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

Idan Lev-Yehudi¹, Moran Barenboim¹, Vadim Indelman²

¹Technion Autonomous Systems Program (TASP), Technion - Israel Institute of Technology, Haifa 32000, Israel
²Department of Aerospace Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel
moranbar, idanlev@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, there have been several practical approaches to obtain probabilistic visual observation models in AI (Eslami et al. 2018) suggested a neural model for scene rendering. Both (Jon schkowski, 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...
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 $\langle X, A, Z, p_T, p_Z, r, \gamma, L, b_0 \rangle$. $X, A, Z$ are the state space, action space and observation space respectively. Throughout the paper, $x \in X, a \in A, z \in Z$ denote individual states, actions and observations respectively. We consider $X$ and $Z$ that are continuous, while $A$ can be either discrete or continuous.

The conditional probability distributions $p_T$ and $p_Z$ are the transition and observation models respectively. $p_T(x' | x, a)$ models the uncertainty of taking action $a \in A$ from state $x \in X$, and $p_Z(z | x)$ models the uncertainty of receiving an observation $z \in Z$ at state $x$. The reward function $r_i : \mathbb{N} \times X \times A \rightarrow \mathbb{R}$ gives the immediate reward of applying action $a$ at state $x$ and time $t$. We denote $r_i : X \times A \rightarrow \mathbb{R}$ 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 time $t$: $H_t \triangleq (b_0, a_0, z_1, \ldots, a_{t-1}, z_t)$. The belief at time $t$ is defined as the conditional distribution of the state given the history $b_t(x_t) \triangleq \mathbb{P}(x_t | H_t)$. We denote $H_{i+1} \triangleq (b_0, a_0, z_1, \ldots, a_{i-1})$ for the same history without the last measurement, and the propagated belief $b_t^\pi(x_t) \triangleq \mathbb{P}(x_t | H_t^\pi)$. The belief reward is the expected state reward: $r_i(b_t, a) \triangleq \mathbb{E}_{x_t \sim b_t} [r_i(x_t, a)]$. A policy $\pi$ is a mapping from the space of all histories to the action space $\pi_t : H_t \rightarrow A$. A policy $\pi$ is a collection of policies from the start time until the
POMDP terminates, i.e. \( \pi \triangleq (\pi_i)_{i=1}^{L-1} \), and for brevity we denote \( \tau_i \) instead of \( \tau_i(H_i) \). 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:

\[
V_t^\pi(b_t) = \mathbb{E}_{x_{t+1:t+L} \sim \mathbb{P}_t} \left[ \sum_{t'=t}^{t+L-1} \gamma^{t'-t} r_{t'}(b_{t'}, \pi_{t'}) \right].
\]

Most often the action value function, defined as

\[
Q_t^\pi(b_t, a_t) = \mathbb{E}_{x_{t+1} \sim \mathbb{P}_t}(r_t(b_t, a_t) + \gamma \mathbb{E}_{z_{t+1} \sim \mathbb{P}_t}[V_{t+1}^\pi(b_{t+1})]),
\]

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 date the belief exactly. Particle filters are used instead to approximated for computational reasons. We denote the simplified value function based on belief updates in planning with either the original or simplified observations models as \( V_t^{\pi, p_Z} \) and \( V_t^{\pi, q_Z} \), respectively, and similarly for \( Q_t^{\pi, p_Z} \) and \( Q_t^{\pi, q_Z} \).

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 \( |V_t^{\pi, p_Z} - V_t^{\pi, q_Z}| \), 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: \( \tau_i \triangleq p_{\tau|x_i} | x_i-1, \pi_i-1 \), \( [\zeta/\xi]_i \triangleq [p_Z/q_Z](z_i | x_i) \), \( [\zeta/\xi]_i^H \triangleq [p_Z/q_Z](z_i | H_i) \). The values in \([\cdot]\) can be replaced with either respective option. The product of densities is denoted \( [\tau/\zeta/\xi]_{i:j} = \prod_{i=1}^{j-1} [\zeta/\xi]_i \). We denote the expectations:

