Confidence-Based Curriculum Learning for Multi-Agent Path Finding

Thomy Phan  
University of Southern California  
Los Angeles, USA  
thomy.phan@usc.edu

Justin Romberg  
Georgia Institute of Technology  
Atlanta, USA  
jrom@ece.gatech.edu

Joseph Driscoll  
Georgia Institute of Technology  
Atlanta, USA  
jdriscoll7@gatech.edu

Sven Koenig  
University of Southern California  
Los Angeles, USA  
skoenig@usc.edu

ABSTRACT

A wide range of real-world applications can be formulated as Multi-Agent Path Finding (MAPF) problem, where the goal is to find collision-free paths for multiple agents with individual start and goal locations. State-of-the-art MAPF solvers are mainly centralized and depend on global information, which limits their scalability and flexibility regarding changes or new maps that would require expensive replanning. Multi-agent reinforcement learning (MARL) offers an alternative way by learning decentralized policies that can generalize over a variety of maps. While there exist some prior works that attempt to connect both areas, the proposed techniques are heavily engineered and very complex due to the integration of many mechanisms that limit generality and are expensive to use. We argue that much simpler and general approaches are needed to bring the areas of MAPF and MARL closer together with significantly lower costs.

In this paper, we propose Confidence-based Auto-Curriculum for Team Update Stability (CACTUS) as a lightweight MARL approach to MAPF. CACTUS defines a simple reverse curriculum scheme, where the goal of each agent is randomly placed within an allocation radius around the agent’s start location. The allocation radius increases gradually as all agents improve, which is assessed by a confidence-based measure. We evaluate CACTUS in various maps of different sizes, obstacle densities, and numbers of agents. Our experiments demonstrate better performance and generalization capabilities than state-of-the-art MARL approaches with less than 600,000 trainable parameters, which is less than 5% of the neural network size of current MARL approaches to MAPF.

1 INTRODUCTION

A wide range of real-world applications like goods transportation in warehouses, search and rescue missions, and traffic management can be formulated as Multi-Agent Path Finding (MAPF) problem, where the goal is to find collision-free paths for multiple agents with individual start and goal locations. Finding optimal solutions w.r.t. flowtime or makespan is NP-hard [29, 39]. Despite the problem complexity, there exist a variety of MAPF solvers that find optimal [32], bounded suboptimal [5], or quick feasible solutions [16]. Most MAPF solvers are centralized and require global information, which limits scalability and flexibility regarding changes that would need expensive replanning. This also limits applicability to partially observable real-time domains [30].

Multi-agent reinforcement learning (MARL) offers an alternative way by learning decentralized policies that can generalize over a variety of maps and make decisions under partial observability [4, 44]. State-of-the-art MARL algorithms are based on centralized training for decentralized execution (CTDE), where training takes place in a laboratory or a simulator with access to global information to learn coordinated policies that can be executed independently under partially observability afterwards [18, 28].

MAPF and MARL have been very active research areas in the last few years with impressive advances on both sides, resulting in...
a variety of sophisticated algorithms [5, 14, 17, 46]. Despite these advances, both fields have been mainly studied independently of each other. However, MARL could benefit MAPF in various ways:

1. **Efficiency:** The learned policies are reactive and decentralized therefore alleviating the computational and communication requirements of centralized MAPF solvers [28].

2. **Generalization:** The learned policies can generalize over a variety of maps thus do not require complete retraining or replanning when being used on new maps [30].

3. **Robustness:** The learned policies make decisions based on actual observations therefore being able to react to local changes, i.e., emerging obstacles or new paths, without requiring replanning of the whole system [24].

On the other hand, MAPF poses an exciting challenge for MARL due to its practical relevance and the following aspects [32, 34]:

1. **Sparse Rewards:** MAPF represents a complex navigation problem, where all agents are only rewarded for reaching their goals. Naive MARL would need exhaustive exploration to obtain informative data, which is time-consuming [10].

2. **Dynamic Constraints:** Agents are not allowed to collide therefore having temporal constraints in addition to static constraints imposed by obstacles and boundaries [34].

3. **Coordination:** MAPF requires coordination of spatially close agents with potentially emergent effects like congestion or circulation. So far, most MARL methods only focus on coordination on a small scale though [18, 51].

We believe that addressing MAPF via MARL can provide a fruitful research direction that would benefit both areas. While there are prior works that attempt to connect these areas, the proposed techniques are heavily engineered and very complex, using extremely neural network size of current MARL approaches to MAPF. Therefore having temporal constraints in addition to static goal is to find a solution that maximizes the utilitarian metric for all agents.

