Sample Efficient Distributionally Robust Multi-Agent Reinforcement Learning via Online Interaction

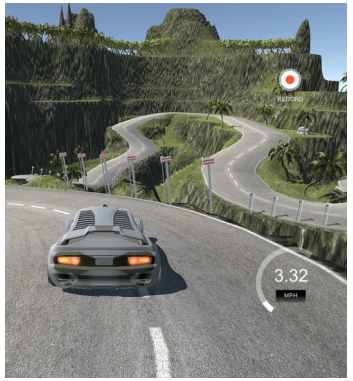
Zain Ulabedeen Farhat, Debamita Ghosh, George K. Atia, Yue Wang

Department of Electrical & Computer Engineering

University of Central Florida

CONTENTS

Motivation: Sim-to-Real Gap



Multiple Agents trained in **simulation** are not **robust** to slight perturbations when deployed in the real world.

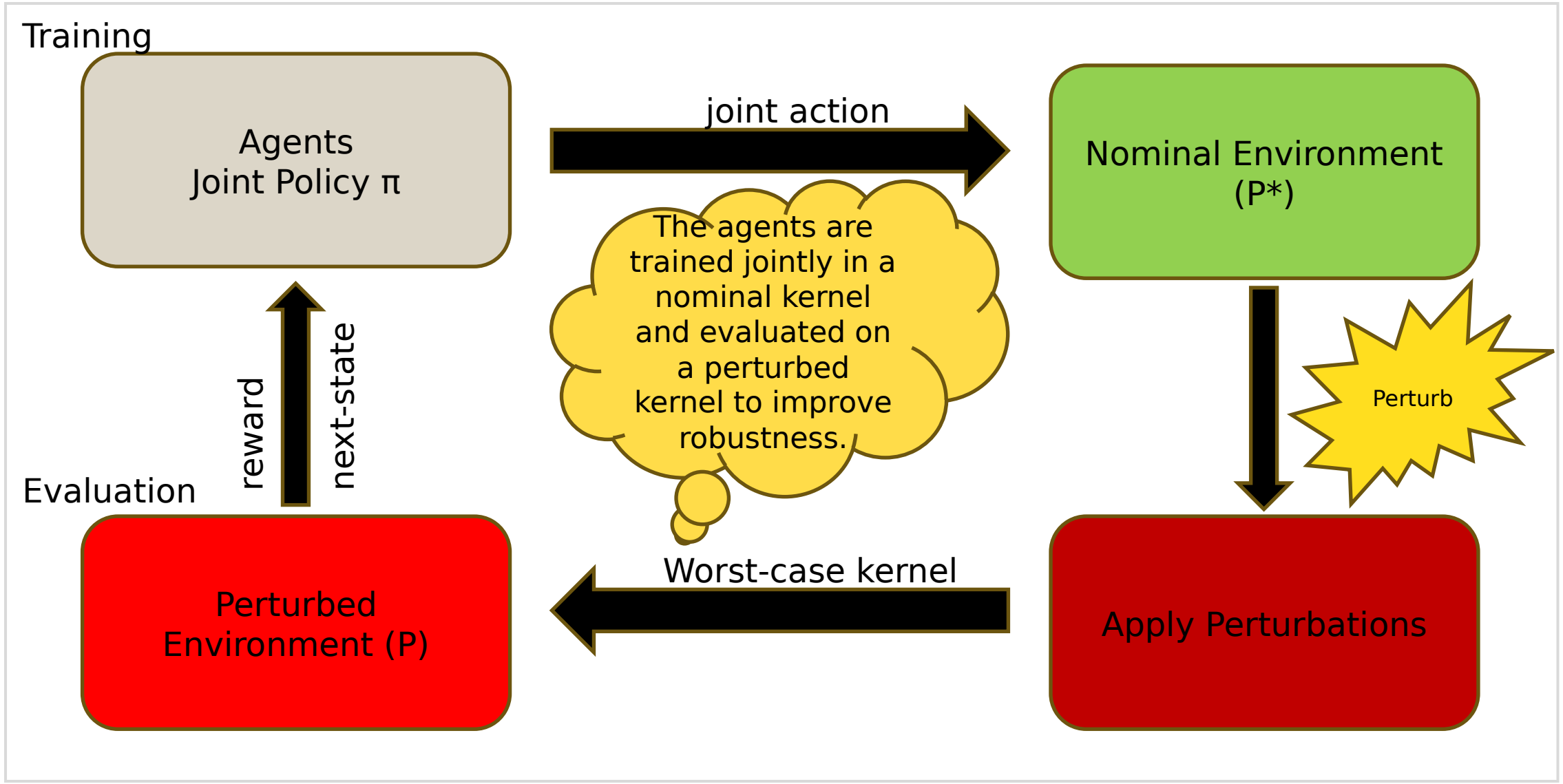
Standard RL thrives in simulation environments where training does not pose a threat or **danger** in the real world. For applications such as healthcare and autonomous driving, simulators are hard to design and data can be very costly to collect, forcing direct interaction with the environment.

