# Chapter 5

## Exercise 5.1:

The figures have a jump in the last two rows because when the player's sum is 20 or 21, he sticks and has a very high probability of winning (because the dealer stops on 17 or higher and might go bust in an attempt to beat 20 or 21).

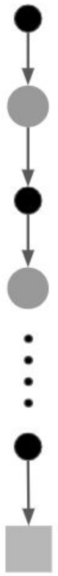
The values drop for the whole row when the dealer has an ace because if that ace is usable, the dealer has lesser chances of going bust (because the ace can count as 1 or 11) and has higher chances of ending up on a high value.

The front rows in case of usable ace value functions are higher because the player also enjoys similar advantages due to the usable ace. He has high chances of ending up on a high value and can use the ace to prevent himself from going bust.

## Exercise 5.2:

In the described example, a state cannot be visited more than once in an episode. Therefore every-visit monte carlo and first-visit monte carlo are the same methods in this case.

## Exercise 5.3:



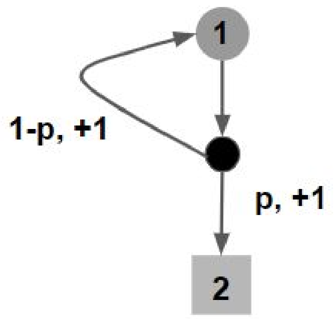
## Exercise 5.4:

Let $c(s,a)$ denote the number of times the state action pair $(s,a)$ is encountered.
The code should be then modified to:

$$c(s,a) \leftarrow c(s,a) + 1$$
$$q(s,a) \leftarrow q(s,a) + [G - q(s,a)]/c(s,a)$$

This is derived from the incremental implementation discussed in chapter 2.

## Exercise 5.5:



Episode lasts 10 steps:

$1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2$
$G_i = 11 - i$

First visit estimator:
$v_\pi(1) = G_1 = 10$
Every visit estimator:

$$v_\pi(1) = [\ \sum_{i=1}^{10} G_i\ ] / 10 = 55/10 = 5.5$$

## Exercise 5.6:

When considering $q_\pi(s, a)$ we are calculation value function when action $a$ is already taken. Therefore, in the estimation of $q_\pi(s_t, a_t)$ we do not need the probability of selecting $a_t\ given\ s_t$.

$$q_\pi(s, a) \leftarrow \frac{\sum_{t \in \tau(s,a)} \rho_{t+1\,:\,T(t)-1} * G_t}{|\tau(s,a)|}$$

## Exercise 5.7:

The weighted importance sampling method is biased especially in the initial few episodes when states are encountered very few times. The bias dies down as the number of episodes increases. This is the reason for the initial rise in the error.

## Exercise 5.8:

The expected squared return

$$E_b[(\prod_{t=0}^{T-1} \frac{\pi(a_t|s_t)}{b(a_t|s_t)} G_o)^2]$$

For multi-visit method is:

$$E_b[(\frac{1}{T-1} \sum_{k=0}^{T-1} \prod_{t=k}^{T-1} \frac{\pi(a_t|s_t)}{b(a_t|s_t)} G_k)^2]$$

In this example $G_i = 1 \, for \ all \ i$

Writing for each episode length:

$E_b =$

$+ 0.5 * 0.1 * (2)^2$ Episode length=1

$+ 0.5 * 0.9 * 0.5 * 0.1 * ([2 + 2 * 2]/2)^2$ Episode length=2

$+ 0.5 * 0.9 * 0.5 * 0.9 * 0.5 * 0.1 * ([2 + 2 * 2 + 2 * 2 * 2]/3)^2$ Episode length=3

And so on…

$$E_b = \sum_{k=1}^{\infty} 0.5^k * 0.9^{k-1} * 0.1 * (1/k * \sum_{i=1}^{k} 2^i)^2$$

$$E_b = 1/9 * \sum_{k=1}^{\infty} 0.45^k * (1/k * \sum_{i=1}^{k} 2^i)^2$$

$$E_b = 1/9 * \sum_{k=1}^{\infty} 0.45^k * (2/k * (2^k - 1))^2$$

$$E_b = 1/9 * \sum_{k=1}^{\infty} 0.45^k * 4/k^2 * [4^k + 1 - 2^{k+1}]$$

$$E_b = 4/9 * [\ \sum_{k=1}^{\infty} 1.8^k * 1/k^2 + \sum_{k=1}^{\infty} 0.45^k * 1/k^2 - 2 * \sum_{k=1}^{\infty} 0.9^k * 1/k^2\ ]$$

It can be seen that $\sum_{k=1}^{\infty} 1.8^k * 1/k^2 \rightarrow \infty for\ large\ k$

while $\sum_{k=1}^{\infty} 0.45^k * 1/k^2 - 2 * \sum_{k=1}^{\infty} 0.9^k * 1/k^2\ ] \rightarrow 0 for\ large\ k$

Therefore $E_b \rightarrow \infty$. Variance is infinite even for the multi-visit method.

