Offline Multi-Agent Reinforcement Learning with Implicit Global-to-Local Value Regularization

Xiangsen Wang[1], Haoran Xu[2], Yinan Zheng[2], Xianyuan Zhan[2]

[1] Beijing Jiaotong University [2] Tsinghua University

Challenges for Offline MARL

Offline Challenges + MARL Challenges:

- Extrapolation error : querying OOD actions can cause extrapolation error accumulation
- Scalability issue: the joint action space grows exponentially as the number of agents increases

It is difficult to incorporate a proper **global-level** offline regularization on the joint action space.

Existing Solutions:

- Value decomposition: **decompose** global value function into local value functions

$$Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) = \sum_i w_i(\boldsymbol{o}) Q_i(o_i, a_i) + b(\boldsymbol{o})$$

$$w_i \geq 0, \ \forall i = 1 \cdots, n$$

- Local offline regularization: apply policy constraints or value regularizations **at the local level**

$$\pi(s) = \arg \max_{a_i + \xi_\phi(s, a_i, \Phi)} Q_\theta(s, a_i + \xi_\phi(s, a_i, \Phi)), \quad \{a_i \sim G_\omega(s)\}_{i=1}^n$$

$$\min_Q \alpha \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} \left[\log \sum_{\mathbf{a}} \exp(Q(\mathbf{s}, \mathbf{a})) - \mathbb{E}_{\mathbf{a} \sim \hat{\pi}_\beta(\mathbf{a}|\mathbf{s})} [Q(\mathbf{s}, \mathbf{a})] \right] + \frac{1}{2} \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim \mathcal{D}} \left[\left(Q - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k \right)^2 \right]$$

Offline MARL Algorithms:

- BCQ-MA
- CQL-MA
- ICQ (NeurIPS 2020)
- OMAR (ICML 2022)
- OMAC (AAMAS 2022)
- …
- …

Problems with Existing Offline MARL Algorithms

☐ Naively combine the value decomposition with local-level offline RL

- Offline regularizations of these methods are completely imposed from the local level without considering the global information.
- Simply enforcing local-level regularization cannot guarantee the induced regularization at the global level still remains valid.
- Existing approaches offer no guarantee whether the optimized local policies are jointly optimal under a given value decomposition scheme.

☐ Offline Multi-Agent Reinforcement Learning with Implicit Global-to-Local Value Regularization (OMIGA):

- Multi-agent POMDP with **global** value regularization
- **Global-to-Local** value and policy decomposition
- Equivalent implicit **local** value regularizations

Organic combination

Provable decomposition

Dec-POMDP with Global Value Regularization

☐ Single-Agent Offline RL:

- Behavior-regularized MDP

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r\left(s_t, a_t\right) - \alpha f\left(\pi\left(a_t|s_t\right), \mu\left(a_t|s_t\right)\right)\right)\right]$$

- Policy evaluation operator

$$\left(\mathcal{T}_f^{\pi}\right) Q(s, a) := r(s, a) + \gamma \mathbb{E}_{s'|s,a}\left[V\left(s'\right)\right],$$

$$V(s) = \mathbb{E}_{a\sim\pi}\left[Q(s, a) - \alpha f\left(\pi\left(a_t|s_t\right), \mu\left(a_t|s_t\right)\right)\right]$$

☐ Multi-Agent Offline RL:

- MA-POMDP with global value regularization

$$\max_{\pi_{tot}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r\left(o_t, a_t\right) - \alpha f\left(\pi_{tot}\left(a_t|o_t\right), \mu_{tot}\left(a_t|o_t\right)\right)\right)\right]$$

- Global policy evaluation operator

$$\left(\mathcal{T}_f^{\pi_{tot}}\right) Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) := r(\boldsymbol{o}, \boldsymbol{a}) + \gamma \mathbb{E}_{\boldsymbol{o}'|\boldsymbol{o},\boldsymbol{a}}\left[V_{tot}\left(\boldsymbol{o}'\right)\right]$$

$$V_{tot}(\boldsymbol{o}) = \mathbb{E}_{\boldsymbol{a}\sim\pi_{tot}}\left[Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) - \alpha \log\left(\frac{\pi_{tot}(\boldsymbol{a}|\boldsymbol{o})}{\mu_{tot}(\boldsymbol{a}|\boldsymbol{o})}\right)\right]$$

the KKT conditions of Lagrangian function

- Establishing relationship among optimal global policy, behavior policy, Q-value function and state-value function

$$\pi_{tot}^*(\boldsymbol{a}|\boldsymbol{o}) = \mu_{tot}(\boldsymbol{a}|\boldsymbol{o}) \cdot \exp\left(\frac{Q_{tot}^*(\boldsymbol{o}, \boldsymbol{a}) - V_{tot}^*(\boldsymbol{o})}{\alpha}\right)$$

