Proximal Policy Optimization

Reinforcement Learning

November 17, 2020

MSU

Reminder: Lower Bound

$J_p(q) \geq J_p^{old}(q) - L_{old}(p, q)$:

Reminder: Lower Bound

$$J(\theta) = J(\theta_{\text{old}}) + L_{\theta_{\text{old}}}(\theta) : \quad \frac{1}{1-\gamma}\, \mathbb{E}_{s \sim d_{\theta_{\text{old}}}(s)}\, \mathbb{E}_{a \sim \pi_{\theta_{\text{old}}}(a|s)}\, \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}\, A_{\theta_{\text{old}}}(s,a)$$

importance sampling correction

data generated by $\pi_{\theta_{\text{old}}}$

do not require fresh critic

Reminder: Lower Bound

$$J(\theta) - J(\theta_{old}) \approx L_{\theta_{old}}(\theta) : \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A^{\pi_{old}}(s,a) \right]$$

importance sampling correction

data generated by π_{old}

do not require fresh critic

Approximation Error Bound:

$$C \, \mathrm{KL}(\pi_{old}, k, \pi)$$

Reminder: Lower Bound

$$J(\theta) - J(\theta_{\text{old}}) = L_{\theta_{\text{old}}}(\theta) + \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_{\pi_{\theta_{\text{old}}}}}\,\mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(a|s)}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}\,A^{\pi_{\theta_{\text{old}}}}(s,a)\right]$$

importance sampling correction

data generated by $\pi_{\theta_{\text{old}}}$

do not require fresh critic

Approximation Error Bound:

$$L_{\theta_{\text{old}}}(\theta) \geq C\,\mathrm{KL}(\theta_{\text{old}}, \theta) \quad \text{where} \quad C = \max$$

Reminder: Lower Bound

$$J(\theta) - J(\theta_{old}) \geq L_{\theta_{old}}(\theta) - \frac{1}{1-\gamma}\underbrace{\mathbb{E}_{s,a\sim\rho_{\theta_{old}}}}_{\text{data generated by } \theta_{old}}\left[\underbrace{\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}}_{\text{importance sampling correction}} A_{\theta_{old}}(s,a)\right]$$

do not require fresh critic

Approximation Error Bound:

$$L_{\theta_{old}}(\theta) \geq C\,\mathrm{KL}(\theta_{old}, \theta) - \epsilon, \quad \epsilon \leq \max$$

$C\,\mathrm{KL}(\theta_{old}, \theta)$

can't compute constant C;

theoretically it is huge;

critic is imperfect;

Reminder: Trust Region Policy Optimization (TRPO)

$$
\begin{cases}
L_{\text{old}}(p, q) \to \max \\
\mathrm{KL}(p^{\text{old}} \| q) \leq \delta
\end{cases}
$$

X robust: prevents large changes;

Reminder: Trust Region Policy Optimization (TRPO)

$
&$L_{\text{old}}(p, q) \tilde{N} \max$

%$\text{KL}(p^{\text{old}}, k, q) \leq$

X robust: prevents large changes;

Reminder: Trust Region Policy Optimization (TRPO)

$$
\begin{cases}
L_{\text{old}}(p, q) \to \max \\
\mathrm{KL}(p^{\text{old}} \| q) \leq \delta
\end{cases}
$$

X robust: prevents large changes;

Reminder: Trust Region Policy Optimization (TRPO)

$$
\begin{cases}
L_{\text{old}}(p, q) \to \max \\
\mathrm{KL}(p^{\text{old}}, q) \le k
\end{cases}
$$

X robust: prevents large changes;

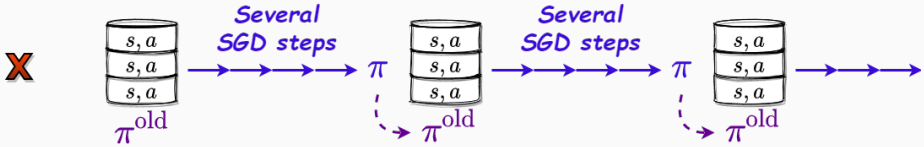
critic and actor can't share backbone;

computationally costly;

complicated :(

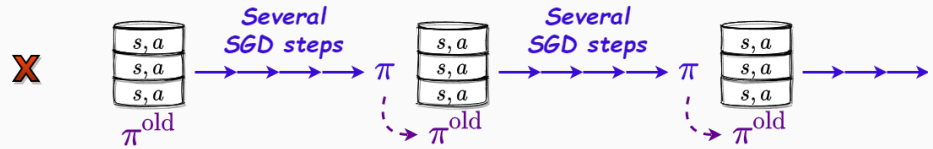
Proximal Policy Optimization (PPO): Pipeline

$$E_s \; d_{\text{old}}(s) E_a \; {}_{\text{old}}(a|s) \frac{(a \mid s)}{{}_{\text{old}}(a \mid s)} A^{\text{old}}(s,a) \quad C \; \text{KL}({}^{\text{old}},) \quad \max$$



