Off-policy Monte Carlo

1 Introduction

In off-policy control, improvement and evaluation are done on a different policy from the one used to select actions. We have:

- target policy π, which is the policy being evaluated and improved;
- behavior policy b, used to select actions.

2 Importance sampling

Let x be a random variable sampled from the behavior policy b. We can estimate the expected value of x with respect to the target distribution π, as follows:

$$\begin{aligned}\mathbb{E}_{X\sim\pi}\left[X\right] &= \sum x\pi\left(x\right)\\ &= \sum x\frac{\pi\left(x\right)}{b\left(x\right)}b\left(x\right)\\ &= \sum x\rho\left(x\right)b\left(x\right),\end{aligned}$$

where $\rho\left(x\right)=\frac{\pi(s)}{b(s)}$ is called the importance sampling ratio. Let $x\rho\left(x\right)$ be a new random variable $X_\rho\left(X\right)$. Then:

$$\begin{aligned}\mathbb{E}_{X\sim\pi}\left[X\right] &= \sum x\rho\left(x\right)b\left(x\right)\\ &= \mathbb{E}_{X\sim b}\left[X_\rho\left(X\right)\right]\\ &\approx \frac{1}{n}\sum_{i=1}^{n}x_i\rho\left(x\right),\end{aligned}$$

where $x_i\sim b$.

3 Off-policy Monte Carlo prediction

In Monte Carlo prediction we estimate the value of each state by computing a sample average over returns starting from that state:

$$V_\pi\left(s\right)=\mathbb{E}_\pi\left[G_t|S_t=s\right].$$

In off-policy Monte Carlo we to estimate the value under the target policy π using returns obtained while following the behavior policy b, thus:

$$V_\pi(s) = \mathbb{E}_b\left[\rho G_t | S_t = s\right],$$

where:

$$\rho = \frac{\Pr(\text{trajectory under } \pi)}{\Pr(\text{trajectory under } b)}.$$

The probability can be given by $\Pr(A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t, A_{t:T})$. Using the Markov property we can split the terms into:

$$\begin{aligned}
&\Pr(A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t, A_{t:T}) = \\
&\quad = \pi(A_t|S_t)\, p(S_{t+1}|S_t, A_t) \ldots \pi(A_{T-1}|S_{T-1})\, p(S_T|S_{T-1}, A_{T-1}) \\
&\quad = \prod_{k=t}^{T-1} \pi(A_k|S_t)\, p(S_{k+1}|S_k, A_k),
\end{aligned}$$

And similarly for the denominator with b instead of π. We can now rewrite the importance sampling ratio as:

$$\begin{aligned}
\rho &= \frac{\Pr(\text{trajectory under } \pi)}{\Pr(\text{trajectory under } b)} \\
&= \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)\, p(S_{k+1}|S_k, A_k)}{b(A_k|S_k)\, p(S_{k+1}|S_k, A_k)} \\
&= \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}.
\end{aligned}$$

4 Incremental implementation

Let $G_1, G_2, \ldots G_{n-1}$ be a sequence of returns all starting from the same state. Let W_k be their corresponding weights, e.g. the importance sampling weight ρ. Then the weighted average of returns is:

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \quad n \geq 2.$$

We use a weighted average in order to avoid the unbounded variance that comes with ordinary importance sampling. Weighted importance sampling is biased, but the bias converges asymptotically to zero. We also keep track of the cumulative sum of the weights given to the first n returns:

$$\begin{aligned}
C_0 &= 0 \\
C_{n+1} &= C_n + W_{n+1}.
\end{aligned}$$

The update rule for V_{n+1} and C_{n+1} are:

$$\begin{aligned} V_{n+1} &= V_n + \frac{W_n}{C_n}\left[G_n - V_n\right], \quad n \geq 1 \\ C_{n+1} &= C_n + W_{n+1}, \end{aligned}$$

where V_1 is arbitrary and $C_0 = 0$.

If π is greedy, then $\pi\left(A_t|S_t\right) = 1$. So, the importance sampling weight is:

$$\rho = \prod_{k=t}^{T-1} \frac{1}{b\left(A_k|S_k\right)},$$

and W can be updated as follows, for each time step in the trajectory:

$$W_{n+1} = W_n \frac{1}{b\left(A_t|S_t\right)}.$$