# The Theoretical and Empirical Analysis of Expected Sarsa

Monica Patel (260728093)

10-Feb-2017

McGill University

# Synopsis of the Paper

- Theoretical and empirical analysis of SARSA and EXPECTED SARSA
- Understanding Convergence conditions of SARSA and EXPECTED SARSA
- Variance analysis of SARSA and EXPECTED SARSA
- Understanding how EXPECTED SARSA exploits knowledge of stochasticity in behavior of policy to perform updates with lower variance.

# Off-Policy and On-Policy methods

There are two types of TD methods using which Optimal policy for a MDP can be learned

- Off-Policy TD methods like Q-learning: Where Behavior policy controlling the agent is different from the policy agent is trying to learn (estimation policy)
  - Advantage of such method is that it allows sufficient exploration to ensure sufficient data gathering while learning.
- On-Policy TD method: Where agent behaves using same policy which it is trying to learn. (Behavioral and estimation policies are same)
  - These methods provide stronger convergence properties and also have potential advantage over off-policy methods in on-line tasks, since same policy is used to decide behavior.

# Understanding updates of TD methods

TD methods seek the optimal action value function Q*(s,a) from which optimal policy can be deduced. Q*(s,a) is found by iteratively updating Q(s,a).
Update rules for off-policy, SARSA and expected SARSA are as follows:

- Off-policy:

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

- SARSA:

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

- Expected SARSA:

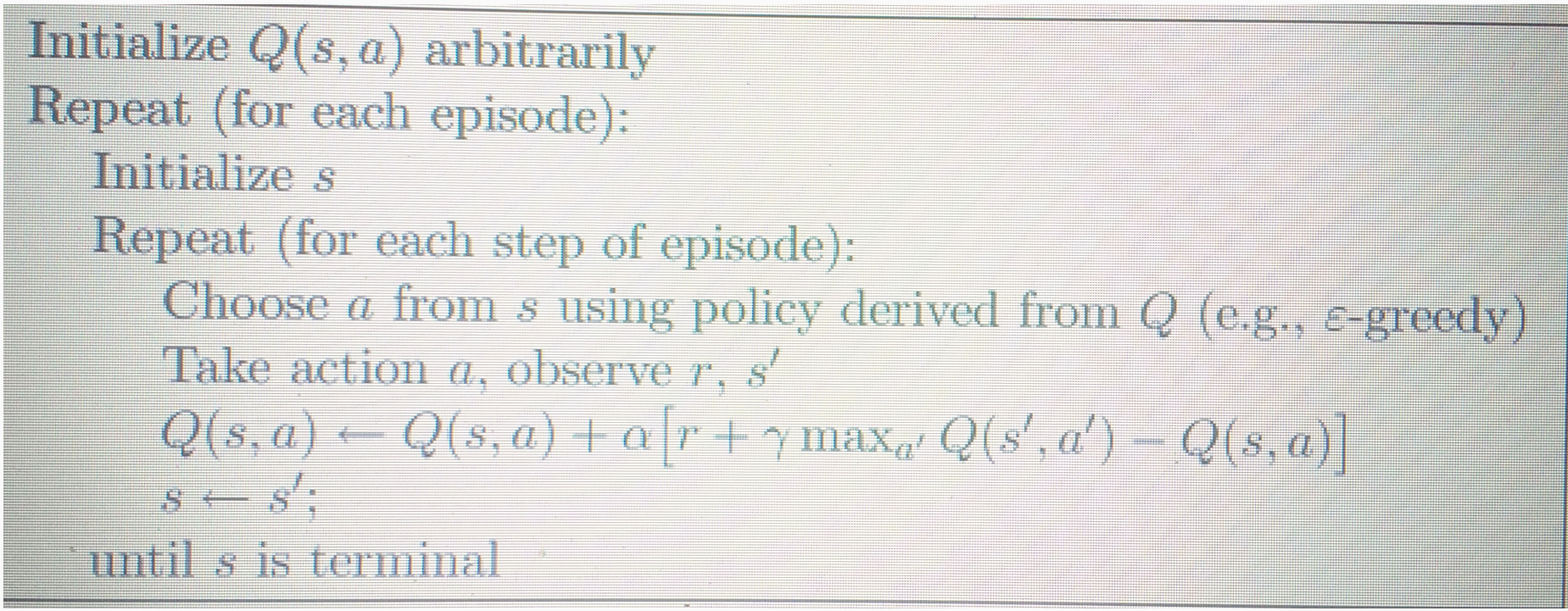
$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \sum_a \pi(s_{t+1}, a) Q(s_{t+1}, a) - Q(s_t, a_t)]$$

