

# Automatic Goal Generation for Reinforcement Learning Agents

Carlos Florensa <sup>\*1</sup> David Held <sup>\*2</sup> Xinyang Geng <sup>\*1</sup> Pieter Abbeel <sup>1,3</sup>

## Abstract

Reinforcement learning (RL) is a powerful technique to train an agent to perform a task; however, an agent that is trained using RL is only capable of achieving the single task that is specified via its reward function. Such an approach does not scale well to settings in which an agent needs to perform a diverse set of tasks, such as navigating to varying positions in a room or moving objects to varying locations. Instead, we propose a method that allows an agent to automatically discover the range of tasks that it is capable of performing in its environment. We use a generator network to propose tasks for the agent to try to accomplish, each task being specified as reaching a certain parametrized subset of the state-space. The generator network is optimized using adversarial training to produce tasks that are always at the appropriate level of difficulty for the agent, thus automatically producing a curriculum. We show that, by using this framework, an agent can efficiently and automatically learn to perform a wide set of tasks without requiring any prior knowledge of its environment, even when only sparse rewards are available. Videos and code available at: <https://sites.google.com/view/goalgeneration4rl>.

## 1. Introduction

Reinforcement learning (RL) can be used to train an agent to perform a task by optimizing a reward function. Recently, a number of impressive results have been demonstrated by training agents using RL: such agents have been trained

<sup>\*</sup>Equal contribution <sup>1</sup>Department of Computer Science, UC Berkeley <sup>2</sup>Department of Computer Science, CMU <sup>3</sup>International Computer Science Institute (ICSI). Correspondence to: Carlos Florensa <[florensa@berkeley.edu](mailto:florensa@berkeley.edu)>, David Held <[dheld@andrew.cmu.edu](mailto:dheld@andrew.cmu.edu)>.

*Proceedings of the 35<sup>th</sup> International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

to defeat a champion Go player (Silver et al., 2016), to outperform humans in 49 Atari games (Guo et al., 2016; Mnih et al., 2015), and to perform a variety of difficult robotics tasks (Lillicrap et al., 2015; Duan et al., 2016; Levine et al., 2016). In each of the above cases, the agent is trained to optimize a single reward function in order to learn to perform a single task. However, there are many real-world environments in which a robot will need to be able to perform not a single task but a diverse set of tasks, such as navigating to varying positions in a room or moving objects to varying locations. We consider the problem of maximizing the average success rate of our agent over all possible goals, where success is defined as the probability of successfully reaching each goal by the current policy.

In order to efficiently maximize this objective, the algorithm must intelligently choose which goals to focus on at every training stage: goals should be at the appropriate level of difficulty for the current policy. To do so, our algorithm allows an agent to generate its own reward functions, defined with respect to target subsets of the state space, called goals. We generate such goals using a Goal Generative Adversarial Network (Goal GAN), a variation of to the GANs introduced by Goodfellow et al. (2014). A goal discriminator is trained to evaluate whether a goal is at the appropriate level of difficulty for the current policy, and a goal generator is trained to generate goals that meet this criteria. We show that such a framework allows an agent to quickly learn a policy that reaches all feasible goals in its environment, with no prior knowledge about the environment or the tasks being performed. Our method automatically creates a curriculum, in which, at each step, the generator generates goals that are only slightly more difficult than the goals that the agent already knows how to achieve.

In summary, our main contribution is a method for automatic curriculum generation that considerably improves the sample efficiency of learning to reach all feasible goals in the environment. Learning to reach multiple goals is useful for multi-task settings such as navigation or manipulation, in which we want the agent to perform a wide range of tasks. Our method also naturally handles sparse reward functions, without needing to manually modify the reward function for every task, based on prior task knowledge. Instead, our

method dynamically modifies the probability distribution from which goals are sampled to ensure that the generated goals are always at the appropriate difficulty level, until the agent learns to reach all goals within the feasible goal space.

## 2. Related Work

The problem that we are exploring has been referred to as “multi-task policy search” (Deisenroth et al., 2014) or “contextual policy search,” in which the task is viewed as the context for the policy (Deisenroth et al., 2013; Fabisch & Metzen, 2014). Unlike the work of Deisenroth et al. (2014), our work uses a curriculum to perform efficient multi-task learning, even in sparse reward settings. In contrast to Fabisch & Metzen (2014), which trains from a small number of discrete contexts / tasks, our method generates a training curriculum directly in continuous task space.

**Intrinsic Motivation:** Intrinsic motivation involves learning with an intrinsically specified objective (Schmidhuber, 1991; 2010). Intrinsic motivation has also been studied extensively in the developmental robotics community, such as SAGG-RIAC (Baranes & Oudeyer, 2010; 2013a), which has a similar objective of learning to explore a parameterized task space. However, our experiments with SAGG-RIAC demonstrate that this approach does not explore the space as efficiently as ours. A related concept is that of competence-based intrinsic motivation (Baldassarre & Mirolli, 2012), which uses a selector to select from a discrete set of experts. Recently there have been other formulations of intrinsic motivation, relating to optimizing surprise (Houthooft et al., 2016; Achiam & Sastry, 2016) or surrogates of state-visitation counts (Bellemare et al., 2016; Tang et al., 2016). All these approaches improve learning in sparse tasks where naive exploration performs poorly. However, these formulations do not have an explicit notion of which states are hard for the learner, and the intrinsic motivation is independent of the current performance of the agent. In contrast, our formulation directly motivates the agent to train on tasks that push the boundaries of its capabilities.

**Skill-learning:** We are often interested in training an agent to perform a collection of tasks rather than a single one, like reaching different positions in the agent’s state-space. Skill learning is a common approach to this problem as it allows the agent to re-use skills, improving learning compared to training for every task from scratch. Discovering useful skills is a challenging task that has mostly been studied for discrete environments (Vigorito & Barto, 2010; Mnih et al., 2016) or for continuous tasks where demonstrations are provided (Konidaris et al., 2011; Ranchod et al., 2015). Recent work overcomes some of these limitations by training low-level modulated locomotor controllers (Heess et al., 2016), or multimodal policies with an information theoretic regularizer to learn a fixed-size set of skills (Florensa et al.,

2017a). Nevertheless, in previous work, learning skills is usually a pre-training step from which useful primitives are obtained and later used to achieve other tasks. Hence, additional downstream training is required to properly compose the skills in a purposeful way. On the other hand, our approach trains policies that learn to achieve multiple goals directly.

**Curriculum Learning:** The increasing interest on training single agents to perform multiple tasks is leading to new developments on how to optimally present the tasks to the agent during learning. The idea of using a curriculum has been explored in many prior works on supervised learning (Bengio et al., 2009; Zaremba & Sutskever, 2014; Bengio et al., 2015). However, these curricula are usually hand-designed, using the expertise of the system designer. Another line of work uses learning progress to build an automatic curriculum (Graves et al., 2017), however it has mainly been applied for supervised tasks. Most curriculum learning in RL still relies on fixed pre-specified sequences of tasks (Karpathy & Van De Panne, 2012). Other recent work assumes access to a baseline performance for several tasks to gauge which tasks are the hardest and require more training (Sharma & Ravindran, 2017), but the framework can only handle a finite set of tasks and cannot handle sparse rewards. Our method trains a policy that generalizes to a set of continuously parameterized tasks, and is shown to perform well even under sparse rewards by not allocating training effort to tasks that are too hard for the current performance of the agent. Finally, an interesting self-play strategy has been proposed that is concurrent to our work (Sukhbaatar et al., 2017); however, they view their approach as simply providing an exploration bonus for a single target task; in contrast, we focus on the problem of efficiently optimizing a policy across a range of goals.

## 3. Problem Definition

### 3.1. Goal-parameterized Reward Functions

In the traditional RL framework, at each timestep  $t$ , the agent in state  $s_t \in \mathcal{S} \subseteq \mathbb{R}^n$  takes an action  $a_t \in \mathcal{A} \subseteq \mathbb{R}^m$ , according to some policy  $\pi(a_t | s_t)$  that maps from the current state  $s_t$  to a probability distribution over actions. Taking this action causes the agent to enter into a new state  $s_{t+1}$  according to a transition distribution  $p(s_{t+1}|s_t, a_t)$ , and receive a reward  $r_t = r(s_t, a_t, s_{t+1})$ . The objective of the agent is to find the policy  $\pi$  that maximizes the expected return, defined as the sum of rewards  $R = \sum_{t=0}^T r_t$ , where  $T$  is a maximal time given to perform the task. The learned policy corresponds to maximizing the expected return for a single reward function.

