Optimal Task Generalisation in Cooperative Multi-Agent Reinforcement Learning

Simon Rosen, Abdel Mfougouon Njupoun, Geraud Nangue Tasse, Steven James & Benjamin Rosman
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
{simon.rosen, ..., geraud.nanguetasse1, steven.james, benjamin.rosman1}@wits.ac.za

NOTE: Pre-camera ready version of paper

Abstract

While task generalisation is widely studied in the context of single-agent reinforcement learning (RL), little research exists in the context of multi-agent RL. The research that does exist usually considers task generalisation implicitly as a part of the environment, and when it is considered explicitly there are no theoretical guarantees. We propose Goal-Oriented Learning for Multi-Task Multi-Agent RL (GOLeMM), a method that achieves provably optimal task generalisation that, to the best of our knowledge, has not been achieved before in MARL. After learning an optimal goal-oriented value function for a single arbitrary task, our method can zero-shot infer the optimal policy for any other task in the distribution given only knowledge of the terminal rewards for each agent for the new task and learnt task. Empirically we show that our method is able to generalise over a full task distribution, while representative baselines are only able to learn a small subset of the task distribution.

1 Introduction

Cooperative multi-agent problems are ubiquitous in real-world applications, such as automated warehouses (Wurman et al., 2008), and train routing (Mohanty et al., 2020). Often, a multi-agent problem can be decomposed into a distribution of tasks, where all tasks share common properties, and solving the problem requires generalising over this task distribution. For example, consider an automated warehouse with multiple autonomous robots and shelves. Each robot is assigned a shelf (or set of shelves) that it must navigate to in the fewest steps. Here there exists a task distribution, where each task is defined by the shelves assigned to each robot, and the common properties across tasks include the warehouse layout and the robot dynamics.

An increasingly prevalent approach to tackling this class of problems is cooperative multi-agent reinforcement learning (MARL) (Zhang et al., 2021), where multiple agents learn via trial and error to accomplish a task while coordinating their actions with each other in the process. While the combination of MARL and deep learning has recently had great success in tackling challenging multi-agent problems (Vinyals et al., 2019; Berner et al., 2019) these methods have been severely hampered by poor sample efficiency. Training in these cases requires monumental amounts of compute not accessible to much of the scientific community.

While there is considerable research in generalising over a task distribution in the context of single-agent RL (Schaul et al., 2015; Yang et al., 2020; Tasse et al., 2021), there exists little research in the multi-agent setting with task generalisation usually only occurring implicitly as a property of the environment and achieved through the use of deep learning. Although Chen et al. (2021) explicitly consider multi-agent task generalisation in the context of agent scaling, there exists no previous literature that explicitly considers provable task generalisation in a multi-agent setting, to the best of our knowledge.

In this work, we propose *Goal Oriented Learning for Multi-Task Multi-Agent Reinforcement Learning* (GOLeMM), an approach that provably generalises over a task distribution after learning a world value function (WVF) (Nangue Tasse et al., 2020), a type of goal-oriented general value function, for a single arbitrary task. We consider a distribution of goal-oriented tasks where, for each task and agent, there exists a set of goal states that the agent wishes to reach and these are constant across the task distribution. However, each task is uniquely defined by the rewards associated with these goals. We prove that the WVF for a particular task encodes how to optimally reach every goal for every agent while considering coordination between agents. We then prove that this knowledge can be transferred to another task allowing for new tasks to be solved without further learning—provided knowledge of the task-specific goal rewards. As a result, our approach achieves optimal task generalisation after learning a single arbitrary task. Additionally, we leverage symmetries present in the WVF to improve the sample efficiency of learning that single task. Since our method only requires learning a single task before it can generalise over the task distribution, it can generalise in substantially fewer steps than representative baselines which must learn each task independently. We show this improved sample efficiency of task generalisation empirically in a domain inspired by an automated warehouse.

2 Background

Collaborative multi-agent tasks are often modelled as collaborative multi-agent Markov decision processes (CM-MDPs) (Guestrin, 2003; Kok & Vlassis, 2006), denoted by $M = (I, \{\mathcal{S}^i\}_{i\in I}, \{\mathcal{A}^i\}_{i\in I}, \mathcal{P}, \{R^i\}_{i\in I})$, where: (i) $I = \{1, ..., n\}$ denotes the set of $n > 1$ agents; (ii) $\mathcal{S}^i$ denotes the specific state space of agent i, with the (joint) state space of the environment defined as $\mathcal{S} \subseteq \mathcal{S}^1 \times ... \times \mathcal{S}^n$; (iii) $\mathcal{A}^i$ denotes the action space of agent i with the joint action space defined as $\mathcal{A} := \mathcal{A}^1 \times ... \times \mathcal{A}^n$, (iv) $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ denotes the deterministic transition dynamics; (v) $R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [R_{min}, R_{max}]$, with $R_{min} < 0 \leq R_{max}$, is the bounded agent reward function that determines the immediate reward received by agent i after the joint action $\mathbf{a} \in \mathcal{A}$ is taken in joint state $\mathbf{s} \in \mathcal{S}$ and the agents transition to joint state $\mathbf{s}' \in \mathcal{S}$. The team (global) reward is then defined by $R(\mathbf{s}, \mathbf{a}, \mathbf{s}') := \sum_{i\in I} R^i(\mathbf{s}, \mathbf{a}, \mathbf{s}')$.

We consider goal-oriented tasks and a special case of episodic tasks where termination occurs at an agent level upon reaching a goal (absorbing) state. For each agent $i \in I$, there exists a set of goals $\mathcal{G}^i \subseteq \mathcal{S}^i$ such that if an agent i experiences a transition to a goal state $g^i \in \mathcal{G}^i$, then it will terminate and receive an associated terminal reward. On subsequent steps, it will receive rewards of zero and it will not be able to change state. We use the notation $\langle s^i, s^{-i}\rangle$ to represent a joint state $\mathbf{s}$ where s^{-i} denotes all elements of $\mathbf{s}$ excluding s^i, $\langle s^i, s^{-i}\rangle$ denotes the concatenation of s^i and s^{-i}, and $\mathcal{S}^{-i} = \{\times \mathcal{S}^j\}_{j\in I\setminus\{i\}}$ denotes the cross product of all agent state spaces except for agent i. Then, for a terminated agent $i \in I$ where $s^i \in \mathcal{G}^i$, we have $\mathcal{P}(\langle s^i, s^{-i}\rangle, \mathbf{a}) = \langle s^i, s^{-i\prime}\rangle$ and $R^i(\langle s^i, s^{-i}\rangle, \mathbf{a}, \langle s^i, s^{-i\prime}\rangle) = 0$ for all $(s^{-i}, s^{-i\prime}) \in \mathcal{S}^{-i} \times \mathcal{S}^{-i}$ and for all $\mathbf{a} \in \mathcal{A}$. Finally, we will use T^i to denote the first timestep upon which the next state of agent i is a goal—that is $s^i_{T^i} \neq s^i_{T^i+1} \in \mathcal{G}^i$. We assume $T^i < 0$ if agent i has already terminated, that is $s^i \in \mathcal{G}^i$.

The set of tasks $\mathcal{M}$ can then be defined such that tasks share the same agent state spaces, agent action spaces, dynamics and non-terminal rewards, which is described by a background CM-MDP $M_B = (I, \{\mathcal{S}^i\}_{i\in I}, \{\mathcal{A}^i\}_{i\in I}, \mathcal{P}, \{R^i_B\}_{i\in I})$, and each task $M \in \mathcal{M}$ is uniquely specified by a set of agent terminal reward functions $\{\hat{R}^i_M\}_{i\in I}$, where $\hat{R}^i_M$ specifies the reward achieved by agent i when experiencing a terminal transition to g^i in the task M. Formally, the task distribution $\mathcal{M}$, which we refer to as the *world*, is defined by

$$\mathcal{M}(M_B) := \{(I, \{\mathcal{S}^i\}_{i\in I}, \{\mathcal{A}^i\}_{i\in I}, \mathcal{P}, \{R^i_M\}_{i\in I}) | \forall i \in I,\ \forall(\mathbf{s}, \mathbf{a}, \mathbf{s}') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$$
$$R^i_M(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \begin{cases} \hat{R}^i_M(s^{i\prime}), \text{ if } s^i \neq s^{i\prime} \in \mathcal{G}^i \\ R^i_B(\mathbf{s}, \mathbf{a}, \mathbf{s}'), \text{otherwise} \end{cases} \quad \}$$

In this work, we assume that there exists a centralised controller (CC) that jointly specifies the actions of each agent. In the centralised setting, a CM-MDP is a special case of an MDP allowing for single-agent reinforcement learning algorithms to be used (Kok & Vlassis, 2006). As such the CC aims to learn an optimal *joint policy* $\pi^* : \mathcal{S} \mapsto Pr(\mathcal{A})$ that when followed maximises the expected sum of team rewards. Under a joint policy $\bar{\pi}$ the expected team return is defined by the value function $V^{\pi}(\mathbf{s}) := \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} R(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})]$.

Similarly, the action value function $Q^{\bar{\pi}}(\mathbf{s}, \mathbf{a}) := \mathbb{E}^{\pi}_{s'}[R(\mathbf{s}, \mathbf{a}, \mathbf{s}') + V^{\pi}(\mathbf{s})]$ defines the expected return from joint state $\mathbf{s}$ when executing joint action $\mathbf{a}$ and following π thereafter. The optimal Q-value function is given by $Q^*(\mathbf{s}, \mathbf{a}) = \max_{\pi} Q^{\pi}(\mathbf{s}, \mathbf{a})$ and the optimal joint policy can be retrieved by acting greedily with respect to Q^* at each joint state: $\pi(\mathbf{s}) \in \arg\max_{\mathbf{a} \in \mathcal{A}} Q^*(\mathbf{s}, \mathbf{a})$. Finally, since we are interested in goal-reaching tasks, we assume the shortest path setting where the optimal joint policy must be *proper*. A joint policy is proper if all agents are guaranteed to eventually reach a terminal agent state while following it. This is equivalent to the definition of proper policies in the single-agent setting (Van Niekerk et al., 2019). Similarly to Nangue Tasse et al. (2020), we assume that the value functions for improper joint policies are unbounded from below.

3 Task Generalisation via World Value Functions

3.1 Multi-Agent World Value Functions

In this section we extend world value functions (WVFs) (Nangue Tasse et al., 2020)[1] to the multi-agent setting and prove that these multi-agent WVFs encode how all agents can jointly reach all joint goals that are reachable under the dynamics of the environment. To achieve this, we first define extended agent reward functions which penalise each agent for reaching goals they did not intend to reach, and then we define the extended team reward function similarly to the original team reward function.

Definition 1. *The extended agent reward function $\bar{R}^i : \mathcal{S} \times \mathcal{G}^i \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ for agent i is defined by*

$$\bar{R}^i(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') := \begin{cases} \tilde{R}_{min} & \textit{if } s^i \neq s^{i'} \in \mathcal{G}^i \setminus \{g^i\}, \\ R^i(\mathbf{s}, \mathbf{a}, \mathbf{s}') & \textit{otherwise} \end{cases}$$

where $\tilde{R}_{min} \leq n((R_{min} - R_{max})D + R_{min})$, $D = \max_{\mathbf{s} \neq \mathbf{s}' \in \mathcal{S}} \max_{\pi \in \Pi_p} \mathbb{E}[T(\mathbf{s}'|\pi, \mathbf{s})]$, T is the number of timesteps required to reach $\mathbf{s}'$ from $\mathbf{s}$ under π and Π_p is the set of proper joint policies.

Definition 2. *The extended team reward function $\bar{R} : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is defined by $\bar{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') := \sum_{i \in I} \bar{R}^i(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')$.*

We wish to learn a goal-oriented joint policy that directs agents to assigned joint goals. We denote this joint policy as a *joint world policy* (JWP) $\bar{\pi} : \mathcal{S} \times \mathcal{G} \mapsto Pr(\mathcal{A})$. We then define the WVF $\bar{Q}^{\bar{\pi}}$ associated with $\bar{\pi}$ similarly to the single-agent definition of WVFs except defined using the team reward function. We note that in the special case where the CM-MDP consists of $n = 1$ agent, this definition of the WVF is analogous to that of Nangue Tasse et al. (2020).

