Explorative Imitation Learning

A Path Signature Approach for Continuous Environments

Nathan Gavenski[1]
Juarez Monteiro, Felipe Meneguzzi[2], Michael Luck[3] and Odinaldo Rodrigues[1]

King's College London [1]
University of Aberdeen [2]
Pontifical Catholic University of Rio Grande do Sul [2]
University of Sussex [3]

Table of contents

Introduction

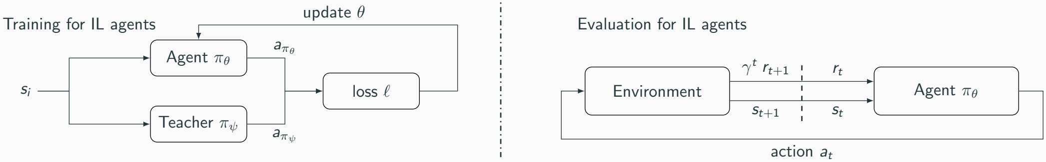
Motivation

- Humans and animals learn from watching others perform a set of actions[1]
- It is more practical for us to reuse prior knowledge in new domains through demonstration than starting fresh without any teacher[2]
- Requiring human intervention for environment-specific tasks can be unfeasible and complicate the process of reusing prior knowledge

[1] Bandura, A. *Social Learning Theory* in Englewood Cliffs (1997)
[2] Rizzolatti, G. and Sinigaglia, C. *The Functional Role of The Parieto-Frontal Mirror Circuit: Interpretations And Misinterpretations* in Nature Reviews (2010)

Imitation Learning

Imitation Learning training and evaluation procedures[3]



Objective: Minimise the loss between agent and expert actions:

$$\arg\min_\theta \sum_{\tau \in \mathcal{T}} \sum_{s \in \tau} \ell(\pi_\psi(s), \pi_\theta(s)).$$

[3] Gavenski et al. *A Survey of Imitation Learning Methods, Environments and Metrics* (2024)

Imitation Learning from Observation

If we assume we do not have access to the expert actions, we need to change the objective function:

$$\arg\min_{\theta} \mathbb{E}_{s_t, s_{t+1} \sim \mathcal{T}_{\pi_\psi}} \ell(s_{t+1}, T(s_t, \pi_\theta(s_t))),$$

Approach: Model the environment with forward or inverse dynamic models, inverse reinforcement learning, or adversarial imitation learning.

Imitation Learning from Observation

If we assume we do not have access to the expert actions, we need to change the objective function:

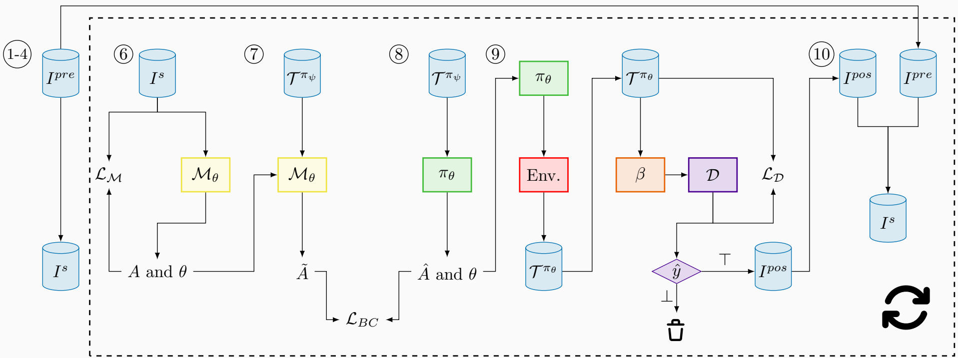
$$\arg\min_{\theta} \mathbb{E}_{s_t, s_{t+1} \sim \mathcal{T}_{\pi_\psi}} \ell(s_{t+1}, T(s_t, \pi_\theta(s_t))),$$

Approach: Model the environment with forward or **inverse dynamic models**, inverse reinforcement learning, or adversarial imitation learning.

Explorative Imitation Learning

Continuous Imitation Learning from Observation (CILO)

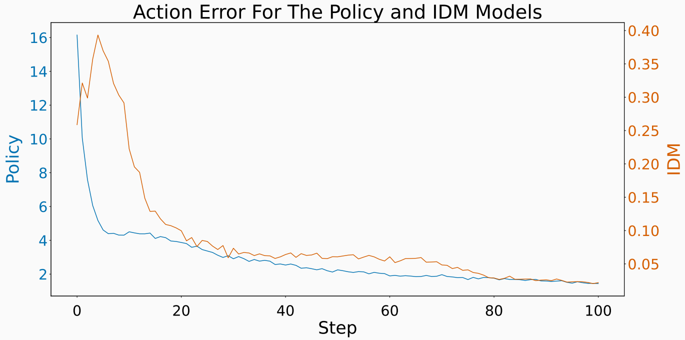
Training Procedure



- Exploration ratio naturally decreases with models' performance
- Sample efficient from appending new samples to its dataset
- Remains goal-aware without any human intervention

Exploration Mechanism

The exploration mechanism relies on the error from the π_θ when using samples from the environment and the $\mathcal{M}_\theta$ error during self-supervision.

