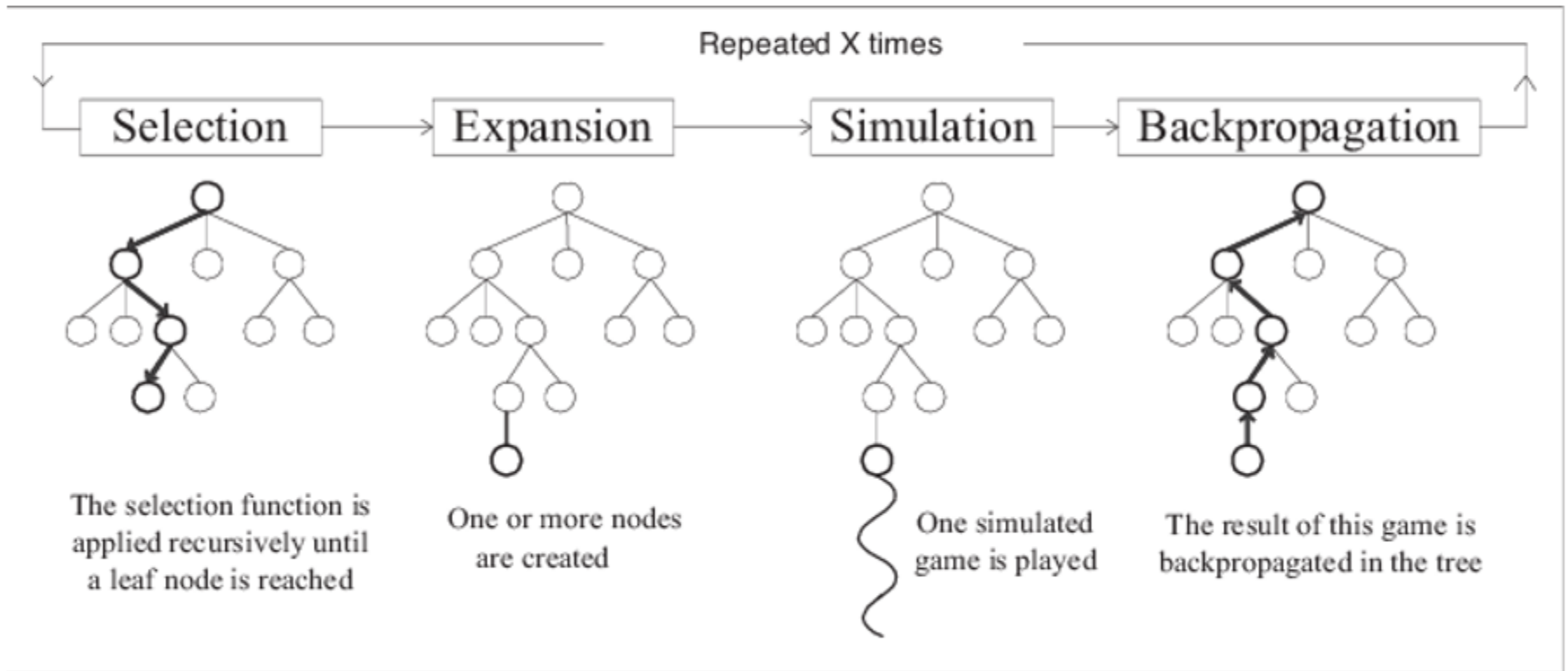
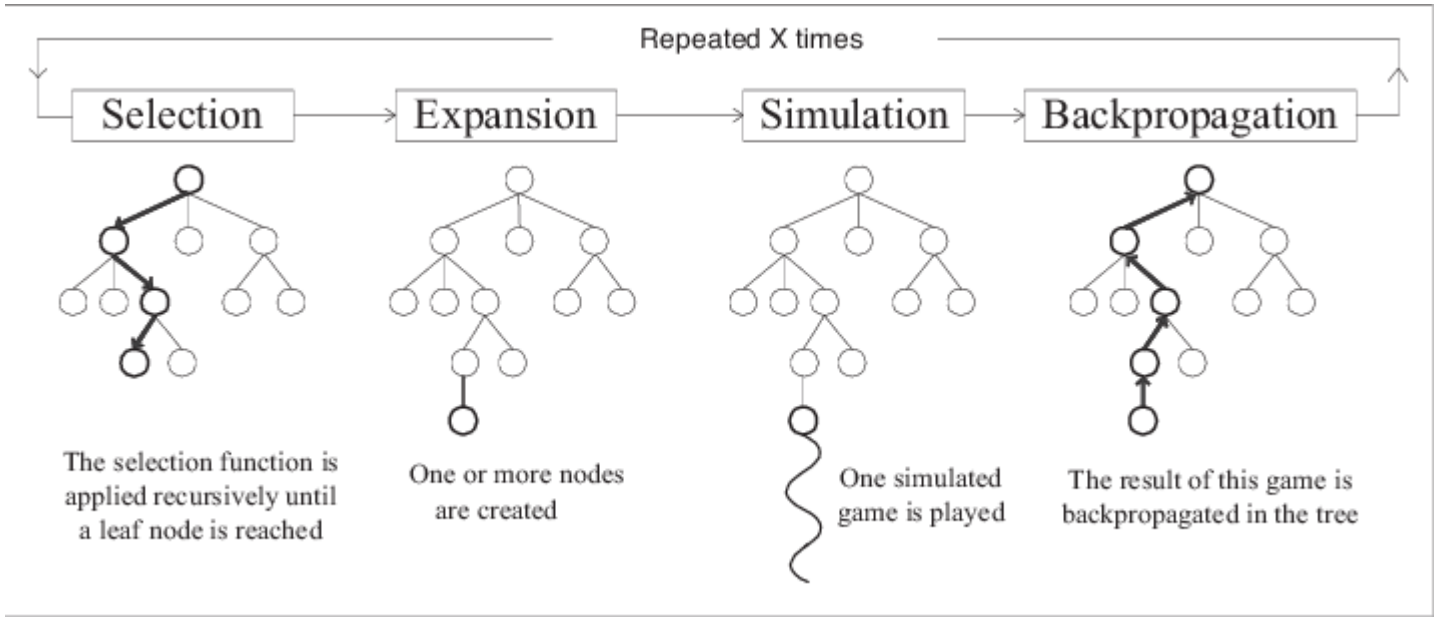
MONTE CARLO TREE SEARCH



