Proximal Curriculum with Task Correlations for Deep Reinforcement Learning

Georgios Tzannetos[1] , **Parameswaran Kamalaruban**[2] and **Adish Singla**[1]
[1]Max Planck Institute for Software Systems
[2]The Alan Turing Institute
{gtzannet, adishs}@mpi-sws.org, kparameswaran@turing.ac.uk

Abstract

Curriculum design for reinforcement learning (RL) can speed up an agent's learning process and help it learn to perform well on complex tasks. However, existing techniques typically require domain-specific hyperparameter tuning, involve expensive optimization procedures for task selection, or are suitable only for specific learning objectives. In this work, we consider curriculum design in contextual multi-task settings where the agent's final performance is measured w.r.t. a *target distribution over complex tasks*. We base our curriculum design on the Zone of Proximal Development concept, which has proven to be effective in accelerating the learning process of RL agents for *uniform distribution over all tasks*. We propose a novel curriculum, PROCURL-TARGET, that effectively balances the need for selecting tasks that are not too difficult for the agent while progressing the agent's learning toward the target distribution via leveraging task correlations. We theoretically justify the task selection strategy of PROCURL-TARGET by analyzing a simple learning setting with REINFORCE learner model. Our experimental results across various domains with challenging target task distributions affirm the effectiveness of our curriculum strategy over state-of-the-art baselines in accelerating the training process of deep RL agents.

1 Introduction

Deep reinforcement learning (RL) has shown remarkable success in various fields such as games, continuous control, and robotics, as evidenced by recent advances in the field [Mnih *et al.*, 2015; Lillicrap *et al.*, 2015; Silver *et al.*, 2017; Levine *et al.*, 2016]. However, despite these successes, the broader application of RL in real-world domains is often very limited. Specifically, training RL agents in complex environments, such as contextual multi-task settings and goal-based tasks with sparse rewards, still presents significant challenges [Kirk *et al.*, 2021; Andrychowicz and others, 2017; Florensa *et al.*, 2017; Riedmiller *et al.*, 2018].

Curriculum learning has been extensively studied in the context of supervised learning [Weinshall *et al.*, 2018; Zhou and Bilmes, 2018; Elman, 1993; Bengio *et al.*, 2009]. Recent research has explored the benefits of using curriculum learning in sequential decision making settings, such as reinforcement learning and imitation learning [Florensa *et al.*, 2017; Riedmiller *et al.*, 2018; Wöhlke *et al.*, 2020; Florensa *et al.*, 2018; Racanière *et al.*, 2020; Klink *et al.*, 2020; Eimer *et al.*, 2021; Kamalaruban *et al.*, 2019; Yengera *et al.*, 2021]. The objective of curriculum design in RL is to speed up an agent's learning process and enable it to perform well on complex tasks by exposing it to a personalized sequence of tasks [Narvekar *et al.*, 2020; Portelas *et al.*, 2021]. To achieve this objective, several works have proposed different curriculum strategies based on different design principles, such as the Zone of Proximal Development (ZPD) [Vygotsky and Cole, 1978; Chaiklin, 2003], Self-Paced Learning (SPL) [Kumar *et al.*, 2010; Jiang *et al.*, 2015], and Unsupervised Environment Design (UED) [Dennis *et al.*, 2020]. However, existing techniques typically require domain-specific hyperparameter tuning, involve expensive optimization procedures for task selection, or are suitable only for specific learning objectives, such as uniform performance objectives.

In this work, we investigate curriculum design in contextual multi-task settings with varying degrees of task similarity, where the agent's final performance is measured w.r.t. a *target distribution over complex tasks*. We base our curriculum design on the Zone of Proximal Development concept, which has proven to be effective in accelerating the learning process of RL agents for *uniform distribution over all tasks* [Florensa *et al.*, 2017; Wöhlke *et al.*, 2020; Florensa *et al.*, 2018; Tzannetos *et al.*, 2023]. We propose a novel curriculum strategy, PROCURL-TARGET, that effectively balances the need for selecting tasks that are neither too hard nor too easy for the agent (according to the ZPD concept) while still progressing its learning toward the target distribution via leveraging task correlations. We have mathematically derived our curriculum strategy by analyzing a specific learning setting. The strengths of our curriculum strategy include its broad applicability to many domains with minimal hyperparameter tuning, computational and sample efficiency, easy integration with deep RL algorithms, and applicability to any target distribution over tasks, not just uniform distribution. Our main results and contributions are as follows:

I. We propose a curriculum strategy, PROCURL-

TARGET, that effectively trades off the suitable task difficulty level for the agent and the progression towards the target tasks (Section 3).

II. We mathematically derive PROCURL-TARGET for the single target task setting with a discrete pool of tasks by analyzing the effect of picking a task on the agent's learning progress (Section 3.1).

III. We propose an extension of PROCURL-TARGET that can be applied to a wide range of task spaces and target distributions. This extension can be seamlessly integrated with deep RL frameworks, making it easy to use and apply in various scenarios (Section 3.2).

IV. We empirically demonstrate that the curricula generated with PROCURL-TARGET significantly improve the training process of deep RL agents in various environments, matching or outperforming existing state-of-the-art baselines (Section 4).[1]

1.1 Related Work

Curriculum strategies based on SPL concept. In the realm of supervised learning, curriculum strategies leveraging the SPL concept attempt to strike a balance between exposing the learner to all available training examples and selecting examples in which it currently performs well [Kumar *et al.*, 2010; Jiang *et al.*, 2015]. In the context of RL, the SPL concept has been adapted by researchers in SPDL [Klink *et al.*, 2020; Klink *et al.*, 2021], SPACE [Eimer *et al.*, 2021], and CURROT [Klink *et al.*, 2022] by controlling the intermediate task distribution with respect to the learner's current training progress. While both SPDL and CURROT involve a setting where the learner's performance is measured w.r.t. a target distribution over the task space (similar to our objective), SPACE operates in a setting where the learner's performance is measured w.r.t. a uniform distribution over the task space. The task selection mechanism varies across these methods. SPDL and CURROT operate by solving an optimization problem at each step to select the most relevant task [Klink *et al.*, 2021; Klink *et al.*, 2022]. On the other hand, SPACE relies on ranking tasks based on the magnitude of differences in current/previous critic values to choose the task for the next step [Eimer *et al.*, 2021]. Furthermore, the work of CURROT [Klink *et al.*, 2022] showcases issues about using KL divergence to measure the similarity between task distributions as used in SPDL – instead, they introduce an alternative approach by posing the curriculum design as a constrained optimal transport problem between task distributions.