\[
\mathbb{E}_\tau[*] \triangleq \int_{x_t} b_t(x_t) \ast dz_t
\]

\[
\mathbb{E}_{\tau_i}[*] \triangleq \int_{x_{i:j}} b_{i:j}(x_{i:j}) \ast dz_{i:j}
\]

\[
\mathbb{E}_{\tau_{i:j}}[p_Z/q_Z][*] \triangleq \int_{z_{i:j}} \mathbb{E}_{\tau_{i-1}}[c[\zeta/\xi]_{i:j}][*] \ast dz_{i:j}
\]

\[
\mathbb{E}_{\tau_{i:j+1}}[p_Z/q_Z][*] \triangleq \int_{z_{i:j}} \mathbb{E}_{\tau_{i-1}}[c[\zeta/\xi]_{i:j+1}][*] \ast dz_{i:j}
\]

\[
\mathbb{E}_{\tau_{i:j+1}}[p_Z/q_Z][*] \triangleq \int_{z_{i:j}} \mathbb{E}_{\tau_{i-1}}[c[\zeta/\xi]_{i:j+1}][*] \ast dz_{i:j}
\]

It holds that \( [\zeta/\xi]_{i:j}^H \mathbb{E}_{\tau_{i-1}}[c[\zeta/\xi]_{i:j}][*] \) by the law of total probability, and it follows that \( \mathbb{E}_{b_{i-1}}[c[\zeta/\xi]_{i:j}][*] \) by Fubini’s theorem (Durrett 2019, 1.7).

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

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

i. The reward at each time step is bounded: \( |r_i(x_i, a_i)| \leq R_i^{\max} \), for time index \( i \). Hence, follows from the triangle inequality that the value function at a certain time step is bounded by \( V_{t-1}^\pi(b_t, a_t) \leq \sum_{i=1}^{L} \gamma^{t-i} R_i^{\max} \leq V_t^{\max} \), and we define \( V_t^{\max} \triangleq \max_{i} V_t^{\max} \).

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 \( \varepsilon \).

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. \( w_i^t \propto \frac{\ell_i(z_i^t)}{w_i^{t-1}} \) for the \( j \)’th observation sample \( z_i^t \). 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: \( \sup_{z_{i+1}} \mathbb{E}[\ell_i(z_i, \pi_i)] \leq d_{\infty}^{\text{max}} \), where \( d_{\infty}^{\text{max}} \in \mathbb{R} \). 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.

\[
\mathbb{E}_{t+1:i}^{p_{t+1} | z}[r_i(b_i, \pi_i)] = \mathbb{E}_{t+1:i}^{p_t | p_{t+1} | z}[r_i(x_i, \pi_i)].
\]

**(6)**

**Proof.** We refer to the supplementary material for all proofs (Lev-Yehudi, Barenboim, and Indelman 2023).

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

\[
\Delta_Z(x) \triangleq \int \rho_Z(z \mid x) - q_Z(z \mid x) \, dz.
\]

**(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 (Tsymbakov 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 \( \Delta_Z \).

**Theorem 1.** Assume current belief is \( b_i \) and a given policy is \( \pi \). Denote with \( b_i^{\text{op}} \), \( b_i^{\text{sp}} \) the future belief at time step \( i \) updated with either \( p_{t+1} \) or \( q_{t+1} \), respectively. Then it holds that

\[
|\mathbb{E}_{t+1:i}^{p_{t+1} | z}[r_i(b_i^{\text{op}} | \pi_i)] - \mathbb{E}_{t+1:i}^{q_{t+1} | z}[r_i(b_i^{\text{sp}} | \pi_i)]| \\
\leq R^{\text{max}}_i \sum_{t=i+1}^{L} \mathbb{E}_{t+1:i}^{p_{t+1} | z}[\Delta_Z(x_i)] ,
\]

**(8)**

i.e. the difference between the expected belief at 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:

\[
|V_t^{\text{op}, \pi}(b_i) - V_t^{\text{sp}, \pi}(b_i)| \\
\leq \sum_{t=i+1}^{L} \mathbb{E}_{t+1:i}^{p_{t+1} | z}[\Delta_Z(x_i)] \sum_{i=1}^{L} \gamma^{t-i} R^{\text{max}}_i \\
= \sum_{i=i+1}^{L} V_i^{\text{max}}, \mathbb{E}_{t+1:i}^{p_{t+1} | z}[\Delta_Z(x_i)].
\]

