Online Distributionally Robust Reinforcement Learning with General Function Approximation

Dr. Debamita Ghosh

Postdoctoral Researcher
Department of Electrical and Computer Engineering
University of Central Florida, FL

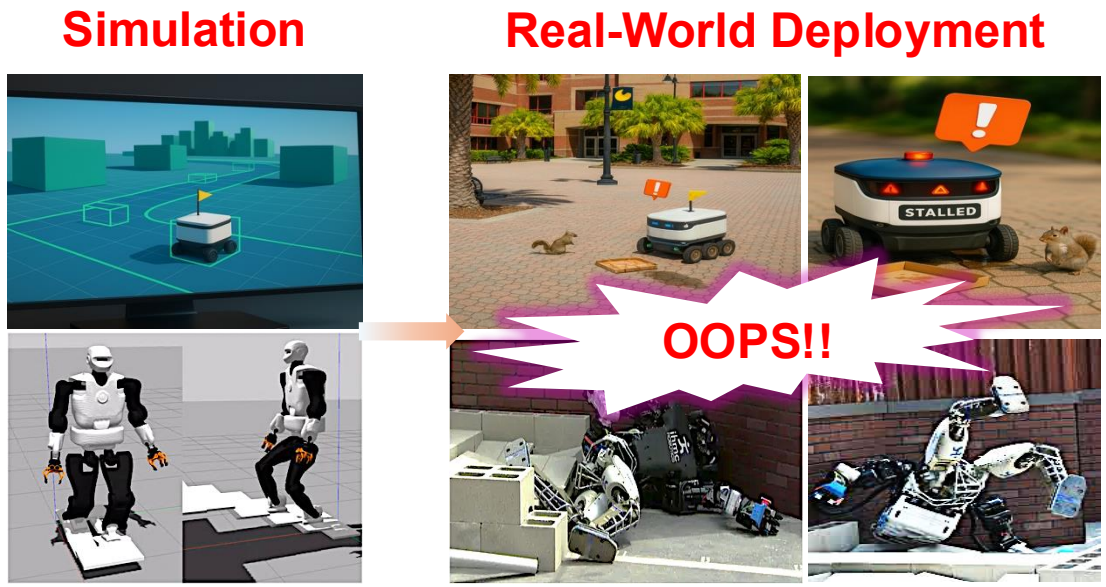
Prof. George K. Atia, *Department of Electrical and Computer Engineering, University of Central Florida, FL*

Prof. Yue Wang, *Department of Electrical and Computer Engineering, University of Central Florida, FL*

ICML Poster Presentation
Seoul, South Korea
July 6-11, 2026

ICML

Motivation: Bridge Sim-to-Real Gap



IHMC's Atlas robot failed on the first day of the DARPA Robotics Challenge Finals.

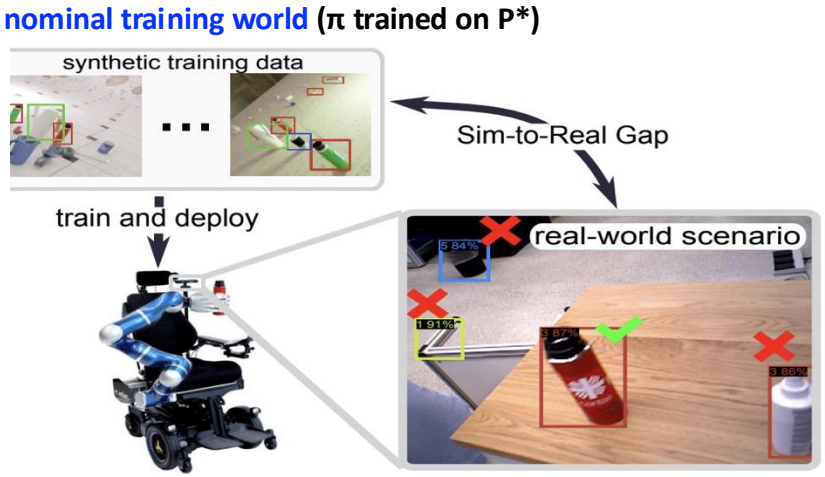


Illustration of a practical problem. *A detector trained only on synthetic images makes errors (red crosses) when deployed in real scenes due to the sim-to-real gap.*

Sim-to-real gap. **Real deployments** face unpredictable **noise**, unmodelled **perturbations** and even **adversarial attacks** that are **not** fully **captured** by the **simulator**.