### 2.2 Multi-Agent Reinforcement Learning

MARL problems can be formulated as a partially observable stochastic game \( M = (D, S, A, P, R, Z, \Omega) \), where \( D = \{1, ..., N\} \) is a set of agents, \( S \) is a set of states \( s_t \), \( A = A_1 \times ... \times A_N \) is the set of joint actions \( a_t = (a_{t,1}, ..., a_{t,N}) \), \( P(s_{t+1}|s_t, a_t) \) is the transition probability, \( R(s_t, a_t) = \{r_{t,1}, ..., r_{t,N}\} \in \mathbb{R}^N \) is the joint reward with \( r_{t,i} \) being the reward of agent \( i \in D \), \( Z \) is the set of local observations \( z_{t,i} \) for each agent \( i \), and \( \Omega(s_{t+1}) = \{z_{t+1,1}, ..., z_{t+1,N}\} \in \mathbb{Z}^N \) is the subsequent joint observation. Each agent \( i \) maintains an action-observation history \( \tau_{t,i} = (Z_1 \times A_1)^T, \tau = (\tau_{1,1}, ..., \tau_{N,N}) \) is the joint policy with local policies \( \pi_i \), where \( \pi_i(a_{t,i}|\tau_{t,i}) \) is the action selection probability of agent \( i \). Local policy \( \pi_i \) can be evaluated with a value function \( Q_i^\pi(s_t, a_t) = \mathbb{E}_[R_{t,i}|s_t, a_t] \) for all states \( s_t \in S \) and \( a_t \in A_t \), where \( R_{t,i} = \sum_{k=0}^{T-1} \gamma^k r_{t+k,i} \) is the return of agent \( i, T > 0 \) is the horizon, and \( \gamma \in [0, 1] \) is the discount factor.

In cooperative MARL, the goal is to find an optimal joint policy \( \pi^* = (\pi_1^*, ..., \pi_N^*) \) that maximizes the utilitarian metric for all states \( s_t \in S \):

\[
Q_{opt}(s_t, a_t) = \sum_{i \in D} Q_i^\pi(s_t, a_t)
\]

#### 2.2.1 Policy Gradient MARL

To learn optimal policies \( \pi_i^* \) in large state spaces, function approximators \( \hat{\pi}_{i,\theta} \) with parameters \( \theta \) are trained with gradient ascent on an estimate of \( J = \mathbb{E}_\pi[R_{t,i}] \). Policy gradient methods use gradients \( g \) of the following form [43]:

\[
g = A_i^\pi(s_t, a_t) \nabla_{\theta} \log \hat{\pi}_{i,\theta}(a_{t,i}|\tau_{t,i})
\]

where \( A_i^\pi(s_t, a_t) = Q_i^\pi(s_t, a_t) - V_i^\pi(s_t) \) is the advantage of agent \( i \) and \( V_i^\pi(s_t) = \mathbb{E}_\pi[R_{t,i}|s_t] \) is its state value function. Actor-critic
approaches often approximate $\hat{A}_i = A^\pi_i$ by replacing $Q_i^\pi(s_t, a_t)$ with $\hat{R}_t$ and $V_i^\pi$ with $\hat{E}_t[A_i^\pi Q_i^\pi]$. $Q_i^\pi$ can be approximated with a critic $\hat{Q}_i$, and parameters $\omega$ using value-based RL [20, 49].

Alternatively, $\pi_{1,\theta}$ can be trained via proximal policy optimization (PPO) by iteratively minimizing the following surrogate loss [31]:

$$L_{PPO}^{\pi_{1,\theta}}(\theta) = \mathbb{E}[\min\{\hat{A}_i(\phi_{1,i}(\theta), \pi_{1,i}(s_t, a_t)), 1 - e, 1 + e\}]$$

where $\phi_{1,i}(\theta) = \frac{\pi_{1,i}(a_t|s_t, \tau_{t,i})}{\pi_{1,\theta}(a_t|s_t, \tau_{t,i})}$ is the policy probability ratio and $e \in [0, 1)$ is a clipping parameter. For simplicity, we omit the parameters $\theta$, $\omega$ and write $\hat{Q}_1, \hat{Q}_i$ for the rest of the paper.

### 2.2.2 Centralized Training Decentralized Execution (CTDE)

For many problems, training takes place in a laboratory or in a simulated environment, where global information is available [28]. Therefore, state-of-the-art MARL algorithms approximate value functions $\hat{Q}_i$, which condition on global states $s_t$ and joint actions $a_t$, and use them as critics in Eq. 2 or 3 [18, 51]. While the value functions $\hat{Q}_i$ are only required during training, the learned policies $\pi_i$ only condition on local histories $\tau_{t,i}$ thus being independently executable. Unlike MAPF, these policies can generalize over a variety of scenarios and thus ideally do not need any centralized retraining or replanning for changes or new maps [30].

$\hat{Q}_i$ can be approximated separately for each agent $i$ while integrating global information, which is done in actor-critic MARL algorithms like MAPPO or MADDPG [18, 51]. However, this approach lacks a multi-agent credit assignment mechanism for agent teams, where all agents optimize the same objective $Q_{tot}$ (Eq. 1).

