SAMPLE EFFICIENT ACTOR-CRITIC WITH EXPERIENCE REPLAY

-Ablation study

0713402 黃亭晅
0713415 林恩衍
0713451 宋沛潔

Background

They want to solve...

1. both discrete and continuous action spaces
2. sample inefficient
3. high bias and high variance

- propose **ACER** (actor critic with experience replay)

ACER for discrete action space

ACER Algorithm

Master algorithm

Algorithm 1 ACER for discrete actions (master algorithm)

// Assume global shared parameter vectors θ and θ_v.
// Assume ratio of replay r.
repeat
    Call ACER on-policy, Algorithm 2.
    $n \leftarrow \text{Possion}(r)$
    for $i \in \{1, \cdots, n\}$ **do**
        Call ACER off-policy, Algorithm 2.
    end for
until Max iteration or time reached.

ACER Algorithm

Algorithm 2 ACER for discrete actions

Reset gradients $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
Initialize parameters $\theta' \leftarrow \theta$ and $\theta'_v \leftarrow \theta_v$.

target network and evaluation network

sampling trajectories

if not On-Policy **then**
  Sample the trajectory $\{x_0, a_0, r_0, \mu(\cdot|x_0), \cdots, x_k, a_k, r_k, \mu(\cdot|x_k)\}$ from the replay memory.
else
  Get state x_0
end if

explore and compute importance weight

for $i \in \{0, \cdots, k\}$ **do**
  Compute $f(\cdot|\phi_{\theta'}(x_i))$, $Q_{\theta'_v}(x_i, \cdot)$ and $f(\cdot|\phi_{\theta_a}(x_i))$.
  if On-Policy **then**
    Perform a_i according to $f(\cdot|\phi_{\theta'}(x_i))$
    Receive reward r_i and new state x_{i+1}
    $\mu(\cdot|x_i) \leftarrow f(\cdot|\phi_{\theta'}(x_i))$
  end if
  $\bar{\rho}_i \leftarrow \min\left\{1, \frac{f(a_i|\phi_{\theta'}(x_i))}{\mu(a_i|x_i)}\right\}$.
end for

$$Q^{ret} \leftarrow \begin{cases} 0 & \text{for terminal } x_k \\ \sum_a Q_{\theta'_v}(x_k, a) f(a|\phi_{\theta'}(x_k)) & \text{otherwise} \end{cases}$$

back propagation using improved TRPO

for $i \in \{k-1, \cdots, 0\}$ **do**
  $Q^{ret} \leftarrow r_i + \gamma Q^{ret}$
  $V_i \leftarrow \sum_a Q_{\theta'_v}(x_i, a) f(a|\phi_{\theta'}(x_i))$
  Computing quantities needed for trust region updating:

$$\boxed{\begin{aligned} g \leftarrow\ & \min\{c, \rho_i(a_i)\} \nabla_{\phi_{\theta'}(x_i)} \log f(a_i|\phi_{\theta'}(x_i))(Q^{ret} - V_i) \\ & + \sum_a \left[1 - \frac{c}{\rho_i(a)}\right]_+ f(a|\phi_{\theta'}(x_i)) \nabla_{\phi_{\theta'}(x_i)} \log f(a|\phi_{\theta'}(x_i))(Q_{\theta'_v}(x_i, a_i) - V_i) \end{aligned}}$$

$$k \leftarrow \nabla_{\phi_{\theta'}(x_i)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_i)) \| f(\cdot|\phi_{\theta'}(x_i))\right]$$

  Accumulate gradients wrt θ': $d\theta' \leftarrow d\theta' + \frac{\partial \phi_{\theta'}(x_i)}{\partial \theta'}\left(g - \max\left\{0, \frac{k^T g - \delta}{\|k\|_2^2}\right\} k\right)$
  Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(Q^{ret} - Q_{\theta'_v}(x_i, a))^2$
  Update Retrace target: $Q^{ret} \leftarrow \bar{\rho}_i \left(Q^{ret} - Q_{\theta'_v}(x_i, a_i)\right) + V_i$
end for

Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$. update target network and evaluation network
Updating the average policy network: $\theta_a \leftarrow \alpha\theta_a + (1 - \alpha)\theta$ update average target network softly

Gradient of actor critic

- original policy gradient:

$$g = E_{x_{0:\infty}, a_{0:\infty}}[\sum_{t \geq 0} A^{\pi}(x_t, a_t) \nabla_\theta log \pi_\theta(a_t|x_t)]$$

- policy gradient with importance sampling:

$$g^{\text{marg}} = \mathbb{E}_{x_t \sim \beta, a_t \sim \mu}\left[\boxed{\rho_t} \nabla_\theta \log \pi_\theta(a_t|x_t) Q^{\pi}(x_t, a_t)\right], \quad \text{where } \rho_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}.$$

Gradient of actor critic

- gradient with correction term and clipped importance weight

$$
\begin{aligned}
g^{\text{marg}} &= \mathbb{E}_{x_t a_t}\left[\rho_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)\right] \\
&= \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}\left[\boxed{\bar{\rho}_t} \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)\right] + \mathop{\mathbb{E}}_{a\sim\pi}\left(\boxed{\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+} \nabla_\theta \log \pi_\theta(a|x_t) Q^\pi(x_t, a)\right)\right]
\end{aligned}
$$

- gradient of actor critic:

$$
\begin{aligned}
\widehat{g}_t^{\text{acer}} = {} & \bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t)\left[\boxed{Q^{\text{ret}}(x_t, a_t)} - \boxed{V_{\theta_v}(x_t)}\right] \\
& + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t)\left[Q_{\theta_v}(x_t, a) - \boxed{V_{\theta_v}(x_t)}\right]\right)
\end{aligned}
$$

$$
Q^{\text{ret}}(x_t, a_t) = r_t + \gamma \bar{\rho}_{t+1}\left[Q^{\text{ret}}(x_{t+1}, a_{t+1}) - Q(x_{t+1}, a_{t+1})\right] + \gamma V(x_{t+1}) \quad \text{, where } \bar{\rho} = \min\{c, \rho_t\}
$$

Trust region policy optimization

01 Algorithm

for $i \in \{k-1, \cdots, 0\}$ **do**

$Q^{ret} \leftarrow r_i + \gamma Q^{ret}$

$V_i \leftarrow \sum_a Q_{\theta'_v}(x_i, a) f(a|\phi_{\theta'}(x_i))$

Computing quantities needed for trust region updating:

$$
\begin{aligned}
g &\leftarrow \min\{c, \rho_i(a_i)\} \nabla_{\phi_{\theta'}(x_i)} \log f(a_i|\phi_{\theta'}(x_i))(Q^{ret} - V_i) \\
&\quad + \sum_a \left[1 - \frac{c}{\rho_i(a)}\right]_+ f(a|\phi_{\theta'}(x_i)) \nabla_{\phi_{\theta'}(x_i)} \log f(a|\phi_{\theta'}(x_i))(Q_{\theta'_v}(x_i, a_i) - V_i) \\
k &\leftarrow \nabla_{\phi_{\theta'}(x_i)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_i)) \| f(\cdot|\phi_{\theta'}(x_i))\right]
\end{aligned}
$$

Accumulate gradients wrt θ': $d\theta' \leftarrow d\theta' + \frac{\partial \phi_{\theta'}(x_i)}{\partial \theta'} \left(g - \max\left\{0, \frac{k^T g - \delta}{\|k\|_2^2}\right\} k\right)$

Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(Q^{ret} - Q_{\theta'_v}(x_i, a))^2$

Update Retrace target: $Q^{ret} \leftarrow \bar{\rho}_i \left(Q^{ret} - Q_{\theta'_v}(x_i, a_i)\right) + V_i$

end for

Trust region policy optimization

02 New TRPO

Original TRPO : $\dfrac{\tilde{\pi}_\theta\left(a|s_n\right)}{q\left(a|s_n\right)}$

$$\begin{aligned}\widehat{g}_t^{\text{acer}} &= \bar{\rho}_t \nabla_{\phi_\theta(x_t)} \boxed{\log f(a_t|\phi_\theta(x))}[Q^{\text{ret}}(x_t,a_t) - V_{\theta_v}(x_t)] \\ &+ \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a)-c}{\rho_t(a)}\right]_+ \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x))[Q_{\theta_v}(x_t,a) - V_{\theta_v}(x_t)]\right).\end{aligned}$$

Trust region policy optimization

02 New TRPO

average policy network :

Distribution f + $\emptyset_\theta$

the relationship between these two terms :

$$\emptyset_\theta : \pi(\cdot\,|x) = f(\cdot\,|\emptyset_\theta(x))$$

Trust region policy optimization

02 New TRPO

$$\begin{aligned}\widehat{g}_t^{\text{acer}} &= \bar{\rho}_t \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x))[Q^{\text{ret}}(x_t, a_t) - V_{\theta_v}(x_t)] \\ &+ \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x))[Q_{\theta_v}(x_t, a) - V_{\theta_v}(x_t)]\right).\end{aligned}$$

Trust region policy optimization

03 Trust region update

KL-divergence :

$$\begin{aligned} \underset{z}{\text{minimize}} \quad & \frac{1}{2}\|\hat{g}_t^{\text{acer}} - z\|_2^2 \\ \text{subject to} \quad & \nabla_{\phi_\theta(x_t)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_t))\|f(\cdot|\phi_\theta(x_t))\right]^T z \le \delta \end{aligned}$$

KKT condition :

Letting $k = \nabla_{\phi_\theta(x_t)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_t))\|f(\cdot|\phi_\theta(x_t))\right]$

$$z^* = \hat{g}_t^{\text{acer}} - \max\left\{0, \frac{k^T \hat{g}_t^{\text{acer}} - \delta}{\|k\|_2^2}\right\} k$$

Trust region policy optimization

03 Trust region update

KL-divergence :

$$\begin{aligned}
\underset{z}{\text{minimize}} \quad & \frac{1}{2}\|\hat{g}_t^{\text{acer}} - z\|_2^2 \\
\text{subject to} \quad & \nabla_{\phi_\theta(x_t)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_t))\|f(\cdot|\phi_\theta(x_t))\right]^T z \leq \delta
\end{aligned}$$

KKT condition :

