Online POMDP planning with Anytime Deterministic Guarantees

Technion
Israel Institute of Technology

TASP | TECHNION AUTONOMOUS SYSTEMS PROGRAM

ANPL | Autonomous Navigation and Perception Lab

Moran Barenboim and Vadim Indelman

moranbar@campus.technion.ac.il , vadim.indelman@technion.ac.il

Motivation

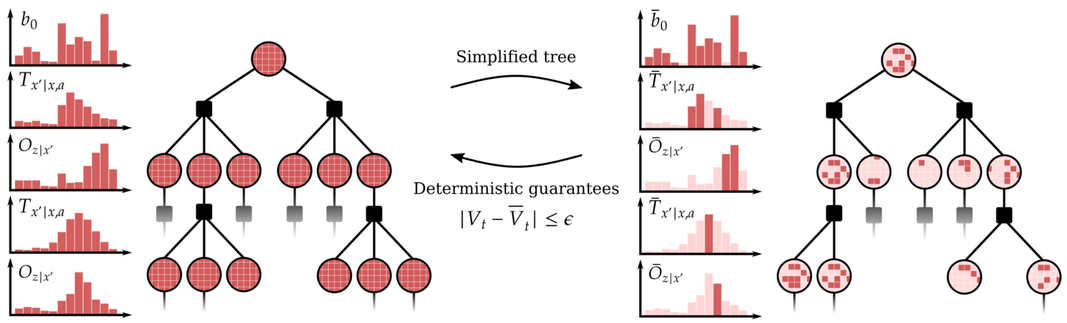
- In an uncertain environment, decisions may be based on a stochastic representation of the world.
- Planning under uncertainty commonly formalized using partially observable Markov decision processes (POMDPs)
- **However**, finding an optimal plan for a POMDP is usually intractable due to high computational burden.
- State-of-the-art (SOTA) solvers rely on online tree search to approximate the optimal plan.
- SOTA solvers are limited to asymptotic and/or probabilistic guarantees relative to the optimal plan.

Contributions

Derive deterministic guarantees and an exploration criteria for online planning with a simplified observation space

Derive guarantees when both the state and observation spaces are simplified (e.g. sampled)

Show how to apply the derivations to provide guarantees for existing SOTA solvers, e.g. POMCP and DESPOT



Problem Formulation

We consider a finite horizon, discrete POMDP and assume the reward function is bounded from above.

The optimal value function denoted as,

$$V_t^{\pi^*}(b_t) = \max_{a_t}\{r(b_t, a_t) + \mathop{\mathbb{E}}_{z_{t+1}|a_t,b_t}\left[V_{t+1}^{\pi^*}(b_{t+1})\right]\}$$

and the state, action and observation spaces are denoted as, $x_t \in \mathcal{X}$, $a_t \in \mathcal{A}$ and $z_t \in \mathcal{Z}$

$b_t \triangleq \mathbb{P}(x_t \mid H_t)$ denotes the belief at time step t.

$r(b_t, \pi_t)$ denotes the reward.

$\pi_t \equiv \pi_t(b_t)$ denotes the policy at a given belief, at time step t.

Approach

1. **Select only a subset of the states and observations at each node,** $\bar{\mathcal{X}}(H_{t+1}^-) \subseteq \mathcal{X}$ and $\bar{\mathcal{Z}}(H_t) \subseteq \mathcal{Z}$

$$\bar{b}_0(x) \triangleq \begin{cases} b_0(x) & , x \in \bar{\mathcal{X}}_0 \\ 0 & , otherwise \end{cases}$$

$$\bar{\mathbb{P}}(x_{t+1} \mid x_t, a_t) \triangleq \begin{cases} \mathbb{P}(x_{t+1} \mid x_t, a_t) & , x_{t+1} \in \bar{\mathcal{X}}(H_{t+1}^-) \\ 0 & , otherwise \end{cases}$$

$$\bar{\mathbb{P}}(z_t \mid x_t) \triangleq \begin{cases} \mathbb{P}(z_t \mid x_t) & , z_t \in \bar{\mathcal{Z}}(H_t) \\ 0 & , otherwise \end{cases}$$

2. **Define a simplified value function, which relies on the definition of the subset,**

$$\begin{aligned} \bar{V}^{\pi}(\bar{b}_t) &\triangleq r(\bar{b}_t, \pi_t) + \bar{\mathbb{E}}\left[\bar{V}(b_t)\right] \\ &= \sum_{x_t} \bar{b}(x_t) r(x_t, \pi_t) + \sum_{z_t} \bar{\mathbb{P}}(z_{t+1} \mid H_{t+1}^-) \bar{V}(\bar{b}(z_{t+1})) \end{aligned}$$

Simplified observation space

Simplified observation and state spaces

3. **Derive bounds relative to to the theoretical, unknown, value function**

$$\left|V^{\pi}(b_t) - \bar{V}^{\pi}(b_t)\right| \leq \mathcal{R}_{\max} \sum_{\tau=t+1}^{T} \left[1 - \sum_{z_{t+1:\tau}} \sum_{x_{t:\tau}} b(x_t) \prod_{k=t+1}^{\tau} \bar{\mathbb{P}}(z_k \mid x_k) \mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})\right] \triangleq \epsilon_z^{\pi}(b_t)$$

4. **Derive new exploration criteria, and guarantee its convergence to the optimal value in finite time**

$$V^{\pi *}(b_t) \leq \mathrm{UDB}^{\pi}(b_t)$$

where,

$$\begin{aligned} \mathrm{UDB}^{\pi}(b_t) &\triangleq \max_{a_t \in \mathcal{A}} [\bar{Q}^{\pi}(b_t, a_t) + \epsilon_z^{\pi}(b_t, a_t)] \\ &= \max_{a_t \in \mathcal{A}} [r(b_t, a_t) + \bar{\mathbb{E}}_{z_{t+1}|b_t,a_t}[\bar{V}^{\pi}(b_{t+1})] + \epsilon_z^{\pi}(b_t, a_t)] \end{aligned}$$

3. **Derive bounds relative to to the theoretical, unknown, value function**

$$\left|V^{\pi}(b_0) - \bar{V}^{\pi}(\bar{b}_0)\right| \leq \mathcal{R}_{\max}\left[1 - \sum_{x} \bar{b}_0(x)\right] + \mathcal{R}_{\max} \sum_{\tau=1}^{T} \left[1 - \bar{\mathbb{E}}_{z_{1:\tau}} \sum_{x} \bar{b}_{\tau}(x)\right] \triangleq \epsilon_{x,z}^{\pi}(b_0)$$

* Notably, this bound only viable at the root node, when t = 0, and thus is not used for exploration.
However, it is being used attached to SOTA algorithms guaranteeing their performance, as shown in the experimental section.

Results

- Attach our bounds to state-of-the-art solvers, POMCP and AR-DESPOT.
- Compute and update the bounds at planning session.
- Use the bounds to select an action to perform after the planning has completed.

Table 1: Performance comparison with and without deterministic bounds.

Algorithm	Tiger POMDP	Laser Tag	Discrete Light Dark	Baby POMDP
DB-DESPOT (ours)	3.74±0.48	−5.29±0.14	−5.29±0.01	−3.92±0.56
AR-DESPOT	2.82±0.55	−5.10±0.14	−61.53±5.80	−5.40±0.85
DB-POMCP (ours)	3.01±0.21	−3.97±0.24	−3.70±0.82	−4.48±0.57
POMCP	2.18±0.76	−3.92±0.27	−4.51±1.15	−5.39±0.63