Curriculum strategies based on UED concept. The UED problem setting involves automatically designing a distribution of environments that adapts to the learning agent [Dennis *et al.*, 2020]. UED represents a self-supervised RL paradigm in which an environment generator evolves alongside a student policy to develop an adaptive curriculum learning approach. This approach can be utilized to create increasingly complex environments for training a policy, leading to the emergence of Unsupervised Curriculum Design. PAIRED [Dennis *et al.*, 2020] is an adversarial training technique that solves the problem of the adversary generating unsolvable environments by introducing an antagonist who works with the environment-generating adversary to design environments in which the protagonist receives a low reward. Furthermore, the connections between UED and another related method called PLR [Jiang *et al.*, 2021b] have been explored in [Jiang *et al.*, 2021a; Parker-Holder *et al.*, 2022], resulting in demonstrated improvements over PAIRED. PLR, originally designed for procedural content generation based environments, samples tasks/levels by prioritizing those with higher estimated learning potential when revisited in the future. TD errors are used to estimate a task's future learning potential. Unlike [Jiang *et al.*, 2021a; Parker-Holder *et al.*, 2022], PLR does not assume control over the environment generation process, requiring only a black box generation process that returns a task given an identifier.

Curriculum strategies based on ZPD concept. Effective teaching provides tasks of moderate difficulty (neither too hard nor too easy) for the learner, as formalized by the ZPD concept [Vygotsky and Cole, 1978; Chaiklin, 2003; Oudeyer *et al.*, 2007; Baranes and Oudeyer, 2013; Zou *et al.*, 2019]. In the context of RL, several curriculum strategies are based on the ZPD concept, such as selecting the next task randomly from a set of tasks with success rates within a specific range [Florensa *et al.*, 2017; Florensa *et al.*, 2018]. However, the threshold values for success rates require tuning based on the learner's progress and domain. A unified framework for performance-based starting state curricula in RL is proposed by [Wöhlke *et al.*, 2020], while [Tzannetos *et al.*, 2023] propose a broadly applicable ZPD-based curriculum strategy with minimal hyperparameter tuning and theoretical justifications. Nonetheless, these techniques are generally suitable only for settings where the learner's performance is evaluated using a uniform distribution over all tasks.

Curriculum strategies based on domain knowledge. In supervised learning, early works involve ordering examples by increasing difficulty [Elman, 1993; Bengio *et al.*, 2009; Schmidhuber, 2013], which has been adapted in hand-crafted RL curriculum approaches [Wu and Tian, 2016]. Recent works on imitation learning have also utilized iterative machine teaching framework to design greedy curriculum strategies [Kamalaruban *et al.*, 2019; Yengera *et al.*, 2021; Liu *et al.*, 2017; Zhu *et al.*, 2018]. However, these approaches require domain-specific expert knowledge.

Other automatic curriculum strategies. Various automatic curriculum generation approaches exist, including: (i) formulating the curriculum design problem as a meta-level Markov Decision Process [Narvekar and Stone, 2019]; (ii) learning to generate training tasks similar to a teacher [Dendorfer *et al.*, 2020; Matiisen *et al.*, 2019; Turchetta *et al.*, 2020]; (iii) using self-play for curriculum generation [Sukhbaatar *et al.*, 2018]; (iv) leveraging disagreement between different agents trained on the same tasks [Zhang *et al.*, 2020]; and (v) selecting starting states based on a single demonstration [Salimans and Chen, 2018]. Interested readers can refer to recent surveys on RL curriculum design [Narvekar *et al.*, 2020; Portelas *et al.*, 2021].

[1] Github repository: https://github.com/machine-teaching-group/ijcai2024-proximal-curriculum-target-rl

Algorithm 1 RL Agent Training as Interaction between Teacher-Student Components

1: **Input:** RL agent's initial policy π_1
2: **for** $t = 1, 2, \dots$ **do**
3: Teacher component picks a task $c_t \in \mathcal{C}$.
4: Student component attempts the task via a trajectory rollout ξ_t using the policy π_t in $\mathcal{M}_{c_t}$.
5: Student component updates the policy to π_{t+1} using the rollout ξ_t.
6: **Output:** RL agent's final policy $\pi_{\text{end}} \leftarrow \pi_{t+1}$.

2 Formal Setup

We formalize our problem setting based on prior work on teacher-student curriculum learning [Matiisen *et al.*, 2019].

Multi-task RL. We consider a multi-task RL setting with a task/context space $\mathcal{C}$, in which each task $c \in \mathcal{C}$ is associated with a learning environment modeled as a contextual Markov Decision Process (MDP), denoted by $\mathcal{M}_c := \big(\mathcal{S}, \mathcal{A}, \gamma, \mathcal{T}_c, R_c, P_c^0\big)$ [Hallak *et al.*, 2015; Modi *et al.*, 2018]. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are shared by all tasks in $\mathcal{C}$, as well as the discount factor γ. Each contextual MDP includes a contextual transition dynamics $\mathcal{T}_c : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a contextual reward function $R_c : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$, where $R_{\max} > 0$, and a contextual initial state distribution $P_c^0 : \mathcal{S} \to [0, 1]$. We denote the space of environments by $\mathcal{M} = \{\mathcal{M}_c : c \in \mathcal{C}\}$. Moreover, we have a target distribution μ over $\mathcal{C}$ that is used for performance evaluation, as further discussed below.

RL agent and training process. We consider an RL agent acting in any environment $\mathcal{M}_c \in \mathcal{M}$ via a contextual policy $\pi : \mathcal{S} \times \mathcal{C} \times \mathcal{A} \to [0, 1]$ that is a contextual mapping from a state to a probability distribution over actions. Given a task $c \in \mathcal{C}$, the agent attempts the task via a trajectory rollout obtained by executing its policy π in the MDP $\mathcal{M}_c$. The trajectory rollout is denoted as $\xi = \big\{(s^{(\tau)}, a^{(\tau)})\big\}_{\tau=0,1,\dots}$ with $s^{(0)} \sim P_c^0$. The agent's performance on task c is measured by the value function $V^\pi(c) := \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \cdot R_c(s^{(\tau)}, a^{(\tau)}) \middle| \pi, \mathcal{M}_c\right]$. The agent training corresponds to finding a policy that performs well w.r.t. the target distribution μ, i.e., $\max_\pi V_\mu^\pi$ where $V_\mu^\pi := \mathbb{E}_{c\sim\mu}\left[V^\pi(c)\right]$. The training process of the agent involves an interaction between two components: a student component that is responsible for policy updates and a teacher component that is responsible for task selection. The interaction happens in discrete steps indexed by $t = 1, 2, \dots$, and is formally described in Algorithm 1. Let π_{end} denote the agent's final policy at the end of teacher-student interaction. The *training objective* is to ensure that the performance of the policy π_{end} is ϵ-near-optimal, i.e., $(\max_\pi V_\mu^\pi - V_\mu^{\pi_{\text{end}}}) \leq \epsilon$.

Student component. We consider parametric policies of the form $\pi_\theta : \mathcal{S} \times \mathcal{C} \times \mathcal{A} \to [0, 1]$, where $\theta \in \Theta \subseteq \mathbb{R}^d$. The agent's policy at step t is given by $\pi_t := \pi_{\theta_t}$. The student component updates the policy parameter based on the following quantities: the current parameter θ_t, the task c_t picked by the teacher component, and the rollout $\xi_t = \left\{(s_t^{(\tau)}, a_t^{(\tau)})\right\}_\tau$. For example, the policy parameter of the REINFORCE agent [Sutton *et al.*, 1999] is updated as follows: $\theta_{t+1} \leftarrow \theta_t + \eta_t \cdot \sum_{\tau=0}^{\infty} G_t^{(\tau)} \cdot \left[\nabla_\theta \log \pi_\theta(a_t^{(\tau)} | s_t^{(\tau)}, c_t)\right]_{\theta=\theta_t}$, where η_t is the learning rate, and $G_t^{(\tau)} = \sum_{\tau'=\tau}^{\infty} \gamma^{\tau'-\tau} \cdot R_{c_t}(s_t^{(\tau')}, a_t^{(\tau')})$.

