Bias-Variance Trade-Off of Expected Sarsa vs. Sarsa

Lucas Berry

Comp 767

March 10th, 2017

Sarsa

The Sarsa algorithm solves the control problem by updating state action pair values. Using the notation from our book the updates have the following form,

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right].$$

Expected Sarsa

Expected Sarsa solves the control problem by updating state action pair values as in Sarsa. Their updates take expectations over the next state action pair values. The updates can be written as,

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t) \Big].$$

Bias

Let $v_t = r_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)$ and $\hat{v}_t = r_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$ denote the target of Expected Sarsa and Sarsa respectively. Using v_t and $\hat{v}_t$ one can show the bias of the two algorithms equal.

$$Bias(s, a) = Q_\pi(s, a) - E(X_t),$$

where X_t is either v_t or $\hat{v}_t$.

Bias

Checking the expectation of v_t,

$$
\begin{aligned}
E[v_t] &= E\left[r_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)\right] \\
&= E[r_{t+1}] + \gamma E\left[\sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)\right] \\
&= E[r_{t+1}] + \gamma E\left[E\left(Q(S_{t+1}, A_{t+1})|S_{t+1}\right)\right] \\
&= E[r_{t+1}] + \gamma E[Q(S_{t+1}, A_{t+1})] \qquad \text{Tower Property} \\
&= E[\hat{v}_t].
\end{aligned}
$$

Thus both algorithms produce the same bias.

Variance

The variance of both estimates can be derived by:

$$Var(s, a) = E\left[(X_t)^2\right] - (E\left[X_t\right])^2 .$$

Using this we will compare the variance of both algorithms.

Variance

Sarsa:

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

Expected Sarsa:

$$Var(s,a) = \sum_{s'} p(s'|s,a) \Bigg[r(s,a,s')^2 + 2\gamma r(s,a,s') \sum_{a'} \pi(a'|s') Q(s',a') + \gamma^2 \left(\sum_{a'} \pi(a'|s') Q(s',a') \right)^2 \Bigg] - \left(E\left[v_t\right]\right)^2 .$$

Variance

Taking the difference of the variances yields,

$$\gamma^2 \sum_{s'} p(s'|s,a) \left[\sum_{a'} \pi(a'|s')(Q(s',a'))^2 - \left(\sum_{a'} \pi(a'|s') Q(s',a') \right)^2 \right].$$

Rewriting the inner term gives,

$$\sum_i w_i x_i^2 - (\sum_i w_i x_i)^2,$$

where $w_i = \pi(a'|s')$ and $x_i = Q(s',a')$.

Variance

Thus the inner term is $Var(Q(s', a')|s')$, therefore the past expression must be greater than or equal to 0. Suggesting the variance of Sarsa is greater than or equal to Expected Sarsa.

