# Policy Evaluation

Policy evaluation goal: given policy $\pi$ find

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \qquad \text{for all } s \in \mathcal{S}, \tag{1}$$

where $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$.

# Policy Evaluation

Policy evaluation goal: given policy $\pi$ find

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \qquad \text{for all } s \in \mathcal{S}, \tag{1}$$

where $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$.

Dynamic Programming or Exhaustive Search: directly estimate (1).

# Policy Evaluation

Policy evaluation goal: given policy $\pi$ find

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \qquad \text{for all } s \in \mathcal{S}, \tag{1}$$

where $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$.

~~Dynamic Programming or Exhaustive Search: directly estimate (1).~~

Approximate expectation with empirical mean.

# Policy Evaluation

Policy evaluation goal: given policy $\pi$ find

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \qquad \text{for all } s \in \mathcal{S}, \tag{1}$$

where $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$.

~~Dynamic Programming or Exhaustive Search: directly estimate (1).~~

Approximate expectation with empirical mean. Let $S(s) =$ total return of state $s$

$$V(s) = \frac{S(s)}{N(s)} \xrightarrow{\text{LLN}} v_\pi(s), \quad \text{as } N(s) \to \infty.$$

# Policy Evaluation

Policy evaluation goal: given policy $\pi$ find

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \qquad \text{for all } s \in \mathcal{S}, \tag{1}$$

where $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$.

~~Dynamic Programming or Exhaustive Search: directly estimate (1).~~

Approximate expectation with empirical mean. Let $S(s) =$ total return of state $s$

$$V(s) = \frac{S(s)}{N(s)} \xrightarrow{\text{LLN}} v_\pi(s), \quad \text{as } N(s) \to \infty.$$

From now on

- common: sample a "future" path starting from current state.
- different: ways to estimate $S(s)$ and update $V(s)$.

# Updates I: Monte Carlo

Let $V_{\text{old}}\left(s\right) = \frac{S(s)}{N(s)}$ and do one more update

$$\begin{aligned} V_{\text{new}}\left(s\right) &= \frac{S\left(s\right) + G_t}{N\left(s\right) + 1} \qquad \text{(MC)} \\ &= \frac{N\left(s\right) V_{\text{old}}\left(s\right) + G_t}{N\left(s\right) + 1} = \frac{\left(N\left(s\right) + 1\right) V_{\text{old}}\left(s\right) + \left(G_t - V_{\text{old}}\left(s\right)\right)}{N\left(s\right) + 1}. \end{aligned}$$

# Updates I: Monte Carlo

Let $V_{\text{old}}(s) = \frac{S(s)}{N(s)}$ and do one more update

$$
\begin{aligned}
V_{\text{new}}(s) &= \frac{S(s) + G_t}{N(s) + 1} \qquad \text{(MC)} \\
&= \frac{N(s)\, V_{\text{old}}(s) + G_t}{N(s) + 1} = \frac{(N(s) + 1)\, V_{\text{old}}(s) + (G_t - V_{\text{old}}(s))}{N(s) + 1}.
\end{aligned}
$$

Simplify

$$
V(S_t) \leftarrow V(S_t) + \frac{1}{N(s) + 1}\left(G_t - V(S_t)\right) \qquad \text{(MC)}
$$

# Updates I: Monte Carlo

Let $V_{\text{old}}\left(s\right)=\frac{S(s)}{N(s)}$ and do one more update

$$V_{\text{new}}\left(s\right)=\frac{S\left(s\right)+G_t}{N\left(s\right)+1} \tag{MC}$$

$$=\frac{N\left(s\right)V_{\text{old}}\left(s\right)+G_t}{N\left(s\right)+1}=\frac{\left(N\left(s\right)+1\right)V_{\text{old}}\left(s\right)+\left(G_t-V_{\text{old}}\left(s\right)\right)}{N\left(s\right)+1}.$$

Simplify

$$V\left(S_t\right)\leftarrow V\left(S_t\right)+\frac{1}{N\left(s\right)+1}\left(G_t-V\left(S_t\right)\right) \tag{MC}$$

