Enhancing Autonomous Reinforcement Learning: A Demonstration-Free Approach via Implicit and Bidirectional Curriculum

Daesol Cho[1], Jigang Kim[1] and H. Jin Kim[1]

***Abstract*— Despite the remarkable accomplishments of reinforcement learning (RL) in learning complex skills solely through interactions with the environment, its prerequisite of easy resets to the initial state at the end of the episode poses a challenge for autonomous learning of embodied agents. Hence, there has been a growing interest in developing autonomous RL (ARL) approaches that are capable of learning from continual, non-episodic interactions. However, existing ARL methodologies are constrained by their reliance on prior data, rendering them ineffective in scenarios where interactions pertinent to the task are infrequent. In contrast, our proposition introduces a demonstration-free ARL algorithm based on an implicit and bidirectional curriculum. Our method, employing a conditionally activated auxiliary agent and a bidirectional goal curriculum, outperforms prior methods, even those that make use of demonstrations.**

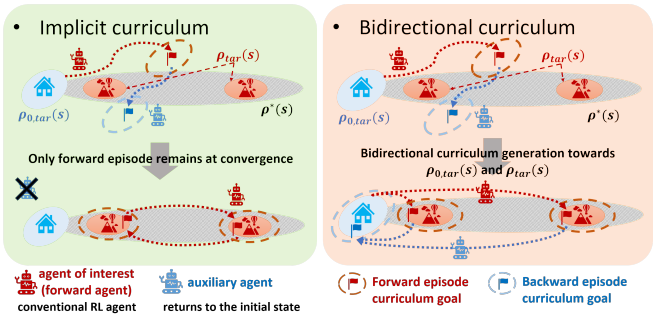


Fig. 1. Our method proposes a bidirectional curriculum for both forward and backward episodes. The auxiliary agent is no longer activated after the agent of interest becomes capable.

I. INTRODUCTION

Reinforcement learning (RL) has enabled interactive agents to learn complex skills across diverse domains without significant prior knowledge [1], [2], [3], [4]. However, prior methods assume an episodic setting where each trial begins from a state drawn from a fixed initial state distribution, and they are not designed to learn autonomously in the real world which involves ongoing, uninterrupted interaction. This challenge is particularly prominent in robotics. In most cases, addressing this challenge involves time-consuming and expensive interventions like human supervision, predefined scripted policies, and specialized experimental setups to reset the environment after each attempt [5], [6], [7]. To overcome these obstacles, it's crucial to develop RL agents that can learn autonomously with minimal external interventions.

Previous works involving RL agents in real-world scenarios primarily include a mechanism to handle resets. Reset mechanisms, which aim to minimize external interventions by requesting resets when necessary [8], [9] are only effective if manual resets can be easily executed. However, within the framework of non-episodic autonomous RL (ARL) [10], the option of manual resets upon request is absent and the agent has to learn continuously without any external interventions. To overcome the challenge of the non-episodic setting, many previous algorithms have depended on some form of pre-existing data with varying degrees of privilege, ranging from expert or sub-optimal trajectories [11], [12] to examples of states of interest [13]. However, the true essence of autonomy demands an agent's ability to learn entirely from scratch, devoid of external interventions or pre-existing data. To that end, we propose an ARL algorithm capable of training a goal-conditioned RL policy without the need for demonstrations, specifically designed for the non-episodic setting.

The ineffectiveness of existing RL algorithms in non-episodic scenarios has been widely acknowledged [14], primarily due to the challenge of insufficient practice opportunities for the evaluation task. A common framework for extending conventional RL to the non-episodic setting is to divide one continual interaction into multiple episodes. Typically, a forward episode pursues the original objective, while a subsequent backward episode targets an auxiliary objective, which serves as a foundation for the forward episode through a favorable initialization. While a common choice for this auxiliary objective is to return to the initial state distribution [8], this isn't always optimal, as it wastes valuable transitions on returning back to the initial state.

