SMoSE: Sparse Mixture of Shallow Experts
for Interpretable Reinforcement Learning in Continuous Control Tasks

Mátyás Vincze[1,2] Laura Ferrarotti[2] Leonardo Lucio Custode[1]
Bruno Lepri[2] Giovanni Iacca[1]

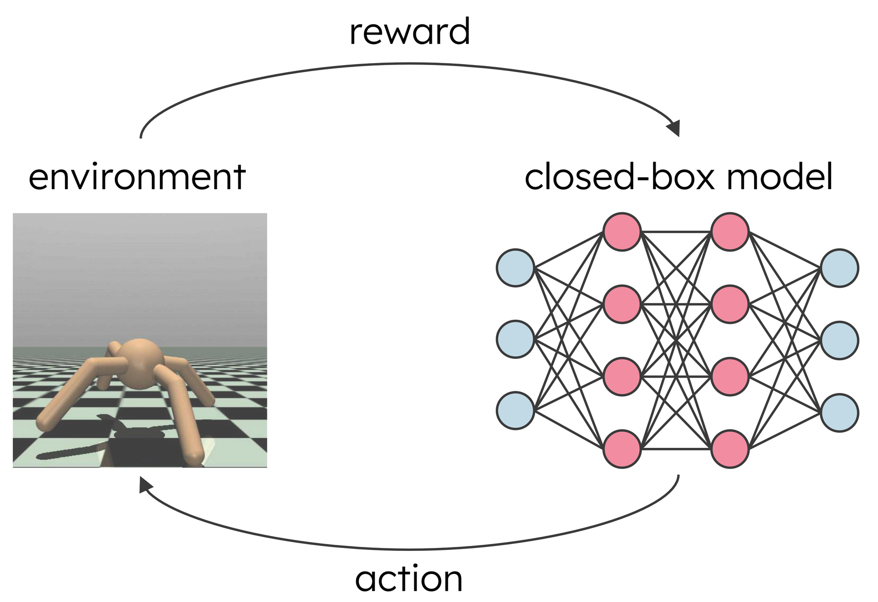
Università degli Studi di Trento [1] FONDAZIONE BRUNO KESSLER [2]

ICT DAYS

Motivation

Unlock safe and efficient RL

State-of-the-art approaches are not interpretable



- Scaling limits interpretability
- Explainability is not enough in most real-world use-cases
- Low-level interpretability is a must to ensure expected behavior

Interpretable approaches do not work in continuous control

- Most solutions require up 10x environment interactions
- No approach has comparable performance to state-of-the-art

paper code

@vinczematyas_ matyas-vincze

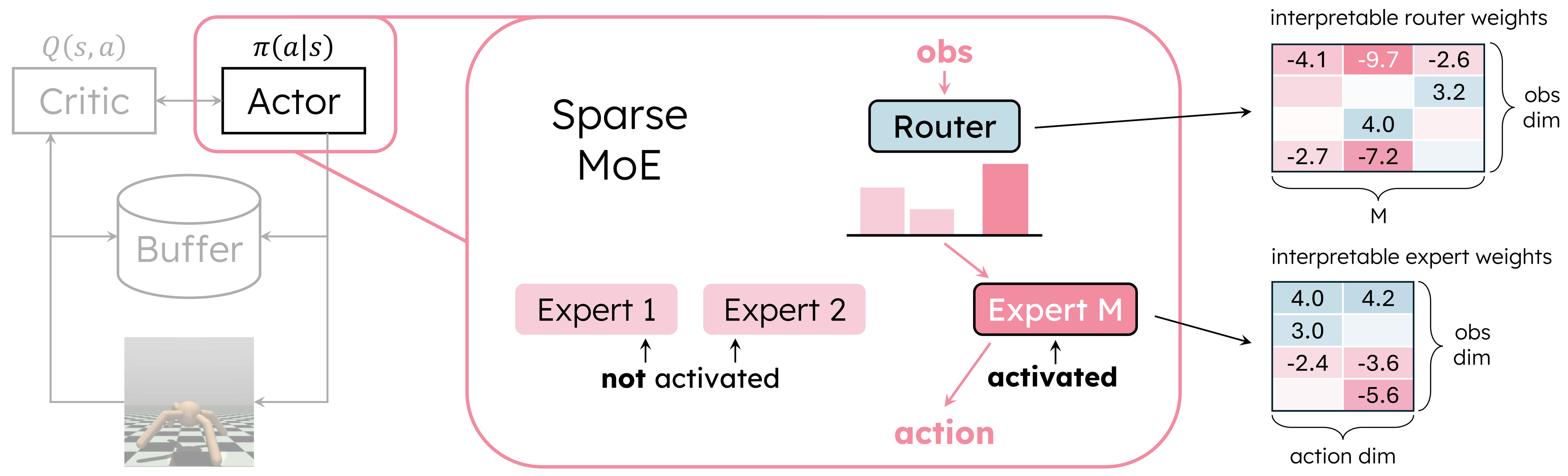
Method

Sparse MoE actor, Linear experts, Post-training distillation

Architecture : Linear router, linear experts

Router partitions the state space while the experts specialize on simple skills.

Per-expert capacity can be minimized as local policy decisions are very simple. The complex critic policy guides the actor to gather useful experience that is than used to learn the efficient policy in a few gradient steps.



