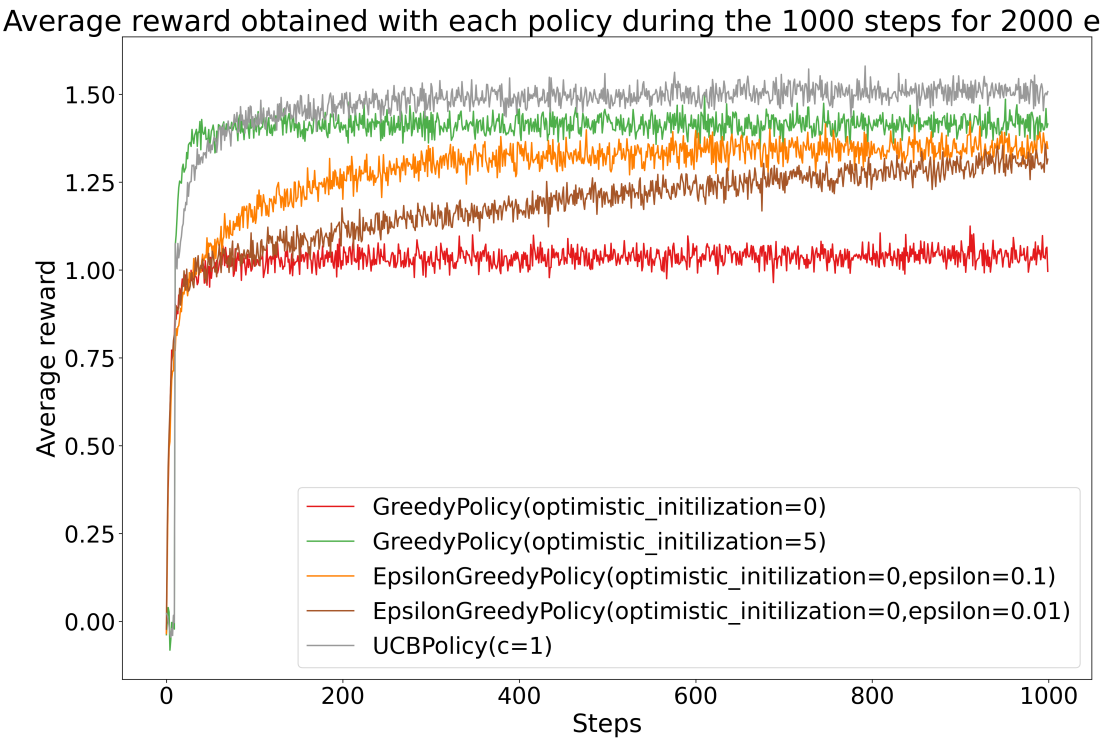
Question 1.



- The **Greedy Policy without optimistic initialization** has the least average reward due to lack of exploration. It chooses always the action with the highest estimated reward, without exploring other actions enough times, not finding the optimal action to maximize the reward as the steps increase. It got stuck on 1 since it kept choosing a non-optimal action.

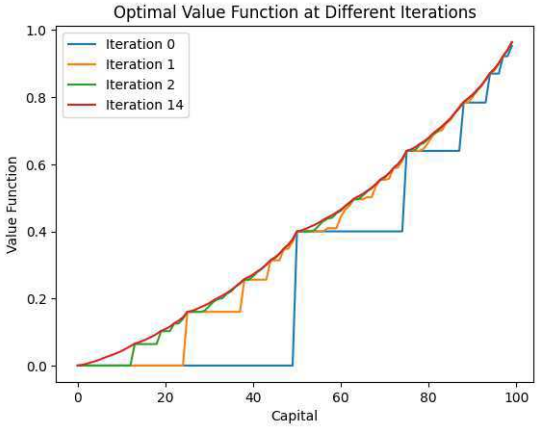
- The **Greedy Policy with optimistic initialization** behaves worse for the initial steps because initial action values are very high, incentivising only exploration. Each action is tried multiple times until the value estimates converge to their approximate values. Then, it picks up the pace, outperforming all but UCB, since it now has a clear idea of which is the optimal action.

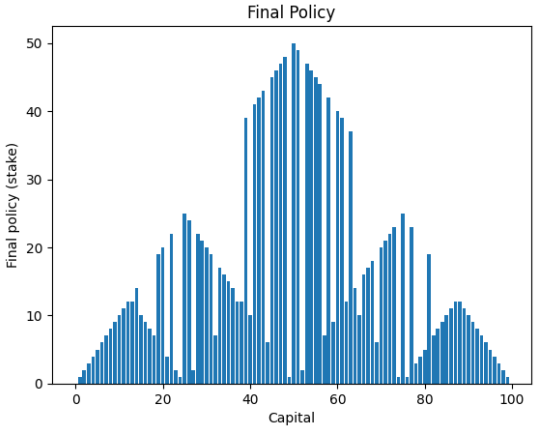
- The **Epsilon Greedy Policy with epsilon = 0.1** explores more (~10% of the times, it chooses exploration) meaning that the average reward increases faster, but also stabilizes around 1.3.

- The **Epsilon Greedy Policy with epsilon = 0.01** explores less (~1% of the times, it chooses exploration) than epsilon = 0.1, which makes it improve more slowly, but continues improving for longer, eventually outperforming the Epsilon Greedy Policy with epsilon = 0.01. So, between these two, it'll will depend on the amount of steps allowed in a given circumstance. However, both are outperformed by Greedy Policy with optimistic initialization and UCB.