Teacher component. At step t, the teacher component selects a task c_t for the student component to attempt via a trajectory rollout, as shown in line 3 in Algorithm 1. The sequence of tasks, also known as the curriculum, that is chosen by the teacher component has a significant impact on the performance improvement of the policy π_t. The primary objective of this work is to develop a teacher component to achieve the training objective in a computationally efficient and sample-efficient manner.

3 Our Curriculum Strategy

In Section 3.1, we mathematically derive a curriculum strategy for the single target task setting with a discrete pool of tasks. Then, in Section 3.2, we present our final curriculum strategy that is applicable in general learning settings. The proofs are provided in the longer version of the paper.

3.1 Curriculum Strategy for Single Target Settings

In this section, we present our curriculum strategy for a setting where the task space $\mathcal{C}$ is a discrete set and the target distribution μ is a delta distribution concentrated on a single target task c_{targ}. To design our curriculum strategy, we investigate the effect of selecting a task c_t at time step t on the agent's performance $V_\mu^{\pi_{\theta_t}}$ and its convergence towards the target performance $V_\mu^* := \max_\pi V_\mu^\pi$. Therefore, we define the training objective improvement at step t and analyze this metric across both general and specific learning scenarios.

Expected improvement in the training objective. At step t, given the current policy parameter θ_t, the task c_t picked by the teacher component, and the student component's rollout ξ_t, we define the improvement in the training objective as:

$$\Delta_t(\theta_{t+1} \big| \theta_t, c_t, \xi_t) \;:=\; (V_\mu^* - V_\mu^{\pi_{\theta_t}}) - (V_\mu^* - V_\mu^{\pi_{\theta_{t+1}}}).$$

Additionally, we define the expected improvement in the training objective at step t due to picking the task c_t as follows [Weinshall *et al.*, 2018; Kamalaruban *et al.*, 2019; Yengera *et al.*, 2021; Graves *et al.*, 2017]:

$$I_t(c_t) \;:=\; \mathbb{E}_{\xi_t | c_t}\left[\Delta_t(\theta_{t+1} | \theta_t, c_t, \xi_t)\right].$$

Based on the above measure, a natural greedy curriculum strategy for selecting the next task c_t is given by:

$$c_t \;\leftarrow\; \arg\max_{c \in \mathcal{C}} I_t(c). \tag{1}$$

We aim to approximate such a curriculum strategy without computing the updated policy $\pi_{\theta_{t+1}}$. To this end, we initially analyze the function $I_t(\cdot)$ for REINFORCE learner model within a general learning setting. Subsequently, we refine and simplify the findings by delving into a specific learning scenario. This analysis enables us to develop an intuitive curriculum strategy by effectively combining the following fundamental factors: (i) the learning potential inherent in the source and target tasks, and (ii) the transfer potential between the source and target tasks, i.e., their similarity.

Gradient alignment approximation of $I_t(\cdot)$. Here, we analyze the function $I_t(\cdot)$ within the context of the REINFORCE learner model operating under a general learning setting. We show that the natural greedy curriculum strategy can be approximated by a simple gradient alignment maximization strategy. Initially, through the application of the first-order Taylor approximation of $V^{\pi_{\theta_{t+1}}}(c_{\text{targ}})$ at θ_t, we approximate the improvement in the training objective as follows:

$$\begin{aligned}\Delta_t(\theta_{t+1}|\theta_t, c_t, \xi_t) \;&=\; V^{\pi_{\theta_{t+1}}}(c_{\text{targ}}) - V^{\pi_{\theta_t}}(c_{\text{targ}}) \\ &\approx\; \langle \theta_{t+1} - \theta_t, g_t(c_{\text{targ}}) \rangle ,\end{aligned}$$

where $g_t(c) := [\nabla_\theta V^{\pi_\theta}(c)]_{\theta=\theta_t}$. Subsequently, by utilizing the parameter update form of the REINFORCE agent (i.e., $\mathbb{E}_{\xi_t|c_t}[\theta_{t+1}] \leftarrow \theta_t + \eta_t \cdot g_t(c_t)$), we approximate the expected improvement in the training objective as follows:

$$\begin{aligned}I_t(c_t) \;&\approx\; \left\langle \mathbb{E}_{\xi_t|c_t}[\theta_{t+1} - \theta_t], g_t(c_{\text{targ}}) \right\rangle \\ &=\; \eta_t \cdot \langle g_t(c_t), g_t(c_{\text{targ}}) \rangle .\end{aligned}$$

Consequently, the natural greedy curriculum strategy in Eq. (1) can be effectively approximated by the following gradient-alignment-based curriculum strategy:

$$c_t \;\leftarrow\; \arg\max_{c} \langle g_t(c), g_t(c_{\text{targ}}) \rangle . \qquad (2)$$

In the following theorem, we demonstrate the effectiveness of employing the above curriculum strategy in accelerating the convergence of the REINFORCE agent.

Theorem 1. *Consider Algorithm 1 with the* REINFORCE *learner model and the curriculum strategy defined in Eq.* (2). *Then, after $t = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ steps, we have:*

$$\mathbb{E}\left[(V^*(c_{\text{targ}}) - V^{\pi_{\theta_t}}(c_{\text{targ}})) \mid \theta_0\right] \leq \epsilon,$$

where $V^*(c) := \max_\pi V^\pi(c)$.

Subsequently, building on the curriculum strategy outlined in Eq. (2), we devise an intuitive curriculum strategy through an analysis of the curriculum objective $\langle g_t(c), g_t(c_{\text{targ}}) \rangle$ within the context of a contextual bandit setting.

