$S = \{1, 2\}$ $\quad R(s,a) = s^2$ $\quad V^\pi(1) = 37$

$V^*(2)$

$V^*(2) = E\sum \gamma^t R(s,a)$ $\quad R(s,a) \leq 4$

$\leq \frac{4}{1-\gamma} = 40$

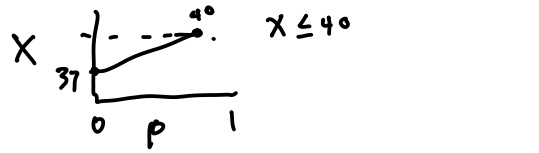
$V^\pi(1) = 37 = R(1, \pi(1))$

$1 + \gamma E[V^\pi(s')]$

$E[V^\pi(s')] = 40$

$p = T(2 | 1, \pi(1))$

$E[V^\pi(s')] = 40 = pX + [1-p]37$



Last time:

What if S is continuous (not LQR)

- Approximate DP
- Direct Policy Opt.

Offline

What if problem to big for offline

Online

"Offline"

Before execution: Find V^*, Find $Q^*(s,a) = R(s,a) + \gamma E[V^*(s')]$

During execution: $\pi^*(s) = \arg\max_a Q^*(s,a)$

"Online"

Before: do nothing

During Execution: Consider action and consequences "From the current state"

Cross Entropy (Direct Policy Optimization) (Offline)

- Genetic
- Sim Annealing
- Cross Entropy

d = initial dist for θ

loop

sample n θ's from d

evaluate

choose m elite θs

d = fit(elite θs)

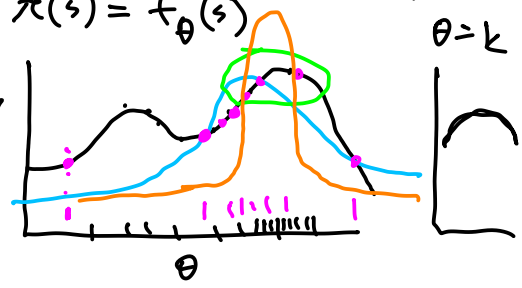
$\pi(s) = f_\theta(s)$

in HW $\theta = k$

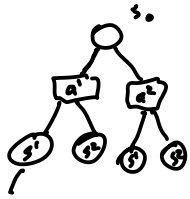
$E[V(s_0)]$ some initial state dist

evaluate θ

run many MC simulations and est. $E[V(s_0)]$



at state s, want to choose action "planning"



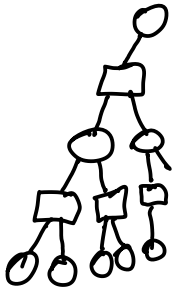
$Q(s_0,a^2) = R(s_0,a^2) + \sum T(s'|s_0,a^2)\, V(s')$

$V(s^1)$ $V(s^2)$

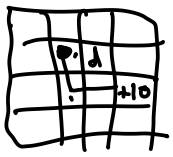
$Q^\pi(s,a) \equiv R(s,a) + \gamma E[V^\pi(s')]$

$V^\pi(s) = \max Q^\pi(s,a)$

Forward Search



$O((|A| \times |S|)^d)$



Branch and Bound

$\bar{Q}(s,a)$ $\frac{+10}{1-\gamma}$

$\underline{V}(s)$

$\gamma^d 10$

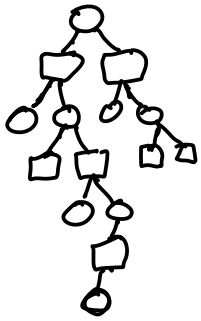
$\bar{Q}(s,a^2) < Q(s,a^1)$

Heuristic Search: Not in book but should work well

Maintain $\bar{V}, \underline{V}$ $\bar{Q}, \underline{Q}$

- expand actions with highest $\bar{Q}$
- states with biggest $\bar{V} - \underline{V}$

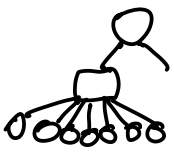
at root argmax $\underline{Q}$

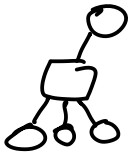


$|S(s,a)| << |S|$

Sparse Sampling

limit # children to n





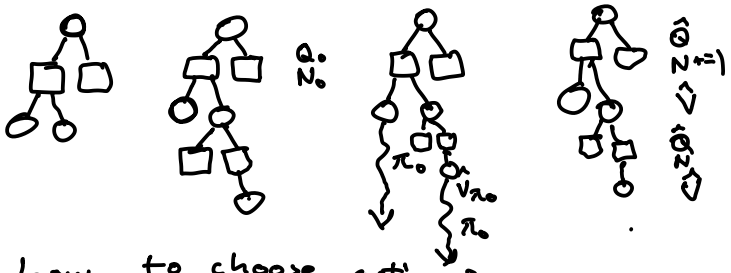
only need G

$O((n|A|)^d)$

MCTS

MCTS

Search Expansion Rollout Backup



how to choose actions?

$\arg\max \hat{Q}$

UCB

$$\arg\max_a \hat{Q}(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}$$

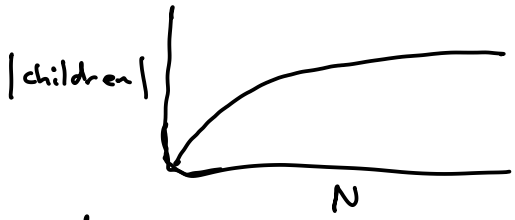
(c: tuning parameter; $\sqrt{\frac{\log N(s)}{N(s,a)}}$: exploration bonus)

MCTS + UCB = UCT

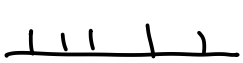
S, A continuous

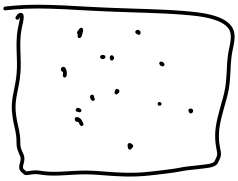
Double Progressive Widening (DPW)

limit |children| to $k N^{\alpha}$ $\alpha < 1$ $\alpha = \frac{1}{2}$



A

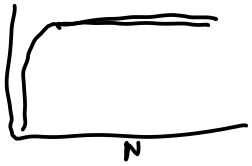




~~only~~ A - only works for 1 dimension

S - $k = \infty$ $\alpha = \frac{1}{30}$

Sparse Sampling with $n = 8$



Sparse UCT

π_θ = MCTS $\theta = c, k_s, \alpha_s, k_a, \alpha_a$

Cross Entrop