# Source of Variance in SARSA

- From the update rule on previous slide we can see that SARSA will not converge to Q* till exploration phase is done, and that its convergence guarantee requires that each state is visited infinitely often.
- Due to this behavior of policy should be stochastic to allow sufficient exploration.
- This introduces substantial variance in updates of SARSA
- There are two main source of variance in such method of update:
  - Environment is stochastic and not known. This source of stochasticity cannot be avoided.
  - Another source of variance is stochasticity in agent's policy which is known to agent.
- Knowledge about stochasticity in agent's policy is exploited by Expected SARSA updates, which can be seen from update rules above.

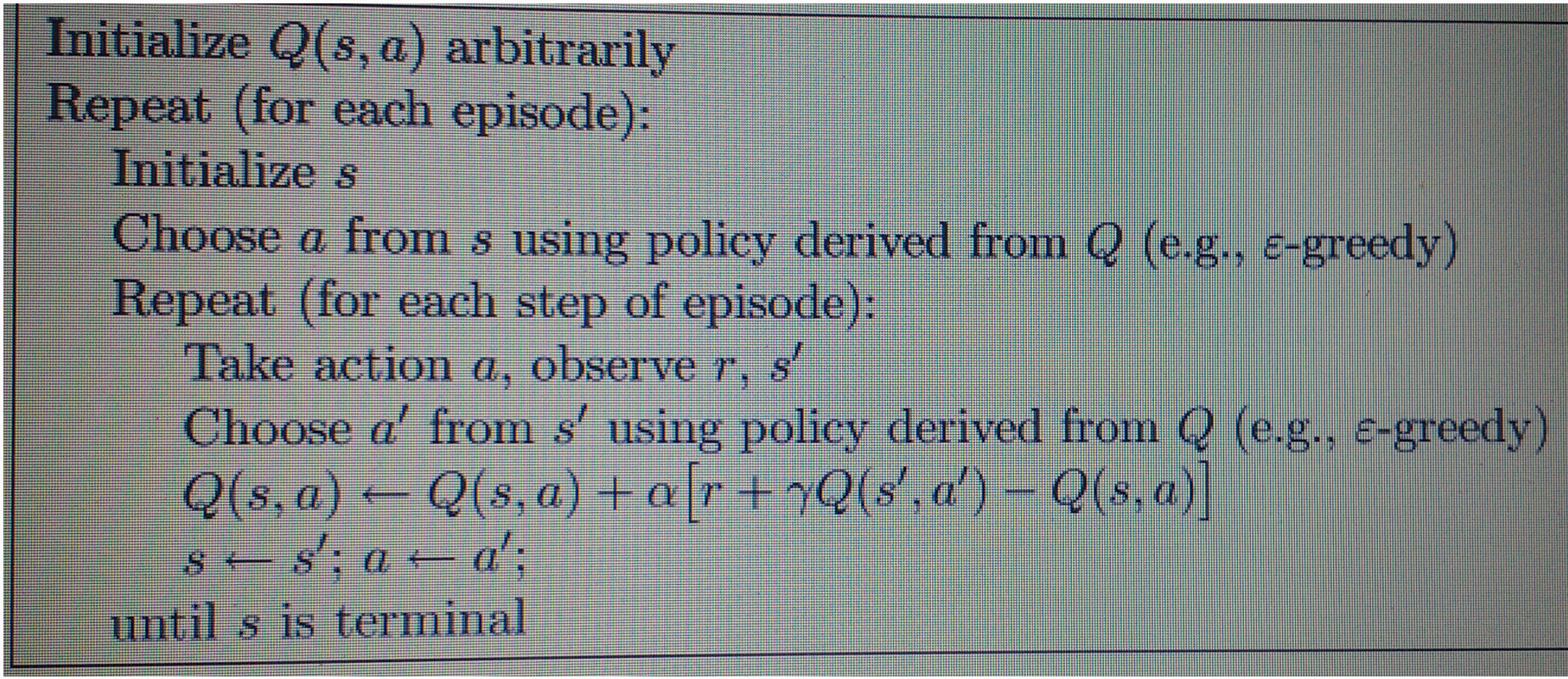
# Q Learning Algorithm

Q Learning algorithm in image below



# SARSA Algorithm

SARSA Algorithm in image below



# Expected SARSA Algorithm

Expected SARSA Algorithm in image below:

**Algorithm 1** Expected Sarsa

Initialize $Q(s, a)$ arbitrarily for all $s$,$a$
**loop** {over episodes}
  Initialize s
  **repeat** {for each step in the episode}
    choose $a$ from $s$ using policy $\pi$ derived from Q
    take action $a$, observe $r$ and $s'$
    $V_{s'} = \sum_a \pi(s', a) \cdot Q(s', a)$
    $Q(s, a) \leftarrow Q(s, a) + \alpha \left[ r + \gamma V_{s'} - Q(s, a) \right]$
    $s \leftarrow s'$
  **until** $s$ is terminal
**end loop**

# Convergence Analysis

Expected SARSA as defined in update rule in above slide converges to the optimal action value Q*(s,a) if following assumptions hold true:

1. S and A are finite
2. The following:

$$\alpha_t(s_t, a_t) \in [0, 1], \quad \sum_t \alpha_t(s_t, a_t) = \infty \quad \sum_t (\alpha_t(s_t, a_t))^2 < \infty \quad w.p.1$$
$$and \quad \forall (s, a) \neq (s_t, a_t), \quad \alpha_t(s_t, a_t) = 0$$

