687 2017-10-31

Note Title 10/31/2017

__Sarsa__ (s,a,r,s',a')

$$\left.\begin{aligned}\delta_t &\leftarrow R_t + \gamma q(S_{t+1}, A_{t+1}) - q(S_t, A_t)\\ q(S_t, A_t) &\leftarrow q(S_t, A_t) + \alpha \delta_t\end{aligned}\right\} \text{on-policy}$$

converges to $\longrightarrow q^{\pi_b}$ under sampling policy π_b

__Q-learning__ (s,a,r,s')

$$\left.\begin{aligned}\delta_t &\leftarrow R_t + \gamma \max_{a'} q(S_{t+1}, a') - q(S_t, A_t)\\ q(S_t, A_t) &\leftarrow q(S_t, A_t) + \alpha \delta_t\end{aligned}\right\} \text{off-policy} \longrightarrow q^*$$

__Init__ $q(s,a)$ arbitrarily

__Repeat__ (for each episode)

  __Init__ $s \sim d_0$

  __Repeat__ (for each time step)

  - choose a from s using the policy derived from q
  - take action a, observe r, s'
  - update $q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a))$
  - $s \leftarrow s'$

  __Until__ s is terminal

Course Recap

① Definitions

MDP

$\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma$

v^*, q^*, v^π, q^π

J, π

② Dynamic Programming

Policy evaluation

+ Bellman Eqn.

Policy Improvement

Policy Iteration

Value Iteration

- Bellman optimality eqn.
- Bellman operator

③ Monte Carlo approximation

④ TD

Same

+ Func. Approx.

⑤ TD (λ)

Same

⑥ Policy Gradient (∇) Theorems

Midterm

Convergence

	TD	Sarsa	Q-learning
Tabular	✓	✓	✓
Linear	✓	✓	X
General	X	X	X

n-Step Returns

- Different potential targets for value function updates.

$G_t^{(n)}$ is the n-step return

$G_t^{(1)} = R_t + \gamma v(S_{t+1}) = G_t^{(TD)}$

$G_t^{(2)} = R_t + \gamma R_{t+1} + \gamma^2 v(S_{t+2})$

$\boxed{\text{Larger } n \text{ gives lower bias, higher variance.}}$

$$G_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k R_{t+k} + \gamma^n v(S_{t+n})$$

$G_t^{(\infty)}$ = Monte Carlo = G_t or $G_t^{(MC)}$

Idea: use more than one (in fact, of <u>all</u>): Use a weighted average, where weights sum to 1.

λ-Returns

$$G_t^{\lambda} \triangleq \left(\lambda^0 G_t^{(1)} + \lambda^1 G_t^{(2)} + \ldots + \lambda^i G_t^{(i+1)} + \ldots\right) \cdot (1-\lambda)$$

where $\lambda \in (0, 1)$

$$= (1-\lambda) \sum_{i=1}^{\infty} \lambda^{i-1} G_t^{(i)} \text{ if } \lambda < 1$$

and $G_t^{(MC)}$ otherwise

This approach is intuitively appealing to reduce variance - but not why we do it.

Andy Barto says "The math works out." In particular, it works backwards for actual computation.

TD γ ... Complex .... @ NIPS 2011
Ω-return @ " 2015
ICML 2016

λ-return is MLE-estimate of $v^\pi(S_t)$ if:

1. $G_t^{(1)}, G_t^{(2)}, \ldots$ are independent
2. Every $G_t^{(i)}$ is normally distributed
3. $\mathrm{Var}(G_t^{(i)}) \propto \frac{1}{\lambda^i}$ $\quad (\sigma^2 = \beta/\lambda^i)$
4. $\mathbb{E}[G_t^{(i)}] = v^\pi(S_t)$ for all i

These assumptions are false, but the rest of the math works out!

To find best, $\frac{d}{dx}$ & set to 0

$$\sum_{i=1}^{\infty} 2(G_t^{(i)} - X)\lambda^i = 0$$

$$\sum_{i=1}^{\infty} G_t^{(i)}\lambda^i = X\sum_{i=1}^{\infty}\lambda^i \Rightarrow X = \left(\sum_{i=1}^{\infty} G_t^{(i)}\lambda^i\right)\frac{(1-\lambda)}{\lambda} = \left(\sum_{i=1}^{\infty} G_t^{(i)}\lambda^{i-1}\right)(1-\lambda)$$

The λ-return.

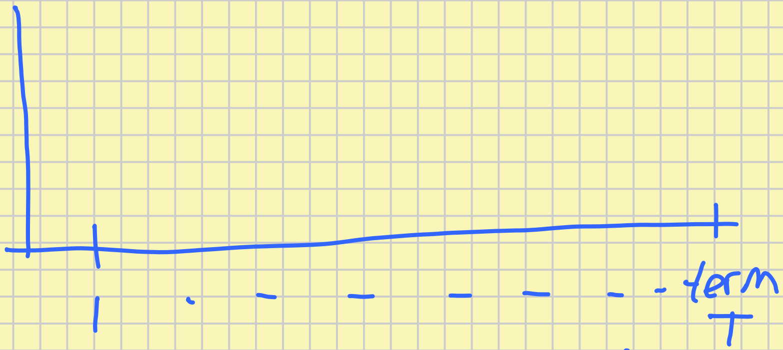
Proof:

$$L(X \mid G_t^{(1)}, G_t^{(2)}, \ldots)$$
$$= \Pr(G_t^{(1)}, G_t^{(2)}, \ldots \mid v^\pi(S_t) = X)$$
$$= \prod_{i=1}^{\infty} \Pr(G_t^{(i)} \mid v^\pi(S_t) = X)$$

$$\arg\max_X L(X) = \arg\max_X \ln L(X)$$

$$\arg\max_X L(X) = \arg\max_X \sum_{i=1}^{\infty} \ln \Pr(G_t^{(i)} \mid v^\pi(S_t) = X)$$
$$= \arg\max_X \sum_{i=1}^{\infty} \ln\left(\frac{1}{\sqrt{2\pi\,\beta/\lambda^i}}\, e^{-(G_t^{(i)}-X)^2 / \frac{2\beta}{\lambda^i}}\right)$$
$$= \arg\max_X \sum_{i=1}^{\infty}\left[\ln\left(\frac{1}{\sqrt{2\pi\,\beta/\lambda^i}}\right) - \frac{(G_t^{(i)}-X)^2\lambda^i}{2\beta}\right]$$
$$= \arg\max_X -\sum_{i=1}^{\infty}(G_t^{(i)}-X)^2\lambda^i$$

What is the λ-return at a terminal state?



$G_t^{(T)} = G_t^{(T+1)} = \cdots$

imagine if we put all this weight on $G_t^{(T)}$

Shows that larger λ gives G's closer to Monte Carlo.

λ-return Algorithm

Use $G_t^{(\lambda)}$ as target. $v(S_t) \leftarrow v(S_t) + \alpha\left(G_t^{(\lambda)} - v(S_t)\right)$ ← "Forward view," which requires waiting until the end of the episode.