Exercise 5

Instructions

1. 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

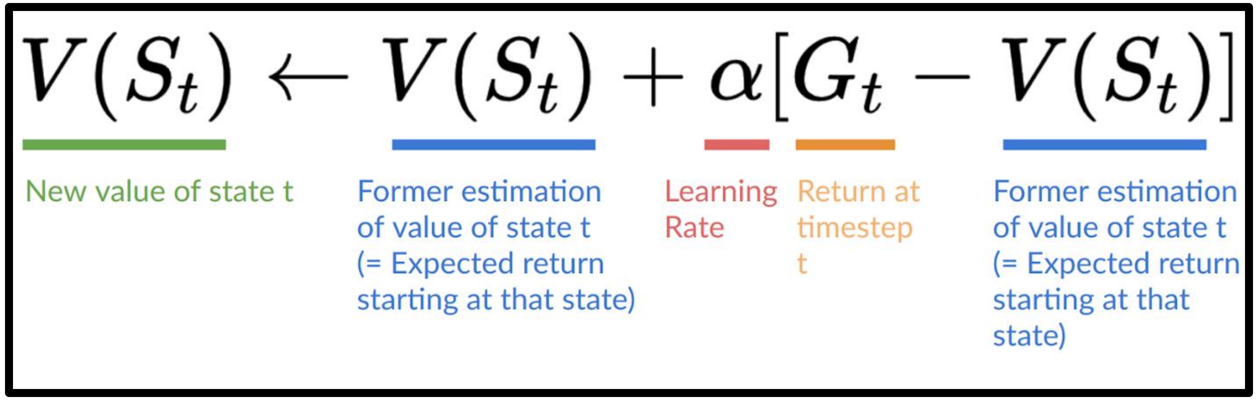
Step	State (Player Sum, Dealer Showing, Usable Ace?)	Action (Hit=1, Stand=0)	Reward G	Next State
1				
2				
…				
END				

Part B: Monte Carlo Update (First-Visit)

- At the end of the episode, compute the **return**

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

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



$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

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)

Part B: Monte Carlo Update (First-Visit)

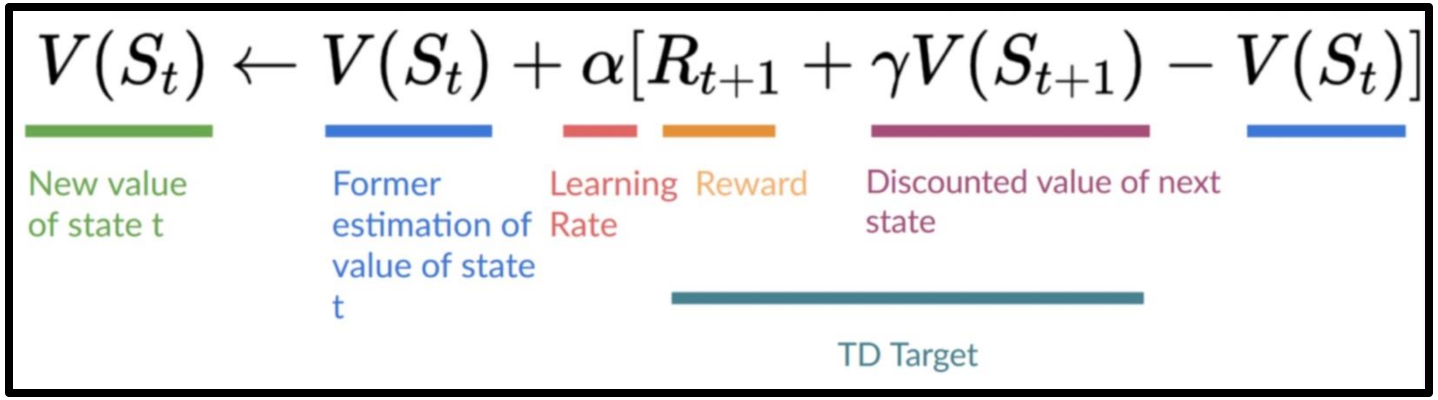
- Record the **sequence of states, actions, and rewards**.

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

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

Part C: TD(0) update

- Update **during the episode** for each transition:



$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

New value of state t
Former estimation of value of state t
Learning Rate
Reward
Discounted value of next state
TD Target

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

Part C: TD(0) update

Step	State s	Reward r	Next State s'	Old $V(s)$	New $V(s)$
1					
2					