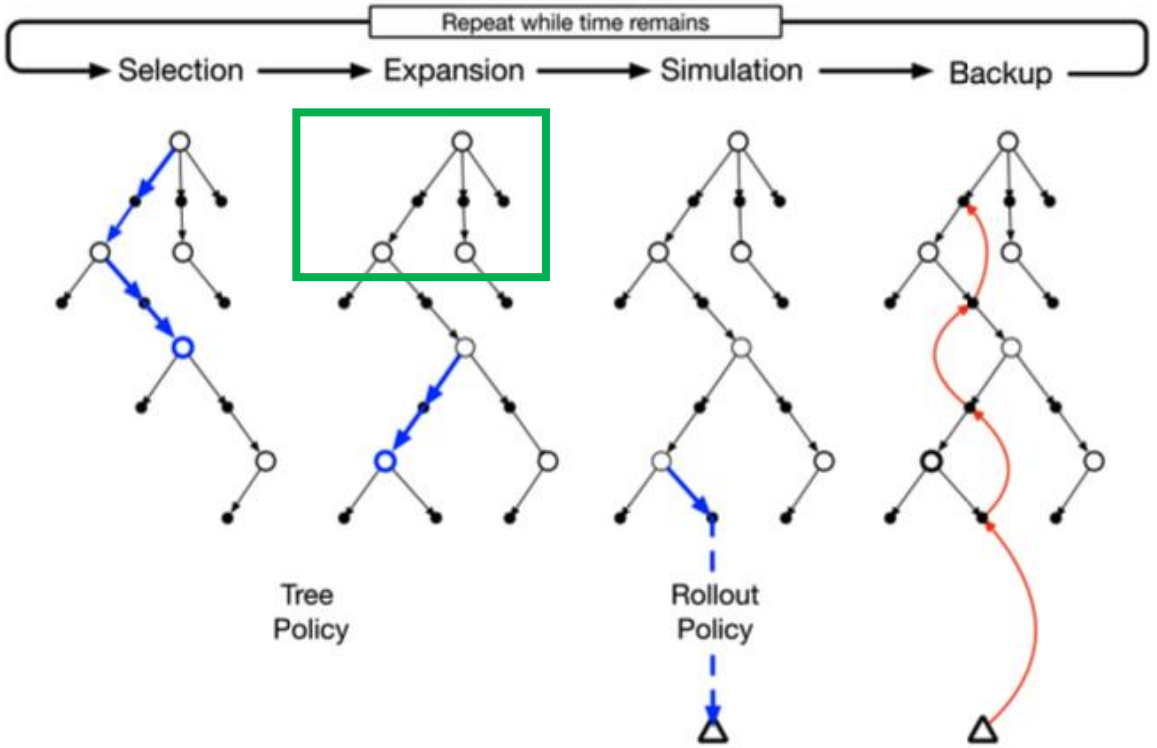
Monte Carlo Tree Search

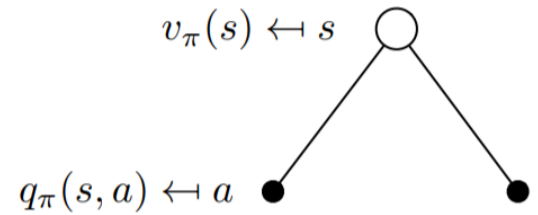
- 1.Tree traversal:
  - $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln(N)}{n}}$
- 2. Node expansion:
- 3. Rollout(random simulation):
- 4. Backpropagation

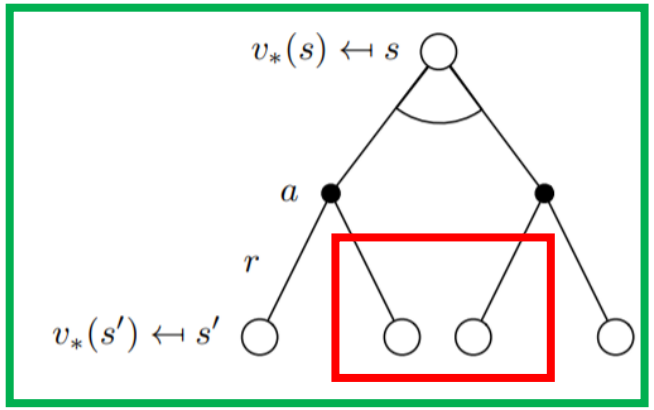


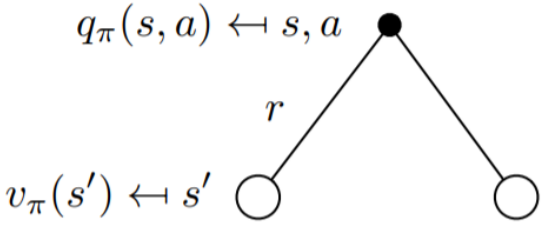
Markov Decision Process

- 1. State 1:
- 2. Take action 1
- 3. State 2
- 4. Reward







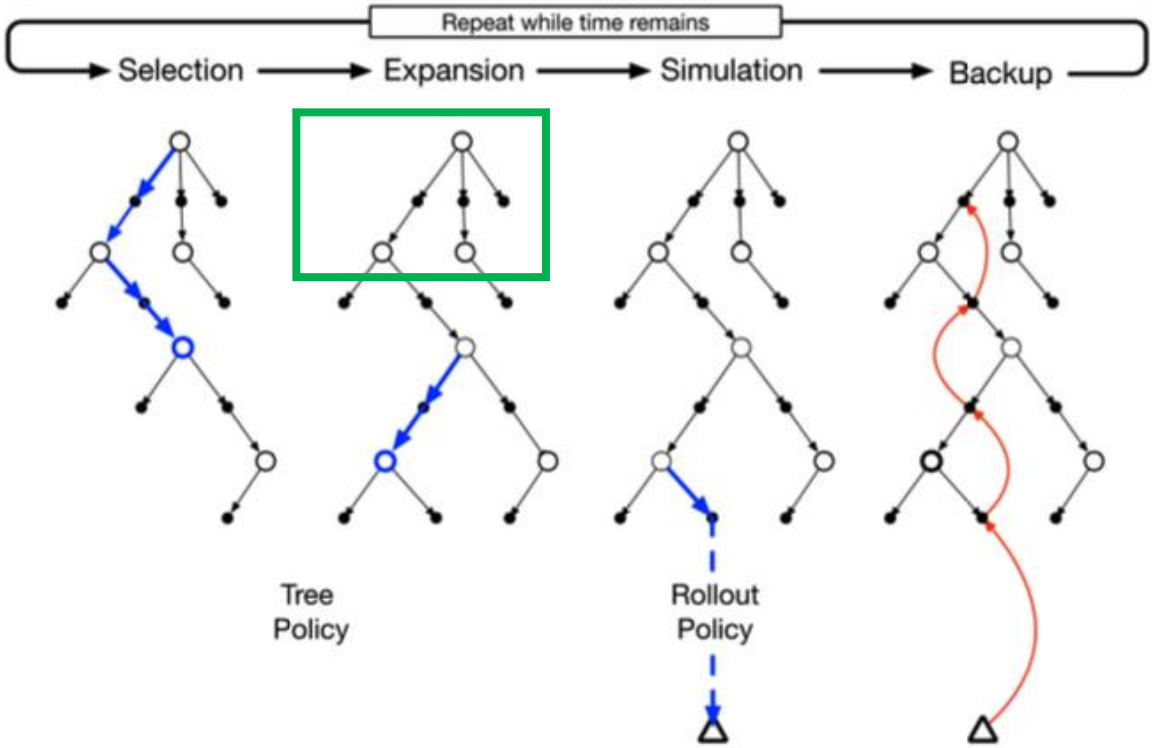


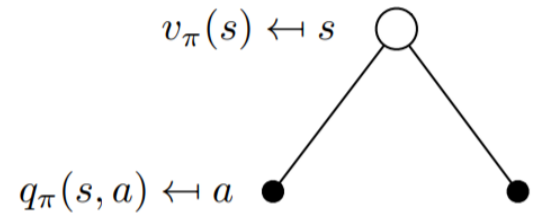
$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in S'} P(s' \mid a, s) V_\pi(s')$$

