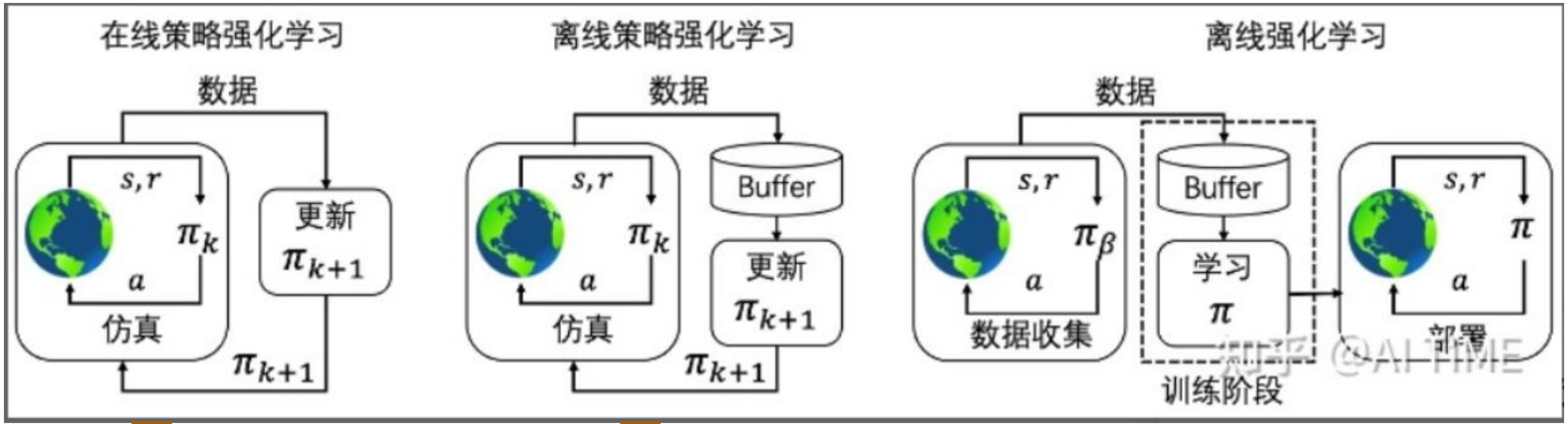
Believe What You See: Implicit Constraint Approach for Offline Multi-Agent Reinforcement Learning

NIPS 2021

Offline RL



Data comes from interacting

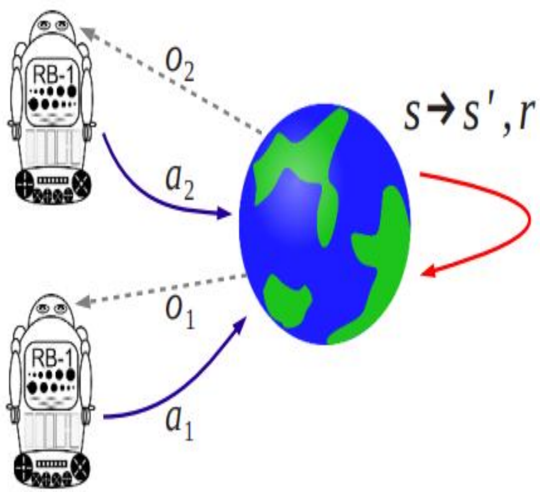
Data comes from past policy interacting

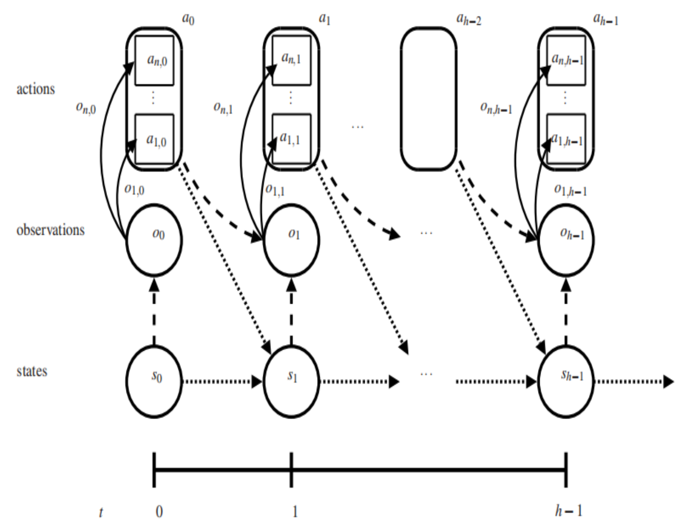
Data comes from unknown behavior-policy

Offline Advantages:

1. Avoid cost of interacting with environment
2. Make use of precious expert demonstration

MULTI-AGENT:dec-POMDP





1. This state emits a joint observation
2. each agent observes its individual component
3. each agent selects an action, together forming the joint action,
4. joint action leads to state transition according to the transition model

Model of Dec-POMDP

$G =< S, A, P, r, \Omega, O, n, \gamma >$

S: set of states

A: set of joint actions

P: state transition function

r: reward function shared by all agents

Ω: set of joint observations

O: observation function

n: agents

γ: discount rate

Extrapolation Error

Definition:

- Extrapolation error is an error in off-policy value learning which is introduced by *the mismatch between the dataset and true state-action visitation of the current policy*.

Cause:

- The extrapolation error mainly attributes the out-of-distribution (OOD) actions in the evaluation of *Qπ(**overestimate the Q value of unknown action**)*

To quantify the effect of OOD actions, we define the state-action pairs within the dataset as seen pairs. Otherwise, we name them as *unseen* pairs.

Extrapolation Error

$$\epsilon_{\mathrm{EXP}}(\tau, a) = \sum_{\tau'} \left(P_M(\tau' \mid \tau, a) - P_{\mathcal{B}}(\tau' \mid \tau, a)\right)\left(r(\tau, a, \tau') + \gamma \sum_{a'} \pi(a' \mid \tau') Q^{\pi}_{\mathcal{B}}(\tau', a')\right).$$

M: true MDP

B: new MDP computed from batch by P_B ($\tau' \mid \tau$, a) = $\mathcal{N}(\tau, a, \tau') / \sum_{\tilde{\tau}} \mathcal{N}(\tau, a, \tilde{\tau})$.

($N(\tau, a, \tau')$ is the number of times the tuple ($s, a, s0$) is observed in$B$)

E-error in ICQ

Define:

$$\epsilon_{\mathbf{MDP}} = [\epsilon_{\mathbf{s}}, \epsilon_{\mathbf{u}}]^{\mathbf{T}}$$

$$\epsilon_{\mathbf{EXT}} = [\mathbf{0}, \epsilon_{\mathbf{b}}]^{\mathrm{T}}$$

$$P_M^\pi = [P_{\mathrm{s,s}}^\pi, P_{\mathrm{s,u}}^\pi; P_{\mathrm{u,s}}^\pi, P_{\mathrm{u,u}}^\pi]$$

$$\begin{bmatrix} \epsilon_{\mathbf{s}} \\ \epsilon_{\mathbf{u}} \end{bmatrix} = \gamma \begin{bmatrix} P_{\mathrm{s,s}}^\pi & P_{\mathrm{s,u}}^\pi \\ P_{\mathrm{u,s}}^\pi & P_{\mathrm{u,u}}^\pi \end{bmatrix} \begin{bmatrix} \epsilon_{\mathbf{s}} \\ \epsilon_{\mathbf{u}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \epsilon_{\mathbf{b}} \end{bmatrix}. \quad (3)$$

Es: seen pair estimate error

Eu: unseen pair estimate error

Theorem 1. *Given a deterministic MDP, the propagation of ϵ_{b} to ϵ_{s} is proportional to $\left\|P_{\mathrm{s,u}}^\pi\right\|_\infty$:*

$$\left\|\epsilon_{\mathrm{s}}\right\|_\infty \leq \frac{\gamma \left\|P_{\mathrm{s,u}}^\pi\right\|_\infty}{(1-\gamma)\left(1-\gamma\left\|P_{\mathrm{s,s}}^\pi\right\|_\infty\right)} \left\|\epsilon_{\mathrm{b}}\right\|_\infty . \quad (4)$$

Implicit Constraint Q-learning

1. Maximization reward
2. Constrain policy to dataset policy

$$\pi_{k+1} = \arg\max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|\tau)}[Q^{\pi_k}(\tau, a)], \quad \text{s.t.} \quad D_{\mathrm{KL}}(\pi \parallel \mu)[\tau] \leq \epsilon. \tag{7}$$

KKT condition

$$\pi^*_{k+1}(a \mid \tau) = \frac{1}{Z(\tau)}\mu(a \mid \tau)\exp\left(\frac{Q^{\pi_k}(\tau, a)}{\alpha}\right), \tag{8}$$

$$\rho(\tau, a) = \frac{\pi^*_{k+1}(a \mid \tau)}{\mu(a \mid \tau)} = \frac{1}{Z(\tau)}\exp\left(\frac{Q^{\pi_k}(\tau, a)}{\alpha}\right). \tag{9}$$

