Implicit Q Learning: Improvements on Antmaze

Reporter:
廖修誼 (111652017)
吳泓諺 (111652040)
王裕昕 (111550066)

Outline

- Part I (Introduction)
- Part II (Interesting findings)
- Part III (Tech. 1 - Distribution model)
- Part IV (Tech. 2 - D2RL)
- Part V (Tech. 3 - Bounus Reward)
- Part VI (Conclusion)

Part I (Introduction)

IQL Setting

- Offline RL
- main insight: no need to evaluate OOD actions
- method: approx. an upper expectile of distribution
- Goal: minimizing the deviation from the behavior policy

Ref: (IQL) https://arxiv.org/abs/2110.06169

IQL algorithm

Algorithm 1 Implicit Q-learning

Initialize parameters ψ, θ, $\hat{\theta}$, ϕ.
TD learning (IQL):
for each gradient step **do**
$\quad \psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$
$\quad \theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$
$\quad \hat{\theta} \leftarrow (1-\alpha)\hat{\theta} + \alpha\theta$
end for
Policy extraction (AWR):
for each gradient step **do**
$\quad \phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$
end for

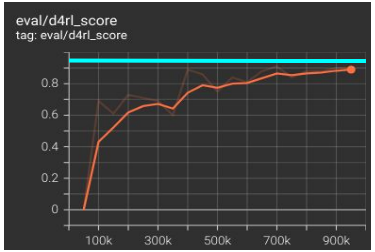
$$L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right]$$

$$L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a)\right)^2\right]$$

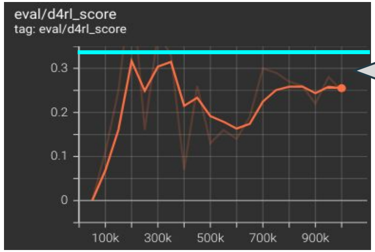
$$L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right)\log\pi_\phi(a \mid s)\right]$$

Ref: (IQL) https://arxiv.org/abs/2110.06169

Result Reproduce (antmaze-v0)

