

# Synthesis of Communication Policies for Multi-Agent Systems Robust to Communication Restrictions

Saleh Soudijani, Rayna Dimitrova

CISPA Helmholtz Center for Information Security, Germany  
{saleh.soudijani, dimitrova}@cispa.de

## Abstract

We study stochastic multi-agent systems in which agents must cooperate to maximize the probability of achieving a common reach-avoid objective. In many applications, during the execution of the system, the communication between the agents can be constrained by restrictions on the bandwidth currently available for exchanging local-state information between the agents. In this paper, we propose a method for computing joint action and communication policies for the group of agents that aim to satisfy the communication restrictions as much as possible while achieving the optimal reach-avoid probability when communication is unconstrained. Our method synthesizes a pair of action and communication policies robust to restrictions on the number of agents allowed to communicate. To this end, we introduce a novel cost function that measures the amount of information exchanged beyond what the communication policy allows. We evaluate our approach experimentally on a range of benchmarks and demonstrate that it is capable of computing pairs of action and communication policies that satisfy the communication restrictions, if such exist.

## 1 Introduction

In cooperative multi-agent systems (MAS), individual agents are required to collaborate to achieve a joint task. One way to achieve this collaboration is to provide a centralized joint policy that the agents must adhere to. Typically, such policies need to correlate the actions of different agents. As a result, their successful execution requires communication between the agents to exchange, for example, local state information, and to coordinate their actions. However, in many real-world settings, agents have to operate in environments where communication could be restricted due to physical limitations, such as limited bandwidth or signal interference. As such restrictions can severely impact the coordination between agents and, hence, their performance, it is imperative that communication restrictions be considered in the design of joint policies for MAS. This requires devising policies that prescribe how the limited resources available for communication should be allocated.

In this paper, we focus on cooperative MAS with joint reach-avoid objectives (which require the agents to reach some target set of joint states while avoiding some unsafe states), *possibly operating under communication restrictions on the number of agents allowed to communicate*. We consider the setting where communicating agents exchange full current state information and jointly select actions. This requires policies that determine which subset of agents should communicate at the current state of system execution, which we term *communication policies*. Clearly, communication policies cannot depend on the full information about agents' local states, as this would defeat their purpose. In this work, we assume that communication policies can use some *public information* about the agents' states, which, in practice, could be very limited or even non-existent. A typical example is a system in which agents know the coarse regions in which other agents are located but not their precise locations. The rationale is that this public information is significantly less costly to communicate and changes less frequently.

We study the problem of synthesizing communication policies, together with joint policies that govern the agents' actions, which we call *action policies*. The challenge is that these two policies should be synthesized in tandem since the joint action policy should be adapted to the communication policy, requiring as little communication as possible beyond that allowed by the communication policy. To address this challenge, we introduce a cost function that, intuitively, measures the *information exchange between agents required by the action policy that goes beyond what is allowed by the communication policy*. We show that this cost function can be used in an upper bound on the performance loss when the action policy is executed under restricted communication and following the communication policy. Based on this bound, we propose a method for synthesizing pairs of action and communication policies that minimizes an over-approximation of the cost function and achieves optimal reach-avoid probability under unrestricted communication.<sup>1</sup>

**Related work.** Decision-theoretic models for MAS [Rizk *et al.*, 2018] such as decentralized MDPs (Dec-MDPs) and decentralized partially observable MDPs (Dec-POMDPs) [Goldman and Zilberstein, 2004] and their

<sup>1</sup>An extended version of this paper is available at: <https://arxiv.org/abs/2505.13311>

respective policy synthesis problems have been extensively studied. A key characteristic of these models is that agents cannot communicate. In contrast, in our setting, agents are allowed to communicate and exchange information about their independent local states and transitions, but this communication must be minimized relative to a communication policy.

A number of MAS models exist where communication is allowed but used sparsely to simplify the policy synthesis task. These include [Guestrin *et al.*, 2001], where a coordination graph representing the dependencies between the agents is given, and [Melo and Veloso, 2011], where the decentralized model is equipped with information about the states in which the agents need to interact. Other methods [Wu *et al.*, 2011] use online planning to use communication dynamically on demand. These approaches enable the synthesis of optimal policies that conform to given communication structures or minimize communication. On the other hand, our method synthesizes optimal action policies equipped with communication policies that make them robust to communication restrictions.

The closest to our work is [Karabag *et al.*, 2022], which proposes a technique for constructing joint policies for cooperative MAS that are robust to temporary or permanent loss of communication. In contrast, the policies we compute must be *robust to communication restrictions*, and thus benefit from *associated communication policies*. Thus, while [Karabag *et al.*, 2022] can use total correlation to synthesize policies minimizing dependency between the agents, we need to develop a cost function whose values depend on the sought communication policy. Similarly to [Karabag *et al.*, 2022], our cost function uses information-theoretic measures, but the challenge is to account for the unknown communication policy.

## 2 Preliminaries

In this section, we review some definitions and concepts.

For  $n \in \mathbb{N}$ , we define  $[n] := \{1, \dots, n\}$ . We denote the set of discrete probability distributions over a set  $X$  with  $\Delta(X)$ .