Alternatively, a common value function $\hat{Q}(\tau_t, a_t) \approx Q_{tot}(s_t, a_t)$ can be learned, which is factorized into $\hat{Q}_1(\tau_{1,N})$ as local utility functions by using a factorization operator $\Psi$ [26, 27]:

$$\hat{Q}((\tau_t, a_t) = \Psi(\hat{Q}_1(\tau_{1,N}), ..., \hat{Q}_N(\tau_{N,N}, a_{1,N}))$$

In practice, $\Psi$ is realized with deep neural networks, such that $(\hat{Q}_1, ..., \hat{Q}_N)$ can be learned end-to-end via backpropagation by minimizing the mean squared temporal difference (TD) error [28, 42]. A factorization operator $\Psi$ is decentralized when satisfying the IGM (Individual-Global-Max) such that [37]:

$$\arg\max_{\pi_i}\hat{Q}(\tau_t, a_t) = \begin{cases} \arg\max_{a_{1,i}}\hat{Q}_1(\tau_{1,i}, a_{1,i}) \\ \vdots \\ \arg\max_{a_{N,i}}\hat{Q}_N(\tau_{N,i}, a_{N,i}) \end{cases}$$

There exists a variety of factorization operators $\Psi$, which satisfy Eq. 5 using monotonicity constraints like QMIX [28] or nonlinear transformation like QPLEX or QTRAN [37, 47].

### 2.3 Curriculum Learning

**Curriculum learning** is a machine learning paradigm, inspired by human learning, to master complex tasks through stepwise solving of easier (sub-)tasks, which are sorted by difficulty [3, 38]. The difficulty can depend on various aspects like the complexity of data samples, the objective function, or the learned model [10, 22]. Curriculum learning has been applied to reinforcement learning (RL) to solve hard exploration problems with sparse rewards or dynamic constraints [21]. The methods are typically based on self-play [35, 45], task graphs with traversal mechanisms [33], or automatic generation of tasks [7, 11].

A key challenge of curriculum learning is to find or generate a suitable sequence of tasks that are neither too easy nor too difficult for the learner to ensure steady and robust progress [10, 11, 33].

We focus on reverse curriculum learning, where we assume explicit goal states as in the MAPF problem (Section 2.1). The curriculum consists of a sequence of tasks, where the (expected) distance between agent and goal gradually increases [2, 10].

### 3 RELATED WORK

**Reverse Curriculum Generation.** Many works on RL-based motion control assume a single goal state, which is easy to specify [1, 19]. A reverse curriculum is defined, where the start state is initialized within a short distance to the goal state. The distance gradually increases with the convergence or performance improvement of the agent [2]. Our work is based on reverse curriculum generation, focusing on multi-agent path finding (MAPF). In MAPF, there are several goal states that are unique per instance $i$ (which can vary for the same map $G$ though). We propose a simple confidence-based approach to adapt the curriculum by considering the uncertainty of performance estimates.

**Curriculum Learning in MARL.** Curriculum learning has been widely used in single-agent or two-player zero-sum games to improve convergence speed or performance. While many of these approaches are based on foundations of multi-agent learning [7, 8, 41], there exist methods particularly designed for MARL based on self-play, agent skills, and population-based training [14, 46, 50]. These methods are typically very complex due to heavily engineered architectures and mechanisms thus requiring a significant amount of compute. As our work focuses on simple and efficient MARL approaches to MAPF, we do not consider such resource and tuning-intensive training regimes.

**MARL for MAPF.** MAPF and MARL are very active research areas with remarkable progress in recent years [14, 16, 46]. Both fields have been mainly studied independently of each other, with only a few works attempting to connect them. The first work in this direction is PRIMAL. PRIMAL and its successor approaches are heavily engineered and very complex, using extremely large neural networks, extensively shaped rewards, and centralized MARL solvers for imitation learning to address the challenging aspects of MAPF, i.e., sparse rewards, dynamic constraints, and coordination [6, 30, 48]. Despite their effectiveness, these approaches are very expensive to use due to significant effort on fine-tuning and enormous computational data requirements. Some recent works proposed manually designed curricula to enhance PRIMAL but still rely on very complex architectures and reward functions [23, 52]. Besides applying MARL to MAPF, there have been other attempts to combine MAPF with machine learning techniques to guide or select centralized search algorithms [12, 13, 15, 25]. Our goal is to provide a suitable foundation to bring the areas of MARL and MAPF closer together with significantly lower costs. Therefore, we propose a simple reverse curriculum scheme to ease applicability and enable faster progress in this direction.
4 MAP AS A STOCHASTIC GAME

To apply MARL techniques to MAP in a general way, we first need to formulate the MAP problem defined in Section 2.1 as a stochastic game \( M \) according to Section 2.2. Similar to prior work [6, 30, 39, 48], we focus on discrete gridworlds but try to keep our formulation general. An adaptation of our methods to arbitrary graphs, e.g., using graph neural networks, is left for future work.