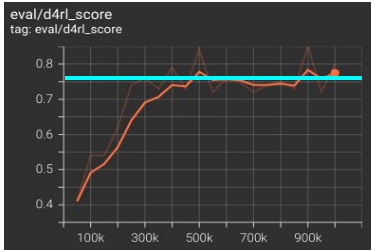


umaze, $\tau = 0.9$

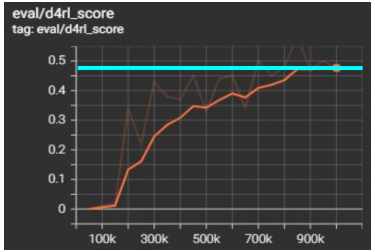


large, $\tau = 0.9$

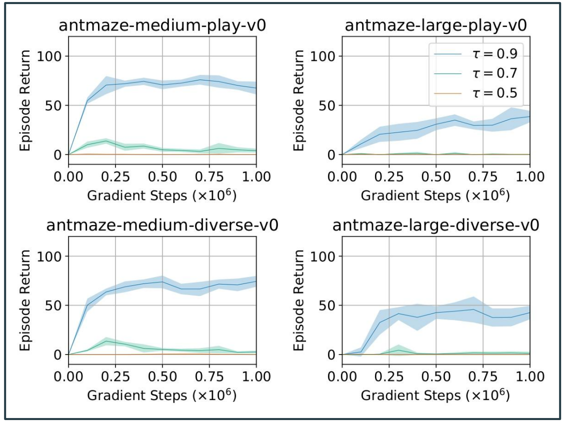
loss hyper parameter



medium, $\tau = 0.9$



large-diverse, $\tau = 0.9$



Experiment

Dataset	BC	10%BC	DT	AWAC	Onestep RL	TD3+BC	CQL	IQL (Ours)
halfcheetah-medium-v2	42.6	42.5	42.6	43.5	**48.4**	**48.3**	44.0	**47.4**
hopper-medium-v2	52.9	56.9	**67.6**	57.0	59.6	59.3	58.5	**66.3**
walker2d-medium-v2	75.3	75.0	74.0	72.4	**81.8**	83.7	72.5	78.3
halfcheetah-medium-replay-v2	36.6	40.6	36.6	40.5	38.1	**44.6**	**45.5**	**44.2**
hopper-medium-replay-v2	18.1	75.9	82.7	37.2	**97.5**	60.9	**95.0**	**94.7**
walker2d-medium-replay-v2	26.0	62.5	66.6	27.0	49.5	**81.8**	77.2	73.9
halfcheetah-medium-expert-v2	55.2	**92.9**	86.8	42.8	**93.4**	**90.7**	**91.6**	86.7
hopper-medium-expert-v2	52.5	**110.9**	**107.6**	55.8	103.3	98.0	**105.4**	91.5
walker2d-medium-expert-v2	**107.5**	**109.0**	**108.1**	74.5	**113.0**	**110.1**	**108.8**	**109.6**
locomotion-v2 total	466.7	**666.2**	**672.6**	450.7	**684.6**	**677.4**	**698.5**	**692.4**
antmaze-umaze-v0	54.6	62.8	59.2	56.7	64.3	78.6	74.0	**87.5**
antmaze-umaze-diverse-v0	45.6	50.2	53.0	49.3	60.7	71.4	**84.0**	62.2
antmaze-medium-play-v0	0.0	5.4	0.0	0.0	0.3	10.6	61.2	**71.2**
antmaze-medium-diverse-v0	0.0	9.8	0.0	0.7	0.0	3.0	53.7	**70.0**
antmaze-large-play-v0	0.0	0.0	0.0	0.0	0.0	0.2	15.8	**39.6**
antmaze-large-diverse-v0	0.0	6.0	0.0	1.0	0.0	0.0	14.9	**47.5**
antmaze-v0 total	100.2	134.2	112.2	107.7	125.3	163.8	303.6	**378.0**
total	566.9	800.4	784.8	558.4	809.9	841.2	1002.1	**1070.4**
kitchen-v0 total	**154.5**	-	-	-	-	-	144.6	**159.8**
adroit-v0 total	104.5	-	-	-	-	-	93.6	**118.1**
total+kitchen+adroit	825.9	-	-	-	-	-	1240.3	**1348.3**
runtime	10m	10m	960m	20m	$\approx$ 20m*	20m	80m	20m

Ref: (IQL) https://arxiv.org/abs/2110.06169

Part II (Interesting findings)

