Towards Safe Reinforcement Learning via OOD Dynamics Detection

Arnaud Gardille [1] Ola Ahmad [2]

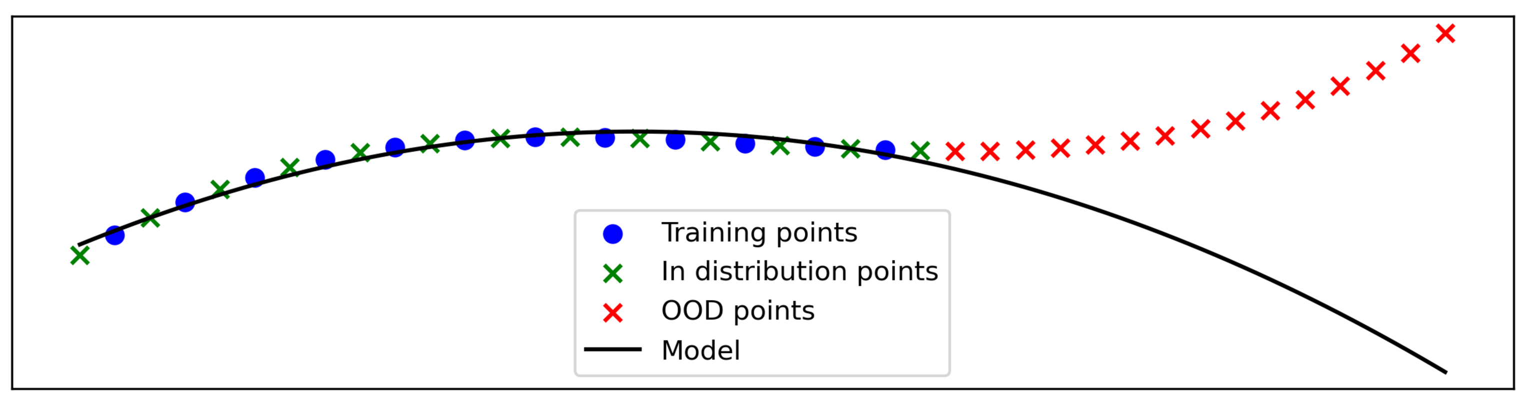
[1]Paris-Saclay University, France [2]Thales Digital Solutions, Montreal, Canada

Why is reinforcement learning rarely used in industry?

- Able to achieve goals in complex environments.
- Difficult to train.
- Lack of interpretability
- Risk of OOD utilisation.

OOD: The intrinsic limitation of machine learning

- ML algorithms fail **out of their training distribution (OOD)**.



In the context of RL: **OOD dynamic**: *when the transitions of the deployment environment differ from those of the agent training environment.*

→ An **effective method to quickly detect OOD dynamic** appears to be a prerequisite to the **use of RL in safety-critical systems**.

Comparison to existing methods:

Deep learning methods can estimate the agent **network's confidence**, which can be used as an OOD metrics [1]. However:

- Require a particular intervention of the agent's network.
- Fail facing **temporally correlated data** *(iid assumption)*.

