Answers

Q2)

1) In the 1k episodes,
5 Bins has total number of states = $5^4 = 625$
50 Bins has total number of states = $(50)^4 = 625 \times 10^4$
$= 6,250,000$

In 1K episodes :-
In the configuration with 5 Bins, the agent can almost visit all of the states, even after considering that it will ~~ter~~ not land in each unique state in episode.

In the configuration with 50 Bins the agent can no way explore all the 6,250,000 steps in just 1000 episodes. Even if it explored 1 unique state each episode (which is not real), it can only explore 1000 states,

So the Q table for 1K episodes is more developed and on way to convergence as each state would have been visited many times. Whereas for 50 Bins configuration, most of the states aren't visited even once, leading to a very immature Q table → less reward.

2) As of 10k episodes, the agent would have explored most of the states many times leading to a more mature Q table, continuously increas-ing reward with episodes.
Whereas for 5bins, the Q table would have converged to a particular table, leading to a stable reward.

~~W~~ The rewards for the 50 bin configuration will eventually overtake the 5bin configuration as the Q table begins to mature.

3) Discretization is generally a good method for simple problems. But the problem arises on edge cases. It assumes a hard boundary at the edge.
Let us take example of pole angle.
Assume it sets a boundary at 20°. Two intervals 20°-24°, 16°-20°. But the ~~strategy~~ action for both the spaces remains same.
So the agent would have to learn ~~a~~ completely from scratch for the observation space 20°-24°, it can't transfer its knowledge even when knowledge is similar, increasing training time.

4) 5000 bins in each dimension → $(5000)^4$ states
$= 625 \times (10^3)^4 = 625 \times 10^{12}$ states. = 625 trillion states

These will take a lot of time (impossible) to train and will require a very large number of episodes to make a good Q table. But if we are able to train this, then it will lead to very high reward. But then also, it will not increase the reward by a high margin because microscopic bins will lead to ~~marginal~~ gain very microscopic adjustements that don't matter much.