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

Idan Lev-Yehudi Moran Barenboim Vadim Indelman

Introduction

- Planning under uncertainty can be formalized as a Partially Observable Markov Decision Process (POMDP)

[1] Wang et al., "DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs"; Deglurkar et al., "Compositional Learning-based Planning for Vision POMDPs".

Introduction

- Planning under uncertainty can be formalized as a Partially Observable Markov Decision Process (POMDP)
- Optimally solving POMDPs is computationally expensive and feasible only for small tasks

[1] Wang et al., "DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs"; Deglurkar et al., "Compositional Learning-based Planning for Vision POMDPs".

Introduction

- Planning under uncertainty can be formalized as a Partially Observable Markov Decision Process (POMDP)
- Optimally solving POMDPs is computationally expensive and feasible only for small tasks
- Visual observations are complex to model in planning[1]

[1] Wang et al., "DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs"; Deglurkar et al., "Compositional Learning-based Planning for Vision POMDPs".

Introduction

- Planning under uncertainty can be formalized as a Partially Observable Markov Decision Process (POMDP)
- Optimally solving POMDPs is computationally expensive and feasible only for small tasks
- Visual observations are complex to model in planning[1]
- Learned observation models are impractical for solving the POMDP in real-time

[1] Wang et al., "DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs"; Deglurkar et al., "Compositional Learning-based Planning for Vision POMDPs".

Contribution

- We explore planning with a simpler observation model while attaining formal guarantees of the solution quality

Contribution

- We explore planning with a simpler observation model while attaining formal guarantees of the solution quality
- Potential of substantial computational improvement for complex models

Contribution

- We explore planning with a simpler observation model while attaining formal guarantees of the solution quality
- Potential of substantial computational improvement for complex models
- Our main contributions:
  - Bound the theoretical loss with observation model discrepancy
  - Probabilistic bound for the empirical simplified performance
  - Practical computation of the bounds in SOTA planners

Continuous POMDP Solvers

- POMCPOW is a SOTA continuous POMDP solver [2]

Algorithm 2 POMCPOW

procedure SIMULATE(s, h, d)
  if $d = 0$ **then**
    return 0
  $a \leftarrow$ ACTIONPROGWIDEN(h)
  $\boxed{s', o, r \leftarrow G(s, a)}$
  if $|C(ha)| \leq k_o N(ha)^{\alpha_o}$ **then**
    $M(hao) \leftarrow M(hao) + 1$
  else
    $o \leftarrow$ select $o \in C(ha)$ w.p. $\frac{M(hao)}{\sum_o M(hao)}$
  append s' to $B(hao)$
  $\boxed{\text{append } \mathcal{Z}(o \mid s, a, s') \text{ to } W(hao)}$
  if $o \notin C(ha)$ **then** ▷ new node
    $C(ha) \leftarrow C(ha) \cup \{o\}$
    $total \leftarrow r + \gamma$ROLLOUT$(s', hao, d-1)$
  else
    $s' \leftarrow$ select $B(hao)[i]$ w.p. $\frac{W(hao)[i]}{\sum_{j=1}^{m} W(hao)[j]}$
    $r \leftarrow R(s, a, s')$
    $total \leftarrow r + \gamma$SIMULATE$(s', hao, d-1)$
  $N(h) \leftarrow N(h) + 1$
  $N(ha) \leftarrow N(ha) + 1$
  $Q(ha) \leftarrow Q(ha) + \frac{total - Q(ha)}{N(ha)}$
  return $total$

[2] Sunberg and Kochenderfer, "Online algorithms for POMDPs with continuous state, action, and observation spaces"

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

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

Simplfying the Observation Model

- We replace p_Z with a cheaper model q_Z
- Simplified Action-value function: $Q^{q_Z}_{\mathbf{P}}$