Further simplification of gradient alignment. We consider the REINFORCE learner model with the following policy parameterization: given a feature mapping $\phi : \mathcal{S} \times \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, for any $\theta \in \mathbb{R}^d$, we parameterize the policy as $\pi_\theta(a|s,c) = \frac{\exp(\langle \theta, \phi(s,c,a) \rangle)}{\sum_{a'} \exp(\langle \theta, \phi(s,c,a') \rangle)}, \forall s \in \mathcal{S}, c \in \mathcal{C}, a \in \mathcal{A}$. In the following, we consider a specific problem instance of contextual MDP setting. Let $\mathcal{M}_c$ be a contextual MDP with a singleton state space $\mathcal{S} = \{s\}$, and an action space $\mathcal{A} = \{a_1, a_2\}$. Any action $a \in \mathcal{A}$ taken from the initial state $s \in \mathcal{S}$ always leads to a terminal state. Let $r : \mathcal{C} \to [0,1]$ be a mapping from task/context space $\mathcal{C}$ to the interval $[0,1]$. For any context $c \in \mathcal{C}$, we denote the optimal and non-optimal actions for that context as a_c^{opt} and a_c^{non}, respectively. The contextual reward function is defined as follows: $R_c(s, a_c^{\text{opt}}) = 1$, and $R_c(s, a_c^{\text{non}}) = 0$, for all $c \in \mathcal{C}$. Further, we define $\psi : \mathcal{C} \to \mathbb{R}^d$ as $\psi(c) := (\phi(s,c,a_c^{\text{opt}}) - \phi(s,c,a_c^{\text{non}}))$. Subsequently, for the REINFORCE agent operating under the above setting, the following proposition quantifies the objective term of the curriculum strategy as per Eq. (2) at step t:

Proposition 1. *For the* REINFORCE *agent with softmax policy parameterization under the contextual bandit setting described above, we have:*

$$\langle g_t(c), g_t(c_{\text{targ}}) \rangle \;=\; \eta_t \cdot Z_t(c) \cdot Z_t(c_{\text{targ}}) \cdot \langle \psi(c), \psi(c_{\text{targ}}) \rangle ,$$

where $Z_t(c) := \frac{V^{\pi_{\theta_t}}(c)}{V^*(c)} \cdot \left(V^*(c) - V^{\pi_{\theta_t}}(c)\right)$ denotes the agent's learning potential on task c at step t.

Our curriculum strategy. Inspired by the above analysis, we propose the following curriculum strategy:

$$c_t \;\leftarrow\; \arg\max_{c \in \mathcal{C}} \underbrace{Z_t(c)}_{Ⓐ} \cdot \underbrace{Z_t(c_{\text{targ}})}_{Ⓑ} \cdot \underbrace{\langle \psi(c), \psi(c_{\text{targ}}) \rangle}_{Ⓒ}, \qquad (3)$$

where $\psi : \mathcal{C} \to \mathbb{R}^d$ is a context representation mapping. At step t, the teacher component picks a task c_t according to Eq. (3). The curriculum strategy involves the following quantities: Ⓐ the agent's learning potential on task c, Ⓑ the agent's learning potential on task c_{targ}, and Ⓒ the correlation between the tasks c and c_{targ}. Term Ⓐ enforces the selection of tasks that are neither too hard nor too easy for the current policy, aligning with the ZPD principle. The combined effect of terms Ⓑ and Ⓒ emphasizes the choice of tasks highly correlated with the target task (which has high learning potential). The curriculum strategy effectively balances these two objectives.

3.2 Curriculum Strategy for General Settings

In this section, we extend the curriculum strategy in Eq. (3) to practical settings of interest, i.e., a general task space $\mathcal{C}$, a general target distribution μ, and $V^*(c)$ values being unknown. We begin by constructing two large discrete sets, $\widehat{\mathcal{C}}_{\text{unif}}$ and $\widehat{\mathcal{C}}_{\text{targ}}$, which are subsets of the original task space $\mathcal{C}$. $\widehat{\mathcal{C}}_{\text{unif}}$ is obtained by sampling contexts from $\mathcal{C}$ according to uniform distribution, while $\widehat{\mathcal{C}}_{\text{targ}}$ is obtained by sampling contexts from $\mathcal{C}$ according to the target distribution μ. For the general setting, we consider the following curriculum strategy:

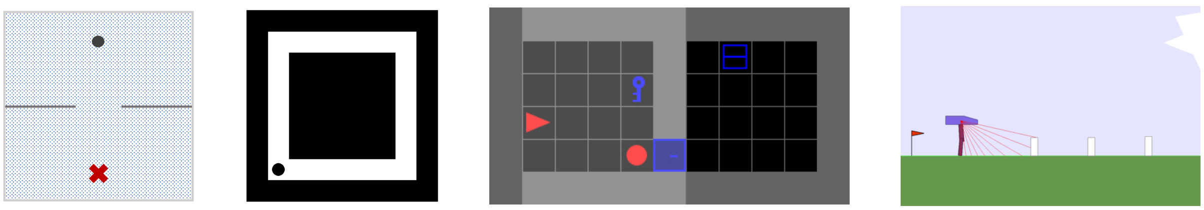
$$(c_{\text{targ}}^t, c_t) \leftarrow \arg\max_{(c_{\text{targ}}, c) \in \widehat{\mathcal{C}}_{\text{targ}} \times \widehat{\mathcal{C}}_{\text{unif}}} Z_t(c) \cdot Z_t(c_{\text{targ}}) \cdot \langle \psi(c), \psi(c_{\text{targ}}) \rangle .$$

Next, we replace $V^*(\cdot)$ with $V_{\max}$, i.e., the maximum possible value that can be achieved for any task in the task space – this value can typically be obtained for a given domain. Further, when training deep RL agents, allowing some stochasticity in task selection is useful. In particular, the $\arg\max$ selection can be problematic in the presence of any approximation errors while computing $V^{\pi_{\theta_t}}(\cdot)$ values. To make the selection more robust, we replace $\arg\max$ selection with softmax selection and sample (c_{targ}^t, c_t) from the distribution given below:

$$\begin{aligned}\mathbb{P}\big[(c_{\text{targ}}^t, c_t) = (c_{\text{targ}}, c)\big] \;\propto\; \exp\Big(\beta \cdot \frac{V^t(c)}{V_{\max}} \cdot \big(V_{\max} - V^t(c)\big) \\ \cdot \frac{V^t(c_{\text{targ}})}{V_{\max}} \cdot \big(V_{\max} - V^t(c_{\text{targ}})\big) \cdot \langle \psi(c), \psi(c_{\text{targ}}) \rangle \Big), \qquad (4)\end{aligned}$$

	PM-S:1T	PM-S:2G	SGR	MINIG	BW
Reward	binary	binary	binary	binary	dense
Context	$\mathbb{R}^3$	$\mathbb{R}^3$	$\mathbb{R}^3$	$\{0,1\}^8$	$\mathbb{R}^2$
State	$\mathbb{R}^4$	$\mathbb{R}^4$	$\mathbb{R}^4$	$[0,255]^{147}$	$\mathbb{R}^{24}$
Action	$\mathbb{R}^2$	$\mathbb{R}^2$	$\mathbb{R}^2$	7	$\mathbb{R}^4$
Target Dist.	Single Task	Double-Mode Gaussian	$\mathbb{R}^2$ Plane	Single Task	Uniform with trivial tasks

(a) Complexity of environments



(b) Illustration of the environments

Figure 1: **(a)** provides a comprehensive overview of the complexity of the environments based on the reward signals, context space, state space, action space, and target distribution. **(b)** showcases the environments by providing an illustrative visualization of each environment (from left to right): PM-S, SGR, MINIG, and BW.

where β is a hyperparameter and $V^t(\cdot)$ values are obtained from the critic network of the RL agent to estimate $V^{\pi_{\theta_t}}(\cdot)$. Finally, the teacher component samples (c^t_{targ}, c_t) from the above distribution and provides the task c_t to the student component – we refer to this selection strategy as PROCURL-TARGET.

