IIIT Hyderabad **Scribed By: Aditya Raj**
Topics in Machine Learning (CS975) **Lecture #18**
Instructor: Dr. Naresh Manwani **October 17, 2017**

Outline. *Off-Policy Monte Carlo control with ordinary importance sampling, weighted importance sampling and discounting aware importance sampling techniques.*

1 Off-Policy Monte Carlo Control

Off-policy methods allow you to assess your strategy observing other players. Formally we use another policy μ to generate data, and estimate v_π or q_π. We call π the target policy and μ the behavior policy. Moreover we do not need the assumption of exploring starts. The behavior policy μ needs to satisfy the condition

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

Every action which is taken under policy π must have a non-zero probability to be taken as well under policy μ. This is called the assumption of coverage. Typically the target policy π would be a greedy (deterministic) policy with respect to the current action-value function.

Given a state S_t, the probability of a subsequent state-action trajectory A_t, S_{t+1}, A_{t+1}, . . . , S_T occurring under policy π is

$$P_?(S_t, A_t, ..., S_T) = \Pi_{k=t}^{T} \pi(A_k|S_k) p(S_{k+1}|S_k, A_k)$$

where p is the state-transition probability.
The relative probability of the trajectory under the target and behavior policies, or the importance sampling, is

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

The trajectory probabilities p depend on the MDP, which are generally unknown, but cancel each other out. We also observe that,

$$\mathbb{E}_{A_k|S_k}\left[\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}\right] = \sum_{A_k} \mu(A_k|S_k) \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} = \sum_{A_k} \pi(A_k|S_k) = 1 \tag{1}$$

- $\mathcal{T}(s)$ = set of all time steps in which state s is visited.
- For a first-visit method: $\mathcal{T}(s)$ = time of first visits to s.
- $T(t)$ = first time of termination following time t
- G_t = return after t up through $T(t)$.
- $\{G_t\}_{t \in \mathcal{T}(s)}$ = returns corresponding to state s.
- $\{\rho_t^{T(t)}\}_{t \in \mathcal{T}(s)}$ = importance sampling ratios.

2 Ordinary Importance Sampling

One way to estimate $v_\pi(s)$ is to scale the return of each episode by the importance sampling ratio and then average it out over all the episodes. In particular, as per our notation, the estimate of $v_\pi(s)$ using ordinary importance sampling can be expressed as

$$\hat{v}_\pi(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{|\mathcal{T}(s)|} \tag{2}$$

where $\rho_t^{T(t)}$ is the importance sampling and G_t is the return of policy.
Taking Expectation of $\hat{v}_\pi(s)$ with respect to probability distribution μ, we see that

$$\begin{aligned}
\mathbb{E}_\mu\Big[\hat{v}_\pi(s)\Big] &= \mathbb{E}_\mu\Big[\frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{|\mathcal{T}(s)|}\Big] \\
&= \frac{1}{|\mathcal{T}(s)|}\sum_{t\in\mathcal{T}(s)} \mathbb{E}_\mu\Big[\rho_t^{T(t)} G_t\Big]
\end{aligned}$$

Now, evaluating the inner expectation term,

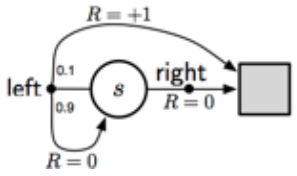
$$\begin{aligned}
\mathbb{E}_\mu\Big[\rho_t^{T(t)} G_t\Big] &= \mathbb{E}_\mu\Big[\Pi_{k=t}^{T(t)-1}\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} G_t\Big] \\
&= \mathbb{E}_\mu\Big[\Pi_{k=t}^{T(t)-1}\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \sum_{l=0}^{T(t)-t-1}\gamma^l R_{t+l+1}\Big] \\
&= \sum_{l=0}^{T(t)-t-1}\gamma^l \mathbb{E}_\mu\Big[\Pi_{k=t}^{T(t)-1}\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} R_{t+l+1}\Big] \\
&= \sum_{l=0}^{T(t)-t-1}\gamma^l \mathbb{E}_\mu\Big[\frac{\pi(A_l|S_l)}{\mu(A_l|S_l)} R_{t+l+1}\Big]\Pi_{k=t,k\neq l}^{T(t)-1}\mathbb{E}_\mu\Big[\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}\Big]
\end{aligned}$$