	Theoretical Values	PB-MDP	Planner
Original Model	$Q^{p_Z}_{\mathbf{P}}$	$Q^{p_Z}_{\mathbf{M_P}}$	$\hat{Q}^{p_Z}_{\mathbf{M_P}}$
Simplified Model	$Q^{q_Z}_{\mathbf{P}}$	$Q^{q_Z}_{\mathbf{M_P}}$	$\hat{Q}^{q_Z}_{\mathbf{M_P}}$

Simplfying the Observation Model

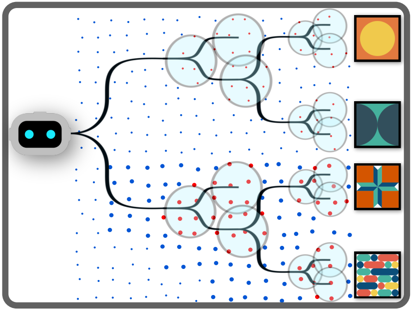
- We replace p_Z with a cheaper model q_Z
- Simplified Action-value function: $Q^{q_Z}_{\mathbf{P}}$

	Theoretical Values	PB-MDP	Planner
Original Model	$Q^{p_Z}_{\mathbf{P}}$	$Q^{p_Z}_{\mathbf{M_P}}$	$\hat{Q}^{p_Z}_{\mathbf{M_P}}$
Simplified Model	$Q^{q_Z}_{\mathbf{P}}$	$Q^{q_Z}_{\mathbf{M_P}}$	$\hat{Q}^{q_Z}_{\mathbf{M_P}}$

- Can we bound $|Q^{p_Z}_{\mathbf{P}} - \hat{Q}^{q_Z}_{\mathbf{M_P}}|$?

Approach to Bounds

- Pre-sample states at which we compute "observation model discrepancy"
- During online, we weight these states according to their likelihood
- We prove convergence guarantees for our estimated bounds



State-Dependent Observation TV-Distance

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

State-Dependent Observation TV-Distance

- Obs. TV-Distance: $\Delta_Z(x) \triangleq \int_{\mathcal{Z}} |p_Z(z \mid x) - q_Z(z \mid x)| \, \mathrm{d}z$
- $m_i(x_i, a) \triangleq V_{i+1}^{\max} \cdot \mathbb{E}_{x_{i+1} \sim p_T(\cdot \mid x_i, a)}[\Delta_Z(x_{i+1})]$

State-Dependent Observation TV-Distance

- Obs. TV-Distance: $\Delta_Z(x) \triangleq \int_{\mathcal{Z}} |p_Z(z \mid x) - q_Z(z \mid x)| \, \mathrm{d}z$
- $m_i(x_i, a) \triangleq V_{i+1}^{\max} \cdot \mathbb{E}_{x_{i+1} \sim p_T(\cdot|x_i, a)}[\Delta_Z(x_{i+1})]$
- $m_i(b_i, a) \triangleq \mathbb{E}_{x_i \sim b_i}[m_i(x_i, a)]$

State-Dependent Observation TV-Distance

- Obs. TV-Distance: $\Delta_Z(x) \triangleq \int_{\mathcal{Z}} |p_Z(z \mid x) - q_Z(z \mid x)| \, \mathrm{d}z$
- $m_i(x_i, a) \triangleq V_{i+1}^{\max} \cdot \mathbb{E}_{x_{i+1} \sim p_T(\cdot \mid x_i, a)}[\Delta_Z(x_{i+1})]$
- $m_i(b_i, a) \triangleq \mathbb{E}_{x_i \sim b_i}[m_i(x_i, a)]$
- It is natural to define cumulative bound function