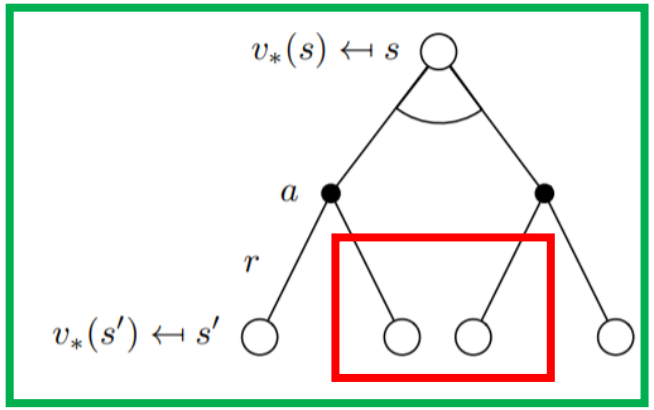
- 采取动作后转移的状态唯一
- 故删除框内的其他状态

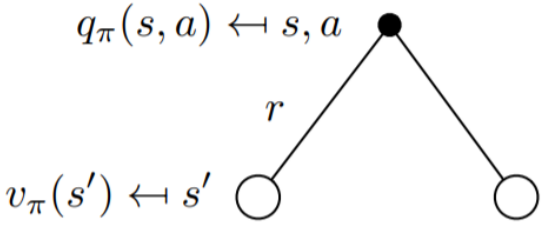
Markov Decision Process

- 1. State 1:
- 2. Take action 1
- 3. State 2
- 4. Reward





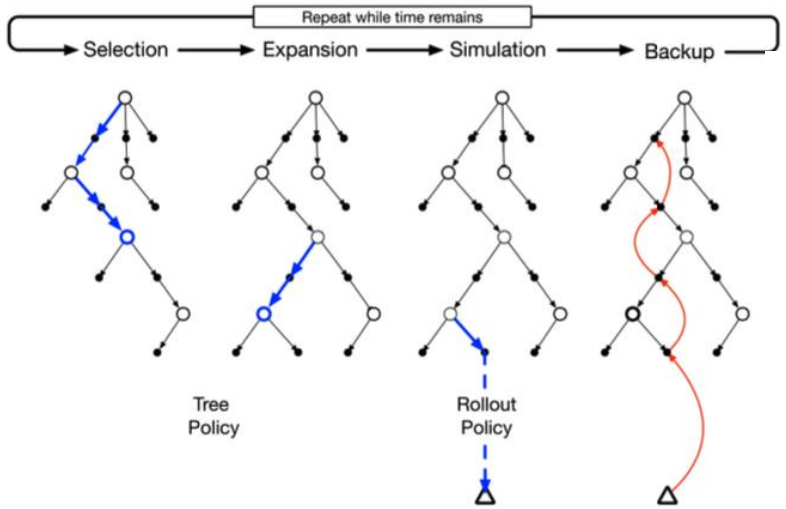


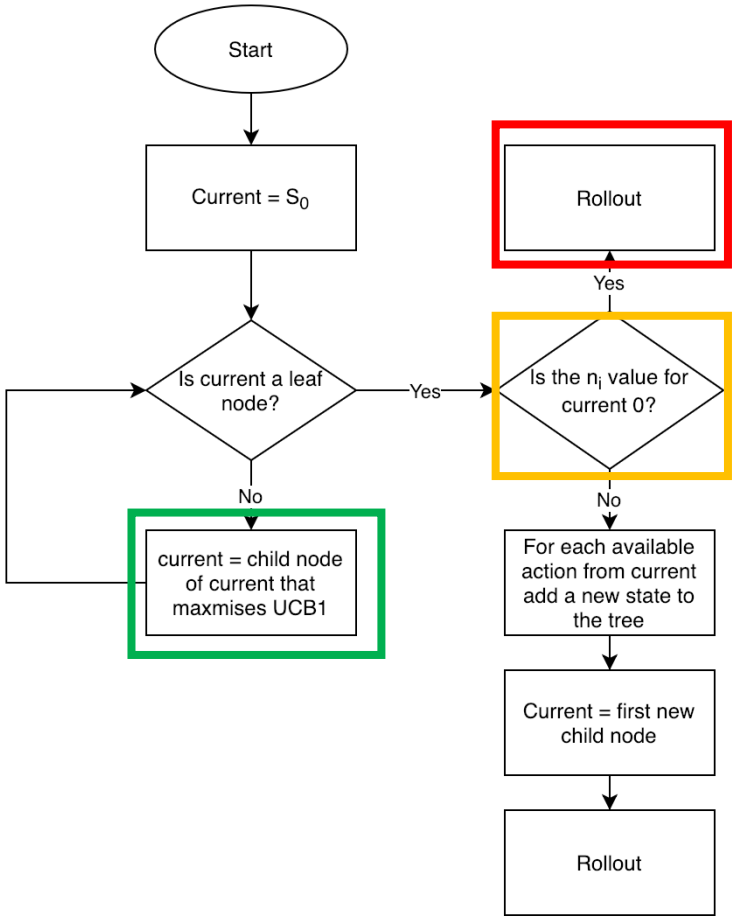


$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s \in S'} P(s' \mid a,s) V_\pi(s')$$

- 计算的是Q值，对动作打分
- 如最终决定游戏结束的是最后一个动作

MCTS Flowchart






```
Loop Forever:

if Si is a terminal state:

   return Value(Si)

Ai = random(available_actions(Si))

Si = Simulate(Si, Ai)

This loop will run forever until you reach a terminal state.
```

UCB Value

UCB1, or upper confidence bound for a node, is given by the following formula:

$$UCB1 = V_i + 2\sqrt{\frac{\ln N}{n_i}}$$

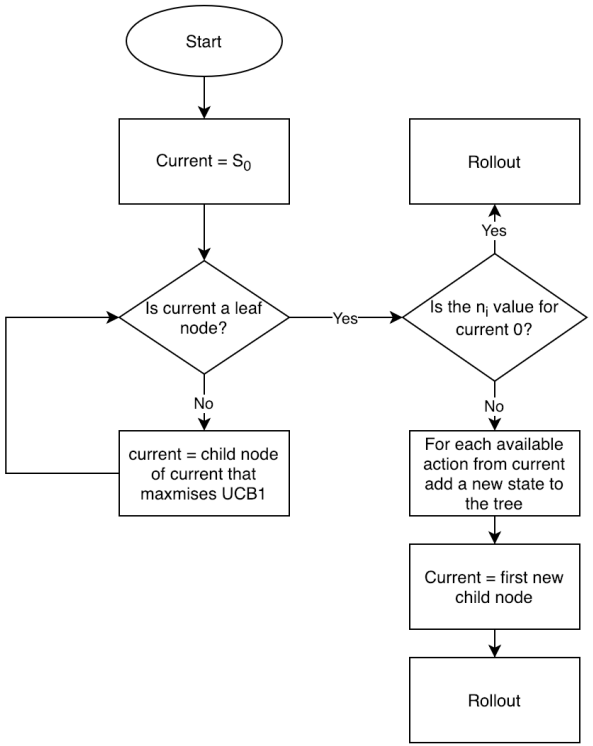
where,

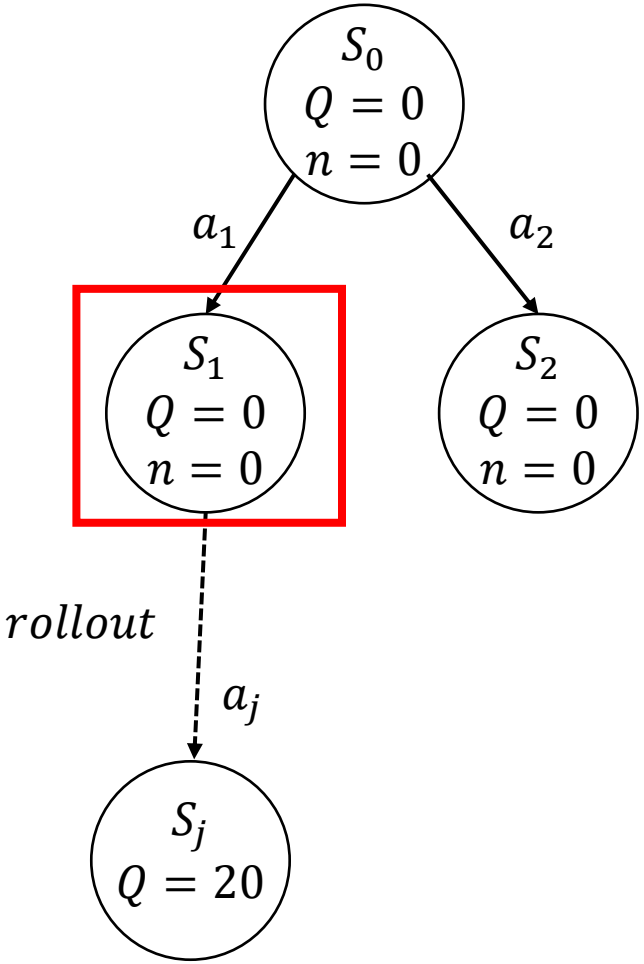
- Vi is the average reward/value of all nodes beneath this node
- N is the number of times the parent node has been visited, and
- ni is the number of times the child node i has been visited

- 判断该节点采样的次数
- 如果未被采样，则仿真到结束
- 否则，继续探索动作添加新的状态并仿真到结束

MDP+MCTS (1/4)

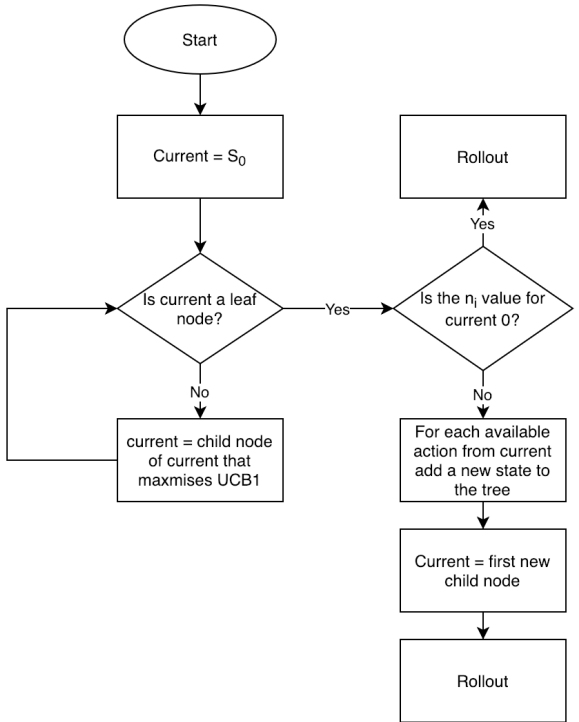
- **Trajectory:**
  - $s_0, a_1, s_1, \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 0, USB(s_1) = USB(S_2) \to \infty$

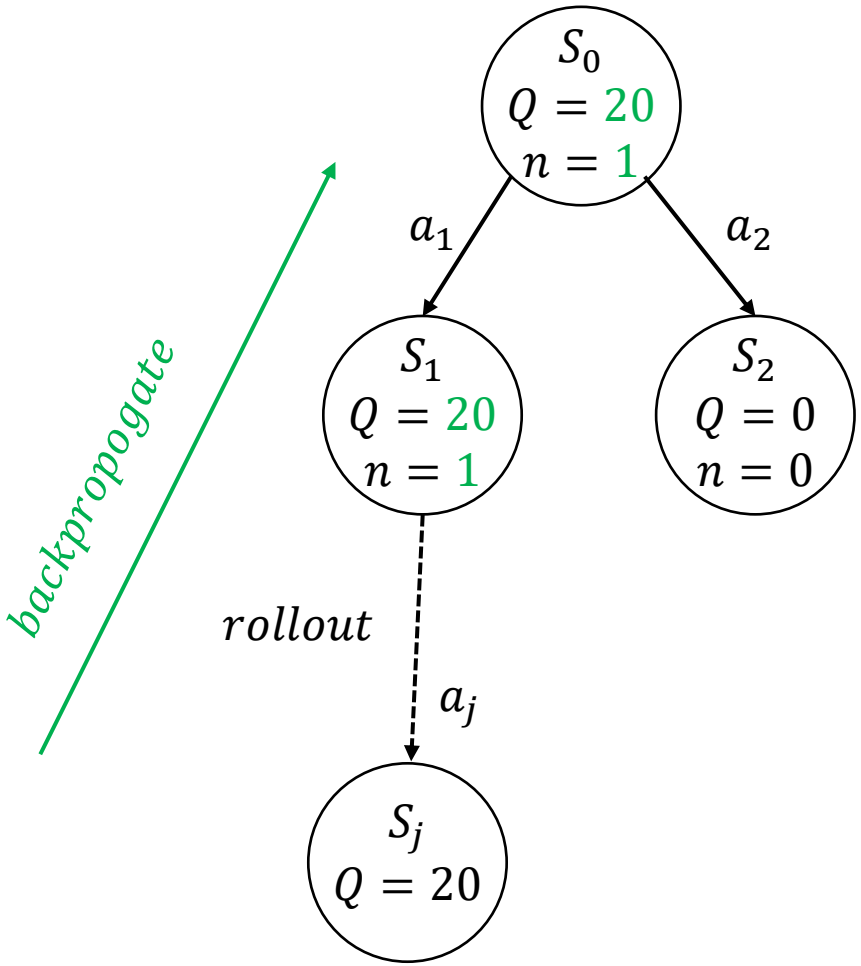




MDP+MCTS (1/4)

