Constrained POMDPs

Bellairs Institute

Holetown, Barbados

April 30, 2013

Dongho Kim, Jaesong Lee, Kee-Eung Kim (KAIST, Korea)
Pascal Poupart (University of Waterloo, Canada)

Motivation

- Multiple objectives
- Performance guarantees

- Spoken dialog system
  - First objective: task completion (e.g., correct booking)
  - Second objective: minimize dialog length

- Operations research
  - Stock holders' goal: maximize profits
  - Employees: maximize wages

Reinforcement Learning

- Dual objective
  - Exploration and Exploitation
- Chatbot for entertainment
  - Exploitation: maximize conversation length
  - Exploration: try new responses
- Exploration needs to be controlled to avoid frustrating initial users with weird responses.

Classic solutions

- Weighted combination of objectives

$$R(s,a) = w_1 R_1(s,a) - w_2 C_2(s,a) - w_3 C_3(s,a)$$

- Pareto optimal solutions

Proposal

- Model each objective separately
  - Maximize one objective (reward)
  - Bound other objectives (costs)

Constrained POMDP

- $CPOMDP = \langle S, A, O, T, Z, R, \{\boldsymbol{C_k}\}, \{\hat{\boldsymbol{c}}_{\boldsymbol{k}}\}, \gamma, b_0 \rangle$
  - S: states, A: actions, O: observations
  - Transition prob: $T(s', s, a) = \Pr(s'|s, a)$
  - Observation prob: $Z(o, s', a) = \Pr(o|s', a)$
  - Rewards: $R(s, a)$
  - **Costs: $\boldsymbol{C_k(s, a)}\ \forall \boldsymbol{k}$**
  - **Constraints: $\mathbf{E}[\sum_t \boldsymbol{\gamma^t C_k(s_t, a_t)}] \leq \hat{\boldsymbol{c}}_{\boldsymbol{k}} \quad \forall \boldsymbol{k}$**
  - Discount: γ
  - Initial belief: b_0

Type of Constraints

- Bound on expected costs
- Bound on probability of reaching some states

- What about hard constraints?
  - Deterministic costs: yes
  - Stochastic costs: no

- What about bounded variance?
  - Not in this framework

Constrained MDP

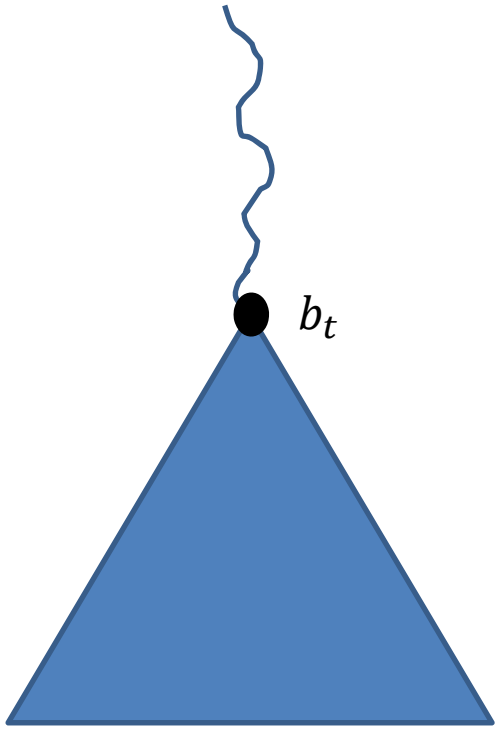
- LP formulation [Altman, 1999]

$$\max_{x} \sum_{s,a} R(s,a)x(s,a)$$
$$\text{s.t.} \quad \sum_{a} x(s',a) = b_0(s') + \gamma \sum_{s,a} \Pr(s'|s,a)\, x(s,a) \;\forall s'$$
$$\sum_{s,a} C_k(s,a)x(s,a) \leq \hat{c}_k \;\; \forall k$$
$$x(s,a) \geq 0 \;\; \forall s,a$$

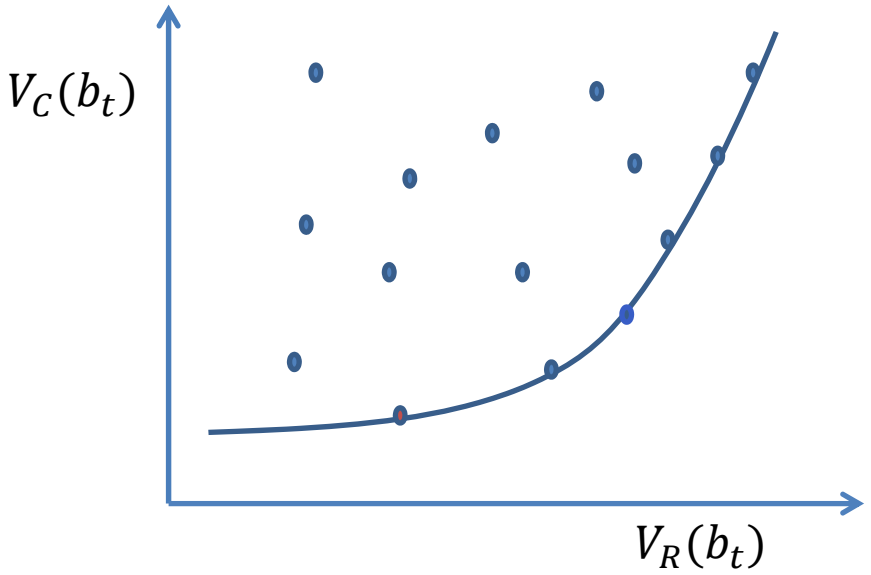
- Is there a *deterministic* optimal policy?
  - Not always: may need to up to K stochastic actions (K = # of cost constraints)

