Scalable Decision-Theoretic Planning in Open and Typed Multiagent Systems

Adam Eck[1], Maulik Shah[2], Prashant Doshi[2], and Leen-Kiat Soh[3]

[1]Oberlin College, [2]University of Georgia, [3]University of Nebraska

aeck@oberlin.edu, mns28652@uga.edu, pdoshi@uga.edu, lksoh@cse.unl.edu

Problem: Scalable Planning in Open Environments

Many-Agent Environments: environments with dozens to thousands of agents

Open Environments: agents join and leave the environment (temporarily or permanently) over time

- ***Real-world Examples***: wildfire fighting, autonomous ridesharing, cybersecurity
- **Problem**: Openness requires agents to not only predict what actions their peers will take in order to choose a best response, but also ***first estimate which peers will even be present to take actions***

- Tracking the presence of neighbors results in a ***more complicated problem model***. The increase in the problem model size is:
  - ***exponential*** if agent presence is considered in the environment state, which is intractable, especially for many-agent environments
  - only ***linear*** if agent presence is considered within the mental models of each agent in an I-POMDP-Lite problem model (or other I-POMDP variants)

- Larger planning problem ***affects scalability*** as the number of agents increases (to many-agents)
  - MCTS algorithms: each trajectory requires sampling actions for all neighbors, so more agents results in fewer sampled trajectories in a fixed time budget (for responsive reasoning)
  - With openness, time is also spent updating estimates of presence of each neighbor, further reducing the number of trajectories possible within a time budget
  - **Frame-action Anonymity** offers some relief by replacing joint actions with counts of actions called **configurations *C*** when environment dynamics ***do* not *depend on* which *agents*** take which actions.
    - **Key observation**: estimating *counts* of actions might not require estimating actions for all agents *individually*

Solution: Many-Agent Planning under Openness

Main principle: ***only model some neighbors*** to counter the increased complexity caused by openness

- In polling and survey theory, social scientists only survey a small randomly sampled portion of the population to estimate the behaviors and opinions of all people

Subject agent estimates action counts for the entire neighborhood by *extrapolating* the estimated actions of modeled neighbors $\hat{N}_\theta(i)$ using the **multinomial distribution**

$$P(C_\theta|s^t, M^t) \sim Multi\left(|N_\theta(i)|, \left\{\hat{p}_{a_1,\hat{N}_\theta(i)}, \dots, \hat{p}_{a_{|A|},\hat{N}_\theta(i)}\right\}\right)$$

$$\hat{p}_{a,\hat{N}_\theta(i)} = \frac{\hat{n}_{\pi(s^t)=a,\hat{N}_\theta(i)}}{|\hat{N}_\theta(i)|}$$

I-POMCP$_O$ Many-Agent MCTS Algorithm: adapts POMCP algorithm to many-agent open environments

- Estimate actions of a few neighbors then extrapolate to all neighbors' behaviors
- ***Better time complexity*** than previous I-POMCP MCTS algorithm (Hua et al., 2015).
- Comparable to Dec-POMDP MCTS algorithms (Amato & Oliehoek, 2015; Best et al., 2019) ***but does not require*** the subject agent to observe other agents' actions nor their observations

