_i(x_i,\pi_i)\right] \tag{37}$$

□

Theorem 1

Theorem 1. *Assume current belief is b_t and a given policy is π. Denote with $b^{p_Z}_i$, $b^{q_Z}_i$ the future belief at time step i updated with either p_Z or q_Z, respectively. Then it holds that*

$$|\mathbb{E}^{p_Z}_{t+1:i}[r_i(b^{p_Z}_i,\pi_i)] - \mathbb{E}^{q_Z}_{t+1:i}[r_i(b^{q_Z}_i,\pi_i)]|$$
$$\leq R^{\max}_i \textstyle\sum_{l=t+1}^{i} \mathbb{E}^{p_T,q_Z}_{t+1:l^-}\left[\Delta_Z(x_l)\right], \tag{38}$$

i.e. the difference between the expected reward at future time i for the original and simplified POMDP is bounded by the maximum reward, times the sum of the expected state-dependent TV-distances between the observation models.

Proof. We note that the two expectations share integration domain. By the linearity of the integral, we can combine integrands that are equal across the domains, and use the distributive property of the integral. We then apply the integral triangle inequality to the difference term of the two different observation models.

$$|\mathbb{E}^{p_Z}_{t+1:i}[r_i(b^{p_Z}_i,\pi_i)] - \mathbb{E}^{q_Z}_{t+1:i}[r_i(b^{q_Z}_i,\pi_i)]| \tag{39}$$
$$= |\mathbb{E}^{p_T,p_Z}_{t+1:i}\left[r_i(x_i,\pi_i)\right] - \mathbb{E}^{p_T,q_Z}_{t+1:i}\left[r_i(x_i,\pi_i)\right]| \tag{40}$$

$$= \left| \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1},\pi_{j-1}\left(H_{j-1}\right)\right) \right.$$

$$\left(\prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right)\right)$$

$$\left. \cdot r(x_i, \pi_i(H_i))\, \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \right| \tag{41}$$

$$\leq \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\left|\prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right)\right|$$

$$\cdot \left|r(x_i, \pi_i(H_i))\right| \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{42}$$

We utilize the assumption of a bounded state reward, of the form $|r(x_i, \pi_i(H_i))| \leq R_i^{\max}$.

$$\leq R_i^{\max} \cdot \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t)$$

$$\prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left|\prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right)\right| \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{43}$$

Next, we add and subtract a mixed term of $i-t-1$ simplified observation factors q_Z and a single original observation factor p_Z. This factor is carefully chosen to pick together common factors such that we'll be left with our desired expectations - computed over the simplified observation model.

$$= R_i^{\max} \cdot \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t)$$

$$\prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right) \tag{44}$$

$$\cdot \left|\prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) p_Z\left(z_i \mid x_i\right)\right.$$

$$\left. + \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) p_Z\left(z_i \mid x_i\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right)\right|$$

$$\mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{45}$$

$$= R_i^{\max} \cdot \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t)$$

$$\prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right) \tag{46}$$

$$\cdot \left|\prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right)\left(p_Z\left(z_i \mid x_i\right) - q_Z\left(z_i \mid x_i\right)\right) + \right.$$

$$\left. p_Z\left(z_i \mid x_i\right)\left(\prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right)\right)\right|$$

$$\mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{47}$$

We once more use the triangle inequality and the linearity of integral to separate the obtained integral into two. We denote these terms as (1) and (2) and then handle each separately.

$$\leq R_i^{\max}$$

$$\cdot (1) \triangleq \left\{ \begin{array}{l} \displaystyle\int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right) \\ \displaystyle\quad \cdot \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) \left|p_Z\left(z_i \mid x_i\right) - q_Z\left(z_i \mid x_i\right)\right| \\ \hfill \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \end{array} \right. \tag{48}$$

$$+R_i^{\max}$$

$$\cdot (2) \triangleq \left\{ \begin{array}{l} \displaystyle\int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right) \\ \displaystyle \cdot p_Z\left(z_i \mid x_i\right) \left|\prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right)\right| \\ \hfill \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \end{array} \right. \tag{49}$$

The first term turns out to become the expected TV-distance w.r.t. trajectories taken with the simplified observation model.

$$(1) = \int\limits_{z_{t+1:i-1}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) \left(\int\limits_{z_i} \left|p_Z\left(z_i \mid x_i\right) - q_Z\left(z_i \mid x_i\right)\right| \mathrm{d}z_i\right)$$

$$\mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i}$$

$$= R_i^{\max} \cdot \mathbb{E}_{t+1:i^-}^{p_T, q_Z}\left[\Delta_Z(x_i)\right] \tag{50}$$

The second term is the recursive term, in which we can integrate out first the last observation factor, and then the last transition factor.

$$(2) = \int\limits_{z_{t+1:i-1}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left|\prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right)\right|$$

$$\cancel{\left(\int\limits_{z_i} p_Z\left(z_i \mid x_i\right) \mathrm{d}z_{i-1}\right)} \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{51}$$

$$= \int\limits_{z_{t+1:i-1}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i-1} p_T\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left| \prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) \right|$$

$$\cancel{\left(\int_{x_i} p\left(x_i \mid x_{i-1}, \pi_{i-1}\left(z_{i-1}\right)\right) \mathrm{d}x_i\right)} \mathrm{d}x_{t:i-1}\, \mathrm{d}z_{t+1:i} \tag{52}$$

$$= \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i-1}} b_t(x_t) \prod_{j=t+1}^{i-1} p\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left| \prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) \right| \mathrm{d}x_{t:i-1}\, \mathrm{d}z_{t+1:i-1} \tag{53}$$

Notice the recursive relationship:

$$\int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left| \prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right) \right| \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{54}$$

$$\leq \mathbb{E}^{p_T, q_Z}_{t+1:i^-}\left[\Delta_Z(x_i)\right]$$

$$+ \int\limits_{z_{t+1:i}} \int\limits_{x_{t:i-1}} b_t(x_t) \prod_{j=t+1}^{i-1} p\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left| \prod_{j=t+1}^{i-1} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i-1} q_Z\left(z_j \mid x_j\right) \right| \mathrm{d}x_{t:i-1}\, \mathrm{d}z_{t+1:i-1} \tag{55}$$

In the base case of $i = t+1$ we get directly

$$\int\limits_{z_{t+1:i}} \int\limits_{x_{t:i}} b_t(x_t) \prod_{j=t+1}^{i} p\left(x_j \mid x_{j-1}, \pi_{j-1}\left(H_{j-1}\right)\right)$$

$$\cdot \left| \prod_{j=t+1}^{i} p_Z\left(z_j \mid x_j\right) - \prod_{j=t+1}^{i} q_Z\left(z_j \mid x_j\right) \right| \mathrm{d}x_{t:i}\, \mathrm{d}z_{t+1:i} \tag{56}$$

$$= \mathbb{E}^{p_T, q_Z}_{t+1:i^-}\left[\Delta_Z(x_i)\right] = \mathbb{E}_{b_t}\left[\mathbb{E}^{p_T}_{t+1}\left[\Delta_Z(x_{t+1})\right]\right] \tag{57}$$

Hence we can see that the sum of the recursion turns out to be

$$\left|\mathbb{E}^{p_Z}_{t+1:i}\left[r_i(b_i^{p_Z}, \pi_i)\right] - \mathbb{E}^{q_Z}_{t+1:i}\left[r_i(b_i^{q_Z}, \pi_i)\right]\right| \tag{58}$$

$$\leq R_i^{\max} \sum_{l=t+1}^{i} \mathbb{E}^{p_T, q_Z}_{t+1:l^-}\left[\Delta_Z(x_l)\right] \tag{59}$$

□

Corollary 1

Corollary 1. *The difference between the original and simplified value functions can be bounded by the following sum of scaled expected TV-distance terms:*

$$\left|V_t^{\pi, p_Z}(b_t) - V_t^{\pi, q_Z}(b_t)\right|$$

$$\leq \textstyle\sum_{i=t+1}^{L} \mathbb{E}^{p_T, q_Z}_{t+1:i^-}\left[\Delta_Z(x_i)\right] \sum_{l=i}^{L} \gamma^{l-t} R_l^{\max} \tag{60}$$

$$= \textstyle\sum_{i=t+1}^{L} V_i^{\max} \cdot \mathbb{E}^{p_T, q_Z}_{t+1:i^-}\left[\Delta_Z(x_i)\right]. \tag{61}$$

Proof. We substitute for the definition of V_t^{π}, and we can see that the immediate reward at time i cancels out.

$$\left|V_t^{\pi, p_Z}(b_t) - V_t^{\pi, q_Z}(b_t)\right| \tag{62}$$

$$= \Bigg|\mathbb{E}^{p_Z}_{t+1:L}\Big[\sum_{i=t}^{L} \gamma^{i-t} r_i(b_i, \pi_i)\Big]$$

$$- \mathbb{E}^{q_Z}_{t+1:L}\Big[\sum_{i=t}^{L} \gamma^{i-t} r_i(b_i, \pi_i)\Big]\Bigg| \tag{63}$$

$$= \Bigg|\cancel{r_t(b_t, \pi_t)} + \mathbb{E}^{p_Z}_{t+1:L}\Big[\sum_{i=t+1}^{L} \gamma^{i-t} r_i(b_i, \pi_i)\Big]$$

$$- \cancel{r_t(b_t, \pi_t)} - \mathbb{E}^{q_Z}_{t+1:L}\Big[\sum_{i=t+1}^{L} \gamma^{i-t} r_i(b_i, \pi_i)\Big]\Bigg| \tag{64}$$

We apply Lemma 1 to substitute expectation over belief rewards to expectation over state rewards, and we rearrange terms so that we get summation over difference terms of expected state rewards.

