

Example Episode: Player Loses

Step	State (Player Sum, Dealer, Usable Ace)	Action <i>A</i>	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

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

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

Step 1

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

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

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

Step 1

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

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

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

$$\underbrace{V(S_t)}_{\text{New value of state } t} \leftarrow \underbrace{V(S_t)}_{\text{Former estimation of value of state } t} + \underbrace{\alpha}_{\text{Learning Rate}} \underbrace{[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]}_{\text{TD Target}}$$

The diagram illustrates the TD(0) update equation with color-coded components: $V(S_t)$ (green) is the new value; $V(S_t)$ (blue) is the former estimation; α (red) is the learning rate; R_{t+1} (orange) is the reward; $\gamma V(S_{t+1}) - V(S_t)$ (purple) is the TD target.

- 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

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

$$\underbrace{V(S_t)}_{\text{New value of state t}} \leftarrow \underbrace{V(S_t)}_{\text{Former estimation of value of state t}} + \underbrace{\alpha}_{\text{Learning Rate}} \underbrace{[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]}_{\text{TD Target}}$$

Diagram illustrating the TD(0) update equation with color-coded components:

- $V(S_t)$ (green underline): New value of state t
- $V(S_t)$ (blue underline): Former estimation of value of state t
- α (red underline): Learning Rate
- R_{t+1} (orange underline): Reward
- $\gamma V(S_{t+1})$ (purple underline): Discounted value of next state
- $V(S_t)$ (blue underline): Former estimation of value of state t
- The entire bracketed term $[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$ is labeled as the TD Target (blue underline).

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 <i>A</i>	Reward <i>G</i>	Next State
1	(17, 6, False)	Hit	0	(21, 6, False)
2	(21, 6, False)	Stand	+1	WIN

Sample Monte Carlo (First-Visit) Update

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

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

Step 1

- State =(17,6,False)
- Number of Visits: $N(17,6,False) = 0$
- $V(17,6,False) = 0$
- Increment visit count: $N(17,6,False) = 1$
- Update value function: $V(17,6,False) = 0 + \frac{1}{1}(1 - 0) = 1$

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

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

Sample Monte Carlo (First-Visit) Update

Step 2

- a. State =(21,6,False)
- b. Number of Visits: $N(21,6,\text{False}) = 0$
- c. $V(21,6,\text{False}) = 0$
- d. Increment visit count: $N(21,6,\text{False}) = 1$
- e. Update value function: $V(21,6,\text{False}) = 0 + \frac{1}{1}(1 - 0) = 1$

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

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

$$\underbrace{V(S_t)}_{\text{New value of state } t} \leftarrow \underbrace{V(S_t)}_{\text{Former estimation of value of state } t} + \underbrace{\alpha}_{\text{Learning Rate}} [\underbrace{R_{t+1}}_{\text{Reward}} + \underbrace{\gamma V(S_{t+1})}_{\text{Discounted value of next state}} - \underbrace{V(S_t)}_{\text{Former estimation of value of state } t}]$$

TD Target

- 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

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

$$\underbrace{V(S_t)}_{\text{New value of state t}} \leftarrow \underbrace{V(S_t)}_{\text{Former estimation of value of state t}} + \underbrace{\alpha}_{\text{Learning Rate}} [\underbrace{R_{t+1}}_{\text{Reward}} + \underbrace{\gamma V(S_{t+1})}_{\text{Discounted value of next state}} - \underbrace{V(S_t)}_{\text{Former estimation of value of state t}}]$$

TD Target

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

Sample Temporal Difference(0) Update

Step 3

Update all previously visited states using

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

The diagram shows the TD(0) update equation with color-coded components:
 - $V(S_t)$ (green bar): New value of state t
 - $V(S_t)$ (blue bar): Former estimation of value of state t
 - α (red bar): Learning Rate
 - R_{t+1} (orange bar): Reward
 - $\gamma V(S_{t+1})$ (purple bar): Discounted value of next state
 - The entire term in brackets is labeled as the TD Target (blue bar).

- 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, <i>False</i>)	0	(21,6, <i>False</i>)	0	0.25
2	(21,6, <i>False</i>)	+1	NONE/WIN	0	+0.5