Constrained Bayesian Reinforcement Learning via Approximate Linear Programming [*]

Jongmin Lee[†], Youngsoo Jang[†], Pascal Poupart[‡], Kee-Eung Kim[†]
[†]School of Computing, KAIST, Republic of Korea
[‡]David R. Cheriton School of Computer Science, University of Waterloo, Canada
{jmlee, ysjang}@ai.kaist.ac.kr, ppoupart@uwaterloo.ca, kekim@cs.kaist.ac.kr

Abstract

In this paper, we consider the safe learning scenario where we need to restrict the exploratory behavior of a reinforcement learning agent. Specifically, we treat the problem as a form of Bayesian reinforcement learning in an environment that is modeled as a constrained MDP (CMDP) where the cost function penalizes undesirable situations. We propose a model-based Bayesian reinforcement learning (BRL) algorithm for such an environment, eliciting risk-sensitive exploration in a principled way. Our algorithm efficiently solves the constrained BRL problem by approximate linear programming, and generates a finite state controller in an off-line manner. We provide theoretical guarantees and demonstrate empirically that our approach outperforms the state of the art.

1 Introduction

In reinforcement learning (RL), the agent interacts with the unknown environment to maximize the long-term return defined by real-valued reward signals [Sutton and Barto, 1998]. Due to the uncertain nature of the environment, the agent faces an *exploration-exploitation* trade-off, a fundamental challenge in RL: the agent has to weigh between the action that yields the best return based on past experience and other actions that facilitate new experiences towards discovering better actions. This paper considers model-based Bayesian reinforcement learning (BRL) [Dearden *et al.*, 1999; Duff, 2002; Poupart *et al.*, 2006], which provides a principled way of optimally balancing between exploration and exploitation in the Bayesian perspective, with the goal of obtaining sample-efficient learning behaviours.

Still, in many situations, the notion of *safety* or *risk avoidance* is crucial and should be considered as another prime objective to the RL agent [Mihatsch and Neuneier, 2002; Hans *et al.*, 2008; García and Fernández, 2012]. For example, a Mars rover has to reach a target position as fast as possible, but at the same time, it should avoid navigating into dangerous ditches, which can potentially render it irrecoverable.

This safe behavior requirement has been captured in various forms, mostly considering the risk of performing very poorly due to the inherent stochasticity of the environment. In this formulation, the classic objective of maximizing expected return may be modified to minimize the variance of returns [Howard and Matheson, 1972], or to maximize the return in the worst case [Iyengar, 2005; Nilim and El Ghaoui, 2005].

In this paper, we consider the constrained MDP (CMDP) [Altman, 1999] as the framework for modeling the safe exploration requirement. CMDP assumes that actions incur costs as well as rewards, where the goal is to obtain a behaviour policy that maximizes the expected cumulative rewards while the expected cumulative costs are bounded. Under these circumstances we can naturally encode the risks of specific behaviours as cost functions and the degree of risk taking as cost constraints respectively. Here, we assume that the reward functions and the cost functions are known to the RL agent and only the transition probabilities are unknown, as in many model-based BRL studies [Poupart *et al.*, 2006; Asmuth *et al.*, 2009; Kolter and Ng, 2009; Araya-López *et al.*, 2012; Kim *et al.*, 2012]. Specifically, following [Kim *et al.*, 2012], we model BRL as a planning problem with the hyperstate constrained partially observable MDP (CPOMDP) [Kim *et al.*, 2011] and adopt constrained approximate linear programming (CALP) [Poupart *et al.*, 2015] to compute Bayes-optimal policies in an off-line manner.

Most of the successful approximate planning algorithms for (constrained) POMDPs confine the whole set of infinitely many beliefs to a finite set. This technique was also adopted in CALP [Poupart *et al.*, 2015] to treat other beliefs as convex combinations of finite samples of beliefs. However, doing so for model-based BRL can be problematic as it is not straightforward to represent a distribution over the transition probabilities as a finite convex combination. As will be described in the later part of the paper, one of our contributions is in introducing the notion of 'slip to ϵ-close beliefs', which enables a theoretical analysis and provides empirical support.

[*]This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2014-0-00147, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)) and was conducted at High-Speed Vehicle Research Center of KAIST with the support of Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD).

2 Background

We model the environment as a constrained Markov decision process (CMDP), defined by a tuple $\langle S, A, T, R, \mathbf{C} = \{C_k\}_{1..K}, \mathbf{c} = \{c_k\}_{1..K}, \gamma, s_0\rangle$ where S is the set of states s, A is the set of actions a, $T(s'|s,a) = \Pr(s'|s,a)$ is the transition probability, $R(s,a) \in \mathbb{R}$ is the reward function which denotes immediate reward incurred by taking action a in state s, $C_k(s,a) \in \mathbb{R}$ is the k^{th} cost function upper bounded by $c_k \in \mathbb{R}$ of k^{th} cost constraint, $\gamma \in [0,1)$ is the discount factor, and s_0 is the initial state. The goal is to compute an optimal policy π^* that maximizes expected cumulative rewards while expected cumulative costs are bounded.

$$\max_{\pi}\ V_R^{\pi}(s_0) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|s_0\right]$$

$$s.t.\ V_{C_k}^{\pi}(s_0) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t C_k(s_t,a_t)|s_0\right] \le c_k\ \ \forall k$$

The optimal policy of a CMDP is generally stochastic and can be obtained by solving the following linear program (LP) [Altman, 1999].

$$\max_{\{y(s,a)\}\forall s,a} \sum_{s,a} R(s,a)y(s,a) \tag{1}$$

$$s.t.\quad \sum_{a'} y(s',a') = \delta(s_0,s) + \gamma\sum_{s,a}T(s'|s,a)y(s,a)\ \ \forall s'$$

$$\sum_{s,a} C_k(s,a)y(s,a) \le c_k \quad \forall k$$

$$y(s,a) \ge 0 \quad \forall s,a$$

where $y(s,a)$ can be interpreted as a discounted occupancy measure of (s,a), and $\delta(x,y)$ is a Dirac delta function that has a value of 1 if $x = y$ and 0 otherwise. Once the optimal solution $y(s,a)$ is obtained, an optimal stochastic policy and the corresponding optimal value are computed as $\pi^*(a|s) = \Pr(a|s) = y(s,a)/\sum_{a'} y(s,a')$ and $V_R^*(s_0) = \sum_{s,a} R(s,a)y(s,a)$ respectively.