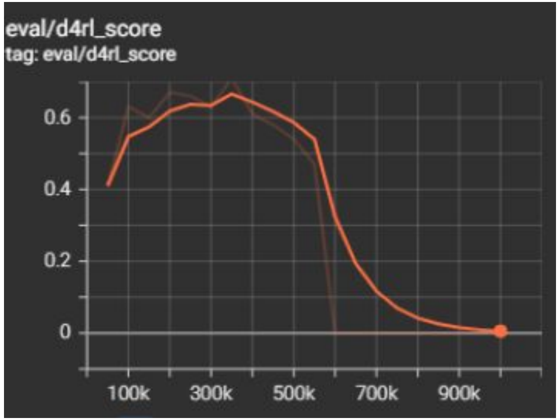
Rethink on the equation

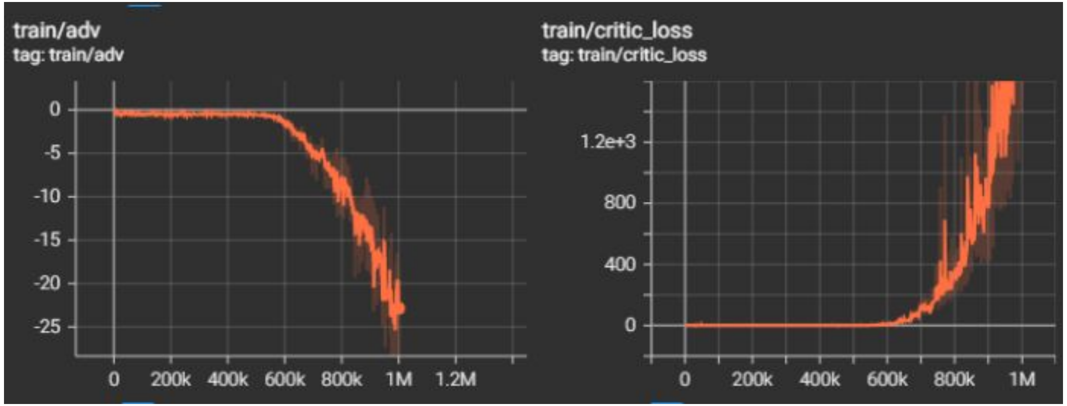
value network: $\boxed{L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right]}$ $\longrightarrow$ **Q < V**

Q network: $\boxed{L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma V_\psi\left(s'\right) - Q_\theta(s,a)\right)^2\right]}$ $\longrightarrow$ **Q -> r + γ * V**

Policy: $\boxed{L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right)\log\pi_\phi(a \mid s)\right]}$

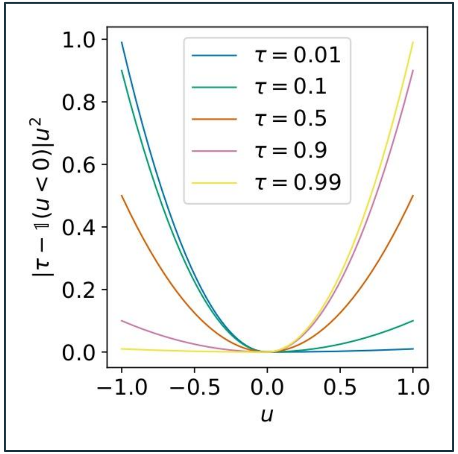
What if there is no double Q in implementation





adv = q - v

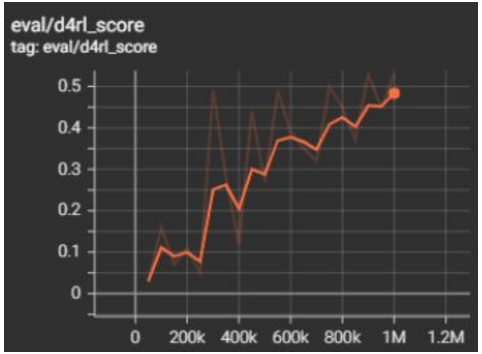
Meaning of τ



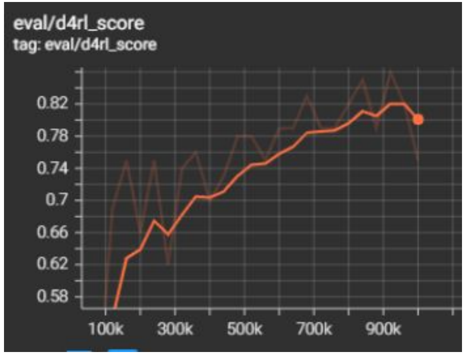
$$L_2^\tau(u) = |\tau - 1(u < 0)|u^2$$

Lemma 2. *For all s, τ_1 and τ_2 such that $\tau_1 < \tau_2$ we get*

$$V_{\tau_1}(s) \leq V_{\tau_2}(s).$$



expectile = 0.8



expectile = 0.9

Ref: (IQL) https://arxiv.org/abs/2110.06169

Evolve from quantile regression loss

The Quantile Regression Loss

- Given that the derivative of $L(x; Z, \tau)$ is $F_Z(x) - \tau$, we can recover the QR loss by integration

