Negatively Correlated Ensemble Reinforcement Learning for Online Diverse Game Level Generation

Ziqi Wang [1] Chengpeng Hu [1] Jialin Liu [1] Xin Yao [2]

[1]Southern University of Science and Technology, Shenzhen

[2]Lingnan University, Hong Kong

Speaker: Ziqi Wang

Our *Learning and Optimisation in Games Lab*: https://aingames.cn/

Contact: wangzq2021@mail.sustech.edu.cn

Background

- Online procedural content generation (PCG), which refers to the real-time and incremental generation of new game content, is an important demand from the game industry.
- Recent works [1], [2] have shown that reinforcement learning (RL) is powerful for online game level generation [1], [2].

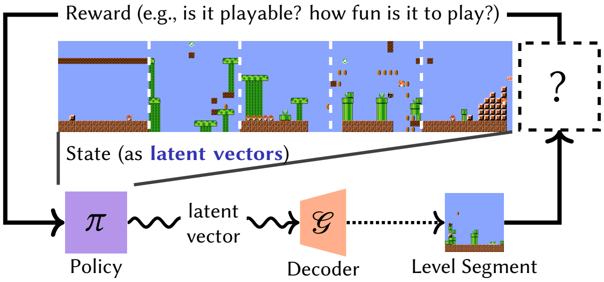


Figure 1: Framework of online PCG through RL, first introduced by Shu *et al.* [1].

[1] T. Shu, J. Liu, and G. N. Yannakakis, "Experience-driven PCG via reinforcement learning: A Super Mario Bros study," in *2021 IEEE Conference on Games*, IEEE, 2021, pp. 1–9

[2] Z. Wang, J. Liu, and G. N. Yannakakis, "The fun facets of Mario: Multifaceted experience-driven PCG via reinforcement learning," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*, ACM, 2022, pp. 1–8

The Issue of Level Diversity

- However, directly applying RL algorithms generates similar levels, i.e., lacking diversity [3].
- Also, regardless of the parameter setting, RL policies tend to generate recurrent patterns.

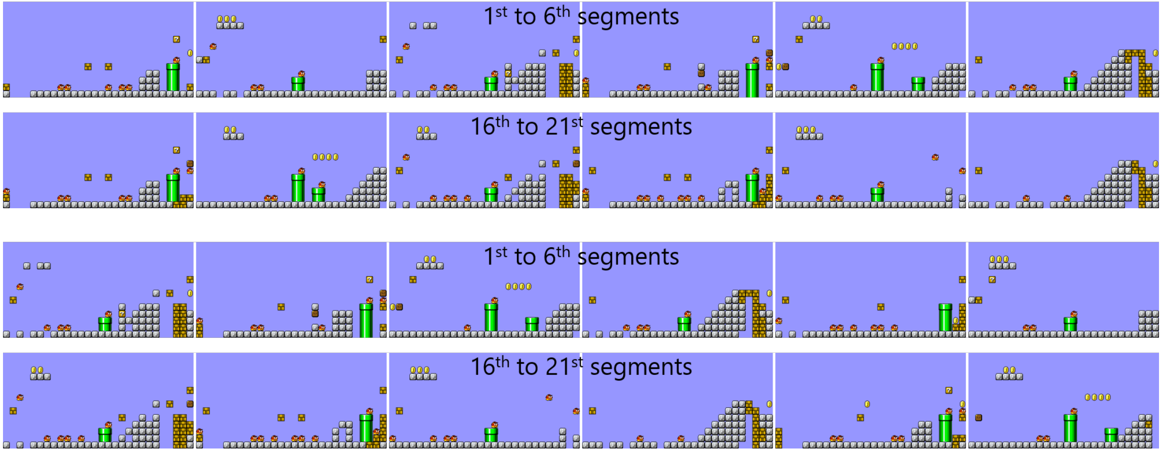


Figure 2: Two levels generated from RL policies trained with different parameters [3].

[3] Z. Wang, T. Shu, and J. Liu, "State space closure: Revisiting endless online level generation via reinforcement learning," *IEEE Transactions on Games*, vol. Early Access, 2023. DOI: 10.1109/TG.2023.3262297

Challenges and Our Approach

What is the challenge in generating diverse levels with reinforcement learning policies?

1. Promising game levels are diverse, but greedy or unimodal stochastic policies can hardly make diverse decisions.
   - **Generate multiple candidate segments** using ensemble and stochastically select one from the candidates.
2. Diversity is a concept about the entire distribution. Using reward functions can hardly express diversity.
   - Encourage the policy to make diverse decisions with a **regularisation regarding the entire decision distribution**.

