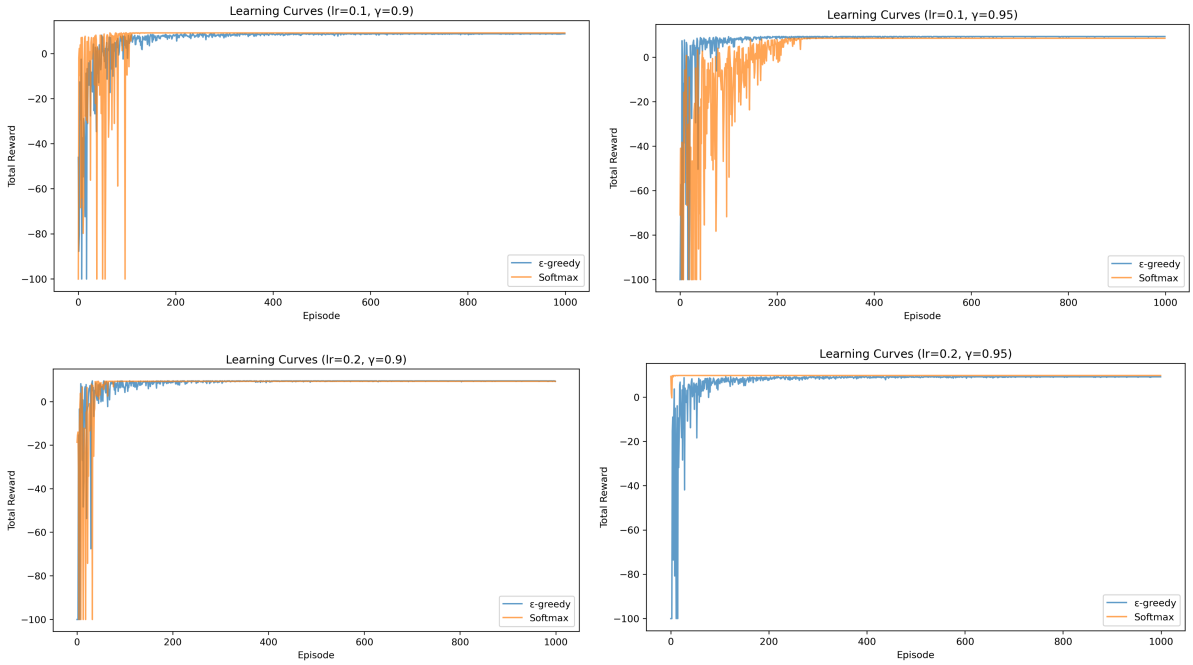

The above plots show the convergence of two exploration methods (Epsilon & Softmax), experimented across 4 runs with different learning rates and discount factors.

The exploration strategy being used influences the learning speed of the agent,policy quality, and stability. In the plots above, higher learning rates like (lr = 0.2) accelerate the learning by allowing a broader exploration earlier, while a higher discount factor (gamma = 0.95) supports long term planning, resulting in better policy quality for the agent. Low gamma values limit foresight, reducing the effectiveness of the policy.
The improved stability and smoother evaluation curves seen at lr=0.2, gamma=0.95 suggest that there is a balance between exploration-exploitation. In contrast, lower learning rates and gamma values slows the learning and produces noisier or suboptimal policies, which indicates inadequate exploration or premature convergence to local optima.