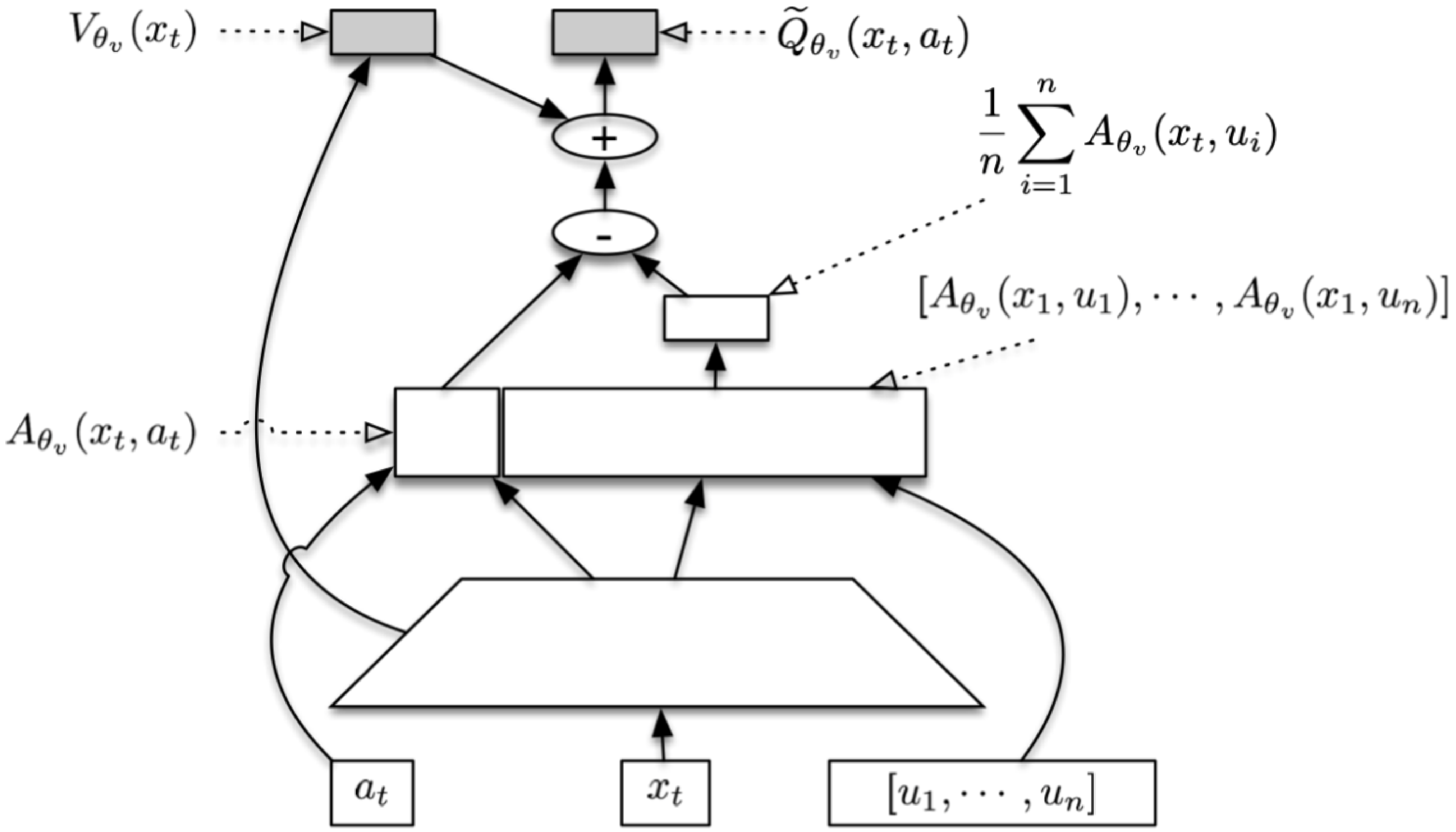
Letting $k = \nabla_{\phi_\theta(x_t)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_t))\|f(\cdot|\phi_\theta(x_t))\right]$

$$z^* = \hat{g}_t^{\text{acer}} - \max\left\{0, \boxed{\frac{k^T \hat{g}_t^{\text{acer}} - \delta}{\|k\|_2^2}}\right\} k$$

ACER for continuous action space

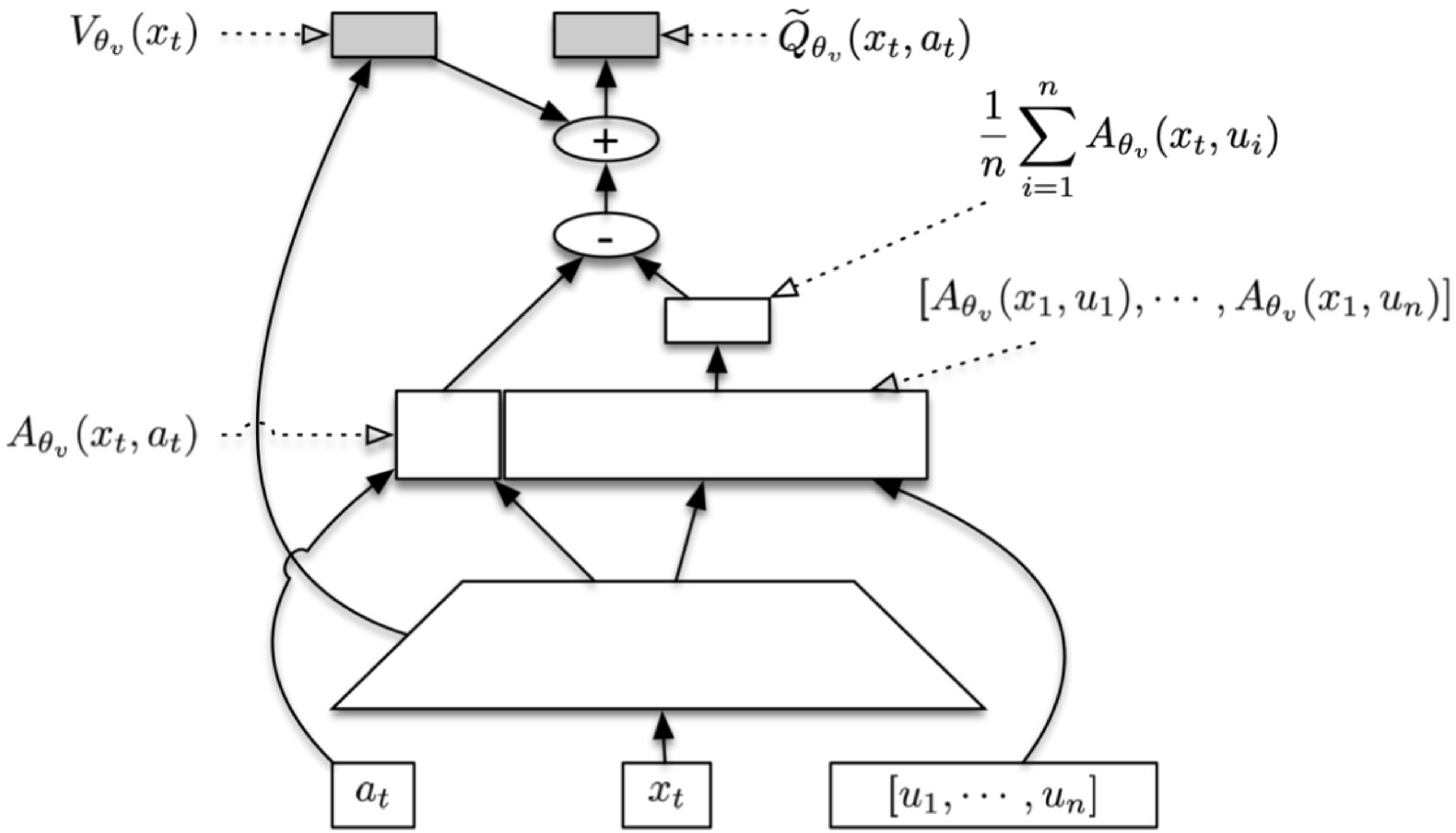
Continuous action space

SDN (Stochastic Dueling Network)



Continuous action space

$$\tilde{Q}_{\theta_v}(x_t, a_t) \sim V_{\theta_v}(x_t) + A_{\theta_v}(x_t, a_t) - \frac{1}{n}\sum_{i=1}^{n} A_{\theta_v}(x_t, u_i), \text{where } u_i \sim \pi(\cdot|x_t)$$



Continuous action space

Value part gradient update

$$V^{target}(x_t) = \min\{1, \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}\}(Q^{ret}(x_t, a_t) - Q_{\theta_v}(x_t)) + V_{\theta_v}(x_t)$$

Trust region policy optimization

04 Continuous trust region updating

average policy network :

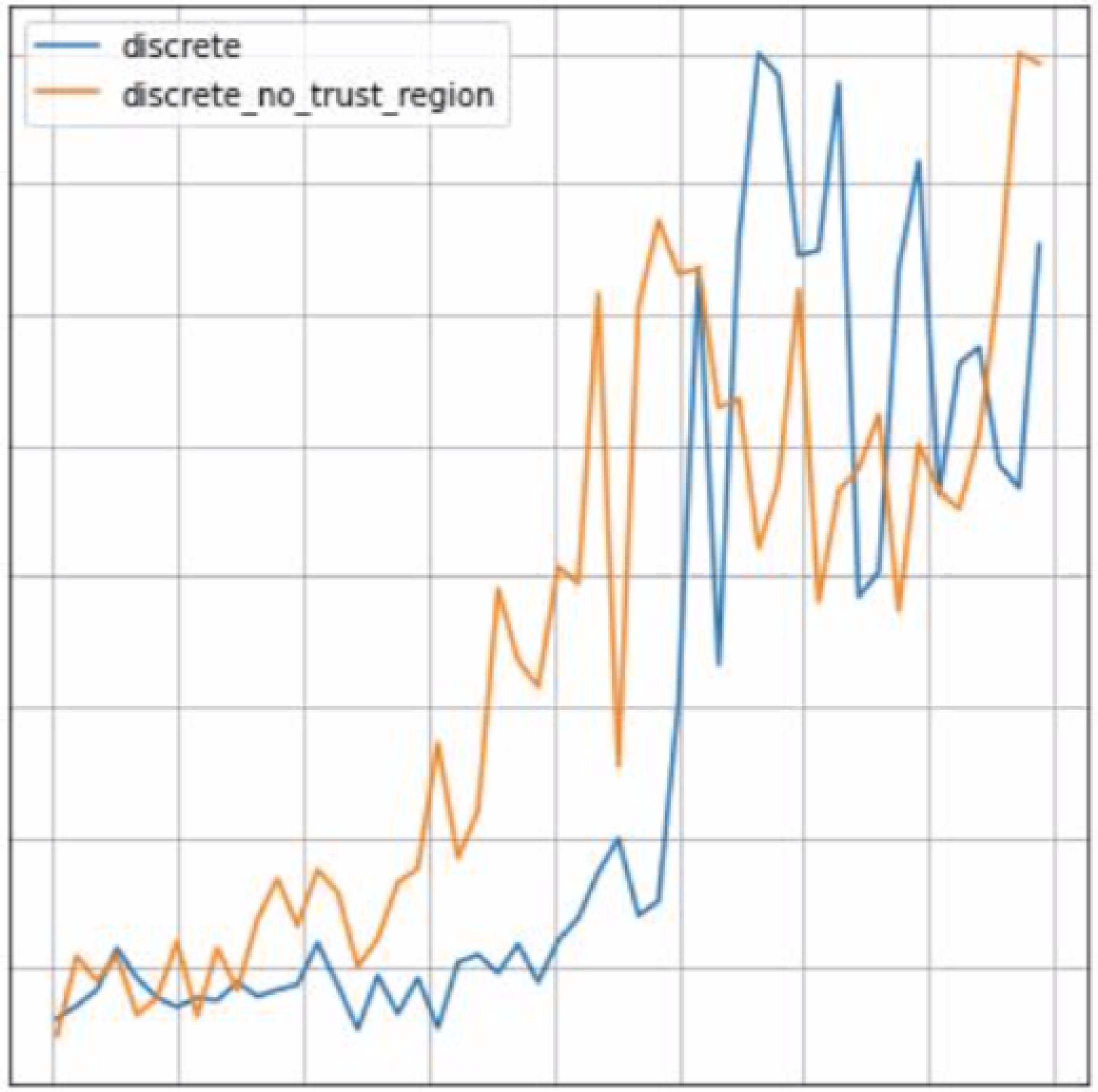
Gaussian distribution f + ϕ_θ

$Q_{ret} \rightarrow Q_{opc}$:

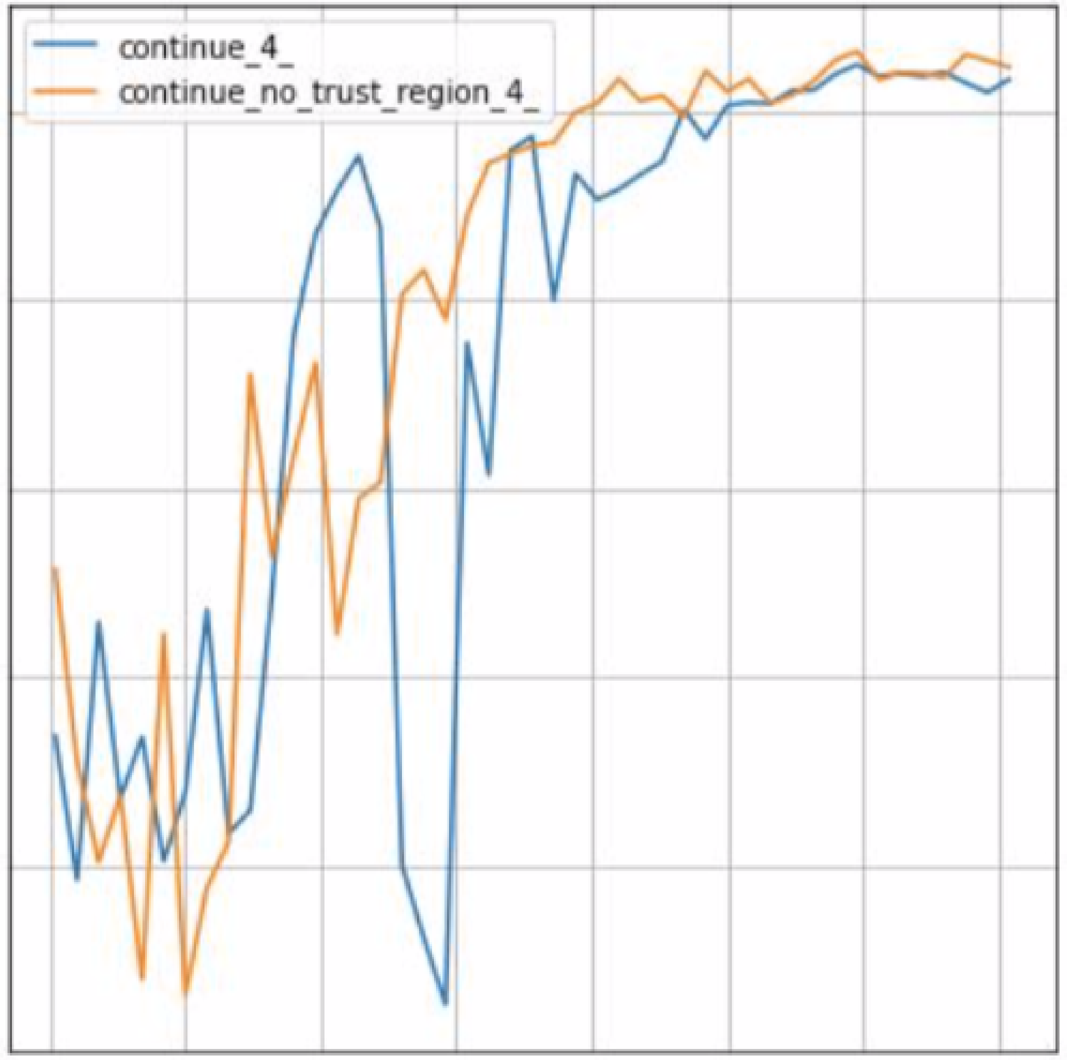
$$\hat{g}_t^{\text{acer}} = \bar{\rho}_t \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x_t))\left(\boxed{Q^{\text{opc}}(x_t, a_t)} - V_{\theta_v}(x_t)\right) + \left[\frac{\rho_t(a'_t) - c}{\rho_t(a'_t)}\right]_+ (\widetilde{Q}_{\theta_v}(x_t, a'_t) - V_{\theta_v}(x_t))\nabla_{\phi_\theta(x_t)} \log f(a'_t|\phi_\theta(x_t)).$$

Ablation study

remove trust region constraint :



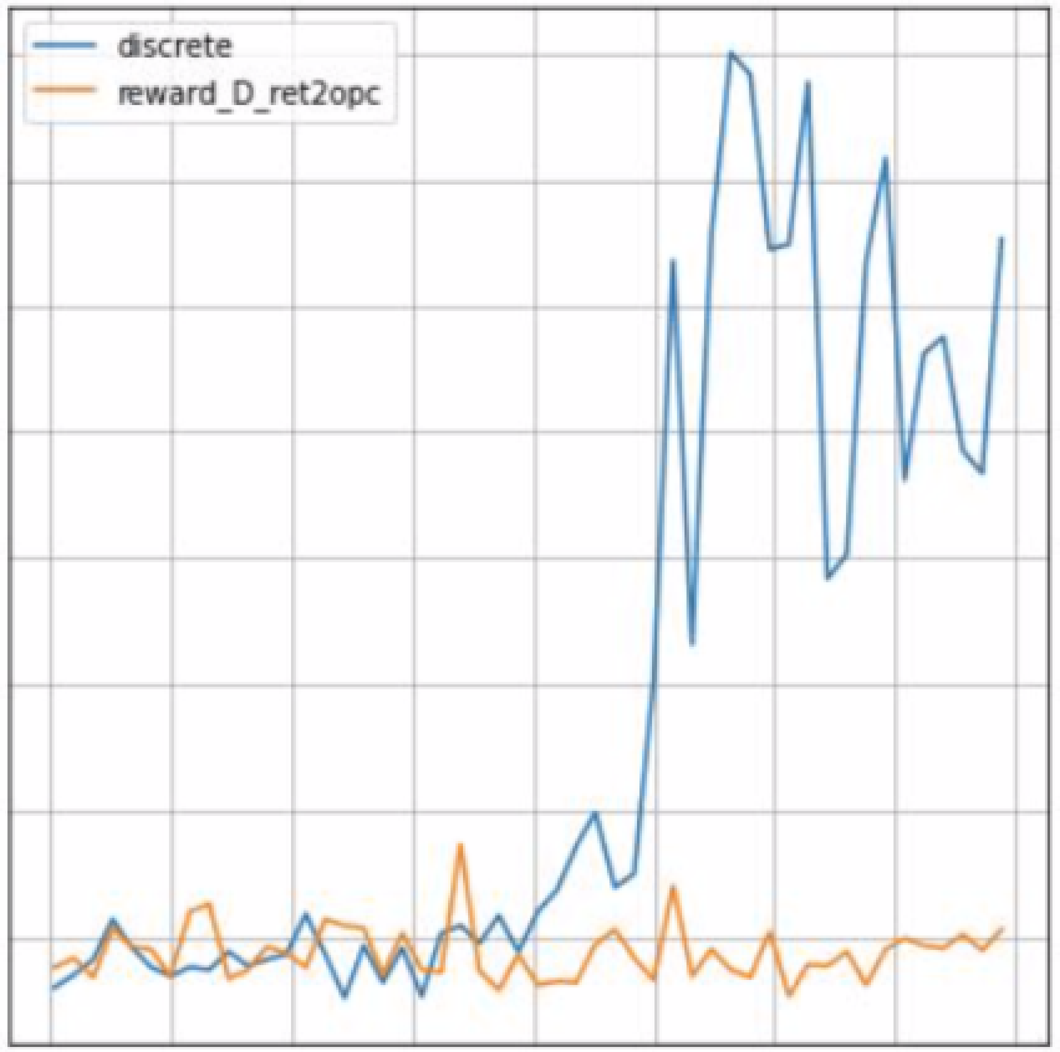
Discrete



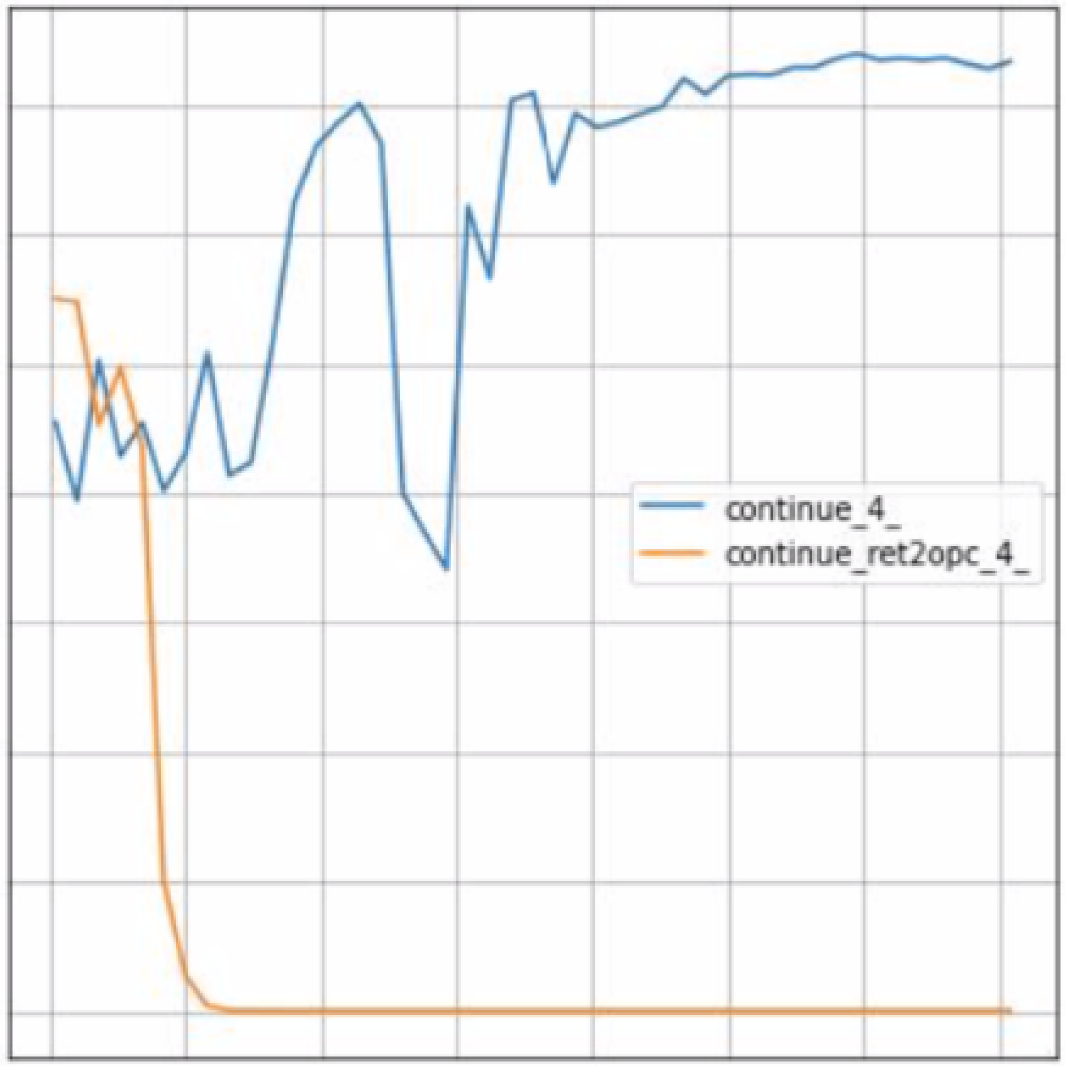
Continous

Ablation study

$Q_{ret} \rightarrow Q_{opc}$:



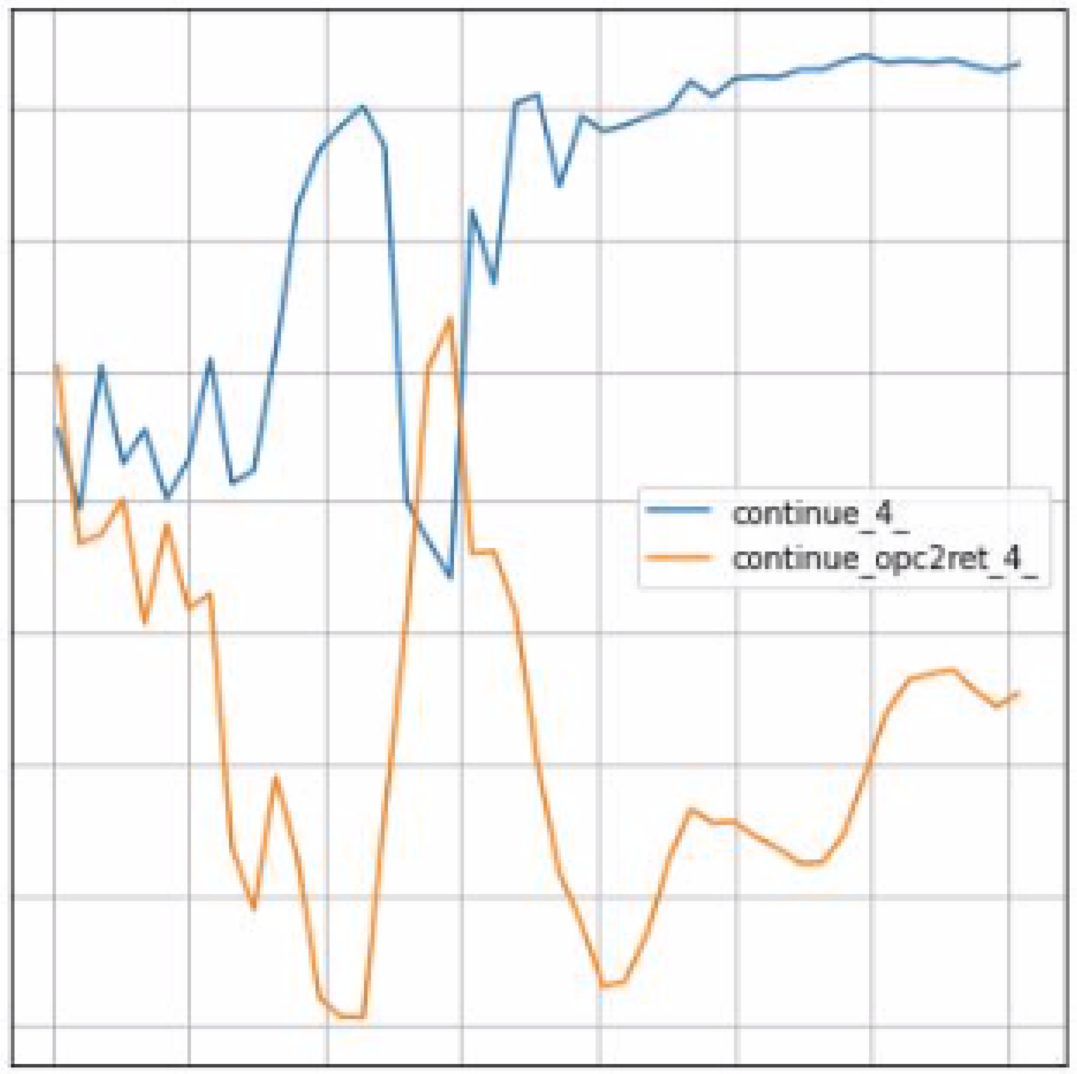
Discrete



Continous

Ablation study

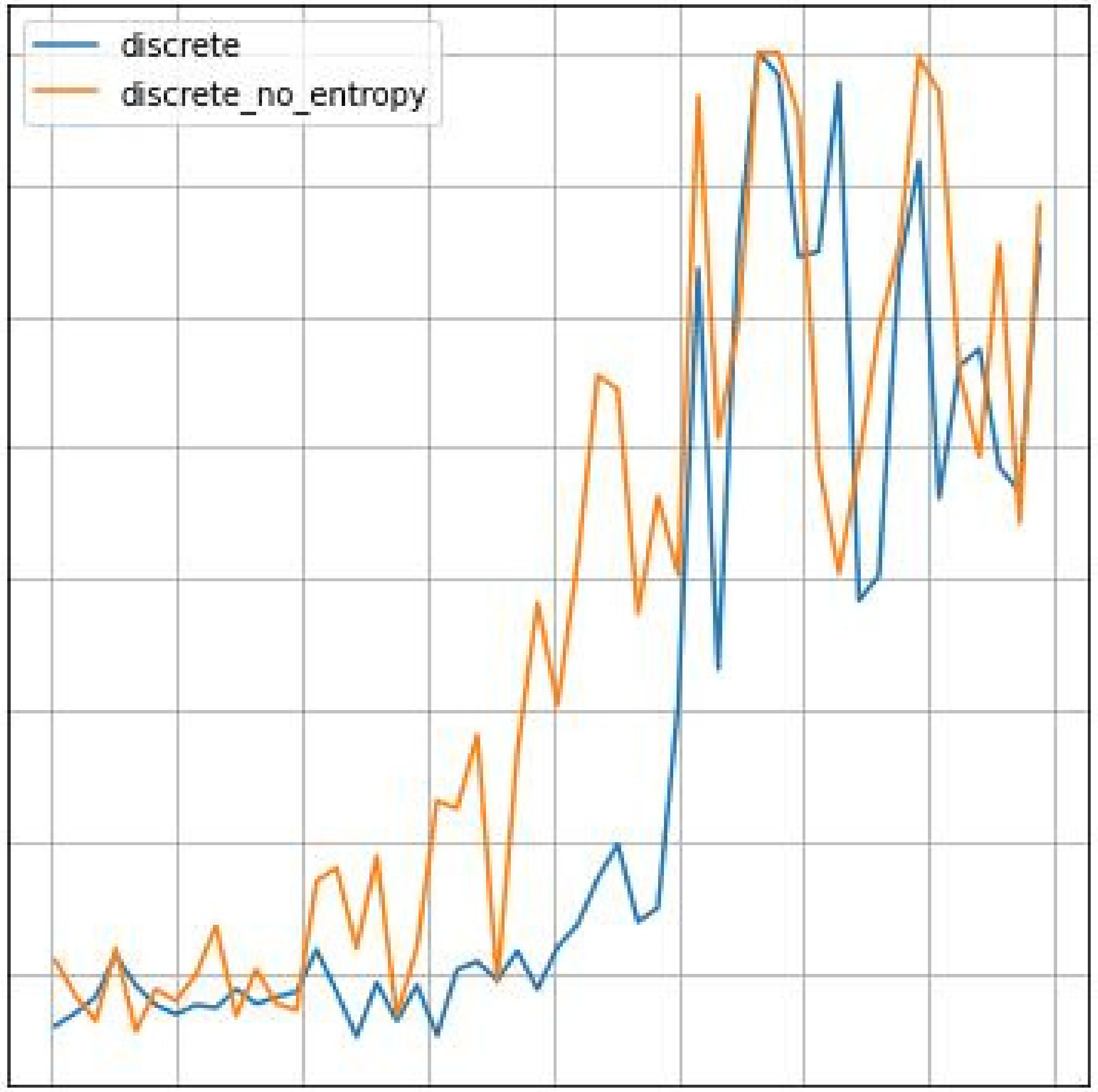
$Q_{opc} \rightarrow Q_{ret}$:



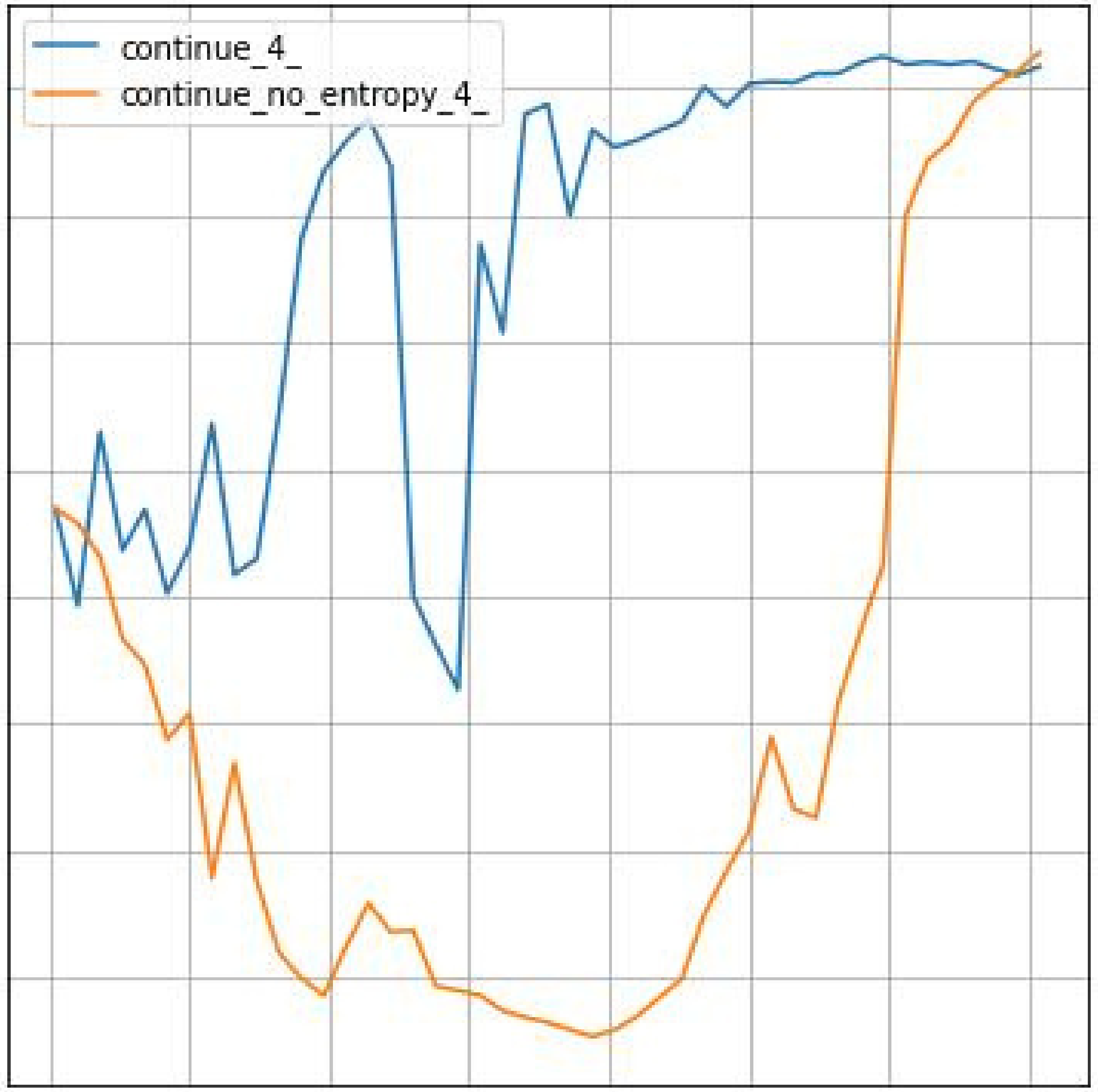
Continous

Ablation study

Remove entropy



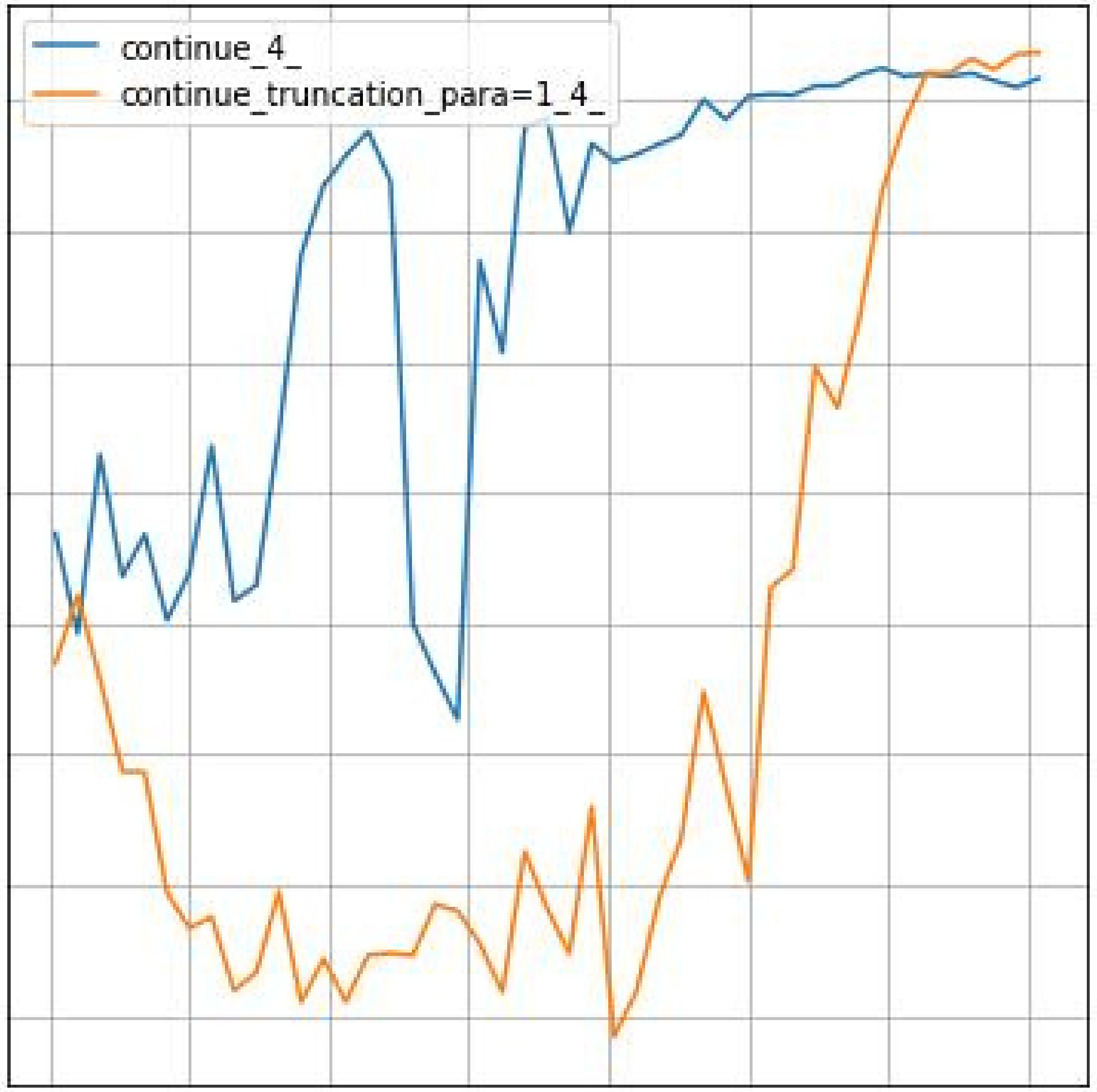
Discrete



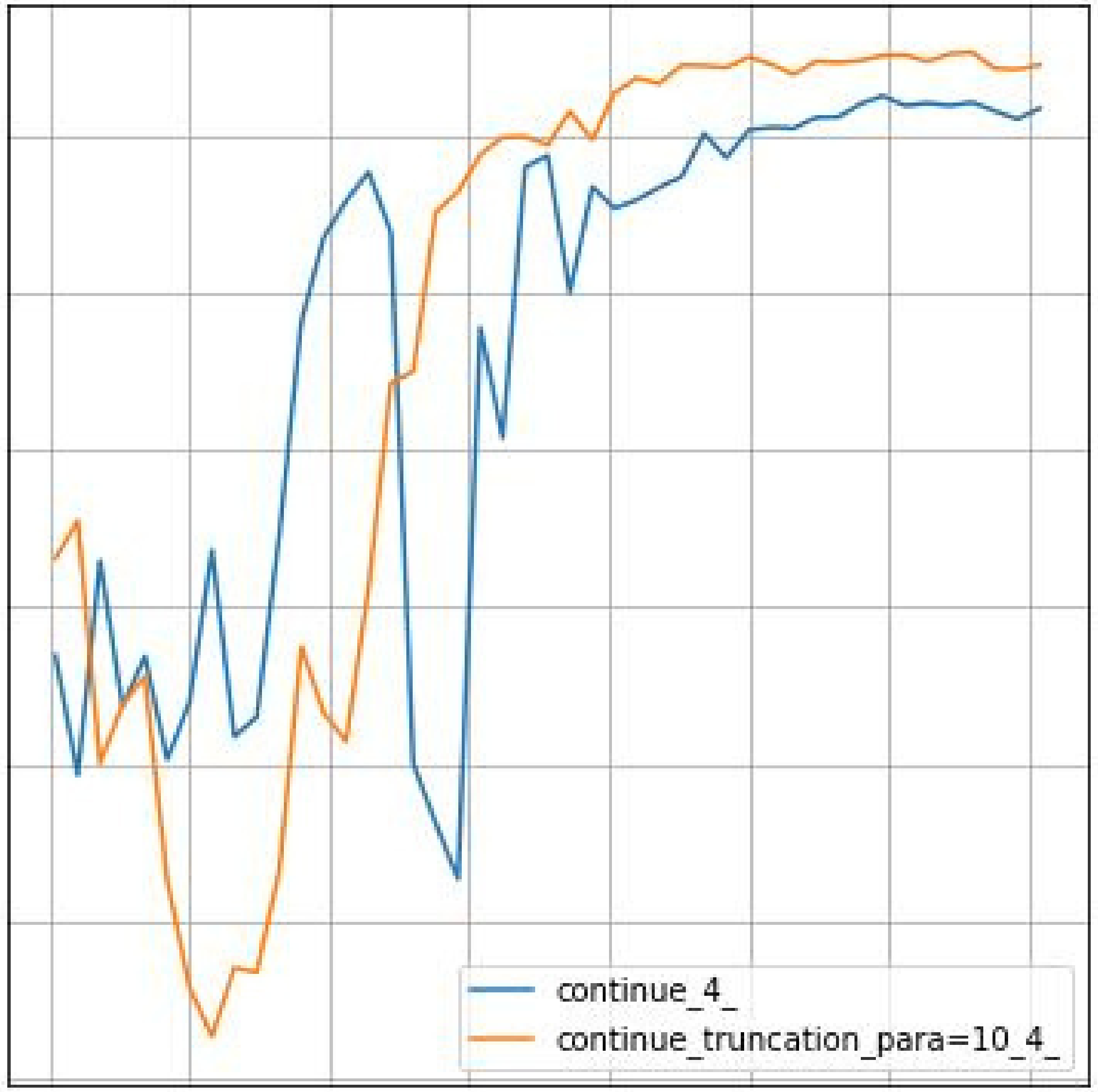
Continous

Ablation study

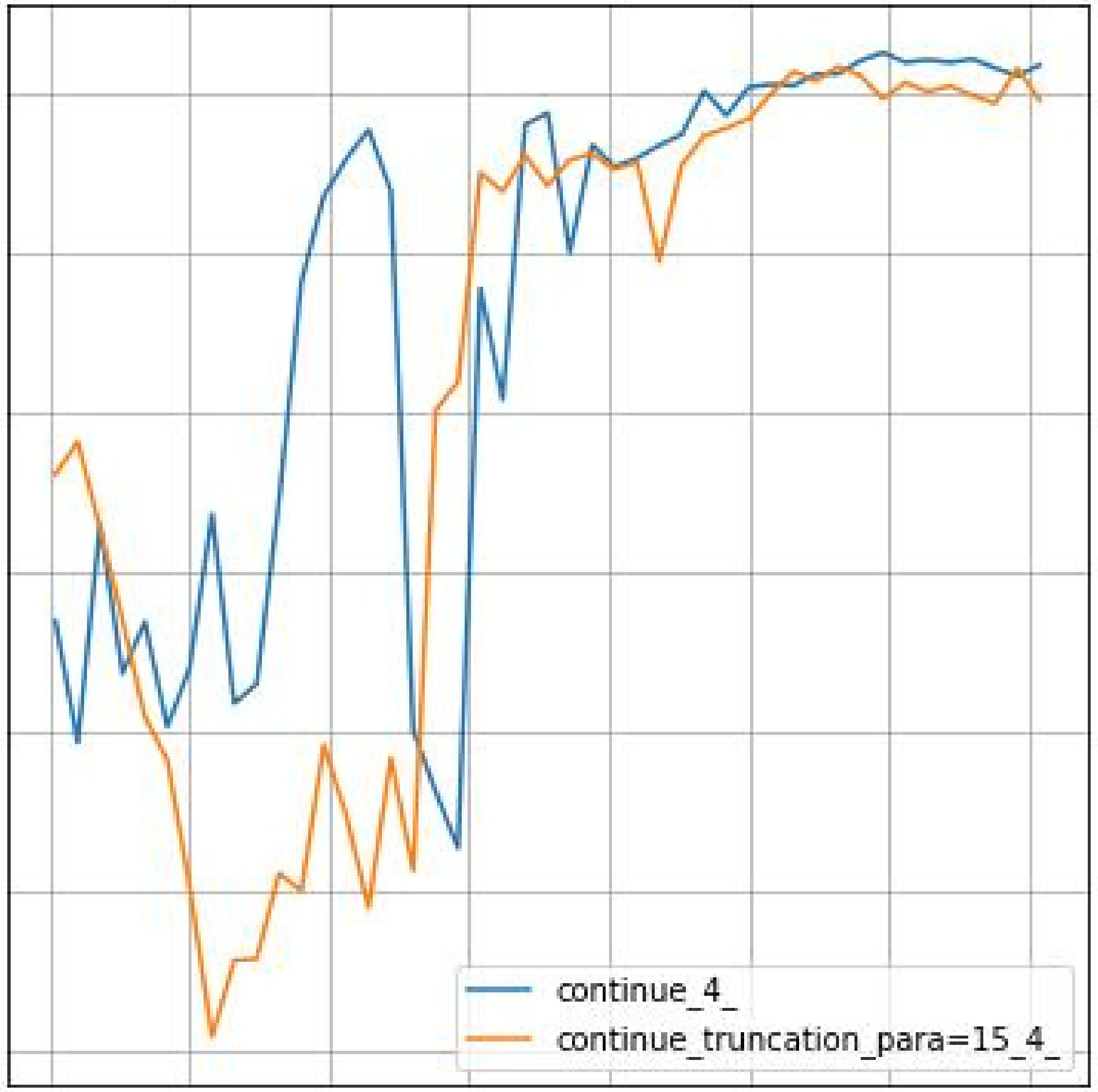
Truncation Param Change -- Continuous



D_TC = 1



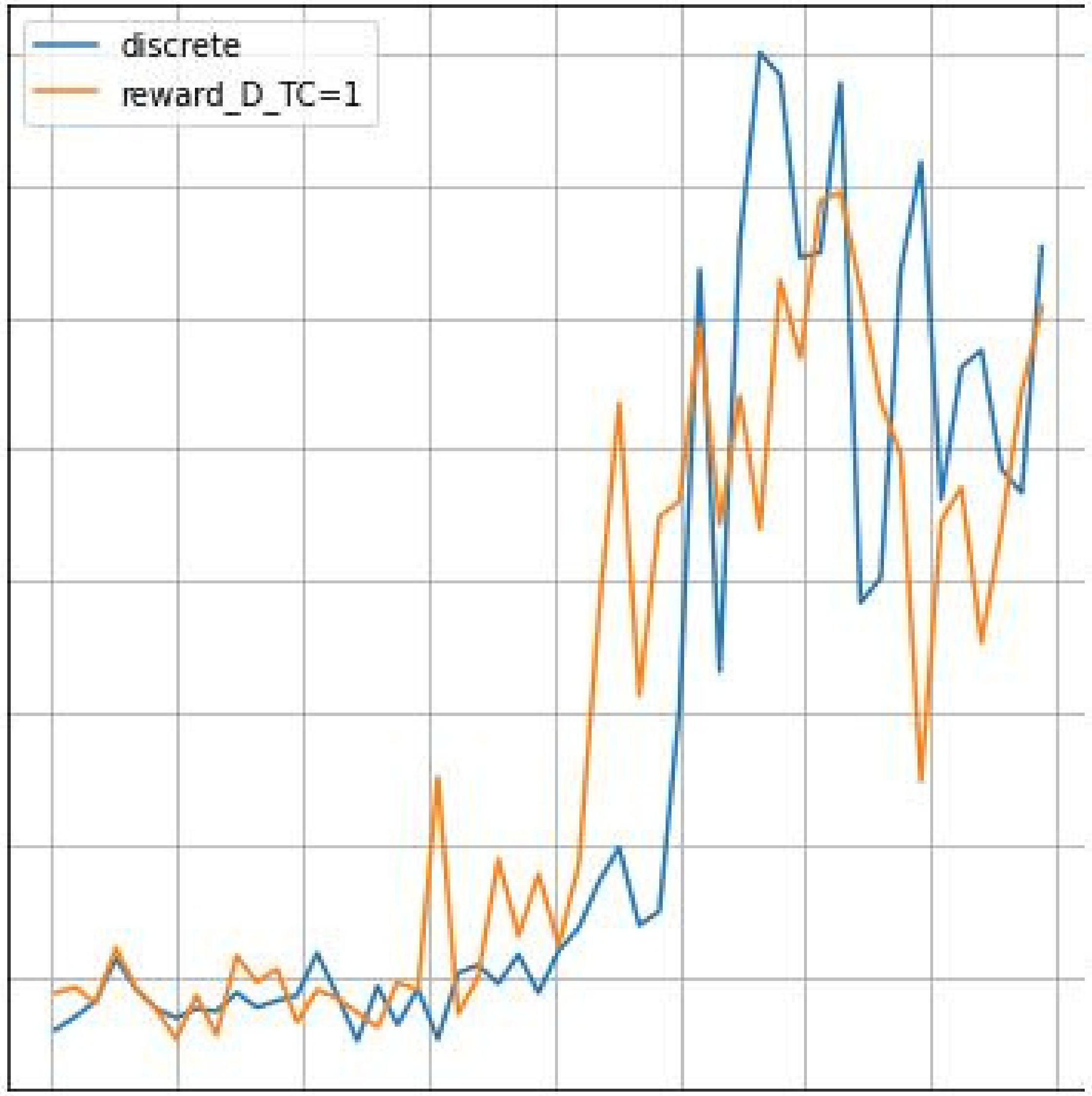
