1 Question 1

1.1 a

Action state: $\mathcal{A} = \{\text{hit,stick}\} = \{\text{h,s}\}$
State Action: $\mathcal{S} = \{\text{X,Y}\}$ s.t $X \in \{4, ..., 21, bust\}, Y \in \{2, .., 11\}$ all state s.t $X \in \{bust, 21\}$ are terminal states.
So, in total we have $19 \cdot 10 = 190$ states.
Transition Probabilities:

$$p(s_{t+1}|s_t, a_t) = \begin{cases} 1, & s_t = s_{t+1}, a_t = stick \\ \frac{3}{14}\mathbb{1}\{X_{s_{t+1}} - X_{s_t} = 10\} + \frac{1}{14}\mathbb{1}\{1 \le X_{s_{t+1}} - X_{s_t} < 10\} + \frac{1}{13}\mathbb{1}\{X_{s_{t+1}} - X_{s_t} = 11\}, & X_{s_{t+1}} \in \{6, .., 21\}, X_{s_t} \in \{4, ..., X_{s_{t+1}} - 2\}, a = hit \\ \mathbb{1}\{22 - X_{st} \le 2\} + \frac{3+10-(22-X_{st})}{13}\mathbb{1}\{2 < 22 - X_{st} \le 10\} + \frac{1}{13}\mathbb{1}\{10 < 22 - X_{st} \le 11\}, & X_{s_{t+1}} = bust, X_{s_t} \in \{11, ..., 20\}, a = hit \\ 0 & else \end{cases}$$

Reward:

$$r(s, a) = \begin{cases} 0, & X_s \in \{4, ..., 20\} \\ 1, & X_s = 21 \\ -1, & X_s = bust \\ -1 \cdot p(win|s) + 1 \cdot p(lose|s) + 0 \cdot p(draw|s), & a = stick \end{cases}$$

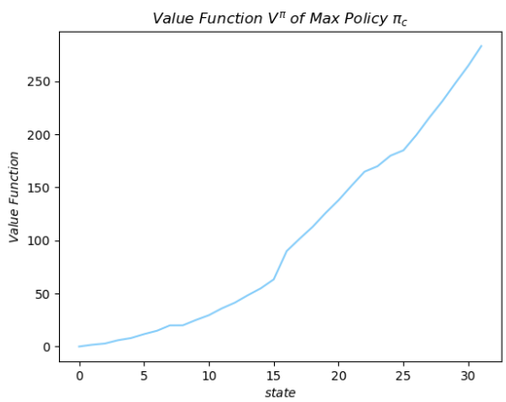
2 Question 2 - part 1

2.1 a

There are 5 actions, and $\sum_{i=1}^{5} \binom{5}{i} = 31$ number of stages adding a terminal stage with no jobs left we have $|S| = 32$

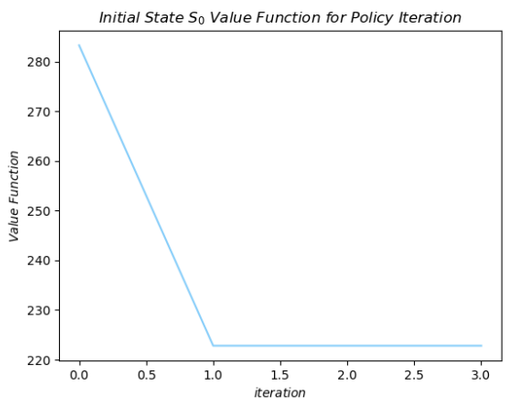
2.2 c

we sorted the values for better visibility.