Definition 3. *The world value function $\bar{Q}^{\bar{\pi}} : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \mapsto \mathbb{R}$ under a world joint policy $\bar{\pi}$ is defined by $\bar{Q}^{\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a}) := \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') + \bar{V}^{\bar{\pi}}(\mathbf{s}, \mathbf{g})]$ where the state world value function $\bar{V}^{\bar{\pi}} : \mathcal{S} \times \mathcal{G} \mapsto \mathbb{R}$ under $\bar{\pi}$ is defined as $\bar{V}^{\bar{\pi}}(\mathbf{s}, \mathbf{g}) := \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}(\mathbf{s}_t, \mathbf{g}, \mathbf{a}_t, \mathbf{s}_{t+1})]$.*

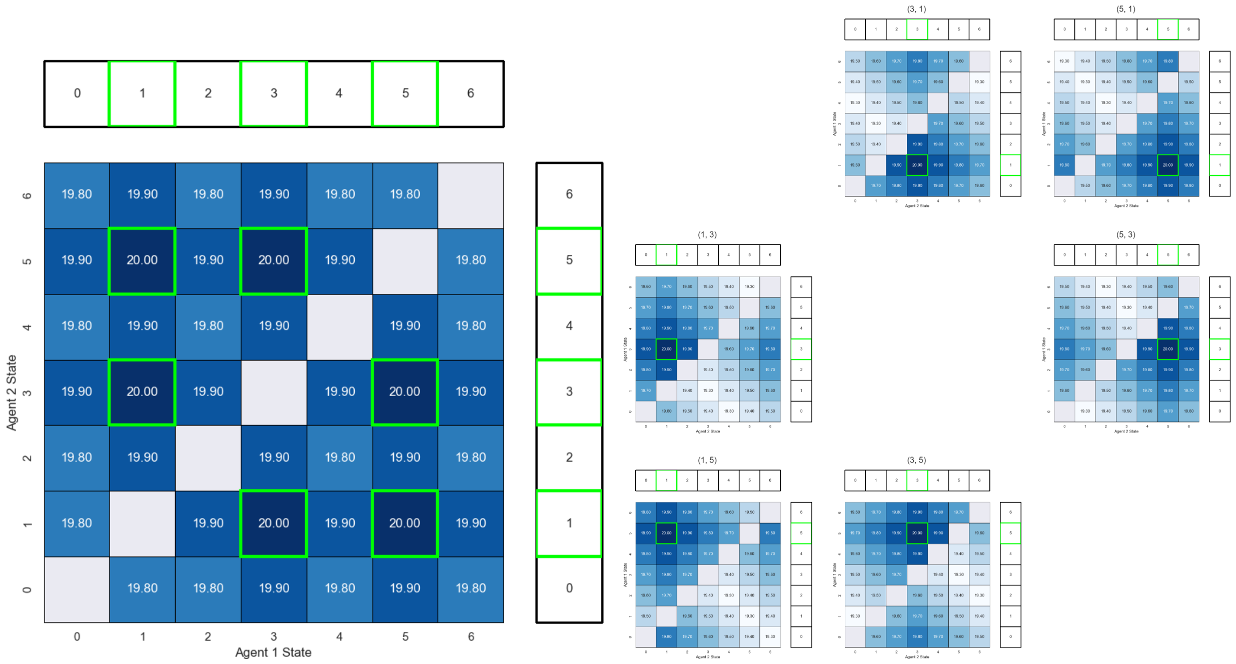
Since the WVF satisfies the Bellman equations the optimal JWP can be retrieved from the optimal WVF by maximising over joint actions: $\bar{\pi}^*_M(\mathbf{s}, \mathbf{g}) \in \arg\max_{\mathbf{a} \in \mathcal{A}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$. The optimal state world value function (SWVF) can be retrieved similarly: $\bar{V}^*_M(\mathbf{s}, \mathbf{g}) = \max_{\mathbf{a} \in \mathcal{A}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$. Additionally since we consider a CC, the optimal WVF can be learnt using any suitable single-agent RL algorithm with minor modification, which we describe in Subsection 3.4. However, the optimal WVF and associated optimal JWP do not immediately solve the given task since they define their own reward function. Thus, we must retrieve the optimal Q-value function from the optimal WVF. We do this by maximising over joint goals on the optimal WVF, which is possible due to the team reward function being retrievable from the extended team reward function—also by maximising over joint goals. These are given in the following theorem.

Theorem 1. *Let R_M and $\bar{R}_M$ be the team reward function and extended reward function for a task $M \in \mathcal{M}$. Let Q^*_M and $\bar{Q}^*_M$ be the corresponding optimal Q-value function and optimal world value function. For all $(\mathbf{s}, \mathbf{a}, \mathbf{s}') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have (i) $R_M(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \max_{\mathbf{g} \in \mathcal{G}} \bar{R}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}')$, and (ii) $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$* [2].

[1] Originally referred to as extended value functions by Nangue Tasse et al. (2020), but later generalised to world value functions (Tasse et al., 2022).

[2] Proofs are provided in the appendix

We can then retrieve the optimal joint policy from the optimal WVF directly with $\pi^*(\mathbf{s}) \in \arg\max_{\mathbf{a}\in\mathcal{A}}\{\max_{\mathbf{g}\in\mathcal{G}} \bar{Q}^*_M(\mathbf{s},\mathbf{g},\mathbf{a})\}$ to solve the current task. We can also retrieve the optimal value function from the optimal SWVF with $V^*_M(\mathbf{s}) = \max_{\mathbf{g}\in\mathcal{G}} \bar{V}^*_M(\mathbf{s},\mathbf{g})$. This is illustrated in Figure 1 for a two-agent ring world.



(a) Optimal value function V^*

(b) Optimal state world value function $\bar{V}^*$

Figure 1: Value functions in a two-agent ring world. (a) shows the heatmap of values associated with a two agent ring world. The ring world can be decomposed into the agent states corresponding to agent 1 and agent 2 which are shown on the top and right of the figure with the coordinates of the associated agent goals highlighted in green. An agent terminates upon taking a 'WAIT' action at the coordinate of a goal. (b) shows the heatmaps corresponding to the joint state values associated with each joint goal in the WVF. We can see that we can retrieve (a) from (b) with $V^*(\mathbf{s}) = \max_{\mathbf{g}\in\mathcal{G}} \bar{V}^*(\mathbf{s},\mathbf{g})$. ***Comment:*** *Explain nuances better*

Next we show that an optimal WVF provably encodes how all agents can optimally reach all joint goals provided that the joint goals are reachable under the dynamics of the environment, which we refer to as *mastery*. Due to interactions between agents and the dynamics of the environment, it may only be possible for a subset of the agents to reach their assigned goal. Thus, we informally define mastery as the property where the number of agents who reach their assigned goal is maximised. More formally, it is the sum of conditional probabilities of each agent reaching their assigned goal when starting at some joint state:

Definition 4. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. Then $\bar{Q}^*_M$ has mastery if for all $(\mathbf{s},\mathbf{g}) \in \mathcal{S}\times\mathcal{G}$ there exists an optimal joint world policy $\bar{\pi}^*_M(\mathbf{s},\mathbf{g}) \in \arg\max_{\mathbf{a}\in\mathcal{A}} \bar{Q}^*_M(\mathbf{s},\mathbf{g},\mathbf{a})$ such that $\bar{\pi}^*_M \in \arg\max_{\bar{\pi}} \sum_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M)$.*

Theorem 2. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. Then $\bar{Q}^*_M$ has mastery.*

Since the optimal WVF has mastery, there exists an associated optimal JWP that will ensure that the maximum number of agents reach their assigned goal—if they are reachable in the environment. This property is powerful because it (along with Theorem 1) shows that the optimal WVF encodes both how the agents can jointly reach every (reachable) goal and how to solve the associated task. This is leveraged in the following section to enable zero-shot task generalisation after learning the optimal WVF for a single task.

3.2 Task Generalisation

In this section, we show that WVFs encode the common properties across the task distribution as specified by the background CM-MDP. This allows for knowledge learnt in one task to be transferred to another, where the learned optimal WVF for one task can then be used to zero-shot infer the optimal policy of any other task in the task distribution—given the goal rewards that define the new task. We first show that we can retrieve the optimal Q-value function from the optimal WVF by maximising over the set of *reachable* joint goals. We will focus on the set of non-terminated agents $\tilde{I}$ whose joint goals are reachable, ignoring the goals of terminated agents since they cannot reach any goal. We use the notation $\mathbf{g}^{\tilde{I}}$ to denote the assigned agent goals of non-terminated agents and define $\mathbf{g}^{\tilde{I}}$ reachable from $\mathbf{s}'$ as meaning that every non-terminated agent $i \in \tilde{I}$ can reach their assigned goal in $\mathbf{g}$ when all agents start at $\mathbf{s}'$. Additionally, by the assumption of all optimal joint policies being proper, for every $\mathbf{s}' = \mathcal{P}(\mathbf{s}, \mathbf{a})$ there will be at least one joint goal where the non-terminated agents can reach their goal and $\mathcal{G}_R$ will never be empty if $\mathbf{s} \notin \mathcal{G}$.

Lemma 1. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ let $\mathcal{G}_R = \{\mathbf{g} \,|\, \forall \mathbf{g} \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}$ is reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$ be the set of joint goals reachable by the non-terminated agents $\tilde{I} = \{i \,|\, \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$. Then $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}_R} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$.*

If we look back at the definition of the task distribution $\mathcal{M}$ we see that the common properties across tasks are defined by the background CM-MDP B which is also the task $M_B \in \mathcal{M}$ in the task distribution with zero terminal rewards. By Theorem 2 the optimal WVF for M_B encodes how to optimally reach every reachable joint goal in the task which is constant across tasks in the task distribution. Thus, we show that the set of optimal WJPs for every task is equal to the set of optimal WJPs for the background CM-MDP except for where a WJP is defined on an unreachable joint goal. What changes across tasks is the set of goal rewards $\{\hat{R}^i_M\}_{i \in I}$ and as such we show that the optimal WVF for a task $M \in \mathcal{M}$ can be decomposed into the sum of the optimal WVF for the background CM-MDP and the the sum of associated goal rewards for that task for agents that have not terminated provided that the given joint goal is reachable.

Theorem 3. *Let $\bar{Q}^*_M$ and $\bar{Q}^*_{M_B}$ be the optimal world value functions for the task $M \in \mathcal{M}$ and the background CM-MDP M_B, and for all $i \in I$. Let $\{\hat{R}^i_M\}_{i \in I}$ be the set of agent terminal reward functions for the task M. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}_r^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$, we have $\bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^*_{M_B}(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} \hat{R}^i_M(g^i) \mathbb{1}_{s^i \notin \mathcal{G}^i}$*

Thus every optimal WVF in the world can be decomposed into what information is shared across the world and what differs. Provided we have knowledge of what changes across tasks, that is the agent terminal rewards, we can leverage the shared information to zero-shot infer the optimal WVF for a new task given the optimal WVF of another task and the associated goal rewards of both tasks.

Theorem 4. *Let $\bar{Q}^*_{M_1}$ be the optimal world value function for a task $M_1 \in \mathcal{M}$. Let $\{\hat{R}^i_{M_1}\}_{i \in I}$ and $\{\hat{R}^i_{M_2}\}_{i \in I}$ be the sets of agent terminal reward functions for the tasks M_1 and $M_2 \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}_r^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$, we have*

$$\bar{Q}^*_{M_2}(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^*_{M_1}(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} (\hat{R}^i_{M_2}(g^i) - \hat{R}^i_{M_1}(g^i)) \mathbb{1}_{s^i \notin \mathcal{G}^i}$$

If we directly combine Theorem 4 and Lemma 1 we can zero-shot infer the optimal Q-value function for a new task M_2 with $Q^*_{M_2}(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}_R} \{\bar{Q}^*_{M_1}(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} (\hat{R}^i_{M_1}(g^i) - \hat{R}^i_{M_2}(g^i)) \mathbb{1}_{s^i \in \mathcal{G}^i}\}$. But, this requires knowledge of the dynamics of the environment in order to compute $\mathcal{G}_R$. Thus, we show that although Theorem 4 does not hold for $\mathcal{G} \setminus \mathcal{G}_R$ we show that using Theorem 4 to infer the value of unreachable joint goals will still provide values that are smaller than that for reachable joint goals. This means that we can optimally infer the Q-value function from the inferred WVF by maximising over the full set of joint goals - thus requiring no extra knowledge of the environment.

Theorem 5. *Let $\bar{Q}^*_{M_1}$ be the optimal world value function for a task $M_1 \in \mathcal{M}$. Let $\{\hat{R}^i_{M_1}\}_{i \in I}$ and $\{\hat{R}^i_{M_2}\}_{i \in I}$ be the sets of agent terminal reward functions for the tasks M_1 and $M_2 \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$*

$$Q^*_{M_2}(\mathbf{s},\mathbf{a}) = \max_{\mathbf{g}\in\mathcal{G}}\{\bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) + \sum_{i\in I}(\hat{R}^i_{M_1}(g^i) - \hat{R}^i_{M_2}(g^i))\mathbb{1}_{s^i\in\mathcal{G}^i}\}$$

After learning the optimal WVF for a single arbitrary task and given knowledge of the task specific goal rewards we can optimally zero-shot infer the optimal Q-value function and thus optimal joint policy for any task in the task distribution leading to optimal task generalisation.

3.3 Symmetry

While WVFs are powerful in that they allow for zero-shot task generalisation after learning a single task, learning that single task has the potential for poor sample efficiency due to the exponential explosion in the size of joint state joint action joint goal space as the number of agents increases. Thus, we propose leveraging symmetries in the environment to reduce the number of values that must be learnt. Particularly we reduce the number of joint goals that must be learnt. First we consider symmetries present from our definition of termination occurring at an agent level and next we consider the special case of homogeneous agents.

3.3.1 Agent Level Termination

Consider a terminated agent $i \in I$ where $s^i \in \mathcal{G}^i$, by our assumptions on agent level termination the reward obtained by agent i will always be zero. This is regardless of the goal agent i was trying to reach, and as such there exists an equivalence relation in the WVF where all the goals of a terminated agent are equivalent. Thus, we substitute the goal followed by agent i with a dummy variable $g^{i,T}$ represented the equivalence class containing all elements of $\mathcal{G}^i$. This is given in the following theorem.