4 Experimental Evaluation

In this section, we validate the effectiveness of our curriculum strategy by conducting experiments in environments selected from the state-of-the-art works of [Klink *et al.*, 2022] and [Romac *et al.*, 2021]. We utilize the PPO method from the Stable-Baselines3 library for policy optimization [Schulman *et al.*, 2017; Raffin *et al.*, 2021]. The implementation details of different curriculum strategies are provided in the longer version of the paper.

4.1 Environments

In our evaluation, we examine four distinct environments detailed in the following paragraphs. These environments are selected to showcase the utility of our curriculum strategy in diverse settings, such as in procedural task generation or with image-based observations, and its effectiveness in handling target distributions with varying characteristics within the context space $\mathcal{C}$. For the first environment, *Point Mass Sparse* (PM-S), we consider two settings. In one setting, the target is concentrated on a single context $c \in \mathcal{C}$, a similar setting as analyzed in Section 3.1. In the second setting, the target distribution exhibits multiple modalities. The second environment, *Sparse Goal Reaching* (SGR), features target distributions with uniform coverage over specific dimensions of the context space and concentrated on one dimension. The third environment, MiniGrid (MINIG), uniquely features a discrete context space and additionally has an image-based state space. Lastly, a fourth environment, Bipedal Walker Stump tracks (BW), has a uniform target distribution spanning the entirety of the context space. Moreover, it shows the applicability of our technique in the procedural task generation domain. A summary and illustration of these environments are presented in Figure 1.

***Point Mass Sparse* (PM-S).** Based on the work of [Klink *et al.*, 2020], we consider a contextual PM-S environment where an agent navigates a point mass through a gate of a given size towards a goal in a two-dimensional space. To heighten the challenge, we replace the original dense reward function with a sparse one, a strategy also considered in [Tzannetos *et al.*, 2023]. Specifically, in the PM-S environment, the agent operates within a goal-based reward setting where the reward is binary and sparse, i.e., the agent receives a reward of 1 only upon successfully moving the point mass to the goal position. The parameters governing this environment, such as the gate's position, width, and the ground's friction coefficient, are controlled by a contextual variable $c \in \mathcal{C} \subseteq \mathbb{R}^3$. This variable comprises *C-GatePosition*, *C-GateWidth*, and *C-Friction*. Our experimental section explores two distinct PM-S environment settings. In the first setting, denoted as PM-S:2G, the target distribution μ takes the form of a bimodal Gaussian distribution. Here, the means of the contextual variables $\big[$*C-GatePosition*, *C-GateWidth*$\big]$ are set to $\big[-3.9, 0.5\big]$ and $\big[3.9, 0.5\big]$ for the two modes, respectively. In the second setting, PM-S:1T, the target distribution μ is concentrated on a single context $c \in \mathcal{C}$. More precisely, the contextual variables $\big[$*C-GatePosition*, *C-GateWidth*, *C-Friction*$\big]$ take on the following values: $\left[0.9, 0.5, 3.5\right]$. To construct our training tasks, we draw 20000 contexts from $\mathcal{C}$ using a uniform distribution, forming $\widehat{\mathcal{C}}_{\text{unif}}$. The set $\widehat{\mathcal{C}}_{\text{targ}}$ is created by sampling 400 contexts from $\mathcal{C}$ according to the target distribution μ. We employ a held-out set sampled from the target μ for evaluation.

Sparse Goal Reaching **(SGR).** Based on the work of [Klink *et al.*, 2022], we consider a sparse-reward, goal-reaching environment in which an agent needs to reach a desired position with high precision. Such environments have previously been studied by [Florensa *et al.*, 2018]. Within this environment, the contexts, denoted as $c \in \mathcal{C} \subseteq \mathbb{R}^3$, encode both the desired 2D goal position and the acceptable tolerance for reaching that goal. Our primary objective centers around achieving as many goals as possible with high precision, indicated by a low tolerance threshold. In this regard, the target distribution μ takes the form of a uniform distribution, but it is restricted to a specific 2D region within $\mathcal{C}$ where the tolerance (*C-Tolerance*) for each context is set at a minimal value of 0.05. Additionally, the presence of walls within the environment renders many of the tasks specified by $\mathcal{C}$ infeasible, necessitating the identification of a feasible task subspace. We generate our training tasks by randomly selecting 9900 contexts from $\mathcal{C}$ using uniform distribution to create $\widehat{\mathcal{C}}_{\text{unif}}$, and by selecting 100 contexts according to the target distribution μ to form $\widehat{\mathcal{C}}_{\text{targ}}$. For evaluation, we employ a separate held-out set sampled from the target distribution μ.

MiniGrid **(MiniG).** We establish the MINIG environment by assembling six diverse Minigrid environments from [Chevalier-Boisvert *et al.*, 2023]: Crossing, Dynamic Obstacles, Four Rooms, Unlock, Unlock Pickup, and Blocked Unlock Pickup. Each environment presents a unique mission, demanding distinct skills such as navigation, goal-reaching, lava avoidance, moving obstacle avoidance, key picking, door unlocking, object picking, and door unblocking. These skills define the discrete context space $\{0, 1\}^8$ of MINIG, with each environment requiring a specific subset of skills for a successful resolution. The type Blocked Unlock Pickup is chosen as the target environment due to its inherent difficulty, making it challenging to solve without a curriculum. Additionally, MINIG includes environments like Crossing and Dynamic Obstacles, featuring skills not pertinent to the target mission. The state space comprises observed images of the grid world, and the action space is discrete. The reward is set at 1 for successful mission completion and 0 otherwise. For training tasks, we select 1000 instances from all six environment types. The first three types (Crossing, Dynamic Obstacles, and Four Rooms) collectively contribute to 75% of the training tasks, while the remaining three types, including samples from the target environment, equally constitute the remaining 25%.

Bipedal Walker Stump Tracks **(BW).** We conduct additional experiments within the TeachMyAgent benchmark for curriculum techniques, as introduced in [Romac *et al.*, 2021]. In this context, we chose a bipedal agent tasked with walking in the Stump Tracks environment, which is an extension of the environment initially proposed in [Portelas *et al.*, 2019]. The state space comprises lidar sensors, head position, and joint positions. The action space is continuous, and the goal is to learn a policy that controls the torque of the agent's motors. The walker is rewarded for going forward and penalized for torque usage. An episode lasts 2000 steps and is terminated if the agent reaches the end of the track or if its head collides with the environment (in which case a reward of -100 is received). Within this environment, the contextual variables $c \in \mathcal{C} \subseteq \mathbb{R}^2$ control the height (*C-StumpHeight*) and spacing (*C-StumpSpacing*) of stumps placed along the track for each task. Our experimental setup is equivalent to the bipedal walker stump track environment with mostly trivial tasks, as described in [Romac *et al.*, 2021]. In this setup, *C-StumpHeight* is constrained to the range $\left[-3; 3\right]$, while *C-StumpSpacing* remains within $\left[0; 6\right]$. Notably, the environment enforces the clipping of negative values for *C-StumpHeight*, setting them to 0. Consequently, half of the tasks have a mean stump height of 0, introducing a significant proportion of trivial tasks (50%). To address the procedural task generation, we randomly draw 1000 tasks from $\mathcal{C}$ to construct the training task set, denoted as $\widehat{\mathcal{C}}_{\text{unif}}$. Additionally, every four epochs, we resample 1000 tasks and update the training set $\widehat{\mathcal{C}}_{\text{unif}}$. The set $\widehat{\mathcal{C}}_{\text{targ}}$ is obtained by sampling 500 tasks from $\mathcal{C}$ according to the target distribution μ, which is uniform in $\mathcal{C}$.

