Proximal Policy Optimization

Reinforcement Learning

November 17, 2020

MSU

Reminder: Lower Bound

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) :=$$

Reminder: Lower Bound

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{data generated by } \pi^{\text{old}}} \overbrace{\frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}}^{\text{importance sampling correction}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\text{do not require fresh critic}}$$

Reminder: Lower Bound

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a\sim\pi^{\text{old}}(a|s)}}_{\text{data generated by } \pi^{\text{old}}} \overbrace{\frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}}^{\text{importance sampling correction}} \underbrace{A^{\pi^{\text{old}}}(s,a)}_{\text{do not require fresh critic}}$$

Approximation Error Bound:

$$C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)$$

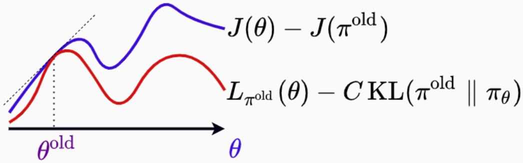
Reminder: Lower Bound

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{data generated by } \pi^{\text{old}}} \overbrace{\frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}}^{\text{importance sampling correction}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\text{do not require fresh critic}}$$

Approximation Error Bound:

$C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)$

$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$



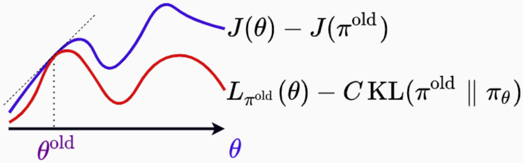
Reminder: Lower Bound

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{data generated by } \pi^{\text{old}}} \overbrace{\frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}}^{\text{importance sampling correction}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\text{do not require fresh critic}}$$

Approximation Error Bound:
$C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)$

- × can't compute constant C;
- × theoretically it is huge;
- × critic is imperfect;

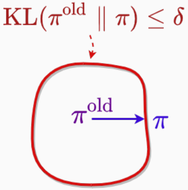
$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$



Reminder: Trust Region Policy Optimization (TRPO)

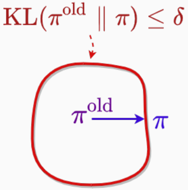
$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \to \max\limits_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \leqslant \delta \end{cases}$$

✓ robust: prevents large changes;



Reminder: Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \to \max\limits_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \leqslant \delta \end{cases}$$

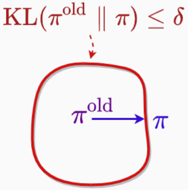


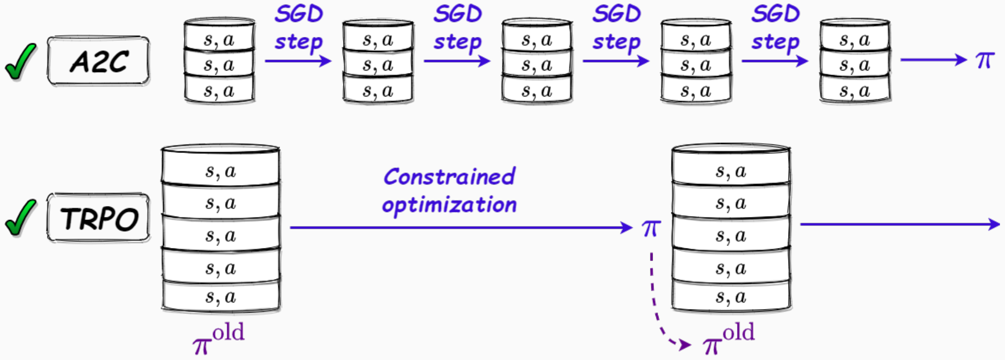
✓ robust: prevents large changes;

Reminder: Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \to \max\limits_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \leqslant \delta \end{cases}$$

✓ robust: prevents large changes;

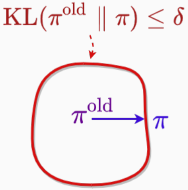




Reminder: Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \to \max\limits_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \leqslant \delta \end{cases}$$

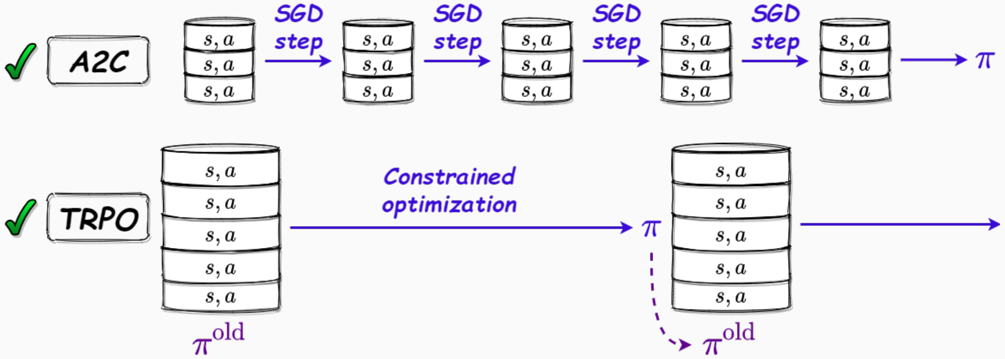
✓ robust: prevents large changes;



× critic and actor can't share backbone;

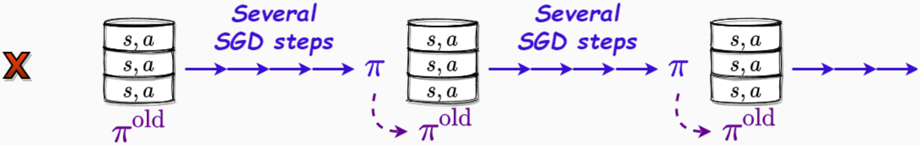
× computationally costly;