In our framework, instead of learning to optimize a single reward function, we consider a range of reward functions

$r^g$  indexed or parametrized by a goal  $g \in \mathcal{G}$ . Each goal  $g$  corresponds to a set of states  $S^g \subset \mathcal{S}$  such that goal  $g$  is considered to be achieved when the agent is in any state  $s_t \in S^g$ . Then the objective is to learn a policy that, given any goal  $g \in \mathcal{G}$ , acts optimally with respect to  $r^g$ . We define a very simple reward function that measures whether the agent has reached the goal:

$$r^g(s_t, a_t, s_{t+1}) = \mathbb{1}\{s_{t+1} \in S^g\}, \quad (1)$$

where  $\mathbb{1}$  is the indicator function. In our case, we use  $S^g = \{s_t : d(f(s_t), g) \leq \epsilon\}$ , where  $f(\cdot)$  is a function that projects a state into goal space  $\mathcal{G}$ ,  $d(\cdot, \cdot)$  is a distance metric in goal space, and  $\epsilon$  is the acceptable tolerance that determines when the goal is reached. However, our method can handle generic binary rewards (as in Eq. (1)) and does not require a distance metric for learning.

Furthermore, we define our MDP such that each episode terminates when  $s_t \in S^g$ . Thus, the return  $R^g = \sum_{t=0}^T r_t^g$  is a binary random variable whose value indicates whether the agent has reached the set  $S^g$  in at most  $T$  time-steps. Policies  $\pi(a_t | s_t, g)$  are also conditioned on the current goal  $g$  (as in Schaul et al. (2015)). The expected return obtained when we take actions sampled from the policy can then be expressed as the probability of success on that goal within  $T$  time-steps, as shown in Eq. (2).

$$\begin{aligned} R^g(\pi) &= \mathbb{E}_{\pi(\cdot | s_t, g)} \mathbb{1}\{\exists t \in [1 \dots T] : s_t \in S^g\} \\ &= \mathbb{P}(\exists t \in [1 \dots T] : s_t \in S^g | \pi, g) \end{aligned} \quad (2)$$

The sparse indicator reward function of Eq. (1) is not only simple but also represents a property of many real-world goal problems: in many settings, it may be difficult to tell whether the agent is getting closer to achieving a goal, but easy to tell when a goal has been achieved (e.g. in a maze). In theory, one could hand-engineer a meaningful distance function for each task that could be used to create a dense reward function. Instead, our method is able to learn simply using the indicator function of Eq. (1).

### 3.2. Overall Objective

We desire to find a policy  $\pi(a_t | s_t, g)$  that achieves a high reward for many goals  $g$ . We assume that there is a test distribution of goals  $p_g(g)$  that we would like to perform well on. For simplicity, we assume that the test distribution samples goals uniformly from the set of goals  $\mathcal{G}$ , although in practice any distribution can be used. The overall objective is then to find a policy  $\pi^*$  such that

$$\pi^*(a_t | s_t, g) = \arg \max_{\pi} \mathbb{E}_{g \sim p_g(\cdot)} R^g(\pi). \quad (3)$$

Recall from Eq. (2) that  $R^g(\pi)$  is the probability of success for each goal  $g$ . Thus the objective of Eq. (3) measures the average probability of success over all goals sampled from  $p_g(g)$ . We refer to the objective in Eq. (3) as the coverage.

### 3.3. Assumptions

Similar to previous work (Schaul et al., 2015; Kupcsik et al., 2013; Fabisch & Metzen, 2014; Deisenroth et al., 2014) we need a continuous goal-space representation such that a goal-conditioned policy can efficiently generalize over the goals. In particular, we assume that:

1. A policy trained on a sufficient number of goals in some area of the goal-space will learn to interpolate to other goals within that area.
2. A policy trained on some set of goals will provide a good initialization for learning to reach close-by goals, meaning that the policy can occasionally reach them but maybe not consistently.

Furthermore, we assume that if a goal is reachable, there exists a policy that does so reliably. This is a reasonable assumption for any practical robotics problem, and it will be key for our method, as it strives to train on every goal until it is consistently reached.

## 4. Method

Our approach can be broken down into three parts: First, we label a set of goals based on whether they are at the appropriate level of difficulty for the current policy. Second, using these labeled goals, we train a generator to output new goals at the appropriate level of difficulty. Finally, we use these new goals to efficiently train the policy, improving its coverage objective. We iterate through each of these steps until the policy converges.

### 4.1. Goal Labeling

As shown in our experiments, sampling goals from  $p_g(g)$  directly, and training our policy on each sampled goal may not be the most sample efficient way to optimize the coverage objective of Eq. (3). Instead, we modify the distribution from which we sample goals during training to be uniform over the set of *Goals of Intermediate Difficulty* (*GOID*):

$$GOID_i := \{g : R_{\min} \leq R^g(\pi_i) \leq R_{\max}\} \subseteq \mathcal{G}. \quad (4)$$

The justification for this is as follows: due to the sparsity of the reward function, for most goals  $g$ , the current policy  $\pi_i$  (at iteration  $i$ ) obtains no reward. Instead, we wish to train our policy on goals  $g$  for which  $\pi_i$  is able to receive some minimum expected return  $R^g(\pi_i) > R_{\min}$  such that the agent receives enough reward signal for learning. On the other hand, with this single restriction, we might sample repeatedly from a small set of already mastered goals. To force our policy to train on goals that still need improvement, we also ask for  $R^g(\pi_i) \leq R_{\max}$ , where  $R_{\max}$  is a hyperparameter setting a maximum level of performance above which we prefer to concentrate on new goals. Note

that from Eq. (2),  $R_{\min}$  and  $R_{\max}$  can be interpreted as **a minimum and maximum probability of reaching a goal over  $T$  time-steps**. Training our policy on goals in  $GOID_i$  allows us to efficiently maximize the coverage objective of Eq. (3). Therefore, we need to approximate the sampling from  $GOID_i$ . We propose to first estimate the label  $y_g \in \{0, 1\}$  that indicates whether  $g \in GOID_i$  for all goals  $g$  used in the previous training iteration, and then use these labels to train a generative model from where we can sample goals to train on the next iteration. We estimate the label of a goal  $g$  by computing the fraction of success among all trajectories that had this goal  $g$  during the previous training iteration, and then check whether this estimate is in between  $R_{\min}$  and  $R_{\max}$ . In all our experiments we use 0.1 and 0.9 respectively, although the algorithm is very robust to these hyperparameters (any value of  $R_{\min} \in (0, 0.25)$  and  $R_{\max} \in (0.75, 1)$  would yield basically the same result, as shown in Appendix C)

## 4.2. Adversarial Goal Generation

In order to sample new goals  $g$  uniformly from  $GOID_i$ , we introduce an adversarial training procedure called “goal GAN”, which is a modification of the procedure used for training Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). The modification allows us to train the generative model both with positive examples from the distribution we want to approximate and negative examples sampled from a distribution that does not share support with the desired one. This improves the accuracy of the generative model despite being trained with few positive samples. Our choice of GANs for goal generation is motivated both from this training from negative examples, as well as their ability to generate very high dimensional samples such as images (Goodfellow et al., 2014) which is important for scaling up our approach to goal generation in high-dimensional goal spaces. Other generative models like Stochastic Neural Networks (Tang & Salakhutdinov, 2013) don’t accept negative examples, and don’t scale to higher dimensions.

