Set-membership Belief State-based Reinforcement Learning for POMDPs

Wei Wei [1] **Lijun Zhang** [1] **Lin Li** [1] **Huizhong Song** [1] **Jiye Liang** [1]

Abstract

Reinforcement learning (RL) has made significant progress in areas such as Atari games and robotic control, where the agents have perfect sensing capabilities. However, 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. The current POMDP RL methods, such as particle-based and Gaussian-based, can only provide a probability estimate of hidden states rather than certain belief regions, which may lead to inefficient and even wrong decision-making. This paper proposes a novel algorithm called Set-membership Belief state-based Reinforcement Learning (SBRL), which 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 on SBM. 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. The effectiveness of the proposed method is verified on a collection of benchmark tasks, and the results show that our method outperforms the state-of-the-art methods.

1. Introduction

Boosted by advanced and rapid developments of reinforcement learning (RL), sequence decisions in stochastic environments have made considerable progress and have been found in many applications, such as Atari games (Mnih et al., 2015), autonomous navigation (Wang et al., 2019), and robotics control (Kurniawati, 2021). However, most remarkable improvements in areas such as autonomous driving or robotics control with such powerful technology are limited to simulation platforms, which have yet not been used in widespread real-world applications (Haarnoja et al., 2018; Richter et al., 2019; Chen et al., 2021a). One of the critical issues in real-world applications is that the agent may have to decide under uncertain or partially observable (i.e., the agent's observations could be noisy or incomplete, which cannot accurately represent the complete state). Thus the decisions based on the observations are likely wrong. In practical tasks, it is ubiquitous that the agents often have to take action continuously, even when their observations of the environment are noisy or incomplete (Somani et al., 2013; Hausknecht & Stone, 2015; Ni et al., 2022). Take autonomous driving as an example. A running autonomous driving car must keep moving despite its radar perception sensors being noisy or partially occluded rather than stop abruptly itself and then move again. In such cases, ignoring the imperfect state information or setting specific action selection rules is inappropriate and could even bring disaster to the system. Consequently, it is necessary to research the decision-making problem with noisy and incomplete observations.

Generally, this type of decision-making problem is usually modeled as partially observable Markov decision processes (POMDPs) (Åström, 1965; Kurniawati et al., 2008; Somani et al., 2013), which is suitable for scenarios where an agent cannot accurately observe the complete state of the environment. There are two main approaches for POMDPs in the existing literature. The first type uses a recurrent neural network (RNN) as a function approximator to learn the representation of the hidden states from the state-transition data (Hausknecht & Stone, 2015; Zhu et al., 2017; Chen et al., 2022). Despite these RNN-based methods being simple and, to some extent, practical, they are often likely suboptimal in complex tasks due to performing inference implicitly requiring a known or learned model. The second type is the belief inference approach, which can characterize the uncertainty of the current hidden state by introducing particle filters, diagonal Gaussians, or other technologies. Compared to the first type, this method can explicitly characterize the uncertainty of the knowledge about the current hidden state. To mention a few, DVRL (Igl et al., 2018) and DPFRL (Ma et al., 2020) introduce particle filter-based methods that use

*Equal contribution [1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006. PR. China. Correspondence to: Jiye Liang <liy@sxu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

sampling particles to approximate the belief states. Still, particle filters are reported to experience the curse of dimensionality and therefore suffer from low sample efficiency and performance (Lee et al., 2020). In addition, many researchers use the Gaussian model (Han et al., 2019; Wang & Tan, 2021) to construct the dynamics and generative model and obtain the current posterior belief state. Unfortunately, due to the reliance on Gaussian assumptions, this approach may lead to poor performance when detailed statistical information is not accurately available or the environmental noise is non-Gaussian. Furthermore, among the particle-based or Gaussian-based probabilistic methods, the common drawback is that they can only provide a probability estimate of hidden states rather than a certain belief region. However, in many real-world applications, accurate belief estimation is crucial since it is the cornerstone of the agent to make the right decision, which motivates us to investigate a new POMDP RL method based on a bounded belief state set.

To address the sequences decision-making problem under uncertain or partially observable, we propose a novel POMDP RL algorithm called Set-membership Belief state-based Reinforcement Learning (SBRL). The algorithm consists of a Set-membership Belief state learning Model (SBM) for learning bounded belief state sets and an RL controller for making decisions based on SBM. We also demonstrate the results on several challenging control tasks, showing that our SBRL algorithm outperforms the state-of-the-art methods under challenging POMDP scenarios.

To summarize the main contributions of this paper can be summarized as follows:

- We propose a set-membership belief state-based reinforcement learning algorithm to solve POMDP tasks by training a set-membership belief state learning model (SBM) and an RL controller network.
- 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 outperforms the state-of-the-art methods under various challenging POMDP scenarios. SBM allows the agent to provide a reasonable basis for the agent to make good decisions.

2. Preliminaries

2.1. POMDPs

A POMDPs can be described as an 8-tuple $< \mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, r, \gamma, b_0 >$, where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{O}$ represent the set of state, action, and observation spaces, respectively. T is a set of conditional transition functions between states.

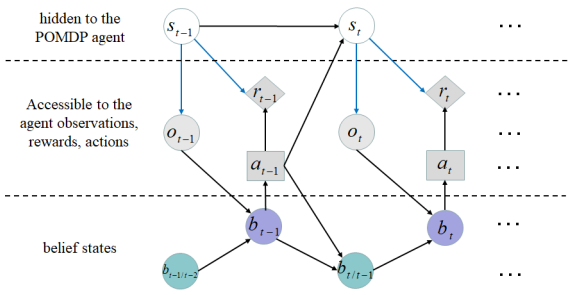


Figure 1. The BIGM of POMDP. The white circles represent the unobservable hidden states s; the grey icons represent observations o, rewards r are accessible, and the agent determines the actions a; the green and purple circles represent the belief states obtained through inference.

Z is a set of conditional observation functions. In addition, $r : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ represents the reward function and outputs a scalar. $\gamma \in (0, 1]$ is the discount factor, and b_0 is the initial belief set.

At each timestep $t-1$, the state of the environment is $s_{t-1} \in \mathcal{S}$. The agent take an action $a_{t-1} \in \mathcal{A}$, which causes the environment to transit to state s_t with $T(s_t|s_{t-1}, a_{t-1})$. The agent then receives an observation $o_t \in \mathcal{O}$, which depends on the new state s_t with $Z(o_t|s_t)$. The agent's goal is to maximize the expected sum of discounted rewards $\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k\right]$. Such a POMDP model can be described using a belief inference graphical model (BIGM), as shown in Figure 1. After taking action a_{t-1} and obtaining observation o_t, the agent needs to update its belief state, defined as a bounded set containing the true state.