Proximal Policy Optimization (PPO): Pipeline

$$\mathrm{E}_{s \sim d^{\pi_{\mathrm{old}}}(s)} \mathrm{E}_{a \sim \pi_{\mathrm{old}}(a|s)} \frac{\pi(a \mid s)}{\pi_{\mathrm{old}}(a \mid s)} A^{\pi_{\mathrm{old}}}(s,a) \quad C \, \mathrm{KL}(\pi_{\mathrm{old}}, \pi) \quad \mathrm{KL}(\pi_{\mathrm{old}}, \pi) \leq \max$$



Clipping Objective

$$L_{old}(p, q) \quad C\,\mathrm{KL}(p^{old}, k) \quad q \tilde{N} \max$$

Default surrogate function:

$$(p, q): \quad \frac{p(a \mid s, q)}{p^{old}(a \mid s, q)}$$

$$L_{old}(p, q): \quad E_{s;a}(p, q, A^{old}(s; a, q))$$

Clipping Objective

$$L_{old}(p, q) \quad C\,\mathrm{KL}(p^{old}, k) \quad q \tilde{N} \max$$

Default surrogate function:

$$(p, q): \quad \frac{p(a \mid s, q)}{p^{old}(a \mid s, q)}$$

$$L_{old}(p, q): \quad E_{s;a}(p, q, A^{old}(s; a, q))$$

Clipping Objective

$$L_{old}(\theta) \quad C\,\mathrm{KL}(\theta_{old}, \theta) \quad \text{Ñ} \max$$

Default surrogate function:

$$\theta : \frac{\pi_\theta(a \mid s)}{\pi_{old}(a \mid s)}$$

$$L_{old}(\theta) : \mathbb{E}_{s,a}\left(\theta, A^{old}(s,a)\right)$$

Clipped surrogate function:

$$^{clip}(\theta) : \mathrm{clip}(\theta, 1 \quad , 1 \quad \theta)$$

$$L^{clip}_{old}(\theta) : \mathbb{E}_{s,a}\left(^{clip}(\theta), A^{old}(s,a)\right)$$

Clipping Objective

$$L_{old}(\theta) \quad C\,\mathrm{KL}(\theta_{old}, \theta) \quad \text{Ñ} \max$$

Default surrogate function:

$$\theta : \frac{\pi_\theta(a \mid s)}{\pi_{old}(a \mid s)}$$

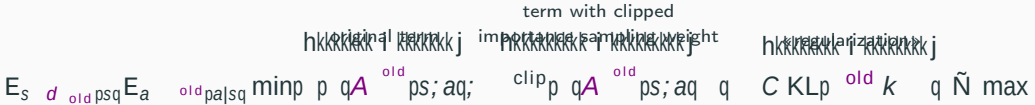
$$L_{old}(\theta) : \mathbb{E}_{s,a}\left(\theta, A^{old}(s,a)\right)$$

Clipped surrogate function:

$$^{clip}(\theta) : \mathrm{clip}(\theta, 1 \quad , 1 \quad \theta)$$

$$L^{clip}_{old}(\theta) : \mathbb{E}_{s,a}\left(^{clip}(\theta), A^{old}(s,a)\right)$$

Recalling lower bound intuition



Recalling lower bound intuition

term with clipped importance sampling weight

original term regularization

$$\mathbb{E}_{s\sim d^{\text{old}}}\mathbb{E}_{a\sim\pi^{\text{old}}(\cdot|s)}\Big[\min\Big(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)}A^{\text{old}}(s,a),\ \text{clip}\Big(\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)},1-\epsilon,1+\epsilon\Big)A^{\text{old}}(s,a)\Big)\Big] - C\,\text{KL}[\pi_{\text{old}},\pi_\theta]\le\delta$$

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

Recalling lower bound intuition

term with clipped importance sampling weight

original term regularization

$$\mathbb{E}_{s\sim d^{\text{old}}}\mathbb{E}_{a\sim\pi^{\text{old}}(a|s)}\Big[\min\Big(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}A^{\text{old}}(s,a),\ \text{clip}\Big(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)},1-\epsilon,1+\epsilon\Big)A^{\text{old}}(s,a)\Big)\Big] - C\,\text{KL}\big[\pi^{\text{old}},\pi_\theta\big] \geq \max$$

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

Recalling lower bound intuition

term with clipped importance sampling weight

original term regularization

$$\mathbb{E}_{s\ d_{old}(s)} \mathbb{E}_{a\ \pi_{old}(a|s)} \min\Big(\rho_{\theta}(s,a) A^{\pi_{old}}(s,a),\ \text{clip}\big(\rho_{\theta}(s,a), 1-\epsilon, 1+\epsilon\big) A^{\pi_{old}}(s,a)\Big) - C\, \mathrm{KL}\big[\pi_{old}(\cdot|s), \pi_{\theta}(\cdot|s)\big] \le \max_{\theta} L_{\theta_{old}}(\theta)$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A^{\pi_{old}}(s,a) \gtrless 0$	$\pi(a\|s) \uparrow$	$\rho_{\theta}(s,a) \gg 1+\epsilon$	

