Off-policy Learning For Prediction

Objectives

- Understand how off-policy learning can help deal with the exploration problem
- Understand importance sampling
- Use importance sampling to estimate the expected value of a target distribution using samples from a different distribution.

FPT Education

FPT UNIVERSITY

Off-Policy Learning

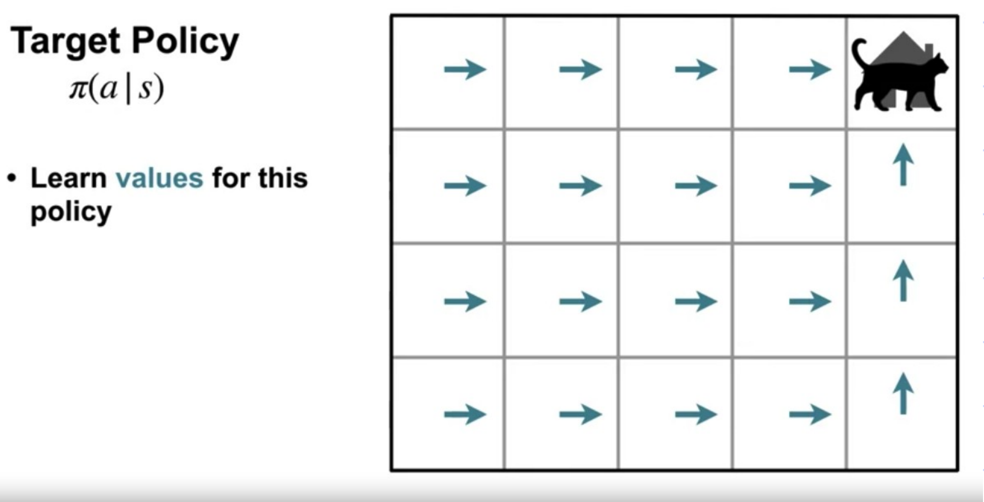
- What is it ?
  - Off-policy learning is a reinforcement learning technique where the agent learns to estimate value functions, state-value function V(s) or action-value function Q(s,a)) from data generated by following a different behavior policy than the one being evaluated.
  - In off-policy learning, the behavior policy determines the agent's actions, while the target policy is the one whose value function the agent aims to estimate.

Off-Policy Learning

- On-Policy: improve and valuate the policy being used to select actions
- Off-policy: improve and evaluate a different policy from the one used to select actions
- Example:
  - Learning the optimal policy involves following a completely random policy, termed the target policy, as it serves as the objective for the agent's learning process

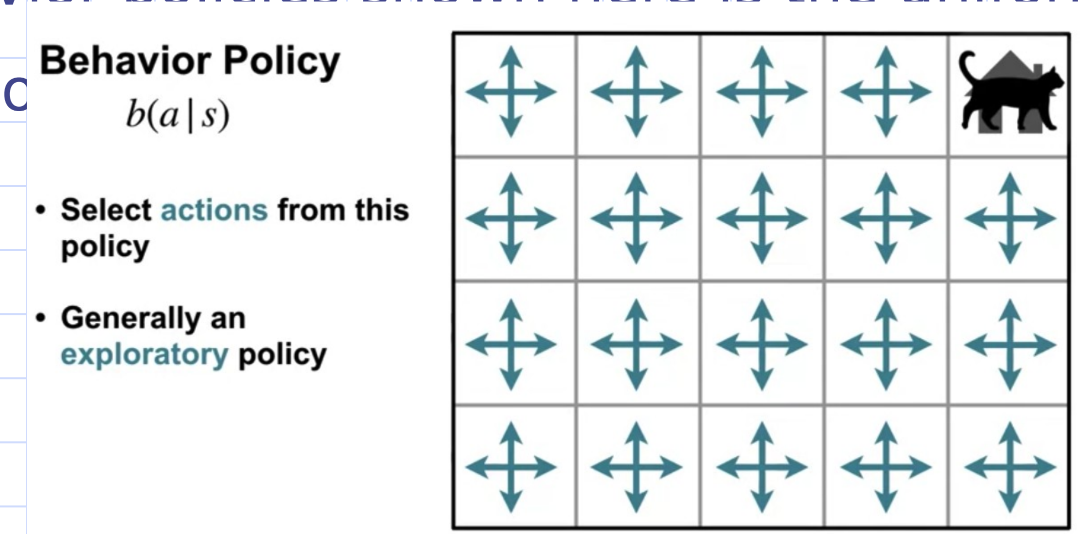
Off-Policy Learning

- The value function that the agent is learning is based on the target policy. One example of a **Target policy** is the optimal policy we call the policy that the agent is using to select actions the behavior policy because it defines our agents E