Theorem 6. *For all $i \in I$. Let $g^{i,T} \in \mathcal{G}^i$. For all $(s^{-i}, g^{-i}) \in \mathcal{S}^{-i} \times \mathcal{G}^{-i}$, for all $(s^i, g^i) \in \mathcal{G}^i \times \mathcal{G}^i$, for all $\mathbf{a} \in \mathcal{A}$ we have: $\bar{Q}^*(\mathbf{s}, \langle g^{-i}, g^{i,T}\rangle, \mathbf{a}) = \bar{Q}^*(\mathbf{s}, \langle g^{-i}, g^i\rangle, \mathbf{a})$.*

Thus, when an agent i is terminated we can reduce the joint goal space that must be considered by a factor of $|\mathcal{G}^i|$. This theorem also supports multiple terminated agents, further reducing the size of the joint goal space.

3.3.2 Homogeneity

Consider tasks where agents are homogeneous and interchangeable such that for all permutations $Perm$ we have $p(Perm(\mathbf{s}), Perm(\mathbf{a})) = Perm(\mathbf{s}')$ and $R^i(Perm(\mathbf{s}), Perm(\mathbf{a}), Perm(\mathbf{s}')) = R^i(\mathbf{s}, \mathbf{a}, \mathbf{s}')\,\forall i \in I$, and for all $(i,j) \in I \times I$ we have $\mathcal{S}^i = \mathcal{S}^j$, $\mathcal{A}^i = \mathcal{A}^j$, and $R^i(\mathbf{s}, \mathbf{a}, \mathbf{s}') = R^j(\mathbf{s}, \mathbf{a}, \mathbf{s}')$. As a result the WVF is invariant under the permutation of the joint state, joint action and joint goal as shown in the following theorem.

Theorem 7. *Let $\bar{Q}^{\bar{\pi}}_M$ be the world value function for a task $M \in \mathcal{M}$ under a joint world policy $\bar{\pi}$. $\bar{Q}^{\bar{\pi}}_M$ is invariant to any permutation $Perm$ applied to the joint state, joint action and joint goal. That is, for all permutations $Perm$ and for all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$ we have: $\bar{Q}^{\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^{\bar{\pi}}(Perm(\mathbf{s}), Perm(\mathbf{g}), Perm(\mathbf{a}))$*

Due to this, there exists a set of equivalence classes where each contains every permutation of a particular joint goal. We define a new set of joint goals $\mathcal{G}_b \subseteq \mathcal{G}$ known as the set of *base joint goals* which contains one joint goal from each equivalence class. As a result of Theorem 7 we only need to learn the values associated with each base joint goal $\mathbf{g}_b \in \mathcal{G}_b$ and can retrieve the values associated with the full set of joint goals using $\bar{Q}(Perm_{\mathbf{g}_b,\mathbf{g}}(\mathbf{s}), Perm_{\mathbf{g}_b,\mathbf{g}}(\mathbf{g}), Perm_{\mathbf{g}_b,\mathbf{g}}(\mathbf{a})) = \bar{Q}(\mathbf{s}, \mathbf{g}_b, \mathbf{a})$ where $Perm_{\mathbf{g}_b,\mathbf{g}}$ is the permutation that converts $\mathbf{g}_b$ to $\mathbf{g}$. The upper bound for the size of the joint goal space is $|\mathcal{G}| = \prod_{i\in I} |\mathcal{G}^i| = |\mathcal{G}^j|$ where j is any agent due to homogeneity and the associated upper bound for the size of the base joint goal space is $|\mathcal{G}_b| = {}^{|\mathcal{G}^j|}P_n$. Thus, the size of the base joint goal space is significantly smaller than that of the joint goal space.

3.4 Learning a WVF

In order to learn the WVF we extend goal-oriented learning (GOL) to the multi-agent setting and leverage the symmetries described in Subsection 3.3. Since we consider a CC, GOL only requires minor modification to be applied to the multi-agent setting. Specifically, we must now compute the extended agent reward functions and extended team reward function. To leverage symmetry from homogeneity, we simply learn the values associated with the set of base joint goals $\mathcal{G}_b$, and to leverage symmetry from agent level termination, we augment the space of joint goals with the dummy variable $g^{i,T}$ when appropriate for terminated agents. The full pseudocode for this with an additional ablation experiment are given in the appendix.

4 Experiments

In this section we empirically demonstrate that given a fixed step quota for pretraining, our method can optimally learn or infer the full task distribution, while comparative methods can only learn a small subset of the tasks. We consider a tabular grid world domain with two homogeneous agents and a varying number of goals with randomised locations. Each goal has an associated reward of 10 or -10, denoting whether that goal is desirable or not. For each number of goals and layout of goals, there exists a task distribution containing all possible assignments of 'desirable' or 'undesirable' rewards to said goals. Figure 2 shows a variety of grid world layouts, and we note that each grid shown is just one task sampled from the associated task distribution for that particular layout.

Since agents are homogeneous, all agents share the same reward function and receive the same terminal rewards upon reaching goals. The reward function is sparse, with agents either receiving a terminal reward or a step reward of -0.01. The action space of each agent is the cardinal directions and a 'STAY' action. The dynamics are such that agents cannot occupy the same position or pass through each other. To achieve a goal, an agent must take the 'STAY' action while at the coordinate of a goal, upon which it terminates and receives the associated reward. In order to align this domain with our problem definition, we augment the agent state space with a done status such that the goal states are represented as (coordinate of goal, done=True) and an agent transitions into this state upon taking the 'STAY' action at the state (coordinate of goal, done=False).

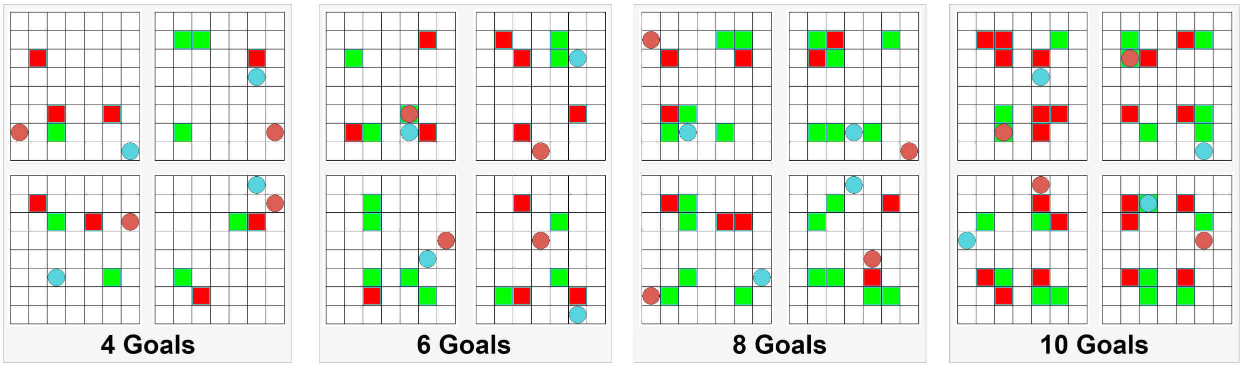
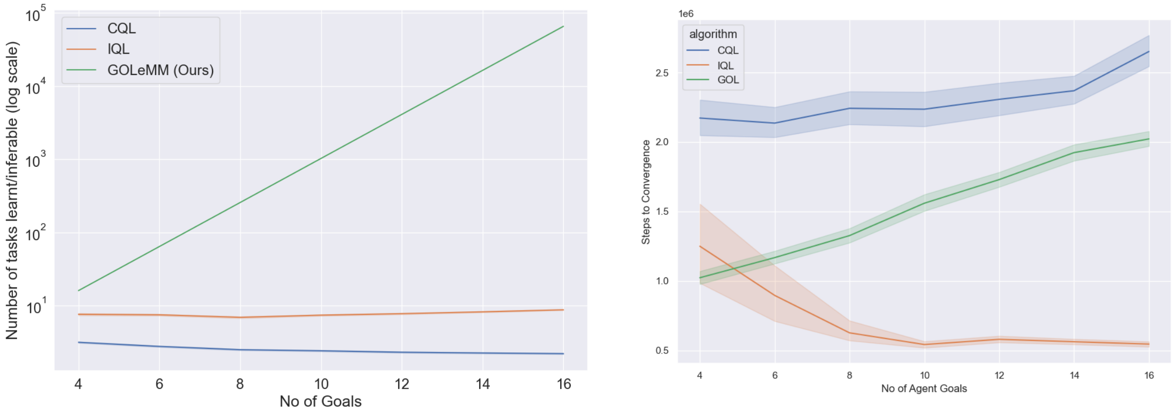


Figure 2: Sample grid world layouts with varying numbers of goals—each represented by colored squares. Squares are green if desirable and red if undesirable. The two agents are represented by colored circles.

In Figure 3(a) we compare the number of tasks learnt or inferred by GOLeMM against those learnt by traditional Q-learning operating on the joint spaces referred to as centralised Q-learning (CQL), and independent Q-learning (IQL). We vary the number of goals and the locations of goals. For each run, we measure how many tasks can be optimally learnt or inferred given a fixed step quota for learning. Here we randomly sample a task from the associated task distribution, optimally learn or infer the task, and then sequentially move onto a new task where optimality is evaluated by comparing evaluated returns against optimal returns as computed by optimal multi-agent path finding. In the case of GOLeMM our method learns the optimal WVF for the first task seen then uses Theorem 5 to zero-shot infer the rest of the tasks and we assume it has knowledge of the goal rewards for each task. Since our method only needs to learn a single task and is then able to zero-shot infer the rest of task distribution we observed that it was able to optimally learn or infer the full task distribution for 99.81% of the runs where in the remaining runs it was only unable to optimally infer 1 or 2 tasks. We suspect this is due to minor sub-optimality in the learnt WVF which was not picked up by our method of evaluating optimality. Since CQL and IQL must learn each task independently they are only able to learn a small subset of the task distribution while ours on average is able to learn the full task distribution whose size is exponential in the number of goals.



(a) Number of tasks learnt/inferable

(b) Steps until convergence

Figure 3: Results in the grid world. (a) The number of tasks learnt/inferable given a fixed step quota for GOLeMM, CQL and IQL as a function of number of goals. The y-axis is plotted on a log scale. (b) shows the steps until optimal convergence of CQL, IQL and GOL as a function of the number of goals in the environment. For each algorithm and each number of goals in (a-b), 512 runs were executed with varying goal locations and the mean was taken to plot the lines along with 95% confidence intervals. The confidence intervals in (a) are too small to be seen due to the high number of runs and the log scale.

Figure 3(b) shows the number of steps until optimal convergence of GOLeMM for learning a single task (labeled GOL) compared to that of IQL and CQL as a function of the number of goals in the environment. We see that our method outperforms a representative centralised method, CQL, but is substantially less sample efficient in learning a single task than a representative decentralised method, IQL. Although our method must learn how to achieve every joint goal in the environment it is still able to learn quicker than CQL which only needs to learn how to achieve the current task. We believe this is due to the goal-oriented exploration inherent in our method and the leveraging of symmetries. IQL is more sample efficient on learning a single task because it requires less values to learn. However even in a setting with minimal coordination between agents like this one, we observed that IQL failed to converge to an optimal solution for a non-negligible percentage of the tasks when the number of goals is low: approxmiately 6.2% for 4 goals, 3.1% for 6 goals, 0.3% for 8 goals, and 0% for the rest. Finally, although GOLeMM has a higher upfront cost of learning a single task than IQL, it is clear that the increased upfront cost of learning a single task is made up for by being able to immediately generalise over the task distribution.

5 Conclusion

In this work we extend WVFs to the multi-agent setting and prove that they capture the common properties across the world. Using this, after learning the WVF for an arbitrary task we are able to zero-shot infer the optimal policy for any other task given knowledge of the terminal rewards. Thus we have proposed a method for optimally generalising over a task distribution in a multi-agent setting which to the best of our knowledge is the first work to do so. Empirically we show that given a fixed step quota our method can infer the entire task distribution, while comparative methods are only able to learn a small percentage with the gap between our method and theirs growing exponentially as the number of goals in the environment increases. Finally, we note that there is number of limitations in the current work. For example, we focused on the centralised setting, but this presents challenges in terms of agent scaling (the joint action space grows exponentially with the number of agents). While these assumptions are similar to those made in prior theoretical works, an exciting direction for future works is to relax them, extending our results to the broader distributions of tasks and RL algorithms.

References

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Jiayu Chen, Yuanxin Zhang, Yuanfan Xu, Huimin Ma, Huazhong Yang, Jiaming Song, Yu Wang, and Yi Wu. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. *Advances in Neural Information Processing Systems*, 34:9681–9693, 2021.

Carlos Ernesto Guestrin. *Planning under uncertainty in complex structured environments*. Stanford University, 2003.

Jelle R Kok and Nikos Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research*, 7:1789–1828, 2006.

Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, Gereon Vienken, Irene Sturm, Guillaume Sartoretti, and Giacomo Spigler. Flatland-rl : Multi-agent reinforcement learning on trains, 2020.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9497–9507, 2020.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. Generalisation in lifelong reinforcement learning through logical composition. In *International Conference on Learning Representations*, 2021.

