ACML 2017 11.15-17 Seoul, Korea
Asian Conference on Machine Learning

IJCAI-17 MELBOURNE

Constrained Bayesian Reinforcement Learning via Approximate Linear Programming

2017. 11. 15

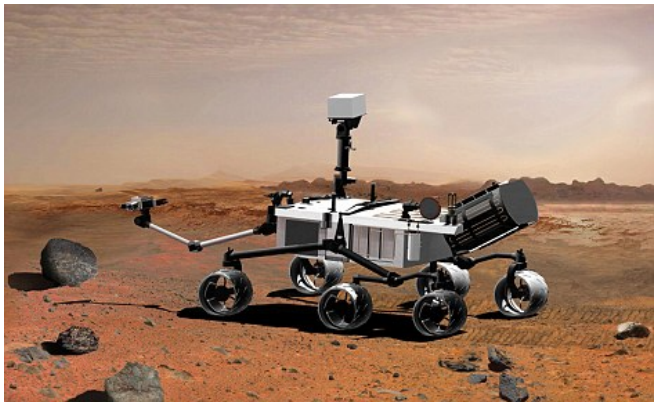
Jongmin Lee (KAIST)
Youngsoo Jang (KAIST)
Pascal Poupart (Univ. of Waterloo)
Kee-Eung Kim (KAIST)

KAIST

UNIVERSITY OF WATERLOO

Safe Reinforcement Learning ?

Helicopter control

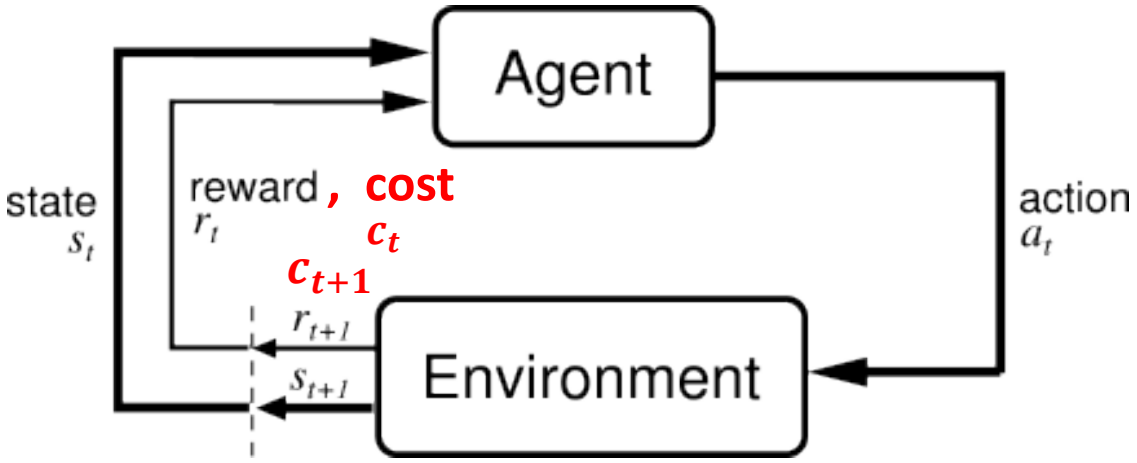


Mars rover

In many situations, ***safety*** concept is crucial to RL agent

⇒ **Can be modelled as cost-constrained RL**

Safe RL ⇒ Cost-constrained RL



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

$$\text{subject to} \quad V_C(s_0) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t) \right] \leq c$$

- Cost function $C(s, a)$:
  ⇒ Risk of the behavior
- Cost constraint c:
  ⇒ Degree of risk taking

Goal

- Propose **model-based Bayesian RL** algorithm for **constrained MDP** (CMDP) environment

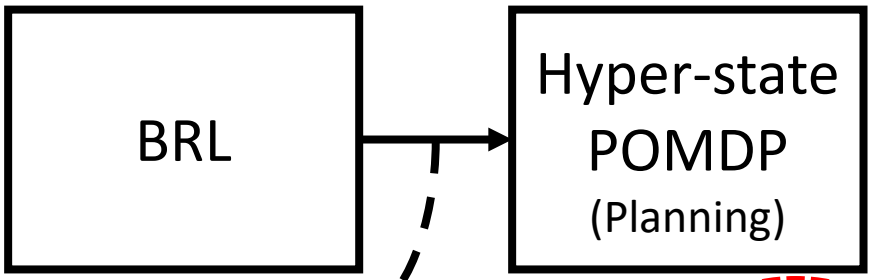
$$\langle S, A, T, R, C, c, \gamma, s_0 \rangle$$