2.2. Set-membership Filter

It is more desirable to use a bounded region containing all possible hidden states instead of using the particle-based or Gaussian-based probabilistic methods to approximate the prior and posterior belief distributions. Therefore, we introduce the set-membership filter (Witsenhausen, 1968; Calafiore, 2005; Yang & Li, 2009) to estimate the hidden state.

Rather than needing the statistics of the distribution itself, the set-membership filter method can provide a series of bounded regions, which guarantees to contain the true state of the system when one obtains the amplitude boundary. More specifically, it limits the boundary of possible hidden states by identifying the noise amplitude, i.e.,

$$\mathcal{W}_t^a = \left\{\omega_t^a : (\omega_t^a)^{\mathrm{T}} (M_t^a)^{-1} \omega_t^a \leq 1\right\},$$

$$\mathcal{W}_t^o = \left\{\omega_t^o : (\omega_t^o)^{\mathrm{T}} (M_t^o)^{-1} \omega_t^o \leq 1\right\},$$

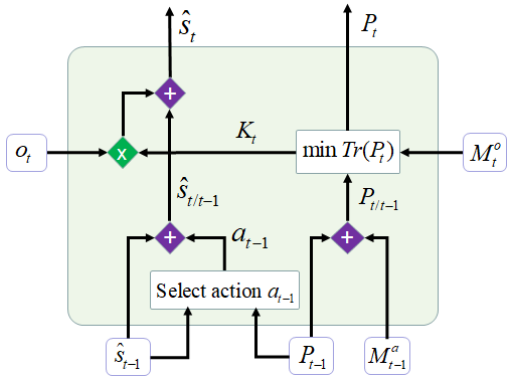


Figure 2. Architecture of set-membership filter.

where ω_t^a and ω_t^o are the noise vectors. $M_t^a = (M_t^a)^{\mathrm{T}} > \mathbf{0}$ and $M_t^o = (M_t^o)^{\mathrm{T}} > \mathbf{0}$ are known matrices with compatible dimensions, which physical considerations sensors and actuators of the agent can obtain. This noise description is often more realistic than a probabilistic description in many applications.

3. The Proposed Method

This section proposes a novel POMDP RL algorithm called Set-membership Belief state-based Reinforcement Learning (SBRL). Similar to prior methods for POMDPs (Igl et al., 2018; Ma et al., 2020; Wang & Tan, 2021), SBRL adopts the belief inference scheme to maintain a belief of the state based on past trajectory data and propagates the belief state recurrently during the execution phase. The difference is that our SBRL algorithm consists of a Set-membership Belief state learning Model (SBM) and an RL controller, in which the SBM can provide a bounded belief region rather than a probability estimate of hidden states. In addition, we prove that our belief estimation method can give a series of belief state sets that always contain the true states under the UBB sequence noise.

3.1. Set-membership Belief state learning Model

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:

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

In addition, we have a belief inference model based on the set-membership filter to depict the true state, and the belief learning approach for a single timestep is as follows.

At the time $t-1$, the hidden state s_{t-1} is assumed to be satisfied the condition

$$s_{t-1} = \hat{s}_{t-1} + E_{t-1}z, ||z|| \leq 1, \tag{2}$$

where $\hat{s}_{t-1}$ is an estimate of s_{t-1} , and E_{t-1} is a matrix used to describe the estimated range. The belief state at time $t-1$ is denoted by $b\left(\hat{s}_{t-1}, diag\{E_{t-1}\}\right)$, abbreviated as b_{t-1}. Then, the Equation (2) can be described in the form of a bounded set as follows:

$$\begin{aligned} s_{t-1} &\in \epsilon(\hat{s}_{t-1}, P_{t-1}) \\ &= \left\{ s_{t-1} : (s_{t-1} - \hat{s}_{t-1})^{\mathrm{T}} P_{t-1}^{-1} (s_{t-1} - \hat{s}_{t-1}) \leq 1 \right\}, \end{aligned} \tag{3}$$

where the shape matrix $P_{t-1} = E_{t-1}E_{t-1}^{\mathrm{T}} > 0$. Then, the agent selects an action at $a_{t-1} \in \mathcal{A}$ based on $\pi_\phi(b_{t-1})$.

In what follows, a prior belief state $b_{t/t-1}$ at timestep t can be directly obtained through a state transition model $s_t = \hat{T}(s_{t-1}, a_{t-1}, \omega_{t-1}^a)$, which modeled by a neural network with parameter ϕ. The ω_{t-1}^a is an additive noise that is confined to a specified bounded set

$$\mathcal{W}_{t-1}^a = \{\omega_{t-1}^a : (\omega_{t-1}^a)^{\mathrm{T}} (M_{t-1}^a)^{-1} \omega_{t-1}^a \leq 1\}, \tag{4}$$

where $M_{t-1}^a = (M_{t-1}^a)^{\mathrm{T}} \geq 0$ is a known matrix with compatible dimensions.

Next, an imputation and filtering operation on between the new observation o_t and $b_{t/t-1}$ is applied, which is of the follow form:

$$\hat{s}_t = \hat{s}_{t/t-1} + K_t o_t, \tag{5}$$

where ω_t^o is the observation noise which is confined to a specified bounded set

$$\mathcal{W}_t^o = \{\omega_t^o : (\omega_t^o)^{\mathrm{T}} (M_t^o)^{-1} \omega_t^o \leq 1\}, \tag{6}$$

where $M_{t-1}^o = (M_{t-1}^o)^{\mathrm{T}} \geq 0$ is a known matrix with compatible dimensions and K_t is a filter parameter to be determined.

In formulas, we want to compute the state estimate $\hat{s}_t$ and the shape matrix P_t with the smallest trace at timestep t, such that the condition

$$(s_t - \hat{s}_t)^{\mathrm{T}} P_t^{-1} (s_t - \hat{s}_t) \leq 1 \tag{7}$$

holds for any ω_t^o obeying Equation (6). Subsequently, 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 $P_t \geq 0, \tau_z \geq 0, \tau_\omega^o \geq 0, \hat{s}_t$

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

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

where $\Phi_t = [-K_t\hat{s}_{t/t-1} \quad (I-K_t)E_{t/t-1} \quad -K_t]$, $\Pi_t = \mathrm{diag}\left(1-\tau_z-\tau_\omega^o, -\tau_z I, -\tau_\omega^o (M_t^o)^{-1}\right)$*, and* I *is the identity matrix with appropriate dimensions.*

The proof of this theorem is reported in Appendix B.