4.2 Curriculum Strategies Evaluated

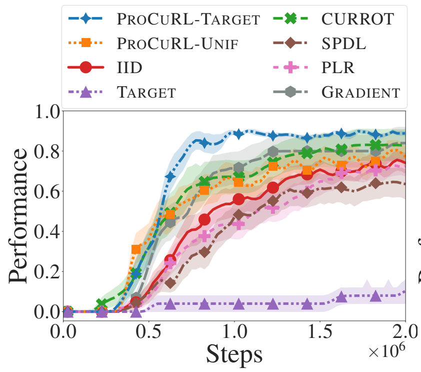
Variants of our curriculum strategy. We consider two curriculum strategies as described next. First, PROCURL-TARGET is based on Eq. (4). Throughout all the experiments, we use the following choice to compute the similarity between $\psi(s)$ and $\psi(c_{\text{targ}})$: $\exp(-\left\|c - c_{\text{targ}}\right\|_2)$. Second, PROCURL-UNIF is a variant of it that does not take into account the target distribution μ and hence ignores the correlations. Specifically, PROCURL-UNIF drops the target task-related terms Ⓑ and Ⓒ derived in Eq. (3), and selects the next task according to the following distribution: $\mathbb{P}\left[c_t = c\right] \propto \exp\left(\beta \cdot \frac{V^t(c)}{V_{\max}} \cdot (V_{\max} - V^t(c))\right)$. We note that this strategy is similar to a ZPD-based curriculum strategy proposed in [Tzannetos *et al.*, 2023] for uniform performance objectives.

State-of-the-art baselines. SPDL [Klink *et al.*, 2020], CURROT [Klink *et al.*, 2022], PLR [Jiang *et al.*, 2021b], and GRADIENT [Huang *et al.*, 2022] are state-of-the-art curriculum strategies for contextual RL. We adapt the implementation of an improved version of SPDL, presented in [Klink *et al.*, 2021], to work with a discrete pool of contextual tasks. PLR [Jiang *et al.*, 2021b] was originally designed for procedurally generated content settings, but we have adapted its implementation for the contextual RL setting operating on a discrete pool of tasks.

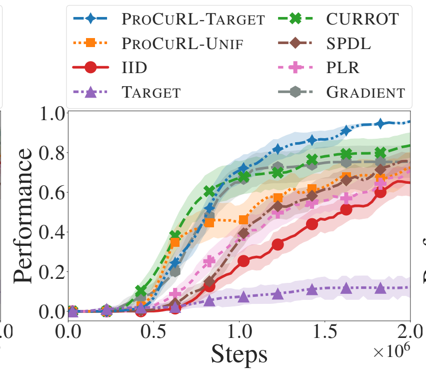
Prototypical baselines. We consider two prototypical baselines: IID and TARGET. The IID strategy samples the next task from $\mathcal{C}$ with a uniform distribution, while the TARGET strategy samples according to the target distribution μ.

4.3 Results

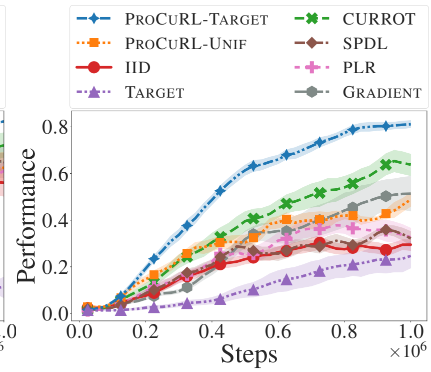
Convergence behavior. As illustrated in Figure 2, the RL agents trained using our curriculum strategy, PROCURL-TARGET, perform competitively w.r.t. those trained with