To mitigate these problems, we consider the **Distributionally Robust RL** setup, which is designed to shrink the **sim-to-real gap**.

Distributionally Robust Markov Game (DRMG)

In the **multi agent** setting, we consider **general-sum Markov games**.

Address the **sim-to-real** gap by optimizing the worst-case performance over an **f-divergence** (KL/TV) **uncertainty set**.


Training
Agents
Joint Policy π
joint action
Nominal Environment
(P*)
The agents are trained jointly in a nominal kernel and evaluated on a perturbed kernel to improve robustness.
Perturb
reward
next-state
Evaluation
Perturbed
Environment (P)
Worst-case kernel
Apply Perturbations

Problem Formulation of DRMGs

- State Space *S*, Action Space *A*, Episodes K, and Horizon H
- Reward: $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$
- Nominal Transition Kernel: $P_h^\star(\cdot|s, a)$
- f-divergence uncertainty set: $\mathcal{P}_{i,h,f}^{\rho_i}(s, \mathbf{a}) = \left\{ P_h \in \Delta(\mathcal{S}) : f\left(P_h, P_h^\star(\cdot|s, \mathbf{a})\right) \leq \rho_i \right\}$
- Robust Value Function: $V_{i,h}^{\pi,\rho_i}(s) = \inf_{\tilde{P} \in \mathcal{P}_i} \mathbb{E}_{\pi,\tilde{\mathcal{P}}} \left[\sum_{t=h}^{H} r_{i,h}(s_t, \mathbf{a}_t) \middle| s_h = s \right]$
- Robust Action-Value Function: $Q_{i,h}^{\pi,\rho_i}(s, \mathbf{a}) = \inf_{\tilde{P} \in \mathcal{P}_i} \mathbb{E}_{\pi,\tilde{\mathcal{P}}} \left[\sum_{t=h}^{H} r_{i,h}(s_t, \mathbf{a}_t) \middle| s_h = s, \mathbf{a}_h = \mathbf{a} \right]$
- Our goal is to achieve one notion of equilibrium: Nash Equlibrium (NE), Coarse Correlated Equilibrium (CCE), or Coarse Equilibrium (CE)

Challenges of DRMGs

Agents are trained in **nominal kernel** but must optimize over the **worst-case kernel**

Exploration is harder under the **worst-case transition kernel** constraint and requires more samples to reach the desired **equilibrium**

In practice, achieving a **Nash equilibrium** is **PPAD-hard**, so we consider weaker notions such as **coarse correlated equilibrium** and **coarse equilibrium**

Online interaction is limited to data collected during exploration, making it **fundamentally harder** than having access to a simulator or a fully covered offline dataset

Hardness of Online DRMGs

- **Support Shift:** worst-case transition is not covered by the support of nominal transition
- Under both **TV** (with support shift) and **KL** (without), scaling with the joint action space is a hard lower bound.
  - For **TV-DRMG**: $\inf_{\mathcal{ALG}} \mathbb{E}[Regret_{NASH}(K)] \geq \Omega\left(\rho K \cdot \min\{H, \prod_{i \in \mathcal{M}} A_i\}\right).$
  - For **KL-DRMG**: $\inf_{\mathcal{ALG}} \mathbb{E}[Regret_{NASH}(K)] \geq \Omega\left(\sqrt{K \prod_{i \in \mathcal{M}} A_i}\right).$

