Example Episode: Player Loses

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

Sample Monte Carlo (First-Visit) Update

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

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)

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

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

- $V(S_t)$: New value of state t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)
- α: Learning Rate
- G_t: Return at timestep t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)

Step 1

a. State =(15,10,F)

b. Number of Visits: $N(15,10,\text{False}) = 0$

c. V(15,10,False) = 0

d. Increment visit count: $N(15,10,\text{False}) = 1$

e. Update value function: $V(15,10,\text{False}) = 0 + \frac{1}{1}(-1 - 0) = -1$

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

Sample Monte Carlo (First-Visit) Update

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

- $V(S_t)$: New value of state t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)
- α: Learning Rate
- G_t: Return at timestep t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)

Step 1

a. State =(19,10,F)

b. Number of Visits: $N(19,10,\text{False}) = 0$

c. V(19,10,False) = 0

d. Increment visit count: $N(19,10,\text{False}) = 1$

e. Update value function: $V(19,10,\text{False}) = 0 + \frac{1}{1}(-1 - 0) = -1$

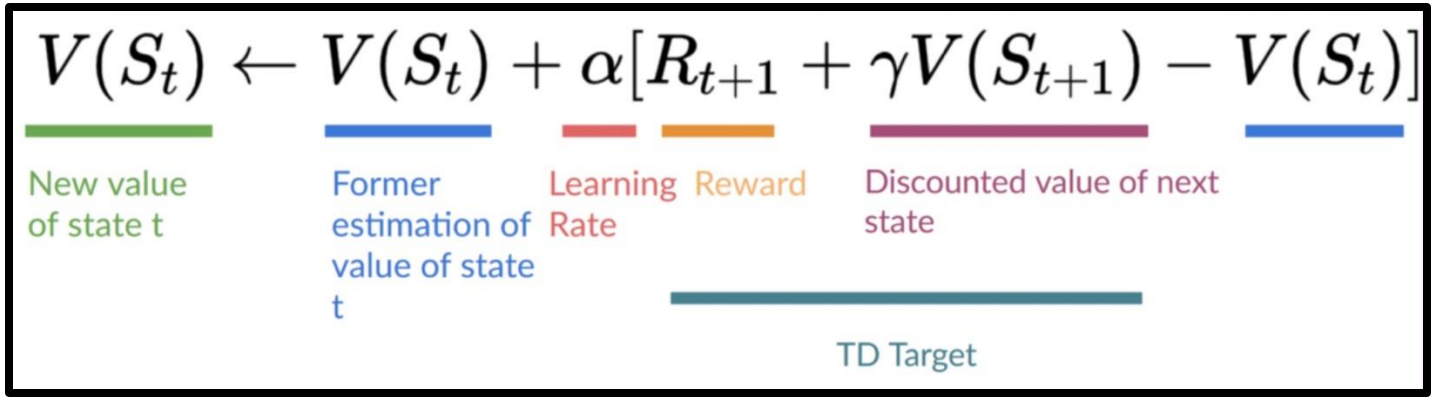
State	**Return G**	N(s)	**Old $V(s)$**	**New $V(s)$**
(15,10,False)	–1	1	0	–1
(19,10,False)	–1	1	0	–1

Sample Temporal Difference(0) Update

Step 1

Start with $V(s) = 0$.

Update each step immediately using



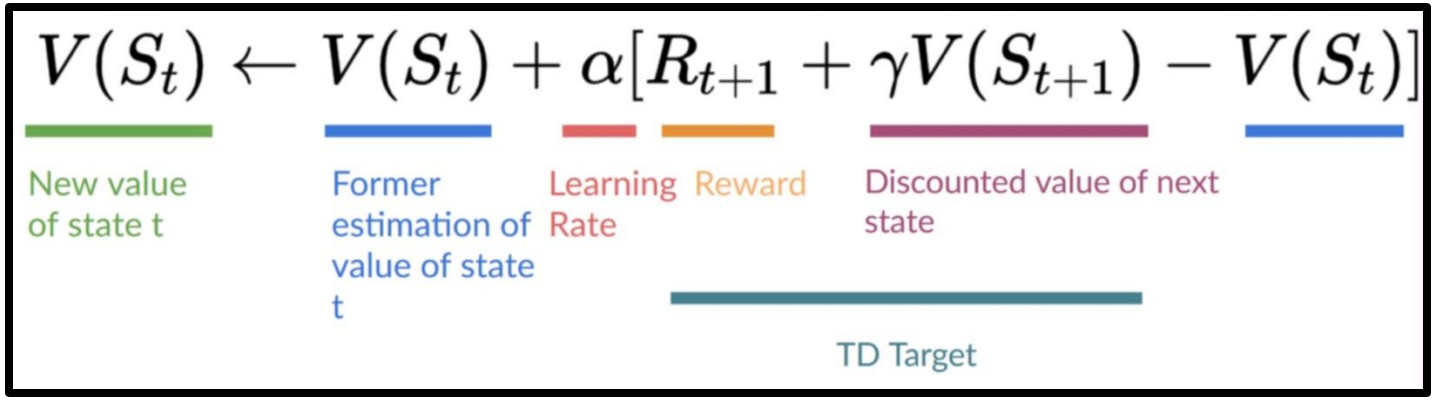
a. State = (15,10,F)
b. Reward r = 0,
c. Next state = (19,10,F)
d. Update: $V(15,10,F) = 0 + 0.5(0 + V(19,10,F) - 0) = 0$