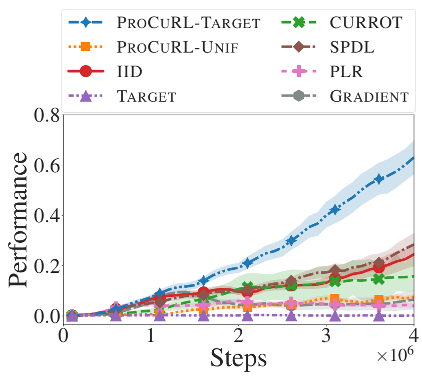
(a) PM-s:1T



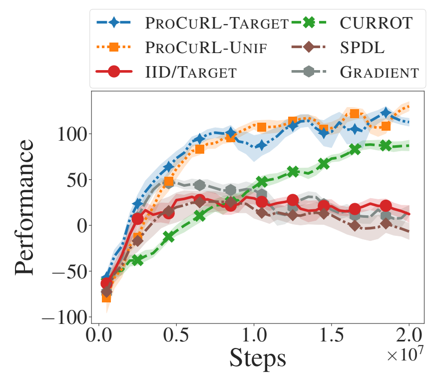
(b) PM-s:2G



(c) SGR



(d) MiniG

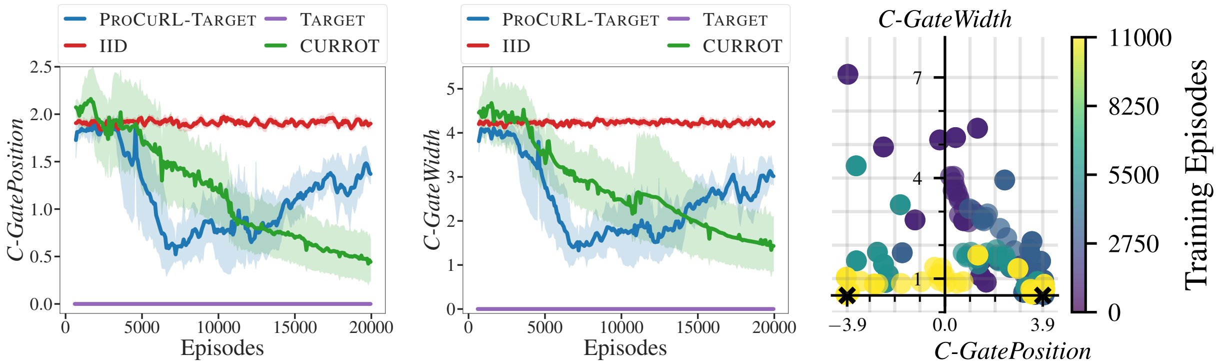


(e) BW

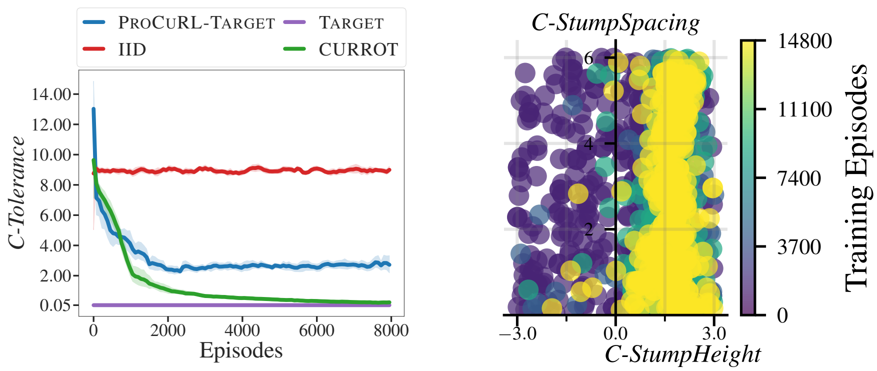
Figure 2: Performance comparison of RL agents trained using different curriculum strategies. The performance is measured as the mean return (± 1 standard error) on the test pool of tasks. The results are averaged over 25 random seeds for PM-s:1T, 25 random seeds for PM-s:2G, 10 random seeds for SGR, 20 random seeds for MiniG, and 10 random seeds for BW. The plots are smoothed across 2 evaluation snapshots that occur over 25000 training steps.

state-of-the-art and prototypical baselines. For PM-s:1T, in Figure 2a, we observe that ProCuRL-Target quickly succeeds in the single target task compared to the other techniques. Although CURROT and Gradient converge slower, they finally perform similarly to the proposed technique. The results for PM-s:2G are presented in Figure 2b, where we can observe that ProCuRL-Target, CURROT and Gradient outperform the other strategies. ProCuRL-Target demonstrates success in handling bimodal target distributions by alternating the selection between the modes of the target distribution. Although it initially has a slower performance than ProCuRL-Unif and CURROT, it quickly matches and surpasses their performance. Despite ProCuRL-Unif not explicitly considering the target distribution in its formulation, it progressively selects more challenging contexts and effectively encompasses the tasks from the target distribution in this scenario. In Figure 2c for SGR, ProCuRL-Target outperforms all the other techniques. ProCuRL-Target selects tasks that are neither too hard nor too easy for the agent's current policy and are also correlated with the target distribution. CURROT stands out among other strategies due to its ability to gradually choose tasks from the target distribution. Importantly, solely selecting target contexts for training is inadequate, as evidenced by the underperformance of Target compared to all other techniques. Similarly, in Figure 2d for MiniG, ProCuRL-Target outperforms all the other techniques by a large margin. For BW, Figure 2e, where the target distribution is uniform, ProCuRL-Target and ProCuRL-Unif achieve the best performance. Although, ProCuRL-Unif, by definition, considers a uniform performance objective, ProCuRL-Target is capable of succesfully handling a uniform target distribution.

Curriculum plots. Figures 3a and 3b display the average distance between the target and the contexts selected from ProCuRL-Target, CURROT, IID, and Target. We observe that ProCuRL-Target and CURROT manage to reduce the average context distance below that of IID, indicating that both techniques gradually prioritize tasks that align with the target. However, it is noteworthy that CURROT continues to decrease the context values to reach the target. Whereas ProCuRL-Target, after succeeding on the target, returns closer to IID sampling. Figure 3c provides a visual representation of the two-dimensional context space for the PM-s:2G setting. The curriculum initially starts from larger *C-GateWidth* values and centered *C-GatePosition* values, gradually shifting towards the two modes of the target in the later stages of training. In Figure 3d, we dis-



(a) PM-S:2G: *C-GatePosition* (b) PM-S:2G: *C-GateWidth* (c) PM-S:2G: 2D context



(d) SGR: *C-Tolerance* (e) BW: 2D context

Figure 3: **(a-b)** present the average distance between the selected contexts *C-GatePosition* and *C-GateWidth* and the target distribution for PM-S:2G. **(c)** presents the two-dimensional context space of PM-S:2G. The target distribution is depicted as a black **x** and encodes the two gates with *C-GateWidth* $= 0.5$ at *C-GatePosition* $= \{-3.9, 3.9\}$. Each colored dot represents the context/task selected by PROCURL-TARGET during training, where brighter colors indicate later training stages. **(d)** presents the average *C-Tolerance* of the selected tasks during different curriculum strategies for SGR. **(e)** presents the two-dimensional context space of BW. The target distribution is uniform. Each colored dot represents the context/task selected by PROCURL-TARGET during training, where brighter colors indicate later training stages.

play the average *C-Tolerance* of selected tasks in SGR. Our findings indicate a consistent trend with Figures 3a and 3b. Both PROCURL-TARGET and CURROT reduce the average *C-Tolerance* below that of IID. However, PROCURL-TARGET does not necessarily converge to the target. Conversely, CURROT persists in reducing the context values to attain convergence with the target. In Figure 3e, we depict the two-dimensional context space for the BW setting. Despite the uniformity of the target distribution of contexts, we observe that in the later stages of training, PROCURL-TARGET disregards trivial tasks characterized by *C-StumpHeight* values smaller than 0. Instead, it focuses on tasks from the remaining task space.

5 Concluding Discussions

We proposed a novel curriculum strategy that strikes a balance between selecting tasks that are neither too hard nor too easy for the agent while also progressing the agent's learning toward the target distribution by utilizing task correlation. We mathematically derived our curriculum strategy through an analysis of a specific learning scenario and demonstrated its effectiveness in various environments through empirical evaluations. Here, we discuss a few limitations of our work and outline a plan on how to address them in future work. First, it would be interesting to extend our curriculum strategy to high-dimensional context spaces in sparse reward environments. However, sampling new tasks in such environments poses a significant challenge due to the estimation of the value of all tasks in the discrete sets $\widehat{\mathcal{C}}_{\text{unif}}$ and $\widehat{\mathcal{C}}_{\text{targ}}$. Second, while our curriculum strategy uses a simple distance measure to capture task correlation, it would be worthwhile to investigate the effects of employing different distance metrics over the context space on curriculum design.

Acknowledgments

Parameswaran Kamalaruban acknowledges support from The Alan Turing Institute. Funded/Co-funded by the European Union (ERC, TOPS, 101039090). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

[Andrychowicz and others, 2017] Marcin Andrychowicz et al. Hindsight Experience Replay. In *NeurIPS*, 2017.

[Baranes and Oudeyer, 2013] Adrien Baranes and Pierre-Yves Oudeyer. Active Learning of Inverse Models with Intrinsically Motivated Goal Exploration in Robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *ICML*, 2009.

[Chaiklin, 2003] Seth Chaiklin. The Zone of Proximal Development in Vygotsky's Analysis of Learning and Instruction. *Vygotsky's Educational Theory in Cultural Context*, page 39, 2003.

[Chevalier-Boisvert *et al.*, 2023] Maxime Chevalier-Boisvert *et al.* Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

[Dendorfer *et al.*, 2020] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixé. Goal-GAN: Multimodal Trajectory Prediction based on Goal Position Estimation. In *ACCV*, 2020.

[Dennis *et al.*, 2020] Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design. In *NeurIPS*, 2020.