Global-to-Local value and policy decomposition

☐ Value decomposition:

$$Q_{tot}(\boldsymbol{o}, \boldsymbol{a}) = \sum_i w_i(\boldsymbol{o}) Q_i(o_i, a_i) + b(\boldsymbol{o})$$

$$V_{tot}(\boldsymbol{o}) = \sum_i w_i(\boldsymbol{o}) V_i(o_i) + b(\boldsymbol{o})$$

$$w_i \geq 0, \ \forall i = 1 \cdots, n$$

☐ Policy decomposition:

$$\begin{aligned} &\pi_{tot}^*(\boldsymbol{a}|\boldsymbol{o}) \\ &= \mu_{tot}(\boldsymbol{a}|\boldsymbol{o}) \cdot exp\left(\frac{\sum_i w_i(\boldsymbol{o})\left(Q_i^*(o_i, a_i) - V_i^*(o_i)\right)}{\alpha}\right) \\ &= \prod_{i=1}^{n} \mu_i(a_i|o_i) \cdot exp\left(\frac{w_i(\boldsymbol{o})}{\alpha}\left(Q_i^*(o_i, a_i) - V_i^*(o_i)\right)\right) \\ &= \prod_{i=1}^{n} \pi_i^*(a_i|o_i) \end{aligned}$$

Local formula:

$$\pi_i^*(a_i|o_i) = \mu_i(a_i|o_i) \cdot exp\left(\frac{w_i(\boldsymbol{o})}{\alpha}\left(Q_i^*(o_i, a_i) - V_i^*(o_i)\right)\right)$$

- The relationship among optimal **global** policy, behavior policy, Q-value function and state-value function
- The relationship among optimal **local** policy, behavior policy, Q-value function and state-value function

Equivalent Implicit Local Value Regularizations

- Self-normalization constraints on local policies:

$$\mathbb{E}_{a_i \sim \mu_i}\left[\exp\left(\frac{1}{\alpha} w_i(\boldsymbol{o})\left(Q_i^*(o_i, a_i) - V_i^*(o_i)\right)\right)\right] = 1$$

- Obtain V by solving the following convex optimization problem:

$$\min_{V_i} \mathbb{E}_{a_i \sim \mu_i}\left[\exp\left(\frac{w_i(\boldsymbol{o})}{\alpha}\left(Q_i(o_i, a_i) - V_i(o_i)\right)\right) + \frac{w_i(\boldsymbol{o}) V_i(o_i)}{\alpha}\right]$$

Hyperparameter α is used to control the degree of regularization.
The higher α encourages the algorithm to stay near the behavioral distribution.
The lower α makes the algorithm more radical and optimistic.

OMIGA Algorithm

☐ Learn the local state-value function:

$$\min_{V_i} \mathbb{E}_{(o_i,a_i)\sim\mathcal{D}}\left[\exp\left(\frac{w_i(\boldsymbol{o})}{\alpha}\left(Q_i(o_i,a_i)-V_i(o_i)\right)\right)+\frac{w_i(\boldsymbol{o})V_i(o_i)}{\alpha}\right]$$

☐ Learn the local Q-value function, the weight, and offset:

$$\min_{\substack{Q_i,w_i,b\\ i=1,\cdots,n}} \mathbb{E}_{(\boldsymbol{o},\boldsymbol{a},\boldsymbol{o}')\sim\mathcal{D}}\left[\left(r(\boldsymbol{o},\boldsymbol{a})+\gamma V_{tot}\left(\boldsymbol{o}'\right)-Q_{tot}(\boldsymbol{o},\boldsymbol{a})\right)^2\right]$$

☐ Learn the local policy:

$$\max_{\pi_i} \mathbb{E}_{(o_i,a_i)\sim\mathcal{D}}\left[\exp\left(\frac{w_i(\boldsymbol{o})}{\alpha}\left(Q_i(o_i,a_i)-V_i(o_i)\right)\right)\cdot\log\pi_i(a_i|o_i)\right]$$

The training process uses **in-sample** learning (without querying OOD action samples).

Algorithm 1 Pseudocode of OMIGA

Require: Offline dataset $\mathcal{D}$. hyperparameter α.

