687 2017-10-17

Properties of First-Visit MonteCarlo

Intuition
1) generate many episodes
2) for each state average the return after the first time the state was visited in each episode

Pseudo-code:

<u>Init</u> $\pi \leftarrow$ policy to evaluate
$v \leftarrow$ arbitrary state-value fn
Returns(s) $\leftarrow$ empty list for each state

<u>Forever</u>
1) Generate an episode with π
2) for each state s in the episode
   $G \leftarrow$ return after first occurrence of s
   Append G to Returns(s)
   $v(s) \leftarrow$ average (Returns(s))

Properties:
- Converges almost surely to v^π if each state is visited infinitely often.

<u>Proof</u>: Consider sequence of estimates $v_k(s)$ for $k \in \mathbb{N}$ $(0, 1, \dots)$ for a particular state s.

$v_k(s) = \frac{1}{k} \sum_{i=1}^{k} G_i$ where G_i is the ith return from s (not return at time i)

$$E[G_i] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, \pi\right] = v^\pi(s)$$

↳ unbiased estimate of $v^\pi(s)$
(Bias$(G_i) = E[G_i] - \mu$ ← mean of distribution
Unbiased means Bias $= 0$.)

Apply Strong Law of Large Numbers:
Khintchin's S.L. of L.N.
Let $x_1, x_2, \dots$ be i.i.d. random variables w/ expected value μ (finite), then $\frac{1}{n} \sum_{i=1}^{n} x_i$ is a sequence of random variables converges to μ almost surely. I.e. $\Pr\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \mu\right) = 1$.

By this Law of Large Numbers, with the x_i being G_i, we have $v_k(s) \to E[G_i]$ almost surely. What we have is convergence of v^k at each s, i.e. *pointwise*. Given that S is finite, $v_k \to v^\pi$, i.e., *uniform* convergence.

Rate of convergence

$$Var(v_k(s)) \approx \frac{1}{k}, \quad Stddev(v_k(s)) \approx \frac{1}{\sqrt{k}}$$

$$Var(A+B) = Var(A) + Var(B) + \cancel{2\,Covar(A,B)} \to 0 \text{ since independent}$$

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

Every-Visit Monte Carlo

Only change: Add return from s at *every* place it occurs in an episode.

$\Rightarrow$ The G_i in the same episode *overlap* so previous theorem does not directly apply.

$\Rightarrow$ *However*, the process still converges. (Likely with same rate, but Phil is not certain about it.)

Policy Improvement: Monte Carlo Control

- Generalized Policy Iteration (GPI)

$$\pi \underset{\text{improve}}{\overset{\text{evaluate}}{\rightleftarrows}} q^\pi$$

MC Policy Iteration

$$\pi_0 \xrightarrow{\text{eval}} q_1 \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{eval}} q_1 \rightarrow \dots$$

- Evaluate with FV-MC
- Use q rather than v
  (v isn't adequate because we don't know P and R.)
- Same properties as PI if FV-MC is guaranteed to converge.
- Improve step is: $\pi'(s) = \arg\max_a q^\pi(s,a)$

What is the fix? More than one way:

- Exploring starts: randomly select s_0 and a_0
- Use stochastic policies

But this requires each s, a action pair to occur infinitely often!

Not quite: This is deterministic, while this isn't!