A ALGORITHM

Algorithm 2 Regioned Episodic Reinforcement Learning (RERL)

1: Initialize π, g^* as terminal state, $\bar{g}^*$ as inital state
2: Initialize region-based memories $\{\mathcal{M}_k\}_{k=1}^N$ by random sample
3: **for** episode $= 1, 2, \ldots, E$ **do**
4: Select region k according to Eq. (5)
5: Collect Z trajectories $\{\tau_z\}_{z=1}^Z \in \mathcal{M}_k$ that maximize $\sum_{z=1}^Z w(x_z, \tau_z)$ according to Eq. (10)
6: Construct intermediate goal g according to Eq. (11), $\bar{g}$ from $s \in \mathcal{M}_k$ with average value
7: **for** $t = 1, 2, \ldots, T$ **do**
8: $a_t \leftarrow \pi(a|s, g, \bar{g})$
9: $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$
10: $r_t \leftarrow r(s, g, \bar{g})$ according to Eq. (2)
11: $\tau \leftarrow \{s_0, a_0, r_0, s_1, \ldots\}$
12: $\mathcal{M}_k \leftarrow \mathcal{M}_k \cup \{\tau\}$
13: Update $\mathcal{M}_k$ according to Eq. (12)
14: Rank $\mathcal{M}_k$ and sample a minibatch b from $\mathcal{M}_k$
15: Update policy π on minibatch b using DDPG or PPO
16: **end for**
17: **end for**

The overall description of our algorithm is shown in Algorithm 2. In the initialization procedure, we set the terminal state as the initial goal and initial state as the initial anti-goal, and sample trajectories into each memory. At each episode e, the agent select one region that is most promising to lead to terminal state in line 4. We construct goal based on the historical trajectories in line 5. We take previous goals in other memories into consideration in the goal generation in line 6. From line 8 to line 13, the agent interacts with environment and update the memory. Our work focuses on how to build efficient exploration and exploitation mechanism that is naturally complementary with policy networks such as deep deterministic policy gradient (DDPG (Duan et al., 2016)) and proximal policy optimization (PPO (Schulman et al., 2017)) in line 15.

B DICUSSIONS

B.1 EXAMPLE FOR GOAL GENERATION

Previous works (Florensa et al., 2018; Vezhnevets et al., 2017) adopt a goal generator to construct immediate intrinsic rewards according to the previous states. However, they often suffer a lot from balancing the efficiency of exploration and exploitation and stability in training. In the first episode, the agent explores two trajectories in the different directions, with the closest one τ_{1a} to the target state labeled as the goal g_1 and the farthest one τ_{1b} as the anti-goal $\bar{g}_1$. In the second episode, the agent evaluates the highest value of states in the regions and selects one according to Eq. (5). The agent does exploration under the guided by g_1 (illustrated as sun icon in blue region) and $\bar{g}_1$ (illustrated as moon icon in blue region). If the agent selects the region 2, following the similar procedures, the agent will explore the region guided by g_2 (illustrated as sun icon in green region) and $\bar{g}_2$ (illustrated as moon icon in green region). Note that goal g will direct the exploration. Hence, in the goal generation, we take the historical goals in the other regions in the consideration by the diversity constraint. However, for the anti-goal generation, there is no need to consider other region data as described in Section 4.

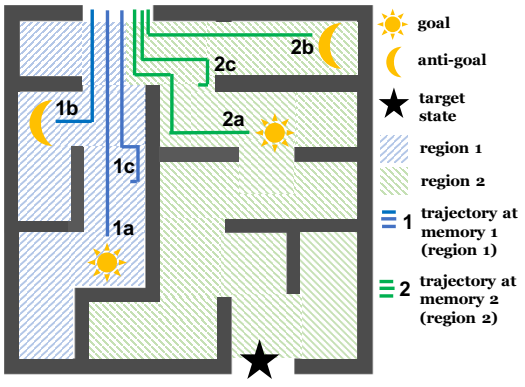


Figure 6: An illustration of exploration strategy.

B.2 RELATIONSHIP TO CURRICULUM LEARNING

In order to better understand why our method can work in complex environments and can excel other traditional methods more intuitively, we further investigate the relationship between our algorithm from Eq. (3) and empirical utility maximization formulation proposed in (Hacohen & Weinshall, 2019). We provide theoretical analysis that under some assumptions, optimizing our objection function can be similar to optimizing a curriculum algorithm under additional constraints.