Replace $\frac{1}{N(s)+1}$ with some $\alpha\in\left(0,1\right)$

# Updates I: Monte Carlo

Let $V_{\text{old}}(s) = \frac{S(s)}{N(s)}$ and do one more update

$$\begin{aligned} V_{\text{new}}(s) &= \frac{S(s) + G_t}{N(s) + 1} \\ &= \frac{N(s)\,V_{\text{old}}(s) + G_t}{N(s) + 1} = \frac{(N(s) + 1)\,V_{\text{old}}(s) + (G_t - V_{\text{old}}(s))}{N(s) + 1}. \end{aligned} \tag{MC}$$

Simplify

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(s) + 1}\left(G_t - V(S_t)\right) \tag{MC}$$

Replace $\frac{1}{N(s)+1}$ with some $\alpha \in (0, 1)$

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right) \tag{$\alpha$-MC}$$

# Updates II: TD-n

Replace $G_t$ with $G_t^{(1)} := \underbrace{R_{t+1} + \gamma V\left(S_{t+1}\right)}$

# Updates II: TD-n

Replace $G_t$ with $G_t^{(1)} := \underbrace{R_{t+1} + \gamma V\left(S_{t+1}\right)}$

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t^{(1)} - V\left(S_t\right)\right), \tag{TD}$$

where $\delta_t := R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right)$ is called the 1-step TD error.

## Updates II: TD-n

Replace $G_t$ with $G_t^{(1)} := \underbrace{R_{t+1} + \gamma V(S_{t+1})}$

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{(1)} - V(S_t) \right), \tag{TD}$$

where $\delta_t := R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ is called the 1-step TD error.

Replace $G_t^{(1)}$ with $G_t^{(2)} := R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{(2)} - V(S_t) \right). \tag{TD-2}$$

# Updates II: TD-n

Replace $G_t$ with $G_t^{(1)} := \underbrace{R_{t+1} + \gamma V\left(S_{t+1}\right)}$

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t^{(1)} - V\left(S_t\right)\right), \tag{TD}$$

where $\delta_t := R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right)$ is called the 1-step TD error.

Replace $G_t^{(2)}$ with $G_t^{(n)} := R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V\left(S_{t+n}\right)$

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t^{(n)} - V\left(S_t\right)\right). \tag{TD-n}$$

# Updates II: TD-n

Replace $G_t$ with $G_t^{(1)} := \underbrace{R_{t+1} + \gamma V(S_{t+1})}$

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{(1)} - V(S_t) \right), \tag{TD}$$

where $\delta_t := R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ is called the 1-step TD error.

Replace $G_t^{(2)}$ with $G_t^{(n)} := R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{(n)} - V(S_t) \right). \tag{TD-n}$$

Replace $G_t^{(n)}$ with $G_t^{(\infty)} := G_t$ and you get again $(MC)$.

$$V(S_t) \leftarrow V(S_t) + \alpha \left( G_t^{(\infty)} - V(S_t) \right). \tag{MC}$$

## Update III: $TD\,(\lambda)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}$, $n = 1, 2, \ldots, \infty$

$$G_t^{\lambda} := (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

# Update III: $TD(\lambda)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}$, $n = 1, 2, \ldots, \infty$

$$G_t^\lambda := (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \overset{\substack{\text{finite horizon } T\\ \downarrow}}{=} (1-\lambda)\sum_{n=1}^{T-1}\lambda^{n-1}G_t^{(n)} + \lambda^{T-1}G_t,$$

for some $\lambda \in [0,1]$ (geometric weights)

# Update III: $TD(\lambda)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}, n = 1, 2, \ldots, \infty$

$$G_t^\lambda := (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)} \overset{\substack{\text{finite horizon } T \\ \downarrow}}{\doteq} (1-\lambda)\sum_{n=1}^{T-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-1} G_t,$$

for some $\lambda \in [0, 1]$ (geometric weights)

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right). \qquad (\text{TD}(\lambda))$$

# Update III: $TD(\lambda)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}$, $n = 1, 2, \ldots, \infty$

$$G_t^\lambda := (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)} \overset{\text{finite horizon } T}{\overset{\downarrow}{=}} (1-\lambda)\sum_{n=1}^{T-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-1} G_t,$$

for some $\lambda \in [0, 1]$ (geometric weights)

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right). \tag{TD($\lambda$)}$$