1: Initialize local state-value network V_i, local action-value network Q_i and its target network $\bar{Q}_i$, and policy network π_i for agent i=1, 2, ... n.
2: Initialize the weight function network w and b.
3: **for** $t = 1, \cdots, \textit{max-value-iteration}$ **do**
4: Sample batch transitions $(\boldsymbol{o}, \boldsymbol{a}, r, \boldsymbol{o}')$ from $\mathcal{D}$
5: Update local state-value function $V_i(o_i)$ for each agent i via Eq. (13).
6: Compute $V_{tot}(\boldsymbol{o}')$, $Q_{tot}(\boldsymbol{o}, \boldsymbol{a})$ via Eq. (9).
7: Update local action-value network $Q_i(o_i, a_i)$, weight function network $w(\boldsymbol{o})$ and $b(\boldsymbol{o})$ with objective Eq. (14).
8: Update local policy network π_i for each agent i via Eq. (15).
9: Soft update target network $\bar{Q}_i(o_i, a_i)$ by $Q_i(o_i, a_i)$ for each agent i.
10: **end for**

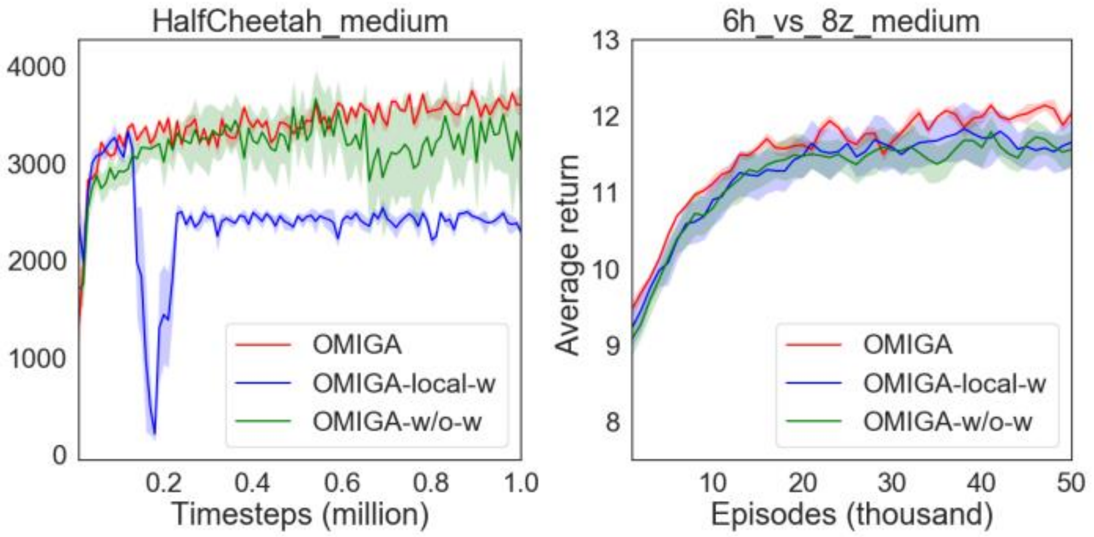
Experimental Results

Multi-agent MuJoCo						
Task	Dataset	BCQ-MA	CQL-MA	ICQ	OMAR	OMIGA(ours)
Hopper	expert	77.85±58.04	159.14± 313.83	754.74± 806.28	2.36± 1.46	**859.63±709.47**
Hopper	medium	44.58±20.62	401.27±199.88	501.79±14.03	21.34±24.90	**1189.26± 544.30**
Hopper	medium-replay	26.53±24.04	31.37±15.16	195.39±103.61	3.30±3.22	**774.18±494.27**
Hopper	medium-expert	54.31±23.66	64.82±123.31	355.44±373.86	1.44±0.86	**709.00±595.66**
Ant	expert	1317.73±286.28	1042.39±2021.65	2050.00±11.86	312.54±297.48	**2055.46±1.58**
Ant	medium	1059.60±91.22	533.90±1766.42	1412.41±10.93	-1710.04±1588.98	**1418.44±5.36**
Ant	medium-replay	950.77±48.76	234.62±1618.28	1016.68±53.51	-2014.20±844.68	**1105.13±88.87**
Ant	medium-expert	1020.89±242.74	800.22±1621.52	1590.18±85.61	-2992.80± 6.95	**1720.33±110.63**
HalfCheetah	expert	2992.71±629.65	1189.54±1034.49	2955.94±459.19	-206.73±161.12	**3383.61±552.67**
HalfCheetah	medium	2590.47±1110.35	1011.35±1016.94	2549.27±96.34	-265.68±146.98	**3608.13±237.37**