We use a “goal generator” neural network  $G(z)$  to generate goals  $g$  from a noise vector  $z$ . We train  $G(z)$  to uniformly output goals in  $GOID_i$  using a second “goal discriminator” network  $D(g)$ . The latter is trained to distinguish goals that are in  $GOID_i$  from goals that are not in  $GOID_i$ . We optimize our  $G(z)$  and  $D(g)$  in a manner similar to that of the Least-Squares GAN (LSGAN) (Mao et al., 2017), which we modify by introducing the binary label  $y_g$  allowing us to train from “negative examples” when  $y_g = 0$ :

$$\begin{aligned} \min_D V(D) &= \mathbb{E}_{g \sim p_{\text{data}}(g)} \left[ y_g (D(g) - b)^2 + \right. \\ &\quad \left. (1 - y_g) (D(g) - a)^2 \right] + \mathbb{E}_{z \sim p_z(z)} [(D(G(z)) - a)^2] \\ \min_G V(G) &= \mathbb{E}_{z \sim p_z(z)} [D(G(z)) - c]^2 \end{aligned} \quad (5)$$

We directly use the original hyperparameters reported in Mao et al. (2017) in all our experiments ( $a = -1$ ,  $b = 1$ , and  $c = 0$ ). The LSGAN approach gives us a considerable improvement in training stability over vanilla GAN, and it has a comparable performance to WGAN (Arjovsky et al., 2017). However, unlike in the original LSGAN paper (Mao et al., 2017), we have three terms in our value function  $V(D)$  rather than the original two. For goals  $g$  for which  $y_g = 1$ , the second term disappears and we are left with only the first and third terms, which are identical to that of the original LSGAN framework. Viewed in this manner, the discriminator is trained to discriminate between goals from  $p_{\text{data}}(g)$  with a label  $y_g = 1$  and the generated goals  $G(z)$ . Looking at the second term, our discriminator is also trained with “negative examples” with a label  $y_g = 0$  which our generator should not generate. The generator is trained to “fool” the discriminator, i.e. to output goals that match the distribution of goals in  $p_{\text{data}}(g)$  for which  $y_g = 1$ .

## 4.3. Policy Optimization

### Algorithm 1 Generative Goal Learning

---

```

Input: Policy  $\pi_0$ 
Output: Policy  $\pi_N$ 
 $(G, D) \leftarrow \text{initialize\_GAN}()$ 
 $goals_{\text{old}} \leftarrow \emptyset$ 
for  $i \leftarrow 1$  to  $N$  do
     $z \leftarrow \text{sample\_noise}(p_z(\cdot))$ 
     $goals \leftarrow G(z) \cup \text{sample}(goals_{\text{old}})$ 
     $\pi_i \leftarrow \text{update\_policy}(goals, \pi_{i-1})$ 
     $returns \leftarrow \text{evaluate\_policy}(goals, \pi_i)$ 
     $labels \leftarrow \text{label\_goals}(returns)$ 
     $(G, D) \leftarrow \text{train\_GAN}(goals, labels, G, D)$ 
     $goals_{\text{old}} \leftarrow \text{update\_replay}(goals)$ 
end for

```

---

Our full algorithm for training a policy  $\pi(a_t | s_t, g)$  to maximize the coverage objective in Eq. (3) is shown in Algorithm 1. At each iteration  $i$ , we generate a set of goals by first using `sample_noise` to obtain a noise vector  $z$  from  $p_z(\cdot)$  and then passing this noise to the generator  $G(z)$ . We use these goals to train our policy using RL, with the reward function given by Eq. (1) (`update_policy`). Any RL algorithm can be used for training; in our case we use TRPO with GAE (Schulman et al., 2015b). Our policy’s empirical performance on these goals (`evaluate_policy`) is used to determine each goal’s label  $y_g$  (`label_goals`), as described in Section 4.1. Next, we use these labels to train our goal generator and our goal discriminator (`train_GAN`), as described in Section 4.2. The generated goals from the previous iteration are used to compute the Monte Carlo estimate of the expectations with respect to the distribution  $p_{\text{data}}(g)$  in Eq. (5). By training on goals within  $GOID_i$  produced by the goal generator, our method efficiently finds

a policy that optimizes the coverage objective. For details on how we initialize the goal GAN (`initialize_GAN`), and how we use a replay buffer to prevent “catastrophic forgetting” (`update_replay`), see Appendix A.

The algorithm described above naturally creates a curriculum. The goal generator is updated along with the policy to generate goals in  $GOID_i$ , for which our current policy  $\pi_i$  obtains an intermediate level of return. Thus such goals are always at the appropriate level of difficulty. However, the curriculum occurs as a by-product via our optimization, without requiring any prior knowledge of the environment or the tasks that the agent must perform.

## 5. Experimental Results

In this section we provide the experimental results to answer the following questions:

- Does our automatic curriculum yield faster maximization of the coverage objective?
- Does our Goal GAN dynamically shift to sample goals of the appropriate difficulty (i.e. in  $GOID_i$ )?
- Can our Goal GAN track complex multimodal goal distributions  $GOID_i$ ?
- Does it scale to higher-dimensional goal-spaces with a low-dimensional space of feasible goals?

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 (CoM) of a dynamically complex quadruped agent. In the first experiment the agent has no constraints (see Fig. 1a) and in the second one the agent is inside a U-maze (see Fig. 1b). To answer the third question, we train a point-mass agent to reach any point within a multi-path maze (see Fig. 1d). To answer the final question, we study how our method scales with the dimension of the goal-space in an environment where the feasible region is kept of approximately constant volume in an embedding space that grows in dimension (see Fig. 1c for the 3D case). We compare our *Goal GAN* method against four baselines. *Uniform Sampling* is a method that does not use a curriculum at all, training at every iteration on goals uniformly sampled from the goal-space. To demonstrate that a straight-forward distance reward can be prone to local minima, *Uniform Sampling with L2 loss* samples goals in the same fashion as the first baseline, but instead of the indicator reward that our method uses, it receives the negative L2 distance to the goal as a reward at every step. We have also adapted two methods from the literature to our setting: *Asymmetric Self-play* (Sukhbaatar et al., 2017) and SAGG-RIAC (Baranes & Oudeyer, 2013b). Finally, we provide an ablation and an oracle for our method to better understand the importance of sampling goals of intermediate difficulty  $g \in GOID_i$ . The ablation *GAN fit all* consists on train-

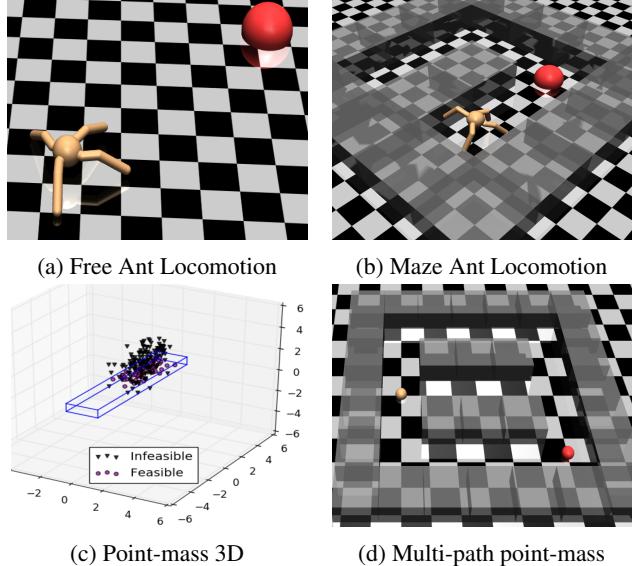


Figure 1. In 1a-1d, the red areas are goals reachable by the orange agent. In 1c any point within the blue frame is a feasible goal (purple balls) and the rest are unfeasible (black triangles).

ing the GAN not only on the goals  $g \in GOID_i$  but rather on every goal attempted in the previous iteration. Given the noise injected at the output of the GAN this generates a gradually expanding set of goals - similar to any hand-designed curriculum. The oracle consists in sampling goals uniformly from the feasible state-space, but only keeping them if they satisfy the criterion in Eq. (4) defining  $GOID_i$ . This *Rejection Sampling* method is orders of magnitude more expensive in terms of labeling, but it serves to estimate an upper-bound for our method in terms of performance.