Training stabilization

- Load balancing with auxiliary loss

$$\mathrm{L}_{\mathrm{aux}} = 0.1 * \left[\begin{array}{c} f_{\mathrm{imp}}(S) = \frac{1}{2}\left(\frac{\mathrm{std}(\mathrm{Imp}(S))}{\mathrm{mean}(\mathrm{Imp}(S))}\right)^2 \\ + \\ f_{\mathrm{load}}(S) = \frac{1}{2}\left(\frac{\mathrm{std}(\mathrm{Load}(S))}{\mathrm{mean}(\mathrm{Load}(S))}\right)^2 \end{array} \right]$$

- Forced expert-space exploration

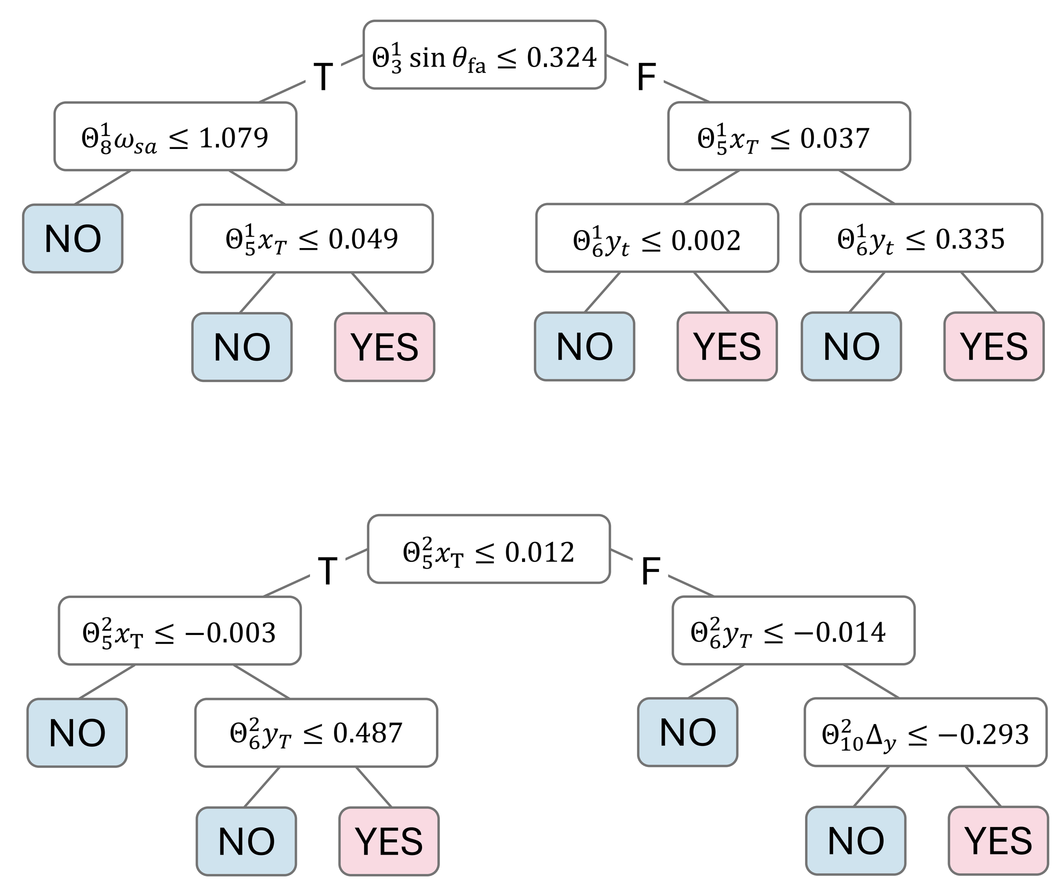
$$\varepsilon \sim \mathcal{N}(0, 1/M^2)$$

$$\mathrm{Load}_m(S) = \sum_{s_k \in S} \mathbb{P}(\varepsilon_{new} \geq \tau(s_k) - \pi_m(s_k))$$

$$\mathrm{Imp}_m(S) = \sum_{s_k \in S} \mathrm{softmax}(\pi_m(s_k|\theta_m, \sigma_m))$$

Router distillation

- Per-expert binary decision tree for “free”



Results

Strong performance on Mujoco tasks

Comparison with interpretable solutions

- Significantly better performance on Mujoco, except in environments where SAC already struggles
- Better sample-efficiency

	Walker2d	Hopper	Ant	HalfCheetah	Reacher	Swimmer
SAC-L	**4358.06**	2636.49	**5255.46**	**11809.87**	**-3.75**	68.59
SAC-M	4020.51	**3224.25**	4894.18	8992.22	-4.02	71.94
SAC-S	2967.14	3076.09	4162.97	7214.3	-4.82	59.42
PPO	3362.16	2311.9	2327.12	2308.29	-6.57	93.26
CGP	1090.00	1150.00	1130.00	6375.00	-68.50	**280.00**
LGP	1080.00	1120.00	1210.00	6388.50	-58.50	278.50
Metric-40	775.00	2005.00	2210.50	2210.50	x	x
Ours	**4224.29**	**2816.08**	**3245.43**	**7310.17**	**-5.49**	**45.4**

Comparison with closed-box solutions

- 99% less active actor parameters compared to SAC-L
- Performance is comparable on all environments
- Matched sample-efficiency