The constrained partially observable Markov decision process (CPOMDP) generalizes the CMDP by allowing partial observability and is defined by the tuple $\langle S, A, O, T, Z, R, \mathbf{C}, \mathbf{c}, \gamma, b_0\rangle$. Additional components are O, Z, and b_0, where O is the set of observations o and $Z(o|s',a) = \Pr(o|s',a)$ is the observation probability of observing o when taking action a and moving to state s', and $b_0(s) = \Pr(s_0 = s)$ is the initial belief at time step 0, respectively. Since the current Markovian state is not directly observable, the agent infers a belief $b_t(s) = \Pr(s_t = s)$ (which is a sufficient statistic for decision making) at every time step using the Bayes rule: upon executing a in b and observing o, the updated belief b^{ao} is

$$b^{ao}(s') \propto Z(o|s',a)\sum_s T(s'|s,a)b(s) \quad \forall s' \tag{2}$$

Consequently, decision making in CPOMDPs can be understood as a process of repeating: 1) choose an action based on the current belief, 2) update the current belief distribution based on the last action and observation using Eq. (2). A CPOMDP is equivalent to a constrained *belief state CMDP* $\langle \bar{S}, A, \bar{T}, \bar{R}, \bar{\mathbf{C}}, \mathbf{c}, \gamma, \bar{s}_0\rangle$. Here $\bar{s}_0 = b_0$ and $\bar{S} = B$ is the set of reachable beliefs starting from b_0. Transition probability $\bar{T}(b'|b,a)$ is constructed from components of the original CPOMDP and is expressed in terms of beliefs.

$$\bar{T}(b'|b,a) = \sum_o \Pr(o|b,a)\Pr(b'|b,a,o) \tag{3}$$

$$= \sum_o \underbrace{\left[\sum_{s,s'} Z(o|s',a)T(s'|s,a)b(s)\right]}_{\Pr(o|b,a)} \underbrace{\left[\delta(b', b^{ao})\right]}_{\Pr(b'|b,a,o)}$$

Similarly, the reward and cost functions are represented as

$$\bar{R}(b,a) = \sum_s b(s)R(s,a) \tag{4}$$

$$\bar{C}_k(b,a) = \sum_s b(s)C_k(s,a) \tag{5}$$

Although the resulting constrained belief MDP can be solved by LP (1) in principle, the cardinality of B is usually very large or even infinite, which makes the problem computationally intractable. To tackle the intractability, several approximate algorithms have been proposed, such as CPBVI, which is based on dynamic programming [Kim *et al.*, 2011], and CALP, which is based on linear programming [Poupart *et al.*, 2015]. CALP has been shown to perform much better than CPBVI.

3 Constrained BRL via Approximate Linear Programming

In many practical situations, too much exploration can severely impact an RL agent and therefore exploration should be restricted. Restrictions can naturally be encoded as constraints on cost functions that quantify the impact of certain behaviours. In this section, we propose a model-based BRL algorithm for the CMDP environment, leading to cost-sensitive exploration in a principled way. To this end, we will take the following approach: we first convert the constrained BRL problem into a hyper-state CPOMDP planning problem. The CPOMDP planning problem is then cast into an equivalent belief-state CMDP planning problem. Finally, we use constrained approximate linear programming (CALP) to efficiently compute the Bayes-optimal policy in an offline manner. These steps are explained in subsequent subsections.

3.1 Constrained BRL as CPOMDP Planning

Model-based BRL computes a full posterior distribution over the transition models and use it to make decisions. We can formulate model-based BRL in a CMDP environment $\langle S, A, T, R, \mathbf{C}, \mathbf{c}, \gamma, s_0\rangle$ as a hyper-state CPOMDP planning problem [Duff, 2002; Poupart *et al.*, 2006; Kim *et al.*, 2012], which is formally defined by the tuple $\langle S^+, A, O^+, T^+, Z^+, R^+, \mathbf{C}^+, \mathbf{c}, \gamma, b_0^+\rangle$. Assuming finite state and action spaces, each component is specifically $S^+ = S \times \{\theta^{sas'}\}$, $O^+ = S$, $T^+(\langle s', \theta'\rangle|\langle s, \theta\rangle, a) = \theta^{sas'}\delta(\theta, \theta')$,

$Z^+(o|\langle s',\theta'\rangle,a) = \delta(o,s')$, $R^+(\langle s,\theta\rangle,a) = R(s,a)$, $C_k^+(\langle s,\theta\rangle,a) = C_k(s,a)$, and $b_0^+ = (s_0,b_0)$.

A belief distribution over S^+ in a hyper-state CPOMDP is a pair (s,b) consisting of a Markovian state s of the original CMDP and the posterior distribution $b(\theta)$ over unknown parameters θ. Here $b(\theta)$ is commonly chosen to be a product of Dirichlet distributions since Dirichlets are conjugate priors of the multinomial transition probabilities:

$$b(\theta) = \prod_{s,a} \mathrm{Dir}(\theta^{sa*}|n^{sa*})$$

When the agent in belief $(\bar{s},b)$ takes an action $\bar{a}$ and observes the successor state $\bar{s}'$, the belief is updated to $(\bar{s}',b')$, where b' is defined as:

$$\begin{aligned} b'(\theta) &= b^{\bar{s}\bar{a}\bar{s}'}(\theta) = \eta b(\theta)\theta^{\bar{s}\bar{a}\bar{s}'} \\ &= \prod_{s,a} \mathrm{Dir}\big(\theta^{sa*}|n^{sa*} + \delta((\bar{s},\bar{a},\bar{s}'),(s,a,s'))\big) \end{aligned} \tag{6}$$

where η is a normalizing constant. The hyper-state CPOMDP can also be easily understood as an equivalent belief-state CMDP $\langle \bar{S}^+, A, \bar{T}^+, \bar{R}^+, \bar{\mathbf{C}}^+, \mathbf{c}, \gamma, \bar{s}_0^+\rangle$. Here $\bar{S}^+ = S \times B$ and $\bar{s}_0^+ = (s_0,b_0)$ where B is the set of possible posterior distributions over $\theta^{sas'}$ from initial prior b_0. Transition probabilities among belief states (s,b) are defined as

$$\begin{aligned} \bar{T}^+(\langle s',b'\rangle|\langle s,b\rangle,a) &= \Pr(s'|s,b,a)\Pr(b'|s,b,a,s') \\ &= \mathbb{E}_b\left[\theta^{sas'}\right]\delta(b',b^{sas'}) \end{aligned} \tag{7}$$

Similarly, the reward function and the cost functions are

$$\bar{R}^+(\langle s,b\rangle,a) = R(s,a)$$
$$\bar{C}_k^+(\langle s,b\rangle,a) = C_k(s,a)$$