Proposed Algorithm: f-MORNAVI

- Three Stage Algorithm:

1. **Nominal Transition Estimation:** estimate transition kernel P (model-based)
2. **Optimistic Robust Planning:** plan with optimism β to guide exploration

$$\overline{Q}_{i,h}^{k,\rho_i}(s,\boldsymbol{a}) = \min\left\{r_{i,h}(s,\boldsymbol{a}) + \sigma_{\widehat{\mathcal{P}}_{i,h,f}^{\rho_i}(s,\boldsymbol{a})}[\overline{V}_{i,h+1}^{k,\rho_i}] + \beta_{i,h,f}^{k}(s,\boldsymbol{a}),\ H\right\}.$$

$$\underline{Q}_{i,h}^{k,\rho_i}(s,\boldsymbol{a}) = \max\left\{r_{i,h}(s,\boldsymbol{a}) + \sigma_{\widehat{\mathcal{P}}_{i,h,f}^{\rho_i}(s,\boldsymbol{a})}[\underline{V}_{i,h+1}^{k,\rho_i}] - \beta_{i,h,f}^{k}(s,\boldsymbol{a}),\ 0\right\}.$$

   I. **Equilibrium subroutine:**

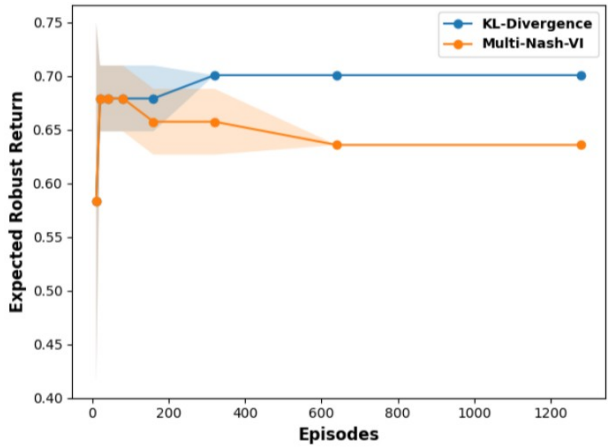
$$\pi_h^k(\cdot|s) \leftarrow \textsc{Equilibrium}\left(\left\{\overline{Q}_{i,h}^{k,\rho_i}(s,\cdot)\right\}_{i\in\mathcal{M}}\right).$$

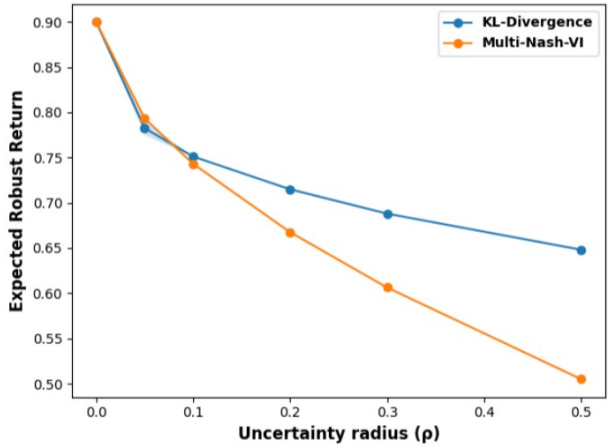
3. **Execution of Policy and Data Collection:** execute policy in environment and collect data