[Eimer *et al.*, 2021] Theresa Eimer, André Biedenkapp, Frank Hutter, and Marius Lindauer. Self-Paced Context Evaluation for Contextual Reinforcement Learning. In *ICML*, 2021.

[Elman, 1993] Jeffrey L Elman. Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition*, 48(1):71–99, 1993.

[Florensa *et al.*, 2017] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse Curriculum Generation for Reinforcement Learning. In *CORL*, 2017.

[Florensa *et al.*, 2018] Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic Goal Generation for Reinforcement Learning Agents. In *ICML*, 2018.

[Graves *et al.*, 2017] Alex Graves, Marc G Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated Curriculum Learning for Neural Networks. In *ICML*, 2017.

[Hallak *et al.*, 2015] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov Decision Processes. *CoRR*, abs/1502.02259, 2015.

[Huang *et al.*, 2022] Peide Huang, Mengdi Xu, Jiacheng Zhu, Laixi Shi, Fei Fang, and Ding Zhao. Curriculum Reinforcement Learning using Optimal Transport via Gradual Domain Adaptation. In *NeurIPS*, 2022.

[Jiang *et al.*, 2015] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-Paced Curriculum Learning. In *AAAI*, 2015.

[Jiang *et al.*, 2021a] Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-Guided Adversarial Environment Design. In *NeurIPS*, 2021.

[Jiang *et al.*, 2021b] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized Level Replay. In *ICML*, 2021.

[Kamalaruban *et al.*, 2019] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive Teaching Algorithms for Inverse Reinforcement Learning. In *IJCAI*, 2019.

[Kirk *et al.*, 2021] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A Survey of Generalisation in Deep Reinforcement Learning. *CoRR*, abs/2111.09794, 2021.

[Klink *et al.*, 2020] Pascal Klink, Carlo D'Eramo, Jan R Peters, and Joni Pajarinen. Self-Paced Deep Reinforcement Learning. In *NeurIPS*, 2020.

[Klink *et al.*, 2021] Pascal Klink, Hany Abdulsamad, Boris Belousov, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. A Probabilistic Interpretation of Self-Paced Learning with Applications to Reinforcement Learning. *JMLR*, 22:182–1, 2021.

[Klink *et al.*, 2022] Pascal Klink, Haoyi Yang, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Curriculum Reinforcement Learning via Constrained Optimal Transport. In *ICML*, 2022.

[Kumar *et al.*, 2010] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-Paced Learning for Latent Variable Models. In *NeurIPS*, 2010.

[Levine *et al.*, 2016] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end Training of Deep Visuomotor Policies. *JMLR*, 17(1):1334–1373, 2016.

[Lillicrap *et al.*, 2015] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous Control with Deep Reinforcement Learning. *CoRR*, abs/1509.02971, 2015.

[Liu *et al.*, 2017] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative Machine Teaching. In *ICML*, 2017.

[Matiisen *et al.*, 2019] 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.

[Mnih *et al.*, 2015] Volodymyr Mnih *et al.* Human-Level Control Through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, 2015.

[Modi *et al.*, 2018] Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov Decision Processes with Continuous Side Information. In *ALT*, 2018.

[Narvekar and Stone, 2019] Sanmit Narvekar and Peter Stone. Learning Curriculum Policies for Reinforcement Learning. In *AAMAS*, 2019.

[Narvekar *et al.*, 2020] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. *JMLR*, 21:1–50, 2020.

[Oudeyer *et al.*, 2007] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

[Parker-Holder *et al.*, 2022] Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving Curricula with Regret-Based Environment Design. *CoRR*, abs/2203.01302, 2022.

[Portelas *et al.*, 2019] Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of Deep RL in continuously parameterized environments. In *CoRL*, 2019.

[Portelas *et al.*, 2021] Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic Curriculum Learning for Deep RL: A Short Survey. In *IJCAI*, 2021.

[Racanière *et al.*, 2020] Sébastien Racanière, Andrew K Lampinen, Adam Santoro, David P Reichert, Vlad Firoiu, and Timothy P Lillicrap. Automated Curricula Through Setter-Solver Interactions. In *ICLR*, 2020.

[Raffin *et al.*, 2021] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *JMLR*, 22(268):1–8, 2021.

[Riedmiller *et al.*, 2018] Martin A Riedmiller *et al.* Learning by Playing Solving Sparse Reward Tasks from Scratch. In *ICML*, 2018.

[Romac *et al.*, 2021] Clément Romac, Rémy Portelas, Katja Hofmann, and Pierre-Yves Oudeyer. Teachmyagent: a Benchmark for Automatic Curriculum Learning in Deep RL. In *ICML*, 2021.

[Salimans and Chen, 2018] Tim Salimans and Richard Chen. Learning Montezuma's Revenge from a Single Demonstration. *CoRR*, abs/1812.03381, 2018.

[Schmidhuber, 2013] Jürgen Schmidhuber. Powerplay: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem. *Frontiers in Psychology*, 4:313, 2013.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347, 2017.

[Silver *et al.*, 2017] David Silver *et al.* Mastering the Game of Go Without Human Knowledge. *Nature*, 550(7676):354–359, 2017.

[Sukhbaatar *et al.*, 2018] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Robert Fergus. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play. In *ICLR*, 2018.

[Sutton *et al.*, 1999] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *NeurIPS*, 1999.

[Turchetta *et al.*, 2020] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. Safe Reinforcement Learning via Curriculum Induction. In *NeurIPS*, 2020.

[Tzannetos *et al.*, 2023] Georgios Tzannetos, Barbara Gomes Ribeiro, Parameswaran Kamalaruban, and Adish Singla. Proximal Curriculum for Reinforcement Learning Agents. *TMLR*, 2023.

[Vygotsky and Cole, 1978] Lev Semenovich Vygotsky and Michael Cole. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[Weinshall *et al.*, 2018] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks. In *ICML*, 2018.

[Wöhlke *et al.*, 2020] Jan Wöhlke, Felix Schmitt, and Herke van Hoof. A Performance-Based Start State Curriculum Framework for Reinforcement Learning. In *AAMAS*, 2020.

[Wu and Tian, 2016] Yuxin Wu and Yuandong Tian. Training Agent for First-Person Shooter Game with Actor-Critic Curriculum Learning. In *ICLR*, 2016.

[Yengera *et al.*, 2021] Gaurav Raju Yengera, Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Curriculum Design for Teaching via Demonstrations: Theory and Applications. In *NeurIPS*, 2021.

[Zhang *et al.*, 2020] Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic Curriculum Learning Through Value Disagreement. In *NeurIPS*, 2020.

[Zhou and Bilmes, 2018] Tianyi Zhou and Jeff Bilmes. Minimax Curriculum Learning: Machine Teaching with Desirable Difficulties and Scheduled Diversity. In *ICLR*, 2018.

[Zhu *et al.*, 2018] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An Overview of Machine Teaching. *CoRR*, abs/1801.05927, 2018.

[Zou *et al.*, 2019] Xiaotian Zou, Wei Ma, Zhenjun Ma, and Ryan S Baker. Towards Helping Teachers Select Optimal Content for Students. In *AIED*, 2019.