Following Section 3, we formulated reinforcement learning problem as a Markov Decision Process (MDP) by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability distribution, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1)$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor for future rewards. Utility function is defined as the expected sum of the immediate and long-time utility $U_\pi(s)$ under the policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1)$, and discount factor $\gamma \in [0, 1)$, which can be formulated as:

$$U_\pi(s) := \mathbb{E}_{s_0=s, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right], \tag{1}$$

where T is the number of steps in the lifetime of the agent. Therefore, we can utilize $U_\pi(s)$ to represent the long-time reward. To formulate the short-time reward, similarly, we define $U_\pi(s_t)$ by $U_\pi(s_t) := \gamma^t r(s_t, a_t)$. In the similar manner with Empirical Risk Minimization (ERM) framework, we choose to maximize the average utility, which is defined as follows:

$$\pi^* = \arg\max_\pi \mathcal{U}(\pi), \text{ where } \mathcal{U}(\pi) := \mathbb{E}(U_\pi) = \frac{1}{T} \sum_{t=1}^{T} U_\pi(s_t) \tag{2}$$

Hindsight Constriant. We define the scoring function, *i.e.*, pacing function in curriculum learning (Bengio et al., 2009) with $\phi : \mathcal{S} \to \mathcal{G} \times \mathcal{G}$ which is a known and tractable mapping. ϕ effectively provides a Bayesian prior $g \in \mathcal{G}$ for data sampling namely exploration, where g denotes goal and $\mathcal{G}$ denotes goal space. We also adopt ϕ^+ and ϕ^- to represent the postive goal (*i.e.*, goal) generation and negative goal (*i.e.*, anti-goal) generation respectively. Based on the analysis above, we can formulate Eq. (1) as

$$\mathcal{U}_g(\pi) = \mathbb{E}_g[U_\pi] = \frac{1}{T} \sum_{t=1}^{T} U_\pi(s_t) \cdot \phi(\cdot|s_t), \tag{3}$$

where $\phi(\cdot|s_t)$ denotes the induced prior probability conditioned on s_t. In order to guarantee the convergence, $\phi(\cdot|s_t)$ should always be a non-increasing function of the difficulty level of s_t. In our algorithm, we define the goal space $\mathcal{G}$ as a subset of state space $\mathcal{S}$ (*i.e.*, hindsight constraint in Section 4), which guarantees each goal/anti-goal is sampled from previous states. This proves the following result:

Proposition 1. The difference between the expected utility function with and without prior g (*i.e.*, $\mathcal{U}_g(\pi)$ and $\mathcal{U}(\pi)$) is the covariance between utility function $U_\pi(s)$ and goal generation $\phi(\cdot|s)$.

The proof of Proposition 1 can be found in Appendix C.3.

Diversity Constraint. However, one should be noted that goal g here is sampled from previous states which guarantees reachability of the goal but also limits potential exploration. To address this issue, we adopt diversity measure $\mathcal{H}_{\text{region}}(\pi)$ to encourage the exploration between different region (diversity constraint in Section 4 is a simple implementation). Combining the aforementioned hindsight and diversity constraints, we define our objective as

$$\pi^* = \arg\max_\pi \mathcal{U}_g(\pi), \text{ under hindsight and diversity constraints.} \tag{4}$$

which can easily derive as the equivalence to Eq. (6).

Proposition 2. The modified optimization landscape induced by curriculum learning has the same global optimum π^* as the original problem.

The proof of Proposition 2 can be found in Appendix C.4.

According to the analysis, we can conclude that our algorithm can be regarded as an novel curriculum learning approach in goal-oriented setting, which can be proved to have the same global optimum as the original problem. In Section 5, we conduct experiment to prove that the goals are generated in the different levels as the curriculum to guide the agent in curriculum learning.

B.3 RELATIONSHIP TO MAXIMUM ENTROPY RL

In this section, we consider multi-goal RL as goal-oriented policy learning (Schaul et al., 2015; Plappert et al., 2018). We further discuss the motivation behind these two constraints, namely hindsight and diversity constraints, and the relationships between our work and inverse maximum entropy reinforcement learning.

Preliminaries. We begin with some notations and previous motivations in maximum entropy reinforcement learning (Eysenbach et al., 2020). The likelihood of a trajectory $\tau := \{s_t\}_{t=0}^{T}$ under policy π can be formulated as $\mathcal{L}(s) = \mathcal{P}(s_0) \cdot \prod_t \mathcal{P}(s_{t+1}|s_t, a_t)\pi(a_t|s_t)$. In the goal-oriented RL, we can rewrite as

$$\mathcal{L}(s, g) = \mathcal{P}(s_0) \cdot \prod_t \mathcal{P}(s_{t+1}|s_t, a_t)\pi(a_t|s_t, g), \tag{5}$$