× complicated :(



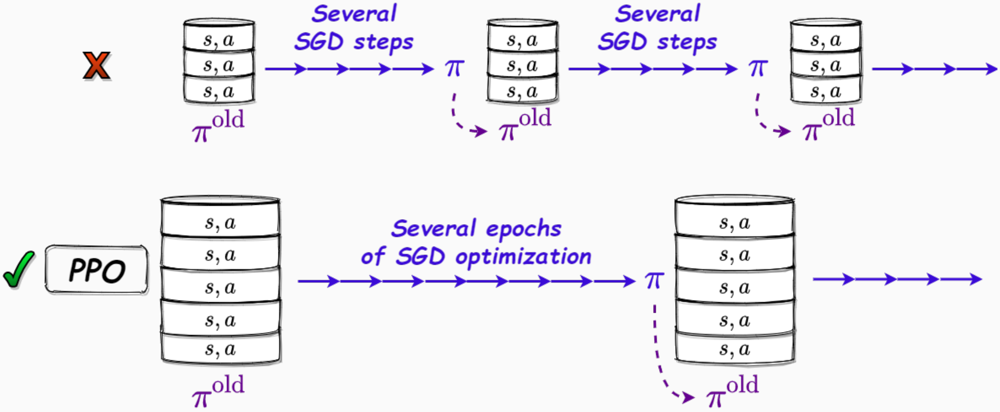
Proximal Policy Optimization (PPO): Pipeline

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)}\frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}A^{\pi^{\text{old}}}(s,a) - C\,\text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$



Proximal Policy Optimization (PPO): Pipeline

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)} A^{\pi^{\text{old}}}(s, a) - C\, \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$



Clipping Objective

$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$

Default surrogate function:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

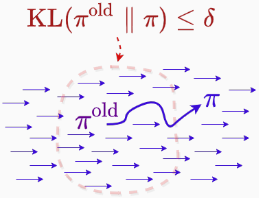
$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho(\theta)A^{\pi^{\text{old}}}(s, a)$$

$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$

Default surrogate function:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho(\theta)A^{\pi^{\text{old}}}(s,a)$$



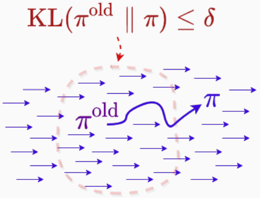
Clipping Objective

$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$

Default surrogate function:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho(\theta)A^{\pi^{\text{old}}}(s, a)$$



Clipped surrogate function:

$$\rho^{\text{clip}}(\theta) := \text{clip}(\rho(\theta), 1 - \epsilon, 1 + \epsilon)$$

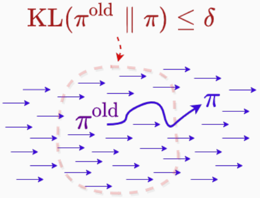
$$L^{\text{clip}}_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s, a)$$

$$L_{\pi^{\text{old}}}(\theta) - C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta) \to \max_\theta$$

Default surrogate function:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

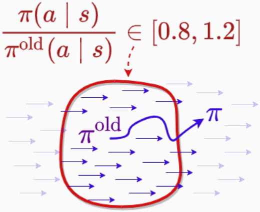
$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho(\theta)A^{\pi^{\text{old}}}(s,a)$$



Clipped surrogate function:

$$\rho^{\text{clip}}(\theta) := \text{clip}(\rho(\theta), 1-\epsilon, 1+\epsilon)$$

$$L^{\text{clip}}_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a}\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)$$



Recalling lower bound intuition

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min\Big(\overbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{original term}},\quad \overbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}^{\substack{\text{term with clipped}\\ \text{importance sampling weight}}}\Big) - \overbrace{C\,\text{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Recalling lower bound intuition

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min(\overbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}^{\text{original term}}, \overbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}^{\substack{\text{term with clipped} \\ \text{importance sampling weight}}}) - \overbrace{C \, \mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s, a) \geqslant 0$			

Recalling lower bound intuition

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min\Big(\overbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{original term}},\ \overbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}^{\substack{\text{term with clipped}\\ \text{importance sampling weight}}}\Big) - \overbrace{C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$		

Recalling lower bound intuition

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min\Big(\overbrace{\rho(\theta) A^{\pi^{\text{old}}}(s,a)}^{\text{original term}}, \overbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s,a)}^{\text{term with clipped importance sampling weight}} \Big) - \overbrace{C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	

Recalling lower bound intuition

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min\Big(\overbrace{\rho(\theta) A^{\pi^{\text{old}}}(s,a)}^{\text{original term}}, \overbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s,a)}^{\substack{\text{term with clipped} \\ \text{importance sampling weight}}}\Big) - \overbrace{C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	0

Recalling lower bound intuition

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min(\underbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}_{\text{original term}}, \underbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}_{\text{term with clipped importance sampling weight}}) - \underbrace{C\,\mathrm{KL}(\pi^{\text{old}} \,\|\, \pi_\theta)}_{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0

Recalling lower bound intuition

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min\big(\overbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{original term}},\ \overbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{term with clipped importance sampling weight}}\big) - \overbrace{C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	0
		$\rho(\theta) < 0.8$	same

Recalling lower bound intuition

$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min\Big(\underbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}_{\text{original term}},\ \underbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}_{\substack{\text{term with clipped}\\ \text{importance sampling weight}}}\Big) - \underbrace{C\,\mathrm{KL}(\pi^{\text{old}}\parallel\pi_\theta)}_{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	0
		$\rho(\theta) < 0.8$	same
$A^{\pi^{\text{old}}}(s,a) < 0$	$\pi_\theta(a \mid s) \Downarrow$	$\rho(\theta) > 1.2$	
		$\rho(\theta) < 0.8$	

Recalling lower bound intuition

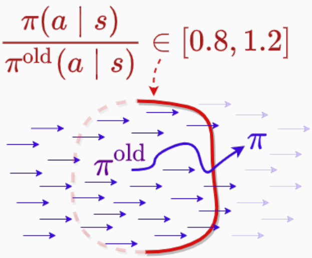
$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min(\overbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{original term}},\ \overbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}^{\substack{\text{term with clipped}\\ \text{importance sampling weight}}}) - \overbrace{C\,\mathrm{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	0
		$\rho(\theta) < 0.8$	same
$A^{\pi^{\text{old}}}(s,a) < 0$	$\pi_\theta(a \mid s) \Downarrow$	$\rho(\theta) > 1.2$	same
		$\rho(\theta) < 0.8$	0