Quantile regression (QR) loss:

$$L_{QR}(x; Z, \tau) = (\tau - 1)\int_{-\infty}^{x} (z - x)dF_Z(z) + \tau \int_{x}^{\infty} (z - x)dF_Z(z)$$

(It is easy to verify that $\frac{d}{dx}L_{QR}(x; Z, \tau) = F_Z(x) - \tau$ by the Leibniz integral rule)

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)\,dt\right) = f(x,b(x))\cdot\frac{d}{dx}b(x) - f(x,a(x))\cdot\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)}\frac{\partial}{\partial x}f(x,t)\,dt$$

Alternative expression of QR loss:

$$\rho_\tau(y) := y(\tau - \mathbb{I}\{y < 0\})$$

$$L_{QR}(x; Z, \tau) = E_Z[\rho_\tau(Z - x)]$$

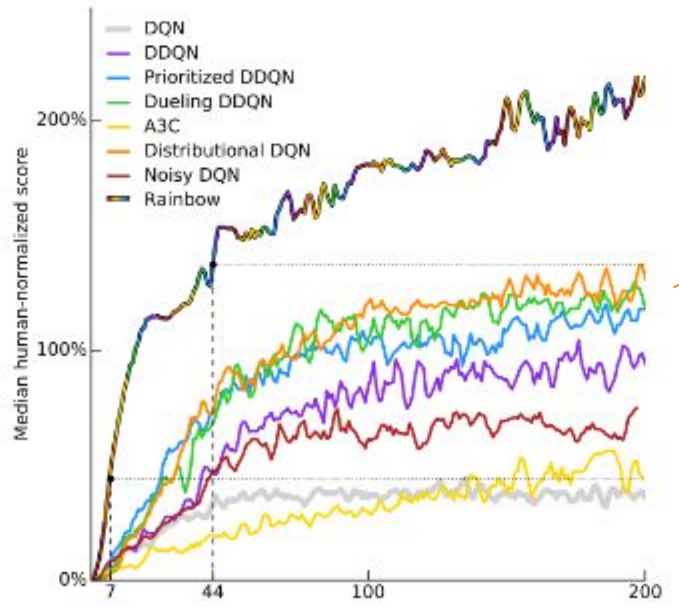
Value loss

$$L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\hat\theta}(s,a) - V_\psi(s)\right)\right]$$

$$L_2^\tau(u) = |\tau - \mathbb{1}(u < 0)|u^2.$$

Part III (Tech. 1 - Distribution model)

Motivation idea



improve performance with Distributionalizing IQL

Comparison : our method & quantile

implicit quantile network

$$\delta_t^{\tau,\tau'} = r_t + \gamma Z_{\tau'}(x_{t+1}, \pi_\beta(x_{t+1})) - Z_\tau(x_t, a_t). \quad (2)$$

Then, the IQN loss function is given by

$$\mathcal{L}(x_t, a_t, r_t, x_{t+1}) = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa} \left(\delta_t^{\tau_i, \tau_j'} \right), \quad (3)$$

Distributional IQL

$$\boxed{L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[L_2^\tau \left(Q_{\hat{\theta}}(s,a) - V_\psi(s) \right) \right]}$$

$$\boxed{L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\left(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a) \right)^2 \right]}$$

$$\text{mean\_loss} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{T} \sum_{k=1}^{T} (q_{ij} - v_{ik})^2 \cdot \left| \tau - 1_{\{q_{ij} - v_{ik} < 0\}} \right| \cdot \text{prob}_{ij} \cdot \text{prob}_{ik}$$

implementation follows from IQN

Guess , benefit of our method :
Two step policy improvement

$$\boxed{L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a)\right)^2\right]}$$

