Question 2

1. **A formal MDP for the given problem:**

   State Space: $S = \{0, \dots, 2k-1\}$

   Action Space: $A = \{CW, CCW\}$

   Transition Probability:

   $\forall i, j \in S:\ P(s_{t+1} = j \mid a_t = CW, s_t = i) = \begin{cases} 1 & j = i+1 \\ 0 & otherwise \end{cases}$

   $\forall i, j \in S:\ P(s_{t+1} = j \mid a_t = CCW, s_t = i) = \begin{cases} 1 & j = i-1 \\ 0 & otherwise \end{cases}$

   Initil State: $s_0 = k$

   immediate reward for a given state and action:

   $\forall\ 1 \le i \le 2k:\ \ r(i, a) = 0 \qquad a \in A$

   $i = 0:\ \ r(i, a) = 1 \qquad a \in A$

   Return $V_\gamma^\pi(k) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \mid s_0 = k]$

2. The intuition for the optimal policy is how we can reach to state 0 as fast as possible at any given time, because only when we at state 0 we will get reward of 1. For any other state the reward is 0.

   The optimal policy:

   $\pi(s) = \begin{cases} CCW & 0 \le s < k \\ CW & s \le k \le 2k \end{cases}$

3. Value Iteration formula: $V_{n+1}(s) = \max_{a \in A}\{r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) \cdot V_n(s')\}$

   we let $V_0(s) = 0\ \forall s \in S$, after one iteration we will get:

   $V_1(0) = \max\{1 + \gamma V_0(1), 1 + \gamma V_0(2k-1)\} = 1$

   if $i = 1,2,3 \dots, 2k-1$:

   $V_1(i) = \max\{\gamma V_0(i-1), 1 + \gamma V_0(i+1)\} = 0$

4. $V_2(0) = \max\{1 + \gamma V_1(1), 1 + \gamma V_1(2k-1)\} = 1$

   if $i = 2,3 \dots, 2k-2$:

   $V_2(i) = \max\{\gamma V_1(i-1), \gamma V_1(i+1)\} = 0$

   if $i = 1$ $(and\ the\ same\ calculation\ for\ 2k-1)$:

   $V_2(1) = \max\{\gamma V_1(1-1), 1 + \gamma V_1(1+1)\} = \max\{\gamma V_1(0), \gamma V_1(2)\} = \gamma$

   so overall the states $1, 2\mathrm{k} - 1$ change their value after two iterations

5. $S = \{0,1,2,3\}$, notice that we can reach to 0 from 0 only on odd number of steps.

   from 1 or 3 we can reach to 0 in 1 step and from 2 we need 2 steps.

   overall:

   1-step:

   $V_1(0) = 1$

   $V_1(i) = \gamma \; i \in \{1,3\} \; for \; all \; i \; in \; this \; calculation$

   $V_1(2) = 0$

   2-steps:

   $V_2(0) = 1$

   $V_2(i) = \gamma$

   $V_2(2) = \gamma^2$

   3-steps:

   $V_3(0) = 1 + \gamma^2$

   $V_3(i) = \gamma$

   $V_3(2) = \gamma^2$

   4-steps:

   $V_4(0) = 1 + \gamma^2$

   $V_4(i) = \gamma + \gamma^3$

   $V_4(2) = \gamma^2$

   5-steps:

   $V_5(0) = 1 + \gamma^2 + \gamma^4$

   $V_5(i) = \gamma + \gamma^3$

   $V_5(2) = \gamma^2 + \gamma^4$

   6-steps:

   $V_6(0) = 1 + \gamma^2 + \gamma^4$

   $V_6(i) = \gamma + \gamma^3 + \gamma^5$

   $V_6(2) = \gamma^2 + \gamma^4$

   6-steps:

   $V_7(0) = 1 + \gamma^2 + \gamma^4 + \gamma^6$

   $V_7(i) = \gamma + \gamma^3 + \gamma^5$

   $V_7(2) = \gamma^2 + \gamma^4 + \gamma^6$

Overall:

$$V^*(0) = \sum_{i=0}^{\infty} \gamma^{2i} = \frac{1}{1-\gamma^2}$$

$$V^*(1) = V^*(3) = \gamma \sum_{i=0}^{\infty} \gamma^{2i} = \frac{\gamma}{1-\gamma^2}$$

$$V^*(2) = \gamma^2 \sum_{i=0}^{\infty} \gamma^{2i} = \frac{\gamma^2}{1-\gamma^2}$$

Question 4

Let M be a MDP define under (S, A, P, s_0) and discounted factor $\gamma \in [0,1]$.
We want to show that for each $v_1, v_2 \in \mathbb{R}^{|S|}$ it holds:

$$\|T(v_1) - T(v_2)\|_\infty \le \gamma \cdot \|v_1 - v_2\|_\infty$$

Proof:

$$\|T(v_1) - T(v_2)\|_\infty = |T(v_1)(s) - T(v_2)(s)| =$$

$$\left|\frac{1}{|A|}\sum_{a \in A}\left(r(s,a) + \gamma \sum_{s' \in S} P(s'|a,s)v_1(s')\right) - \frac{1}{|A|}\sum_{a \in A}\left(r(s,a) + \gamma \sum_{s' \in S} P(s'|a,s)v_2(s')\right)\right| =$$

$$\frac{1}{|A|}\gamma \cdot \left|\sum_{a \in A}\sum_{s' \in S}\Big(P(s'|a,s) \cdot \big(v_1(s') - v_2(s')\big)\Big)\right| \le$$

$$\frac{1}{|A|}\gamma \cdot \sum_{a \in A}\sum_{s' \in S}(P(s'|a,s) \cdot |v_1(s') - v_2(s')|) \le$$

$$\frac{1}{|A|}\gamma \cdot \sum_{a \in A}\sum_{s' \in S}(P(s'|a,s) \cdot \|v_1 - v_2\|_\infty\ ) = \frac{1}{|A|}\gamma \cdot |A| \cdot 1 \cdot \|v_1 - v_2\|_\infty = \gamma \cdot \|v_1 - v_2\|_\infty$$

when the first inequality holds using triangle inequality and the second inequality holds using the definition of $\|\cdot\|_\infty$

Overall:

$$\|T(v_1) - T(v_2)\|_\infty \le \gamma \cdot \|v_1 - v_2\|_\infty$$

Programming Part – Question 2

	iteration	chg actions	V[0]
0	0	1	0.0
1	1	6	0.0
2	2	5	0.0
3	3	5	0.0563
4	4	1	0.17956
5	5	0	0.18047
6	6	0	0.18047
7	7	0	0.18047
8	8	0	0.18047
9	9	0	0.18047
10	10	0	0.18047
11	11	0	0.18047
12	12	0	0.18047
13	13	0	0.18047
14	14	0	0.18047
15	15	0	0.18047
16	16	0	0.18047
17	17	0	0.18047
18	18	0	0.18047
19	19	0	0.18047