### 5.1. Ant Locomotion

We test our method in two challenging environments of a complex robotic agent navigating either a free space (Free Ant, Fig. 1a) or a U-shaped maze (Maze Ant, Fig. 1b). Duan et al. (2016) describe the task of trying to reach the other end of the U-turn, and they show that standard RL methods are unable to solve it. We further extend the task to ask to be able to reach any given point within the maze, or within the  $[-5, 5]^2$  square for Free Ant. The reward is still a sparse indicator function being 1 only when the  $(x, y)$  CoM of the Ant is within  $\epsilon = 0.5$  of the goal. Therefore the goal space is 2 dimensional, the state-space is 41 dimensional, and the action space is 8 dimensional (see Appendix B.1).

We first explore whether, by training on goals that are generated by our Goal GAN, we are able to improve our policy’s training efficiency, compared to the baselines described above. In Figs. 2a-Fig. 2b we see that our method leads to faster training compared to the baselines. The *Uniform Sampling* baseline does very poorly because too many samples

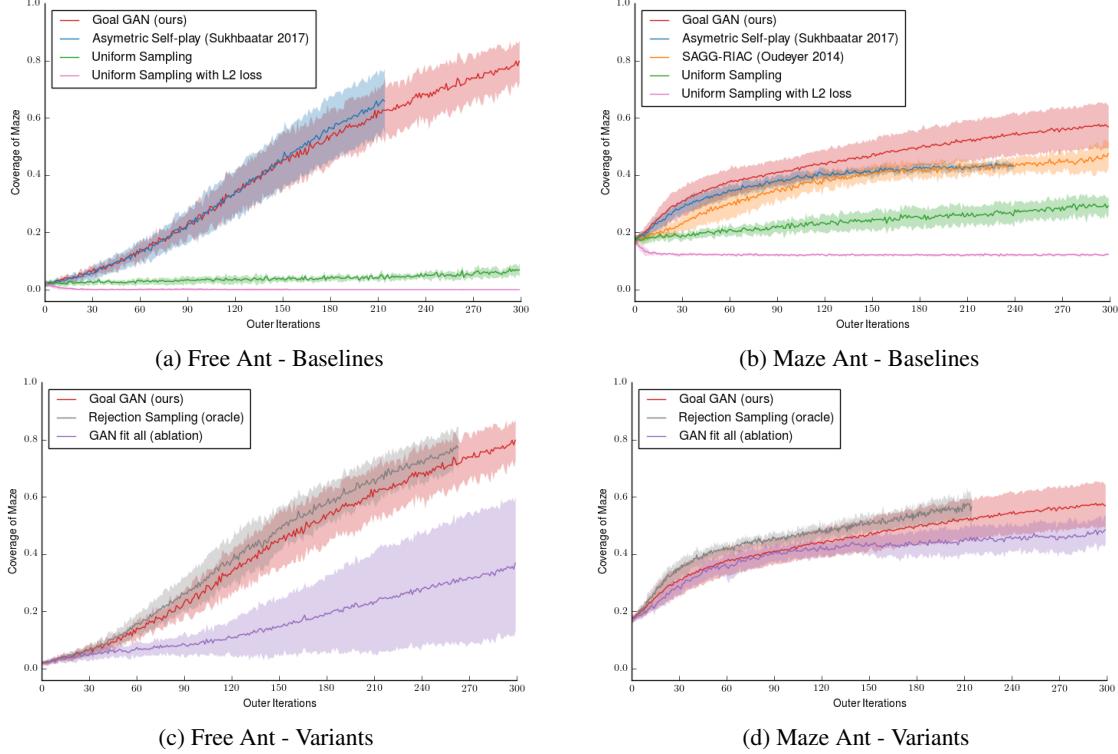


Figure 2. Learning curves comparing the training efficiency of our Goal GAN method and different baselines (first row) and variants (second row), for the Free Ant (left column) and the Maze Ant (right column). The y-axis indicates the average return over all feasible goals. The x-axis shows the number of times that new goals have been sampled. All plots average over 10 random seeds.

are wasted attempting to train on goals that are infeasible or not reachable by the current policy - hence not receiving any learning signal. If an L2 loss is added to try to guide the learning, the agent falls into a poor local optima of not moving to avoid further negative rewards. The two other baselines that we compare against perform better, but still do not surpass the performance of our method. In particular, Asymmetric Self-play needs to train the goal-generating policy (Alice) at every outer iteration, with an amount of rollouts equivalent to the ones used to train the goal-reaching policy. This additional burden is not represented in the plots, being therefore at least half as sample-efficient as the plots indicate. SAGG-RIAC maintains an ever-growing partition of the goal-space that becomes more and more biased towards areas that already have more sub-regions, leading to reduced exploration and slowing down the expansion of the policy’s capabilities. Details of our adaptation of these two methods to our problem, as well as further study of their failure cases, is provided in the Appendices F.1 and F.2.

To better understand the efficiency of our method, we analyze the goals generated by our automatic curriculum. In these Ant navigation experiments, the goal space is two dimensional, allowing us to study the shift in the probability distribution generated by the Goal GAN (Fig. 3 for the Maze Ant) along with the improvement of the policy

coverage (Fig. 4 for the Maze Ant). We have indicated the difficulty of reaching the generated goals in Fig. 3. It can be observed in these figures that the location of the generated goals shifts to different parts of the maze, concentrating on the area where the current policy is receiving some learning signal but needs more improvement. The percentage of generated goals that are at the appropriate level of difficulty ( $GOID_i$ ) stays around 20% even as the policy improves. The goals in these figures include a mix of newly generated goals from the Goal GAN as well as goals from previous iterations that we use to prevent our policy from “forgetting” (Appendix A.1). Overall it is clear that our Goal GAN dynamically shift to sample goals of the appropriate difficulty. See Appendix D and Fig. 9-10 therein for the analogous analysis of Free Ant, where we observe that Goal GAN produces a growing ring of goals around the origin.

It is interesting to analyze the importance of generating goals in  $GOID_i$  for efficient learning. This is done in Figs. 2c-2d, where we first show an ablation of our method *GAN fit all*, that disregards the labels. This method performs worse than ours, because the expansion of the goals is not related to the current performance of the policy. Finally, we study the *Rejection Sampling oracle*. As explained in Section 4.1, we wish to sample from the set of goals  $GOID_i$ , which we approximate by fitting a Goal GAN to the distribution of good

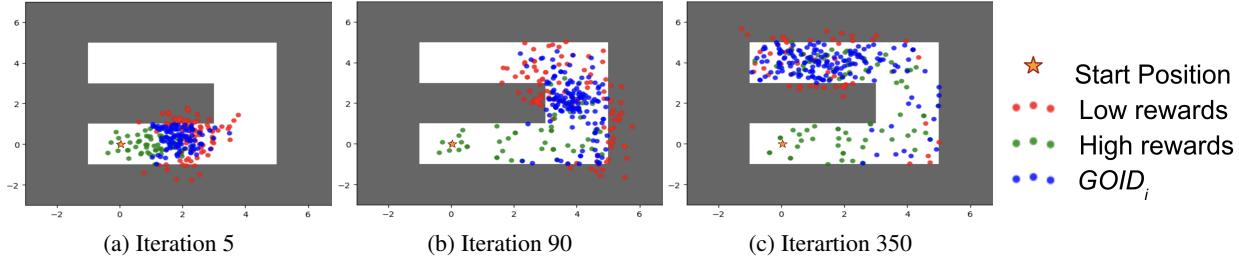


Figure 3. Goals that, at iterations  $i$ , our algorithm trains on - 200 sampled from Goal GAN, 100 from replay. Green goals satisfy  $\bar{R}^g(\pi_i) \geq R_{\max}$ . Blue ones have appropriate difficulty for the current policy  $R_{\min} \leq \bar{R}^g(\pi_i) \leq R_{\max}$ . The red ones have  $R_{\min} \geq \bar{R}^g(\pi_i)$ .

