Sarsa vs. Expected Sarsa

Michael Noseworthy

Review

- Sarsa Update

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

- Expected Sarsa Update

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

Variance (Sarsa)

$$
\begin{aligned}
Var(v_t) &= \mathbb{E}\left\{\left(r+\gamma Q\left(s',a'\right)\right)^2\right\} - \left(\mathbb{E}\left\{v_t\right\}\right)^2 \\
&= \mathbb{E}\left\{r^2 + 2\gamma r Q\left(s',a'\right) + \gamma^2 Q^2\left(s',a'\right)\right\} - \left(\mathbb{E}\left\{v_t\right\}\right)^2 \\
&= \sum_{s'} p(s'|s,a)\left[\gamma^2 \sum_{a'} \pi(s',a') Q^2\left(s',a'\right) + \left(r_{sa}^{s'}\right)^2\right] \\
&+ \left[2\gamma \left(r_{sa}^{s'}\right)^2 \sum_{a'} \pi(s',a') Q\left(s',a'\right)\right] - \left(\mathbb{E}\left\{v_t\right\}\right)^2
\end{aligned}
$$

Variance (Expected Sarsa)

$$
\begin{aligned}
Var(v_t) &= \mathbb{E}\left\{\left(r + \gamma \sum_{a'} \pi(s', a') Q(s', a')\right)^2\right\} - \left(\mathbb{E}\left\{v_t\right\}\right)^2 \\
&= \mathbb{E}\left\{r^2 + 2\gamma r \sum_{a'} \pi(s', a') Q(s', a') + \gamma^2 \left(\sum_{a'} \pi(s', a') Q(s', a')\right)^2\right\} - \left(\mathbb{E}\left\{v_t\right\}\right)^2 \\
&= \sum_{s'} p(s'|s, a) \left[\gamma^2 \left(\sum_{a'} \pi(s', a') Q(s', a')\right)^2 + (r_{sa}^{s'})^2\right] \\
&+ \left[2\gamma (r_{sa}^{s'})^2 \sum_{a'} \pi(s', a') Q\left(s', a'\right)\right] - \left(\mathbb{E}\left\{v_t\right\}\right)^2
\end{aligned}
$$

Variance Comparison (1)

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

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

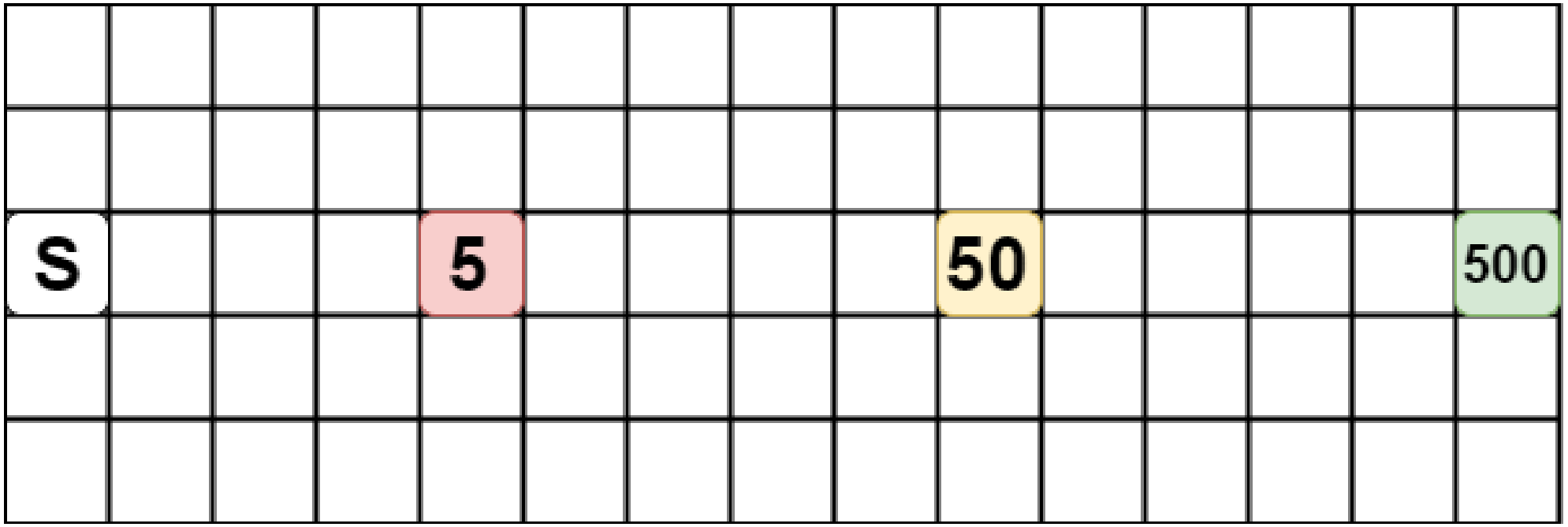
Variance Comparison (2)

$$var(v_t) - var(\hat{v}_t) = \gamma^2 \sum_{s'} p(s'|s,a) \left[\sum_{a'} \pi(s',a') Q^2(s',a') - \left(\sum_{a'} \pi(s',a') Q(s',a') \right)^2 \right]$$