Algorithm 1 I-POMCP$_\mathcal{O}$: Open Many-Agent MCTS

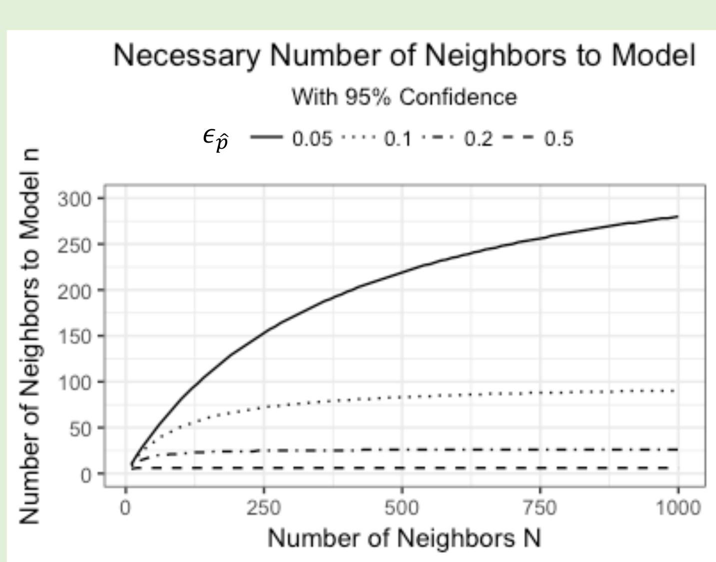
Note: T is the tree (initially empty), p is a path from the root of the tree (with $p = \emptyset$ signifying the root), B_p is the particle filter signifying the set of state-model pairs encountered at the node at p in the tree, PF is the root particle filter, N is count of the number of visits to each node in the tree initialized to some constant $\nu \geq 0$, Q is the Q function initialized to 0, c a constant from UCB-1.

```
procedure I-POMDP-MCTS(PF, τ)
  time ← 0
  while time < τ do
    s^0, M^0 ← SampleParticle(PF)
    UpdateTree(s^0, M^0, 0, ∅)
    Increment time
  return argmax_{a∈A_i} Q(∅, a)

procedure UPDATETREE(s^t, M^t, t, p)
  if t ≥ H then
    return 0
  B_p ← B_p ∪ {(s^t, M^t)}
  if p ∉ T then
    T ← T + leafnode(p)
    return Rollout(s^t, M^t, t)
  else
    C^t ← SampleConfiguration(s^t, M^t)
    a_i^t ← argmax_{a∈A_i} Q(p, a) + c·sqrt((log N_p) / N_{p↦a})
    s^{t+1}, M^{t+1}, o_i^t, r_i^t ← Simulate(s^t, M^t, a_i^t, C^t)
    N_p ← 1 + N_p
    N_{p↦a_i^t} ← 1 + N_{p↦a_i^t}
    p' ← p + (a_i^t, o_i^t)
    R ← r_i^t + γ·UpdateTree(s^{t+1}, M^{t+1}, t + 1, p')
    Q(p, a_i^t) ← Q(p, a_i^t) + R − Q(p, a_i^t)/N_{p↦a_i^t}
    return R

procedure ROLLOUT(s^t, M^t, t)
  R ← 0, t' ← t
  while t < H do
    C^t ← SampleConfiguration(s^t, M^t)
    a_i^t ← SampleAction(A_i)
    s^{t+1}, M^{t+1}, o_i^t, r_i^t ← Simulate(s^t, M^t, a_i^t, C^t)
    R ← R + γ^{t−t'} · r_i^t, t ← t + 1
  return R

procedure SAMPLECONFIGURATION(s^t, M^t)
  C(a, θ) ← 0, n̂_{π(s^t)=a, N̂_θ(i)} ← 0   ∀a, θ
  for M_{j,l−1} ∈ M^t do
    a ~ π_{j,l−1}(s^t)
    n̂_{π(s^t)=a, N̂_{θ_j}(i)} ← n̂_{π(s^t)=a, N̂_{θ_j}(i)} + 1
  for θ ∈ Θ do
    for a ∈ A do
      p̂_{a,N̂(i)} ← n̂_{π(s^t)=a, N̂_θ(i)} / |N̂_θ(i)|
    for j ∈ N_θ(i) do
      a ~ Cat(p̂_{a_1,N̂_θ}, p̂_{a_2,N̂_θ}, ..., p̂_{a_|A|,N̂_θ})
      C(a, θ) ← C(a, θ) + 1
  return C
```

Theoretical Results

Theorem 1 With confidence $1 - \alpha$, the ***error*** incurred by the subject agent in its estimate ***of the proportion*** $\hat{p}_{a,\hat{N}_\theta(i)}$ of its neighbors of frame θ that will perform action a is ***bounded by the given*** $\epsilon_{\hat{p}}$ so long as it models the following number of neighbors:

$$n_\theta = |\hat{N}_\theta(i)| \geq \frac{N\left(\frac{t_{n-1,\frac{\alpha}{2}}}{\epsilon_{\hat{p}}}\right)^2}{N - 1 + \left(\frac{t_{n-1,\frac{\alpha}{2}}}{\epsilon_{\hat{p}}}\right)^2}$$