In both settings, the set of agents \( D \) is equivalent. Given a map \( G = (\mathcal{V}, \mathcal{E}) \), the state space \( \mathcal{S} \) is defined by the joint locations of all agents \( s_t = (s_{t,1}, ..., s_{t,N}) \in \mathcal{S} \subset \mathcal{V}^N \), where each location in \( s_t \) is unique such that \( s_{t,i} \neq s_{t,j} \) for each agent pair \( i, j \in D \) with \( i \neq j \). The individual action space \( \mathcal{A}_i \) of each agent \( i \) is defined by the maximum degree of map \( G \) plus a wait action. In 4-neighborhood gridworlds as displayed in Fig. 2, each agent would be able to wait or move in all cardinal directions. The state transitions are deterministic, where a valid move action will change the location of the corresponding agent. Attempts to move over non-existent edges or cause vertex or edge conflicts, i.e., collisions, are automatically treated as wait action. The individual reward \( r_t, i \) is defined by +1 if agent \( i \) reaches its goal \( s_{\text{goal},i} \), 0 when agent \( i \) is staying at its goal location \( v_{\text{goal},i} \), and -1 otherwise. Each agent \( i \) can partially observe the state \( s_t \) through a local neighborhood around its location \( s_{t,i} \). For gridworlds, we assume a \( 7 \times 7 \) field of view (FOV) similar to PRIMAL, which is illustrated in Fig. 2 [30]. The features of an observation \( z_{t,i} \), i.e., obstacles, other agents and their goals, and the direction and goal location of agent \( i \), are encoded as a multi-channel image. The direction channel encodes the Manhattan distance to the goal \( v_{\text{goal},i} \) and indicates the direction to it.

![Environment Map](image)

**Figure 2:** Example for an individual observation of the red agent in a gridworld domain. Agents are represented as colored circles, their goals as similarly-colored squares, and obstacles as black squares. Each agent \( i \) has a limited field of view (FOV) of the environment map, which is centered around its location encoded by five channels: locations of obstacles, location of other agents’ goals, locations of nearby agents, and location of the goal \( s_{\text{goal},i} \) if within the FOV, and the Manhattan distance and direction of agent \( i \) to its goal.

When the discount factor is \( y = 1 \), the negated return \(-R_{t,i}\) of each agent \( i \) is equivalent to its travel distance \( l(p_t) \) from time step \( t \) if \( v_{\text{goal},i} \) was reached, and horizon \( T \) otherwise. Therefore, maximizing \( Q_{tot} = \sum_{i \in D} Q_i = -E[l(p)] \) in MARL (Section 2.2 and Eq. 1) would be equivalent to minimizing the expected flowtime in MAP w.r.t. any instance \( l \) on map \( G \) (Section 2.1).

Since any time step is penalized with -1 anyway (unless an agent reaches or occupies its goal), all agents are discouraged from unnecessary delays, which includes collision attempts. Unlike prior work, we do not need additional penalties for particular situations like collisions, blocking, or waiting which could fundamentally change the actual objective and lead to unintended side-effects [36].

Thus, our problem formulation is simpler and more general, which allows us to solve it in a black-box manner that is more intuitive for standard MARL methods [28, 47, 51]. However, the simplicity of our formulation notably increases difficulty since the reward is sparse in contrast to PRIMAL and related approaches.

5 CONFIDENCE-BASED CURRICULUM

5.1 Training Scheme

We assume a separate training phase to learn coordinated local policies \( \hat{\pi}_i \) for decentralized execution. We train \( \hat{\pi}_i \) via policy gradient methods according to Eq. 2 or 3. The critics \( \hat{Q}_i \) are trained via CTDE methods to exploit global information during training using either independent learning like MAPPO or value factorization like QMIX or QPLEX as illustrated in Fig. 3 [28, 47, 51]. Since the value factorization based actor-critic scheme has been used in a variety of prior work [24, 40], we do not claim novelty here, but propose it as a basic approach to train cooperative policies via credit assignment mechanisms [28, 37, 47].

![Common actor-critic scheme](image)

**Figure 3:** Common actor-critic scheme as used in various prior work on cooperative MARL [24, 40]. A separate critic is trained for each actor using some centralized factorization operator \( \Psi \) like QMIX or QPLEX [27].