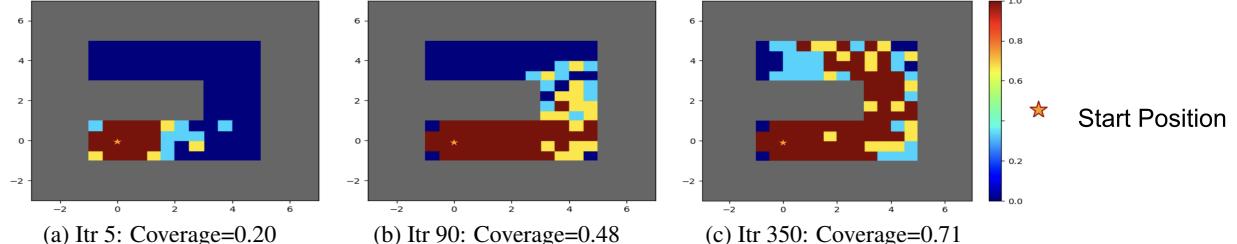


Figure 4. Visualization of the policy performance (same policy training as in Fig. 3). For illustration purposes, each grid cell is colored according to the expected return achieved when fixing its center as goal: Red indicates 100% success; blue indicates 0% success.

goals observed in the previous policy optimization step. We evaluate now how much this approximation affects learning by comparing the learning performance of our Goal GAN to a policy trained on goals sampled uniformly from  $GOID_i$  by using rejection sampling. This method is orders of magnitude more sample inefficient, but gives us an upper bound on the performance of our method. Figs. 2c-2d demonstrate that our performance is quite close to the performance of this much less efficient baseline.

## 5.2. Multi-path point-mass maze

In this section we show that our Goal GAN method is efficient at tracking clearly multi-modal distributions of goals  $g \in GOID_i$ . To this end, we introduce a new maze environment with multiple paths, as can be seen in Fig. 1d. To keep the experiment simple we replace the Ant agent by a point-mass, which actions are the velocity vector (2 dim). As in the other experiments, our aim is to learn a policy that can reach any feasible goal corresponding to  $\epsilon$ -balls in state space, like the one depicted in red.

Similar to the experiments in Figures 3 and 4, here we show the goals that our algorithm generated to train the Multi-path point-mass agent. Figures 5 and 6 show the results. It can be observed that our method produces a multi-modal distribution over goals, tracking all the areas where goals are at the appropriate level of difficulty. Note that the samples from the regularized replay buffer are responsible for the trailing spread of “High Reward” goals and the Goal GAN is responsible for the more concentrated nodes (see only Goal GAN samples in Appendix Fig. 11). A clear benefit

of using our Goal GAN as a generative model is that no prior knowledge about the distribution to fit is required (like the number of modes). Finally, having several possible paths to reach a specific goal does not hinder the learning of our algorithm that consistently reaches full coverage in this problem (see Appendix Fig. 12).

## 5.3. N-dimensional Point Mass

In many real-world RL problems, the set of feasible states is a lower-dimensional subset of the full state space, defined by the constraints of the environment. For example, the kinematic constraints of a robot limit the set of feasible states that the robot can reach. In this section we use an N-dimensional Point Mass to demonstrate the performance of our method as the embedding dimension increases.

In this experiments, the full state-space of the  $N$ -dimensional Point Mass is the hypercube  $[-5, 5]^N$ . However, the Point Mass 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  $L_2$  distance between points in higher dimensions. The fraction of the volume of the feasible space decreases as  $N$  increases (e.g. 0.00023:1 for  $N = 6$ ).

We compare the performance of our method to the baselines in Fig. 7. The uniform sampling baseline has poor

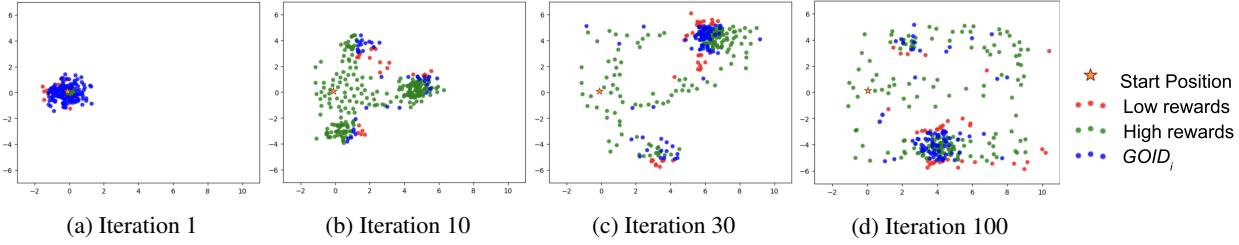


Figure 5. Goals that, at iterations  $i$ , our algorithm trains on - 200 sampled from Goal GAN, 100 from replay. Green goals satisfy  $\bar{R}^g(\pi_i) \geq R_{\max}$ . Blue ones have appropriate difficulty for the current policy  $R_{\min} \leq \bar{R}^g(\pi_i) \leq R_{\max}$ . The red ones have  $R_{\min} \geq \bar{R}^g(\pi_i)$ .

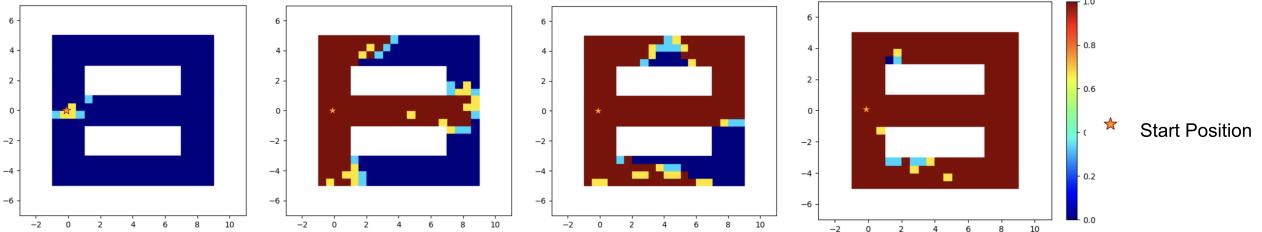


Figure 6. Visualization of the policy performance (same policy training as in Fig. 5). For illustration purposes, each grid cell is colored according to the expected return achieved when fixing its center as goal: Red indicates 100% success; blue indicates 0% success.

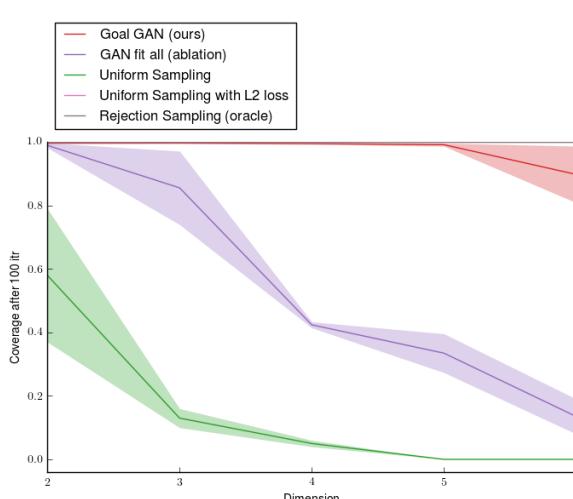


Figure 7. Final goal coverage obtained after 200 outer iterations on the N-dim point mass environment. All plots average over 5 random seeds.

performance as the number of dimensions increases because the fraction of feasible states within the full state space decreases as the dimension increases. Thus, sampling uniformly results in sampling an increasing percentage of unfeasible goals, leading to poor learning signal. In contrast, the performance of our method does not decay as much as the state space dimension increases, because our Goal GAN always generates goals within the feasible portion of the state space. The *GAN fit all* variation of our method suffers from the increase in dimension because it is not encouraged to track the narrow feasible region. Finally, the oracle and the baseline with an L2 distance reward have perfect per-

formance, which is expected in this task where the optimal policy is just to go in a straight line towards the goal. Even without this prior knowledge, the Goal GAN discovers the feasible subset of the goal space.

