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 System Program (TASP) [2]Department of Aerospace Engineering, Technion

POMDPs with Visual Observations

Planning under uncertainty can be formalized as a Partially Observable Markov Decision Process (POMDP). **Optimally solving POMDPs is computationally infeasible** except for only the smallest tasks.

Visual observations are complex to model for planning. **Learned observation models are impractical** for use in solving the POMDP in real-time due to the many samples required in POMDP solvers.

Contribution

We explore **planning with a simpler observation model** with formal guarantees of the solution quality for computational reasons.

Our main contributions:

- Bound the theoretical loss with observation model discrepancy
- Probabilistic bound for the empirical simplified performance
- Practical computation of the bounds in state-of-the-art planners

Problem Formulation

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 state, action and observation spaces
- p_T, p_Z are probabilistic transition and observation models
- $r_t \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is a bounded reward function at time t
- γ is the reward discount for future time steps
- L is the time limit (horizon)
- b_0 is the starting distribution (belief) of states

During planning we replace the **original observation model p_Z** with a **simplified observation model q_Z**.

The original and simplified action-value functions:

$$Q^{p_Z}_{\mathbf{P}}(b_t, a) \triangleq r_t(b_t, a) + \mathbb{E}^{p_Z}_{t+1:L}[\sum_{i=t+1}^{L} \gamma^{i-t} r_i(b_i, \pi_i)]$$

$$Q^{q_Z}_{\mathbf{P}}(b_t, a) \triangleq r_t(b_t, a) + \mathbb{E}^{q_Z}_{t+1:L}[\sum_{i=t+1}^{L} \gamma^{i-t} r_i(b_i, \pi_i)]$$

We denote the **original POMDP as $\mathbf{P}$**, and its **particle-belief MDP as $\mathbf{M_P}$** (approximating with a finite number of particles).

Bounds Using Total-Variation Distance

State-dependent total-variation distance between observation models:

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

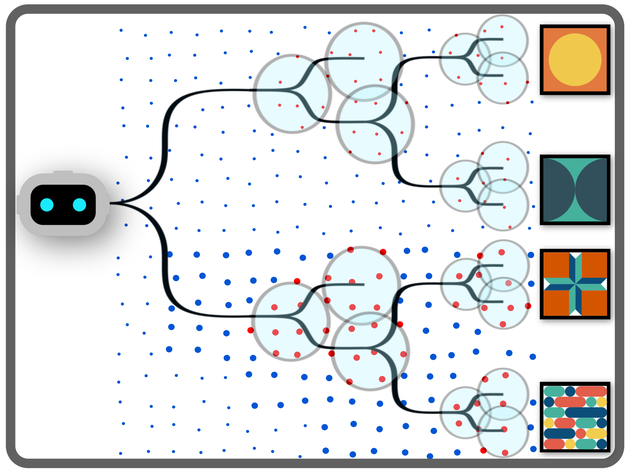
m_i quantifies a one-timestep bound over the loss in value function (intuitive):

$$m_i(x_i, a) \triangleq V^{\max}_{i+1} \cdot \mathbb{E}_{x_{i+1} \sim p_T(\cdot \mid x_i, a)}[\Delta_Z(x_{i+1})]$$

$$m_i(b_i, a) \triangleq \mathbb{E}_{x_i \sim b_i}[m(x_i, a)]$$

We extend with the definition of the **cumulative bound function Φ**:

$$\Phi_{\mathbf{P}}(b_t, a) \triangleq m_t(b_t, a) + \mathbb{E}^{q_Z}_{t+1:L-1}[\sum_{i=t+1}^{L-1} m_i(b_i, \pi_i)]$$



An illustration of our bounds approach. The scattered dots are the pre-sampled states, and the dot size is relative to $\Delta_Z(x)$. For the two policies, we compute the bound as a summation over Δ_Z weighted by the transition model. The bottom policy chooses actions that give higher weights to states with greater Δ_Z, resulting in looser bounds.

Deterministic Value Loss Bound (Theorem 2)

For every belief b_t, action a, policy π, observation models p_Z and q_Z, the following bound holds deterministically:

$$|Q^{p_Z}_{\mathbf{P}}(b_t, a) - Q^{q_Z}_{\mathbf{P}}(b_t, a)| \leq \Phi_{\mathbf{P}}(b_t, a)$$

Generalized PB-MDPs Convergence (Theorem 3, Informal)