Markov decision processes (MDPs) provide a framework for modeling and analysis of sequential decision processes.

**Definition 1.** A Markov decision process (MDP) is a tuple  $M = (\mathcal{S}, \mathcal{A}, P, s_{init})$  where  $\mathcal{S}$  is a finite set of states,  $\mathcal{A}$  is a finite set of actions,  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is a partial transition probability function, and  $s_{init} \in \mathcal{S}$  is an initial state.

For simplicity we sometimes write  $P(s, a, s')$  instead of  $P(s, a)(s')$ , for  $s, s' \in \mathcal{S}$  and  $a \in \mathcal{A}$ . We denote with  $\mathcal{A}(s) := \{s \in \mathcal{S} \mid \exists s' \in \mathcal{S}. P(s, a, s') > 0\}$  the set of actions enabled in  $s \in \mathcal{S}$ . We assume that  $\mathcal{A}(s) \neq \emptyset$  for every  $s \in \mathcal{S}$ .

A *path* in an MDP  $M$  is a finite or infinite sequence  $\tau = s_0 a_1 s_1 \dots s_{t-1} a_t s_t, \dots$  of alternating states and actions such that  $P(s_t, a_{t+1}, s_{t+1}) > 0$  for all  $t \in \mathbb{N}$ .

A *policy* for an MDP  $M = (\mathcal{S}, \mathcal{A}, P, s_{init})$  is a function  $\pi : (\mathcal{S} \cdot \mathcal{A})^* \cdot \mathcal{S} \rightarrow \Delta(\mathcal{A})$ , that maps each finite path ending in a state to a probability distribution over actions and is such that if  $\pi(s_0 a_1 \dots s_t)(a) > 0$ , then  $a \in \mathcal{A}(s_t)$ . A policy is called *positional*, if its decisions depend solely on the current state. Formally, we can represent a positional policy  $\pi$  as a

function  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ . For simplicity, we write  $\pi(\tau, a)$  and  $\pi(s, a)$  instead of  $\pi(\tau)(a)$  and  $\pi(s)(a)$ , respectively.

**Definition 2.** A Markov chain is a triple  $C = (\mathcal{S}, P, s_{init})$  where  $\mathcal{S}$  is the set of states,  $P : \mathcal{S} \rightarrow \Delta(\mathcal{S})$  is the transition probability function, and  $s_{init} \in \mathcal{S}$  is the initial state.

Given an MDP  $M$ , a policy  $\pi$  for  $M$  induces an (potentially infinite-state) Markov chain. We denote this Markov chain with  $M_\pi$ , which is defined as  $M_\pi = ((\mathcal{S} \cdot \mathcal{A})^* \cdot \mathcal{S}, P_{M_\pi}, s_{init})$ , where for every  $\tau = s_0 a_1 \dots s_{t-1} a_t s_t \in (\mathcal{S} \cdot \mathcal{A})^* \cdot \mathcal{S}$ ,  $a \in \mathcal{A}$  and  $s \in \mathcal{S}$  we have that  $P_{M_\pi}(\tau, \tau \cdot a \cdot s) = \pi(\tau, a) \cdot P(s_t, a, s)$ . For a positional policy  $\pi$  for an MDP  $M$ , the induced Markov chain has a finite set of states. Formally,  $M_\pi = (\mathcal{S}, P_{M_\pi}, s_{init})$ , where for every  $s \in \mathcal{S}$  and  $s' \in \mathcal{S}$  we have  $P_{M_\pi}(s, s') = \sum_{a \in \mathcal{A}(s)} \pi(s, a) \cdot P(s, a, s')$ .

A Markov chain  $C = (\mathcal{S}, P, s_{init})$  can be seen as a sequence of discrete stochastic variables  $(S_t, t \in \mathbb{N})$ , which generates a stationary process  $S$  where  $\mathbb{P}(S_t = s)$  is the probability of the chain visiting state  $s \in \mathcal{S}$  at time  $t$ . The *occupancy measure of a state*  $s$  is  $\nu_s := \sum_{t=0}^{\infty} \mathbb{P}(S_t = s)$ .

Given a policy  $\pi$  for an MDP  $M = (\mathcal{S}, \mathcal{A}, P, s_{init})$ , we denote with  $\nu_{s,a}$  the *occupancy measure of the state-action pair*  $(s, a)$ , i.e., the expected number of times that action  $a$  is taken at state  $s$ , defined as  $\nu_{s,a} := \sum_{t=0}^{\infty} \mathbb{P}(S_t = s, A_t = a)$ . By definition, we have that  $\nu_{s,a} = \sum_{t=0}^{\infty} \mathbb{P}(S_t = s, A_t = a) = \sum_{t=0}^{\infty} \mathbb{P}(S_t = s) \cdot \mathbb{P}(A_t = a \mid S_t = s) = \pi(s, a) \cdot \nu_s$ .

### Entropy of Stochastic Processes

The entropy is a measure of uncertainty about the outcome of a random variable [Shannon and Weaver, 1949].