Geraud Nangue Tasse, Benjamin Rosman, and Steven James. World value functions: Knowledge representation for learning and planning. *arXiv preprint arXiv:2206.11940*, 2022.

Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. Composing value functions in reinforcement learning. In *International conference on machine learning*, pp. 6401–6409. PMLR, 2019.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Peter R Wurman, Raffaello D'Andrea, and Mick Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI magazine*, 29(1):9–9, 2008.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. *Advances in Neural Information Processing Systems*, 33:4767–4777, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

A Appendix

A.1 Proofs & Supporting Lemmas

A.1.1 World Value Functions

Theorem 1. *Let R_M and $\bar{R}_M$ be the team reward function and extended reward function for a task $M \in \mathcal{M}$. Let Q^*_M and $\bar{Q}^*_M$ be the optimal Q-value function and optimal world value function for a task $M \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}, \mathbf{s}') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have:*

(i) $R_M(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \max_{\mathbf{g} \in \mathcal{G}} \bar{R}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}')$*, and*

(ii) $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$.

Proof.

(i) Let $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$ with $\tilde{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') = \sum_{i \in I} \tilde{R}^i(\mathbf{s}, \mathbf{g}^i, \mathbf{a}, \mathbf{s}')$.

$$\begin{aligned}
\max_{\mathbf{g} \in \mathcal{G}} \tilde{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') &= \max_{\mathbf{g} \in \mathcal{G}} \sum_{i \in I} \tilde{R}^i(\mathbf{s}, \mathbf{g}^i, \mathbf{a}, \mathbf{s}') \\
&= \sum_{i \in I} \max_{\mathbf{g}^i \in \mathcal{G}^i} \tilde{R}^i(\mathbf{s}, \mathbf{g}^i, \mathbf{a}, \mathbf{s}') \\
&= \sum_{i \in I} R^i(\mathbf{s}, \mathbf{a}, \mathbf{s}') && \text{By (Nangue Tasse et al., 2020)} \\
&= R(\mathbf{s}, \mathbf{a}, \mathbf{s}').
\end{aligned}$$

(ii) Each $\mathbf{g} \in \mathcal{G}$ can be thought of as defining a CM-MDP $M_{\mathbf{g}} = (I, \{\mathcal{S}^i\}_{i \in I}, \{\mathcal{A}^i\}_{i \in I}, \mathcal{P}, \{R^i_{M_{\mathbf{g}}}\}_{i \in I})$ with the team reward function defined as $R_{M_{\mathbf{g}}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \tilde{R}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') = \sum_{i \in I} \tilde{R}^i_M(\mathbf{s}, \mathbf{g}^i, \mathbf{a}, \mathbf{s}')$ and optimal action-value function $Q^*_{M_{\mathbf{g}}}(\mathbf{s}, \mathbf{a}) = \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$. By the use of $R(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \max_{\mathbf{g} \in \mathcal{G}} \tilde{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}')$ and from (Van Niekerk et al., 2019) (2019, Corollary 1), we have that $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}} Q^*_{M_{\mathbf{g}}}(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$.

□

Theorem 2. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. Then $\bar{Q}^*_M$ has mastery.*

Proof.
Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. Let $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Let $\tilde{I} = \{i | \forall i \in I \text{ where } s^i \notin \mathcal{G}^i\}$.

For all $i \in I \setminus \tilde{I}$ and for all $\bar{\pi} \in \bar{\Pi}$ where $\bar{\Pi}$ is the set of all optimal joint world policies:

$$\begin{aligned}
P(s^i_{T^i+1} = g^i | \mathbf{s}, \mathbf{g}, \bar{\pi}, M) &= \begin{cases} 1 & \text{if } s^i = g^i \\ 0 & \text{otherwise} \end{cases} \\
&= P(s^i_{T^i+1} = g^i | \mathbf{s}, \mathbf{g}, M)
\end{aligned}$$

Since each agent $i \in I \setminus \tilde{I}$ has already terminated at the goal $s^i \in \mathcal{G}^i$. Therefore

$$
\begin{aligned}
\arg\max_{\bar{\pi}} \sum_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M) &= \arg\max_{\bar{\pi}} \big\{ \sum_{i\in \tilde{I}} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M) \\
&\quad + \sum_{i\in I\setminus\tilde{I}} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M) \big\} \\
&= \arg\max_{\bar{\pi}} \big\{ \sum_{i\in \tilde{I}} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M) \big\}
\end{aligned}
$$

Thus, we only need to consider the non terminated agents in $\tilde{I}$.

Let $\bar{\pi}^*(\mathbf{s},\mathbf{g}) \in \bar{Q}^*(\mathbf{s},\mathbf{g},\mathbf{a})$ be a deterministic optimal world joint policy for the task M. Let $\bar{\pi}^{\mathbf{g}}$ be a deterministic world joint policy that satisfies $\bar{\pi}^{\mathbf{g}} \in \arg\max_{\bar{\pi}} \sum\limits_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M)$. Similarly to before this implies $\bar{\pi}^{\mathbf{g}} \in \arg\max_{\bar{\pi}} \sum\limits_{i\in \tilde{I}} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M)$. We use proof by contradiction and assume $\sum\limits_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi}^*,M) < \sum\limits_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi}^{\mathbf{g}},M)$, which we show leads to a contradiction in the form of $\bar{V}^*(\mathbf{s},\mathbf{g}) < \bar{V}^{\bar{\pi}^{\mathbf{g}}}(\mathbf{s},\mathbf{g})$. By this contradiction we we must have $\bar{\pi}^* \in \arg\max_{\bar{\pi}} \sum\limits_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi},M)$. This contradiction is shown below.

Let $G^{\star,i}_{T^i-1} = \mathbb{E}^{\bar{\pi}^*}[\sum\limits_{t=0}^{T^i-1} \bar{R}^i(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})]$ and $G^{g,i}_{T^i-1} = \mathbb{E}^{\bar{\pi}^{\mathbf{g}}}[\sum\limits_{t=0}^{T^i-1} \bar{R}^i(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})]$. Let $P^{i,*} = P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi}^*,M)$ and let $P^{i,\mathbf{g}} = P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\bar{\pi}^{\mathbf{g}},M)$. Let $P^* = \sum\limits_{i\in\tilde{I}} P^{i,*}$ and $P^{\mathbf{g}} = \sum\limits_{i\in\tilde{I}} P^{i,\mathbf{g}}$. Let $m = |\tilde{I}|$.

We first show that the following inequality holds

$$
\sum_{i\in\tilde{I}} (G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) + \sum_{i\in\tilde{I}} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) < (P^* - P^{\mathbf{g}})\tilde{R}_{min}
$$

We do this by first establishing (i) an upper bound on $\sum\limits_{i\in\tilde{I}} (G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) + \sum\limits_{i\in\tilde{I}} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}})$ and (ii) a lower bound on $(P^* - P^{\mathbf{g}})\tilde{R}_{min}$.

(i) D corresponds to the maximum length of a trajectory under an optimal joint policy, thus $T^i \leq D$. Since $R_{min} \leq \bar{R}^i(\mathbf{s},\mathbf{g},\mathbf{a},\mathbf{s}') \leq R_{max}$, we have $G^{\star,i}_{T^i-1} \leq R_{max}(D-1)$ and $-G^{g,i}_{T^i-1} \leq -R_{min}(D-1)$. Therefore,

$$
\begin{aligned}
&G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1} \leq (R_{max} - R_{min})(D-1) \\
\implies &\sum_{i\in\tilde{I}} (G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) \leq \sum_{i\in\tilde{I}} (R_{max} - R_{min})(D-1) \\
\implies &\sum_{i\in\tilde{I}} (G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) \leq \sum_{i\in\tilde{I}} (R_{max} - R_{min})(D-1) \leq m(R_{max} - R_{min})(D-1)
\end{aligned}
$$

Let $\tilde{I}^+ = \{i | \forall i \in \tilde{I} \text{ where } P^{i,*} - P^{i,\mathbf{g}} \geq 0\}$.

$$
\begin{aligned}
&\hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq R_{max}(P^{i,*} - P^{i,\mathbf{g}}) \forall i \in \tilde{I}^+ \\
\implies &\sum_{i\in\tilde{I}^+} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq R_{max} \sum_{i\in\tilde{I}^+} (P^{i,*} - P^{i,\mathbf{g}})
\end{aligned}
$$

Let $\tilde{I}^- = \{i | \forall i \in \tilde{I}$ where $P^{i,*} - P^{i,\mathbf{g}} < 0\}$.

$$\begin{aligned}
&\hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq R_{min}(P^{i,*} - P^{i,\mathbf{g}})\, \forall i \in \tilde{I}^- \\
\implies &\sum_{i \in \tilde{I}^-} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq R_{min} \sum_{i \in \tilde{I}^-} (P^{i,*} - P^{i,\mathbf{g}})
\end{aligned}$$

Therefore, $\sum_{i \in \tilde{I}} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq R_{max} \sum_{i \in \tilde{I}^+} (P^{i,*} - P^{i,\mathbf{g}}) + R_{min} \sum_{i \in \tilde{I}^-} (P^{i,*} - P^{i,\mathbf{g}})$.

$$\begin{aligned}
&P^{i,*} - P^{i,\mathbf{g}} \leq 1\, \forall i \in \tilde{I}^+ \\
\implies &\sum_{i \in \tilde{I}^+} (P^{i,*} - P^{i,\mathbf{g}}) \leq m && \text{since } |\tilde{I}^+| \leq |\tilde{I}| \leq m \\
\implies R_{max} &\sum_{i \in \tilde{I}^+} (P^{i,*} - P^{i,\mathbf{g}}) \leq mR_{max}
\end{aligned}$$

$$\begin{aligned}
&P^{i,*} - P^{i,\mathbf{g}} \geq -1\, \forall i \in \tilde{I}^- \\
\implies &\sum_{i \in \tilde{I}^-} (P^{i,*} - P^{i,\mathbf{g}}) \geq -m \\
\implies R_{min} &\sum_{i \in \tilde{I}^-} (P^{i,*} - P^{i,\mathbf{g}}) \leq mR_{min}
\end{aligned}$$

Thus, $\sum_{i \in \tilde{I}} \hat{R}^i_M(P^{i,*} - P^{i,\mathbf{g}}) \leq m(R_{max} - R_{min})$. From this and before, we get

$$\sum_{i \in \tilde{I}} (G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) + \sum_{i \in \tilde{I}} \hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) \leq m(R_{max} - R_{min})D$$

ii) Since $\bar{\pi}^*$ and $\bar{\pi}^{\mathbf{g}}$ are assumed to be deterministic, then for all $i \in \tilde{I}$ $P^{i,*} \in \{0, 1\}$ and $P^{i,\mathbf{g}} \in \{0, 1\}$. Thus, $P^* \in \mathbb{N}$ and $P^{\mathbf{g}} \in \mathbb{N}$. Since by assumption $P^* < P^{\mathbf{g}}$ we have $P^* - P^{\mathbf{g}} \leq -1$.

$$\begin{aligned}
&\tilde{R}_{min} \leq n((R_{min} - R_{max})D + R_{min}) \\
\implies &n((R_{min} - R_{max})D + R_{min})(P^* - P^{\mathbf{g}}) \leq \tilde{R}_{min}(P^* - P^{\mathbf{g}})
\end{aligned}$$

$$\begin{aligned}
&R_{min} - R_{max} < 0 \\
\implies &n(R_{min} - R_{max})D < 0 \\
\implies &n((R_{min} - R_{max})D + R_{min}) < n(R_{min} - R_{max})D < 0
\end{aligned}$$

Since $n((R_{min} - R_{max})D + R_{min}) < 0$ and $P^* - P^{\mathbf{g}} \leq -1$,

$$\begin{aligned}
&-n((R_{min} - R_{max})D + R_{min}) \leq n((R_{min} - R_{max})D + R_{min})(P^* - P^{\mathbf{g}}) \\
\implies &n((R_{max} - R_{min})D - R_{min}) \leq n((R_{min} - R_{max})D + R_{min})(P^* - P^{\mathbf{g}})
\end{aligned}$$

Thus, $n((R_{max} - R_{min})D - R_{min}) \leq \tilde{R}_{min}(P^* - P^{\mathbf{g}})$.