In this work, we consider an auxiliary agent that is conditionally activated to return to the initial state. This stems from our observation that providing a strong anchor is crucial, particularly when dealing with tasks that involve infrequent interactions that are unlikely to occur by chance in a non-episodic context. In our proposed approach, the agent of interest initially relies on the auxiliary agent but gradually reduces this dependence as training progresses through an implicit curriculum. As the agent of interest gains proficiency, consecutive forward episodes can be conducted without the auxiliary agent's involvement, enabling more transitions to be dedicated to training the agent of interest, leading to enhanced sample efficiency. While the auxiliary agent serves as a strong initial support, additional guidance is required for effectively training the agent of interest. Since the agent of interest must learn without prior data, we generate curriculum goals that

[1] Seoul National University, Artificial Intelligence Institute of Seoul National University (AIIS), Automation and Systems Research Institute (ASRI) dscho1234@snu.ac.kr, jgkim2020@snu.ac.kr, hjinkim@snu.ac.kr

do not rely on demonstrations or predetermined curricula. Specifically, we introduce a bidirectional goal curriculum approach that concurrently selects suitable goals for both episodes.

Our primary contribution is in introducing a demonstration-free ARL algorithm via implicit and bi-directional curriculum. Evaluations demonstrate the superior performance of our approach compared to existing methods. Additional analyses indicate that both the suggested implicit curriculum (utilizing the auxiliary agent) and the explicit curriculum (bidirectional goal curriculum) are well-formed and crucial for achieving successful learning in the demonstration-free, non-episodic scenario.

II. Preliminary

A. Autonomous Reinforcement Learning

We assume an ergodic environment for the demonstration-free, non-episodic setting, similar to many previous works on autonomous RL (ARL). We consider the Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho_0)$, where $\mathcal{S}$ denotes the state space, $\mathcal{G}$ the goal space, $\mathcal{A}$ the action space, $\mathcal{P}(s'|s,a)$ the transition dynamics, γ the discount factor, and ρ_0 the initial state distribution of the evaluation setting. The learning algorithm $\mathbb{A}$ is defined as $\mathbb{A} : \{s_j, a_j, r_j, s_{j+1}\}_{j=0}^{t} \mapsto \{a_t, \pi_t(\cdot|s)\}$, which maps the collected data until time t to an action a_t to be applied during the non-episodic training and its current best guess of the optimal evaluation policy $\pi_t(\cdot|s)$.

Typical implementations of RL algorithms (episodic) involve thousands or millions of sampling $s_0 \sim \rho_0(s)$, which require manual resets at the end of every episode. However, under the ARL framework (non-episodic), the initial state $s_0 \sim \rho_0(s)$ is sampled only once at the beginning and the agent interacts with the environment through the actions a_t determined by the algorithm $\mathbb{A}$ until $t \to \infty$.

ARL defines the *Deployed Policy Evaluation metric*, which measures how fast the policy π_t improves in terms of the evaluation performance for a given task:

$$\mathbb{D}(\mathbb{A}) = \sum_{t=0}^{\infty} \Big[J(\pi^*) - J(\pi_t) \Big] \tag{1}$$

where $J(\pi) = \mathbb{E}_{\rho_0, \pi, \mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$, and π^* is the optimal policy. The goal of algorithm $\mathbb{A}$ is to minimize $\mathbb{D}(\mathbb{A})$ by learning as fast as possible.

B. Surrogate Objective for Curriculum RL

We replace the original RL objective with a surrogate objective to be utilized for curriculum generation in Section III and describe it in detail. Let $\mathcal{T}$ be the joint distribution of some initial state s_0 and goal g. Then, the original objective $\max_\pi J(\pi)$ can be represented as,

$$\max_{\pi} V^{\pi}(\mathcal{T}) := \mathop{\mathbb{E}}_{(s_0, g) \sim \mathcal{T}} \left[V^{\pi}(s_0, g) \right] \tag{2}$$

where $V^{\pi}(s_0, g)$ is the goal-conditioned value function.