Corollary 1 The ***maximum error in the multinomial distribution*** of the configuration of other agents' actions ***$P(C|s^t, M^t)$ is bounded by***:

$$\epsilon_{P(C)} = |P^*(C|s^t, M^t) - P(C|s^t, M^t)|$$
$$< \frac{\prod_\theta |N_\theta(i)|!}{\prod_{\theta,a} C(a,\theta)!}\left[\prod_{\theta,a}\left(\hat{p}_{a,\theta} - \epsilon_{\hat{p}}\right)^{C(a,\theta)} - \prod_{\theta,a}\hat{p}_{a,\theta}^{C(a,\theta)}\right]$$

when the subject agent models only n_θ neighbors (from Theorem 1) to achieve at most $\epsilon_{\hat{p}}$ error in its estimates of the action probabilities parameterizing the multinomial distribution.

Theorem 2 The ***regret*** of the subject agent from modeling only a subset of its neighbors ***is bounded by***:

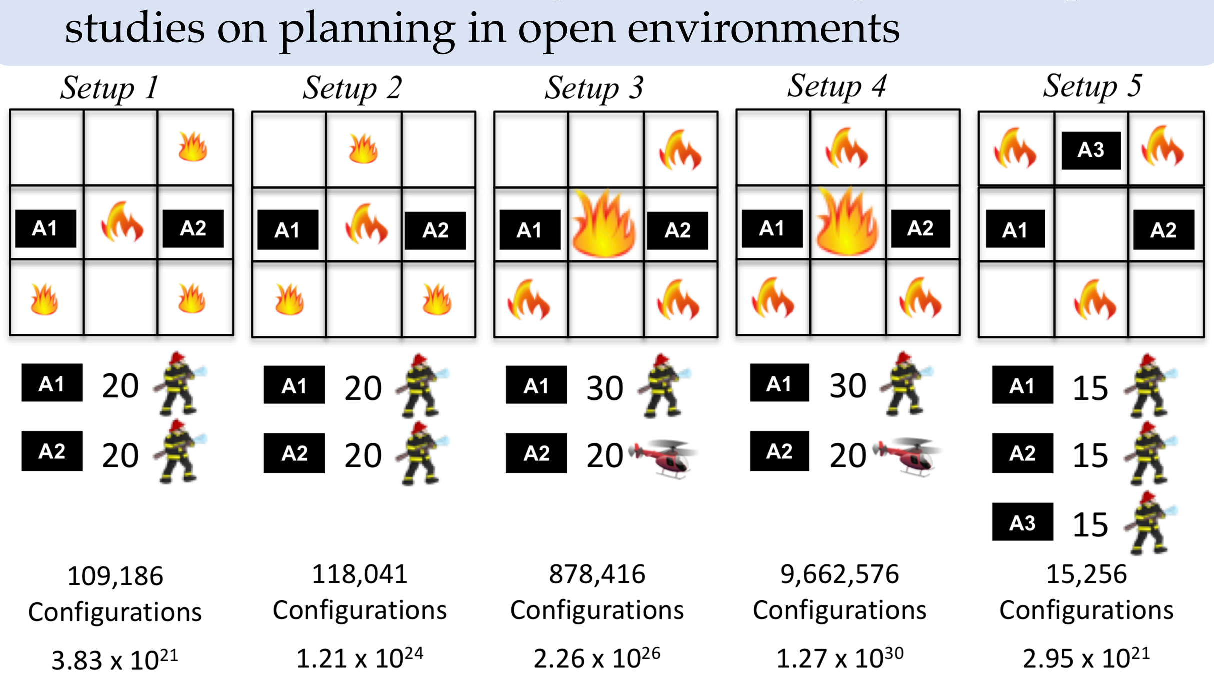
$$\left\|V_{i,k}^* - J_{i,k}\right\|_\infty$$
$$\leq 2\epsilon_{P(C)} \cdot |\mathbb{C}| \cdot R_{max}\left[\gamma^{k-1} + \frac{1}{1-\gamma}\left(1 + 3\gamma\frac{|\Omega_i|}{1-\gamma}\right)\right]$$