Implicit Constraint Q-learning

Standard policy evaluation(off policy)

$$(\mathcal{T}^{\pi}Q)(\tau,a) \triangleq Q(\tau,a) + \mathbb{E}_{\tau'}[r + \gamma\mathbb{E}_{a'\sim\pi}[Q(\tau',a')] - Q(\tau,a)]. \quad (5)$$

$$(\mathcal{T}^{\pi}Q)(\tau,a) = Q(\tau,a) + \mathbb{E}_{\tau'}[r + \gamma\mathbb{E}_{a'\sim\mu}[\rho(\tau',a')Q(\tau',a')] - Q(\tau,a)], \quad (6)$$

According to Equation 9, it gives the ICQ(re-weight the target value function)

$$\mathcal{T}_{\mathrm{ICQ}}Q(\tau,a) = r + \gamma\mathbb{E}_{a'\sim\mu}\left[\frac{1}{Z(\tau')}\exp\left(\frac{Q(\tau',a')}{\alpha}\right)Q(\tau',a')\right]. \quad (10)$$

$Z(\tau) = \sum_{\tilde{a}} \mu(\tilde{a} \mid \tau)\exp\left(\frac{1}{\alpha}Q^{\pi_k}(\tau,\tilde{a})\right)$ is the normalizing partition function

Thus we obtain a SARSA-like algorithm which not uses any unseen pairs.

Convergence or No?

Theorem 2. *Let* $\mathcal{T}_{\text{ICQ}}^k Q_0$ *denote that the operator* $\mathcal{T}_{\text{ICQ}}$ *are iteratively applied over an initial state-action value function* Q_0 *for* k *times. Then, we have* $\forall(\tau, a)$, $\limsup_{k\to\infty} \mathcal{T}_{\text{ICQ}}^k Q_0(\tau, a) \leq Q^*(\tau, a)$,

$$\liminf_{k\to\infty} \mathcal{T}_{\text{ICQ}}^k Q_0(\tau, a) \geq Q^*(\tau, a) - \frac{\gamma(|A|-1)}{(1-\gamma)} \max\left\{\frac{1}{(\frac{1}{\alpha}+1)C+1}, \frac{2Q_{\max}}{1+C\exp(\frac{1}{\alpha})}\right\}, \quad (13)$$

where $|A|$ *is the action space,* $|A_\tau|$ *is the action space for state* τ, $C \triangleq \inf_{\tau\in S} \inf_{2\leq i\leq |A_\tau|} \frac{\mu(a_{[1]}|\tau)}{\mu(a_{[i]}|\tau)}$ *and* $\mu(a_{[1]} \mid \tau)$ *denotes the probability of choosing the expert action according to behavioral policy* μ. *Moreover, the upper bound of* $\mathcal{T}_{\text{BCQ}}^k Q_0 - \mathcal{T}_{\text{ICQ}}^k Q_0$ *decays exponentially fast in terms of* α.

ICQ操作符从理论上可以证明收敛到一簇稳定解

Implicit Constraint Q-learning

Minimizing:

$$\mathcal{J}_Q(\phi) = \mathbb{E}_{\tau,a,\tau',a'\sim\mathcal{B}}\left[r + \gamma\frac{1}{Z(\tau')}\exp\left(\frac{Q(\tau',a';\phi')}{\alpha}\right)Q(\tau',a';\phi') - Q(\tau,a;\phi)\right]^2, \quad (14)$$

Policy learning(minimizing KL distance):

$$\begin{aligned}
\mathcal{J}_\pi(\theta) &= \mathbb{E}_{\tau\sim\mathcal{B}}\left[D_{\mathrm{KL}}\left(\pi^*_{k+1}\|\pi_\theta\right)[\tau]\right] = \mathbb{E}_{\tau\sim\mathcal{B}}\left[-\sum_a \pi^*_{k+1}(a\mid\tau)\log\frac{\pi_\theta(a\mid\tau)}{\pi^*_{k+1}(a\mid\tau)}\right] \\
&\overset{(a)}{=} \mathbb{E}_{\tau\sim\mathcal{B}}\left[\sum_a \frac{\pi^*_{k+1}(a\mid\tau)}{\mu(a\mid\tau)}\mu(a\mid\tau)\left(-\log\pi_\theta(a\mid\tau)\right)\right] \\
&\overset{(b)}{=} \mathbb{E}_{\tau,a\sim\mathcal{B}}\left[-\frac{1}{Z(\tau)}\log(\pi(a\mid\tau;\theta))\exp\left(\frac{Q(\tau,a)}{\alpha}\right)\right],
\end{aligned}$$