To address the coordination problem as mentioned in the introduction, we suggest to optimize individual utilities \( \hat{Q}_i \approx \hat{Q}_i^\Psi \) under consideration of the utilitarian metric \( Q_{tot} \) in Eq. 1. For that, the individual utilities \( \hat{Q}_i \) can be learned end-to-end through a factorization operator \( \Psi \) like QMIX or QPLEX to consider multi-agent credit assignment from a cooperative perspective [26–28, 29, 47]. The factorization operator \( \Psi \) approximates the expected sum of individual returns by minimizing the factorization loss \( \mathcal{L}_{\Psi} \) defined by:

\[
\mathcal{L}_{\Psi} = \mathbb{E} \left[ \Psi(\hat{Q}_1(\tau_{1,i}, a_{1,i}), ..., \hat{Q}_N(\tau_{N,i}, a_{N,i})) - \sum_{i \in D} R_{t,i} \right]^2
\]

The local policies \( \hat{\pi}_i \) are then trained according to Eq. 2 or 3 using counterfactual advantages \( \hat{A}_i \) defined by [40]:

\[
\hat{A}_i(\tau_{t,i}, a_{t,i}) = R_{t,i} - \sum_{a' \in \mathcal{A}_i} \hat{\pi}_i(a' | \tau_{t,i}) \hat{Q}_i(\tau_{t,i}, a')
\]


1561
where the return \( R_{t,i} \) represents the negative travel distance \(-l(p_i)\) of agent \( i \) from time step \( t \) according to Section 4. The advantage \( A_t \) incentivizes the optimization of travel distances under implicit consideration of other agents through \( Q_t \) and \( \Psi \) w.r.t. Eq. 1 and 6.

### 5.2 Reverse Curriculum Scheme

The training scheme described above represents a general approach to learn coordinated policies \( \hat{\pi}_t \) [24, 40]. However, sparse rewards and dynamic constraints in our MAPF setting (Section 4) pose particular challenges that require a suitable curriculum to learning meaningful policies [2, 10]. Unlike prior work that relied on complex reward functions with various penalties and expensive expert data for imitation learning, we propose Confidence-based Auto-Curriculum for Team Update Stability (CACTUS) to enhance the training scheme of Section 5.1 without significant costs.

At the beginning of every episode \( m \), each agent \( i \) starts at a random location \( v_{start,i} \in \mathcal{V} \) and needs to navigate to an assigned goal location \( v_{goal,i} \in \mathcal{V} \) which is randomly placed within an allocation radius \( R_{alloc} \). The curriculum update scheme is illustrated in Fig. 1. CACTUS characterizes the potential difficulty of generated instances \( I \) as larger allocation radii may require the agents to move and explore over longer distances to locate their respective goals. Thus, our reverse curriculum scheme starts with a small allocation radius of \( R_{alloc} = 1 \) and gradually increments \( R_{alloc} \) with improving performance, which is measured by the completion rate \( \rho_m^{complete} = \frac{N_{m,goal}^{complete}}{N} \), where \( N_{m,goal}^{complete} = |\{ i \in \mathcal{D}[v_{start,i} = v_{goal,i}] \}| \) is the number of agents that successfully reached their respective goals in episode \( m \).

CACTUS uses a statistical approach to decide whether to increment \( R_{alloc} \) or not. After each epoch of \( K \) episodes \( m \), we measure the average completion rate \( \mu = \frac{1}{K} \sum_{m=1}^{K} \rho_m^{complete} \) and its standard deviation \( \sigma = \sqrt{\frac{1}{K-1} \sum_{m=1}^{K} (\rho_m^{complete} - \mu)^2} \).

Assuming that the completion rates \( \rho_m^{complete} \) follow a normal distribution, CACTUS increments \( R_{alloc} \) by 1, if \( \mu - \sigma > U \), where \( U \in (0, 1) \) is the curriculum decision threshold and \( \eta > 0 \) is a deviation factor to specify the confidence level. For example, if \( U = 75\% \) and \( \eta = 2 \) then \( R_{alloc} \) would be incremented only if all agents achieve an average completion rate over 75% with a confidence level of around 97%. Note that we only regard one-tailed tests here, where we assume no upper limit to the average completion rate of agents (except for \( \mu = 100\% \), where \( \sigma \) would be zero). The curriculum update scheme is illustrated in Fig. 1.

The complete formulation of CACTUS is given in Algorithm 1. \( G \) is a set of training maps or a map generator, \( \mathcal{D} : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R} \) is a vertex distance function, \( U \) is the curriculum decision threshold, and \( \eta \) is the deviation factor.

### 5.3 Conceptual Discussion

CACTUS represents a simple reverse curriculum scheme inspired by prior work [2, 10]. Our work focuses on the MAPF problem, where we have multiple agents with different start and goal locations that can vary per instance \( I \). To ensure generalization over a variety of MAPF instances and maps, our scheme adjusts the random allocation of goals around the random start locations.

