Learning Progress Driven Multi-Agent Curriculum

Wenshuai Zhao, Zhiyuan Li, Joni Pajarinen

July 2025

ICML
International Conference On Machine Learning

Aalto University

Outline:

- ❖ Background
- ❖ Motivation
- ❖ Method
- ❖ Experiments
- ❖ Conclusion

Background: Homotopy Optimization Methods

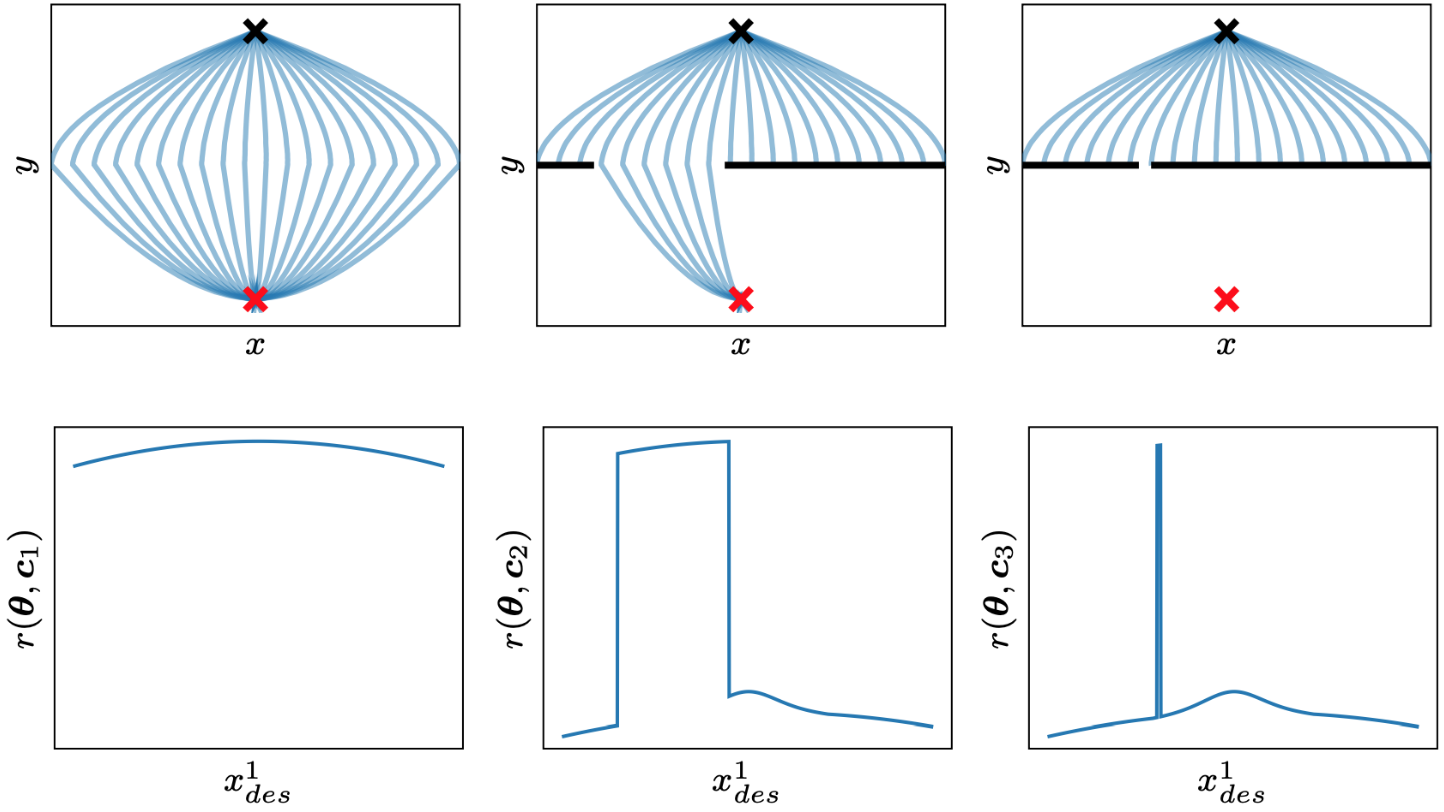


Figure 1: Gate task and visualization of the point-mass trajectories with their reward [1].

[1] Klink, Pascal, et al. "Self-paced contextual reinforcement learning." Conference on Robot Learning. PMLR, 2020.