Our approach relies on the following generalizability condition [15], [16], [17], [18] that is characterized by the Lipschitz continuity-based assumption:

$$|V^{\pi}(\mathcal{T}') - V^{\pi}(\mathcal{T})| \leq L \cdot D(\mathcal{T}, \mathcal{T}') \tag{3}$$

where L is the Lipschitz constant and $D(\mathcal{T}, \mathcal{T}') = \inf_{\mu \in \Gamma(\mathcal{T}, \mathcal{T}')} (\mathbb{E}_{\mu}[d((s_0, g), (s_0', g'))])$ is the Wasserstein distance based on the distance metric $d(\cdot, \cdot)$. $\Gamma(\mathcal{T}, \mathcal{T}')$ denotes the set of all possible transport plans μ.

Under Eq (3), optimizing Eq (2) can be relaxed into the following lower-bound maximization,

$$\max_{\mathcal{T}, \pi} \Big[V^{\pi}(\mathcal{T}) - L \cdot D(\mathcal{T}, \mathcal{T}^*) \Big] \tag{4}$$

where $(s_0^*, g^*) \sim \mathcal{T}^*$ is the joint distribution of the target initial state s_0^* and target goal state g^*. Intuitively, it maximizes the policy performance and closeness to $\mathcal{T}^*$, which results in a task curriculum with increasing difficulty.

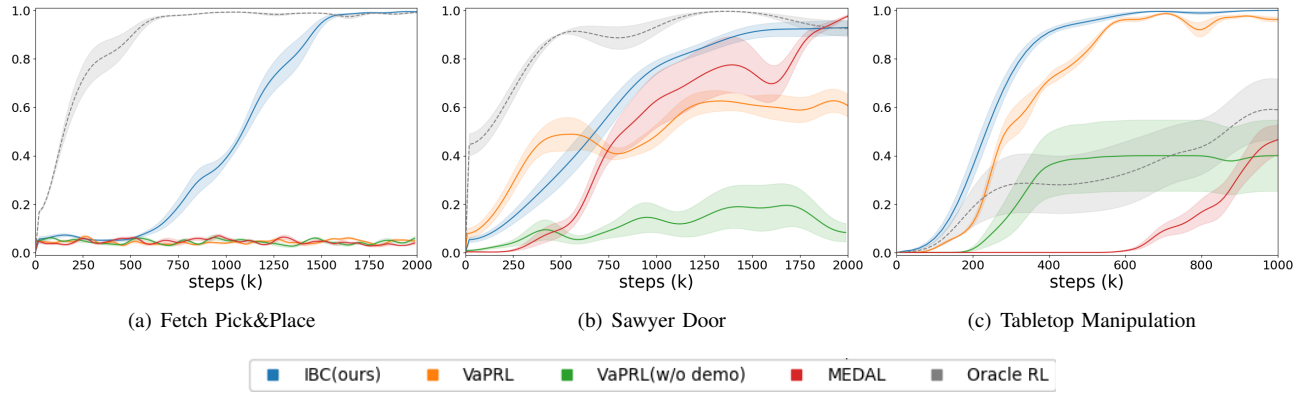
III. Method

A. Non-Episodic RL with an Auxiliary Agent

During non-episodic training, we alternate between the two agents such that the auxiliary agent guides the forward agent only when necessary. Specifically, we conditionally activate the auxiliary agent when the forward agent has failed at the given goal state such that the auxiliary agent gradually disappears as the forward agent improves which results in better sample efficiency. Let us consider the hypothetical setting where the forward agent is fully capable and the auxiliary agent does not intervene at all. Under this setting, the forward agent repeatedly attempts its target goal states $s_{g^*} \sim \rho_{tar}(s)$ without resets. Thus, the agent is no longer restricted by $\rho_0(s)$ unlike in episodic settings and we can consider a better initial state distribution by appropriately designing $\rho_{tar}(s)$.