where the initial state is sampled as $s_0 \sim \mathcal{P}(s_0)$ and subsequent states are governed by a dynamic distribution $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$. As we discuss in Appendix B.2, goal-oriented RL can be regarded as regular RL with prior knowledge g generated by mapping function ϕ based s. Hence, the target joint distribution over goals and states is

$$\mathcal{X}(s, g) = \frac{\phi(\cdot|s)}{\mathrm{Z}(g, \bar{g})} \cdot \mathcal{P}(s_0) \prod_t \mathcal{P}(s_{t+1}|s_t, a_t) e^{r(s_t, g_t, \bar{g}_t)}. \tag{6}$$

where $\mathcal{X} : \mathcal{S} \times \mathcal{G} \times \mathcal{G} \to [0, 1]$ be the joint distribution over state $s \in \mathcal{S}$, goal $g \in \mathcal{G}$ and anti-goal $\bar{g} \in \mathcal{G}$; and $\mathrm{Z}(g, \bar{g})$ is the factor of normalization.

Diversity Constraint. We can express the multi-goals RL objective as the reverse KL divergence between the joint state-goal distributions:

$$\max_{\pi} -\mathcal{H}(s, g) = \max_{\pi} -\mathcal{D}_{\mathrm{KL}}(\mathcal{L}(s, g) \| \mathcal{X}(s, g)) \tag{7}$$

where the joint distribution of likelihood $\mathcal{L}$ and prior information g of a trajectory τ is defined as $\mathcal{L}(s, g) := \mathcal{L}(s|g) \cdot \phi(\cdot|s)$. Then, we can rewrite Eq. (7) as maximizing the expected (entropy-regularized) reward of a goal-conditioned policy $\mathcal{L}(s|g)$:

$$\mathbb{E}_{g \sim \phi(\cdot|s),\ s \sim \mathcal{L}(\cdot|g)} \left[\left(\sum_{t=0}^{T} r(s_t, a_t|g, \bar{g}) - \log \pi(s_t, a_t|g, \bar{g}) \right) - \log \mathrm{Z}(g, \bar{g}) \right]. \tag{8}$$

Hindsight Constraint. Since the distribution over goals g is fixed, we can ignore the $\log \mathrm{Z}(g, \bar{g})$ term for optimization. A less common but more intriguing choice is to factor $\mathcal{L}(s, g) = \phi(\cdot|s) \cdot \mathcal{B}(s)$, where $\mathcal{B}(\tau)$ is represented non-parametrically as a distribution over previously-observed states. Therefore, $\phi(\cdot|s)$ is formulated as a hindsight relabeling distribution. In this implementation, we sample goals from previous states in the region-based memory to present $\mathcal{B}(s)$.

C PROOFS

C.1 PROOF OF PROPOSITION 1

Proposition 3. *Given the global distribution $\mathcal{X}$ and several region-based distributions $\mathcal{X}_k$, where $k = 1, 2, \ldots, N$ and N is the number of regions, we have*

$$\forall \pi,\ \max_{x \sim \mathcal{X}} V^{\pi}(x) \geq \max_{x \in \{x_1, x_2 \ldots, x_N\}} V^{\pi}(x),\ \textit{where}\ x_k = \arg\max_{x_k \sim \mathcal{X}_k} V^{\pi}(x_k). \tag{9}$$

In this section, we provide the proof of Proposition 1. The motivation of Proposition 1 is to find a relaxed lower bound of $V^{\pi}(x),\ x \sim \mathcal{X}$ based on the definition of the region.

Proof. By Eq. (3), $\forall \pi$ we have

$$\begin{aligned}
\max_{x\sim\mathcal{X}} V^{\pi}(x) &= \max_{x\sim\mathcal{X},\ \mathcal{X}:\mathcal{S}\times\mathcal{G}\times\mathcal{G}} \mathbb{E}_{s\in\mathcal{S};g,\bar{g}\in\mathcal{G},\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t|g,\bar{g})\right] \\
&\geq \max_{x\sim\{x_1,x_2,\ldots,x_N\}}\Big\{\max_{x_1\sim\mathcal{X}_1,\ \mathcal{X}_1:\mathcal{S}_1\times\mathcal{G}_1\times\mathcal{G}_1} \mathbb{E}_{s\in\mathcal{S}_1;g,\bar{g}\in\mathcal{G}_1,\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})\right],\ldots, \\
&\qquad \ldots,\max_{x_N\sim\mathcal{X}_N,\ \mathcal{X}_N:\mathcal{S}_N\times\mathcal{G}_N\times\mathcal{G}_N} \mathbb{E}_{s\in\mathcal{S}_N;g,\bar{g}\in\mathcal{G}_N,\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})\right]\Big\} \\
&\geq \max_{x\sim\{x_1,x_2,\ldots,x_N\}} V^{\pi}(x),\ \text{where } x_i=\arg\max_{x_i\sim\mathcal{X}_i} V^{\pi}(x_k),\ k=1,2,3,\ldots,N.
\end{aligned} \tag{10}$$

