PA2

Rishabh Samra

April 2019

1.2.1 SARSA

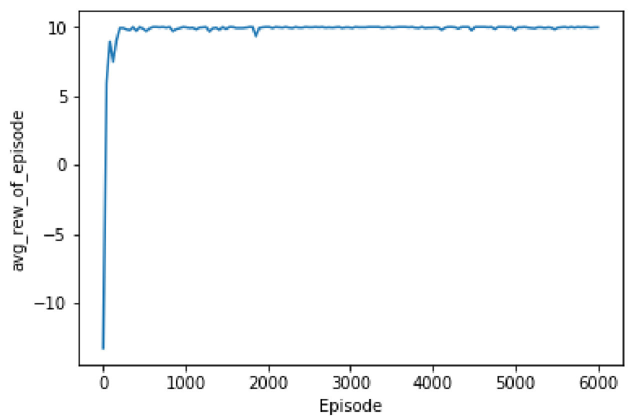


Figure 1: Comparision of Average Rewards for Environment A

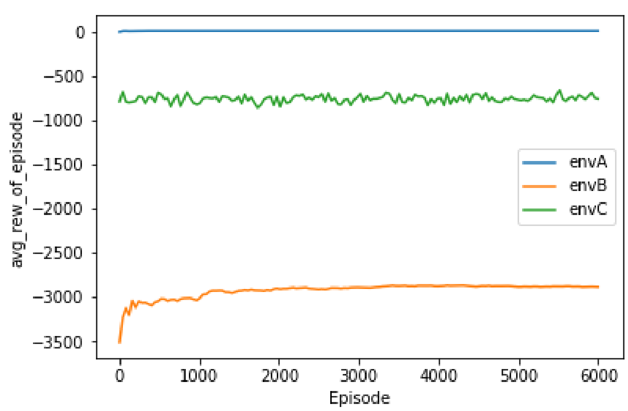


Figure 2: Comparision of Average Rewards

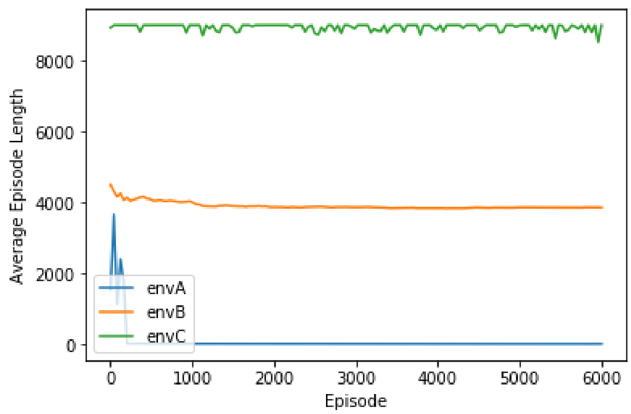


Figure 3: Comparision of Average Episodes

0 = Up ;1 = Right ;2= Down ;3 = Left Optimal Policies:

$$EnvA = \begin{matrix}
2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 3 & 1 & 1 & 1 & 1 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 3 & 0 & 0 & 0 \\
1 & 0 & 0 & 3 & 3 & 0 & 0 & 2 & 0 & 1 & 1 & 0 \\
0 & 3 & 0 & 0 & 0 & 1 & 3 & 1 & 1 & 0 & 2 & 0 \\
0 & 3 & 0 & 3 & 3 & 3 & 1 & 1 & 1 & 0 & 3 & 1 \\
0 & 2 & 3 & 3 & 0 & 2 & 2 & 3 & 2 & 0 & 1 & 3 \\
0 & 3 & 3 & 2 & 2 & 0 & 1 & 3 & 1 & 2 & 0 & 1 \\
0 & 2 & 3 & 2 & 2 & 0 & 1 & 3 & 2 & 0 & 2 & 3 \\
0 & 3 & 3 & 3 & 3 & 2 & 1 & 0 & 2 & 3 & 1 & 3 \\
0 & 3 & 0 & 0 & 3 & 3 & 1 & 3 & 0 & 1 & 0 & 2
\end{matrix}$$

