Constrained Bayesian Reinforcement Learning via Approximate Linear Programming

Jongmin Lee[1], Youngsoo Jang[1], Pascal Poupart[2], and Kee-Eung Kim[1]

[1] School of Computing, KAIST, Republic of Korea
[2] David R. Cheriton School of Computer Science, University of Waterloo, Canada

Abstract. In this paper, we highlight our recent work [9] considering 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 (BRL) in an environment that is modeled as a constrained MDP (CMDP) where the cost function penalizes undesirable situations. We propose a model-based 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 [13]. 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) [2, 3, 12], 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 [10, 6, 4, 5]. 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.

In this paper, we consider the constrained MDP (CMDP) [1] 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. Specifically, following [7], we model BRL as a planning problem with the hyper-state constrained partially observable MDP (CPOMDP) [8] and adopt constrained approximate linear programming (CALP) [11] 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 [11] 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.

2 Background

We model the environment as a 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.

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

The optimal policy of a CMDP is generally stochastic and can be obtained by solving the following linear program (LP) [1].

$$\max_{\{y(s,a)\}\forall s,a} \sum_{s,a} R(s,a)y(s,a) \tag{1}$$
$$s.t. \quad \sum_{a'} y(s',a') = \delta(s_0,s) + \gamma\sum_{s,a} T(s'|s,a)y(s,a) \quad \forall s'$$
$$\sum_{s,a} C_k(s,a)y(s,a) \le c_k \quad \forall k \quad \text{and} \quad y(s,a) \ge 0 \quad \forall s,a$$

where $y(s,a)$ can be interpreted as a discounted occupancy measure of (s,a), and $\delta(x,y) = 1$ 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 $\pi^*(a|s) = y(s,a)/\sum_{a'} y(s,a')$ 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 $b_t(s) = \Pr(s_t = s)$ at every time step using the Bayes rule: upon executing a in b and observing o, the updated belief b^{ao} is $b^{ao}(s') \propto Z(o|s', a) \sum_s T(s'|s, a) b(s) \quad \forall s'$.

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.

$$\bar{T}(b'|b, a) = \sum_o \left[\sum_{s,s'} Z(o|s', a) T(s'|s, a) b(s) \right] \delta(b', b^{ao}) \tag{2}$$

Similarly, the reward and cost functions are represented as $\bar{R}(b, a) = \sum_s b(s) R(s, a)$ and $\bar{C}_k(b, a) = \sum_s b(s) C_k(s, a)$. 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 [8], and CALP, which is based on linear programming [11]. CALP has been shown to perform much better than CPBVI.

3 Constrained BRL via Approximate LP

3.1 Constrained BRL as CPOMDP Planning

Model-based BRL computes a full posterior distribution over the transition models and uses 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 [3, 12, 7], 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 $b'(\theta) = b^{\bar{s}\bar{a}\bar{s}'}(\theta) = \eta b(\theta) \theta^{\bar{s}\bar{a}\bar{s}'} = \prod_{s,a} \mathrm{Dir}\big(\theta^{sa*} | n^{sa*} + \delta((\bar{s}, \bar{a}, \bar{s}'), (s, a, s'))\big)$ 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

$$\bar{T}^+(\langle s', b' \rangle | \langle s, b \rangle, a) = \Pr(s'|s, b, a) \Pr(b'|s, b, a, s') = \mathbb{E}_b \left[\theta^{sas'} \right] \delta(b', b^{sas'}) \quad (3)$$

Similarly, the reward function and the cost functions are $\bar{R}^+(\langle s, b \rangle, a) = R(s, a)$ and $\bar{C}_k^+(\langle s, b \rangle, a) = C_k(s, a)$. In theory, this belief-state CMDP can be solved using the following LP, which is an extension of (1) and the one in [11] to treat hyper belief states:

$$\max_{\{y(s,b,a)\} \forall s,a} \sum_{s,b,a} R(s, a) y(s, b, a) \quad (4)$$

$$s.t. \quad \sum_{a'} y(s', b', a') = \delta((s_0, b_0), (s, b)) + \gamma \sum_{s,b,a} T(s', b'|s, b, a) y(s, b, a) \quad \forall s', b'$$

$$\sum_{s,b,a} C_k(s, a) y(s, b, a) \leq c_k \quad \forall k \quad \text{and} \quad y(s, b, a) \geq 0 \quad \forall s, b, a$$

3.2 Approximate Linear Programming

The main challenge in solving the linear program (4) 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 (4) 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 $\forall b \in \widehat{B},\ s, s' \in S,\ a \in A, \quad \min_{b' \in \widehat{B}} \| b' - b^{sas'} \|_1 \leq \epsilon$ 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 $T(s', b'|s, b, a)$ in Eq. (3), we relax $\delta(b', b^{sas'})$ to $W(b'|b^{sas'})$ that has non-zero probability only for ϵ-close beliefs:

$$\widehat{T}(s', b'|s, b, a) = \Pr(s'|s, b, a) \widehat{\Pr}(b'|s, b, a, s') = \mathbb{E}_b \left[\theta^{sas'} \right] W(b'|b^{sas'}), \quad (5)$$

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

$$W(b'|b^{sas'}) = \begin{cases} \kappa K(b', b^{sas'}) & \text{if } \| b' - b^{sas'} \|_1 \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where κ is the normalizing constant $\sum_{b' \in \widehat{B}} W(b'|b^{sas'}) = 1$ and $K(b, b') \geq 0$ 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. Thus, we approximate the original LP (4) by using a finite set of beliefs $\widehat{B}$

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 (6)
  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 (4) 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: (π, W): finite state controller, $\widehat{V}^*_R(s_0, b_0)$: approximate Bayes-optimal value

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.

We remark that [11] takes a similar approach to solving CPOMDPs by considering finitely sampled beliefs. Specifically, they approximate the transitions by relaxing $\delta(b', b^{ao})$ in (2) 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 the main result that bounds the error in the value function due to approximate LP incurred by taking a finite set of beliefs and using the 'slip to ϵ-close belief' approximation. The full proof is provided in [9].

Theorem 1. *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:*

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

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
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

Table 1: Experimental results. The results with (*) are from [7].

where τ represents how much we can reduce the cost constraint without making the problem infeasible intuitively.

5 Experiments

We conducted experiments on 3 discrete state domains (chain, maze, and cliff) and 1 continuous domain (cartpole). The detailed domain description and the experimental setup are presented in [9]. In this paper, we only show some parts of the whole experimental results due to the page limit.

Table 1 summarizes the experimental results for discrete domains, comparing our algorithm CBRL-ALP to the previous state-of-the-art approach CBEETLE [7]. Overall, our method outperforms CBEETLE in computation speed by an order of magnitude, while yielding good policies. Besides, we can see from the table that agent starts to trade-off between reward and cost as we lower c.

6 Conclusion

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.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2016-0-00464) supervised by the IITP (Institute for Information & communications Technology Promotion) 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).

References

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

[2] 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.

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

[4] 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.

[5] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

[6] 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.

[7] 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.

[8] 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.

[9] Jongmin Lee, Youngsoo Jang, Pascal Poupart, and Kee-Eung Kim. Constrained Bayesian reinforcement learning via approximate linear programming. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

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

[11] 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.

[12] 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.

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