# Robust Reinforcement Learning in a Sample-Efficient Setting

Siemen Herremans[1,2,3], Ali Anwar[1,3], Siegfried Mercelis[1,3]

[1] IDLab – Faculty of Applied Engineering at University of Antwerp, [2] Research Foundation Flanders, [3] imec

ICML International Conference On Machine Learning

Paper

*Evaluation code available.*

*Relevant training code snippets (JAX) in paper.*

## 0. Abstract Summary


- Robust Adversarial RL often requires strong practical assumptions, such as a parameterizable simulator at training time.
- By moving the adversary to act on the learned model, we avoid these assumptions and gain the sample efficiency of model-based methods.
- Since we explicitly have access to predicted nominal and adversarial transitions, we can interpret the adversarial transitions.
- Our quantitative results demonstrate that we **meet or outperform state-of-the-art robust RL methods in a variety of locomotion tasks, often with fewer assumptions and always with less samples.**


## 1. Background: Robust Adversarial RL

**Robust Adversarial Reinforcement Learning** (RARL) considers an agent that operates in a set $\mathcal{P}$ of Markov Decision Processes (MDPs). Notably, the agent is oblivious to which environment it is operating in. The goal of the agent in this Robust MDP is (by definition) to maximize its expected utility when the MDP is adversarially selected from the set. This adversarial selection happens at every timestep, to keep the problem tractable ($(s,a)$-rectangularity assumption).

$$J_{robust}(\pi) = \min_{P \in \mathcal{P}} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

With the corresponding optimal policy:

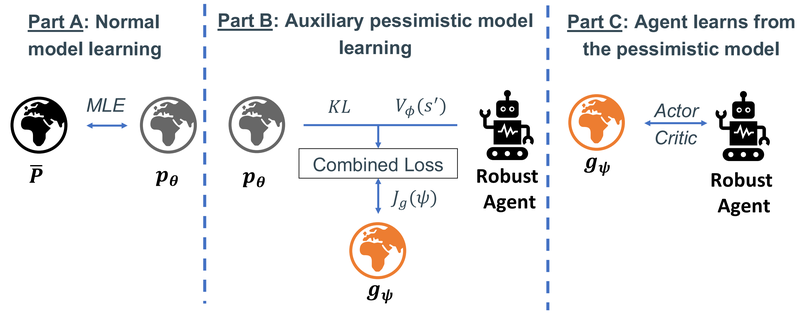
$$\pi^*_{\mathcal{P}} = \underset{\pi \in \Pi}{\operatorname{argmax}} J_{robust}(\pi)$$

Many options to choose $\mathcal{P}$ exist, in this work we opt for **the Kullback-Leibler (KL) uncertainty set** due its property to blow up outside of the support of the nominal transition function $\bar{P}(\cdot)$. Additionally, as the KL-divergence bounds the total variation distance, we lower-bound the nominal performance (see Section 3).

$$\mathcal{P}^{s,a}_{KL} = \{P(s',r|s,a) \in \Delta_{\mathcal{S}\times\mathcal{R}} \mid D_{KL}(P(s',r|s,a) \,||\bar{P}(s',r|s,a)) \leq \epsilon^{s,a}\}$$

## 2. Auxiliary Pessimistic Transition Model Learning

Starting from a model-based setting, where $p_\theta$ approximates the nominal transition function $\bar{P}$, an auxiliary model $g_\psi$ is trained to serve as an adversarial version of the nominal model.




Where the auxiliary pessimistic model is trained by minimizing the likelihood of high-value transitions (based on Rigter et al., 2022):

$$J_g(\psi) = \mathbb{E}_{(.)}\left[D_{KL}\left(g_\psi(\cdot)||p_\theta(\cdot)\right) + \eta \cdot ln\left(g_\psi(s',r|s,a,p_\theta(\cdot))(r + \gamma\, V^{\theta,\phi}_\psi(s')\right)\right]$$

Minimize the likelihood of high-value transitions

Remain reasonably close to the nominal model

## 3. Impact on performance bound

Janner et al. (2019) provided a lower bound on the return of a policy in the real environment $G[\pi]$ , given its predicted performance in the learned model. This bound depends on the model error $\epsilon_m$ and the off-policy error $\epsilon_\pi$ . Our work demonstrates that this bound can be extended to include the Total Variation (TV) distance $\epsilon_{m^{aux}} = \max_t \mathbb{E}_{s\sim\pi,t}\left[D_{TV}(g_\psi(\cdot), p_\theta(\cdot))\right]$ . Note that the KL-divergence bounds the TV distance.

$$G[\pi] \geq \hat{G}_{aux}[\pi] - \left[\frac{2\gamma r_{max}(\epsilon_m + 2\epsilon_\pi + \epsilon_{m^{aux}})}{(1-\gamma)^2} + \frac{4r_{max}\epsilon_\pi}{(1-\gamma)}\right]$$