⇒ Risk-sensitive exploration in a principled way

Assumption:

- Reward/cost function is known
- Only transition function is unknown

Constrained BRL to Belief-State CMDP



$S^+ = S \times \{\theta^{sas'}\}$

Belief-state $= (s, b(\theta))$

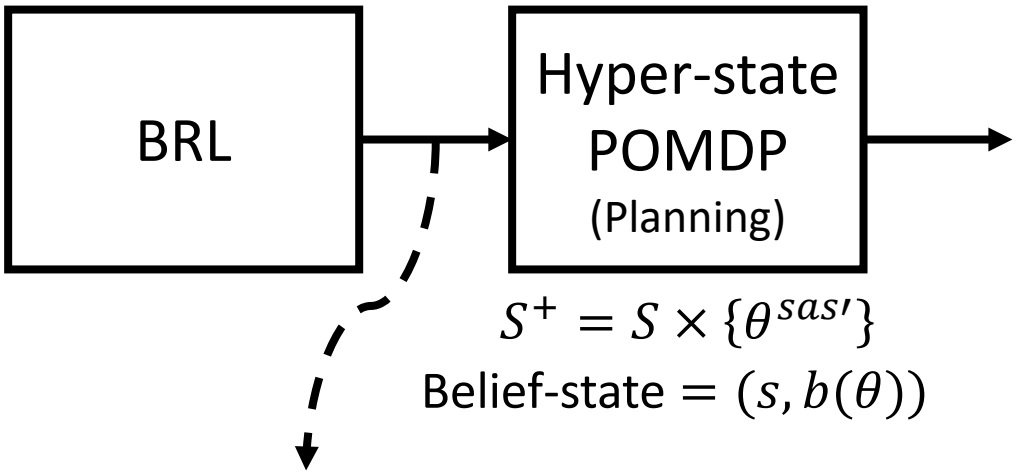
(Unknown) transition probability
$T(s'|s,a) = \theta^{sas'}$

[Duff, 2002]
- Formulate model-based Bayesian RL as POMDP

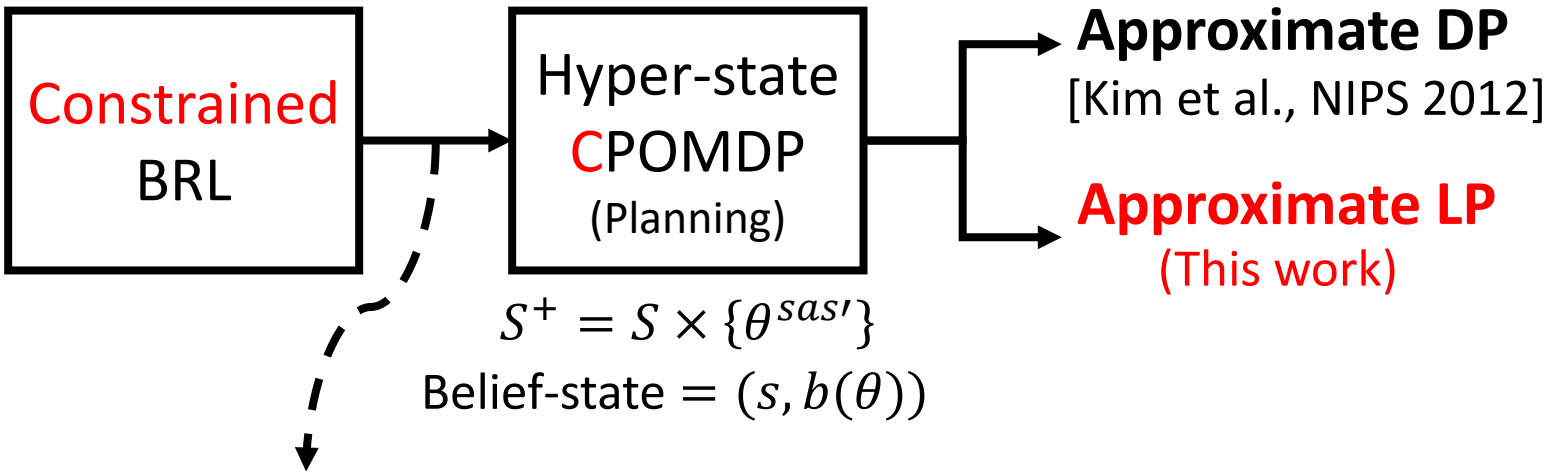
$$b(\boldsymbol{\theta}) = Pr(\boldsymbol{\theta}) = \prod_{s,a} Dir(\theta^{sa*}|n^{sa*})$$

