Enhancing autonomy and efficiency in goal-conditioned reinforcement learning

Yang, Z.

Chapter 1

Introduction

A child who learns to walk integrates sensory feedback from the environment with motor commands from the brain. We, as humans, first perceive the surrounding environment, gather visual, proprioceptive and tactile information and adjust our movements accordingly to avoid falls or setbacks to learn to walk in a stable, fast fashion. The aforementioned learning process is an example of reinforcement learning (RL): we adjust our behavior based on the feedback we collect from the external environment (Sutton & Barto, 2018). If we walk steadily and reach the desired locations sooner to get food for survival, which is positive feedback, we reinforce the behaviors that create such fast movements. In contrast, when we fall, we receive negative feedback. The impact of hitting the ground creates pain. As a result, we avoid the behaviors that lead to the fall. Although humans generally take 10 to 18 months to learn to walk by following this reinforcement learning circle, we do not require external help and can learn innately. This is the appeal of RL, that the learning occurs through trial and error, in a self-improving way. In this thesis, we will discuss a computational way to implement reinforcement learning, which is a popular subfield of artificial intelligence.

As an illustration, let's imagine there is a humanoid robot called Bob that is instructed to learn to walk. Bob is equipped with all different types of sensors for capturing the surrounding information and more importantly, it is programmed with a reward function that acts like our brain to score every action that Bob takes. The reward will be higher when the movement is stable and swift, and lower if Bob falls down. Bob dedicates itself to learning day and night, sometimes falling to the ground and requiring human assistance for repairs. After several days of effort, Bob finally masters the ability to walk.

With walking mastered, Bob is instructed to learn to cook, cycle, play soccer, and more. There are many more skills it can acquire and goals it can achieve. However, the skills Bob mastered for walking might not be useful for new tasks. For example, cooking requires Bob to stand and use its arm, rather than walk. Bob must develop new skills for newly encountered tasks. This way, if the same task is commanded in the future, Bob can execute the previously learned corresponding skill set. To learn multiple tasks, Bob begins to adjust its behaviors based on the feedback given by the reward function that changes with each new goal. Over time, Bob learns different skill sets to accomplish different goals. This process is called goal-conditioned reinforcement learning (GCRL), which is a generalization of standard RL and aims to learn multiple tasks. In the example, GCRL enables Bob to learn more than just one task, with each task linked to a specific reward function. In GCRL, the behavior learned is conditioned on the given goal and adapts when the goal changes.

To complete the GCRL loop (Colas et al., 2022), Bob should follow three primary steps and one optional step:

1. **Define the goal space**. Bob must first understand which goals are achievable within its capabilities. For instance, Bob could set goals such as learning to walk or cook. However, goals like flying would be unrealistic since Bob lacks wings.

2. **Select a goal**. Once the goal space is defined, Bob needs to choose a goal that is interesting to pursue. For example, if Bob has already mastered walking, it might find cooking more interesting and challenging as a new goal.

3. **Reach the goal**. After selecting a goal, Bob must devise a strategy to achieve it. This could involve various learning methods such as learning through trail and error, or imitating human behaviors.

4. **Post-explore** (optional). Even after achieving a goal, Bob can choose to continue exploring further and push beyond the initial goal to discover additional capabilities or improvements. For example, after mastering walking, Bob might continue to practice and eventually learn to run, finding that running is an extension of walking.