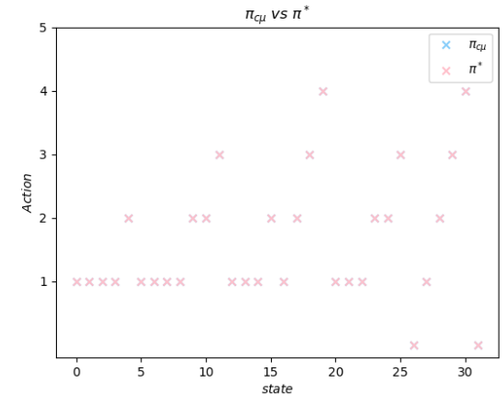
2.3 d

Because the initial policy is close to the optimal policy we converged in 2 iterations:

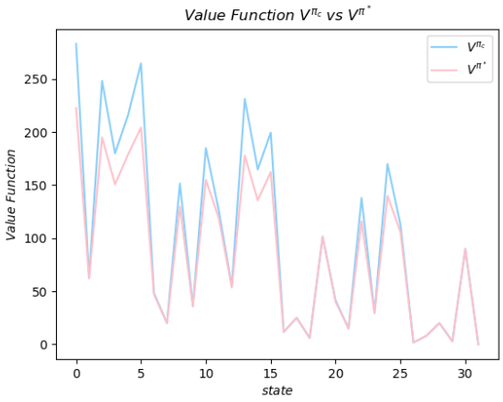


2.4 e

policy comparison:



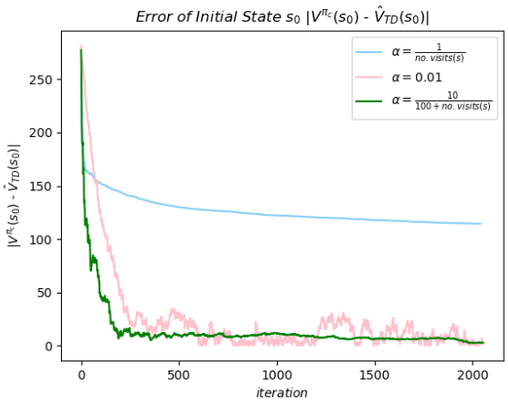
value comparison

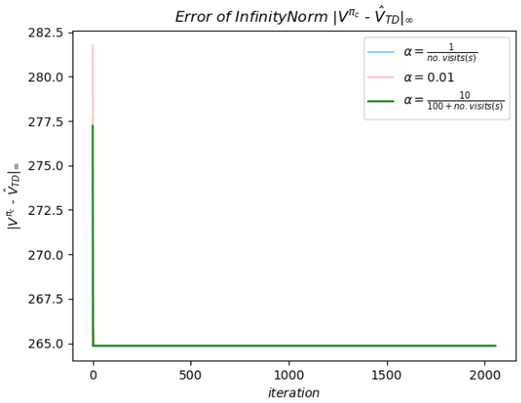


3 Question 2 - part 2

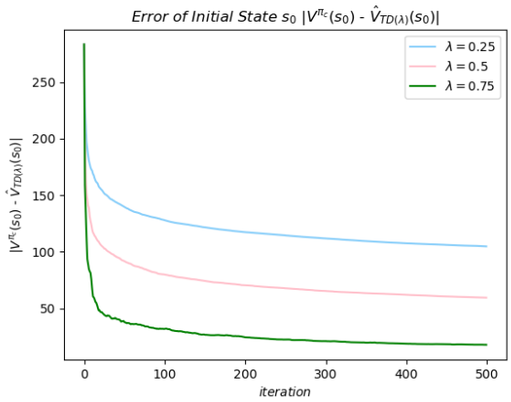
3.1 g

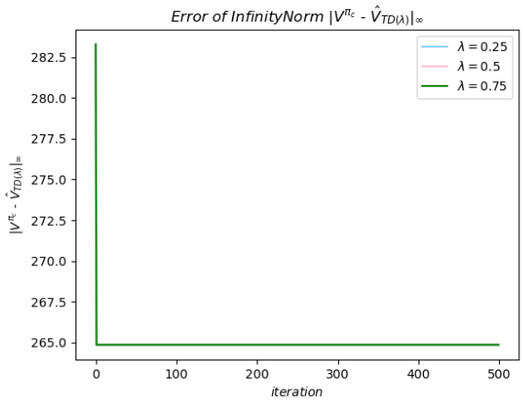
We can see that in all steps sizes the infinity norm does not converge to 0. Probably because some states have low probability then others.
$\alpha_n = \frac{1}{number\ of\ visits}$:
doesnt converge to 0 because the step size is too small, meaning the number of visits is too high.
$\alpha_n = 0.01$:
converges to 0, but in a slow and noisy way.
$\alpha_n = \frac{10}{100+(number\ of\ visits)}$:
converges to 0, relatively fast and smooth.