Posterior distribution over transition model updated by (s,a,s') observations

Constrained BRL to Belief-State CMDP



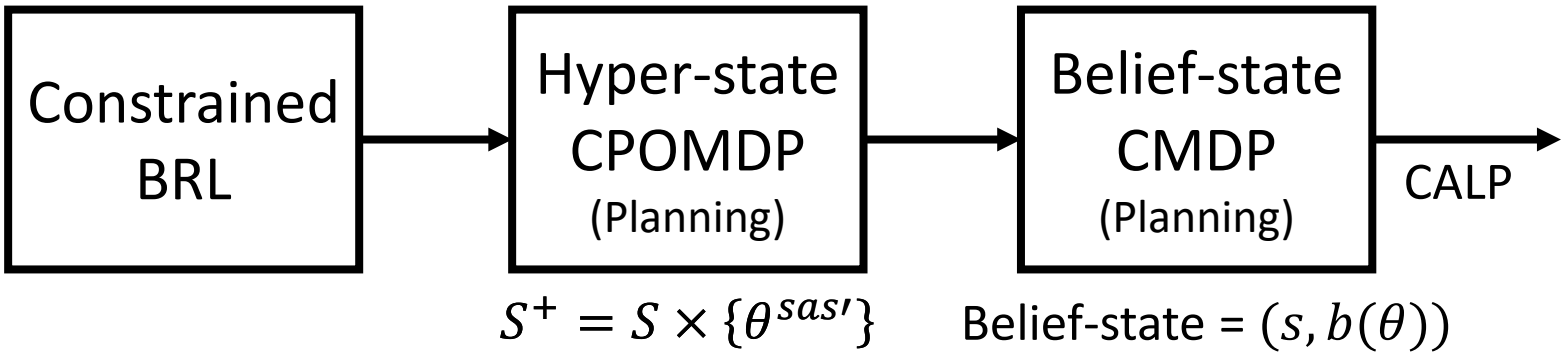
Constrained BRL to Belief-State CMDP



[Kim et al., NIPS 2012]
- Formulate **constrained** model-based Bayesian RL as **C**POMDP
- Solve CPOMDP based on approximate **dynamic programming**

[Poupart et al., AAAI 2015]
- State-of-the-art CPOMDP solver using **approximate linear programming**
- Think of CPOMDP as belief-state CMDP
- Construct approximate transition among sampled finite beliefs using **convex combination weight** (Difficult to apply to BRL directly)

Constrained BRL to Belief-State CMDP



1. **Constrained BRL** problem ⇒ hyper-state **CPOMDP planning** problem

2. **CPOMDP planning** problem ⇒ **belief-state CMDP planning** problem

3. Adopt constrained **approximate linear programming** (CALP) to efficiently compute the Bayes-optimal policy in an offline manner

Background: LP Formulation of CMDP

$$\underset{\{y(s,a)\}\ \forall s,a}{\text{maximize}} \sum_{s,a} y(s,a)R(s,a)$$

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

$$\sum_{s,a} y(s,a)C(s,a) \leq c$$

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

$$\max_{\pi} V_R^{\pi}(s_0)$$

$$\text{s.t. } V_C^{\pi}(s_0) \leq c$$

- $y(s,a)$: (discounted) occupancy frequencies of (s,a) pair.

