# HYPOLE: Hyperproperty-Guided Multi-Agent Reinforcement Learning under Partial Observation

Arshia Rafieioskouei , Tzu-Han Hsu , Matthew Lucas , Borzoo Bonakdarpour

ICML International Conference On Machine Learning


**Under the veil of uncertainty, HYPOLE guides agents to learn effective tactics.**

## Research Question

**Partial Observability** | **Multi-Agent** | **Tactics** → **Formal Specification**

How can **decentralized policies** be learned from formal specifications in partially observable multi-agent settings?

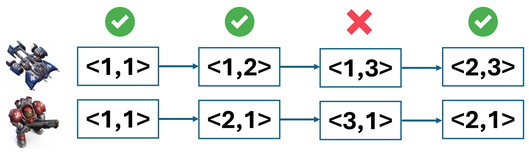
## Formal Specification vs. Shaped Rewards

- **Mathematical rigor**
- **Expressiveness** in specifying objectives and constraints
- The ability to define **tactics** for achieving objectives

## Specifications as Hyperproperties

- **Hyperproperties** characterize requirements over sets of execution traces, enabling the specification of behaviors in multi-agent systems.

*"Is the distance between Marine and Medivac always less than 3?"*




- We encode tactics as **HyperLTL** formulas, enabling specification-guided MARL under partial observability.

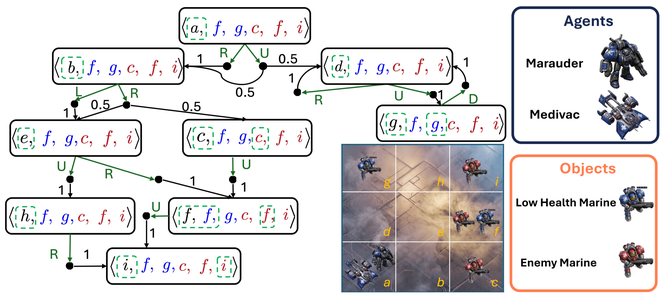
## Problem Statement

Given a POMDP $\mathcal{M}$ and a HyperLTL formula $\varphi = \mathbb{Q}_1\tau_1 \ldots \mathbb{Q}_n\tau_n.\ \psi$, our goal is to identify a tuple of $n$ policies $\langle \pi_1^\star, \ldots, \pi_n^\star \rangle$, such that:

$$\langle \pi_i^\star \rangle_{i\in\{1,\ldots,n\}} \in \left[ \arg\max_{\langle \pi_i \rangle} \mathbb{P}\Big[ \langle \mathsf{Traces}\big( \bigcup_{h\in(\mathcal{H}_i\sim\mathcal{D}_{\pi_i})} \{\arg\max_{\zeta\in\mathcal{Z}^*} \Phi(h,\zeta)\}\big)\rangle \models \varphi \Big] \right]_{i\in\{1,\ldots,n\}}$$

where $\mathcal{D}_{\pi_1}, \ldots, \mathcal{D}_{\pi_n}$ are the distributions over a set of histories drawn by policies $\pi_1, \ldots, \pi_n$, and $\bigcup_{h\in(\mathcal{H}_i\sim\mathcal{D}_{\pi_i})}\{\arg\max_{\zeta\in\mathcal{Z}^*}\Phi(h,\zeta)\}$ is a set of paths $\mathcal{Z}_i$ associated with policy $\pi_i$.

## Illustrative Example




HyperLTL Formula: $\forall\tau_1\exists\tau_2.\ \mathrm{Dist}(\mathtt{pos}_{\tau_1}, \mathtt{pos}_{\tau_2}) < 3\ \mathcal{U}\ (\mathtt{pos}_{\tau_1} = \mathtt{i})$

Samples $\mathcal{Z}_{\tau_1}$ :

Samples $\mathcal{Z}_{\tau_2}$ :

$h_2^1 = \langle a\rangle \xrightarrow{U} \langle d\rangle \xrightarrow{U} \langle g,g\rangle \xrightarrow{D} \langle d\rangle \xrightarrow{R} \langle d\rangle$

$h_2^2 = \langle a\rangle \xrightarrow{U} \langle b\rangle \xrightarrow{L} \langle e\rangle \xrightarrow{R} \langle f,f,f\rangle \xrightarrow{U} \langle i,i\rangle$

Computing the satisfaction probability of HyperLTL specifications:

$\mathsf{Traces}(\{\arg\max_{\zeta\in\mathcal{Z}^*}\Phi(h_1^1,\zeta)\},\mathcal{Z}_{\tau_2}) \models \varphi_{\exp}$
$\mathsf{Traces}(\{\arg\max_{\zeta\in\mathcal{Z}^*}\Phi(h_1^2,\zeta)\},\mathcal{Z}_{\tau_2}) \models \varphi_{\exp}$
$\Big\}$ $h_2^2$ serves as a witness for $\tau_2$ $\Longrightarrow$ $\mathbb{P} = 1$

*Note that changing the formula to $\forall\forall\psi$ reduces the satisfaction probability to 0.75, showing that the $\forall\exists\psi$ formula expands the policy search space.*

## Our Solution

**Step 1:** We apply **Skolemization** to resolve **quantifier alternations** in a HyperLTL formula.

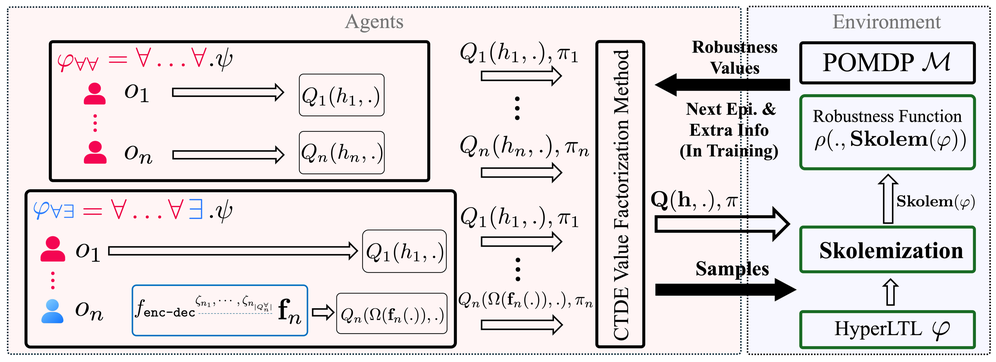
$$\mathbf{Skolem}(\varphi) = \underbrace{\exists \mathbf{f}_i(\tau_{i_1}, \ldots, \tau_{i_{|\mathbb{Q}_i^\forall|}})}_{\text{for each } i\in\mathbb{Q}^\exists}. \quad \underbrace{\forall\tau_j.}_{\text{for each } j\in\mathbb{Q}^\forall} \quad \mathbf{Skolem}(\psi)$$

**Step 2:** We use **quantitative semantics** to construct robustness functions that transform skolemized HyperLTL formula into real-valued rewards (real-valued robustness values).

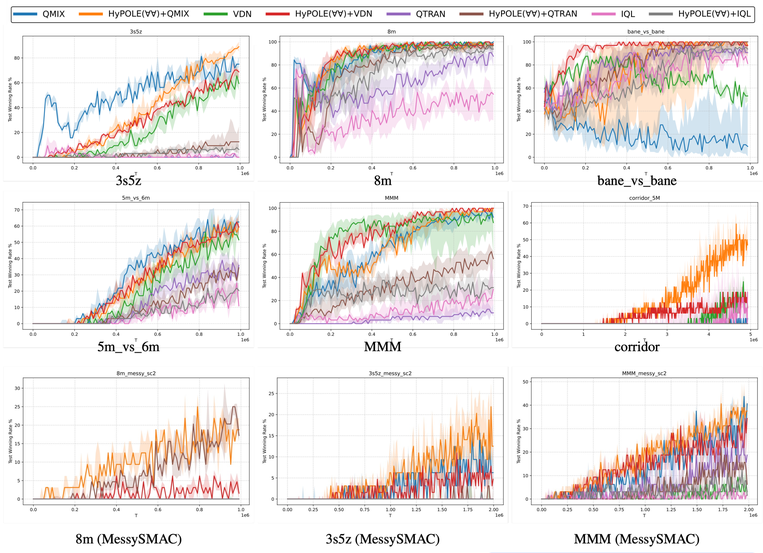
**Step 3:** We use **CTDE algorithms** (VDN/QMIX/QTRAN) to learn decentralized policies:

1. We lift a shared POMDP environment to a Dec-POMDP setting compatible with CTDE methods.
2. We learn optimal policies for agents with universal quantifiers. For each $j \in \mathbb{Q}^\forall$, $\pi_j^\star(h_{j[0:k]})$.
3. Since existentially quantified agents require the traces of preceding universally quantified agents as input to their Skolem functions, we train an `encoder-decoder` model on replay-buffer samples during training and use it to construct policies for each $i \in \mathbb{Q}^\exists$, $\pi_i^\star(\Omega(\mathbf{f}_i(\mathsf{Tr}(\zeta_{i_1[0:k]}), \ldots, \mathsf{Tr}(\zeta_{i_{|\mathbb{Q}_i^\forall|}[0:k]}))))$.
4. CTDE algorithms aim to approximate the joint value function $\mathbf{Q}^*(\mathbf{h},\mathbf{a}) = \max_{\boldsymbol{\pi}} \mathbf{Q}^{\boldsymbol{\pi}}(\mathbf{h},\mathbf{a})$. If the CTDE algorithm satisfies the IGM property and learns the optimal joint value function, then the induced decentralized policies $\langle \pi_i \rangle_{i\in\mathbb{Q}^\exists}$ and $\langle \pi_j \rangle_{j\in\mathbb{Q}^\forall}$ are optimal as well.

## HYPOLE Overview

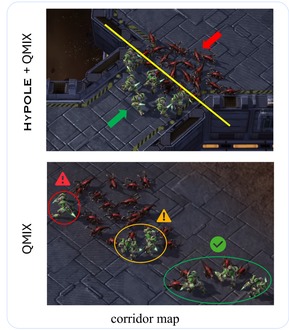



## HYPOLE vs. Shaped Rewards



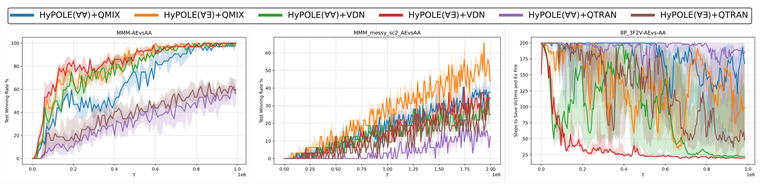

8m (MessySMAC) | 3s5z (MessySMAC) | MMM (MessySMAC)

**Take Away:**
Our experiments show that encoding effective tactics as HyperLTL formulas improves MARL baselines over shaped rewards. **HYPOLE**+IQL outperforms QTRAN on 8m, while **HYPOLE**+QMIX and **HYPOLE**+VDN achieve up to 90% and 55% gains in win rate on the challenging bane_vs_bane map. On corridor, **HYPOLE**+QMIX achieves up to a 60% gain by encoding the defensive tactic of holding the choke point in HyperLTL. Even in the noisier and more stochastic MessySMAC environment, **HYPOLE** improves MARL baselines by up to 25% on 8m, where the original baselines struggle to achieve nonzero win rates. Notably, **HYPOLE**+VDN achieves up to a 30% gain.




corridor map

## ∀∀ vs. ∀∃ Specifications

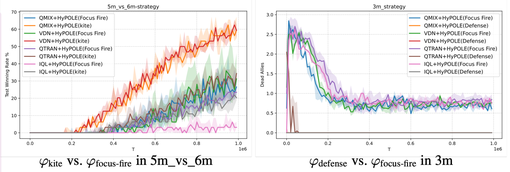



MMM (SMAC) | MMM (MessySMAC) | 8P_3F2V (WildFire)

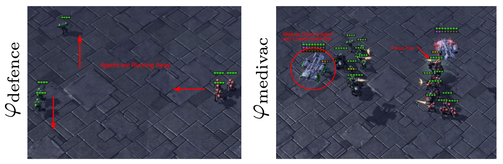
**Take Away:**
Changing the HyperLTL quantifier structure from ∀∀ to ∀∃ improves MARL when agents have strong interdependencies. In MMM, assigning an existential quantifier to the medivac agent yields up to 40% win-rate gains for **HYPOLE**( ∀∃ )+QMIX in SMAC and up to 20% in MessySMAC compared to **HYPOLE**(∀∀ )with similar gains observed for QTRAN.

## Tactics as HyperLTL Formulas

**Take Away:**
**HYPOLE** can encode different tactics through HyperLTL specifications. In 5m_vs_6m, the kiting specification consistently outperforms the focus-fire specification, indicating that kiting is the more effective tactic for this scenario. In 3m, a defensive specification encourages agents to retreat from enemies and reduce casualties.




$\varphi_{\text{kite}}$ vs. $\varphi_{\text{focus-fire}}$ in 5m_vs_6m | $\varphi_{\text{defense}}$ vs. $\varphi_{\text{focus-fire}}$ in 3m




**Caution**
**HYPOLE** is sensitive to the design of the HyperLTL formula. When we remove the subformula that encourages agents to shoot enemies from the focus-fire specification, the win rate drops to zero, highlighting the importance of encoding effective tactics.