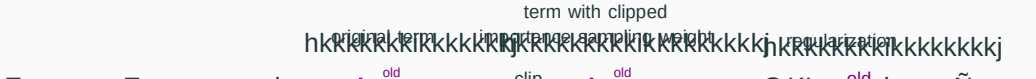
Recalling lower bound intuition

term with clipped
original term importance sampling weight regularization

$$\mathbb{E}_{s\ d_{old}(s)}\mathbb{E}_{a\ \pi_{old}(a|s)}\min\Big(\underbrace{\frac{p\ (a|s)}{p_{old}(a|s)}A_{old}(s,a)}_{\text{original term}};\ \underbrace{\text{clip}\Big(\frac{p\ (a|s)}{p_{old}(a|s)},1-\epsilon,1+\epsilon\Big)A_{old}(s,a)}_{\text{term with clipped importance sampling weight}}\Big)\quad \underbrace{C\ KL\big(p_{old}\ (\cdot|s),p\ (\cdot|s)\big)}_{\text{regularization}} \le \delta$$

Advantage Sign	Direction	Bad ratio case	Gradient
$A_{old}(s,a) \geq 0$	$\pi(a\|s) \uparrow$	$\frac{p(a\|s)}{p_{old}(a\|s)} > 1+\epsilon$	0

Recalling lower bound intuition



Advantage Sign	Direction	Bad ratio case	Gradient
$A^{old}(s,a) \geq 0$	$\pi(a \mid s) \downarrow$	$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} > 1+\epsilon$	0
		$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} < 1-\epsilon$	

Recalling lower bound intuition

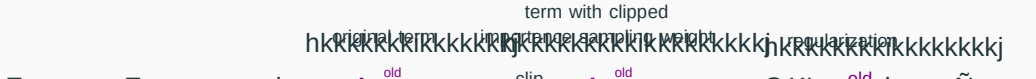
term with clipped

original term importance sampling weight regularization

$$E_{s\ d\ old\ psq} E_{a\ old\ pa|sq} \min\left(p\ q \frac{A^{old}\ ps; aq}{}; \text{clip}\ p\ q \frac{A^{old}\ ps; aq}{}\ q\ C\ KL\left(p^{old}\ k\ q\right) \tilde{N} \max\right)$$

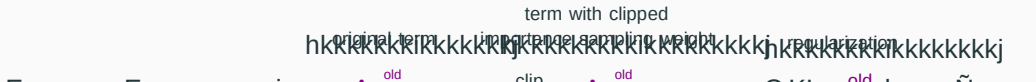
Advantage Sign	Direction	Bad ratio case	Gradient
$A^{old}\ ps; aq \geq 0$	$pa \mid sq$ ò	p q ¡ 1:2 p q 0:8	0 same

Recalling lower bound intuition



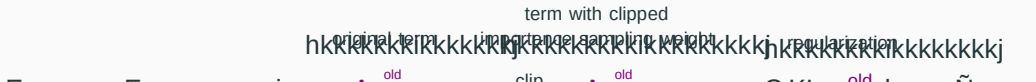
Advantage Sign	Direction	Bad ratio case	Gradient
$A^{old}(s,a) \geq 0$	$\pi_\theta(s) \geq \pi_{old}(s)$	$\frac{\pi_\theta(a\|s)}{\pi_{old}(a\|s)} > 1.2$	0
		$\frac{\pi_\theta(a\|s)}{\pi_{old}(a\|s)} < 0.8$	same
$A^{old}(s,a) < 0$	$\pi_\theta(s) \leq \pi_{old}(s)$	$\frac{\pi_\theta(a\|s)}{\pi_{old}(a\|s)} > 1.2$	
		$\frac{\pi_\theta(a\|s)}{\pi_{old}(a\|s)} < 0.8$	

Recalling lower bound intuition



Advantage Sign	Direction	Bad ratio case	Gradient
$A^{old}(s,a) \geq 0$	$\pi(a \mid s) \uparrow$	$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} > 1+\epsilon$	0
		$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} < 1+\epsilon$	same
$A^{old}(s,a) < 0$	$\pi(a \mid s) \downarrow$	$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} < 1-\epsilon$	same
		$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} > 1-\epsilon$	0

Recalling lower bound intuition



Advantage Sign	Direction	Bad ratio case	Gradient
$A^{old}(s,a) \geq 0$	$\pi(a \mid s) \uparrow$	$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} > 1+\epsilon$	0
		$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} < 1+\epsilon$	same
$A^{old}(s,a) < 0$	$\pi(a \mid s) \downarrow$	$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} < 1-\epsilon$	same
		$\frac{\pi(a \mid s)}{\pi_{old}(a \mid s)} > 1-\epsilon$	0

