AIL 722: Reinforcement Learning

Lecture 18: Monte Carlo Control (Part 2)

Raunak Bhattacharyya

ScAI | YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap

- Estimating Q-Values
- Exploring Starts
- Stochastic exploration policies
    - Can we find expected returns for the target policy using expected returns obtained from a source policy?

Outline

- Importance sampling for prediction
- Weighted IS and incremental algorithm
- Off-policy MC control
- TD control

Importance Sampling

$$\mathbb{E}_p\left[z(x)\right] = \int z(x)p(x)dx$$

$$r = \mathbb{E}_p\left[z(x)\right]$$

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} z(x_i)$$

$$\mathbb{E}_p\left[z(x)\right] = \int z(x)p(x)dx,$$

$$= \int z(x)\frac{p(x)}{q(x)}q(x)dx$$

$$= \mathbb{E}_q\left[z(x)\frac{p(x)}{q(x)}\right]$$

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} z(x_i)\frac{p(x_i)}{q(x_i)}$$

Back to Prediction: Off-Policy

We are trying to estimate the expected return for π

Target policy

We have a policy b

Behavior policy

Assume coverage:

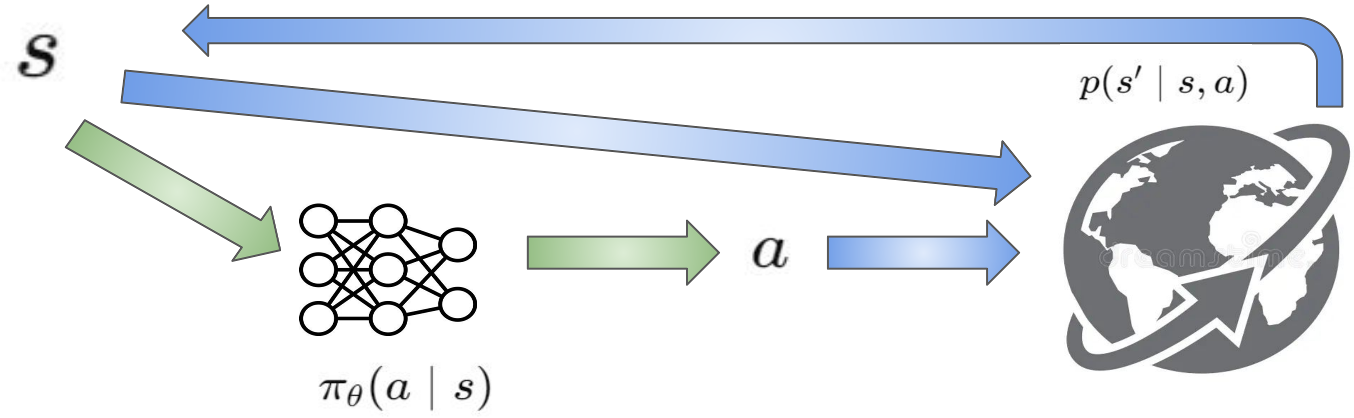
$$\pi(a \mid s) > 0 \implies b(a \mid s) > 0$$

Implies that b must be stochastic in states where it is not identical to pi

Because every action taken under pi has to be taken under b

What is the IS ratio?

RL Objective



$$p_\theta(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t)$$

$p_\theta(\tau)$

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_t r(s_t, a_t)\right]$$

Off-Policy Prediction via Importance Sampling

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T-1} \sim \pi\}$$

$$= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1})$$

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

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

$$= \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

Use for prediction?

Section 5.5, Reinforcement Learning: An Introduction, Sutton & Barto

Expected Returns

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$$

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big], \qquad n \geq 1$$

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}}$$

$$C_{n+1} \doteq C_n + W_{n+1}$$

Can we build an incremental estimation algorithm?

Monte Carlo Prediction using Importance Sampling

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \in \mathbb{R}$ (arbitrarily)
  $C(s,a) \leftarrow 0$

Loop forever (for each episode):
  $b \leftarrow$ any policy with coverage of π
  Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
    $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

Is this first-visit or every-visit?

Section 5.5, Reinforcement Learning: An Introduction, Sutton & Barto

Monte Carlo Control using Importance Sampling

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \in \mathbb{R}$ (arbitrarily)
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(s, a)$ (with ties broken consistently)

Loop forever (for each episode):
  $b \leftarrow$ any soft policy
  Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

Follow behavior policy while learning about and improving the target policy

Section 5.5, Reinforcement Learning: An Introduction, Sutton & Barto

Summary & Announcements

- Summary
  - Importance sampling return estimation
  - Off-policy MC control

- Announcements
  - Sign up for Assgn 1 viva slots
    - To be held this Sat, 7/9/24

Viva sign up link