The architecture of the set-membership filter is demonstrated in Figure 2. Computing the optimal solution to the optimization problem in Theorem 3.1 essentially requires $O(n^3)$ operations, which is challenging to be solved by the existing optimization toolbox in the high dimensional state space. Thus, we also provide a parameterized calculation method to make model training more convenient and practical.

Firstly, according to the Equation (1) and Equation (5), we can deduce it as follows

$$\begin{aligned} s_t &= \hat{s}_t + E_t z \\ &= \hat{s}_{t/t-1} + K_t o_t + E_{t/t-1} z + K_t \omega_t^o \\ &= \hat{s}_{t/t-1} + K_t(\hat{s}_{t/t-1} + E_{t/t-1} z + \omega_t^o) \\ &\quad + E_{t/t-1} z + K_t \omega_t^o \\ &= (I+K_t)\hat{s}_{t/t-1} + (I+K_t)E_{t/t-1} z + 2K_t\omega_t^o. \end{aligned} \tag{10}$$

Then, by using Lemma A.1 and Lemma A.2 (the details are reported in Appendix A), we can draw

$$\begin{aligned} s_t &= (I+K_t)\hat{s}_{t/t-1} + (I+K_t)E_{t/t-1} z + 2K_t\omega_t^o \\ &\in \epsilon\left((I+K_t)\hat{s}_{t/t-1}, (I+K_t)^2 P_{t/t-1}\right) \oplus \epsilon\left(0, 4K_t^2 M_t^o\right) \\ &\subseteq \epsilon\left(\hat{s}_t, P_t\right), \end{aligned} \tag{11}$$

where $\hat{s}_t = (I+K_t)\hat{s}_{t/t-1}$, $P_t = (1-\eta)^{-1}(I+K_t)^2 P_{t/t-1} + 4\eta^{-1}K_t^2 M_t^o$, and $\eta \in [0,1]$ is a scalar parameter.

As a result, the new belief $b\left(\hat{s}_t, diag\{E_t\}\right)$ is obtained. Finally, the belief b_t is employed by the policy π to decide an action

$$\pi(b_t) \doteq \pi\left(\hat{s}_t, diag\{E_t\}\right). \tag{12}$$

Remark 3.2. Different from the probability-based belief state distribution, the belief state obtained based on set-membership filtering are formalized as a strictly bounded region of the estimated midpoint and shape matrix. It has significant advantages in the following two scenarios:

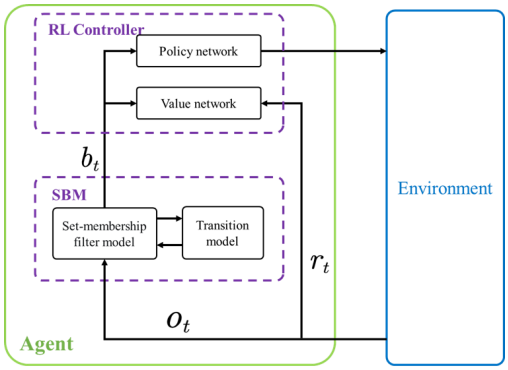


Figure 3. Overview of SBRL. The SBRL consists of two parts: a Set-membership Belief state learning Model (SBM) for learning bounded belief state sets and an RL controller for making decisions based SBM.

- Unknown-but-bounded(UBB) noise: In this situation, sensor and state transition noise distributions are multimodal 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(Ma et al., 2021; Chen et al., 2021b) or robot control(Zhao et al., 2021), many safe RL works pursue agents to learn a zero-violation policy. Our approach also has significant implications for safe RL under partially observable, which is an inevitable challenge when RL is geared toward real-world applications.

3.2. POMDP RL Framework based on SBM

To show the advantage of SBM in POMDP tasks, we integrate SBM into a POMDPs RL framework and propose the Set-membership Belief-based Reinforcement Learning (SBRL) algorithm based PPO (Schulman et al., 2017) to learn the optimal policy. The detail of SBRL can be found in Appendix C, and the overview of SBRL is shown in Figure 3.

In our algorithm, both actor and critic make decisions based on the bounded belief sets $b\left(\hat{s}_t, diag\{E_t\}\right)$ obtained by the SBM rather than the probabilistic belief state distribution. The algorithm may reduce the impact of the inaccuracy observation and decrease the precision requirements of sensors.

In addition, we construct an observation generation model to predict future observations in the decision-making process, which be represented by recurrent neural networks parameterized by $q_\theta(s_t|o_{\leq t}, a_{\leq t})$. For simplification, let $\psi \doteq (\phi, \theta, K, \eta)$ denote all the parameters to be learned for

Algorithm 1 SBRL algorithm

Input: Initial belief $b\left(\hat{s}_1, diag\{E_1\}\right)$
Parameter: Buffer $\mathcal{B}$, interacting step T, max epoches N

while not converged **do**
  // Execution Phase
  Start with $b\left(\hat{s}_1, diag\{E_1\}\right)$
  Receive an observation o_1 from the environment
  for $t = 1$ to T **do**
    Select action a_t based on (12)
    Receive a reward r_t and a new observation o_{t+1}
    Update the next belief state $b\left(\hat{s}_{t+1}, diag\{E_{t+1}\}\right)$ by (5) and (8)
    Add experience to buffer $\mathcal{B} = \mathcal{B} \cup \{(o_t, a_t, r_t)_{t=1}^{T}\}$
    $b\left(\hat{s}_t, diag\{E_t\}\right) \leftarrow b\left(\hat{s}_{t+1}, diag\{E_{t+1}\}\right)$
  end for
  // Training Phase
  for $k = 1$ to T **do**
    Compute A^t and V^t for $t = 1 \dots T$
    Train the networks by (14) with $\{(o_t, a_t, r_t)_{t=1}^{T}\}$
  end for
end while

the belief computation. Then, the belief inference network can be jointly optimized by maximizing the Evidence Lower BOund (ELBO):

$$
\begin{aligned}
L^m(\psi) &= \log p(o_{1:T}|a_{1:T}) \\
&\geq \mathbb{E}_{q(s_{1:T}|o_{1:T},a_{1:T})}\left[\prod_{t=1}^{T} \log p(o_t|s_t)\right. \\
&\left. + \log T(s_t|s_{t-1}, a_{t-1}) - \log q(s_t|o_{\leq t}, a_{\leq t})\right].
\end{aligned} \tag{13}
$$