After the goal is achieved (step 3), Bob generally returns to step 2 to select a new goal and repeat the process. Or it goes back to step 1 since the newly acquired experience can also change / grow the goal space. Alternatively, Bob can opt for the post-exploration phase (step 4) to further explore new possibilities. Intuitively, when a goal is accomplished, Bob reaches the limit of its current abilities. By continuing

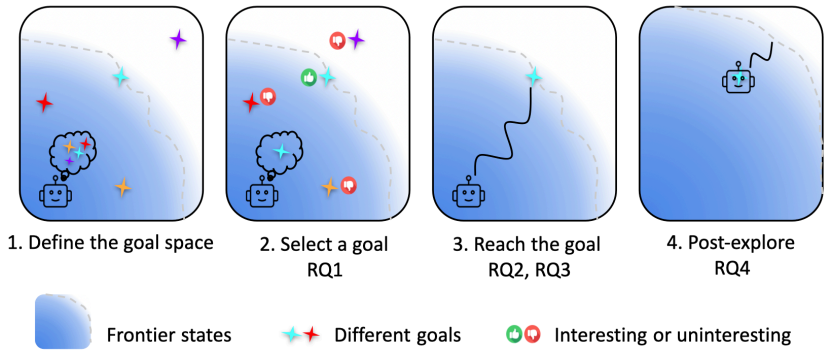


Figure 1.1: Four steps in the GCRL framework. We investigate four research questions (RQ) in total. RQ1 in Chapter 3 proposes a new goal selection mechanism to deal with the reset-free RL setting. RQ2 in Chapter 4 and RQ3 in Chapter 5 propose two non-parametric methods to learn to solve the given task faster. RQ4 in Chapter 6 investigates the role of post-exploration.

to push beyond these limits, it is likely to uncover new and exciting insights, thereby learning new skills.

GCRL (Liu et al., 2022) is an extension of conventional RL that adapts it to a multiple-goal setting. As illustrated by the example of Bob, GCRL enables the agent to learn to achieve multiple goals rather than just one. In our complex society, everyone must manage various roles and responsibilities, and ideally, we would want an AI agent to emulate this versatility, becoming a generalist agent proficient in various fields.

Recent developments demonstrate that leveraging the GCRL framework can enable an agent to achieve remarkable feats. For instance, an agent trained using GCRL is able to play StarCraft at the level of best human professional players (Vinyals et al., 2019), a game that requires rapid decision-making and precise control; perform hundreds of tasks (Mendonca et al., 2021), including object manipulation and locomotion; or solve Rubik's cube (Akkaya et al., 2019) with a robot hand. It also shows tremendous potential in other fields, such as finance (Zheng et al., 2022), healthcare (He et al., 2023) and energy management (Yi et al., 2022).

Training a generalist agent using GCRL takes four steps (illustrated in Figure 1.1), each of which are researched individually or combined to improve the GCRL loop. We now discuss each of these four steps and identify what is still missing.

As shown in Figure 1.1, the first step of the GCRL framework is to define the goal space. The goal space defines the space of interest that the agent should learn to achieve ultimately. It can be defined in several ways: the same as the state space, a

sub-space of the entire state space (Ecoffet et al., 2021; Plappert et al., 2018), or an embedding space of the original state space (Mendonca et al., 2021; Péré et al., 2018). After having a goal space, we need to specify a goal sampling strategy to sample goals from the defined goal space (see the step 2 in Figure 1.1).

A default approach for sampling goals is uniform sampling, which assumes that every goal is equally important to learn. However, previous research also shows that biasing the sampling towards frontier states (Pitis et al., 2020; Pong et al., 2019) or goals on which the agent has learning progress (Florensa et al., 2018; Portelas et al., 2020) can lead to better performance.

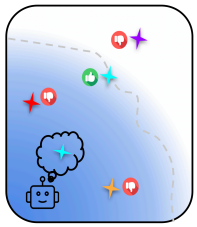
When a goal is selected (see the step 3 in Figure 1.1), the agent can be trained to reach the goal using goal-conditioned policies (Rahman et al., 2023; Schaul, Horgan, et al., 2015; Z. Zhang et al., 2023) or online planning (Hansen et al., 2022; Sancaktar et al., 2022). Afterwards, if the agent achieves the selected goal or exhausts the assigned budget for reaching it, it can opt for performing additional exploration (Ecoffet et al., 2021; E. S. Hu et al., 2023), which we call post-exploration. As illustrated in the fourth step of Figure 1.1, by engaging in post-exploration, the robot ventures into new areas and expands the frontier of its capabilities.

