ELL729 Coding Assignment 2

Suraj Joshi

December 19, 2020

1 Optimal Values and Optimal Policy for no discounting case

1.1 Assumptions

The problem is formulated as an MDP with states from 0 to 100. The reward is 1 if transitioning from an amount less than 100 to 100 and 0 otherwise. At any state s $<$ 100, the possible investments a $\in$ {0,...,min(s,100-s)}. On investing a, the possible next states are min(s+a,100) with probability p (provided as an input parameter) and max(s-a,0) with probability 1-p. The discounting factor is set to 1 (no discounting) and iterations are run till convergence.The discount factor (α) is set to 1 for no discounting case. The iterations are

1.2 Results

(1) Optimal Value Estimate vs. State for p = 0.4:

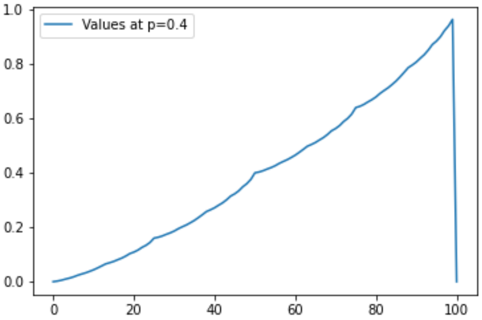


Figure 1: Optimal Value Estimate vs. State for $\alpha = 1, p = 0.4$

(2) Optimal Investment vs. State for p = 0.4:

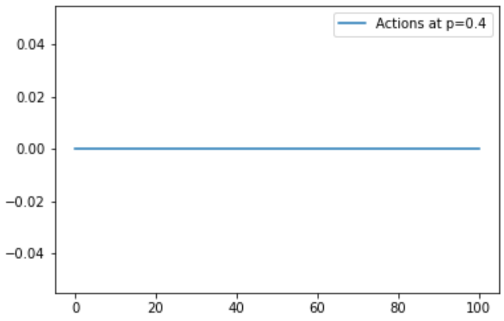


Figure 2: Optimal Investment vs. State for $\alpha = 1, p = 0.4$

- The optimal value increases from 0 to 99. The optimal value at 100 is 0 as starting from 100 the only action is 0 and the reward for repeatedly arriving at 100 is 0 (can also ignore 100 completely as it is not a starting state). The optimal actions are 0 for every stage. This means that though there were some non-zero actions in the first few to reach the optimal values, after the optimal values have converged, 0 is the optimal action for every state.
- Mathematically it descibes the meaning of $\alpha = 1$ which means that rewards are as valuable now as they are in the distant future. Hence taking the action 0 and reaching the state of 100 takes infintely many weeks, but the investment banker is fine with it because all that matters is reaching the state 100 and not how long it takes.

(3) Optimal Value Estimate vs. State for p = 0.25

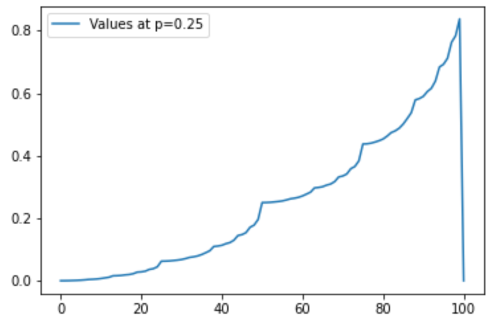


Figure 3: Optimal Value Estimate vs. State for $\alpha = 1, p = 0.25$

(4) Optimal Investment vs. State for p = 0.25

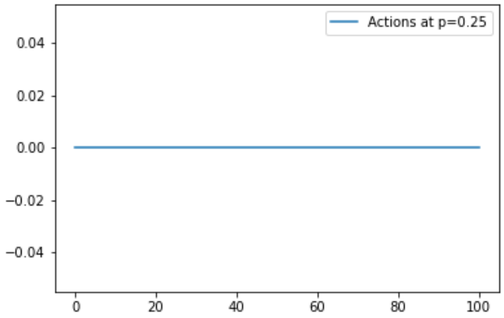


Figure 4: Optimal Investment vs. State for $\alpha = 1, p = 0.25$

- Similar to the above case, except that the values grow in a different manner. There are steeper jumps at the points where the plot for p = 0.4 has smaller jumps. Note that in both the figures, the peaks are at the same states, and that at state 50, the value converges to p itself. This is because at the state 50 one can immediately get 100 (a reward of 1) with probability p and 0 with probability 1-p (a reward of 0) giving the expected value of p. Note that this is from action 50 even though the optimal action is 0. This is because there are multiple optimal actions possible for several states at convergence (will be shown later). These can be displayed by allowing equal values to be overwritten by using >= instead of > to compare rewards or to iterate from the largest investment to the smallest investment.