Now, using equation (1) we see that

$$\begin{aligned}
\mathbb{E}_\mu\Big[\rho_t^{T(t)} G_t\Big] &= \sum_{l=0}^{T(t)-t-1}\gamma^l \mathbb{E}_\pi\Big[R_{t+l+1}\Big] \\
&= \mathbb{E}_\pi\Big[\sum_{l=0}^{T(t)-t-1}\gamma^l R_{t+l+1}|S_t = s\Big] \\
&= \mathbb{E}_\pi\Big[G_t|S_t = s\Big] \\
&= v_\pi(s)
\end{aligned}$$

Now, putting the value back in original equation, we get

$$\begin{aligned}
\mathbb{E}\Big[\hat{v}_\pi(s)\Big] &= \frac{1}{|\mathcal{T}(s)|}\sum_{t\in\mathcal{T}(s)} v_\pi(s) \\
&= \frac{1}{|\mathcal{T}(s)|}\Big[|\mathcal{T}(s)|v_\pi(s)\Big] \\
&= v_\pi(s)
\end{aligned}$$



$$\pi(\text{left}|s) = 1$$

$$\mu(\text{left}|s) = \frac{1}{2}$$

Figure 1: Example

Thus, the average using ordinary importance sampling is unbiased estimator of $v_\pi(s)$. However, the variance of ratios can be unbounded in general. This makes ordinary importance sampling unbounded. This is one of the problems of ordinary importance sampling.

Consider the above example with trajectory $s, left, 0, s, left, 0, s, left, 0, s, left, +1$
We know that

$$Var(X) = \mathbb{E}[X^2] - [\mathbb{E}[X]]^2$$

$\mathbb{E}[X^2]$ can make the variance unbounded, since $\mathbb{E}[X]$ is bounded.

$$\begin{aligned}
\mathbb{E}_\mu[((\hat{v}_\pi(s))^2] &= \mathbb{E}_\mu\Big[\Big(\Pi_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}G_0\Big)^2\Big] \\
&= \underbrace{\frac{1}{2}\times 0.1\times\Big(\frac{1}{0.5}\Big)^2}_{\text{length 1 trajectory}} + \underbrace{\frac{1}{2}\times 0.9\times\frac{1}{2}\times 0.1\times\Big(\frac{1}{0.5\times 0.5}\Big)^2}_{\text{length 2 trajectory}} + ... \\
&= 0.1\sum_{k=0}^{\infty}(0.9)^k\times 2^k\times 2 \\
&= \infty
\end{aligned}$$

Thus, we see that the variance is not bounded using ordinary importance sampling.

3 Weighted Importance Sampling

Using the weighted importance sampling, the state value can be estimated as follows

$$\hat{v}_\pi(s) = \frac{\sum_{t\in\mathcal{T}(s)}\rho_t^{T(t)}G_t}{\sum_{t\in\mathcal{T}(s)}\rho_t^{T(t)}} \tag{3}$$

The above term is zero if the denominator is zero. In this case the variance never becomes unbounded like ordinary importance sampling. However, this is a biased estimator of $\hat{v}_\pi(s)$ as we can see that

$$\mathbb{E}_\mu\Big[\hat{v}_\pi(s)\Big] = \Big[\frac{\sum_{t\in\mathcal{T}(s)}\rho_t^{T(t)}G_t}{\sum_{t\in\mathcal{T}(s)}\rho_t^{T(t)}}\Big] = v_\mu(s)$$

Algorithm 1 Off-Policy MC prediction, for estimating $Q \approx q_\pi$

Input:
an arbitrary target policy π
Initialize:
for all $s \in S, a \in A(s)$ **do**
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
Repeat Forever:
$b \leftarrow$any policy with coverage of π
Generate an episode using $b : S_0, A_0, R_1, ...S_{T-1}, A_{T-1}, R_T, S_T$
$G \leftarrow 0$
$W \leftarrow 1$
for $t = T-1, T-2, ...0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}\Big[G - Q(S_t, A_t)\Big]$
    $W \leftarrow W\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$
    if $W = 0$ **then**
        ExitForLoop