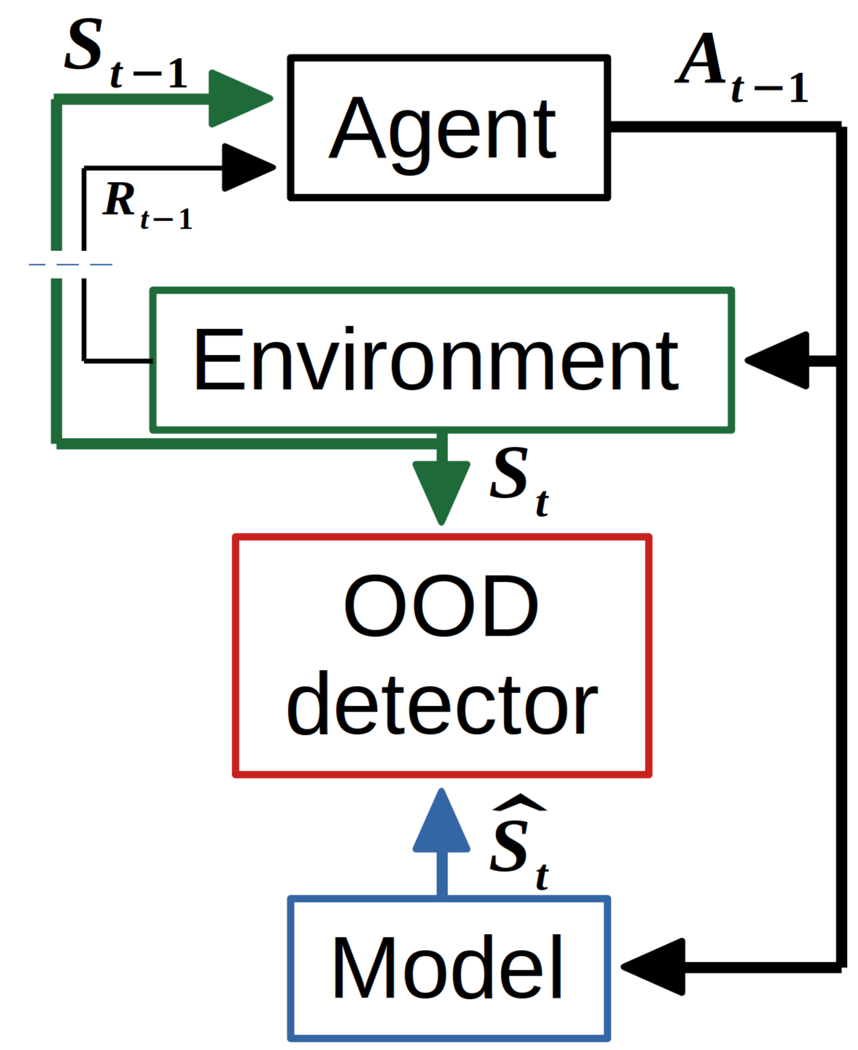
Assumptions

- Environments modeled as MDP: $\mathcal{T}(S_{t-1}, A_{t-1}) = S_t$.
- We have a model $\mathcal{M} \approx \mathcal{T}_{\text{train}}$ of the training environment.
- Prediction error: $\mathcal{M} - \mathcal{T}_{\text{train}} \sim \mathcal{N}(0, \cdot)$
- $\mathcal{M}$ will be a biased estimator of $\mathcal{T}_{\text{OOD}}$.

Proposed method

- Sample models estimation: $M(S_{t-1}, A_{t-1}) = \hat{S}_t$
- Observe real transition: $\mathcal{T}(S_{t-1}, A_{t-1}) = S_t$

Then **update the statistical test**:



Method architecture

Test whether (τ): $\mathcal{M} - \mathcal{T}_{\text{train}} \sim \mathcal{N}(0, \widehat{\sigma})$

$$\tau = \frac{\hat{\mu}}{\widehat{\sigma}/\sqrt{T}} \tag{1}$$

Improvement: train $\mathcal{M}'(S_{t-1}, A_{t-1}) \approx \sigma(S_{t-1}, A_{t-1})$
→ Replace $\widehat{\sigma}$ with $\mathcal{M}'$