Clipped Critic Loss

$$\text{Loss}_\phi(p, q) := p \cdot y \quad V_\phi(p, q)^2$$

Clipped Critic Loss

$$\text{Loss}_{p,q}: \; p,y \quad V_{p,q} - q^2$$

$$p,y \quad V^{\text{old}} \quad V^{\text{old}} \quad V_{p,q} - q^2$$

Clipped Critic Loss

$$\text{Loss}_{\phi}(s, q) = \left(\gamma \, V_{\phi}(s') + r - V_{\phi}(s, q)\right)^2$$

$$\gamma \, V_{\phi^{\text{old}}}(s') + r - V_{\phi^{\text{old}}}(s, q)$$

$$\text{Loss}^{\text{clip}}_{\phi}(s, q) = \left(\gamma \, V_{\phi^{\text{old}}}(s') + r - \text{clip}\left(V_{\phi}(s, q) - V_{\phi^{\text{old}}}(s, q), -\epsilon, \epsilon\right)\right)^2$$

Clipped Critic Loss

$$\text{Loss}_V(\theta) = \left(r_t(\theta)\hat{A}_t - V_\theta(s_t) - V_{\text{target}}\right)^2$$

$$\left(r_t(\theta)\hat{A}_t - V_\theta(s_t) - V_t^{\text{old}} - V_t^{\text{old}} - V_{\text{target}}\right)^2$$

$$\text{Loss}_V^{\text{clip}}(\theta) = \left(r_t(\theta)\hat{A}_t - V_t^{\text{old}} - \text{clip}\left(V_\theta(s_t) - V_t^{\text{old}}, -\epsilon, +\epsilon\right) - V_{\text{target}}\right)^2$$

$$\max\left(\text{Loss}_V(\theta), \text{Loss}_V^{\text{clip}}(\theta)\right)$$

Bias-Variance trade-o

Given rollout $s; r; s^{1}; r^{1}; s^{2}; r^{2} : : : s^{pM q}$ from policy and approximation of $V\ psq$

Bias-Variance trade-o

Given rollout $s; r; s^1; r^1; s^2; r^2 : : : s^{pM q}$ from policy and approximation of $V\ psq$ perform credit assingment for state-action pairs; a (was this decision good or bad?)

Bias-Variance trade-o

Given rollout $s; r; s^1; r^1; s^2; r^2 : : : s^{pM q}$ from policy and approximation of $V\ psq$ perform credit assingment for state-action pairs; a (was this decision good or bad?)

For Actor:

$r\ :\ p\ qr\ \log\ pa \mid sq\ loomoon\ ps; aq$

advantage estimator

For Critic:

$loomoon\ Vo\ :\ ps; aq\ V\ psq$

target for regression

Bias-Variance trade-o

Given rollout $s; r; s^1; r^1; s^2; r^2 : : : s^{pM q}$ from policy and approximation of $V\ psq$ perform credit assingment for state-action pairs; a (was this decision good or bad?)

For Actor:

$$r\ :\quad p\ q r\quad \log\quad pa \mid sq \underbrace{\quad ps; aq}_{\text{loomoon}}$$

advantage estimator

For Critic:

$$\underbrace{\quad V_o\ :}_{\text{loomoon}}\quad ps; aq\quad V\ psq$$

target for regression

	$ps; aq$	Bias	Variance
Monte Carlo	$_{p8q}\ ps; aq : \quad r \quad r^1 \quad {}^2 r^2 \quad V\ psq$	0	high
1-step	$_{p1q} ps; aq : \quad r \quad V\ ps^1 q \quad V\ psq$	high	low

Bias-Variance trade-o

Given rollout $s; r; s^1; r^1; s^2; r^2 : : : s^{pM q}$ from policy and approximation of $V\ psq$ perform credit assingment for state-action pairs; a (was this decision good or bad?)

For Actor:

$r\ :\ p\ qr\ \log\ pa \mid sq$ loomoon $ps; aq$

advantage estimator

For Critic:

∇_θ loomoon $ps; aq\ \ V\ psq$

target for regression

	$ps; aq$	Bias	Variance
Monte Carlo	$_{p8q}\ ps; aq : \ r\ \ r^1\ \ ^2r^2\ \ V\ psq$	0	high
N-step	$_{pN q}ps; aq : \ r\ \ r^1\ \ ^N V\ ps^{pN q}q\ \ V\ psq$	intermediate	intermediate
1-step	$_{p1q}ps; aq : \ r\ \ V\ ps^1q\ \ V\ psq$	high	low

Bias-Variance trade-o

Given rollout $s; r; s^1; r^1; s^2; r^2 : : : s^{pM q}$ from policy and approximation of $V\ psq$ perform credit assingment for state-action pairs; a (was this decision good or bad?)

For Actor:

$$r \ : \quad p\ qr \quad \log \quad pa \mid sq \quad \underbrace{loomoon}_{\text{advantage estimator}} ps; aq$$

For Critic:

$$\underbrace{loomoon}_{\text{target for regression}} \quad ps; aq \quad V \quad psq$$

	$ps; aq$	Bias	Variance
Monte Carlo	$_{p8q}\ ps; aq : \ r \quad r^1 \quad {}^2r^2 \quad V\ psq$	0	high
N-step	$_{pN q}ps; aq : \ r \quad r^1 \quad {}^N V\ ps^{pN q}q \quad V\ psq$	intermediate	intermediate
1-step	$_{p1q}ps; aq : \ r \quad V\ ps^1q \quad V\ psq$	high	low

Problem: hard to choose N.