For $\lambda = 0$, it holds that $G_t^0 = G_t^{(1)} + \sum_{n=2}^{T-1} \cancel{\lambda^{n-1} G_t^{(n)}} + \cancel{\lambda^{T-1} G_t} = G_t^{(1)}$ so,

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^{(1)} - V(S_t)\right). \tag{TD(0) $\equiv$ TD}$$

## Update III: $TD(\lambda)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}$, $n = 1, 2, \dots, \infty$

$$G_t^\lambda := (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \overset{\substack{\text{finite horizon } T\\ \downarrow}}{\doteq} (1-\lambda)\sum_{n=1}^{T-1}\lambda^{n-1}G_t^{(n)} + \lambda^{T-1}G_t,$$

for some $\lambda \in [0,1]$ (geometric weights)

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right). \qquad (\mathrm{TD}(\lambda))$$

For $\lambda = 0$, it holds that $G_t^0 = G_t^{(1)} + \sum_{n=2}^{T-1}\cancel{\lambda^{n-1}G_t^{(n)}} + \cancel{\lambda^{T-1}G_t} = G_t^{(1)}$ so,

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^{(1)} - V(S_t)\right). \qquad (\mathrm{TD}(0) \equiv \mathrm{TD})$$

For $\lambda = 1$, it holds that $G_t^1 = \cancel{(1-\lambda)\sum_{n=1}^{T-1}\lambda^{n-1}G_t^{(n)}} + G_t = G_t$ so,

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right). \qquad (\mathrm{TD}(1) \equiv \mathrm{MC})$$

# Update III: $TD\left(\lambda\right)$

Replace $G_t$ with arbitrary convex combinations of $G_t^{(n)}$, $n = 1, 2, \ldots, \infty$

$$G_t^{\lambda} := (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)} \overset{\substack{\text{finite horizon } T\\ \downarrow}}{=} (1-\lambda)\sum_{n=1}^{T-1}\lambda^{n-1}G_t^{(n)} + \lambda^{T-1}G_t,$$

for some $\lambda \in [0,1]$ (geometric weights)

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t^{\lambda} - V\left(S_t\right)\right). \qquad (\text{TD}(\lambda))$$

For $\lambda = 0$, it holds that $G_t^0 = G_t^{(1)} + \cancel{\sum_{n=2}^{T-1}\lambda^{n-1}G_t^{(n)}} + \cancel{\lambda^{T-1}G_t} = G_t^{(1)}$ so,

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t^{(1)} - V\left(S_t\right)\right). \qquad (\text{TD}(0) \equiv \text{TD})$$

For $\lambda = 1$, it holds that $G_t^1 = \cancel{(1-\lambda)\sum_{n=1}^{T-1}\lambda^{n-1}G_t^{(n)}} + G_t = G_t$ so,

$$V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha\left(G_t - V\left(S_t\right)\right). \qquad (\text{TD}(1) \equiv \text{MC})$$

# Offline Updates, Forward View

Common methodology up to now. Initialize: $S_t$ and $V(s), s \in \mathcal{S}$

- Forward view: take some action from $S_t$ and observe rewards (far or near) in the future.
- Offline updates: When you finish with the forward view (end of episode or $n$ steps + estimate), return to $S_t$ and update $V(S_t)$ according to this information.

# Online Updates, Backward View

Obvious twists/improvements

- Backward view: once I learn something about $V(S_t)$, then implicitly, I learned something about $S_{t-1}$, so I will update $S_{t-1}$ as well.
- Online updates: update knowledge (here values of states) in visited states as it comes in, i.e., online.

In sum:

- Collect information as before.
- Allow information to propagate to previously visited states.

# Backward View: Intuition

Starting from a state $S_t$ and value estimates $V(s), s \in \mathcal{S}$

- Take an action and observe some reward ($G_t$, $G_t^{(1)}$ etc.).
- Update $V(S_t)$ accordingly.
- Observation: recent actions (states) before $S_t$ are also responsible for this reward, with *more recent states being more responsible.*
- Observation: frequent actions (states) before $S_t$ are also responsible for the outcome, with *more frequent states being more responsible.*

