Off-policy Evaluation

The goal is to evaluate a $\overbrace{\text{target}}^{\text{policy of interest}}$ policy using samples from another policy, called a $\underbrace{\text{behavior policy}}_{\text{policy generating behavior}}$.

$$\boxed{\begin{array}{l} \pi^t : \text{target policy} \\ \pi^b : \text{behavior policy} \end{array}}$$

Benefits

Enables using samples from
- old policies
- exploratory policies
- other agents / demonstrations

Note: Consider a sample finite trajectory from an MDP

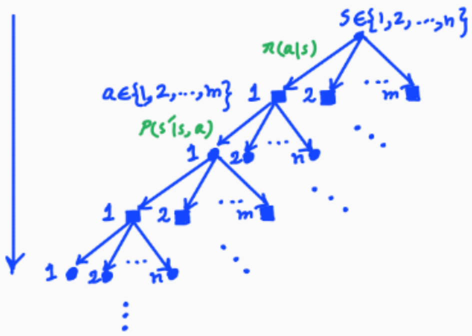
rewards conform to the deterministic reward function

$$r_0 = R(s_0, a_0) \quad r_1 = R(s_1, a_1) \quad r_{T-1} = R(s_{T-1}, a_{T-1})$$

$$\tau_T = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

generated based on an initial distribution μ_0 and following a stationary, stochastic policy π.

T steps



* The probability of τ_T being realized is

$$\mathbb{P}(\tau_T | \mu_0, \pi, P) = \mathbb{P}(\overbrace{s_0, a_0, s_1, a_1, \dots, s_{T-1}, a_{T-1}, s_T}^{\tau_T} | \mu_0, \pi, P)$$

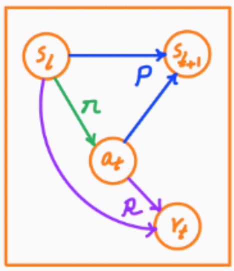
$$= \mathbb{P}(s_T | s_0, a_0, s_1, a_1, \dots, s_{T-1}, a_{T-1}, \mu_0, \pi, P)$$

$P(s_T | s_{T-1}, a_{T-1})$

$$\mathbb{P}(a_{T-1} | s_0, a_0, s_1, a_1, \dots, s_{T-1}, \mu_0, \pi, P)$$

$\pi(a_{T-1} | s_{T-1})$

$$\mathbb{P}(\underbrace{s_0, a_0, s_1, a_1, \dots, s_{T-1}}_{\tau_{T-1}} | \mu_0, \pi, P)$$



$$= \vdots$$

repeat the factorization

$$\underbrace{\tau_T \to \tau_{T-1} \to \tau_{T-2} \to \dots \to \tau_1 \to \tau_0}_{\text{unroll trajectory over time}}$$

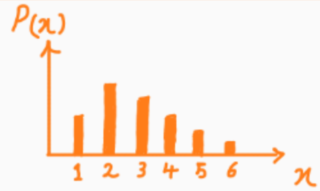
$$= \mu_0(s_0) \prod_{t=0}^{T-1} \pi(a_t | s_t) P(s_{t+1} | s_t, a_t)$$

Importance sampling: A general technique for evaluating properties of a target distribution, e.g., the expectation of a function over that distribution based on samples from another distribution.

Let X be a discrete RV distributed according to probability measure P.

$$\mathbb{E}_{X\sim P}[f(X)] = \sum_{x\in\mathcal{X}} p(x)\, f(x)$$

$$\boxed{\mathbb{E}_{X\sim P}[X] = \sum_{x\in\mathcal{X}} p(x)\, x}$$



$$\downarrow f(x) = x^2$$

$$\mathbb{E}_{X\sim P}[X^2] = \sum_{x=1}^{6} p(x)\, x^2$$

Having n i.i.d. samples $x_1, x_2, \dots, x_n$ of X, drawn from $P(X)$, a simple Monte Carlo method can estimate this expected value by the empirical mean

$$\hat{\mathbb{E}}_{X\sim P}[f(X)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Now, consider another distribution over X captured by probability measure Q.

$$\mathbb{E}_{X\sim P}[f(X)] = \sum_{x\in\mathcal{X}} p(x)\, f(x) = \sum_{x\in\mathcal{X}} Q(x)\,\frac{P(x)}{Q(x)}\, f(x)$$

Likelihood ratio / weight $w(x)$

$$= \mathbb{E}_{X\sim Q}\Big[\underbrace{\frac{P(x)}{Q(x)}\, f(x)}_{g(x)}\Big]$$

requirement:
$Q(x) > 0$ if $p(x) f(x) \neq 0$

Having n i.i.d. samples $x_1, x_2, \dots, x_n$ of X, drawn from $Q(X)$, a simple Monte carlo method can estimate this expected value by the empirical mean

$w(x_i)$ should be known

$$\hat{\mathbb{E}}_{X\sim P}\left[f(X)\right] = \hat{\mathbb{E}}_{X\sim Q}\left[\frac{P(X)}{Q(X)} f(X)\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{P(x_i)}{Q(x_i)} f(x_i)$$

