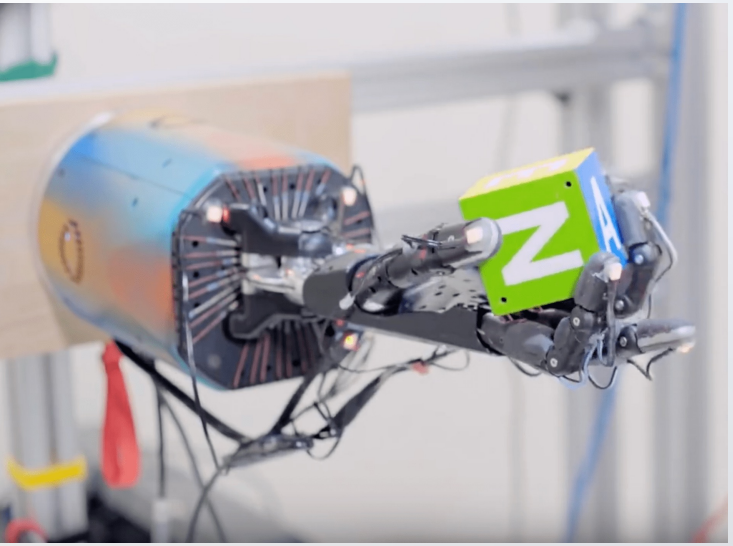


Physics-**I**nformed Model and **H**ybrid **P**lanning for Efficient Dyna-Style Reinforcement Learning

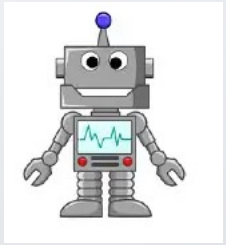
Zakariae El Asri **Olivier Sigaud** **Nicolas Thome**

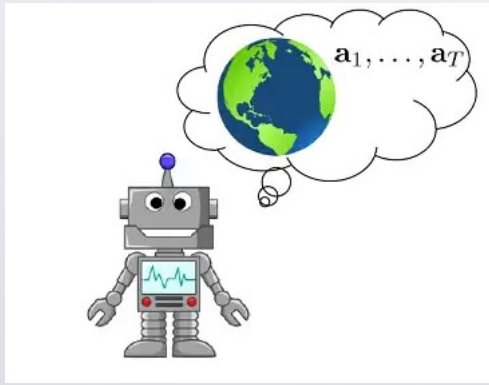
Sorbonne Université, CNRS, ISIR, Paris, France

Introduction

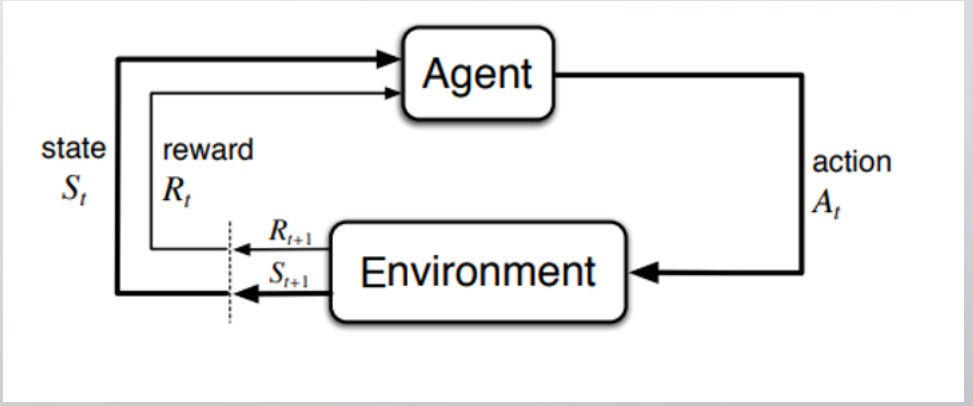
Markov Decision Process (MDP) : $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$

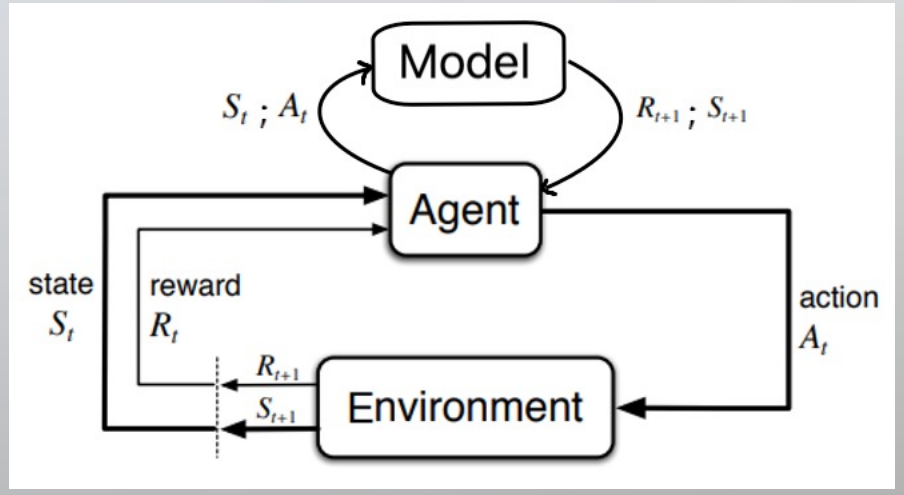
Objective in RL: maximize $\sum_{t=t_0}^{\infty} \gamma^{t-t_0}.r_t$