**Definition 3.** For a discrete random variable  $X$ , its support  $V$  defines a countable sample space from which  $X$  takes a value  $v \in V$  according to a probability mass function (pmf)  $p(v) := \mathbb{P}(X = v)$ . The entropy of  $X$  is defined as  $H(X) := -\sum_{v \in V} p(v) \log p(v)$ . By convention,  $0 \log 0 = 0$ .

The entropy is always non-negative. It vanishes for a deterministic  $X$  (i.e., if  $X$  is completely determined).

Let  $(X_1, X_2)$  be a pair of random variables with joint pmf  $p(v_1, v_2)$  and support  $V_1 \times V_2$ . The *joint entropy* of  $(X_1, X_2)$  is defined by  $H(X_1, X_2) := -\sum_{v_1 \in V_1} \sum_{v_2 \in V_2} p(v_1, v_2) \log p(v_1, v_2)$ .

The *conditional entropy* of a random variable  $X_1$  given  $X_2$  is defined as  $H(X_1 \mid X_2) := -\sum_{v_1 \in V_1} \sum_{v_2 \in V_2} p(v_1, v_2) \log p(v_1 \mid v_2)$ .

The joint and conditional entropy definitions extend to the collection of  $n$  random variables [Cover and Thomas, 2006].

The *entropy of a Markov chain*  $C = (\mathcal{S}, P, s_{init})$  is defined as the joint entropy over all random variables  $S_t$  for  $t \in \mathbb{N}$ . That is,  $H(C) := H(S_0, S_1, S_2, \dots) = \sum_{t=0}^{\infty} H(S_t \mid S_{t-1} \dots S_0)$ . The entropy of a Markov chain is in general infinite. The entropy of a Markov chain is finite if and only if it is absorbing [Biondi *et al.*, 2014]. In this paper, we restrict our analysis to absorbing Markov chains.

[Biondi *et al.*, 2014] showed that the entropy of a Markov chain can be characterized in terms of the occupancy measure of the states and their so-called *local entropy*. The local entropy  $L(s)$  of a state  $s$  in a Markov chain is the entropy of the probability distribution over the next states defined by  $P$ , formally,  $L(s) := H(S_{t+1} \mid S_t = s) =$

–  $\sum_{s' \in \mathcal{S}} P(s, s') \log P(s, s')$ . Then, as shown in [Biondi *et al.*, 2014], the entropy  $H(C)$  can be expressed as  $H(C) = \sum_{s \in \mathcal{S}} L(s) \cdot \nu_s$ , where  $\nu_s$  is the occupancy measure of  $s$ .

### Multi-Agent Markov Decision Processes

*Multi-agent Markov decision processes (MMDPs)* describe sequential decision-making tasks in which multiple agents select actions in order to collaboratively maximize a given common reward-based optimization criterion. A joint policy prescribes actions for all agents. During the execution of such a policy, all agents have access to the joint state of the system.

**Definition 4.** Formally, a Multi-agent Markov decision process (MMDP) is a tuple  $M = (N, \mathcal{S}, \mathcal{A}, P, s_{init})$  where:

- $N$  is the number of agents;
- $\mathcal{S} = \mathcal{S}^1 \times \mathcal{S}^2 \times \dots \times \mathcal{S}^N$  is a finite set of global states;
- $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2 \times \dots \times \mathcal{A}^N$  is a finite set of joint actions;
- $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is the joint transition probability function such that for every  $s = \langle s^1, \dots, s^N \rangle \in \mathcal{S}$ ,  $a = \langle a^1, \dots, a^N \rangle \in \mathcal{A}$  and  $u = \langle u^1, \dots, u^N \rangle \in \mathcal{S}$  it holds that  $P(s, a)(u) := \prod_{i=1}^N P^i(s^i, a^i)(u^i)$ , where  $\mathcal{S}^i$ ,  $\mathcal{A}^i$  and  $P^i : \mathcal{S}^i \times \mathcal{A}^i \rightarrow \Delta(\mathcal{S}^i)$  are the local states, actions and transition probability function of agent  $i \in [N]$ ;
- $s_{init} = \langle s_{init}^1, \dots, s_{init}^N \rangle \in \mathcal{S}$  is the initial state.

We identify agents with the elements of the set  $[N]$ . Each agent  $i \in [N]$  is modeled by an MDP  $M^i = (\mathcal{S}^i, \mathcal{A}^i, P^i, s_{init}^i)$ . Note that the transition probability functions  $P^i$  of the agents are independent and the joint transition probability function in Definition 4 is defined as their product. We denote with  $\bar{i}$  the set  $[N] \setminus \{i\}$  of agents other than  $i$ .

All agents have access to a centralized view and operate in the environment by executing a positional joint policy  $\pi_{act} : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ . We denote with  $\Pi_{act}^{pos}(M) := \mathcal{S} \rightarrow \Delta(\mathcal{A})$  the set of positional joint policies for an MMDP  $M$ .

In this paper, we consider MMDPs with joint *reach-avoid objectives*. Such an objective is defined as a pair  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$  of sets of states such that  $\mathcal{S}_{target} \cap \mathcal{S}_{avoid} = \emptyset$ . It requires that the agents maximize the probability of reaching a joint state in  $\mathcal{S}_{target}$ , while avoiding  $\mathcal{S}_{avoid}$ .