In contrast to prior work [23, 52], CACTUS does not separate learning of different skills like navigation and collision avoidance. As illustrated in Fig. 1, all agents first need to focus on reaching their respective goals, which are allocated in close proximity within \( R_{alloc} \). With increasing \( R_{alloc} \), the allocation areas of different agents may overlap, automatically causing agents to interact with each other thus increasing coordination pressure. \( R_{alloc} \) is only incremented when agents are able to coordinate and reach their goals with sufficiently high confidence, which is checked with the hyperparameters \( U \) and \( \eta \). Thus, CACTUS offers an adaptive approach to solving MAPF problems via MARL without requiring explicit separation of agent skills [23, 46, 52], extensive engineering of rewards, or expensive acquisition of expert data [6, 30, 48].

Since the goals are randomly initialized within allocation radius \( R_{alloc} \) around the agents’ start locations, they can still be allocated in closer proximity to the agents which alleviates catastrophic forgetting of easier tasks that the agents have mastered before.

---

\( \text{Algorithm 1 Confidence-Based Curriculum Learning for MAPF} \)

1. procedure **DRIVE**(\( G, \mathcal{D}, \mathcal{U}, \eta \))
2. Initialize parameters of \( \hat{\pi}_t, Q_t \) for each agent \( i \in \mathcal{D} \) and \( \Psi \)
3. Set \( R_{alloc} = 1 \)
4. for epoch \( x \leftarrow 1, X \) do
   5. for episode \( m \leftarrow 1, E \) do
      6. Randomly select or generate map \( G \) from \( \mathcal{G} \)
      7. Sample \( s_0 \) → Set start locations \( v_{start,i} \)
      8. for agent \( i \in \mathcal{D} \) do
         9. Set \( t_0,i \) based on \( \Omega(t_0) \)
         10. Set \( V_{goal,i} \leftarrow \{ v \in \mathcal{V} | \mathcal{DIST}(v, v_{start,i}) \leq R_{alloc} \} \)
         11. Randomly select goal location \( v_{goal,i} \) from \( V_{goal,i} \)
         12. Goal locations must be unique, i.e, \( v_{goal,i} \neq v_{goal,j} \) if \( i \neq j \)
      13. for time step \( t \leftarrow 0, T - 1 \) do
         14. for agent \( i \in \mathcal{D} \) do
            15. \( a_t \leftarrow (a_{t,1}, ..., a_{t,N}) \)
            16. Execute joint action \( a_t \)
            17. \( s_{t+1} \leftarrow T(s_t, a_t) \)
            18. \( z_{t+1} \leftarrow Q(s_{t+1}) \)
            19. \( e_t \leftarrow (\tau_t, a_t, r_t, z_{t+1}) \)
            20. Store experience sample \( e_t \)
            21. \( \tau_{t+1} \leftarrow (\tau_t, a_t, z_{t+1}) \)
            22. \( \rho_m^{complete} \leftarrow |\{ i \in \mathcal{D}[v_{start,i} = v_{goal,i}] \}|/N \)
            23. Train \( \Psi \) and \( \hat{\pi}_t, Q_t \) for each agent \( i \in \mathcal{D} \) with all \( e_t \)
            24. Calculate \( \mu \) and \( \sigma \) with all \( \rho_m^{complete} \)
            25. if \( \mu - \eta \sigma \geq U \) then → Curriculum update decision
               26. \( R_{alloc} \leftarrow R_{alloc} + 1 \)
      27. return \( (\hat{\pi}_1, ..., \hat{\pi}_N) \)
6 EXPERIMENTAL SETUP

6.1 Maps and Instances

The training maps are randomly generated according to [30] and have different shapes $K \times K$ defined by map size $K \in \{10, 40, 80\}$. The obstacle density $\delta \in \{0, 0.1, 0.2, 0.3\}$ defines the fraction of non-occupiable locations in the maps. All agents start at random locations with randomly assigned goals according to an allocation radius $R_{alloc}$. If $R_{alloc} = \infty$, then the goals can be placed anywhere on the training map.

The test maps are provided by [30]. For each map configuration of size $K$ and obstacle density $\delta$, there are 100 pre-generated test instances $I$ with fixed start and goal locations for all agents to ensure a fair comparison between different MARL approaches.

We always set $g = 1$ as suggested in Section 4.

6.2 Algorithms and Training

We implement PPO$^2$ for policy learning according to Eq. 3 and QMIX, QPLEX, and MAPPO for critic learning. We implement a purely RL-based version of PRIMAL using the same reward function as defined in [30]. In addition, we employ a naive baseline, called No Curriculum, only consisting of PPO and QMIX$^2$ without any shaped reward. PRIMAL and No Curriculum are trained with $R_{alloc} = \infty$ (Section 6.1).

We train all algorithms on different training maps as explained in Section 6.1 with $N = 8$ agents for 5000 epochs consisting of $E = 32$ episodes. The maps of size $K = 10$ are sampled twice as often as the other map sizes as proposed in [30]. Each episode terminates after all agents reach their goal or after $T = 256$ time steps. All algorithms use parameter sharing, i.e., where all agents use the same policy and individual critic network $\pi_i$ and $\hat{Q}_i$ respectively.