$$EnvB = \begin{matrix}
2 & 0 & 1 & 0 & 3 & 3 & 3 & 2 & 1 & 2 & 2 & 3 \\
3 & 2 & 3 & 2 & 1 & 3 & 3 & 0 & 1 & 0 & 2 & 0 \\
0 & 2 & 2 & 1 & 0 & 0 & 2 & 2 & 2 & 0 & 2 & 0 \\
0 & 0 & 3 & 0 & 0 & 2 & 0 & 2 & 3 & 1 & 1 & 1 \\
3 & 2 & 1 & 1 & 1 & 1 & 3 & 3 & 1 & 1 & 2 & 3 \\
1 & 3 & 1 & 1 & 0 & 1 & 2 & 2 & 2 & 1 & 0 & 1 \\
3 & 0 & 2 & 3 & 1 & 3 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 2 & 0 & 2 & 0 & 0 & 2 & 2 & 2 & 3 & 1 & 2 \\
2 & 2 & 3 & 3 & 0 & 3 & 0 & 1 & 0 & 0 & 3 & 1 \\
3 & 2 & 0 & 1 & 1 & 3 & 1 & 0 & 2 & 1 & 3 & 1 \\
1 & 1 & 3 & 3 & 0 & 0 & 3 & 3 & 2 & 1 & 1 & 0 \\
2 & 3 & 0 & 1 & 2 & 3 & 0 & 0 & 2 & 3 & 1 & 1
\end{matrix}$$

Q.3 SARSA(λ)

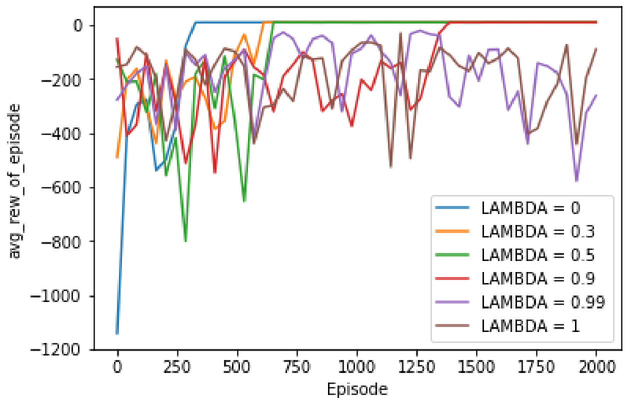


Figure 4: Average Rewards for 2000 episodes

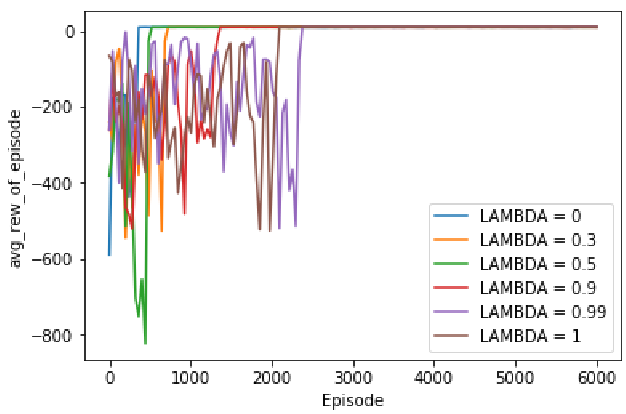


Figure 5: Average Rewards for 6000 episodes

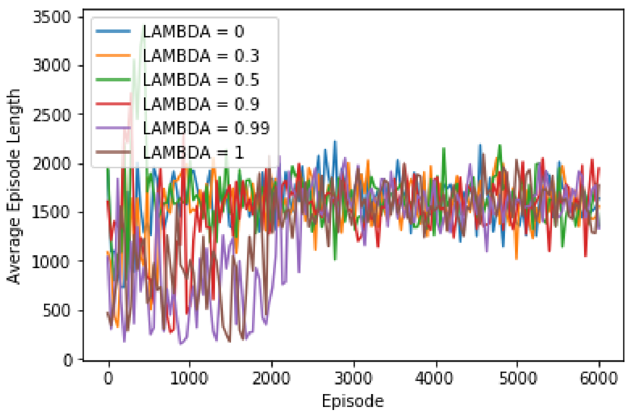


Figure 6: Average Episodes for 6000 episodes

1.3 Policy Gradient

Here preference matrix Theta is defined as:-

$$\begin{bmatrix} \theta_x(N,0) & \theta_x(N,1) & \dots & \theta_x(N,11) & \\ \theta_y(N,0) & \theta_y(N,1) & \dots & \theta_y(N,11) & \\ \theta_x(E,0) & \theta_x(E,1) & \dots & \theta_x(E,11) & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_y(W,0) & \theta_y(W,1) & \dots & \theta_y(W,11) & \end{bmatrix}$$

