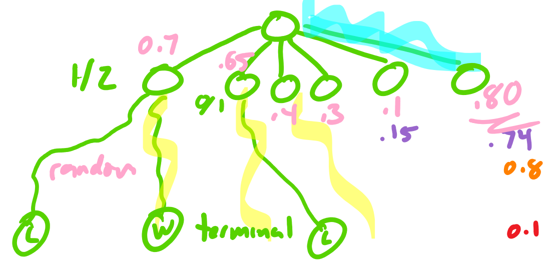
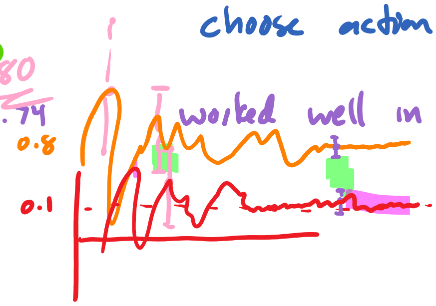
Monte Carlo Techniques

Flat Monte Carlo: for each action
- simulate to terminal using random play
- sample many times
- choose action w/ highest observed avg

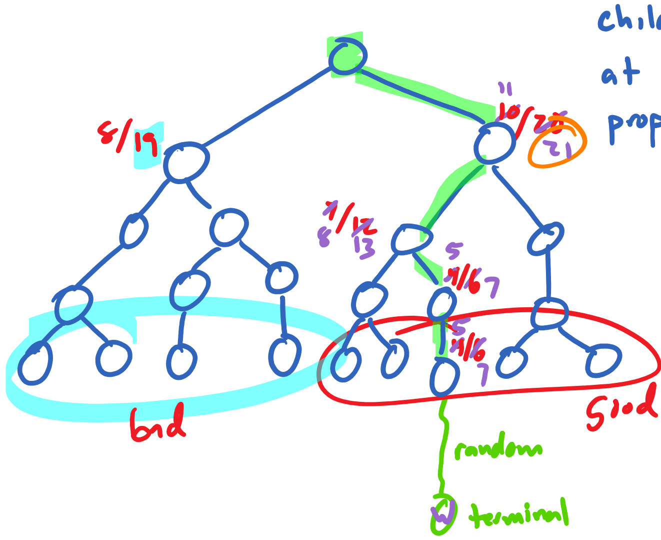




worked well in Scrabble, Bridge

Combine with UCB: choose action max $\bar{r}_j + \sqrt{\frac{2 \ln T}{n_j}}$

Combine with tree search: (Flat UCB)
- build tree to fixed depth
- traverse to a leaf using UCB on children at each level
- at leaf, do random playout
- propagate result back up, accumulating stats in nodes



Grow Tree asymmetrically: Monte Carlo Tree Search

Multi-Armed Bandit

Given unknown probability distributions $R_1, \ldots, R_K$
with means $\mu_1, \ldots, \mu_K$ $\quad \mu^* = \max_i \mu_i$

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

Regret = difference between obtained reward and best possible expected reward

$$\rho_T = T \cdot \mu^* - \sum_{t=1}^{T} \hat{r}_t$$

$\hat{r}_t$ = reward you got at time t (from playing i_t)

"optimal" means zero average regret

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

Ex:

Arm 1 prob	Arm 1 payout	Arm 2 prob	Arm 2 payout	Arm 3 prob	Arm 3 payout
$\frac{1}{3}$	2	$\frac{1}{4}$	3	$\frac{1}{100}$	200
$\frac{2}{3}$	0	$\frac{1}{4}\cdot\frac{1}{100}$	$\frac{1}{2}$	$\frac{99}{100}$	0
		$\frac{1}{2}$	0		

$\mu_1 = \frac{1}{3}\cdot 2 + \frac{2}{3}\cdot 0 = \frac{2}{3}$

$\mu_2 = \frac{1}{4}\cdot 3 + \frac{1}{4}\cdot\frac{1}{2} + \frac{1}{2}\cdot 0 = \frac{7}{8}$

$\mu_3 = 2$

uniform rotation: cycle through each arm 1 2 3 1 2 3 . . . . .

$$\lim_{T\to\infty}\frac{\rho_T}{T} = \lim_{T\to\infty}\frac{T\cdot\mu^* - (\hat{r}_1 + \hat{r}_2 + \hat{r}_3 + \cdots)}{T}$$

$$= \lim_{T\to\infty}\frac{T\mu^* - (\hat{r}_1 + \hat{r}_4 + \hat{r}_7 + \cdots) - (\hat{r}_2 + \cdots) - (\hat{r}_3 + \cdots)}{T}$$

$$= \mu^* - \underbrace{\left(\frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3\right)}_{< \frac{1}{3}\mu^* + \frac{1}{3}\mu^* + \frac{1}{3}\mu^* = \mu^*}$$

$$> 0$$

greedy : play each machine once, then machine w/ highest initial payoff

1 2 3 1 1 1 1 . . . . prob I do this = $\frac{1}{3}\cdot\frac{3}{4}\cdot\frac{99}{100} = \frac{99}{400} > 0$

1 2 3 2 2 . . . . .

1 2 3 3 3 3 . . . . .

$\lim_{T\to\infty}\frac{\rho_T}{T} = \mu^* - \mu^1 = 2 - \frac{2}{3} > 0$

so $P\left(\lim_{T\to\infty}\frac{\rho_T}{T} = 0\right) < 1$

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

ε-greedy : play one round, then play best observed w/ prob $1-\varepsilon$
play randomly w/ prob ε

$$\lim_{T\to\infty}\frac{\rho_T}{T} > \underbrace{\frac{\varepsilon}{k}}_{>0} \cdot \frac{(\mu^* - \mu^{\hat{i}})}{>0} > 0$$

$\bar{r}_2 \qquad \bar{r}_5$

tunable parameter

exploration

total number of plays

zero regret

Choose i that maximizes $\bar{r}_i + \sqrt{\frac{2 \ln T}{\ldots}}$

zero regret

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

exploration

total number of plays

UCB

upper confidence bound

mean observed reward of arm j

exploitation

number of times played arm j

Monte Carlo Tree Search

Until out of time

(for example, UCB)
tree policy

traverse tree   root ⇝ leaf

MCTS+UCB
=
UCT

expand   if leaf is expandable, add its children
↳ nonterminal and non-zero visits

simulate   play to terminal pos (from arb. selected added child if expanded)
default policy (for ex. random)

update   backpropagate stats along path through tree
from point simulation started ⇝ root

Select move from root based on current stats (choose child w/ highest avg or highest visit count)