Visit Distribution Corrections

A lower-variance approach to off-policy learning

Eric Graves

Tea Time Talk, August 19, 2019

Motivation

What is off-policy learning?

- Learning about a policy without following it exactly.

Motivation

What is off-policy learning?

- Learning about a policy without following it exactly.

Why is it interesting?

- can learn an optimal policy from suboptimal data.
- can improve sample efficiency.
- can learn offline when safety is critical.

Motivation

What is off-policy learning?

- Learning about a policy without following it exactly.

Why is it interesting?

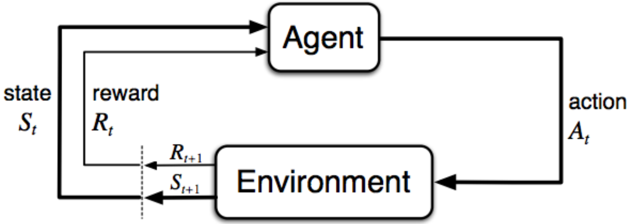
- can learn an optimal policy from suboptimal data.
- can improve sample efficiency.
- can learn offline when safety is critical.

But doesn't it have crazy variance problems or something?

- Importance sampling on **policies** has variance issues.
- Importance sampling on **visit distributions** doesn't!

Outline

The Agent-Environment Interface



On each time step t, the agent receives the environment's current state S_t and uses policy π to select an action $A_t \sim \pi(\cdot \mid S_t)$. On the next time step, the agent receives a reward R_{t+1} and observes the environment's new state S_{t+1}.

Trajectories, Returns, and Values

- The sequence of states, actions, and rewards forms a trajectory $\tau = S_0, A_0, R_1, S_1, A_1, R_2, \ldots.$
- The (possibly discounted) sum of rewards from time $\boldsymbol{t}$ is called the return:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- The value of a state $\boldsymbol{s}$ under policy $\boldsymbol{\pi}$ is the expected return when starting in $\boldsymbol{s}$ and following $\boldsymbol{\pi}$ thereafter:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right], \forall s \in \mathcal{S}$$

The Goal of Off-Policy Learning

Prediction: learn the value function for fixed target policy $\boldsymbol{\pi}$ while following fixed behaviour policy **b**.

Control: learn $\boldsymbol{\pi}$ itself while following **b**.

- Following $\boldsymbol{b}$ gives: $v_b(s) = \mathbb{E}_b[G_t \mid S_t = s]$.
- However, we want: $v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$.
- To learn the value function for $\boldsymbol{\pi}$ while following $\boldsymbol{b}$, we need to correct for the discrepancy between the policies.

Importance Sampling

- Consider the bandit case where there is only one state.
- We want to know what the expected reward would be under $\boldsymbol{\pi}$, but we only have samples from $\boldsymbol{b}$.
- We can correct for the discrepancy in policies like so:

$$\mathbb{E}_\pi[r] = \sum_{a \in \mathcal{A}} \pi(a) r = \sum_{a \in \mathcal{A}} \frac{\pi(a)}{b(a)} b(a) r = \mathbb{E}_b \left[\frac{\pi(a)}{b(a)} r \right]$$

- We often refer to $\frac{\pi(a)}{b(a)}$ as $\boldsymbol{\rho}$.

Importance Sampling on Policies

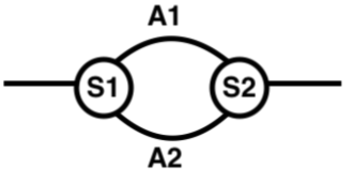
- A straightforward extension to correct returns:

$$
\begin{aligned}
&\mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_b\left[\frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_{t+1} + \frac{\pi(A_t|S_t)}{b(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{b(A_{t+1}|S_{t+1})} \gamma R_{t+2} + \ldots \mid S_t = s\right] \\
&= \mathbb{E}_b\left[\sum_{k=0}^{T-1} \left(\prod_{j=0}^{k} \frac{\pi(A_{t+j}|S_{t+j})}{b(A_{t+j}|S_{t+j})}\right) \gamma^k R_{t+k+1}\right]
\end{aligned}
$$

- Using importance sampling in this way can suffer from exponentially high variance.

An Intuitive Example

- To see why, consider the following example:



- Both actions **A1** and **A2** lead to the same next state **S2**.
- Therefore the probability of visiting **S2** is the same under both policies, and the reward does not need to be corrected.

Visit Distributions

- The value of a policy can be alternatively expressed as:

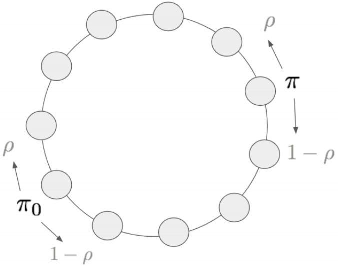
$$\mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)]$$

- Then importance sampling can be done on the state-action visit distribution $d_\pi(s,a)$:

$$\mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d_b}\left[\frac{d_\pi(s,a)}{d_b(s,a)} r(s,a)\right]$$

Another Intuitive Example

- Consider the following example:



- However, the two policies are symmetric, and have identical stationary state distributions.
- Therefore we only need to correct using the stationary state-action densities induced by each policy.

Key Takeaways

1. We can use importance sampling on visit distributions instead of on policies themselves to achieve lower-variance off-policy learning.
2. More broadly, it's always beneficial to think carefully about whether a given issue we're facing is a property of the problem we're trying to solve, or a property of our chosen solution method.