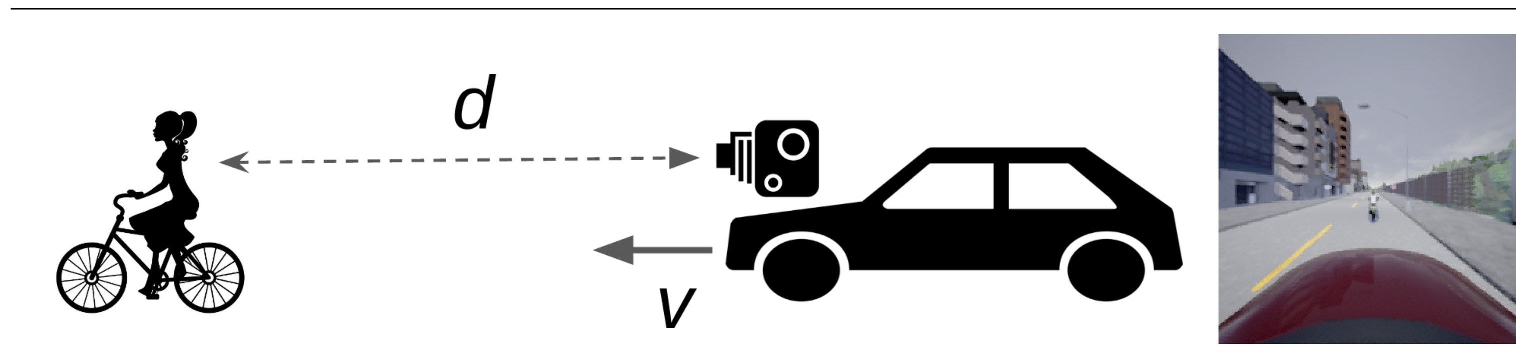
What to use as a model of the training environment?

- If possible, use $\mathcal{T}_{\text{train}}$.

We need to be able to sample from any (S_{t-1}, A_{t-1}).

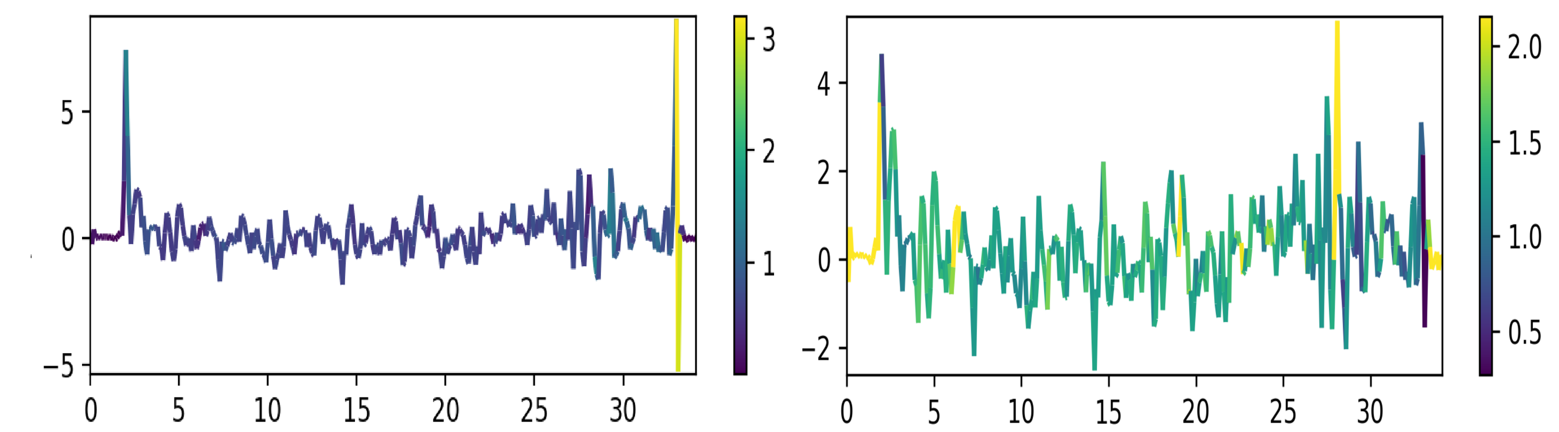
- Train a supervised model $M(S_{t-1}, A_{t-1}) \approx S_t$ using an in-distribution trajectory datasets.