$$= \Bigg|\sum_{i=t+1}^{L} \gamma^{i-t} \mathbb{E}^{p_T, p_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right]$$

$$- \sum_{i=t+1}^{L} \gamma^{i-t} \mathbb{E}^{p_T, q_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right]\Bigg| \tag{65}$$

$$= \left|\sum_{i=t+1}^{L} \gamma^{i-t} \left(\mathbb{E}^{p_T, p_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right] - \mathbb{E}^{p_T, q_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right]\right)\right| \tag{66}$$

We now use the triangle inequality to separate into $L - t$ summands on which we can apply Theorem 1:

$$\leq \sum_{i=t+1}^{L} \gamma^{i-t} \left|\mathbb{E}^{p_T, p_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right] - \mathbb{E}^{p_T, q_Z}_{t+1:i}\left[r_i(x_i, \pi_i)\right]\right| \tag{67}$$

$$\overset{\textit{Thm. 1}}{\leq} \sum_{i=t+1}^{L} \gamma^{i-t} R_i^{\max} \sum_{l=t+1}^{i} \mathbb{E}^{p_T, q_Z}_{t+1:l^-}\left[\Delta_Z(x_l)\right] \tag{68}$$

We switch the order of summation such that for every expected-Δ_Z term we sum all of its $R^{\max}$ terms:

$$\begin{aligned}
&= \gamma R_{t+1}^{\max}\Big(\mathbb{E}^{p_T, q_Z}_{t+1:t+1^-}\left[\Delta_Z(x_{t+1})\right]\Big) \\
&+ \gamma^2 R_{t+2}^{\max}\Big(\mathbb{E}^{p_T, q_Z}_{t+1:t+1^-}\left[\Delta_Z(x_{t+1})\right] \\
&+ \mathbb{E}^{p_T, q_Z}_{t+1:t+2^-}\left[\Delta_Z(x_{t+2})\right]\Big) \\
&+ \dots \\
&+ \gamma^{L-t} R_L^{\max}\Big(\mathbb{E}^{p_T, q_Z}_{t+1:t+1^-}\left[\Delta_Z(x_{t+1})\right] \\
&+ \dots + \mathbb{E}^{p_T, q_Z}_{t+1:t+L^-}\left[\Delta_Z(x_{t+L})\right]\Big)
\end{aligned} \tag{69}$$

$$
\begin{aligned}
&= \mathbb{E}^{p_T,q_Z}_{t+1:t+1^-}[\Delta_Z(x_{t+1})]\left(\gamma R^{\max}_{t+1} + \cdots + \gamma^{L-t} R^{\max}_L\right) \\
&+ \mathbb{E}^{p_T,q_Z}_{t+1:t+2^-}[\Delta_Z(x_{t+2})]\left(\gamma^2 R^{\max}_{t+2} + \cdots + \gamma^{L-t} R^{\max}_L\right) \\
&+ \dots \\
&+ \mathbb{E}^{p_T,q_Z}_{t+1:t+L^-}[\Delta_Z(x_{t+L})]\gamma^{L-t} R^{\max}_L
\end{aligned} \tag{70}
$$

$$
= \sum_{i=t+1}^{L} \mathbb{E}^{p_T,q_Z}_{t+1:i^-}[\Delta_Z(x_i)] \sum_{l=i}^{L} \gamma^{l-t} R^{\max}_l \tag{71}
$$

By identifying the sum of discounted maximum rewards as the maximum value, we arrive at the required:

$$
= \sum_{i=t+1}^{L} V_i^{\max} \cdot \mathbb{E}^{p_T,q_Z}_{t+1:i^-}[\Delta_Z(x_i)]. \tag{72}
$$

□

Theorem 2

Theorem 2. *Under the conditions of Theorem 1, the difference between the original and simplified value function can be bounded by*

$$
|V_t^{\pi,p_Z}(b_t) - V_t^{\pi,q_Z}(b_t)| \leq M_t^{\pi}(b_t). \tag{73}
$$

In addition, the respective difference in action value function can be bounded by the action cumulative bound,

$$
|Q_t^{\pi,p_Z}(b_t,a) - Q_t^{\pi,q_Z}(b_t,a)| \leq \Phi_t^{\pi}(b_t,a). \tag{74}
$$

Proof. First we prove for (73):

$$
\begin{aligned}
M_t^{\pi}(b_t) &= \mathbb{E}^{q_Z}_{t+1:L-1}[\textstyle\sum_{i=t}^{L-1} m_i(b_i,\pi_i)] && (75)\\
&= \textstyle\sum_{i=t+1}^{L} \mathbb{E}^{q_Z}_{t+1:i-1}[m_{i-1}(b_{i-1},\pi_{i-1})] && (76)\\
&\overset{\textit{Lem. 1}}{=} \textstyle\sum_{i=t+1}^{L} \mathbb{E}^{p_T,q_Z}_{t+1:i-1}[m_{i-1}(x_{i-1},\pi_{i-1})] && (77)\\
&= \textstyle\sum_{i=t+1}^{L} V_i^{\max} \cdot \mathbb{E}^{p_T,q_Z}_{t+1:i-1}[\mathbb{E}_i^{p_T}[\Delta_Z(x_i)]] && (78)\\
&= \textstyle\sum_{i=t+1}^{L} V_i^{\max} \cdot \mathbb{E}^{p_T,q_Z}_{t+1:i^-}[\Delta_Z(x_i)] && (79)
\end{aligned}
$$

And the proof is clear by seeing that this is the exact term that bounds (61). For proving (74), note that for every action a we can define a policy π^a by performing a and then continuing with π. For every policy π it holds that $Q_t^{\pi}(b_t,\pi_t) = V_t^{\pi}(b_t)$, and specifically for π^a. Therefore:

$$
\begin{aligned}
&|Q_t^{\pi,p_Z}(b_t,a) - Q_t^{\pi,q_Z}(b_t,a)| = && (80)\\
&\left|V_t^{\pi^a,p_Z}(b_t) - V_t^{\pi^a,q_Z}(b_t)\right| && (81)\\
&\leq M_t^{\pi^a}(b_t) = \Phi_t^{\pi}(b_t,a). && (82)
\end{aligned}
$$

□

Theorem 3

The proof of Theorem 3 is an adaptation of Lemma 2 from (Lim et al. 2022), with relatively small modifications.

We first rely on the Particle Likelihood SN Estimator Convergence lemma from (Lim et al. 2022) (originally Lemma 1):

Algorithm 1: Sparse Sampling-ω-π

Global Variables: γ, C, L, π
Algorithm: Estimate$V^{\pi}(\bar{b},t)$.
Input: Particle belief set $\bar{b} = \{(x_i,w_i)\}$, depth t, policy π.
Output: A scalar $\hat{V}^{\pi}_{\omega,t}\left(\bar{b}_t\right)$ that is an estimate of $V_t^{\pi}\left(\bar{b}_t\right)$.

```
1: if $t \geq L$ then
2:     return 0
3: end if
4: $\hat{Q}_t^{\pi}(\bar{b},\pi(\bar{b})) \leftarrow$ Estimate$Q^{\pi}(\bar{b},\pi(\bar{b}),t)$
5: for all $a \in \mathcal{A}$ do
6:     $\hat{Q}_t^{\pi}(\bar{b},a) \leftarrow$ Estimate$Q^{\pi}(\bar{b},a,t)$
7: end for
8: return $\hat{V}_t^{\pi}(\bar{b}) \leftarrow \hat{Q}_t^{\pi}(\bar{t},\pi(\bar{b}))$
```

Algorithm: Estimate$Q^{\pi}\left(\bar{b},a,t\right)$.
Input: Particle belief set $\bar{b} = \{(x_i,w_i)\}$, action a, depth t, policy π.
Output: A scalar $\hat{Q}^{\pi}_{\omega,t}\left(\bar{b},a\right)$ that is an estimate of $Q_t^{\pi}(b,a)$.

```
1: for all $i = 1,\dots,C$ do
2:     $\bar{b}'_i, \rho \leftarrow$ GenPF$(\bar{b},a)$
3:     $\hat{V}^{\pi}_{t+1}(\bar{b}'_i) \leftarrow$ Estimate$V^{\pi}(\bar{b}'_i,t+1)$
4: end for
5: return $\hat{Q}_t^{\pi}(\bar{b},a) \leftarrow \rho + \frac{1}{C}\sum_{i=1}^{C}\gamma \cdot \hat{V}^{\pi}_{t+1}(\bar{b}'_i)$
```

Lemma 2 (Particle Likelihood SN Estimator Convergence)**.** *Suppose a function f is bounded by a finite constant $\|f\|_{\infty} \leq f_{\max}$, and a particle belief state $\bar{b}_t = \{x_t^i, w_t^i\}_{i=1}^{C}$ at depth t that represents $\bar{b}_t$ with particle likelihood weighting that is recursively updated as $w_t^i = w_{t-1}^i \cdot p_Z(z \mid x_{t+1})$ for an observation sequence $\{z_n^i\}_{n=1}^{t}$. Then, for all $t = 0,\dots,L$, the following weighted average is the SN estimator of f under the belief b_t corresponding to observation sequence $\{z_n\}_{n=1}^{t} : \tilde{\mu}_{\bar{b}_t}[f] = \frac{\sum_{i=1}^{C} w_t^i f(x_t^i)}{\sum_{i=1}^{C} w_t^i}$ and the following concentration bound holds with probability at least $1 - 3\exp\left(-C \cdot k^2_{\max}(\lambda,C)\right)$*

$$
\left|\mathbb{E}_{s\sim b_t}[f(x)] - \tilde{\mu}_{\bar{b}_t}[f]\right| \leq \lambda \tag{83}
$$

$$
k_{\max}(\lambda,C) \triangleq \frac{\lambda}{f_{\max} d_{\infty}^{\max}} - \frac{1}{\sqrt{C}} \tag{84}
$$

$$
d_{\infty}\left(\mathcal{P}^t || \mathcal{Q}^t\right) = \operatorname*{ess\,sup}_{x\sim\mathcal{Q}^t} w_{\mathcal{P}^t/\mathcal{Q}^t}(x) \leq d_{\infty}^{\max} \tag{85}
$$

where $\mathcal{P}^t$ is the target distribution and $\mathcal{Q}^t$ is the distribution of the particle filter.