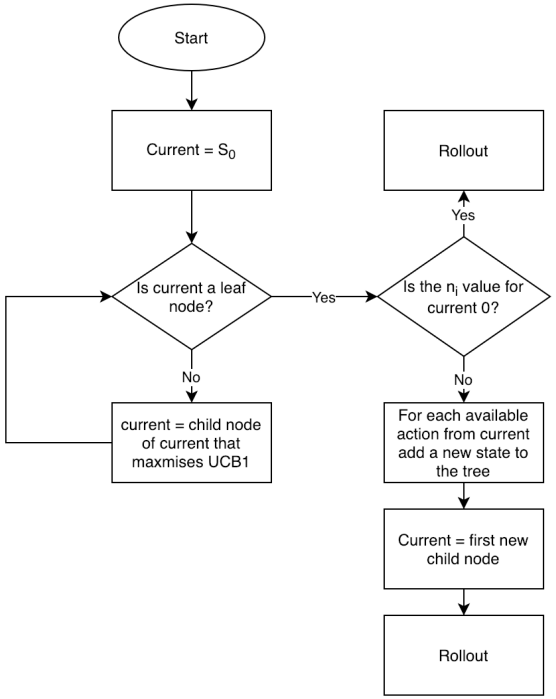
- **Trajectory:**
  - $s_0, a_1, s_1, \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 1, USB(s_1) = 20$

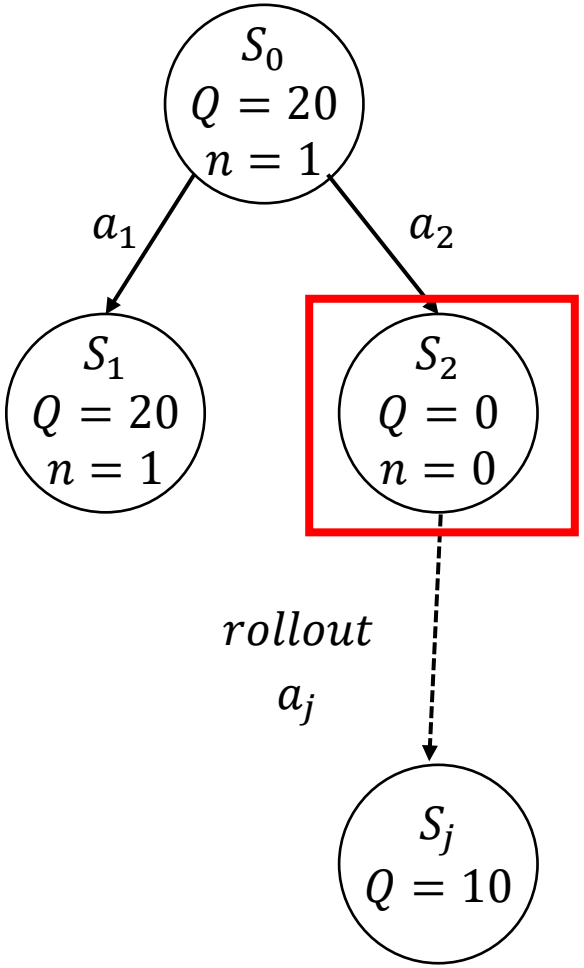




MDP+MCTS (2/4)

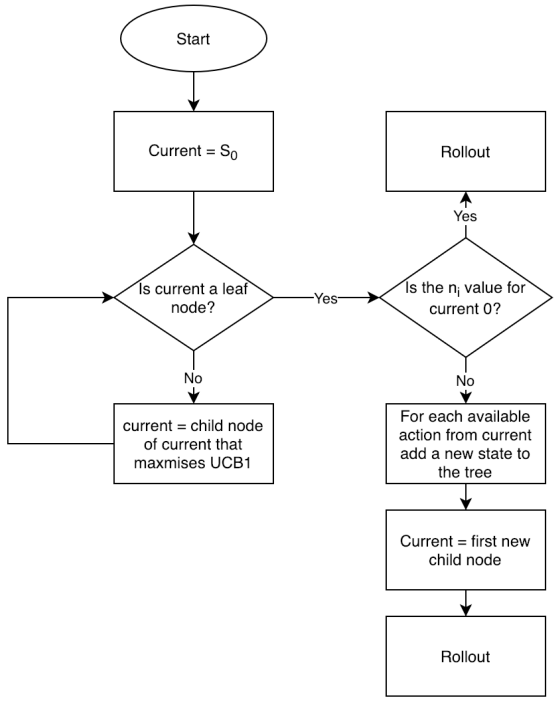
- **Trajectory:**
  - $s_0, a_2, s_2, \dots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 1, USB(s_1) = 20 + 2\sqrt{\frac{\ln 1}{1}}$
  - $n = 0, USB(s_2) \to \infty$

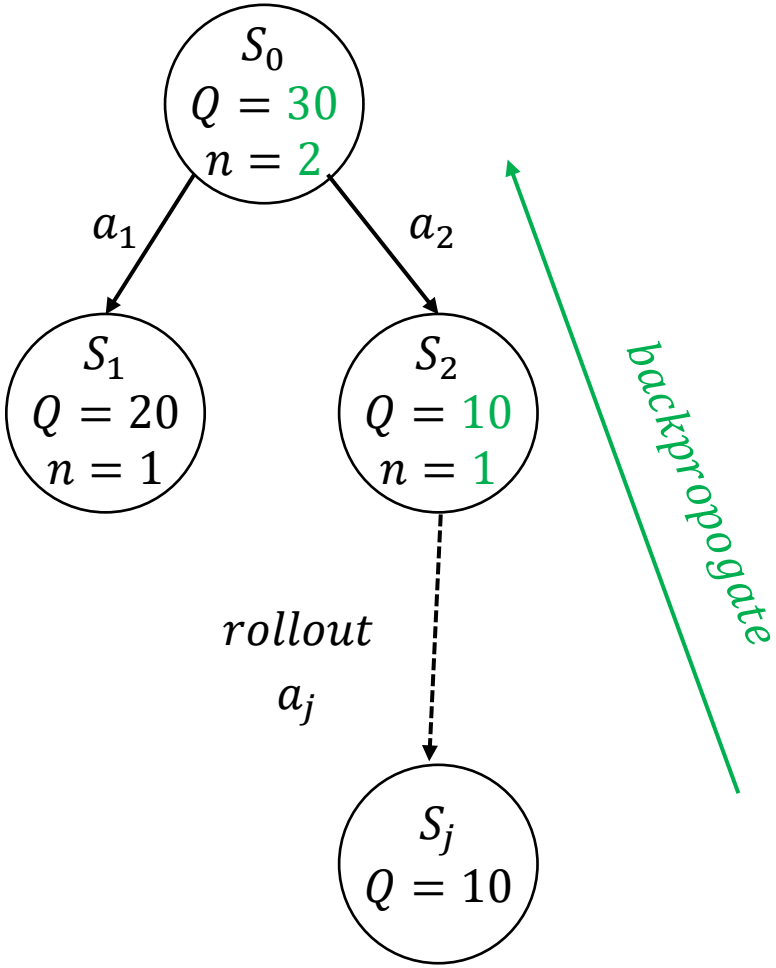




MDP+MCTS (2/4)