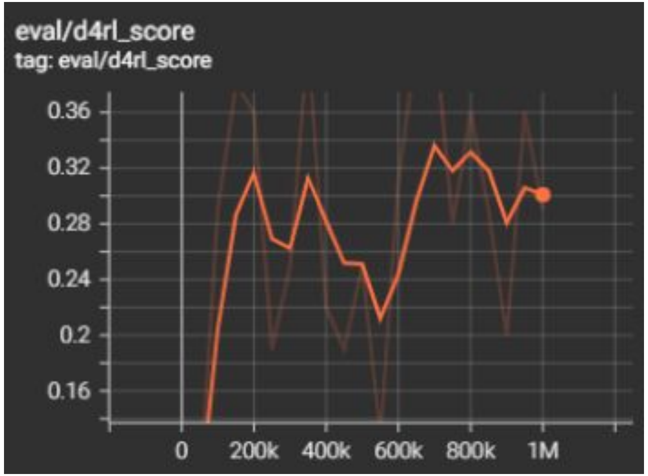
$$\boxed{L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right]}$$

$$\delta = r + \gamma G_{\theta'}(\tau'; s', a') - G_\theta(\tau; s, a).$$

$$\mathcal{L}_\kappa(\delta;\tau) = \begin{cases} |\tau - \mathbb{1}(\delta < 0)| \cdot \delta^2/(2\kappa) & \text{if } |\delta| \leq \kappa \\ |\tau - \mathbb{1}(\delta < 0)| \cdot (|\delta| - \kappa/2) & \text{otherwise .} \end{cases}$$

However, experiment result is …

antmaze-medium-play-v0 – expectile 0.9



antmaze-large-play-v0, expectile 0.9

Part IV (Tech. 2 - D2RL)

Motivation idea

- The problem of choosing architecture designs has been largely ignored.
- Information loss when forwarding through layers.
- The effective rank of the feature matrix is low.
- Add skip connections from the input.

Ref: https://arxiv.org/pdf/2010.09163

Skip Connection

original:

nn.Linear(in_dim, hidden_dim)

nn.Linear(hidden_dim, hidden_dim)

nn.Linear(hidden_dim, out_dim)

with skip connections:

nn.Linear(in_dim, hidden_dim)

nn.Linear(in_dim + hidden_dim, hidden_dim)

nn.Linear(hidden_dim, out_dim)

Ref: https://arxiv.org/pdf/2010.09163

Results — srank

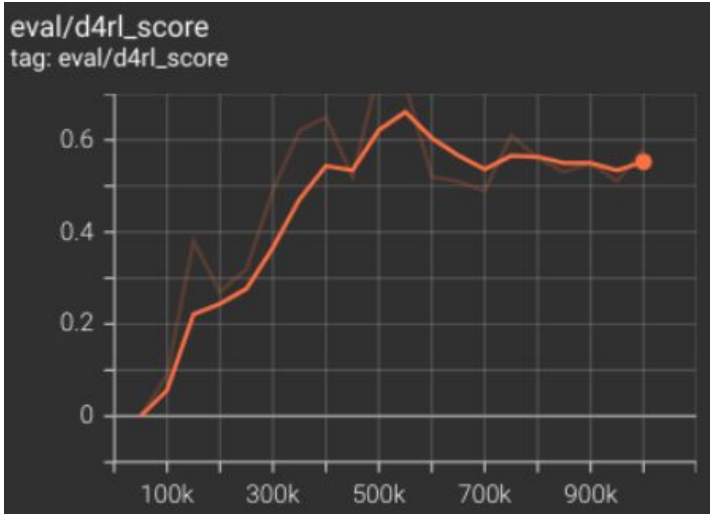
$$srank_\delta(\Phi) = \min\{k : \frac{\sum_{i=1}^{k} \sigma_i(\Phi)}{\sum_{i=1}^{d} \sigma_i(\Phi)} \geq 1 - \delta\}$$

	antmaze-large		halfcheetah-expert	
1M-steps	IQL	IQL+SC	IQL	IQL+SC
Policy	227	**232**	226	**232**
Q-network	223	**231**	225	**233**