In theory, this belief-state CMDP can be solved using the following LP, which is an extension of (1) and the one in [Poupart *et al.*, 2015] to treat hyper belief states:

$$\max_{\{y(s,b,a)\}\forall s,a} \sum_{s,b,a} R(s,a)y(s,b,a) \tag{8}$$

$$\begin{aligned} s.t.\quad & \sum_{a'} y(s',b',a') = \delta((s_0,b_0),(s,b)) \\ &\qquad + \gamma\sum_{s,b,a} T(s',b'|s,b,a)y(s,b,a) \quad \forall s',b' \\ & \sum_{s,b,a} C_k(s,a)y(s,b,a) \le c_k \quad \forall k \\ & y(s,b,a) \ge 0 \quad \forall s,b,a \end{aligned}$$

3.2 Approximate Linear Programming

The main challenge in solving the linear program (8) lies in the fact that the number of beliefs $|S \times B|$ is infinite, yielding infinitely many variables and constraints in the LP. We thus approximate (8) using finitely sampled beliefs. In order to facilitate a formal analysis, we assume a finite set of beliefs $\widehat{B} \subset B$ that covers the entire belief space fairly well. More formally, we assume that there exists a constant ϵ such that

$$\forall b \in \widehat{B},\ s,s' \in S,\ a \in A, \quad \min_{b' \in \widehat{B}} \|b' - b^{sas'}\|_1 \le \epsilon$$

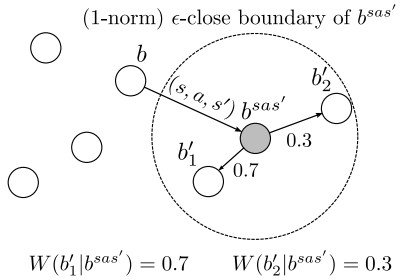


Figure 1: Constructing the approximate transition function $\widehat{T}$. White circles represent the beliefs in $\widehat{B}$, and the gray circle denotes the successor belief $b^{sas'}$, which is not in $\widehat{B}$. Here $\Pr(b'|s,b,a,s') = \delta(b',b^{sas'})$ in Eq (9) is relaxed to $W(b'|b^{sas'}) \in [0,1]$ with non-zero probabilities only for ϵ-close beliefs in the neighborhood of $b^{sas'}$.

where $\|\cdot\|_1$ denotes total variance distance

$$\|b' - b^{sas'}\|_1 = \int |b'(\theta) - b^{sas'}(\theta)|d\theta$$

Since $\widehat{B}$ does not completely cover B for $\epsilon > 0$, we need to re-define the transition function $T(s',b'|s,b,a)$ among (s,b) and $(s',b') \in S \times \widehat{B}$. From the original, exact transition probability $T(s',b'|s,b,a)$ defined over $(s,b) \in S \times B$:

$$\begin{aligned} T(s',b'|s,b,a) &= \Pr(s'|s,b,a)\Pr(b'|s,b,a,s') \\ &= \mathbb{E}_b\left[\theta^{sas'}\right]\delta(b',b^{sas'}), \end{aligned} \tag{9}$$

we relax $\delta(b',b^{sas'})$ to $W(b'|b^{sas'})$ that has non-zero probability only for ϵ-close beliefs:

$$\begin{aligned} \widehat{T}(s',b'|s,b,a) &= \Pr(s'|s,b,a)\widehat{\Pr}(b'|s,b,a,s') \\ &= \mathbb{E}_b\left[\theta^{sas'}\right]W(b'|b^{sas'}), \end{aligned} \tag{10}$$

where W is defined as a probability distribution over $\widehat{B}$.

$$W(b'|b^{sas'}) = \begin{cases} \kappa K(b',b^{sas'}) & \text{if } \|b' - b^{sas'}\|_1 \le \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where κ is the normalizing constant $\sum_{b' \in \widehat{B}} W(b'|b^{sas'}) = 1$ and $K(b,b') \ge 0$ is a similarity measure between two beliefs $b(\theta)$ and $b'(\theta)$. This relaxation can be interpreted as “slipping” to one of the ϵ-close successor beliefs with probability W. Figure 1 depicts the construction process of W. Thus, we approximate the original LP (8) by using a finite set of beliefs $\widehat{B}$ and replacing T by $\widehat{T}$. Algorithm 1 describes the overall process of computing the approximate Bayes-optimal policy.

The policy (π, W) obtained from Algorithm 1 constitutes a finite state controller with $|S||\widehat{B}|$ nodes and is executed in the real environment as follows: the initial node of the controller is set to (s_0,b_0). At every time step, sample an action $a \sim \pi(a|s,b)$ based on the current node (s,b) for execution. Then, observe the next state s' from the environment and sample $b' \sim W(b'|b^{sas'})$. Finally, the new node of the controller is set to (s',b') and repeat.

Algorithm 1 Constrained BRL via Approximate LP

```
Input: $S, A, R, C, c, \gamma, s_0, \widehat{B}, b_0$.
  for each $s, s' \in S,\ b \in \widehat{B},$ and $a \in A$ do
    Compute $W(b'|b^{sas'})\ \forall b'$ by Eq (11)
  end for
  $\widehat{T}(s', b'|s, b, a) \leftarrow \mathbb{E}_b[\theta^{sas'}] \cdot W(b'|b^{sas'})\ \forall s, b, a, b, b'$
  $y \leftarrow$ solve LP (8) with $\widehat{B}$ and $\widehat{T}(s', b'|s, b, a)$
  for each $s \in S,\ b \in \widehat{B},$ and $a \in A$ do
    $\pi(a|s, b) \leftarrow y(s, b, a) / \sum_{a'} y(s, b, a')$
  end for
  $\widehat{V}^*_R(s_0, b_0) \leftarrow \sum_{s,b,a} y(s, b, a) R(s, a)$
  $\widehat{V}^*_{C_k}(s_0, b_0) \leftarrow \sum_{s,b,a} y(s, b, a) C_k(s, a)\ \forall k$
Output: $(\pi, W)$: finite state controller, $\widehat{V}^*_R(s_0, b_0)$: approximate Bayes-optimal value
```

We remark that Poupart *et al.* [2015] take a similar approach to solving CPOMDPs by considering finitely sampled beliefs. Specifically, they approximate the transitions by relaxing $\delta(b', b^{ao})$ in (3) as interpolation weights $w(b', b^{ao})$ such that $\sum_{b' \in \widehat{B}} w(b', b^{ao}) b' = b^{ao}$, $\sum_{b' \in \widehat{B}} w(b', b^{ao}) = 1$ and $w(b', b^{ao}) \geq 0$. This approach cannot be directly adopted in the Bayesian learning setting since a belief is no longer a finite-dimensional probability vector but rather a probability density function. There is no straightforward way to approximate an arbitrary Dirichlet using a convex combination of finitely many Dirichlets.