Interestingly, a previous work [19] provides theoretical grounds that $\rho_0(s)$ close to $\rho^*(s)$ enables efficient training in RL, where $\rho^*(s)$ denotes the state marginal distribution of the optimal policy π^*. If we set $\rho_{tar}(s)$ to be a subset of $\rho^*(s)$ from the optimal policy that achieves the evaluation goal g_{eval}, we can approximately satisfy this ideal initial state distribution. Note that the target goal s_{g^*} achieved by the forward agent policy π_f from the previous rollout becomes the initial state for the next rollout.

In practice, it suffices for $\rho_{tar}(s)$, which is only used for bidirectional curriculum and not for RL, to contain a minimal number of key points that roughly outline the task to be adequate for the goal curriculum generation. This is because the curriculum goals effectively "fill in the blanks" by proposing past states from the replay buffer that are close to $\rho_{tar}(s)$. Typically, specifying $\rho_{tar}(s)$ requires only a handful of samples ($\sim$ 10) from $\rho_0(s)$ and g_{eval} combined to approximate $\rho^*(s)$. For some tasks, it suffices to specify $\rho_{tar}(s)$ with a single example from $\rho_0(s)$ and g_{eval} each. Unlike previous ARL methods, we do not require demonstrations with thousands of transitions or access to the expert policy.



(a) Fetch Pick&Place (b) Sawyer Door (c) Tabletop Manipulation

Fig. 2. Comparison of evaluation success rates of various algorithms. Shading indicates standard deviation across 5 seeds.

Until now, we have considered the setting where π_f has converged and is fully capable. However, most of the rollouts by π_f before convergence will lead the agent to an arbitrary state rather than s_{g^*}, leading to highly-varying initial states for the next rollout which results in unstable learning. For this reason, we need an auxiliary agent that provides an anchor and guides the forward agent. More precisely, the auxiliary agent tries to bring the forward agent back to the set of target initial states $s_0^* \sim \rho_{0,tar}(s)$. Even though $\rho_{0,tar}(s)$ can be an arbitrary set of states that are useful for the repeated practice of the forward agent, we set $\rho_{0,tar}(s)$ to include the environmental initial state distribution $\rho_0(s)$. This is because providing a strong anchor is crucial in practice and the evaluation will be performed from $\rho_0(s)$.

B. Bidirectional Curriculum Generation

While our non-episodic training process involving an auxiliary agent, $\rho_{0,tar}(s)$, and $\rho_{tar}(s)$ approximately satisfies the ideal initial state condition, it might not be sufficient for autonomous training in environments where target states are difficult to be achieved from scratch. Thus, we need to find intermediate goals that can guide the learning of the agent. To find such goals without relying on demonstrations, the candidates must be obtained from past trajectories with highly varying initial states due to non-episodic training. We propose a bidirectional goal curriculum based on the surrogate problem (Eq (4)) for both forward and auxiliary agents without relying on demonstrations in the non-episodic setting.

For autonomous curriculum generation, we sample the candidates for $\mathcal{T}$ from past states in the replay buffer $\mathcal{B}$. To prevent a degenerate solution in the curriculum selection process, a diversity constraint is incorporated such that for every trajectory $\tau = (s_0, ..., s_{t_{final}}) \in \mathcal{B}$, at most one state can be chosen for $\mathcal{T}$. Then, Eq (4) is transformed as follows,

$$\begin{aligned} &\max_{\pi_f, \mathcal{T}} \left[V^{\pi_f}(\mathcal{T}) - L \cdot D\left(\mathcal{T}, \mathcal{T}^*\right) \right] \\ &\text{s.t.} \sum_t \mathbb{1}\left[\left(s_0, \phi_f\left(s_t\right)\right) \in \mathcal{T}\right] \leq 1, \quad s_0, s_t \in \tau, \forall \tau \in \mathcal{B} \end{aligned} \tag{5}$$

where $\phi(\cdot)$ is a mapping function that abstracts the state space into the goal space. To solve Eq (5), we iteratively update $\mathcal{T}$ and policies π_f, π_a until π_f achieves a desirable evaluation performance. The policy optimization is simply achieved by applying off-the-shelf RL algorithms such as SAC [20]. The optimization of $\mathcal{T}$ is defined by the Wasserstein Barycenter problem augmented with a value bias term.