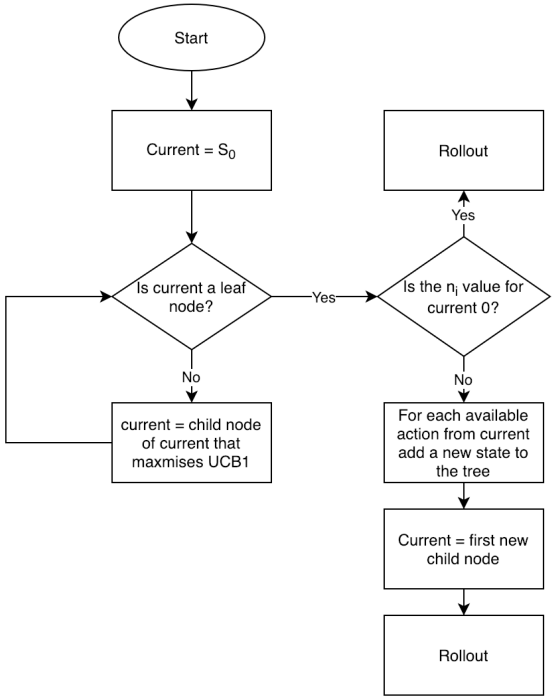
- **Trajectory:**
  - $s_0, a_2, s_2, \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$

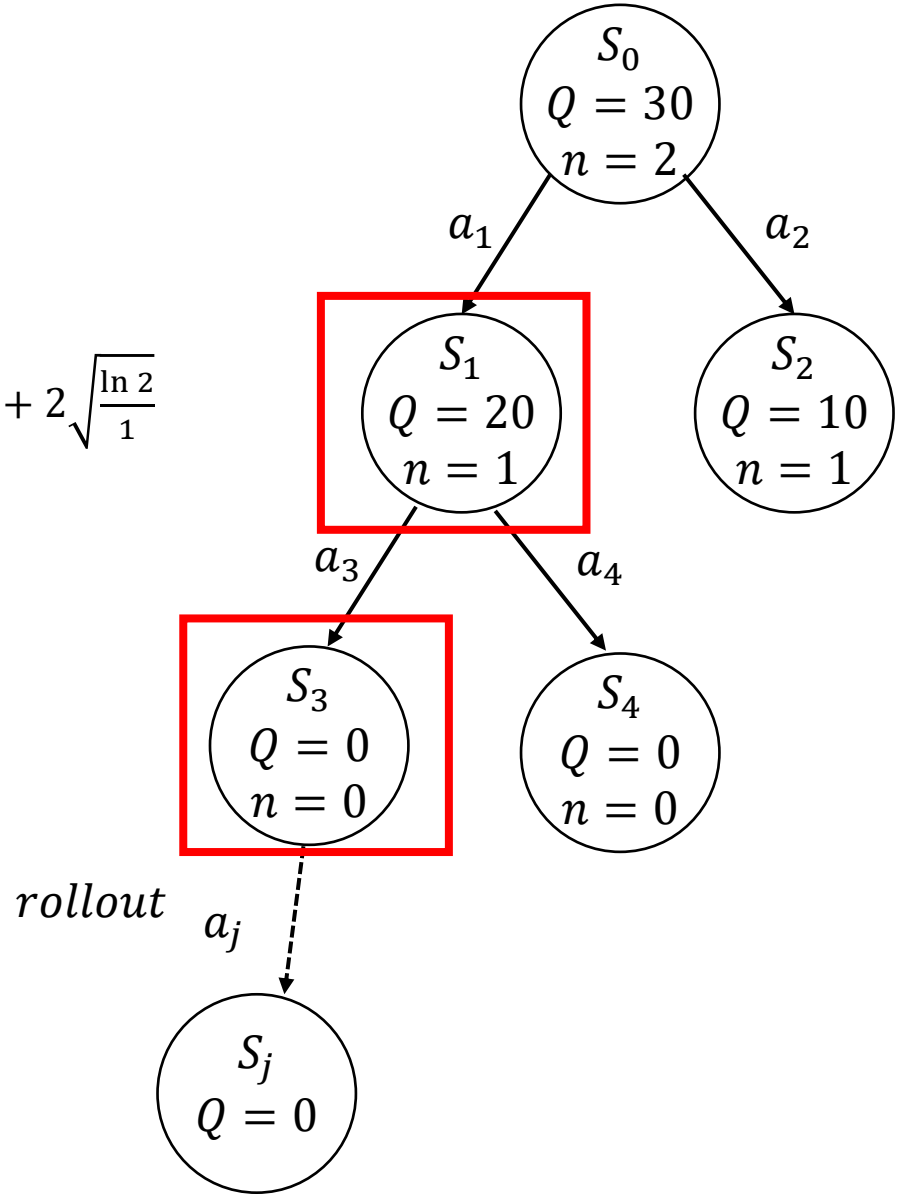




MDP+MCTS (3/4)

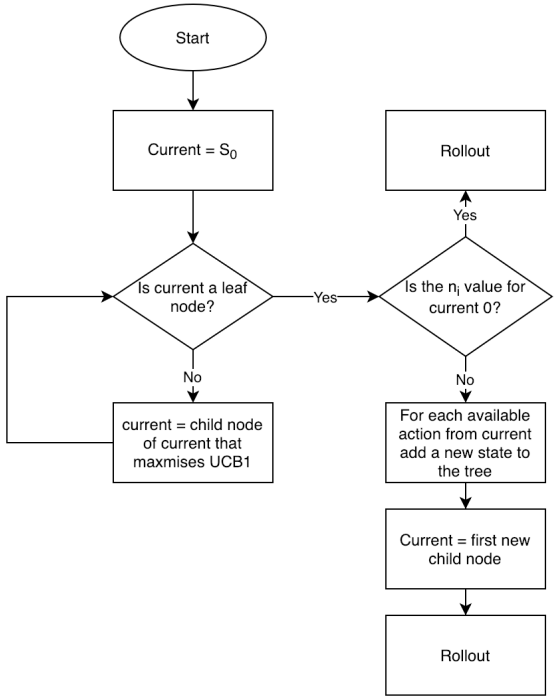
- **Trajectory:**
  - $s_0, a_1, s_1, a_3, s_3 \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 1, USB(s_1) = 20 + 2\sqrt{\frac{\ln 2}{1}} > USB(S_2) = 10 + 2\sqrt{\frac{\ln 2}{1}}$
  - $n = 0, USB(s_3) = USB(S_4) \to \infty$