Constrained POMDP

- Idea: compute value function for each objective ($V_R(b_t)$ and $V_C(b_t)$)

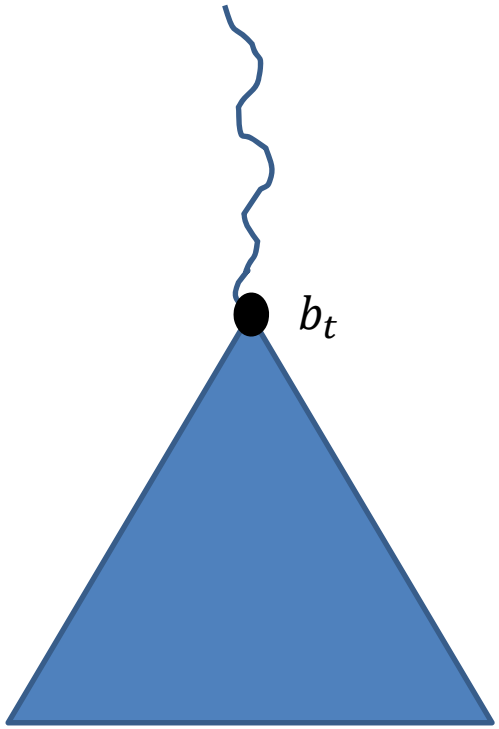


Max $V_R(b_t)$ or Min $V_C(b_t)$

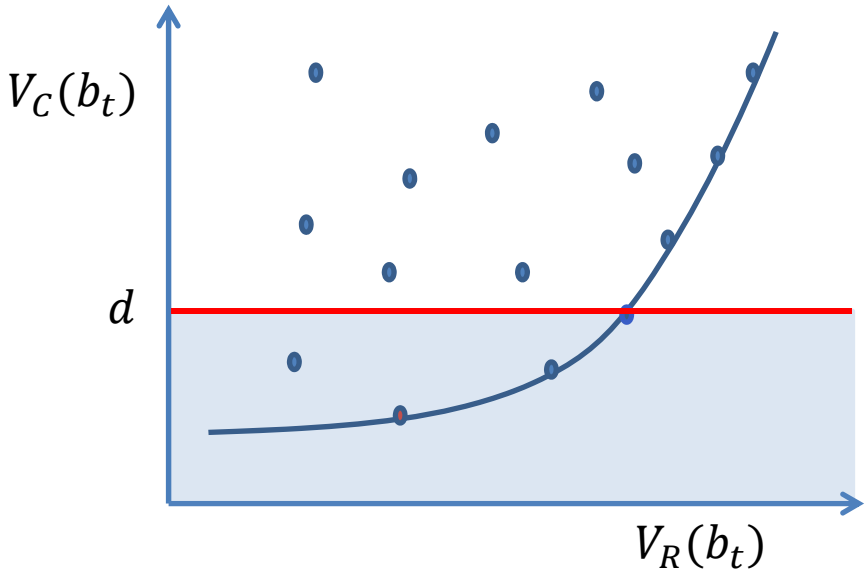


Constrained POMDP

- Admissible cost: $d = (c - \sum_{i=0}^{t} \gamma^i b(s_i) C(s_i, a_i))/\gamma^t$



Max $V_R(b_t)$ s.t. $V_C(b_t) \leq d$



Point-based Value Iteration

- Sample set of reachable $\langle b, d \rangle$ pairs
- Compute $\langle V_R, V_C \rangle$ pair for each $\langle b, d \rangle$ pair by linear programming

$$\max_{\Pr(a|b)} \underbrace{\sum_a \Pr(a|b) \left[R(b,a) + \gamma \sum_o \Pr(o|b,a)\, V_R(b_{ao}) \right]}_{V_R(b)}$$

$$s.t. \underbrace{\sum_a \Pr(a|b) \left[C(b,a) + \gamma \sum_o \Pr(o|b,a)\, V_C(b_{ao}) \right]}_{V_C(b)} \leq d$$