Results

- competitive performance similar to original IQL
- outperforms IQL on antmaze-large
- prevent the reduction of effective ranks

better convergence
on antmaze-large



Part V (Tech. 3 - Bonus Reward)

Motivation idea

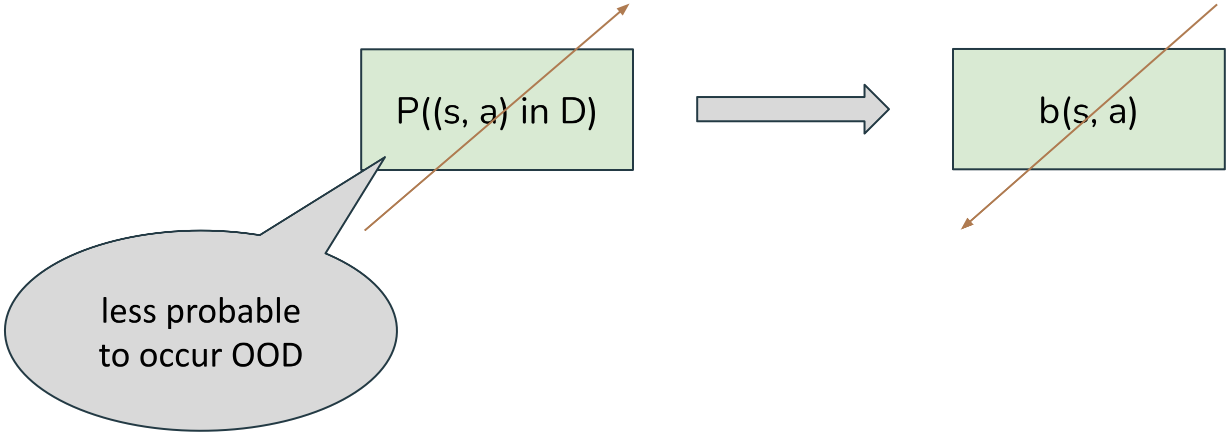
Offline Reinforcement Learning as Anti-Exploration

Shideh Rezaeifar*[1], Robert Dadashi*[2], Nino Vieillard[2,3], Léonard Hussenot[2,4], Olivier Bachem[2], Olivier Pietquin[2], and Matthieu Geist[2]

[1]University of Geneva
[2]Google Research, Brain Team
[3]Univ. Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France
[4]Univ. Lille, CNRS, Inria Scool, UMR 9189 CRIStAL

Ref: (anti-exploration) https://arxiv.org/abs/2106.06431

(subtractive) Bounus Reward Property

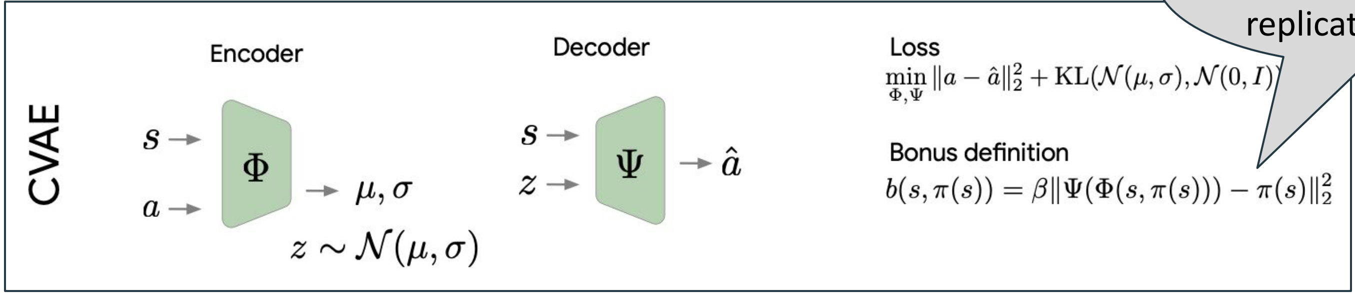


Ref: (anti-exploration) https://arxiv.org/abs/2106.06431

Train Bounus Reward (CVAE)

Algorithm 1 CVAE training.