Given a joint policy  $\pi_{act}$  for an MMDP  $M$  with a reach-avoid objective  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$ , we denote with  $\mathbb{P}_{M_{\pi_{act}}}((-\mathcal{S}_{avoid})\mathcal{U}\mathcal{S}_{target})$  the probability of reaching  $\mathcal{S}_{target}$  while avoiding  $\mathcal{S}_{avoid}$  in the Markov Chain  $M_{\pi_{act}}$ .

Given a reach-avoid objective in an MMDP  $M$ , the optimal joint policy synthesis problem requires finding a policy  $\pi^*$  such that  $\mathbb{P}_{M_{\pi^*}}((-\mathcal{S}_{avoid})\mathcal{U}\mathcal{S}_{target}) = \sup_{\pi} \mathbb{P}_{M_{\pi}}((-\mathcal{S}_{avoid})\mathcal{U}\mathcal{S}_{target})$ . The optimal policy that maximizes the reach-avoid probability is denoted as  $\pi^*$ , and the optimal value is denoted as  $v^*(M, \mathcal{S}_{target}, \mathcal{S}_{avoid})$ .

Since a joint policy in an MMDP has access to the full state, it is essentially a policy in the product MDP. Thus, for reach-avoid objectives, it suffices to consider positional joint policies. An optimal joint policy  $\pi_{act}$  for the reach-avoid objective  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$  in the MMDP  $M$  can be computed using standard techniques by solving a linear program.

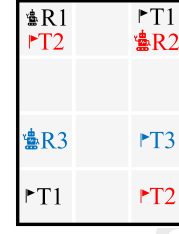


Figure 1: Environment #1 for a robots navigation problem, with robots  $R1$ ,  $R2$ , and  $R3$  and their respective targets ( $T1$ ,  $T2$ ,  $T3$ ).

### 3 Problem Formulation

Implementing joint policies in environments with uncertainty requires effective coordination among the agents. Achieving this coordination often depends on establishing robust communication. However, agents may face limitations in their communication capabilities, including restrictions on the type and amount of information they can share. Additionally, at certain time steps, the environment might limit the number of agents allowed to communicate. Let us consider an example.

**Example 1.** Figure 1 depicts a simple robot navigation problem where three robots must coordinate to reach their respective targets while avoiding collision. The target locations of robots  $R1$ ,  $R2$ , and  $R3$  are labeled with  $T1$ ,  $T2$ , and  $T3$ , respectively. Each of  $R1$  and  $R2$  has two potential target locations. Once each of the robots has reached one of their targets, the team's task is complete.

To maximize the probability of reaching their targets while avoiding collision, the robots must communicate, exchanging information about their current locations and actions. What communication is necessary depends on the executed joint policy and the resulting execution paths. For example, if  $R1$  and  $R2$  decide to swap their locations to reach their respective targets at the top of the grid, communication between these two agents is essential to avoid collisions, while  $R3$  can navigate to its target independently, disregarding the positions of the other robots. On the other hand, if one of  $R1$  or  $R2$  decides to reach its target at the bottom of the grid, communication between this robot and  $R3$  will be needed. Finally, if both  $R1$  and  $R2$  decide to reach their targets at the bottom of the grid, then all three robots must communicate.

Consider the scenario where the communication is constrained, and at any given time at most two robots are allowed to communicate. In order to be robust to this restriction, the joint policy should minimize the need for communication between all three robots at the same time, and should be equipped with a communication policy that prescribes which pair of robots should communicate at a given time.

We study the cooperative execution of a joint policy under specific conditions. Consider a scenario where agents can, without restriction, share some *public information*, such as, for example, their current region. Additionally, they can share precise state and local information, but there is a limitation on the number of agents permitted to do this at any given time.

We now extend the policy execution in this scenario, which we refer to as '*restricted communication*', and present our problem formulation. We formalize our problem as a Markov

game with one and a half players: the multi-agent system, and the stochastic environment. The objective of the system is to reach the set of target states while avoiding the "avoid" states. The game is played in a sequence of rounds, starting at an initial state. At each step, public information is freely exchanged between all agents, and based on this information the agents collaboratively select a subset of agents for further information exchange, which includes the sharing of agents' local states. Communication is established within this subset, with the selected agents sharing information and making collective decisions. Conversely, the remaining agents cannot communicate and share local state information. Consequently, the remaining agents must act independently and make decisions solely based on locally available information and estimates of other agents' states. After all agents execute the respective policies to select actions, the process transitions to the next state according to the probabilistic transition relation.

**Definition 5.** We formulate the team's decision problem as a cooperative Markov game represented as a tuple  $\widehat{M} = (M, \mathcal{O}^1, \dots, \mathcal{O}^N, \mathcal{L}^1, \dots, \mathcal{L}^N, K)$  where

- for each  $i \in [N]$ ,  $\mathcal{O}^i$  and  $\mathcal{L}^i$  are finite sets of respectively publicly observable and local states of agent  $i$ ;
- $M = (N, \mathcal{S}, \mathcal{A}, P, s_{init})$  is an MMDP such that  $\mathcal{S}^i = \mathcal{O}^i \times \mathcal{L}^i$  for each agent  $i \in [N]$ ;
- $K \in \{0, 1, \dots, N\}$  is the number of agents allowed to communicate at each point in time.

