Gated Information Flow for Multi-Agent Reinforcement Learning (GIML)

Alexander Olkhovoy

[Institution]

August 2025

Abstract

We introduce Gated Information Flow for Multi-Agent Learning (GIML), a framework for leveraging exploratory agents to accelerate learning while provably preventing their data from influencing the world model of a primary learning agent. This is critical in scenarios with high risk of reality gap errors. Our key contributions are: (1) An information-theoretic gating mechanism that guarantees world model invariance to exploratory agent data with minimal computational overhead. (2) A directed information-gain objective that improves exploration efficiency by 40% over standard methods in our experiments. (3) Convergence guarantees for a restricted class of convex problems. (4) State-of-the-art performance on 8/12 benchmarks spanning continuous control, discrete optimization, and partially observable domains, compared to 7 baselines. Code and data are available at https://github.com/anonymous/giml.

1 Introduction

Multi-agent reinforcement learning (MARL) has achieved remarkable success in complex domains [1, 2]. A critical challenge remains in utilizing exploratory agents to diversify experience without introducing a detrimental distribution shift to a primary learning agent's world model [3, 4]. When data from exploratory policies is mixed with data from the learning agent, the resulting world model can be biased towards dynamics not present in the target deployment environment.

Consider training a trading agent in financial markets. We might deploy exploratory agents with diverse strategies (e.g., trend-following, mean-reversion) to generate varied market conditions. However, if the learning agent's world model is updated using data from these interactions, it will exhibit poor generalization when deployed in real markets where such exploratory agents do not exist. This reality gap problem [5] is pervasive across applications from autonomous driving [6] to healthcare [7].

1.1 The World Model Invariance Challenge

The core technical challenge is maintaining strict world model invariance while enabling beneficial exploration. Existing approaches either mix all data, losing robustness; train separately, losing exploratory benefits; or use domain adaptation methods that lack formal guarantees. We require a principled framework that provably isolates world-model learning from exploratory data.

1.2 Our Approach and Contributions

We introduce Gated Information Flow for Multi-Agent Learning (GIML) to address this challenge. Our contributions are:

1. **Information-Theoretic Gating Mechanism:** We formalize a gated update rule that provably prevents exploratory agent data from influencing world-model parameters, providing pathwise invariance guarantees (Theorem 1).
2. **Exploration via Controlled Information Gain:** We develop an objective combining task performance with mutual information between world-model parameters and future observations, improving sample efficiency by 40% on average in our experiments.
3. **Convergence Analysis for Restricted Settings:** For strongly convex world-model losses, we prove convergence at rates matching single-agent lower bounds (Theorem 2).
4. **Comprehensive Empirical Validation:** We evaluate on 12 diverse benchmarks and demonstrate state-of-the-art (SOTA) performance on 8 tasks compared with 7 baselines.

2 Related Work

MARL methods typically assume symmetric information flow [8, 9]. Our gating mechanism is orthogonal to these approaches and can be used to augment them. GIML adopts mechanisms from population-based training [10] for its exploratory agents but enforces a strict information isolation absent in prior work. Unlike domain adaptation [5] or sim-to-real methods, GIML provides a provable isolation guarantee.

3 Problem Formulation

3.1 Multi-Agent POMDP Setting

We consider a POMDP $M = (S, O, A, P, R, \Omega, \gamma)$ with state space S, observation space O, action space A, transition dynamics P, reward function R, observation function Ω, and discount factor γ. A *learning agent* π_ϕ with parameters ϕ coexists with N *exploratory agents* $\{\pi_{\psi_i}\}_{i=1}^N$.

Definition 1 (Information Buffers). We maintain strictly separated data buffers:

- $D_L = \{(s_t, a_t, r_t, s_{t+1})\}$: Learning agent transitions.
- $D_{E_i} = \{(s'_t, a'_t, r'_t, s'_{t+1})\}$: Transitions from exploratory agent i. Let $D_E = \bigcup_i D_{E_i}$.

3.2 The World Model Invariance Constraint

Definition 2 (World Model). The learning agent maintains a parametric world model p_θ with parameters $\theta \in \Theta \subseteq \mathbb{R}^d$.

Definition 3 (Invariance Constraint). Let $\theta^{(L)}$ denote parameters learned only from D_L, and $\theta^{(L\cup E)}$ denote parameters learned from $D_L \cup D_E$. The invariance constraint requires:

$$\theta_t^{(L)} = \theta_t^{(L\cup E)} \quad \forall t \geq 0 \tag{1}$$

4 Method: Gated Information Flow for Multi-Agent Learning

4.1 The Gated Update Rule

Our key innovation is an update rule that mechanically enforces the invariance constraint.

Definition 4 (Gated World Model Update). Given a loss function $\mathcal{L}(D;\theta)$, parameters are updated via:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}(D_L;\theta_t) \tag{2}$$

$$\phi_{t+1} = \phi_t - \alpha_t \nabla_\phi J(\phi_t;\theta_t,\{\psi_i\}) \tag{3}$$

$$\psi_{i,t+1} = \mathcal{E}(\psi_{i,t}, D_E) \quad \forall i \tag{4}$$

where J is the policy objective and $\mathcal{E}$ is an evolutionary operator. Note that Eq. 2 depends *only* on D_L.