For the next lemma, we define a new theoretical algorithm Sparse Sampling-ω-π, shown in Algorithm 1. The algorithm estimates the value function of a policy π according to Sparse Sampling-ω (Lim et al. 2023).

The algorithm has two variations, indicated with the blue line in 4, or with the red lines in 5-6. The blue variation is for estimating the value of a single policy π, whereas the red variation expands the entire action space (if it is finite), to estimate the values of each possible policy via the Q-value.

Lemma 3 (Sparse Sampling-ω-π Q-Value Coupled Convergence)**.** *For a given policy π, for all $t = 0, \ldots, L$ and actions a, the following bounds hold with probability at least $1 - 5\left(4C\right)^{L+1}\left(\exp\left(-C \cdot \acute{k}^2\right) + \delta_r\left(\nu, N_r\right)\right)$:*

$$\left|Q^{\pi}_{\mathbf{P},t}\left(b_t, a\right) - \hat{\hat{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t, a\right)\right| \leq \alpha_t, \tag{86}$$

$$\alpha_t = \lambda + \nu + \gamma\alpha_{t+1},\ \alpha_L = \lambda + \nu, \tag{87}$$

$$\left|Q^{\pi}_{\mathbf{M_P},t}\left(\bar{b}_t, a\right) - \hat{\hat{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t, a\right)\right| \leq \beta_t, \tag{88}$$

$$\beta_t = \nu + \gamma\left(\lambda + \beta_{t+1}\right), \beta_L = \nu, \tag{89}$$

$$k_{\max}\left(\lambda, C\right) = \frac{\lambda}{4V_{\max}d^{\max}_{\infty}} - \frac{1}{\sqrt{C}}, \tag{90}$$

$$\acute{k} = \min\left\{k_{\max}, \lambda/4\sqrt{2}V_{\max}\right\} \tag{91}$$

Under the assumption that the immediate reward estimate is probabilistically bounded such that $\mathbb{P}(\left|r^i_t - \tilde{r}^i_t\right| \geq \nu) \leq \delta_r(\nu, N_r)$, for a number of samples parameter N_r.

If we require the bound to hold for all possible policies that can be extracted from a given belief tree simultaneously, then the probability becomes at least $1 - 5\left(4\left|A\right|C\right)^{L+1}\left(\exp\left(-C \cdot \acute{k}^2\right) + \delta_r\left(\nu, N_r\right)\right)$.

Proof. We prove for observation model p_Z without loss of generality.

POMDP Value Convergence We split the difference between the SN estimator and $Q^{\pi}_{\mathbf{P},t}$ into two terms, the reward estimation error (A) and the next-step value estimation error (B):

$$\left|Q^{\pi}_{\mathbf{P},t}\left(b_t, a\right) - \hat{\hat{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t, a\right)\right| \tag{92}$$

$$\leq \underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[R\left(x_t, a\right) \mid b_t\right] - \frac{\sum_{i=1}^{C} w^i_t \tilde{r}^i_t}{\sum_{i=1}^{C} w^i_t}\right|}_{(A)} \tag{93}$$

$$+\gamma\underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[V^{\pi}_{\mathbf{P},t+1}\left(b_t a z\right) \mid b_t\right] - \frac{1}{C}\sum_{i=1}^{C}\hat{\hat{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right|}_{\text{(B)}} \tag{94}$$

Where I_i is a RV sampled from the probability mass $p_{w,t}\left(I = i\right) = \left(w^i_t / \sum_j w^j_t\right)$, and the particle belief $\bar{b}'^{[I_i]}_{t+1}$ is updated with an observation generated from $x^{I_i}_t$.

To prove the base case $t = L$, note that only term (A) is needed to be bounded, since $t = L$ corresponds to the leaf node of Sparse Sampling-ω-π and no further next step value estimation is performed.

We split term (A) into two terms:

$$\underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[R\left(x_t, a\right) \mid b_t\right] - \frac{\sum_{i=1}^{C} w^i_t \tilde{r}^i_t}{\sum_{i=1}^{C} w^i_t}\right|}_{(A)} \tag{95}$$

$$\leq \underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[R\left(x_t, a\right) \mid b_t\right] - \frac{\sum_{i=1}^{C} w^i_t r^i_t}{\sum_{i=1}^{C} w^i_t}\right|}_{(1)\text{Importance sampling error}} \tag{96}$$

$$+ \underbrace{\left|\frac{\sum_{i=1}^{C} w^i_t r^i_t}{\sum_{i=1}^{C} w^i_t} - \frac{\sum_{i=1}^{C} w^i_t \tilde{r}^i_t}{\sum_{i=1}^{C} w^i_t}\right|}_{(2)\text{Reward approximation error}} \tag{97}$$

Term (1) is a particle likelihood weighted average of $R\left(\cdot, a\right)$, and we will use the SN concentration bounds from Lemma 2. We bound R with $R_{\max}$ and augment λ to $\frac{R_{\max}}{4V_{\max}}\lambda$, in order to obtain the same uniform $t_{\max}$ factor with the following terms. This also covers the base case since $\frac{R_{\max}}{4V_{\max}}\lambda \leq \lambda = \alpha_L$. Hence the bound will hold with probability at least $1 - 3\exp\left(-C \cdot t^2_{\max}\left(\lambda, C\right)\right)$.

For term (2), we use the triangle inequality to bound the apply the assumption of a probabilistic bound on the state reward to a probabilistic bound on the belief reward. Next we use the monotonicity of the weighted mean, in the context that if all terms have an upper bound ν, then follows that the average itself is upper bounded by ν. Finally we use the inverse of the union bound (Boole's inequality) to lower bound with assumed bound for each individual reward term.

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{C} w^i_t r^i_t}{\sum_{i=1}^{C} w^i_t} - \frac{\sum_{i=1}^{C} w^i_t \tilde{r}^i_t}{\sum_{i=1}^{C} w^i_t}\right| \leq \nu\right) \tag{98}$$

$$\overset{\text{Triangle inequality}}{\geq} \tag{99}$$

$$\mathbb{P}\left(\left(\sum_{i=1}^{C} w^i_t\right)^{-1}\sum_{i=1}^{C} w^i_t\left|\left(r^i_t - \tilde{r}^i_t\right)\right| \leq \nu\right) \tag{100}$$

$$\overset{\text{Weighted mean monotonicity}}{\geq} \mathbb{P}\left(\bigcap_{i=1}^{C}\left|\left(r^i_t - \tilde{r}^i_t\right)\right| \leq \nu\right) \tag{101}$$

$$\overset{\text{Union bound}}{\geq} 1 - \sum_{i=1}^{C}\mathbb{P}\left(\left|\left(r^i_t - \tilde{r}^i_t\right)\right| \geq \nu\right) \tag{102}$$

$$\overset{\text{Assumption}}{\geq} 1 - \sum_{i=1}^{C}\delta_r\left(\nu, N_r\right) = 1 - C\delta_r\left(\nu, N_r\right) \tag{103}$$

Term (B) is repeatedly separated into four terms:

$$\underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[V^{\pi}_{t+1}\left(b_t a z\right) \mid b_t\right] - \frac{1}{C}\sum_{i=1}^{C}\hat{\hat{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right|}_{(B)} \tag{104}$$

$$\leq \underbrace{\left|\mathbb{E}_{\mathbf{P}}\left[V^{\pi}_{t+1}\left(b_t a z\right) \mid b_t\right] - \frac{\sum_{i=1}^{C} w^i_t \boldsymbol{V}^{\pi}_{t+1}\left(b_t, a\right)^{[i]}}{\sum_{i=1}^{C} w^i_t}\right|}_{(1)\text{ Importance sampling error}} \tag{105}$$

$$+\underbrace{\left|\frac{\sum_{i=1}^{C} w_t^i \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]}}{\sum_{i=1}^{C} w_t^i}-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{\left[I_i\right]}\right|}_{\text{(2) MC weighted sum approximation error}} \tag{106}$$

$$+\underbrace{\left|\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{\left[I_i\right]}-\frac{1}{C} \sum_{i=1}^{C} V_{t+1}^{\pi}\left(b_t a z^{\left[I_i\right]}\right)\right|}_{\text{(3) MC next-step integral approximation error}} \tag{107}$$

$$+\underbrace{\left|\frac{1}{C} \sum_{i=1}^{C} V_{t+1}^{\pi}\left(b_t a z^{\left[I_i\right]}\right)-\frac{1}{C} \sum_{i=1}^{C} \hat{\hat{V}}_{\omega, t+1}^{\pi}\left(\bar{b}_{t+1}^{\prime\left[I_i\right]}\right)\right|}_{\text{(4)Function estimation error}} \tag{108}$$

$$\leq \frac{1}{4} \lambda+\frac{1}{4} \lambda+\frac{1}{2} \lambda+\alpha_{t+1} \tag{109}$$

Note the notations used:

$$Z_{t+1} \triangleq p_Z\left(z \mid x_{t+1}\right) \tag{110}$$

$$T_{t, t+1} \triangleq p_T\left(x_{t+1} \mid x_t, a\right) \tag{111}$$

$$T_{t, t+1}^{[i]} \triangleq p_T\left(x_{t+1} \mid x_t^i, a\right) \tag{112}$$

For the rest of this proof, we define $V_t^\pi\left(b_t\right)$ to be the value function attained from time t given initial belief b_t and by following policy π. Additionally, we define the following:

$$\boldsymbol{V}_{t+1}^{\pi}\left(b_t, a, x_t\right) \triangleq \tag{113}$$

$$\int_{\mathcal{X}} \int_{\mathcal{Z}} V_{t+1}^{\pi}\left(b_t a z\right)\left(Z_{t+1}\right)\left(T_{t, t+1}\right) \mathrm{d} x_{t+1} \mathrm{~d} z \tag{114}$$

$$\boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]} \triangleq \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a, x_{t, i}\right)= \tag{115}$$

$$\int_{\mathcal{X}} \int_{\mathcal{Z}} V_{t+1}^{\pi}\left(b_t a z\right) Z_{t+1} T_{t, t+1}^{[i]} \mathrm{d} x_{t+1} \mathrm{~d} z \tag{116}$$

$$\mathbb{E}_{\mathbf{P}}\left[V_{t+1}^{\pi}\left(b_t a z\right) \mid b_t\right] \triangleq \tag{117}$$