Experimental Results



Emergency breaking scenario on a realistic autonomous driving simulator:

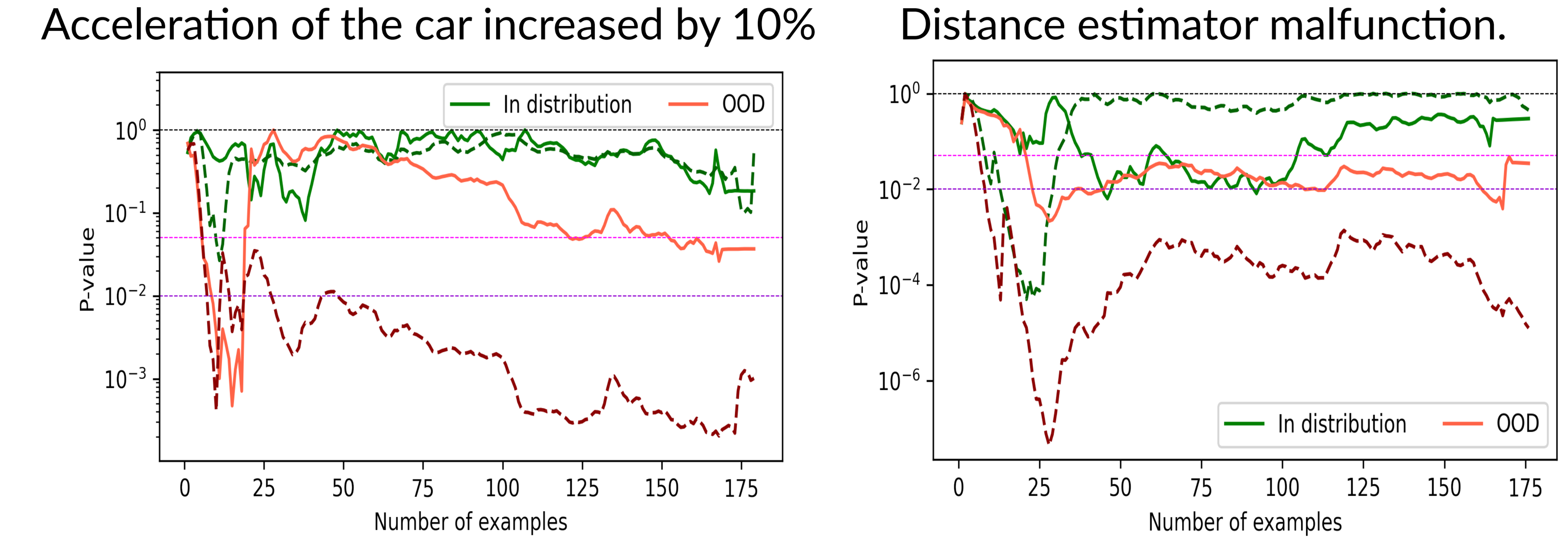
- **State**: speed v_t and perceived distance to the frontal obstacle d_t.
- **Action**: intensity of beaking or acceletation.
- **Reward** = right speed + huge penalty if collision.
- Trained **models**: MLP trained with MSE for regression.
- Distance estimator: ResNet18 fin-tuned for regression.



Model's error in speed prediction normalized globally (left) and locally (right).

Impact of local normalization on the expressiveness of the error:

Acceleration of the car increased by 10% Distance estimator malfunction.



Evolution of the p-values as a function of sample size T:

Conclusion and Discussions

- ✔ Dynamic OOD detected with high confidence.
- ✔ Method completely independent from the agent.
- ✗ Requires a model of the environment.
- ✗ Stability should be improved.

Perspectives

- Investigate replacing Student's t-test by the martingales method of [1]11.
- Other agents may choose among several decisions.
  → predict a gaussian mixture!
- Explore a decision transformer as an integrated model.
- How to create efficient models for high-dimensional states?
- Evaluate our method on the benchmark proposed in [4].

References

[1] Feiyang Cai and Xenofon Koutsoukos. Real-time out-of-distribution detection in learning-enabled cyber-physical systems, 2020. URL `https://arxiv.org/abs/2001.10494`.

[2] Jianyu Chen, Zhuo Xu, and Masayoshi Tomizuka. End-to-end autonomous driving perception with sequential latent representation learning. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[3] Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks, 2018. URL `https://arxiv.org/abs/1802.04865`.

[4] Aaqib Parvez Mohammed and Matias Valdenegro-Toro. Benchmark for out-of-distribution detection in deep reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021. URL `https://openreview.net/forum?id=bvC9rzKqi1b`.

arnaud.gardille@gmail.com