Monte Carlo Methods

Rupam Mahmood

February 12, 2020

Monte Carlo version of classical policy iteration (with construction of greedy policies)

$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

Here, we use:

Action value estimates

Deterministic policies

Exploring starts

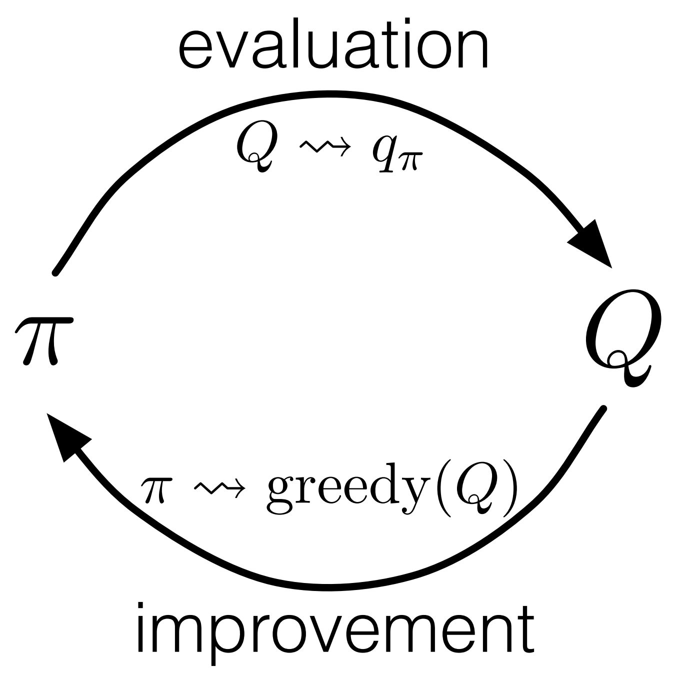
Requiring infinite episodes per iteration

Monte Carlo control with generalized policy iteration removes the requirement of using infinite episodes

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
- $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
- $Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
- $\mathit{Returns}(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
- Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
- Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
    - Append G to $\mathit{Returns}(S_t, A_t)$
    - $Q(S_t, A_t) \leftarrow \text{average}(\mathit{Returns}(S_t, A_t))$
    - $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$



Monte Carlo control without exploring start

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

  $\pi \leftarrow$ an arbitrary ε-soft policy

  $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

  $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

  Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

  $G \leftarrow 0$

  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

    $G \leftarrow \gamma G + R_{t+1}$

    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

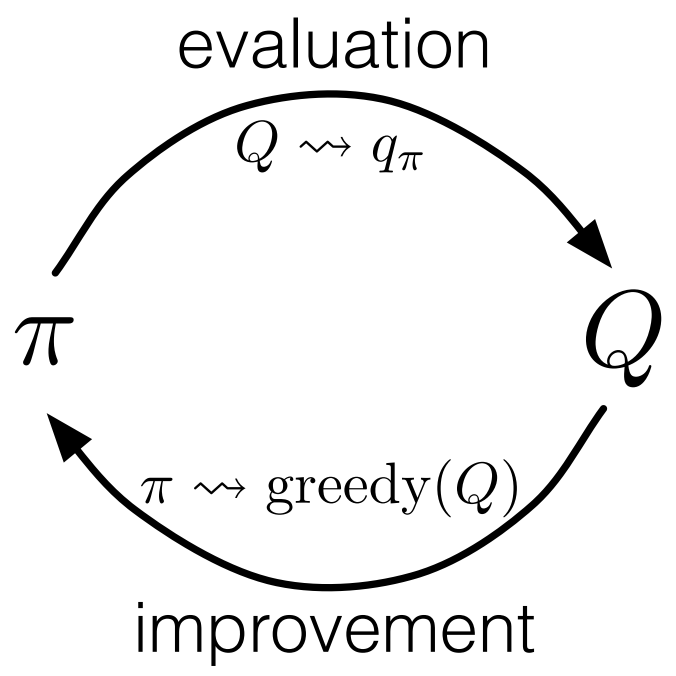
      Append G to $Returns(S_t, A_t)$

      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

      $A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

      For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$



Unbiased and consistent estimation

Say $X_i \sim p$ **is an iid random variable**

The sample average $Z_n = \frac{\sum_{i=1}^{n} X_i}{n}$ **is an estimate of** $E_{X \sim p}[X] = \sum_x xp(x)$