The intuition behind the proposition is easy to understand. Since we have partitioned the whole distribution $\mathcal{X}$ into several region-based distributions $\{\mathcal{X}_k\}_{k=1}^{N}$. We effectively avoid the agent switching among regions meanwhile removing these trajectories out of original candidate trajectory family. □

C.2 Proof of Proposition 4

Proposition 4. *Denote the Bellman backup operator in Q learning with goal as* $\mathcal{B}:\mathbb{R}^{|S|\times|A|\times|G|}\rightarrow\mathbb{R}^{|S|\times|A|\times|G|}$ *and a mapping* $Q: S\times A\times G\rightarrow\mathbb{R}^{|S|\times|A|\times|G|}$ *with* $|S|<\infty$ *and* $|A|<\infty$. *Repeated application of the operator* $\mathcal{B}$ *for our goal-oriented state-action value estimate* $\hat{Q}$ *converges to a unique optimal value* $\hat{Q}^*$.

Proof. The proof of Proposition 4 is done in two main steps. The first step is to show that our goal $g\in\mathcal{G}$ can converge to the terminal state. In the second step, we prove that given goal g, our goal-oriented approach can converge to unique optimal value Q^*. In other words, we need to prove that $g\rightarrow g^*$ in the first step and $Q\rightarrow Q^*$ in the second step.

Step I. Since our algorithm aims to find the high-value previous states for goal generation. At the beginning of the task, the terminal state will be regarded as the final goal, since it has the highest value. Hence, the terminal state, if it has been visited once, will be assigned as the goal. Assume that the agent can conduct plenty of exploration. Then, we can say that the generated goal g will keep approaching to the terminal state g^*.

Step II. Note that the proof of convergence for our goal-oriented RL is quite similar to Q-learning (Bellman, 1966; Bertsekas et al., 1995; Sutton & Barto, 2018). The differences between our approach and Q-learning are that Q-value $Q(s,a,g,\bar{g})$ is also conditioned on goal g and anti-goal $\bar{g}$. As introduced in Section 4, anti-goal $\bar{g}$ works like a reward shaping technique, which is proposed to avoid local optima (Trott et al., 2019). Hence, we omit $\bar{g}$ in the following proof. We provide detailed proof as follows:

We can obtain goal $g\in G$ approaching the terminal state from Step I. Based on that, our estimated goal-conditioned action-value function $\hat{Q}$ can be defined as

$$\mathcal{B}\hat{Q}(s,a,g)=R(s,a,g)+\gamma\cdot\max_{a'\in A}\sum_{s'\in S}P(s'|s,a)\cdot\hat{Q}(s',a',g). \tag{11}$$

For any action-value function estimates $\hat{Q}^1,\hat{Q}^2$, we study that

$$\begin{aligned}
&|\mathcal{B}\hat{Q}^1(s,a,g)-\mathcal{B}\hat{Q}^2(s,a,g)| \\
&=\gamma\cdot|\max_{a'\in A}\sum_{s'\in S}P(s'|s,a)\cdot\hat{Q}^1(s',a',g)-\max_{a'\in A}\sum_{s'\in S}P(s'|s,a)\cdot\hat{Q}^2(s',a',g)| \\
&\leq\gamma\cdot\max_{a'\in A}|\sum_{s'\in S}P(s'|s,a)\cdot\hat{Q}^1(s',a',g)-\sum_{s'\in S}P(s'|s,a)\cdot\hat{Q}^2(s',a',g)| \\
&=\gamma\cdot\max_{a'\in A}\sum_{s'\in S}P(s'|s,a)\cdot|\hat{Q}^1(s',a',g)-\hat{Q}^2(s',a',g)| \\
&\leq\gamma\cdot\max_{s\in S,a\in A}|\hat{Q}^1(s,a,g)-\hat{Q}^2(s,a,g)|
\end{aligned} \tag{12}$$

Combining Step I and II, we can conclude that our goal-conditioned estimated state-action value $\hat{Q}$ can converge to unique optimal value Q^* leading to the terminal state g^*. □

