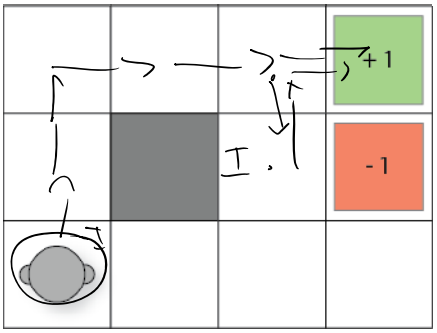
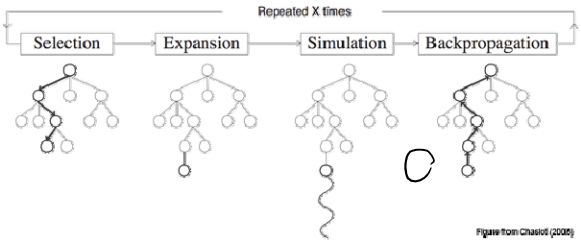
Monte-Carlo Tree Search

Monday, 21 September 2020 12:44 PM

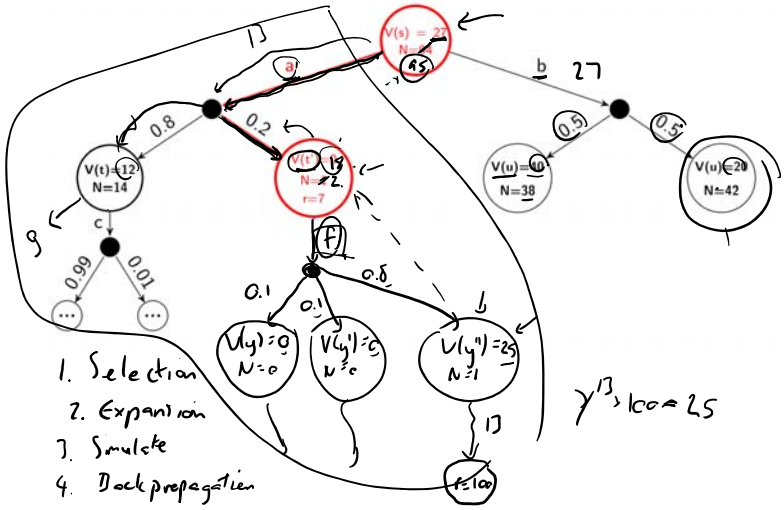


Online planning
Offline planning

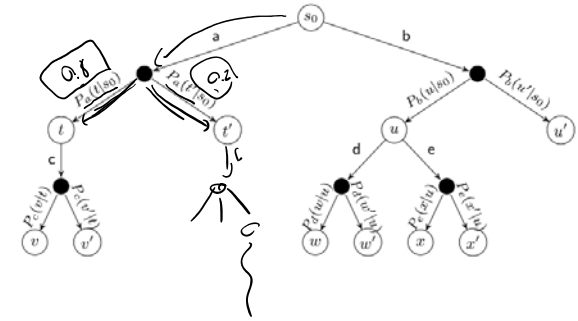


Monte-Carlo Tree Search (MCTS) overview

1. *V(s)* is the estimate of the value of a state.
2. 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.

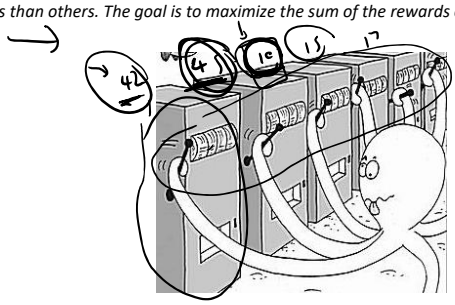


Simulation-based MCTS



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.



FOMO

Exploration vs exploitation

1. ϵ-greedy: ϵ in [0.1] (typically around 0.1) , choose the best with probability $1-\epsilon$, and other choose randomly
2. ϵ-decreasing: as above but ϵ decreases over time
3. Softmax: choose proportionally

$$V(a) = \frac{e^{Q(s,a)/\tau}}{\sum_{b=1}^{n} e^{Q(s,b)/\tau}}$$

Upper confidence bounds (UCB)

$$\pi(s) = \underset{a \in A}{\operatorname{argmax}} \; Q(s,a) + \underbrace{\sqrt{\frac{2 \ln N(s)}{N(s,a)}}}_{\text{exploration}} \qquad C_p > 0$$

exploit

$Q(s,a)$ estimated Q value

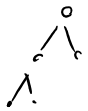
$N(s)$ number of visits to s

$N(s,a)$ number of times a executed in s

→ Upper confidence tree (UCT)

UCT = UCB + MCTS (almost!)

$$C_p = \frac{1}{\sqrt{2}}$$



Exercise

Your phone-stealing habits continue, but you are doing well and opening up to a new market of people. Each day, you can sell a bag of 20 iPhones or Samsung phones, but the price varies with each buyer, and you don't know the probability that people will buy them. You decide to alternate: iPhones one day, Samsung the next. After 100 days, you notice the following average return per day:

- Samsung: $400
- iPhone: $250

To help with your supply, you need to decide what you want to do the today. Assuming today will look similar to the previous 100 days, what should you do?

UCT playing Mario Brothers: A MCTS-based Mario-playing controller

UCT playing Freeway: UCT Freeway - atari 2600

Value/policy iteration vs. MCTS

	Value/policy iteration	MCTS
Cost	High	Low
Coverage/ Robustness	High	Low