**COMP: 767 - Reinforcement Learning**
Assignment 2
Problem 2(b)

---

# **Paper**: Reinforcement Learning with Function Approximation Converges to a Region

Samin Yeasar Arnob
ID: 260800927

## Summary

In the paper trajectory based version of SARSA is considered. Where in general, for a trajectory $\alpha$ and policy $\pi$ are fixed. At the begining of each trajectory it selects it's $\epsilon$-greedy policy for it's current Q function and keep it fixed along the trajectory and it's important factor to consider as it plays a role in convergence as in this paper author is considering convergence for trajectory. As mentioned in the [9] convergence doesn't depend whether we're doing a per step update of our weights or after one trajectory as long as learning rate $\alpha$ decays to zero and policy is fixed along the trajectory.
But in the paper it's proved that for the weights of the approximation of the value function to converge to an optimal the policy need to be fixed when we do updates of the weights.
SARSA(0) does TD(0) update for sample policy. TD(0) update is 2-norm contraction and thus converge to a fixed point if applied repeatedly. Considering the convergence in a bounded region in the paper SARSA update rule described as "contraction mapping plus bounded amount of slop"
The conclusion of the paper can be discussed in simple manner if we consider approximating value function similar to projecting on a hyper-plane. Wher we start with a random weight $w$ and want to converge to optimal $w^*$. And as we update our weights we converge to $V^*$. But in this case the oriantation of the hyper-plane also depends on the policy $\pi$. Everytime we update our value-function we update our policy. Policy controls our state visitaion frequency $d^\pi$ and thus the representation of out hyperplane depends on $V^\pi$ and $d^\pi$. So we need to constraint our update of policy over value function updates such a way that while we're still updating our value function, the plane oriantation doesn't change much or value funciton update will fall into a loop.

### 0.1 Problem with SARSA

Gordon (2000) SARSA linear FA converges to a region it means algorithm converges subset of policies and jump around that policies and doesn't converge to any optimal one. Switch between those policies and Gordon referred this as chattering. If's possible for the subset to contain both best and worst policy and thus we can see fluctuation in performance (but still not as bad as Qlearning divergence)

## 0.2 Ways to solve the issue

We can put a constraint (example- KL divergence) to control how much we allow for policy to change. We also can work on proper gradient update and exploration for proper policy update.

## 0.3 Compare with Qlearning divergence

From the similaraity in SARSA(0) and Qlearning it can assumed that both should have same convergence gaurentees. lemann biard: Qlearning with function approximation diverge even when states are updated fixed update policy. But this paper shows, even though Qlearning diverges to an exploratory policy, SARSA converges to a bounded region (set of policies).
The reason behind this is update rule:

$$\begin{aligned} w \leftarrow & w + \alpha * [U_t - \hat{q}(s_t a_t; w)] \bigtriangledown_w \hat{q}(s_t, a_t; w) \\ & SARSA : R_{t+1} + \gamma \hat{q}(s_{t+1} a_{t+1}, w) \\ & Qlearning : R_{t+1} + \gamma max_a \hat{q}(s_{t+1} a, w) \end{aligned} \tag{1}$$

When acting greedy with respect current Q function, SARSA takes next action considering current policy. Whereas, Q learning takes actions that gives maximum Q value. Most of the cases both of these algothims gives similar returns but we find discrepancy when we take exploratory action in these two algorithms.
Moreover, we have seen one of the crusial assumption for converging is to keep policy fixed while updating weights and for Q-learning that will note be possible as it's $max_a$ operation doesn't follow current policy.

## 0.4 Proof discussion

### 0.4.1 Lischitz Continous

$f : R^n \rightarrow R^n$ is called locally Lipchitz at $u_0 \in R^n$ if there exists a neighbour $B(u_0, C)$ with $C > 0$ and constant $K > 0$ such that

$$||f(u_1) - f(u_2)|| \leq K||u_1 - u_2|| \forall u_1, u_2 \in B(u_0, C)$$

Where $C$ is the radius of the bound and $u_0$ Another way to say this if f(u) is lipchitz continuous then u is bounded by a region.

### 0.4.2 Lemma

**Lemma 1**: $J(w_t) \rightarrow J^*$ with probability 1
Let the $J$ be a differentiable function, bounded below by $J*$ and let $\bigtriangledown J$ lipschitz continous
$w_{t+1} = w_t - \alpha_t * s_t$ [1] [1] is decreasing as
**Lemma 2**: Update rule can be written as:

$$w \leftarrow w - \alpha * [A_t^\pi * w_t - R_{t+1}^\pi - \epsilon_t][2] \tag{2}$$

**Lemma 3**: $\bigtriangledown_w H$ is lipchitsz continous then $w$ are bounded bt region

### 0.4.3 Discussing Theorem

In theorem $J(w)$ considered as non-negative potential function decreases on each update as long as $\alpha_t$ is small enough. And $J(w)$ start out large enough compared to $\alpha_t$ and $J(w) = ||w - w^\pi||$
In equation [2] $A_t$ and $R_{t+1}$ depends on policy, thus to ensure $J(w)$ term is decreasing we need to keep the $s_t = [A_t * w_t - R_{t+1} - \epsilon_t]$ value bounded and thus keeping the policy fixed. Thus it's mentioned in the paper "To apply Lemma 1 under the assumption that we keep policy constant rather than changing it begining of each trajectory"
**Lemma 2**, shows $A_t$ is positive defnite
**Lemma 1**, if $s_t$ is positive then $w_{t+1} = w_t - \alpha_t * s_t$ is decreasing and that's one of the condition for $J \rightarrow J^*$ In the proof it shows for $s_t = [A_t^\pi * w_t - R_{t+1}^\pi - \epsilon_t]$ and $J(w) = |||w - w_\pi||^2$, $\mathbf{E}(\bigtriangledown J(w_t)^T s_t | w_t) \geq 2\rho J(w_t)$
Here $J(w_t)$ is a nonnegative potential function and $\rho$ eigen value of $A$ (from lemma2) which is positive definite and thus makes $-s_t$ descending. According to lemma 1 that's one of the prior assumption to $J$ converging to $J*$. When $J(w)$ converges to $J(w*)$, so does the weights, $w$
Now to consider multiple policies , author consider a region bounded with region $C$ where arbitrary weights are $u$ centred around $w$ In similar to $J(w)$ author defines $H(w) = max(0, ||w - u|| - C)^2$
Within the circle $C > ||w - u||$ and thus ,

$$H(w) = max(0, -ve)^2 = 0 \bigtriangledown H = 0$$

Outside the Circle $C < ||w - u||$

$$H(w) = max(0, +ve)^2 = 0$$
$$\bigtriangledown H = +ve$$

and thus,

$$\bigtriangledown H(w_t)^T \mathbf{E}(\bigtriangledown s_t | w_t) \geq d(w_t)(\rho ||w_t - w_\pi||^2) - ||w_\pi - u|| \times ||A|| \times ||w_t - w_\pi|| \tag{3}$$

where right term in the ineqality is positive

In **lemma 3** it's proved that $\bigtriangledown H$ is Lipschitz continous Thus whatever we have proved using $J(w)$ can be proved for $H(w)$ but with a weaker conditioned where the convergence is bounded within region. And thus when $H(w)$ converge with probability 1, weight $w$ also convrge to it's optimal with probability 1 in a bounded region and for fixed policy