$$y^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}(s_t = s, a_t = a)\right]$$

$$\delta(s_1,s_2) = \begin{cases} 1 & \text{if } s_1 = s_2 \\ 0 & \text{if } s_1 \neq s_2 \end{cases}$$

- Optimal policy $\pi^*(a|s) = \frac{y^*(s,a)}{\sum_{a'} y^*(s,a')}$

LP Formulation of Belief-State CMDP

* $y(s,a) \rightarrow y(s,b,a)$ and $T(s'|s,a) \rightarrow T(s',b'|s,b,a)$

$$\underset{\{y(s,b,a)\}\forall s,b,a}{\text{maximize}} \sum_{s,b,a} R(s,a)y(s,b,a)$$

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

$$\sum_{s,b,a} C(s,a)y(s,b,a) \leq c$$

$$y(s,b,a) \geq 0 \quad \forall s,b,a$$

$$\max_{\pi} V_R^{\pi}(s_0,b_0)$$

$$\text{s.t. } V_C^{\pi}(s_0,b_0) \leq c$$

- $y(s,b,a)$: occupancy measure of $(\langle s,b\rangle, a)$
- $T(\langle s',b'\rangle|\langle s,b\rangle, a)$: transition probability among belief-states $\langle s,b\rangle$
- Policy is computed as: $\pi^*(a|s,b) = y(s,b,a)/\sum_{a'} y(s,b,a')$

LP Formulation of Belief-State CMDP

$$
\begin{aligned}
&\underset{\{y(s,b,a)\}\forall s,b,a}{\text{maximize}} && \sum_{s,b,a} R(s,a)y(s,b,a) \\
&\text{subject to} && \sum_{a'} y(s',b',a') = \delta((s_0,b_0),(s,b)) + \gamma \sum_{s,b,a} T(s',b'|s,b,a)y(s,b,a) \quad \forall s',b' \\
& && \sum_{s,b,a} C(s,a)y(s,b,a) \leq c \\
& && y(s,b,a) \geq 0 \quad \forall s,b,a
\end{aligned}
$$

- Problem: $|S \times B| = \infty$
- Proposed approach
  - Finitely approximate the set of beliefs: $S \times \hat{B}$
  - Reconstruct approximate transition $\hat{T}(s',b'|s,b,a)$
  - Solve approximate CMDP via LP

'Slip to ϵ-close beliefs' approximation

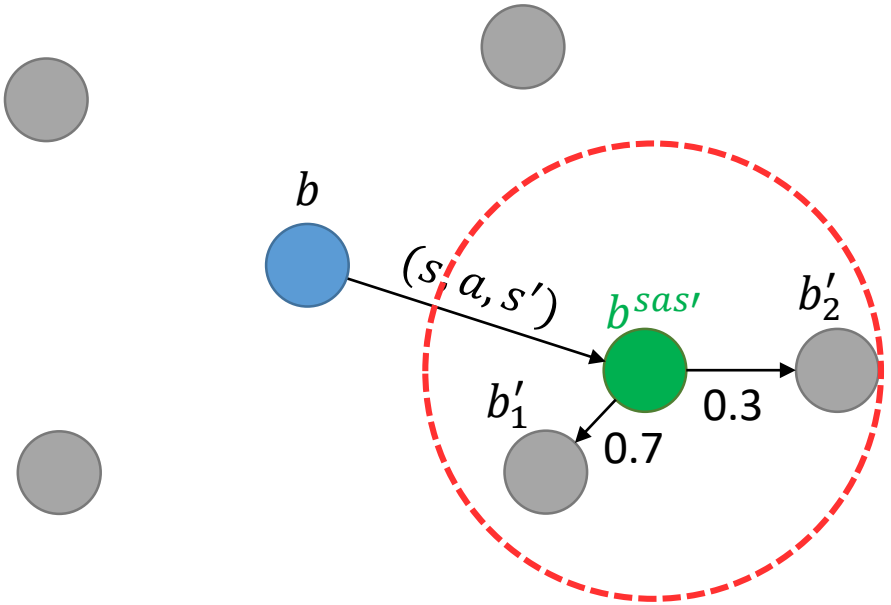
Approximate $\widehat{T}$: Slip to ϵ-close beliefs