C.3 Proof of Proposition 1

In this section, we provide the proof of Proposition 1. From Eq. (3), $\mathcal{U}_g(\pi)$ is a function of π which is determined by the correlation between $U_\pi(s)$ and $\phi(g)$ (*i.e.*, $\phi(\cdot|s)$). We can rewrite Eq. (3) as

$$\begin{aligned}\mathcal{U}_g(\pi) &= \frac{1}{T}\{\sum_{t=1}^{T}(U_\pi(s_t) - \mathbb{E}[U_\pi])(\phi(g_t) - \mathbb{E}[\phi]) + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\ &= \frac{1}{T}\{\text{Cov}[U_\pi, \phi] + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\ &= \frac{1}{T}\{\mathcal{U}(\pi) + \text{Cov}[U_\pi, \phi]\}\end{aligned} \tag{13}$$

This derivation can be found in Appendix C.6. We can find that curriculum learning changes the landscape of the optimization function over the policy π from $\mathcal{U}(\pi)$ to $\mathcal{U}_g(\pi)$. Intuitively, the above equation also suggests that if the induced goal g, which defines a latent variable over the goal space $\mathcal{G}$, is positively correlated with the optimal utility $U_{\pi^*}(s)$, and more so than with any other $U_\pi(s)$, then the gradients in the direction of the optimal policy π in the new optimization landscape may be overall steeper.

Hence, this is necessary to design task-related goals. However, it is infeasible to obtain appropriate goals through handcrafted design and manual generation. In this paper, we introduce hindsight and diversity constraints to help the agent learn from achieved task-related information (previous states) and unknown task-related information (unexplored states) respectively.

C.4 Proof of Proposition 2

In this section, we provide the proof of Proposition 2. In order to prove that the modified optimization function in the state space related parameter space π has the property that the global maximum at π^* is more pronounced, we derive the objective function based on Proposition 1. We can assume that optimal policy π^* maximizes the covariance between $\phi(g)$ (*i.e.*, $\phi(\cdot|s)$) and utility $U_\pi(s)$, namely

$$\arg\max_\pi \mathcal{U}(\pi) = \arg\max_\pi \text{Cov}[U_\pi, \phi] = \pi^* \tag{14}$$

The proof of the assumption can be found in Appendix C.3. We introduce Lemma 1 here, the proof of which can be found in Appendix C.5.

Lemma 1. (Florensa et al. (2017)) For any curriculum satisfying Eq. (14):

1. $\pi^* = \arg\max_\pi \mathcal{U}(\pi) = \arg\max_\pi \mathcal{U}(\pi^*)$
2. $\mathcal{U}_g(\pi^*) - \mathcal{U}_g(\pi) \geq \mathcal{U}(\pi^*) - \mathcal{U}(\pi), \quad \forall\pi$

Lemma 1 has proposed two claims. The first one presents that the problem of maximizing the covariance between $\phi(g)$ and utility $U_\pi(s)$ shares the same optimal solution with the original problem. In addition, the modified optimization function in the original parameter space without goal g has the property that the global maximum with goal g is more pronounced.

C.5 Proof of Lemma 1

In this section, we provide the proof of Lemma 1. Claim 1 in Lemma 1 can be derived directly from Eq. (14), while for the claim 2, we have

Proof.

$$\begin{aligned}\mathcal{U}_g(\pi^*) - \mathcal{U}_g(\pi) &= \mathcal{U}_g(\pi^*) - \mathcal{U}(\pi) - \text{Cov}[U_\pi, g] \\ &\geq \mathcal{U}_g(\pi^*) - \mathcal{U}(\pi) - \text{Cov}[\mathcal{U}_{\pi^*}, g] \\ &= \mathcal{U}(\pi^*) - \mathcal{U}(\pi)\end{aligned} \tag{15}$$

□

C.6 Detailed Derivation of Eq. (13)

In this section, we provide the detailed derivation of Eq. (13). We begin from the formulation of $\mathcal{U}_g(\pi)$ in Eq. (13) and try to obtain that in Eq. (3).

Proof.

$$\begin{aligned}
\mathcal{U}_g(\pi) &= \frac{1}{T}\{\sum_{t=1}^{T}(U_\pi(s_t) - \mathbb{E}[U_\pi])(\phi(g_i) - \mathbb{E}[\phi]) + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\
&= \frac{1}{T}\{\sum_{t=1}^{T}(U_\pi(s_t)\phi(g_t)) - \sum_{t=1}^{T}(U_\pi(s_t)\mathbb{E}[\phi]) - \sum_{t=1}^{T}(\phi(g_t)\mathbb{E}[U_\pi]) + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi] + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\
&= \frac{1}{T}\{\sum_{t=1}^{T}(U_\pi(s_t)\phi(g_t)) - T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi] - \sum_{t=1}^{T}(\phi(g_t)\mathbb{E}[U_\pi]) + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi] + T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\
&= \frac{1}{T}\sum_{t=1}^{T}(U_\pi(s_t)\phi(g_t)) + \frac{1}{T}\{T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi] - \sum_{t=1}^{T}(\phi(g_t)) \cdot \mathbb{E}[U_\pi]\}
\end{aligned} \tag{16}$$