- Also note the positioning of the jumps i.e. points between regions of high growth and regions of low growth. The largest such jump (or steepest) is at 50. This is because of being able to take the action of 50 as mentioned earlier. The other most noticable jumps occur at points (e.g. 75) where there is an action that leads to the state 100 (and 1 reward) and/or in case of failure leads to another jump point itself allowing another shot at reaching 100. Other smaller jump points are placed recursively in the similar manner. These are much more visible in the case of 0.25 as compared to the plot of optimal actions at p = 0.4, due to the lower probablity of success leading to larger difference between optimal values between the largest jump states and other states.

- Another observable point of difference is how high the highest point of the optimal value curve goes in each case. In the case of p = 0.4, it is very close to 1 while in the case of p = 0.25, it is slightly above 0.8 and more distant from 0.1. This is again due to the probability of success being higher in the first case. On taking further iterations, they will go closer to one, but for the same number of iterations, a greater p value will be closer to 1 than a lower p value.

(5) Optimal Value Estimate vs. State for p = 0.55

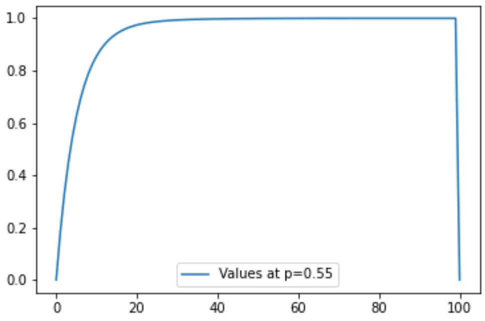


Figure 5: Optimal Value Estimate vs. State for $\alpha = 1, p = 0.55$

(6) Optimal Investment vs. State for p = 0.55

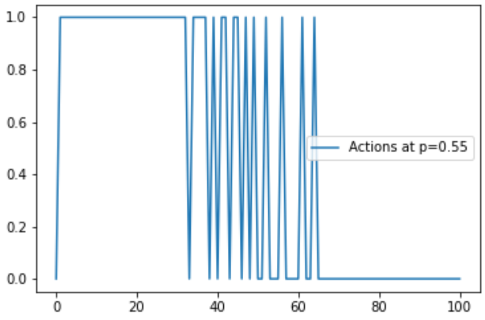
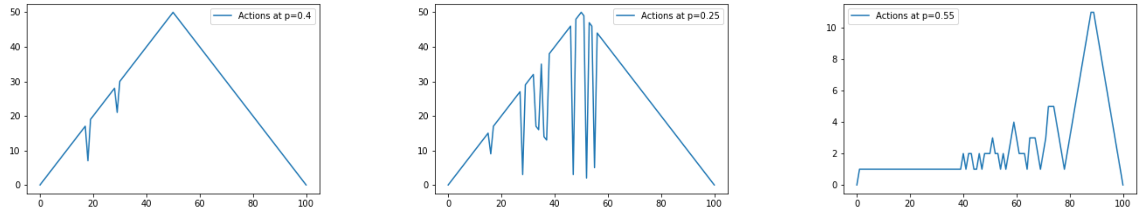


Figure 6: Optimal Investment vs. State for $\alpha = 1, p = 0.55$

Note the difference in the optimal value curve compared to the earlier values of p. The curve climbs to one and almost flattens out, quite unlike the other 2 where the values grow slowly and sublinearly to 1. The reason for this type of increase is because the probability of a successful investment is greater than 0.5 i.e. it is more likely to earn than to lose. Hence the banker values lower states more than he did in the earlier two cases because they no longer require the property of being having an action leading to a jump to be valuable (also the reason for the absence of similar jumps as in the other 2 cases), their associated reward increases on account of being closer to 100. So the ones closest to 100 have the highest (expected) reward, the states below them have slightly lower (expected) rewards due to every higher state having greater rewards and hence every state is followed by a state with a greater optimal value. The optimal actions oscillate between 0 and 1. Although this is common to the earlier cases as well, in this case the rate of growth is higher because it is more likely to win from a lower state as well as from the current state (and this can be extended to any streak of losing). Hence 'losing' is more rewarding on average leading to states closer to 0 having high optimal values boosting the rest of the graph upwards. This is because of incomplete convergence due to either rounding error or limited number of iterations. On increasing to a greater number of iterations we will observe fewer spikes to 1 and more zeros. The optimal actions are again close to 0 because of the same reason as before.

