Notes on Importance Sampling and Policy Learning in RL

Reinforcement Learning Concepts

October 2025

1 On-Policy vs. Off-Policy Learning (The Context)

Reinforcement Learning methods are fundamentally categorized by the relationship between the policy used to gather data and the policy being learned.

1.1 On-Policy Learning

- **Definition:** The **target policy** (π) is the same as the **behavior policy** (μ). $\boldsymbol{\pi = \mu}$.
- **Mechanism:** The agent evaluates and improves the policy it is currently using for exploration (e.g., an ϵ-greedy policy). The value function learned reflects the value of this exploratory policy.
- **Key Algorithm: Sarsa** (State-Action-Reward-State-Action).

1.2 Off-Policy Learning

- **Definition:** The **target policy** (π) (often the optimal, greedy one) is different from the **behavior policy** (μ) (the exploratory one). $\boldsymbol{\pi \neq \mu}$.
- **Mechanism:** The agent gathers data from an exploratory policy (μ) but uses that data to learn the value of a potentially different, more optimal policy (π).
- **Key Algorithms: Q-Learning** and **Off-Policy Monte Carlo (using Importance Sampling)**.

2 Importance Sampling (IS) in Off-Policy Monte Carlo

Importance Sampling (IS) is a statistical technique necessary to make **Off-Policy Monte Carlo** estimation possible. It corrects the statistical bias that arises when using samples generated by one policy (μ) to estimate the expected returns of another policy (π).

2.1 The Importance Sampling Ratio ($\boldsymbol{\rho}$)

The core of the technique is the $\boldsymbol{\rho}$ ratio, a scaling factor applied to the observed returns (G_t). It quantifies the relative probability of observing the specific sequence of actions under the target policy compared to the behavior policy.

The ratio for a trajectory segment starting at time t and ending at $T-1$ is:

$$\rho_{t:T-1} = \frac{P(\text{Observed Path } |\pi)}{P(\text{Observed Path } |\mu)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

Coverage Condition: For the estimate to be defined, the behavior policy μ must have a non-zero probability of taking any action that the target policy π might take. If $\mu(A|S) = 0$ and $\pi(A|S) > 0$, the ratio is undefined.

2.2 Importance Sampling Estimators

When estimating the state value $V(S_t)$ by averaging N observed returns, two estimators are common:

4.A Ordinary Importance Sampling (IS)

$$V(S_t) \approx \frac{1}{N} \sum_{i=1}^{N} \rho_{t:T-1}^{(i)} G_t^{(i)} \tag{1}$$

Property: Unbiased (correct in expectation). **Drawback:** High variance, especially for long episodes, due to the unbounded product of ratios.

4.A Weighted Importance Sampling (WIS)

$$V(S_t) \approx \frac{\sum_{i=1}^{N} \rho_{t:T-1}^{(i)} G_t^{(i)}}{\sum_{i=1}^{N} \rho_{t:T-1}^{(i)}} \tag{2}$$

Property: Biased, but consistent (converges to the true value as $N \to \infty$). **Advantage:** Significantly **lower variance** than Ordinary IS, making it the preferred choice for off-policy MC.

3 Why Q-Learning is Preferred (TD Advantage)

Temporal Difference (TD) control algorithms, most notably **Q-Learning**, are the dominant off-policy methods in practice because they inherently avoid the need for Importance Sampling:

- **Q-Learning Solution:** By using a one-step backup (bootstrapping) and targeting the maximum next action-value ($Q(S_{t+1}, a)$), Q-Learning learns the optimal policy's value without referencing the actual next action taken by the exploratory policy μ.
- **Stability:** Since TD methods rely only on the single reward R_{t+1} and the estimated value of the next state $V(S_{t+1})$, they maintain much lower variance than any MC method that uses the full, weighted return $\rho_{t:T-1}G_t$.