Policies that look excellent in simulation can fail catastrophically after deployment.

Vanilla RL
Optimizes expected return under the training dynamics only.

Distributionally Robust RL (DR-RL)
Optimizes performance under a divergence ball around the nominal dynamics.

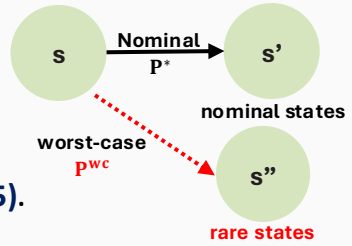
Online challenge
Data come from P, while robust value depends on worst-case dynamics.*

Key Challenges: Why is Online Robust RL difficult?

Purely **online** **robust reinforcement learning** with **general function approximation** is **fundamentally hard**.

Off-Dynamics Learning

Data come from **nominal dynamics.**
Robustness is evaluated under **worst-case dynamics.**



Key Challenge:

- **Distribution mismatch**
- **Information bottleneck**

(Lu et al. 2024, He et al. 2025).

Existing methods need generative model or rich offline data or strong coverage assumptions **(Shi et al. 2023; Shi and Chi 2024).**

Robust Bellman Learning

Robust Bellman update requires optimization over uncertainty sets.

$$\left(\mathcal{T}_h^{\phi,\sigma} V\right)(s,a) = r(s,a) + \inf_{P \in \mathcal{U}_{\phi,\sigma}} \mathbb{E}_P[V]$$

Key Challenge:

- Worst-case operator is **not directly observable.**
- **Inner optimization** required at every **(s,a).**

Existing Standard fitted RL/UCB assume nominal Bellman updates **(Jin et al., 2021)**. Robust RL requires learning a worst-case Bellman operator online **(Panaganti et al., 2022; 2024).**

Scalability

Real-world robust RL requires:

- **neural networks**
- **continuous states**
- **large action spaces**

Key Challenge:

- **Tabular bonuses** scale with: $|\mathcal{S}||\mathcal{A}|$.
- **Linear function approximation methods** require **restrictive linearly structure assumptions.**

No general intrinsic-complexity theory existed for online DR-RL.

Can we learn a robust policy purely online, with general function approximation, and provable regret?

- Shi, L.; Li, G.; Wei, Y.; Chen, Y.; Geist, M.; and Chi, Y. 2023. The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model. Advances in Neural Information Processing Systems, 36: 79903–79917.
- Shi, L.; and Chi, Y. 2024. Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity. Journal of Machine Learning Research, 25(200): 1–91.
- Lu, M.; Zhong, H.; Zhang, T.; and Blanchet, J. 2024. Distributionally Robust Reinforcement Learning with Interactive Data Collection: Fundamental Hardness and Near-Optimal Algorithm. *The Thirty-eighth Annual Conference on Neural Information Processing Systems.*
- He, Y.; Liu, Z.; Wang, W.; and Xu, P. 2025. Sample Complexity of Distributionally Robust Off-Dynamics Reinforcement Learning with Online Interaction. *In Forty-second International Conference on Machine Learning.*
- *Jin, C., Liu, Q., and Miryoosefi, S. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. Advances in neural information processing systems, 34:13406–13418, 2021.*
- *Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust Reinforcement Learning using Offline Data. Advances in neural information processing systems, 35:32211–32224, 2022.*
- *Panaganti, K., Wierman, A., and Mazumdar, E. Model-Free Robust ϕ-Divergence Reinforcement Learning Using Both Offline and Online Data. arXiv preprint arXiv:2405.05468, 202*

Algorithm: Robust Fitted Learning with ϕ-Divergence (RFL-ϕ)