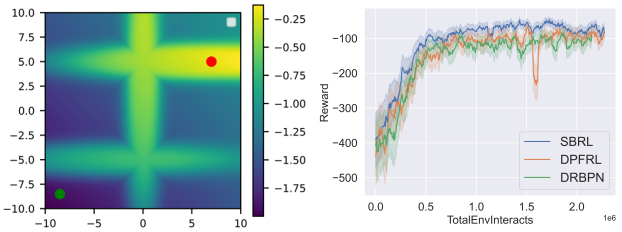
The detailed derivations can be found in Appendix D. All three networks, the policy network, the value network, and the belief inference network, can be trained jointly. The overall loss function is

$$
L^{\text{SBM}}(\zeta, \xi, \psi) = -L^p(\zeta) + \lambda_v L^v(\xi) - \lambda_m L^m(\psi), \tag{14}
$$

where λ_v and λ_m are the coefficients to trade-off the losses. We adopt the normalized advantage values and rewards to train policy and value network; therefore, all three terms have similar magnitude across different tasks. Our SBRL algorithm is presented in Algorithm 1.

4. Experiments

We empirically evaluate our method for several challenging control tasks in this section. Our experiments aim to answer the following questions: First, can SBRL algorithm achieve good results in both partially observable and uncertain environments? Second, can SBM maintain accurate belief states to provide a reasonable basis for agents' decision-making under uncertain and partially observable environments?



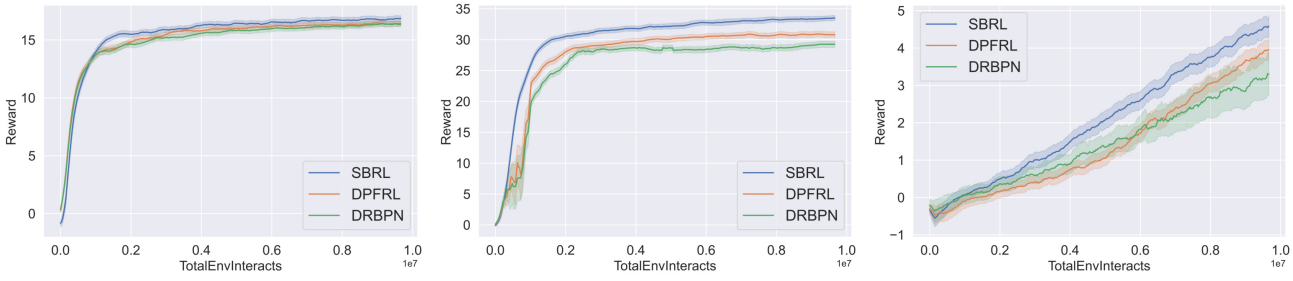
(a) Mountain Hike Task (b) Performance

Figure 4. Mountain Hike Task. An agent navigates on the map from the start position (white dot) to the goal (green dot with the shaded area as the threshold). Partial observation is introduced by a bounded random noise and appended with a long noise vector of length l. The reward $r(x, y)$ for position (x, y) is given by the heat map.

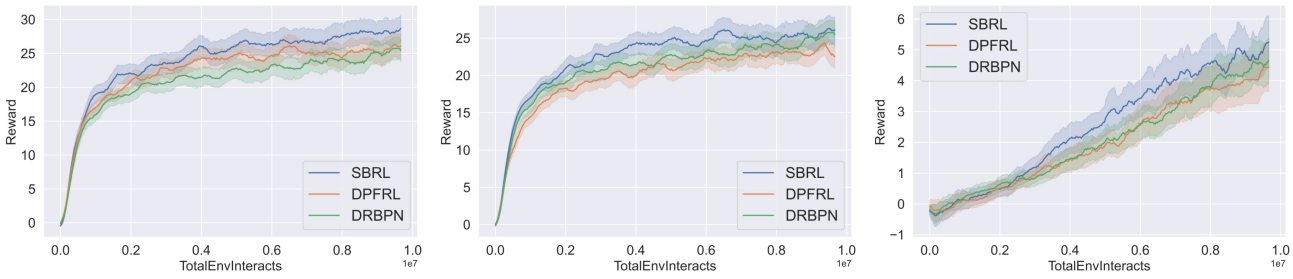
To answer the first question, we conduct experiments on Mountain Hike, several variants of Safety Gym, and Flickering Atari games. We train SBRL and baselines with similar network architecture and hyperparameters as the original DPFRL implementation. We compare our method against the following algorithms: 1) DPFRL (Ma et al., 2020), which performs belief inference relying on a particle filter and learns the environment models simultaneously; 2) DRBPN (Wang & Tan, 2021), which employs the Gaussian model to learn belief states. Concerning the second question, we give two ablation experiments on the variant of Safety Gym, which takes PPO-ISSA (Zhao et al., 2021) as the nominal policy because PPO-ISSA can make a Zero-Violation policy on the standard Safety Gym environment. All reported results are averages over three random seeds, and the curves are smoothed over time.

4.1. Mountain Hike

Experimental Setup : Mountain Hike is a continuous control environment with observation uncertainty where an agent navigates on a fixed 20×20 map, introduced by (Igl et al., 2018) to demonstrate the benefit of belief tracking for POMDP RL. (Ma et al., 2020) concatenates the original observation vector with a random noise vector to make the environment more challenging. The main difference between our environment setup and (Ma et al., 2020) is that both state-transition noise and observation noise are set to be bounded. More specifically, the state space and action space in Mountain Hike are defined as $\mathcal{S} = \mathcal{A} = \mathbb{R}^2$, where $s_t = [x_t, y_t]$ and $a_t = [\delta x_t, \delta y_t]$. Transitions of the agent are stochastic with an additive bounded random noise $s_{t+1} = s_t + a_t + \omega_t^a$, where $\omega_t^a \sim \mathcal{U}(-1, 1)$. The observation space is $\mathcal{O} = \mathbb{R}^{2+l}$, where l is a predefined



(a) Point-Goal1-v0 (b) Car-Goal1-v0 (c) Doggo-Goal1-v0



(d) Point-Button1-v0 (e) Car-Button1-v0 (f) Doggo-Button1-v0

Figure 5. 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.

Table 1. Max average return on Flickering Atari Games. The optimal value for each task is bolded. $\pm$ corresponds to a single standard deviation.

