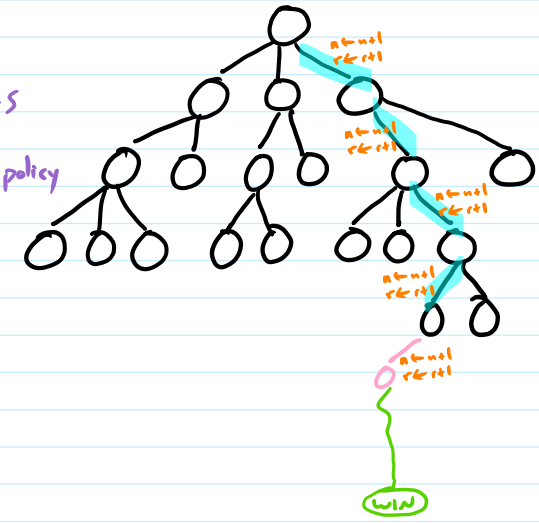
Monte Carlo Tree Search

UCT : UCB for trees
=
MCTS + UCB for tree policy

converges to minimax!



search tree: nodes = positions
branches = moves
children = resulting pos

$\bar{r}_j$ = observed score over all playouts after traversing tree to child j

n_j = # times we traversed tree through child j

T = # times through current node

Until out of time

traverse tree root → expandable node (not all children in tree) — tree policy

expand node (add missing child)

play from new node (playout) — default policy: random / heuristic (fast) (long games hurt)

update statistics along path through tree
- # times played
- total score

Advantages: always have move ready (gets better as time ?)

no domain knowledge except rules

Multi-Armed Bandit

Given unknown probability distributions $R_1, \dots, R_K$ with means $\mu_1, \dots, \mu_K$ — let $\mu^* = \max_i \mu_i$

Choose indices $i_1, i_2, \dots$ to optimize payout

Regret = difference between observed reward and best possible expectation

$$\rho_T = T \cdot \mu^* - \sum_{t=1}^{T} \hat{r}_t \qquad \hat{r}_t = \text{reward obtained on play } t$$

"optimal" means $P\left(\lim_{T\to\infty} \frac{\rho_T}{T} = 0\right) = 1$

	$\mu_1 = \frac{2}{3}$ Arm 1		$\mu_2 = \frac{7}{8}$ Arm 2		$\mu_3 = 2$ Arm 3	
Ex:	prob $\frac{1}{3}$	payout 2	prob $\frac{1}{4}$	payout $\frac{5}{4}$	prob $\frac{1}{100}$	payout 200
	$\frac{2}{3}$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{99}{100}$	0
			$\frac{1}{2}$	0		

uniform rotation: expected regret $\frac{4}{3}$ $\frac{7}{8}$ 0 $\frac{4}{3}$ $\frac{7}{8}$ 0 $\frac{4}{3}$ $\frac{7}{8}$ 0 ...

1 2 3 1 2 3 1 2 3 ...

$$\lim_{T\to\infty} \frac{\rho_T}{T} = \frac{59}{24}$$

expected regret for $T = 3n$ $= \frac{\frac{4}{3}n + \frac{7}{8}n + 0n}{3n} = \frac{59}{24}$

greedy: play each arm once (or n times), then arm with best observed reward forever

1 2 3 3 3 3 3 3 3 ...

w/prob $\frac{1}{100}$: 2 0 200

expected regret: $\frac{4}{3}$ $\frac{5}{4}$ 0 0 0 0 0 0 0 ...

$$\lim \frac{\rho_T}{T} = \frac{\frac{4}{3} + \frac{7}{8}}{T} = 0$$

w/prob $\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{99}{100}$: 1 2 3 1 1 1 1 1 ...

2 0 0

expected regret: $\frac{4}{3}$ $\frac{4}{3}$ ...

$$\lim_{T\to\infty} \frac{\rho_T}{T} = \frac{4}{3} \qquad \text{so } P\left(\lim \frac{\rho_T}{T} = 0\right) \neq 1$$

ε-greedy: play each once, then play random w/prob ε, arm with best avg reward so far otherwise

play best observed machine w/ prob $(1-\varepsilon) + \frac{\varepsilon}{K}$ — expected regret 0 if best observed = actual best

each other machine: w/prob $\frac{\varepsilon}{K}$ $(\mu^* - \mu_i) > 0$

expected regret $(1-\varepsilon) + \frac{\varepsilon}{K}\cdot 0 + \underbrace{\sum_{i \text{ not optimal}} \frac{\varepsilon}{K}\cdot(\mu^* - \mu_i)}_{\text{positive}} > 0$

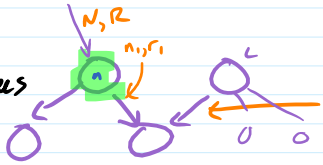
zero regret (optimal!)

Choose arm j that maximizes $\bar{r}_j + \sqrt{\frac{2 \ln T}{n_j}}$

- $\bar{r}_j$: avg observed reward for arm j (exploitation term)
- T: total plays over all arms
- n_j: number of plays for option j
- 2: tunable parameter
- $\sqrt{\frac{2 \ln T}{n_j}}$: exploration term

UCB: upper confidence bound

But many games are DAGs, not trees



standard UCT uses n_1, r_1, N in UCB formula

but can use n_1, n_2, r_1, r_2 for exploit term
n_1, N for explore term

(using n_1, n_2, r_1, r_2 in both affects convergence)