Combining Optimism with Pessimism for Robust and Efficient Model-Based Deep Reinforcement Learning

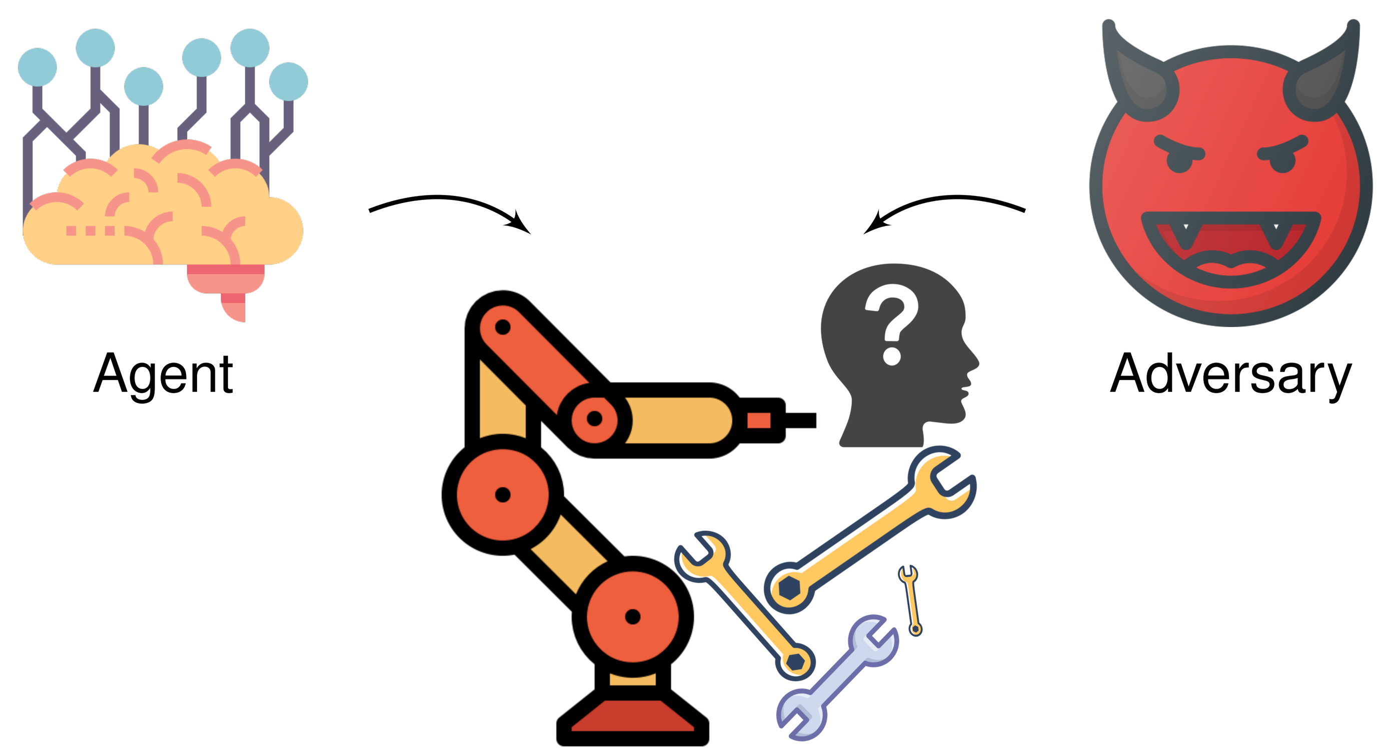
Sebastian Curi, Ilija Bogunovic, Andreas Krause

ETH zürich

Paper Code

RH-UCRL: A sample efficient algorithm that provably outputs a robust policy.