- $\mathcal{F}$: value functional class
- $\mathcal{G}$: dual functional class
- β: confidence parameter
- σ: robustness radius

Interaction with Nominal Env. & Data update

Agent

$\left(\text{robust policy } \pi^{k} = \pi^{f^{(k)}}\right)$

Action $a_h^{(k)} \sim \pi^{k}\left(\cdot \middle| s_h^{(k)}\right)$

Reward $r_h^{(k)}$ next-state $\sim P_h^*(\cdot | s_h^{(k)}, a_h^{(k)})$

Nominal Environment P^*

trajectories

Data update

Update online data buffer

$$\mathcal{D}_h^{(k)} \leftarrow \mathcal{D}_h^{(k-1)} \cup \left\{\left(s_h^{(k)}, a_h^{(k)}, r_h^{(k)}, s_{h+1}^{(k)}\right)\right\}$$

Dual Robust Bellman Learning (Our Core)

1. Learn dual function $\left(\text{for any } f_{h+1} \in \mathcal{F}_{h+1}^{(k)}\right)$

$$\underline{g}_{f_{h+1}}^{(k)} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \widehat{\text{DualLoss}}_h^{(k)}(g; f_{h+1})$$

2. Approximate robust Bellman backup

$\widehat{\mathcal{T}}_{\underline{g}_{f_{h+1}}^{(k)}}^{\phi,\sigma} f_{h+1}$ (dual-based robust Bellman operator)

3. Compute loss for candidate f_h

$\mathcal{L}_h^{(k)}\left(f_h, f_{h+1}, \underline{g}_{f_{h+1}}^{(k)}\right)$ -- robust fitting loss on $\mathcal{D}_h^{(k)}$

Confidence set over functional class $\mathcal{F}$

Build confidence set for step h:

$$\mathcal{F}_h^{(k)} = \left\{ \begin{array}{c} f_h \in \mathcal{F}_h: \\ \mathcal{L}_h^{(k)}\left(f_h, f_{h+1}, \underline{g}_{f_{h+1}}^{(k)}\right) \\ - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(k)}\left(f'_h, f_{h+1}, \underline{g}_{f_{h+1}}^{(k)}\right) \\ \leq \beta, \forall h \in [H] \end{array} \right\}$$

Optimistic Robust Planning & Greedy Policy

Optimistic robust value estimation

$$f^{(k)} = \underset{f \in \mathcal{F}^{(k-1)}}{\operatorname{argmax}} f_1(s_1, \pi_1^f(s_1))$$

Greedy robust policy

$$\pi^{(k)} = \pi^{f^{(k)}}$$

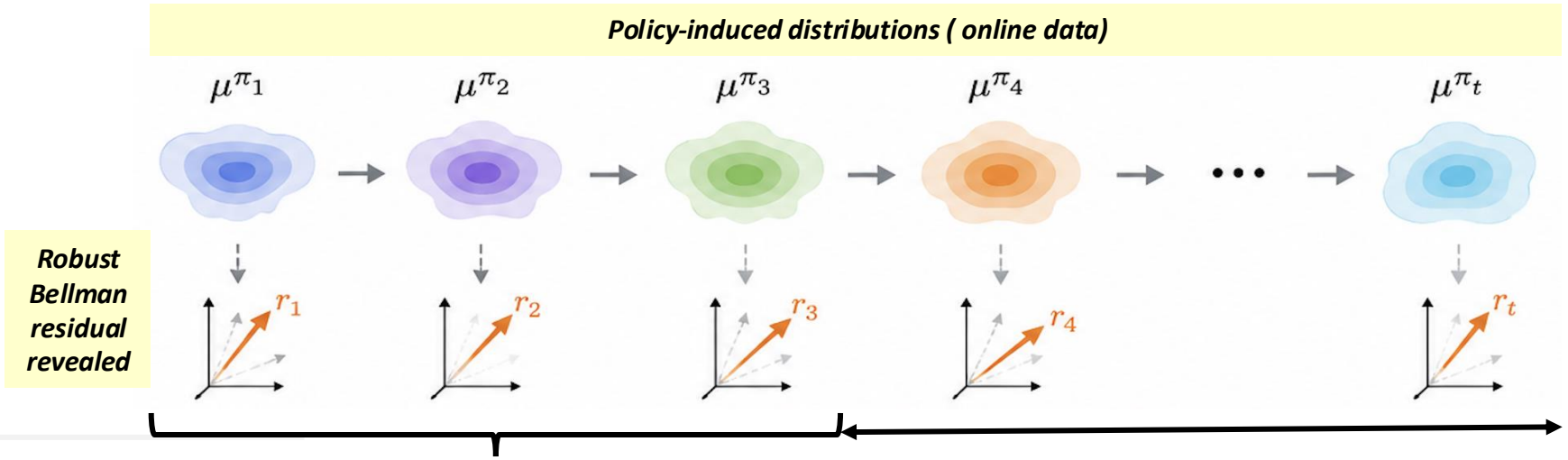
- Choose the **most optimistic value function** in the **confidence set** and **act greedily**.

