EPSRC Engineering and Physical Sciences Research Council

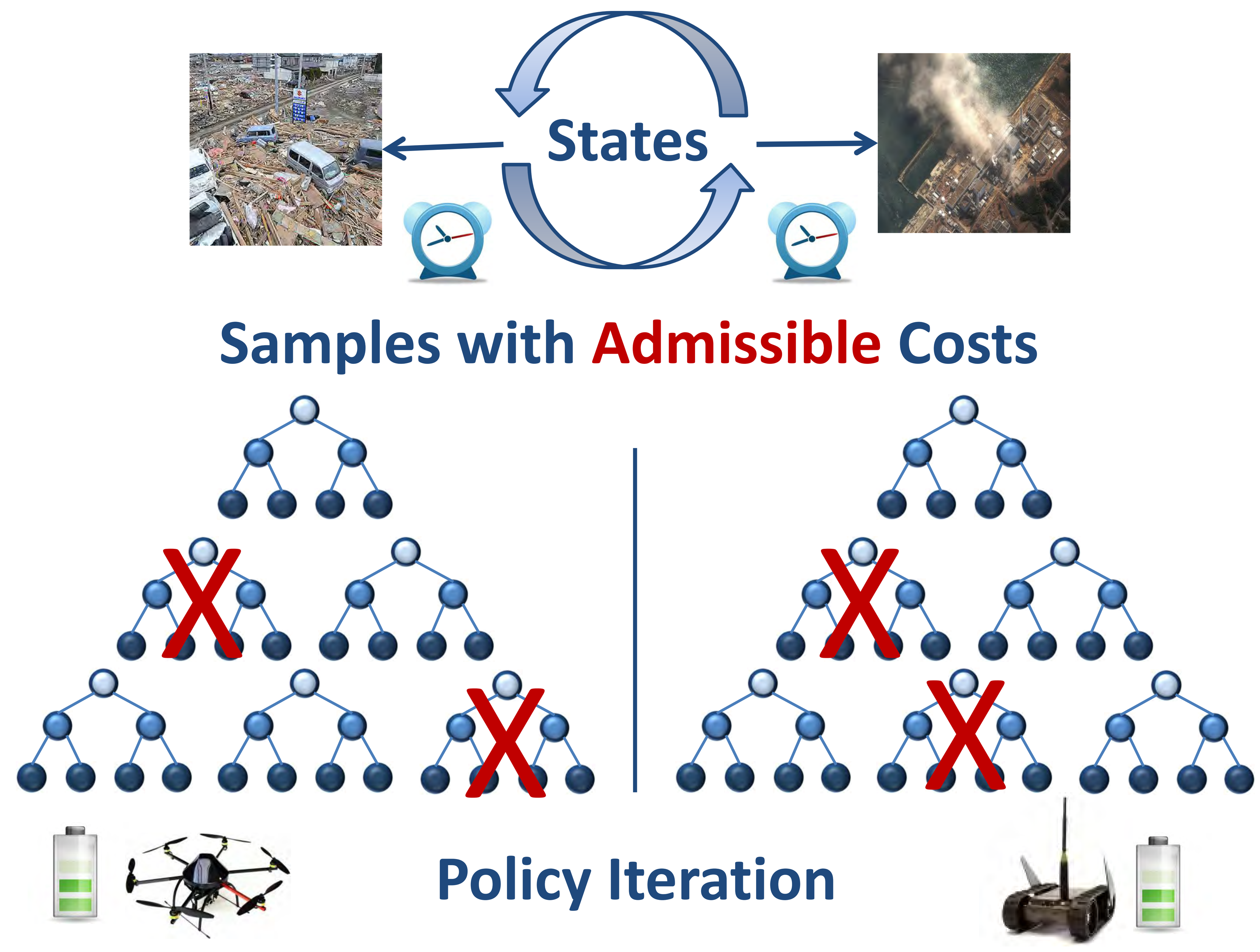
ORCHID

Sample-Based Policy Iteration for Constrained DEC-POMDPs

Feng Wu, Nicholas R Jennings, and Xiaoping Chen
Agents, Interaction, and Complexity Research Group
School of Electronics and Computer Science
University of Southampton