ICQ-MA

Value function decompose:

$$\mathcal{J}_{\boldsymbol{\pi}}(\theta) = \sum_i \mathbb{E}_{\tau^i, a^i \sim \mathcal{B}} \left[-\frac{1}{Z^i(\tau^i)} \log(\pi^i(a^i \mid \tau^i; \theta_i)) \exp\left(\frac{w^i(\boldsymbol{\tau}) Q^i(\tau^i, a^i)}{\alpha} \right) \right]$$

Value function estimate:

$$\mathcal{J}_Q(\phi, \psi) = \mathbb{E}_{\mathcal{B}} \left[\sum_{t \geq 0} (\gamma\lambda)^t \left(r_t + \gamma \frac{1}{Z(\boldsymbol{\tau}_{t+1})} \exp\left(\frac{Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1})}{\alpha} \right) Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}) - Q(\boldsymbol{\tau}_t, \boldsymbol{a}_t) \right) \right] \tag{19}$$

where $Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}) = \sum_i w^i(\boldsymbol{\tau}_{t+1}; \psi') Q^i(\tau^i_{t+1}, a^i_{t+1}; \phi'_i) - b(\boldsymbol{\tau}_{t+1}; \psi')$.

Value estimate with *λ return*:

$$(\mathcal{T}^{\lambda}_{\mathrm{ICQ}} Q)(\boldsymbol{\tau}, \boldsymbol{a}) \triangleq Q(\boldsymbol{\tau}, \boldsymbol{a}) + \mathbb{E}_{\mu} \left[\sum_{t \geq 0} (\gamma\lambda)^t \left(r_t + \gamma \rho(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}) Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}) - Q(\boldsymbol{\tau}_t, \boldsymbol{a}_t) \right) \right], \tag{20}$$

ICQ IN SINGLE AGENT

Algorithm 1: Implicit Constraint Q-Learning in Single-Agent Tasks.

Input: Offline buffer $\mathcal{B}$, target network update rate d.

Initialize critic network $Q^\pi(\cdot;\phi)$ and actor network $\pi(\cdot;\theta)$ with random parameters.
Initialize target networks: $\phi' = \phi, \theta' = \theta$.
for $t = 1$ **to** T **do**
  Sample trajectories from $\mathcal{B}$.
  Train policy according to $\mathcal{J}_\pi(\theta) = \mathbb{E}_{\tau\sim\mathcal{B}}\left[-\frac{1}{Z(\tau)}\log(\pi(a \mid \tau;\theta))\exp\left(\frac{Q^\pi(\tau,a)}{\alpha}\right)\right]$.
  Train critic according to

$$\mathcal{J}_Q(\phi) = \mathbb{E}_{\tau\sim\mathcal{B}}\left[r + \gamma\frac{1}{Z(\tau')}\exp\left(\frac{Q(\tau',a';\phi')}{\alpha}\right)Q(\tau',a';\phi') - Q(\tau,a;\phi)\right]^2.$$

  if $t \bmod d = 0$ **then**
    Update target networks: $\phi' = \phi, \theta' = \theta$.
  end
end

ICQ In multi-agent

Algorithm 2: Implicit Constraint Q-Learning in Multi-Agent Tasks.

Input: Offline buffer $\mathcal{B}$, target network update rate d.

Initialize critic networks $Q^i(\cdot;\phi_i)$, actor networks $\pi^i(\cdot;\theta_i)$ and Mixer network $M(\cdot;\psi)$ with random parameters.

Initialize target networks: $\phi' = \phi$, $\theta' = \theta$, $\psi' = \psi$.

for $t = 1$ **to** T **do**

  Sample trajectories from $\mathcal{B}$.

  Train individual policy according to

  $$\mathcal{J}_{\boldsymbol{\pi}}(\theta) = \sum_i \mathbb{E}_{\tau^i, a^i \sim \mathcal{B}}\left[-\frac{1}{Z^i(\tau^i)} \log(\pi^i(a^i \mid \tau^i; \theta_i)) \exp\left(\frac{w^i(\boldsymbol{\tau})Q^i(\tau^i, a^i)}{\alpha}\right)\right].$$

  Train critic according to $\mathcal{J}_Q(\phi, \psi) =$

  $$\mathbb{E}_{\mathcal{B}}\left[\sum_{t\geq 0}(\gamma\lambda)^t\left[r_t + \gamma\frac{\exp\left(\frac{1}{\alpha}Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}; \phi', \psi')\right)}{Z(\boldsymbol{\tau}_{t+1}; \phi', \psi')}Q(\boldsymbol{\tau}_{t+1}, \boldsymbol{a}_{t+1}; \phi', \psi') - Q(\boldsymbol{\tau}_t, \boldsymbol{a}_t; \phi, \psi)\right]\right]^2.$$

  if $t \bmod d = 0$ **then**

    Update target networks: $\phi' = \phi$, $\theta' = \theta$, $\psi' = \psi$.