Since $\mathbb{E}[\phi] := \frac{1}{T}\sum_{t=1}^{T}(\phi(g_t))$, we have

$$\begin{aligned}
\mathcal{U}_g(\pi) &= \frac{1}{T}\sum_{t=1}^{T}(U_\pi(s_t)\phi(g_t)) + \frac{1}{T}\{T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi] - T \cdot \mathbb{E}[U_\pi]\mathbb{E}[\phi]\} \\
&= \frac{1}{T}\sum_{t=1}^{T}U_\pi(s_t)\phi(g_t)
\end{aligned} \tag{17}$$

□

D Experiment

D.1 Modified Environments

Ant Locomotion. In this part, we introduce two environments based on Ant Locomotion, namely Free Ant and Ant Maze. The ant is a quadruped with 8 actuated joint, 2 for each leg. The environment is implemented in Mujoco. Besides the coordinates of the center of mass, the joint angles and joint velocities are also contained in the observation of the agent. Considering the high degrees of freedom, navigation in this quite complex task requires motor coordination. More details can be found in Duan et al. (2016), and the only difference is that in our goal-oriented version of Ant, we extend the observation with the goals. The reward is still a sparse indicator function being 1 only when the center of mass (x, y) of the Ant is within $\epsilon = 0.5$ positions corresponding to ϵ-balls in state space. For the Free Ant experiments, the objective is to reach any position in the square $[-5, 5]^2$. Therefore the goal space is 2 dimensional, the state-space is 41 dimensional, and the action space is 8 dimensional. As for the Ant Maze environment, the agent is constrained to move within the maze environment, U-maze in this case and the size of all the blocks in the maze is 8×8. The maze consists of totally 18 blocks.

Multi-Path Point Maze. All the experiment setting is similar with the Ant Maze environment. We replace the Ant agent by a Point-Mass and change the maze into a multi-path one. The action of the Point-Mass is a velocity vector, namely in 2 dimension.

N-dimensional Point-Mass Maze. In the N-dimensional Point-Mass maze experiment, the agent can only move within a small subset of this state space. In the two-dimensional case, the set of feasible states corresponds to the $[-5, 5] \times [-1, 1]$ rectangle, making up 20% of the full space. For $N > 2$, the feasible space is the Cartesian product of this 2D strip with $[-\epsilon, \epsilon]^{N-2}$, where $\epsilon = 0.3$. In this higher-dimensional environment, our agent receives a reward of 1 when it moves within $\epsilon_N = 0.3\frac{\sqrt{N}}{\sqrt{2}}$ of the goal state, to account for the increase in average $L2$ distance between points in higher dimensions. In this experiments, the full state-space of the N-dimensional Point Mass is the hypercube $[-5, 5]^N$.

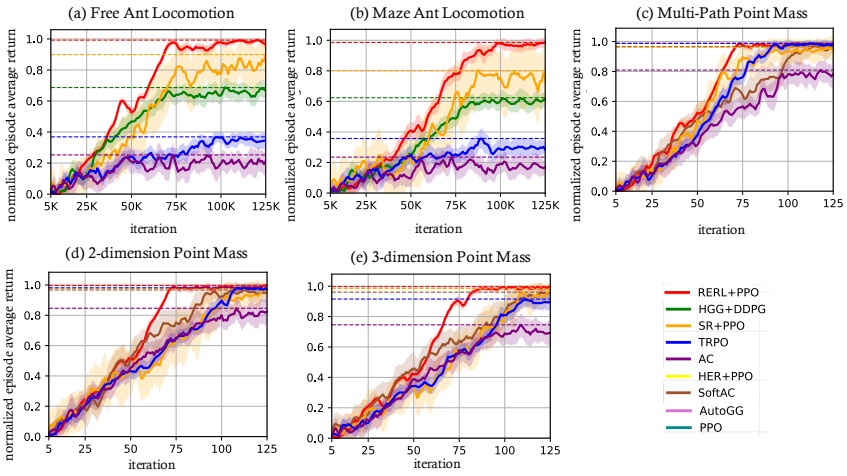


Figure 7: Learning curves of RERL, HGG, HER, SR and AutoGG on various environments, where the solid curves depict the mean, the shaded areas indicate the standard deviation, and dashed horizontal lines show the asymptotic performance.

