Lecture - 12

* First Visit Monte-Carlo (FMC)

intuition

① Generate many episodes
② For each state, average the discounted return after the first time it was visited in each episode

Pseudocode:

Initialize:

$\pi \leftarrow$ policy to be evaluated
$V \leftarrow$ arbitrary state action function
Returns(s) $\leftarrow$ an empty list for all $s \in S$

Repeat forever

① Generate an episode using π
② For each state s appearing in episode

$G \leftarrow$ Returns following the first occurance of s
Append G to Returns(s)
$V(s) \leftarrow$ average(Returns(s))

Properties of first Visit MC

- Converges almost surely to V^{π} if each state is visited infinitely often

Proof: Consider the seq. of estimates $V_k(s)$ for a particular state s.

$$V_k(s) = \frac{1}{k} \sum_{i=1}^{k} G_i$$

where, G_i is the i^{th} return from s not the return for i^{th} time

$$\mathbb{E}[G_i] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k} \mid S_t = s, \pi\right]$$

$$= V^{\pi}(s)$$

Bias $(G_i) = \mathbb{E}[G_i] - \mu$

$V_1(s) = G_1$

$V_2(s) = \frac{G_1 + G_2}{2}$

⋮

- Strong law of large numbers

Let $X_1 \dots X_n$ be iid random variables with expected value $\mu < \infty$, then

$\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)$ is a sequence of random variables that

converges ~~to~~ almost surely to

$$\frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \mu$$

$$\therefore \Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i = \mu\right) = 1$$

$\therefore$ By LLN with

$x_i \leftarrow G_i$, we have that

$$V_k(s) \longrightarrow V^{\pi}(s)$$

since, $|s|$ is finite $V_k \longrightarrow V^{\pi}$

How quickly?

$$Var\left(V_k(s)\right) \propto 1/k$$

$$\hookrightarrow Var\left(\frac{1}{k}\sum_{j=1}^{k} G_i\right)$$

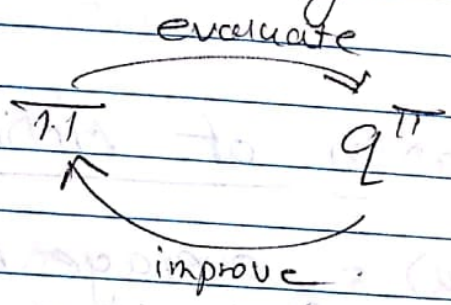
$$= \frac{1}{k^2} Var\left(\sum_{i=1}^{k} G_i\right)$$

$$= \frac{1}{k^2}\left[Var(G_1) + Var(G_2) + \cdots + Var(G_k)\right]$$

$$= \frac{1}{k^2} \cdot k\, Var(G_i) = \frac{1}{k} Var(G_i)$$

Monte-Carlo Control

- Generalized Policy Iteration (GPI)



① MC policy iteration

$\pi_0 \xrightarrow{\text{eval}} q_0 \xrightarrow{\text{improve}} \pi_2 \xrightarrow{\text{eval}} q_1 \cdots$

- Evaluation using ~~f~~-MC
- Use q rather than V.

need to see state-action value infinitely often