# Backward View: Intuition

Starting from a state $S_t$ and value estimates $V(s), s \in \mathcal{S}$

- Take an action and observe some reward ($G_t$, $G_t^{(1)}$ etc.).
- Update $V(S_t)$ accordingly.
- Observation: recent actions (states) before $S_t$ are also responsible for this reward, with *more recent states being more responsible.*
- Observation: frequent actions (states) before $S_t$ are also responsible for the outcome, with *more frequent states being more responsible.*

Update frequent/recent states too using eligibility traces

$$E_0(s) = 0$$
$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbf{1}_t(s), \quad \text{for all} s \in \mathcal{S}.$$

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. Initialize: $S_1 = S, V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| | | | | |

$t = 1$

- $E_1(S) = \gamma\lambda E_0(s) + \mathbf{1}(S_1 = S) = 0 + 1 = 1.$

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. Initialize: $S_1 = S, V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| | | | | |

$t = 1$

- $E_1(S) = \gamma\lambda E_0(s) + \mathbf{1}(S_1 = S) = 0 + 1 = 1.$

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. Initialize: $S_1 = S$, $V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |

$t = 1$

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4$, $\gamma = 1$, $\alpha = 0.5$, $\lambda = 0.9$. Initialize: $S_1 = S$, $V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |

$R_2 = 3$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |

$t = 1$

- Take action at $S_1 = S$: observe $S_2 = S'$ and $R_2 = 3$.
- Calculate TD error: $\delta_1 = R_2 + \gamma V(S') - V(S) = 3 + 0 - 0 = 3$.

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4$, $\gamma = 1$, $\alpha = 0.5$, $\lambda = 0.9$. Initialize: $S_1 = S$, $V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |

$R_2 = 3$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |

$t = 1$

- Take action at $S_1 = S$: observe $S_2 = S'$ and $R_2 = 3$.
- Calculate TD error: $\delta_1 = R_2 + \gamma V(S') - V(S) = 3 + 0 - 0 = 3$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha \delta_1 E_1(s) = \begin{cases} 0 + 0.5 \cdot 3 \cdot 1 = 1.5, & \text{if } s = S, \\ 0, & \text{otherwise} \end{cases}$$

# Backward View: Example $t = 0, 1$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. Initialize: $S_1 = S, V(s) = 0$ for all $s \in \mathcal{S}$.

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |

$t = 1$

- Take action at $S_1 = S$: observe $S_2 = S'$ and $R_2 = 3$.
- Calculate TD error: $\delta_1 = R_2 + \gamma V(S') - V(S) = 3 + 0 - 0 = 3$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha\delta_1 E_1(s) = \begin{cases} 0 + 0.5 \cdot 3 \cdot 1 = 1.5, & \text{if } s = S, \\ 0, & \text{otherwise} \end{cases}$$

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1.5 | 0 |

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| | | | | |

$t = 2$

- $E_2(S) = \gamma\lambda E_1(S) + \mathbf{1}(S_2 = S) = 0.9 \cdot 1 + 0 = 0.9.$

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| | | | | |

$t = 2$

- $E_2(S) = \gamma\lambda E_1(S) + \mathbf{1}(S_2 = S) = 0.9 \cdot 1 + 0 = 0.9.$
- $E_2(S') = \gamma\lambda E_1(S') + \mathbf{1}(S_2 = S') = 0 + 1 = 1.$

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |

$t = 2$

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

$R_3 = -2.5$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |

$t = 2$

- Take action at $S_2 = S'$: observe $S_3 = S$ and $R_3 = -2.5$.

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

$\xrightarrow{} R_3 = -2.5$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| | | | | |

$t = 2$

- Take action at $S_2 = S'$: observe $S_3 = S$ and $R_3 = -2.5$.
- Calculate TD error: $\delta_2 = R_3 + \gamma V(S) - V(S') = -2.5 + 1.5 - 0 = -1$.

