ComaDICE: Offline Cooperative Multi-Agent Reinforcement Learning with Stationary Distribution Shift Regularization

- **The Viet Bui**, *Singapore Management University*, Singapore
- **Tien Anh Mai**, *Singapore Management University*, Singapore
- **Thanh Hong Nguyen**, *University of Oregon*, USA

ICLR
International Conference On Learning Representations

Introduction

- **Offline MARL Challenge:** The fundamental difficulty is learning from a fixed dataset $\mathcal{D}$ collected by some *behavior policy* μ_{tot}. When we try to optimize a new *learning policy* π_{tot} , it might explore state-action pairs (s, a) that are rare or absent in $\mathcal{D}$. Standard RL value estimation (like Q-learning) struggles here, often overestimating values for these out-of-distribution (OOD) pairs, leading to poor performance. This is the *distributional shift* problem.
- **Our Approach:** Instead of just penalizing OOD *actions* (like many prior methods), ComaDICE aims to align the overall *state-action visitation frequency* of the learned policy with the behavior policy. This frequency is captured by the **stationary distribution** $\rho^{\pi_{tot}}(s, a)$.

Preliminaries

Model: Cooperative MARL as a Partially Observable Markov Decision Process (POMDP):

$$M = \langle S, A, P, r, Z, O, n, N, \gamma \rangle$$

Goal: Maximize expected joint return $E[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$

Offline Dataset: $\mathcal{D}$ collected by behavior policy μ_{tot}

Stationary Distribution (Occupancy Measure): Probability of visiting state-action (s, a) under policy π_{tot}: $\rho^{\pi_{tot}}(s, a) = (1-\gamma) \sum_{t=0}^{\infty} P(s_t = s, a_t = a)$

Core Idea

Optimize expected return regularized by the f-divergence between learning (π_{tot}) and behavior (μ_{tot}) stationary distributions:

$$\max_{\pi_{tot}} \underbrace{E_{(s,a)\sim\rho^{\pi_{tot}}}[r(s,a)]}_{\text{Maximize Expected Return}} - \underbrace{\alpha D^f(\rho^{\pi_{tot}} || \rho^{\mu_{tot}})}_{\text{Regularize Distribution Shift}}$$

$D^f(\rho^{\pi_{tot}} || \rho^{\mu_{tot}}) = E_{(s,a)\sim\rho^{\mu_{tot}}}\left[f\left(\frac{\rho^{\pi_{tot}}(s,a)}{\rho^{\mu_{tot}}(s,a)}\right)\right]$ is the f-divergence (f is convex),

α balances reward maximization and distribution matching

Mathematical Formulation & Derivations

Closed-Form Solution: Inner max over w^{tot} has a solution, simplifying to minimization over ν^{tot} only:

$$\min_{\nu^{tot}} \tilde{\mathcal{L}}(\nu^{tot}) = (1-\gamma) E_{s_0}[\nu^{tot}(s_0)] + E_{\rho^{\mu_{tot}}}\left[-\alpha f^*\left(\frac{A_\nu^{tot}(s,a)}{\alpha}\right)\right]$$

f^* is the convex conjugate of f

Optimal $w^{tot*}(s,a) = \max\{0, (f')^{-1}(A_\nu^{tot}(s,a)/\alpha)\}$

Practical Algorithm & Losses

- Local value nets $\nu_i(\psi_\nu)$
- Q-nets $q_i(\psi_q)$
- Policy nets $\pi_i(\eta_i)$
- Mixing nets $\mathcal{M}_\theta$

MSE Loss for Q-function:

$$\mathcal{L}_q(\psi_q) = E_{\mathcal{D}}\left[\left(\mathcal{M}_\theta[q-\nu] - (r + \gamma \mathcal{M}_\theta[\nu'] - \mathcal{M}_\theta[\nu])\right)^2\right]$$