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathcal{Z}} V_{t+1}^{\pi}\left(b_t a z\right)\left(Z_{t+1}\right)\left(T_{t, t+1}\right) p\left(x_t \mid b_t\right) \mathrm{d} x_{t: t+1} \mathrm{~d} z \tag{118}$$

$$=\int_{\mathcal{X}} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right) p\left(x_t \mid b_t\right) \mathrm{d} x_t \tag{119}$$

$$=\frac{\int_{\mathcal{X}^{t+1}} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)\left(Z_{1: t}\right)\left(T_{1: t}\right) p\left(x_0 \mid b_0\right) \mathrm{d} x_{0: t}}{\int_{\mathcal{X}^{t+1}}\left(Z_{1: t}\right)\left(T_{1: t}\right) p\left(x_0 \mid b_0\right) \mathrm{d} x_{0: t}} \tag{120}$$

(1) Importance sampling error This term is the difference between the conditional expectation $\mathbb{E}_{\mathbf{P}}\left[V_{t+1}^{\pi}\left(b_t a z\right) \mid b_t\right]$ and its SN estimator. We have $\left\|\boldsymbol{V}_{t+1}^{\pi}\right\|_{\infty} \leq V_{\max}$, therefore we can apply the SN concentration inequality from Lemma 2 to bound it by the augmented $\lambda / 4$:

$$\mathbb{P}\left(\left|\mathbb{E}_{\mathbf{P}}\left[V_{T+1}^{\pi}\left(b_t a z\right) \mid b_t\right]-\frac{\sum_{i=1}^{C} w_t^i \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]}}{\sum_{i=1}^{C} w_t^i}\right| \leq \frac{\lambda}{4}\right) \tag{121}$$

$$\geq 1-3 \exp \left(-C \cdot k_{\max }^2(\lambda, C)\right) \tag{122}$$

(2) Monte Carlo weighted sum approximation error First, we assume that all variables $\left\{s_t^i, w_t^i\right\}, b_t, a$ are given, and only I is random. Note that $\left\|\boldsymbol{V}_{t+1}^{\pi}\left(b_t, a, \cdot\right)\right\|_{\infty} \leq V_{\max}$.

We will define the discrete probability mass defined by the weights at depth t: $p_{w, t}(I=i) \triangleq \left(w_t^i / \sum_j w_t^j\right)$, and for convenience denote $\boldsymbol{V}(i) \triangleq \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]}$. The term $\frac{\sum_{i=1}^{C} w_t^i \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]}}{\sum_{i=1}^{C} w_t^i}$ is equivalent to the expectation of $\boldsymbol{V}(I)$ w.r.t. $p_{w, t}(I=i)$. The term $\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{\left[I_i\right]}$ is equivalent to a Monte Carlo average of the previous quantity with C samples. Therefore:

$$\left|\frac{\sum_{i=1}^{C} w_t^i \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{[i]}}{\sum_{i=1}^{C} w_t^i}-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}_{t+1}^{\pi}\left(b_t, a\right)^{\left[I_i\right]}\right| \tag{123}$$

$$\Rightarrow\left|\mathbb{E}_{p_{w, t}(I=i)}[\boldsymbol{V}(I)]-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}\left(I_i\right)\right| \tag{124}$$

This is the form of the double-sided Hoeffding-type bound on the function values $\boldsymbol{V}(I)$. Hence, we can choose λ such that for an arbitrary fixed set $\left\{s_t^i, w_t^i\right\}, b_t, a$:

$$\mathbb{P}\left(\left|\mathbb{E}_{p_{w, t}(I=i)}[\boldsymbol{V}(I)]-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}\left(I_i\right)\right| \leq \lambda \mid\left\{s_t^i, w_t^i\right\}, b_t, a\right) \tag{125}$$

$$\geq 1-2 \exp \left(-C \lambda^2 / 2 V_{\max }^2\right) \tag{126}$$

We will use the two following well-known facts:

1. The probability of an event A is equal to the expectation of the indicator $\mathbf{1}_A$, i.e. $\mathbb{P}(A)=\mathbb{E}\left(\mathbf{1}_A\right)$.
2. The tower property, also known as the law of total expectation: for any two random variables X, Y defined on the same probability space, holds that $\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X \mid Y]]$.

Therefore

$$\mathbb{P}\left(\left|\mathbb{E}_{p_{w, d}(I=i)}[\boldsymbol{V}(I)]-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}\left(I_i\right)\right| \leq \lambda\right) \tag{127}$$

$$\overset{\text{Expt. of indicator}}{=} \mathbb{E}\left[\mathbf{1}_{\mathbb{P}\left(\left|\mathbb{E}_{p_{w, d}(I=i)}[\boldsymbol{V}(I)]-\frac{1}{C} \sum_{i=1}^{C} \boldsymbol{V}\left(I_i\right)\right| \leq \lambda\right)}\right] \tag{128}$$

$$\overset{\text{Tower property}}{=} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\mathbb{P}\left(\left|\mathbb{E}_{p_{w,d}(I=i)}[\boldsymbol{V}(I)]-\frac{1}{C}\sum_{i=1}^{C}\boldsymbol{V}(I_i)\right|\leq\lambda\right)} \mid \{s_{d,i}, w_{d,i}\}, b_d, a\right]\right] \tag{129}$$

$$\overset{\text{Expt. of indicator}}{=} \mathbb{E}\left[\mathbb{P}\left(\left|\mathbb{E}_{p_{w,d}(I=i)}\left[\boldsymbol{V}(I)\right]-\frac{1}{C}\sum_{i=1}^{C}\boldsymbol{V}(I_i)\right| \leq \lambda \mid \{s_{d,i}, w_{d,i}\}, b_d, a\right)\right] \tag{130}$$

$$\geq \mathbb{E}\left[1-2\exp\left(-C\lambda^2/2V_{\max}^2\right)\right] \tag{131}$$

$$= 1-2\exp\left(-C\lambda^2/2V_{\max}^2\right) \tag{132}$$

We choose to bound term (2) with augmented $\lambda/4$, and this holds with probability at least $1-2\exp\left(-C\lambda^2/32V_{\max}^2\right)$:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{C} w_t^i \boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[i]}}{\sum_{i=1}^{C} w_t^i} - \frac{1}{C}\sum_{i=1}^{C}\boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[I_i]}\right| \leq \frac{\lambda}{4}\right) \tag{133}$$

$$\geq 1-2\exp\left(-C\lambda^2/32V_{\max}^2\right) \tag{134}$$

(3) Monte Carlo next-step integral approximation error First we define $\Delta_{t+1}(b_t,a)^{[I_i]} \triangleq \boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[I_i]} - V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right)$. Note that by rearranging the summation we can write

$$\left|\frac{1}{C}\sum_{i=1}^{C}\boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[I_i]} - \frac{1}{C}\sum_{i=1}^{C}V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right)\right| \tag{135}$$

$$= \left|\frac{1}{C}\sum_{i=1}^{C}\Delta_{t+1}(b_t,a)^{[I_i]}\right| \tag{136}$$

Note that $V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right)$ is simply a single sample Monte Carlo approximation of $\boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[I_i]}$, as the random vector (x_{t+1,I_i}, z_{I_i}) is jointly generated using the generative model according to the correct probability $Z_{t+1}\cdot T_{t,t+1}$ given x_t^i. This can be seen by the following:

$$\mathbb{E}_{\left(s_{d+1,I_i}, o_{I_i}\right)}\left[V_{d+1}^{\pi}\left(b_d a o^{[I_i]}\right)\right] \tag{137}$$

$$= \int_{\mathcal{X}}\int_{\mathcal{O}} V_{d+1}^{\pi}(b_d a o)\, Z_{d+1} T_{d,d+1}^{[i]}\, \mathrm{d}s_{d+1}\, \mathrm{d}z \tag{138}$$

$$= \boldsymbol{V}_{d+1}^{\pi}(b_d,a)^{[I_i]} \tag{139}$$

Hence follows from the tower property that $\mathbb{E}\left[\Delta_{t+1}\right]=0$. Define $\Delta_{t+1}(b_t,a)^{[I_i]} \triangleq \boldsymbol{V}_{t+1}^{\pi}(b_t,a)^{[I_i]} - V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right)$. Note from the triangle inequality that $\left\|\Delta_{t+1}\right\|_{\infty} \leq 2V_{\max}$. Since I_i are i.i.d. follows that $\Delta_{t+1}(b_t,a)^{[I_i]}$ are i.i.d. too, and we can directly use another Hoeffding's bound:

$$\mathbb{P}\left(\left|\frac{1}{C}\sum_{i=1}^{C}\Delta_{t+1}(b_t,a)^{[I_i]}\right| \leq \frac{\lambda}{2}\right) \tag{140}$$

$$= \mathbb{P}\left(\left|\frac{1}{C}\sum_{i=1}^{C}\Delta_{t+1}(b_t,a)^{[I_i]} - \mathbb{E}\left[\Delta_{t+1}\right]\right| \leq \frac{\lambda}{2}\right) \tag{141}$$