4 Theoretical Analysis

In this section, we provide a formal analysis of the error bound incurred by taking a finite set of beliefs and using the 'slip to ϵ-close belief' approximation scheme to re-define the transition probabilities. In order to facilitate our analysis, we first show the result on the single objective case, i.e. reward-only POMDPs.

Lemma 1. *Suppose that the reward function is bounded in* $[-R_{\max}, R_{\max}]$. *For any* $b, \widehat{b} \in B$ *such that* $\|b - \widehat{b}\|_1 \leq \epsilon$, *the following inequality holds.*

$$|V^*_R(s, b) - V^*_R(s, \widehat{b})| \leq \frac{R_{\max}}{1-\gamma}\epsilon \overset{\text{def}}{=} \epsilon_1$$

Proof. It is well known that the optimal value of POMDP is piecewise linear and convex function, and can be expressed as an inner product between a belief distribution b and a set of alpha functions $\Gamma^* = \{\alpha_1, \cdots, \alpha_{|\Gamma|}\}$.

$$\begin{aligned}
|V^*_R(s, b) - V^*_R(s, \widehat{b})| &= \left| \max_{\alpha_1 \in \Gamma^*_s} \langle \alpha_1, b \rangle - \max_{\alpha_2 \in \Gamma^*_s} \langle \alpha_2, \widehat{b} \rangle \right| \\
&\leq \max_{\alpha \in \Gamma^*_s} \| \langle \alpha, b - \widehat{b} \rangle \|_1 \\
&\leq \max_{\alpha \in \Gamma^*_s} \|\alpha\|_\infty \|b - \widehat{b}\|_1 \quad \text{(Hölder's inequality)} \\
&\leq \frac{R_{\max}}{1-\gamma}\epsilon \overset{\text{def}}{=} \epsilon_1 \qquad (\because \|b - \widehat{b}\|_1 \leq \epsilon)
\end{aligned}$$

□

Note that this lemma is same as lemma 1 in [Hsu *et al.*, 2007] with a simpler proof.

Lemma 2. *Suppose that for all* $b \in B$, *there exists* $\widehat{b} \in \widehat{B}$ *such that* $\|b - \widehat{b}\|_1 \leq \epsilon$. *Let* H *be an exact value-function mapping, and* $\widehat{H}$ *be a value-function mapping of approximate POMDP with 'slip to* ϵ*-close belief' approximation. Suppose that* V^* *is an unique fixed point solution of* $HV = V$. *Then, the following inequality holds:*

$$\|HV^* - \widehat{H}V^*\| \leq \gamma\epsilon_1$$

where ϵ_1 *is defined in Lemma 1.*

Proof. For any $s \in S, b \in \widehat{B}, a \in A$, and $s' \in S$,

$$\begin{aligned}
&\sum_{b' \in \widehat{B}} \widehat{T}(s', b'|s, b, a) V^*(s', b') \\
&= \mathbb{E}_b[\theta^{sas'}] \sum_{b' \in \widehat{B}: \|b' - b^{sas'}\|_1 \leq \epsilon} W(b'|b^{sas'}) V^*(s', b') \\
&\geq \mathbb{E}_b[\theta^{sas'}] \sum_{b' \in \widehat{B}: \|b' - b^{sas'}\|_1 \leq \epsilon} W(b'|b^{sas'}) [V^*(s', b^{sas'}) - \epsilon_1] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma 1)} \\
&= \mathbb{E}_b[\theta^{sas'}] \Big[V^*(s', b^{sas'}) - \epsilon_1 \Big] \qquad (\because \sum_{b'} W(b'|b^{sas'}) = 1)
\end{aligned}$$

Therefore, for any $s \in S$ and $b \in \widehat{B}$,

$$\begin{aligned}
&(\widehat{H}V^*)(s, b) \\
&= \max_a \left[R(s, a) + \gamma \sum_{s'} \sum_{b' \in \widehat{B}} \widehat{T}(s', b'|s, b, a) V^*(s', b') \right] \\
&\geq \max_a \left[R(s, a) + \gamma \sum_{s'} \mathbb{E}_b[\theta^{sas'}] \Big[V^*(s', b^{sas'}) - \epsilon_1 \Big] \right] \\
&= (HV^*)(s, b) - \gamma\epsilon_1
\end{aligned}$$

Similarly, we can obtain:

$$\begin{aligned}
&(\widehat{H}V^*)(s, b) \\
&\leq \max_a \left[R(s, a) + \gamma \sum_{s'} \mathbb{E}_b \Big[\theta^{sas'} \Big] \Big[V^*(s', b^{sas'}) + \epsilon_1 \Big] \right] \\
&= (HV^*)(s, b) + \gamma\epsilon_1
\end{aligned}$$

As a consequence, we get the result $\|HV^* - \widehat{H}V^*\| \leq \gamma\epsilon_1$. □

Theorem 1. *Suppose that the reward function is bounded in* $[-R_{\max}, R_{\max}]$, *and for all* $b \in B$, *there exists* $\widehat{b} \in \widehat{B}$ *such that* $\|b - \widehat{b}\|_1 \leq \epsilon$. *Let* V^* *be the fixed point solution* $V^* = HV^*$ *(optimal value function of original POMDP), and* $\widehat{V}^*$ *be the fixed point solution* $\widehat{V}^* = \widehat{H}\widehat{V}^*$ *(optimal value function of approximate POMDP with 'slip to* ϵ*-close belief' approximation). Then, the following inequality holds.*

$$|V^*(s_0, b_0) - \widehat{V}^*(s_0, b_0)| \leq \frac{\gamma R_{\max}}{(1-\gamma)^2}\epsilon$$

Proof.

$$\begin{aligned}
\|V^* - \widehat{V}^*\| &= \|HV^* - \widehat{H}\widehat{V}^*\| \\
&\le \|HV^* - \widehat{H}V^*\| + \|\widehat{H}V^* - \widehat{H}\widehat{V}^*\| \\
&\le \gamma\epsilon_1 + \|\widehat{H}V^* - \widehat{H}\widehat{V}^*\| \qquad \text{(Lemma 2)} \\
&\le \gamma\epsilon_1 + \gamma\|V^* - \widehat{V}^*\| \\
\therefore \|V^* - \widehat{V}^*\| &\le \frac{\gamma\epsilon_1}{1-\gamma} = \frac{\gamma R_{\max}}{(1-\gamma)^2}\epsilon
\end{aligned}$$

□

In the above, $\widehat{H}$ denotes the approximate Bellman backup that arises from finitely sampled beliefs $\widehat{B}$ and 'slip to ϵ-close belief' transition approximation scheme.