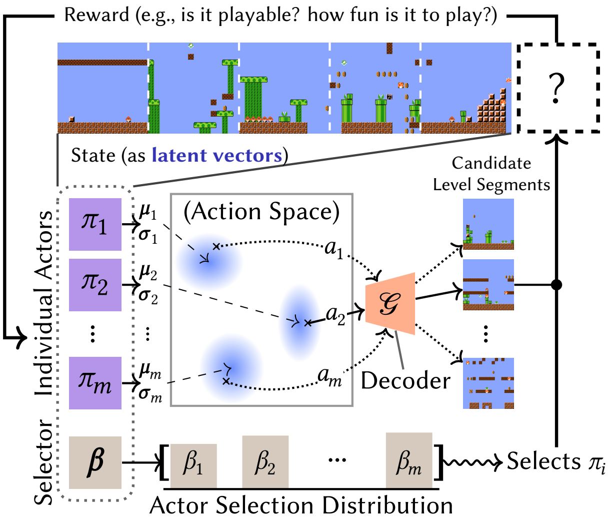


Figure 3: Generating a new level segment with our method

Regularising the Ensemble Policy

To diversify the decision distribution of our ensemble policy, we propose the *negative correlation regularisation*:

$$\varrho^{\pi}(s) = \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i(s)\beta_j(s) \min\left\{\omega(\pi_i(\cdot|s), \pi_j(\cdot|s)),\ \bar{\omega}\right\}, \tag{1}$$

where $\omega(\cdot, \cdot)$ denotes the 2-Wasserstein distance, $\bar{\omega}$ is a hyperparameter.

- **If $\rho^{\pi}(s)$ is optimised, then** $\forall i, j, \omega(\pi_i(\cdot|s), \pi_j(\cdot|s)) \geq \bar{\omega}, \beta_i(s) = \beta_j(s)$

Then the entire objective function is

$$J^{\pi} = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty}(R_t + \lambda\varrho^{\pi}(S_t))\right] \tag{2}$$

Note $\rho^{\pi}(s)$ is not dependent on the actual action but evaluates the entire decision distribution, thus standard value functions and loss functions are applicable for optimising it. Here raise the question:

- ***How to adapt the value function and loss function, so that we can optimise the regularisation?***

Optimising the Regularisation

Regularisation state-action value:

$$Q^{\pi}_{\varrho}(s,a) \doteq \mathbb{E}_{\pi}\left[\sum_{k=1}^{\infty} \gamma^{k} \varrho^{\pi}(S_{t+k}) \,\middle|\, S_t = s, A_t = a\right] \tag{3}$$

Key Difference: The counter k start from 1 instead of 0, because $\varrho^{\pi}(S_{t+k})$ at $k=0$ is independent with a.

Policy Improvement for Regularisation:

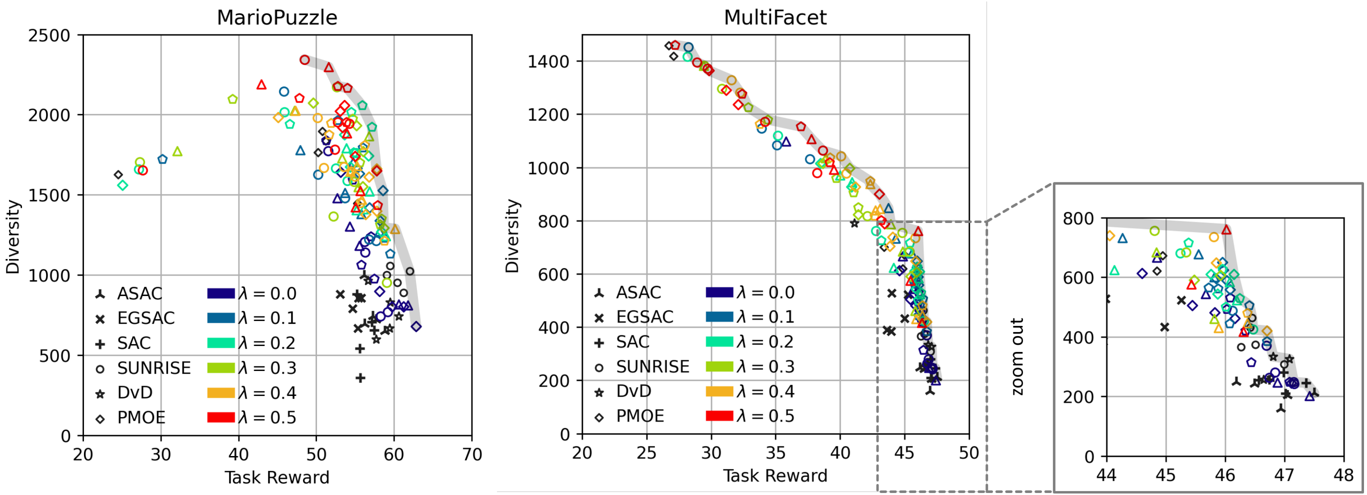
$$\forall s \in \mathcal{S},\ \pi_{\text{new}}(\cdot|s) \leftarrow \operatorname{argmax}_{\pi(\cdot|s)\in\Pi}\left[\varrho^{\pi}(s) + \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi_{\text{old}}}_{\varrho}(s,a)\right]\right] \tag{4}$$