Update equations for backup is defined as follows:-

$$\theta_{t+1} = \theta_t + \alpha\nabla(\log(\pi(A_t|S_t)$$

$$\pi(A_t|S_t = softmax(\theta_x(A,i)), Here A = N, S, E, W$$

Let $\alpha_i = \theta_x(A,i)$

$$\nabla_{\theta x}(\log(\pi(A_t|S_t))) = \frac{\sum_i e^{\alpha_i}}{e^{\alpha 1}} * \frac{\sum_i e^{\alpha_i} * e^{\alpha 1} - e^{2\alpha 1}}{(\sum_A e^{\alpha_i})^2}$$

$= 1$- (e$^{\alpha 1}\overline{\sum_A e^{\alpha_i})}$

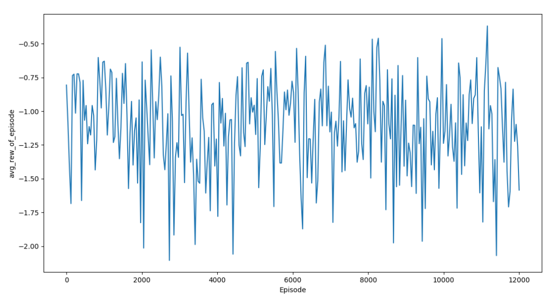
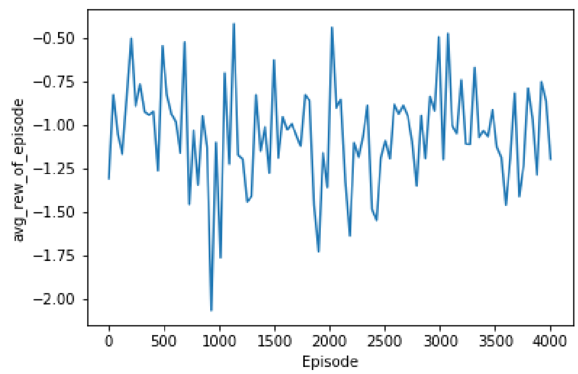
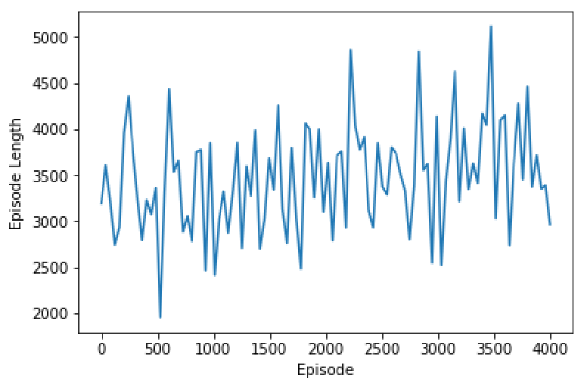


