Bias-Variance tradeoff
SARSA(0) vs Expected SARSA(0)

A Theoretical and Empirical Analysis of Expected Sarsa
van Seijen, van Hasselt, Whiteson, and Weiring (2009)

COMP 767 – Reinforcement Learning
February 10th

Code:

~Nicolas Angelard-Gontier

Theoretical Analysis

- Expected SARSA shares the same convergence guarantees as Sarsa and thus finds the optimal policy in the limit.
- Expected SARSA has same bias and lower variance in its updates than SARSA ~> *alpha* can be increased to speedup learning.
- SARSA update rule:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha [r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

- Expected SARSA update rule prevent stochasticity from the policy to increase Variance:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha [r_{t+1} + \gamma E\{Q(s_{t+1}, a_{t+1})\} - Q(s_t, a_t)]$$

$$= Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \sum_a \pi(a, s_{t+1}) Q(s_{t+1}, a) - Q(s_t, a_t)]$$

Theoretical Analysis

- Expected SARSA shares the same convergence guarantees as Sarsa and thus finds the optimal policy in the limit.
- Expected SARSA has same bias and lower variance in its updates than SARSA ~> *alpha* can be increased to speedup learning.
- SARSA update rule:

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[\boxed{r_{t+1}+\gamma Q(s_{t+1},a_{t+1})}-Q(s_t,a_t)]$$

- Expected SARSA update rule <u>prevent stochasticity from the policy to increase Variance</u>:

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[r_{t+1}+\gamma E\{Q(s_{t+1},a_{t+1})\}-Q(s_t,a_t)]$$

$$=Q(s_t,a_t)+\alpha\boxed{[r_{t+1}+\gamma\sum_a \pi(a,s_{t+1})Q(s_{t+1},a)}-Q(s_t,a_t)]$$

Theoretical Analysis (Bias)

- For both algorithms: $Bias(s,a) = Q^{\pi}(s,a) - E\{X_t\}$

  with X being the target of either algorithm:

  $X_t = r_t + \gamma \sum_a \pi_t(a, s_{t+1}) Q_t(s_{t+1}, a)$ for E.Sarsa

  $\hat{X}_t = r_t + \gamma Q_t(s_{t+1}, a_{t+1})$ for Sarsa

- These two targets are similar in Expectation:

$$E\{\hat{X}_t\} = \sum_{s'} P(s'|s,a)[R(s'|s,a) + \gamma \sum_{a'} \pi(a',s') Q(a',s')]$$
$$= E\{X_t\}$$

  So $E\{\hat{X}_t\} = E\{X_t\}$ and both algorithms have the same bias.

Theoretical Analysis (Variance)

- For both algorithms: $Var(s,a) = E\{X_t^2\} - E\{X_t\}^2$

  with X being the target of either algorithm:

  $X_t = r_t + \gamma \sum_a \pi_t(a, s_{t+1}) Q(s_{t+1}, a)$ for Expected Sarsa

  $\hat{X}_t = r_t + \gamma Q(s_{t+1}, a_{t+1})$ for Sarsa

Theoretical Analysis (Variance)

- For both algorithms: $Var(s,a) = E\{X_t^2\} - \cancel{E\{X_t\}^2}$

  with X being the target of either algorithm:

  $X_t = r_t + \gamma \sum_a \pi_t(a, s_{t+1}) Q(s_{t+1}, a)$ for Expected Sarsa

  $\hat{X}_t = r_t + \gamma Q(s_{t+1}, a_{t+1})$ for Sarsa

- Sarsa: $E\{\hat{X}_t^2\} = \sum_{s'} P(s'|s,a)[A]$

$$A = R(s'|s,a)^2 + \gamma^2 \sum_{a'} \pi(a',s') Q_t(s',a')^2 + 2\gamma R(s'|s,a) \sum_{a'} \pi(a',s') Q_t(s',a')$$

- E.Sarsa: $E\{X_t^2\} = \sum_{s'} P(s'|s,a)[B]$

$$B = R(s'|s,a)^2 + \gamma^2 \left(\sum_{a'} \pi(a',s') Q_t(s',a')\right)^2 + 2\gamma R(s'|s,a) \sum_{a'} \pi(a',s') Q_t(s',a')$$

Theoretical Analysis (Variance)

- For both algorithms: $Var(s,a)=E\{X_t^2\}-\cancel{E\{X_t\}^2}$

  with X being the target of either algorithm:

  $$X_t=r_t+\gamma\sum_a \pi_t(a,s_{t+1})Q(s_{t+1},a) \text{ for Expected Sarsa}$$

  $$\hat{X}_t=r_t+\gamma Q(s_{t+1},a_{t+1}) \text{ for Sarsa}$$

- Sarsa: $E\{\hat{X}_t^2\}=\sum_{s'} P(s'|s,a)[A]$

$$A=\cancel{R(s'|s,a)^2}+\gamma^2\sum_{a'}\pi(a',s')Q_t(s',a')^2+\cancel{2\gamma R(s'|s,a)\sum_{a'}\pi(a',s')Q_t(s',a')}$$