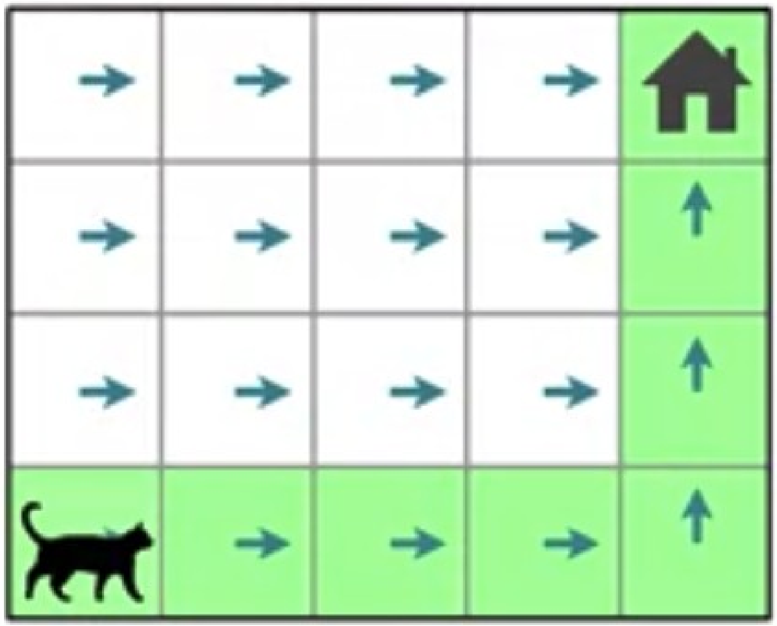
Off-Policy Learning

- The behavior policy is usually denoted by B.
- The behavior policy is in charge of selecting actions for the agent.
- The behavior policies shown here is the uniform random po



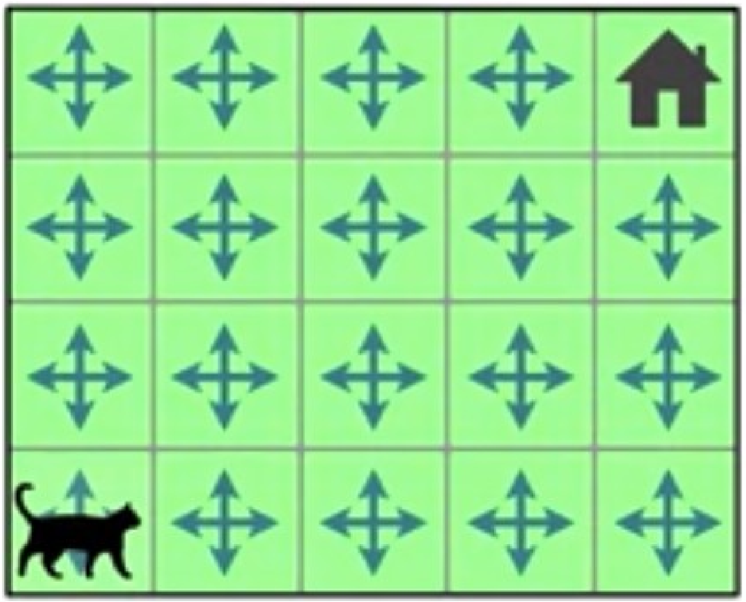
Off-Policy Learning

- If our agent behaves according to the Target policy it might only experience a small number of states.