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

$\searrow R_3 = -2.5$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |

$t = 2$

- Take action at $S_2 = S'$: observe $S_3 = S$ and $R_3 = -2.5$.
- Calculate TD error: $\delta_2 = R_3 + \gamma V(S) - V(S') = -2.5 + 1.5 - 0 = -1$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha \delta_2 E_2(s) = \begin{cases} 0 + 0.5 \cdot (-1) \cdot 1 = -0.5, & \text{if } s = S', \\ 1.5 + 0.5 \cdot (-1) \cdot 0.9 = 1.05, & \text{if } s = S \\ 0, & \text{otherwise} \end{cases}$$

# Backward View: Example $t = 2$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_2 = S'$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| $t = 2$ | 0 | $-0.5$ | 1.05 | 0 |

$t = 2$

- Take action at $S_2 = S'$: observe $S_3 = S$ and $R_3 = -2.5$.
- Calculate TD error: $\delta_2 = R_3 + \gamma V(S) - V(S') = -2.5 + 1.5 - 0 = -1$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha \delta_2 E_2(s) = \begin{cases} 0 + 0.5 \cdot (-1) \cdot 1 = -0.5, & \text{if } s = S', \\ 1.5 + 0.5 \cdot (-1) \cdot 0.9 = 1.05, & \text{if } s = S \\ 0, & \text{otherwise} \end{cases}$$

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| $t = 2$ | 0 | $-0.5$ | 1.05 | 0 |

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |
| $t=2$ | 0 | 1 | 0.9 | 0 |
| $t=3$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1.5 | 0 |
| $t=2$ | 0 | $-0.5$ | 1.05 | 0 |
| | | | | |

$t = 3$

- $E_3(S) = \gamma\lambda E_2(S) + \mathbf{1}(S_3 = S) = 0.9 \cdot 0.9 + 1 = 1.81.$

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |
| $t=2$ | 0 | 1 | 0.9 | 0 |
| $t=3$ | | | | |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1.5 | 0 |
| $t=2$ | 0 | $-0.5$ | 1.05 | 0 |

$t = 3$

- $E_3(S) = \gamma\lambda E_2(S) + \mathbf{1}(S_3 = S) = 0.9 \cdot 0.9 + 1 = 1.81.$
- $E_3(S') = \gamma\lambda E_2(S') + \mathbf{1}(S_3 = S') = 0.9 \cdot 1 + 0 = 0.9.$

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |
| $t = 3$ | 0 | 0.9 | 1.81 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| $t = 2$ | 0 | $-0.5$ | 1.05 | 0 |
| | | | | |

$t = 3$

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9.$ $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |
| $t = 3$ | 0 | 0.9 | 1.81 | 0 |

$R_4 = 3$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| $t = 2$ | 0 | $-0.5$ | 1.05 | 0 |

$t = 3$

- Take action at $S_3 = S$: observe $S_4 = S'$ and $R_4 = 3$.

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1 | 0 |
| $t = 2$ | 0 | 1 | 0.9 | 0 |
| $t = 3$ | 0 | 0.9 | 1.81 | 0 |

$R_4 = 3$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t = 0$ | 0 | 0 | 0 | 0 |
| $t = 1$ | 0 | 0 | 1.5 | 0 |
| $t = 2$ | 0 | $-0.5$ | 1.05 | 0 |

$t = 3$

- Take action at $S_3 = S$: observe $S_4 = S'$ and $R_4 = 3$.
- Calculate TD error: $\delta_3 = R_4 + \gamma V(S') - V(S) = 3 - 0.5 - 1.05 = 1.45$.

# Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |
| $t=2$ | 0 | 1 | 0.9 | 0 |
| $t=3$ | 0 | 0.9 | 1.81 | 0 |

$R_4 = 3$

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1.5 | 0 |
| $t=2$ | 0 | $-0.5$ | 1.05 | 0 |

$t = 3$

- Take action at $S_3 = S$: observe $S_4 = S'$ and $R_4 = 3$.
- Calculate TD error: $\delta_3 = R_4 + \gamma V(S') - V(S) = 3 - 0.5 - 1.05 = 1.45$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha\delta_3 E_3(s) = \begin{cases} -0.5 + 0.5 \cdot 1.45 \cdot 0.9 = 0.15, & \text{if } s = S', \\ 1.05 + 0.5 \cdot 1.45 \cdot 1.81 = 2.36, & \text{if } s = S \\ 0, & \text{otherwise} \end{cases}$$