- E.Sarsa: $E\{X_t^2\}=\sum_{s'} P(s'|s,a)[B]$

$$B=\cancel{R(s'|s,a)^2}+\gamma^2\left(\sum_{a'}\pi(a',s')Q_t(s',a')\right)^2+\cancel{2\gamma R(s'|s,a)\sum_{a'}\pi(a',s')Q_t(s',a')}$$

$$E\{\hat{X}_t\}-E\{X_t\}=$$

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

$$\sim\sum_a \pi_a Q_a^2-\left(\sum_a \pi_a Q_a\right)^2$$

Theoretical Analysis (Variance)

Var(Sarsa) – Var(E. Sarsa) ~ $\sum_a \pi_a Q_a^2 - \left(\sum_a \pi_a Q_a\right)^2$

Let $\bar{Q} = \sum_a \pi_a Q_a$ be the weighted mean

Note that:

$$\sum_a \pi_a \left(Q_a - \bar{Q}\right)^2 = \sum_a \pi_a Q_a^2 - 2\sum_a \pi_a Q_a \bar{Q} + \sum_a \pi_a \bar{Q}^2$$

Theoretical Analysis (Variance)

Var(Sarsa) – Var(E. Sarsa) ~ $\sum_a \pi_a Q_a^2 - \left(\sum_a \pi_a Q_a\right)^2$

Let $\bar{Q} = \sum_a \pi_a Q_a$ be the weighted mean

Note that:

$$\sum_a \pi_a \left(Q_a - \bar{Q}\right)^2 = \sum_a \pi_a Q_a^2 - 2\overbrace{\sum_a \pi_a Q_a}^{\bar{Q}}\bar{Q} + \overbrace{\sum_a \pi_a}^{1}\bar{Q}^2$$

Theoretical Analysis (Variance)

Var(Sarsa) – Var(E. Sarsa) ~ $\sum_a \pi_a Q_a^2 - \left(\sum_a \pi_a Q_a\right)^2$

Let $\bar{Q} = \sum_a \pi_a Q_a$ be the weighted mean

Note that:

$$\sum_a \pi_a (Q_a - \bar{Q})^2 = \sum_a \pi_a Q_a^2 - 2\underbrace{\sum_a \pi_a Q_a}_{\bar{Q}} \bar{Q} + \underbrace{\sum_a \pi_a}_{1} \bar{Q}^2$$

$$= \sum_a \pi_a Q_a^2 - 2\bar{Q}^2 + \bar{Q}^2$$

$$= \sum_a \pi_a Q_a^2 - \bar{Q}^2$$

The more Q_a deviate from the weighted mean $\sum_a \pi_a Q_a$, the larger this difference will be.

Occurs when big difference in values of *Q(s,a)* (for fixed *s*) and when there is much exploration.

Empirical Analysis

Race grid world example:

Rewards:
GOAL = +100
WALL = -10
STEP = -1

Actions: V in [0,3]

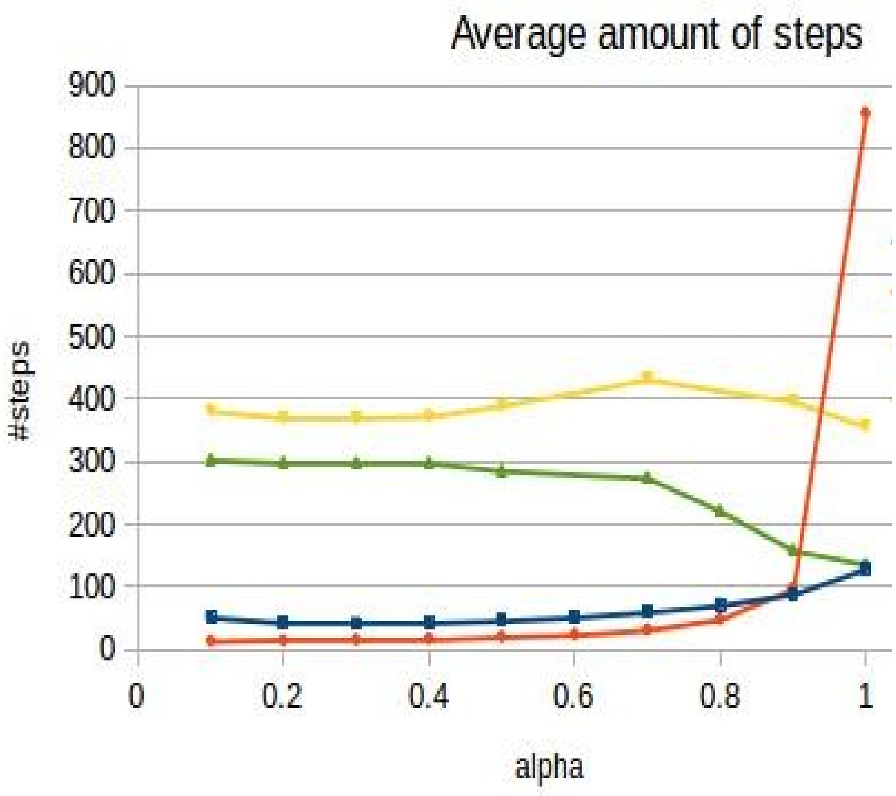
	V – 1	V + 0	V + 1
RIGHT	0	1	2
UP	3	3	5
LEFT	6	7	8