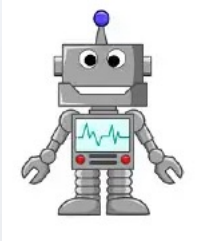


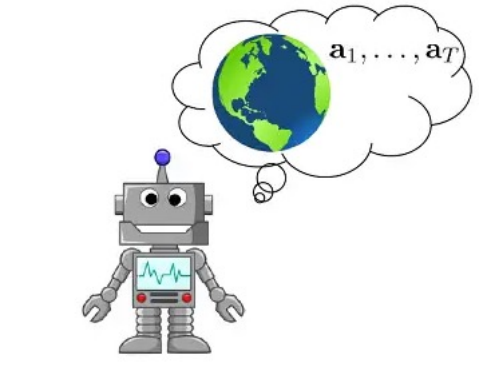
Model-Free RL Vs Model-Based RL





Introduction

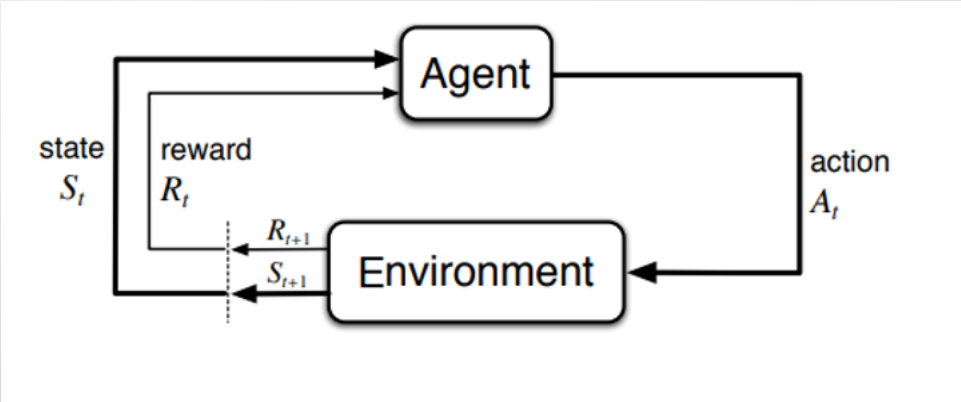


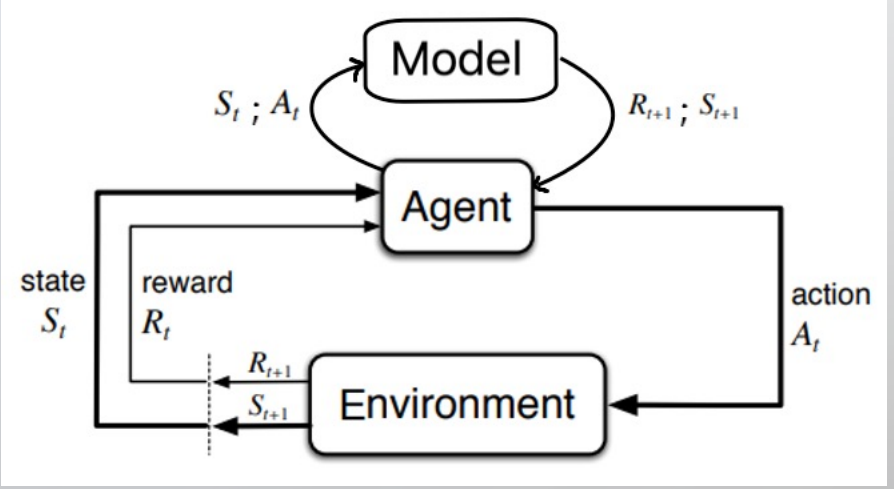


Model-Free RL

Vs

Model-Based RL





- Asymptotic performance
- Sample efficiency
- Time efficiency

- Asymptotic performance
- Sample efficiency
- Time efficiency

Introduction

GOAL

Asymptotic performance

Sample efficiency

Time efficiency

Model-Free RL

- Asymptotic performance
- Sample efficiency
- Time efficiency

Model-Based RL

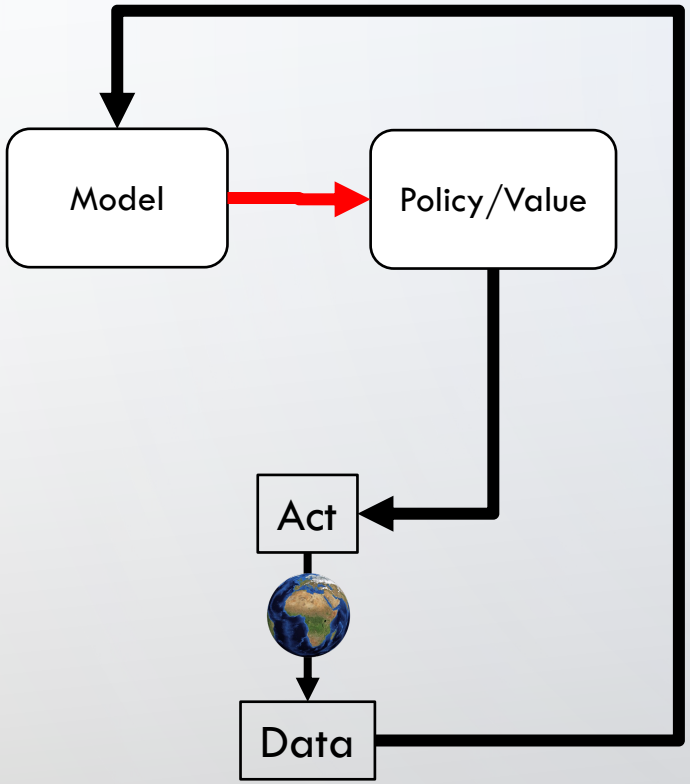
- Asymptotic performance
- Sample efficiency
- Time efficiency

PhIHP (our method)

- Asymptotic performance
- Sample efficiency
- Time efficiency

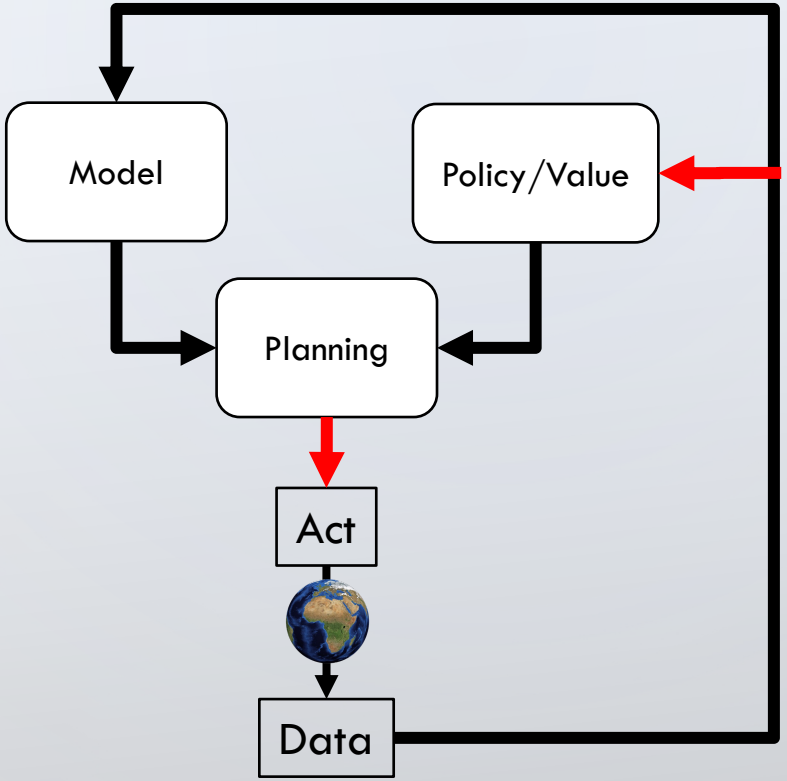
Introduction

Dyna style RL (LOOP [1])