Inspired by [18], we enforce $\mathcal{T}$ and $\mathcal{T}^*$ to be a set of K particles ($|\widehat{\mathcal{T}}| = |\widehat{\mathcal{T}}^*| = K$) where $(s_0, g)^i \sim \widehat{\mathcal{T}}$, and $(s_0^*, \phi(s_{g^*}))^i \sim \widehat{\mathcal{T}}^*$, rather than parameterizing their distribution. Then, to address the Wasserstein Barycenter problem (Eq (5)) in the combinatorial setting, we assign candidates for $\widehat{\mathcal{T}}$ to $\widehat{\mathcal{T}}^*$ via the following bipartite matching problem:

$$\min_{\tau^i = \{s_t^i, \forall t\} \in \mathcal{B}} \sum_{\left(s_0^*, s_{g^*}\right)^i} w\left(\left(s_0^*, s_{g^*}\right)^i, \tau^i\right) \tag{6}$$

where $w(\cdot, \cdot)$ becomes

$$\begin{aligned} w\left(\left(s_0^*, s_{g^*}\right)^i, \tau^i\right) &:= c \left\| \phi_a(s_0^{*,i}) - \phi_a(s_0^i) \right\|_2 \\ &+ \min_t \left(\left\| \phi_f(s_{g^*}^i) - \phi_f(s_t^i) \right\|_2 - \frac{1}{L} V^{\pi_f}(s_0^i, \phi_f(s_t^i)) \right), \end{aligned} \tag{7}$$

when we define the distance metric $d((s, g), (s', g'))$ from Eq (3) as $c\left\|\phi_a(s) - \phi_a(s')\right\|_2 + \left\|g - g'\right\|_2$ (c is a hyperparameter). With the costs w defined according to Eq (7), we can construct a bipartite graph $\mathbf{G}(\{\mathbf{V}_a, \mathbf{V}_b\}, \mathbf{E})$. Let $\mathbf{V}_a$ be the set of nodes representing candidates for $\widehat{\mathcal{T}}$ and $\mathbf{V}_b$ be the set of nodes for $\widehat{\mathcal{T}}^*$. The weights of the edges are defined as $\mathbf{E}(v_a, v_b) = -w(v_a, v_b)$, where $v_a \in \mathbf{V}_a$ and $v_b \in \mathbf{V}_b$.

To solve the bipartite matching problem, the Minimum Cost Maximum Flow algorithm is utilized to find K edges with the minimum combined cost of connecting $\mathbf{V}_a$ and $\mathbf{V}_b$ [21]. The resulting K forward curriculum goals will be proposed towards a region of the state space considered to be close to $s_{g^*} \sim \rho_{tar}(s)$ and within the capability of the forward agent as indicated by the value bias term. Similarly, the K auxiliary curriculum goals will be proposed towards a region considered to be close to $s_0^* \sim \rho_{0,tar}(s)$.

IV. Experiment

We include three sparse reward environments to evaluate our method. Two environments – Tabletop Manipulation,

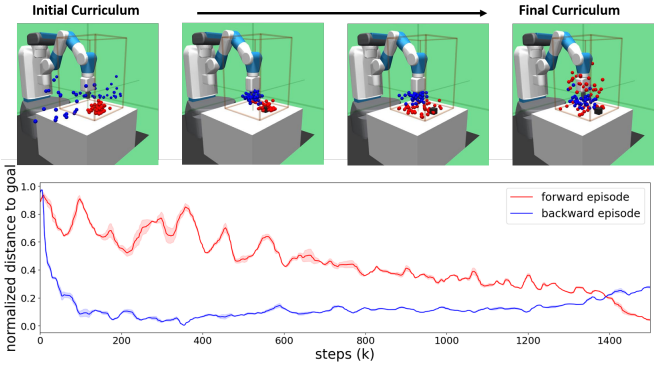


Fig. 3. Visualization of the curriculum goals and their average normalized distance to assigned target goals (Fetch Pick&Place). The red and blue dots indicate the curriculum goals for the forward and auxiliary agents, respectively.