Although extensive research has already been conducted on each step of the GCRL framework, in this thesis, we have identified several drawbacks in current methods: 1) Current goal selection strategies are designed for conventional episodic RL settings, and they fail in other scenarios, for example, in settings where there is no reset. 2) After an interesting goal (e.g. the frontier state) is selected, an under-trained goal-conditioned policy might be lagging behind and not able to reach it, which slows down the learning process. 3) While performing additional exploration after reaching a goal (post-exploration) enhances performance, a direct comparison with the agent without post-exploration is lacking. In this thesis, we investigate the aforementioned problems of existing methods and propose various new approaches to overcome these drawbacks, thereby improving the GCRL framework.

1.1 Research questions

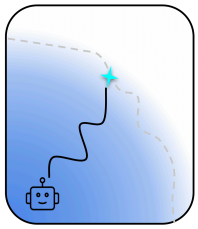
We research four questions in total (see Figure 1.1), one on step 2 (we propose a goal selection strategy to tackle the RL setting where there is no reset), two on step 3 (we propose two non-parametric methods to reach the goal faster) and one on step 4 (we investigate the role of post-exploration). The research questions we aim to answer in the thesis are as follows:

Q1 **Can RL agents learn without access to a 'reset'?[Chapter 3]**



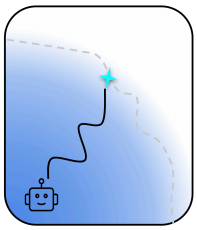
(step 2) RL agents typically need to be reset to their initial states after achieving a goal, starting the process over again. For example, once a quadruped robot completes a loading task in the factory, it will be moved to a designated location and its angles of joints will be restored to their default positions, preparing it for the next task. Additionally, if the robot gets stuck along the way, human intervention is required to free it so it can continue its task. Ideally, an RL agent should be able to operate autonomously and manage to escape from deadlock states, restore its joints, and move itself to designated locations as needed. In the GCRL framework, the behaviors of agents largely depend on commanded goals, making intelligent goal selection crucial for effective exploration and navigation. We demonstrate that by selecting goals appropriately, GCRL agents can operate without the access to the reset method. On the contrary, they can perform reset behaviors by themselves. This process occurs without human intervention, thus being more autonomous.

Q2 **Can non-parametric methods be applied to tasks with a continuous action space to achieve faster learning? [Chapter 4]**



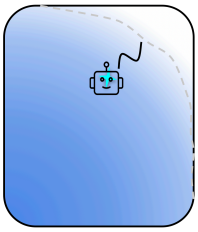
(step 3) Learning signals in RL are slowly back-propagated, especially when combined with parametric function approximations (e.g. deep neural networks). Rather than using RL to learn to achieve goals, previous research shows that non-parametric methods such as episodic control (Blundell, Uria, Pritzel, Li, Ruderman, Leibo, Rae, et al., 2016) can learn faster in Atari (Bellemare et al., 2013) and Labyrinth tasks, which all have discrete action spaces. We instead study if episodic control methods can also be applied to tasks with a continuous action space and, if so, how we can implement it to achieve faster learning, as it does in tasks with a discrete action space.

Q3 **Can non-parametric and parametric methods be combined to achieve both fast learning and optimality? [Chapter 5]**



(step 3) Non-parametric methods (episodic control) overwrite the previous solution by the newly encountered better solution. Consequently, it keeps track of the best solution quickly, even if such a solution might occur with low probability. Unlike RL, in tasks with stochasticity, episodic control is unable to store the optimal solution. In essence, RL is slow but capable of finding the optimal solution, while episodic control is fast but tends to be non-optimal. We investigate whether it is possible to combine episodic control with RL methods to leverage the advantages of both. Specifically, can we achieve the speed of episodic control while also attaining the optimality of RL?