Unfortunately, this result cannot be naively extended to constrained POMDPs. To make the connection between single objective POMDPs and constrained POMDPs, we should look at the dual form of the LP in (8) given by:

$$\min_{\substack{\{\mathcal{V}(s,b)\}\forall s,b \\ \{\lambda_k\}\forall k}} \sum_{s,b} \delta\big((s_0,b_0),(s,b)\big)\mathcal{V}(s,b) + \sum_k c_k\lambda_k \tag{12}$$

$$\begin{aligned}
s.t. \quad \mathcal{V}(s,b) \ge R(s,a) &- \sum_k C_k(s,a)\lambda_k \\
&+ \gamma \sum_{s',b'} T(s',b'|s,b,a)\mathcal{V}(s',b') \quad \forall s,b,a \\
\lambda_k \ge 0 \ \forall k
\end{aligned}$$

If we fix $\boldsymbol{\lambda}$ in the above formulation, the problem becomes the single objective POMDP with new reward function $R(s,a) - \boldsymbol{\lambda}^\top\mathbf{C}(s,a)$. The solution can be obtained through the dynamic programming with the backup operator $H_{\boldsymbol{\lambda}}$ defined as:

$$\begin{aligned}
H_{\boldsymbol{\lambda}}\mathcal{V}(s,b) = \max_a \Bigg[& R(s,a) - \boldsymbol{\lambda}^\top\mathbf{C}(s,a) \\
& + \gamma \sum_{s',b'} T(s',b'|s,b,a)\mathcal{V}(s',b') \Bigg],
\end{aligned}$$

which is a contraction mapping and has an unique fixed point solution $\mathcal{V}^*_{\boldsymbol{\lambda}}$. Then, the LP dual problem can be reduced to $\min_{\boldsymbol{\lambda}} \left[\mathcal{V}^*_{\boldsymbol{\lambda}}(s_0,b_0) + \boldsymbol{\lambda}^\top\mathbf{c}\right]$. Moreover, if there is an optimal solution y^* to primal LP, there must exist the corresponding dual optimal solution $\mathcal{V}^*$ and $\boldsymbol{\lambda}^*$, and duality gap is zero (i.e. $V^*_R(s_0,b_0,\mathbf{c}) = \mathcal{V}^*_{\boldsymbol{\lambda}^*}(s_0,b_0) + \boldsymbol{\lambda}^{*\top}\mathbf{c}$) by the strong duality theorem. We can now apply the result in Theorem 1 with a fixed $\boldsymbol{\lambda}$ since $\mathcal{V}^*_{\boldsymbol{\lambda}}$ is a value function of single objective POMDP.

Lemma 3. *Suppose that reward function and cost functions of the CMDP environment are bounded in* $[0, R_{\max}]$ *and* $[0, C_{\max}]$ *respectively. Let* $\mathcal{V}^*_{\boldsymbol{\lambda}}$ *be the optimal value function of the POMDP with reward function of* $R(s,a) - \boldsymbol{\lambda}^\top\mathbf{C}(s,a)$ *and* $\widehat{\mathcal{V}}^*_{\boldsymbol{\lambda}}$ *be the optimal value of the approximate POMDP with the same reward function and 'slip to* ϵ*-close belief' approximation. Then, the following inequality holds for all* $\boldsymbol{\lambda} \ge 0$*:*

$$|\mathcal{V}^*_{\boldsymbol{\lambda}}(s_0,b_0) - \widehat{\mathcal{V}}^*_{\boldsymbol{\lambda}}(s_0,b_0)| \le \frac{\gamma(R_{\max} + \|\boldsymbol{\lambda}\|_1 C_{\max})}{(1-\gamma)^2}\epsilon$$

Proof. For all b, a and $\boldsymbol{\lambda} \ge \mathbf{0}$,

$$\begin{aligned}
\left|R(s,a) - \boldsymbol{\lambda}^\top\mathbf{C}(s,a)\right| &\le |R_{\max}| + |\boldsymbol{\lambda}^\top\mathbf{C}(s,a)| \\
&\le R_{\max} + \|\boldsymbol{\lambda}\|_1 C_{\max}
\end{aligned}$$

Therefore, by simply applying Theorem 1 with the reward range of $[-R_{\max} - \|\boldsymbol{\lambda}\|_1 C_{\max}, R_{\max} + \|\boldsymbol{\lambda}\|_1 C_{\max}]$, we get the result. □

Lemma 3 indicates that the approximation error can depend on the magnitude of $\boldsymbol{\lambda}$, and the following Lemma 4 shows that $\|\boldsymbol{\lambda}\|_1$ can be bounded.

Lemma 4. *Suppose that reward function is bounded in* $[0, R_{\max}]$ *and there exists* $\tau > 0$ *and policy* π *such that* $V^\pi_{\mathbf{C}}(s_0,b_0) + \tau\mathbf{1} \le \mathbf{c}$*. Then, the following inequality holds:*

$$\|\boldsymbol{\lambda}^*\|_1 \le \frac{R_{\max}}{\tau(1-\gamma)}$$

Proof. Let $\mathcal{C} = \{\mathbf{c} \mid \text{there exists a policy } \pi \text{ such that } V^\pi_{\mathbf{C}}(s_0,b_0) \le \mathbf{c}\}$. For any $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$,

$$\begin{aligned}
& V^*_R(s,b,\mathbf{c}') - V^*_R(s,b,\mathbf{c}) \\
&= \mathcal{V}^*_{\boldsymbol{\lambda}'^*}(s,b) + \boldsymbol{\lambda}'^{*\top}\mathbf{c}' - \mathcal{V}^*_{\boldsymbol{\lambda}^*}(s,b) - \boldsymbol{\lambda}^{*\top}\mathbf{c} \\
& \qquad\qquad\qquad\qquad \text{(Dual variables of LP)} \\
&\le \mathcal{V}^*_{\boldsymbol{\lambda}^*}(s,b) + \boldsymbol{\lambda}^{*\top}\mathbf{c}' - \mathcal{V}^*_{\boldsymbol{\lambda}^*}(s,b) - \boldsymbol{\lambda}^{*\top}\mathbf{c} \\
& \qquad\qquad (\because \boldsymbol{\lambda}'^* \text{ is optimal solution to min. problem}) \\
&= \boldsymbol{\lambda}^{*\top}(\mathbf{c}' - \mathbf{c})
\end{aligned}$$

where $\boldsymbol{\lambda}'^* = \arg\min_{\boldsymbol{\lambda}'}[\mathcal{V}^*_{\boldsymbol{\lambda}'}(s_0,b_0) + \boldsymbol{\lambda}'^\top\mathbf{c}']$, and $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}}[\mathcal{V}^*_{\boldsymbol{\lambda}}(s_0,b_0) + \boldsymbol{\lambda}^\top\mathbf{c}]$. For any $\tau > 0$ such that $\mathbf{c} - \tau\mathbf{1} \in \mathcal{C}$, let $\mathbf{c}' = \mathbf{c} - \tau\mathbf{1}$. Then,

$$\begin{aligned}
V^*_R(s,b,\mathbf{c} - \tau\mathbf{1}) - V^*_R(s,b,\mathbf{c}) &\le \boldsymbol{\lambda}^{*\top}(-\tau\mathbf{1}) = -\tau\|\boldsymbol{\lambda}^*\|_1 \\
\therefore \|\boldsymbol{\lambda}^*\|_1 &\le \frac{V^*_R(s,b,\mathbf{c}) - V^*_R(s,b,\mathbf{c} - \tau\mathbf{1})}{\tau} \\
&\le \frac{R_{\max}}{\tau(1-\gamma)}
\end{aligned}$$