Recalling lower bound intuition

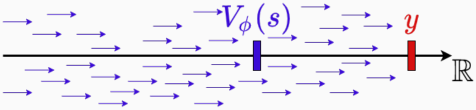
$$\mathbb{E}_{s\sim d_{\pi^{\text{old}}}(s)}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)} \min\Big(\overbrace{\rho(\theta)A^{\pi^{\text{old}}}(s,a)}^{\text{original term}},\ \overbrace{\rho^{\text{clip}}(\theta)A^{\pi^{\text{old}}}(s,a)}^{\substack{\text{term with clipped}\\ \text{importance sampling weight}}}\Big) - \overbrace{C\,\text{KL}(\pi^{\text{old}} \parallel \pi_\theta)}^{\text{«regularization»}} \to \max_\theta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi^{\text{old}}}(s,a) \geqslant 0$	$\pi_\theta(a \mid s) \Uparrow$	$\rho(\theta) > 1.2$	0
		$\rho(\theta) < 0.8$	same
$A^{\pi^{\text{old}}}(s,a) < 0$	$\pi_\theta(a \mid s) \Downarrow$	$\rho(\theta) > 1.2$	same
		$\rho(\theta) < 0.8$	0



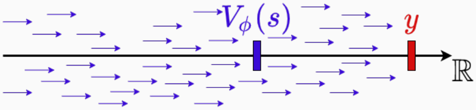
Clipped Critic Loss

$\text{Loss}(\phi) := (y - V^{\pi}(\phi))^2 =$



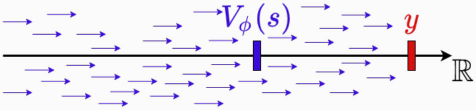
Clipped Critic Loss

$$\text{Loss}(\phi) := (y - V^{\pi}(\phi))^2 =$$
$$= (y - V^{\text{old}} + V^{\text{old}} - V^{\pi}(\phi))^2$$

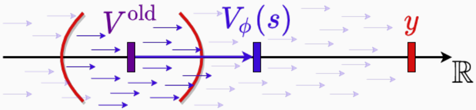


Clipped Critic Loss

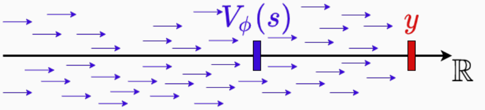
$$\text{Loss}(\phi) := (y - V^{\pi}(\phi))^2 =$$
$$= (y - V^{\text{old}} + V^{\text{old}} - V^{\pi}(\phi))^2$$



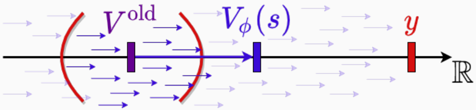
$$\text{Loss}^{\text{clip}}(\phi) := (y - V^{\text{old}} + \text{clip}(V^{\text{old}} - V^{\pi}(\phi), -\hat{\epsilon}, \hat{\epsilon}))^2$$



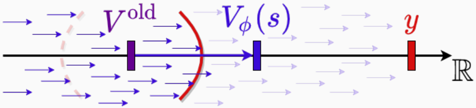
$$\text{Loss}(\phi) := (y - V^{\pi}(\phi))^2 =$$
$$= (y - V^{\text{old}} + V^{\text{old}} - V^{\pi}(\phi))^2$$



$$\text{Loss}^{\text{clip}}(\phi) := (y - V^{\text{old}} + \text{clip}(V^{\text{old}} - V^{\pi}(\phi), -\hat{\epsilon}, \hat{\epsilon}))^2$$



$$\max(\text{Loss}(\phi), \text{Loss}^{\text{clip}}(\phi))$$



Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \dots s^{(M)}$ from policy π and approximation of $V^{\pi}(s)$

Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \dots s^{(M)}$ from policy π and approximation of $V^\pi(s)$ perform **credit assingment** for state-action pair s, a (was this decision good or bad?)

Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \dots s^{(M)}$ from policy π and approximation of $V^\pi(s)$ perform **credit assingment** for state-action pair s, a (was this decision good or bad?)

For Actor:

$$\nabla := \rho(\theta)\nabla_\theta \log \pi_\theta(a \mid s) \underbrace{\Psi(s, a)}_{\text{advantage estimator}}$$

For Critic:

$$\underbrace{y_Q}_{\text{target for regression}} := \Psi(s, a) + V^\pi(s)$$

Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \ldots s^{(M)}$ from policy π and approximation of $V^\pi(s)$ perform **credit assignment** for state-action pair s, a (was this decision good or bad?)

For Actor:

$$\nabla := \rho(\theta)\nabla_\theta \log \pi_\theta(a \mid s) \underbrace{\Psi(s, a)}_{\text{advantage estimator}}$$

For Critic:

$$\underbrace{y_Q}_{\text{target for regression}} := \Psi(s, a) + V^\pi(s)$$

	$\Psi(s, a)$	**Bias**	**Variance**
Monte Carlo	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \cdots - V^\pi(s)$	0	high
1-step	$\Psi_{(1)}(s, a) := r + \gamma V^\pi(s') - V^\pi(s)$	high	low

Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \dots s^{(M)}$ from policy π and approximation of $V^\pi(s)$ perform **credit assingment** for state-action pair s, a (was this decision good or bad?)