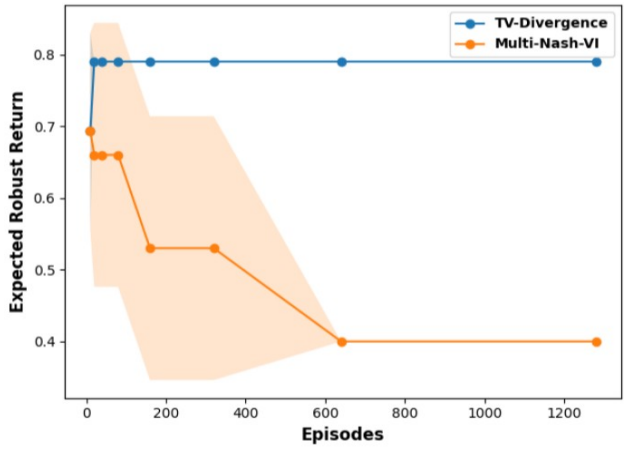
Theoretical Guarantees

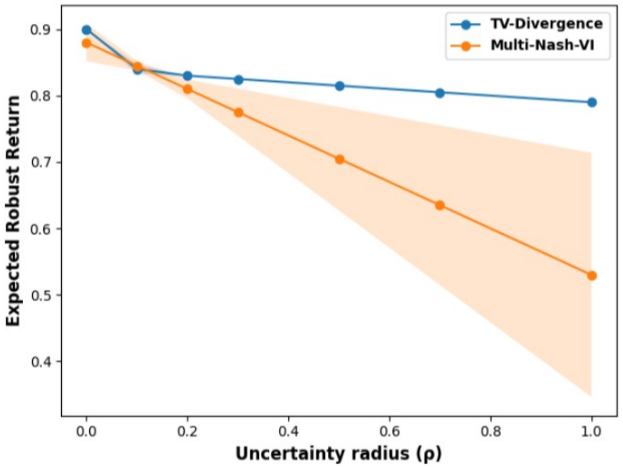
Algorithm	Setting	Uncertainty Set	Sample Complexity
(Shi et al. 2024b)	Generative	TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}H^3S(\prod_{i\in\mathcal{M}}A_i)\min\left\{\sigma_{\min}^{-1},H\right\}\right)$
(Jiao and Li 2024)	Generative	Contamination	$\tilde{\mathcal{O}}(\epsilon^{-2}H^3S(\sum_{i\in\mathcal{M}}A_i)\min\left\{\sigma_{\min}^{-1},H\right\})$
(Shi et al. 2024a)	Generative	TV (fictitious)	$\tilde{\mathcal{O}}\left(\epsilon^{-4}H^6S(\sum_{i\in\mathcal{M}}A_i)\min\left\{\sigma_{\min}^{-1},H\right\}\right)$
(Blanchet et al. 2023)	Offline	KL	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\sigma_{\min}^{-2}C_u^\star H^4\exp(H)S^2(\prod_{i\in\mathcal{M}}A_i)\right)$
		TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}C_u^\star H^4S^2(\prod_{i\in\mathcal{M}}A_i)\right)$
(Li et al. 2025)	Offline	TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}C_p^\star H^4S(\sum_{i=1}^m A_i)\min\left\{\{f(H,\sigma_i)\}_{i\in\mathcal{M}},H\right\}\right)$
(Ma et al. 2023)	Online	KL	$\tilde{\mathcal{O}}(\epsilon^{-2}H^5S(\max_i\{A_i\})^2)$ (with an oracle)
Our work	Online	TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}H^3S(\prod_{i\in\mathcal{M}}A_i)\min\left\{\sigma_{\min}^{-1},H\right\}\right)$
		KL	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\sigma_{\min}^{-2}(P_{\min}^\star)^{-1}H^4\exp(2H^2)S\left(\prod_{i\in\mathcal{M}}A_i\right)\right)$
Lower bound (Shi et al. 2024b)	Generative	TV	$\Omega\left(\epsilon^{-2}H^3S(\max_{i\in\mathcal{M}}A_i)\min\left\{\sigma_{\min}^{-1},H\right\}\right)$