## 6. Conclusions and Future Work

We propose a new paradigm in RL where the objective is to train a single policy to succeed on a variety of goals, under sparse rewards. To solve this problem we develop a method for automatic curriculum generation that dynamically adapts to the current performance of the agent. The curriculum is obtained without any prior knowledge of the environment or of the tasks being performed. We use generative adversarial training to automatically generate goals for our policy that are always at the appropriate level of difficulty (i.e. not too hard and not too easy). In the future we want to combine our goal-proposing strategy with recent multi-goal approaches like HER (Andrychowicz et al., 2017) that could greatly benefit from better ways to select the next goal to train on. Another promising line of research is to build hierarchy on top of the multi-task policy that we obtain with our method by training a higher-level policy that outputs the goal for the lower level multi-task policy (Heess et al., 2016; Florensa et al., 2017a). The hierarchy could also be introduced by replacing our current feed-forward neural network policy by an architecture that learns to build implicit plans (Mnih et al., 2016; Tamar et al., 2016), or by leveraging expert demonstrations to extract sub-goals (Zheng et al., 2016), although none of these approaches tackles yet the multi-task learning problem formulated in this work.

## 7. Acknowledgments

This work was supported in part by Berkeley Deep Drive, ONR PECASE N000141612723, Darpa FunLoL. Carlos Florensa was also supported by a La Caixa Fellowship.

## References

- Achiam, J. and Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning. In *2016 Deep Reinforcement Learning Workshop at NIPS*, 2016.
- Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. In *Stat*, 2017.
- Baldassarre, G. and Mirolli, M. Temporal-difference competence-based intrinsic motivation (td-cb-im): A mechanism that uses the td-error as an intrinsic reinforcement for deciding which skill to learn when. *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 255–276, 2012.
- Baranes, A. and Oudeyer, P.-Y. Intrinsically motivated goal exploration for active motor learning in robots: A case study. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1766–1773. IEEE, 2010.
- Baranes, A. and Oudeyer, P.-Y. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013a.
- Baranes, A. and Oudeyer, P.-Y. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1), 2013b.
- Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in Neural Information Processing Systems*, 2016.
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *International Conference on Machine Learning*, pp. 41–48. ACM, 2009.
- Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1–2):1–142, 2013.
- Deisenroth, M. P., Englert, P., Peters, J., and Fox, D. Multi-task policy search for robotics. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 3876–3881. IEEE, 2014.
- Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. *International Conference on Machine Learning*, 2016.
- Fabisch, A. and Metzen, J. H. Active contextual policy search. *Journal of Machine Learning Research*, 15(1): 3371–3399, 2014.
- Florensa, C., Duan, Y., and Abbeel, P. Stochastic neural networks for hierarchical reinforcement learning. In *International Conference in Learning Representations*, 2017a.
- Florensa, C., Held, D., Wulfmeier, M., and Abbeel, P. Reverse curriculum generation for reinforcement learning. *Conference on Robot Learning (CoRL)*, 2017b.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Graves, A., Bellemare, M. G., Menick, J., Munos, R., and Kavukcuoglu, K. Automated Curriculum Learning for Neural Networks. *arXiv preprint arXiv:1704.03003*, 2017.
- Guo, X., Singh, S., Lewis, R., and Lee, H. Deep learning for reward design to improve monte carlo tree search in atari games. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.
- Heess, N., Wayne, G., Tassa, Y., Lillicrap, T., Riedmiller, M., and Silver, D. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv: 1610.05182*, 2016.
- Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. Variational information maximizing exploration. *Advances in Neural Information Processing Systems*, 2016.
- Karpathy, A. and Van De Panne, M. Curriculum learning for motor skills. In *Canadian Conference on Artificial Intelligence*, pp. 325–330. Springer, 2012.
- Konidaris, G., Kuindersma, S., Grupen, R. A., and Barto, A. G. Autonomous skill acquisition on a mobile manipulator. In *AAAI*, 2011.
- Kupcsik, A. G., Deisenroth, M. P., Peters, J., Neumann, G., and Others. Data-efficient generalization of robot skills with contextual policy search. In *AAAI Conference on Artificial Intelligence*, pp. 1401–1407, 2013.

- Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv: 1509.02971*, 2015.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. On the effectiveness of least squares generative adversarial networks. *arXiv preprint arXiv:1712.06391*, 2017.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.
- Mnih, V., Agapiou, J., Osindero, S., Graves, A., Vinyals, O., Kavukcuoglu, K., et al. Strategic attentive writer for learning macro-actions. *ArXiv preprint ArXiv:1606.04695*, 2016.
- Ranchod, P., Rosman, B., and Konidaris, G. Nonparametric bayesian reward segmentation for skill discovery using inverse reinforcement learning. In *International Conference on Intelligent Robots and Systems*, pp. 471–477. IEEE, 2015.
- Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.
- Schmidhuber, J. Curious model-building control systems. In *International Joint Conference on Neural Networks*, pp. 1458–1463. IEEE, 1991.
- Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015a.
- Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-Dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representation*, 2015b.
- Sharma, S. and Ravindran, B. Online Multi-Task Learning Using Biased Sampling. *arXiv preprint arXiv: 1702.06053*, 2017.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- Sukhbaatar, S., Kostrikov, I., Szlam, A., and Fergus, R. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play. *arXiv preprint arXiv: 1703.05407*, 2017.
- Tamar, A., Wu, Y., Thomas, G., Levine, S., and Abbeel, P. Value iteration networks. In *Advances in Neural Information Processing Systems*, pp. 2154–2162, 2016.
- Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. #exploration: A study of count-based exploration for deep reinforcement learning. *CoRR*, abs/1611.04717, 2016.
- Tang, Y. and Salakhutdinov, R. Learning Stochastic Feedforward Neural Networks. *Advances in Neural Information Processing Systems*, 2:530–538, 2013.
- Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- Vigorito, C. M. and Barto, A. G. Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Transactions on Autonomous Mental Development*, 2(2): 132–143, 2010.
- Zaremba, W. and Sutskever, I. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.
- Zheng, S., Yue, Y., and Hobbs, J. Generating long-term trajectories using deep hierarchical networks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*. 2016.

## A. Implementation details

### A.1. Replay buffer

In addition to training our policy on the goals that were generated in the current iteration, we also save a list (“regularized replay buffer”) of goals that were generated during previous iterations (`update_replay`). These goals are also used to train our policy, so that our policy does not forget how to achieve goals that it has previously learned. When we generate goals for our policy to train on, we sample two thirds of the goals from the Goal GAN and we sample the one third of the goals uniformly from the replay buffer. To prevent the replay buffer from concentrating in a small portion of goal space, we only insert new goals that are further away than  $\epsilon$  from the goals already in the buffer, where we chose the goal-space metric and  $\epsilon$  to be the same as the ones introduced in Section 3.1.

### A.2. Goal GAN Initialization

In order to begin our training procedure, we need to initialize our goal generator to produce an initial set of goals (`initialize_GAN`). If we initialize the goal generator randomly (or if we initialize it to sample uniformly from the goal space), it is likely that, for most (or all) of the sampled goals, our initial policy would receive no reward due to the sparsity of the reward function. Thus we might have that all of our initial goals  $g$  have  $\bar{R}^g(\pi_0) < R_{\min}$ , leading to very slow training.

To avoid this problem, we initialize our goal generator to output a set of goals that our initial policy is likely to be able to achieve with  $\bar{R}^g(\pi_i) \geq R_{\min}$ . To accomplish this, we run our initial policy  $\pi_0(a_t | s_t, g)$  with goals sampled uniformly from the goal space. We then observe the set of states  $S^v$  that are visited by our initial policy. These are states that can be easily achieved with the initial policy,  $\pi_0$ , so the goals corresponding to such states will likely be contained within  $S_0^I$ . We then train the goal generator to produce goals that match the state-visitation distribution  $p_v(g)$ , defined as the uniform distribution over the set  $f(S^v)$ . We can achieve this through traditional GAN training, with  $p_{\text{data}}(g) = p_v(g)$ . This initialization of the generator allows us to bootstrap the Goal GAN training process, and our policy is able to quickly improve its performance.

## B. Experimental details

### B.1. Ant specifications

The ant is a quadruped with 8 actuated joints, 2 for each leg. The environment is implemented in Mujoco (Todorov et al., 2012). Besides the coordinates of the center of mass, the joint angles and joint velocities are also included in the observation of the agent. The high degrees of freedom make

navigation a quite complex task requiring motor coordination. More details can be found in Duan et al. (2016), and the only difference is that in our goal-oriented version of the Ant we append the observation with the goal, the vector from the CoM to the goal and the distance to the goal. For the Free Ant experiments the objective is to reach any point in the square  $[-5m, 5m]^2$  on command. The maximum time-steps given to reach the current goal are 500.