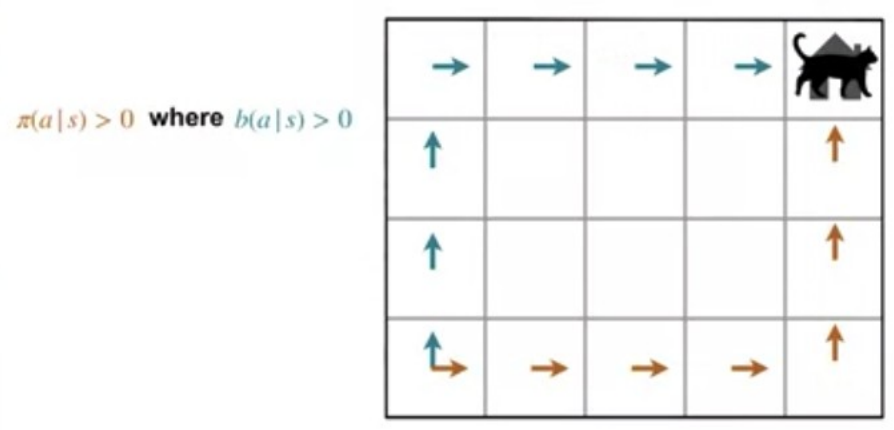
Off-Policy Learning

- If our aging can behave according to a policy that favors exploration.
- It can experience a much larger number of states.



Off-Policy Learning

- One key rule of off policy learning is that the behavior policy must cover the target policy.
  - If the target policy says the probability of selecting an action a given State s is greater than zero then the behavior policy must say the probability of selecting that action

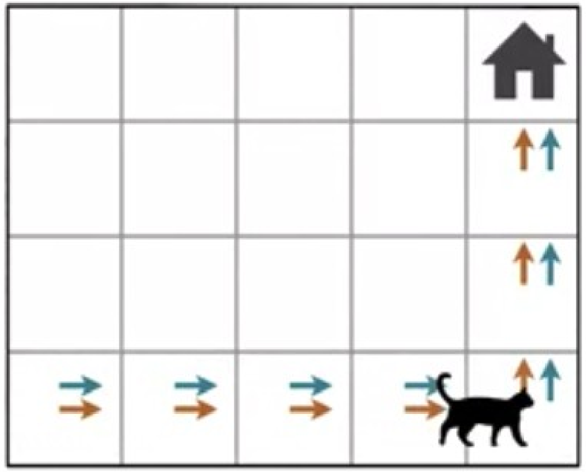


Off-Policy Learning

- It's worth noting that off policy learning is a strict generalization of on policy learning on policies the specific case where the target policy is equal to the behavior policy.

$\pi(a \mid s) > 0$ **where** $b(a \mid s) > 0$

On-Policy: $\pi(a \mid s) = b(a \mid s)$



Importance Sampling

- We have some random variable x that's being sampled from a probability distribution b.
- We want to estimate the expected value of x but with respect to the target distribution Pi.
- Because x is drawn from b, we cannot simply use the sample average to compute the expectation under Pi.
- This sample average will gi value under b instead.

Sample: $x \sim b$

Estimate: $\mathbb{E}_{\pi}[X]$

Importance Sampling

- Importance sampling ratio

$$\mathbb{E}_\pi[X] \doteq \sum_{x \in X} x\pi(x)$$
$$= \sum_{x \in X} x\pi(x)\frac{b(x)}{b(x)}$$
$$= \sum_{x \in X} x\boxed{\frac{\pi(x)}{b(x)}}b(x)$$

Importance sampling ratio

- We can write the importance sampling ratio as Rho of x.