In the above definition, we consider MMDPs of a special form where the states of each agent are factored into a *public* part  $\mathcal{O}^i$  that can be observed by all the other agents and *local* part  $\mathcal{L}^i$  that cannot be directly observed by the other agents. In order for agent  $i$ 's local state information to become known to another agent, agent  $i$  must communicate that information to that agent. We consider a setting where there can be a disruption in the communication, resulting in the restriction that only  $K$  out of the  $N$  agents are allowed to communicate. When  $K = N$  we are in the full-communication case where any joint policy can be executed due to the unrestricted communication between the agents. In the other extreme, when  $K = 0$  no communication is allowed. When communication is restricted, that is,  $K < N$ , we assume that agents rely on the notion of *imaginary play* introduced in [Karabag et al., 2022] in order to estimate the current local states of other agents. Furthermore, when  $0 < K < N$ , agents need to agree at each step of the execution on a subset of at most  $K$  agents that will communicate, that is, exchange state information, at this step. This is done by choosing the so-called *communication actions* prescribed by a *communication policy*. The communication policy is a joint policy that is guaranteed to be implementable because it relies only on the public part of the agents' states, which can be always shared by all agents.

In Definition 5, the number  $K$  of agents allowed to exchange information when communication is restricted is fixed. Our results can be extended to the case when  $K$  changes dynamically in the course of the execution.

We define the set of *communication actions* consisting of the sets of agents of size exactly  $K$ , that is  $\mathcal{A}_{comm} := \{c \subseteq [N] \mid |c| = K\}$ . While we could allow communication ac-

tions selecting sets with fewer than  $K$  agents, such actions are dominated by those with maximal allowed cardinality.

For  $c \in \mathcal{A}_{comm}$ , the set of remaining agents is  $\bar{c} := [N] \setminus c$ .

We denote by  $\mathcal{O} := \mathcal{O}^1 \times \dots \times \mathcal{O}^N$  the set of joint public states. A *communication policy* is a function of the form  $\pi_{comm} : \mathcal{O}^+ \rightarrow \Delta(\mathcal{A}_{comm})$ . We let  $\Pi_{comm}(\mathcal{O}, K)$  be the class of all communication policies for given  $\mathcal{O}$  and  $K$ . We denote with  $\Pi_{comm}^{pos}(\mathcal{O}, K) := \mathcal{O} \rightarrow \Delta(\mathcal{A}_{comm})$  the set of positional communication policies for  $\widehat{M}$ .

The agents operate in the environment by executing a pair of joint policies  $\pi = (\pi_{comm}, \pi_{act})$  where  $\pi_{comm} \in \Pi_{comm}$  and  $\pi_{act} \in \Pi_{act}$ . We refer to  $\pi_{act}$  as an *action policy* and to  $\pi_{comm}$  as a *communication policy*. Each agent maintains a *local imaginary copy* of the current local states of the other agents. At each decision step, all agents first share the public parts of their states. The agents execute jointly the policy  $\pi_{comm}$  to select a subset  $c$  of  $K$  agents that will communicate with each other at the current step. The subset of agents eligible for communication shares the local part of their states with each other, and each agent updates their local imaginary copy based on the received information. Thus, the agents in  $c$  have accurate knowledge of each other's current local state, while this can be inaccurate for the rest of the agents. Subsequently, the agents in  $c$  jointly execute the action policy  $\pi_{act}$  to determine a joint action. Each of the other agents executes  $\pi_{act}$  independently. Its state gets updated accordingly, and the local imaginary copies of the states of the agents in  $\bar{c}$  are sampled from the respective distributions. After that, the system proceeds to the next decision step. This continues until a state in  $\mathcal{S}_{target} \cup \mathcal{S}_{avoid}$  is reached. At each step, the agents selected for communication operate in a centralized manner.

### Execution under Restricted Communication

The evolution of the system given a pair of communication and action policies can be formalized as follows. Given  $\widehat{M} = (M, \mathcal{O}^1, \dots, \mathcal{O}^N, \mathcal{L}^1, \dots, \mathcal{L}^N, K)$  with  $M = (N, \mathcal{S}, \mathcal{A}, P, s_{init})$ , a pair of positional joint policies  $\pi = (\pi_{comm}, \pi_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  induces a Markov chain  $\widehat{M}_\pi = (\widehat{\mathcal{S}}, \widehat{P}, \widehat{s}_{init})$  defined as follows.

- The set of states is  $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}^1 \times \dots \times \widehat{\mathcal{S}}^N$ , where for each agent  $i \in [N]$  we have  $\widehat{\mathcal{S}}^i := \mathcal{S}^i \times \prod_{j \in \bar{i}} \mathcal{L}^j$ .