1.3 Other Observations

Plots of Optimal Actions when we allow equal valued actions to overwrite one another from 0 to min(state, 100 - state) for different p values.



(a) (b) (c)

Figure 7: a) Optimal Actions vs. State for p = 0.4, b) Optimal Actions vs. State for p = 0.25, c) Optimal Actions vs. State for p = 0.55 (note the different scale for p = 0.55)

This demonstrates that at a particular state, there are several optimal values and not only 0. However each one of these optimal values gives the same expected reward at each state as 0 does, and hence after iterating over action 0, no other action can overwrite it. Also note that each one of the plots loosely follows optimal action = state for states between 0 to 50, and optimal action = 50 - state for states between 51 and 100 (very loosely for the last case) with some spiky noise.

2 Optimal Values and Optimal Policy, Discount factor = 0.9

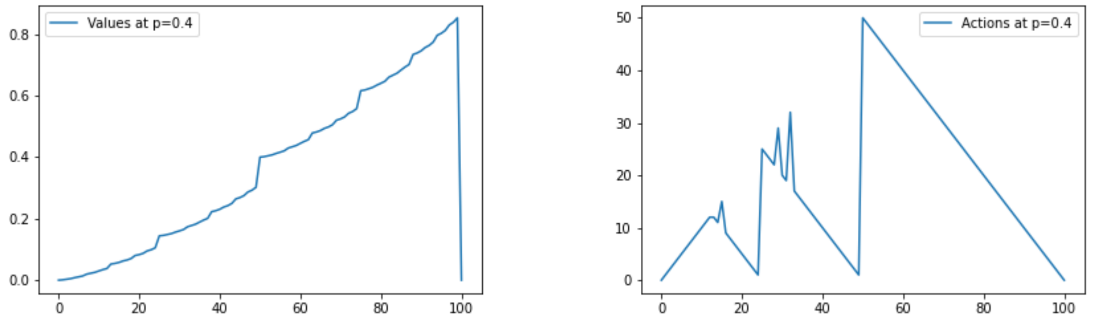
2.1 Assumptions

The MDP model does not change, only this time the discount factor $\alpha = 0.9$. The possible investments at each stage and the reward function is identical.

2.2 Results

Optimal Value and Policy Estimates for p values 0.4, 0.25 and 0.55:

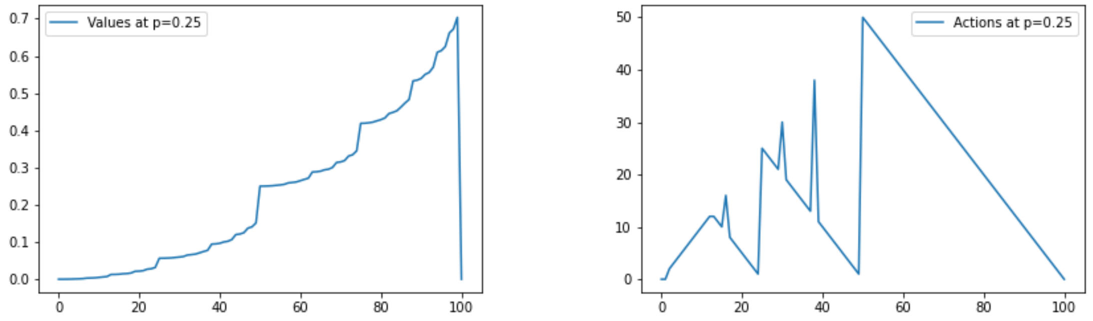
- Here, the optimal values again go from 0 to towards 1. p = 0.25 is again the most far off from 1, followed by p = 0.4 and lastly p = 0.55 which is closest to 1 at state 99. We can again observe significant jumps at some states.
- For the case of p = 0.4 and p = 0.25, we have that the jumps are in the same states as in the no discount case. This is because the logic is the same,



(a)

(b)

