Set-membership Belief State-based Reinforcement Learning for POMDPs

Wei Wei, Lijun Zhang, Lin Li, Huizhong Song, and Jiye Liang

School of Computer and Information Technology, Shanxi University,
Taiyuan 030006, P.R. China

Overview

1 Abstract

2 Preliminary

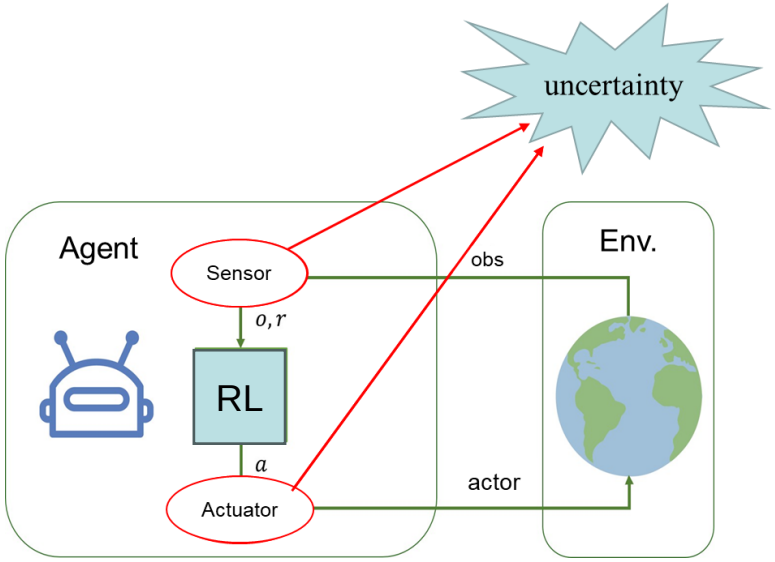
3 Proposed Method

4 Experiment Results

Abstract

Background:

- **Real-world sequential decision-making tasks are very challenging.**
  - In many real-world sequential decision-making tasks, the observation data could be noisy or incomplete due to the intrinsic low quality of the sensors or unexpected malfunctions; that is, the agent's perceptions are rarely perfect.



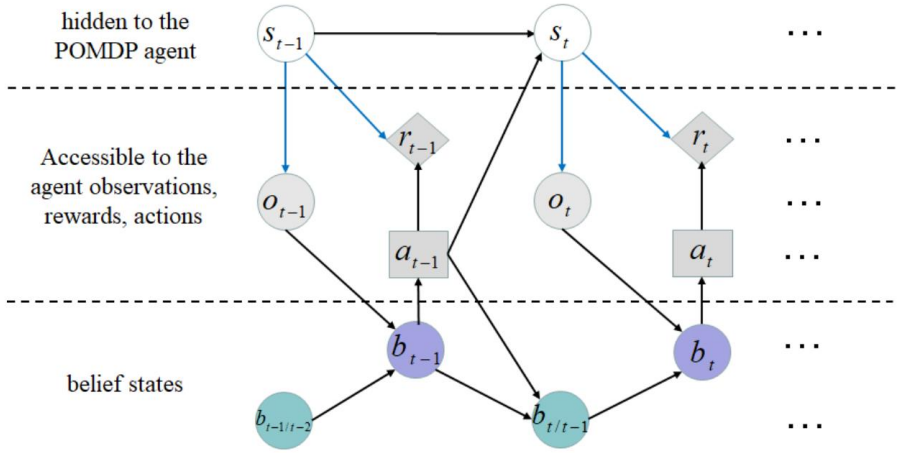
Motivation:

- **Avoid decision errors due to probability estimation.**
  - The current POMDP RL methods, such as particle-based and Gaussian-based, can only provide a probability estimate of hidden states, which may lead to inefficient and even wrong decision-making.

Main contribution:

- We propose a set-membership belief state-based reinforcement learning algorithm to solve POMDP.
- We prove that our belief estimation method can provide a series of belief state sets that always contain the true states under the unknown-but-bounded (UBB) noise.
- Extensive experiments on benchmark tasks show that our SBRL algorithm performs well.

Preliminary



The (belief inference graphical model)BIGM of POMDP. The white circles represent the unobservable hidden states s; the grey icons represent observations o, rewards r are accessible, and the agent determines the actions a; the green and purple circles represent the belief states obtained through inference.