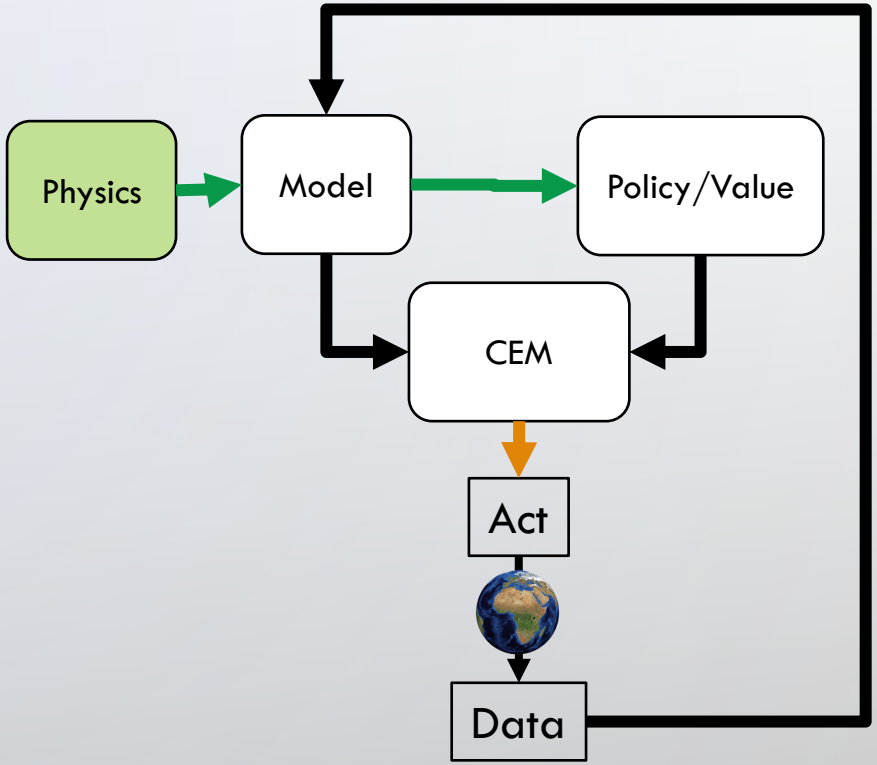
- Asymptotic performance
- Sample efficiency
- Time efficiency

Hybrid RL (TD-MPC [2])



- Asymptotic performance
- Sample efficiency
- Time efficiency

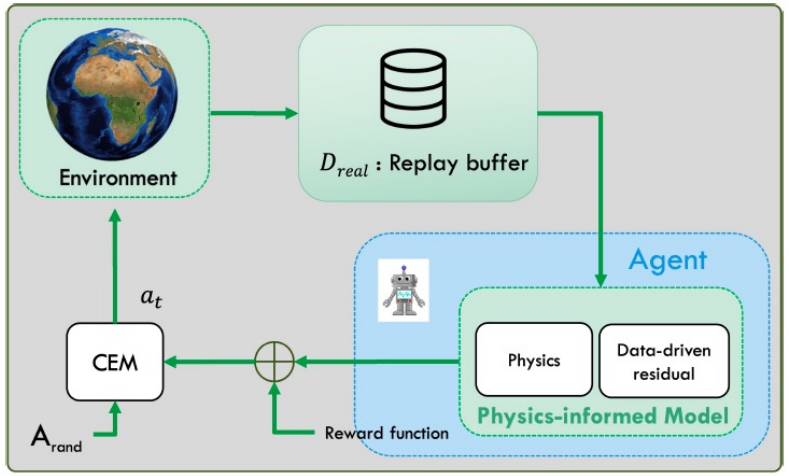
PhIHP (Ours)



- Asymptotic performance
- Sample efficiency
- Time efficiency

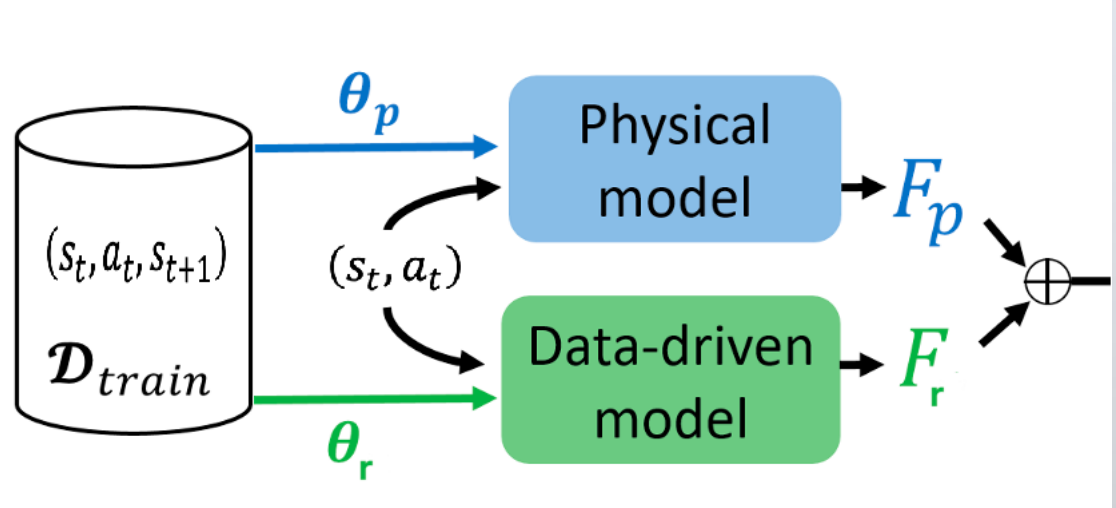
[1] Harshit Sikchi, Wenxuan Zhou, and David Held. Learning off-policy with online planning. CoRL 2022.
[2] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. ICML, 2022.