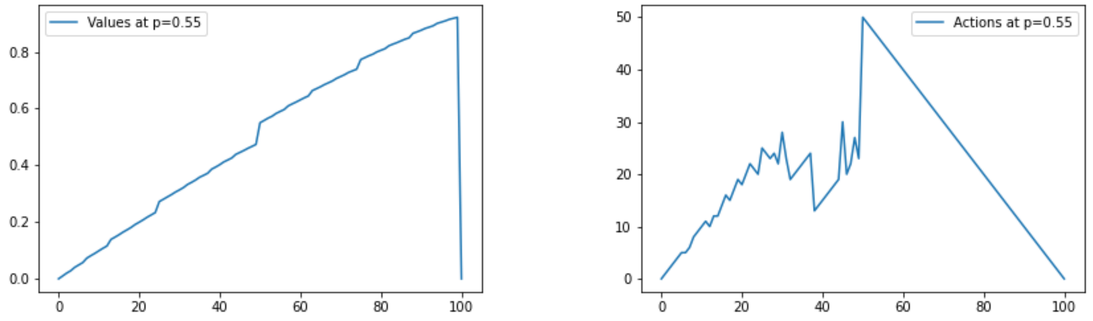
Figure 8: a) Optimal Value Estimate for p = 0.4, b) Optimal Policy Estimate for p = 0.4



(a)

(b)

Figure 9: a) Optimal Value Estimate for p = 0.25, b) Optimal Policy Estimate for p = 0.25



(a)

(b)

Figure 10: a) Optimal Value Estimate for p = 0.55, b) Optimal Policy Estimate for p = 0.55

an action gives value to a state partly because of reward from a successful investment as well as from a failed investment. So 50 has the highest jump from the immediately previous state because the prospect of investing 50 and immediately moving to 100 is more rewarding than the best prospects available at 49. Moreover these are actually reflected in the optimal actions as well because with a discount factor < 1, the time taken to reach 100 affects the banker's reward. Hence the presence of second spike at 75 along with the corresponding optimal value 25, a success immediately gives a reward 1, while a failure gives another shot at achieving the state 100. The explanation for exactly 25 is that for actions below 25, the reward for a successful investment is 0 immediately after (even though we could get 1 reward later from that state) hence even though the states more than 50 allow for actions that lead to 100 in 1 step, they cannot be more valuable than also allowing a possibly immediate gain (due to the discount factor). 25 is the greatest possible investment hence there is no action above 25. These two factors combined explain not only the peaks but also patterns of optimal actions that are observed.

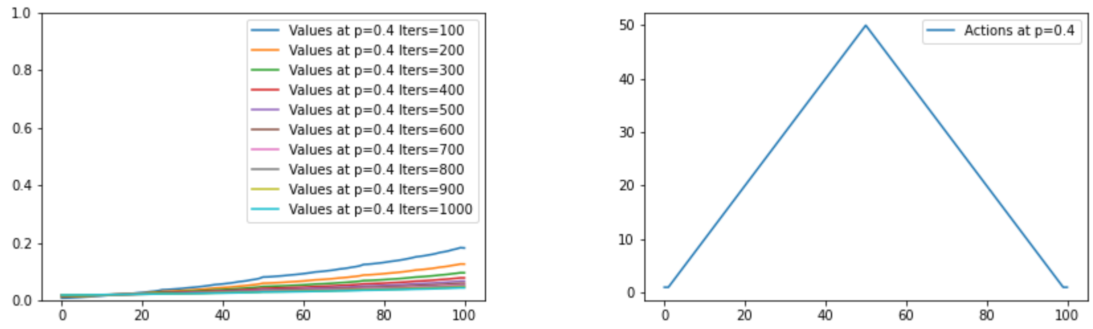
- Note that the general shape of the optimal action curve is quite similar with peaks at 12/13, 25, 50 and downwards slopes before the next peak. These can be explained in a similar way, a peak allows easy access to either the next peak or the rewarding state of 100. For states which are not peaks the goal is to take an action that takes them to the next peak as soon as possible. As the probability is skewed towards failure, they choose to aim towards the peak behind them rather than the one immediately in front of them. Some spikes observed can be due to rounding errors or incomplete convergence, but the general shape shows follows the logic explained.

