ROB311 Quiz 3

Hanhee Lee

March 29, 2025

Contents

Turn-Taking Multi-Agent Decision Algorithms

1 Zero-Sum Turn-Based Games

Summary: In a **zero-sum turn-based games**, we assume that

- **Agents and Environment:**
  - there are two agents, called the **maximizer** and **minimizer**
  - the environment is always in one of a discrete set of states, $\mathcal{S}$
  - a subset of the states, $\mathcal{T} \subseteq \mathcal{S}$, are terminal states
  - there is only one decision maker for each non-terminal state, $s \in \mathcal{S} \setminus \mathcal{T}$
  - For each non-terminal state, $s \in \mathcal{S} \setminus \mathcal{T}$, the decision-maker has a discrete set of actions, $\mathcal{A}(s)$
- **Decision Process:** At time-step t, the decision-maker will:
  - **Observe:** Observe the state s_t
  - **Select:** Select an action $a_t \in \mathcal{A}(s_t)$
  - **Move:** Make the move (s_t, a_t)
- **State Transitions:**
  - Environment transitions to a deterministic state, s_{t+1}, based on a stationary fn,

    $$s_{t+1} = \text{tr}(s_t, a_t)$$

  - Once a terminal state is reached (if $s_{t+1} \in \mathcal{T}$), the maximizer obtains a reward for the final transition based on a reward fn, $r(\cdot, \cdot, \cdot)$:

    $$r(s_t, a_t, s_{t+1}) = \text{maximizer's reward for reaching state } s_{t+1}$$

    $$-r(s_t, a_t, s_{t+1}) = \text{minimizer's reward for reaching state } s_{t+1}$$

Warning:

- Maximizer is trying to maximize the reward of agent 1
- Minimizer is trying to minimize the reward of agent 1 (i.e. maximize the reward of agent 2)

1.1 α/β Pruning

Motivation: Don't explore the entire game tree by pruning branches that are unreachable under perfect play.

Definition: For each state s:

- α_s: Maximum value at s thus far (initially $-\infty$)
- β_s: Minimum value at s thus far (initially $+\infty$)

1.1.1 α Cuts

Definition: If the **maximizer** is the turn-taker at s, then α_s increases to the maximum value of s's successors as they are explored, and $\beta_s = \beta_{\text{parent}(s)}$.

- If α_s increases beyond β_s, then s unreachable under perfect play.

1.1.2 β Cuts

Definition: If the **minimizer** is the turn-taker at s, then β_s decreases to the minimum value of s's successors as they are explored, and $\alpha_s = \alpha_{\text{parent}(s)}$.

- If β_s decreases beyond α_s, then s unreachable under perfect play.

1.2 Monte-Carlo Tree Search (MCTS) Algorithm

Algorithm:

1. Selection: Traverse using an alternate policy until a node has unexplored children.

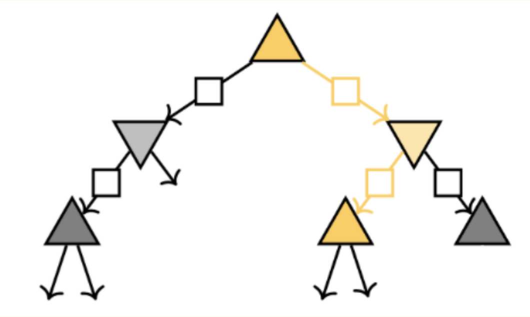

Figure 1

   - Our Agent (Upper Triangle): Uses UCB to choose the next node to explore
   - Other Agent (Down Triangle): Can't control their actions, so this agent picks w/ their own heuristic.
   - Square Boxes: Estimated values (i.e. n and $\hat{q}$)
   - Ends when there is at least one action that hasn't been explored yet. In this case, two actions ahven't been explored.
   - Can skip expansion and simulation if the most recently expanded node is a terminal state.