$\Pr(s'|s,b,a)\Pr(b'|s,b,a,s')$

(Exact transition) $T(s',b'|s,b,a) = \mathbb{E}_b\left[\theta^{sas'}\right]\delta(b', b^{sas'})$

(Approximate transition) $\hat{T}(s',b'|s,b,a) = \mathbb{E}_b\left[\theta^{sas'}\right]W(b'|b^{sas'})$

(non-zero only for ϵ-close belief)



Ex)

$W\left(b'_1 \middle| b^{sas'}\right) = 0.7$

$W\left(b'_2 \middle| b^{sas'}\right) = 0.3$

ϵ-close boundary of $b^{sas'}$

Algorithm

1. Finite belief set S × $\hat{B} = \{(s, b)\}$ is given.
2. Construct approximate transition $\hat{T}(s', b'|s, b, a)$
3. Solve the following LP: $(s, b) \in S \times \hat{B}$ and $(s', b') \in S \times \hat{B}$

$$
\begin{aligned}
&\underset{\{y(s,b,a)\}\forall s,b,a}{\text{maximize}} && \sum_{s,b,a} R(s,a)y(s,b,a) \\
&\text{subject to} && \sum_{a'} y(s',b',a') = \delta((s_0,b_0),(s,b)) + \gamma \sum_{s,b,a} \hat{T}(s',b'|s,b,a)y(s,b,a) \quad \forall s', b' \\
& && \sum_{s,b,a} C(s,a)y(s,b,a) \leq c \\
& && y(s,b,a) \geq 0 \quad \forall s, b, a
\end{aligned}
$$

4. Policy is computed as $\hat{\pi}^*(a|s, b) := \frac{y(s,b,a)}{\sum_{a'} y(s,b,a')}$

Theoretical Analysis

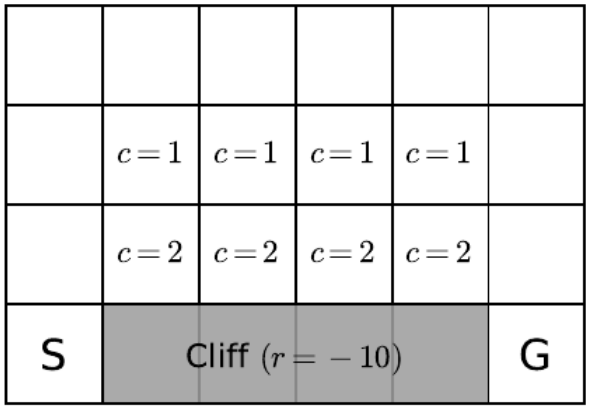
- Error incurred by 'slip to ϵ-close beliefs' approximation can be bounded by the coverage of sampled beliefs

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

  - V_R^*: Bayes-optimal value function
  - $\widehat{V}_R^*$: Approximate Bayes-optimal value function

Experiment – Domain

Discrete State Domain – Cliff



(c) Cliff

Reward:

+20 to goal position

-10 to cliff

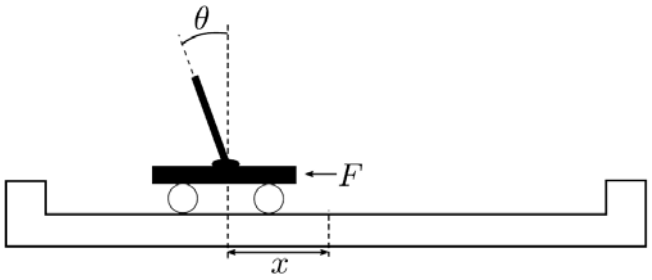
Cost:

0 to low-risk area

1 to medium-risk area

2 to high-risk area

Continuous State Domain – Cart pole



(d) Cart pole

Reward:

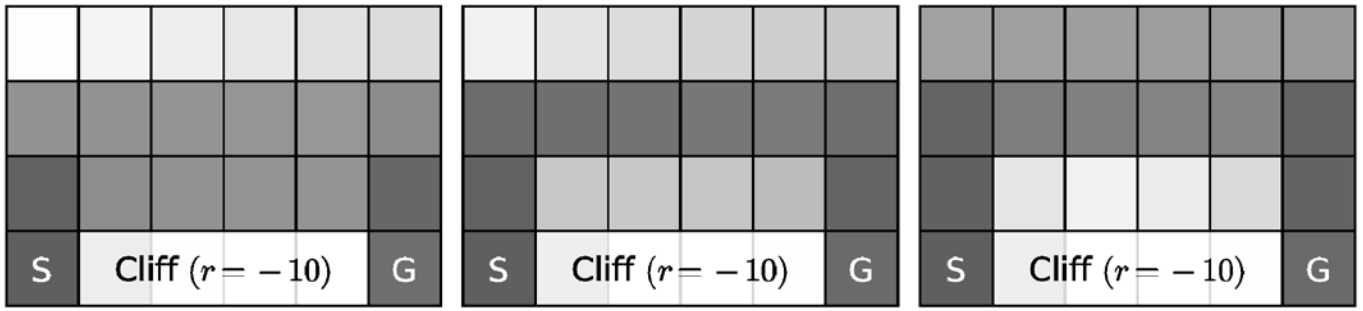
-1 if the pole falls down

Cost:

$|x|$ = (deviation from origin)

Experiment – Visual Result

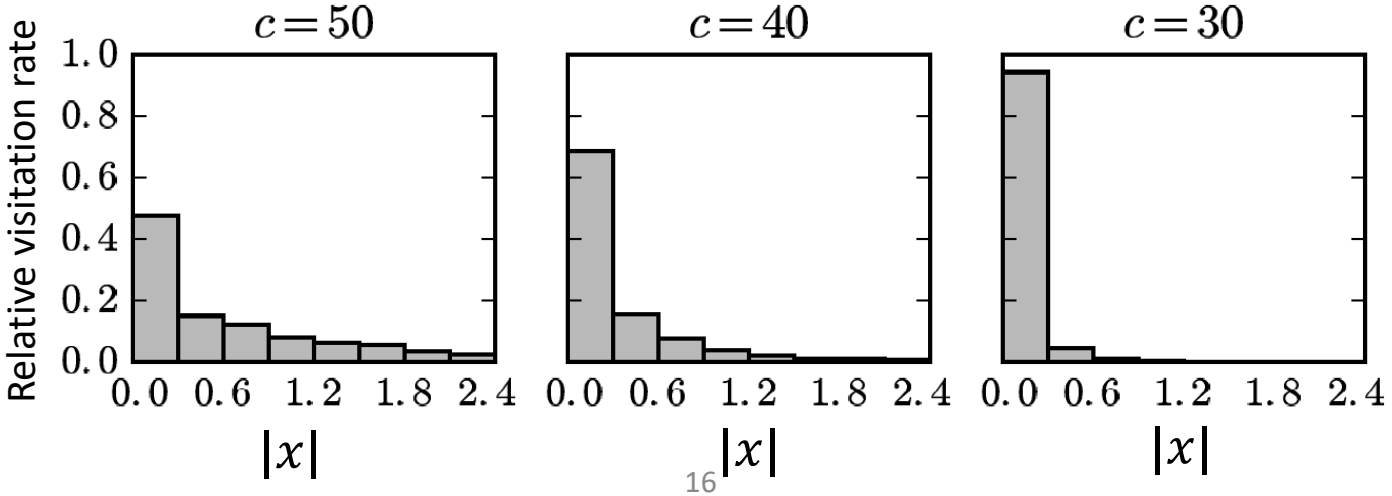
Cliff



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

* Darkness: proportional to visitation frequency

Cart pole



Experiment – Numerical Result

Cliff

c	algorithm	avg discounted total reward	avg discounted total cost	time (min)
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**

* CBEETLE: [Kim et al., NIPS 2012] / CBRL-ALP: Ours

Cart pole

c	avg discounted total reward	avg discounted total cost	$\hat{V}_C^*(s_0, b_0)$
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

Conclusion

- Risk-sensitive behaviors can naturally be encoded into CMDP framework.

- We proposed model-based BRL algorithm for CMDP environment.
  - Outperforms the previous approach by orders of magnitude in computation time

Thank you !