### B.2. Ant Maze Environment

The agent is constrained to move within the maze environment, which has dimensions of 6m x 6m. The full state-space has an area of size 10 m x 10 m, within which the maze is centered. To compute the coverage objective, goals are sampled from within the maze according to a uniform grid on the maze interior. The maximum time-steps given to reach the current goal are 500.

### B.3. Point-mass specifications

For the N-dim point mass of Section 5.3, in each episode (rollout) the point-mass has 400 timesteps to reach the goal, where each timestep is 0.02 seconds. The agent can accelerate in up to a rate of  $5 \text{ m/s}^2$  in each dimension ( $N = 2$  for the maze). The observations of the agent are  $2N$  dimensional, including position and velocity of the point-mass.

### B.4. Goal GAN design and training

After the generator generates goals, we add noise to each dimension of the goal sampled from a normal distribution with zero mean and unit variance. At each step of the algorithm, we train the policy for 5 iterations, each of which consists of 100 episodes. After 5 policy iterations, we then train the GAN for 200 iterations, each of which consists of 1 iteration of training the discriminator and 1 iteration of training the generator. The generator receives as input 4 dimensional noise sampled from the standard normal distribution. The goal generator consists of two hidden layers with 128 nodes, and the goal discriminator consists of two hidden layers with 256 nodes, with relu nonlinearities.

### B.5. Policy and optimization

The policy is defined by a neural network which receives as input the goal appended to the agent observations described above. The inputs are sent to two hidden layers of size 32 with tanh nonlinearities. The final hidden layer is followed by a linear  $N$ -dimensional output, corresponding to accelerations in the  $N$  dimensions. For policy optimization, we use a discount factor of 0.998 and a GAE lambda of 0.995. The policy is trained with TRPO with Generalized Advantage Estimation implemented in rllab (Schulman et al., 2015a;b; Duan et al., 2016). Every “`update_policy`” consists of 5

iterations of this algorithm.

### C. Study of GoalGAN goals

To label a given goal (Section 4.1), we could empirically estimate the expected return for this goal  $\bar{R}^g(\pi_i)$  by performing rollouts of our current policy  $\pi_i$ . The label for this goal is then set to  $y_g = \mathbb{1}\{R_{\min} \leq \bar{R}^g(\pi_i) \leq R_{\max}\}$ . Nevertheless, having to execute additional rollouts just for labeling is not sample efficient. Therefore, we instead use the rollouts that were used for the most recent policy update. This is an approximation as the rollouts where performed under  $\pi_{i-1}$ , but as we show in Figs. 8a-8b, this small “delay” does not affect learning significantly. Indeed, using the true label (estimated with three new rollouts from  $\pi_i$ ) yields the *Goal GAN true label* curves that are only slightly better than what our method does. Furthermore, no matter what labeling technique is used, the success rate of most goals is computed as an average of at most four attempts. Therefore, the statement  $R_{\min} \leq \bar{R}^g(\pi_i)$  will be unchanged for any value of  $R_{\min} \in (0, 0.25]$ . Same for  $\bar{R}^g(\pi_i) \leq R_{\max}$  and  $R_{\max} \in (0.75, 1)$ . This implies that the labels estimates (and hence our automatic curriculum generation algorithm) is almost invariant for any value of the hyperparameters  $R_{\min}$  and  $R_{\max}$  in these ranges.

In the same plots we also study another criteria to choose the goals to train on that has been previously used in the literature: learning progress (Baranes & Oudeyer, 2013b; Graves et al., 2017). Given that we work in a continuous goal-space, estimating the learning progress of a single goal requires estimating the performance of the policy on that goal before the policy update and after the policy update (potentially being able to replace one of these estimations with the rollouts from the policy optimization, but not both). Therefore the method does require more samples, but we deemed interesting to compare how well the metrics allow to automatically build a curriculum. We see in the Figs. 8a-8b that the two metrics yield a very similar learning, at least in the case of Ant navigation tasks with sparse rewards.

### D. Goal Generation for Free Ant

Similar to the experiments in Figures 3 and 4, here we show the goals that were generated for the Free Ant experiment in which a robotic quadruped must learn to move to all points in free space. Figures 9 and 10 show the results. As shown, our method produces a growing circle around the origin; as the policy learns to move the ant to nearby points, the generator learns to generate goals at increasingly distant positions.

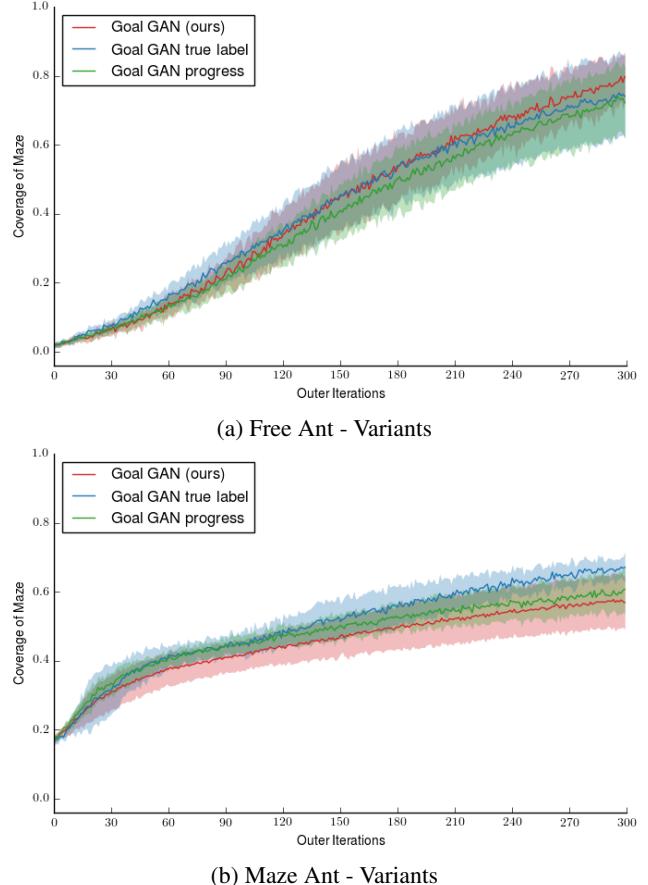


Figure 8. Learning curves comparing the training efficiency of our method and different variants. All plots are an average over 10 random seeds.

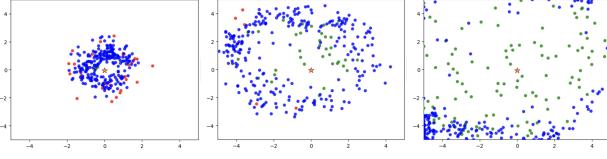
### E. Learning for Multi-path point-mass

To clearly observe that our GoalGAN approach is capable of fitting multimodal distributions, we have plotted in Fig. 11 only the samples coming from the GoalGAN (i.e. no samples from the replay buffer). Also, in this environment there are several ways of reaching every part of the maze. This is not a problem for our algorithm, as can be seen in the full learning curves in Fig. 12, where we see that all runs of the algorithm reliably reaches full coverage of the multi-path maze.

### F. Comparisons with other methods

#### F.1. Asymmetric self-play (Sukhbaatar et al., 2017)

Although not specifically designed for the problem presented in this paper, it is straight forward to apply the method proposed by Sukhbaatar et al. (2017) to our problem. An interesting study of its limitations in a similar setting can be found in (Florensa et al., 2017b).