So is X_i

Then we have $E_{X_i \sim p}[Z_n] = E_{X \sim p}[X]$**; unbiasedness of** Z_n

And we have $P\left(\lim_{n \to \infty} Z_n = E_{X \sim p}[X]\right) = 1 \iff Z_n \overset{a.s.}{\to} E_{X \sim p}[X]$**; consistency of** Z_n

On the other hand, we have $E_{X_i \sim p}[X_i] = E_{X \sim p}[X]$**, but not** $X_i \overset{a.s.}{\to} E_{X \sim p}[X]$

When samples are from a different distribution ...

Say $X_i \sim d$ **is an iid random variable (note the difference in distribution)**

Let's call d **the data distribution, and** p **the target distribution**

The sample average $Z_n = \frac{\sum_{i=1}^{n} X_i}{n}$ **is a *bad* estimate of** $E_{X \sim p}[X] = \sum_x xp(x)$

Because now we have $E_{X_i \sim d}[Z_n] = E_{X \sim d}[X] \neq E_{X \sim p}[X]$

And $Z_n \overset{a.s.}{\to} E_{X \sim d}[X] \neq E_{X \sim p}[X]$

When samples are from a different distribution …

Obviously, $X_i \sim d$ **is a worse estimate of** $E_{X \sim p}[X]$

How about $Y_i = \frac{p(X_i)}{d(X_i)} X_i$**, where** $X_i \sim d$**?**

If d **provides adequate coverage of** $p : p(x) > 0$ **implies** $d(x) > 0$,

$$E_{X_i \sim d}\left[Y_i\right] = E_{X_i \sim d}\left[\frac{p(X_i)}{d(X_i)} X_i\right] = \sum_x \frac{p(x)}{d(x)} x d(x)$$

$$= \sum_x x p(x) = E_{X \sim p}[X]$$

When samples are from a different distribution, we can use importance sampling correction

$\frac{p(X_i)}{d(X_i)}$ **is known as the importance sampling ratio**

It can be used to correct the discrepancy between target and data distributions

The following importance sampling estimator is an unbiased and consistent estimator of $E_{X \sim p}[X]$

$$Z_n = \frac{\sum_{i=1}^{n} Y_i}{n}, \text{ where } Y_i = \frac{p(X_i)}{d(X_i)} X_i \text{ and } X_i \sim d$$

Importance sampling for off-policy prediction

We want to estimate v_π whereas samples are from a different policy $b \neq \pi$

We call b the behavior policy, and π the target policy

Then the importance sampling ratio for a trajectory corresponding to return G_t is

$$\rho_{t:T-1} \doteq \frac{P(A_t, S_{t+1}, A_{t+1}, \cdots, S_T \mid S_t, A_{t:T-1} \sim \pi)}{P(A_t, S_{t+1}, A_{t+1}, \cdots, S_T \mid S_t, A_{t:T-1} \sim b)} = \frac{\prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k)}$$

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

Importance sampling for off-policy prediction

Sample average estimator for on-policy prediction: $V(s) \doteq \dfrac{\sum_{t \in \mathscr{T}(s)} G_t}{|\mathscr{T}(s)|}$

$\mathscr{T}(s)$ **contains all time steps in which state** s **is visited**

G_t **denotes the return after** t **up through** $T(t)$

$T(t)$ **denotes the first time of termination after** t

Importance sampling estimator for off-policy prediction: $V(s) \doteq \dfrac{\sum_{t \in \mathscr{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathscr{T}(s)|}$