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

Example Episode: Player Wins

Step	State (Player Sum, Dealer, Usable Ace)	Action A	Reward G	Next State
1	(17, 6, False)	Hit	0	(21, 6, False)
2	(21, 6, False)	Stand	+1	WIN

Sample Monte Carlo (First-Visit) Update

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

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)

State	Return G	N(s)	Old $V(s)$	New $V(s)$
(17,6,False)	+1	1	0	0
(21,6,False)	+1	1	0	0

Sample Monte Carlo (First-Visit) Update

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

- $V(S_t)$: New value of state t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)
- α: Learning Rate
- G_t: Return at timestep t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)

Step 1

a. State =(17,6,False)

b. Number of Visits: $N(17,6,\text{False}) = 0$

c. V(17,6,False) = 0

d. Increment visit count: $N(17,6,\text{False}) = 1$

e. Update value function: $V(17,6,\text{False}) = 0 + \frac{1}{1}(1-0) = 1$

State	Return G	N(s)	Old $V(s)$	New $V(s)$
(17,6,False)	+1	1	0	1
(21,6,False)	+1	0	0	0

Sample Monte Carlo (First-Visit) Update

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

- $V(S_t)$: New value of state t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)
- α: Learning Rate
- G_t: Return at timestep t
- $V(S_t)$: Former estimation of value of state t (= Expected return starting at that state)

Step 2

a. State =(21,6,False)

b. Number of Visits: $N(21,6,\text{False}) = 0$

c. V(21,6,False) = 0

d. Increment visit count: $N(21,6,\text{False}) = 1$

e. Update value function: $V(21,6,\text{False}) = 0 + \frac{1}{1}(1-0) = 1$

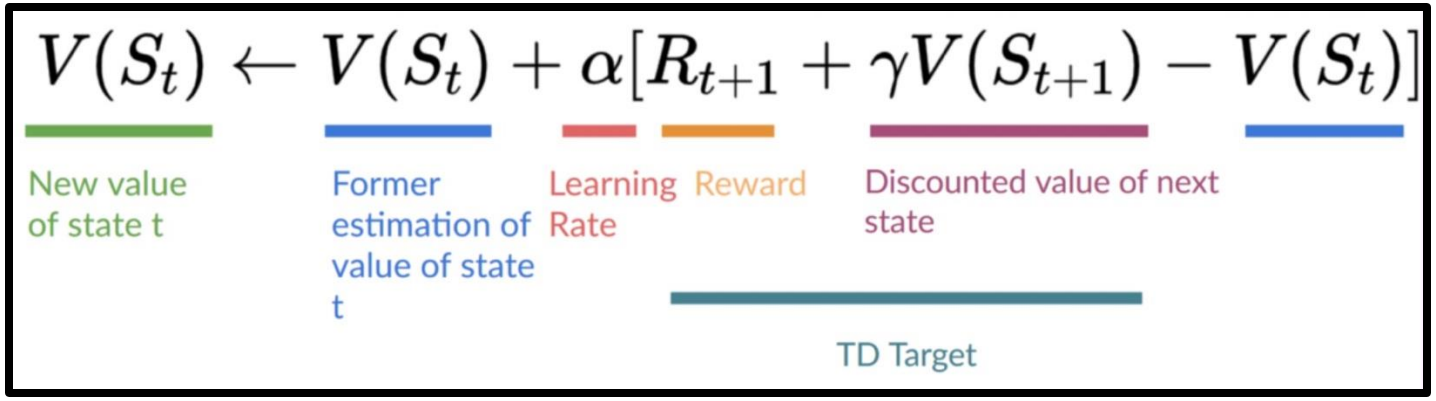
State	Return G	N(s)	Old $V(s)$	New $V(s)$
(17,6,False)	+1	1	0	1
(21,6,False)	+1	1	0	1

Sample Temporal Difference(0) Update

Step 1

Start with $V(s) = 0$.