Value Function Loss: Sample-based version of $\tilde{\mathcal{L}}$

$$\tilde{\mathcal{L}}(\psi_\nu, \theta) = (1-\gamma) E_{s_0}\left[\mathcal{M}_\theta[\nu_{s_0}]\right] + E_{(s,a)}\left[\alpha f^*\left(\frac{\mathcal{M}_\theta[q-\nu]}{\alpha}\right)\right]$$

Policy Loss:

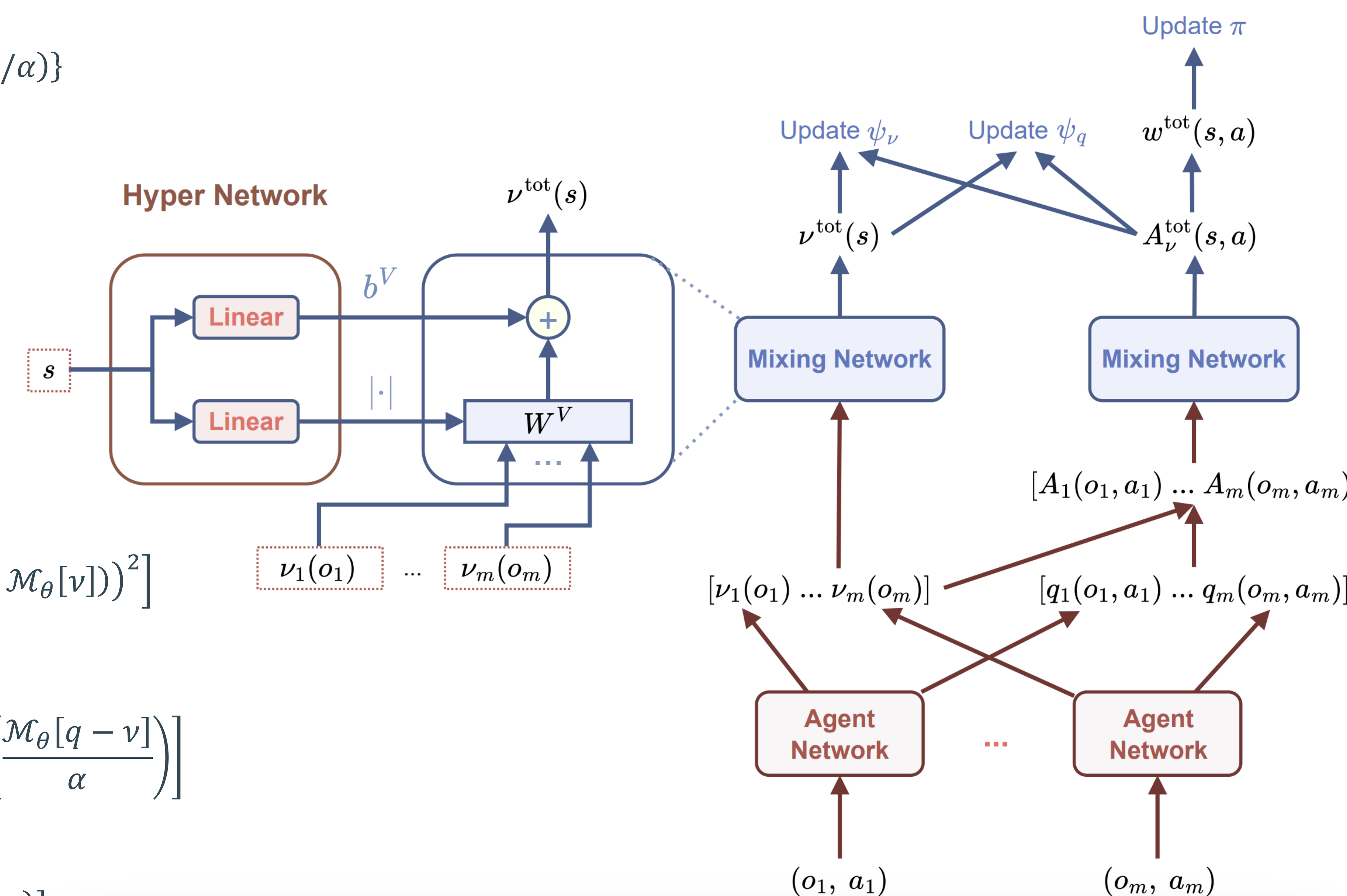
$$\mathcal{L}_\pi(\eta_i) = E_{\mathcal{D}}[w^{tot*}(s,a) \log \pi_i(a_i | s_i; \eta_i)]$$

