CURRICULUM AS SELECTIVE DATA ACQUISITION: TOWARD RELIABLE GENERALIZATION IN GOAL-CONDITIONED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

We study curriculum learning in goal-conditioned reinforcement learning (GCRL) through the lens of data selection. Instead of sampling all goals uniformly, we bias sampling toward underachieved goals, thereby shifting the state–goal distribution seen by the agent. Using universal value function approximators (UVFAs) with potential-based reward shaping in GridWorld, we compare uniform and curriculum-guided training. Our results show that curricula alter goal coverage, reduce approximation error, and improve success on difficult edge goals. These findings highlight curriculum learning as a principled mechanism for selective data acquisition, suggesting a pathway toward more persistent and open-ended agents.

1 INTRODUCTION

Goal-conditioned reinforcement learning (GCRL) provides a flexible paradigm for training agents to solve multiple tasks within the same environment by conditioning policies or value functions on a goal state (Schaul et al., 2015). A persistent challenge in this setting is that many goals are difficult to reach under uniform sampling, leading to extremely sparse rewards and poor signal for function approximation (Andrychowicz et al., 2017). This challenge is magnified in open-ended learning (OEL), where agents must continually acquire and refine skills over an unbounded set of goals or tasks (Clune, 2019). Our own motivation for this paper stems directly from recent work by Hughes et al. (2024), who highlight the need for algorithmic paths toward persistent, open-ended learning. We view curriculum learning as one such path, offering a tractable starting point for shaping state–goal distributions.

Curriculum learning has been widely explored as a remedy for sparsity and exploration issues, typically by sequencing goals from easy to hard (Bengio et al., 2009; Florensa et al., 2017; Portelas et al., 2020). Prior work has developed handcrafted curricula (Bengio et al., 2009; Florensa et al., 2017), teacher–student frameworks (Matiisen et al., 2019; Narvekar et al., 2020), and automated goal-generation strategies (Held et al., 2018; Portelas et al., 2020). While these approaches differ in implementation, they share a core intuition: curricula act as a mechanism to ensure agents remain within their "zone of proximal development," preventing stagnation on trivial tasks and collapse on impossible ones (Matiisen et al., 2019).

Despite this progress, much of the literature treats curriculum as an exploration heuristic or as a way to overcome reward sparsity. Far less attention has been paid to its effect on the *distribution of training data* itself. In particular, curricula can be seen as a form of selective data acquisition, biasing the state–goal visitation distribution toward goals that are currently underachieved. This reframing highlights a structural rather than incidental role for curricula: by reshaping the data distribution, they change the inductive biases of the learned function approximator. By focusing on how curricula reshape state–goal distributions, we explicitly link curriculum design in GCRL to the broader questions of persistence and adaptability central to OEL (Hughes et al., 2024).

In this work, we investigate this perspective empirically using Universal Value Function Approximators (UVFAs; (Schaul et al., 2015)) trained in GridWorld. We compare uniform goal sampling to curriculum-biased sampling, analyzing how distributional shifts in data affect function approximation and downstream policy success. We show that curricula concentrate data in informative

regions of the state–goal space, reduce approximation error on a shared evaluation set, and improve policy success particularly on harder-to-reach goals. These findings suggest that curriculum learning should be understood not only as an exploration strategy, but also as a structural mechanism for guiding data acquisition—one that provides a concrete entry point into the larger challenge of scaling toward lifelong and open-ended learning.

2 METHODS

2.1 ENVIRONMENT SETUP

We use a deterministic GridWorld navigation environment where an agent must reach a goal location specified at the start of each episode. Each state is defined by the agent's current position, and each task is defined by a desired goal cell. Episodes terminate either upon reaching the goal or when a maximum horizon H is reached. This setting provides full observability, yet exposes the challenges of goal-conditioned reinforcement learning: large goal spaces, varying difficulty across cells (interior vs. edge), and sparse terminal rewards.

2.2 UNIVERSAL VALUE FUNCTION APPROXIMATORS (UVFAS)

We employ Universal Value Function Approximators (UVFAs; Schaul et al., 2015), which generalize value estimation across states and goals.

- Input: concatenation of agent state (x, y) and (g_x, g_y)
- Architecture: a multilayer perceptron (MLP) with ReLU activations and hidden dimension 64.
- Output: scalar estimate of the value function $V(s, g)$
- Training objective: mean squared error regression against pseudo-reward targets (see below).

This formulation allows us to assess not only policy performance but also how curricula affect function approximation quality across the entire state–goal space.