Backward view: idea

N-step update:

$$V(s_t) \leftarrow V(s_t) + \alpha_{N(s_t)}(s_t)\,\delta_t$$

Backward view: idea

N-step update:

$$V(s_t) \leftarrow V(s_t) + \alpha_{(N)}(s_t, a_t)$$

$$V(s_t) \leftarrow V(s_t) + \alpha \underbrace{\left[r_t + \gamma V(s_{t+1}) - V(s_t) \right]}_{\alpha_{(1)}(s_t, a_t)}$$

Backward view: idea

N-step update:

$$V(s) \leftarrow V(s) + \alpha_{(N)}(s,a)$$

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

$$V(s) \leftarrow V(s) + \alpha \underbrace{\quad\quad}_{\alpha_{(2)}(s,a)} \left(r + V(s') - V(s) \right)$$

Backward view: idea

N-step update:

$$V(s_t) \leftarrow V(s_t) + \alpha\left(R_t^{(n)} - V(s_t)\right), \quad R_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n})$$

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

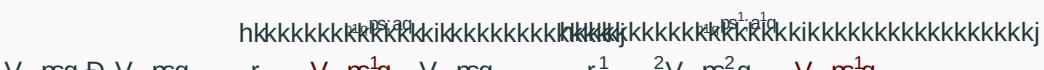
$$V(s_t) \leftarrow V(s_t) + \alpha\left(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right) + \alpha\left(r_{t+2} + \gamma V(s_{t+2}) - V(s_{t+1})\right)$$

Backward view: idea

N-step update:

$V(s_t) \leftarrow V(s_t) + \alpha\left(R_t^{(n)} - V(s_t)\right)$

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

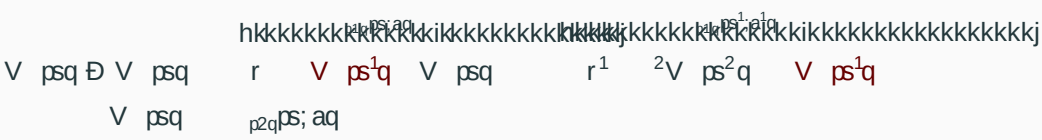


Backward view: idea

N-step update:

$$V(s_t) \leftarrow V(s_t) + \alpha\left(R_t^{(n)} - V(s_t)\right)$$

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



Backward view: idea

N-step update:

$$V(s_t) \leftarrow V(s_t) + \alpha\left(R_t^{(n)} - V(s_t)\right)$$

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

$$R_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_{t+k+1} + \gamma^n V(s_{t+n})$$

$$R_t^{(1)} + \gamma R_{t+1}^{(1)} + \gamma^2 V(s_{t+2}) = r_{t+1} + \gamma V(s_{t+1}) + \gamma\left(r_{t+2} + \gamma V(s_{t+2})\right) + \gamma^2 V(s_{t+2})$$

$$= r_{t+1} + \gamma r_{t+2} + \gamma^2 V(s_{t+2})$$

$$= R_t^{(2)}$$

Eligibility Traces

Eligibility Traces

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

Eligibility Traces

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

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

$$\Delta V(s) = V(s) + \delta e(s) \quad (1)$$

Eligibility Traces

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

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

$$\Delta V(s) = \delta \, e(s) \quad (1)$$

Online Monte-Carlo updates:

- $\delta = e(s) = 0$ at the start of each episode

Eligibility Traces

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

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

$$\forall s: V(s) \leftarrow V(s) + \alpha \delta e(s) \quad (1)$$

Online Monte-Carlo updates:

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

Eligibility Traces

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

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

$$\forall s: V(s) \leftarrow V(s) + \alpha \delta e(s) \quad (1)$$

Online Monte-Carlo updates:

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

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0; 1; 2 : : :$
  ^ take action a_k , observer $r_k; s_{k+1}$

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $eps(s) = 0$

observe s_0
for $k = 0; 1; 2 : : :$
- take action a_k, observer $r_k; s_{k+1}$
- $\delta_{p1q} : r_k + V(s_{k+1}) - V(s_k)$

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2, \ldots$
- take action a_k, observer r_k, s_{k+1}
- $\delta_{(1)} : r_k + V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2, \ldots$
- take action a_k, observer r_k, s_{k+1}
- $\delta_{t+1} = r_k + V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \delta_{t+1}$

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2, \ldots$
- take action a_k, observe r_k, s_{k+1}
- $\delta_k = r_k + V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha \delta_k e(s)$
- $\forall s: e(s) \leftarrow \gamma e(s)$

TD(1) and TD(0)

TD(1)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2, \ldots$
- take action a_k, observer r_k, s_{k+1}
- $\delta = r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha \delta e(s)$
- $\forall s: e(s) \leftarrow \gamma \lambda e(s)$

TD(0)

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

observe s_0
for $k = 0, 1, 2, \ldots$
- take action a_k, observer r_k, s_{k+1}
- $\delta = r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha \delta e(s)$
- $\forall s: e(s) \leftarrow 0 \cdot \gamma \lambda e(s)$

TD( )

TD( )

TD( )

TD( )

TD( )

Input: policy
Initialize $V(s)$ arbitrarily
Initialize $e(s) = 0$

Observe s_0
for $k = 0, 1, 2, \ldots$:
- take action a_k, observer r_k; s_{k+1}
- $\delta \leftarrow r_k + V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha\delta e(s)$
- $\forall s: e(s) \leftarrow \gamma\lambda e(s)$

Backward view vs Forward view

Forward View

Give credit topresent from knownfuture

is this decision good or bad based on the outcome?