$$\geq 1-2\exp\left(-C\lambda^2/32V_{\max}^2\right) \tag{142}$$

(4) Function estimation error From the inductive hypothesis, for each possible $az^{[I_i]}$, we have the bound $\left|Q_{\mathbf{P},t}^{\pi}\left(b_t a z^{[I_i]}, a'\right) - \hat{\hat{Q}}_{\omega,t}^{\pi}\left(\bar{b}_t a z^{[I_i]}, a'\right)\right| \leq \alpha_{t+1}$ holding with high probability for all actions a', hence in particular for $a' = \pi\left(b_t a z^{[I_i]}\right)$. Then follows:

$$\left|\frac{1}{C}\sum_{i=1}^{C}V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right) - \frac{1}{C}\sum_{i=1}^{C}\hat{\hat{V}}_{\omega,t+1}^{\pi}\left(\bar{b}_{t+1}'^{[I_i]}\right)\right| \tag{143}$$

$$\leq \frac{1}{C}\sum_{i=1}^{C}\left|V_{t+1}^{\pi}\left(b_t a z^{[I_i]}\right) - \hat{\hat{V}}_{\omega,t+1}^{\pi}\left(\bar{b}_{t+1}'^{[I_i]}\right)\right| \tag{144}$$

$$\leq \frac{1}{C}\sum_{i=1}^{C}\alpha_{t+1} = \alpha_{t+1} \tag{145}$$

Combining the bounds Thus, each of the terms are bound by $(A) \leq \frac{R_{\max}}{4V_{\max}}\lambda$ and $(B) \leq \frac{1}{4}\lambda + \frac{1}{4}\lambda + \frac{1}{2}\lambda + \alpha_{t+1}$, which uses the SN concentration bounds twice and Hoeffding's bound twice. Combining together:

$$\left|Q_{\mathbf{P},t}^{\pi}(b_t,a) - \hat{\hat{Q}}_{\omega,t}^{\pi}\left(\bar{b}_t,a\right)\right| \tag{146}$$

$$\leq \frac{R_{\max}}{4V_{\max}}\lambda + \nu + \gamma\left[\frac{1}{4}\lambda + \frac{1}{4}\lambda + \frac{1}{2}\lambda + \alpha_{t+1}\right] \tag{147}$$

$$\leq \frac{1-\gamma}{4}\lambda + \nu + \gamma\lambda + \gamma\alpha_{t+1} \tag{148}$$

$$\leq \lambda + \nu + \gamma\alpha_{t+1} = \alpha_t \tag{149}$$

We obtain the worst case union bound on the probability that all inequalities simultaneously hold. We used the SN concentration bound twice, Hoeffding's bound twice, and assumed reward bound once. For the SN and Hoeffding's bounds, we can bound the worst case probability of either by the following

$$\max\left\{3\exp\left(-C\cdot k_{\max}^2(\lambda,C)\right), 2\exp\left(-C\lambda^2/32V_{\max}^2\right)\right\} \tag{150}$$

$$\leq 3\exp\left(-C\cdot \acute{k}^2\right), \tag{151}$$

which we multiply by the union factor bound $(4C)^{L+1}$ since we want the function estimates to be within the bounds for the specific action chosen by the policy, and all child nodes (used C times in the function estimation error), and we used either SN concentration bound or Hoeffding's bound 4 times in total. For the reward approximation bound, we multiply by a factor of C^{L+1} to account for the branching factor. Hence, in total, we have shown that for all levels t the worst case union bound probability of all bad events is bounded by

$$\mathbb{P}\left(\left|Q_{\mathbf{P},t}^{\pi}(b_t,a) - \hat{\hat{Q}}_{\omega,t}^{\pi}\left(\bar{b}_t,a\right)\right| \leq \alpha_t\right) \tag{152}$$

$$\geq 1-3(4C)^{L+1}\left(\exp\left(-C\cdot \acute{k}^2\right)\right) - C^{L+1}\alpha(\nu,N) \tag{153}$$

$$\geq 1 - 3\left(4C\right)^{L+1}\left(\exp\left(-C\cdot \acute{k}^{2}\right)+\alpha\left(\nu,N\right)\right) \tag{154}$$

When giving a union bound for all possible actions as well (i.e. for all policies) then the action branching factor becomes $|A|^{L+1}$, once from bounding the root level, and another L times for all depth levels. Thus, in this case the probability becomes at least $1 - 3\left(4\left|A\right|C\right)^{L+1}\left(\exp\left(-C\cdot \acute{k}^{2}\right)+\alpha\left(\nu,N\right)\right)$.

PB-MDP Value Convergence Similarly to the previous convergence bound, we split the difference into two terms, the reward estimation error (A) and the next-step value estimation error (B):

$$\left|Q^{\pi}_{\mathbf{M_P},t}\left(\bar{b}_t,a\right)-\hat{\tilde{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t,a\right)\right| \tag{155}$$

$$\leq \underbrace{\left|\rho\left(\bar{b}_t,a\right)-\tilde{\rho}\left(\bar{b}_t,a\right)\right|}_{\text{(A)}} \tag{156}$$

$$\underbrace{+\gamma\left|\mathbb{E}_{\mathbf{M_P}}\left[V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}_{t+1}\right)\mid \bar{b}_t,a\right] -\frac{1}{C}\sum_{i=1}^{C}\hat{\tilde{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right|}_{\text{(B)}}. \tag{157}$$

Term (A) can be bounded like the reward approximation error of term (A) in the previous case.

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{C}w^i_t r^i_t}{\sum_{i=1}^{C}w^i_t}-\frac{\sum_{i=1}^{C}w^i_t \tilde{r}^i_t}{\sum_{i=1}^{C}w^i_t}\right|\leq\nu\right)\geq 1-C\delta_r\left(\nu,N_r\right) \tag{158}$$

For the inductive step, we prove that the difference (B) is bounded for all $t=0,\ldots,L$. We split it into two terms:

$$\left|\mathbb{E}_{\mathbf{M_P}}\left[V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}_{t+1}\right)\mid \bar{b}_d,a\right]-\frac{1}{C}\sum_{i=1}^{C}\hat{\tilde{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right| \tag{159}$$

$$\leq \underbrace{\left|\mathbb{E}_{\mathbf{M_P}}\left[V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}_{t+1}\right)\mid \bar{b}_t,a\right]-\frac{1}{C}\sum_{i=1}^{C}V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right|}_{\text{(1)MC transition approximation error}} \tag{160}$$

$$+\underbrace{\left|\frac{1}{C}\sum_{i=1}^{C}V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)-\frac{1}{C}\sum_{i=1}^{C}\hat{\tilde{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right|}_{\text{(2) Function approximation error}} \tag{161}$$

$$\leq \underbrace{\lambda}_{(1)}+\underbrace{\beta_{t+1}}_{(2)}. \tag{162}$$

(1) MC transition approximation error This term is a Monte Carlo estimate of the integration over the transition estimate $\tau\left(\bar{b}_{t+1}\mid\bar{b}_t,a\right)$. The value function and its estimate are both bounded by $V_{\max}$, therefore we can invoke Hoeffding's bound to obtain the following probabilistic bound:

$$\begin{aligned}\mathbb{P}\Bigg(\Bigg|\mathbb{E}_{\mathbf{M_P}}&\left[V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}_{t+1}\right)\mid\bar{b}_t,a\right]-\\ &\frac{1}{C}\sum_{i=1}^{C}V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}^{[I_i]}_{t+1}\right)\Bigg|\leq\lambda\Bigg)\\ &\geq 1-2\exp\left(-C\lambda^2/2V^2_{\max}\right).\end{aligned} \tag{163}$$

(2) Function approximation error From the inductive hypothesis, for each $\bar{b}'^{[I_i]}_{t+1}$, its PB-MDP Q-value it's sparse sampling-ω-π estimate at step $t+1$ is bounded by β_{t+1} for all actions. In particular, this also applies for $a=\pi\left(\bar{b}'^{[I_i]}_{t+1}\right)$. Thus follows

$$\left|\frac{1}{C}\sum_{i=1}^{C}V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)-\frac{1}{C}\sum_{i=1}^{C}\hat{\tilde{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right| \tag{164}$$

$$\leq\frac{1}{C}\sum_{i=1}^{C}\left|V^{\pi}_{\mathbf{M_P},t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)-\hat{\tilde{V}}^{\pi}_{\omega,t+1}\left(\bar{b}'^{[I_i]}_{t+1}\right)\right| \tag{165}$$

$$\leq\frac{1}{C}\sum_{i=1}^{C}\beta_{d+1}=\beta_{t+1} \tag{166}$$

Combining the bounds By applying similar logic of ensuring that every particle belief node and satisfies the concentration inequalities, we combine one Hoeffding's inequality with one immediate reward approximation error. The terms are bounded by $(A)\leq\nu$ and $(B)\leq\lambda+\beta_{t+1}$, therefore:

$$\left|Q^{\pi}_{\mathbf{M_P},t}\left(\bar{b}_t,a\right)-\hat{\tilde{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t,a\right)\right| \tag{167}$$

$$\leq\nu+\gamma\left(\lambda+\beta_{t+1}\right)=\beta_t, \tag{168}$$

and with the union bound, we get the following probabilistic bound

$$\mathbb{P}\left(\left|Q^{\pi}_{\mathbf{M_P},t}\left(\bar{b}_t,a\right)-\hat{\tilde{Q}}^{\pi}_{\omega,t}\left(\bar{b}_t,a\right)\right|\leq\beta_t\right) \tag{169}$$

$$\geq 1-2\cdot C^{L+1}\left(\exp\left(-C\lambda^2/2V^2_{\max}\right)\right)-C^{L+1}\delta_r\left(\nu,N_r\right) \tag{170}$$

$$\geq 1-2\cdot C^{L+1}\left(\exp\left(-C\lambda^2/2V^2_{\max}\right)+\delta_r\left(\nu,N_r\right)\right). \tag{171}$$

For the case of bounding for all policies simultaneously, the probabilities become at least $1-2\cdot\left(\left|A\right|C\right)^{L+1}\left(\exp\left(-C\lambda^2/2V^2_{\max}\right)+\delta_r\left(\nu,N_r\right)\right)$

Combining Both Concentration Bounds In order to enable the two concentration inequalities to simultaneously hold, we bound the worst case union probability:

$$3\left(4C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^2\right)+\delta_r\left(\nu,N_r\right)\right) \tag{172}$$

$$+2\cdot C^{L+1}\left(\exp\left(-C\lambda^2/2V^2_{\max}\right)+\delta_r\left(\nu,N_r\right)\right) \tag{173}$$

$$\leq 3\left(4C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^2\right)+\delta_r\left(\nu,N_r\right)\right) \tag{174}$$

$$+2\left(4C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^{2}\right)+\delta_{r}\left(\nu,N_{r}\right)\right) \tag{175}$$
$$=5\left(4C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^{2}\right)+\delta_{r}\left(\nu,N_{r}\right)\right) \tag{176}$$

Therefore, we conclude that the Sparse Sampling-ω-π Q-value estimate concentration inequalities approximation error, for both the original POMDP and its PB-MDP approximation, are bounded by α_t, β_t at every belief node, respectively, with probability at least $1-5\left(4C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^{2}\right)+\delta_{r}\left(\nu,N_{r}\right)\right)$. If we require the concentration inequality to simultaneously hold for all policies, then the probability becomes $1-5\left(4\left|A\right|C\right)^{L+1}\left(\exp\left(-C\cdot\acute{k}^{2}\right)+\delta_{r}\left(\nu,N_{r}\right)\right)$.