D_TC = 10



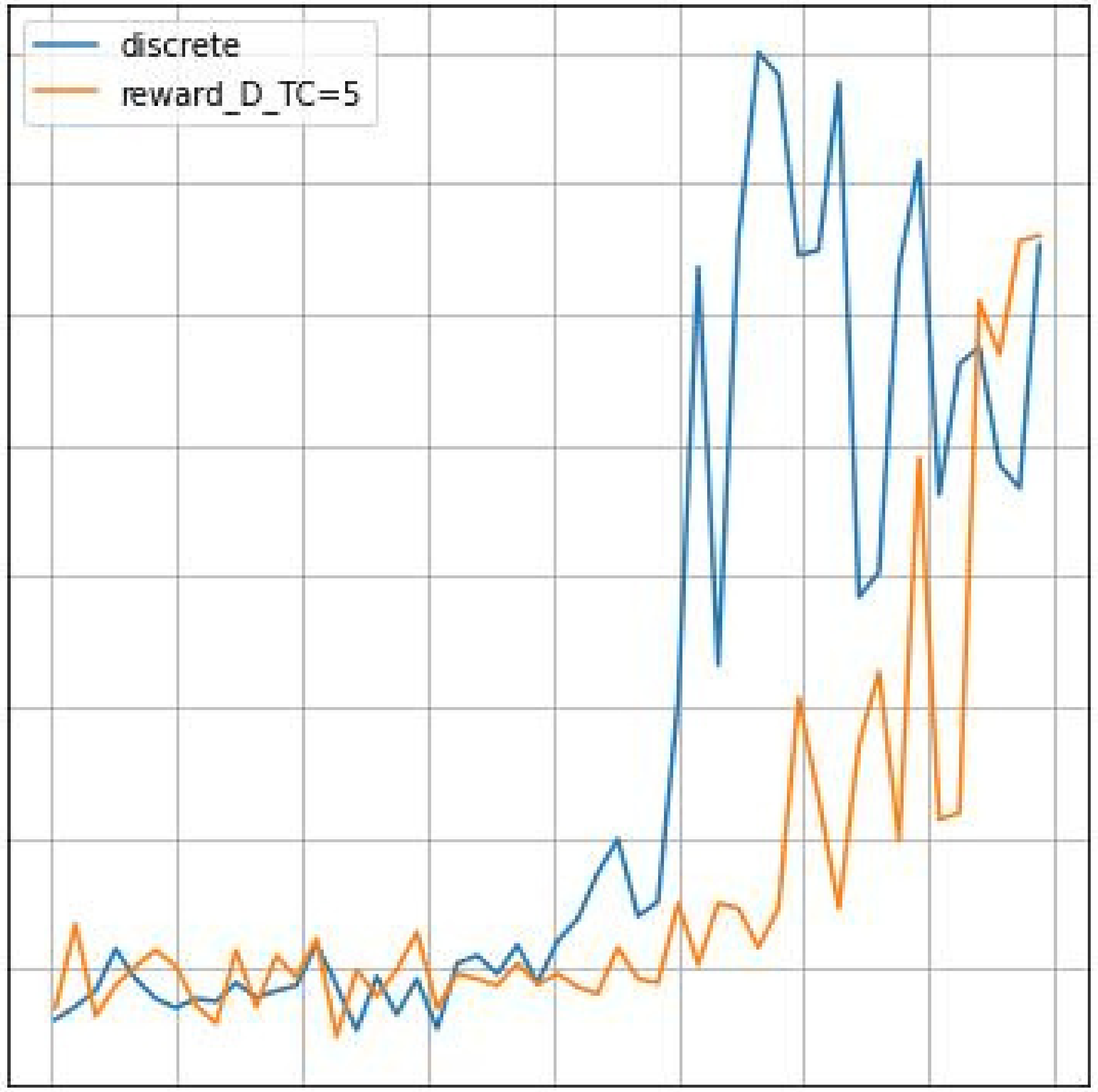
D_TC = 15

Ablation study

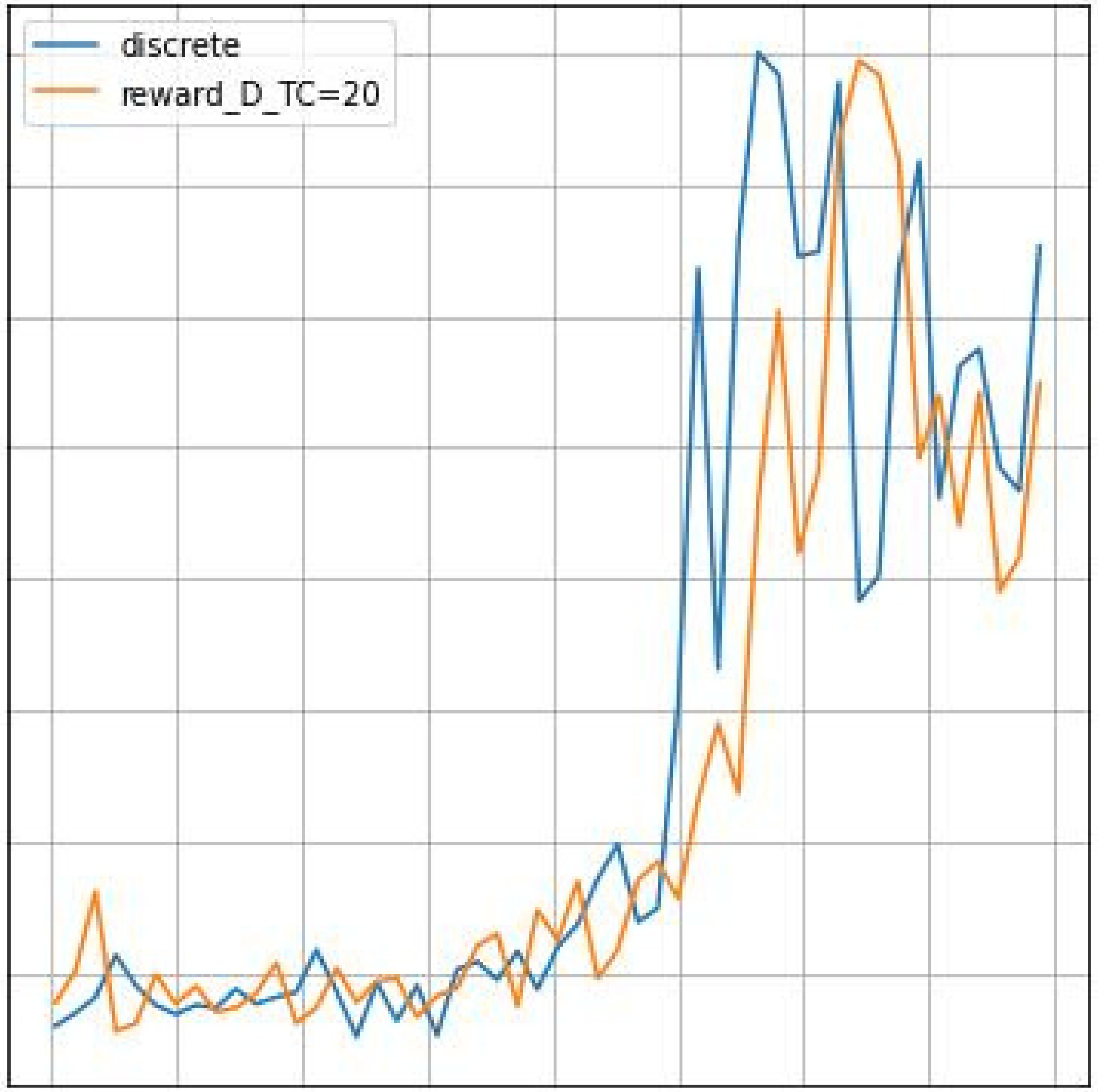
Truncation Param Change -- Discrete



D_TC = 1



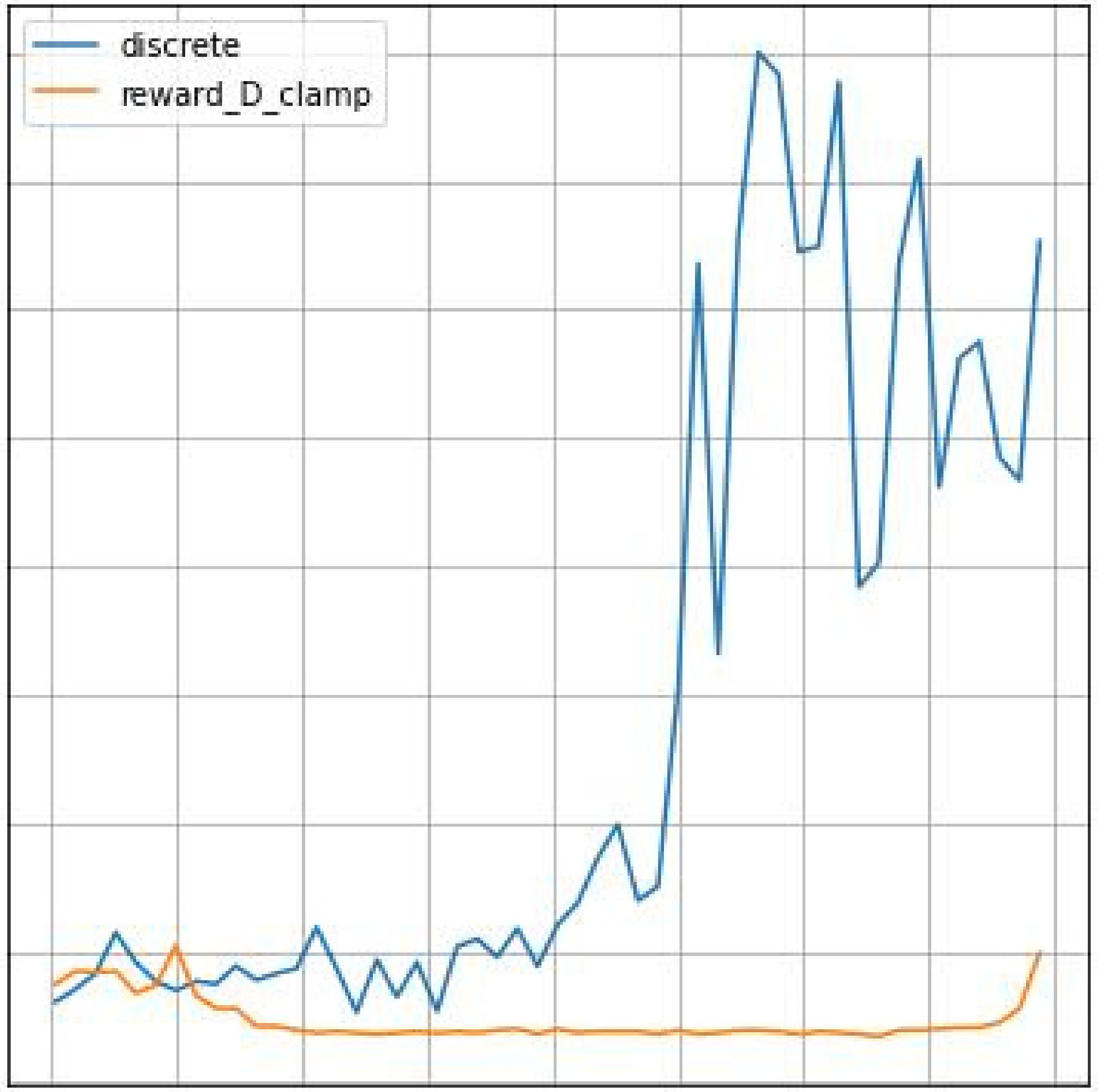
D_TC = 5



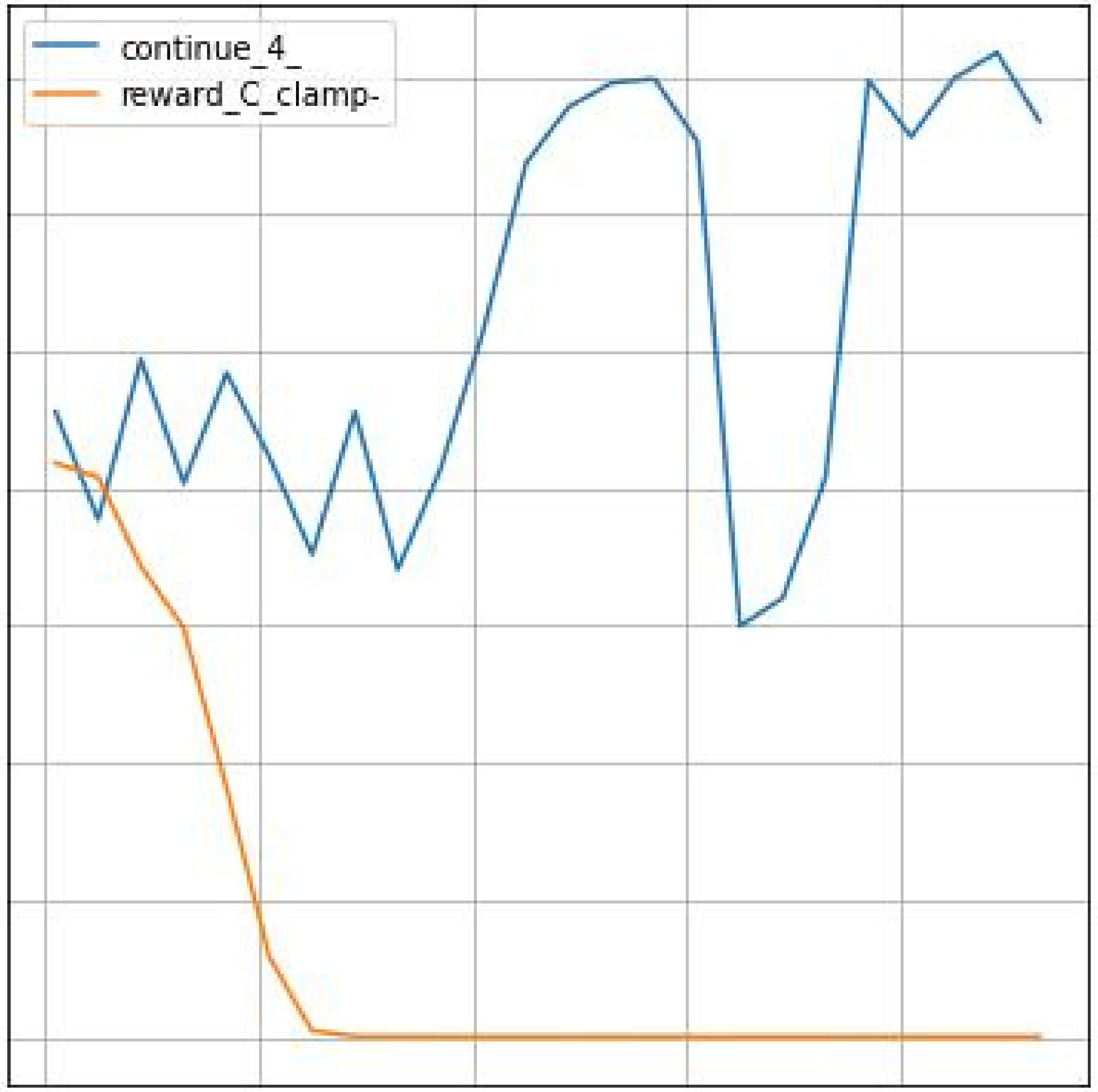
D_TC = 20

Ablation study

remove clamping



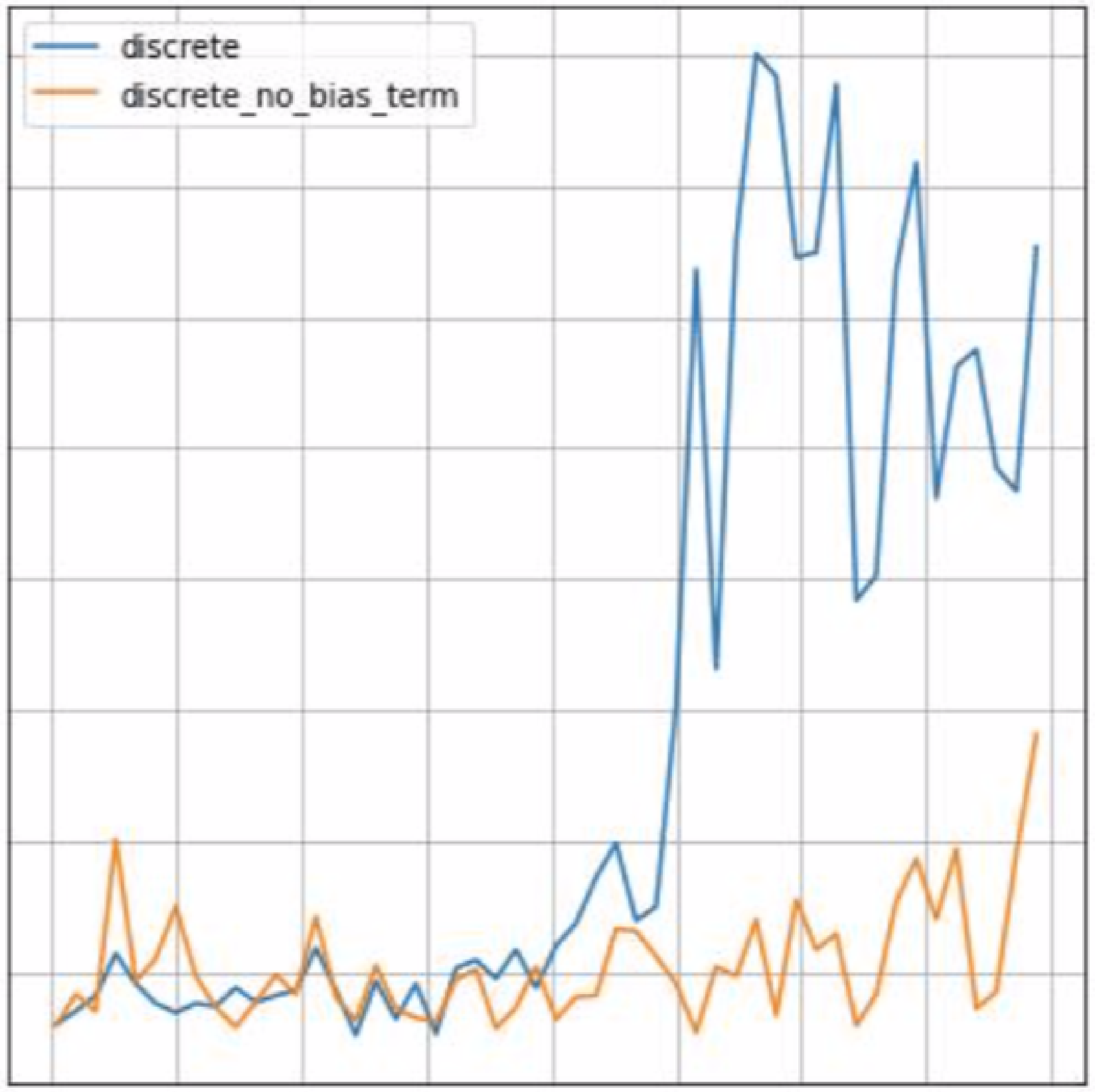
