Agent-based Reinforcement Learning in Colonel Blotto

Joseph Christian G. Noel

1 Introduction

2 Colonel Blotto

2.1 Definition

Colonel Blotto is a constant-sum game where two or more players distribute resources (or troops) over several fronts in a battlefield. A player wins or loses a front depending on whether they have allocated more or less resources to it than their opponent has, and the winner of the game is who has won the most fronts.

2.2 Strategy

3 Reinforcement Learning

3.1 Definition

Reinforcement Learning (RL) is a form of agent-based modeling. In RL an agent learns by performing actions which changes the state of an environment. After each action, the agent may also receive a "reward" whose value depends on how close the agent is to what it wants to achieve. The goal for the agent is to maximize the cumulative reward that it receives over all the actions that it takes. After a series of actions, the agent eventually reaches a goal state or terminal state, which signifies the end of an episode. The environment is then reset and the agent starts again from an initial state and the process then repeats itself.

Formally, an RL model is a set of states S, a set of actions A, and transition rules between states depending on the action taken. For state $s \in S$ at time t, an agent performs an action $a \in A$, moves to a new state s' and receives a reward r_t. The goal of the agent is to maximize the expected reward Rt,

$$R_t = \sum_{k=0}^{\infty} \lambda^k r_{t+k} \tag{1}$$

where $0 \leq \lambda \leq 1$ is a discount factor for handling infinite horizons.

$\pi(s, a)$ is a probability mapping of an agent taking action a while in state s. A proper policy is one where there is a non-zero probability of reaching a terminal state. There is always an optimal policy π^* that is better than or equal to all other policies when it comes maximizing the cumulative rewards.

3.2 Markov Decision Process

A common formulation of the Reinforcement Learning problem is as a Markov Decision Process (MDP). In this class of RL problems, the information from the history of all states, actions, and rewards before time t is encapsulate in the current state at time t. This is the Markov property, and tasks which exhibit this property are called Markov decision processes. Formally, the Markov property states that for transition probability function Pr,

$$Pr(s_{t+1}, r_{t+1}|s_t, a_t) = Pr(s_{t+1}, r_{t+1}|s_t, a_t, r_t, s_{t-1}, a_{t-1}, r_{t-1}, ..., s_0, a_0, r_0) \tag{2}$$