Numerical Experiments: Fully Cooperative DRMG

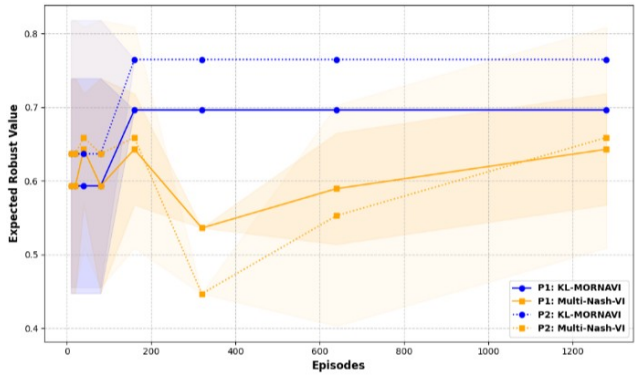


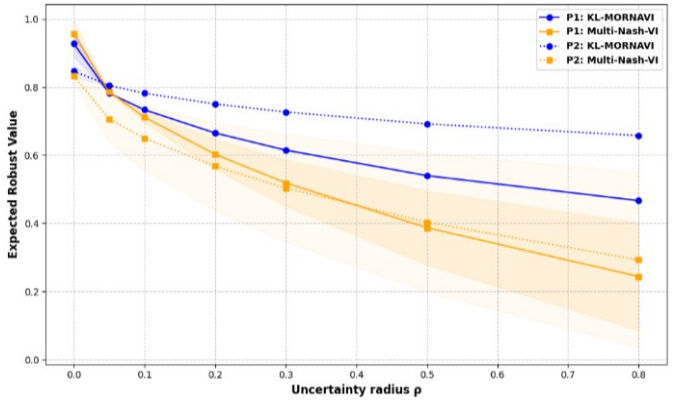


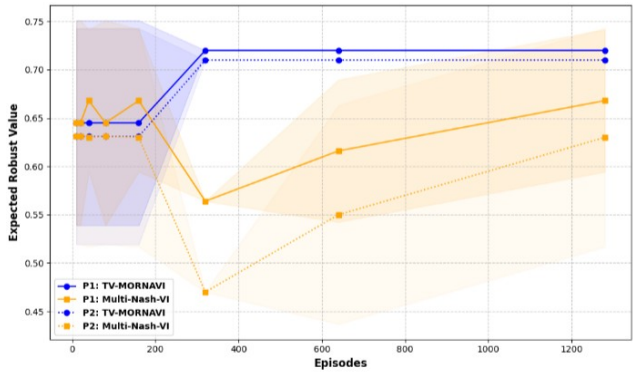


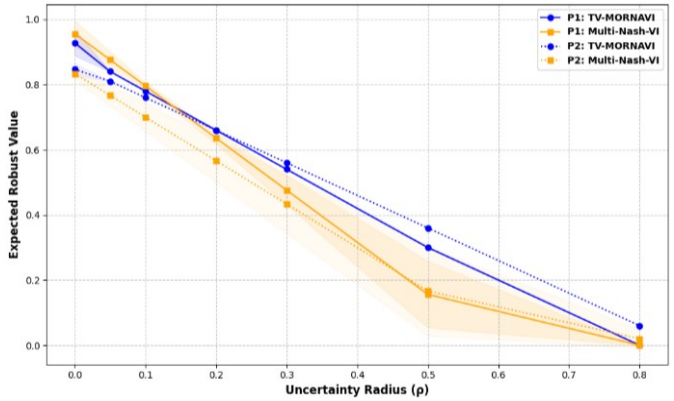


Numerical Experiments: General-Sum DRMG









Conclusion: Future Directions

- A simple **model-based distributionally robust** algorithm built on optimism
- First **provable guarantees** for **online DRMGs**
- **Numerical validation** over **KL** and **TV** uncertainty sets
- **Future Directions:**
  - Extend to function approximation
  - Eliminate joint action dependency
  - Scalable MARL

Thank You

- **Special Thanks:**
  - **Dr. Yue Wang** & **Dr. George K. Atia** — for guidance and support
  - **Dr. Debamita Ghosh** — postdoctoral researcher and co-first author, whose leadership was central to this work

Funding Source: DARPA HR0011-24-9-0427 & NSF CCF-2106339

Email: {za464241, de881780, george.atia, yue.wang}@ucf.edu