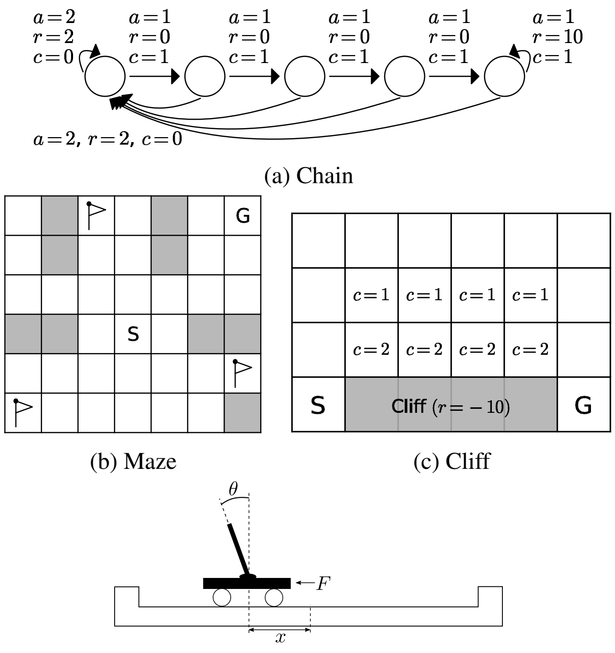
□

Remark. Intuitively, τ represents how much we can reduce the cost constraint without making the problem infeasible. For example, when the cost constraint $\mathbf{c}$ is set to 100, while the minimum cost value of the given CMDP environment is 30 (i.e. $\min_\pi V^\pi_{\mathbf{C}}(s_0,b_0) = 30$), $\tau = 100 - 30 = 70$.

We can now show the main result that bounds the error in the value function due to approximate LP:

Theorem 2. *Suppose that reward function and cost functions of CMDP environment are bounded in* $[0, R_{\max}]$ *and* $[0, C_{\max}]$ *respectively. Let* $V^*_R(s_0,b_0,\mathbf{c})$ *be an optimal value of the original CPOMDP with cost constraint* $\mathbf{c}$*, and* $\widehat{V}^*_R(s_0,b_0,\mathbf{c})$ *be an optimal value function of approximate CPOMDP with cost constraint* $\mathbf{c}$ *and 'slip to* ϵ*-close beliefs' approximation. Then, the following inequality holds:*

$$|V^*_R(s_0,b_0,\mathbf{c}) - \widehat{V}^*_R(s_0,b_0,\mathbf{c})| \le \frac{\gamma\,(\tau - \tau\gamma + C_{\max})\,R_{\max}}{\tau(1-\gamma)^3}\epsilon$$



(a) Chain

(b) Maze

(c) Cliff

(d) Cart pole

Figure 2: Domains

Proof.

$$\begin{aligned} &|V_R^*(s_0,b_0,\mathbf{c}) - \widehat{V}_R^*(s_0,b_0,\mathbf{c})| \\ &= \left|\left(\mathcal{V}^*_{\boldsymbol\lambda^*}(s_0,b_0)+{\boldsymbol\lambda^*}^\top\mathbf{c}\right) - \left(\widehat{\mathcal{V}}^*_{\boldsymbol\lambda'^*}(s_0,b_0)+{\boldsymbol\lambda'^*}^\top\mathbf{c}\right)\right| \\ &\le \max_{\boldsymbol\lambda\in\{\boldsymbol\lambda^*,\boldsymbol\lambda'^*\}} \left|\mathcal{V}^*_{\boldsymbol\lambda}(s_0,b_0)-\widehat{\mathcal{V}}^*_{\boldsymbol\lambda}(s_0,b_0)\right| \\ &\le \max_{\boldsymbol\lambda\in\{\boldsymbol\lambda^*,\boldsymbol\lambda'^*\}} \left(\frac{\gamma(R_{\max}+\|\boldsymbol\lambda\|_1 C_{\max})}{(1-\gamma)^2}\epsilon\right) && \text{(Lemma 3)} \\ &\le \frac{\gamma\,(\tau-\tau\gamma+C_{\max})\,R_{\max}}{\tau(1-\gamma)^3}\epsilon && \text{(Lemma 4)}\end{aligned}$$

where $\boldsymbol\lambda^* = \arg\min_{\boldsymbol\lambda}[\mathcal{V}^*_{\boldsymbol\lambda}(s_0,b_0)+\boldsymbol\lambda^\top\mathbf{c}]$, and $\boldsymbol\lambda'^* = \arg\min_{\boldsymbol\lambda'}[\widehat{\mathcal{V}}^*_{\boldsymbol\lambda'}(s_0,b_0)+{\boldsymbol\lambda'}^\top\mathbf{c}]$. □

5 Experiments

We conducted experiments on 3 discrete state domains and 1 continuous state domain depicted in Figure 2. The chain domain [Dearden *et al.*, 1998] has 5 states and 2 actions. The agent receives large reward of 10 by executing action 1 at the fifth state, and small of 2 by executing action 2 at any state. With probability 0.2 agent slips, which means that the opposite transition occurs. We prepared a constrained version of this domain by allocating cost 1 for action 1 and cost 0 for action 2, which makes the excessively taking action 1 may violate the cost constraint, following [Kim *et al.*, 2012]. The maze domain [Strens, 2000] has 33 rooms and three flags. There are 5 actions {left, right, up, down, stay} that allow moving to neighboring rooms. Every action except for stay fails with probability 0.1. If an action fails, the agent randomly slips into one of the directions orthogonal to the intended one. The reward is given by the number of captured flags when the agent arrives at the goal. Since there are 33 rooms and 8 possible status of captured flags, there are 264 states in total. We defined the cost 1 to every action except staying at the starting cell. The cliff domain [Sutton and Barto, 1998] has 24 states and 4 actions {left, right, up, down}. Every action fails with probability 0.1. If an action fails, the agent randomly slips into one of the not intended directions. The agent receives large reward of 20 once reaching the goal state, and reward of -10 when it falls into cliff. We assign cost of 2 to the cliff and the locations one unit away from the cliff (high-risk area) and cost of 1 to the locations two units away from the cliff (medium-risk area) and cost of 0 elsewhere (low-risk area). The cart pole [Sutton and Barto, 1998] domain is the classical control problem to keep the pole upright. The state is encoded $s = [x, \dot{x}, \theta, \dot{\theta}]$, where x is the position, $\dot{x}$ is the velocity, θ is the angle, and $\dot{\theta}$ is the angular velocity. There are two actions {left, right} that apply a force of -1 or $+1$ to the cart. A reward of -1 is received if the cart or pole is out of the range: $|x| \geq 2.4$ or $|\theta| \geq 12°$. Cost function is defined by $C(s,a) = |x|$ to encourage the agent to stay in the middle.

