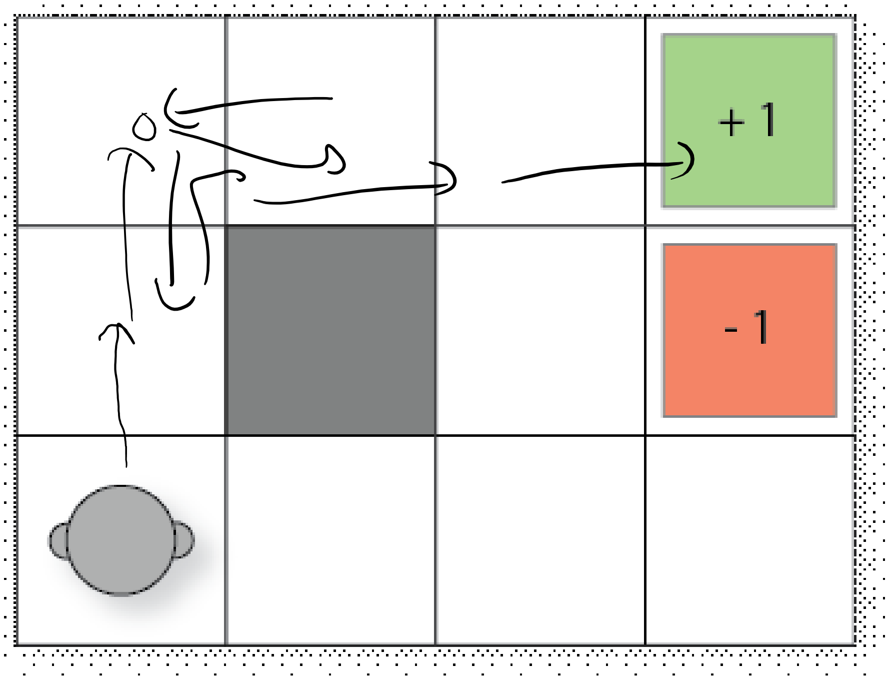
Monte-Carlo Tree Search

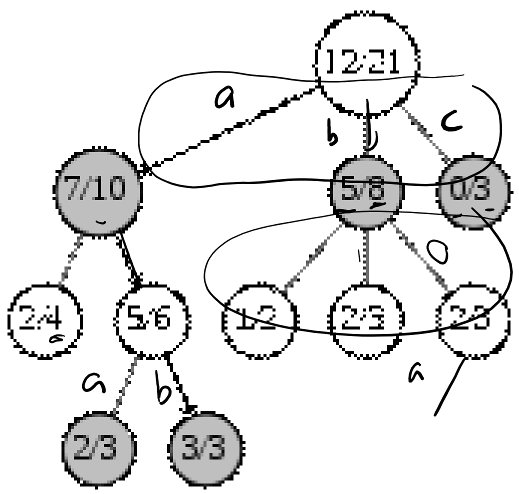
Tuesday, 11 September 2018 13:58 PM



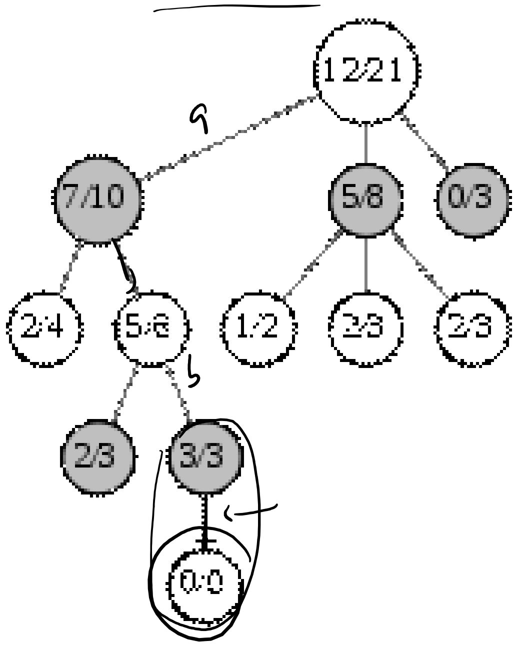

Monte-Carlo Tree Search (MCTS) overview

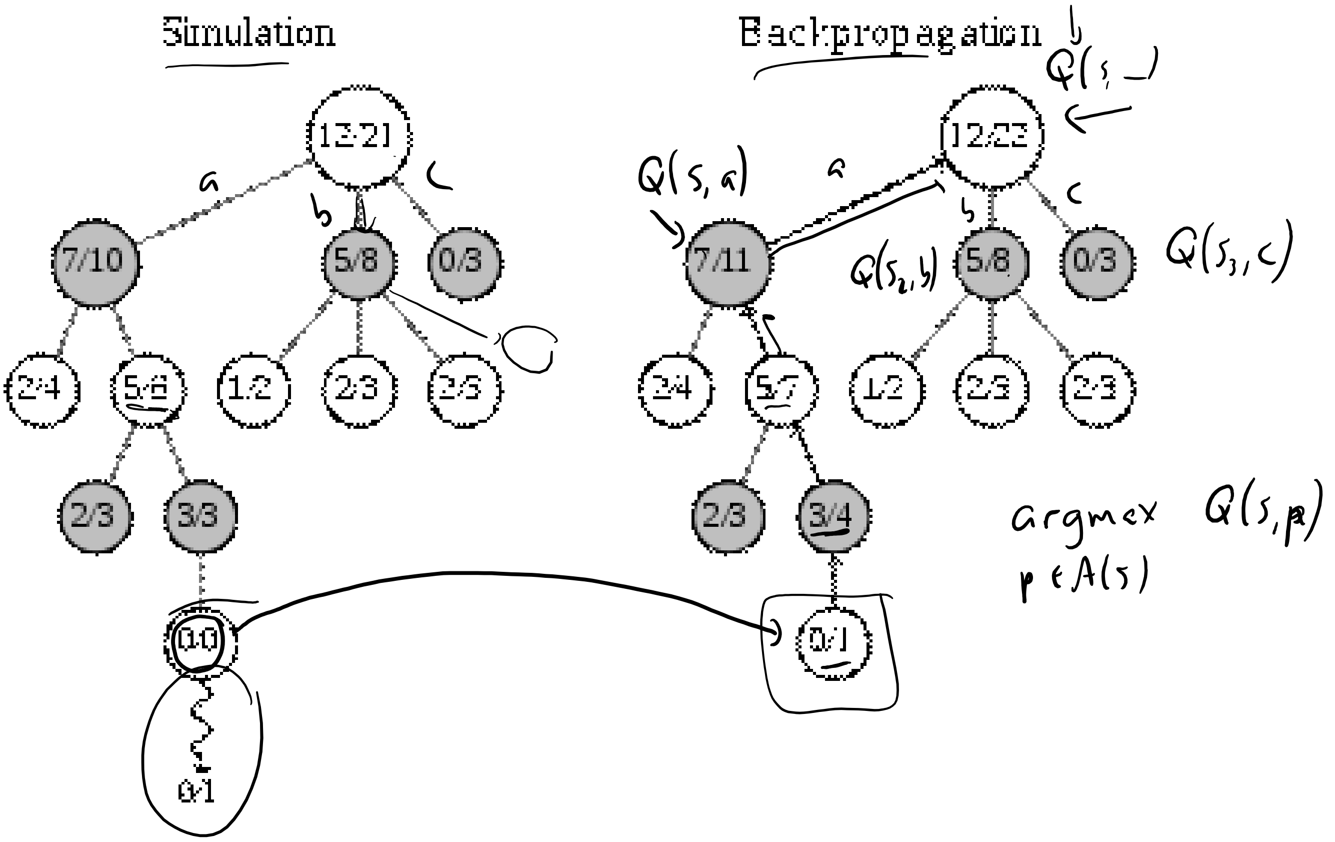
1. *Q(s,a)* is the Q-function: an estimate of the value of applying *a* in state *s.*
2. *Q(s,a)* is both the estimate but we will also use it as an heuristic.
3. The search tree is incrementally built.
4. MCTS is an *anytime* algorithm: we terminate whenever and give the best answer so far.

Selection



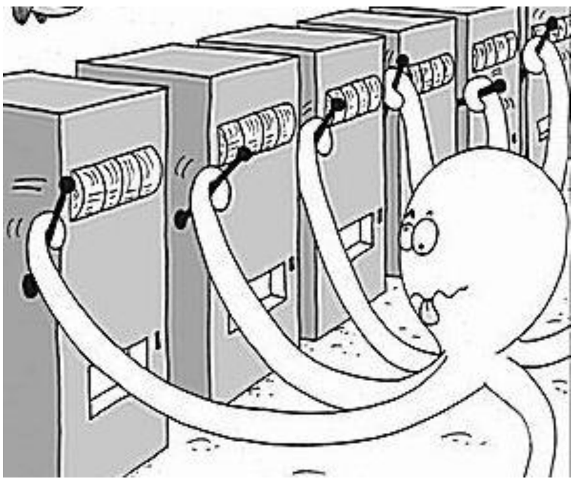

Expansion






Multi-armed bandits

*Imagine that you have N number of slot machines (or poker machines in Australia), which are sometimes called one-armed bandits. Over time, each bandit pays a random reward from an unknown probability distribution. Some bandits pay higher rewards than others. The goal is to maximize the sum of the rewards of a sequence of lever pulls of the machine.*