Background: RL vs. Contextual RL

RL

$$\max_{\boldsymbol{\omega}} J(\boldsymbol{\omega}) = \max_{\boldsymbol{\omega}} \mathbb{E}_{p_0(\mathbf{s}_0), p(\mathbf{s}_{i+1}|\mathbf{s}_i, \mathbf{a}_i), \pi(\mathbf{a}_i|\mathbf{s}_i, \boldsymbol{\omega})} \left[\sum_{i=0}^{\infty} \gamma^i r(\mathbf{s}_i, \mathbf{a}_i) \right]$$

$$V_{\boldsymbol{\omega}}(\mathbf{s}) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s}, \boldsymbol{\omega})} \left[r(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{p(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \left[V_{\boldsymbol{\omega}}(\mathbf{s}') \right] \right]$$

Contextual RL

$$\max_{\boldsymbol{\omega}} J(\boldsymbol{\omega}, \mu) = \max_{\boldsymbol{\omega}} \mathbb{E}_{\mu(\mathbf{c})} \left[J(\boldsymbol{\omega}, \mathbf{c}) \right] = \max_{\boldsymbol{\omega}} \mathbb{E}_{\mu(\mathbf{c}), p_{0,\mathbf{c}}(\mathbf{s})} \left[V_{\boldsymbol{\omega}}(\mathbf{s}, \mathbf{c}) \right]$$

$$V_{\boldsymbol{\omega}}(\mathbf{s}, \mathbf{c}) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s}, \mathbf{c}, \boldsymbol{\omega})} \left[r_{\mathbf{c}}(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{p_{\mathbf{c}}(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \left[V_{\boldsymbol{\omega}}(\mathbf{s}', \mathbf{c}) \right] \right]$$

Background: SPRL

Self-paced reinforcement learning (SPRL) is one of SOTA curriculum reinforcement learning (CRL) method

$$\min_{\nu} \quad D_{\mathrm{KL}}(p(\mathbf{c}|\nu) \parallel \mu(\mathbf{c}))$$
$$\text{s.t.} \quad \mathbb{E}_{p(\mathbf{c}|\nu)}[J(\theta, \mathbf{c})] \geq V_{\mathrm{LB}} \text{ and } D_{\mathrm{KL}}(p(\mathbf{c}|\nu) \parallel p(\mathbf{c}|\nu')) \leq \epsilon$$

where $\mathbb{E}_{p(\mathbf{c}|\nu)}[J(\theta, \mathbf{c})]$ is the objective and maximized by

$$\max_{\nu_{k+1}} \frac{1}{M} \sum_{i=1}^{M} \frac{p(\mathbf{c}_i|\nu_{k+1})}{p(\mathbf{c}_i|\nu_k)} V_\theta(\mathbf{s}_{i,0}, \mathbf{c}_i)$$

The idea of generating tasks based on reward/ return/ value is shared in most existing single-agent CRL methods, such as *Goal-GAN* [2], *CURROT* [3]

[2] Florensa, Carlos, et al. "Automatic goal generation for reinforcement learning agents." International conference on machine learning. PMLR, 2018.
[3] Klink, Pascal, et al. "Curriculum reinforcement learning via constrained optimal transport." International Conference on Machine Learning. PMLR, 2022.

Background: Curriculum MARL

CRL for multi-agent learning (by controlling the number of agents as the curriculum context) is still in early stage, e.g. via prior knowledge.

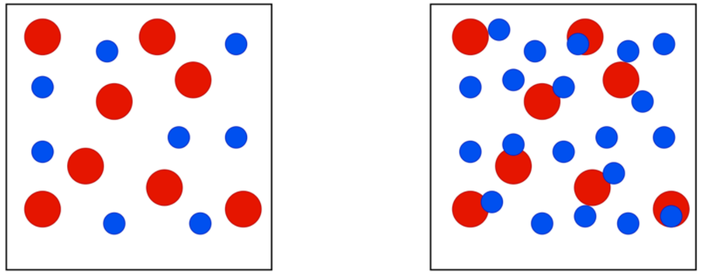
- *DyMA-CL* [4]: manually designed, from few to more.
- *EPC* [5]: in the order N⟶2N, with evolutionary selection.
- *VACL* [6]: in a presumed order to change number of agents.

- We abstract these works as a *Linear* baseline

[4] Wang, Weixun, et al. "From few to more: Large-scale dynamic multiagent curriculum learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.
[5] Long, Qian, et al. "Evolutionary population curriculum for scaling multi-agent reinforcement learning." arXiv preprint arXiv:2003.10423 (2020).
[6] Chen, Jiayu, et al. "Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems." Advances in Neural Information Processing Systems 34 (2021): 9681-9693.

