Monte Carlo ES is on-policy method

Assume,
Randomised policy setting,

27/oct

With E.S	W/o ES
- Episodic tasks - multiple episodes - Each episode with a random initial state - Greedy wrt current value of $\hat{}$ $\pi \rightarrow V_\pi$ - average over samples. Problem: One typically requires several runs or episodes to be played out before convergence.	- Episodic tasks - multiple episodes - do not randomize over initial states Alternative: Need to select all actions infinitely often On-policy methods: learn the value of the policy that is being played out. off-policy method: Two policies behavior policy — target policy learn value of ^{n} for target policy given the behavior policy.

On policy methods: Work with randomized policies

$$\pi(a|s) > 0 \quad \forall \; a \in A(s)$$

other possiblity: have deterministic policies but pertrub them to make ϵ-greedy policies.

We consider ϵ greedy policies here

& each non-greedy action is selected w.p. $\frac{\epsilon}{|A(s)|}$

Greedy action is selected w.p. $(1-\epsilon) + \frac{\epsilon}{|A(s)|}$

w.p. $\frac{\epsilon}{|A(s)|}$, the greedy action can also be selected by the ϵ-greedy policy.

Where $|A(s)| \equiv$ cardinality of $A(s)$

On-policy first visit MC control for ϵ-soft policies, estimates $\pi \simeq \pi_*$

- Algorithm parameter: small $\epsilon > 0$
- Initialize

$\pi \leftarrow$ an arbitrary ϵ soft policy.

$Q(s,a)$ arbitrarily $\leftarrow \forall \; s \in S, a \in A(s)$

Returns$(s,a) \leftarrow$ empty list $\forall \; s \in S, a \in A(s)$

Repeat (for each episode)

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1 \ldots R_T$

$G \leftarrow 0$

loop (for $t = T-1, T-2 \ldots 0$)

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

Append G to Returns (S_t, A_t)

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

$A^* \leftarrow \arg\max_a Q(S_t, a)$

for all $a \in A(S_t)$

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A(S_t)|} & \text{if } a = A^* \\ \frac{\epsilon}{|A(S_t)|} & \text{if } a \neq A^* \end{cases}$$

Q) Do we get a better ϵ-greedy policy each time?

Suppose q_π is the q-value function estimated for policy π

Let π' be the greedy policy wrt q_π

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

$$= \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \max_a q_\pi(s,a) \rightarrow (*)$$

$$\geq \frac{\epsilon}{|A(s)|} \sum_a q(s,a) + (1-\epsilon) \sum_a \left(\frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{1-\epsilon} \right) q_\pi(s,a) \quad \text{---------- (+)}$$

Note that $\sum_a \left(\pi(a|s) - \frac{\epsilon}{|A(s)|} \right) = \sum_a \pi(a|s) - \frac{\epsilon}{|A(s)|} \sum_a 1$

$$= 1 - \epsilon$$

Thus, $0 \leq \left(\frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{(1-\epsilon)} \right) \triangleq \beta_a \leq 1$

$$\sum_a \frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{(1-\epsilon)} = 1$$

Further, $\max_b q_\pi(s,b) \geq q_\pi(s,a)$

$$\Rightarrow \sum_a \beta_a \left(\max_b q_\pi(s,b) \right) \geq \sum_a \beta_a q_\pi(s,a)$$

Thus $\max_b q_\pi(s,b) \geq \sum_a \beta_a (q_\pi(s,a))$

From (+),

$$\frac{\epsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \left(\sum_a \frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{1-\epsilon} \right) q_\pi(s,a)$$

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

$$\geq V_\pi(s)$$

thus $q_\pi(s, \pi'(s)) \geq v_\pi(s)$

$\Rightarrow$ the policy π' gives an improvement in values i.e., π' is a better policy than π.

Off policy prediction via importance sampling

On-policy methods	Off-policy methods
→ low variance & faster convergence.	→ higher variance & slower convergence (since data is from a policy different from the one being learned)
→ less general methods	→ more general. (On policy is subset of off policy.

Prediction Problem: Assume that both behavior & target policy are fixed.

Problem: & Estimate V_π or q_π both information ~~are~~ on episodes is available for another policy $b \neq \pi$

Assumptions: If policy π is s.t $\pi(a|s) > 0$ for ~~state~~ some $a \in A(s)$ ~~then~~ policy 'b' should also have $b(a|s) > 0$ for that $a \in A(s)$

(Assumption of coverage.)

Importance Sampling:

Given a state S_t, the probability of the subsequent state-action trajectory $A_t, S_{t+1}, A_{t+1}, S_{t+2} \ldots S_T$ occuring under policy π is:

$$p(A_t, S_{t+1}, A_{t+1}, S_{t+2} \ldots S_T \mid S_t, A_{t:T-1} \sim \pi)$$

$$= \pi(A_t \mid S_t)\, p(S_t, A_t, S_{t+1})\, \pi(A_{t+1} \mid S_{t+1})\, p(S_{t+1}, A_{t+1}, S_{t+2}) \ldots$$

$$\ldots \pi(A_{T-1} \mid S_{T-1})\, p(S_{T-1}, A_{T-1}, S_T)$$

$$= p(A_t \mid S_t, S_{t+1}, A_{t+1} \ldots S_T, A_{t:T-1} \sim \pi)\, p(S_{t+1}, A_{t+1} \ldots S_T \mid S_t, A_{t \ldots T-1} \sim \pi)$$

$$\pi(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_{t+1} \ldots S_T, A_{t \ldots T-1} \sim \pi)\, p(A_{t+1} \ldots S_T \mid S_{t+1}, A_{t \ldots T-1} \sim \pi)$$

$$p(S_t, A_t, S_{t+1})\, \pi(A_{t+1} \mid S_{t+1}) \ldots$$

Thus,

$$p(A_t, S_{t+1}, A_{t+1} \ldots S_T \mid S_t, A_{t:T-1} \sim \pi)$$

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

However if we use Behavior policy b instead of π, then,

$$p(A_t, S_{t+1}, S_{t+2} \ldots S_T \mid S_t, A_{t:T-1} \sim b)$$

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

Importance Sampling (I S) ratio:

$$\rho_{t:T-1} = \frac{P(A_t, S_{t+1}, \ldots S_T | S_t, A_{t:T-1} \sim \pi)}{P(A_t S_{t+1}, \ldots S_T | S_t, A_{t:T-1} \sim b)}$$

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

$$= \prod_{k=t}^{T-1} \left(\frac{\pi(A_k|S_k)}{b(A_k|S_k)} \right)$$

* We need to compute expected returns under policy π
* ——— " ——— policy b

$E[G_t | S_t = s] = V_b(s)$ (We want $V_\pi(s)$

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

Suppose $\mathcal{T}(s)$ = set of times or instances when state s is visited

Then $V(s) \triangleq \dfrac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T-1} G_t}{|\mathcal{T}(s)|}$