3. Policy is greedy in limit of exploration (Sufficient Exploration allowed)
4. Reward is bounded.

The Theorem is proved using same Lemma used for convergence analysis of SARSA. It shows that SARSA and expected SARSA converge with same conditions.

# Variance Analysis

The target for variance analysis for SARSA and expected SARSA are taken as follows respectively:

$$v_t = r_t + \gamma Q_t(s_{t+1}, a_{t+1})$$

$$\overline{v_t} = r_t + \gamma \sum_a \pi(s_{t+1}, a) Q_t(s_{t+1}, a_t)$$

$$Bias(s, a) = Q^{\pi}(s, a) - E(X), x \text{ is } v \text{ or } \overline{v}$$

*But the Bias for SARSA and Expected SARSA is same*

*So the difference between variance reduces to expression of form*

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

*Unbiased estimate of variance of this can be given by*

$$\sum_i w_i (x_i - \overline{x})^2 \div 1 - \sum_i w_i^2$$

*Looking at numerator*

$$\sum_i w_i x_i^2 - \overline{x}^2$$

# Variance Analysis Continued

If in the above analysis w corresponds to pi and x corresponds to Q values, We can see that, more xi deviate from weighted mean, variance will be larger. That is when there is much exploration and large difference between Q values of different actions.

# Proposed Hypotheses

The paper propose two hypotheses for two different type of problem:

1. Problem where the optimal e-soft policy is better than the e-soft policy based on Q*(s,a)
2. Problem where the optimal e-soft policy is equal to the e-soft policy based on Q*(s,a)

States Hypotheses are as follows:

1. Expected SARSA will out perform Q learning method in Type 1 problem.
2. Expected SARSA will out perform SARSA in both Type 1 and Type 2 problem. Size of performance difference will depend on what type of stochasticity is high, if environment stochasticity is high, difference will be small but if policy stochasticity is high difference will be large.

