Comparison between Sarsa and Expected Sarsa

Tom Bosc

10/02/17

Introduction

- TD methods combines model-free learning and bootstrapping.
- Sarsa: On-policy
- Q-learning: Off-policy
- Expected Sarsa: On-policy but also generalizes Q-learning

Update rules

Update rules:

- Sarsa:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$$

- Expected Sarsa:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \sum_{a'} \pi(a'|s') Q(s',a') - Q(s,a)]$$

Corresponding estimators (consider a, s, s' fixed here).

- Sarsa: $\hat{v}_t = r + \gamma Q(s', A')$
- Expected Sarsa: $v_t = r + \gamma \sum_{a'} \pi(a'|s') Q(s', a')$

Bias

Check that the 2 estimators are unbiased. Expectation over the policy (variable A').

$$E[v_t] = E[r+\gamma \sum_{a'} \pi(a'|s')Q(s',a')] = E[r]+\gamma E[V(s')] = E[r]+\gamma V(s')$$

$$E[\hat{v}_t] = E[r + \gamma Q(s', A')] = E[r] + \gamma V(s')$$

So $E[v_t] = E[\hat{v}_t]$.
Now, compute the variance.

Variance

Compute the variance:

$$Var(\hat{v}_t - v_t) = E[\hat{v}_t{}^2] - E[\hat{v}_t]^2 - (E[v_t^2] - E[v_t]^2)$$

$$Var(\hat{v}_t - v_t) = E[\hat{v}_t{}^2] - E[v_t^2]$$

$$Var(\hat{v}_t - v_t) = E[(r + \gamma Q(s', A'))^2] - E[(r + \gamma \sum_{a'} \pi(a'|s')Q(s', a'))^2]$$

$$Var(\hat{v}_t - v_t) = E[r] - E[r] + 2\gamma(E[Q(s', A')] - E[\sum_a Q(s', a)\pi(a|s')]) +$$

$$\gamma^2(E[Q(s', A')^2] - E[(\sum_a \pi(a|s')Q(s', a))^2])$$

$$Var(\hat{v}_t - v_t) = \gamma^2(E[Q(s', a')^2] - E[(\sum_a \pi(a|s')Q(s', a))^2])$$

Variance (2)

$$Var(\hat{v}_t - v_t) = \gamma^2(E[Q(s', a')^2] - E[(\sum_a \pi(a|S')Q(S', a))^2])$$

$$Var(\hat{v}_t - v_t) = \gamma^2(\sum_{a'} \pi(a|s')Q(s', a')^2] - E[(\sum_a \pi(a|s')Q(s', a))^2])$$

Variance (3)

The inner term is of the form:

$$\sum_i w_i x_i^2 - (\sum_i w_i x_i)^2 \quad , \tag{11}$$

where the w and x correspond to the π and Q values. When $w_i \geq 0$ for all i and $\sum_i w_i = 1$, we can give an unbiased estimate of the variance of the weighed values $w_i x_i$ as follows:

$$\frac{\sum_i w_i (x_i - \bar{x})^2}{1 - \sum_i w_i^2} \quad , \tag{12}$$

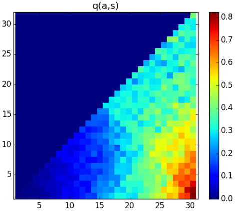
where $\bar{x}$ is the weighted mean $\sum_i w_i x_i$. Taking the numerator of this fraction and rewriting this gives us:

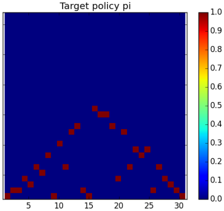
$$\begin{aligned} \sum_i w_i (x_i - \bar{x})^2 &= \sum_i w_i x_i^2 - 2 \sum_i w_i x_i \bar{x} + \sum_i w_i \bar{x}^2 \\ &= \sum_i w_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \sum_i w_i x_i^2 - \bar{x}^2 \quad , \end{aligned}$$

which is exactly the same quantity as given in (11). This

Experiment: Gambler's problem

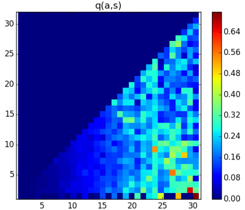
Goal is to reach 32. $p_h = 0.3$
MC with exploring starts, 100k episodes.

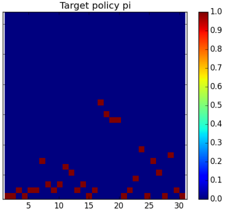




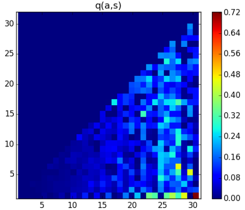
Experiment: Gambler's problem

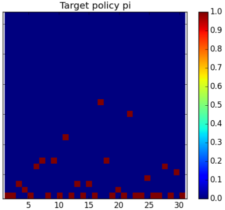
Sarsa, 300k updates, $\alpha = 0.1$, $\epsilon = 0.05$.





Expected Sarsa, 100k updates, $\alpha = 0.1$, $\epsilon = 0.05$.





Bibliography

- A Theoretical and Empirical Analysis of Expected Sarsa, Seijen et al., 2009