Motivation

Sample-based policy iteration

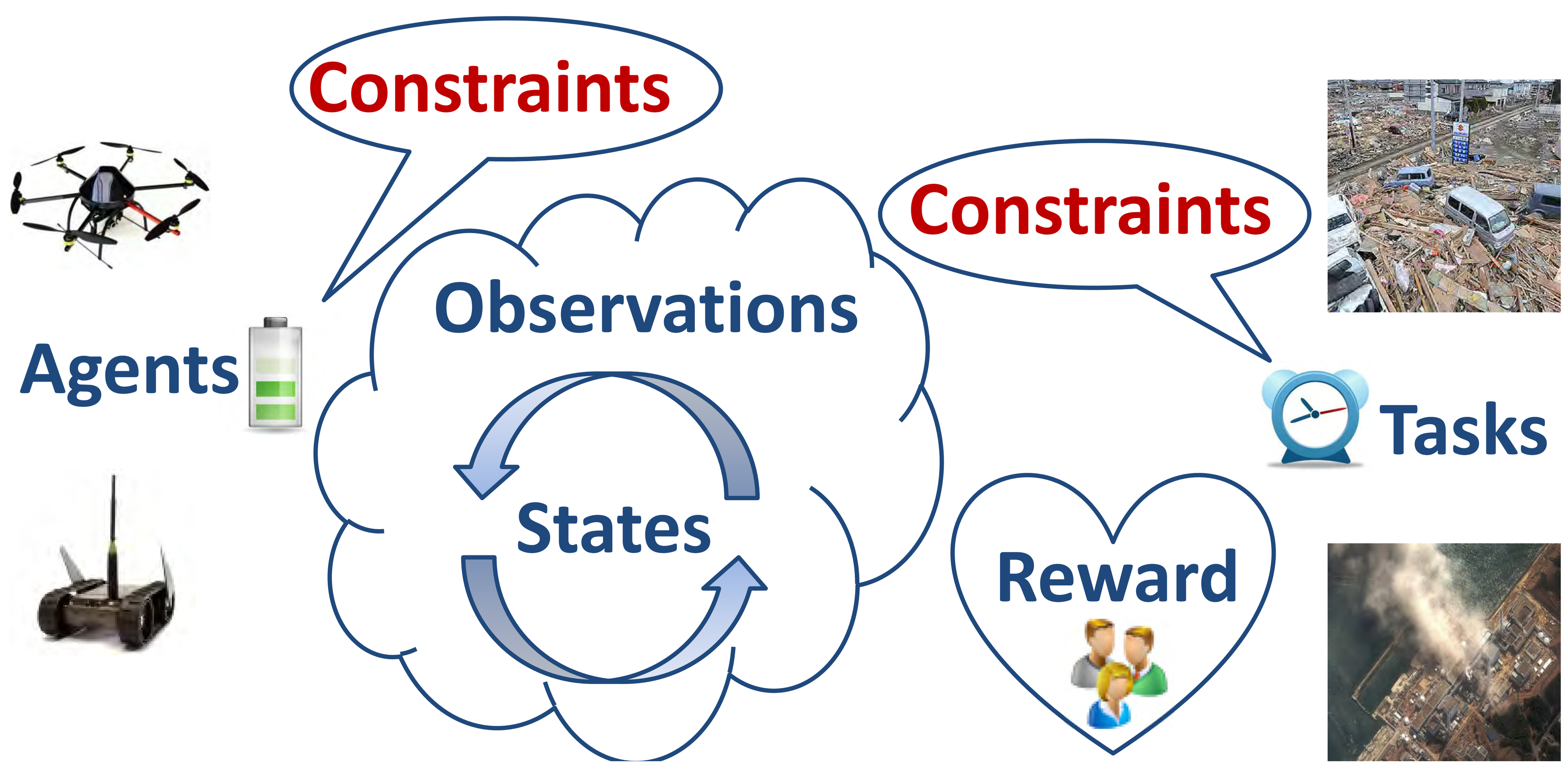


Constrained DEC-POMDPs

DEC-POMDP: $< I, A, R, S, T, \Omega, O>$

\+
- C_k: Constraint function
- c_k: Limited budget

$$E\left[\sum C_k(s,\vec{a})\right] \le c_k$$



Technical details

- Belief and cost sampling:
  1. Reachable belief: $b(s) = \frac{1}{w}\sum_{j=1}^{N}\{w_j : s_j = s\}$
  2. Admissible cost: $d_k = \sum_t C_k(s^t, \vec{a}^t)$