Problem Setting: Zero-Sum Markov Game

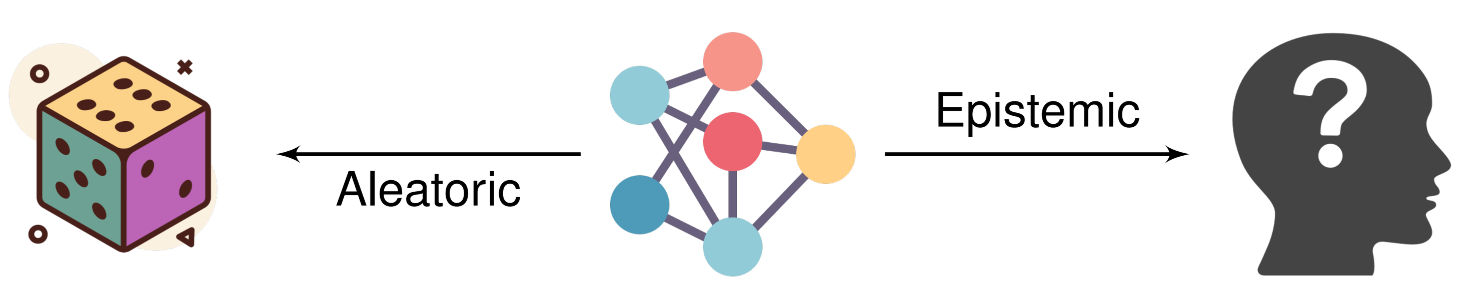


How to output a **single robust policy** that works well for every tool?

We simulate an adversary that has the ability to choose the tool during training, but we cannot control it during testing.

How to **explore with the adversary** is crucial for sample efficiency!

Model Learning: Aleatoric vs. Epistemic Uncertainty

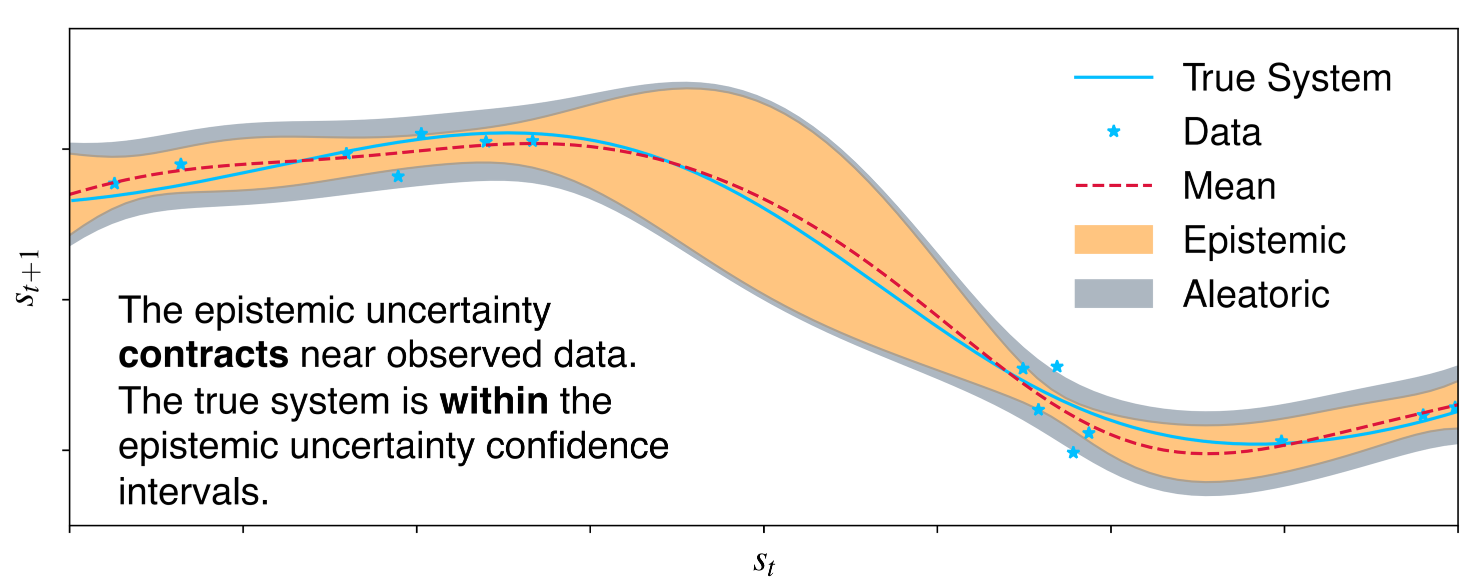


- *Aleatoric*: inherent stochasticity from the system (e.g. sensor noise).
- *Epistemic*: data scarcity (e.g. unknown mass of the robot links).