Set-membership filtering has significant advantages in the following two POMDP scenarios:

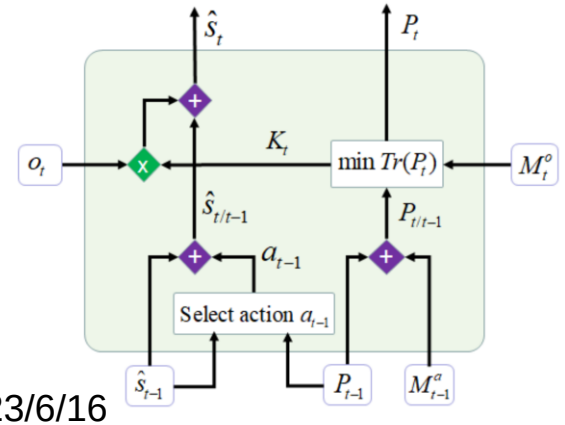
- **Unknown-but-bounded(UBB) noise:** In this situation, sensor and state transition noise distributions are multi-modal and imprecise due to complex factors, making it impossible to model the noise accurately.
- **Safety-critical environment:** To meet the application requirements of safety-critical systems such as autonomous driving or robot control many safe RL works pursue agents to learn a zero-violation policy.

Proposed Method

To accurately describe the hidden state, we propose the Set-membership Belief state learning Model (SBM) to provide a series of belief state sets that always contain the true states under the UBB sequence noise. Specifically, the SBM model consists of the following components:

$$\begin{aligned} \textit{State transition model} &: s_t = T(s_{t-1}, a_{t-1}, \omega^a_{t-1}) \\ \textit{Observation model} &: o_t = Z(s_t, \omega^o_t) \\ \textit{Reward model} &: r_t = R(s_t). \end{aligned}$$

And the architecture of the set-membership filter is the following.



The result for the updated bounded belief and the existence conditions are developed, which are given in the following theorem.

Theorem 3.1. *If Equation (4), Equation (5), and Equation (6) hold, the updated bounded optimized belief set for the state s_t can be computed by solving the following semidefinite program (SDP) in the variables $P_t \geq 0, \tau_z \geq 0, \tau^o_\omega \geq 0, \hat{s}_t$*

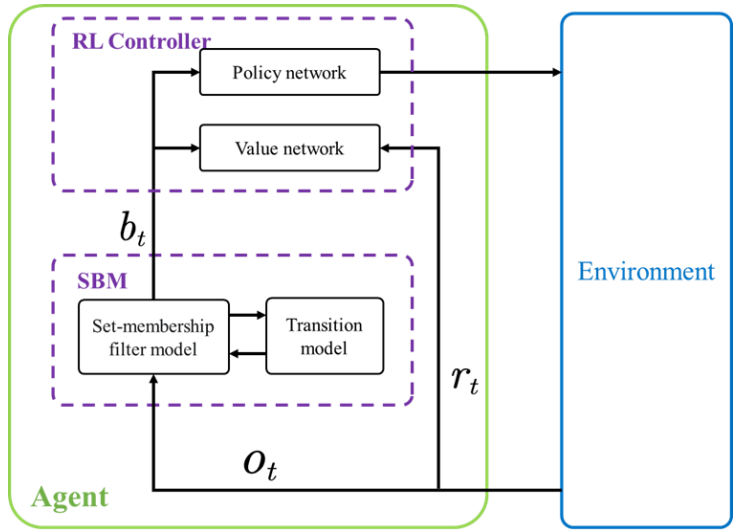
$$\begin{aligned} &\min \operatorname{Tr}(P_t) \\ &\textit{subject to } \tau_z \geq 0, \tau^o_\omega \geq 0 \end{aligned} \tag{8}$$

$$\begin{bmatrix} -P_t & \Phi_t \\ (\Phi_t)^{\mathrm{T}} & -\Pi_t \end{bmatrix} \leq 0, \tag{9}$$

where $\Phi_t = [-K_t\hat{s}_{t/t-1} \quad (I-K_t)E_{t/t-1} \quad -K_t]$, $\Pi_t = \operatorname{diag}\left(1-\tau_z-\tau^o_\omega, -\tau_z I, -\tau^o_\omega (M^o_t)^{-1}\right)$, and I is the identity matrix with appropriate dimensions.

The we integrate SBM into a POMDPs RL framework and propose the Set-membership Belief-based Reinforcement Learning (SBRL) algorithm, which can be trained jointly. The overall loss function is