ENV	SBRL	DPFRL	DVRL
PONG	**18.72±0.83**	15.40±0.76	18.17±2.67
CHOPPER	8,002±139.1	**8,086±159.1**	6,602±449
MSPACMAN	**3,143±245.2**	3,028±545.3)	2,221±199
CENTIPEDE	4,429±191.4	**4,849±291.4**	4,240±116
BEAMRIDER	**4,120±99.4**	3,940±107.4	1,663±183
FROSTBITE	**298.5±6.18**	293.5±5.06	297±7.85
BOWLING	30.51±0.26	**33.89±0.34**	29.53±0.23
ICEHOCKEY	**-4.12±0.05**	-4.06±0.02	-4.88±0.17
DDUNK	-7.32±1.12	-11.25±1.25	**-5.95±1.25**
ASTEROIDS	1,892±153	**1,948±202.6**	1,539±73

constant. Observations are $o_t = [o_t^s, o_t^n]$, where $o_t^s = s_t + \omega_t^o$ and $o_t^n \subseteq \mathbb{R}^l$ is sampled from a uniform distribution $\mathcal{U}(-10, 10)$. The reward for each step is given by $r_t = r_(x_t, y_t) - 0.01\|a_t\|$, where $r_(x_t, y_t)$ is shown in Figure 4(a). The results are shown in Figure 4(b) when $l = 100$.

Evaluation : Figure 4(b) shows SBRL achieves superior performance and learns faster than the DPFRL and DRBPN, demonstrating the ability of SBRL under the observation uncertainty environment.

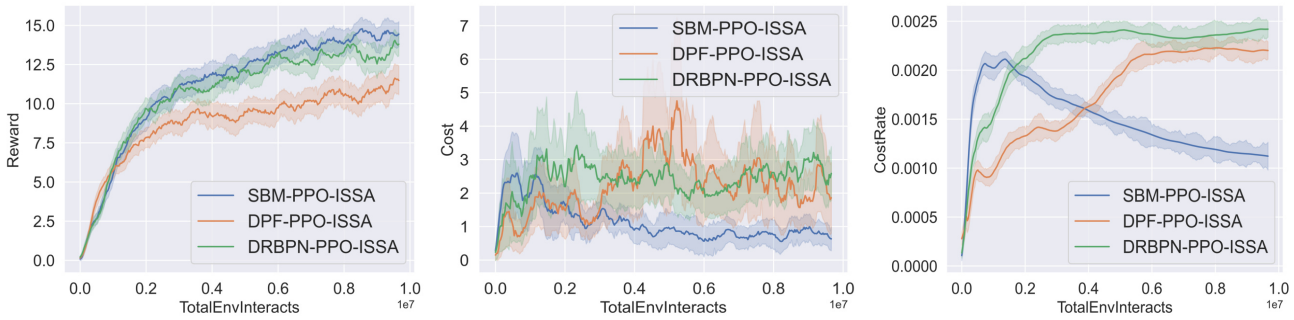
4.2. Safety Gym with noise

Experimental Setup : Safety Gym is a state-of-the-art high-dimensional continuous control environment (Ray et al., 2019), where an agent can only observe part of the environment state through its sensors. In our experiment, we choose 6 games from Safety Gym: Robot-Goal1-v0, Car-Goal1-v0, Doggo-Goal1-v0, Robot-Button1-v0, Car-Button1-v0, and Doggo-Button1-v0 to evaluate SBRL algorithm. It is worth noting that we added bounded random observation noise to the above benchmark environment: $o_t = s_t + \omega_t^o$, where ω_t^o is less than 15% of s_t. We conduct tests under different environments, and Figure 5 presents the experimental results.

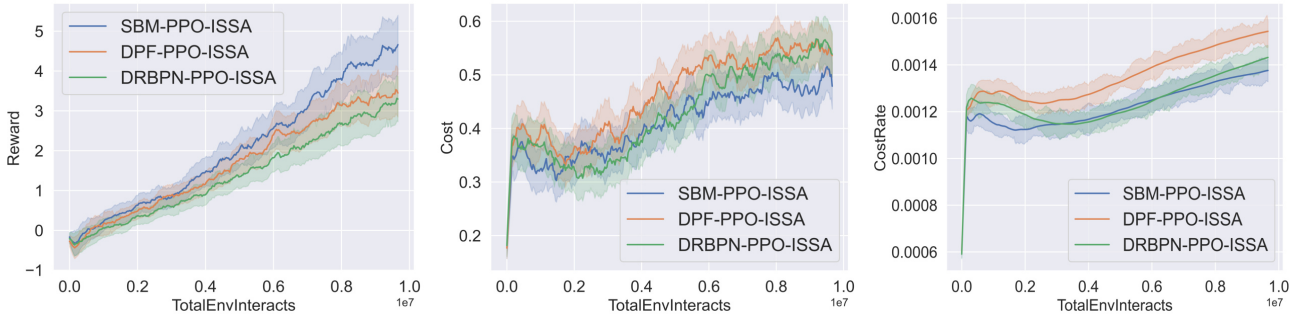
Evaluation : From Figure 5, we can see that compared to DPFRL and DVRL, SBRL can also achieve better performance under the partially observable environment with uncertainty. In addition, when the environment becomes complex, such as in Figure 5(c) and Figure 5(f), the advantages of the SBRL algorithm are more prominent. These results demonstrate the ability of SBRL in the robot navigation task with noise and incomplete observation.

4.3. Flickering Atari Games

Our algorithm is designed specifically for the decision problems under the environment with UBB noise, but consider-



(a) Point-Goal1-v0



(b) Doggo-Button1-v0

Figure 6. 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.

ing that many works (Hausknecht & Stone, 2015; Igl et al., 2018; Ma et al., 2020) were tested in Flickering Atari environments, which is another partially observable environment and does not violate the UBB noise setting. So, we added these experiments to show that our algorithm is still valid in the environment with the missing observations.

Experimental Setup : Flickering Atari Games are the variant of the Atari Games, and here image observations are single frames randomly replaced by a blank frame with a probability of 0:5. We test our algorithm on the same subset of games on which DPFRL and DVRL (Igl et al., 2018) were evaluated. The comparisons about max average return are summarized in Table 1.

Evaluation: From Table 1, we can see that SBRL and DPFRL significantly outperform DVRL in almost all games, and SBRL beats DPFRL in 6 out of 10 games (Pong, MsPacman, BeamRider, Frostbite, IceHockey, and DDunk) and performs comparably in 4 other games (Chopper, Centipede, Bowling, Asteroids). These results show that SBRL can also outperform or achieve performance similar to advanced algorithms, where the observation image of the agent is probabilistically missing. The simulation results are in line with our expectations.