Repeat over episodes k

Theoretical Guarantee: Why this matters ?

Robust Bellman-Eluder (BE) dimension: $\mathrm{d}_{\mathrm{BE}}^{\mathrm{rob}} = \underbrace{\dim_{\mathrm{DE}}}_{\text{Eluder dimension}} \Big(\underbrace{\big(\mathcal{I} - \mathcal{T}_{\mathrm{h}}^{\phi,\sigma}\big)\mathcal{F}}_{\text{Robust Bellman residual class}} \Big)$

- **intrinsic exploration complexity** under **robust Bellman updates.**
- Counts how many **policy distributions** reveal **new, statistically, independent Bellman errors.**



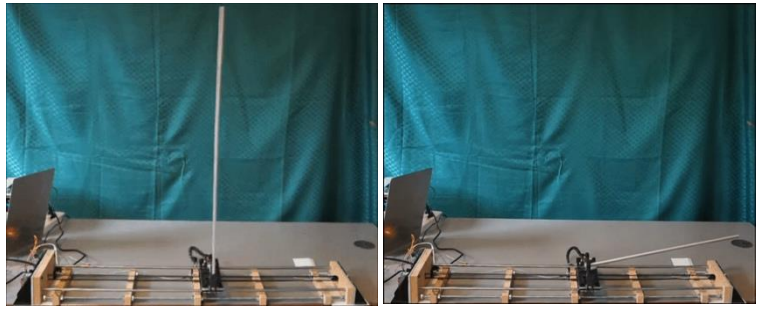
Why it matters?

- **Smaller** $\mathrm{d}_{\mathrm{BE}}^{\mathrm{rob}} \Longrightarrow$ **fewer essentially** different robust Bellman errors.
- **Exploration** scales with $\mathrm{d}_{\mathrm{BE}}^{\mathrm{rob}}$ **not** $|\mathcal{S}|\,|\mathcal{A}| \Longrightarrow$ **sample efficient.**
- **First regret guarantee** for **purely online DR-RL** with **general function approximation**:

$$\mathbf{Regret(K)} \leq \widetilde{\mathbf{O}}\left(\sqrt{\mathbf{d}_{\mathbf{BE}}^{\mathbf{rob}}\mathbf{H}^{\mathbf{2}}\mathbf{B}_{\boldsymbol{\phi}}^{\mathbf{2}}(\boldsymbol{\sigma})\mathbf{K}} + \boldsymbol{\varepsilon}^{\mathbf{dual}}\right).$$

- **Recovers near-optimal tabular** and **linear** robust RL **guarantees.**

Experiments: Cartpole



CartPole policies trained nominally, then tested under action and dynamics perturbations.