□

Theorem 3 (Generalized PB-MDP Convergence)**.** *Assume that the immediate state reward estimate is probabilistically bounded such that $\mathbb{P}(|r_i^j-\tilde{r}_i^j|\geq\nu)\leq\delta_r(\nu,N_r)$, for a number of reward samples N_r and state sample x_i^j. Assume that $\delta_r(\nu,N_r)\to 0$ as $N_r\to\infty$. For all policies π, $t=0,\dots,L$ and $a\in\mathcal{A}$, the following bounds hold with probability of at least $1-5(4C)^{L+1}(\exp(-C\cdot\acute{k}^2)+\delta_r(\nu,N_r))$:*

$$|Q^{\pi,[pz/qz]}_{\mathbf{P},t}(b_t,a)-Q^{\pi,[pz/qz]}_{\mathbf{M_P},t}(\bar{b}_t,a)|\leq\alpha_t+\beta_t, \tag{177}$$

where,

$$\alpha_t=(1+\gamma)\lambda+\gamma\alpha_{t+1},\ \alpha_L=\lambda\geq 0, \tag{178}$$
$$\beta_t=2\nu+\gamma\beta_{t+1},\ \beta_L=2\nu\geq 0, \tag{179}$$
$$k_{\max}(\lambda,C)=\frac{\lambda}{4V_{\max}d_{\infty}^{\max}}-\frac{1}{\sqrt{C}}>0, \tag{180}$$
$$\acute{k}=\min\{k_{\max},\lambda/4\sqrt{2}V_{\max}\}. \tag{181}$$

If we require the bound to hold for all possible policies that can be extracted from a given belief tree simultaneously, then under the assumption of a finite action space, the probability is at least $1-5(4|\mathcal{A}|C)^{L+1}(\exp(-C\cdot\acute{k}^2)+\delta_r(\nu,N_r))$.

Proof. We prove for the observation model p_Z without loss of generality.

Under the same conditions and probability as Lemma 3, we bound the difference directly between the theoretical action value function and the particle-belief approximation using the triangle inequality,

$$\left|Q^{\pi}_{\mathbf{P},t}(b_t,a)-Q^{\pi}_{\mathbf{M_P},t}(\bar{b}_t,a)\right|\leq\alpha_t+\beta_t. \tag{182}$$

We define the following recursive bounds

$$A_t\triangleq(1+\gamma)\lambda+\gamma A_{d+1},\ A_L\triangleq\lambda \tag{183}$$
$$B_t\triangleq 2\nu+\gamma B_{t+1},\ B_L\triangleq 2\nu \tag{184}$$

and follows that $A_t+B_t=\alpha_t+\beta_t$ and $A_L+B_L=\alpha_L+\beta_L$, hence follows that $\left|Q^{\pi}_{\mathbf{P},t}(b_t,a)-Q^{\pi}_{\mathbf{M_P},t}(\bar{b}_t,a)\right|\leq A_t+B_t$. By renaming $\alpha_t\triangleq A_t$ and $\beta_t\triangleq B_t$ we get the required result. □

Corollary 2

Corollary 2. *For arbitrary precision ε and accuracy δ we can choose constants λ,ν,C,N_r such that the following holds with probability of at least $1-\delta$:*

$$|Q^{\pi,[pz/qz]}_{\mathbf{P},t}(\bar{b}_t,a)-Q^{\pi,[pz/qz]}_{\mathbf{M_P},t}(\bar{b}_t,a)|\leq\varepsilon. \tag{185}$$

Proof. We prove for the case of bounding for a belief tree with a finite action space. The case of a single policy can be proven similarly by removing the factors related $|\mathcal{A}|$.

Let $\varepsilon>0$ and let $\lambda>0$. We denote $L_{+1}\triangleq L+1$.

Let $\nu,\lambda=\frac{1}{4L_{+1}}\varepsilon$.

The conditions necessary for Theorem 3 are the following:

$$k_{\max}(\lambda,C)=\frac{\lambda}{4V_{\max}d_{\infty}^{\max}}-\frac{1}{\sqrt{C}}>0, \tag{186}$$
$$\delta\geq 5(4|\mathcal{A}|C)^{L_{+1}}(\exp(-C\cdot\acute{k}^2)+\delta_r(\nu,N_r)) \tag{187}$$
$$\acute{k}=\min\{k_{\max},\lambda/4\sqrt{2}V_{\max}\}. \tag{188}$$

Denote $A_1\triangleq\frac{\lambda}{4V_{\max}d_{\infty}^{\max}}$, $A_2\triangleq\lambda/4\sqrt{2}V_{\max}$, and note $A_1,A_2>0$. We obtain $k_{\max}(\lambda,C)=A_1-\frac{1}{\sqrt{C}}$.

We would like to choose a particle count C large enough such that $k_{\max}(\lambda,C)$ is larger than $A_1/2$. Hence we denote the solution for the following equation with $C_{A_1/2}$:

$$(A_1-\frac{1}{\sqrt{C_{A_1/2}}})^2\cdot C_{A_1/2}=\frac{A_1}{2} \tag{189}$$
$$\Rightarrow C_{A_1/2}\triangleq\frac{2\sqrt{2}\sqrt{\frac{1}{(A_1)^3}}(A_1)^2+A_1+2}{2(A_1)^2}. \tag{190}$$

Denote the following constants:

$$K_1\triangleq\max\{C_{A_1/2},\frac{1}{(A_1)^2},\frac{2}{(A_2)^2}\} \tag{191}$$
$$K_2\triangleq\min\{\frac{A_1}{2},2\}>0, \tag{192}$$
$$K_3\triangleq 5(4|\mathcal{A}|)^{L_{+1}} \tag{193}$$

We choose an auxiliary particle count $\tilde{C}\in\mathbb{N}$, such that the particle count C satisfies

$$C>K_1\cdot\tilde{C}\geq K_1. \tag{194}$$

Condition (186) is satisfied because

$$k_{\max}(\lambda,C)=A_1-\frac{1}{\sqrt{C}}>A_1-\frac{1}{\sqrt{\frac{1}{(A_1)^2}}}, \tag{195}$$
$$A_1-\sqrt{(A_1)^2}=0. \tag{196}$$

Additionally,

$$k_{\max}(\lambda,C)^2\cdot C>\frac{A_1}{2}, \tag{197}$$
$$(A_2)^2\cdot C>(A_2)^2\cdot\frac{2}{(A_2)^2}=2. \tag{198}$$

We obtain that

$$\acute{k}^2 \cdot C = \tag{199}$$

$$\min\{k_{\max}(\lambda, C)^2, (A_2)^2\} \cdot C > \tag{200}$$

$$\min\{\frac{A_1}{2}, 2\} \cdot \tilde{C} = K_2 \cdot \tilde{C}. \tag{201}$$

Therefore, condition (187) will be satisfied if the following is satisfied:

$$\delta \geq K_3(K_1)^{L+1} \cdot \tilde{C}^{L+1}(\exp(-K_2 \cdot \tilde{C}) + \delta_r(\nu, N_r)). \tag{202}$$

We perform the change of variables $X \triangleq K_2 \cdot \tilde{C}$, i.e. $\tilde{C} = \frac{X}{K_2}$:

$$K_3(K_1)^{L+1} \cdot \tilde{C}^{L+1} \exp(-K_2 \cdot \tilde{C}) \tag{203}$$

$$= K_3(\frac{K_1}{K_2})^{L+1} \cdot X^{L+1} \exp(-X). \tag{204}$$

The exponential function grows faster than any polynomial, and specifically for $P(X) = K_3(\frac{K_1}{K_2})^{L+1} \cdot X^{L+1}$. Therefore, we can choose $X' \in \mathbb{R}$ such that $\forall X \in \mathbb{N} > X'$:

$$K_3(\frac{K_1}{K_2})^{L+1} \cdot X^{L+1} \exp(-X) \leq \frac{\delta}{2}. \tag{205}$$

By choosing an auxiliary particle count $\tilde{C} > \frac{X'}{K_2}$ we satisfy (205).

