Off-Policy Learning using Importance Sampling

Sanjay Thakur
260722338
MSc CS, McGill University

What is off-policy learning

- Learning the action values from a policy using samples drawn from some other policy is called off-policy learning.
- **Target Policy** is the policy being learned about (denoted as Π).
- **Behavior Policy** is the policy to generate behavior (denoted as μ).

Why do we need off-policy learning

- The on policy approach seek to learn action values conditional on subsequent optimal behavior.
- But they behave non-optimally to explore all actions.
- Hence, it learns action values not for the optimal policy, but for a near-optimal policy that still explores.

Off-policy learning using importance sampling

- Importance sampling is a general technique for estimating expected values under one distribution given samples from another.

- We multiply the return received from a trajectory with weights, where weight is defined as the ratio of probabilities of that trajectory being selected by the target policy to that of the behavior policy.

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

Source: Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto

Two ways to do off-policy learning using importance sampling

Ordinary Importance Sampling

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)} G_t}{|\mathcal{T}(s)|}$$

- Less Bias
- High Variance

Weighted Importance Sampling

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)}}$$

- Less Variance
- High Bias

Source: Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto

Demonstration

- Solve a prediction problem with fixed target and behavior policy.

- We'll estimate state action values of the target policy, i.e. $Q(s,a)_{\Pi}$ using the episodes drawn from the behavior policy.

Demonstration

Environment taken is 4 X 4 GridWorld

Start State

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

End State

Demonstration(1)

Target Policy	Behavior Policy
ACTION UP :: 0.1	ACTION UP :: 0.25
ACTION RIGHT :: 0.4	ACTION UP :: 0.25
ACTION DOWN :: 0.4	ACTION UP :: 0.25
ACTION LEFT :: 0.1	ACTION UP :: 0.25

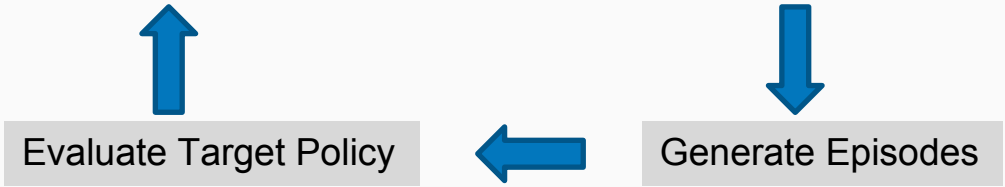
Demonstration(1)

Target Policy	Behavior Policy
ACTION UP :: 0.1	ACTION UP :: 0.25
ACTION RIGHT :: 0.4	ACTION UP :: 0.25
ACTION DOWN :: 0.4	ACTION UP :: 0.25
ACTION LEFT :: 0.1	ACTION UP :: 0.25

Generate Episodes

Demonstration(1)

Target Policy	Behavior Policy
ACTION UP :: 0.1	ACTION UP :: 0.25
ACTION RIGHT :: 0.4	ACTION UP :: 0.25
ACTION DOWN :: 0.4	ACTION UP :: 0.25
ACTION LEFT :: 0.1	ACTION UP :: 0.25



Demonstration(2)

DEMO TIME

Additional Uses

- They can often be applied to learn from data generated by a conventional non-learning controller, or from a human expert.
- Off-policy learning is also seen by some as key to learning multi-step predictive models of the world's dynamics (Sutton, 2009, Sutton et al., 2011).

Questions