Our method – Leveraging Physics in Model Learning



(a) Learn a physics-informed model

Training Strategy:



$$\left.\frac{\mathrm{d}\hat{s}_t}{\mathrm{d}t}\right|_{t=t'} = (F_p^{\theta_p} + F_r^{\theta_r})(s_t, a_{t'}) \rightarrow \text{ODE Solver} \rightarrow \hat{s}_{t+1}$$

$$\text{Loss} = \sum_i \|f(s_i, a_i) - s_i'\|^2 + \lambda\|F_a\| \quad \text{s.a} \quad f(s_i, a_i) = (F_a + F_p)(s_i, a_i)$$

The quality of the physics-informed model

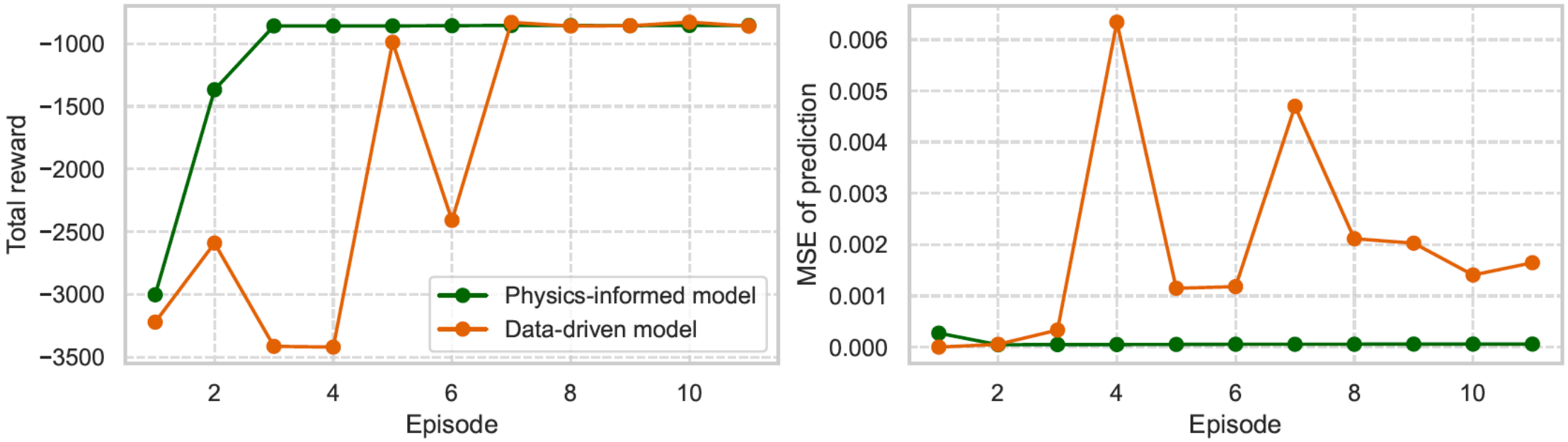
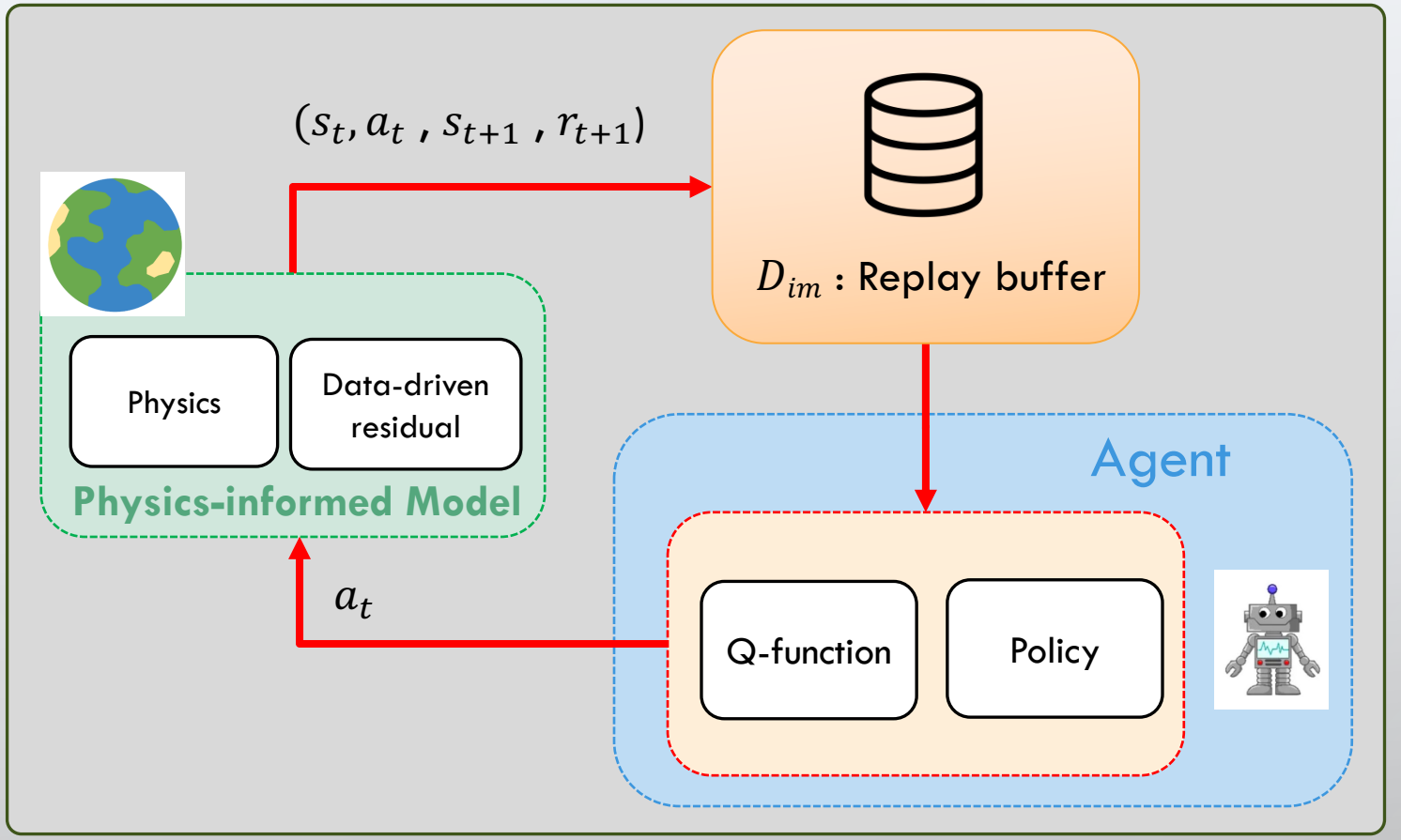


Figure 6: A data-driven model still poorly predicts the next states even when its asymptotic performance matches that of the physics-informed model. Figure obtained with 10 episodes of model training on Pendulum swingup.