4.4. Ablation Studies

We additionally test the performance of SBM in the Safety Gym with noise. In our experiment, using the PPO-ISSA agent as the nominal policy, which can realize zero-violation under the standard Safe Gym environments, and estimating the current belief state of the agent through SBM, this structure is called SBM-PPO-ISSA. Similarly, the comparison algorithms DPFRL and DRBPN adopt the same nominal strategy and are combined to form DPF-PPO-ISSA and DRBPN-PPO-ISSA, respectively. We conduct tests on Robot-Goal1-v0 and Doggo-Button1-v0 with noise. The experimental results are indicated in Figure 6, showing that SBM-PPO-ISSA has fewer constraint violations than DPF-PPO-ISSA and DRBPN-PPO-ISSA in all environments and gets slightly higher rewards than the other two methods. These experimental results demonstrate the ability of SBM to track a latent state, which is consistent with the above-mentioned analysis.

5. Related Work

5.1. POMDPs

POMDPs (Åström, 1965; Kurniawati et al., 2008; Somani et al., 2013) provide a principled and generic framework for modeling complex planning and decision problems in stochastic domains, which is suitable for scenarios where an agent cannot accurately observe the complete hidden state of the environment. Unfortunately, the modeling advantages of POMDPs come at cost-precise solutions that are computationally very expensive and thus only work in practice for elementary problems. Thus, in recent years, many researchers have tried to introduce the value function approximation method, such as point-based methods (Kurniawati et al., 2008; Shani et al., 2013) and Monte-Carlo sampling in the belief space (Silver & Veness, 2010; Kurniawati & Yadav, 2016), to solve approximate solutions of POMDPs. However, these studies only focus on the POMDP problems in discrete spaces. Subsequently, to track belief states in POMDPS with continuous state and action spaces, some works (Silver & Veness, 2010; Wu et al., 2021) use Monte Carlo algorithms like particle filters to maintain sample sets extracted from belief states. Other continuous space POMDP solvers often approximate the belief states with distributions like diagonal Gaussians (Lee et al., 2020), Gaussian mixture (Tschiatschek et al., 2018), and categorical distribution (Hafner et al., 2020) and solve the problem analytically using gradients.

5.2. Belief inference

Recent researches in model-based belief inference provide promising methods to deal with high-dimensional continuous control problems under partially observable or uncertain. Learning effective latent dynamics models to solve challenging continuous control problems is becoming feasible through advances in deep generative modeling and latent variable models. Among these, some works propose particle filter-based methods that use samples to approximate the belief states (Igl et al., 2018; Ma et al., 2020; Wu et al., 2021). However, particle filters are reported to experience the curse of dimensionality and therefore suffer from low sample efficiency and performance (Lee et al., 2020).

In addition, many researchers use the Gaussian model (Han et al., 2019; Wang & Tan, 2021) to construct the dynamics and generative model to obtain the current posterior belief state analytically. However, they assume the belief states that obey diagonal Gaussian distributions. Such assumptions impose substantial restrictions on belief inference and lead to limitations in practice, including mode collapse, posterior collapse, and object vanishing in reconstruction uses a Gaussian mixture to approximate the belief states. More recently, to learn general continuous belief states for POMDPs, FORBES (Chen et al., 2022) incorporates Normalizing Flows into the variational inference step to construct flexible belief states. However, this approach relies on iteratively applying the transformations at each time step to learn general belief state distribution, increasing computational complexity.

Specifically, different from the above method of probabilistic belief inference, we depart from the Bayesian approach and propose a new methodology for the POMDP task which requires no assumption on the noise statistics.

5.3. Set-membership filter

Set-membership filter (Granichin et al., 2021) serves as a well-appreciated robust filtering scheme, which ensures the true states are confined in some optimized region with high confidence at each time step, even in the presence of unknown-but-bounded process and observation noises. For the past two decades, many researchers have attempted to solve the set-membership filter problems with various methods. For example, a convex optimization approach was applied to deal with the robust set-membership filter for the systems with norm-bounded uncertainty in the system matrices (Calafiore, 2005). The base point and truncation errors are introduced by the linearization of nonlinear functions, which are respectively confined to bounded ellipsoids (Yang & Li, 2009). To guarantee satisfactory filtering performance, the authors introduce a parameter-dependent set-membership filter (Zou et al., 2021) to generate a time-varying ellipsoidal region containing the true state. The applications of the set-membership filter include the system over WSNs (Ding et al., 2020), photovoltaic grid-connected generation system (Zhang et al., 2020), and autonomous ground vehicles (Mousavinejad et al., 2021).

6. Conclusion

Uncertain and partially observable is a tremendous challenge for reinforcement learning when applied to real-world environments. In this paper, we propose a Set-membership belief state learning model to learn accurate belief states. In addition, we prove that our belief estimation method can provide a series of belief state sets that always contain the true states under the UBB noise. Furthermore, we integrate the SBM into a POMDP RL framework and propose a novel algorithm called Set-membership Belief state-based Reinforcement Learning (SBRL), which combines the strength of both the set-membership filter and end-to-end RL. Extensive experimental results show that the proposed method significantly outperforms the state-of-the-art methods under uncertain environments and is better or comparable to current state-of-the-art methods in partially observable environments.

Acknowledgements

This work were supported in part by the National Key Research and Development Program of China (under grant 2020AAA0106100), in part by the National Natural Science Project of China (under grant 62276160), and in part by the Basic Research Program of Shanxi Province (under grant No.202203021211294).

References

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. *Linear Matrix Inequalities in System and Control Theory*. SIAM, 1994.

Calafiore, G. Reliable localization using set-valued nonlinear filters. *IEEE Transactions on Systems, Man, and Cybernetics-part A: Systems and Humans*, 35(2):189–197, 2005.

Chen, J., Li, S. E., and Tomizuka, M. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2021a.

Chen, J., Li, S. E., and Tomizuka, M. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5068–5078, 2021b.

Chen, X., Mu, Y. M., Luo, P., Li, S., and Chen, J. Flow-based recurrent belief state learning for pomdps. In *ICML*, pp. 3444–3468, 2022.

Ding, D., Wang, Z., and Han, Q. A set-membership approach to event-triggered filtering for general nonlinear systems over sensor networks. *IEEE Transactions on Automatic Control*, 65:1792–1799, 2020.

Durieu, C., Polyak, B. T., and Walter, E. Trace versus determinant in ellipsoidal outer-bounding, with application to state estimation. *IFAC Proceedings Volumes*, 29(1): 3975–3980, 1996.