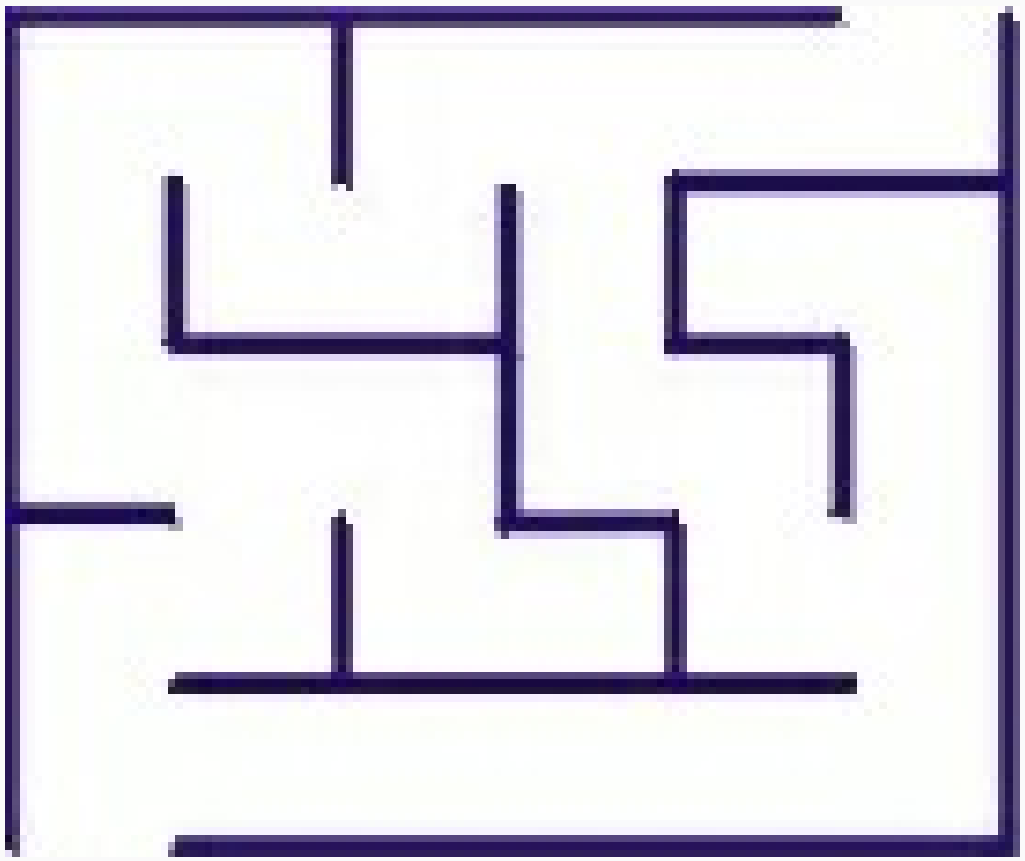
# MAZE GRID WORLD

In order to verify above given hypotheses both SARSA and expected SARSA were tested on the maze grid world with following parameters.

- reward: -2 for hitting wall (if it hits the wall it stays in same state), -0.1 for unoccupied grid cells, 100 for goal states
- Stochasticity of environment: moves agent in random direction with probability 10%
- Policy is epsilon greedy with e $= 0.05$
- Discount factor $= 0.997$
- Learning rate $= 0.24$ for SARSA, 0.27 for expected SARSA and 0.28 for Q-learning.
- Episode is completed when goal is reached or after taking 10,000 steps.

# Maze Grid World Picture

Start at lower left corner and goal at upper right corner.



# Results

We can see that Expected SARSA out performs SARSA and Q learning performs almost equal to expected SARSA. The Following is obtained over only 30 episodes, if data is obtained over more number of episodes we can see the convergence of the algorithms. Following Figure 14 average reward over per episode.