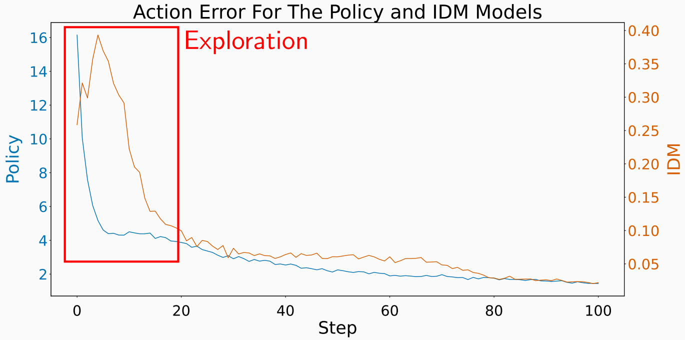


$$\tilde{a}_{\mathbb{M}_\theta} = \frac{1}{\varepsilon\sqrt{2\pi}} e^{-\frac{(s_t^e - \mathbb{M}_\theta(S))}{2\varepsilon^2}}$$

$$\varepsilon = \|a - \hat{a}\|^p$$

Exploration Mechanism

When the error is **high** it acts as an exploration phase, where the models can diverge **more** from the initial prediction

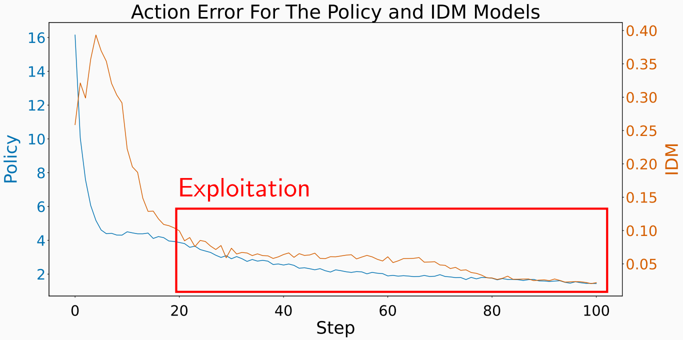


$$\tilde{a}_{\mathbb{M}_\theta} = \frac{1}{\varepsilon\sqrt{2\pi}} e^{-\frac{\left(s_t^e - \mathbb{M}_\theta(S)\right)}{2\varepsilon^2}}$$

$$\varepsilon = \left\| a - \hat{a} \right\|^p$$

Exploration Mechanism

And when the error is **small** it acts as an exploitation phase, where the models can diverge **less** from the initial prediction

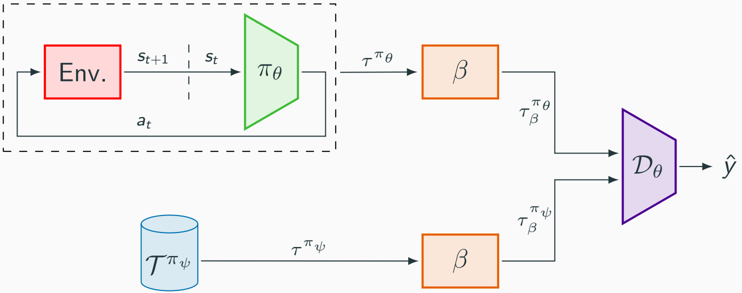


$$\tilde{a}_{\mathbb{M}_\theta} = \frac{1}{\varepsilon\sqrt{2\pi}} e^{-\frac{(s_t^e - \mathbb{M}_\theta(S))}{2\varepsilon^2}}$$

$$\varepsilon = \|a - \hat{a}\|^p$$

Goal-Aware Function

- CILO uses path-signatures[4] β as a deterministic encoding mechanism to represent different trajectories.



[4] For more information on path-signature, we refer to our supplementary material.

Goal-Aware Function

- We assume the expert **always** reaches the goal
- Include in the training dataset **agent's** trajectories that the discriminator classifies as being from the expert
- This allows the expansion of the initial dataset with additional trajectories that are **most similar** to the expert's
- Even though the discriminator might not be optimal, resulting in dissimilar trajectories being added, it allows for trajectories that are **better** than the initial **random** ones;

Experimental Results

Results

Comparison with the state-of-the-art in MuJoCo environments.