(a) Iteration 10      (b) Iteration 100      (c) Iterartion 300

Figure 9. Goals that our algorithm trains on (200 sampled from the Goal GAN, 100 from the replay). “High rewards” (green) are goals with  $\bar{R}^g(\pi_i) \geq R_{\max}$ ;  $GOID_i$  (blue) have appropriate difficulty for the current policy  $R_{\min} \leq \bar{R}^g(\pi_i) \leq R_{\max}$ . The red ones have  $R_{\min} \geq \bar{R}^g(\pi_i)$

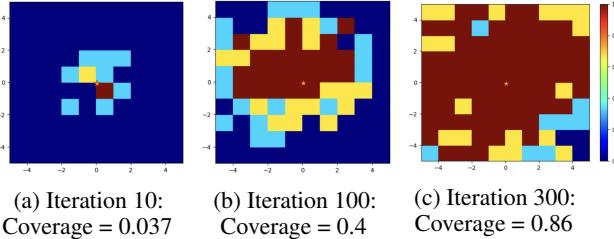


Figure 10. Visualization of the policy performance for different parts of the state space (same policy training as in Fig. 9). For illustration purposes, the feasible state-space is divided into a grid, and a goal location is selected from the center of each grid cell. Each grid cell is colored according to the expected return achieved on this goal: Red indicates 100% success; blue indicates 0% success.

## F.2. SAGG-RIAC (Baranes & Oudeyer, 2013b)

In our implementation of this method, we use TRPO as the “Low-Level Goal-Directed Exploration with Evolving Context”. We therefore implement the method as batch: at every iteration, we sample  $N_{new}$  new goals  $\{y_i\}_{i=0\dots N_{new}}$ , then we collect rollouts of  $t_{max}$  steps trying to reach them, and perform the optimization of the parameters using all the collected data. The detailed algorithm is given in the following pseudo-code.

**UpdateRegions( $\mathbf{R}, y_f, \Gamma_{y_f}$ )** is exactly the Algorithm 2 described in the original paper, and **Self-generate** is the ”Active Goal Self-Generation (high-level)” also described in the paper (Section 2.4.4 and Algorithm 1), but it’s repeated  $N_{new}$  times to produce a batch of  $N_{new}$  goals jointly. As for the competence  $\Gamma_{y_g}$ , we use the same formula as in their section 2.4.1 (use highest competence if reached close enough to the goal) and  $C(y_g, y_f)$  is computed with their equation (7). The `collect_rollout` function resets the state  $s_0 = s_{reset}$  and then applies actions following the goal-conditioned policy  $\pi_\theta(\cdot, y_g)$  until it reaches the goal or the maximum number of steps  $t_{max}$  has been taken. The final state, transformed in goal space,  $y_f$  is returned.

As hyperparameters, we have used the recommended ones in the paper, when available:  $p_1 = 0.7$ ,  $p_2 = 0.2$ ,  $p_3 = 0.1$ .

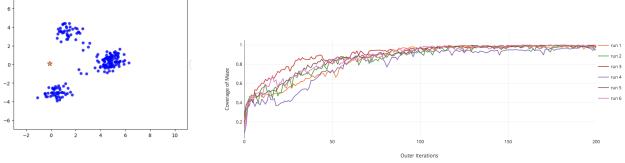


Figure 11. Iteration 10 Goal GAN samples (Fig. 5b without replay samples)

Figure 12. Learning curves of our algorithm on Multi-path Point-mass Maze, consistently achieving full coverage

## Algorithm 2 Generative Goal with Sagg-RIAC

**Hyperparameters:** window size  $\zeta$ , tolerance threshold  $\epsilon_{max}$ , competence threshold  $\epsilon_C$ , maximum time horizon  $t_{max}$ , number of new goals  $N_{new}$ , maximum number of goals  $g_{max}$ , mode proportions  $(p_1, p_2, p_3)$

**Input:** Policy  $\pi_{\theta_0}(s_{start}, y_g)$ , goal bounds  $B_Y$ , reset position  $s_{rest}$

**Output:** Policy  $\pi_{\theta_N}(s_{start}, y_g)$

$\mathbf{R} \leftarrow \{(R_0, \Gamma_{R_0})\}$  where  $R_0 = Region(B_Y)$ ,  $\Gamma_{R_0} = 0$   
**for**  $i \leftarrow 1$  **to**  $N$  **do**

$goals \leftarrow$  **Self-generate**  $N_{new}$  goals:  $\{y_j\}_{j=0\dots N_{new}}$   
      $paths = []$

**while**  $number\_steps\_in(paths) < batch\_size$  **do**

        Reset  $s_0 \leftarrow s_{rest}$   
          $y_g \leftarrow \text{Uniform}(goals)$   
          $y_f, \Gamma_{y_f}, path \leftarrow \text{collect\_rollout}(\pi_{\theta_i}(\cdot, y_g), s_{reset})$   
          $paths.append(path)$   
         **UpdateRegions( $\mathbf{R}, y_f, 0$ )**  
         **UpdateRegions( $\mathbf{R}, y_g, \Gamma_{y_g}$ )**

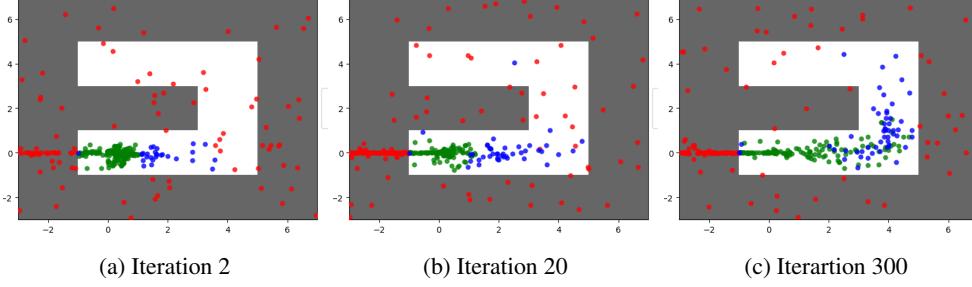
**end while**

$\pi_{\theta_{i+1}} \leftarrow \text{train } \pi_{\theta_i} \text{ with TRPO on collected } paths$

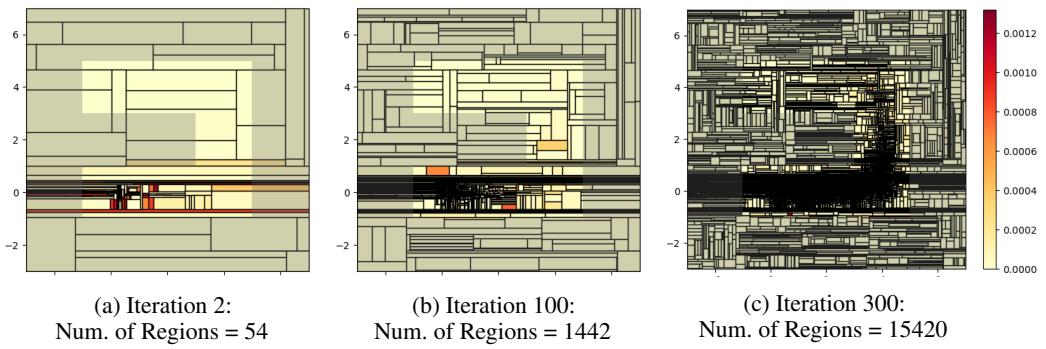
**end for**

For the rest, the best performance in an hyperparameter sweep yields:  $\zeta = 100$ ,  $g_{max} = 100$ . The noise for mode(3) is chosen to be Gaussian with variance 0.1, the same as the tolerance threshold  $\epsilon_{max}$  and the competence threshold  $\epsilon_C$ .

As other details, in our tasks there are no constraints to penalize for, so  $\rho = \emptyset$ . Also, there are no sub-goals. The reset value  $r$  is 1 as we reset to  $s_{start}$  after every reaching attempt. The number of explorative movements  $q \in \mathbb{N}$  has a less clear equivalence as we use a policy gradient update with a stochastic policy  $\pi_\theta$  instead of a SSA-type algorithm.



*Figure 13.* Goals sampled by SAGG-RIAC (same policy training as in Fig. 14). “High rewards” (in green) are goals with  $\bar{R}^g(\pi_i) \geq R_{\max}$ ;  $GOID_i$  (in blue) are those with the appropriate level of difficulty for the current policy ( $R_{\min} \leq \bar{R}^g(\pi_i) \leq R_{\max}$ ). The red ones have  $R_{\min} \geq \bar{R}^g(\pi_i)$



*Figure 14.* Visualization of the regions generated by the SAGG-RIAC algorithm