Sawyer Door – are from established ARL benchmark, EARL [10], and the other environment – Fetch Pick&Place environment – is a modified version of existing MuJoCo-based OpenAI Gym environments [22], [23] for the ARL setting.

We compare with other previous methods designed for the ARL framework, which can be summarized as follows:

MEDAL [11] – a backward agent that minimizes the distance between its state marginal distribution and the expert state distribution.

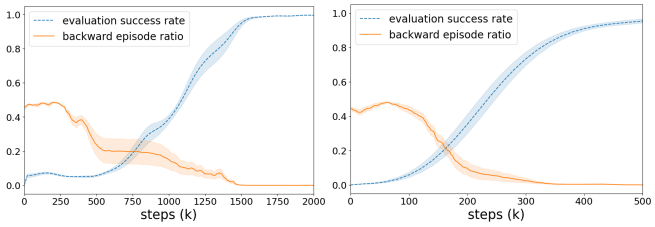
VaPRL [24] – value-based subgoal curricula towards the initial state distribution $\rho_0(s)$ during the backward episode; amenable to demonstration-free setting, but reports on the version with demonstration data.

oracle RL – a standard RL baseline such as SAC [20] in an episodic setting with goal relabeling technique [25] common for sparse reward environments.

A. Results and Analyses

We follow the evaluation setting similar to the EARL benchmark [10]. Specifically, the agent interacts with the environment after initially being spawned at $s_0 \sim \rho_0(s)$ and occasionally being reset to $s_0 \sim \rho_0(s)$ after hundreds of thousands of steps. Since we focus on minimizing the deployed policy evaluation metric, $\mathbb{D}(\mathbb{A})$, we report on $J(\pi_t)$ in 10k training step intervals by averaging returns from the policy over multiple evaluation episodes.

a) Evaluation results.: As shown in Figure 2, the proposed method achieves state-of-the-art performance against other baselines, without requiring any demonstration data and even achieving comparable success rates to the oracle RL. Although some prior works such as VaPRL and MEDAL utilize nearly expert-level demonstration data, they have difficulty in environments where the task-relevant interactions are very sparse in the non-episodic setting or the evaluation goals g_{eval} are uniformly spread over some region rather than a few points such as Fetch environments. For a fair comparison with our method, we also evaluated a version of VaPRL without demonstrations; it performed noticeably worse than the original VaPRL.



(a) Fetch Pick&Place (b) Tabletop Manipulation

Fig. 4. Episode ratio of the auxiliary agent and evaluation success rate.

To validate whether the intervention of the auxiliary agent vanishes as training proceeds, we plot the episode ratio of the auxiliary agent within the latest 1k episodes. As shown in Figure 4, the auxiliary agent does not intervene when the forward agent is fully trained.

b) Bidirectional curriculum.: To validate whether the bidirectional curriculum goals are properly interpolated and eventually converge to the desired target distributions, we evaluate the progress of the curriculum goals qualitatively and quantitatively. To do so, we visualize the forward and auxiliary curriculum goals and plot the corresponding normalized distance averaged over target goals assigned by bipartite matching (Section III-B).

The plots in Figure 3 demonstrate that the average distance to goals consistently decreases as training proceeds, which indicates that the curriculum goals for both forward and auxiliary agents have properly converged to their respective target states. The visualizations in Figure 3 provide further validation. Specifically, the forward curriculum goals gradually converge toward the $\rho_{tar}(s)$, which encompasses a region in the air and on the table for the Fetch Pick & Place. The auxiliary curriculum goals also converge to the target goal states $\rho_{0,tar}(s)$, initially. However, there is a gradual shift of the auxiliary curriculum goals towards $\rho^*(s)$ after initial convergence which is reflected in the slight increase in average distance to goals for the backward episode ($\rho_{0,tar}(s)$). This is because the candidates for the backward curriculum goals, which eventually become the initial states for the forward agent, are obtained from both $\rho_{0,tar}(s)$ and $\rho_{tar}(s) \subset \rho^*(s)$ when the forward agent remains at intermediate proficiency ($\sim$50%) for prolonged timesteps during training.

