687 2017-10-19

Note Title 10/19/2017

First Visit MC:

If each state is visited infinitely often then $\hat{v}_k \xrightarrow{a.s.} v^\pi$.

If each state-action pair is visited infinitely often, then $\hat{q}_k \xrightarrow{a.s.} q^\pi$. ↑

Doing this: **MC Estimation of Action-values:**

$\hat{q}(s,a)$ = average return from the first time action $\underline{a}$ taken in state $\underline{s}$ in each episode.

Problem: What if π never chooses $\underline{a}$ in state $\underline{s}$?

Is $\hat{q}(s,a)$ still defined? **YES**

→ $\hat{q}(s,a)$ will not be estimated!

→ What to do?

Solution 1: Exploring starts:

Randomize s_0 and a_0 s.t. every (s,a) pair has a non-zero probability.

→ Not possible for all systems!

Solution 2: Stochastic Policy π

Non-zero probability for every action in each state.

Monte Carlo Control with Exploring Starts

- Avoid generating infinite # of episodes in the evaluation step by improving after a single episode.
- Accumulate returns over all episodes.

Pseudocode:

```
Init: For all $s \in \mathcal{S}$, $a \in \mathcal{A}$,
        $q(s,a) \leftarrow$ arbitrary value
        $\pi(s) \leftarrow$ " action
        Returns$(s,a) \leftarrow$ empty list
Repeat Forever
    ① Generate an episode using
      exploring starts using $\pi$
    ② For each $(s,a)$ in the episode,
        $G \leftarrow$ return following first occurrence
            of $(s,a)$ in the episode
        Append $G$ to Returns$(s,a)$
        $q(s,a) \leftarrow$ mean(Returns$(s,a)$)
    ③ For each $s$ in the episode, set
        $\pi(s) \in \arg\max_a q(s,a)$
```

If this converges, it does so to an optimal π.

Proof: Suppose it converges to a *sub*optimal π. Then q has converged to the actual values for π. Then policy improvement *must* move π to a *better* policy.

Note: It *can* *fail* to converge.

Monte Carlo Control with Stochastic Policies:

↳ ε-greedy MC Control:

```
Init: $q(s,a) \leftarrow$ arb. values
      $\pi(s,a) \leftarrow$ arb. s.t. $\pi(s,a)$
      $Returns(s,a) \leftarrow$ empty
Repeat Forever:
    ① Generate an episode using $\pi$
    ② For each $(s,a)$ in the episode:
        $G \leftarrow$ Return following first occurrence
            of $(s,a)$ in the episode
        Append $G$ to $Returns(s,a)$
        $q(s,a) \leftarrow mean(Returns(s,a))$
    ③ For each $s$ in the episode
        $a^* \in \arg\max_a q(s,a)$
        $\pi(s,a) \leftarrow \begin{cases} 1-\epsilon+\frac{\epsilon}{|a|} & \text{if } a=a^* \\ \frac{\epsilon}{|a|} & \text{if } a\neq a^* \end{cases}$
```

Properties:

- By variations on the Policy Improvement Thm (Sutton & Barto, Vol. I, Sect. 5.4), converges to an ε-greedy policy.
- In fact to the optimal ε-greedy policy (not necessarily optimal among all policies).

TD (Temporal Difference) Learning:

Properties:

- Like MC, learns directly from experiences: does not need P and R.
- Like DP (dynamic pgming): update estimates in terms of other estimates.

But first, another MC algorithm:

Consider updates: $f(x) \leftarrow f(x) + \alpha\,(Y - f(x))$

"stepsize" target (usually an estimate)

$f_w(x)$

weights

$$L(w) = \mathbb{E}\left[\tfrac{1}{2}\left(Y - f_w(x)\right)^2\right]$$

$$w \leftarrow w + \alpha\,\underbrace{\mathbb{E}\left[\left(Y - f_w(x)\right)\cdot \frac{\partial f_w(x)}{\partial w}\right]}_{\text{Gradient of } L(w)}$$

gradient descent

estimate from one sample: stochastic gradient descent

$$w \leftarrow w + \alpha\,(Y - f_w(x))\,\frac{\partial f_w(x)}{\partial w_x} = \begin{bmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{bmatrix}$$

1 only in the x^{th} spot

$$v(s_t) \leftarrow v(s_t) + \alpha\,(G_t - v(s_t))$$

$$\sum_{k=0}^{\infty} \gamma^k R_{t+k}$$

stochastic gradient descent on

$$L(v) = \mathbb{E}\left[\left(G_t - v(s_t)\right)^2\right]$$

But have to wait until the end of the episode!

Use the Bellman Equation:

$$v^\pi(s) = \mathbb{E}\left[R_t + \gamma v^\pi(S_{t+1}) \mid S_t = s\right]$$

$$\Downarrow$$

$$\boxed{v(s_t) \leftarrow v(s_t) + \alpha\left(\underline{R_t + \gamma v(S_{t+1}) - v(s_t)}\right)}$$

Called the TD Update

so $f(x) \leftarrow f(x) + \alpha\,(Y - f(x))$

$\boxed{\text{The TD Error: } \delta_t}$

$$v(s) \leftarrow v(s) + \alpha\,\delta_t$$

"Reward Prediction Error"

MC is a stochastic gradient descent.

Is TD stochastic gradient descent?

If it is: $f(v) = \mathbb{E}\left[\frac{1}{2}\left(R_t + \gamma v(S_{t+1}) - v(S_t)\right)^2\right]$

$= \mathbb{E}\left[\frac{1}{2}\delta_t^2\right]$

$\nabla f(v) = \delta_t \cdot \frac{\partial \delta_t}{\partial v}$ (ignoring expectation)

$= \left(R_t + \gamma v(S_{t+1}) - v(S_t)\right)\left(\gamma \frac{\partial v(S_{t+1})}{\partial v} - \frac{\partial v(S_t)}{\partial v}\right.$

$$S_{t+1} \rightarrow \begin{bmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{bmatrix} \qquad \begin{bmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{bmatrix} \leftarrow S_t$$

$v(S_t) \leftarrow v(S_t) + \alpha \delta_t$

$v(S_{t+1}) \leftarrow v(S_{t+1}) - \alpha \gamma \delta_t$

Seems <u>close</u> to s.g.d., but it is not properly sampled. (Will cover later in the course.)

<u>Some properties</u>:

- TD offers an update at each step.
- Converges a.s. to correct v.