# Review-38

## A Theoretical and Empirical Analysis of Expected Sarsa

Temporal Difference (TD) methods such as Sarsa and Q-learning present suffer at the cost of high stochasticity ans approximation errors in their estimates. In light of model-free Reinforcement Learning (RL) algorithms, the work explores the suitable alternative of Expected Sarsa. Contrary to Sarsa and Q-learning, Expected Sarsa exploits knowledge about stochasticity in the behavior policy to perform updates with lower variance. This allows the method to scale towards higher learning rates and speed up learning. Under similar settings of Sarsa, Expected Sarsa converges and even outperforms prior TD learning methods based on specific hypotheses. Experiments carried out on various domains validate the outcomes of analysis.

Offline planning techniques such as dynamic programming have been pivotal in formulating TD methods. While Sarsa and Q-learning demonstrate suitable performance, state-based $Q$-value estimates in these methods yield high variance and slower learning. Expected Sarsa stabilizes learning of estimates by computing expectations of $Q$-values under the agent's current policy. Similar to Sarsa, Expected Sarsa adopts the on-policy learning framework and effectively explores while executing its policy. In comparison to Q-learning, the method reduces approximation errors by doing away with the greedy selection of agent's actions during its updates. Theoretically, Expected Sarsa demonstrates suitable convergence properties towards the optimal value function under similar settings of Sarsa. This is shown by proving that agent's policy is greedy in the limit of infinite exploration as a result of the contraction mapping formed from the stochastic approximation when $\gamma < 1$.

Empirically, Expected Sarsa presents monotonically increasing performance for higher values of learning rate on the cliff-walking and windy-gridworld tasks. These validate the proposed hypotheses under which the method would perform better in comparison to Sarsa and Q-learning. Furthermore, Expected Sarsa presents equivalent performance to Q-learning under environment and policy stochasticity scenarios when Sarsa tends to diverge at higher learning rates. However, in other domains such as the Maze task, Expected Sarsa and Sarsa present a sudden drop in performance when combined with function approximation. This drop in the case of Sarsa is justified as a result of all estimates accumulating similar values during a start state. However, the question arises that since Sarsa conducts on-policy learning, action-value estimates are updated concurrently and a performance drop should not be observed as a result of simultaneous exploration. Thus, sub-optimal policy selection in Sarsa requires a more reasonable justification with respect to its average performance. Moreover, similar drop in performance in the case of Expected Sarsa is not justified irrespective of its knowledge over policy actions.

Expected Sarsa demonstrates suitable convergence guarantees in comparison to Sarsa and empirical performance equivalent to Q-learning. The TD scheme can be further extended towards backward-view updates such as TD($\lambda$) which would throw light on future variants of Sarsa. Furthermore, the scheme can be stabilized along the lines of double Q-learning by keeping track of dual value estimates in case of function approximation.

TD methods such as Sarsa and Q-learning present high variance and slower learning as a result of stochastic $Q$-value estimates. In light of these limitations, the work explores Expected Sarsa, a stable variant of Sarsa which utilizes knowledge about stochasticity of actions by weighing them under the current policy. Expected Sarsa, similar to Sarsa, presents suitable convergence guarantees in the limit of infinite exploration. The scheme demonstrates empirical performance equivalent to Q-learning on a variety of evaluated tasks.