- For the case of p = 0.55, the difference is that the optimal value curve, although it has similar peaks as the earlier two, is again over (has a higher slope) the straight line joining (0,0) and (100,1) than the other two. Also, the jumps are not followed by slow increases in value (compared to p = 0.25, p = 0.4). Note also that in the optimal actions curve for p = 0.55, one of the valleys is completely eliminated, and the other is much higher. This means that due to expecting a win more often than a loss, the banker is incentivised to take greater risk in more stages than before and aim for the next peak in value instead of playing it safe. The second half of the action curve is again exactly the same straight line as there is no change in strategy here.

3 Optimal Values and Optimal Policies for Average Cost Formulation

3.1 Assumptions

To satisfy the unichain condition for convergence of iterations in average cost, the actions possible at some states is changed. At 0, the next state is 1 with probability 1 (and the action is also set to 1 directly to for a reasonable action) and at 100, the



(a) (b)

Figure 11: a) Optimal Value Estimate for p = 0.4, b) Optimal Policy Estimate for p = 0.4

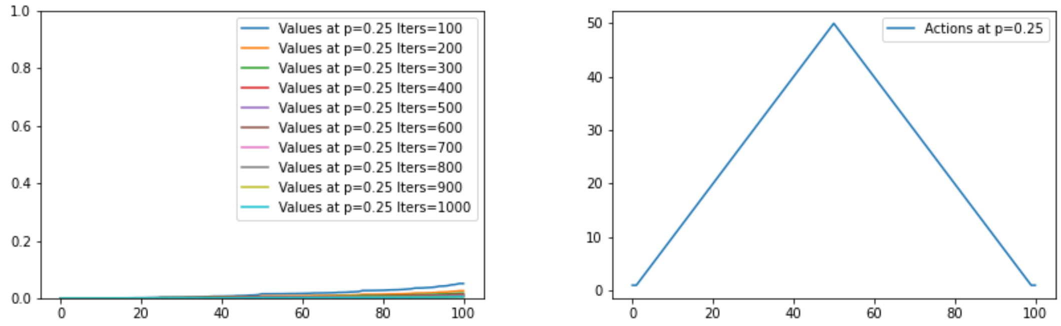


Figure 12: a) Optimal Value Estimate for p = 0.25, b) Optimal Policy Estimate for p = 0.25

next state is 99 with probability 1 (again the optimal action at 100 is fixed to 1). The reward function now does not require an exception for reaching 100 from 100 itself and hence can be simplified to reward(state) = 1 if state = 100 and 0 otherwise. The actions for the rest of the states are identical as in sections 1 and 2. Again iterations are carried till convergence. There is no discount factor and the accumulated value is divided by the number of iterations to get the average value.

3.2 Results

Optimal Value and Policy Estimates for p values 0.4, 0.25 and 0.55:

- The plots for the optimal value estimates show that with an increase in the number of iterations, the Average Reward (or Cost) from each state moves towards a common value. This common value increases with an increase in the probablity p of a successful investment. The optimal actions plot for both p = 0.25 and p = 0.4 are identical. They are equal to the state s for states between 1 and 50 and equal to 50 - s for states s from 51 to 99. In both the cases, the optimal action is to invest all that you can. The reason for this is that

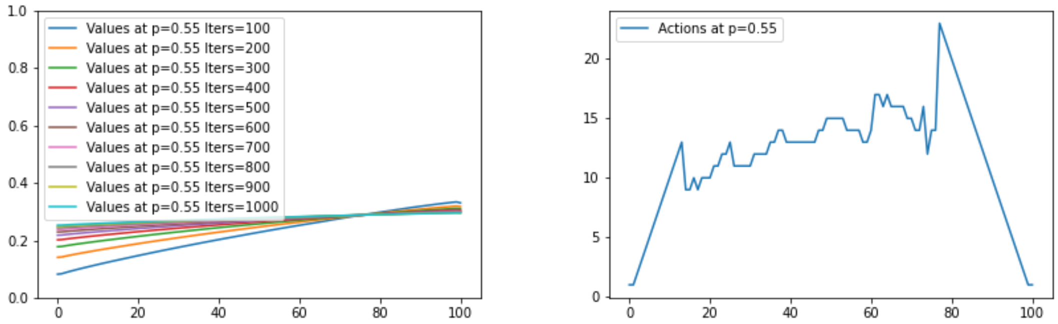


Figure 13: a) Optimal Value Estimate for p = 0.55, b) Optimal Policy Estimate for p = 0.55