Importance sampling for off-policy evaluation: Let G_T be the sample return corresponding to sample trajectory τ_T generated by π^b.

$$w(\tau_T) = \frac{P(\tau_T \mid \mu_0, \pi^t, P)}{P(\tau_T \mid \mu_0, \pi^b, P)} = \frac{\mu_0(s_0)\prod_{t=0}^{T-1} \pi^t(a_t|s_t)\, P(s_{t+1}|s_t, a_t)}{\mu_0(s_0)\prod_{t=0}^{T-1} \pi^b(a_t|s_t)\, P(s_{t+1}|s_t, a_t)}$$

P is unknown but it cancels out

$$= \prod_{t=0}^{T-1} \frac{\pi^t(a_t|s_t)}{\pi^b(a_t|s_t)}$$

Off-policy MC evaluation:

the return of the trajectory

$$\mathbb{E}_{\pi^t, P}\left[G_T\right] \approx \hat{\mathbb{E}}_{\pi^b, P}\left[\overbrace{\prod_{t=0}^{T-1} \frac{\pi^t(a_t|s_t)}{\pi^b(a_t|s_t)}}^{w} G_T\right] = \frac{1}{n}\sum_{i=1}^{n} w_i G_i$$

$$\hat{V}^{\pi_t}(s) \longleftarrow \hat{V}^{\pi_t}(s) + \alpha\left(\prod_{t=0}^{T-1} \frac{\pi^t(a_t|s_t)}{\pi^b(a_t|s_t)} G_T - \hat{V}^{\pi_t}(s)\right)$$

$t = t_s$ if the trajectory does not start from s

Off-policy TD evaluation:

$$\mathbb{E}_{\pi^t, p}\left[R(S_t, A_t) + \gamma V(S_{t+1}) \mid S_t = s \right]$$

$$= \mathbb{E}_{\substack{A_t \sim \pi^t(s) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}}\left[R(S_t, A_t) + \gamma V(S_{t+1}) \mid S_t = s \right]$$

$$\approx \hat{\mathbb{E}}_{\substack{A_t \sim \pi^b(s) \\ S_{t+1} \sim P(\cdot \mid S_t, A_t)}}\left[\frac{\pi^t(A_t \mid S_t)}{\pi^b(A_t \mid S_t)} \left(R(S_t, A_t) + \gamma V(S_{t+1}) \right) \mid S_t = s \right]$$

$$\hat{V}^{\pi_t}(s) \longleftarrow \hat{V}^{\pi_t}(s) + \alpha \left(\frac{\pi^t(A_t \mid S_t)}{\pi^b(A_t \mid S_t)} \left(R(S_t, A_t) + \gamma \hat{V}^{\pi_t}(S_{t+1}) \right) - \hat{V}^{\pi_t}(S_t) \right)$$

Off-policy Optimization

The goal is to optimize a target policy using samples from another policy, called a behavior policy.

We can use a generalized policy iteration approach, where the policy evaluation step is performed using importance sampling.

off-policy MC method with soft exploration

Iterative method:

- Initialize $\hat{Q}$
- Initialize π^t as a greedy policy with respect to $\hat{Q}$
- Initialize $\mathcal{G}(s,a) = \emptyset$ for all $s \in \mathcal{S}, a \in \mathcal{A}$
- Repeat
    - Build π^b as a soft policy with respect to $\hat{Q}$
    - Generate a sample trajectory τ using π^b

policy evaluation

- For $(s,a) \in \mathcal{S} \times \mathcal{A}$ s.t. $(s,a) \in \tau$:
    - $t_{s,a} =$ first time (s,a) is visited
    - $G = \prod_{t=t_{s,a}}^{T} \frac{\pi^t(a_t|s_t)}{\pi^b(a_t|s_t)} \left(\sum_{t=t_{s,a}}^{T} \gamma^{t-t_{s,a}} R(s_t, a_t) \right)$
    - $\mathcal{G}(s,a) \leftarrow \mathcal{G}(s,a) \cup \{G\}$
    - $N(s,a) \leftarrow N(s,a) + 1$
    - $\hat{Q}(s,a) = \frac{1}{N(s,a)} \sum_{G \in \mathcal{G}(s,a)} G$

policy improvement

- Build π^t as a greedy policy with respect to $\hat{Q}$

off-policy TD method (Q learning)

$\boxed{\text{Q learning does not use importance sampling}}$

Iterative method:

- Initialize $\hat{Q}$ s.t. $\hat{Q}(s,a)=0$ for terminal states
- Initialize π^t as a greedy policy with respect to $\hat{Q}$
- Repeat
  - Sample $s_0 \in S$
  - Repeat until the episode terminates
    - Build π^b as a soft policy with respect to $\hat{Q}$ (e.g., ε-greedy)
    - Take action a according to $\pi^b(a|s)$ at s_0
    - observe reward $R(s,a)$ and next state s'
    - policy evaluation and improvement: $\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha\left[R(s,a) + \gamma \max_{a' \in A} \hat{Q}(s',a') - \hat{Q}(s,a)\right]$
    - policy improvement: Build π^t as a greedy policy with respect to $\hat{Q}$
    - Update the current state