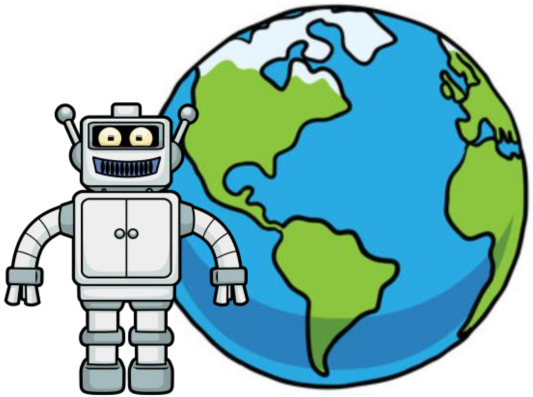
Reward-Free Curricula for Training Robust World Models

Marc Rigter, Minqi Jiang, Ingmar Posner

Research Question

How should we train a general world model that is robust to both:

- The task (i.e. reward function)
- A range of environments



→ Reward-free pretraining setting

Preliminaries (UPOMDP)

Underspecified POMDP:

$$\mathcal{U} = \{\mathcal{P}_\theta\}_{\theta \in \Theta}$$

where each $\mathcal{P}_\theta$ is a standard POMDP with different dynamics.

P_1 P_2 P_3 P_4 P_5 P_6 $\cdots$

Preliminaries (Regret)

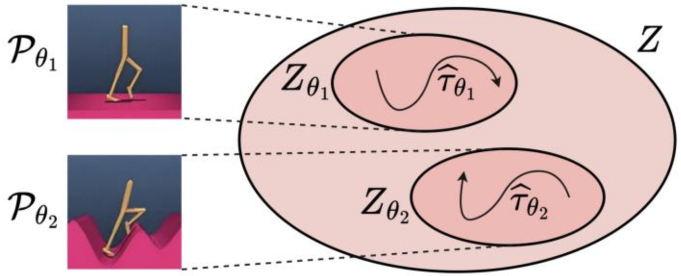
Given:

- Environment, $\mathcal{P}_\theta$
- Reward function, R
- Policy, π

$$\text{REGRET}(\pi, \mathcal{P}_\theta^R) := V(\pi^*_{\theta,R}, \mathcal{P}_\theta^R) - V(\pi, \mathcal{P}_\theta^R)$$

Problem Definition

A world model, W learns approximation $\widehat{\mathcal{P}}_\theta$ for all $\theta \in \Theta$ in shared latent state space Z.



Assume an optimal planner in world model:

$$\widehat{\pi}^*_{\theta,R} = \arg\max_{\pi} V(\pi, \widehat{\mathcal{P}}^R_\theta)$$

Find the world model that minimizes the maximum regret:

$$W^* = \arg\min_{W} \max_{\theta,R} \textsc{Regret}(\widehat{\pi}^*_{\theta,R}, \mathcal{P}^R_\theta)$$

Analysis

Assume π_θ^{expl} seeks the maximum world model error in each environment.

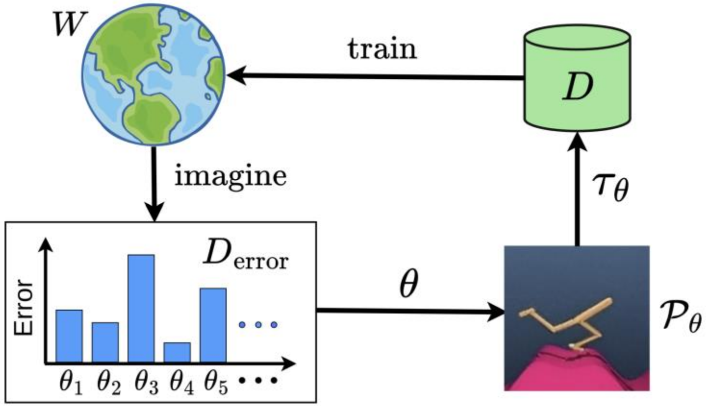
Then the maximum regret is bounded by:

$$\max_{\theta,R} \textsc{Regret}(\widehat{\pi}^*_{\theta,R}, \mathcal{P}_\theta^R) \leq \max_\theta \frac{4\gamma}{(1-\gamma)^2} \underbrace{\mathbb{E}_{z,a\sim d(\pi_\theta^{\text{expl}}, \widehat{\mathcal{P}}_\theta)} \Big[\mathrm{TV}\big(\widehat{T}(\cdot|z,a), T(\cdot|z,a)\big)\Big]}_{\text{Latent dynamics error in } \mathcal{P}_\theta \text{ under } \pi_\theta^{\text{expl}}}$$

→ Minimise maximum latent dynamics error across environments under π_θ^{expl}

Practical Algorithm (WAKER)

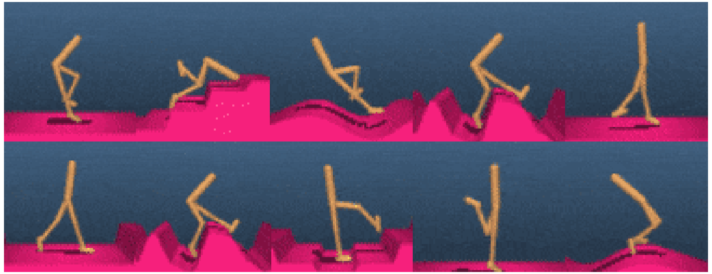
Idea: Sample more data from environments that have high epistemic uncertainty under $\pi_{\theta}^{\text{expl}}$.