2. Expansion: Expand an unexplored child; initialize $n(a)$ and $\hat{q}(s, a)$.

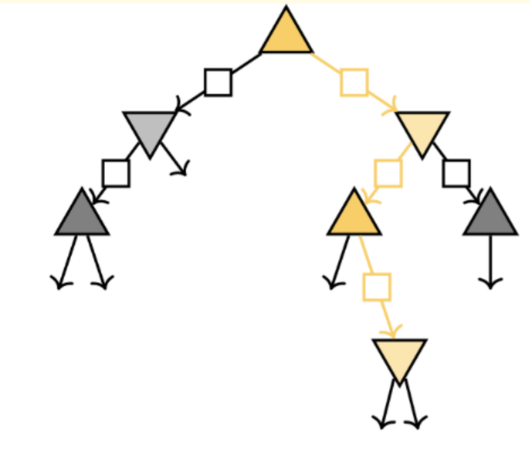

Figure 2

   - $\hat{q}(s, a)$ is initialized to 0 and $n(a)$ is initialized to 1 b/c we've visited this node once.
   - Randomly pick an unexplored action unless there is only one action left.
   - Can skip similuation if the most recently expanded node is a terminal state.

3. Simulation: Traverse using the random policy until a terminal node is reached.

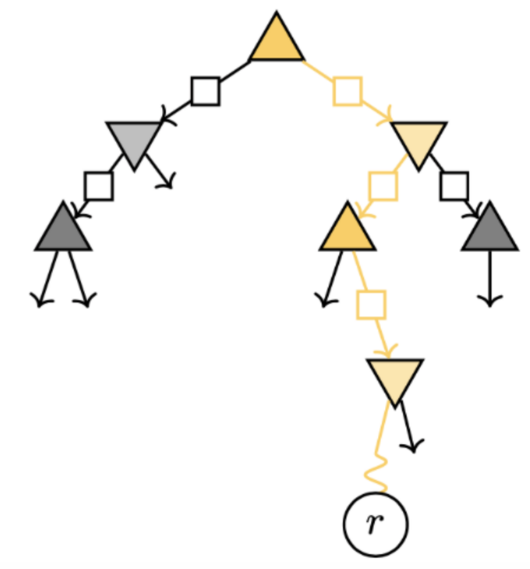

Figure 3

- Using random policy to simulate the game until a terminal state is reached (i.e. reward is obtained)

4. Back-propogation: Get the reward and reverse; update $n(a)$ and $\hat{q}(s,a)$.

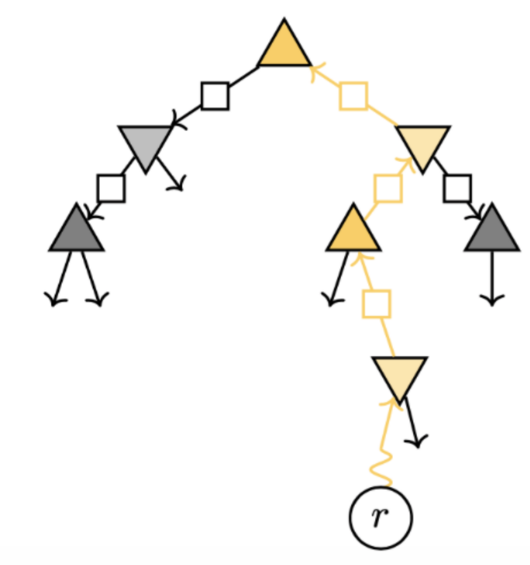

Figure 4

- Go up the path in yellow and update the values of $n(a)$ and $\hat{q}(s,a)$ for OUR agent only (i.e. the upper triangle)

Warning:

- Works for more than 2 agents.
- Don't need to know anyone else's reward function.
- Has to be turn taking but can be not alternating (i.e. immediate switch between agents)
- Can augment simultaneous actions
- Communication
- Works fo rnon-zero sum games.

1.3 Examples

1.3.1 Zero Sum Turn-Based Games

Example:

- **Given:** Cavemen is injured from his hunt. He has extra food, but needs medicine.
  - He meets another caveman who is willing to trade.