2.3 POTENTIAL-BASED REWARD SHAPING (PBRS)

To provide dense learning signals, we adopt Potential-Based Reward Shaping (PBRS; Ng et al., 1999). The formula is defined as follows:

Define a potential $\phi(s, g) = -d(s, g)$ where d is the Manhattan distance between state and goal

The shaped reward is defined as:

$$r_t = \lambda[\gamma\phi(s_t + 1, g) - \phi(s_t, g)] - c$$

where discount $\gamma = .99$, shaping coefficient $\lambda = .5$ and step cost $c = .01$. A terminal bonus of $+1$ is added on successful episodes.

Targets are constructed as discounted returns-to-go under this shaped reward. For evaluation, we negate returns so that greedy action selection corresponds to $\arg\max$ over predicted values.

2.4 CURRICULUM DESIGN

We compare two data acquisition strategies:

1. Uniform (NoCurr): goals are sampled uniformly from all valid grid cells
2. Edge-Weighted Curriculum (Curr): sampling distribution is biased toward harder-to-reach goals, defined as those on the grid periphery. Empirically, edge cells are less frequently reached under uniform sampling, leading to underrepresentation in training data

In all cases, we collect fixed-size datasets per seed and train UVFAs with identical architectures, isolating the effect of curriculum-induced distributional shifts.

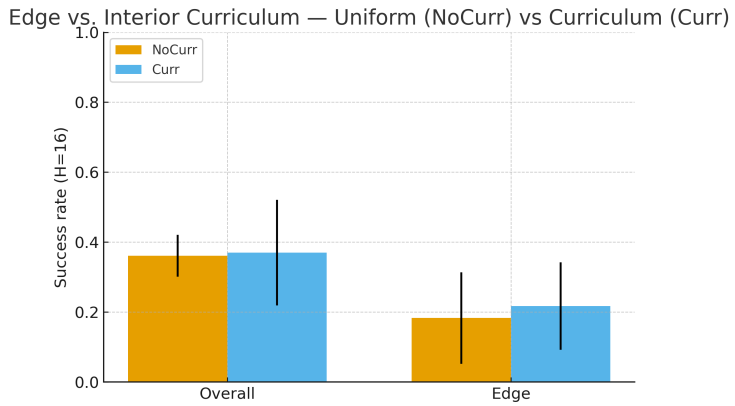

Figure 1: Edge vs. Interior Curriculum. Success rates at horizon $H = 16$ for agents trained with uniform sampling (NoCurr) versus edge-biased curriculum (Curr). Results are averaged across three seeds; bars show mean $\pm$ one standard deviation. Curriculum improves performance on harder edge goals while maintaining comparable performance overall.

2.5 Training Protocol

Data collection: For each seed, we roll out 1000 episodes with greedy action selection under PBRS shaping. Each trajectory is stored as a JSONL file and converted into a PBRS dataset (*.npz).

UVFA training: Models are trained for 50 epochs using Adam with learning rate $10^{-}3$ and batch size 256. Each run is repeated across three seeds for robustness.

Evaluation: Trained UVFAs are evaluated zero-shot on held-out goals with varying horizons ($H \in \{30, 20, 16, 12\}$).Success is measured as the fraction of goals achieved within horizon H, reported separately for interior and edge subsets.

3 Results

3.1 Baseline: Uniform vs. Curriculum Sampling

We first compare universal value function approximators (UVFAs) trained on uniformly sampled goals (NoCurr) with those trained using a manual curriculum that upweights edge goals (Curr). All agents were trained on $N = 1000$ episodes per seed (three seeds, max steps $= 30$). Evaluation was performed using greedy policies at varying horizons $H \in \{30, 20, 16, 12, 10\}$.

Success rates Across seeds, the curriculum models showed modest but consistent improvements on harder edge goals, with comparable overall performance. At $H = 16$, uniform (NoCurr) achieved 0.361 ± 0.060 overall and 0.183 ± 0.131 on edge goals, whereas curriculum (Curr) achieved 0.370 ± 0.151 overall and 0.217 ± 0.125 on edge goals (Fig. 2). While not universally stronger in aggregate, the curriculum condition tended to improve performance on the harder subset, consistent with the idea that selective sampling reshapes the state–goal distribution.

Distributional shifts. We confirm that edge-biased curricula shift the training distribution (Fig. 2), with increased density of trajectories targeting harder edge goals. These shifts translate into modest but measurable improvements in function approximation and policy success in these regions, supporting our hypothesis that curriculum should be viewed as selective data acquisition. While gains are not uniform across all goals, the evidence suggests that even simple hand-crafted curricula can systematically bias training data toward underachieved subsets and thereby improve performance where uniform sampling struggles.

