Numerical Experiment

BY JACK YANSONG LI
University of Illinois Chicago
Email: yli340@uic.edu

1 Simplified Overcooked Environment (5 by 5 is needed)

We consider a simplified version of the Overcooked environment. The environment is a 3×3 grid with two players (player 1: robot player, player 2: human player) working together to deliver food (take food in $(3, 2)$, deliver food in $(1, 2)$). The action space for both players are $A^{(1)} = A^{(2)} = A$ and $A_{\text{joint}} = A \times A$. The transition kernel $\mathcal{T}$ is deterministic, i.e., $\mathcal{T}: S \times A \to S$. The reward r is defined as

$$r(s, a_{\text{joint}}) = \begin{cases} 0 & \text{else} \\ 1 & \text{if} \quad \mathcal{T}(s, a_{\text{joint}}) = s_{\text{abs}} \end{cases}.$$

The following is an example of a state in simplified Overcooked environment:

```
state = {'player_positions': [(1,2), (3,2)],  'player_has_food': [False, False],
'table_position': (2, 2), 'table_has_food': True}
```

In this simplified Overcooked environment, the state space consists of the following components:

1. Player positions: 8 possible positions (eliminate table position) for each player (3×3 grid), so $8 \times 8 = 64$ combinations.
2. Player has food: 2 possibilities for each player (True or False), so $2 \times 2 = 4$ combinations.
3. Table has food: 2 possibilities (True or False), so 2 combinations.
4. An abosrbing state s_{abs}.

Total state space size: $|S| = 64 \times 4 \times 2 + 1 = 513$. Also, $|A| = 5$.

Assumption 1. The initial state s_1^k for each episode k is fixed, i.e., $s_1^k = s_1$ for any $k \in [K]$.

Assumption 2. Player 2 adopts a stationary policy denoted as $\pi^{(2)}$, which belongs to the set $\mathcal{H}$.

2 Comparative Analysis of Regret

We evaluate and compare the regret observed in the following algorithms:

1. MEX.
2. MAB.
3. Q-UCB.
4. Model-based RL + UCB (MBRL+UCB).
5. Optimistic posterior sampling.

For all these 5 algorithms, we provide them a finite hypothesis set $\mathcal{H}_{\text{fin}}$.

2.1 Finite Hypothesis Set

For this experiment, we let $\pi^* \in \mathcal{H}_{\text{fin}}$. Then, we generate the subsequent graph utilizing the following parameters for 5 algorithms:

1. Horizontal axis: Episode number, denoted by k.

2. Vertical axis: $\sum_{i=1}^{k} V(\psi(\pi^i), \pi) - V(\psi(\pi^*), \pi^*)$, where ψ represents the value iteration algorithm.

2.2 Infinite Hypothesis Set

For the purposes of this experiment, we do not employ a true infinite hypothesis set. Instead, we arbitrarily select a policy from $\mathcal{H}_{\text{fin}}$, denote it as π^*_{det}, and choose $\pi^* \notin \mathcal{H}_{\text{fin}}$ such that $\|\pi^*_{\text{det}} - \pi^*\| \leq \varepsilon$.

Keeping ε constant, we construct the following graph utilizing the following parameters, applied across 5 different algorithms:

1. Horizontal axis: Episode number, denoted by k.

2. Vertical axis: $\sum_{i=1}^{k} V(\psi(\pi^i), \pi^i) - V(\psi(\pi^*), \pi^*)$, where ψ represents the value iteration algorithm.

Moreover, maintaining a fixed total episode count denoted by K, we develop another graph utilizing the subsequent parameters, again applied across 5 different algorithms:

1. Horizontal axis: The covering distance, ε

2. Vertical axis: $\sum_{i=1}^{K} V(\psi(\pi^i), \pi^i) - V(\psi(\pi^*), \pi^*)$, where ψ represents the value iteration algorithm.

3 Theoretical Analysis

We seek to build the following examples

Example 3. An example where the regret of MEX is smaller that of MAB/MBRL-UCB when K is sufficiently large, or conversely.