Figure 7: Average Rewards for 12000 episodes

Learning Rate = 0.01

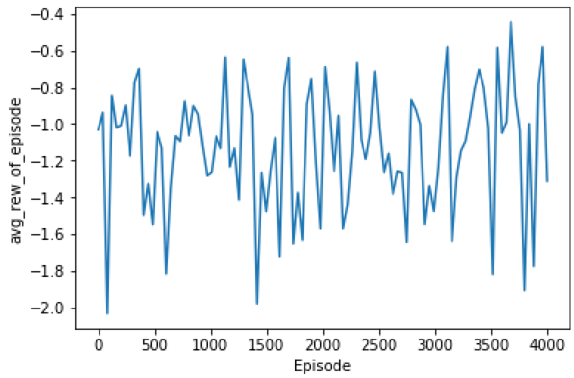


(a) Average rewards

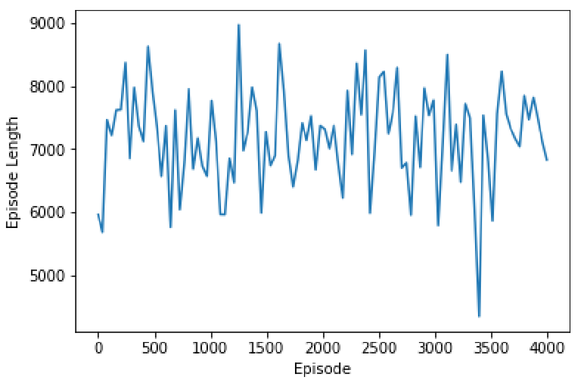


(b) Average episode length

Learning Rate = 0.05

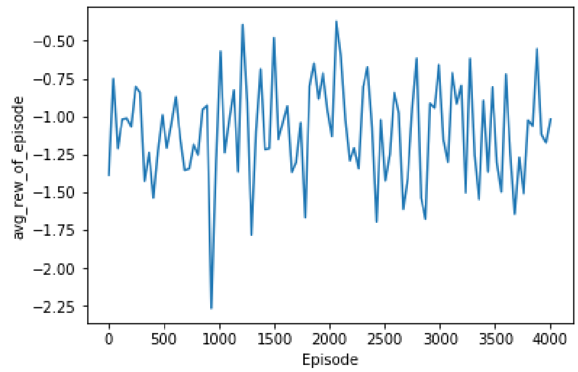


(a) Average rewards

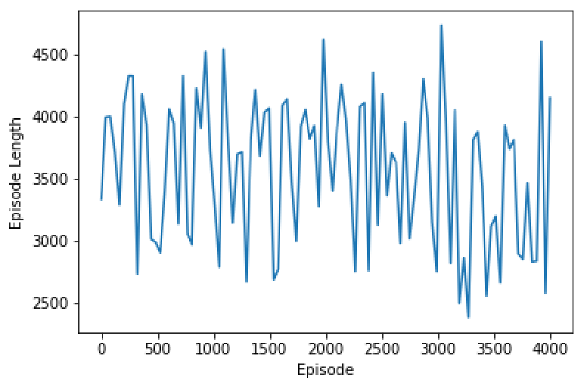


(b) Average episode length

Learning Rate = 0.001



(a) Average rewards



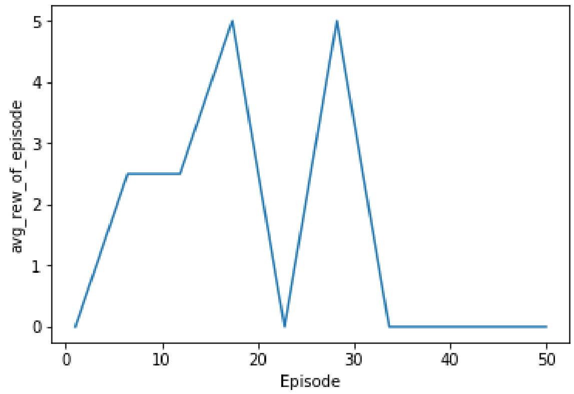
(b) Average episode length

Ans:-Policy parameterization is preferred over value function in problems involving stochastic policies since they are better solved by policy iteration.It is better to introduce some stochasticity because while using value based method if we act greedily every time it would take a lot of time to reach destination. **Note:If the program had ran for more episodes the plot could have been converged but due to lack of time at last moment could not ran it for more.The last policy I got is as follows:**

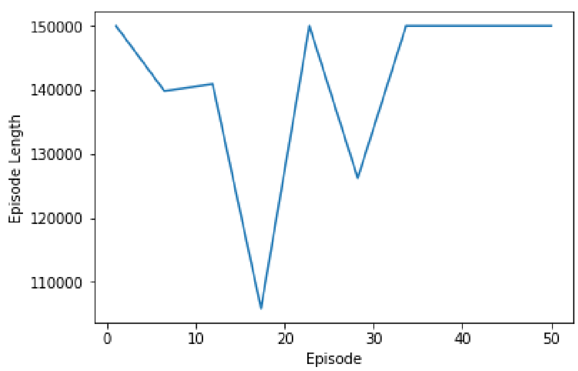
$$OptimalPolicy = \left[\begin{matrix} 1 & 1 & 3 & 0 & 0 & 1 & 1 & 1 & 3 & 1 & 1 & 1 \\ 3 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 3 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\ 3 & 2 & 3 & 0 & 3 & 0 & 3 & 2 & 3 & 2 & 1 & 3 \\ 3 & 2 & 3 & 0 & 2 & 3 & 2 & 2 & 3 & 2 & 1 & 3 \\ 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 2 & 1 & 1 \\ 3 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 \\ 3 & 2 & 3 & 0 & 0 & 0 & 0 & 2 & 3 & 0 & 1 & 3 \\ 1 & 1 & 3 & 0 & 0 & 0 & 0 & 2 & 3 & 0 & 1 & 3 \\ 3 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 & 2 & 2 & 3 \\ 3 & 2 & 3 & 0 & 2 & 3 & 2 & 2 & 3 & 2 & 1 & 3 \end{matrix}\right.$$