D.2 Evaluation Details

We adopt HGG (Ren et al., 2019) incorporating with DDPG (Lillicrap et al., 2015), SR (Trott et al., 2019) accompanying with PPO (Schulman et al., 2017) as these models are originally proposed. All curves presented in this paper are plotted from 12 runs with random task initializations and seeds. Following the regular procedure in goal-oriented RL, an episode is considered successful if and only if the agent obtain 1 as the reward according to Eq. (2) where δ stays the same for all the approaches. However, in the practice, we conduct reward as the $r(s_t, a_t|g, \bar{g}) = \min[0, -d(\phi(g_{t+1}|s_{t+1}), g) + d(\phi(\bar{g}_{t+1}|s_{t+1}), \bar{g})]$ to accelerate the training process.

D.3 Implementation Details

Almost all hyper-parameters using DDPG (Lillicrap et al., 2015), TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), Soft-AC (Haarnoja et al., 2018) are kept the same as benchmark results. Specifically, we list our hyper-parameters as here. number of MPI workers: 1; buffer size: 10^4 trajectories; number of regions N: 5 in agent level; batch size: 256, number of trajectories Z: 50, Lipschitz constant L: 5; learning rate: 10^{-5} in the network level; discount factor: 0.99; interpolation factor in polyak averaging (if there is): 0.995; scale of additive Gaussian noise: 0.2; probability of HER (Andrychowicz et al., 2017) experience replay: 0.8.

E Results

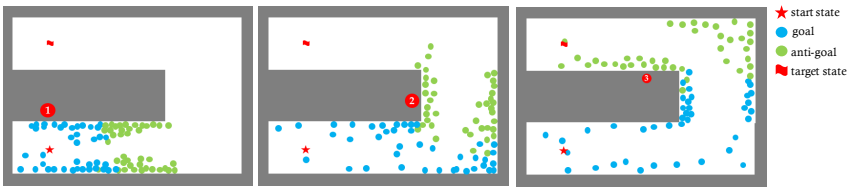
E.1 Additional Evaluation on Standard Tasks

In this section, we provide additional results on comparison between RERL and various baselines.

In order to answer the first two questions, we demonstrate our method in two challenging robotic locomotion tasks, where the goals are the (x, y) position of the center of mass of dynamically complex quadruped agent. In the first example, the agent has no constraints, and in the second one, the agent is inside a U-maze (see Section 5 for details). Results in Figure 7(a)(b) demonstrate that the performance of our approach exceeds that of the strong baselines mentioned in Section 5. To answer the third question, we train an ant agent to reach any position within a multi-path maze. As shown in Figure 7(c), our approach obtains better performance even at multi-path environment where goal distribution are naturally more complex than previous environments. To answer the fourth question, we investigate how our method performs with the dimension of goal-space in an environment where the goal space grows in dimension within the feasible region, *e.g.*, 2D and 3D. As shown in Figure 7(d), our approach outperforms strong baselines in both low- and high-dimensional environments.

E.2 Additional Results on Visualization of Generated Goals

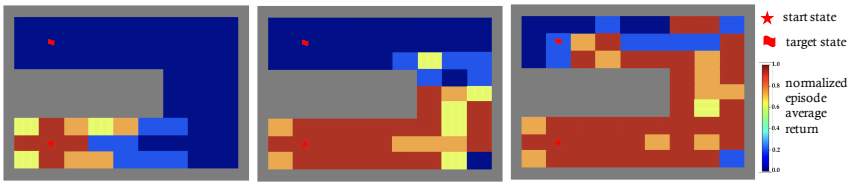
To answer the final question, we conduct a visualization study on generated goals to investigate whether goals can encourage the agent to the target state, and anti-goals can prevent the agent from the local optima. The visualization of goals can also represent the effect of diversity and hindsight constraints through exploration and reachability of generated goals.



(a) iteration=50 (b) iteration=150 (c) iteration=400

Figure 8: Generated goals and anti-goals visualized as the blue and green points respectively.

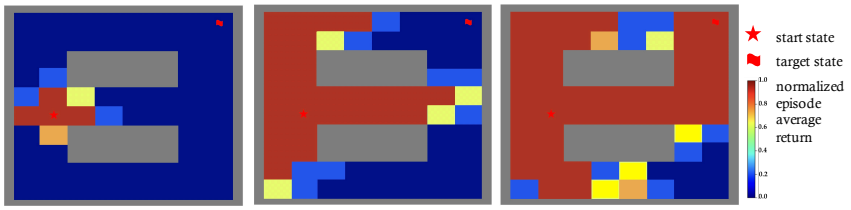
Results in Figure 8 show that the hindsight constraint helps the agent aim at feasible positions while our diversity constraint encourages the agent to approach the target state. Specifically, from (1) and (2), one can note that the agent is pulled by its goal, and pushed by its anti-goal and goals from the other regions. Hence, once a region leading to a wrong direction, it also can encourage exploration via diversity constraint.