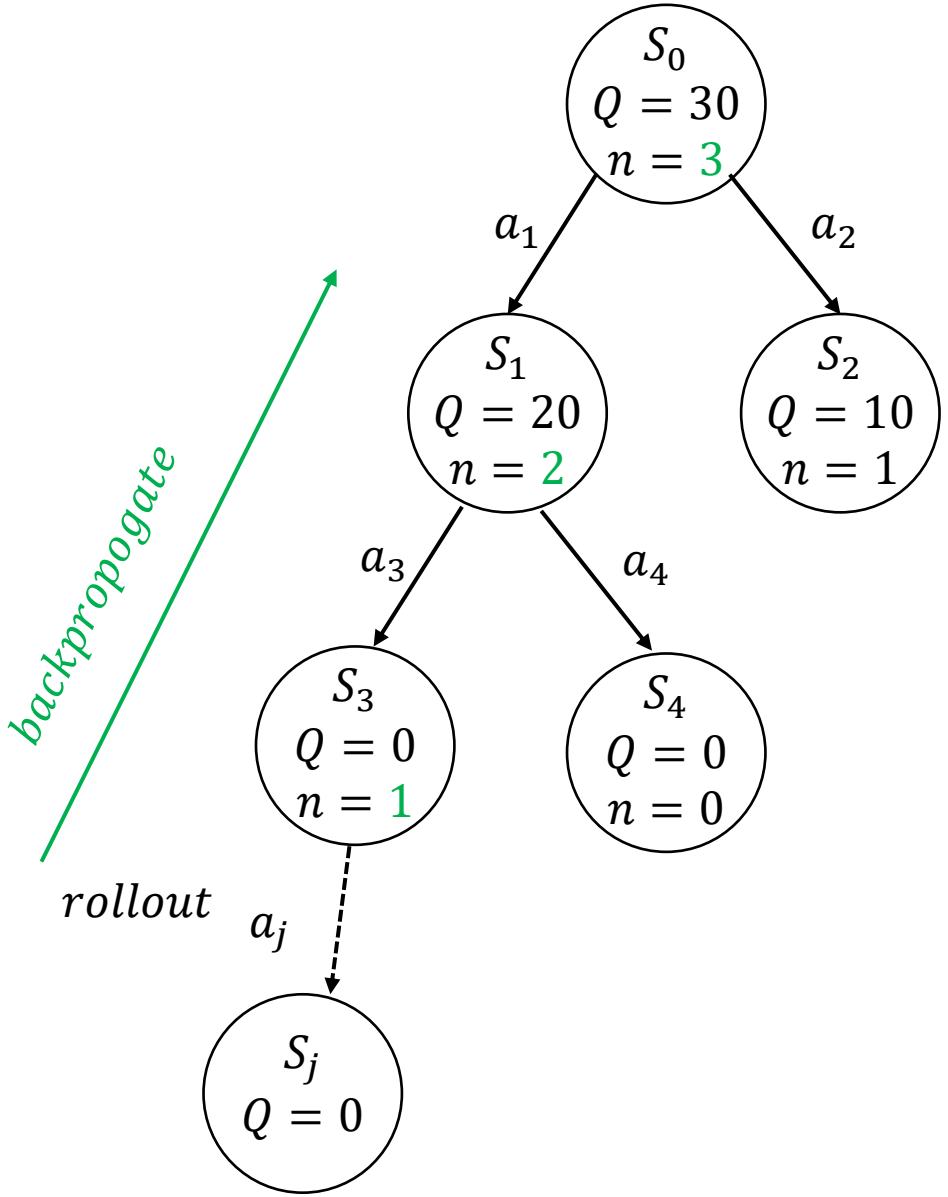




MDP+MCTS (3/4)

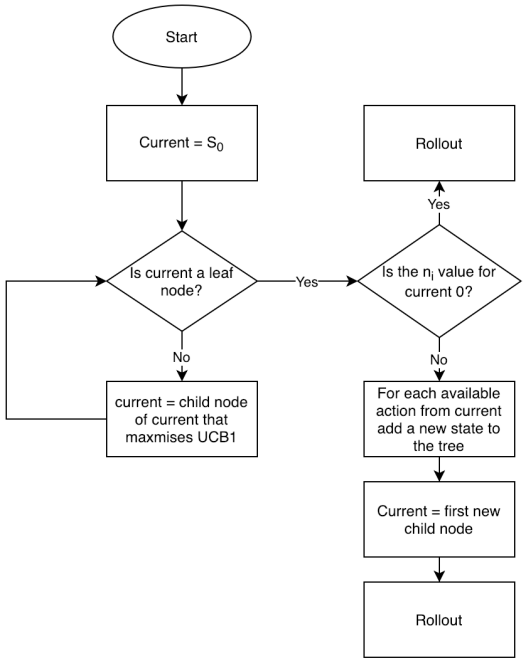
- **Trajectory:**
  - $s_0, a_1, s_1, a_3, s_3 \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 0, USB(s_3) = USB(S_4) \to \infty$

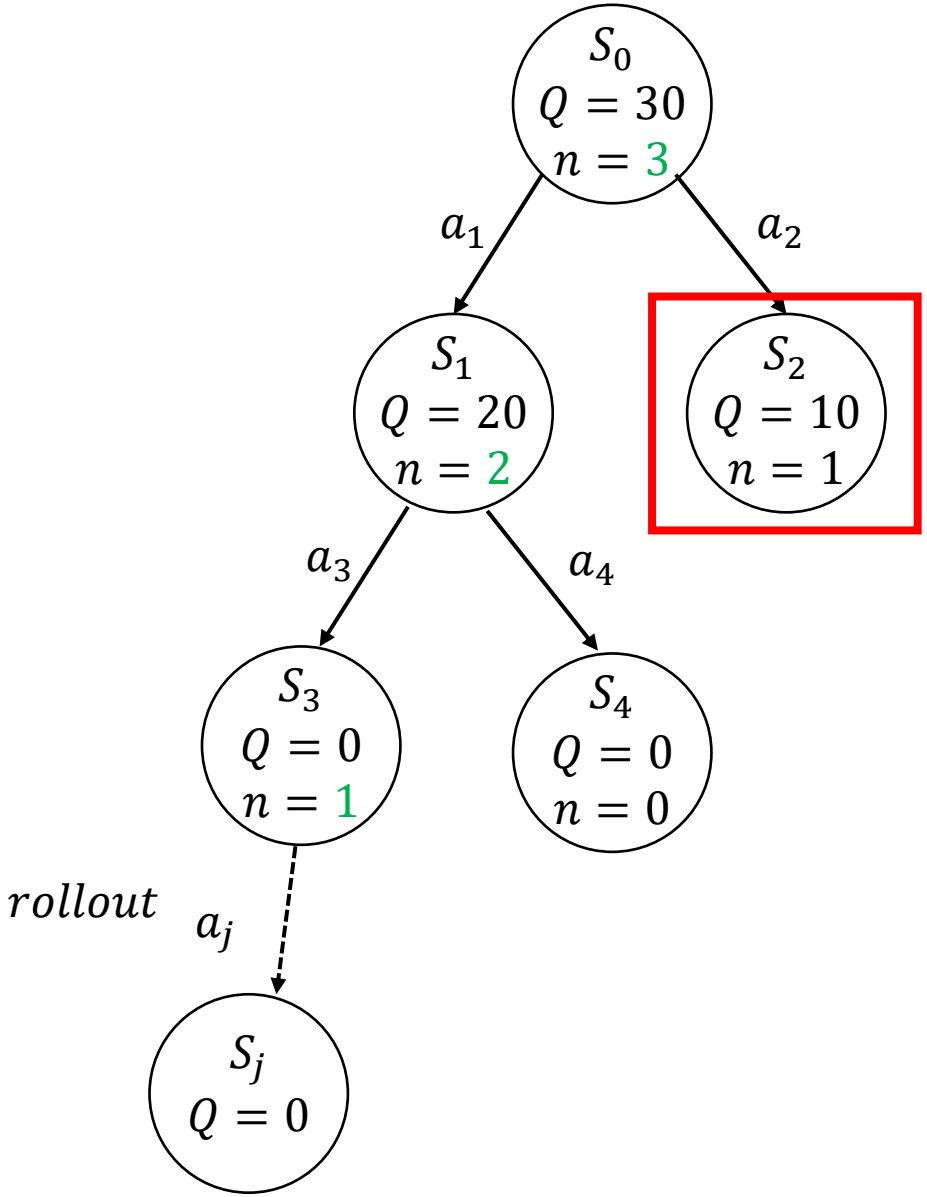




MDP+MCTS (3/4)