**(9)**

**Equivalence to State-Action Local Bound**

Corollary 1 requires computing the theoretical expectations \( \mathbb{E}_{t+1:i}^{p_{t+1} | z}[\Delta_Z(x_i)] \), 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,

\[
m_i(x_i, \pi_i) \triangleq V_{i+1}^{\text{max}}, \mathbb{E}_{t+1:i}^{p_{t+1} | z}[\Delta_Z(x_i)],
\]

**(11)**

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

**(12)**

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

1. Cumulative bound for a given policy and initial belief,

\[
M_t^{\text{op}}(b_i) \triangleq \mathbb{E}_{t+1:L-1}^{q_{t+1}}[\sum_{i=1}^{L-1} m_i(b_i, \pi_i)], \text{ analogous to } V_t^{\text{op}}(b_i).
\]

**(13)**

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, \( \Phi_t^{\text{op}}(b_i, a) \triangleq m_i(b, a) + \mathbb{E}_{t+1}^{q_{t+1}}[M_t^{\text{op}}(b_{i+1})] \), analogous to \( Q_t^{\text{op}}(b_i, a) \).

**(14)**

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

\[
|V_t^{\text{op}, \pi}(b_i) - V_t^{\text{sp}, \pi}(b_i)| \leq M_t^{\text{op}}(b_i).
\]

**(15)**

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

\[
|Q_t^{\text{op}, \pi}(b_i, a) - Q_t^{\text{sp}, \pi}(b_i, a)| \leq \Phi_t^{\text{op}}(b_i, a).
\]

**(16)**

Computing \( M_t^{\text{op}}(b_i) \) 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 \( \Phi_t^{\text{op}}(b_i, a) \) 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 (Sztysic and Indelman 2022).
Online Estimator of Local Bound

We now show how we compute \( m_i \) during online planning without requiring access to \( p_{X|Z} \). We do this by offline pre-sampling \( x_n \) \( \sim Q_0(x) \), named delta states, and evaluating \( \Delta_Z(x_n^2) \) for each. During online planning, we simply reweight \( \Delta_Z \) using the importance sampling formalism for state sample \( x^i \). We use this to redefine \( m_i \) and also to define the empirical estimate \( \hat{m}_i \).

\[
m_i(x^i, a) = V^{\max}_{i+1} \mathbb{P}_{x^i \sim Q_0} \left[ \frac{p_{x^i|x^i,a}}{Q_0(x^i)} \Delta_Z(x^i) \right],
\]

(15)

\[
\hat{m}_i(x^i, a) \triangleq V^{\max}_{i+1} \frac{1}{N_\Delta} \sum_{i=1}^{N_\Delta} \frac{p_{x^i|x^i,a}}{Q_0(x_n^2)} \Delta_Z(x_n^2).
\]

(16)

Based on this, we define \( \hat{\Phi}_t^\pi(h_t, a) \) when computed with \( \hat{m}_i \) instead of \( m_i \). We explicitly define this notation to differentiate \( \hat{\Phi} \) from \( \Phi^\pi \). 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 \( \hat{\Phi} \triangleq \hat{\Phi}^\pi \).

The support of \( Q_0 \) must cover all of \( X \) 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 \( X \) 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 is possible to choose constants \( \lambda, \nu, C, N_r \) such that the following holds with probability of at least 1 - \( \delta \):

\[
|Q_{\pi|\nu|C}^\pi(h_t, a) - Q_{\pi|\nu|C}^\pi(h_t, a)| \leq \varepsilon.
\]

(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 2. Assuming that \( P \) is an MDP planner that can approximate \( Q_{\nu} \)-values with arbitrary precision \( \varepsilon \) at an accuracy 1 - \( \delta \), we denote the precision and accuracy of the action value and action cumulative bound functions:

\[
\mathbb{P}(\|Q_{\pi|\nu|C}^\pi(h_t, a) - \hat{Q}_{\pi|\nu|C}^\pi(h_t, a)\| \leq \varepsilon_Q) \geq 1 - \delta_Q
\]