Granichin, O. N., Erofeeva, V., Ivanskiy, Y., and Jiang, Y. Simultaneous perturbation stochastic approximation-based consensus for tracking under unknown-but-bounded disturbances. *IEEE Transactions on Automatic Control*, 66: 3710–3717, 2021.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Han, D., Doya, K., and Tani, J. Variational recurrent models for solving partially observable control tasks. *arXiv preprint arXiv:1912.10703*, 2019.

Hausknecht, M. and Stone, P. Deep recurrent q-learning for partially observable mdps. In *AAAI*, 2015.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., and Whiteson, S. Deep variational reinforcement learning for pomdps. In *ICML*, pp. 2117–2126, 2018.

Kurniawati, H. Partially observable markov decision processes (pomdps) and robotics. *arXiv preprint arXiv:2107.07599*, 2021.

Kurniawati, H. and Yadav, V. An online pomdp solver for uncertainty planning in dynamic environment. *Robotics Research*, 114:611–629, 2016.

Kurniawati, H., Hsu, D., and Lee, W. S. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, 2008.

Kurzhanskiy, A. A. and Varaiya, P. Ellipsoidal toolbox (et). In *IEEE Conference on Decision and Control*, pp. 1498–1503, 2006.

Lee, A. X., Nagabandi, A., Abbeel, P., and Levine, S. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *NIPS*, pp. 741–752, 2020.

Ma, H., Liu, C., Li, S. E., Zheng, S., Sun, W., and Chen, J. Learn zero-constraint-violation policy in model-free constrained reinforcement learning. *arXiv preprint arXiv:2111.12953*, 2021.

Ma, X., Karkus, P., Hsu, D., Lee, W. S., and Ye, N. Discriminative particle filter reinforcement learning for complex partial observations. In *ICLR*, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Mousavinejad, E., Ge, X., Han, Q., Lim, T. J., and Vlacic, L. B. An ellipsoidal set-membership approach to distributed joint state and sensor fault estimation of autonomous ground vehicles. *IEEE/CAA Journal of Automatica Sinica*, 8:1107–1118, 2021.

Ni, T., Eysenbach, B., and Salakhutdinov, R. Recurrent model-free rl can be a strong baseline for many pomdps. In *ICML*, pp. 16691–16723, 2022.

Ray, A., Achiam, J., and Amodei, D. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7:1, 2019.

Richter, F., Orosco, R. K., and Yip, M. C. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, G., Pineau, J., and Kaplow, R. A survey of point-based pomdp solvers. In *AAMAS*, pp. 1–51, 2013.

Silver, D. and Veness, J. Monte-carlo planning in large pomdps. In *NIPS*, 2010.

Skelton, R. E., Iwasaki, T., and Grigoriadis, K. A unified algebric approach to control design, 1997.

Somani, A., Ye, N., Hsu, D., and Lee, W. S. Despot: Online pomdp planning with regularization. In *NIPS*, 2013.

Tschiatschek, S., Arulkumaran, K., Stühmer, J., and Hofmann, K. Variational inference for data-efficient model learning in pomdps. *arXiv preprint arXiv:1805.09281*, 2018.

Wang, C., Wang, J., Shen, Y., and Zhang, X. Autonomous navigation of uavs in large-scale complex environments: A deep reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 68:2124–2136, 2019.

Wang, Y. and Tan, X. Deep recurrent belief propagation network for pomdps. In *AAAI*, pp. 10236–10244, 2021.

Witsenhausen, H. S. Sets of possible states of linear systems given perturbed observations. *IEEE Transactions on Automatic Control*, 13:556–558, 1968.

Wu, C., Yang, G., Zhang, Z., Yu, Y., Li, D., Liu, W., and Hao, J. Adaptive online packing-guided search for pomdps. In *NIPS*, pp. 28419–28430, 2021.

Yang, F. and Li, Y. Set-membership filtering for discrete-time systems with nonlinear equality constraints. *IEEE Transactions on Automatic Control*, 54(10):2480–2486, 2009.

Zhang, Y., Xia, N., Han, Q., and Yang, F. Set-membership global estimation of networked systems. *IEEE Transactions on Cybernetics*, 52:1454–1464, 2020.

Zhao, W., He, T., and Liu, C. Model-free safe control for zero-violation reinforcement learning. In *ICRL*, 2021.

Zhu, P., Li, X., Poupart, P., and Miao, G. On improving deep reinforcement learning for pomdps. *arXiv preprint arXiv:1704.07978*, 2017.

Zou, L., Wang, Z., Geng, H., and Liu, X. Set-membership filtering subject to impulsive measurement outliers: A recursive algorithm. *IEEE/CAA Journal of Automatica Sinica*, 8:377–388, 2021.

A. Lemma

Lemma A.1. *((Kurzhanskiy & Varaiya, 2006)) For the affine transformation $x \mapsto Ax + b$ with known matrix A and vector b, if $x \in \epsilon(a, P)$, we have*

$$A\epsilon(a, p) + b = \epsilon(Aa + b, APA^{\mathrm{T}}).$$

Lemma A.2. *((Durieu et al., 1996)) The elementwise sum of given ellipsoids $\epsilon(a_i, P_i)$, $i = 1, 2, \cdots, m$, can be enclosed by a bounding ellipsoid*

$$\epsilon(a_1, P_1) \oplus \epsilon(a_2, P_2) \oplus \cdots \oplus \epsilon(a_m, P_m) \subseteq \epsilon(a, P),$$

with the center $a = \sum_{i=1}^{m} a_i$ and the shape matrix

$$P = \left(\sum_{i=1}^{m} \rho_i\right)\left(\sum_{i=1}^{m} \rho_i^{-1} P_i\right) \qquad \forall \rho_i > 0.$$

