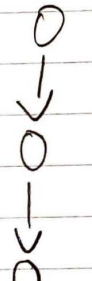
Input π_i to be eval

Initialize:
$V(s) \in \mathbb{R}$ arbitrarily for all $s \in S$
Returns$(s) \leftarrow$ empty list $\forall s \in S$

Loop forever (for each episode)
gen an ep following π: $S_0, A_0, R_1, S_1, A_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each state in the ep, $t = T-1, T-2, \ldots, 0$
$G \leftarrow \gamma G + R_{t+1}$
Unless S_t in $S_0, S_1, \ldots S_{t-1}$:
Append G to Returns(S_t)
$V(S_t) \leftarrow$ average(Returns(S_t))

Simulate many games using a policy with an MC approach then average the returns for each state

No need to pre-comp probs



policy improvement - use greedy

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a))$$

$$= \max_a q_{\pi_k}(s,a)$$

$$\geq q_{\pi_k}(s, \pi_k(s))$$

$$\geq V_{\pi_k}(s)$$

$$\pi_0 \to q_{\pi_0} \to \pi_1 \to q_{\pi_1} \to \quad \to \pi_* \to q_*$$

Init

$\pi(s) \in A(s)$ arbitrarily $\forall s \in S$

$Q(s,a) \in \mathbb{R}$ $\forall s \in S, a \in A(s)$

Returns $(s,a) \leftarrow$ empty list

Loop Forever (for each ep)

Choose $S_0 \in S$, $A_0 \in A(S_0)$ randomly $(p(s_0, A_0) > 0 \ \forall A_0, S_0)$

generate an ep π: $S_0 A_0 R_1 \cdots S_{T-1} A_{T-1} R_T$

$G \leftarrow 0$

Loop for each step $t = T-1, \ldots$

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t in our ep

Append G to Returns (S_t, A_t)

$Q(S_t, A_t) \leftarrow$ average (Returns (S_t, A_t))

$\pi(S_t) \leftarrow \text{argmax}_a Q(S_t, a)$

On policy first visit

$A^* \leftarrow \text{argmax}_a Q(S_t, a)$

$\forall a \in A(S_t)$

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|A(S_t)| \\ \varepsilon/|A(S_t)| \end{cases}$$

$$q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s) q_\pi(s,a)$$

$$= \frac{\varepsilon}{|A(s)|} \sum q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a)$$

$$\geq \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum \frac{\pi(a|s) - \frac{\varepsilon}{|A(s)|}}{1-\varepsilon} q_\pi(s,a)$$

$$= \frac{\varepsilon}{|A(s)|} \sum q_\pi(s,a) - \frac{\varepsilon}{|A(s)|} \sum q_\pi(s,a) + \sum \pi(a|s) q_\pi(s,a)$$

$$= v_\pi(s)$$

Off policy MC

Need to behave non-optimally to learn but want to act optimally

Two policies
- Target
- behavior

High var low convergence more power

b (behavior), π (target)

$\pi(a|s) > 0 \Rightarrow b(a|s) > 0$

$Pr\{A_t, S_{t+1}, A_{t+1} \dots S_T | S_t, A_{t:T-1} \sim \pi\}$

$= \pi(A_t|S_t) p(S_{t+1}|S_t, A_t) \pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$

$= \prod_{k=t}^{T-1} \pi(A_k|S_k) p(S_{k+1}|S_k, A_k)$

Importance sampling ratio - ~~estimating expected values~~ of one distro from another

$$\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k) p(S_{k+1}|S_k, A_k)}{b(A_k|S_k) p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

$E[G_t | S_t = s] = v_b(s)$ - won't give us π

$E[\rho_{t:T-1} G_t | S_t = s] = v_\pi(s)$

$$V(\tau) = \frac{\sum_{t \in \tau(\tau)} P_{t:T(t)-1} \, G_t}{|\tau(\tau)|}$$

τ given here

weighted version

$$V(\tau) = \frac{\sum_{t \in \tau(\tau)} \rho_{t:T(t)-1} \, G_t}{\sum_{t \in \tau(\tau)} \rho_{t:T(t)-1}}$$

Want

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum W_k}$$

$$V_{n+1} = V_n + \frac{W_n}{C_n}\left[G_n - V_n\right] \qquad (C_n \text{ is sum of weights})$$

$$C_{n+1} = C_n + W_{n+1}$$

MCTS

1.) Selection: Follow the tree policy to a leaf

2.) Expansion: expand the tree by adding a new leaf node to the tree

3.) Simulation: From the selected node or new child run a MC episode with the rollout policy

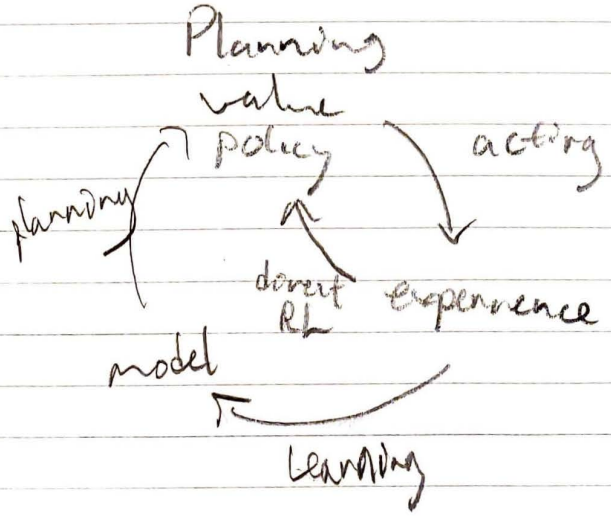
4.) Backup: Return the generated episode and its return

Rollout Algos

Estimate action values by averaging the rewards of many trajectories that start with a given action.

Unlike MC we don't estimate q_* or q_π

Instead they only get values for only the actions associated with the current state.



Tab Dyna Q

Init $Q(s,a)$, $Model(s,a)$ $\forall s \in S, a \in A(s)$

Loop:

$S \leftarrow$ current state

$A \leftarrow \varepsilon$-greedy (S,Q)

Take action A: observe resultant reward R, S'

$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$

$Model(S,A) \leftarrow R, S'$ (assumes deterministic env)

Loop n times:

$S \leftarrow$ random previous state

$A \leftarrow$ random action previously taken in S

$R, S' \leftarrow Model(S,A)$

$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$