For the choice of λ the following holds:

$$\alpha_0 \leq \sum_{k=0}^{L} 2\lambda \cdot \gamma^k \leq 2\lambda L_{+1} = \frac{2 \cdot \varepsilon(L_{+1})}{4L_{+1}} = \frac{\varepsilon}{2}. \tag{206}$$

For the given choice of the particle count C, we remind the assumption in Theorem 3 that for all $\nu > 0$ it holds that $\delta_r(\nu, N_r) \to 0$ as $N_r \to \infty$. Therefore, we can choose N_r' such that the following holds:

$$K_3 C^{L+1} \cdot \delta_r(\nu, N_r) \leq \frac{\delta}{2} \tag{207}$$

For the choice of ν the following holds:

$$\beta_0 \leq \sum_{k=0}^{L} 2\nu \cdot \gamma^k \leq 2\nu L_{+1} = \frac{2 \cdot \varepsilon(L_{+1})}{4L_{+1}} = \frac{\varepsilon}{2}. \tag{208}$$

In summary, for the choices of:

$$C > \max\{K_1, \frac{K_1}{K_2} X'\} \tag{209}$$

$$N_r > N_r' \tag{210}$$

we have that the following statements hold,

$$K_3 C^{L+1} \cdot \exp(-C \cdot \acute{k}^2) \leq \frac{\delta}{2} \tag{211}$$

$$K_3 C^{L+1} \cdot \delta_r(\nu, N_r) \leq \frac{\delta}{2}, \tag{212}$$

And therefore from Theorem 3, with probability of at least $1 - K_3 C^{L+1}(\exp(-C \cdot \acute{k}^2) + \delta_r(\nu, N_r)) \geq 1 - \delta$ the following bound holds:

$$\left|Q^{\pi}_{\mathbf{P},t}(b_t, a) - Q^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a)\right| \leq \alpha_t + \beta_t \tag{213}$$

$$\leq \alpha_0 + \beta_0 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \tag{214}$$

□

Corollary 3

Corollary 3. *Assuming that $\mathcal{P}$ is an MDP planner that can approximate Q-values with arbitrary precision $\varepsilon^{\mathcal{P}}$ at an accuracy $1 - \delta^{\mathcal{P}}$, we denote the precision and accuracy of the action value and action cumulative bound functions:*

$$\mathbb{P}(|Q^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a) - \hat{Q}^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \leq \varepsilon^{\mathcal{P}}_Q) \geq 1 - \delta^{\mathcal{P}}_Q \tag{215}$$

$$\mathbb{P}(|\Phi^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a) - \hat{\Phi}^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a)| \leq \varepsilon^{\mathcal{P}}_\Phi) \geq 1 - \delta^{\mathcal{P}}_\Phi \tag{216}$$

From Corollary 2 it holds that we can choose constants λ, ν, C, N_r such that the following holds,

$$\mathbb{P}(|Q^{\pi,qz}_{\mathbf{P},t}(b_t, a) - Q^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \leq \varepsilon_Q) \geq 1 - \delta_Q, \tag{217}$$

$$\mathbb{P}(|\Phi^{\pi}_{\mathbf{P},t}(b_t, a) - \tilde{\Phi}^{\pi}_{\mathbf{M_P},d}(\bar{b}_t, a)| \leq \varepsilon_\Phi) \geq 1 - \delta_\Phi. \tag{218}$$

Then with probability of at least $1 - (\delta_Q + \delta^{\mathcal{P}}_Q + \delta_\Phi + \delta^{\mathcal{P}}_\Phi)$

$$|Q^{\pi,pz}_{\mathbf{P},t}(b_t, a) - \hat{Q}^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \leq \tag{219}$$

$$\hat{\Phi}^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a) + \varepsilon_Q + \varepsilon^{\mathcal{P}}_Q + \varepsilon_\Phi + \varepsilon^{\mathcal{P}}_\Phi. \tag{220}$$

Proof. We combine all probabilistic bounds with the triangle inequality and the union bound, to conclude that with probability of at least $1 - (\delta_Q + \delta^{\mathcal{P}}_Q + \delta_\Phi + \delta^{\mathcal{P}}_\Phi)$:

$$|Q^{\pi,pz}_{\mathbf{P},t}(b_t, a) - \hat{Q}^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \tag{221}$$

$$\leq |Q^{\pi,pz}_{\mathbf{P},t}(b_t, a) - Q^{\pi,qz}_{\mathbf{P},t}(b_t, a)| \, (\textit{Thm. 2}) \tag{222}$$

$$+|Q^{\pi,qz}_{\mathbf{P},t}(b_t, a) - Q^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \, (\textit{eq.}\ (217)) \tag{223}$$

$$+|Q^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a) - \hat{Q}^{\pi,qz}_{\mathbf{M_P},t}(\bar{b}_t, a)| \, (\textit{eq.}\ (215)) \tag{224}$$

$$\leq \Phi^{\pi}_{\mathbf{P},t}(b_t, a) + \varepsilon_Q + \varepsilon^{\mathcal{P}}_Q \quad (\textit{eq.}\ (218) + \textit{eq.}\ (216)) \tag{225}$$

$$\leq \hat{\Phi}^{\pi}_{\mathbf{M_P},t}(\bar{b}_t, a) + \varepsilon_Q + \varepsilon^{\mathcal{P}}_Q + \varepsilon_\Phi + \varepsilon^{\mathcal{P}}_\Phi. \tag{226}$$

□

Further Implementation Details

Simulative Setting

In the 2D beacons environment, the state and observation spaces are defined as the whole plane: $\mathcal{X} = \mathcal{Z} = \mathbb{R}^2$. The planning horizon is $L = 15$.

We define the 2D rectangle $\textit{Rect}((x_1, y_1), (x_2, y_2))$ as the axis-aligned rectangle starting from the bottom-left corner (x_1, y_1) to the top-right corner (x_2, y_2):

$$\textit{Rect}((x_1, y_1), (x_2, y_2)) \triangleq \tag{227}$$

$$\left\{(x, y) \in \mathbb{R}^2 \mid x_1 \leq x \leq x_2, y_1 \leq y \leq y_2\right\} \tag{228}$$

The outer walls are defined by $\mathcal{X}_{\textit{collision}} \triangleq \neg\textit{Rect}((-2, 0), (12, 6))$. The goal region is defined by $\mathcal{X}_{\textit{goal}} = \textit{Rect}((4, -1.5), (6, 0))$. The prior is the following Gaussian mixture with 2 components:

$$b_0(x_0) = \frac{1}{2}\mathcal{N}(\mu_1, \Sigma_0) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma_0), \tag{229}$$

$$\mu_1 = (1, 2), \mu_2 = (9, 2), \tag{230}$$

$$\Sigma_0 = \text{diag}(\sigma^2_{x_0}, \sigma^2_{y_0}), \tag{231}$$

$$\sigma_{x_0} = 0.5, \sigma_{y_0} = 0.25. \tag{232}$$

The action space is the discrete space of the following 4 actions: $\mathcal{A} = \{(0,1),(0,-1),(1,0),(-1,0)\}$. The transition model is the following Gaussian model:

$$p_T(x' \mid x, a) = \mathcal{N}(x + a, \Sigma_T), \tag{233}$$

$$\Sigma_T = \mathrm{diag}(\sigma_T^2, \sigma_T^2), \tag{234}$$

$$\sigma_T^2 = 0.15. \tag{235}$$

In the arena the 6 beacons are arranged in a horizontal row in equal distances along the 2D line from $(0,4)$ to $(10,4)$, such that their locations are:

$$(x_1^{\textit{beacon}}, y_1^{\textit{beacon}}) = (0,4), \tag{236}$$

$$(x_2^{\textit{beacon}}, y_2^{\textit{beacon}}) = (2,4), \tag{237}$$

$$(x_3^{\textit{beacon}}, y_3^{\textit{beacon}}) = (4,4), \tag{238}$$

$$(x_4^{\textit{beacon}}, y_4^{\textit{beacon}}) = (6,4), \tag{239}$$

$$(x_5^{\textit{beacon}}, y_5^{\textit{beacon}}) = (8,4), \tag{240}$$

$$(x_6^{\textit{beacon}}, y_6^{\textit{beacon}}) = (10,4). \tag{241}$$

The range of a beacon is $R^{\textit{beacon}} = 1$. The light and dark regions are defined as the following:

$$\mathcal{X}_{\textit{light}} = \bigcup_{i=1}^{6} B((x_i^{\textit{beacon}}, y_i^{\textit{beacon}}); R^{\textit{beacon}}) \tag{242}$$

$$\mathcal{X}_{\textit{dark}} = \mathcal{X} \setminus \mathcal{X}_{\textit{light}} \tag{243}$$

The original and simplified observation models in the dark region are the same Gaussian model:

$$p_Z(z \mid x) = q_Z(z \mid x) = \mathcal{N}(x, \Sigma_{\textit{dark}}), \tag{244}$$

$$\Sigma_{\textit{dark}} = \mathrm{diag}(\sigma_{\textit{dark}}^2, \sigma_{\textit{dark}}^2), \tag{245}$$

$$\sigma_{\textit{dark}} = 5. \tag{246}$$

The original model in the light region is defined as a Gaussian mixture model made of rings of components with increasing number of components but decaying weights, centered around a center component. Its definition is the following:

$$N_\sigma = 3, k_r = 10, k_\theta = 25 \tag{247}$$

$$\sigma_{\textit{light}} = 0.3, \tag{248}$$

$$\Sigma_{p_Z} = \mathrm{diag}(\sigma_{\textit{light}} \frac{N_\sigma}{k_r}) \tag{249}$$

$$N_i^\theta = \max\{1, i \cdot k_\theta\}, \tag{250}$$

$$r_i = N_\sigma \cdot i, \tag{251}$$

$$w_i = \exp(-\frac{(i\frac{N_\sigma}{k_r})^2}{2}), \tag{252}$$

$$\tilde{w}_i = w_i / (\sum_{k=1}^{k_r} \sum_{j=1}^{N_i^\theta} w_k), \tag{253}$$

$$v_{i,j}^\theta = i\frac{N_\sigma}{k_r} \textit{Rot}(\frac{2\pi j}{N_i^\theta}) \begin{bmatrix} \sigma_{\textit{light}} \\ \sigma_{\textit{light}} \end{bmatrix} \tag{254}$$

$$p_Z(z \mid x \in \mathcal{X}_{\textit{light}}) = \sum_{i=1}^{k_r} \sum_{j=1}^{N_i^\theta} \tilde{w}_i \mathcal{N}(x + \cdot\, v_{i,j}^\theta, \Sigma_{p_Z}) \tag{255}$$

Algorithm 2: Empirical TV-Distance $\hat{\Delta}_Z$

Algorithm: Estimate Δ_Z.

