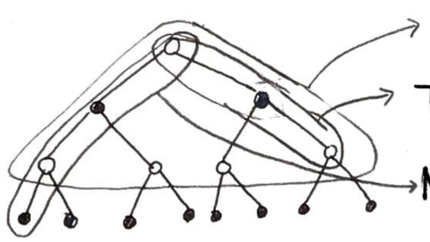
Model-Free Prediction < MC / TD (Model Free: no knowledge of MDP transitions)

- Basic : MC vs TD vs DP



→ DP

→ TD $V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$.

→ MC $V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$

* Monte-Carlo Policy Evaluation : random sampling → average sample returns,

- learn V_π from episodes of experience under policy π (offline)
  ↳ uses empirical mean return ⇐ 끝까지 가서 G_t 구하고 난 후 update ⇒ Sampling의 평균

$$V(S_t) \longleftarrow V(S_t) + \alpha(G_t - V(S_t))$$

< First-Visit Monte-Carlo : first time-step (처음) t that state S is visited in an episode.
  Every-Visit Monte-Carlo : every time-step (아무때나) t that state S is visited in an episode

To evaluate state S

$$N(S) \longleftarrow N(S) + 1$$

$$S(s) \longleftarrow S(s) + G_t$$

$$V(s) = S(s) / N(s) \text{ (mean return)}$$

⇒ $V(s) \rightarrow V_\pi(s)$ as $N(s) \rightarrow \infty$

episode by episode, updated

→ MC vs TD

MC	TD ↗ MC ⊃ TD라 생각
learn before knowing final outcome	must wait until end of episode before return is known.
high variance, zero bias ↳ noisy 하지만 후에 수렴.	low variance, some bias ↳ 안정적으로 빨리 수렴.
shallow, sample backup	deep, sample backup.
× bootstrap (다 계산 후 수정)	o bootstrap (추정 후 나아감)
o Samples	o samples
	n-step에 n→∞면 MC로 귀결.

+) (Bootstrap : update involves an estimate.
  Sample : update samples an expectation

* Temporal Difference Learning : n-step sample → average / weight returns

- learn v_π online from experience under policy π

value $V(S_t)$

TD(0) : $V(S_t) \leftarrow V(S_t) + \alpha \, (\overbrace{\overbrace{R_{t+1} + \gamma V(S_{t+1})}^{\text{TD target}} - V(S_t)}^{\text{TD error}})$

n-step TD : $V(S_t) \longleftarrow V(S_t) + \alpha \, (G_t^{(n)} - V(S_t))$

← n step 까지의 return (n-return ⇒ λ-return)

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

forward-view TD(λ) : $V(S_t) \longleftarrow V(S_t) + \alpha \, (G_t^{\lambda} - V(S_t))$

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

Backward-view TD(λ) : $V(S_t) \longleftarrow V(S_t) + \alpha \, \delta_t \, E_t(S)$

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

* $\boxed{\text{Backward View TD}(\lambda)}$

→ Eligibility Trace.

$E_0(s) = 0$

$E_t(s) = \underbrace{\gamma\lambda E_{t-1}(s)}_{\text{Frequency}} + \underbrace{1(S_t = s)}_{\text{Recency}}$

heuristic

→ 정의 $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$: TD Error

$V(s) \leftarrow V(s) + \alpha \underbrace{\delta_t}_{\text{TD error}} \underline{E_t(s)}$ eligibility trace.

if (TD(0))

$\lambda = 0$ $\quad E_t(s) = 1(S_t = s)$

(only current state is updated) $\quad V(s) \leftarrow V(s) + \alpha \underbrace{\delta_t}_{\text{TD error}} \underbrace{\cancel{E_t(s)}}_{1}$

☆ if (TD(1))

$\lambda = 1$ $\quad E_t(s) = \gamma E_{t-1}(s) + 1(S_t = s) = \begin{cases} 0 & \text{if } t < k \\ \gamma^{t-k} & t \geq k \end{cases}$

$\sum_{t=1}^{T-1} \alpha \delta_t E_t(s) = \alpha \sum_{t=1}^{T-1} \underbrace{\gamma^{t-k}}_{\text{discount}} \underbrace{\delta_t}_{\text{TD error}} = \alpha(G_k - V(S_k))$

$\therefore V(s) \leftarrow V(s) + \alpha(G_k - V(S_k))$ → Every-Visit MC와 동일!

즉, offline인 MC가 (online) Backward View TD로 환원될수있다.