Definition (Set of Plausible Models) $\mathcal{M}_t = \{\tilde{f}, |\tilde{f} - \mu_t| \leq \beta_t \sigma_t\}$

Assumption (Well-Calibrated Models) $f \in \mathcal{M}_t \quad \forall t = 0, 1, \ldots$

- GP models are calibrated under certain conditions (Srinivas et al., 2010).
- Bayesian NN models can be recalibrated empirically (Malik et al., 2019).



Definitions

Performance

$$J(\pi, \bar{\pi}; \tilde{f}) = \sum_{h=1}^{H} r_h(s_h, a_h, \bar{a}_h), \quad \text{s.t.} \quad s_{h+1} = \tilde{f}(s_h, \pi(s_h), \bar{\pi}(s_h))$$

Optimal Policy

$$\pi^* = \arg\max_{\pi \in \Pi} \min_{\bar{\pi} \in \bar{\Pi}} J(\pi, \bar{\pi}; f)$$

Goal: Output a policy $\hat{\pi}$

$$\min_{\bar{\pi} \in \bar{\Pi}} J(\hat{\pi}, \bar{\pi}; f) \geq \min_{\bar{\pi} \in \bar{\Pi}} J(\pi^*, \bar{\pi}; f) - \epsilon$$

Given Output Precision

RH-UCRL

I) Optimistic and Pessimistic Performance through Hallucination

We construct optimistic and pessimistic estimates of the policies by optimizing w.r.t. the set of plausible models.
To make the optimization practical, we use hallucination as in Curi et al. (2020), and reparameterize the set of plausible models with a hallucinated control input.

$$J_t^{(p)}(\pi, \bar{\pi}) = \min_{\tilde{f} \in \mathcal{M}_t} J(\pi, \bar{\pi}; \tilde{f}) = \min_{\eta} J(\pi, \bar{\pi}; \mu_t + \beta_t \sigma_t \eta)$$

$$J_t^{(o)}(\pi, \bar{\pi}) = \max_{\tilde{f} \in \mathcal{M}_t} J(\pi, \bar{\pi}; \tilde{f}) = \max_{\eta} J(\pi, \bar{\pi}; \mu_t + \beta_t \sigma_t \eta)$$

II) Agent and Adversary Policy Selection

At the beginning of each episode, the agent and adversary use the optimistic and pessimistic estimates to select their policies.

$$\pi_t = \arg\max_{\pi \in \Pi} \min_{\bar{\pi} \in \bar{\Pi}} J_t^{(o)}(\pi, \bar{\pi}),$$

$$\bar{\pi}_t = \arg\min_{\bar{\pi} \in \bar{\Pi}} J_t^{(p)}(\pi_t, \bar{\pi})$$

III) Algorithm Output

After T episodes, the algorithm outputs the policy that maximizes the sequence of pessimistic estimates.

$$\hat{\pi} = \arg\max_{1,\ldots,T} J_t^{(p)}(\pi_t, \bar{\pi}_t)$$

Theoretical Results: Simple Regret

After

$$T = \tilde{O}\left(\frac{H^3 \beta_T^H C^{2H} \Gamma_T}{\epsilon^2}\right)$$

episodes, **RH-UCRL** outputs a robust policy $\hat{\pi}$ that satisfies the requirement,