## 4. Practical Algorithm: Robust MBPO (RMBPO)

We implement the robust model in MBPO (Janner et al., 2019). The proposed modifications are:

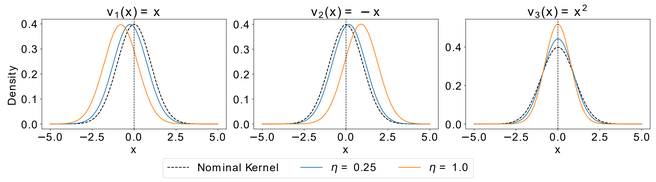
- *Learning the auxiliary model as described in Section 2.*
- *Using the auxiliary model as the environment for Soft Actor-Critic.*

**Algorithm 1** RMBPO (Additions in blue)

Initialize policy $\pi_\phi$, predictive model $p_\theta$ , auxiliary model $g_\psi$,
environment dataset $\mathcal{D}_{env}$, model dataset $\mathcal{D}_{model}$
**for** N epochs **do**
  **while** improving on holdout set **do**
    Update model parameters $\theta$ on environment data $\mathcal{D}_{env}$ via maximum likelihood
  **end while**
  **while** improving on holdout set **do**
    Update auxiliary model parameters $\psi$ according to Eq. 7: $\psi \leftarrow \psi - \lambda_a \hat{\nabla}_\psi J_g(\psi, \mathcal{D}_{env}, p_\theta, \pi_\phi)$
  **end while**
  **for** E steps **do**
    Take action in environment according to $\pi_\phi$; add to $\mathcal{D}_{env}$
    **for** M model rollouts **do**
      Sample $s_t$ uniformly from $\mathcal{D}_{env}$
      On-policy rollout according to Eq. 5 starting from $s_t$ using policy $\pi_\phi$; add to $\mathcal{D}_{model}$
    **end for**
    Perform (soft) actor-critic updates on $\phi$ using samples from $D_{model}$.
  **end for**
**end for**

## 5. Toy experiment

To demonstrate how the loss function biases the transition distribution towards low-value transitions, we fit a standard Gaussian distribution to 3 value functions: $v_1(x) = x, v_2(x) = -x, v_3(x) = x^2$, using the loss function from Section 2. As expected, the pessimism is proportional with $\eta$.

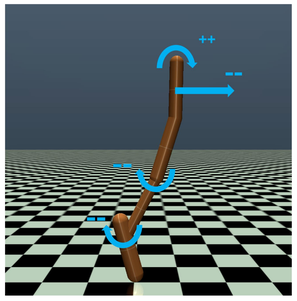



## 6. Interpretability

When comparing the 4 largest state modifications between the nominal model and the pessimistic model in Hopper-v4, we see:

- *Increased torso angular velocity*
- *Decreased actuator angular velocity*
- *Lower forward-moving velocity*

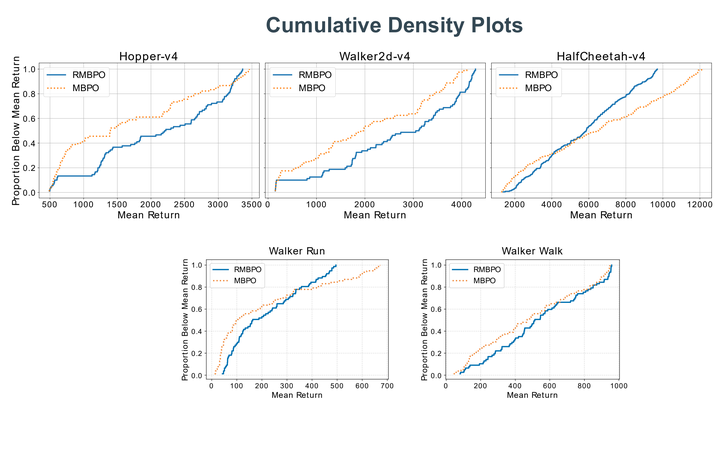
All these aspects are expected, as this makes the robot more prone to falling, and slower, therefore minimizing the value.




## 7. Representative Results

We evaluate the robustness of RMBPO by **training it in the nominal environment and evaluating it under a sweep of two simultaneous perturbations**. In all environments, RMBPO is competitive with state-of-the-art baselines using less data (see paper Fig. 3, Fig. 4 for more results).

Next, we generate an **empirical cumulative density plot** of these experiments which confirms that RMBPO successfully reduces the number of low-return episodes compared to the MBPO. Sometimes at the expected trade-off of a lower nominal performance.


## References

Janner, M., Fu, J., Zhang, M., & Levine, S. (2019). When to Trust Your Model: Model-Based Policy Optimization. *NeurIPS* (Vol. 32).

Rigter, M., Lacerda, B., & Hawes, N. (2022). Rambo-rl: Robust adversarial model-based offline reinforcement learning. NeurIPS (Vol. 35).