Value Factorization for MARL (CTDE)

Decomposition: Using local functions (ν_i, q_i) and a mixing network $\mathcal{M}_\theta$

$$\nu^{tot}(s) = \mathcal{M}_\theta[\nu(s)],\ A_\nu^{tot}(s,a) = \mathcal{M}_\theta[q(s,a) - \nu(s)]$$

Convexity: The objective $\tilde{\mathcal{L}}(\nu, \theta)$ is convex in ν if $\mathcal{M}_\theta$ has non-negative weights and convex activations (e.g., linear, ReLU).



Experiments & Results

Instances		BCQ	CQL	ICQ	OMIGA	OptDICE	AlberDICE	ComaDICE (ours)
Hopper	expert	77.9 ± 58.0	159.1 ± 313.8	754.7 ± 806.3	859.6 ± 709.5	655.9 ± 120.1	844.6 ± 556.5	**2827.7 ± 62.9**
	medium	44.6 ± 20.6	401.3 ± 199.9	501.8 ± 14.0	**1189.3 ± 544.3**	204.1 ± 41.9	216.9 ± 35.3	822.6 ± 66.2
	m-replay	26.5 ± 24.0	31.4 ± 15.2	195.4 ± 103.6	774.2 ± 494.3	257.8 ± 55.3	419.2 ± 243.5	**906.3 ± 242.1**
	m-expert	54.3 ± 23.7	64.8 ± 123.3	355.4 ± 373.9	709.0 ± 595.7	400.9 ± 132.5	515.1 ± 303.4	**1362.4 ± 522.9**
Ant	expert	1317.7 ± 286.3	1042.4 ± 2021.6	2050.0 ± 11.9	2055.5 ± 1.6	1717.2 ± 27.0	1896.8 ± 33.7	**2056.9 ± 5.9**
	medium	1059.6 ± 91.2	533.9 ± 1766.4	1412.4 ± 10.9	1418.4 ± 5.4	1199.0 ± 26.8	1304.3 ± 2.6	**1425.0 ± 2.9**
	m-replay	950.8 ± 48.8	234.6 ± 1618.3	1016.7 ± 53.5	1105.1 ± 88.9	869.4 ± 62.6	1042.8 ± 80.8	**1122.9 ± 61.0**
	m-expert	1020.9 ± 242.7	800.2 ± 1621.5	1590.2 ± 85.6	1720.3 ± 110.6	1293.2 ± 183.1	1780.0 ± 23.6	**1813.9 ± 68.4**
Half Cheetah	expert	2992.7 ± 629.7	1189.5 ± 1034.5	2955.9 ± 459.2	3383.6 ± 552.7	2601.6 ± 461.9	3356.4 ± 546.9	**4082.9 ± 45.7**
	medium	2590.5 ± 1110.4	1011.3 ± 1016.9	2549.3 ± 96.3	**3608.1 ± 237.4**	305.3 ± 946.8	522.4 ± 315.5	2664.7 ± 54.2
	m-replay	-333.6 ± 152.1	1998.7 ± 693.9	1922.4 ± 612.9	2504.7 ± 83.5	-912.9 ± 1363.9	440.0 ± 528.0	**2855.0 ± 242.2**
	m-expert	3543.7 ± 780.9	1194.2 ± 1081.0	2834.0 ± 420.3	2948.5 ± 518.9	-2485.8 ± 2338.4	2288.2 ± 759.5	**3889.7 ± 81.6**