5.1 Experimental Setup

Finite State Domains

For all the finite state domains, we used two kinds of structural priors, "tied" and "semi". In both tied and semi, it is assumed that the transition dynamics are known except for the slip probabilities. The tied prior assumes that the slip probability is independent of state and action, so there is only one unknown parameter. The semi prior assumes action-dependent slip probabilities, so the number of unknown parameters is 2 for chain, 5 for maze, and 4 for cliff, respectively. We used uninformative Dirichlet priors. Approximate transitions in (10) were constructed by

$$W(b'|b^{sas'}) \propto \exp\left(-\frac{d(b', b^{sas'})}{2\sigma^2}\right) \text{ and } \tag{13}$$

$$d(b_1,b_2) = \frac{1}{2}\left(\mathbb{KL}(b_1 \parallel b_2) + \mathbb{KL}(b_2 \parallel b_1)\right),$$

where σ is set to 0.5. $\widehat{B}$ were collected by uniformly random policy executed in the environment for 50 time steps.

In the chain and cliff domains, we report the results of 200 trials of the first 2000 times steps with discount factor $\gamma = 0.99$. In the maze domain, we report the results of 100 tirals of the first 1000 time steps with discount factor $\gamma = 0.95$.

Continuous State Domain

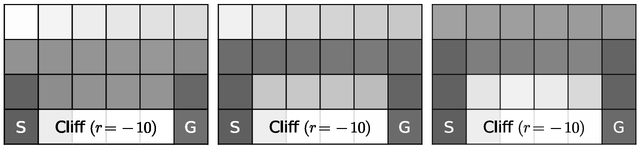
In cart pole, the environment dynamics are modeled as a linear dynamical system, as in [Tziortziotis *et al.*, 2013].

$$\begin{aligned} s_{t+1} &= A_{a_t}\phi(s_t) + \epsilon_{a_t} \\ \epsilon_{a_t} &\sim \mathcal{N}(\mathbf{0}, V_{a_t}) \\ s_{t+1}|s_t, a_t &\sim \mathcal{N}(A_{a_t} f(s_t), V_{a_t}) \end{aligned}$$

Here $\phi(s) = [s, 1]^\top$ and the transition model is parameterized by $\{(A_a, V_a)\}_{\forall a}$. We use matrix-normal prior for A and inverse-Wishart prior for V. Given the samples

domain	c	algorithm	avg discounted total reward	avg discounted total cost	time (min)
chain-tied	100	CBEETLE	355.85 ± 4.55	99.64 ± 0.08	1.2
		CBRL-ALP	339.77 ± 8.01	91.26 ± 2.44	0.1
	75	CBEETLE	305.02 ± 3.82	74.96 ± 0.04	2.1
		CBRL-ALP	315.22 ± 7.14	71.46 ± 1.75	0.1
	50	CBEETLE	243.54 ± 3.29	50.03 ± 0.10	9.7
		CBRL-ALP	289.86 ± 6.25	48.37 ± 1.10	0.1
	25	CBEETLE	218.54 ± 1.94	25.03 ± 0.04	34.8
		CBRL-ALP	235.06 ± 6.03	23.72 ± 1.12	0.1
chain-semi	100	CBEETLE	355.12 ± 4.16	98.42 ± 0.14	1.9
		CBRL-ALP	327.10 ± 8.91	83.51 ± 3.39	0.2
	75	CBEETLE	298.03 ± 4.09	75.00 ± 0.05	3.8
		CBRL-ALP	307.22 ± 7.87	66.19 ± 2.42	0.2
	50	CBEETLE	237.71 ± 3.49	50.10 ± 0.10	16.7
		CBRL-ALP	276.01 ± 7.65	44.36 ± 1.89	0.2
	25	CBEETLE	214.08 ± 2.14	24.94 ± 0.06	89.9
		CBRL-ALP	226.74 ± 6.32	22.07 ± 1.73	0.2
maze-tied	20	CBEETLE$^{(*)}$	1.02 ± 0.02	19.04 ± 0.02	242.5
		CBRL-ALP	1.03 ± 0.02	19.09 ± 0.03	39.3
	18	CBEETLE$^{(*)}$	0.93 ± 0.04	17.96 ± 0.46	733.1
		CBRL-ALP	0.96 ± 0.02	17.92 ± 0.22	41.0
cliff-tied	100	CBEETLE	121.21 ± 4.94	91.88 ± 0.54	173.8
		CBRL-ALP	166.20 ± 2.32	64.75 ± 3.57	1.5
	50	CBEETLE	52.98 ± 3.77	44.41 ± 0.50	180.0
		CBRL-ALP	160.89 ± 1.57	44.72 ± 0.90	1.5
	30	CBEETLE	-104.52 ± 4.58	54.64 ± 0.97	206.8
		CBRL-ALP	150.19 ± 1.41	25.99 ± 0.83	1.5
cliff-semi	100	CBEETLE	51.55 ± 5.57	91.25 ± 0.63	485.0
		CBRL-ALP	161.05 ± 1.88	51.65 ± 2.81	7.9
	50	CBEETLE	-78.80 ± 5.89	63.03 ± 0.73	594.7
		CBRL-ALP	158.44 ± 1.59	42.28 ± 1.00	7.9
	30	CBEETLE	-117.80 ± 9.29	43.09 ± 1.28	657.4
		CBRL-ALP	146.91 ± 1.39	22.44 ± 0.86	7.9

Table 1: Experimental result of chain, maze and cliff domains. The results with (*) are from [Kim *et al.*, 2012].



(a) $c = 100$ (b) $c = 50$ (c) $c = 30$

Figure 3: Visualizing behaviours under different c in cliff/tied domain. Darkness represent the visitation frequency.