  where $w_j = \prod_{i\in I} p(q_i^t|q_i^{t-1}, o_i^t)$ and $w = \sum_j w_j$.
- Policy improvement:

  $$\max_{x,y} \sum_{\vec{a}} \prod_{i\in I} x_{a_i|q_i}[R_{b,\vec{a}} + \sum_{s',\vec{o}} P_{s',\vec{o}\,|b,\vec{a}} \sum_{\vec{q}\,'} \prod_{i\in I} y_{q_i'|q_i,a_i,o_i} V_{s',\vec{q}\,'}] \textbf{ s.t.}$$

  1. The cost constraints:

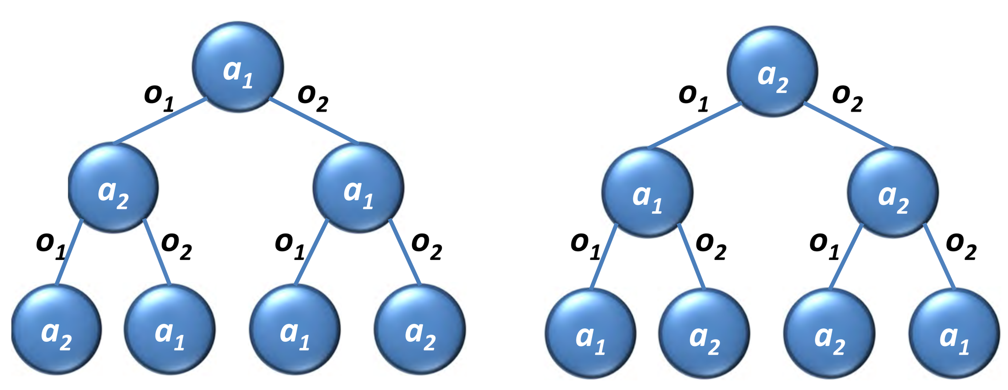
  $$\sum_{\vec{a}} \prod_{i\in I} x_{a_i|q_i}[C_{b,\vec{a}} + \sum_{s',\vec{o}} P_{s',\vec{o}\,|s',\vec{a}} \sum_{\vec{q}\,'} \prod_{i\in I} y_{q_i'|q_i,a_i,o_i} U_{s',\vec{q}\,'}] \le c_k - d_k$$

  2. The probability constraints:

  $$\sum_{a_i\in A_i} x_{a_i|q_i} = 1; \forall a_i \in A_i, o_i \in \Omega_i, \sum_{q_i'\in Q_i} y_{q_i'|q_i,a_i,o_i} = x_{a_i|q_i}$$

  where $x_{a_i|q_i}, y_{q_i'|q_i,a_i,o_i}$ are variables of each agent i's policy, b is the sampled joint belief state, and d_k is the admissible costs.
- Repeat the above two steps until converge.

Policy tree representation



- A local policy of agent i, $q_i : \Omega_i^* \to A_i$, and a joint policy is a set of local policies, $\vec{q} = \langle q_1, q_2, \cdots, q_n\rangle$, one for each agent.
- The value of a joint policy $\vec{q}$ is defined as