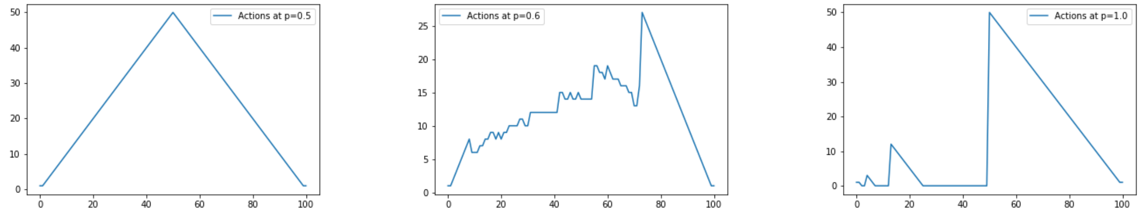
- For p = 0.5, the optimal actions follow the triangle of actions of the earlier states in the beginning and at the very end, however in the middle the optimal investments are quite varied. If the optimal actions for p >= 0.5 are plotted in intervals of 0.1, i.e. p = 0.5, p = 0.51, p = 0.52 and so on, a pattern emerges where the optimal action decreases almost symmetrically on either side of state 50, and then much more asymmetrically in order to reach the optimal actions at p = 1. Plotting the optimal values and optimal policy for p = 1, we can observe that the peaks correspond to where the values jump as well.

- The actions for p = 0.55 correspond to an approximation of the intermediate policy. The final slope begins from the state where the optimal value exceeds the ideal average reward per state, while before it, there are approximately regular peaks which correspond to the largest jumps in the final iteration before termination. The reason for these peaks is that when the probability to win is more than the probability to lose, the investments try to reach the state with the greatest peak with the available actions.

- The beginning part has every state investing the entire amount as a win leads to an increase in value, while losing leads to the value at 0. Due to the shape of the value curve which increases slower with increase in state, and the weightage of the winning amount being greater in the expectation, investing the entire amount available leads to the highest payoff for the first few states. Beyond that, the optimal action changes with the growing values even on a small loss, combined with slower increasing reward with large wins, leading to the shape observed. Beyond the state where all the value curves after different iterations intersect, the optimal investment is always (100 - state) as the extra reward from reaching 100 overcomes any other expected reward which misses out on the 1 reward for once reaching 100.

- For the case of p = 1, some improvements are unreachable from a given state and the optimal action stays at 0. However when the improvements are significant and reachable within the available actions, the alternating peaks and valleys graph can be seen.

3.3 Other Observations



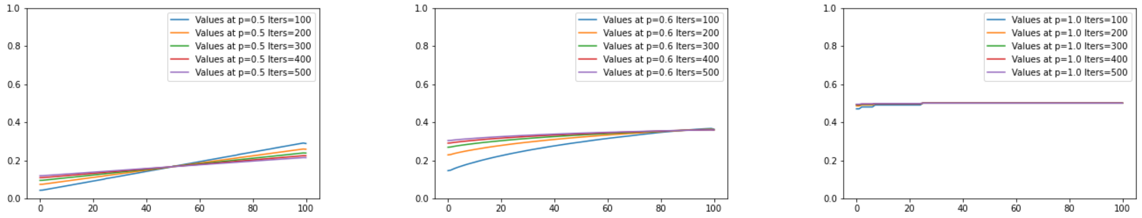
(a) (b) (c)

Figure 14: a) Optimal Actions vs. State for p = 0.5, b) Optimal Actions vs. State for p = 0.6, c) Optimal Actions vs. State for p = 1

- We can see that the case of p = 0.55 is somewhere in between the case of p = 0.5 and p = 0.6. Moreover, the optimal actions at p = 1 indicate the behaviour. The plots of optimal values vs. states for the same cases as above, we can see how the trend of values evolves. The jumps in value seen in the p = 1 curve correspond exactly with the jumps in actions seen in the optimal investments plot for p = 1, offering visual justification for the explanation in section 3.2.

- Note also that the optimal value after complete convergence can be estimated by the value at the intersection point as we can see that after any number of iterations, the value at that state remains unchanged and from the fact that in the average cost problem, the optimal value for all the states must be the same.

- Another observation from the value graphs is that the optimal value per state increases with increase in p from 0 to 1. As p increases, the state for which the value remains fixed moves from 0 to 100. The maximum optimal value is 0.5 at p = 1.



(a) (b) (c)

Figure 15: a) Optimal Values vs. State for p = 0.5, b) Optimal Values vs. State for p = 0.6, c) Optimal Values vs. State for p = 1