To test the effect of curriculum design choices, we compared two variants against uniform sampling (NoCurr). The *baseline curriculum* biased sampling toward edge goals with a fixed proportion,

Figure 2: Edge vs. Interior Curriculum. Training distributions and success rates at horizon $H = 16$ for agents trained with uniform sampling (NoCurr) versus edge-biased curriculum (Curr). Results are averaged across three seeds; bars show mean $\pm$ one standard deviation. Curriculum biases data toward harder edge goals, yielding modest improvements in those regions while maintaining comparable performance overall.

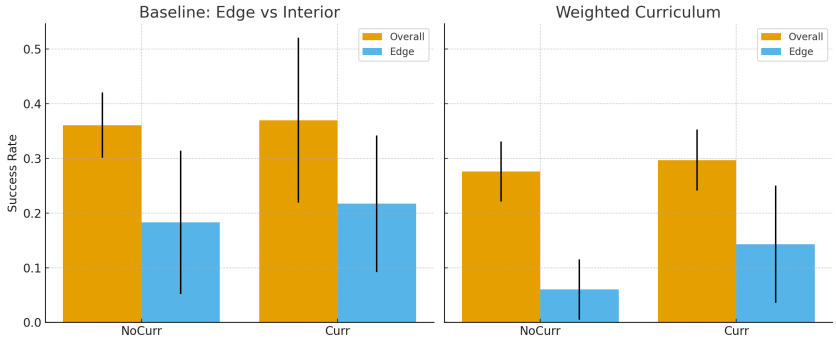

Figure 3: Curriculum variants. Success rates at horizon $H = 16$ for agents trained with uniform sampling (NoCurr), baseline curriculum (Curr), and weighted curriculum (Curr-W). Results are averaged across three seeds; bars show mean $\pm$ one standard deviation. Both curricula improve edge-goal success, with weighting amplifying the effect.

while the *weighted curriculum* further increased edge sampling to match their empirical difficulty under NoCurr. This weighting scheme was intended to strengthen the effect of curriculum as selective data acquisition, especially on harder goals.

3.2 Curriculum Variants: Baseline vs. Weighted

Figure 3 summarizes the results. Overall performance remained comparable across conditions, but both curriculum variants improved success on edge goals, with the weighted curriculum showing the strongest gains (Δedge $\approx +0.18$). These findings highlight how curricula reshape the state–goal distribution: by allocating more data to underachieved regions, they systematically improve function approximation where it matters most.

3.3 Summary

Overall, our experiments support the interpretation of curriculum as selective data acquisition. By biasing training toward underachieved goals—here instantiated as edge-aligned subsets—curricula reshape the state–goal visitation distribution and improve approximation in targeted regions of the universal value function approximator (UVFA). While gains were modest in aggregate, both the baseline and weighted curricula consistently provided benefits on harder edge goals, with the weighted variant amplifying these improvements. This underscores the role of curriculum as a structural mechanism for guiding data rather than an incidental exploration heuristic.

As shown in Table 1, curriculum improves overall success by +0.02 on average and edge-goal success by +0.08. These gains, though modest in absolute terms, consistently favor the curriculum condition on harder subsets. This provides evidence for our central claim: curricula act not merely as exploration heuristics but as structural mechanisms for data acquisition that enhance learning in regions where UVFA generalization is weakest.

4 Discussion

Our findings suggest that curriculum learning in goal-conditioned reinforcement learning (GCRL) is best interpreted as selective data acquisition rather than merely an exploration heuristic. By biasing

Setting (H=16)	**Uniform (NoCurr)**	**Curriculum (Curr)**	Δ (Curr–NoCurr)
Overall Success	0.276 ± 0.055	0.297 ± 0.056	+0.021
Edge-Goal Success	0.060 ± 0.055	0.143 ± 0.107	+0.083

Table 1: Pc

training toward goals that are harder to achieve under uniform sampling, curricula reshape the state–goal visitation distribution and improve value approximation in targeted regions of the space. This effect is particularly evident in subsets of goals at the periphery or in empirically defined "zones of proximal development," where uniform sampling struggles to provide sufficient coverage.

At the same time, our experiments show that the benefits of curricula are not uniform. Improvements are strongest on hard-to-reach goals but less consistent across easier subsets. In some cases, the curriculum bias may even reduce performance on goals already well-represented under uniform sampling. This reinforces the idea that curricula act as structural biases: their effectiveness depends on how well the sampling emphasis aligns with the learning bottlenecks of the agent.