- Let  $\widehat{s}_1 = \langle \widehat{s}_1^1, \dots, \widehat{s}_1^N \rangle \in \widehat{\mathcal{S}}$  and  $\widehat{s}_2 = \langle \widehat{s}_2^1, \dots, \widehat{s}_2^N \rangle \in \widehat{\mathcal{S}}$  where  $\widehat{s}_1^i = \langle \langle o_1^i, l_1^i \rangle, \langle l_1^{i,j} \rangle_{j \in \bar{i}} \rangle$  and  $\widehat{s}_2^i = \langle \langle o_2^i, l_2^i \rangle, \langle l_2^{i,j} \rangle_{j \in \bar{i}} \rangle$  for all  $i \in [N]$ .

$$\text{Let } \widehat{P}(\widehat{s}_1, \widehat{s}_2) = \sum_{c \in \mathcal{A}_{comm}} \sum_{a \in \mathcal{A}} \sum_{\substack{a_j \in \mathcal{A} \\ \text{for all } j \in \bar{c}}} q \cdot$$

$\prod_{i \in c} p_{i,a,c} \cdot \prod_{j \in \bar{c}} p_{j,a_j,c}$  where the quantities  $q$ , and  $p_{i,a,c}$ , and  $p_{j,a_j,c}$  are defined below. We pick the smallest  $i_{min} \in c$  and define the element  $l_c^k \in \mathcal{L}^k$  for  $k \in [N]$ , where  $l_c^k = l_1^k$  if  $k \in c$  and  $l_c^k = l_{i_{min},k}^k$  otherwise.

Then, we define

$$q := \pi_{comm}(o_1^1, \dots, o_1^N)(c) \cdot \pi_{act}(\langle \langle o_1^1, l_c^1 \rangle, \dots, \langle o_1^N, l_c^N \rangle \rangle)(a) \cdot \prod_{j \in \bar{c}} \pi_{act}(\langle \langle o_1^j, l_c^j \rangle, \dots, \langle o_1^j, l_c^j \rangle \rangle)(a_j).$$

For  $i \in c$  and  $a = \langle a^1, \dots, a^N \rangle \in \mathcal{A}$ , let

$$p_{i,a,c} := \frac{P^i(\langle o_1^i, l_1^i \rangle, a^i, \langle o_2^i, l_2^i \rangle)}{\prod_{k \in \bar{i}} \sum_{l \in \mathcal{L}^k} P^k(\langle o_1^k, l_1^k \rangle, a^k, \langle o_2^k, l_2^k \rangle)}.$$

For  $j \in \bar{c}$  and  $a_j = \langle a^1, \dots, a^N \rangle \in \mathcal{A}$ , let

$$p_{j,a_j,c} := \frac{P^j(\langle o_1^j, l_1^j \rangle, a^j, \langle o_2^j, l_2^j \rangle)}{\prod_{k \in \bar{j}} \sum_{l \in \mathcal{L}^j} P^k(\langle o_1^k, l_1^k \rangle, a^k, \langle o_2^k, l_2^k \rangle)}.$$

- $\hat{s}_{init} = \langle \hat{s}_{init}^1, \dots, \hat{s}_{init}^N \rangle$ , where  $s_{init}^i = \langle o_{init}^i, l_{init}^i \rangle$   
and  $\hat{s}_{init}^i = \langle \langle o_{init}^i, l_{init}^i \rangle, \langle l_{init}^i, j_{init}^i \rangle \rangle$ .

For  $U \subseteq \mathcal{S}$ , we define  $\hat{U} \subseteq \hat{\mathcal{S}}$  as  $\hat{U} = \{ \langle \langle s^i, m^i \rangle \rangle_{i \in [N]} \in \hat{\mathcal{S}} \mid \langle s^i \rangle_{i \in [N]} \in U \}$ . We thus lift  $\mathcal{S}_{target}$  and  $\mathcal{S}_{avoid}$  to  $\hat{\mathcal{S}}_{target}$  and  $\hat{\mathcal{S}}_{avoid}$ . For a pair  $\pi = (\pi_{comm}, \pi_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  of communication and action policies,  $\mathbb{P}_{\hat{M}_\pi}((-\hat{\mathcal{S}}_{avoid})\mathcal{U}\hat{\mathcal{S}}_{target})$  is the probability of reaching  $\mathcal{S}_{target}$  while avoiding  $\mathcal{S}_{avoid}$ , executing  $\pi_{comm}$  and  $\pi_{act}$ .

Our goal is to compute a pair  $(\pi_{comm}^*, \pi_{act}^*)$  of positional communication and action policies such that  $\pi_{act}^*$  is optimal for  $M$  under unrestricted communication, and among all pairs with optimal action policies,  $\pi^*$  is optimal for  $\hat{M}$ .