  end

end

Experiments

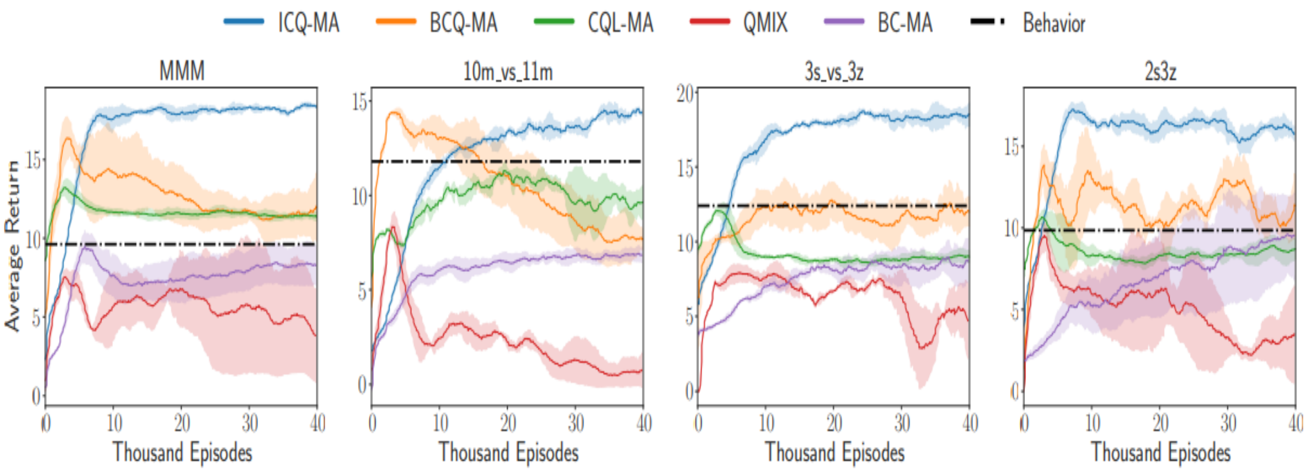


Figure 4: Performance comparison in offline StarCraft II tasks.

Baselines comparison

Dataset type	Environment	ICQ (ours)	BC	BCQ	CQL	AWR	BRAC-p
fixed	antmaze-umaze	**85.0 ± 2.7**	65.0	78.9	74.0	56.0	50.0
play	antmaze-medium	**80.0 ± 1.3**	0.0	0.0	61.2	0.0	0.0
play	antmaze-large	**51.0 ± 4.8**	0.0	6.7	15.8	0.0	0.0
diverse	antmaze-umaze	65.0±3.3	55.0	55.0	84.0	**70.3**	40.0
diverse	antmaze-medium	**65.0 ± 3.9**	0.0	0.0	53.7	0.0	0.0
diverse	antmaze-large	**44.0 ± 4.2**	0.0	2.2	14.9	0.0	0.0
expert	adroit-door	**103.9 ± 3.6**	101.2	99.0	-	102.9	-0.3
expert	adroit-relocate	**109.5 ± 11.1**	101.3	41.6	-	91.5	-0.3
expert	adroit-pen	**123.8 ± 22.1**	85.1	114.9	-	111.0	-3.5
expert	adroit-hammer	**128.3 ± 2.5**	125.6	107.2	-	39.0	0.3
human	adroit-door	6.4±2.4	0.5	-0.0	**9.1**	0.4	-0.3
human	adroit-relocate	**1.5 ± 0.7**	-0.0	-0.1	0.35	-0.0	-0.3
human	adroit-pen	**91.3 ± 10.3**	34.4	68.9	55.8	12.3	8.1
human	adroit-hammer	2.0±0.9	1.5	0.5	**2.1**	1.2	0.3
medium	walker2d	71.8±10.7	66.6	53.1	**79.2**	17.4	77.5
medium	hopper	55.6±5.7	49.0	54.5	**58.0**	35.9	32.7
medium	halfcheetah	42.5±1.3	36.1	40.7	**44.4**	37.4	43.8
med-expert	walker2d	**98.9 ± 5.2**	66.8	57.5	98.7	53.8	76.9
med-expert	hopper	109.0±13.6	**111.9**	110.9	111.0	27.1	1.9
med-expert	halfcheetah	**110.3 ± 1.1**	35.8	64.7	104.8	52.7	44.2