(24)

\[
\mathbb{P}(\|\Phi_{\pi|\nu|C}^\pi(h_t, a) - \hat{\Phi}_{\pi|\nu|C}^\pi(h_t, a)\| \leq \varepsilon_\Phi) \geq 1 - \delta_\Phi
\]

(25)

From Corollary 2 it holds that we can choose constants \( \lambda, \nu, C, N_r \) such that the following holds.

\[
\mathbb{P}(\|Q_{\pi|\nu|C}^\pi(h_t, a) - Q_{\pi|\nu|C}^\pi(h_t, a)\| \leq \varepsilon_Q) \geq 1 - \delta_Q,
\]

(26)

\[
\mathbb{P}(\|\Phi_{\pi|\nu|C}^\pi(h_t, a) - \hat{\Phi}_{\pi|\nu|C}^\pi(h_t, a)\| \leq \varepsilon_\Phi) \geq 1 - \delta_\Phi.
\]

(27)

Then with probability of at least 1 - (\( \delta_Q + \delta_\Phi + \delta_P + \delta_\Phi \))

\[
|Q_{\pi|\nu|C}^\pi(h_t, a) - \hat{Q}_{\pi|\nu|C}^\pi(h_t, a)| \leq \varepsilon_P
\]

\[
\hat{\Phi}_{\pi|\nu|C}^\pi(h_t, a) + \varepsilon_Q + \varepsilon_P + \varepsilon_\Phi + \varepsilon_\Phi.
\]

(28)

In summary, any planner that can approximate the PB-MDP values \( Q_{\pi|\nu|C}^\pi(h_t, a) \) with high probability can also approximate the cumulative bound \( \Phi_{\pi|\nu|C}^\pi(h_t, a) \) with high probability, since it is mathematically formulated like a state reward. Therefore, we can construct a bound for the theoretical value \( Q_{\pi|\nu|C}^\pi(h_t, a) \) with high probability from online calculations that do not involve access to \( p_{X|Z} \).

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. \(\max_a [L_E] / [U_E](Q_{\pi, \sigma_Z}^{\pi}(b_t, a))\), we will obtain the same action choice by computation of \(\max_a \tilde{Q}^{\pi, \sigma_Z}_{M, t}(\tilde{b}_t, a) [-//+] \tilde{\Phi}_{M, t}(\tilde{b}_t, a)\).

**Implementation**

We verify the computability of our approach in a 2D beacons POMDP. We integrate the computation of \(\tilde{m}_i\) and \(\tilde{\Phi}_{M, t}\) 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 \(X_{\text{light}}\) is defined by circles centered at each of the 6 beacons located at the top of the arena, and \(X_{\text{dark}} = X \setminus X_{\text{light}}\). The observation model in the light region \(p_Z(z \mid x \in X_{\text{light}})\) 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 \(p_Z(z \mid x \in X_{\text{light}}) \sim N(x, \Sigma_{\text{dark}})\).

We refer readers to (Bar-Shalom, Li, and Kirubarajan 2004, 1.4.16) on approximating GMMs with a single Gaussian. In the dark region, \(p_Z(z \mid x \in X_{\text{dark}}) = q_Z(z \mid x \in X_{\text{dark}}) = N(x, \Sigma_{\text{dark}})\). 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 set is \(A = \{\{\pm 1, 0\}, (0, \pm 1)\}\). 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: \(r_t(x) = R_{\text{hit}} \cdot 1_{x \in X_{\text{goal}}} + R_{\text{miss}} \cdot 1_{x \notin X_{\text{goal}}} + R_{\text{collision}} \cdot 1_{x \in X_{\text{collision}}}\). In all time steps, \(R_{\text{hit}} = 100\), \(R_{\text{collision}} = -50\). The miss reward is \(R_{\text{miss}} = -50\) 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 \(x_{\pi, n}^{\Delta}\) 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 choose 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 Koskama-Hlwaka inequality (Lemieux 2009, 5.6). However, it is generally hard to compute, and infinite for several very simple functions.