## Backward View: Example $t = 3$

Space $|\mathcal{S}| = 4, \gamma = 1, \alpha = 0.5, \lambda = 0.9$. $S_3 = S$

| $E_t(s)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1 | 0 |
| $t=2$ | 0 | 1 | 0.9 | 0 |
| $t=3$ | 0 | 0.9 | 1.81 | 0 |

| $V(S_t)$ | | $S'$ | $S$ | |
|---|---|---|---|---|
| $t=0$ | 0 | 0 | 0 | 0 |
| $t=1$ | 0 | 0 | 1.5 | 0 |
| $t=2$ | 0 | $-0.5$ | 1.05 | 0 |
| $t=3$ | 0 | 0.15 | 2.36 | 0 |

$t = 3$

- Take action at $S_3 = S$: observe $S_4 = S'$ and $R_4 = 3$.
- Calculate TD error: $\delta_3 = R_4 + \gamma V(S') - V(S) = 3 - 0.5 - 1.05 = 1.45$.
- Update $V(s)$ for all $s \in S$

$$V(s) \leftarrow V(s) + \alpha\delta_3 E_3(s) = \begin{cases} -0.5 + 0.5 \cdot 1.45 \cdot 0.9 = 0.15, & \text{if } s = S', \\ 1.05 + 0.5 \cdot 1.45 \cdot 1.81 = 2.36, & \text{if } s = S \\ 0, & \text{otherwise} \end{cases}$$

## Remarks I

For $\lambda = 0$, it holds that $E_t(s) = \cancel{\gamma \cdot 0 \cdot E_{t-1}(s)} + \mathbf{1}(S_t = s) = \mathbf{1}(S_t = s)$, so

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t = V(S_t) + \alpha\left(G_t^{(1)} - V(S_t)\right). \tag{TD(0)}$$

# Remarks I

For $\lambda = 0$, it holds that $E_t(s) = \cancel{\gamma \cdot 0 \cdot E_{t-1}(s)} + \mathbf{1}(S_t = s) = \mathbf{1}(S_t = s)$, so

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t = V(S_t) + \alpha\left(G_t^{(1)} - V(S_t)\right). \tag{TD(0)}$$

For general $\lambda \in (0, 1]$, we have the following

**Theorem**
*The offline forward and backward TD($\lambda$) accumulate the same error, i.e.,*

$$\sum_{t=1}^{T} \alpha\delta_t E_t(s) = \alpha \sum_{t=1}^{T} \left(G_t^{\lambda} - V(S_t)\right) \mathbf{1}_t(s).$$

*Special case, $\lambda = 1$, then TD(1) and MC accumulate the same error.*

# Remarks II

**Proof.**

Step 1: Show that

$$\sum_{k=t}^{T} (\gamma\lambda)^{k-t} \delta_k = G_t^{\lambda} - V(S_t),$$

using telescoping sum (slides), where $\delta_k := R_{k+1} + \gamma V(S_{t+1}) - V(S_t)$.

Step 2: By definition of $E_t$, it holds that

$$\begin{aligned}
E_t(s) &= (\gamma\lambda) E_{t-1}(s) + \mathbf{1}_t(s) \\
&= (\gamma\lambda)\left[(\gamma\lambda) E_{t-2}(s) + \mathbf{1}_{t-1}(s)\right] + \mathbf{1}_t(s) \\
&= (\gamma\lambda)^2 E_{t-2}(s) + (\gamma\lambda)\mathbf{1}_{t-1}(s) + \mathbf{1}_t(s) = \dots \\
&= (\gamma\lambda)^t \overset{=0}{\cancel{E_0(s)}} + \sum_{j=1}^{t} (\gamma\lambda)^{t-j} \mathbf{1}_j(s) = \sum_{j=1}^{t} (\gamma\lambda)^{t-j} \mathbf{1}_j(s), \forall s \in S.
\end{aligned}$$

# Remarks II

So,

$$\begin{aligned}
\sum_{t=1}^{T} \alpha \delta_t E_t(s) &= \alpha \sum_{t=1}^{T} \delta_t \sum_{k=1}^{t} (\gamma\lambda)^{t-k} \mathbf{1}_k(s) && \mid \text{change summation order} \\
&= \alpha \sum_{k=1}^{T} \left( \sum_{t=k}^{T} (\gamma\lambda)^{t-k} \delta_t \right) \mathbf{1}_k(s) && \mid \text{use Step 1} \\
&= \alpha \sum_{k=1}^{T} \left( G_k^{\lambda} - V(S_k) \right) \mathbf{1}_k(s) .
\end{aligned}$$