which is ***linear*** in the error $\epsilon_{P(C)}$ in the agent's estimation of configuration likelihoods caused by modeling some neighbors only and ***proportional to only $\sqrt{1/n_\theta}$ in the worst case*** (due to the fact that $\epsilon_{P(C)}$ is at worst linear in $\epsilon_{\hat{p}}$ and $\epsilon_{\hat{p}}$ is proportional to $\sqrt{1/n_\theta}$)

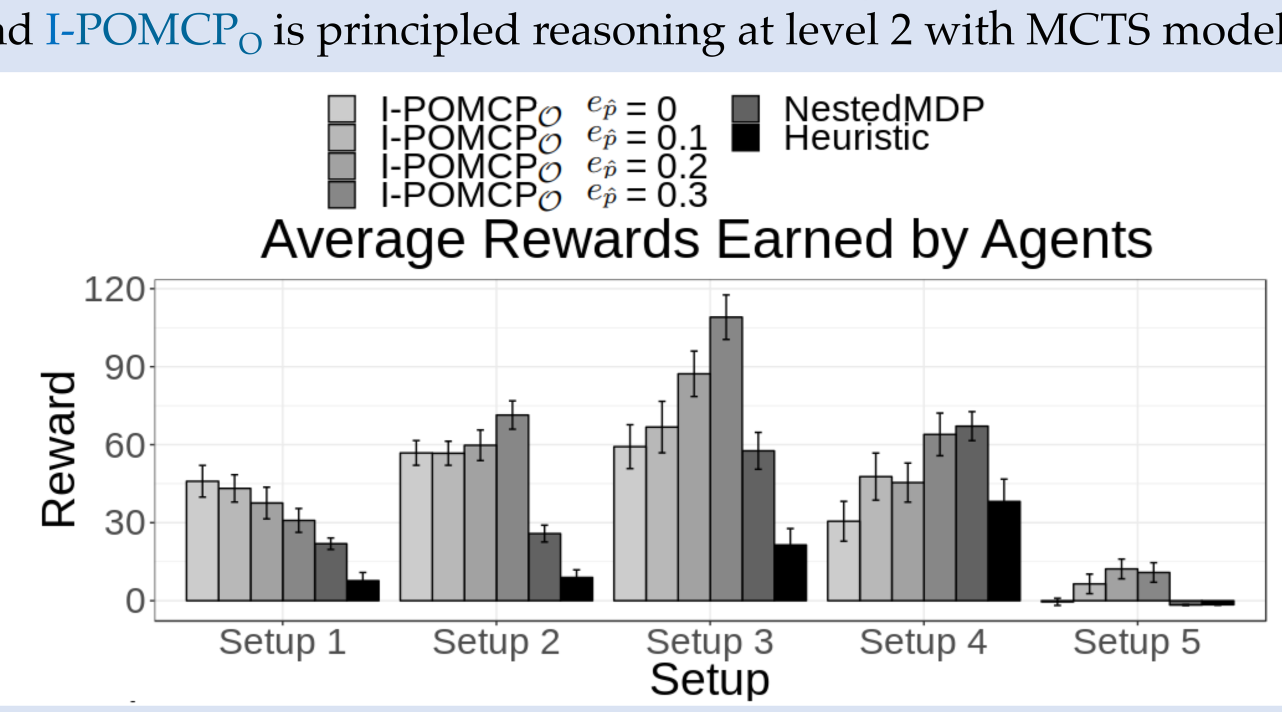
Experimental Results

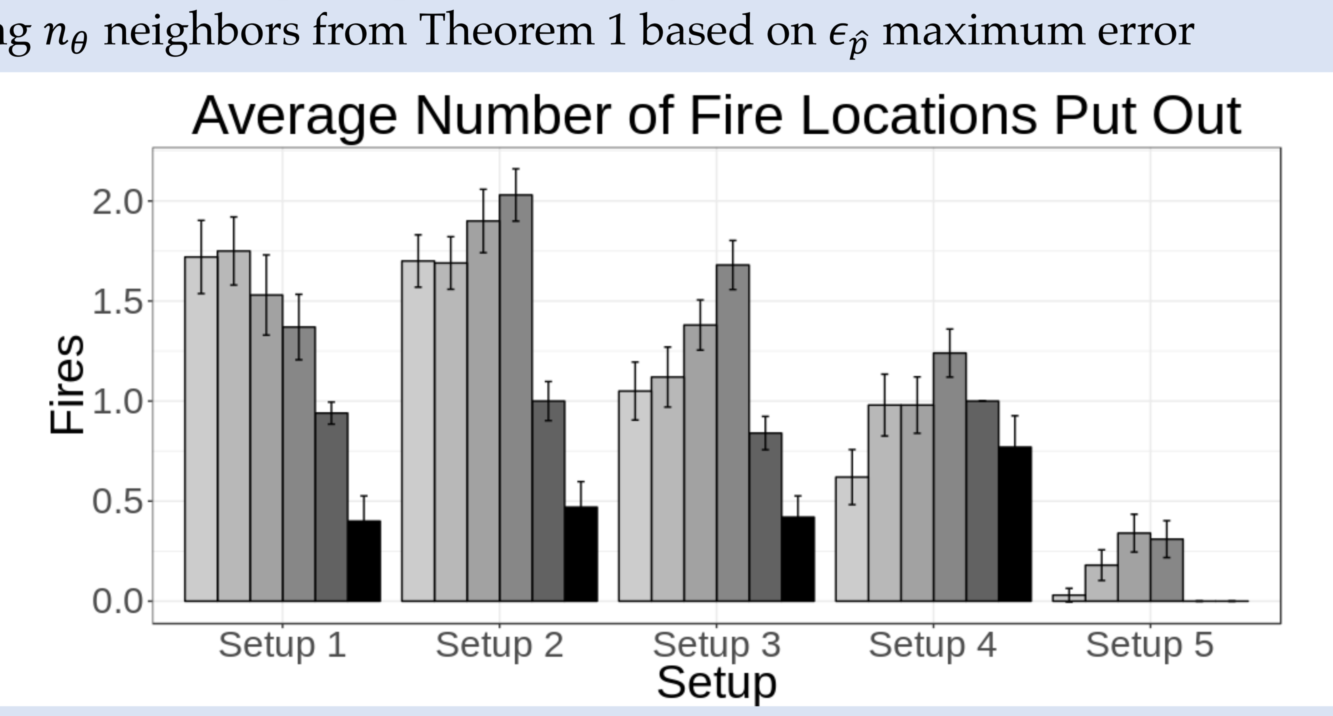
Wildfire Problem: ground firefighter and helicopter agents with different capabilities cooperate to put out nearby fires.

- **Goal:** maximize rewards earned for putting out fires (and minimize costs for fires burning out)
- Agents can ***run out of suppressant***, requiring them to temporarily leave the environment to recharge.
- Consider an order of magnitude more agents than prior studies on planning in open environments



Agent Reasoning Models: Heuristic randomly chooses active fires, NestedMDP is principled reasoning with value iteration at level 1, and I-POMCP$_O$ is principled reasoning at level 2 with MCTS modeling n_θ neighbors from Theorem 1 based on $\epsilon_{\hat{p}}$ maximum error





- I-POMCP$_O$ ***outperformed all baselines*** (statistically significantly higher rewards in 4 of 5 setups) due to ***better coordination*** between agents by working together to put out fires in more locations
- In the more complicated Setups 2-4, ***modeling fewer neighbors*** n_θ due to higher allowable error $\epsilon_{\hat{p}}$ ***led to improved performance*** due to more sampled trajectories in the fixed time budget. This implies that the approximation error caused by modeling only some neighbors ***was less than*** the approximation error from sampling fewer trajectories in MCTS.

 This work was partially supported by the National Science Foundation under Grant No. 1909513 and the UGA Institute for Artificial Intelligence Research.