V. Conclusion

In this work, we considered a non-episodic RL setting where the agent should learn how to perform the given task autonomously without any external interventions such as manual resets and prior data. We proposed a demonstration-free autonomous learning algorithm based on implicit and bidirectional curriculum generation. We have shown that our method outperforms previous methods, both in terms of sample efficiency and final average success rate. However, our method still requires minimal human input for specifying sparse rewards. We aim to further develop our approach by transitioning to a reward-free setting to enable more autonomous training of the agent.

REFERENCES

[1] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[2] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," in *International Conference on Learning Representations*, 2019.

[3] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[4] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, *et al.*, "Magnetic control of tokamak plasmas through deep reinforcement learning," *Nature*, vol. 602, no. 7897, pp. 414–419, 2022.

[5] V. Kumar, E. Todorov, and S. Levine, "Optimal control with learned local models: Application to dexterous manipulation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 378–383.

[6] S. Ha, P. Xu, Z. Tan, S. Levine, and J. Tan, "Learning to walk in the real world with minimal human effort," *arXiv preprint arXiv:2002.08550*, 2020.

[7] A. Nagabandi, K. Konolige, S. Levine, and V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Conference on Robot Learning*. PMLR, 2020, pp. 1101–1112.

[8] B. Eysenbach, S. Gu, J. Ibarz, and S. Levine, "Leave no trace: Learning to reset for safe and autonomous reinforcement learning," *arXiv preprint arXiv:1711.06782*, 2017.

[9] J. Kim, J. hyeon Park, D. Cho, and H. J. Kim, "Automating reinforcement learning with example-based resets," *IEEE Robotics and Automation Letters*, 2022.

[10] A. Sharma, K. Xu, N. Sardana, A. Gupta, K. Hausman, S. Levine, and C. Finn, "Autonomous reinforcement learning: Formalism and benchmarking," *arXiv preprint arXiv:2112.09605*, 2021.

[11] A. Sharma, R. Ahmad, and C. Finn, "A state-distribution matching approach to non-episodic reinforcement learning," *arXiv preprint arXiv:2205.05212*, 2022.

[12] A. S. Chen, A. Sharma, S. Levine, and C. Finn, "You only live once: Single-life reinforcement learning," *arXiv preprint arXiv:2210.08863*, 2022.

[13] H. Zhu, J. Yu, A. Gupta, D. Shah, K. Hartikainen, A. Singh, V. Kumar, and S. Levine, "The ingredients of real-world robotic reinforcement learning," *arXiv preprint arXiv:2004.12570*, 2020.

[14] J. D. Co-Reyes, S. Sanjeev, G. Berseth, A. Gupta, and S. Levine, "Ecological reinforcement learning," *arXiv preprint arXiv:2006.12478*, 2020.

[15] C. Florensa, D. Held, X. Geng, and P. Abbeel, "Automatic goal generation for reinforcement learning agents," in *International conference on machine learning*. PMLR, 2018, pp. 1515–1528.

[16] Y. Luo, H. Xu, Y. Li, Y. Tian, T. Darrell, and T. Ma, "Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees," *arXiv preprint arXiv:1807.03858*, 2018.

[17] K. Asadi, D. Misra, and M. Littman, "Lipschitz continuity in model-based reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 264–273.

[18] Z. Ren, K. Dong, Y. Zhou, Q. Liu, and J. Peng, "Exploration via hindsight goal generation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[20] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[21] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, 1st ed. Englewood Cliffs, NJ: Prentice Hall, 1993.

[22] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[23] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[24] A. Sharma, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Autonomous reinforcement learning via subgoal curricula," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 474–18 486, 2021.

[25] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," *arXiv preprint arXiv:1707.01495*, 2017.