Chapter 5 : Monte Carlo Methods

06/08/19

Monte-Carlo methods :

→ require experience only

→ no assumption of complete knowledge of env.

→ based on averaging sample returns

* updates occur only on episode completion to have well defined returns

→ random sampling

5.1 Monte Carlo Prediction

prediction of state value function, $v_\pi(s)$

- → first visit MC (average of all returns following first)
- → every visit MC (average of all returns)

both of these converge to $v_\pi(s)$ as number of visits $\to \infty$

Blackjack

Objective : maximize sum constraint to sum ≤ 21

Rules : Face card = 10, ace = 1 or 11, number cards = number

2 cards dealt to both dealer and player, one of dealer's card is face up

Scoring : natural ← has 21 at begining

Actions : hit, stick, going bust

↖ changes turns

Win : sum closest to 21 unless someone goes bust

Exercise 5.1:

- value function jumps as probability of winning from those states is very high
- dip in left now as getting usable aces is hard and unusable aces merely add 1 to the sum
- higher sum with usable aces and hence a greater chance of winning

Exercise 5.2:

- results should be same as both converge to $v_\pi(s)$

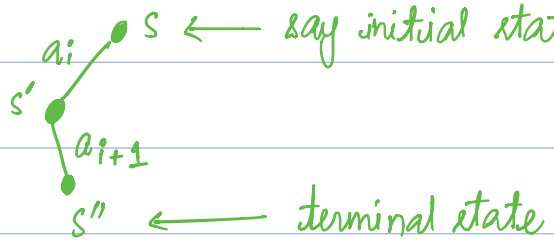
Note: Estimates for each state are independant
Compute required to get value of a single state is independant of the number of states

5.2 MC Estimation of Action Values

model-free ← action value is used
task is to estimate $q_\pi(s,a)$
two types → every visit MC
→ first visit MC

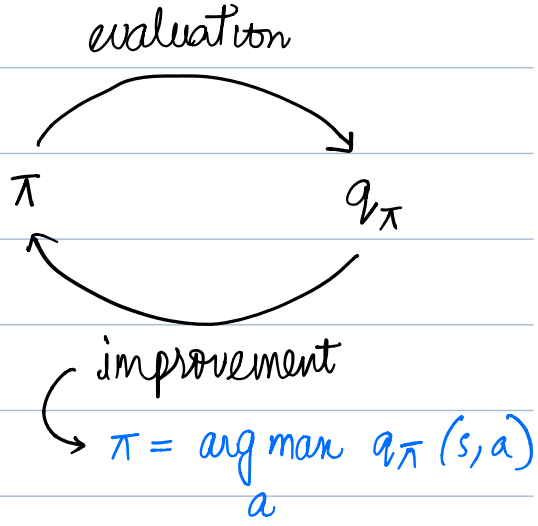
exploring starts: every s-a pair has non-zero probability of being selected at the start

Exercise 5.3 : backup diagram for MC estimation of q_π



5.3 : MC Control

GPI (generalized policy iteration)
↳ has both π and q_π



assumptions for convergence :
- have exploring starts
- infinite number of episodes
  ↳ two ways to overcome this :
  1. measurements, approximations are made to obtain bounds on mag. and prob. of error and accordingly sufficient steps are taken
     - this approach however requires too many steps for non-trivial problems
  2. in MC policy iteration eval. and improvement is alternated on an episode-by-episode basis. Policy is improved at all states visited during the episode after the end of the episode