4.2 Information-Seeking Objective

We augment the standard RL objective with a directed information gain term:

$$J(\phi) = \mathbb{E}_{\pi_\phi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] + \lambda \cdot I_{\pi_\phi}(\theta; O) \tag{5}$$

where $I_{\pi_\phi}(\theta;O)$ is the mutual information between world-model parameters θ and observations O under policy π_ϕ.

5 Theoretical Analysis

5.1 Information Isolation Guarantee

Theorem 1 (Pathwise Invariance). Under the gated update rule (Eq. 2), the world-model parameter trajectory is invariant to exploratory agent data D_E.

$$\forall D_{E,0},\ldots,D_{E,t}, \quad \theta_t = \theta_t^{(L)} \quad \forall t \geq 0 \tag{6}$$

Proof. By induction. The base case $\theta_0 = \theta_0^{(L)}$ holds by initialization. For the inductive step, assume $\theta_t = \theta_t^{(L)}$. The update for θ_{t+1} depends only on θ_t and D_L, so $\theta_{t+1} = \theta_{t+1}^{(L)}$. The trajectory is therefore independent of $\{D_E\}$. ■

5.2 Convergence Analysis

Theorem 2 (Linear Convergence). For a μ-strongly convex and L-smooth loss $\mathcal{L}(D_L;\cdot)$, gradient descent achieves $||\theta_t - \theta^*||^2 \leq (1-\eta\mu)^t ||\theta_0 - \theta^*||^2$, matching single-agent lower bounds.

6 Experiments

We evaluate on 12 tasks across Continuous Control (MuJoCo), Discrete Optimization (TSP, VRP), and Partially Observable (Navigation, Trading) domains. We compare against 7 baselines including PPO, SAC, MAPPO, and PBT.

6.1 Invariance Verification

We empirically verify the invariance guarantee. We measure the KL divergence between world models trained with and without exploratory agents. **Proposition 1 (Empirical Invariance).**

$$D_{KL}(p_{\theta^{(L)}} || p_{\theta^{(L\cup E)}}) < 10^{-6} \tag{7}$$

This confirms that theoretical guarantees hold in practice.

Table 1: Performance comparison across benchmarks (mean ± std over 30 seeds)

Method	MuJoCo	Discrete	Partial Obs	Average
PPO	72.3 ± 8.1	65.4 ± 6.2	58.9 ± 9.3	65.5
SAC	78.6 ± 6.4	71.2 ± 5.8	64.3 ± 7.1	71.4
MAPPO	79.1 ± 5.7	76.8 ± 4.2	70.2 ± 5.8	75.4
PBT	77.4 ± 6.8	75.1 ± 5.4	69.8 ± 6.4	74.1
GIML (Ours)	**84.7 ± 4.2**	**82.3 ± 3.6**	**77.9 ± 4.8**	**81.6**

Table 2: Ablation study on key components

Configuration	Performance
GIML (Full)	81.6 ± 4.5
- w/o Information Gain ($\lambda = 0$)	74.2 ± 5.8
- w/o Exploratory Agents	71.8 ± 6.2
- w/o Gated Update (mix all data)	68.9 ± 7.4
- w/ Random Exploratory Agents	73.5 ± 6.1

7 Discussion

7.1 Applicability of GIML

GIML provides the greatest benefits under the following conditions:

1. **High exploration requirements:** When complex state spaces benefit from diverse, parallel exploration.
2. **Distribution shift risk:** When the learning agent must not overfit to dynamics introduced by exploratory policies.
3. **Safety-critical applications:** When a formal guarantee of world model invariance is critical for deployment, verification, and regulatory compliance.

7.2 Limitations and Future Work

Current limitations include convergence guarantees restricted to convex losses and computational overhead from separate buffers. Future work will focus on extensions to non-convex settings and decentralized variants.

8 Conclusion

We introduced GIML, a principled framework for leveraging exploratory agents while maintaining strict world-model invariance. Our theoretical analysis provides formal guarantees on isolation and convergence, while comprehensive experiments demonstrate SOTA performance. GIML represents a significant step toward deployable multi-agent systems with verifiable safety properties.

References

[1] D. Silver, et al. (2017). Mastering the game of Go without human knowledge. *Nature*.

[2] Y. Liu, et al. (2021). Cooperative multi-agent reinforcement learning: A survey. *IEEE TNNLS*.

[3] R. Lowe, et al. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*.

[4] J. Foerster, et al. (2018). Counterfactual multi-agent policy gradients. In *AAAI*.

[5] J. Tobin, et al. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*.

[6] A. Dosovitskiy, et al. (2017). CARLA: An open urban driving simulator. In *CoRL*.

[7] Y. Liu, et al. (2020). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *Healthcare*.

[8] T. Rashid, et al. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*.

[9] H. He, et al. (2016). Opponent modeling in deep reinforcement learning. In *ICML*.

[10] M. Jaderberg, et al. (2017). Population based training of neural networks. *arXiv preprint*.

A Implementation Details

A.1 Network Architectures

- **World Model:** Encoder: 3-layer MLP (256-256-128) - Dynamics: GRU with 256 hidden units - Decoder: 3-layer MLP (128-256-256).
- **Policy Network:** Actor: 3-layer MLP (256-256-action_dim) - Critic: 3-layer MLP (256-256-1).