Find the Optimal Strategy for the Rock-Paper-Scissors Game

Jingpei Lu
Jacobs School of Engineering
University of California, San Diego
jil360@ucsd.edu

I. Problem Formulation

A. Optimization Problem for Rock-Paper-Scissors (RPS) Game

Imaging we are challenge your friend to a 100-game match of rock-paper-scissors, and our goal is to find a strategy that can win as many games as possible. Assume Our opponent's move is randomize but still biased towards one of the three options. He plays his preferred move 50% of the time and each of the other two options, 25% of the time. For each game, we can have the score $\{-1, 0, +1\}$ depending on the result. The score is +1 for winner, -1 for loser and 0 for draw. Our goal is to maximize our cumulative score s by predicting opponent's preference of rock, paper and scissor at time t.

B. Formulate the RPS Problem as a Markov Decision Process (MDP)

To model this problem as a Markov Decision Process, we should clearly define the states space $\mathcal{X}$, control spaces $\mathcal{U}$, the initial state x_o, the transition model $\psi(x_t, u_t, z_t)$, the planning horizon T, the stage reward $l(x_t, u_t)$ and terminal rewards $q(x_T)$. We can define the MDP as below.

- The state space is $x_t \in \mathcal{X} := \{Pr_t \times SD\}$, which is the cartesian product between the opponent's preference at time t and the score difference (SD). The score difference is defined as

  $$SD = your\ score - oppnent's\ score$$

  Since the opponent's preference cannot be directly observed, we should convert this state space $\mathcal{X}$ into a believe space $\mathcal{B}$ in order to formulate it as an MDP. Then, our state is $x_t \in \mathcal{B} := \{b_t \in [0,1]^3 \times SD | 1^T b_t = 1\}$, where $b_t = [Pr_r, Pr_p, Pr_s]$ is the opponent's preference of rock, paper and scissor at time t.

- The control is $u_t \in \mathcal{U} := \{R_t, P_t, S_t\}$,which is our move at time t.

- The initial state x_o is $\{b_0, SD\}$, where $b_0 = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$, SD = 0.

- The planning horizon is $T = 100$.

- The transition model is $b_{t+1} = \psi(b_t, u_t, z_t)$, where $z_t \in \{R_t, P_t, S_t\}$ is the observation of opponent's move at time t, and the model will be described later.

- The stage reward is $l(x_t, u_t) = \{-1, 0, +1\}$, depending on result of the game at time t. It is +1 if we win, -1 if we lose, 0 if draw.

- The terminal reward $q(x_T)$ is the same as the stage reward at time $t = T$.

After we have this set up of MDP, our problem becomes how to obtain the optimal policy

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} V_T^\pi(x_T)$$

that maximize our reward in the planning horizon T. The value function is defined as

$$V_t^\pi(x_t) = \max\{l(x_t, u_t) + V_{t-1}^\pi(x_{t-1})\}$$

II. Technical Approach

To solve this problem, we will use Forward Dynamic Programming Algorithm to find the optimal strategy by estimating the opponent's preference. As we described, we will use transition model to update the belief state. Since in this problem, our control u_t doesn't affect our believe state, so we can simply the model as

$$b_{t+1} = \psi(b_t, z_t)$$

where $b_{t+2} \in [0,1]^3$, and

$$b_{t+1}[0] = Pr(Rock|b_t, z_t)$$
$$b_{t+1}[1] = Pr(Paper|b_t, z_t)$$
$$b_{t+1}[2] = Pr(Scissor|b_t, z_t)$$

We then can update them as

$$Pr(x|b_t, z_t) \sim \frac{Pr\ (x|z_t)\Pr(b_t|x, z_t)}{\Pr(b_t|z_t)}$$

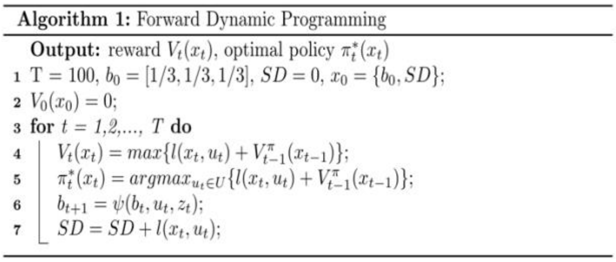
For example, if we want to update $b_{t+1}[0]$, then