Q4 **What benefits can post-exploration introduce compared to not having it? [Chapter 6]**



(step 4) Optionally, when the selected goal is achieved, RL agents can choose to explore additionally and hopefully step into the unknown area. We call the exploration after the goal has been achieved 'post-exploration'. Go-Explore (Ecoffet et al., 2021) implements the post-exploration idea as an RL agent and shows strong performance on extremely hard exploration tasks. However, it is not clear why the agent with post-exploration outperforms the one without and whether it can be extended to more general cases. We investigate why post-exploration is helpful and conduct a more explicit comparison between the agent with it and without by turning it on and off in the same RL algorithm within a more general GCRL framework, under both tabular and deep RL settings in discrete navigation and continuous control tasks.

1.2 Outline of this thesis

The stucture of this thesis is as follows. In Chapter 2, we provide the necessary background for understanding the work presented: RL and GCRL as well as its four main algorithmic design choices. Chapter-specific preliminaries will be introduced in each chapter separately.

Then, in Chapter 3, we propose a model-based reset-free (MoReFree) agent to tackle the reset-free RL setting where there is no access to a reset mechanism. By using world models and commanding the GCRL agent on three different goal distributions, MoReFree can eliminate the need for reset and result in a more autonomous agent.

MoReFree outperforms state-of-the-art methods tailored for reset-free RL setting with regard to data efficiency and asymptotic performance.

In Chapter 4, we propose a non-parametric method called continuous episodic control (CEC) that can learn to achieve goals faster. Since the original non-parametric method episodic control (Blundell, Uria, Pritzel, Li, Ruderman, Leibo, Rae, et al., 2016) only works for tasks with a discrete action space, we extend it to be able to solve tasks with a continuous action space and examine it on several robotic control tasks. Results show that CEC has a superior learning speed than conventional deep RL methods while obtaining competitive asymptotic performance.

Non-parametric methods (episodic control) overwrite the existing solutions with better ones as they are encountered. These methods learn faster but are non-optimal in stochastic tasks. In Chapter 5, we combine episodic control with RL to form one united agent, which switches between the two agents to gain benefits from both sides, i.e. the speed of episodic control and optimality and generalizability of RL.

In Chapter 6, we present the work on post-exploration, where we systematically investigate how additional exploration after the selected goal is achieved (post-exploration) enhances the performance under tabular and deep RL settings in discrete navigation and continuous control tasks. We find that the agent with post-exploration discovers more of the state space compared to the one without.

Finally, in Chapter 7 we reflect on the work presented in this thesis and the field of goal-conditioned RL as a whole, and propose potential directions for future work. At the end, we summarize the main conclusion.

1.3 List of publications

The chapters in this thesis are based on the following publications. Publications are edited only for style cohesion; all content remains unchanged from the published versions. The last one publication is related to the research, but not part of this thesis.

Chapter	Publication
3	Zhao Yang, Thomas M. Moerland, Mike Preuss, Aske Plaat, Edward S. Hu (2024). *World Models Increase Autonomy in Reinforcement Learning.* In submission
4	Zhao Yang, Thomas M. Moerland, Mike Preuss, Aske Plaat (2023). *Continuous Episodic Control.* In proceedings of the 2023 IEEE Conference on Games.
5	Zhao Yang, Thomas M. Moerland, Mike Preuss, Aske Plaat (2023). *Two-Memory Reinforcement Learning.* In proceedings of the 2023 IEEE Conference on Games.
6	Zhao Yang, Thomas M. Moerland, Mike Preuss, Aske Plaat (2023). *First Go, then Post-Explore: the Benefits of Post-Exploration in Intrinsic Motivation.* In proceedings of the 15th International Conference on Agents and Artificial Intelligence.
-	Zhao Yang, Mike Preuss, Aske Plaat (2021). *Transfer Learning and Curriculum Learning in Sokoban.* In proceedings of the 33rd Benelux Conference on Artificial Intelligence. Communications in Computer and Information Science, vol 1530 (pp. 187-200). Springer, Cham.