For Actor:

$$\nabla := \rho(\theta) \nabla_\theta \log \pi_\theta(a \mid s) \underbrace{\Psi(s, a)}_{\text{advantage estimator}}$$

For Critic:

$$\underbrace{y_Q}_{\text{target for regression}} := \Psi(s, a) + V^\pi(s)$$

	$\Psi(s, a)$	**Bias**	**Variance**
Monte Carlo	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \dots - V^\pi(s)$	0	high
N-step	$\Psi_{(N)}(s, a) := r + \gamma r' + \dots + \gamma^N V^\pi(s^{(N)}) - V^\pi(s)$	intermediate	intermediate
1-step	$\Psi_{(1)}(s, a) := r + \gamma V^\pi(s') - V^\pi(s)$	high	low

Bias-Variance trade-off

Given rollout $s, r, s', r', s'', r'' \dots s^{(M)}$ from policy π and approximation of $V^\pi(s)$ perform **credit assignment** for state-action pair s, a (was this decision good or bad?)

For Actor:

$$\nabla := \rho(\theta)\nabla_\theta \log \pi_\theta(a \mid s) \underbrace{\Psi(s, a)}_{\substack{\text{advantage} \\ \text{estimator}}}$$

For Critic:

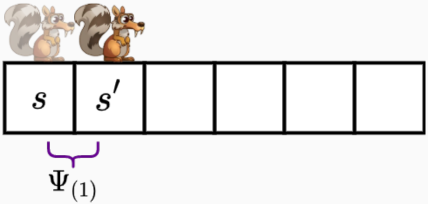
$$\underbrace{y_Q}_{\substack{\text{target} \\ \text{for regression}}} := \Psi(s, a) + V^\pi(s)$$

	$\Psi(s, a)$	**Bias**	**Variance**
Monte Carlo	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \dots - V^\pi(s)$	0	high
N-step	$\Psi_{(N)}(s, a) := r + \gamma r' + \dots + \gamma^N V^\pi(s^{(N)}) - V^\pi(s)$	intermediate	intermediate
1-step	$\Psi_{(1)}(s, a) := r + \gamma V^\pi(s') - V^\pi(s)$	high	low

Problem: hard to choose N.

N-step update:

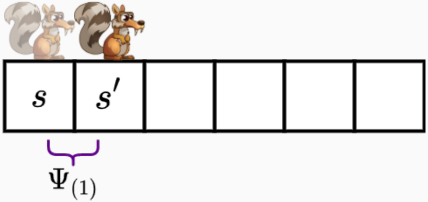
$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\Psi_{(N)}(s, a)$$



Backward view: idea

***N*-step update:**

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\Psi_{(N)}(s,a)$$

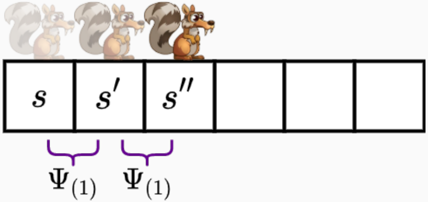


$$V^\pi(s) \leftarrow V^\pi(s) + \alpha \overbrace{\left(r + \gamma V^\pi(s') - V^\pi(s)\right)}^{\Psi_{(1)}(s,a)}$$

N-step update:

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\Psi_{(N)}(s,a)$$

How to turn 1-step update into 2-step?

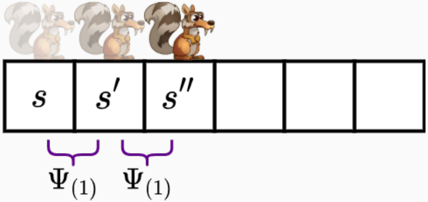


$$V^\pi(s) \leftarrow V^\pi(s) + \alpha \overbrace{\left(r + \gamma V^\pi(s') - V^\pi(s)\right)}^{\Psi_{(\mathbf{1})}(s,a)}$$

***N*-step update:**

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\Psi_{(N)}(s,a)$$

How to turn 1-step update into 2-step?



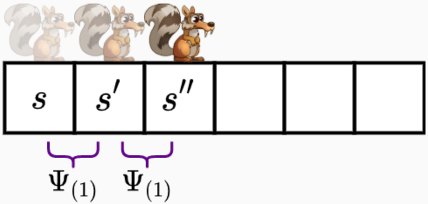
$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\overbrace{\left(r + \gamma V^\pi(s') - V^\pi(s)\right)}^{\Psi_{(1)}(s,a)} + \alpha\overbrace{\left(\gamma r' + \gamma^2 V^\pi(s'') - \gamma V^\pi(s')\right)}^{\gamma\Psi_{(1)}(s',a')}$$

Backward view: idea

N-step update:

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha \Psi_{(N)}(s, a)$$

How to turn 1-step update into 2-step?



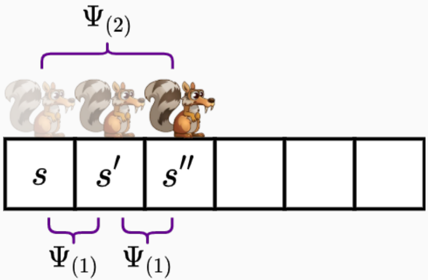
$$V^\pi(s) \leftarrow V^\pi(s) + \alpha \overbrace{\left(r + \gamma V^\pi(s') - V^\pi(s)\right)}^{\Psi_{(1)}(s,a)} + \alpha \overbrace{\left(\gamma r' + \gamma^2 V^\pi(s'') - \gamma V^\pi(s')\right)}^{\gamma \Psi_{(1)}(s',a')} =$$

Backward view: idea

N-step update:

$$V^\pi(s) \leftarrow V^\pi(s) + \alpha\Psi_{(N)}(s, a)$$

How to turn 1-step update into 2-step?



