I-POMDP Solver v2

May 26th

Forest structure

Level 2

the same hierarchical tree structure as I-NTMCP

$\Rightarrow \sum_{d=0}^{L} (n-1)^d = \mathcal{O}(n^L)$ trees for n agents at max. level L

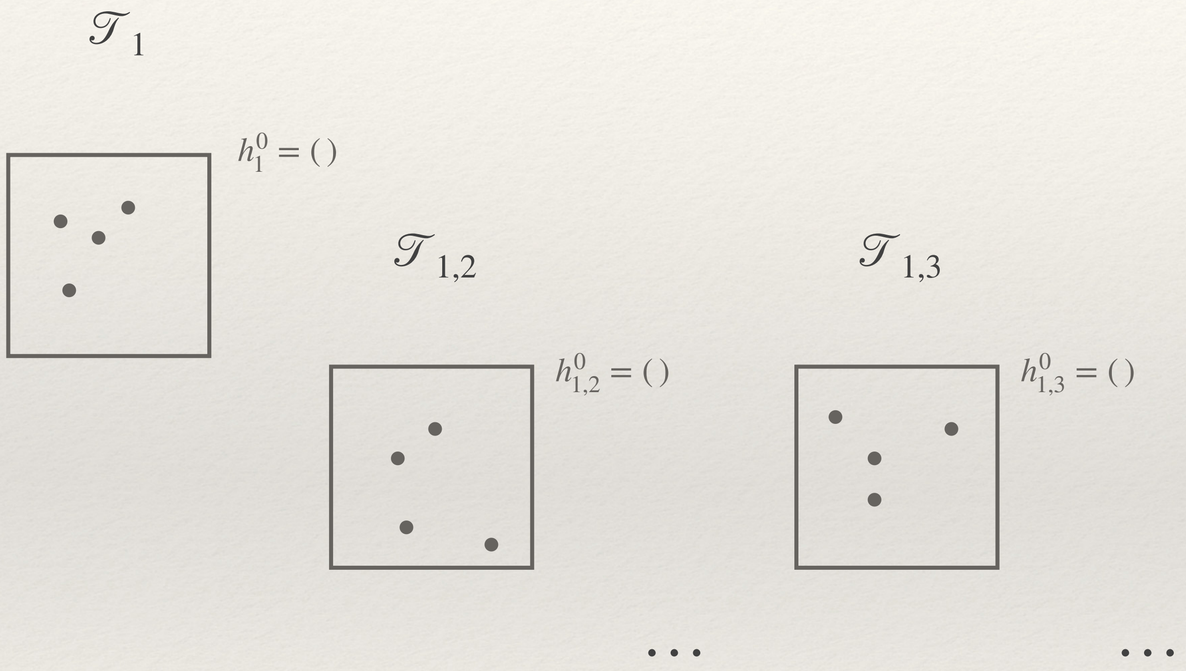
Level 1

Represents 1's beliefs about 3's beliefs about 2's beliefs

Level 0

● = root node

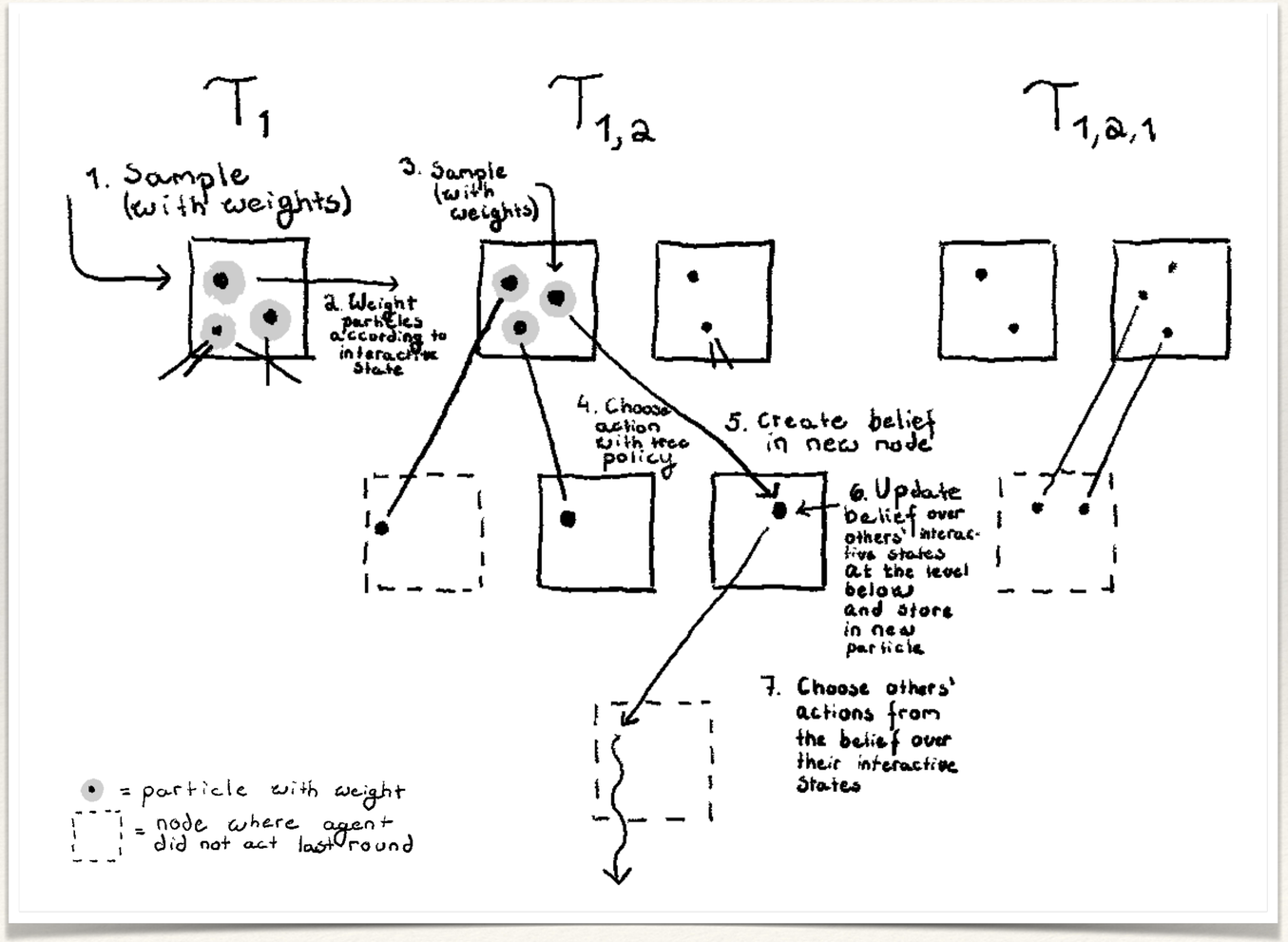
Initialisation



- Each node corresponds to an agent action history
- A node contains a set of particles
- A particle is a tuple $p = \langle s, h, b_{l-1} \rangle$ where
  - s is an environment state
  - h is a joint action history
  - $b_{l-1} = \langle b_2, b_3 \rangle$ is a belief distribution over particles at the level below for each other agent
    - note that h uniquely determines which root node the belief corresponds to at the level below
- Particles are therefore **interactive states** (with additional joint action history to structure the search process)
- No separate beliefs at root nodes!

Planning

- planning consists of expanding each tree a given numbers of times
- bottom up: first we plan at level 0, then level 1, and so on



Tree policy

- The tree policy resembles MCTS's UCT method, but has to be adapted slightly
- When the particles p in a node N all have weights $b(p)$, the particles represent a belief b. Calculate:

$N_+(b) = \sum_{p \in \mathcal{T}(h_k),\, b(p)>0} n(p)$ = total number of times particles with > 0 weight have been expanded with some action (other than "no turn")