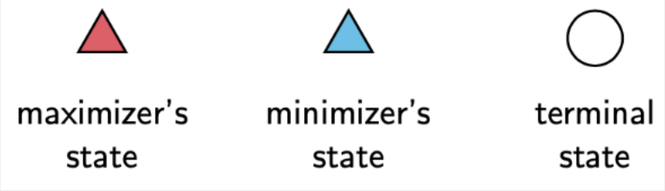

Figure 5: States

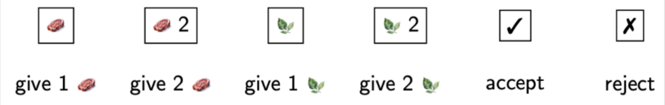

Figure 6: Actions

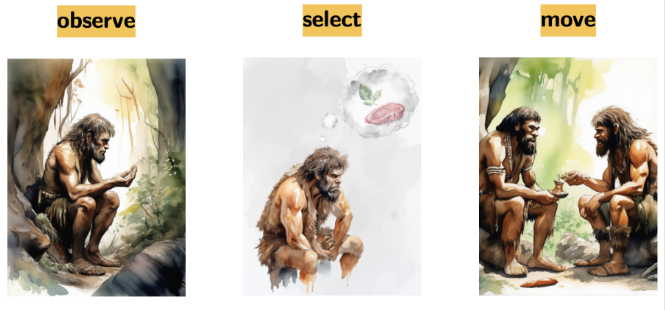

Figure 7: Decision Process

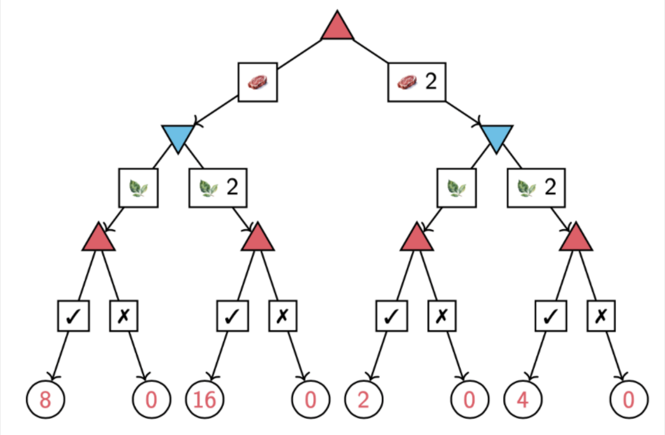

Figure 8: Game Tree

- States
  - Red triangle: Maximizing agent
  - Blue triangle: Minimizing agent
  - White circles with #s: terminal states
  - Rewards: In red b/c it's for the maximizer. The minimizer's reward is the negative of the maximizer's reward.

> - Actions: Square boxes are actions
>
> - **Solution:** Backtracking through the game tree, we can find the optimal path for the maximizer and minimizer.
>   - **Maximizer Turn:** LL: Accept to get reward of 8, L: Accept to get reward of 16, R: Accept to get reward of 2, RR: Accept to get reward of 4
>   - **Minimizer Turn:** LL: 1 medicine to make maximizer get reward of 8, R: 1 medicine to make maximizer get reward of 2
>   - **Maximizer Turn:** 1 food to make maximizer get reward of 8 b/c going right will make maximizer get reward of 2
>   - **Optimal Path:** Therefore, the optimal path will be LLL b/c the maximizer will get a reward of 8, while the minimizer will reduce the reward from 16 to 8.
>   - Assume boths agents play optimally, this will be the path taken.

1.3.2 α Cuts

Example:

- Explored 14, 12 and now $\beta_{\text{parent}(s)} = \beta_s = 5$, so this will be compared for α_s until $\alpha_s > \beta_s$ b/c then s unreachable under perfect play.
- Iterate:
  - $\alpha_s = -\infty < \alpha'_s = 2 \rightarrow \alpha_s = 2$, but $\alpha_s = 2 < \beta_s = 5$
  - $\alpha_s = 2 < \alpha'_s = 4 \rightarrow \alpha_s = 4$, but $\alpha_s = 4 < \beta_s = 5$
  - $\alpha_s = 4 < \alpha'_s = 9 \rightarrow \alpha_s = 9$, and $\alpha_s = 9 > \beta_s = 5$, therefore, prune all the other branches that haven't been explored yet in the children of s paths

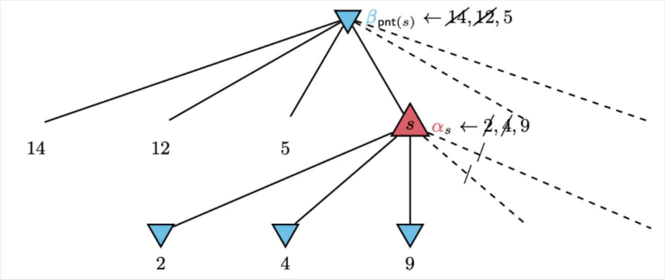
$\beta_{\text{pnt}(s)} \leftarrow \cancel{14}, \cancel{12}, 5$
14
12
5
s
$\alpha_s \leftarrow \cancel{2}, \cancel{4}, 9$
2
4
9

Figure 9

1.3.3 β Cuts

Example:

- Explored 4,6, and now $\alpha_{\text{parent}(s)} = \alpha_s = 7$, so this will be compared for β_s until $\beta_s < \alpha_s$ b/c then s unreachable under perfect play.
- Iterate:
  - $\beta_s = +\infty > \beta'_s = 9 \rightarrow \beta_s = 9$, but $\beta_s = 9 > \alpha_s = 7$
  - $\beta_s = 9 > \beta'_s = 8 \rightarrow \beta_s = 5$, but $\beta_s = 8 > \alpha_s = 7$
  - $\beta_s = 8 > \beta'_s = 3 \rightarrow \beta_s = 3$, and $\beta_s = 3 < \alpha_s = 7$, therefore, prune all the other branches that haven't been explored yet in the children of s paths

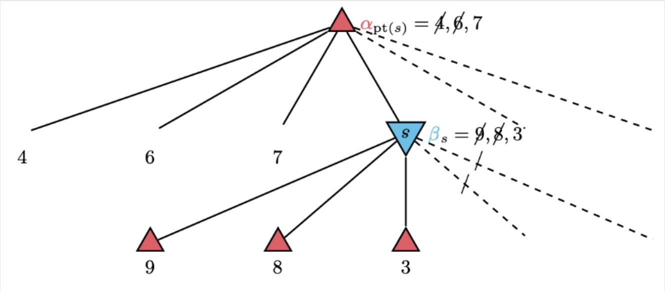
$\alpha_{\text{pt}(s)} = \cancel{4}, \cancel{6}, 7$
4
6
7
s
$\beta_s = \cancel{9}, \cancel{8}, 3$
9
8
3

Figure 10

1.3.4 Alpha Beta Pruning

Process:

1.

Example: Alpha-Beta Pruning Practice

1.

1.3.5 Monte-Carlo Tree Search (MCTS) Algorithm

Example:

1. **Given:** Consider a simplified two-player turn-based game tree. You are currently at the root node S_0, which has three possible actions a_1, a_2, a_3. The current statistics of its children are as follows:

Action	$N(s_0, a)$	$\bar{X}(s_0, a)$
a_1	10	0.6
a_2	5	0.8
a_3	0	–

   - $N(s_0, a)$: Number of times action a has been selected at state s_0
   - $\bar{X}(s_0, a)$: Average reward obtained from action a at state s_0
   - $\text{UCB} = \bar{X}(s_0, a) + \sqrt{\dfrac{\ln(t)}{N(s_0, a)}}$
     - t: Total number of actions taken at s_0

2. **Problems:**
   - If we were to use the UCB algorithm, which nodes get selected during the selection phase? Which node gets expanded during the expansion phase?
   - Suppose from the expanded node, simulation is performed until termination. A reward of +1 is obtained. Update the statistics at s_0 accordingly.
   - Then, repeat the question, assuming a reward of -1 is attained after the simulation phase.

3. **Solution:**

   (a) **Selection 1:** s_0 since we traverse until a node has unexplored children (i.e. s_3 is unexplored)

   (b) **Expansion 1:** s_3 is automatically expanded since it is the only unexplored child of s_0 w/ $N(s_0, a_3) = 1$ and $\bar{X}(s_0, a_3) = 0$

   (c) **Simulation 1:** Get a reward of +1

   (d) **Back Propogation 1:** For this edge from s_0 to s_3, we update the statistics as follows:
   - $N(s_0, a_3) = 1$
   - $\bar{X}(s_0, a_3) = \dfrac{1}{1} = 1$

   (e) **Selection 2:** s_0 and choose the action with the highest UCB value for s_1, s_2, and s_3:
   - $UCB(s_0, a_1) = 0.6 + \sqrt{\dfrac{\ln(16)}{10}} = 1.13$
   - $UCB(s_0, a_2) = 0.8 + \sqrt{\dfrac{\ln(16)}{5}} = 1.54$
   - $UCB(s_0, a_3) = 1 + \sqrt{\dfrac{\ln(16)}{1}} = 2.67$. Therefore, choose s_3 as part of the selection phase and assume it has unexplored children.

   (f) **Expansion 2:** Not enough info but assume we expand an unexplored child.

   (g) **Simulation 2:** Get a reward of -1

   (h) **Back Propogation 2:** For this edge from s_0 to s_3, we update the statistics as follows:
   - $N(s_0, a_3) = 2$
   - $\bar{X}(s_0, a_3) = \dfrac{1 + (-1)}{2} = 0$