Input: Number of state samples N_Δ, number of observation per state sample N_Z, total variation threshold $\Delta_{\textit{Thresh}}$, Q_0 state proposal distribution, p_Z original measurement model, q_Z simplified observation model.

Output: Number of kept states $N_\Delta^{\textit{eff}}$, a set of states $\{x_n^\Delta\}_{n=1}^{N_\Delta^{\textit{eff}}}$, TV-distance estimates $\{\hat{\Delta}_Z(x_n^\Delta)\}_{n=1}^{N_\Delta^{\textit{eff}}}$.

```
$X = list(), D = list()$
for all $n = 1, \dots, N_\Delta$ do
    $x_n^\Delta \sim Q_0$
    for all $j = 1, \dots, N_Z$ do
        $z_n^j \sim (p_Z + q_Z)/2$
    end for
    $\hat{\Delta}_Z(x_n^\Delta) = \sum_{j=1}^{N_Z} 2 \cdot \frac{|p_Z(z_j^n|x_n^\Delta) - q_Z(z_j^n|x_n^\Delta)|}{p_Z(z_j^n|x_n^\Delta) + q_Z(z_j^n|x_n^\Delta)}$
    if $\hat{\Delta}_Z(x_n^\Delta) > \Delta_{Thresh}$ then
        $X \leftarrow x_n^\Delta, D \leftarrow \hat{\Delta}_Z(x_n^\Delta)$
    end if
end for
return $(\#X), X, D$
```

where $\textit{Rot}(\varphi)$ is the 2D rotation matrix corresponding to angle φ. The simplified observation model is the following single component Gaussian:

$$q_Z(z \mid x \in \mathcal{X}_{\textit{light}}) = \mathcal{N}(x, \Sigma_{\textit{light}}), \tag{256}$$

$$\Sigma_{\textit{light}} = \mathrm{diag}(\sigma_{\textit{light}}^2, \sigma_{\textit{light}}^2), \tag{257}$$

$$\sigma_{\textit{light}} = 0.3. \tag{258}$$

The reward function is time and state dependent only, and defined as the sum of 3 indicators:

$$r_t(x) = R_{\textit{hit}} \cdot \mathbf{1}_{x \in \mathcal{X}_{\textit{goal}}} \tag{259}$$

$$+ R_{\textit{miss}} \cdot \mathbf{1}_{x \notin \mathcal{X}_{\textit{goal}}} \tag{260}$$

$$+ R_{\textit{collide}} \cdot \mathbf{1}_{x \in \mathcal{X}_{\textit{collision}}} \tag{261}$$

For all time steps $t > 0$ we set $R_{\textit{hit}} = 100$, $R_{\textit{collide}} = -50$. We set $R_{\textit{miss}} = -50$ if $t = L$, and otherwise -1 if $t > 0$. In the first time step the reward is 0, i.e. $r_0 = 0$. Additionally, there is no discount factor, i.e. $\gamma = 1$. From these definitions, it holds that $R_i^{\max} = 100$ and $V_t^{\max} = 100 + (15 - t)$ for $1 \leq t \leq 15$.

Bound Estimate

We now describe the parameters describing the computation of $\hat{\Delta}_Z$ and $\tilde{m}_i$. In Algorithm 2 we describe the process of sampling the delta states and estimating $\hat{\Delta}_Z$. In Algorithm 3 we describe the procedures for computing $\tilde{m}_i$ for a particle belief and for a state sample.

We chose the number of delta states $N_\Delta = 2048$, the number of observation samples for estimating $\hat{\Delta}_Z$ is $N_Z = 256$. The number of particles in the belief $C = 250$, and $N_X = 30$ is the number of particles used for computing $\hat{\tilde{m}}_i$. We chose $\Delta_{\textit{Thresh}} = 10^{-4}$, the threshold for filtering delta states with low delta. The threshold distance

Algorithm 3: Empirical Bounds $\tilde{m}_i$

```
**Algorithm:** Estimate $m_i\left(\bar{b}, a, i\right)$.
**Parameters**: Number of state samples $N_X$.
**Input**: Particle belief set $\bar{b}_i = \{(x_i, w_i)\}$, action $a$, time step $i$.
**Output**: A scalar $\hat{\tilde{m}}_i(\bar{b}_i, a)$ that is an estimate of $m_i(\bar{b}_i, a)$.
for all $j = 1, \ldots, N_X$ do
    $x_i^j \sim \bar{b}_i$
end for
return $\frac{1}{N_x} \sum_{j=1}^{N_X}$ Estimate $m_i\left(x_i^j, a, i\right)$

**Algorithm:** Estimate $m_i\left(x, a, i\right)$.
**Parameters**: Transition model $p_T$, Distance threshold $d_T$, $Q_0$ state proposal distribution.
**Input**: State $x_i$, action $a$, time step $i$.
**Output**: A scalar $\tilde{m}_i(x_i, a)$ that is an estimate of $m_i(x_i, a)$.
$Neighborhood = list()$
{Neighborhood search can be implemented with KD-Tree}
for all $n = 1, \ldots, N_{\Delta}^{\mathit{eff}}$ do
    if $\|x_i - x_n^{\Delta}\| \leq d_T$ then
        $Neighborhood \leftarrow x_n^{\Delta}$
    end if
end for
$m \leftarrow 0$
for all $x_n^{\Delta} \in Neighborhood$ do
    $m \leftarrow m + V_{i+1}^{\max} \cdot \frac{p_T(x_n^{\Delta} | x_i, a)}{N_{\Delta} Q_0(x_n^{\Delta})} \hat{\Delta}_Z(x_n^{\Delta})$
end for
return $m$
```

for the transition model is $d_T = 0.6 = 4 \cdot \sigma_T$. After prefiltering of $\hat{\Delta}_Z(x_n^{\Delta}) > \Delta_{\mathit{Thresh}}$, we had $N_{\Delta}^{\mathit{eff}} = 251$, sample mean of $\hat{\mathbb{E}}[\hat{\Delta}_Z] = 8.31 \cdot 10^{-2}$, and sample variance $\hat{\sigma}[\hat{\Delta}_Z] = 0.67 \cdot 10^{-2}$. Minimum and maximum values are $\min \hat{\Delta}_Z = 7.06 \cdot 10^{-2}$, $\max \hat{\Delta}_Z = 12.08 \cdot 10^{-2}$.

Solver

We now describe details regarding the implementation of the PFT-DPW solver (Sunberg and Kochenderfer 2018).

We implement the particle filter belief with 2 key points to note. The first is that we use resampling for stability of the particle filter, as when the particle filter algorithm was running without resampling we ran into particle depletion issues in a few scenarios. The second is that during planning with a particle belief, since our scenario may end prematurely, we often get a situation where only some particles terminate and others do not. To solve this problem, we note that in general, we may describe the reward of a belief with the law of total expectation:

$$Q_t^{\pi}(b_t, a) = r_t(b_t, a) + \gamma \mathbb{E}_{z_{t+1}}[V_{t+1}^{\pi}(b_{t+1})] \tag{262}$$

$$r_t(b_t, a) \tag{263}$$

$$+\gamma \mathbb{E}_{z_{t+1}}[V_{t+1}^{\pi}(b_{t+1}) \mid T(b_{t+1})]\mathbb{P}(T(b_{t+1})) \tag{264}$$

$$+\gamma \mathbb{E}_{z_{t+1}}[V_{t+1}^{\pi}(b_{t+1}) \mid \neg T(b_{t+1})]\mathbb{P}(\neg T(b_{t+1})) \tag{265}$$

where $T(b_{t+1})$ is a random variable indicating that a belief has terminated, whether because of states being in terminal states (due to entering goal or collision states), or due to reaching the time limit. The future value for all terminated states is 0, therefore the action value is equal to:

$$Q_t^{\pi}(b_t, a) = r_t(b_t, a) \tag{266}$$

$$+\gamma \mathbb{E}_{z_{t+1}}[V_{t+1}^{\pi}(b_{t+1}) \mid \neg T(b_{t+1})]\mathbb{P}(\neg T(b_{t+1})) \tag{267}$$

We implemented $\mathbb{P}(\neg T(b_{t+1}))$ as summing the total particle weights that are in goal or collision states after sampling from the transition model. To implement $\mathbb{E}_{z_{t+1}}[V_{t+1}^{\pi}(b_{t+1}) \mid \neg T(b_{t+1})]$, we added a parameter controlling the normalized sum of the weights of a particle belief, which we multiply after each transition by $\mathbb{P}(\neg T(b_{t+1}))$. This conditions all future beliefs of the same branch by the same factor, and trickles down the belief tree recursively.

The rollout policy is based on maximum likelihood transitions and observations. It steers the empirical mean of the belief towards the goal by executing each time step the action that has maximal inner product with the goal's center.

We implement in PFT-DPW a secondary reward that is stored in all posterior belief nodes. We modified the posterior nodes in the planner's belief tree to include $\hat{\Phi}_{\mathbf{M_P},t}$ values in addition to $\hat{Q}$. In addition, we hooked on the generative model of the environment to additionally compute m_i in the simplified planning, such that $(\bar{b}_{i+1}, r_{i+1}, m_i) \sim G(\bar{b}_i, a_i)$.

The parameters of PFT-DPW were taken to be the following:

Parameter	Notation (PFT-DPW)	Value
No. of simulations	n	500
Exploration bonus	c	50
Action prog. widening mult. const.	k_a	1.1
Action prog. widening exp. const.	α_a	0.24
Obs. prog. widening mult. const.	k_o	1.1
Obs. prog. widening exp. const.	α_o	0.19

Additionally, the number of particles in the particle filter was $C = 250$.