1: Initialize CVAE networks Φ and Ψ
2: **for** step $i = 0$ to N **do**
3: Sample a minibatch of k state-action pairs $\{(s_t, a_t), t = 1, .., k\}$ from $\mathcal{D}$
4: Train Φ and Ψ using $\mathcal{L}_{\Phi,\Psi}$, see Eq. (5)



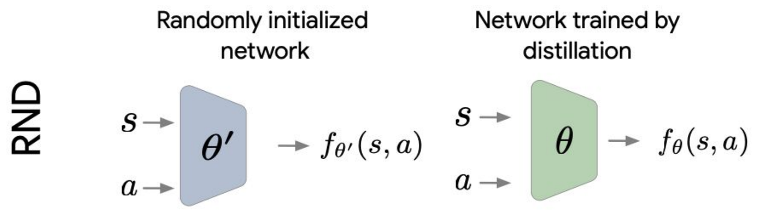
depends on complexity of replicating action

Ref: (anti-exploration) https://arxiv.org/abs/2106.06431

Train Bounus Reward (RND)

Algorithm 1 CVAE training.

1: Initialize CVAE networks Φ and Ψ
2: **for** step $i = 0$ to N **do**
3: Sample a minibatch of k state-action pairs $\{(s_t, a_t), t = 1, .., k\}$ from $\mathcal{D}$
4: Train Φ and Ψ using $\mathcal{L}_{\Phi,\Psi}$, see Eq. (5)



Loss
$$\min_{\theta} \|f_\theta(s,a) - f_{\theta'}(s,a)\|_2^2$$

Bonus definition
$$b(s,\pi(s)) = \beta \|f_\theta(s,\pi(s)) - f_{\theta'}(s,\pi(s))\|_2^2$$

depends on complexity of replicating action

Ref: (anti-exploration) https://arxiv.org/abs/2106.06431

Apply Bounus Reward to IQL

Algorithm 1 Implicit Q-learning

Initialize parameters ψ, θ, $\hat{\theta}$, ϕ.
TD learning (IQL):
for each gradient step **do**
$\psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$
$\theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$
$\hat{\theta} \leftarrow (1-\alpha)\hat{\theta} + \alpha\theta$
end for
Policy extraction (AWR):
for each gradient step **do**
$\phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$
end for

$$L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(\cancel{Q_{\hat{\theta}}(s,a)} - V_\psi(s)\right)\right]$$

$$\boxed{L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(\left(Q_{\hat{\theta}}(s,a) - b(s,a)\right) - V_\psi(s)\right)\right]}$$

$$L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma \cancel{V_\psi(s')} - Q_\theta(s,a)\right)^2\right]$$

$$\boxed{L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\mathbb{E}_{a'=\mu_{theta(s')},\epsilon\sim N(0,\sigma I)}\left[\left(r(s,a) + \gamma\left(Q_\psi\left(s', a'+\epsilon\right) - b\left(s', a'\right)\right) - Q_\theta(s,a)\right)\right]^2\right]}$$

$$L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(\cancel{Q_{\hat{\theta}}(s,a)} - V_\psi(s)\right)\right)\log\pi_\phi(a \mid s)\right]$$

$$\boxed{L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\hat{\theta}}(s,a) - b(s,a)\right) - V_\psi(s)\right)\log\pi_\phi(a \mid s)\right]}$$

Ref: (IQL) https://arxiv.org/abs/2110.06169

Part VI (Conclusion)

Recall

- Part I (Introduction)
- Part II (interesting findings)
- Part III (Tech. 1 - Distribution model): modify update formulas to distribution form
- Part IV (Tech. 2 - D2RL): modify neural layers
- Part V (Tech. 3 - bonus reward): modify reward

Thank you for your attention