$$\sum_{x \in X} x\rho(x)b(x)$$

Importance Sampling

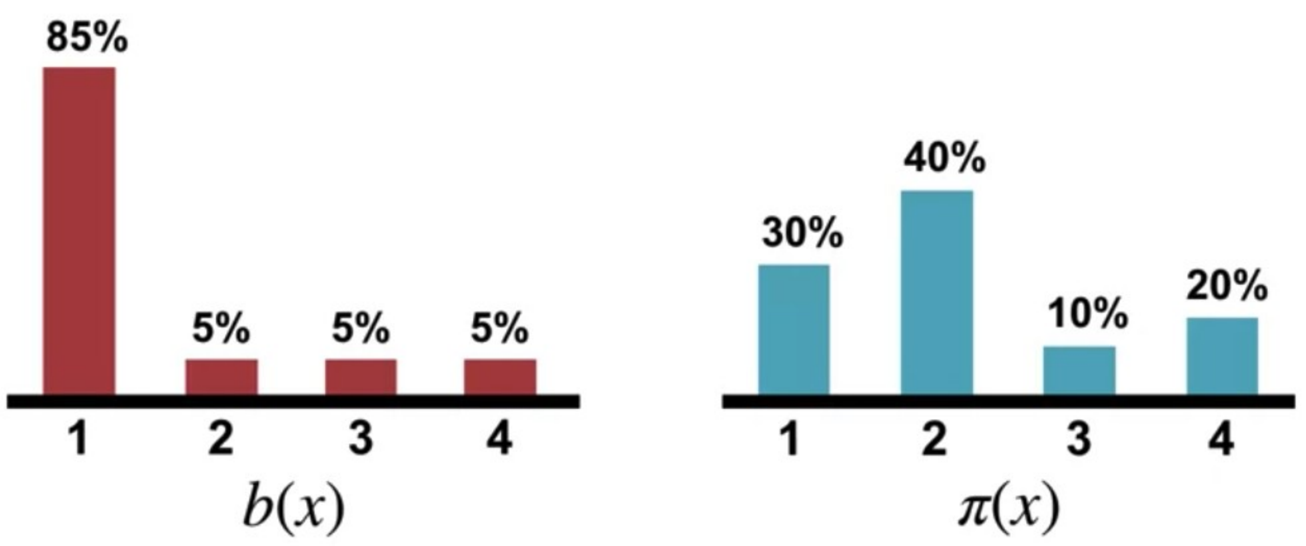
- How do we use it to estimate the expectation from data?
  - We just need to compute a weighted sample average with the importance sampling ratio as the weightings
  - We can now estimate the expected value of x under distributio erage, the samples drawn fro

$$\mathbb{E}_{\pi}[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i \rho(x_i)$$

$$x_i \sim b$$

Importance Sampling

- We have two rather different distributions: b and Pi. We'll draw samples according to b and try to estimate the expected value under Pi.

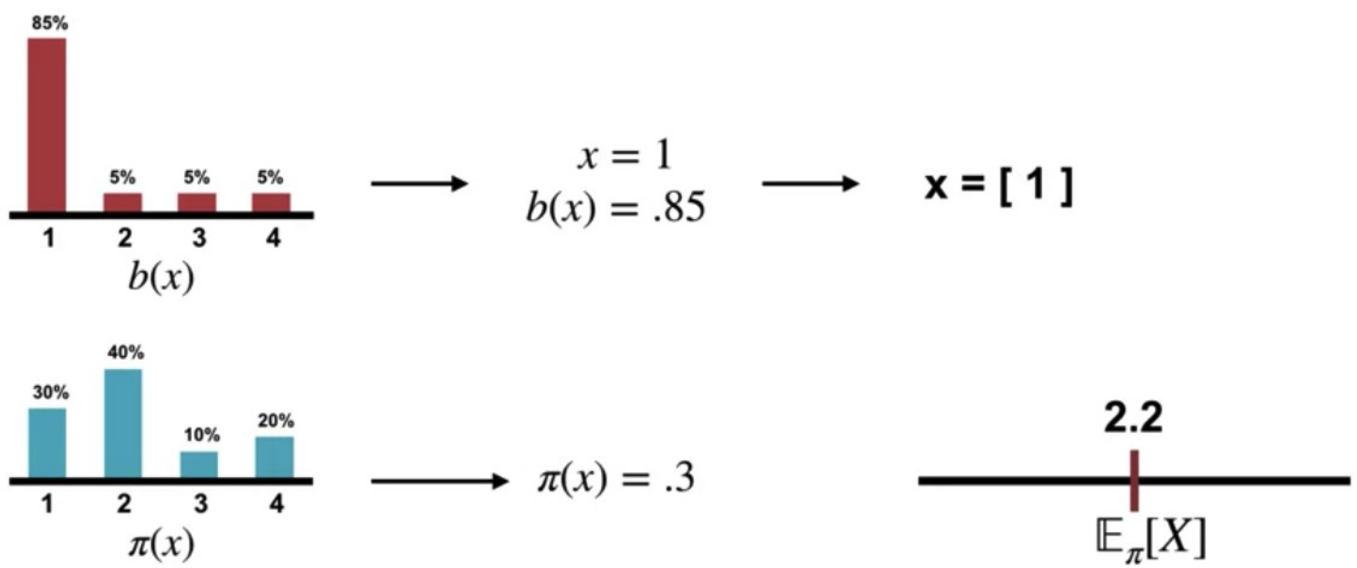


Importance Sampling

- Draw samples according to b and try to estimate the expected value under Pi.
  - On the right: track of our current estimate for the expected value under Pi.
  - For reference, The true expected value is in the middle of the