$$Pr(Rock|b_t, z_t) \sim \frac{Pr\ (Rock|z_t)\Pr(b_t|Rock, z_t)}{\Pr(b_t|z_t)}$$

In this equation, $Pr\ (Rock|z_t)$ is the preference of Rock given the single observation at time t. We know our opponent move his preferred move 50% of the time, so we can define this as $Pr\ (Rock|z_t = Rock) = \frac{1}{2}$, and $Pr\ (Rock|z_t = Paper) = \frac{1}{4}, Pr\ (Rock|z_t = Scissor) = \frac{1}{4}$. $\Pr(b_t|Rock, z_t)$ is just our belief of opponent's preference of Rock at time t, which is $b_t[0]$. The denominator $\Pr(b_t|z_t) = \Pr(b_t|Rock, z_t) + \Pr(b_t|Paper, z_t) + \Pr(b_t|Scissor, z_t)$ is always 1. After we have $Pr(Rock|b_t, z_t)$, $Pr(Paper|b_t, z_t)$, and $Pr(Scissor|b_t, z_t)$, we should normalize them so that $1^T b_{t+1} = 1$ still hold.

A. Forward Dynamic Programming Algorithm

For each game, we can obtain the optimal action based on our belief at time t. Then, we can update our belief state using the transition model as we specified above and update the score differential based on the result of the game. The implementation of the algorithm is described as below.



```
Algorithm 1: Forward Dynamic Programming
  Output: reward $V_t(x_t)$, optimal policy $\pi_t^*(x_t)$
T = 100, $b_0 = [1/3, 1/3, 1/3]$, $SD = 0$, $x_0 = \{b_0, SD\}$;
$V_0(x_0) = 0$;
for $t$ = 1,2,..., $T$ do
    $V_t(x_t) = max\{l(x_t, u_t) + V_{t-1}^{\pi}(x_{t-1})\}$;
    $\pi_t^*(x_t) = argmax_{u_t \in U}\{l(x_t, u_t) + V_{t-1}^{\pi}(x_{t-1})\}$;
    $b_{t+1} = \psi(b_t, u_t, z_t)$;
    $SD = SD + l(x_t, u_t)$;
```

III. Result

In our simulation, we played 50 100-game matches with the randomly generated opponent's move. The opponent's moves consist 50% of his/her preferred move, and 25% of each other option. We simulated the game using 3 strategies: optimal policy, deterministic policy, and stochastic policy. Optimal policy is the one we obtained using Forward Dynamic Programming. Deterministic policy is iterating rock, paper, scissors, rock, paper, scissors, rock, and so on. Stochastic policy is choosing among the three options uniformly at random for each game. We plotted the mean and standard deviation over 50 100-game matches played by the three strategies with the number of games on the x axis, and the game score differential on the y axis.

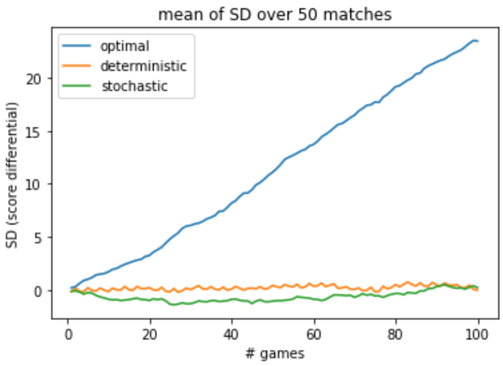


Figure 1

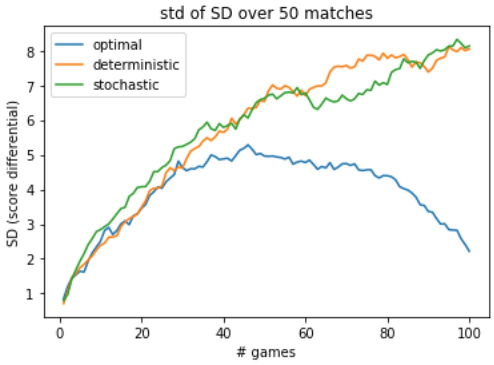


Figure 2

The Figure 1 shows the mean over 50 100-game matches played by the three strategies and the Figure 2 shows the standard deviation over 50 100-game matches played by the three strategies.

We can see that our optimal strategy outperforms the other two strategies. We can reach around 22 score differential on average at the end of each 100-game match.