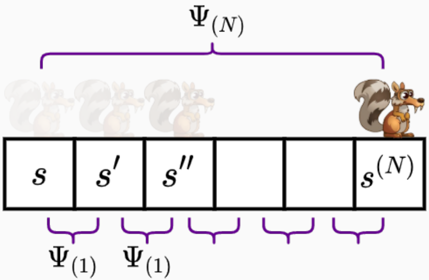
$$V^\pi(s) \leftarrow V^\pi(s) + \alpha \overbrace{\left(r + \gamma V^\pi(s') - V^\pi(s)\right)}^{\Psi_{(1)}(s,a)} + \alpha \overbrace{\left(\gamma r' + \gamma^2 V^\pi(s'') - \gamma V^\pi(s')\right)}^{\gamma\Psi_{(1)}(s',a')} =$$
$$= V^\pi(s) + \alpha\Psi_{(2)}(s, a)$$

N-step update:

$$V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha\Psi_{(N)}(s, a)$$

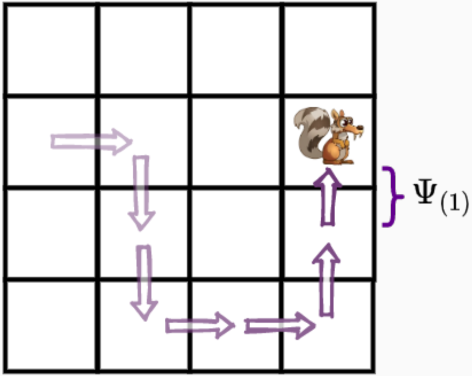
How to turn 1-step update into 2-step?

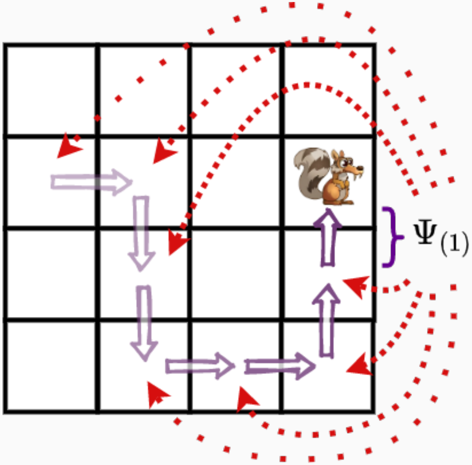
$$\Psi_{(N)}(s, a) = \sum_{t=0}^{N} \gamma^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$



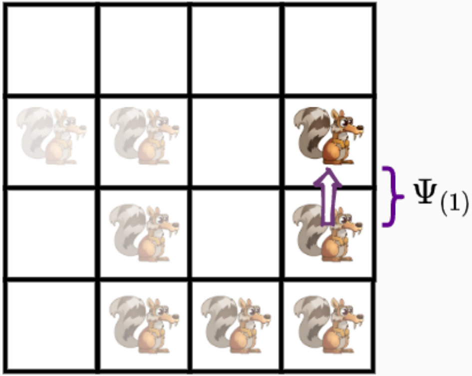
$$V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha \overbrace{\left(r + \gamma V^{\pi}(s') - V^{\pi}(s)\right)}^{\Psi_{(1)}(s,a)} + \alpha \overbrace{\left(\gamma r' + \gamma^2 V^{\pi}(s'') - \gamma V^{\pi}(s')\right)}^{\gamma\Psi_{(1)}(s',a')} =$$

$$= V^{\pi}(s) + \alpha\Psi_{(2)}(s, a)$$


$\Psi_{(1)}$



Use 1-step TD-error to update $V^{\pi}(s)$ for **all** states

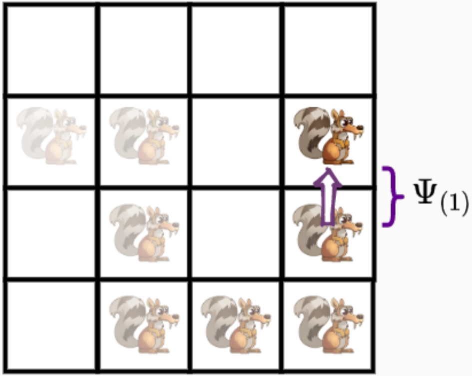


Use 1-step TD-error to update $V^\pi(s)$ for **all** states

Define **eligibility trace** $e(s)$ as a coefficient of update:

$$\forall s\colon\ V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s)\Psi_{(1)}$$

Eligibility Traces



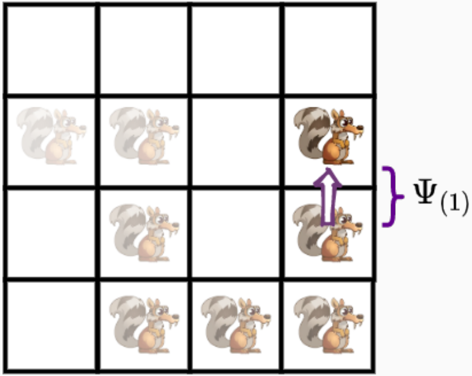
Use 1-step TD-error to update $V^{\pi}(s)$ for **all** states

Define **eligibility trace** $e(s)$ as a coefficient of update:

$$\forall s\colon\ V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha e(s)\Psi_{(1)}$$

Online «Monte-Carlo» updates:

- $\forall s\colon\ e(s) := 0$ at the start of each episode



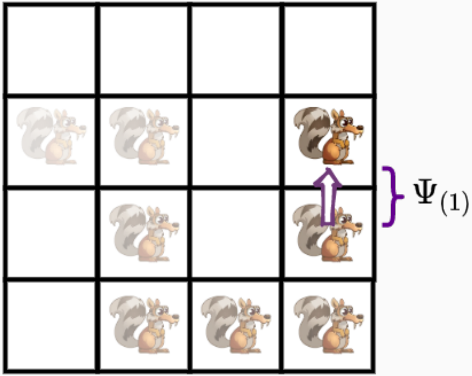
Use 1-step TD-error to update $V^\pi(s)$ for **all** states

Define **eligibility trace** $e(s)$ as a coefficient of update:

$$\forall s\colon\ V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s)\Psi_{(1)}$$

Online «Monte-Carlo» updates:

- $\forall s\colon\ e(s) := 0$ at the start of each episode
- $e(s) \leftarrow e(s) + 1$ after visiting s



Use 1-step TD-error to update $V^\pi(s)$ for **all** states

Define **eligibility trace** $e(s)$ as a coefficient of update:

$$\forall s\colon V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s)\Psi_{(1)}$$