For each sampled \(x_{\pi, n}^{\Delta}\) we estimate \(\Delta_Z\) via observation samples, based on assumption iii. We perform the following importance sampling estimation w.r.t. \((p_Z + q_Z)/2\):

\[
\hat{\Delta}_Z(x_{\pi, n}^{\Delta}) = \sum_{j=1}^{N_Z} \frac{p_Z(z_j^n \mid x_{\pi, n}^{\Delta}) - q_Z(z_j^n \mid x_{\pi, n}^{\Delta})}{p_Z(z_j^n \mid x_{\pi, n}^{\Delta}) + q_Z(z_j^n \mid x_{\pi, n}^{\Delta})} \sum_{j=1}^{N_Z} \frac{p_Z(z_j^n \mid x_{\pi, n}^{\Delta}) - q_Z(z_j^n \mid x_{\pi, n}^{\Delta})}{p_Z(z_j^n \mid x_{\pi, n}^{\Delta}) + q_Z(z_j^n \mid x_{\pi, n}^{\Delta})}
\]

where \(\{z_j^n\}_{j=1}^{N_Z} \overset{i.i.d.}{\sim} (p_Z + q_Z)/2\). 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 \(x_{\pi, n}^{\Delta}\) for which \(\Delta_Z(x_{\pi, n}^{\Delta}) > \Delta_{\text{thresh}}\). The second optimization is to only consider sampled states within a truncation distance of \(d_T\) from state particle \(x_t^n\) in (16). We implemented this by keeping all \(x_{\pi, n}^{\Delta}\) in a KD-tree for efficient radius queries (Maneewongvatana and Mount 1999).

We chose \(\Delta_{\text{thresh}}\) and \(d_T\) such that the error in computing \(\tilde{m}_i\) is at most \(V_{\text{max}}^{\pi}, 10^{-4}\). Lastly, since the runtime complexity of \(\tilde{m}_i\) grows linearly with \(C\), the number of particles in the belief \(\tilde{b}_i\), we limit the number of particles used to \(N_x\) by performing MC estimation w.r.t. the particle belief: \(\tilde{m}_i(\tilde{b}_i, a) = \frac{1}{N_x} \sum_{j=1}^{N_x} \tilde{m}_i(x_t^n, a)\) where \(\{x_t^n\}_{j=1}^{N_x} \overset{i.i.d.}{\sim} \tilde{b}_i\).

**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 \(\hat{Q}_t^{[p_z/q_z]}\) and the plan session duration. In simplified planning, we also record the expected action cumulative bound \(\tilde{\Phi}_{M, t}\). The policy is maximization of the estimated action value function based on the original or simplified model, i.e. \(\pi_t^{[p_z/q_z]} \triangleq \max_a \hat{Q}_t^{[p_z/q_z]}(\tilde{b}_i, a)\), 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 \(\tilde{\Phi}_{M, t}\) 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 and evaluating the PDF of the original observation model, leading to longer planning times even when additionally computing \(\tilde{\Phi}\) in the simplified planning.
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 \( \{ x^\Delta_n \}_{n=1}^{N_{\text{iter}}} \) are indicated by the tri-downs, with color relative to estimated TV-distance of the simplified observation model \( 7.06 \leq \Delta_{x_n} \leq 12.08 \). 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 \( ^1 \). 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 \( ^1 \).

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 \( q_Z \).

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

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 \( \pi^{LB}_t = \arg\max_a \{ Q^t_{LB} (b_t, a) - \Phi_t (b_t, a) \} \) and the upper bound policy as \( \pi^{UB}_t = \arg\max_a \{ Q^t_{UB} (b_t, a) + \Phi_t (b_t, a) \} \). 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. \( \pi^{LB} \) tends to steer away from the light region, hence it differs mostly in the earlier time steps, whereas \( \pi^{UB} \) 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.

**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


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.


Sunberg, Z.; and Kochenderfer, M. 2018. Online algorithms for POMDPs with continuous state, action, and observation