Lemma A.3. *(S-Procedure (Skelton et al., 1997)) Let $Y_0(\eta), Y_1(\eta), \cdots, Y_p(\eta)$ be quadratic functions of $\eta \in \mathbb{R}^n$*

$$Y_i(\eta) = \eta^T T_i \eta, i = 0, 1, \cdots, p$$

with $T_i = T_i^T$. Then, the implication

$$Y_1(\eta) \leq 0, \cdots, Y_p(\eta) \leq 0 \Rightarrow Y_0(\eta) \leq 0$$

holds if there exist $\tau_1, \cdots, \tau_p \geq 0$ such that

$$\eta^T \left(T_0 - \sum_{i=1}^{p} \tau_i T_i\right) \eta \leq 0.$$

Lemma A.4. *(Schur Complements (Boyd et al., 1994)) Given constant matrices L_1, L_2, L_3 where $L_1 = L_1^T$ and $L_2 = L_2^T < 0$, then*

$$L_1 - L_3^T L_2^{-1} L_3 \leq 0$$

if and only if

$$\begin{bmatrix} L_1 & L_3^T \\ L_3 & L_2 \end{bmatrix} \leq 0$$

or equivalently

$$\begin{bmatrix} L_1 & L_3 \\ L_3^T & L_2 \end{bmatrix} \leq 0.$$

B. The proof of Theorem 1

First, we write the estimation error $s_t - \hat{s}_t$, taking into account Equation (5)

$$\begin{aligned} & s_t - \hat{s}_t \\ &= \hat{s}_{t/t-1} + E_{t/t-1} z - \hat{s}_{t/t-1} - K_t o_t \\ &= E_{t/t-1} z - K_t(\hat{s}_{t/t-1} + E_{t/t-1} z + \omega_t^o) \\ &= -K_t \hat{s}_{t/t-1} + (I - K_t) E_{t/t-1} z - K_t \omega_t^o \\ &= \Phi_t \xi, \end{aligned} \tag{15}$$

where, $\xi = [1 \quad z^{\mathrm{T}} \quad (\omega^o)^{\mathrm{T}}]^{\mathrm{T}}$, and

$$\Phi_t = [-K_t \hat{s}_{t/t-1} \quad (I - K_t) E_{t/t-1} \quad -K_t]. \tag{16}$$

Thus, $(s_t - \hat{s}_t)^{\mathrm{T}} (P_t)^{-1} (s_t - \hat{s}_t) \leq 1$ can be rewritten as

$$\xi^{\mathrm{T}} \left[(\Phi_t)^{\mathrm{T}} (P_t)^{-1} \Phi_t - \mathrm{diag}(1, \mathbf{0}, \mathbf{0}) \right] \xi \leq 0, \tag{17}$$

all the inequality conditions can be expressed as follows:

$$\xi^{\mathrm{T}} \operatorname{diag}(-1, I, \mathbf{0})\xi \leq 0, \tag{18}$$

$$\xi^{\mathrm{T}} \operatorname{diag}(-1, \mathbf{0}, (M_t^o)^{-1})\xi \leq 0, \tag{19}$$

where I is the identity matrix or vector with appropriate dimensions.

By using the S-procedure Lemma A.3, the sufficient conditions such that the Equation (17) hold is that there exists positive scalars of τ_z, τ_ω^o such that

$$\xi^{\mathrm{T}}[(\Phi_t)^{\mathrm{T}} (P_t)^{-1} \Phi_t - \operatorname{diag}(1, \mathbf{0}, \mathbf{0}) - \tau_z \operatorname{diag}(-1, I, \mathbf{0}) - \tau_\omega^o \operatorname{diag}(-1, \mathbf{0}, (M_t^o)^{-1})]\xi \leq 0. \tag{20}$$

Equation (20) is written in the following compact form:

$$\xi^{\mathrm{T}}[(\Phi_t)^{\mathrm{T}} (P_t)^{-1} \Phi_t - \operatorname{diag}(1 - \tau_z - \tau_\omega^o, -\tau_z I, -\tau_\omega^o (M_t^o)^{-1}))]\xi \leq 0. \tag{21}$$

By denoting

$$\Pi_t = \operatorname{diag}\left(1 - \tau_z - \tau_\omega^o, -\tau_z I, -\tau_\omega^o (M_t^o)^{-1})\right), \tag{22}$$

Equation (21) is written as

$$\xi^{\mathrm{T}} \left[(\Phi_t)^{\mathrm{T}} (P_t)^{-1} \Phi_t - \Pi_t\right] \xi \leq 0. \tag{23}$$

Then, the statement of the Theorem 3.1 then follows by straightforward application of the Lemma A.4 to the above matrix inequality.

C. RL controller

The RL controller follows an actor-critic framework, which makes decisions based on belief states by a modifying PPO. The parameter ζ of the policy π_ζ is learned by optimizing a clipped "surrogate" objective

$$L^p(\zeta) = \mathbb{E}_{b_t, a_t}\left[min\left(\frac{\pi_\zeta(a_t|b_t)}{\pi_{\zeta_{old}}(a_t|b_t)}\hat{A}_t, clip\left(\frac{\pi_\zeta(a_t|b_t)}{\pi_{\zeta_{old}}(a_t|b_t)}, 1-\delta, 1+\delta\right)\hat{A}_t\right)\right], \tag{24}$$

where δ is the clipping parameter, $\hat{A}_t = \sum_{k=0}^{T} (\gamma\lambda)^k \left[r_t + \gamma \hat{V}_\phi(b_{t+k+1} - \hat{V}_\phi(b_{t+k})\right]$ is the Generalized Advantage Estimator with a trade-off coefficient λ; and $\hat{V}_\phi$ is the approximated value function, which is trained by $L^v(\xi) = \mathbb{E}_{b_t}||\hat{V}_\xi(b_t) - \hat{V}_\phi||$, where $\hat{V}_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$ is the discounted accumulated reward from timestep t onwards.

D. Evidence Lower BOund (ELBO)

The belief inference network can be optimized by maximizing the Evidence Lower BOund (ELBO). The detailed derivation is as follow: The variational bound for latent dynamics models $p(o_{1:T}, s_{1:T}|a_{1:T}) = \Pi_t p(s_t|s_{t-1}, a_{t-1})p(o_t|s_t)$ and a variational posterior $q(s_{1:T}|o_{1:T}, a_{1:T}) = \Pi_t q(s_t|o_{\leq t}, a_{\leq t})$ follows from importance weighting and Jensen's inequality as shown,

$$\begin{aligned}
L^m(\psi) &= \log p(o_{1:T}|a_{1:T}) \\
&= \log \mathbb{E}_{p(s_{1:T}|a_{1:T}} \left[\prod_{t=1}^{T} p(o_t|s_t)\right] \\
&= \log \mathbb{E}_{q(s_{1:T}|o_{1:T}, a_{1:T})} \left[\prod_{t=1}^{T} \frac{p(o_t|s_t)T(s_t|s_{t-1}, a_{t-1})}{q(s_t|o_{\leq t}, a_{\leq t})}\right] \\
&\geq \mathbb{E}_{q(s_{1:T}|o_{1:T}, a_{1:T})} \left[\prod_{t=1}^{T} \log p(o_t|s_t) + \log T(s_t|s_{t-1}, a_{t-1}) - \log q(s_t|o_{\leq t}, a_{\leq t})\right].
\end{aligned} \tag{25}$$