Algorithm	Metric	Ant	Pendulum	Swimmer	Hopper	HalfCheetah
Random	AER	-65.11 ± 106.16	5.70 ± 3.26	0.73 ± 11.44	17.92 ± 16.02	-293.13 ± 82.12
	$\mathcal{P}$	0	0	0	0	0
Expert	AER	5544.65 ± 76.11	1000 ± 0	259.52 ± 1.92	3589.88 ± 2.43	7561.78 ± 181.41
	$\mathcal{P}$	1	1	1	1	1
CILO	AER	$\mathbf{6092 \pm 801.2}$	$\mathbf{1000 \pm 0}$	$\mathbf{334.6 \pm 3.45}$	$\mathbf{3589 \pm 178.2}$	$\mathbf{7100.6434 \pm 90.1775}$
	$\mathcal{P}$	$\mathbf{1.0974 \pm 0.1372}$	$\mathbf{1 \pm 0}$	$\mathbf{1.2901 \pm 0.0128}$	$\mathbf{0.9998 \pm 0.0487}$	$\mathbf{0.9413 \pm 0.0115}$
OPOLO	AER	5508.6807 ± 930.7590	$\mathbf{1000 \pm 0}$	253.3297 ± 3.4771	3428.6405 ± 420.3285	7004.65 ± 568.66
	$\mathcal{P}$	0.9935 ± 0.1659	$\mathbf{1 \pm 0}$	0.9761 ± 0.0134	0.9549 ± 0.1177	0.9291 ± 0.0724
MobILE	AER	995.5 ± 25.65	111.7 ± 31.25	130.7 ± 24.36	2035 ± 192.95	4721.5 ± 364.5
	$\mathcal{P}$	0.1891 ± 0.0047	0.1066 ± 0.0313	0.5022 ± 0.0968	0.5647 ± 0.0531	0.5647 ± 0.0454
BCO	AER	1529 ± 980.86	521 ± 178.9	257.38 ± 4.28	1845.66 ± 628.41	3881.10 ± 938.81
	$\mathcal{P}$	0.2842 ± 0.1724	0.5675 ± 0.1785	0.9917 ± 0.0166	0.5177 ± 0.1765	0.5117 ± 0.1217

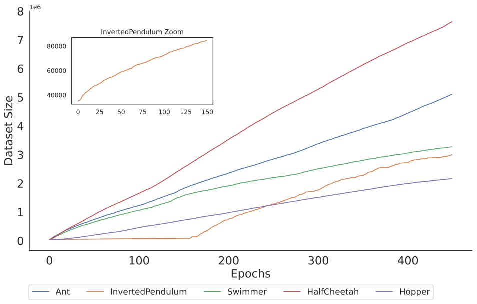
All datasets are available at `https://github.com/NathanGavenski/IL-Datasets`

Sample and Space Efficiency

Sample efficiency of CILO for Ant

Trajectories	AER	$\mathcal{P}$
1	1003 ± 1999	0.18
10	$\mathbf{6091 \pm 801.2}$	$\mathbf{1.1}$
100	6026 ± 725.86	1.09

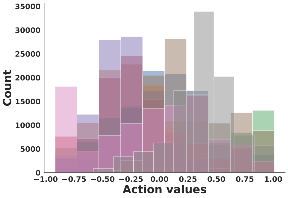
Growth of the dataset size.

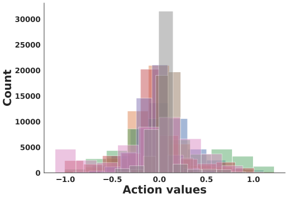


Size of $I^s \times$ epochs for all environments.

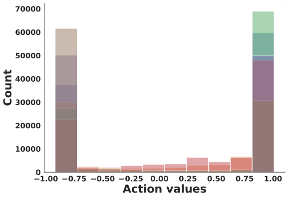
Different Distributions

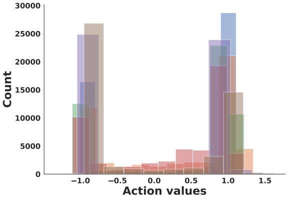
Ant





π_ψ π_θ





HalfCheetah

- Actions for 1st joint
- Actions for 2nd joint
- Actions for 3rd joint
- Actions for 4th joint
- Actions for 5th joint
- Actions for 6th joint
- Actions for 7th joint
- Actions for 8th joint

Conclusion

- CILO **does not require** prior domain knowledge or information about the expert's actions to learn a policy
- It has sample efficiency **superior or equal** to the state-of-the-art imitation learning from observation alternatives
- It implements new **model-agnostic mechanisms**, allowing them to be used in other IL methods
- It **approximates** (sometimes surpassing) expert performance

Questions?

GitHub Repo

nathangavenski.github.io
nathan.schneider_gavenski@kcl.ac.uk
https://github.com/NathanGavenski

King's College London
University of Aberdeen
University of Sussex
Safe & Trusted AI
UKRI Centre for Doctoral Training