We denote CACTUS (X) as CACTUS using MARL algorithm X for critic learning. Unless stated otherwise, CACTUS always uses PPO for policy and X-QMIX for critic learning.

In addition, we run CBSH as a slow but optimal MAPF solver with a runtime limit of 5 minutes [9] and MAPF-LNS [16] as a fast anytime MAPF solver with a runtime limit of 1 minute.

6.3 Neural Networks and Hyperparameters

For CACTUS and No Curriculum, we use deep neural networks to implement $\pi_i$ and $\hat{Q}_i$ for each agent $i$ and factorization operator $\Psi$ for QMIX and QPLEX. The neural networks are updated after every $E = 32$ episodes using ADAM with a learning rate of 0.001.

Since all regarded maps are gridworlds, the observations are encoded as multi-channel image as illustrated in Fig. 2. We implement all neural networks as multilayer perceptron (MLP) and flatten the multi-channel images before feeding them into the networks. $\pi_i$ and $\hat{Q}_i$ have two hidden layers of 64 units with ELU activation. The output of $\pi_i$ has $|A_i|$ units with softmax activation. The output of $\hat{Q}_i$ has $|A_i|$ linear units. The hypernetworks of QMIX as well as the critic of MAPPO have two hidden layers of 128 units with ELU activation and one or $|A_i|$ linear output units respectively. For PRIMAL, we use the same architecture as proposed in [30].

Unless stated otherwise, CACTUS always uses a threshold of $U = 75\%$ and a deviation factor of $\eta = 2$, which corresponds to a confidence level of about 97% in one-tailed tests.

Computing Infrastructure. All training and test runs are performed on a x86_64 GNU/Linux (Ubuntu 18.04.5 LTS) machine with i7-8700 @ 3.2GHz CPU (8 cores) and 64 GB RAM. Due to the simplicity of CACTUS, we do not need any GPU or distributed HPC infrastructure in contrast to [6, 30, 48].

7 RESULTS

For each experiment, all respective algorithms are run 10 times to report the average progress and the 95% confidence interval. We evaluate the training progress and generalization of trained policies with the pre-generated test instances $I$ explained in Section 6.1.

7.1 Simplicity of CACTUS

To demonstrate the simplicity of CACTUS, we first quantify the training time, the training data w.r.t. the number of episodes, the number of trainable parameters, and the reward complexity and compare them with the original PRIMAL as specified in [30].

An overview is given in Table 1. In almost all aspects, CACTUS only requires 5% or less of the effort of the original PRIMAL therefore being clearly the simpler and more efficient MARL approach to MAPF. Unlike PRIMAL, CACTUS is only run on CPU while still requiring significantly less training time. Furthermore, CACTUS does not depend on any expert data, i.e., recommendations of a centralized MAPF solver, which saves a significant amount of compute. While the reward function of PRIMAL requires four penalties for very specific situations, CACTUS is trained with a very simple reward function that penalizes any time step unless the goal is reached without considering any specific case (Section 4).

Table 1: Comparison of the original PRIMAL and CACTUS w.r.t. various numbers in our experiments. The last column provides the amount of effort relative to the original PRIMAL. The numbers of the original PRIMAL are from [30]. Unlike [30], we do not use any GPU or expert data for training.

<table>
<thead>
<tr>
<th>Training Time</th>
<th>PRIMAL (original [30])</th>
<th>CACTUS</th>
<th>Rel. to PRIMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td># Training Episodes</td>
<td>$\approx 20$ days</td>
<td>$\approx 1$ day</td>
<td>$\approx 5%$</td>
</tr>
<tr>
<td># Parameters</td>
<td>$\approx 3.8$ million</td>
<td>160,000</td>
<td>$\approx 4.2%$</td>
</tr>
<tr>
<td># Reward Penalties</td>
<td>$\approx 13$ million</td>
<td>579,979</td>
<td>$\approx 4.5%$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>25%</td>
</tr>
</tbody>
</table>

Fig. 4 compares the number of trainable parameters and neural network architectures of PRIMAL and CACTUS. In CACTUS, the critic with the mixing network has the majority of trainable parameters, which are only required during training. The actor size is negligible in CACTUS, which enables significantly faster inference than PRIMAL. The network architecture of CACTUS is also much simpler than PRIMAL since it is only based on MLPs thus does not depend on specialized hardware or significant computational effort for fine-tuning and training.
We evaluate the effect of CACTUS using QMIX, QPLEX, and MAPPO w.r.t. training epochs (left) and training time (right). The performance is evaluated on all pre-generated test instances. CACTUS outperforms PRIMAL always achieves an average completion rate of 100% in all test instances. All CACTUS variants clearly outperform PRIMAL.