$\epsilon = 0.1$

Exploration vs exploitation

1. $\epsilon$-greedy: exploit best action with probability 1-$\epsilon$ and random action with probability
2. $\epsilon$-decreasing: $\epsilon$-greedy, but decrease $\epsilon$ over time
3. Softmax: select action with probability proportional to *Q(s,a)* so far

## Upper confidence bounds (UCB)

$$arg\max a \in A \left[ \underbrace{Q(s,a)}_{\text{exploit}} + \underbrace{\sqrt{\frac{2\ln N(S)}{N(a,s)}}}_{\text{exploit}} \right]$$

## Upper confidence tree (UCT)

UCT = UCB + MCTS (almost!)

$$arg\max a \in A \left[ Q(s,a) + 2\text{Cp}\sqrt{\frac{2\ln N(S)}{N(a,s)}} \right]$$

Cp is an exploration constant > 0

$Cp < \frac{1}{2} \Rightarrow$ exploit

$Cp = \frac{1}{2} \Rightarrow$ UCT

$Cp > \frac{1}{2} \Rightarrow$ explore

## UCT playing Mario Brothers: [A MCTS-based Mario-playing controller]


## UCT playing Freeway: [UCT Freeway - atari 2600]


Value/policy iteration vs. MCTS

Value/policy it => full policy
Computational