**Problem 1.** Given a cooperative Markov game  $\hat{M} = (M = (N, \mathcal{S}, \mathcal{A}, P, s_{init}), \mathcal{O}^1, \dots, \mathcal{O}^N, \mathcal{L}^1, \dots, \mathcal{L}^N, K)$  as in Definition 5 and a reach-avoid objective  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$  for  $M$ , find a pair of positional policies  $\pi^* = (\pi_{comm}^*, \pi_{act}^*) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  such that  $\mathbb{P}_{M_{\pi_{act}^*}}((-\mathcal{S}_{avoid})\mathcal{U}\mathcal{S}_{target}) = v^*(M, \mathcal{S}_{target}, \mathcal{S}_{avoid})$  and for every  $\pi = (\pi_{comm}, \pi_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  for which we have  $\mathbb{P}_{M_{\pi_{act}}}((-\mathcal{S}_{avoid})\mathcal{U}\mathcal{S}_{target}) = v^*(M, \mathcal{S}_{target}, \mathcal{S}_{avoid})$ , it also holds that

$$\mathbb{P}_{\hat{M}_{\pi^*}}((-\hat{\mathcal{S}}_{avoid})\mathcal{U}\hat{\mathcal{S}}_{target}) \geq \mathbb{P}_{\hat{M}_\pi}((-\hat{\mathcal{S}}_{avoid})\mathcal{U}\hat{\mathcal{S}}_{target}).$$

We restrict **Problem 1** to *positional* communication and action policies. By considering a cooperative Markov game where  $|\mathcal{O}| = 1$  but different sets of agents need to communicate over time, it is easy to see that communication policies with memory are strictly more powerful. However, for the sake of efficient synthesis, we focus on positional policies.

## 4 Policy Synthesis

For the rest of this section, we fix a cooperative Markov game  $\hat{M} = (M, \mathcal{O}^1, \dots, \mathcal{O}^N, \mathcal{L}^1, \dots, \mathcal{L}^N, K)$  with  $M = (N, \mathcal{S}, \mathcal{A}, P, s_{init})$  as in Definition 5, and a reach-avoid objective  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$ . Our goal is to compute a joint action policy that is both optimal and robust to communication restrictions. To this end, we introduce a cost function based on entropy for information sharing among agents. We first present this cost function, followed by our approach for computing a pair of action and communication policies.

### 4.1 Cost Function for Information Sharing Relative to a Communication Policy

Recall that we denote  $\mathcal{O} := \mathcal{O}^1 \times \dots \times \mathcal{O}^N$  and  $\mathcal{L} := \mathcal{L}^1 \times \dots \times \mathcal{L}^N$ . For each communication action  $c \in \mathcal{A}_{comm}$ , we define  $\mathcal{O}^c := \prod_{i \in c} \mathcal{O}^i$  and  $\mathcal{L}^c := \prod_{i \in c} \mathcal{L}^i$ .

Let  $\pi = (\pi_{comm}, \pi_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  be a pair of positional joint communication and action policies.

The cost function  $D_{(\pi_{comm}, \pi_{act})}$  which we define, measures the information exchange between agents required by the action policy that goes beyond what is allowed by the communication policy. It considers all agents  $i$  and all possible coalitions  $c$  of  $K$  agents. With each, it associates a value  $G^i(\pi_{comm}, \pi_{act})$  or  $G^c(\pi_{comm}, \pi_{act})$ , respectively. These are sums of entropy over time, with additional conditioning on the global observations  $\mathcal{O}$  and weighted by the probability that at the respective time step the process is “relevant”.

$$\begin{aligned} D_{(\pi_{comm}, \pi_{act})} &:= \sum_{i \in [N]} G^i(\pi_{comm}, \pi_{act}) \\ &+ \sum_{c \in \mathcal{A}_{comm}} G^c(\pi_{comm}, \pi_{act}) \\ &- \sum_{t=1}^{\infty} H(A_t S_t | S_0 A_1 S_1 \dots A_{t-1} S_{t-1}), \end{aligned}$$

where

$$\begin{aligned} G^i(\pi_{comm}, \pi_{act}) &:= \sum_{t=1}^{\infty} \sum_{\substack{o \in \mathcal{O} \\ l^i \in \mathcal{L}^i}} w'(o, i) p'(o, l^i) L'(i, o, l^i) \\ G^c(\pi_{comm}, \pi_{act}) &:= \sum_{t=1}^{\infty} \sum_{\substack{o \in \mathcal{O}^c \\ l^c \in \mathcal{L}^c}} w''(o, c) p''(o, l^c) L''(c, o, l^c) \\ w'(o, i) &:= \sum_{c \in \mathcal{A}_{comm}, i \notin c} \pi_{comm}(o)(c) \\ w''(o, c) &:= \pi_{comm}(o)(c) \\ p'(o, l^i) &:= \mathbb{P}(O_{t-1} = o, L_{t-1}^i = l^i) \\ p''(o, l^c) &:= \mathbb{P}(O_{t-1} = o, L_{t-1}^c = l^c) \\ L'(i, o, l^i) &:= - \sum_{\substack{a^i \in \mathcal{A}^i \\ o_1^i \in \mathcal{O}^i, l_1^i \in \mathcal{L}^i}} p'(a^i, o_1^i, l_1^i, o, l^i) \\ L''(c, o, l^c) &:= - \sum_{\substack{a^c \in \mathcal{A}^c \\ o_1^c \in \mathcal{O}^c, l_1^c \in \mathcal{L}^c}} p''(a^c, o_1^c, l_1^c, o, l^c) \\ p'(a^i, o_1^i, l_1^i, o, l^i) &:= \\ &\mathbb{P}(A_t^i = a^i, O_t^i = o_1^i, L_t^i = l_1^i \mid O_{t-1} = o, L_{t-1}^i = l^i) \cdot \\ &\log(\mathbb{P}(A_t^i = a^i, O_t^i = o_1^i, L_t^i = l_1^i \mid O_{t-1} = o, L_{t-1}^i = l^i)); \\ p''(a^c, o_1^c, l_1^c, o, l^c) &:= \\ &\mathbb{P}(A_t^c = a^c, O_t^c = o_1^c, L_t^c = l_1^c \mid O_{t-1} = o, L_{t-1}^c = l^c) \cdot \\ &\log(\mathbb{P}(A_t^c = a^c, O_t^c = o_1^c, L_t^c = l_1^c \mid O_{t-1} = o, L_{t-1}^c = l^c)). \end{aligned}$$