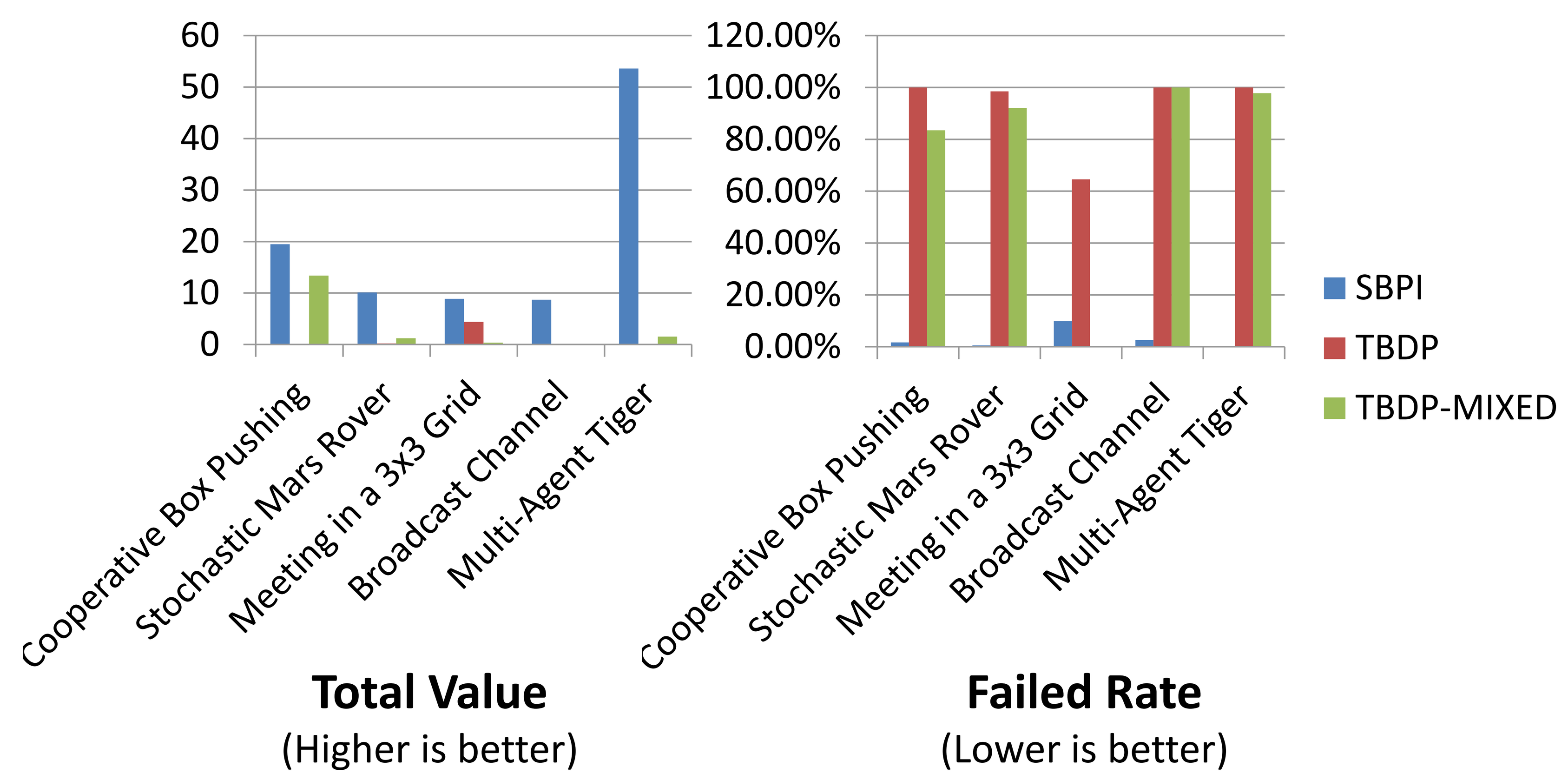
  $$V(s,\vec{q}) = R(s,\vec{a}) + \sum_{s'\in S} P(s'|s,\vec{a}) \sum_{\vec{o}\in\vec{\Omega}} O(\vec{o}|s',\vec{a}) V(s',\vec{q}_{\vec{o}})$$

- The goal is to find a joint policy $\vec{q}^*$ that maximizes

  $$V(b,\vec{q}^*) = \max_{\vec{q}} \sum_s b^0(s) V(s,\vec{q})$$

.

Experiments



Results of DEC-POMDP Common Benchmark Problems
(TBDP: [Wu et al., AAAI 2010])

BAE SYSTEMS ACFR SECURE University of Southampton