Motivation

Two Issues of ***reward-based*** curriculum learning methods for multi-agent learning, when controlling the number of agents as curriculum



(a) 8 agents (b) 20 agents

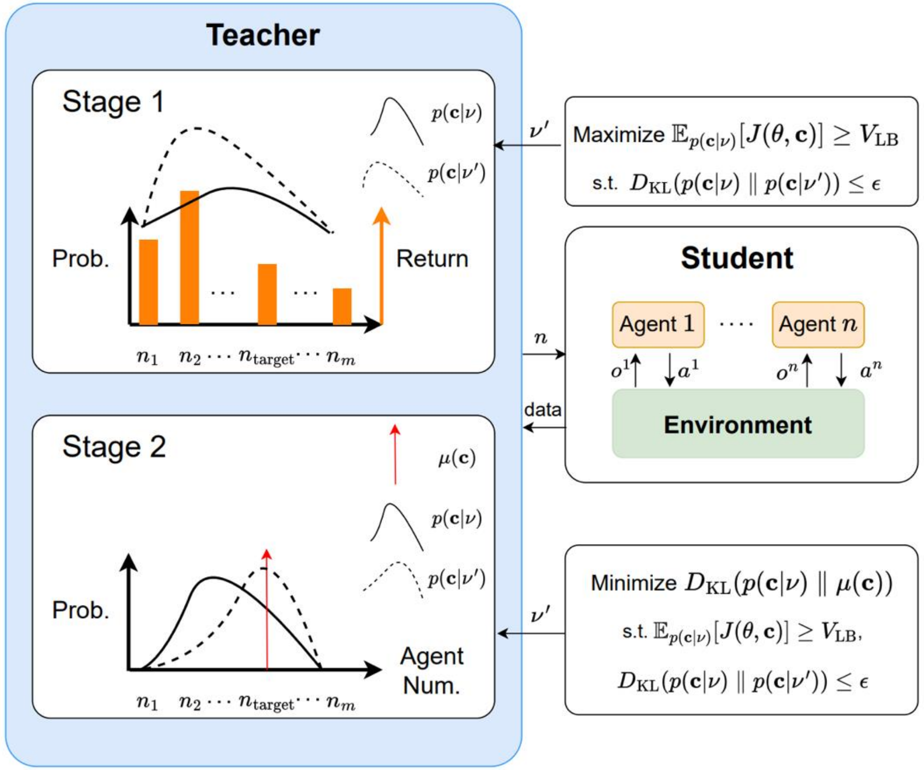
Figure 2: Simple-spread task, where the common reward is computed by the sum of minimum distances of each landmark to agents. With more agents, the task becomes eaiser to get higher rewards.

$$\max_{\nu_{k+1}} \frac{1}{M} \sum_{i=1}^{M} \frac{p(\mathbf{c}_i|\nu_{k+1})}{p(\mathbf{c}_i|\nu_k)} V_\theta(\mathbf{s}_{i,0}, \mathbf{c}_i), \tag{3}$$

- High estimation variance
- Increased credit assignment difficulty

Method

We propose a ***learning progess*** based curriculum learning method: SPMARL



Two-stage optimization

Main idea:

- Value loss indicates the policy change well.
- On tasks with higher value loss, the policy can be improved more.

$$LP(c) = \frac{1}{2}\mathbb{E}_{s,\mathbf{a}\sim\pi(\mathbf{a}|s,\mathbf{c})}[\|R(s,\mathbf{a}) - V(s)\|^2]$$

The new objective maximized by

$$\max_{\nu_{k+1}} \frac{1}{M}\sum_{i=1}^{M}\frac{p(\mathbf{c}_i|\nu_{k+1})}{p(\mathbf{c}_i|\nu_k)} LP_\theta(\mathbf{c}_i)$$