→ The more complex an environment is, the more data is gathered

Experiments

- First, reward-free world model learning.
- Second, reward functions provided and policies optimized in world model.



Exploration Policy		Plan2Explore						Random Exploration		
Environment Sampling		WAKER-M	WAKER-R	DR	GE	HE-Oracle	RW-Oracle	WAKER-M	WAKER-R	DR
Clean Up	Sort	0.711 ± 0.09	0.643 ± 0.07	0.397 ± 0.12	0.426 ± 0.09	0.240 ± 0.13	0.482 ± 0.14	0.007 ± 0.01	0.010 ± 0.01	0.000 ± 0.0
	Sort-Rev.	0.741 ± 0.06	0.586 ± 0.06	0.395 ± 0.10	0.490 ± 0.07	0.230 ± 0.11	0.537 ± 0.10	0.000 ± 0.0	0.000 ± 0.0	0.000 ± 0.0
	Push	0.716 ± 0.11	$0. 702 \pm 0.08$	0.590 ± 0.093	0.628 ± 0.12	0.262 ± 0.16	0.596 ± 0.14	0.124 ± 0.09	0.058 ± 0.06	0.023 ± 0.03
Car Clean Up	Sort	0.894 ± 0.04	0.815 ± 0.11	0.665 ± 0.14	0.641 ± 0.09	0.041 ± 0.07	0.624 ± 0.15	0.433 ± 0.11	0.378 ± 0.09	0.337 ± 0.07
	Sort-Rev.	0.914 ± 0.079	0.880 ± 0.080	0.659 ± 0.16	0.646 ± 0.13	0.043 ± 0.06	0.567 ± 0.16	0.408 ± 0.10	0.408 ± 0.09	0.269 ± 0.11
	Push	0.906 ± 0.05	0.888 ± 0.04	0.796 ± 0.10	0.807 ± 0.12	0.046 ± 0.06	0.777 ± 0.11	0.584 ± 0.12	0.526 ± 0.15	0.373 ± 0.12
Terrain Walker	Walk	818.0 ± 15.3	805.3 ± 42.0	748.9 ± 39.5	741.2 ± 43.6	543.2 ± 85.3	791.6 ± 32.3	243.9 ± 26.7	224.8 ± 41.9	224.3 ± 25.3
	Run	312.6 ± 19.9	303.0 ± 16.1	279.9 ± 18.1	300.1 ± 17.4	223.3 ± 18.6	305.5 ± 15.2	120.4 ± 14.7	104.2 ± 9.7	114.1 ± 12.2
	Flip	955.0 ± 11.9	937.7 ± 10.5	936.1 ± 10.2	946.0 ± 9.5	962.9 ± 5.7	952.4 ± 11.6	878.9 ± 18.4	850.7 ± 40.5	849.9 ± 27.4
	Stand	941.2 ± 12.3	945.4 ± 16.6	936.5 ± 17.5	938.6 ± 16.3	829.3 ± 66.5	923.4 ± 22.0	585.1 ± 31.8	581.5 ± 68.8	591.3 ± 65.5
	Walk-Back.	752.5 ± 24.8	722.1 ± 33.5	729.6 ± 39.2	700.4 ± 23.7	418.5 ± 94.3	712.2 ± 18.7	369.9 ± 13.1	311.5 ± 49.8	311.2 ± 48.4
Terrain Hopper	Hop	342.0 ± 35.2	301.3 ± 42.1	278.7 ± 43.0	267.6 ± 48.6	222.8 ± 23.1	345.5 ± 29.2	8.6 ± 7.4	9.1 ± 6.9	10.2 ± 7.4
	Hop-Back.	330.7 ± 24.9	284.3 ± 27.6	299.1 ± 26.6	285.6 ± 36.6	204.1 ± 27.3	324.0 ± 41.7	2.9 ± 5.6	2.7 ± 3.1	12.0 ± 13.3
	Stand	639.8 ± 68.3	699.0 ± 76.4	661.9 ± 51.5	625.0 ± 81.0	507.2 ± 89.7	656.6 ± 82.1	9.0 ± 7.2	25.7 ± 27.3	18.0 ± 10.2

Table 1: Robustness evaluation: $\mathrm{CVaR}_{0.1}$ of policies evaluated on 100 randomly sampled environments.

→ WAKER improves robustness of policies a range of downstream tasks

Reward-Free Curricula for Training Robust World Models

Marc Rigter, Minqi Jiang, Ingmar Posner