1.4 Function Approximation

SARSA

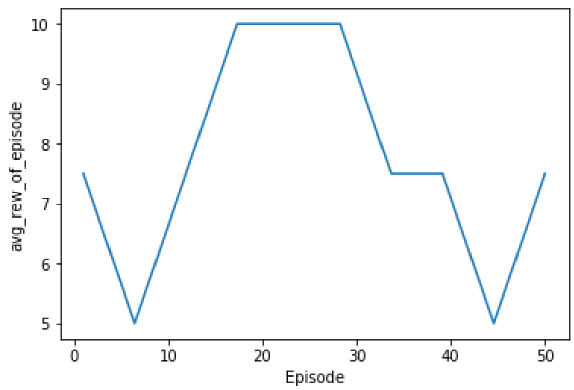


(a) Average rewards

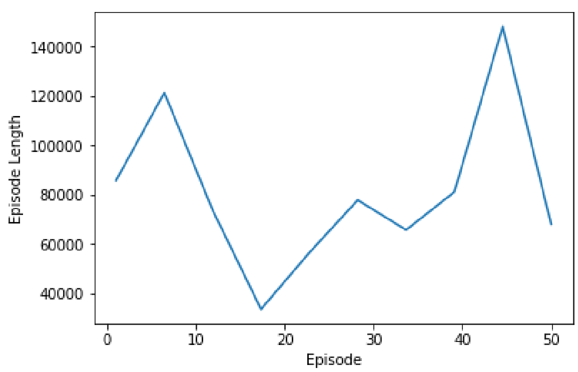


(a) Episode Length

Lambda = 0.3



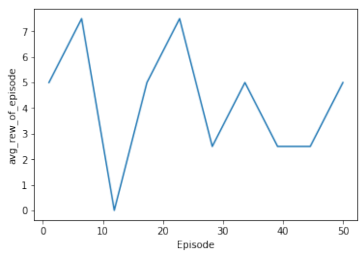
(a) Average rewards



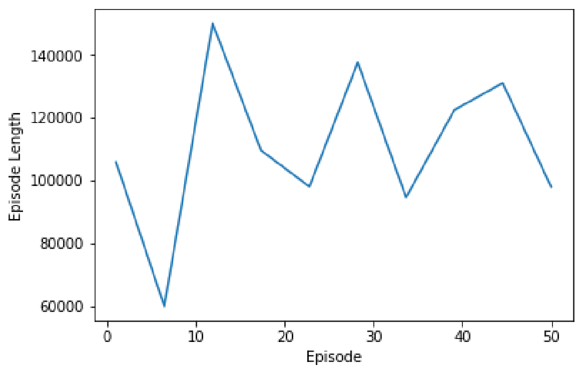
(b) Average episode length

$$OptimalPolicy: \begin{bmatrix} 1 & 2 \\ 2 & 2 \\ 0 & 0 \\ 3 & 0 \\ 0 & 0 \\ 2 & 2 \\ 3 & 2 \\ 2 & 0 \\ 0 & 0 \\ 3 & 2 \\ 1 & 0 \\ 1 & 2 \\ 0 & 0 \\ 2 & 1 \\ 0 & 2 \\ 1 & 0 \\ 3 & 3 \\ 2 & 3 \\ 3 & 0 \\ 2 & 0 \end{bmatrix}$$

Lambda = 0.9



(a) Average rewards



(b) Average episode length