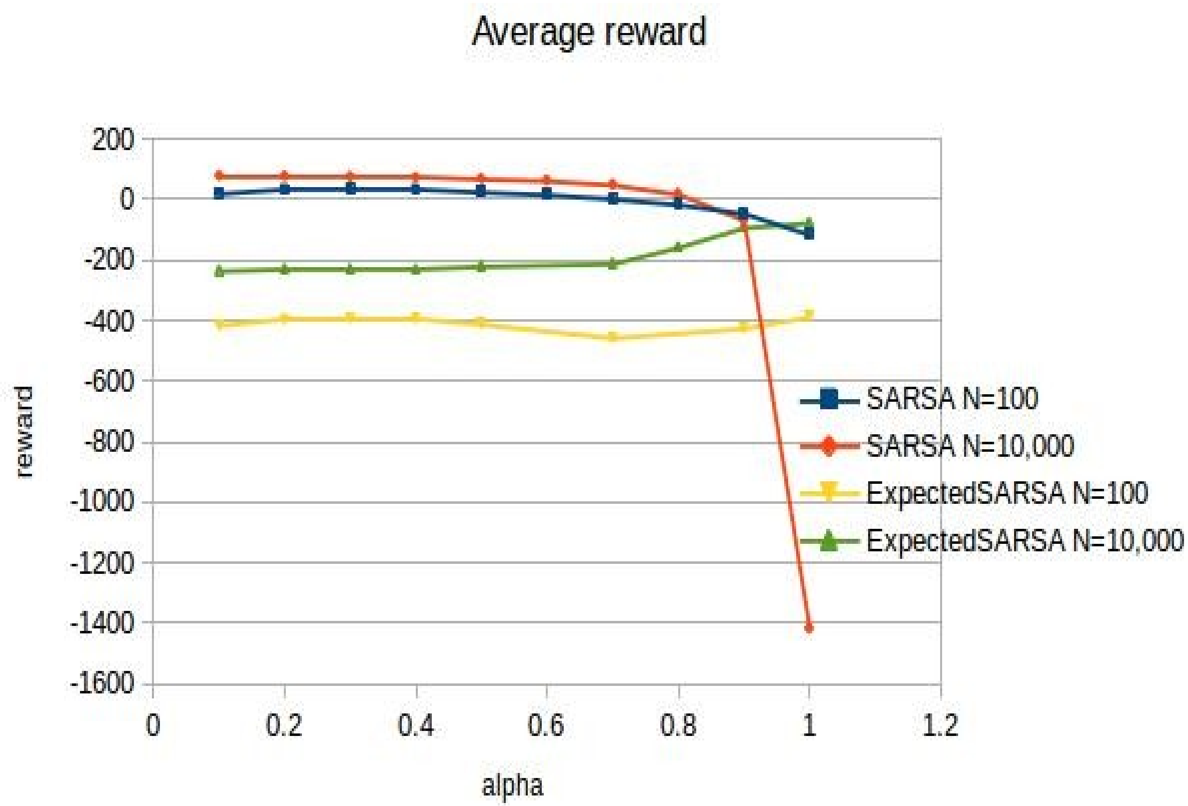
Crash:
return to S and V=0

Policy Stochastisity: “epsilon-greeddy”: with proba **epsilon**, do a random non-optimal action.
Environment Stochasticity: with probability **beta**, do not update velocity no matter the action.