Our method – Policy learning through Imagination

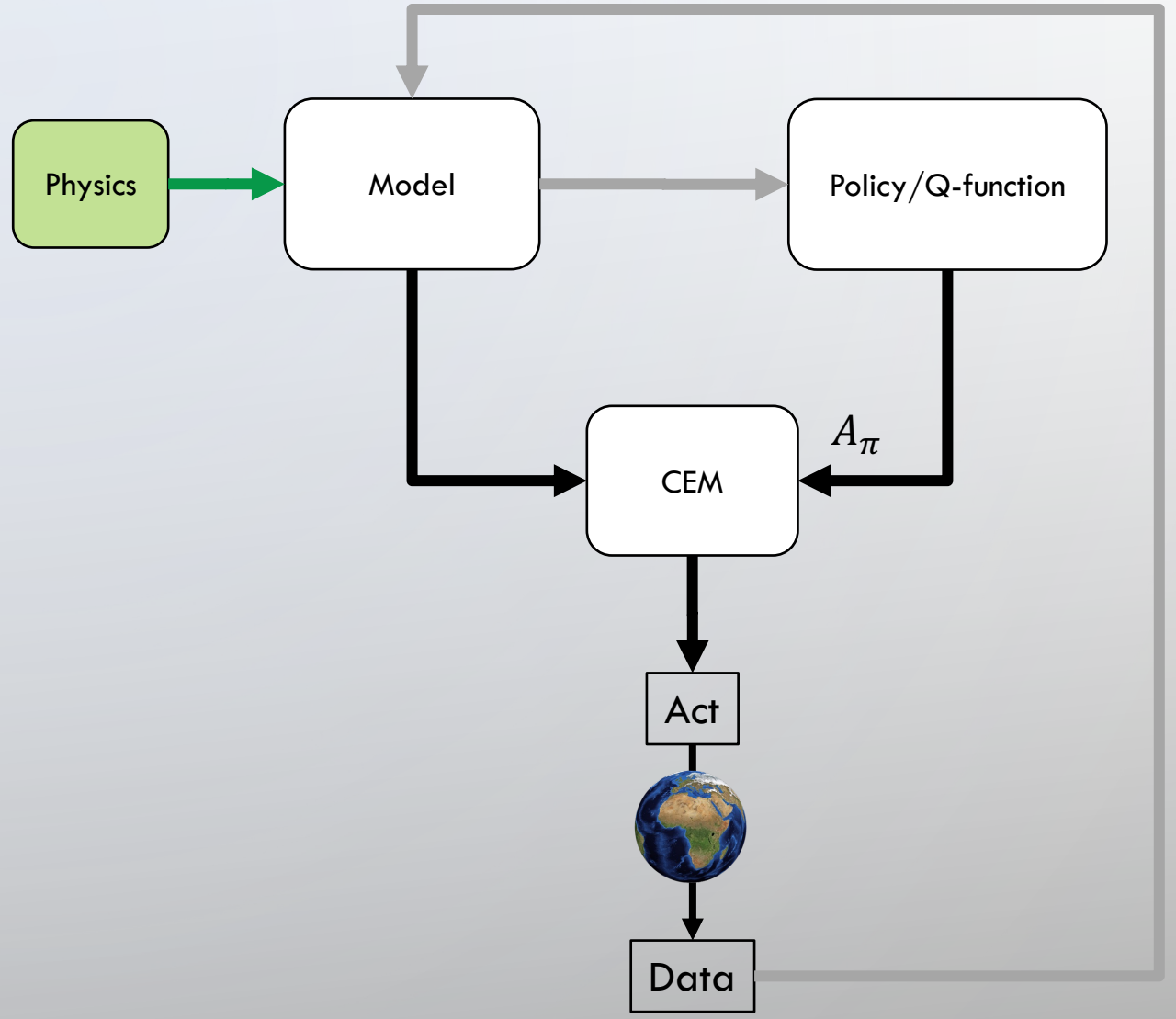


Our method – Hybrid planning

$$A^* = \arg\max_{A \in \mathcal{A}^H} \Big(\sum_{t=t_0}^{H} \gamma^{t-t_0} R(s_t, a_t) + \alpha \cdot \gamma^{H-t_0} Q(s_H) \Big)$$

local solution

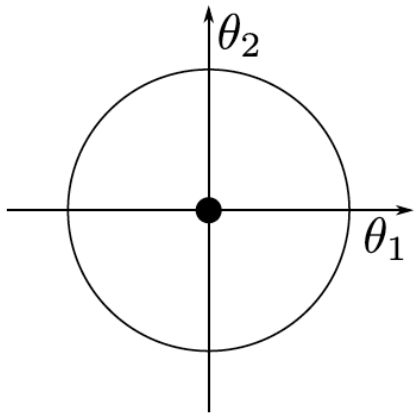
long-term reward



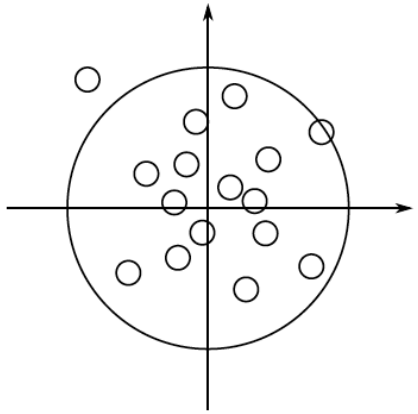
A_π: Informative candidates,
reduce population size and iterations in CEM

Q-function: reduce the planning horizon H

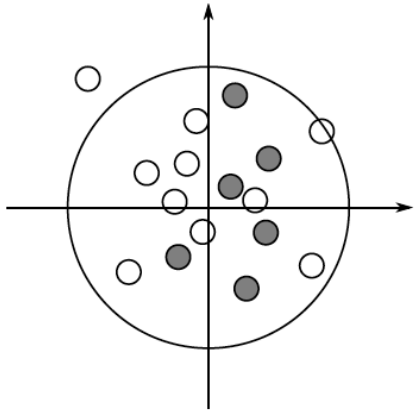
Reminder: Cross-Entropy Method (CEM)



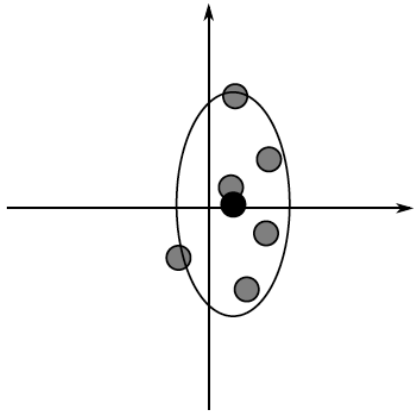
1.Start with the normal distribution $N(\mu,\sigma^2)$



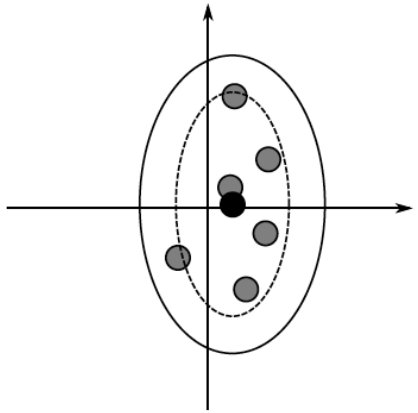
2. Generate N vectors with this distribution