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

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

TV-Distance Loss Bounds

Theorem 2

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

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

TV-Distance Loss Bounds

Theorem 2

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

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

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

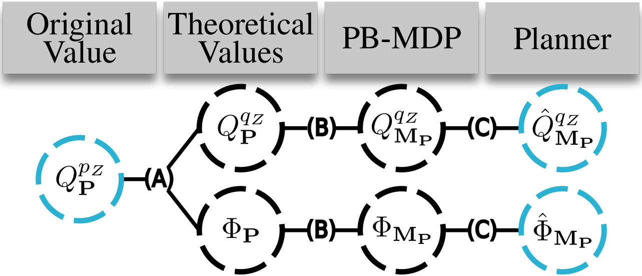
Empirical Concentration Inequalities

Corollary 3

For arbitrary $\varepsilon, \delta > 0$ there exists a number of particles for which

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

with probability of at least $1 - \delta$ for any guaranteed planner



- (A) is given by Theorem 2, (B) is given by Theorem 3, (C) is given by any planner with performance guarantees

Practical Computation of Bounds

- Computation of m_i is impractical
  - Importance Sampling
  - Separate calculations to offline/online

Online

$$\tilde{m}_i(x_i, a) \triangleq V_{i+1}^{\max} \frac{1}{N_\Delta} \sum_{i=1}^{N_\Delta} \frac{p_T(x_n^\Delta | x_i, a)}{Q_0(x_n^\Delta)} \Delta_Z(x_n^\Delta)$$

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

Offline

Practical Computation of Bounds

- Computation of m_i is impractical
  - Importance Sampling
  - Separate calculations to offline/online

Online

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

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

Offline

- Optimizations:
  - Considering state-samples based on a KD-Tree and a truncation distance
  - Computing a Monte Carlo estimate of $\tilde{m}_i$.

Practical Computation of Bounds

- Computation of m_i is impractical
  - Importance Sampling
  - Separate calculations to offline/online

Online

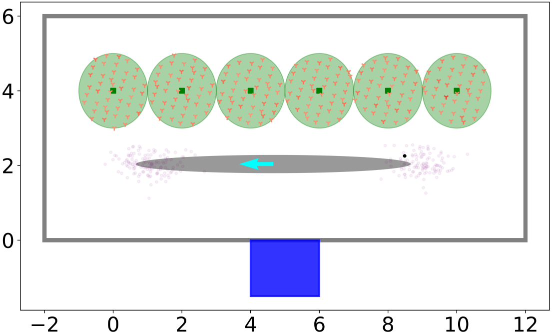
$$\tilde{m}_i(x_i, a) \triangleq V_{i+1}^{\max} \frac{1}{N_\Delta} \sum_{i=1}^{N_\Delta} \frac{p_T(x_n^\Delta | x_i, a)}{Q_0(x_n^\Delta)} \Delta_Z(x_n^\Delta)$$

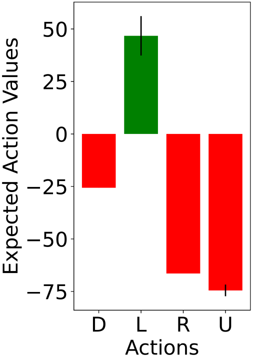
$$\{x_n^\Delta\}_{n=1}^{N_\Delta} \sim Q_0(x)$$

Offline

- Optimizations:
  - Considering state-samples based on a KD-Tree and a truncation distance
  - Computing a Monte Carlo estimate of $\tilde{m}_i$.
- In the paper we discuss embedding $\tilde{m}_i$ into POMDP solvers

Results in Simulation





- We show in our simulative setup that even with bounds calculation we achieve a significant speedup

Deglurkar, Sampada et al. "Compositional Learning-based Planning for Vision POMDPs". In: *Learning for Dynamics and Control Conference*. PMLR. 2023, pp. 469–482.

Sunberg, Zachary and Mykel Kochenderfer. "Online algorithms for POMDPs with continuous state, action, and observation spaces". In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 28. 1. 2018.

Wang, Yunbo et al. "DualSMC: Tunneling Differentiable Filtering and Planning under Continuous POMDPs". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 4190–4198. DOI: 10.24963/ijcai.2020/579. URL: https://doi.org/10.24963/ijcai.2020/579.