Online «Monte-Carlo» updates:

- $\forall s\colon e(s) := 0$ at the start of each episode
- $e(s) \leftarrow e(s) + 1$ after visiting s
- $\forall s\colon e(s) \leftarrow \gamma e(s)$ after each step

TD(1) and TD(0)

TD(1)

Input: policy π

Initialize $V^\pi(s)$ arbitrarily

Initialize $e(s) = 0$

observe s_0

for $k = 0, 1, 2 \ldots$

- take action $a_k \sim \pi$, observe r_k, s_{k+1}

TD(1) and TD(0)

TD(1)

Input: policy π
Initialize $V^\pi(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2 \ldots$

- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$

TD(1) and TD(0)

TD(1)

Input: policy π
Initialize $V^\pi(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2 \ldots$

- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$

TD(1) and TD(0)

TD(1)

Input: policy π
Initialize $V^\pi(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2 \ldots$

- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s\colon\ V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s) \Psi_{(1)}$

TD(1) and TD(0)

TD(1)

Input: policy π

Initialize $V^\pi(s)$ arbitrarily

Initialize $e(s) = 0$

observe s_0

for $k = 0, 1, 2 \ldots$

- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s\colon V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s\colon e(s) \leftarrow \gamma e(s)$

TD(1) and TD(0)

TD(1)

Input: policy π
Initialize $V^\pi(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2 \ldots$

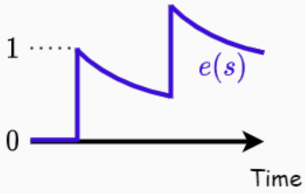
- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s\colon V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s\colon e(s) \leftarrow \gamma e(s)$

TD(0)

Input: policy π
Initialize $V^\pi(s)$ arbitrarily
Initialize $e(s) = 0$

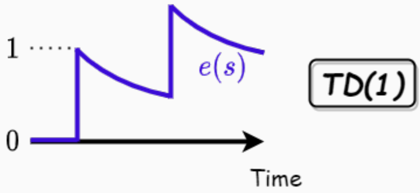
observe s_0
for $k = 0, 1, 2 \ldots$

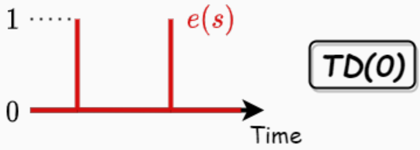
- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s\colon V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s\colon e(s) \leftarrow 0 \cdot \gamma e(s)$



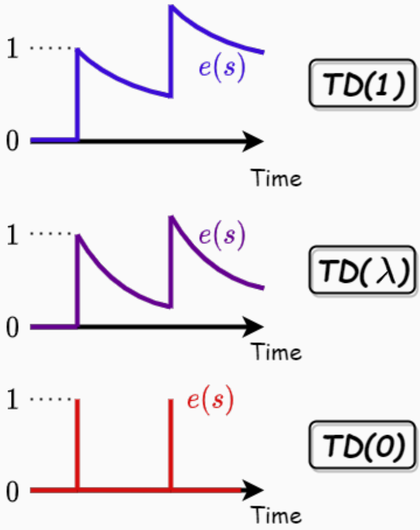
TD(1)

TD(λ)

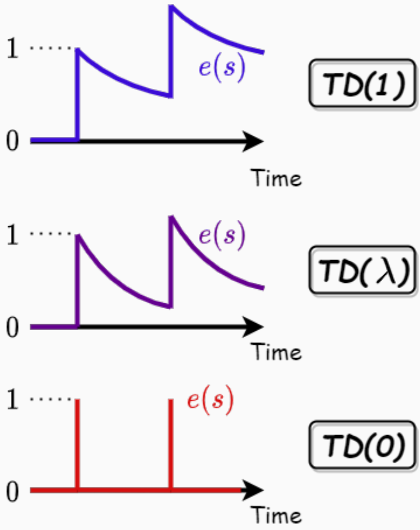




TD(λ)



TD(λ)



TD (λ)

Input: policy π

Initialize $V^\pi(s)$ arbitrarily

Initialize $e(s) = 0$

observe s_0

for $k = 0, 1, 2 \ldots$

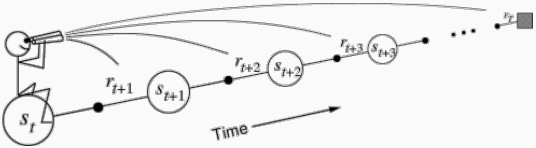
- take action $a_k \sim \pi$, observe r_k, s_{k+1}
- $\Psi_{(1)} := r_k + \gamma V^\pi(s_{k+1}) - V^\pi(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s\colon\ V^\pi(s) \leftarrow V^\pi(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s\colon\ e(s) \leftarrow \lambda \gamma e(s)$

Backward view vs Forward view

Forward View

Give credit to **present** from known **future**

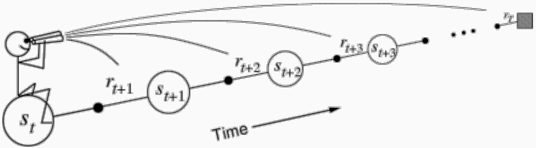
«is this decision good or bad based on the outcome?»

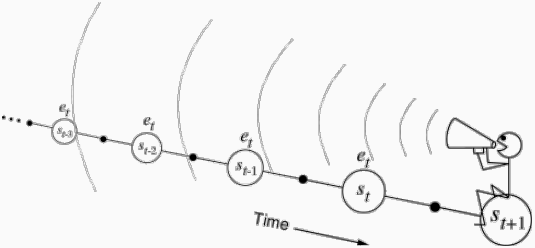


Forward View

Give credit to **present** from known **future**

«is this decision good or bad based on the outcome?»





Backward View

Update **past** credits with **present** information

«which decisions in the past to blame?»