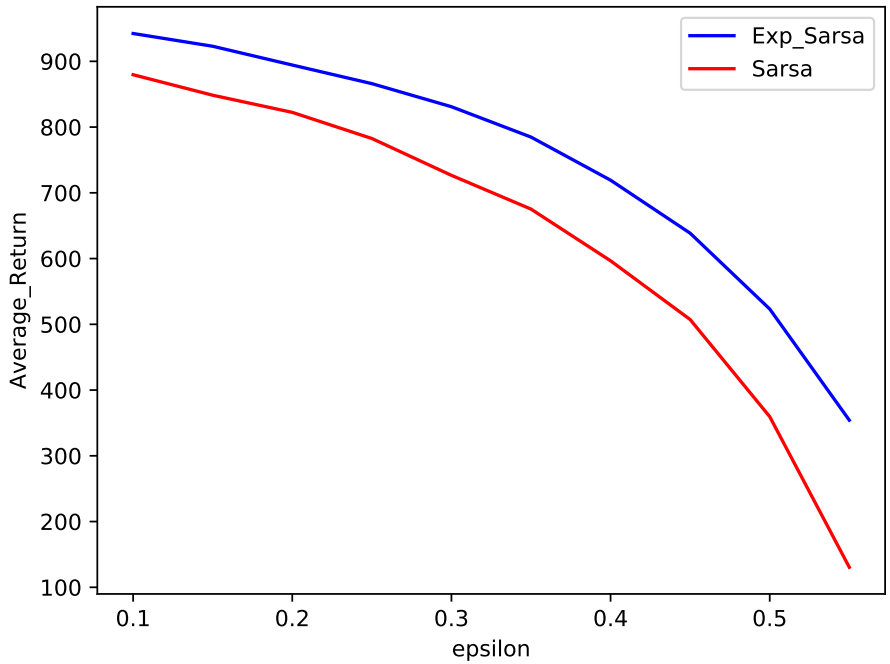
Experiment

St is the start, W are the walls and G is the goal.

			W	W	W
					G
			W	W	W
St			W	W	W

Varying Exploration

Increasing the Stochasticity in the policy creates a bigger difference between the two methods. The following graph and table depicts this

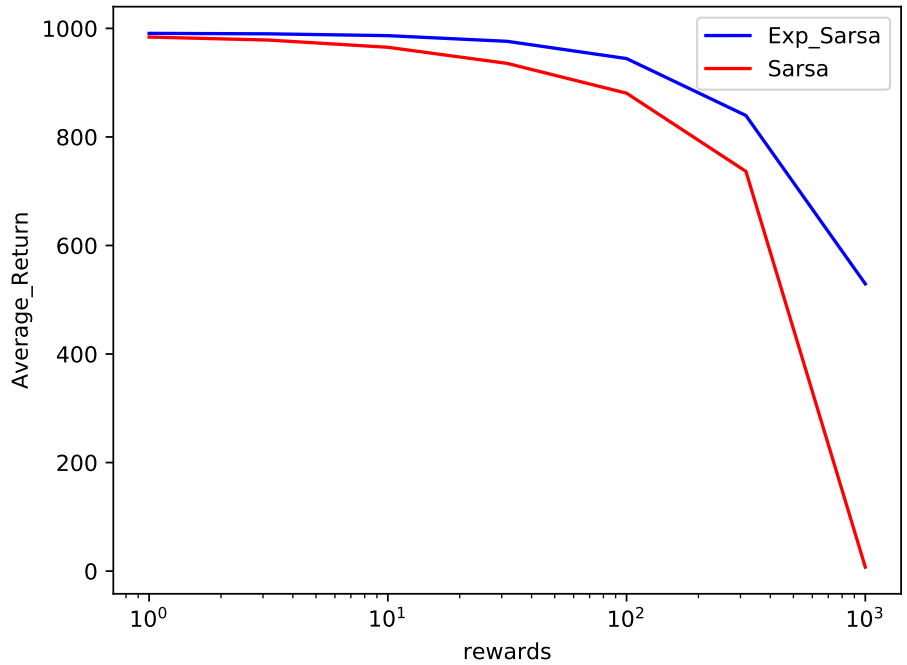


Epsilon	ExpSar-Sar
0.1	62.70
0.15	71.99
0.2	74.46
0.25	84.44
0.30	104.37
0.35	109.61
0.40	122.61
0.45	131.30
0.5	163.72
0.55	223.69

Table: Difference between the two curves

Varying Penalty for Hitting the Wall

Increasing the penalty for hitting the wall should create different Q-values thus translating in a larger variance. The following graph and table depicts this



Penalty	ExpSar-Sar
-1	6.80
$-1(10^{.5})$	11.47
$-1(10^{1})$	21.45
$-1(10^{1.5})$	40.69
$-1(10^{2})$	63.76
$-1(10^{2.5})$	102.95
$-1(10^{3})$	521.90

Table: Difference between the two curves

Notes

All code had 50 runs with 100 episodes and then averaged with $\alpha = .6$. For the first graph and table the penalty for hitting a wall was set to -100 and for the second $\epsilon = .1$.