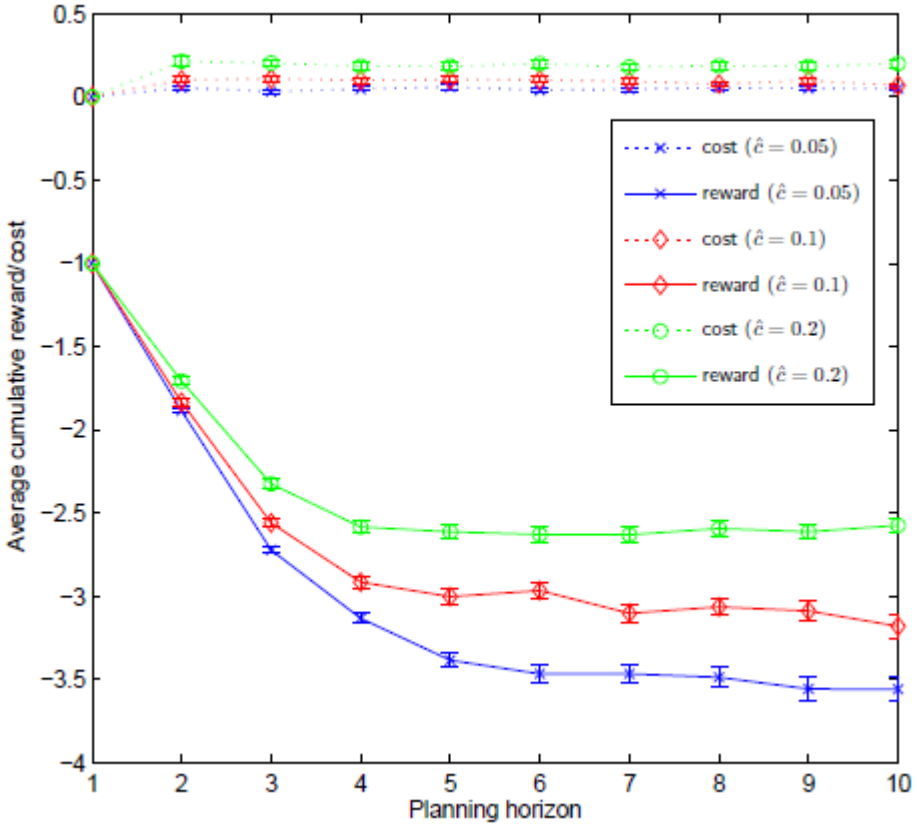
Cost-sensitive Bayesian RL

- CMDP = $\langle S, A, T, R, \{C_k\}, \{c_k\}, \gamma, s_0 \rangle$
- Cost-sensitive BRL = $\langle \hat{S}, A, \hat{O}, \hat{T}, \hat{R}, \{\hat{C}_k\}, \{c_k\}, \hat{Z}, \gamma, b_0 \rangle$
  - $\hat{S} = \langle S, \theta \rangle$ where θ's are parameters of T
  - $\hat{O} = S$
  - $\hat{T}: \Pr(s', \theta | s, \theta, a) = \Pr(s' | s, \theta, a)\, \delta(\theta = \theta')$
  - $\hat{R}(s, \theta, a) = R(s, a)$, $\hat{C}_k(s, \theta, a) = C_k(s, a)$
  - $\hat{Z}: \Pr(s_i' | s_j', \theta', a) = \delta(s_i' = s_j')$

- C-Beetle: point-based value iter. at $\langle s, b(\theta), d \rangle$ triples

Empirical Results

- Ticketing spoken dialog system



Empirical Results

- Cost-sensitive Bayesian RL

Table 1: Experimental results for the chain and maze problems.

problem	algorithm	$\hat{c}$	utopic value	avg discounted total reward	avg discounted total cost	time (minutes)
chain-tied $\|S\| = 5$ $\|A\| = 2$	BEETLE	–	354.77	351.11±8.42	–	1.0
	CBEETLE	100	354.77	354.68±8.57	100.00±0	2.4
		75	325.75	287.70±8.17	75.05±0.14	2.4
		50	296.73	264.97±7.06	49.96±0.09	44.3
		25	238.95	212.19±4.98	25.12±0.13	80.59
chain-semi $\|S\| = 5$ $\|A\| = 2$	BEETLE	–	354.77	351.11±8.42	–	1.6
	CBEETLE	100	354.77	354.68±8.57	100.00±0	3.7
		75	325.75	287.64±8.16	75.05±0.14	3.8
		50	296.73	256.76±7.23	50.09±0.14	70.7
		25	238.95	204.84±4.51	25.01±0.16	139.3
maze-tied $\|S\| = 264$ $\|A\| = 5$	BEETLE	–	1.03	1.02±0.02	–	159.8
	CBEETLE	20	1.03	1.02±0.02	19.04±0.02	242.5
		18	0.97	0.93±0.04	17.96±0.46	733.1

Conclusion

- Summary
  - Constraint-based planning and RL
  - Approach to deal with multiple objectives
  - Performance guarantees

- Future work:
  - Online constraint-based planning
  - Demonstrate on a real-world dialog problem

Other work related to this workshop

- Chatbot lifelong reinforcement learning
  - In collaboration with Kik Interactive Inc
- Performance guarantees in POMDP planning
  - GapMin: 3 significant digit matching bounds
- Planning for mobile/embedded devices
  - Resource constrained policy execution