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) =$$

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1 - \lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty}(\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda)\sum_{N=1}^{\infty}\lambda^{N-1}\Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$				

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$		

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$				

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$			

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^{t-1}\gamma - (\gamma\lambda)^t$		

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^{t-1}\gamma - (\gamma\lambda)^t$	$\Psi_{(t)}(s, a)$	

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^{t-1}\gamma - (\gamma\lambda)^t$	$\Psi_{(t)}(s, a)$	$(1-\lambda)(\lambda^{t-1}\gamma^t)$
$V^\pi(s)$				

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

$$\sum_{t=0}^{\infty}(\gamma\lambda)^t\Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda)\sum_{N=1}^{\infty}\lambda^{N-1}\Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^{t-1}\gamma - (\gamma\lambda)^t$	$\Psi_{(t)}(s, a)$	$(1-\lambda)(\lambda^{t-1}\gamma^t)$
$V^\pi(s)$	$\Psi_{(1)}(s, a)$	-1		

Forward view for TD(λ)

Equivalent forms of TD(λ) updates

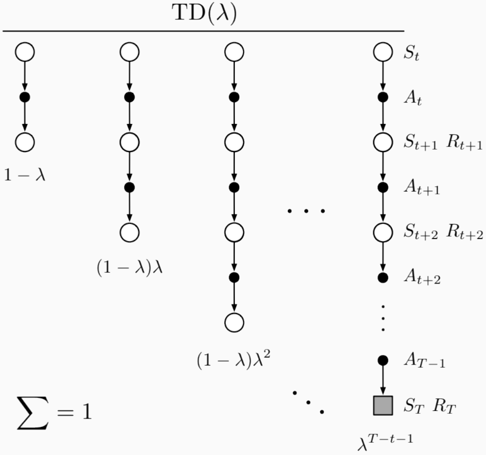
$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1-\lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

Term	Left side	Left side coeff.	Right side	Right side coeff.
$r^{(t)}$	$\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^t$	$\Psi_{(N)}(s, a)$ for $N > t$	$\gamma^t(1-\lambda)(\lambda^t + \lambda^{t+1} + \lambda^{t+2} + \dots)$
$V^\pi(s^{(t)})$ $(t > 0)$	$\Psi_{(1)}(s^{(t-1)}, a^{(t-1)})$ $\Psi_{(1)}(s^{(t)}, a^{(t)})$	$(\gamma\lambda)^{t-1}\gamma - (\gamma\lambda)^t$	$\Psi_{(t)}(s, a)$	$(1-\lambda)(\lambda^{t-1}\gamma^t)$
$V^\pi(s)$	$\Psi_{(1)}(s, a)$	-1	all	$-(1-\lambda)(1 + \lambda + \lambda^2 + \dots)$

Generalized Advantage Estimation (GAE)

$$\Psi^{\text{GAE}}(s, a) := \sum_{t=0}^{T} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

(trace is zeroed when future is not available)



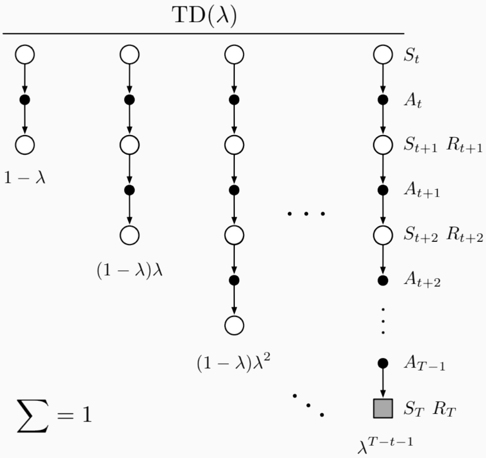
Generalized Advantage Estimation (GAE)

$$\Psi^{\mathrm{GAE}}(s, a) := \sum_{t=0}^{T} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

(trace is zeroed when future is not available)

How to compute in practice:

$$\Psi^{\mathrm{GAE}}(s_t, a_t) = \Psi_{(1)}(s_t, a_t) + \\ + \lambda\gamma(1 - \mathrm{done}_{t+1})\Psi^{\mathrm{GAE}}(s_{t+1}, a_{t+1})$$



Proximal Policy Optimization: implementation matters

Key elements:

- ✓ Clipped policy loss
- ✓ Clipped critic loss
- ✓ GAE

Pipeline details:

- ! Advantage normalization in mini-batches
- No KL regularization
- Entropy loss

Other hacks:

- ! Reward normalization[1] and clipping
- Observations normalization and clipping[2]
- Orthogonal initialization of layers
- ϵ (clipping parameter) annealing

Standard tricks:

- Adam, learning rate annealing
- Tanh activation functions
- ! Gradient clipping

[1] divided by running std of collected cumulative rewards
[2] can be critical in continuous control

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta), V^{\pi}_{\phi}(s)$;

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta), V_\phi^\pi(s)$;

for $k = 0, 1, 2 \ldots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \ldots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V_\phi^\pi(s_t)$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta), V_\phi^\pi(s)$;

for $k = 0, 1, 2 \dots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \dots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V_\phi^\pi(s_t)$
- compute 1-step errors: $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \mathrm{done}_{t+1}) V_\phi^\pi(s_{t+1}) - V_\phi^\pi(s_t)$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta), V_\phi^\pi(s)$;

for $k = 0, 1, 2 \dots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \dots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V_\phi^\pi(s_t)$
- compute 1-step errors: $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \mathrm{done}_{t+1}) V_\phi^\pi(s_{t+1}) - V_\phi^\pi(s_t)$
- compute GAE advantage estimations: $\Psi^{\mathrm{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- for t from $N - 2$ to 0:
  - $\Psi^{\mathrm{GAE}}(s_t, a_t) :=$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta), V_\phi^\pi(s)$;