Importantly, our weighted curriculum experiment provides further evidence for this interpretation. By explicitly rebalancing the goal distribution to upweight harder regions, we observed amplified gains on edge goals compared to the baseline curriculum. This suggests that the magnitude and direction of curriculum effects depend critically on how the sampling distribution is shaped. Rather than treating curricula as one-size-fits-all exploration strategies, they should be viewed as tunable mechanisms for structuring data acquisition in line with task difficulty and representational limits.

4.1 Limitations and Future Work

Our study has several limitations. First, we evaluate curricula in relatively small GridWorld environments with hand-designed goal distributions. While this setting allows clear analysis of distributional shifts, it limits direct applicability to more complex domains. Second, our curricula remain manually specified, with the edge–interior and weighted sampling schemes serving as simple proxies for more principled strategies. As a result, gains were modest and sometimes inconsistent across seeds.

Future work should focus on more robust manual curricula that better capture the "zone of proximal development" (ZPD), as well as the development of automated approaches that adaptively adjust sampling distributions in response to an agent's progress (e.g., teacher–student or adversarial frameworks). Another promising direction is testing curriculum-driven selective data acquisition in environments with more complex goals or continuous control settings, where distributional bias may have stronger effects on function approximation. Ultimately, advancing toward automated and generalizable curriculum mechanisms offers a more practical pathway to open-ended learning (Hughes et al., 2024).

5 Conclusion

We conclude that curriculum learning provides a structural mechanism for shaping the training distribution in goal-conditioned reinforcement learning (GCRL). By reallocating data toward underachieved goals, curricula improve value approximation and policy success in targeted regions of the state–goal space. Although our experiments are preliminary and limited to small GridWorld settings, they support reframing curriculum as selective data acquisition rather than a mere exploration aid. Using universal value function approximators (UVFAs) as our testbed, we showed how curriculum biases reshape state–goal visitation and guide function approximation. Looking forward, the integration of curricula with UVFAs offers a promising pathway toward more persistent and open-ended agents, connecting this line of work with recent efforts in lifelong learning and open-ended systems (**?**). This perspective motivates future research on more robust manual strategies, automated curriculum generation, and generalization to richer goal spaces and environments.

See Bengio et al. (2009) for early work on curricula.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Andrés Campero, Roberta Raileanu, Heinrich Kuttler, Razvan Pascanu, Tim Rocktäschel, and Martin Riedmiller. Learning with amigo: Adversarially motivated intrinsic goals. In *International Conference on Learning Representations*, 2021.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minigrid: A minimalistic gridworld environment for openai gym. `https://github.com/Farama-Foundation/Minigrid`, 2018.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Guillaume Lample, Alexandre Olteanu, Sungjin Lin, Sandeep Subramanian, Ricky Lo, Thien Huu Nguyen, Nicolas Le Roux, Alessandro Sordoni, et al. Babyai: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*, 2019.

Jeff Clune. Ai-gas: Artificial intelligence as a general-purpose technology (review/position). *arXiv preprint arXiv:1901.05511*, 2019.

Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer. Curious: Intrinsically motivated modular multi-goal reinforcement learning. In *International Conference on Machine Learning*, 2019.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pp. 482–495, 2017.

Stéphane Forestier, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *IEEE Transactions on Cognitive and Developmental Systems*, 9(2):151–163, 2017.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *International Conference on Machine Learning*, 2017.

David Held, Xinyang Geng, Carlos Florensa, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. 2018.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024.

Vincenzo Lomonaco, Davide Maltoni, Lorenzo Pellegrini, Andrea Cossu, Antonio Carta, Gabriele Graffieti, Tyler L. Hayes, Matthias Lange, Marc Masana, Jary Pomponi, et al. Avalanche: An end-to-end library for continual learning. In *Advances in Neural Information Processing Systems*, 2021.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740, 2019.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, 1999.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Raphaël Portelas, Cédric Colas, Louis Manella Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep reinforcement learning: A short survey. *IJCAI*, pp. 4819–4825, 2020.

Sébastien Racanière, Aleksandar Botev, David Reichert, Razvan Pascanu, Oriol Vinyals, Raia Hadsell, and Nicolas Heess. Automated curricula through self-play. In *International Conference on Learning Representations*, 2020.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.

DeepMind Team. Open-ended learning leads to generally capable agents. *Nature*, 600:595–602, 2021. arXiv preprint arXiv:2109.07438.

First Wang and Others. Title placeholder for wang et al. 2024. *arXiv preprint*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

A Appendix