Finally, we evaluate the generalization capabilities of policies trained with CACTUS using QMIX, QPLEX, and MAPPO as well as PRIMAL, and No Curriculum. All policies are trained for 5000 epochs consisting of 32 episodes before being evaluated on all test instances with map size $K \in \{10, 40, 80\}$, obstacle density $\delta \in \{0.0, 0.1, 0.2, 0.3\}$, and $N = 8$ agents. The right plot shows CACTUS variants with $U = 0.25$. Shaded areas show the 95% confidence interval.

### 7.4 Generalization

The generalization results w.r.t. different numbers of agents $N$ are shown in Fig. 7. CACTUS (QMIX) generalizes best compared to all other MARL approaches. In test instances with low obstacle density, CACTUS (QMIX) always achieves an average completion rate over 70% when scaling up to $N = 64$ agents. If $\delta = 0$, then CACTUS (QMIX) is able to complete at least 50% of all agents when scaling up to $N = 128$. CACTUS (QPLEX) always outperforms PRIMAL on average except in instances with low obstacle density and map size $K = 40$, where the number of agents exceeds 32. PRIMAL is always outperformed by CACTUS (QMIX) but consistently outperforms CACTUS (MAPPO) and No Curriculum. All approaches perform poorly when the agent number is $N \geq 256$ or $\delta = 0.2$.

Compared to the centralized MAPF solvers CBSH, CACTUS (QMIX) can only outperform CBSH, when the obstacle density $\delta$ is low. MAPF-LNS is the best performing approach, always achieving a completion rate of 100% in all test instances.

### 7.5 Limitation in Structured Maps

In addition, we tested CACTUS and the learning baselines in structured maps with rooms and narrow corridors such as mazes. While
variants as well as MARL and MAPF baselines to all test unstructured maps in Section 7.2, all approaches perform significantly worse with none of them reaching an average completion rate above 25%. CACTUS (QMIX) consistently outperforms all other approaches, which generally fail to complete more than 10% of all agent tasks.

8 DISCUSSION

In this paper, we presented CACTUS as a lightweight MARL approach to MAPF. CACTUS defines a simple reverse curriculum scheme, where the goal of each agent is randomly placed within an allocation radius around the agent’s start location. The allocation radius increases gradually as all agents improve, which is assessed by a confidence-based measure.

Our results confirm the necessity of adequate curricula, as standard MARL methods without any curriculum would likely fail to learn any meaningful policy in our MAPF setting (Section 4). This is due to the sparse reward and the dynamic constraints of the problem formulation. The shaped reward function of PRIMAL is helpful in improving performance over standard MARL. Due to the many specifically defined penalties, PRIMAL policies are rather conservative, which leads to a performance plateau that cannot be overcome without imitation learning on suitable expert data. However, CACTUS with QMIX or QPLEX clearly outperforms PRIMAL with 95% less training time and learnable parameters without any additional reward shaping or expert data. CACTUS with MAPPO performs poorly, which indicates the importance of adequate credit assignment mechanisms, e.g., value factorization, to address the coordination problem in MAPF.

CACTUS is robust w.r.t. the choice of hyperparameters as any configuration with decision threshold $U \geq 25\%$ and deviation factor $\eta \geq 1$ outperforms PRIMAL as shown in Fig. 6. However, the decision threshold $U$ should be sufficiently high (at least 50%) to ensure an adequate difficulty level for the agents. Choosing a confidence level that is too high, e.g., using $\eta > 3$, could result in slow curriculum updates, where agents may overfit on easy tasks.

Despite the relatively restrictive time and data budget compared to the original PRIMAL, CACTUS generalizes quite well with value factorization over different numbers of agents $N$ and map sizes $K$. CACTUS with QMIX scales up to instances with 8 to 16 times more agents than used during training, in contrast to standard MARL, which generally fails to learn any meaningful policy in the MAPF problem. While generalizing better than the alternative approaches, CACTUS still has some limitations regarding high obstacle density, large map sizes and structured maps with separate rooms and narrow corridors, where there is still potential for improvement. Despite the ability of efficient decentralized decision-making, CACTUS is not competitive to centralized MAPF solvers due to its limited scalability.

Nevertheless, CACTUS clearly demonstrates how simple and well-defined curricula can enhance MARL techniques for MAPF without relying on extremely large neural networks, extensively shaped reward functions, or centralized MAPF solvers for imitation learning. Therefore, we hope to provide a suitable foundation to enable faster progress in connecting the areas of MARL and MAPF with significantly lower costs.

ACKNOWLEDGMENTS

The research at the University of Southern California was supported by the National Science Foundation (NSF) under grant numbers 1817189, 1837779, 1935712, 2121028, 2112533, and 2321786 as well as a gift from Amazon Robotics. The research at the Georgia Institute of Technology was supported by NSF under grant number 2112533. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring organizations, agencies, or the U.S. government.