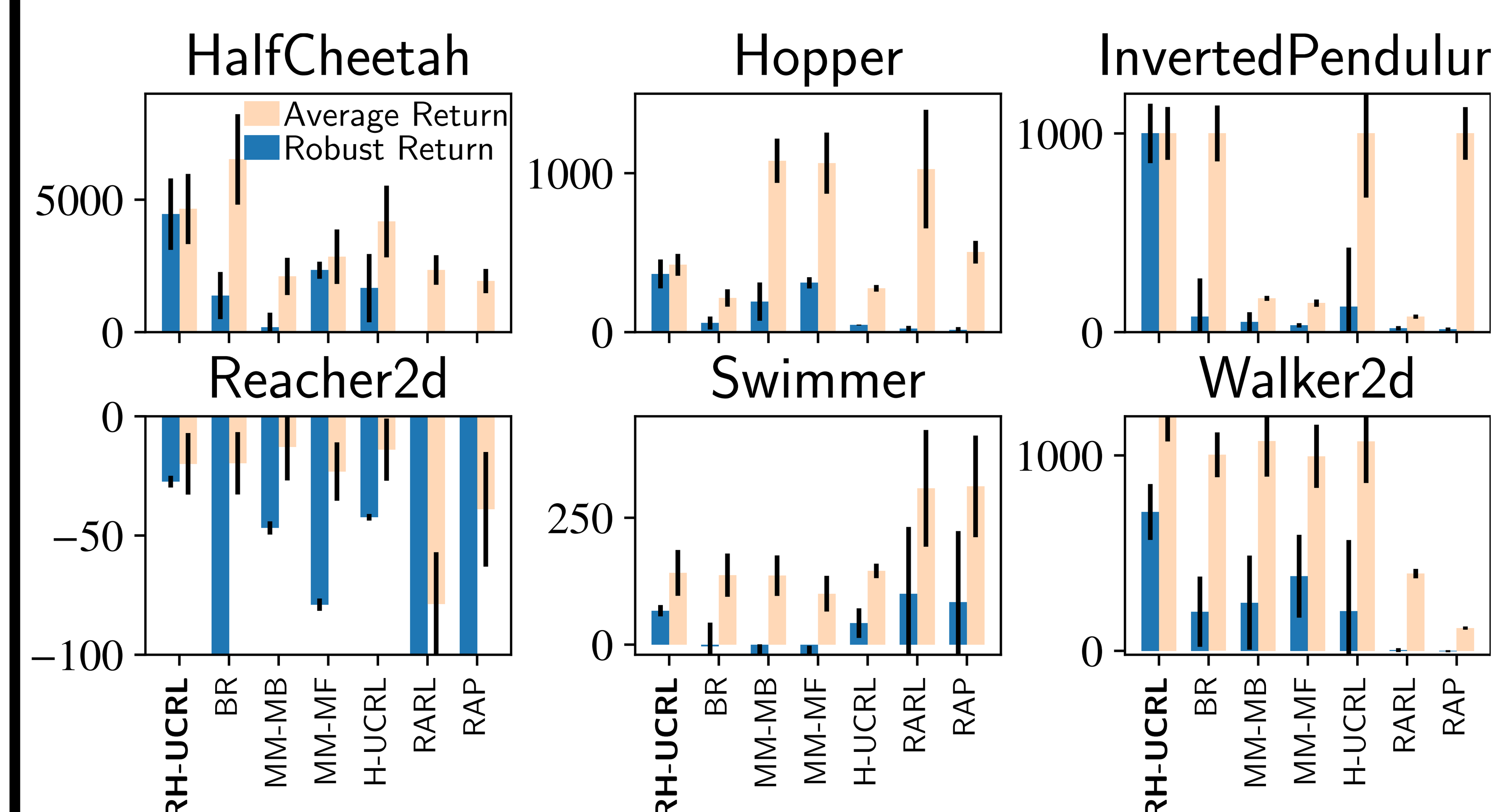
$$\min_{\bar{\pi} \in \bar{\Pi}} J(\hat{\pi}, \bar{\pi}; f) \geq \min_{\bar{\pi} \in \bar{\Pi}} J(\pi^*, \bar{\pi}; f) - \epsilon.$$

- β_T *: A scalar that enlarges the confidence intervals for calibration.*
- Γ_T *: A measure of complexity of the model-class we are trying to learn.*

In the main paper, we also provide an analysis of cumulative regret.

Experimental Results

We train the different algorithms for 200 episodes. Next, we freeze the agent policy and train the adversary for another 200 episodes. The average return is the return without an adversary. The robust return is the return with the fully trained adversary.



- RH-UCRL **outperforms** the other algorithms in terms of **robust** return.
- RH-UCRL has **good performance** in terms of **average** return..

References

.Curi, S., Berkenkamp, F., & Krause, A. (2020). Efficient Model-Based Reinforcement Learing through Optimistic Policy Search and Planning. NeuRIPS.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., & Ermon, S. (2019). Calibrated Model-Based Deep Reinforcement Learning. ICML.