actor: policy gradient: eligibility trace

starting from s_t ⟶ maximize expected return $V(s_t)$, but "single episode"

(1) $\Delta\theta \sim \frac{\partial}{\partial\theta} \ln \pi(a_t|s_t,\theta)\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots\right]$

$+\gamma \frac{\partial}{\partial\theta} \ln \pi(a_{t+1}|s_{t+1},\theta)\left[0 + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots\right]$

$+\gamma^2 \frac{\partial}{\partial\theta} \ln \pi(a_{t+2}|s_{t+2},\theta)\left[0 + 0 + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots\right]$

$+\gamma^3 \frac{\partial}{\partial\theta} \ln \pi(a_{t+3}|s_{t+3},\theta)\left[0 + 0 + 0 + \gamma^3 r_{t+3} + \dots\right]$

$+ \dots$

same episode but starting from s_{t+1} ⟶ maximize $V(s_{t+1})$

(2) $\Delta\theta \sim \frac{\partial}{\partial\theta} \ln \pi(a_{t+1}|s_{t+1},\theta)\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots\right]$

$+\gamma \frac{\partial}{\partial\theta} \ln \pi(a_{t+2}|s_{t+2},\theta)\left[0 + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots\right]$

$+\gamma^2 \frac{\partial}{\partial\theta} \ln \pi(a_{t+3}|s_{t+3},\theta)\left[0 + 0 + \gamma^2 r_{t+3} + \dots\right]$

$+ \dots$

same episode but starting from s_{t+2}

(3) $\Delta\theta \sim \frac{\partial}{\partial\theta} \ln \pi(a_{t+2}|s_{t+2},\theta)\left[r_{t+2} + \gamma r_{t+3} + \dots\right]$

$+\gamma \frac{\partial}{\partial\theta} \ln \pi(a_{t+3}|s_{t+3},\theta)\left[0 + r_{t+3} + \dots\right] + \dots$

Total parameter change is the sum of (1) + (2) + (3) + ...

let us now reorder terms: at time $(t+2)$ reward r_{t+2} was delivered

$\alpha_{eff} = \alpha_0 \cdot [1 + \gamma + \gamma^2 + \dots]$

$\Delta\theta = \alpha_0 \cdot r_{t+2}\left[\frac{\partial}{\partial\theta} \ln \pi(a_{t+2}|s_{t+2},\theta)\left\{1 + \gamma + \gamma^2 + \dots\right\}\right.$

$+ \frac{\partial}{\partial\theta} \ln \pi(a_{t+1}|s_{t+1},\theta)\,\gamma\left\{1 + \gamma + \dots\right\}$

$\left. + \frac{\partial}{\partial\theta} \ln \pi(a_t|s_t,\theta)\,\gamma^2\left\{1 + \dots\right\}\right]$

update weights at arbitrary time t':

$\Delta\theta = \alpha^{eff} \cdot r_t\left[\frac{\partial}{\partial\theta} \ln \pi(a_{t'}|s_{t'},\theta) + \gamma \frac{\partial}{\partial\theta} \ln \pi(a_{t'-1}|s_{t'-1},\theta) + \gamma^2 \dots\right.$