Algorithm 2 Off-Policy MC control, for estimating $\pi \approx \pi_*$

Initialize:
for all $s \in S, a \in A(s)$ **do**
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow argmax_a Q(S_t, a)$ (with ties broken consistently)
Repeat Forever:
$b \leftarrow$any soft policy
Generate an episode using $b : S_0, A_0, R_1, ...S_{T-1}, A_{T-1}, R_T, S_T$
$G \leftarrow 0$
$W \leftarrow 1$
for $t = T-1, T-2, ...0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}\Big[G - Q(S_t, A_t)\Big]$
    $\pi(S_t) \leftarrow argmax_a Q(S_t, a)$ (with ties broken consistently)
    if $A_t \neq \pi(S_t)$ **then**
        ExitForLoop
    $W \leftarrow W\frac{1}{b(A_t|S_t)}$

4 Incremental Implementation: Off-Policy Every Visit Monte Carlo Evaluation and Control

Monte Carlo methods can be implemented incrementally, on an episode by episode basis. We use weighted average in which each return G_n is weighted by W_n. We want to compute

$$\hat{v}_{n+1}(s) = \frac{\sum_{k=1}^{n+1} w_k G_k}{\sum_{k=1}^{n+1} w_k}$$

Incremental update equation becomes

$$\begin{aligned}
\hat{v}_{n+1}(s) &= \frac{\sum_{k=1}^{n+1} w_k G_k}{\sum_{k=1}^{n+1} w_k} \\
&= \frac{\sum_{k=1}^{n} w_k G_k + w_{n+1} G_{n+1}}{\sum_{k=1}^{n+1}} \\
&= \frac{\hat{v}_n(s) \sum_{k=1}^{n} w_k + w_{n+1} G_{n+1}}{\sum_{k=1}^{n+1} w_k} \\
&= \frac{\hat{v}_n(s) \sum_{k=1}^{n+1} w_k - \hat{v}_n(s) w_{n+1} + w_{n+1} G_{n+1}}{\sum_{k=1}^{n+1} w_k} \\
&= \hat{v}_n(s) + \frac{w_{n+1}(G_{n+1} - \hat{v}_n(s))}{\sum_{k=1}^{n+1} w_k} \\
\hat{v}_{n+1}(s) &= \hat{v}_n(s) + \frac{w_{n+1}(G_{n+1} - \hat{v}_n(s))}{C_{n+1}}
\end{aligned}$$

where $C_{n+1} = \sum_{k=1}^{n+1} w_k$

5 Discounted Aware Importance Sampling

So far we have weighted returns without taking into account that they are discounted sum. This may not be the best thing to do. For example, if $\gamma = 0$, then G_0 will be weighted by

$$\rho_0^{T-1} = \Pi_{k=0}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

but we it need only be weighted by

$$\rho_0 = \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

The other factors

$$\Pi_1^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

are irrelevant as after the first reward the return has been found.
Define, Flat partial return

$$\bar{G}_t = \sum_{k=1}^{h} R_{t+k}$$

where $0 \leq t < h \leq T$.
Now let γ be the probability of non termination at any state. We thus calculate return at the state as

$$\begin{aligned} G_t &= (1-\gamma)R_{t+1} \\ &\quad + (1-\gamma)\gamma(R_{t+1} + R_{t+2}) \\ &\quad + (1-\gamma)\gamma^2(R_{t+1} + R_{t+2} + R_{t+3}) \\ &\quad + ... \\ &\quad + ... \\ &\quad + (1-\gamma)\gamma^{T-t-1}(R_{t+1} + ...R_T) \\ G_t &= (1-\gamma)\sum_{h=t+1}^{T-1}\left[\gamma^{h-t-1}\bar{G}_t^h + \gamma^{T-t-1}\bar{G}_t^T\right] \end{aligned}$$

Now putting this G_t value in equation (2), we find the discounted ordinary importance sampling ratio. We get

$$\hat{v}(s) = \frac{\sum_{t=\mathcal{T}(s)}\left[(1-\gamma)\sum_{h=t+1}^{T(t)-1}\gamma^{h-t-1}\rho_t^h\bar{G}_t^h + \gamma^{T(t)-t-1}\rho_t^{T(t)}\bar{G}_t^{T(t)}\right]}{|\mathcal{T}(s)|} \tag{4}$$

where

$$\rho_t^h = \Pi_{k=t}^h \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

References

[1] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT press* 1998.