Policy Gradient for Regularisation:

$$\frac{\partial J^{\theta}_{\varrho}}{\partial\theta} = \int_{\mathcal{S}} d^{\pi}(s)\left(\frac{\partial\varrho^{\theta}(s)}{\partial\theta} + \int_{\mathcal{A}} Q^{\pi}_{\varrho}(s,a)\frac{\partial\pi^{\theta}(a|s)}{\partial\theta}\,\mathrm{d}a\right)\mathrm{d}s \tag{5}$$

Performance under Varied Parameter Settings

We tested our proposed method and other RL algorithms across two tasks, following the works of [1] and [2], respectively. The two tasks are varied in reward function and observation space.



Diversity is evaluated by the **average Hamming distance** of levels generated by the policy.

Generated Levels

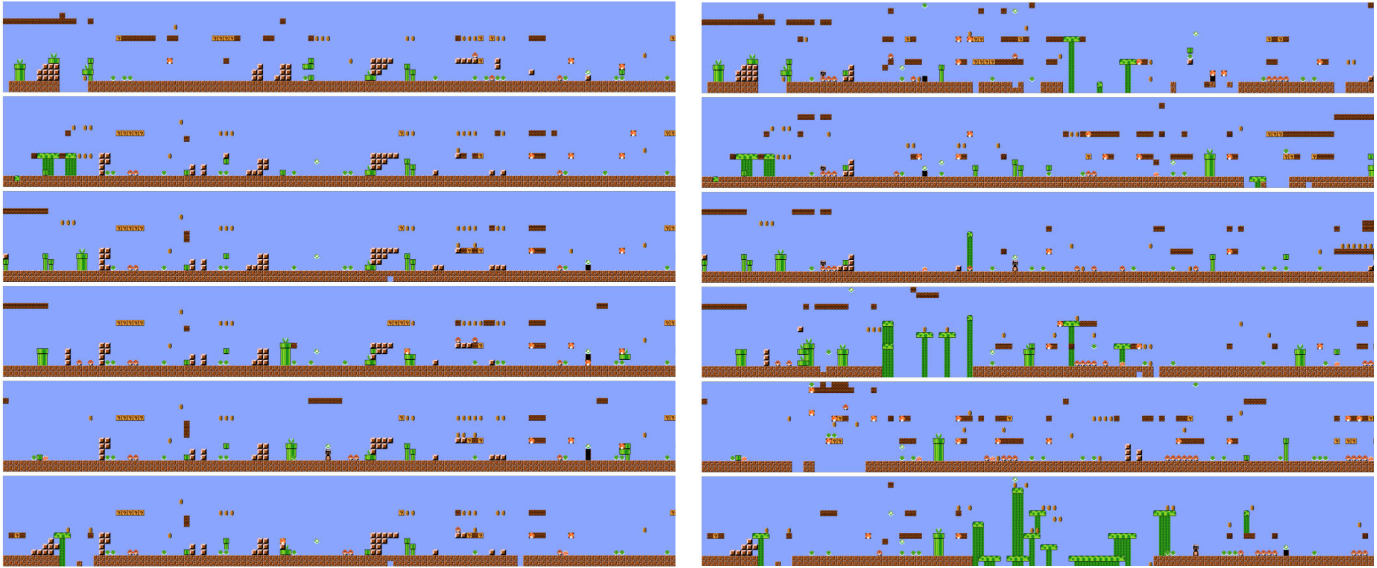


Figure 4: Levels generated by two policies trained **without (left)** and **with (right)** our approach, from the same set of initial conditions.

Conclusion

Method

1. We introduced an ensemble RL approach that is capable of **modeling a diverse decision distribution**.
2. We proposed a novel **negative correlation regularisation** to promote the diversity of levels online generated by our ensemble RL policy.

Result

- By using different λ values, our approach produces **a wide range of trade-offs** between task reward and diversity.

Thank You!