Variance of Weighted Sum: $\sum_i w_i x_i^2 - \left(\sum_i w_i x_i \right)^2$

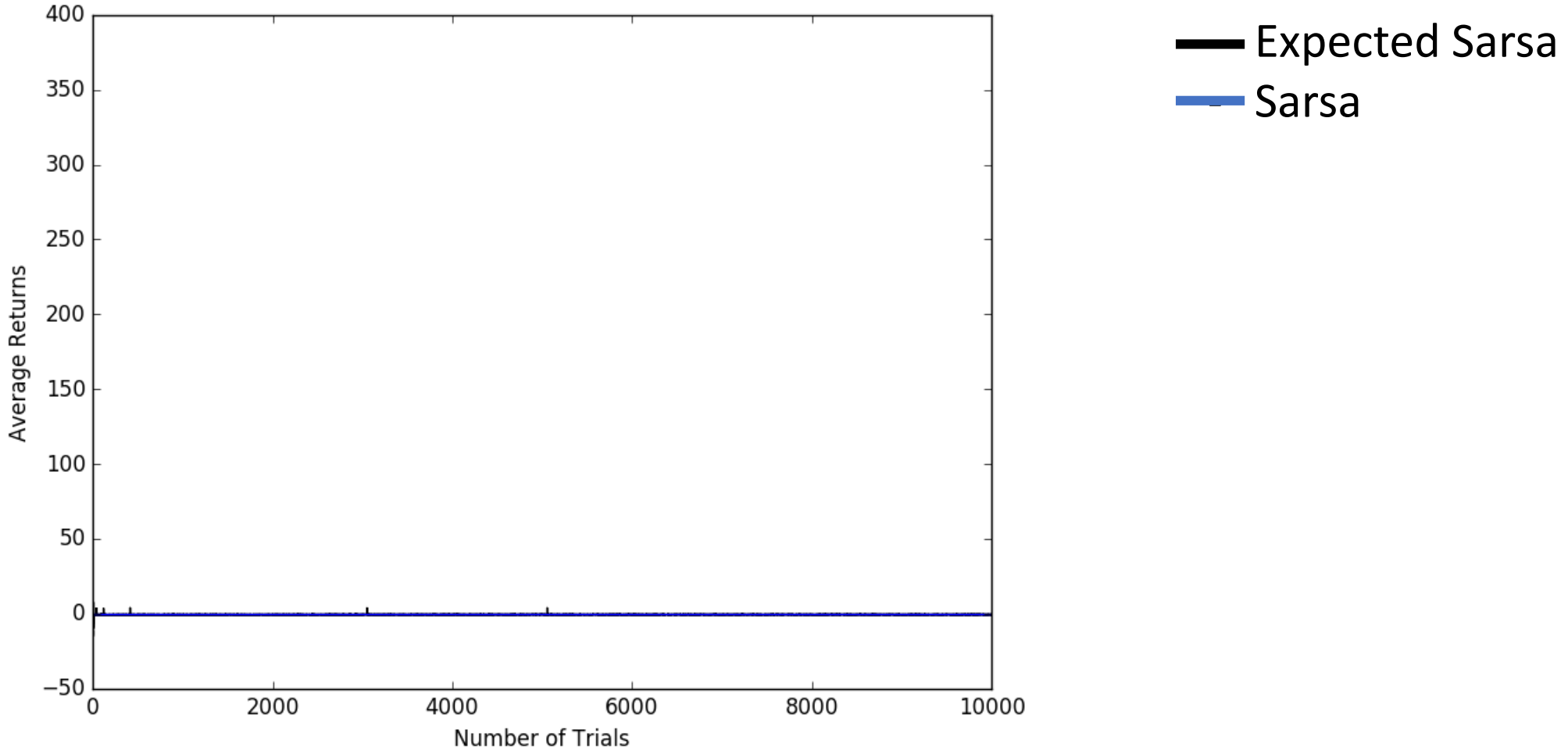
1) Large difference between Q-values of different actions
2) **Lots of exploration**

“Treasure” World



- Each time-step reward is -1
- Finding a “treasure” ends the episode
- Exploration is required to find the farthest episode

Results (1) (0.1-greedy)



Results (2) (0.5-greedy)