Backward view vs Forward view

Forward View

Give credit topresent from knownfuture

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( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p1-qp\,s^{pt\,q}; a^{pt\,q}q$$

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p_{1-q} p s^{pt\,q}; a^{pt\,q} q \quad p1 \quad q^{\sum_{N=1}^{\infty}} p_{N\,q} p s; a q$$

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p_{1}(q)\, p(s^{pt}(q); a^{pt}(q)) \quad p_{1} \quad q \sum_{N=1}^{\infty} p_{N}(q)\, p(s; a)$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{pt}(q)$				

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; {}_{p1-q}p\mathrm{s}^{pt\,q\cdot}; a^{pt\,q}q \quad p1 \quad q \sum_{N=1}^{\infty} {}_{pN\,q}p\mathrm{s}; aq$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{pt\,q}$	${}_{p1-q}p\mathrm{s}^{pt\,q\cdot}; a^{pt\,q}q$	$p \; q^{t}$		

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; {}_{p1-q}p s^{p t q}; a^{p t q} q \quad p1 \quad q \sum_{N=1}^{\infty} {}_{pN-q}p s; a q$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{p t q}$	${}_{p1-q}p s^{p t q}; a^{p t q} q$	$p \; q^{t}$	${}_{pN-q}p s; a q$ for $N \; ¡ \; t$	

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p1-q p s^{p t q}; a^{p t q} q \quad p1 \quad q \sum_{N=1}^{\infty} p N q p s; a q$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{p t q}$	$p1-q p s^{p t q}; a^{p t q} q$	$p \; q^{t}$	$p N q p s; a q$ for $N \neg t$	$t p1 \quad qp^{t} \quad t \quad 1 \quad t \quad 2 \quad : : : q$
$V \; p s^{p t q} q$ ($t \neg 0$)				

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p_{1q}(s^{pt\,q}; a^{pt\,q}q \quad p1 \quad q \sum_{N=1}^{\infty} p_{Nq}(s; aq$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{pt\,q}$	$p_{1q}(s^{pt\,q}; a^{pt\,q}q$	$p \quad q^{t}$	$p_{Nq}(s; aq$ for $N \; ¡ \; t$	$^{t}p1 \quad qp^{t} \quad t \quad 1 \quad t \quad 2 \quad : : : q$
$V \; (s^{pt\,q}q$ $(t \; ¡ \; 0)$	$p_{1q}(s^{pt \quad 1q}; a^{pt \quad 1q}q$ $p_{1q}(s^{pt\,q}; a^{pt\,q}q$			

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p_{1-q}(s^{pt\,q}; a^{pt\,q}) \quad p1 \quad q \sum_{N=1}^{\infty} p_{N\,q}(s; aq)$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{pt\,q}$	$p_{1-q}(s^{pt\,q}; a^{pt\,q})$	$p \; q^{t}$	$p_{N\,q}(s; aq)$ for $N \; ¡ \; t$	$^{t}p1 \quad qp^{t} \quad t \quad 1 \quad t \quad 2 \quad : : : q$
$V \; (s^{pt\,q})$ $(t \; ¡ \; 0)$	$p_{1-q}(s^{pt \quad 1-q}; a^{pt \quad 1-q})$ $p_{1-q}(s^{pt\,q}; a^{pt\,q})$	$p \quad q^{t \quad 1} \quad p \quad q^{t}$		

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} p \; q^{t} \; p_{1-q} p s^{p t q}; a^{p t q} q \quad p 1 \quad q \sum_{N=1}^{\infty} p_{N q} p s; a q$$

Term	Left side	Left side coe .	Right side	Right side coe .
$r^{p t q}$	$p_{1-q} p s^{p t q}; a^{p t q} q$	$p \quad q^{t}$	$p_{N q} p s; a q$ for $N \; ¡ \; t$	$^{t} p 1 \quad q p^{t} \quad t \quad 1 \quad t \quad 2 \quad : : : q$
$V \; p s^{p t q} q$ $(t \; ¡ \; 0)$	$p_{1-q} p s^{p t \quad 1 q}; a^{p t \quad 1 q} q$ $p_{1-q} p s^{p t q}; a^{p t q} q$	$p \quad q^{t \quad 1} \quad p \quad q^{t}$	$p_{t q} p s; a q$	

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} (1-\lambda)\lambda^t \, R_{t+1}^{(n)} \qquad (1-\lambda)\sum_{n=1}^{N-1} \lambda^{n-1} R_{t+1}^{(n)}$$