Exercise 5.4:

$$R^t(s) \leftarrow R^t(s) + \frac{1}{t+1}\left[R - R^t(s)\right]$$

5.4: MC Control w/o Exploring Starts

- overcoming the first assumption made for convergence
- on-policy v/s off-policy
  - ↳ evaluate/improve policy that is used to make decisions

  eg: MC ES
- for on-policy methods, policy is usually soft
  - ↳ $\pi(a|s) > 0 \quad \forall a, s$

$$\begin{cases} A^* \leftarrow \arg\max_a q(s,a) \\ \pi(a|s) \leftarrow \begin{cases} 1-\varepsilon + \varepsilon/|A(s)| & \text{if } a = A^* \\ \varepsilon/|A(s)| & \text{otherwise} \end{cases} \end{cases} \quad \forall a \in A(s)$$

policy improvement theorem: assures that any ε-greedy policy wrt. q_π is an improvement over ε-soft policy π

$$\begin{aligned} q_\pi(s, \pi'(s)) &= \sum_a \pi'(a|s)\, q_\pi(s,a) \\ &= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a) \\ &\geq \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \varepsilon/|A(s)|}{1-\varepsilon}\, q_\pi(s,a) \\ &= \sum_a \pi(a|s)\, q_\pi(s,a) = v_\pi(s) \end{aligned}$$

Thus, $\pi' \geq \pi$ (ie $v_{\pi'}(s) \geq v_\pi(s) \quad \forall s \in S$)

5.5: Off-policy Prediction via Importance Sampling

dilemma: action values are learned conditional to subsequent optimal behaviour

↳ this restricts exploration

solution: have 2 policies, target and behaviour

"off-policy" methods — target: optimize; behaviour: facilitate exploration

- learning happens from data "off" target policy

off policy methods:

- slow convergence
- high variance
- more powerful, generally applicable

prediction problem: both target and behaviour policies are fixed

ie, estimate v_π, q_π given b (behaviour policy)

assumption of coverage:

$\pi(a|s) > 0$ implies that $b(a|s) > 0$

π may be deterministic

b is stochastic for states where it is non-identical to π

importance sampling :

- general method of estimating expected values under one distrib. given samples from another
- importance sampling ratio : returns are weighted according to the relative trajectories occuring under the off-policy

For initial state S_t, the probability of subsequent state action trajectory $A_t, S_{t+1}, A_{t+1}, \dots, S_T$ under policy π is :

$$\begin{aligned} &Pr\{A_t, S_{t+1}, A_{t+1}, \dots, S_T \mid S_t ; A_{t:T-1} \sim \pi\} \\ &= \pi(A_t|S_t)\, p(S_{t+1}|S_t, A_t)\, \pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1}) \\ &= \prod_{K=t}^{T} \pi(A_K|S_K)\, p(S_{K+1}|S_K, A_K) \end{aligned}$$

$$\therefore \rho_{t:T-1} = \frac{Pr\{\dots ; A_{t:T-1} \sim \pi\}}{Pr\{\dots ; A_{t:T-1} \sim b\}} = \prod_{K=t}^{T-1} \frac{\pi(A_K|S_K)}{b(A_K|S_K)}$$

this is importance sampling ratio

now, $E[G_t \mid S_t] = v_b(s)$ is transformed into :

$$v_\pi(s) = E[\rho_{t:T-1}\, G_t \mid S_t]$$

Let $\mathcal{T}(s)$ ← set of all time steps when s is visited

$T(t)$ ← first time of termination follow t

G_t ← return following t till $T(t)$

∴ under ordinary importance sampling,

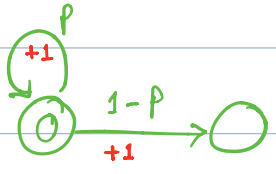
$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} \, G_t}{|\mathcal{T}(s)|}$$

∴ under weighted importance sampling,

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

- ordinary importance sampling (OIS) is unbiased whereas weighted importance sampling (WIS) is biased.
- bias converges to zero asymptotically
- variance of OIS is unbounded, for WIS it is bounded to 1

Exercise 5.5 : (?)



$G_t = \{9, 8, 7, 6, 5, 4, 3, 2, 1, 0\}$

$Pr\{\text{trajectories}\}$: $\rho^k (1-\rho)^{10-k}$

$\mathcal{T}(s)$ ← differs for every visit and first visit

$|\mathcal{T}(s)| = 10$ $\quad$ $|\mathcal{T}(s)| = 8$ or 9