Next we show that $m(R_{max} - R_{min})D < n((R_{max} - R_{min})D - R_{min})$.

$$\begin{aligned} m(R_{max} - R_{min})D &\leq n(R_{max} - R_{min})D && \text{since } m \leq n \text{ and } (R_{max} - R_{min})D \geq 0 \\ \implies m(R_{max} - R_{min})D + nR_{min} &< n(R_{max} - R_{min})D && \text{since } nR_{min} < 0 \\ \implies m(R_{max} - R_{min})D &< n((R_{max} - R_{min})D - R_{min}) \end{aligned}$$

Thus, $\sum_{i\in\tilde{I}}(G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) + \sum_{i\in\tilde{I}}\hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}})< (P^* - P^{\mathbf{g}})\tilde{R}_{min}$.

$$\begin{aligned} &\sum_{i\in\tilde{I}}(G^{\star,i}_{T^i-1} - G^{g,i}_{T^i-1}) + \sum_{i\in\tilde{I}}\hat{R}^i_M(g^i)(P^{i,*} - P^{i,\mathbf{g}}) < (P^* - P^{\mathbf{g}})\tilde{R}_{min} \\ &\qquad \implies \sum_{i\in\tilde{I}}(G^{\star,i}_{T^i-1} + P^{i,*}\hat{R}^i_M(g^i)) - P^*\tilde{R}_{min} < \sum_{i\in\tilde{I}}(G^{g,i}_{T^i-1} + P^{i,\mathbf{g}}\hat{R}^i_M(g^i)) - P^{\mathbf{g}}\tilde{R}_{min} \end{aligned}$$

Let $m = |\tilde{I}|$.

$$\begin{aligned} &\sum_{i\in\tilde{I}}(G^{\star,i}_{T^i-1} + P^{i,*}\hat{R}^i_M(g^i)) - P^*\tilde{R}_{min} + m\tilde{R}_{min} < \sum_{i\in\tilde{I}}(G^{g,i}_{T^i-1} + P^{i,\mathbf{g}}\hat{R}^i_M(g^i)) - P^{\mathbf{g}}\tilde{R}_{min} + m\tilde{R}_{min} \\ &\implies \sum_{i\in\tilde{I}}(G^{\star,i}_{T^i-1} + P^{i,*}\hat{R}^i_M(g^i) + (1 - P^{i,*})\tilde{R}_{min}) < \sum_{i\in\tilde{I}}(G^{g,i}_{T^i-1} + P^{i,\mathbf{g}}\hat{R}^i_M(g^i) + (1 - P^{i,\mathbf{g}})\tilde{R}_{min}) \end{aligned}$$

Since $m\tilde{R}_{min} - P^*\tilde{R}_{min} = \sum_{i\in\tilde{I}}(1 - P^{i,*})\tilde{R}_{min}$ and $m\tilde{R}_{min} - P^{\mathbf{g}}\tilde{R}_{min} = \sum_{i\in\tilde{I}}(1 - P^{i,\mathbf{g}})\tilde{R}_{min}$. Substituting, we have

$$\sum_{i\in\tilde{I}}(\mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1}\bar{R}^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] + P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}^*, M)\hat{R}^i_M(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}^*, M)\tilde{R}_{min})$$

$$<$$

$$\sum_{i\in\tilde{I}}(\mathbb{E}^{\bar{\pi}^{\mathbf{g}}}[\sum_{t=0}^{T^i-1}\bar{R}^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] + P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}^{\mathbf{g}}, M)\hat{R}^i_M(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}^{\mathbf{g}}, M)\tilde{R}_{min})$$

$$\implies \bar{V}^*_M(\mathbf{s}, \mathbf{g}) < \bar{V}^{\bar{\pi}^{\mathbf{g}}}_M(\mathbf{s}, \mathbf{g})$$

Which is a contradiction. Thus, we must have $\bar{\pi}^* \in \arg\max_{\bar{\pi}} \sum_{i\in I} P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}, M)$. Therefore we have proved that $\bar{Q}^*_M$ has mastery.

□

A.1.2 Task Generalisation

Supporting Lemmas

Lemma 2. *Let $\bar{V}^{\bar{\pi}}_M$ and $\bar{Q}^{\bar{\pi}}_M$ be the state world value function and world value function for a task $M \in \mathcal{M}$ under a joint world policy $\bar{\pi}$. For all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$*

(i) $\bar{V}^{\bar{\pi}}_M(\mathbf{s}, \mathbf{g}) = \sum_{i\in I} \bar{V}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g})$*, where* $\bar{V}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}) := \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})]$ *for all $i \in I$, and*

(ii) $\bar{Q}^{\bar{\pi}}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \sum_{i\in I} \bar{Q}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$*, where* $\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) := \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \bar{V}^{i,\bar{\pi}}_M(\mathbf{s}', \mathbf{g})]$ *for all $i \in I$.*

Proof. For all $(\mathbf{s},\mathbf{g},\mathbf{a}) \in \mathcal{S}\times\mathcal{G}\times\mathcal{A}$:

(i)

$$\begin{aligned}
\bar{V}^{\bar{\pi}}_M(\mathbf{s},\mathbf{g}) &:= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}_M(\mathbf{s}_t,\mathbf{g},\mathbf{a}_t,\mathbf{s}_{t+1})] \\
&= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty}\sum_{i\in I} \bar{R}^i_M(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})] && \text{By 2} \\
&= \sum_{i\in I}\mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^i_M(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})] \\
&= \sum_{i\in I}\bar{V}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g})
\end{aligned}$$

(ii)

$$\begin{aligned}
\bar{Q}^{\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) &:= \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}_M(\mathbf{s},\mathbf{a},\mathbf{g},\mathbf{s}') + \bar{V}^{\bar{\pi}}_M(\mathbf{s}',\mathbf{g})] \\
&= \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\sum_{i\in I}\bar{R}^i_M(\mathbf{s},\mathbf{a},\mathbf{g},\mathbf{s}') + \sum_{i\in I}\bar{V}^{i,\bar{\pi}}_M(\mathbf{s}',\mathbf{g})] && \text{By 2 and (i)} \\
&= \sum_{i\in I}\{\mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^i_M(\mathbf{s},\mathbf{a},\mathbf{g},\mathbf{s}') + \bar{V}^{i,\bar{\pi}}_M(\mathbf{s}',\mathbf{g})]\} \\
&= \sum_{i\in I}\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a})
\end{aligned}$$

□

Lemma 3. *Let $\bar{V}^{\bar{\pi}}_M$ and $\bar{Q}^{\bar{\pi}}_M$ be the state world value function and world value function for a task $M \in \mathcal{M}$ under a joint world policy $\bar{\pi}$. For all $(\mathbf{s},\mathbf{g},\mathbf{a}) \in \mathcal{S}\times\mathcal{G}\times\mathcal{A}$*

(i) $\bar{V}^{\bar{\pi}}_M(\mathbf{s},\mathbf{g}) = \sum_{i\in\tilde{I}}\bar{V}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g})$,

where $\bar{V}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g}) := \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty}\bar{R}^i_M(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})]$ *for all* $i\in\tilde{I}$, *and*

(ii) $\bar{Q}^{\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) = \sum_{i\in\tilde{I}}\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a})$,

where $\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) := \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^i_M(\mathbf{s},g^i,\mathbf{a},\mathbf{s}') + \bar{V}^{i,\bar{\pi}}_M(\mathbf{s}',\mathbf{g})]$ *for all* $i\in\tilde{I}$,

where $\tilde{I} = \{i | \forall i \in I \text{ where } s^i \notin \mathcal{G}\}$.

Proof. For all $(\mathbf{s},\mathbf{g},\mathbf{a}) \in \mathcal{S}\times\mathcal{G}\times\mathcal{A}$:

(i) For all $i \in I\setminus\tilde{I}$

$$\begin{aligned}
\bar{V}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g}) &= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty}\bar{R}^i_M(\mathbf{s}_t,g^i,\mathbf{a}_t,\mathbf{s}_{t+1})] \\
&= 0
\end{aligned}$$

Since by assumptions on agent level termination if $s_0^i \in \mathcal{G}^i$, that is if $i \in I \setminus \tilde{I}$, then $\bar{R}_M^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) = 0$ for all $t \in [0, \infty)$. Therefore we have

$$\begin{aligned}
\bar{V}_M^{\bar{\pi}}(\mathbf{s}, \mathbf{g}) &= \sum_{i \in I} \bar{V}_M^{i,\bar{\pi}}(\mathbf{s}, \mathbf{g}) \\
&= \sum_{i \in \tilde{I}} \bar{V}_M^{i,\bar{\pi}}(\mathbf{s}, \mathbf{g})
\end{aligned}$$

(ii) For all $i \in I \setminus \tilde{I}$

$$\begin{aligned}
\bar{Q}_M^{i,\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a}) &= \mathbb{E}_{\mathbf{s}'}^{\bar{\pi}}[\bar{R}_M^i(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}' + \bar{V}_M^{i,\bar{\pi}}(\mathbf{s}', \mathbf{g}))] \\
&= 0 && \text{Since } s^i \in \mathcal{G}^i \text{ and from (i)}
\end{aligned}$$

Therefore, $\bar{Q}_M^{\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \sum_{i \in \tilde{I}} \bar{Q}_M^{i,\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a})$

□

Lemma 4. *Let $\bar{Q}_M^{\bar{\pi}}$ and $\bar{V}_M^{\bar{\pi}}$ be the world value function and world state value function for a task $M \in \mathcal{M}$ under a world joint policy $\bar{\pi}$. Let $\{\hat{R}_M^i\}_{i \in I}$ be the set of agent terminal reward functions.*

For all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$:

(i)

$$\begin{aligned}
\bar{V}_M^{\bar{\pi}}(\mathbf{s}, \mathbf{g}) = &\sum_{i \in \tilde{I}} \Big\{ \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}_M^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] + P(s_{T^i+1}^i = g^i \,|\, \mathbf{s}, \mathbf{g}, \bar{\pi}, M) \hat{R}_M^i(g^i) \\
&+ P(s_{T^i+1}^i \neq g^i \,|\, \mathbf{s}, \mathbf{g}, \bar{\pi}, M) \tilde{R}_{min} \Big\}
\end{aligned}$$

(ii)

$$\begin{aligned}
\bar{Q}_M^{\bar{\pi}}(\mathbf{s}, \mathbf{g}, \mathbf{a}) = &\sum_{i \in \tilde{I}} \Big\{ \bar{R}_M^i(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') \mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}_M^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] \\
&+ P(s_{T^i+1}^i = g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) \hat{R}_M^i(g^i) \\
&+ P(s_{T^i+1}^i \neq g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) \tilde{R}_{min} \Big\}
\end{aligned}$$

where $\tilde{I} = \{i | \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$.

Proof. For all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$:
Let $\tilde{I} = \{i | \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$.

(i) For all $i \in \tilde{I}$:

$$\begin{aligned}
\bar{V}_M^{i,\bar{\pi}}(\mathbf{s}, \mathbf{g}) &= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}_M^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] \\
&= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i} \bar{R}_M^i(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})]
\end{aligned}$$

Since by assumptions on agent level termination $\bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) = 0$ for all $t \in [T^i, \infty)$

$$\begin{aligned}
\bar{V}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}) &= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] + \mathbb{E}^{\bar{\pi}}[\bar{R}^i_M(\mathbf{s}_{T^i}, g^i, \mathbf{a}_{T^i}, \mathbf{s}_{T^i+1})] \\
&= \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] + P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}, M)\hat{R}^i_M(g^i) \\
&+ P(s^i_{T^i+1} \neq g^i|\mathbf{s}, \mathbf{g}, \bar{\pi}, M)\tilde{R}_{min}
\end{aligned}$$

since

$$\bar{R}^i_M(\mathbf{s}_{T^i}, g^i, \mathbf{a}_{T^i}, \mathbf{s}_{T^i+1}) = \begin{cases} \hat{R}^i_M(g^i) & \text{if } s^i_{T^i+1} = g^i \\ \tilde{R}_{min} & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned}
V^{\bar{\pi}}_M(\mathbf{s}, \mathbf{g}) = &\sum_{i \in \tilde{I}} \Big\{ \mathbb{E}^{\bar{\pi}}\Big[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})\Big] + P(s^i_{T^i+1} = g^i \,|\, \mathbf{s}, \mathbf{g}, \bar{\pi}, M)\hat{R}^i_M(g^i) \\
&+ P(s^i_{T^i+1} \neq g^i \,|\, \mathbf{s}, \mathbf{g}, \bar{\pi}, M)\tilde{R}_{min} \Big\}
\end{aligned}$$

(ii) For all $i \in \tilde{I}$:

$$\begin{aligned}
\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) &= \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \bar{V}^{i,\bar{\pi}}_M(\mathbf{s}', \mathbf{g})] \\
&= \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})]] && \text{By determinism assumption} \\
&= \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})]
\end{aligned}$$

Consider CASE 1 where $s^{i'} \notin \mathcal{G}^i$, then

$$\begin{aligned}
\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = &\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] \\
&+ P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M)\hat{R}^i(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M)\tilde{R}_{min}
\end{aligned}$$

Consider CASE 2 where $s^{i'} \in \mathcal{G}^i$, then:

$$\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') = \begin{cases} \hat{R}^i_M(g^i) & \text{if } s^{i'} = g^i \\ \tilde{R}_{min} & \text{otherwise} \end{cases}$$

and

$$P(s^i_{T^i+1} = g^i|\mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) = \begin{cases} 1 & \text{if } s^{i'} = g^i \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned}\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) = P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\hat{R}^i_M(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\tilde{R}_{min} \\ + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})]\end{aligned}$$