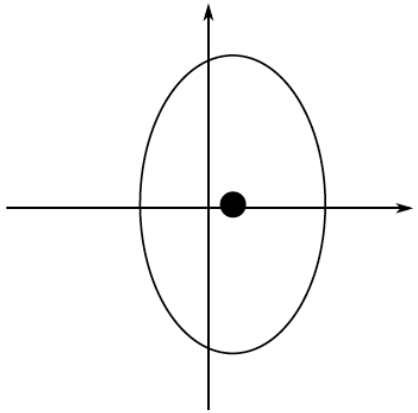
3. Evaluate each vector and select a proportion ρ of the best ones. These vectors are represented in grey



4. Compute the mean and standard deviation of the best vectors



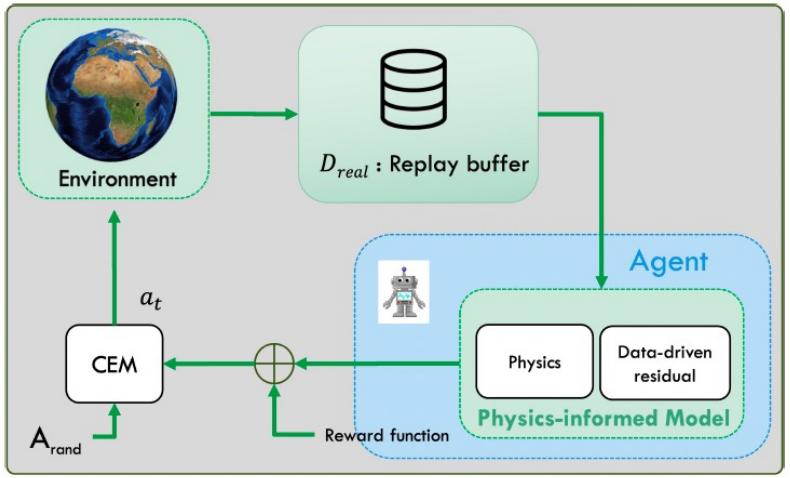
5. Add a noise term to the standard deviation, to avoid premature convergence to a local optimum



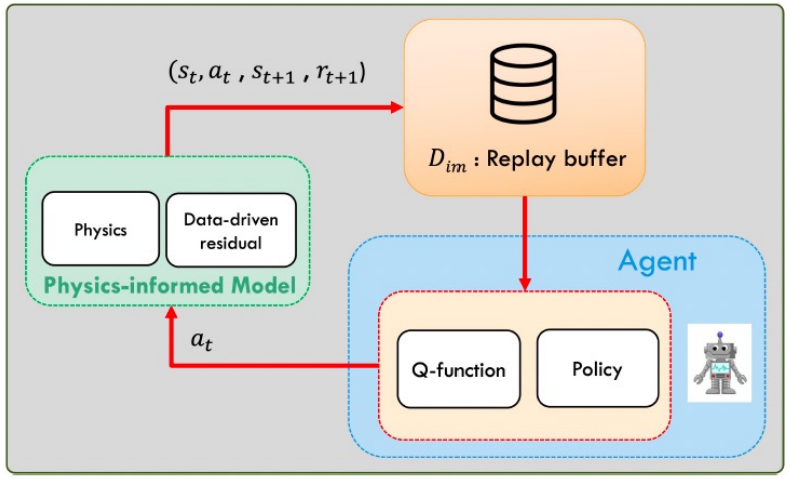
6. This mean and standard deviation define the normal distribution of next iteration

Marin, and Sigaud, O. Towards fast and adaptive optimal control policies for robots: A direct policy search approach, Robotica 2012

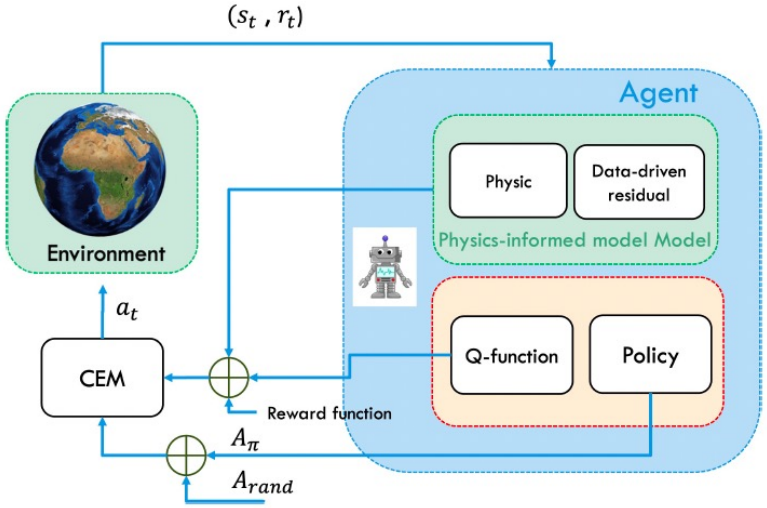
Our method - Summary



(a) Learn a physics-informed model



(b) Learn an actor/critic offline



(c) Behaviour at inference time

- Sample efficiency
- Reduced bias

- Sample efficiency
- Time efficiency

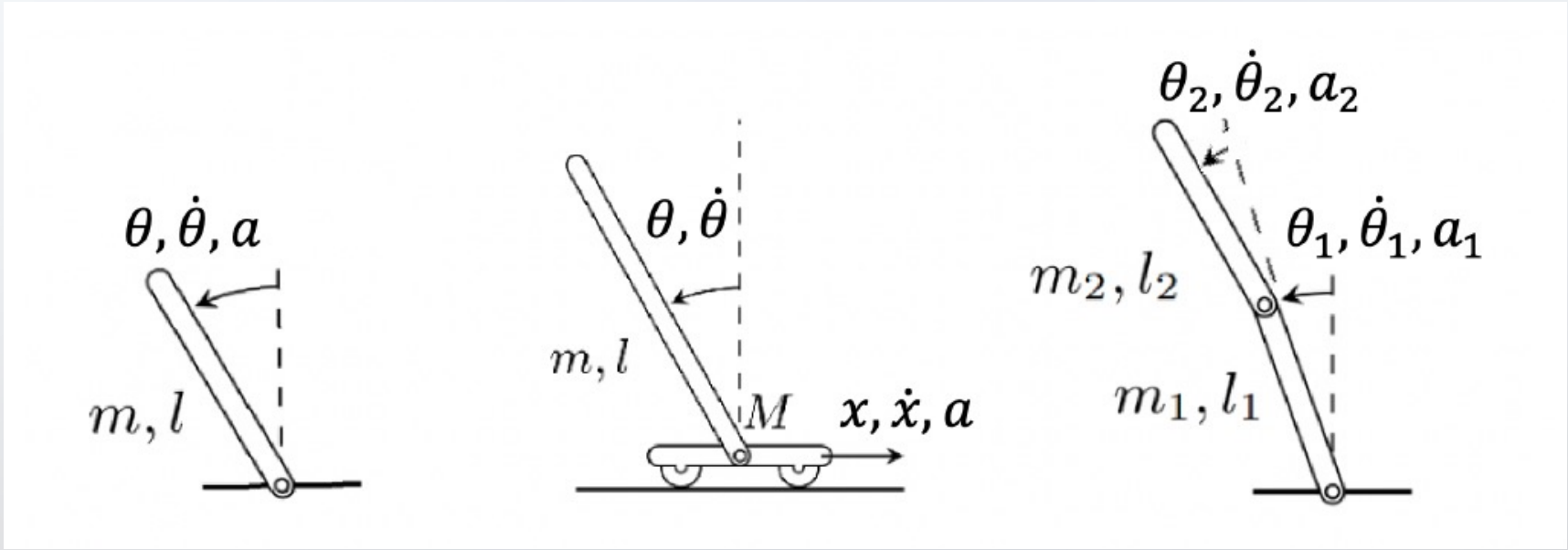
- Sample efficiency
- Time efficiency
- Asymptotic performance