Experiments: Simple-Spread

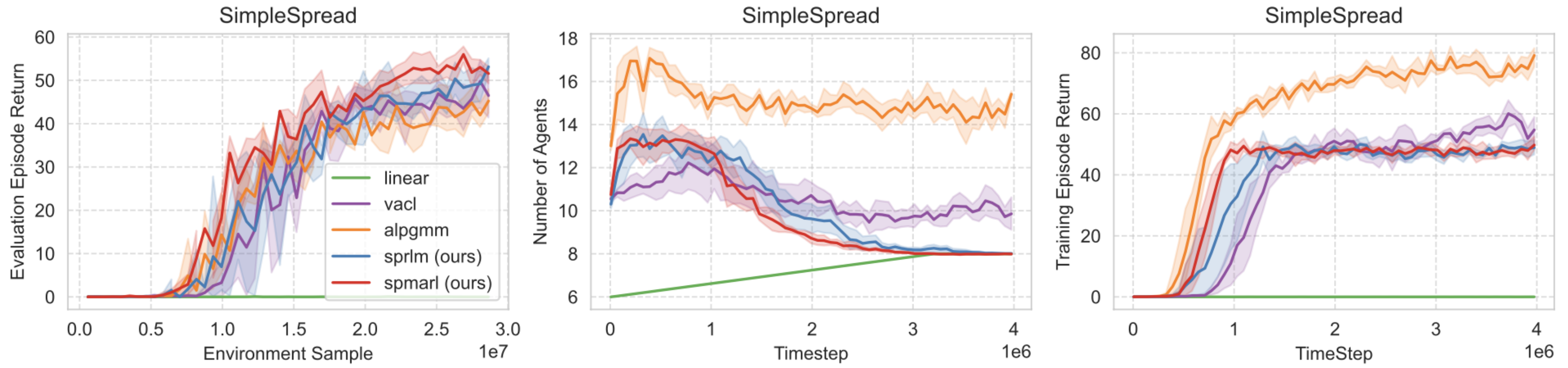


Figure 2. Comparison on the *Simple-Spread* task, where the target is set with 8 agents and 8 landmarks. The plots are averaged over 5 random seeds and the shadow area denotes the 95% confidence intervals. The **left** figure shows the evaluation returns on the target task with 8 agents. Note that the x-axis represents the samples collected from the environment, which is proportional to the number of agents. The **middle** figure presents the generated curriculum from different methods, where SPMARL and SPRLM first generate more agents and then converge to the target 8 agents while ALPGMM and VACL always generates more agents. The **right** figure shows the episode returns on the training tasks. The ALPGMM algorithm achieves the highest because it samples tasks with more than 14 agents.