Strong robustness under perturbations

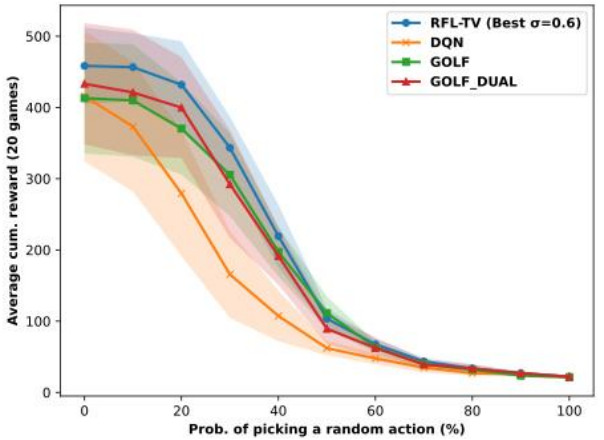
- Up to 2 ×higher reward than DQN; competitive with tabular OPROVI-TV.

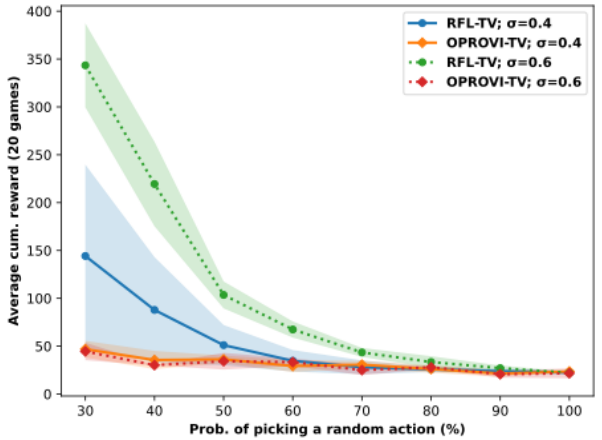
Stable under severe dynamics shifts

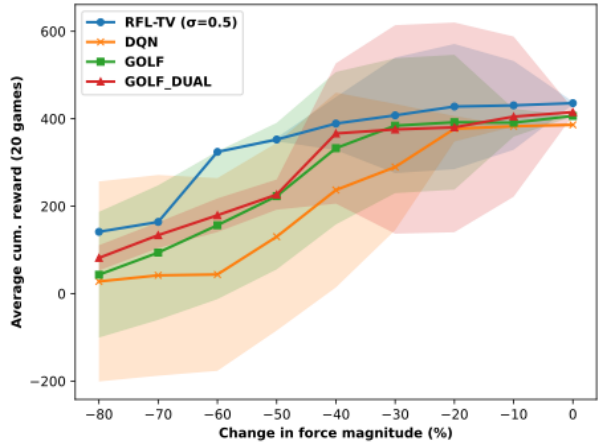
- Robust even under $-80\% \sim -40\%$ force perturbations; non-robust FA baselines collapse.

Action Perturbation

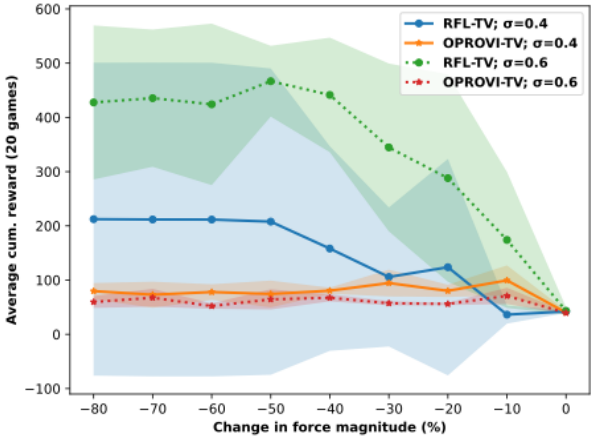
Force-magnitude Perturbation







RFL-TV vs. Functional Approximation (FA) Algorithms



RFL-TV vs. OPROVI-TV (Tabular).

Key Takeaway: Robust fitted Bellman learning achieves strong robustness under severe distribution shifts without relying on tabular representations.

THANK YOU!

ANY QUESTIONS?

de881780@ucf.edu

yue.wang@ucf.edu

george.atia@ucf.edu

Poster Q&A & Follow-Up:

https://forms.gle/FRMRDh4Toj2GY5wC7