Importance Sampling

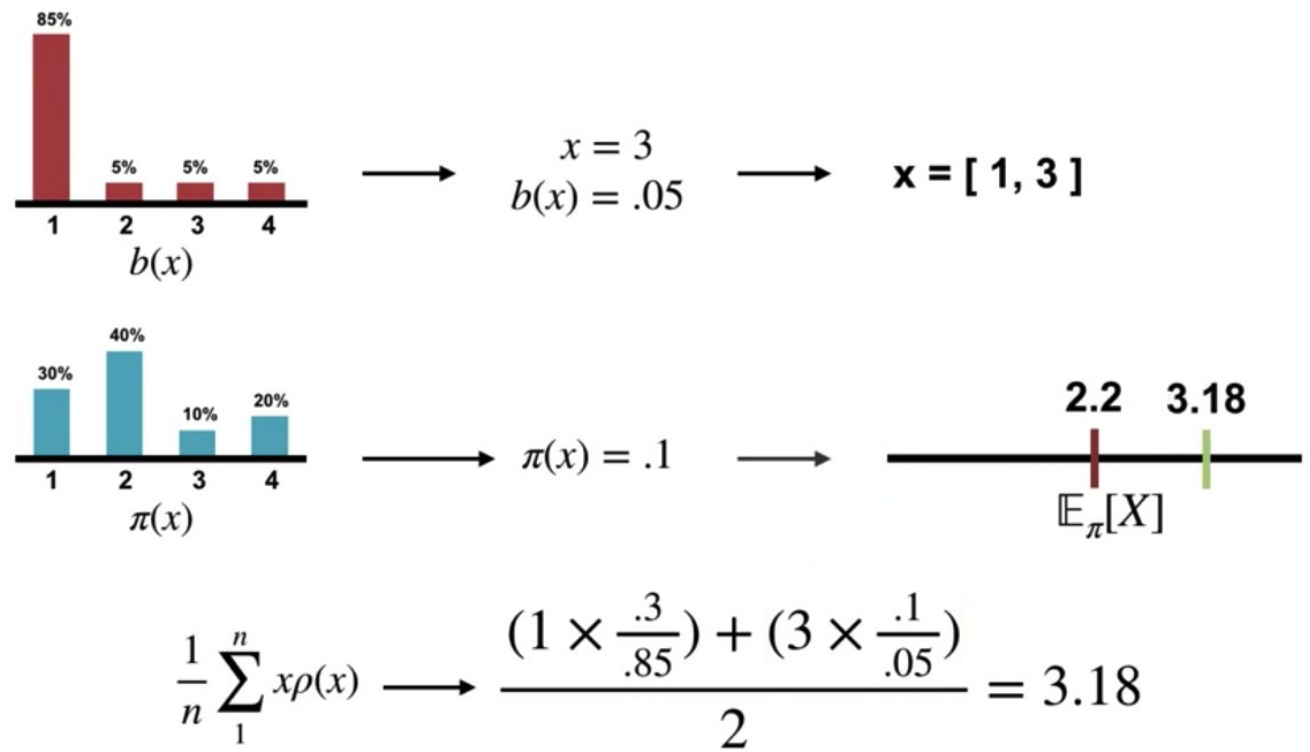
- Sample from b(x=1), we get estimate of 0.35



$$\frac{1}{n}\sum_{1}^{n} x\rho(x) \longrightarrow 1 \times \frac{.3}{.85} = 0.35$$

Importance Sampling

- Sample from b(x=3), we get estimate of 3.18



Summary

- Understand how off-policy learning can help deal with the exploration problem
- Understand importance sampling
- Use importance sampling to estimate the expected value of a target distribution using samples from a different distribution.

FPT UNIVERSITY

Q & A