Since $\mathbb{E}^{\bar{\pi}}[\sum\limits_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] = \mathbb{E}^{\bar{\pi}}[\sum\limits_{t=0}^{\infty} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] = 0$ when $s^i_0 \in \mathcal{G}^i$. Combining CASE 1 and CASE 2 we get

$$\begin{aligned}\bar{Q}^{i,\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) &= (\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] \\ &+ P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\hat{R}^i(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\tilde{R}_{min})\mathbb{1}_{s^i \notin \mathcal{G}^i} \\ &+ (P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\hat{R}^i_M(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\tilde{R}_{min} \\ &+ \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})])\mathbb{1}_{s^i \in \mathcal{G}^i} \\ &= \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] \\ &+ P(s^i_{T^i+1} = g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\hat{R}^i(g^i) + P(s^i_{T^i+1} \neq g^i|\mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\tilde{R}_{min}\end{aligned}$$

By Lemma 3 we get:

$$\begin{aligned}Q^{\bar{\pi}}_M(\mathbf{s},\mathbf{g},\mathbf{a}) = \sum_{i \in \tilde{I}} \Big\{ &\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}}\Big[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})\Big] \\ &+ P(s^i_{T^i+1} = g^i \,|\, \mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\hat{R}^i_M(g^i) \\ &+ P(s^i_{T^i+1} \neq g^i \,|\, \mathbf{s},\mathbf{g},\mathbf{a},\bar{\pi},M)\tilde{R}_{min}\Big\}\end{aligned}$$

□

Lemma 5. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. For all $(\mathbf{s},\mathbf{g},\mathbf{a}) \in \mathcal{S} \times \mathcal{G}_r \times \mathcal{A}$, where $\mathcal{G}_R = \{\mathbf{g} \,|\, \forall \mathbf{g} \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}$ is reachable from $\mathbf{s}' = p(\mathbf{s},\mathbf{a})\}$ is the set of joint goals reachable by the non-terminated agents $\tilde{I} = \{i \,|\, \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$ and $m = |\tilde{I}|$, then*

$$mR_{min}D \leq \bar{Q}^*_M(\mathbf{s},\mathbf{g},\mathbf{a}) \leq mR_{max}D$$

Proof. ... □

Lemma 6. *Let $\bar{\pi}^*$ be the optimal world joint policy associated with a task $M \in \mathcal{M}$. Let $\bar{R}^i_M$ be the extended reward function for agent i for the task M. For all $i \in I$ and $(\mathbf{s}, g^i, \mathbf{a}) \in \mathcal{S} \times \mathcal{G}^i \times \mathcal{A}$, we have*

$$R_{min}(D-1) \leq \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] \leq R_{max}(D-1)$$

where $\mathbf{s}_0 = \mathbf{s}$ and $\mathbf{a}_0 = \mathbf{a}$.

Proof. ... □

Lemma 7. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ let $\mathcal{G}_R = \{\mathbf{g} \,|\, \forall \mathbf{g} \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}$ is reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$ be the set of joint goals reachable by the non-terminated agents $\tilde{I} = \{i \,|\, \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$. Then,*

$$Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}_R} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$$

Proof. Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$ and $\bar{\pi}^*$ be an associated deterministic optimal joint world policy. Let $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Let $\mathcal{G}_R = \{\mathbf{g} \,|\, \forall \mathbf{g} \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}$ is reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$ be the set of joint goals reachable by the non-terminated agents $\tilde{I} = \{i \,|\, \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$. Let $\mathbf{g}_r \in \mathcal{G}_r$ and $\mathbf{g}_u \in \mathcal{G} \setminus \mathcal{G}_r$. Let $m = |\tilde{I}|$.

Let

$$\begin{aligned}\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i,'} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1})] + P(s^i_{T^i+1} = g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M)\hat{R}^i_M(g^i) \\ + P(s^i_{T^i+1} \neq g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M)\tilde{R}_{min}\end{aligned}$$

.

By Lemma 4 we have $\bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \sum_{i \in \tilde{I}} \bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$.

We first consider $\mathbf{g}_r$. Let $G^{i,r}_{T^i-1} = \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i_r, \mathbf{a}_t, \mathbf{s}_{t+1})]$ and let $P^{i,r} = P(s^i_{T^i+1} = g^i_r | \mathbf{s}, \mathbf{g}_r, \mathbf{a}, \bar{\pi}^*, M)$.

Since $\mathbf{g}^{\tilde{I}}_r$ is reachable, by mastery $\sum_{i \in \tilde{I}} P^{i,r} = m$ where $m = |\tilde{I}|$. Thus, $P^{i,r} = 1$ for all $i \in \tilde{I}$. From this we get,

$$\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) = \bar{R}^i_M(\mathbf{s}, g^i_r, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i,'} \notin \mathcal{G}^i} + G^{i,r}_{T^i-1} + \hat{R}^i_M(g^i_r)$$

We consider two cases.
Case 1: $s^{i,'} \in \mathcal{G}^i$
$G^{i,r}_{T^i-1} = 0$ since $T^i = 0$ if $s^i \neq s^{i,'} \in \mathcal{G}^i$. Therefore, $\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) = \hat{R}^i_M(g^i_r)$ which means $\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \geq R_{min}$.

Case 2: $s^{i,'} \notin \mathcal{G}^i$

$$\begin{aligned}\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) &= \bar{R}^i_M(\mathbf{s}, g^i_r, \mathbf{a}, \mathbf{s}') + G^{i,r}_{T^i-1} + \hat{R}^i_M(g^i_r) \\ &= \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i_r, \mathbf{a}, \mathbf{s}_{t+1})] + \hat{R}^i_M(g^i_r)\end{aligned}$$

$T^i - 1 \leq D - 1$ since D is the maximum length of any trajectory. Thus, $\mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i_r, \mathbf{a}, \mathbf{s}_{t+1})] \geq R_{min}(D-1)$. Since $\hat{R}^i_M(g^i_r) \geq R_{min}$ we get $\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \geq DR_{min}$.

Combining case 1 and case 2 we get

$$\begin{aligned}&\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \geq DR_{min} \\ \implies &\sum_{i \in \tilde{I}} \bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \geq mDR_{min} \\ \implies &\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \geq mDR_{min}\end{aligned}$$

Next we consider $\mathbf{g}_u$. Let $G^{i,u}_{T^i-1} = \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i_u, \mathbf{a}_t, \mathbf{s}_{t+1})]$ and $P^{i,u} = P(s^i_{T^i+1} = g^i_r | \mathbf{s}, \mathbf{g}_u, \mathbf{a}, \bar{\pi}^*, M)$.

Since $\bar{\pi}^*$ is deterministic then $P^{i,u} \in \{0, 1\}$.

Since $\mathbf{g}_u^{\tilde{I}}$ is unreachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})$ then there exists $J \subseteq \tilde{I}$ where $P^{i,r} = 0$ for all $i \in J$ and $J \neq \{\}$. That is, J represents the non-terminated agents that will reach the wrong goal.

For all $i \in \tilde{I} \setminus J$, we have $\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) = \bar{R}^i_M(\mathbf{s}, g^i_u, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i,\prime} \notin \mathcal{G}^i} + G^{i,u}_{T^i-1} + \hat{R}^i_M(g^i_u)$ which implies $\bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq DR_{max}$ similarly to before.

Let $j \in J$. Then, $\bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) = \hat{R}^j_M(\mathbf{s}, g^i_u, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{j,\prime} \notin \mathcal{G}^j} + G^{j,u}_{T^j-1} + \tilde{R}_{min}$. We consider two cases.

Case 1: $s^{j,'} \in \mathcal{G}^j$

$$\begin{aligned}
&\bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) = \tilde{R}_{min} \\
\implies &\bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq n((R_{min} - R_{max})D + R_{min})
\end{aligned}$$

Case 2: $s^{j,'} \notin \mathcal{G}^j$

$$\begin{aligned}
\bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &= \bar{R}^j_M(\mathbf{s}, g^j_u, \mathbf{a}, \mathbf{s}') + G^{j,u}_{T^j-1} + \tilde{R}_{min} \\
&= \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{T^j-1} \bar{R}^j_M(\mathbf{s}_t, g^j_u, \mathbf{a}_t, \mathbf{s}_{t+1})] + \tilde{R}_{min} \\
\implies \bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq (D-1)R_{max} + \tilde{R}_{min}
\end{aligned}$$

Combining case 1 and case 2, we have $\bar{Q}^{j,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq (D-1)R_{max} + \tilde{R}_{min}$. Next we establish the upper limit on $\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a})$.

$$\begin{aligned}
&\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) = \sum_{i \in \tilde{I} \setminus J} \bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) + \sum_{i \in J} \bar{Q}^{i,*}_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \\
\implies &\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq (m-l)DR_{max} + l((D-1)R_{max} + \tilde{R}_{min}) && \text{where } l = |J| \\
\implies &\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq (mD-l)R_{max} + l\tilde{R}_{min} \\
\implies &\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq (mD-l)R_{max} + l(n(R_{min} - R_{max})D + R_{min}) \\
\implies &\bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \leq (mD - l - lnD)R_{max} + ln(D+1)R_{min}
\end{aligned}$$

Next we show $mDR_{min} > (mD - l - lnD)R_{max} + ln(D+1)R_{min}$.

$$\begin{aligned}
m &\leq ln && \text{since } m \leq n \text{ and } l \geq 1 \\
\implies mD - lnD &\leq 0 && \text{since } D > 0 \\
\implies mD - lnD - l &< 0 && \text{since } -l < 0 \\
\implies (mD - lnD - l)R_{max} &\leq 0 && \text{since } R_{max} \geq 0
\end{aligned}$$

$$\begin{aligned} & mD \leq lnD \\ \implies & mD < lnD + ln && \text{since } ln > 0 \\ \implies & mDR_{min} > ln(D+1)R_{min} && \text{since } R_{min} < 0 \\ \implies & mDR_{min} > (mD - lnD - l)R_{max} && \implies \bar{Q}^*_M(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) > \bar{Q}^*_M(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \end{aligned}$$

From this we get, for all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in (\mathcal{S} \setminus \mathcal{G}) \times \mathcal{G} \times \mathcal{A}$

$$\begin{aligned} Q^*_M(\mathbf{s}, \mathbf{a}) &= \max_{\mathbf{g} \in \mathcal{G}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) && \text{by Theorem 1} \\ &= \max\{\max_{\mathbf{g} \in \mathcal{G}_r} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}), \max_{\mathbf{g} \in \mathcal{G} \setminus \mathcal{G}_r} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})\} \\ &= \max_{\mathbf{g} \in \mathcal{G}_r} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) \end{aligned}$$

In the case of $(\mathbf{s}, \mathbf{a}) \in \mathcal{G} \times \mathcal{A}$ we get $\bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = 0$ for all $\mathbf{g} \in \mathcal{G}$, which leads to $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}_r} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = 0$.

Thus, for all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$ we have $Q^*_M(\mathbf{s}, \mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}_r} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$.

□

Theorem 3. *Let $\bar{Q}^*_M$ and $\bar{Q}^*_{M_B}$ be the optimal world value functions for the task $M \in \mathcal{M}$ and the background CM-MDP M_B, and for all $i \in I$. Let $\{\hat{R}^i_M\}_{i \in I}$ be the set of agent terminal reward functions for the task M. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}_r^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$, we have*

$$\bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^*_{M_B}(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} \hat{R}^i_M(g^i) \mathbb{1}_{s^i \notin \mathcal{G}^i}$$

Proof. Let $\{\hat{R}^i_M\}_{i \in I}$ be the extended reward functions for the task M and the background CM-MDP, M_B. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}_r^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$ with $\tilde{I} = \{i$ for all $i \in I$ where $\mathbf{s}^i \notin \mathcal{G}^i\}$.

By mastery, there exists $\bar{\pi}^*_M(\mathbf{s}, \mathbf{a}) \in \arg\max_{\mathbf{a} \in \mathcal{A}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$ where $\bar{\pi}^*_M \in \arg\max_{\bar{\pi}} \sum_{i \in I} p(\mathbf{s}^i_{T^i+1} = \mathbf{g}^i | \mathbf{s}, \mathbf{a}, \bar{\pi}, M)$. Let $\bar{\pi}^*_M(\mathbf{s}, \mathbf{a}) \in \arg\max_{\mathbf{a} \in \mathcal{A}} \bar{Q}^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a})$ be an optimal joint world policy for the task M that satisfies $\bar{\pi}^*_M \in \arg\max_{\bar{\pi}} \sum_{i \in I} p(\mathbf{s}^i_{T^i+1} = \mathbf{g}^i | \mathbf{s}, \mathbf{a}, \bar{\pi}, M)$.

Similarly, let $\bar{\pi}^*_B(\mathbf{s}, \mathbf{a}) \in \arg\max_{\mathbf{a} \in \mathcal{A}} \bar{Q}^*_B(\mathbf{s}, \mathbf{g}, \mathbf{a})$ be an optimal joint world policy for the background CM-MDP that satisfies $\bar{\pi}^*_B \in \arg\max_{\bar{\pi}} \sum_{i \in I} p(\mathbf{s}^i_{T^i+1} = \mathbf{g}^i | \mathbf{s}, \mathbf{a}, \bar{\pi}, B)$.