HalfCheetah	medium-replay	-333.64±152.06	1998.67±693.92	1922.42±612.87	-235.42±154.89	**2504.70±83.47**
HalfCheetah	medium-expert	**3543.70±780.89**	1194.23±1081.06	2833.99±420.32	-253.84± 63.94	2948.46± 518.89
SMAC						
Task	Dataset	BCQ-MA	CQL-MA	ICQ	OMAR	OMIGA(ours)
5m_vs_6m	good	7.76±0.15	8.08±0.21	7.87±0.30	7.40±0.63	**8.25±0.37**
5m_vs_6m	medium	7.58±0.10	7.78±0.10	7.77±0.3	7.08±0.51	**7.92±0.57**
5m_vs_6m	poor	**7.61±0.36**	7.43±0.10	7.26±0.19	7.27±0.42	7.52±0.21
2c_vs_64zg	good	19.13±0.27	18.48±0.95	18.82±0.17	17.27±0.78	**19.15±0.32**
2c_vs_64zg	medium	15.58±0.37	12.82±1.61	15.57±0.61	10.20±0.20	**16.03±0.19**
2c_vs_64zg	poor	12.46±0.18	10.83±0.51	12.56±0.18	11.33±0.50	**13.02±0.66**
6h_vs_8z	good	12.19±0.23	10.44±0.20	11.81±0.12	9.85±0.28	**12.54±0.21**
6h_vs_8z	medium	11.77±0.16	11.29±0.29	11.13±0.33	10.36±0.16	**12.19±0.22**
6h_vs_8z	poor	10.84±0.16	10.81±0.52	10.55±0.10	10.63±0.25	**11.31±0.19**
corridor	good	15.24±1.21	5.22±0.81	15.54±1.12	6.74±0.69	**15.88±0.89**
corridor	medium	10.82±0.92	7.04±0.66	11.30±1.57	7.26±0.71	**11.66±1.30**
corridor	poor	4.47±0.94	4.08±0.60	4.47±0.33	4.28±0.49	**5.61±0.35**

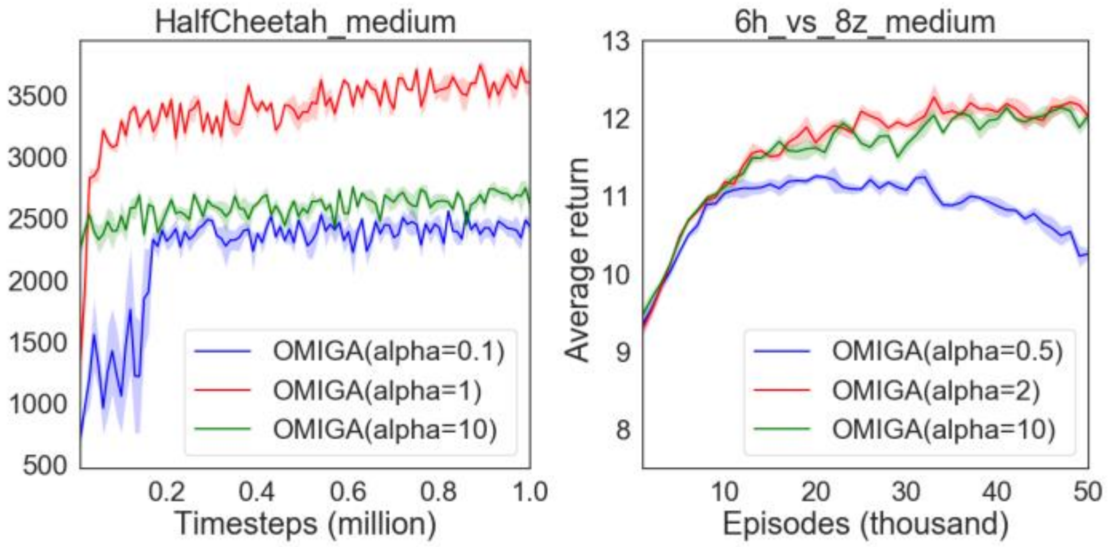
Strong results on Multi-agent MuJoCo and SMAC benchmark datasets

Comparative Evaluation

Analyses on Policy Learning with Global Information



Analyses on the Regularization Hyperparameter



Summary

- We present a new offline multi-agent RL algorithm with implicit global-to-local value regularization (OMIGA), which provides a principled framework to convert global-level value regularization into equivalent implicit local value regularizations.

- OMIGA bridges multi-agent value decomposition and policy learning with offline regularizations, which can guarantee that the learned local policies are jointly optimal at the global level.

More details are available on https://arxiv.org/abs/2307.11620

Thanks