for $k = 0, 1, 2 \ldots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \ldots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V_\phi^\pi(s_t)$
- compute 1-step errors: $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \mathrm{done}_{t+1}) V_\phi^\pi(s_{t+1}) - V_\phi^\pi(s_t)$
- compute GAE advantage estimations: $\Psi^{\mathrm{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- for t from $N-2$ to 0:
  - $\Psi^{\mathrm{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \mathrm{done}_{t+1})\Psi^{\mathrm{GAE}}(s_{t+1}, a_{t+1})$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta)$, $V^{\pi}_{\phi}(s)$;

for $k = 0, 1, 2 \dots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \dots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V^{\pi}_{\phi}(s_t)$
- compute 1-step errors: $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \mathrm{done}_{t+1})V^{\pi}_{\phi}(s_{t+1}) - V^{\pi}_{\phi}(s_t)$
- compute GAE advantage estimations: $\Psi^{\mathrm{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- for t from $N-2$ to 0:
  - $\Psi^{\mathrm{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \mathrm{done}_{t+1})\Psi^{\mathrm{GAE}}(s_{t+1}, a_{t+1})$
- compute critic targets: $y(s_t) := \Psi^{\mathrm{GAE}}(s_t, a_t) + V^{\pi}_{\phi}(s_t)$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $\pi(a \mid s, \theta)$, $V_\phi^\pi(s)$;

for $k = 0, 1, 2 \dots$

- collect several rollouts $s_0, a_0, r_0, s_1, \mathrm{done}_1, a_1 \dots s_N, \mathrm{done}_N$ using $\pi(a \mid s, \theta)$;
  store probabilities of selected actions as $\pi^{\mathrm{old}}(a_t \mid s_t) := \pi(a_t \mid s_t, \theta)$
  store critic output as $V^{\mathrm{old}}(s_t) := V_\phi^\pi(s_t)$
- compute 1-step errors: $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \mathrm{done}_{t+1})V_\phi^\pi(s_{t+1}) - V_\phi^\pi(s_t)$
- compute GAE advantage estimations: $\Psi^{\mathrm{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- for t from $N-2$ to 0:
  - $\Psi^{\mathrm{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \mathrm{done}_{t+1})\Psi^{\mathrm{GAE}}(s_{t+1}, a_{t+1})$
- compute critic targets: $y(s_t) := \Psi^{\mathrm{GAE}}(s_t, a_t) + V_\phi^\pi(s_t)$
- construct dataset of $(s_t, a_t, \Psi^{\mathrm{GAE}}(s_t, a_t), y(s_t), \pi^{\mathrm{old}}(a_t \mid s_t), V^{\mathrm{old}}(s_t))$

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

- go through dataset $\mathrm{n\_epochs}$ times, sampling mini-batches of size B; for each mini-batch:

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

- go through dataset $\mathrm{n\_epochs}$ times, sampling mini-batches of size B; for each mini-batch:
  - normalize $\Psi^{\mathrm{GAE}}(s, a)$ in the batch by subtracting mean and dividing by std

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

- go through dataset $\mathrm{n\_epochs}$ times, sampling mini-batches of size B; for each mini-batch:
  - normalize $\Psi^{\mathrm{GAE}}(s, a)$ in the batch by subtracting mean and dividing by std
  - compute importance sampling weights:

$$\rho(s, a, \theta) := \frac{\pi(a \mid s, \theta)}{\pi^{\mathrm{old}}(a \mid s)}, \qquad \rho^{\mathrm{clip}}(s, a, \theta) = \mathrm{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

- go through dataset n_epochs times, sampling mini-batches of size B; for each mini-batch:
  - normalize $\Psi^{\text{GAE}}(s, a)$ in the batch by subtracting mean and dividing by std
  - compute importance sampling weights:

$$\rho(s, a, \theta) := \frac{\pi(a \mid s, \theta)}{\pi^{\text{old}}(a \mid s)}, \qquad \rho^{\text{clip}}(s, a, \theta) = \text{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$

  - update actor:

$$L_1(s, a, \theta) := \rho(s, a, \theta)\Psi^{\text{GAE}}(s, a), \qquad L_2(s, a, \theta) := \rho^{\text{clip}}(s, a, \theta)\Psi^{\text{GAE}}(s, a)$$

$$\theta \leftarrow \theta + \alpha \nabla_\theta \frac{1}{B} \sum_{s,a} \min(L_1(s, a, \theta), L_2(s, a, \theta))$$

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

- go through dataset n_epochs times, sampling mini-batches of size B; for each mini-batch:
  - normalize $\Psi^{\text{GAE}}(s, a)$ in the batch by subtracting mean and dividing by std
  - compute importance sampling weights:

$$\rho(s, a, \theta) := \frac{\pi(a \mid s, \theta)}{\pi^{\text{old}}(a \mid s)}, \qquad \rho^{\text{clip}}(s, a, \theta) = \text{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$

  - update actor:

$$L_1(s, a, \theta) := \rho(s, a, \theta)\Psi^{\text{GAE}}(s, a), \qquad L_2(s, a, \theta) := \rho^{\text{clip}}(s, a, \theta)\Psi^{\text{GAE}}(s, a)$$

$$\theta \leftarrow \theta + \alpha \nabla_\theta \frac{1}{B} \sum_{s,a} \min(L_1(s, a, \theta), L_2(s, a, \theta))$$

  - update critic:

$$\text{Loss}_1(s, \phi) := \left(y(s) - V_\phi^\pi(s)\right)^2$$

$$\text{Loss}_2(s, \phi) := \left(y(s) - V^{\text{old}}(s) - \text{clip}(V_\phi^\pi(s) - V^{\text{old}}(s), \hat{\epsilon}, -\hat{\epsilon})\right)^2$$

$$\phi \leftarrow \phi - \alpha \nabla_\phi \frac{1}{B} \sum_{s} \max(\text{Loss}_1(s, \phi), \text{Loss}_2(s, \phi))$$

Literature

- Proximal Policy Optimization Algorithms;
- Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO;
- High-Dimensional Continuous Control Using Generalized Advantage Estimation;
- Sutton, Barto — Reinforcement Learning, an Introduction, ch. 12;