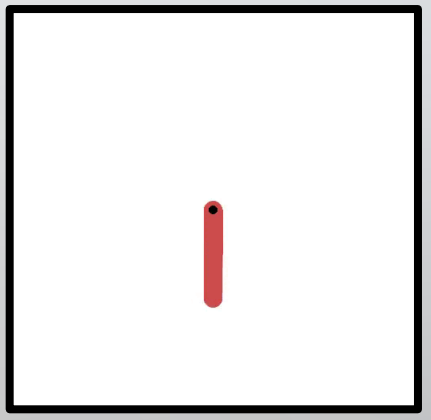
Environments

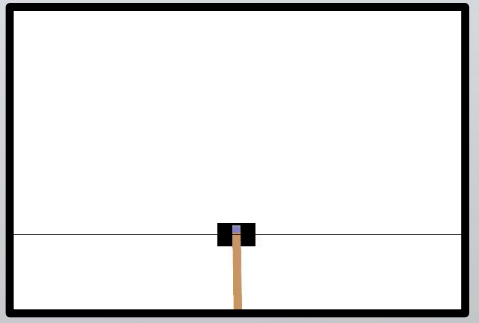
Pendulum

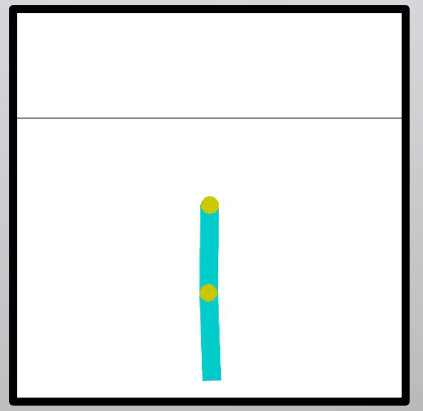
CartPole

Acrobot

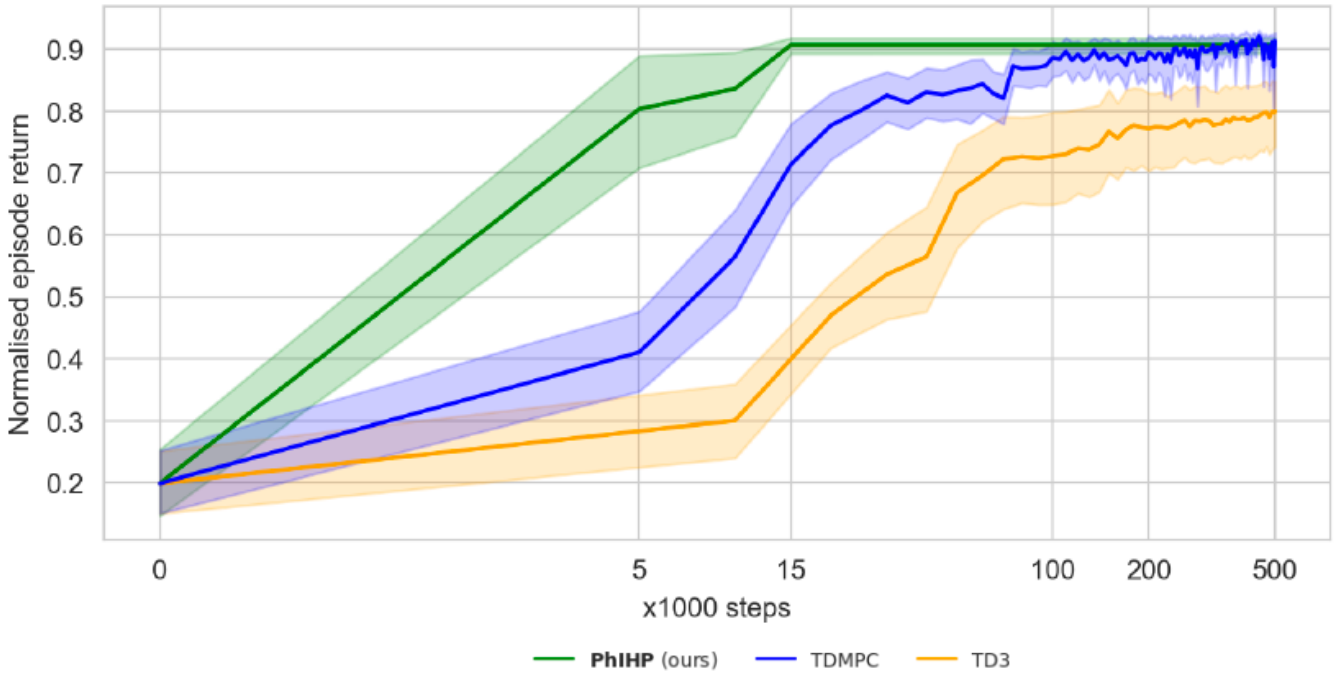








Results – Comparison to the state of the art



(a) Learning curves, the x-axis uses a symlog scale.

Figure 3: Comparison of PhIHP *vs* baselines aggregated on 6 control tasks (10 runs). a) PhIHP shows excellent sample efficiency and better asymptotic performance.