Term	Left side	Left side coe□.	Right side	Right side coe□.
r^{ptq}	$_{p1q}ps^{ptq}; a^{ptq}q$	p q^t	$_{pNq}ps; aq$ for $N \; ¡ \; t$	^{t}p1 qp^t t 1 t 2 $:::$q
V ps^{ptq}q ($t \; ¡ \; 0$)	$_{p1q}ps^{pt \; 1q}; a^{pt \; 1q}q$ $_{p1q}ps^{ptq}; a^{ptq}q$	p q$^{t \; 1}$ p q^t	$_{ptq}ps; aq$	p1 qp$^{t \; 1}$ ^{t}q
V psq				

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} \gamma^t \, r_{t+1} + \gamma^{t+1} V(s_{t+1}; a^{t+1}) - V(s_t; a^t) \qquad V(s_t; a^t) + \sum_{N=1}^{\infty} \gamma^N \, r_{t+N} + \gamma^{t+N} V(s_{t+N}; a^{t+N}) - V(s_{t+N-1}; a^{t+N-1})$$

Term	Left side	Left side coe□.	Right side	Right side coe□.
r^{t+1}	$r_{t+1} + \gamma V(s^{t+1}; a^{t+1})$	γ^t	$r_{t+N} + \gamma V(s^{t+N}; a^{t+N})$ for $N \geq t$	$\gamma^t - \gamma^{t+1} - \gamma^{t+2} \cdots \gamma^{t+N-1}$
$V(s^{t+1})$ ($t \geq 0$)	$r_{t+1} + \gamma V(s^{t+1}; a^{t+1})$ $r_{t+1} + \gamma V(s^{t+1}; a^{t+1})$	γ^t $-\gamma^{t-1}$	$r_{t+N} + \gamma V(s^{t+N}; a^{t+N})$	$\gamma^t - \gamma^{t-1}$
$V(s)$	$r_{t+1} + \gamma V(s^{t+1}; a^{t+1})$	1		

Forward view for TD( )

Equivalent forms of TD( ) updates

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right] \alpha_t \qquad V(s_t) \leftarrow V(s_t) + \sum_{N=1}^{\infty} (\gamma\lambda)^{N-1} \delta_{t+N} \alpha_{t+N}$$

Term	Left side	Left side coe□.	Right side	Right side coe□.
r_{t+1}	$r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$	$(\gamma\lambda)^t \alpha_t$	$\delta_{t+N} \alpha_{t+N}$ for $N \geq t+1$	$(\gamma\lambda)^t (1 - \gamma\lambda) (\gamma\lambda)^{t+1} (\gamma\lambda)^{t+2} \cdots$
$V(s_{t+1})$ ($t \geq 0$)	$r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ $r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$	$(\gamma\lambda)^t \alpha_t$ $(\gamma\lambda)^{t-1} \alpha_{t-1}$	$(\gamma\lambda)^t \alpha_t$	$(1 - \gamma\lambda) (\gamma\lambda)^t$
$V(s_t)$	$r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$	1	all	$(1 - \gamma\lambda) (1 + \gamma\lambda (1 + \gamma\lambda (1 + \cdots)))$

Generalized Advantage Estimation (GAE)

$$\text{GAE}(s, a) := \sum_{t=0}^{T} (\gamma\lambda)^t \, \delta_{t+1}(s_t, a_t)$$

(trace is zeroed when future is not available)

Generalized Advantage Estimation (GAE)

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

(trace is zeroed when future is not available)

How to compute in practice:

$$\text{GAE}(s_t, a_t) \quad \delta_{t+1}(s_t, a_t)$$

$$(1 - \text{done}_{t-1}) \, \text{GAE}(s_{t-1}, a_{t-1})$$

Proximal Policy Optimization: implementation matters

Key elements:

- X Clipped policy loss
- X Clipped critic loss
- X 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
- • (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 $p_a | s; q; V\ p_{sq};$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$
- compute 1-step errors: $_{p1q} p s_t; a_t q : \quad r_t \quad p1 \quad \text{done}_t \; _1 q V \; p s_t \; _1 q \quad V \; p s_t q$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$
- compute 1-step errors: $_{p 1 q} p s_t; a_t q : \quad r_t \quad p 1 \quad$ done$_t \; _1 q V \; p s_t \; _1 q \quad V \; p s_t q$
- compute GAE advantage estimations: $^{\text{GAE}} p s_N \; _1; a_N \; _1 q : \quad _{p 1 q} p s_N \; _1; a_N \; _1 q$
- for t from $N \quad 2$ to 0:
  - $^{\text{GAE}} p s_t; a_t q :$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$
- compute 1-step errors: $_{p1q} p s_t; a_t q : \quad r_t \quad p1 \quad$ done$_t \; _1 q V \; p s_t \; _1 q \quad V \; p s_t q$
- compute GAE advantage estimations: $^{GAE} p s_N \; _1; a_N \; _1 q : \quad _{p1q} p s_N \; _1; a_N \; _1 q$
- for t from $N \quad 2$ to 0:
  - $^{GAE} p s_t; a_t q : \quad _{p1q} p s_t; a_t q \quad p1 \quad$ done$_t \; _1 q \quad ^{GAE} p s_t \; _1; a_t \; _1 q$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$
- compute 1-step errors: $_{p1q} p s_t; a_t q : \quad r_t \quad p1 \quad \text{done}_{t \;\; 1} q V \; p s_{t \;\; 1} q \quad V \; p s_t q$
- compute GAE advantage estimations: $^{GAE} p s_{N \;\; 1}; a_{N \;\; 1} q : \quad _{p1q} p s_{N \;\; 1}; a_{N \;\; 1} q$
- for t from $N \quad 2$ to 0:
  - $^{GAE} p s_t; a_t q : \quad _{p1q} p s_t; a_t q \quad p1 \quad \text{done}_{t \;\; 1} q \quad ^{GAE} p s_{t \;\; 1}; a_{t \;\; 1} q$
- compute critic targets: $y p s_t q : \quad ^{GAE} p s_t; a_t q \quad V \; p s_t q$