$\{(s_t, a_t, s_{t+1})\}_{t=1..N}$, the posteriors can be obtained in a closed form [Minka, 1998].

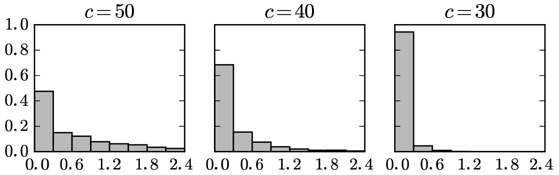
To construct finite set of beliefs $\widehat{S} \times \widehat{B}$, we collected 10000 states by taking random actions and quantized them into 256 representative states via K-means clustering, which constitutes $\widehat{S}$. Then, $\widehat{B}$ was constructed by gathering beliefs during the first 100 steps at every 10 step-interval. $\Pr(s'|s, b, a)$ in (10) was replaced by

$$\widehat{\Pr}(s'|s, b, a) \propto \int_{A,V} \Pr(s'|s, A_a, V_a) b(A_a, V_a) dA dV \; \forall s' \in \widehat{S}$$

which is proportional to the posterior predictive distribution. $W(b'|b^{sas'})$ was defined by (13) with $\sigma = 1$. The results are averaged over 100 trials of the first 10000 steps with discount factor $\gamma = 0.99$.

5.2 Results

Table 1 summarizes the experimental results for discrete domains, comparing our algorithm CBRL-ALP to the previous state-of-the-art approach CBEETLE [Kim *et al.*, 2012]. Overall, our method outperforms in computation speed by an order of magnitude, while yielding good policies. In addition, CBEETLE generally slowed down as c became tighter, but CBRL-ALP was barely affected by the cost constraint. Note that our algorithm (and CBEETLE) generally achieves lower average total reward in the semi prior compared to the tied prior. This is a natural result since the agent has more uncertainty (more unknown parameters) regarding the dynamics and as such, it has to act more conservatively to meet the cost constraint.



c	avg discounted total reward	avg discounted total cost	$\widehat{V}_C^*(s_0, b_0)$
50	-0.51 ± 0.11	30.18 ± 3.01	49.48 ± 0.26
40	-1.12 ± 0.28	15.99 ± 2.05	40.00 ± 0.00
30	-3.19 ± 0.48	7.03 ± 0.68	30.00 ± 0.00

Figure 4: Results of the cart pole domain. Above: histogram of deviations from the origin, where the the horizontal axis denotes $|x|$ and the vertical axis denotes the visitation rate.

Figure 3 depicts the risk-sensitive behaviour in the cliff domain, where the darkness mark the visitation frequency of each state. When $c = 100$ (weakly constrained), agent tries to reach the goal as fast as possible to maximize rewards. As we lower c, it starts to trade-off between reward and cost, making the agent take a detour to be safer.

Figure 4 summarizes the experimental result of the cart pole domain. As we lower c, the agent start to trade-off reward for cost, which makes the agent be more forced to stay in the middle by tolerating the fall downs. We can also see from the table that the total cost estimated by the algorithm mathces c. As for the computation time, it took 105 minutes to construct $\widehat{T}(s', b'|s, b, a)$ and 0.3 minute to solve the LP on a single CPU machine. We remark that most of the computation time is spent on computing the transition probabilies, which should be easily parallelizable using GPUs.

6 Conclusion and Future work

In this paper, we presented CBRL-ALP, a model-based BRL algorithm in CMDP environment to deal with the *safe exploration* in a principled way. We showed that the constrained BRL problem can be solved efficiently via approximate linear programming. Our theoretical analysis shows that the algorithm computes approximate Bayes-optimal value functions and the approximation error can be bounded by the coverage of sampled beliefs. Experimental results show the cost-sensitive behaviours and effectiveness of our algorithms empirically, outperforming the previous state-of-the-art approach, CBEETLE by orders of magnitude in computation time.

As for the future work, we plan to focus on developing scalable algorithm that can be applied to more challenging domains. Extension to online version of the algorithm or application of function approximation would be promising.

References

[Altman, 1999] Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

[Araya-López *et al.*, 2012] Mauricio Araya-López, Vincent Thomas, and Olivier Buffet. Near-optimal BRL using optimistic local transition. In *Proceedings of the 29th International Conference on Machine Learning*, pages 97–104, 2012.

[Asmuth *et al.*, 2009] John Asmuth, Lihong Li, Michael L. Littman, Ali Nouri, and David Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26, 2009.

[Dearden *et al.*, 1998] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 761–768, 1998.

[Dearden *et al.*, 1999] Richard Dearden, Nir Friedman, and David Andre. Model based Bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

[Duff, 2002] Michael O'Gordon Duff. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2002.

[García and Fernández, 2012] Javier García and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45(1):515–564, 2012.

[Hans *et al.*, 2008] Alexander Hans, Daniel Schneega, Anton M. Schäfer, and Steffen Udluft. Safe exploration for reinforcement learning. In *Proceedings of the European Symposium on Artificial Neural Network*, pages 143–148, 2008.

[Howard and Matheson, 1972] Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

[Hsu *et al.*, 2007] David Hsu, Wee Sun Lee, and Nan Rong. What makes some POMDP problems easy to approximate? In *Advances in Neural Information Processing Systems 20*, pages 689–696, 2007.

[Iyengar, 2005] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[Kim *et al.*, 2011] Dongho Kim, Jaesong Lee, Kee-Eung Kim, and Pascal Poupart. Point-based value iteration for constrained pomdps. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 1968–1974, 2011.

[Kim *et al.*, 2012] Dongho Kim, Kee-Eung Kim, and Pascal Poupart. Cost-sensitive exploration in Bayesian reinforcement learning. In *Advances in Neural Information Processing Systems 25*, pages 3068–3076, 2012.

[Kolter and Ng, 2009] J. Zico Kolter and Andrew Y. Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520, 2009.

[Mihatsch and Neuneier, 2002] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290, 2002.

[Minka, 1998] Thomas P. Minka. Bayesian linear regression. Technical report, Microsoft research, 1998.

[Nilim and El Ghaoui, 2005] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[Poupart *et al.*, 2006] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 697–704, 2006.

[Poupart *et al.*, 2015] Pascal Poupart, Aarti Malhotra, Pei Pei, Kee-Eung Kim, Bongseok Goh, and Michael Bowling. Approximate linear programming for constrained partially observable Markov decision processes. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3342–3348, 2015.

[Strens, 2000] Malcolm J. A. Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950, 2000.

[Sutton and Barto, 1998] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[Tziortziotis *et al.*, 2013] Nikolaos Tziortziotis, Christos Dimitrakakis, and Konstantinos Blekas. Linear Bayesian reinforcement learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.