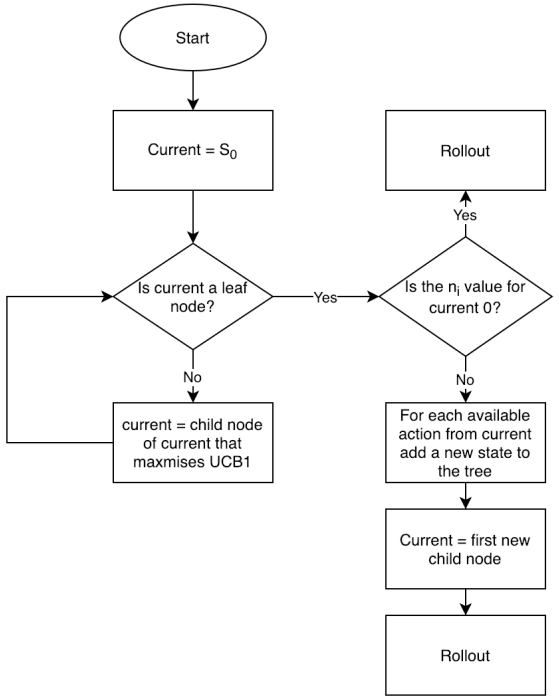
- **Trajectory:**
  - $s_0, a_1, s_1, a_3, s_3 \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 2, USB(s_1) = 10 + 2\sqrt{\frac{\ln 3}{2}} = 11.48$
  - $n = 1, USB(s_2) = 10 + 2\sqrt{\frac{\ln 3}{1}} = 12.10$

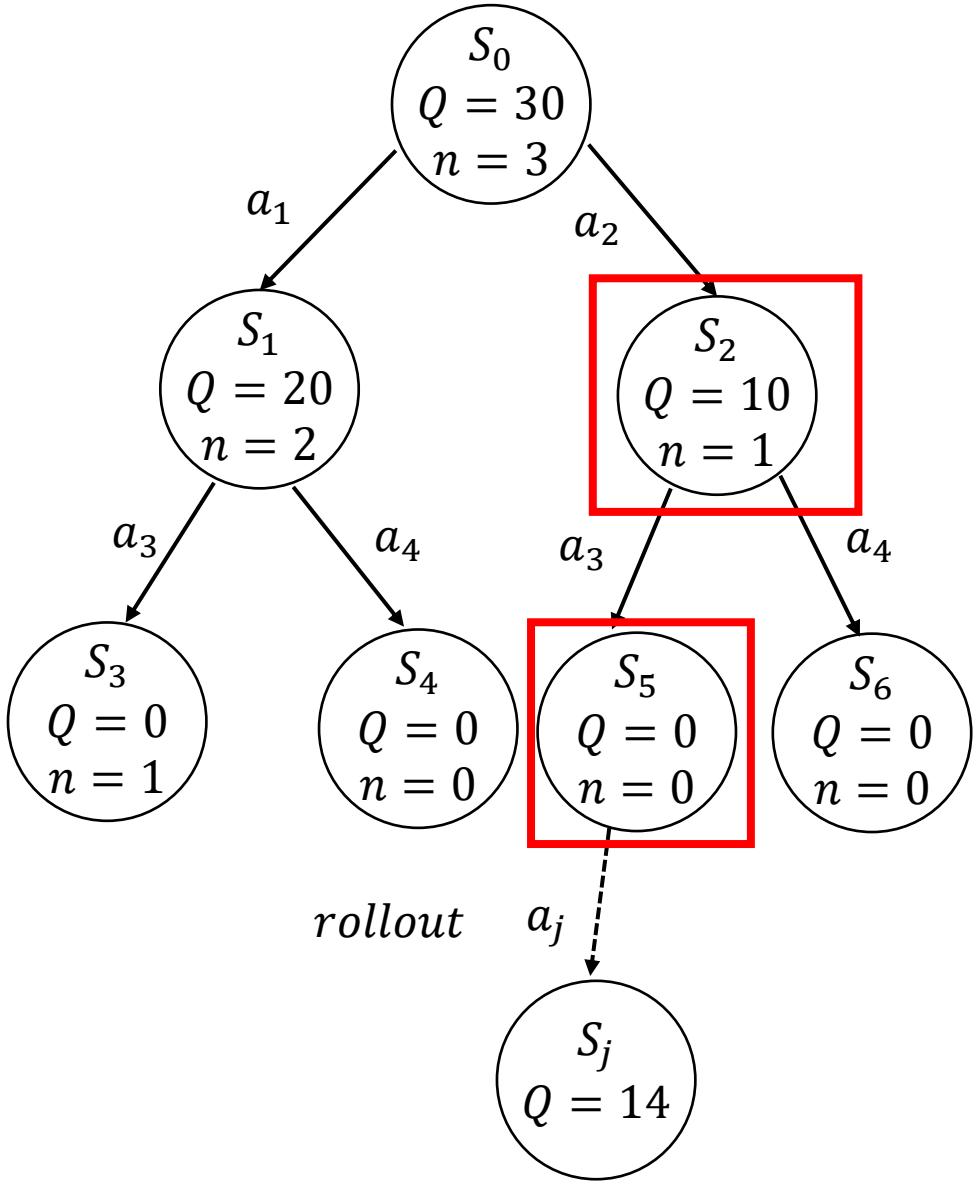




MDP+MCTS (4/4)

- **Trajectory:**
  - $s_0, a_2, s_2, a_3, s_5 \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 0, USB(s_5) = USB(S_6) \to \infty$





MDP+MCTS (4/4)

- **Trajectory:**
  - $s_0, a_2, s_2, a_3, s_5 \ldots, a_j, s_j, r$
- $USB(s_i) = \frac{Q}{n} + c\sqrt{\frac{\ln N}{n}} = \bar{Q} + c\sqrt{\frac{\ln N}{n}}$
  - $n = 2, USB(s_1) = 10 + 2\sqrt{\frac{\ln 4}{2}} = 11.66$
  - $n = 2, USB(s_2) = 12 + 2\sqrt{\frac{\ln 4}{2}} = 13.66$