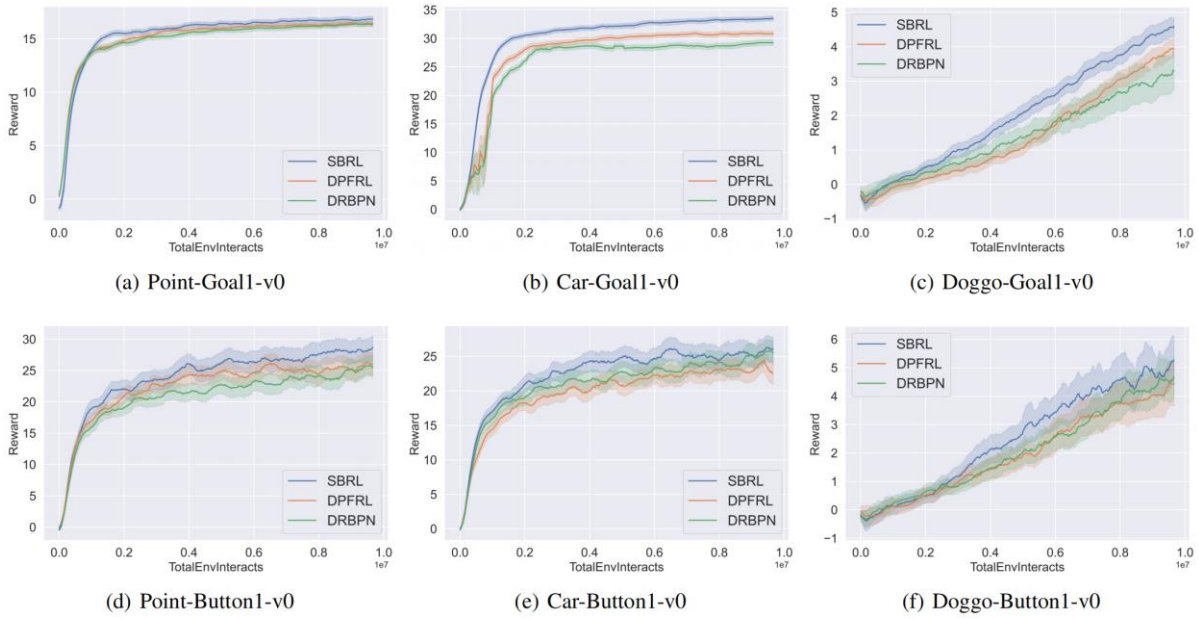
$$L^{\mathrm{SBRL}}(\zeta, \xi, \psi) = -L^p(\zeta) + \lambda_v L^v(\xi) - \lambda_m L^m(\psi)$$



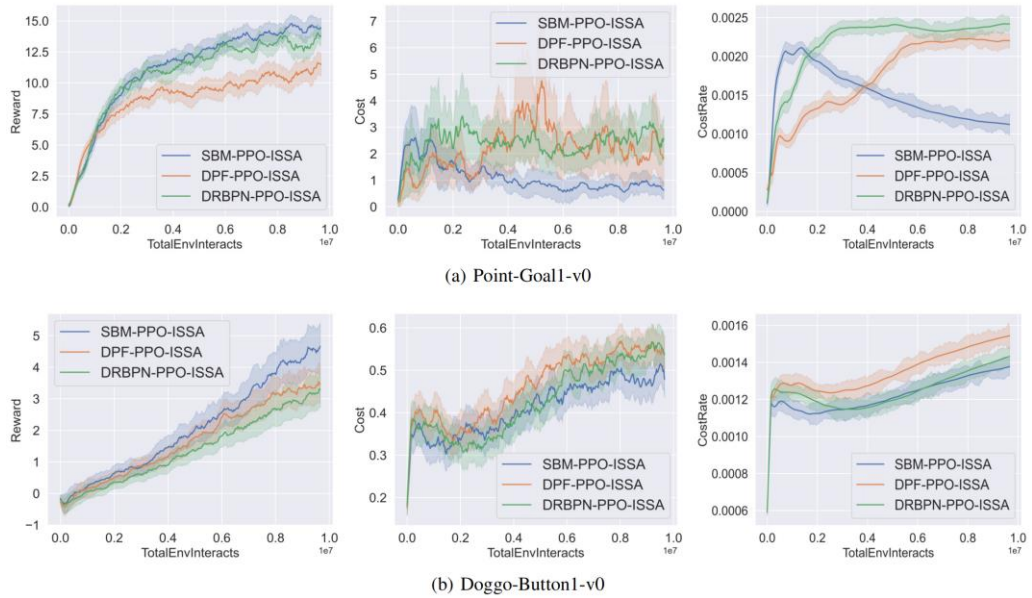
Overview of SBRL. The SBRL consists of two parts: a Set-membership Belief state learning Model (SBM) for learning bounded belief state sets and an RL controller for making decisions based SBM.

Experiment Results

We empirically evaluate our method for several challenging control tasks. The main experimental results are shown as follows.



Safe Gym with noise. Average episodic return of SBRL and baseline methods in the 6 benchmark environments. The curves are smoothed uniformly for better visualization.



Ablation Studies. The first row is the average episodic return, episodic cost, and overall cost rate on Point-Goal1-v0; the second row is the average episodic return, episodic cost, and overall cost rate on Doggo-Button1-v0.

Q & A

Welcome to join the discussion!