## **Exercise 5.9:**

$$c(s) \leftarrow c(s) + 1$$
$$v_\pi(s) \leftarrow v_\pi(s) + [G - v_\pi(s)]/c(s)$$

Where $c(s)$ denotes the number of times state '$s$' was encountered (counting only the first visit in this case).

## Exercise 5.10:

$$v_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$v_{n+1} = \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}$$

$$\sum_{k=1}^{n-1} W_k G_k = v_n * (\sum_{k=1}^{n} W_k - W_n)$$

$$\sum_{k=1}^{n} W_k G_k = W_n G_n + v_n * (\sum_{k=1}^{n} W_k - W_n)$$

$$v_{n+1} = \frac{W_k G_k + v_n * (\sum_{k=1}^{n} W_k - W_n)}{\sum_{k=1}^{n} W_k} = v_n + \frac{W_n}{\sum_{k=1}^{n} W_k}[G_n - v_n]$$

Therefore,

$$v_{n+1} = v_n + \frac{W_n}{C_n}[G_n - v_n], \; C_n = C_{n-1} + W_n$$

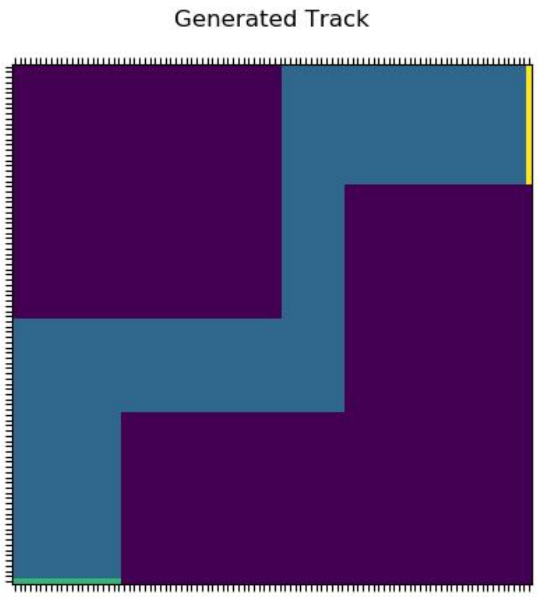
## Exercise 5.11:

The policy is updated to $\pi(s) \leftarrow arg\, max_a\, q(s, a)$ with ties broken consistently. Therefore, $\pi(a_t|s_t) = 1$ (since $a_t$ is guaranteed to be $\pi(s)$). Thus, the update is correct.

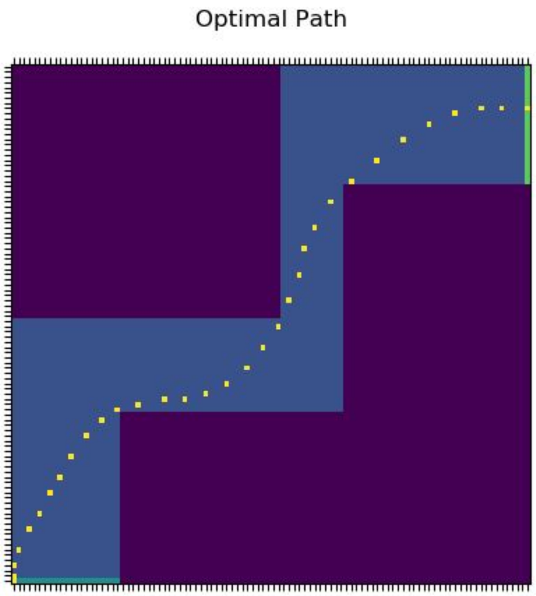
# **Exercise 5.12:**

Code: Check Github

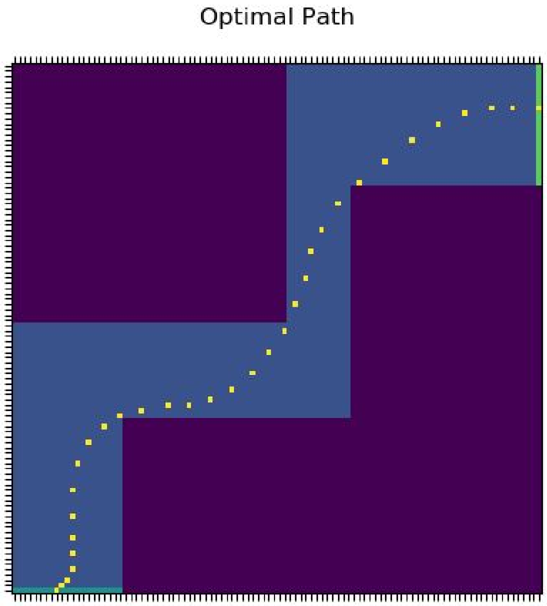
Results:



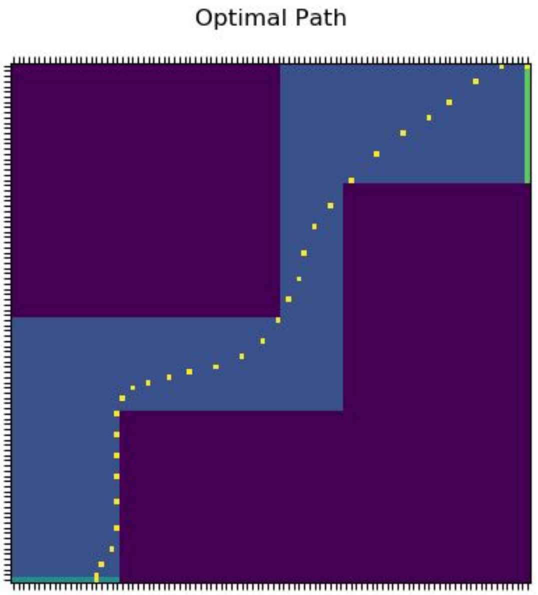
Generated Track



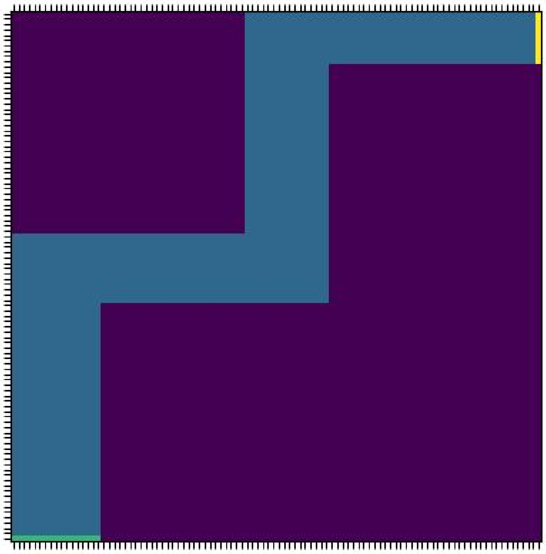
Optimal Path



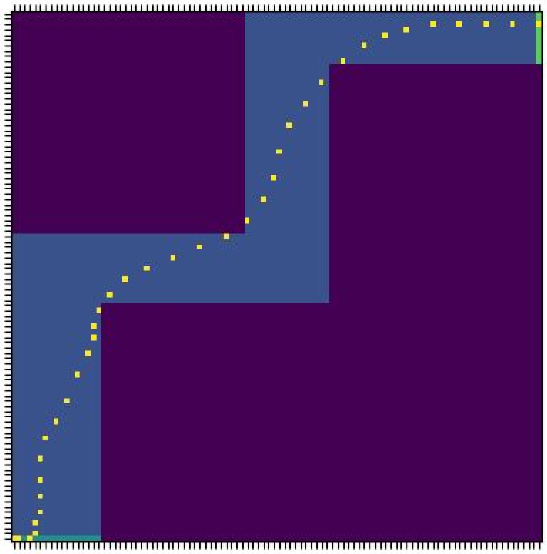
Optimal Path



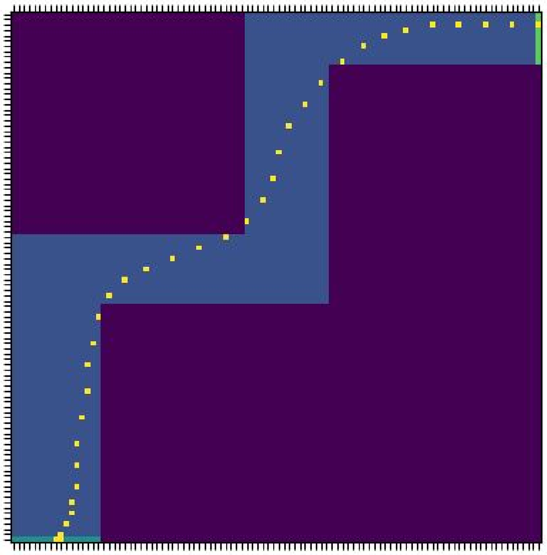
Optimal Path



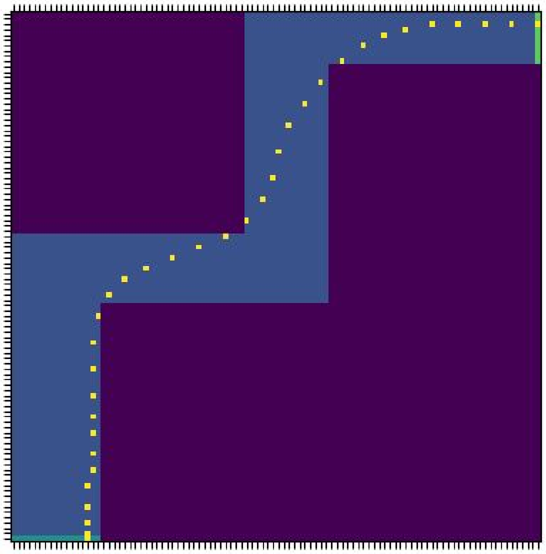
Generated Track



Optimal Path



Optimal Path



Optimal Path