Experiments: XOR

		Player 2: A	Player 2: B
Player 1	A	0	1
	B	1	0

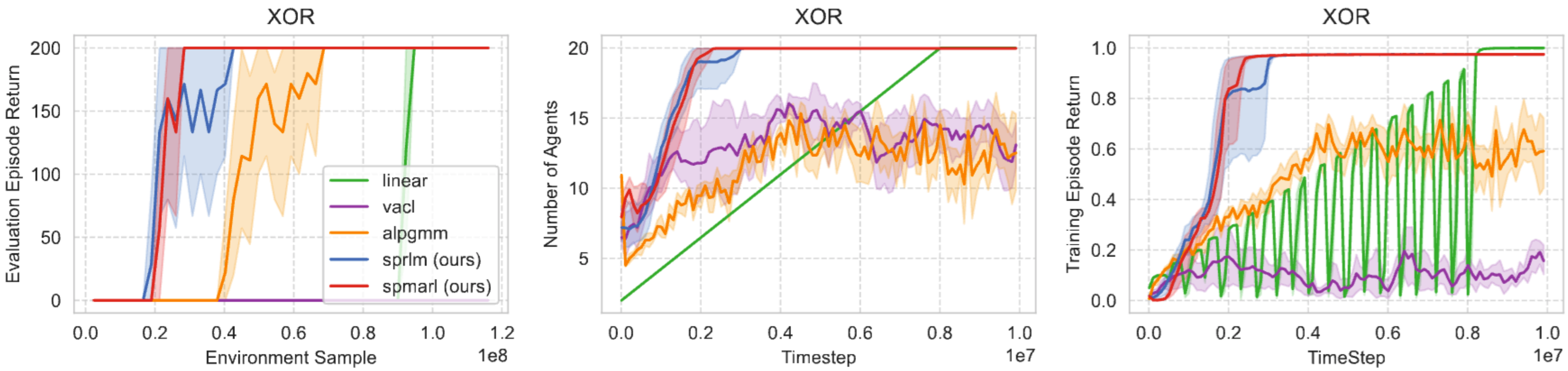
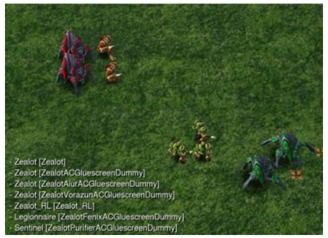
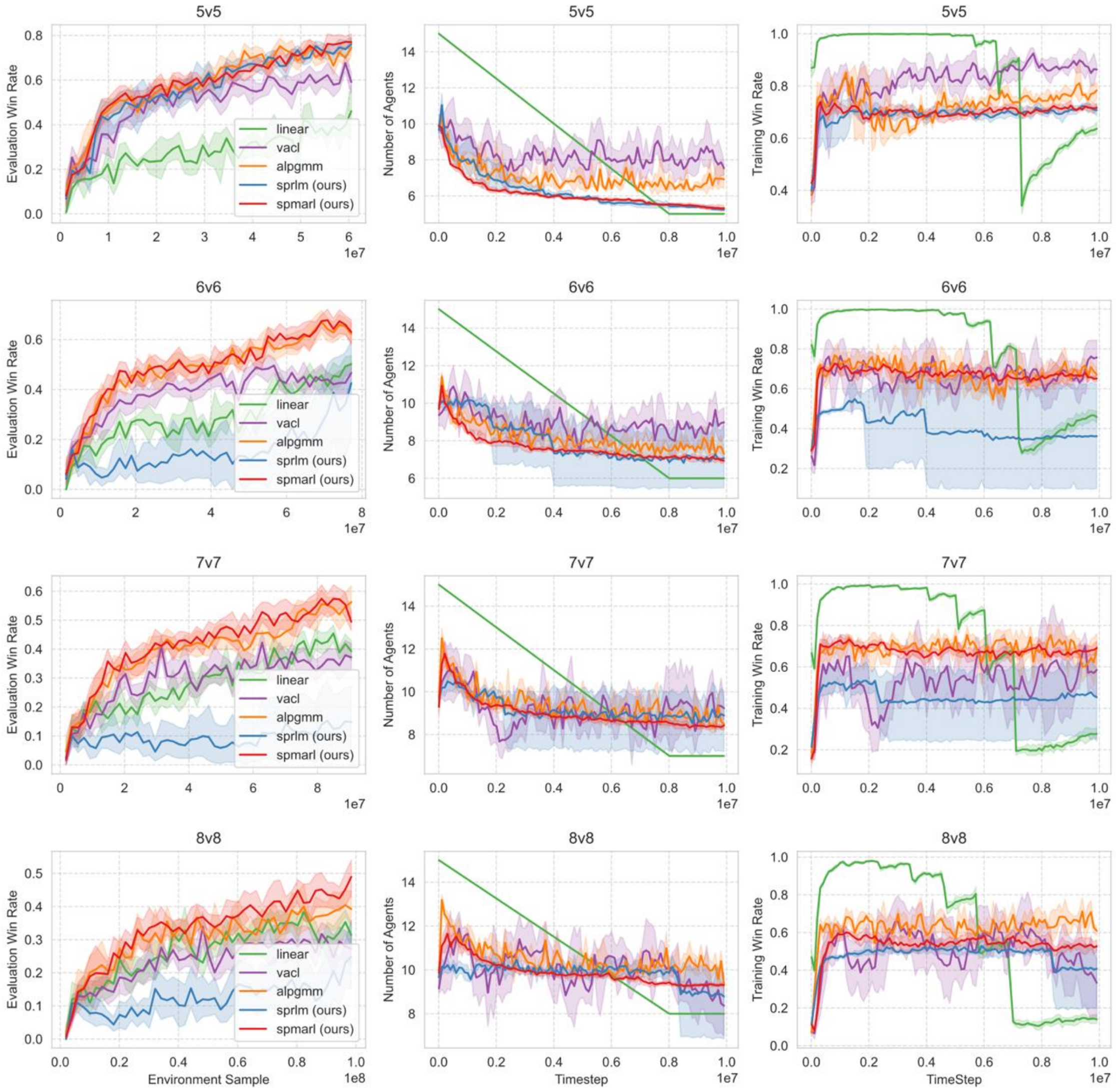


Figure 5. Comparison on the 20-player *XOR* game where each agent needs to output different actions to succeed. While the linear curriculum from few to more (*linear*) and *alpgmm* successfully achieve optima eventually, SPRLM and SPMARL demonstrate a faster convergence.

Experiments: SMAC v2





Conclusion

- We identify two issues related to the general reward-based automatic CRL methods and propose learning-progress based curriculum learning.
- While not maximizing the reward, our method, SPMARL, generates tasks with higher rewards faster than the naïve application of SPRL which maximize the reward over the number of agents.