Full Pipeline: pt.I

Proximal Policy Optimization (PPO)

Initialize $p a \mid s; \; q; V \; p s q;$

for $k \quad 0; 1; 2 : : :$

- collect several rollouts $s_0; a_0; r_0; s_1;$ done$_1; a_1 : : : s_N;$ done$_N$ using $p a \mid s; \; q;$
  store probabilities of selected actions as $^{\text{old}} p a_t \mid s_t q : \quad p a_t \mid s_t; \; q$
  store critic output as $V^{\text{old}} p s_t q : \quad V \; p s_t q$
- compute 1-step errors: $_{p1q} p s_t; a_t q : \quad r_t \quad p1 \quad \text{done}_{t \;\; 1} q V \; p s_{t \;\; 1} q \quad V \; p s_t q$
- compute GAE advantage estimations: $^{\text{GAE}} p s_{N \;\; 1}; a_{N \;\; 1} q : \quad _{p1q} p s_{N \;\; 1}; a_{N \;\; 1} q$
- for t from $N \quad 2$ to 0:
  - $^{\text{GAE}} p s_t; a_t q : \quad _{p1q} p s_t; a_t q \quad p1 \quad \text{done}_{t \;\; 1} q \quad ^{\text{GAE}} p s_{t \;\; 1}; a_{t \;\; 1} q$
- compute critic targets: $y p s_t q : \quad ^{\text{GAE}} p s_t; a_t q \quad V \; p s_t q$
- construct dataset of $p s_t; a_t; \quad ^{\text{GAE}} p s_t; a_t q; y p s_t q; \quad ^{\text{old}} p a_t \mid s_t q; V^{\text{old}} p s_t q q$

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:

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 $^{GAE}(s, a)$ in the batch by subtracting mean and dividing by std

Full Pipeline: pt.II

Proximal Policy Optimization (PPO) -- cont.

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

$$p s; a; \ q : \ \frac{p a \mid s; \ q}{^{old}p a \mid s q}; \qquad ^{clip}p s; a; \ q \quad \text{clip} p \ p s; a; \ q; 1 \quad ; 1 \quad q$$

Proximal Policy Optimization (PPO) -- cont.

- go through dataset n_epochs times, sampling mini-batches of size B; for each mini-batch:
  - normalize $^{GAE}ps; aq$ in the batch by subtracting mean and dividing by std
  - compute importance sampling weights:

$$ps; a; q : \frac{pa \mid s; q}{^{old}pa \mid sq}; \qquad ^{clip}ps; a; q \quad clipp\ ps; a; q; 1 \quad ; 1 \quad q$$

  - update actor:

$$L_1ps; a; q : \quad ps; a; q \ ^{GAE}ps; aq; \qquad L_2ps; a; q : \quad ^{clip}ps; a; q \ ^{GAE}ps; aq$$

$$Ð \qquad r \ \frac{1}{B} \ \sum_{s;a} \ minpL_1ps; a; q; L_2ps; a; qq$$

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 $^{GAE}ps; aq$ in the batch by subtracting mean and dividing by std
  - compute importance sampling weights:

$$ps; a; q : \frac{pa \mid s; q}{^{old}pa \mid sq}; \qquad ^{clip}ps; a; q \quad clipp\ ps; a; q; 1 \quad ; 1 \quad q$$

  - update actor:

$$L_1ps; a; q : \quad ps; a; q \quad ^{GAE}ps; aq; \qquad L_2ps; a; q : \quad ^{clip}ps; a; q \quad ^{GAE}ps; aq$$

$$Ð \qquad r \ \frac{1}{B} \ \sum_{s;a} \ minpL_1ps; a; q; L_2ps; a; qq$$

  - update critic:

$$Loss_1ps; q : \quad ypsq \quad V \ psq \ ^2$$

$$Loss_2ps; q : \quad ypsq \quad V^{old}psq \quad clippV \ psq \quad V^{old}psq; ^; \ ^q \ ^2$$

$$Ð \qquad r \ \frac{1}{B} \ \sum_{s} \ maxpLoss_1ps; q; Loss_2ps; qq$$