My lemma 4, we have

$$\begin{aligned} Q^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = & \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') \mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] \\ & + P(s^i_{T^i+1} = g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) \hat{R}^i_M(g^i) \\ & + P(s^i_{T^i+1} \neq g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) \tilde{R}_{min} \Big\} \end{aligned}$$

Since $\mathbf{g}^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})$, by mastery

$$P(s^i_{T^i+1} = g^i \,|\, \mathbf{s}, \mathbf{g}, \mathbf{a}, \bar{\pi}, M) = 1 \text{ for all } i \in \tilde{I}.$$

Therefore

$$Q^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_M(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] + \hat{R}^i_M(g^i) \Big\}.$$

$\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}') = \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')$ for all $\mathbf{s}^{i'} \notin G^i$.
$\bar{R}^i_M(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} = \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i}$ for all $\mathbf{s}^{i'} \in \mathcal{S}^i$.

It follows that

$$\begin{aligned} Q^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) =& \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_B(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] + \hat{R}^i_M(g^i) \Big\} \\ =& \max_{\bar{\pi}} \Big\{ \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} + \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_B(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] + \hat{R}^i_M(g^i) \Big\} \Big\} \\ =& \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} \Big\} + \max_{\bar{\pi}} \Big\{ \sum_{i \in \tilde{I}} \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_B(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] \Big\} + \sum_{i \in \tilde{I}} \hat{R}^i_M(g^i) \end{aligned}$$

Similarly

$$\begin{aligned} Q^*_B(\mathbf{s}, \mathbf{g}, \mathbf{a}) =& \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} \Big\} + \max_{\bar{\pi}} \Big\{ \sum_{i \in \tilde{I}} \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_B(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] \Big\} + \sum_{i \in \tilde{I}} \hat{R}^i_B(g^i) \\ =& \sum_{i \in \tilde{I}} \Big\{ \bar{R}^i_B(\mathbf{s}, g^i, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{i\prime} \notin \mathcal{G}^i} \Big\} + \max_{\bar{\pi}} \Big\{ \sum_{i \in \tilde{I}} \mathbb{E}^{\bar{\pi}} \Big[\sum_{t=0}^{T^i - 1} \bar{R}^i_B(\mathbf{s}_t, g^i, \mathbf{a}_t, \mathbf{s}_{t+1}) \Big] \Big\} \end{aligned}$$

since $\hat{R}^i_B(g^i) = 0$ for all $(i, g^i) \in I \times G^i$.

Therefore

$$\begin{aligned} Q^*_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) =& Q^*_B(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in \tilde{I}} \hat{R}^i_M(g^i) \\ =& Q^*_B(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} \hat{R}^i_M(g^i)\mathbb{1}_{s^i \notin \mathcal{G}^i}. \end{aligned}$$

□

Theorem 4. *Let $\bar{Q}^*_{M_1}$ be the optimal world value function for a task $M_1 \in \mathcal{M}$. Let $\{\hat{R}^i_{M_1}\}_{i \in I}$ and $\{\hat{R}^i_{M_2}\}_{i \in I}$ be the sets of agent terminal reward functions for the tasks M_1 and $M_2 \in \mathcal{M}$. For all $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}_r$ reachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})\}$, we have*

$$\bar{Q}^*_{M_2}(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^*_{M_1}(\mathbf{s}, \mathbf{g}, \mathbf{a}) + \sum_{i \in I} (\hat{R}^i_{M_2}(g^i) - \hat{R}^i_{M_1}(g^i))\mathbb{1}_{s^i \notin \mathcal{G}^i}$$

Proof. For all $(\mathbf{s},\mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and for all $\mathbf{g} \in \{\mathbf{g}_r \,|\, \forall \mathbf{g}_r \in \mathcal{G}$ where $\mathbf{g}_r^{\tilde{I}}$ reachable from $\mathbf{s}' = p(\mathbf{s},\mathbf{a})\}$ and where $\tilde{I} = \{i \,|\, \forall i \in I$ where $\mathbf{s}^i \notin G^i\}$, by theorem **??**, we have

$$\begin{aligned} Q^*_B(\mathbf{s},\mathbf{g},\mathbf{a}) =& Q^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) - \sum_{i \in \tilde{I}} \hat{R}^i_{M_1}(g^i)\mathbb{1}_{s^i \notin \mathcal{G}^i} \\ =& Q^*_{M_2}(\mathbf{s},\mathbf{g},\mathbf{a}) - \sum_{i \in \tilde{I}} \hat{R}^i_{M_2}(g^i)\mathbb{1}_{s^i \notin \mathcal{G}^i}. \end{aligned}$$

It follows that

$$\begin{aligned} Q^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) - \sum_{i \in \tilde{I}} \hat{R}^i_{M_1}(g^i)\mathbb{1}_{s^i \notin \mathcal{G}^i} = Q^*_{M_2}(\mathbf{s},\mathbf{g},\mathbf{a}) - \sum_{i \in \tilde{I}} \hat{R}^i_{M_2}(g^i)\mathbb{1}_{s^i \notin \mathcal{G}^i}, \text{ therefore} \\ \bar{Q}^*_{M_2}(\mathbf{s},\mathbf{g},\mathbf{a}) = \bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) + \sum_{i \in I}(\hat{R}^i_{M_2}(g^i) - \hat{R}^i_{M_1}(g^i))\mathbb{1}_{s^i \notin \mathcal{G}^i}. \end{aligned}$$

□

Theorem 5. *Let $\bar{Q}^*_{M_1}$ be the optimal world value function for a task $M_1 \in \mathcal{M}$. Let $\{\hat{R}^i_{M_1}\}_{i \in I}$ and $\{\hat{R}^i_{M_2}\}_{i \in I}$ be the sets of agent terminal reward functions for the tasks M_1 and $M_2 \in \mathcal{M}$. For all $(\mathbf{s},\mathbf{a}) \in \mathcal{S} \times \mathcal{A}$*

$$Q^*_{M_2}(\mathbf{s},\mathbf{a}) = \max_{\mathbf{g} \in \mathcal{G}}\{\bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) + \sum_{i \in I}(\hat{R}^i_{M_2}(g^i) - \hat{R}^i_{M_1}(g^i))\mathbb{1}_{s^i \in \mathcal{G}^i}\}$$

Proof. Let $M_1 \in \mathcal{M}$ and $M_2 \in \mathcal{M}$. Let $\bar{Q}^*_{M_1}$ be the optimal world value function for M_1 and let $\bar{\pi}^*_1$ be an associated deterministic optimal world joint policy. Let $\{\hat{R}^i_{M_1}\}_{i \in I}$ and $\{\hat{R}^i_{M_2}\}_{i \in I}$ be the sets of agent terminal reward functions for the tasks M_1 and M_2.

Let $(\mathbf{s},\mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Let $\mathcal{G}_R = \{\mathbf{g} \,|\, \forall \mathbf{g} \in \mathcal{G}$ where $\mathbf{g}^{\tilde{I}}$ is reachable from $\mathbf{s}' = p(\mathbf{s},\mathbf{a})\}$ be the set of joint goals reachable by the non-terminated agents $\tilde{I} = \{i \,|\, \forall i \in I$ where $s^i \notin \mathcal{G}^i\}$ and let $m = |\tilde{I}|$. Let $\mathbf{g}_r \in \mathcal{G}_r$ and let $\mathbf{g}_u \in \mathcal{G} \setminus \mathcal{G}_r$. Let $\tilde{\bar{Q}}^*_{M_2}(\mathbf{s},\mathbf{g},\mathbf{a}) = \bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g},\mathbf{a}) + \sum\limits_{i \in I}(\hat{R}^i_{M_2}(\mathbf{g}^i) - \hat{R}^i_{M_1}(\mathbf{g}^i))\mathbb{1}_{s^i \notin \mathcal{G}^i}$ for all $i \in I$ and for all $\mathbf{g} \in \mathcal{G}$.

We first consider $\mathbf{g}_r$.

$$\begin{aligned} & \bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g}_r,\mathbf{a}) + \sum_{i \in \tilde{I}}(\hat{R}^i_{M_2}(\mathbf{g}^i_r) - \hat{R}^i_{M_1}(\mathbf{g}^i_r))\mathbb{1}_{s^i \notin \mathcal{G}^i} = \bar{Q}^*_{M_2}(\mathbf{s},\mathbf{g}_r,\mathbf{a}) && \text{by Theorem 4} \\ \implies & \bar{Q}^*_{M_1}(\mathbf{s},\mathbf{g}_r,\mathbf{a}) + \sum_{i \in \tilde{I}}(\hat{R}^i_{M_2}(\mathbf{g}^i_r) - \hat{R}^i_{M_1}(\mathbf{g}^i_r))\mathbb{1}_{s^i \notin \mathcal{G}^i} \geq mR_{min}D && \text{by Lemma 5} \end{aligned}$$

Next we consider $\mathbf{g}_u$. For all $i \in \tilde{I}$ let $\tilde{\bar{Q}}^{i,*}_{M_2}(\mathbf{s},\mathbf{g}_u,\mathbf{a}) = \bar{Q}^{i,*}_{M_2}(\mathbf{s},\mathbf{g}_u,\mathbf{a}) + \hat{R}^i_{M_2}(g^i_u) - \hat{R}^i_{M_1}(g^i_u)$. Therefore, $\tilde{\bar{Q}}^*_{M_2}(\mathbf{s},\mathbf{g}_u,\mathbf{a}) = \sum\limits_{i \in \tilde{I}} \tilde{\bar{Q}}^{i,*}_{M_2}(\mathbf{s},\mathbf{g}_u,\mathbf{a})$.

Let $G^{i,u}_{T^i-1} = \mathbb{E}^{\bar{\pi}^*}[\sum\limits_{t=0}^{T^i-1} \bar{R}^i_M(\mathbf{s}_t, g^i_u, \mathbf{a}_t, \mathbf{s}_{t+1})]$ and $P^{i,u} = P(s^i_{T^i+1} = g^i_r | \mathbf{s}, \mathbf{g}_u, \mathbf{a}, \bar{\pi}^*, M)$.

By Lemma 3 and Lemma 4 we get

$$\begin{aligned} \tilde{\bar{Q}}^{i,*}_{M_2}(\mathbf{s},\mathbf{g}_u,\mathbf{a}) =& \bar{R}^i_{M_1}(\mathbf{s},\mathbf{g}_u,\mathbf{a},\mathbf{s}')\mathbb{1}_{s^{i,'} \notin \mathcal{G}^i} + G^{i,u}_{T^i-1} + P^{i,u}\hat{R}^i_{M_1}(g^i_u) + (1 - P^{i,u})\tilde{R}_{min} \\ &+ \hat{R}^i_{M_2}(g^i_u) - \hat{R}^i_{M_1}(g^i_u) \end{aligned}$$

Since $\bar{\pi}^*_{M_2}$ is deterministic then $P^{i,u} \in \{0, 1\}$. Since $\mathbf{g}_u^{\tilde{I}}$ is unreachable from $\mathbf{s}' = p(\mathbf{s}, \mathbf{a})$ then there exists $J \subseteq \tilde{I}$ where $P^{i,u} = 0$ for all $i \in J$ and $J \neq \{\}$. Let $j \in J$ and let $j' \in \tilde{I} \setminus J$.

$$\begin{aligned}
\tilde{\tilde{Q}}^{j',*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &= \bar{R}^{j'}_{M_1}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{j'} \notin \mathcal{G}^{j'}} + G^{j',u}_{T^{j'}-1} + \hat{R}^{j'}_{M_1}(g^{j'}_u) \\
&\quad + \hat{R}^{j'}_{M_2}(g^{j'}_u) - \hat{R}^{j'}_{M_1}(g^{j'}_u) \\
&= \bar{R}^{j'}_{M_1}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{j'} \notin \mathcal{G}^{j'}} + G^{j',u}_{T^{j'}-1} + \hat{R}^{j'}_{M_2}(g^{j'}_u) \\
&= \begin{cases} \bar{R}^{j'}_{M_1}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}, \mathbf{s}') + G^{j',u}_{T^{j'}-1} + \hat{R}^{j'}_{M_2}(g^{j'}_u) & \text{if } s^{j'} \notin \mathcal{G}^{j'} \\ \hat{R}^{j'}_{M_2}(g^{j'}_u) & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathbb{E}^{\bar{\pi}^*}_{M_1}[\sum_{t=0}^{T^{j'}-1} \bar{R}^{j'}_{M_1}(\mathbf{s}_t, g^{j'}_u, \mathbf{a}_t, \mathbf{s}_{t+1})] + \hat{R}^{j'}_{M_2}(g^{j'}_u) & \text{if } s^{j'} \notin \mathcal{G}^{j'} \\ \hat{R}^{j'}_{M_2}(g^{j'}_u) & \text{otherwise} \end{cases} \\
\implies \tilde{\tilde{Q}}^{j',*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq \begin{cases} R_{max}(D-1) + R_{max} & \text{if } s^{j'} \notin \mathcal{G}^{j'} \\ R_{max} & \text{otherwise} \end{cases} \qquad \text{by Lemma 6} \\
\implies \tilde{\tilde{Q}}^{j',*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq R_{max}D \\
\implies \sum_{i \in \tilde{I} \setminus J} \tilde{\tilde{Q}}^{j',*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq (m-l)R_{max}D
\end{aligned}$$

