**Q1. Plot the converged policy and value function for this grid world.**

***ipython solution link*** -

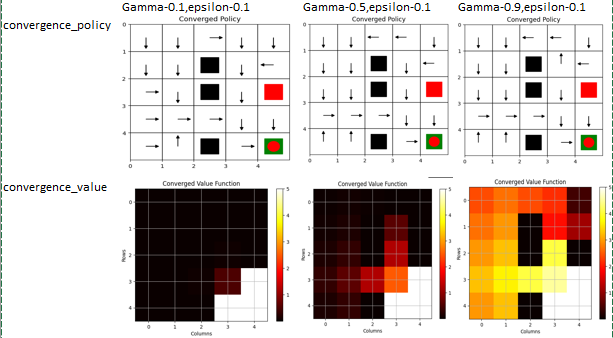
<https://drive.google.com/file/d/19-g9yx7eh1OO4fVT4kHCbpin3FrtINwb/view?usp=sharing>

**Q2. Do it for Gamma = 0.1, 0.5 and 0.9, take epsilon = 0.1**

The discount factor determines the importance of future rewards. A factor of 0 makes the agent short-sighted by only considering current rewards, while a factor approaching 1 will make it strive for a long-term high reward.

From the below convergence policy fig, at different gamma values depicts -

1. Lower value of gamma, prioritized immediate rewards more than the long term reward.
2. Whereas a larger value of gamma, prioritized long term rewards more than the immediate reward.
3. A lower value of gamma, hinders the model to learn the optimal policy. Whereas a larger value of gamma promotes better optimal policy.
4. Lower gamma leads to - immediate convergence, which might not be better whereas a larger value of gamma promotes gradual convergence.



**Q3. For gamma = 0.9, plot the no. of steps to reach the goal across**

**episodes for epsilon = 0.1, 0.3 and 0.5.**

***Epsilon*** - greedy exploration, where the agent chooses a random action with probability epsilon and the action with the highest Q-value with probability (1 - epsilon). This allows the agent to balance exploration (trying out new actions) and exploitation (choosing the best known action).

Based upon the observations, As the value of epsilon increases -

1. model tries to exploration more
2. take more random walks
3. which inturn increases the number of steps required to converse to the Goal stage
4. which leads to gradual decrease in below graph color shades and doesn’t get stuck at local optima
5. For all the values of epsilon, observe that initially the number of steps to reach the goal is higher and gradually it decreases as initially the model explores more and as the learning increases it exploits more.
6. higher epsilon leads to more steps initially as the agent explores more, but would lead to fewer steps in the long run as the agent learns a more optimal policy.
7. for epsilon 0.1 - Policy Convergence = we see the effect of local maxim, around the terminal(red) block.

| **Epsilon** | 0.1 | 0.5 | 0.9 |
| --- | --- | --- | --- |
| **Exploration (random action)** | least | medium | most |
| **Exploitation (choose highest Q value)** | more | medium | least |
| **Steps to converse** | least | medium | most |
| **rate of convergence** | faster | medium | slower |
| **chances to get stuck at local maxim** | most | medium | least |