Update each step immediately using



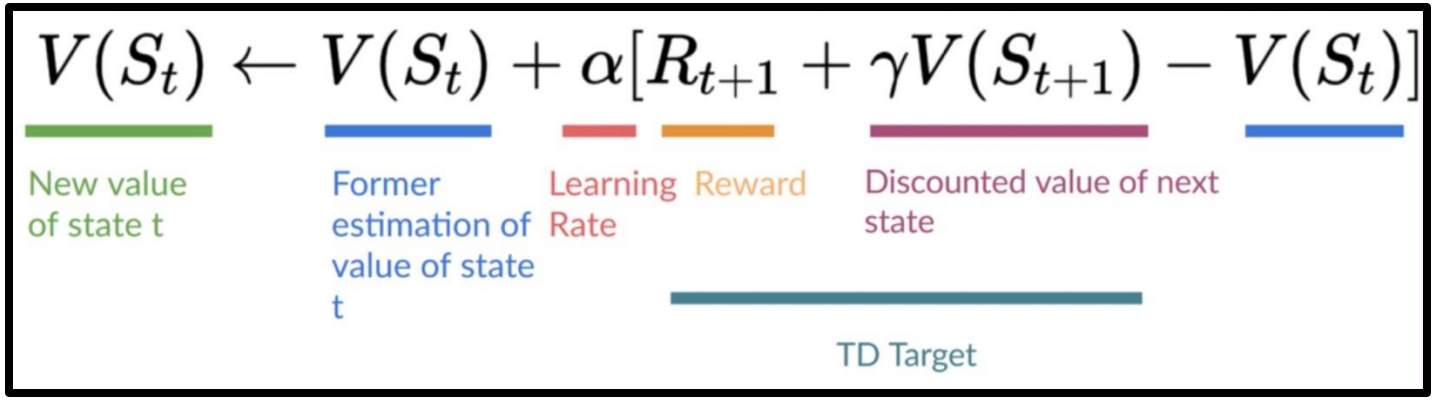
a. State = (17,6,False)
b. Reward r = 0,
c. Next state = (21,6,False)
d. Update: $V(17,6,False) = 0 + 0.5(0 + V(21,6,False) - 0) = 0$

Step	State s	Reward r	Next State s'	Old $V(s)$	New $V(s)$
1	$(17,6,False)$	0	$(21,6,False)$	0	0
2	-	-	-	-	-

Sample Temporal Difference(0) Update

Step 2

Update next state immediately using



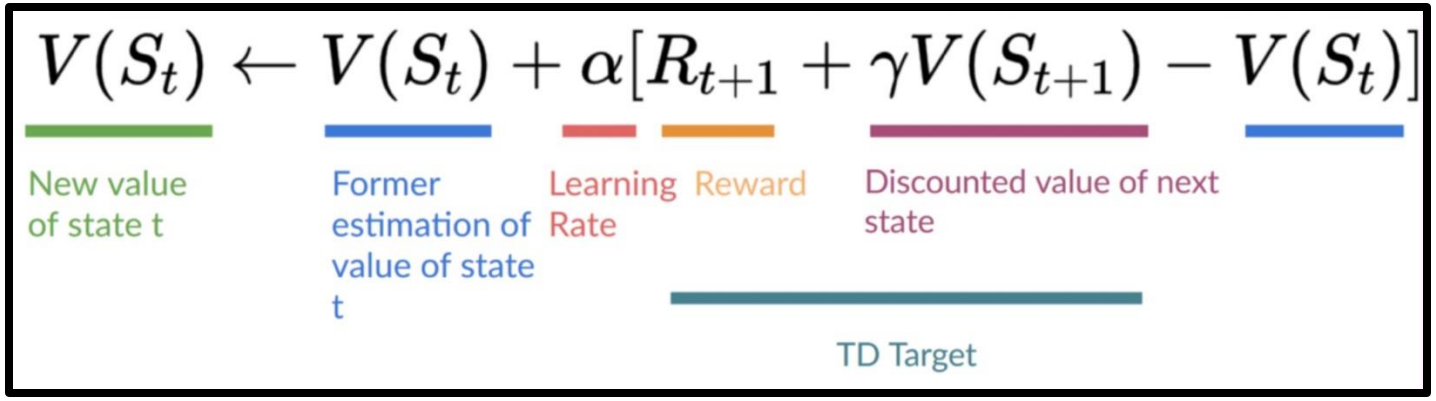
a. State = $(21,6,False)$
b. Reward r = $+1$,
c. Next state = NONE/WIN
d. Update value function: $V(21,6,F) = 0 + 0.5(1 + 0 - 0) = 0$

Step	State s	Reward r	Next State s'	Old $V(s)$	New $V(s)$
1	$(17,6,False)$	0	$(21,6,False)$	0	0
2	$(21,6,False)$	+1	NONE/WIN	0	+0.5

Sample Temporal Difference(0) Update

Step 3

Update all previously visited states using



a. State = $(17,6,False)$
b. Reward r = $+1$,
c. Next state = $(21,6,False)$
d. Update value function: $V(17,6,F) = 0 + 0.5(1 + 0.5 - 0) = 0.25$

Step	State s	Reward r	Next State s'	Old $V(s)$	New $V(s)$
1	$(17,6,False)$	0	$(21,6,False)$	0	0.25
2	$(21,6,False)$	+1	NONE/WIN	0	+0.5