$$\begin{aligned}
\tilde{\tilde{Q}}^{j,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &= \bar{R}^{j}_{M_1}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}, \mathbf{s}')\mathbb{1}_{s^{j} \notin \mathcal{G}^{j}} + G^{j,u}_{T^{j}-1} + \tilde{R}_{min} \\
&\quad + \hat{R}^{j}_{M_2}(g^{j}_u) - \hat{R}^{j}_{M_1}(g^{j}_u) \\
&= \begin{cases} \mathbb{E}^{\bar{\pi}^*}_{M_1}[\sum_{t=0}^{T^{j}-1} \bar{R}^{j}_{M_1}(\mathbf{s}_t, g^{j}_u, \mathbf{a}_t, \mathbf{s}_{t+1})] + \tilde{R}_{min} & \\ \quad + \hat{R}^{j}_{M_2}(g^{j}_u) - \hat{R}^{j}_{M_1}(g^{j}_u) & \text{if } s^{j} \notin \mathcal{G}^{j} \\ \tilde{R}_{min} + \hat{R}^{j}_{M_2}(g^{j}_u) - \hat{R}^{j}_{M_1}(g^{j}_u) & \text{otherwise} \end{cases} \\
\implies \tilde{\tilde{Q}}^{j,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq \begin{cases} (R_{max})(D-1) + \tilde{R}_{min} + R_{max} - R_{min} & \text{if } s^{j} \notin \mathcal{G}^{j} \\ \tilde{R}_{min} + R_{max} - R_{min} & \text{otherwise} \end{cases} \qquad \text{similarly to before} \\
\implies \tilde{\tilde{Q}}^{j,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq R_{max}D - R_{min} + \tilde{R}_{min} \\
\implies \sum_{i \in J} \tilde{\tilde{Q}}^{i,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq l(R_{max}D - R_{min} + \tilde{R}_{min})
\end{aligned}$$

$$\begin{aligned}
\tilde{\tilde{Q}}^{*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &= \sum_{i \in \tilde{I}} \tilde{\tilde{Q}}^{i,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \\
&= \sum_{i \in \tilde{I} \setminus J} \tilde{\tilde{Q}}^{i,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) + \sum_{i \in J} \tilde{\tilde{Q}}^{i,*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) \\
\implies \tilde{\tilde{Q}}^{*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq (m-l)R_{max}D + l(R_{max}DR_{min} + \tilde{R}_{min}) \\
\implies \tilde{\tilde{Q}}^{*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &\leq (m-ln)DR_{max} + (ln(D+1) - l)R_{min}
\end{aligned}$$

Now we show that $(m - ln)DR_{max} + (ln(D+1) - l)R_{min} \leq mDR_{min}$ which implies $\tilde{\tilde{Q}}^{*}_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) < \tilde{\tilde{Q}}^{*}_{M_2}(\mathbf{s}, \mathbf{g}_r, \mathbf{a})$.

$$\begin{aligned} m - ln &\leq 0 && \text{since } m \leq n \text{ and } l \geq 1 \\ \implies (m - ln)DR_{max} &\leq 0 \end{aligned}$$

$$\begin{aligned} ln &\geq m \\ \implies lnD &\geq mD \\ \implies lnD + ln - l &\geq mD && \text{since } ln \geq l \\ \implies R_{min}(lnD + ln - l - mD) &< 0 && \text{since } R_{min} < 0 \\ \implies (ln(D+1) - l)R_{min} &< mDR_{min} \\ \implies (m - ln)DR_{max} + (ln(D+1) - l)R_{min} &< mDR_{min} && \text{since } (m - ln)DR_{max} \leq 0 \\ \implies \tilde{\bar{Q}}^*_{M_2}(\mathbf{s}, \mathbf{g}_u, \mathbf{a}) &< \tilde{\bar{Q}}^*_{M_2}(\mathbf{s}, \mathbf{g}_r, \mathbf{a}) \end{aligned}$$

Therefore,

$$\begin{aligned} \max_{\mathbf{g} \in \mathcal{G}} \{\tilde{\bar{Q}}^*_{M_2}(\mathbf{s}, \mathbf{g}, \mathbf{a})\} &= \max_{\mathbf{g} \in \mathcal{G}_r} \{\tilde{\bar{Q}}^*_{M_2}(\mathbf{s}, \mathbf{g}, \mathbf{a})\} \\ &= \max_{\mathbf{g} \in \mathcal{G}_r} \{\bar{Q}^*_{M_2}(\mathbf{s}, \mathbf{g}, \mathbf{a})\} && \text{by Theorem 4} \\ &= Q^*_{M_2}(\mathbf{s}, \mathbf{a}) && \text{by Lemma 1} \end{aligned}$$

□

A.1.3 Symmetry

Theorem 6. *For all $i \in I$. Let $g^{i,T} \in \mathcal{G}^i$. For all $(s^{-i}, g^{-i}) \in \mathcal{S}^{-i} \times \mathcal{G}^{-i}$, for all $(s^i, g^i) \in \mathcal{G}^i \times \mathcal{G}^i$, for all $\mathbf{a} \in \mathcal{A}$ we have*

$$\bar{Q}^*(\mathbf{s}, < g^{-i}, g^{i,T} >, \mathbf{a}) = \bar{Q}^*(\mathbf{s}, < g^{-i}, g^i >, \mathbf{a})$$

Proof. For all $i \in I$, for all $(s^i) \in \mathcal{G}^i$, for all $(s^{-i}, g^{-i}, \mathbf{a}) \in \mathcal{S}^{-i} \times \mathcal{G}^{-i} \times \mathcal{A}$, and for all $(g^{i,T}, g^i) \in \mathcal{G}^i \times \mathcal{G}^i$:

$$\begin{aligned} \bar{Q}^*(\mathbf{s}, \langle g^{i,T}, g^{-i} \rangle, \mathbf{a}) &= \max_{\bar{\pi}} \{\bar{Q}^{\bar{\pi}}(\mathbf{s}, \langle g^{i,T}, g^{-i} \rangle, \mathbf{a})\} \\ &= \max_{\bar{\pi}} \{\sum_{j \in \tilde{I}} \{\bar{Q}^{j,\bar{\pi}}(\mathbf{s}, \langle g^{i,T}, g^{-i} \rangle, \mathbf{a})\}\} \\ &= \max_{\bar{\pi}} \{\sum_{j \in \tilde{I}} \{\mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^j(\mathbf{s}, g^j, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^j(\mathbf{s}_t, g^j, \mathbf{a}, \mathbf{s}_{t+1})]]\}\} \end{aligned}$$

Similarly,

$$\bar{Q}^*(\mathbf{s}, \langle g^i, g^{-i} \rangle, \mathbf{a}) = \max_{\bar{\pi}} \{\sum_{j \in \tilde{I}} \{\mathbb{E}^{\bar{\pi}}_{\mathbf{s}'}[\bar{R}^j(\mathbf{s}, g^j, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}}[\sum_{t=0}^{\infty} \bar{R}^j(\mathbf{s}_t, g^j, \mathbf{a}, \mathbf{s}_{t+1})]]\}\}$$

Since $s^i \in \mathcal{G}^i$ then $i \notin \tilde{I}$. Therefore,

$$\bar{Q}^*(\mathbf{s}, \langle g^{i,T}, g^{-i} \rangle, \mathbf{a}) = \bar{Q}^*(\mathbf{s}, \langle g^i, g^{-i} \rangle, \mathbf{a})$$

□

Theorem 7. *Let $\bar{Q}^*_M$ be the optimal world value function for a task $M \in \mathcal{M}$. $\bar{Q}^*_M$ is invariant to any permutation $Perm$ applied to the joint state, joint action and joint goal. That is, for all permutations $Perm$ and for all $(\mathbf{s}, \mathbf{g}, \mathbf{a}) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}$ we have*

$$\bar{Q}^*(\mathbf{s}, \mathbf{g}, \mathbf{a}) = \bar{Q}^*(Perm(\mathbf{s}), Perm(\mathbf{g}), Perm(\mathbf{a}))$$

Proof. Let $(\mathbf{s}, \mathbf{a}, \mathbf{g}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{G}$. Since the agents are homogeneous we have $\mathcal{P}(Perm(\mathbf{s}), Perm(\mathbf{a})) = \mathcal{P}(\mathbf{s}')$ and $\bar{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') = \bar{R}(Perm(\mathbf{s}), Perm(\mathbf{g}), Perm(\mathbf{a}), Perm(\mathbf{s}'))$. Thus,

$$\begin{aligned}
\bar{Q}^*(\mathbf{s}, \mathbf{g}, \mathbf{a}) &= \mathbb{E}^{\bar{\pi}^*}_{\mathbf{s}'}[\bar{R}(\mathbf{s}, \mathbf{g}, \mathbf{a}, \mathbf{s}') + \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{\infty} \bar{R}(\mathbf{s}_t, \mathbf{g}, \mathbf{a}_t, \mathbf{s}_{t+1})]] \\
&= \mathbb{E}^{\bar{\pi}^*}_{Perm(\mathbf{s}')}[\bar{R}(Perm(\mathbf{s}), Perm(\mathbf{g}), Perm(\mathbf{a}), Perm(\mathbf{s}')) \\
&+ \mathbb{E}^{\bar{\pi}^*}[\sum_{t=0}^{\infty} \bar{R}(Perm(\mathbf{s}_t), Perm(\mathbf{g}), Perm(\mathbf{a}_t), Perm(\mathbf{s}_{t+1}))]] \\
&= \bar{Q}^*(Perm(\mathbf{s}), Perm(\mathbf{g}), Perm(\mathbf{a}))
\end{aligned}$$

□

A.2 Goal-Oriented Learning

Algorithm 1: Goal-Oriented Learning

```
Input      : Task $M \in \mathcal{M}$ and joint goal library $\bar{\mathcal{D}}_G$
Initialise: World value function $\bar{Q}_M$
foreach episode do
    Observe an initial joint-state $\mathbf{s} \in \mathcal{S}$
    Sample a joint goal $\mathbf{g} \in \mathcal{G}$
    while episode is not done do
        $\mathbf{a} \leftarrow \begin{cases} \arg\max_{\mathbf{a} \in \mathcal{A}} \bar{Q}_M(\mathbf{s}, \mathbf{g}, \mathbf{a}) & \text{probability } 1 - \varepsilon \\ \text{a random joint-action} & \text{probability } \varepsilon \end{cases}$
        Take joint-action $\mathbf{a}$, observe agent rewards $\{r^i\}_{i \in I}$ and next joint-state $\mathbf{s}'$
        $\mathcal{D}'_G \leftarrow$ augment according to Theorem 6 and Theorem 7
        for $\mathbf{g}' \in \mathcal{D}'_G$ do
            for $i \in I$ do
                if $s^i \neq s^{i'} \in \mathcal{G} \setminus \{\mathbf{g}'\}$ then
                    $r^{i'} \leftarrow \tilde{R}_{min}$
                else
                    $r^{i'} \leftarrow r^i$

            $r' \leftarrow \sum_{i \in I} r^{i'}$
            $\bar{Q}_M(\mathbf{s}, \mathbf{g}', \mathbf{a}) \xleftarrow{\alpha} \left[ r' + \max_{\mathbf{a}' \in \mathcal{A}} \bar{Q}_M(\mathbf{s}', \mathbf{g}', \mathbf{a}') \right] - \bar{Q}_M(\mathbf{s}, \mathbf{g}', \mathbf{a})$
        $\bar{s} \leftarrow \bar{s}'$
```

A.3 Ablation

In this experiment, we consider the same domain, but set all goals to be desirable. We perform an ablation to compare the effects of leveraging various symmetries on the sample efficiency of learning as a function of the number of goals. Sample efficiency is quantified by the number of learning steps required to optimally solve the task. We consider a task to be optimally solved if the current evaluated return is equal to the optimal return as computed by optimal multi-agent pathfinding. The algorithms we consider are GOL leveraging no symmetries, GOL only leveraging symmetry from homogeneity, GOL only leveraging symmetry from agent level termination and GOL that leverages all symmetries. Figure 4 demonstrates how leveraging symmetries drastically improves the sample efficiency of learning the WVF, which is to be expected since fewer values need to be learnt.

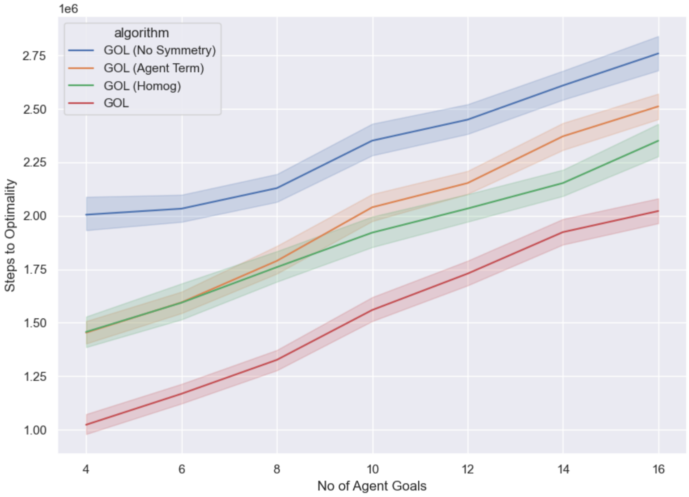


Figure 4: Ablation showing effect of leveraging various symmetries on steps required until optimal convergence of learning WVF. (Less is better) For each number of goals and variant of algorithm, 512 runs were executed with the mean plotted and 95% confidence intervals shown.