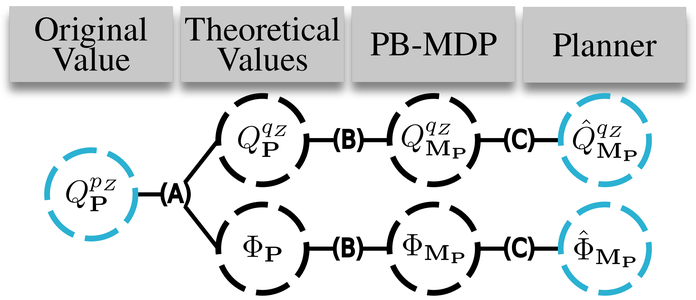
For **every bounded state-action function** (r_i/m_i), its finite-sample cumulative function ($Q^{q_Z}_{\mathbf{M_P}}/\Phi_{\mathbf{M_P}}$) has **probabilistic concentration bounds** from its theoretical counterpart ($Q^{q_Z}_{\mathbf{P}}/\Phi_{\mathbf{P}}$) under certain regularity conditions of the POMDP.

Empirical Concentration Inequalities (Corollary 3)

For arbitrary $\varepsilon, \delta > 0$ there exist numbers of particles and state samples for which

$$|Q^{p_Z}_{\mathbf{P}}(b_t, a) - \hat{Q}^{q_Z}_{\mathbf{M_P}}(\bar{b}_t, a)| \leq \hat{\Phi}_{\mathbf{M_P}}(\bar{b}_t, a) + \varepsilon$$

with probability of at least $1 - \delta$ for any planner with performance guarantees, where $\bar{b}_t$ is the particle-approximation of the belief b_t.



Summary of Corollary 3 for probably approximately bounding $|Q^{p_Z}_{\mathbf{P}} - \hat{Q}^{q_Z}_{\mathbf{M_P}}| \leq \hat{\Phi}_{\mathbf{M_P}}$. (A) is Theorem 2, connecting theoretical value functions with the theoretical local state bound. (B) is Theorem 3, connecting theoretical action value functions with their PB-MDP approximation. (C) is given by any planner with performance guarantees.

Practical Computation of Bounds

It is impractical to calculate m_i explicitly. Therefore, we propose **estimating m_i with samples**, and **separating calculations to offline/online stages**.

During offline we pre-sample states at which we compute $\Delta_Z(x)$, and the proposal distribution likelihood.

During online, we only reweight the states based on the transition model, and perform the summation.

Online

$$\tilde{m}_i(x_i, a) \triangleq V^{\max}_{i+1} \frac{1}{N_\Delta} \sum_{i=1}^{N_\Delta} \frac{p_T(x^\Delta_n \mid x_i, a)}{Q_0(x^\Delta_n)} \Delta_Z(x^\Delta_n)$$

$$\{x^\Delta_n\}_{n=1}^{N_\Delta} \sim Q_0(x)$$

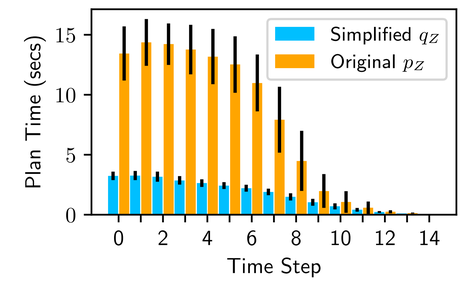
Offline

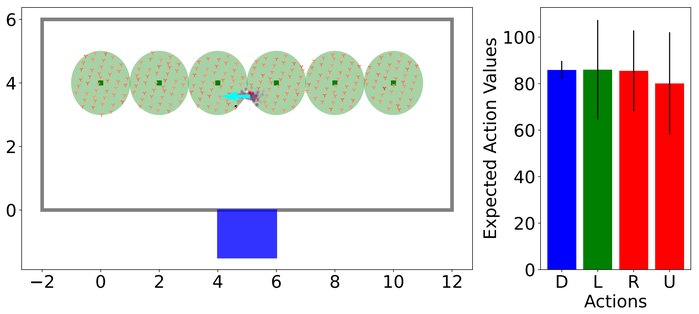
To optimize run-time:

- Discard x^Δ_n with low Δ_Z values
- Consider x^Δ_n based on KD-Tree and truncation distance
- Monte Carlo estimate - sample belief particles

Results in Simulation

We show that even with bounds calculation, we achieve a **significant speedup when planning with the simplified model**.





We set up a 2D light-dark simulation, with different original and simplified observation models in the light region. The initial belief is multi-modal, so the agent has to move up to the light region for better localizing before moving to the goal region (blue rectangle). We plan with the simplified observation model and compute $\hat{\Phi}_{\mathbf{M_P}}$. The simplified value policy chooses the left action, whereas the lower bound policy chooses down.

[idanlev, moranbar]@campus.technion.ac.il
vadim.indelman@technion.ac.il