Exercise 5

Instructions

- Play Blackjack in small groups. One student acts as the dealer, the others are players.
- 2. Use a **fixed policy**:
 - Hit if your total < 20, otherwise Stand.
- 3. For each episode (a full hand until win/loss/draw):
 - a. Record the **sequence of states, actions, and rewards**.
 - b. Compute **MC updates** (after the episode).
 - c. Compute **TD(0) updates** (during the episode).
- 4. Compare how the two methods update the value table.

Part A: Record an Episode

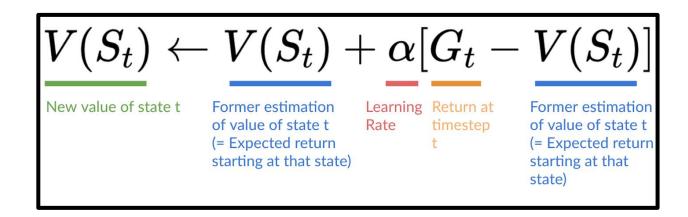
Step	State (Player Sum, Dealer Showing, Usable Ace?)	Action (Hit=1, Stand=0)	Reward <i>G</i>	Next State
1				
2				
•••				
END				

Part B: Monte Carlo Update (First-Visit)

At the end of the episode, compute the return

$$G_t = R_{t+1} + R_{t+2} + \dots R_T$$

• For each state visited **first time** in the episode:



Part B: Monte Carlo Update (First-Visit)

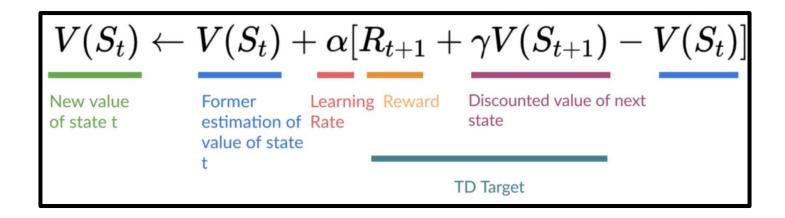
• Record the sequence of states, actions, and rewards.

State S	Return G	Visit Count $N(s)$	OldV(s)	New $V(s)$

• Use $\alpha = \frac{1}{n}$ for manual calculations.

Part C: TD(0) update

Update during the episode for each transition:



• Take $\gamma = 1.0$, choose $\alpha = 0.5$ for manual calculations.

Part C: TD(0) update

Step	State s	Reward r	Next State s'	$\mathbf{Old}V(s)$	New $V(s)$
1					
2					

Example Episodes

Step	State (Player Sum, Dealer, Usable Ace)	Action A	Reward <i>G</i>	Next State
1	(15, 10, False)	Hit	0	(19, 10, F)
2	(19, 10, False)	Hit	– 1	BUST

Sample Monte Carlo (First-Visit) Update

State	Return G	N(s)	Old V(s)	New $V(s)$
(15,10,False)	-1	1	0	0
(19,10,False)	-1	1	0	0

Sample Monte Carlo (First-Visit) Update

Step 1

- a. State =(15,10,F)
- b. Number of Visits: N(15,10,False) = 0
- c. V(15,10,False) = 0
- d. Increment visit count: N(15,10,False) = 1
- e. Update value function: $V(15,10,\text{False}) = 0 + \frac{1}{1}(-1 0) = -1$

State	Return G	N(s)	$\mathbf{Old}V(s)$	$\mathbf{New}V(s)$
(15,10,False)	- 1	1	0	– 1
(19,10,False)	– 1	0	0	0

$V(S_t) \leftarrow$	$-V(S_t)$ -	$+ \underline{\alpha}$	G_t –	$-V(S_t)]$
New value of state t	Former estimation of value of state t (= Expected return starting at that state)	Learning Rate	Return at timestep t	Former estimation of value of state t (= Expected return starting at that state)

Sample Monte Carlo (First-Visit) Update

Step 1

- a. State =(19,10,F)
- b. Number of Visits: N(19,10,False) = 0
- c. V(19,10,False) = 0
- d. Increment visit count: N(19,10,False) = 1
- e. Update value function: $V(19,10,\text{False}) = 0 + \frac{1}{1}(-1 0) = -1$

State	Return G	N(s)	$\mathbf{Old}V(s)$	$\mathbf{New}V(s)$
(15,10,False)	– 1	1	0	– 1
(19,10,False)	- 1	1	0	– 1

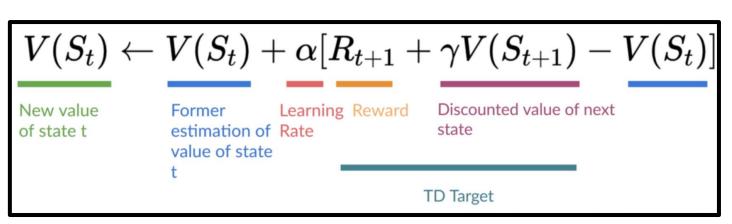
$V(S_t) \leftarrow$	- $V(S_t)$ -	$+ \underline{\alpha}$	G_t –	$-V(S_t)]$
New value of state t	Former estimation of value of state t (= Expected return starting at that state)	Learning Rate	Return at timestep t	Former estimation of value of state t (= Expected return starting at that state)

Sample Temporal Difference(0) Update

Step 1

Start with V(s) = 0.

Update each step immediately using



- a. State = (15,10,F)
- b. Reward r = 0,
- c. Next state = (19,10,F)
- d. Update: V(15,10,F) = 0 + 0.5(0 + V(19,10,F) 0) = 0

Step	State s	Reward r	Next State s'	OldV(s)	New $V(s)$
1	(15,10,F)	0	(19,10,F)	0	0
2	-	-	-	-	-

Sample Temporal Difference(0) Update

Step 2

Update each step immediately using

- a. State = (19,10,F)
- b. Reward r = -1,
- c. Next state = BUST
- d. Update value function: V(15,10,F) = 0 + 0.5(-1 + 0 0) = 0

Step	State s	Reward r	Next State s'	OldV(s)	New $V(s)$
1	(15,10,F)	0	(19,10,F)	0	0
2	(19,10,F)	-1	BUST	0	-0.5

$V(S_t)$ \leftarrow	$-\ V(S_t)$ -	$+ lpha [R_{t+1}]$	$+ rac{\gamma V(S_{t+1})}{2} - V(S_t)$
New value of state t	Former estimation of value of state		Discounted value of next state
	t		TD Target