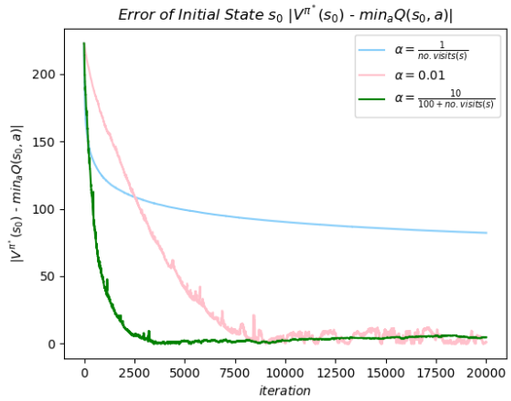
3.2 h

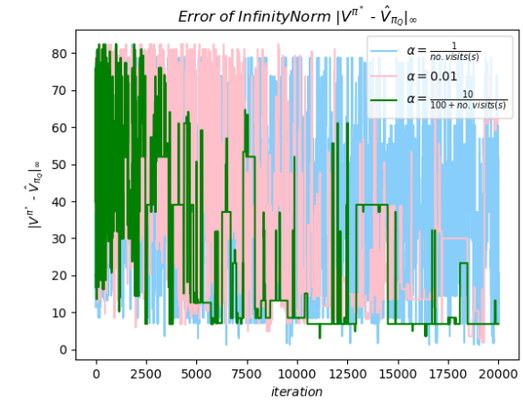




In this section we chose $\alpha_n = \frac{1}{num\ of\ visits}$ because it is smoother. We can see that higher values of λ lead to higher convergence rate.

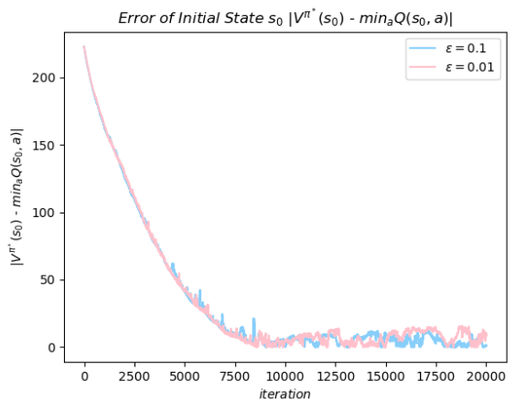
3.3 i

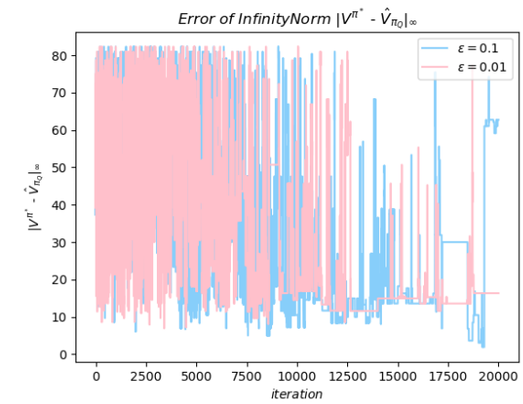




We can see that for $\alpha_n = \frac{1}{num\ of\ visits}$ the Q learning doesnt converge, and $\alpha_n = \frac{10}{100+num\ of\ visits}$ has the highest convergence rate. The infinity norm is quite noisy.

3.4 j





We can see that changing epslion doesnt make much a difference.