(a) iteration=50 (b) iteration=150 (c) iteration=400

Figure 9: Each grid cell in U-maze is colored according to the expected return success rate when fixing its center as the target state.

As illustrated in Figure 9, the generated goals are approaching as the training proceeds, and at an appropriate success rate level, which is accorded with the curriculum in the curriculum learning (see Appendix B.2 for details).



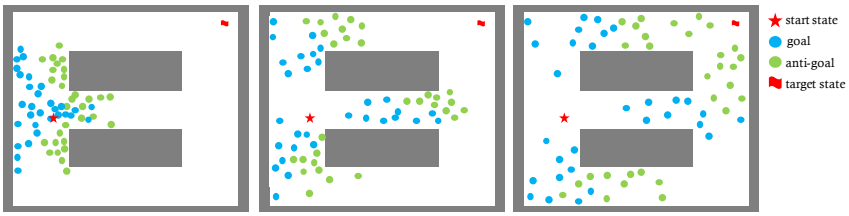
(a) iteration=5 (b) iteration=15 (c) iteration=40

Figure 10: Each grid cell in Multi-path maze is colored according to the expected return success rate when fixing its center as the target state.

Results showed in Figure 10 and 11 are similar with that in Figure 9 and 8 respectively, which actually can confirm the analysis above.

E.3 Experiment on Comparison with Explicit Curriculum Learning

In Florensa et al. (2017), GOID is defined as a goal set as $\text{GOID}(\pi) = \{g : \alpha \leq f(\pi, g) \leq 1 - \alpha\}$ where $f(\pi, g)$ represents the average success rate in a small region closed by goal g. In order to construct GOID set, we follow its definition and sample generated goals from $\text{GOID}(\pi)$ via rejection sampling.



(a) iteration=5 (b) iteration=15 (c) iteration=40

Figure 11: Generated goals and anti-goals visualized as the blue and green points respectively.

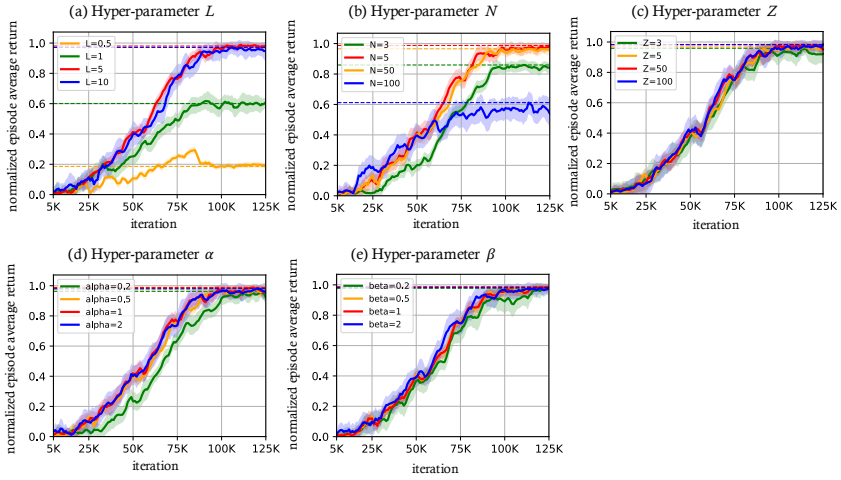


Figure 12: Learning curves of ablation study on parameters: L, N, Z, α and β, where the solid curves depict the mean, the shaded areas indicate the standard deviation, and dashed horizontal lines show the asymptotic performance.

E.4 Additional Results on Ablation Study

In this section, we set up a set of ablation tests on several hyper-parameters used in the RERL. The selection of Lipschitz constant L is task dependent, since it is highly related with scale of value function and goal distance. For the robotics tasks tested in this paper (*i.e.*, Ant Maze Locomotion), as showed in Figure 12(a), we find that the performance of RERL is reasonable as long as L is not too small. Similar as L, the selection of the number of regions N is also theoretically task-specific. We test a few choices on Ant Maze Locomotion and find a range of N that works well. As Figure 12(b) illustrates, it appears that the RERL is reasonable as long as N is not too large. As for the number of trajectories Z, we plot the curve on different Z in Figure 12(c) and find that for the simple tasks, the choice of Z is not critical. Parameters α and β together define the trade-off between value function, diversity and hindsight constraints. Results in Figure 12(d)(e) show that the choice of α and β is indeed robust.