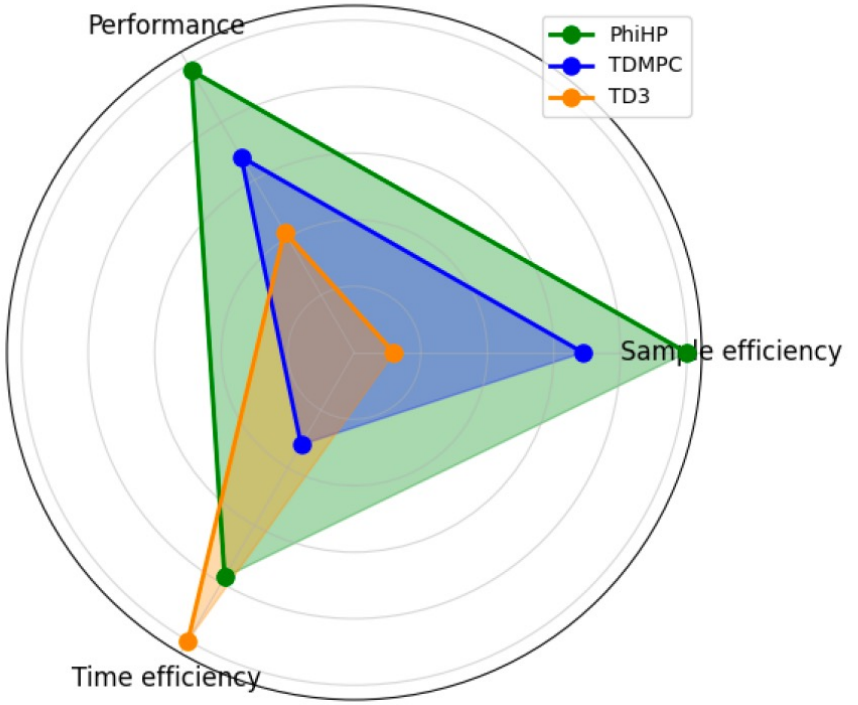


Figure 1: PhIHP includes a Physics-Informed model and hybrid planning for efficient policy learning in RL. PhIHP improves the compromise over state-of-the-art methods, model-free TD3 and hybrid TD-MPC, between sample efficiency, time efficiency, and performance. Results averaged over 6 tasks (Towers et al., 2023).

Ablation study – Impact of learning through imagination & hybrid planning

PhIHP

PhIHP w/o "imagination"

PhIHP w/o "imagination and physics"

PhIHP w/o "policy"

PhIHP w/o "policy and physics"

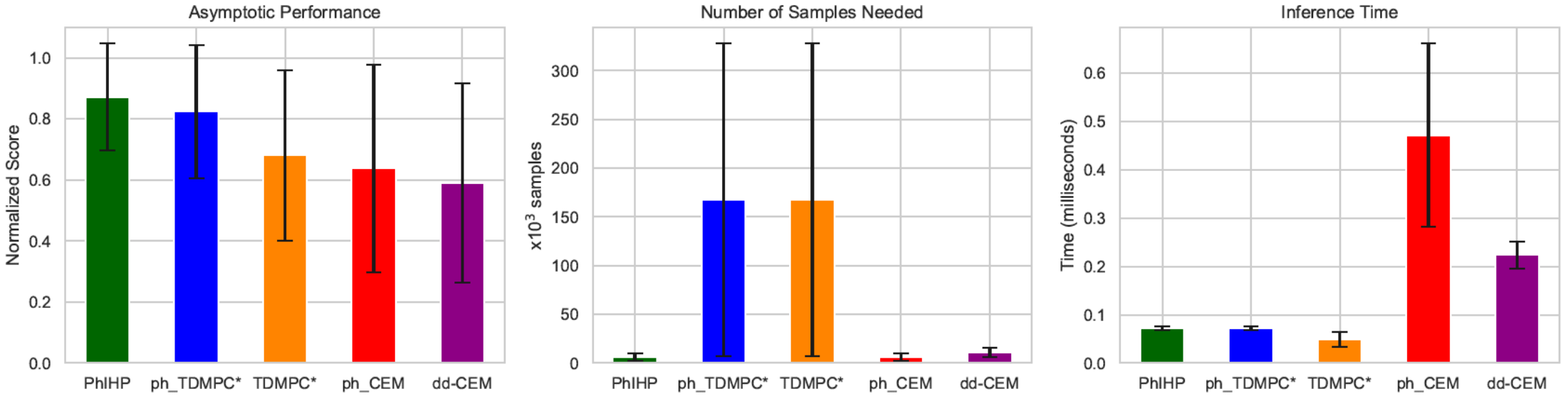


Figure 5: Comparison of PhIHP and its variants on the 3 main metrics. The figures illustrate the aggregated results of running all algorithms on 6 classic control tasks. Histograms and bars represent mean and std. over 10 runs.

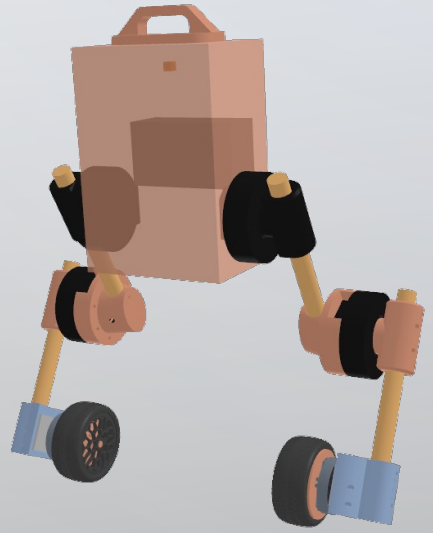
Conclusion

Due to Physics prior, PhiHP achieves the best compromise

GOAL
Asymptotic performance
Sample efficiency
Time efficiency

Perspectives: Applying PhiHP to Upkie[1]:

More challenging control tasks.



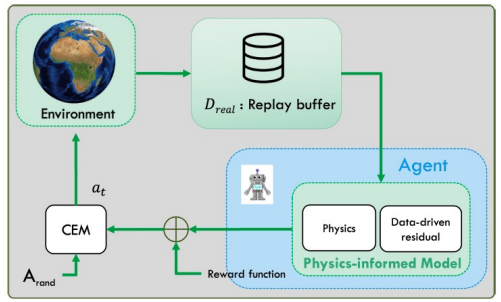
A real robotic application

[1] Stéphane Caron, Nicolas Perrin-Gilbert, Viviane Ledoux, Ü Bora Gökbakan, Pierre-Guillaume Raverdy, and Antonin Raffin. Upkie open source wheeled biped robot, 2024.

Thank you for your attention

Appendix

Our method – Process of Model learning



(a) Learn a physics-informed model

$D = \{(s_t, a_t, s_{t+1})\}$

Random actions

Fit $f(s,a)$

Plan with f H

Execute A_{best} 1

Append (s_t, a_t, s_{t+1}) on D

MPC with CEM