Discrete



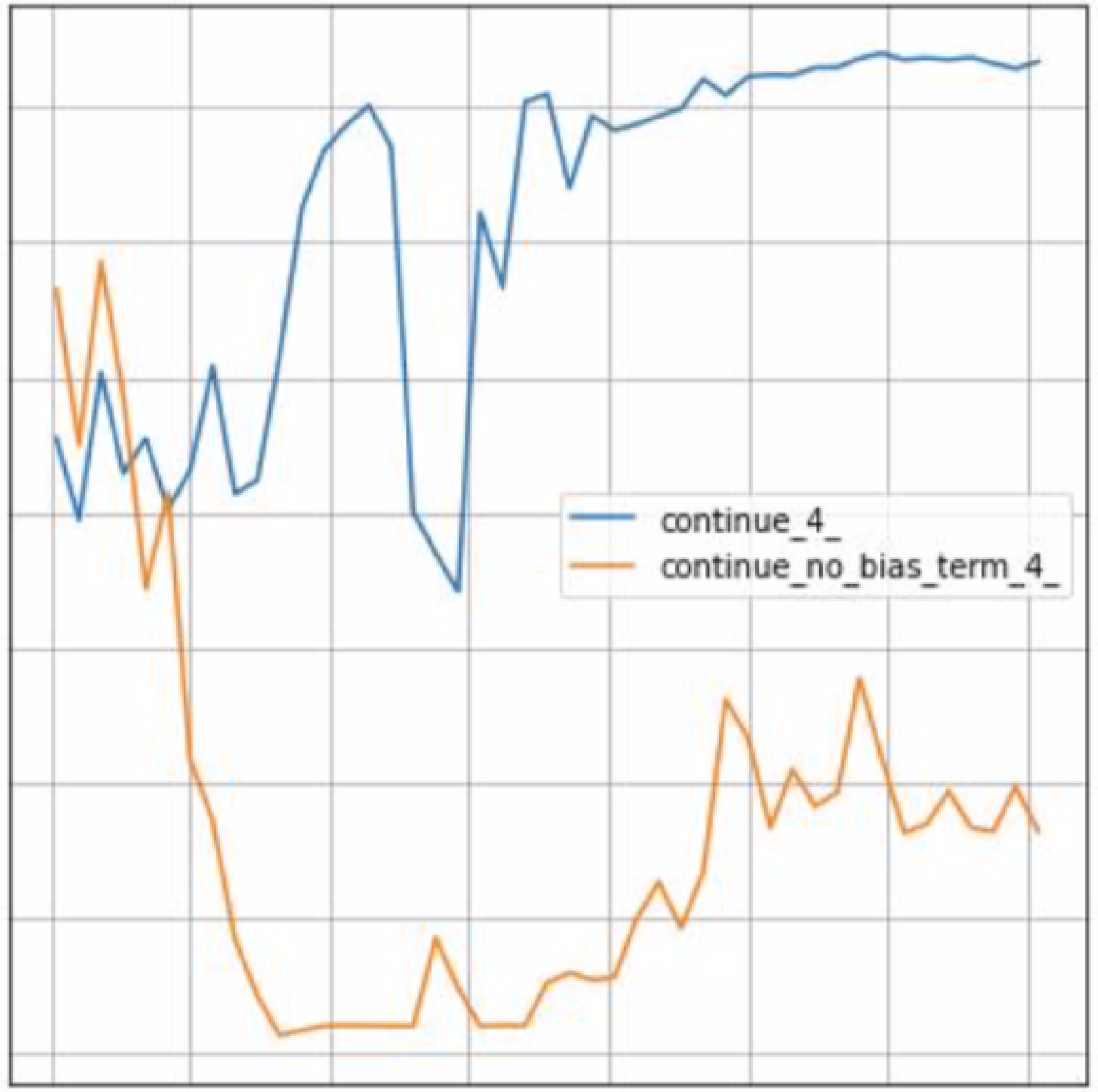
Continous

Ablation study

no bias term



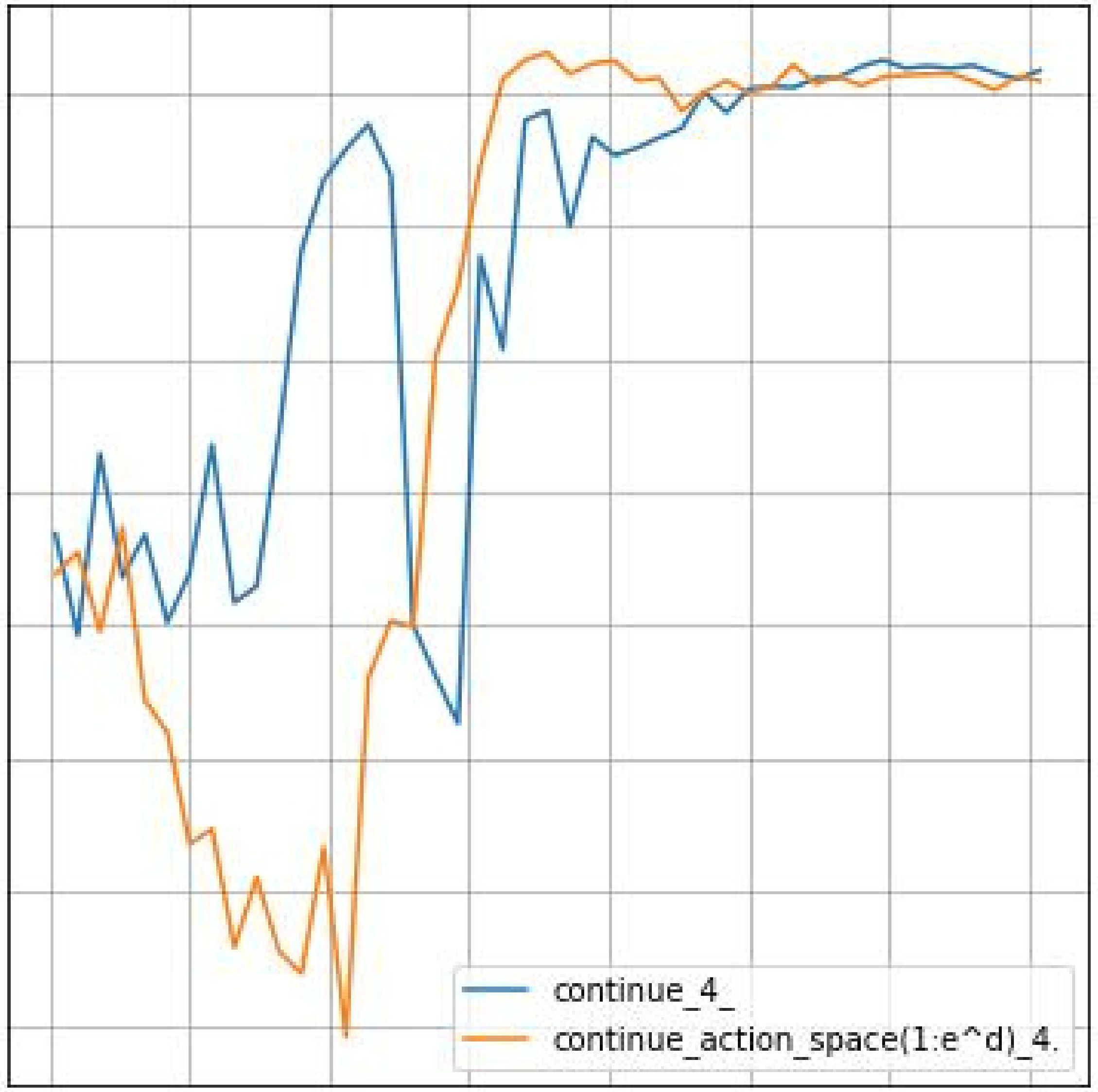
Discrete



Continous

Ablation study

d -> e^d



$$\bar{\rho}_t = \min\left\{1, \left(\frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}\right)^{\frac{1}{d}}\right\}$$

Continous

Conclusion on Ablation study

- Trust Region method don't really have large effect for the environment(mountain car & carpole) we test on
- replacing Qopt with Qret or replacing in reverse leads to worse performance
- clamping and bias term makes great effect on training
- Entropy term provides faster learning

Thank you for listening!