$W(b) = \sum_{p \in \mathcal{T}(h_k)} b(p)n(p)$ = weighted number of times b has been expanded

$W(b,a) = \sum_{p \in \mathcal{T}(h_k)} b(p)n(p,a)$ = weighted number of times b has been expanded with a

$N_+(b,a) = \frac{W(b,a)}{W(b)} N_+(b)$ (same thing as $\tilde{W}(b,a)$ before)

$$Q(b,a) = \frac{1}{\sum_{p \in \mathcal{T}(h_k),\, n(p,a)>0} b(p)} \sum_{p \in \mathcal{T}(h_k)} b(p)V(p,a)$$

- Then the chosen action is one that maximises $Q(b,a) + c\sqrt{\frac{\ln N_+(b)}{N_+(b,a)}}$ (c is exploration constant)

Tree policy for opponents' trees

- A softargmax policy is used for sampling actions from the opponents' trees when expanding a tree
  - the opponents' trees (one level lower) only approximate the optimal actions, and the optimal action given by the tree might be different from what the agent actually does
- The probability of choosing action a under belief b in the lower-level tree is

$$\mathbb{P}(a \mid b) \propto \exp\left(\frac{N_+(b,a)}{\sqrt{N_+(b)}}\right) = \exp\left(\frac{W(b,a)}{W(b)}\sqrt{N_+(b)}\right)$$

- For example: if proportions of action weights are (0.7, 0.2, 0.1), the action probabilities are (0.74, 0.15, 0.11) if $N_+(b) = 10$ and (0.99, 0.01, 0.00) if $N_+(b) = 100$

Belief update

- Agent takes an action and receives an observation
- Belief update happens top down: first the top-level tree, then the trees below them, etc.