- The **UCB Policy** just doesn't perform better than the Epsilon Greedy Policies on the first steps, but it surpasses them in few steps. This happens because it has to try all actions at least once. It grows very fast in the beginning and gets slower to increase in the end because it uses a natural logarithm that ensures all actions will eventually be selected, but it will mostly choose actions with higher confidence (visited more times), allowing it a good balance between clever exploration (clever because it uses a confidence value based on the times it tried a given action) and exploitation (still exploits actions with higher confidence and reward). Here, we used c = 1 instead of c = 2 as in the book, which decreases the level of exploration, but gets better average rewards overall.

Question 2.





- In the first iteration, the value function is initialized with zeros for all states. As the iteration progresses, the values start to change. However, since it's the first iteration, the gambler has not yet refined its estimates, and the values are relatively uniform, which the exception to points where the gambler can easily win the game with prob_heads.
- In the second iteration, the value function starts showing higher values in states like 25 capital, where the gambler can also get closer to beating the game. More states show increases in the value function, in particular when the gambler is closer to the goal.
- In the third iteration, the trend is already quite close to the optimal value function, which is expected since as we get closer to the optimal value function, the improvements get smaller and smaller.
- The final iteration shows the optimal value function, which provides the maximum expected cumulative reward for each state.

- The final policy plot details how much the gambler should stake with a given current capital value. There are two smaller spikes at 25 and 75, and the largest one at 50. If we are at state 50, we have 40% chance of winning by betting all 50, while if we bet only 1 each time, we would need to win 50 times in a row, which is very unlikely. When we are at 25, we should bet 25 to try to reach 50 capital, since this would only require me to win twice in a row to beat the game. With capital 75, we would only need to bet 25 and win once to beat the game. However, if I have between 25 and 50, I can never beat the game in less than two rounds, and since I have a little more capital than 25, I can bet less. When I have more than 75, I also shouldn't risk more money than what is required to beat the game, thus the stakes should also be smaller than 75. If I have between 51 and 74, I don't want to risk getting much worse than 50 capital, since if I get to 75, I can already bet freely, knowing that I'll either get to around 50 or win the game.
- What this means is, since the odds are against us, we should try to beat the game with as few rounds as possible, and bet more when we are at an advantageous position (have more capital).
- If prob_heads > 0.5, the odds would be in our favour, so we wouldn't need to bet more when we are at an advantageous positions, we could just bet 1 every time, and eventually, we would beat the game.

NOTE: The plot is quite different from the one presented in the book, which can be influenced by numerous factors, such as the precisions used, the discount factor gamma, and the precision of the converge value (number of iterations).

Question 3) Given two vectors u and v

$$T(u) = r_\pi + \gamma P_\pi u \quad \text{and} \quad T(v) = r_\pi + \gamma P_\pi v$$

$$\|T(u) - T(v)\| = \|(r_\pi + \gamma P_\pi u) - (r_\pi + \gamma P_\pi v)\| =$$

$$= \|\gamma P_\pi (u - v)\| =$$

$$= \gamma \|P_\pi\| . \|(u - v)\| \leq$$

$\|P_\pi\| \leq 1$ ↳ $\leq \gamma \|u - v\| \longrightarrow T$ is a contraction

Following Banach fixed-point theorem, since T is a contraction, it has a unique fixed point where $Tx^* = x^*$.

Since $v_\pi = r_\pi + \gamma P_\pi v_\pi$, the value iteration converges to v_π

Question 4 For two vectors u and v:

$$\| T^{(\lambda)}(u) - T^{(\lambda)}(v) \|_\infty \leq \gamma \| u - v \|_\infty$$

$$\| T^{(\lambda)}(u) - T^{(\lambda)}(v) \|_\infty =$$

$$= \| \sum_{n=0}^{\infty} (\lambda\gamma P_\pi)^n \left[r_\pi + \gamma P_\pi u - u - (r_\pi + \gamma P_\pi v - v) \right] \|_\infty =$$

$$= \| \sum_{n=0}^{\infty} (\lambda\gamma P_\pi)^n \left[r_\pi + \gamma P_\pi (u - v) \right] \|_\infty \leq$$

triangle inequality $|a+b| \leq |a| + |b|$

$$\leq \sum_{n=0}^{\infty} \| (\lambda\gamma P_\pi)^n \|_\infty \cdot \| \left[\gamma P_\pi (u - v) \right] \|_\infty \leq$$

$\| P_\pi \| \leq 1$ and λ and γ are scalars, so the regular norm applies

$$\leq \sum_{n=0}^{\infty} |\lambda\gamma|^n \cdot \| \gamma P_\pi (u - v) \|_\infty =$$

$$= \frac{1}{1 - |\lambda\gamma|} \cdot \| \gamma P_\pi (u - v) \|_\infty \leq$$

$|\lambda\gamma| < 1$

$$\leq \| \gamma P_\pi (u - v) \|_\infty \leq$$

$\| P_\pi \| \leq 1$

$$\leq \gamma \| u - v \|_\infty =$$

sup-norm

$$= \gamma \max | u - v |$$

By the Banach fixed point theorem $T^{(\lambda)}$ is a contraction with respect to the sup-norm.