Empirical Analysis

epsilon = 0.1 <-- policy stochasticity
beta = 0.1 <-- environment stochasticity

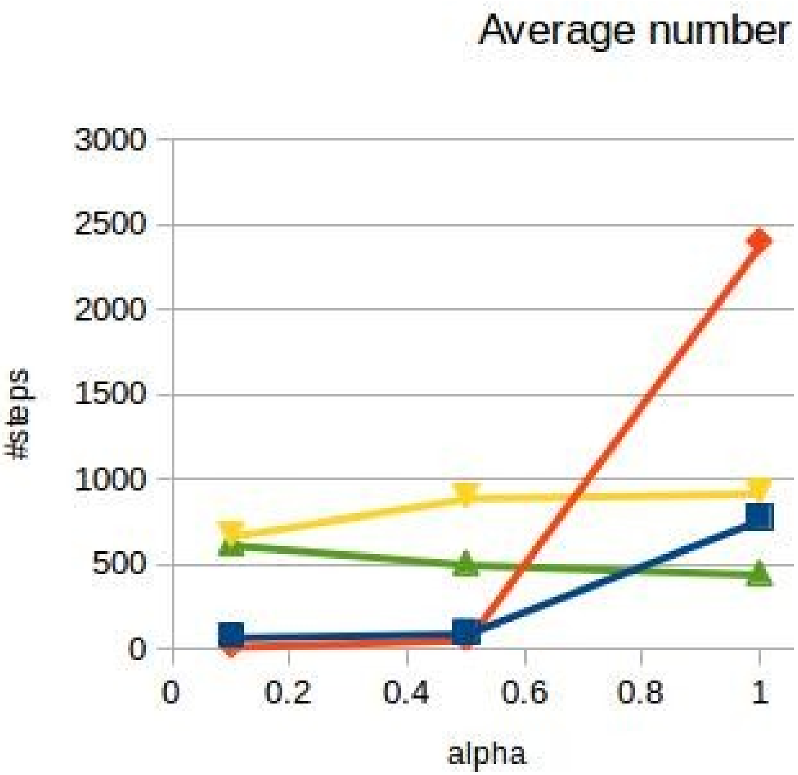


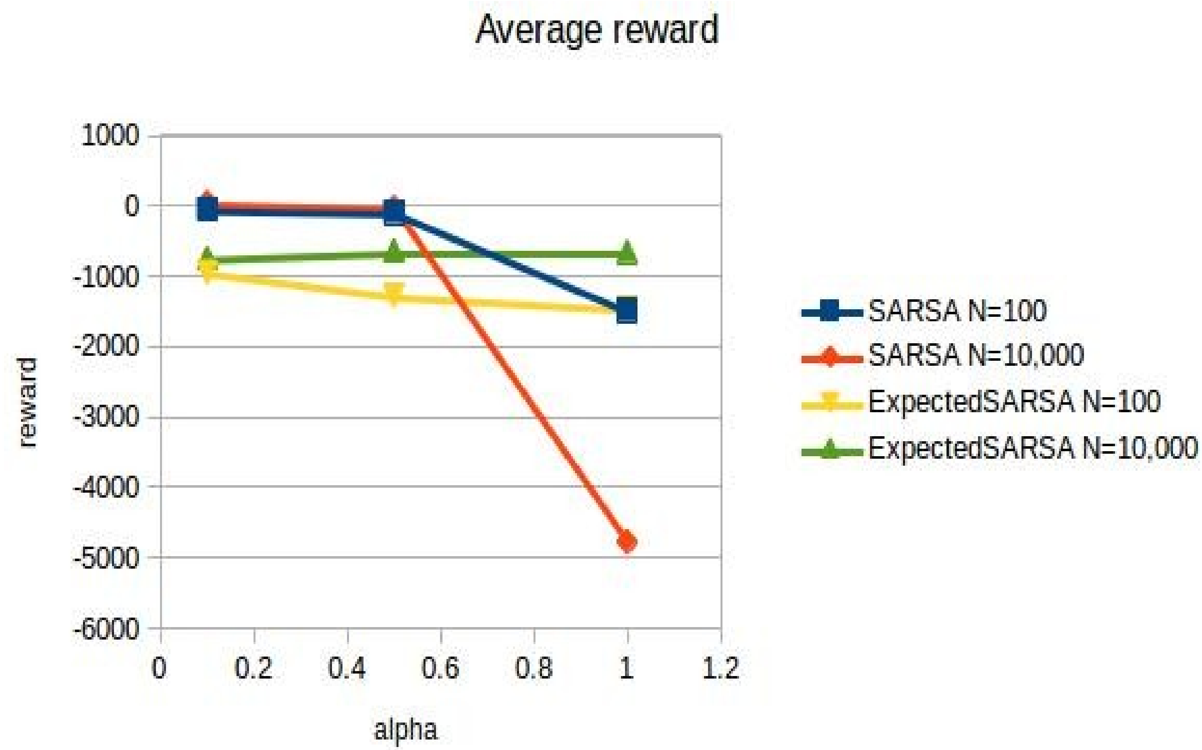


N=100 – average over 1000 tries
N=10,000 – average over 10 tries

Empirical Analysis

epsilon = 0.3 <-- policy stochasticity
beta = 0.1 <-- environment stochasticity





N=100 – average over 1000 tries
N=10,000 – average over 10 tries