Note that if  $K = 0$ , that is,  $\mathcal{A}_{comm} = \{\emptyset\}$ , then  $\pi_{comm}$  is a constant function and no communication between any agents is allowed. In such case,  $D_{(\pi_{comm}, \pi_{act})}$  is the total correlation from [Karabag *et al.*, 2022]. When  $K > 0$ , only the correlation between agents that are outside of what is allowed by  $\pi_{comm}$  contributes to the value of  $D_{(\pi_{comm}, \pi_{act})}$ . If  $D_{(\pi_{comm}, \pi_{act})}$  is 0, this means that all dependencies between the agents in  $\pi_{act}$  are covered by the respective agents being allowed by  $\pi_{comm}$  to communicate at the necessary points in time.

The cost function  $D_{(\pi_{comm}, \pi_{act})}$  has a key property, namely it allows us to provide an upper bound on the performance loss under restricted communication. This is established in the next theorem. Complete proofs are provided in the extended version of this paper [Soudijani and Dimitrova, 2025].



**Theorem 1.** For any cooperative Markov game  $\widehat{M}$  with MMDP  $M$ , reach-avoid objective  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$ , and  $\pi = (\pi_{comm}, \pi_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$ , it holds that

$$\mathbb{P}_{M_{\pi_{act}}}((\neg \mathcal{S}_{avoid}) \mathcal{U} \mathcal{S}_{target}) - \mathbb{P}_{\widehat{M}_{\pi}}((\neg \widehat{\mathcal{S}}_{avoid}) \mathcal{U} \widehat{\mathcal{S}}_{target}) \leq \sqrt{1 - \exp(-D_{(\pi_{comm}, \pi_{act})})}.$$

Due to the form of  $D_{(\pi_{comm}, \pi_{act})}$ , in our method for computing a pair of communication and action policies, described in the rest of the section, we will use a proxy function.

## 4.2 Policy Synthesis

Our approach proceeds in two steps.

### Optimistic Optimal Value for Reach-Avoid Probability

As we require the action policy to be optimal under unrestricted communication, in the first step, we compute the optimal value  $v^*(M, \mathcal{S}_{target}, \mathcal{S}_{avoid})$  for the reach-avoid probability assuming unrestricted communication.

### Minimizing the Cost of Communication

In the second step, we use the value  $v^*(M, \mathcal{S}_{target}, \mathcal{S}_{avoid})$  as a threshold in the computation of a pair  $(\pi_{comm}, \pi_{act})$  of policies. This threshold constrains the action policy  $\pi_{act}$  to be optimal under unrestricted communication. Additionally, we formulate an objective function based on the cost function  $D_{(\pi_{comm}, \pi_{act})}$ . To this end, we provide a proxy to  $D_{(\pi_{comm}, \pi_{act})}$ , expressed in terms of occupancy measures. The term  $\sum_{t=1}^{\infty} H(A_t S_t | S_0 A_1 S_1 \dots A_{t-1} S_{t-1})$  can be expressed in terms of occupancy measure using existing results [Biondi *et al.*, 2014]. For the other two terms in  $D_{(\pi_{comm}, \pi_{act})}$ , we provide upper bounds.

Consider a pair of joint policies  $\pi = (\pi_{comm}, \pi_{act})$  and the Markov chain  $M_{\pi_{act}}$  induced from the MMDP  $M$ . This Markov chain generates a stationary process  $X$ , which is the joint path of the agents. The entropy  $H(X)$  of  $X$  has a closed form expression in terms of  $\nu_{s,a}$ .

**Proposition 1.** The entropy of the joint state-action process until reaching the target can be expressed in terms of the state-action occupancy measure  $\nu_{s,a}$  as

$$\begin{aligned} H(S_0) + \sum_{t=1}^{\infty} H(A_t S_t | S_0 A_1 S_1 \dots A_{t-1} S_{t-1}) &= \\ &- \left( \sum_{s,a'} \nu_{s,a'} \cdot \log \left( \frac{\nu_{s,a'}}{\sum_b \nu_{s,b}} \right) \right) \\ &- \left( \sum_{s,a',s'} \nu_{s,a'} \cdot P(s, a', s') \cdot \log P(s, a', s') \right). \end{aligned}$$

The path of a single agent  $i$  or a group of agents  $c$  follows a hidden Markov model where  $X$  is the underlying process and  $X^i$ , or  $X^c$ , respectively, is the observed process. Therefore, the terms  $\bar{G}^i(\pi_{comm}, \pi_{act})$  and  $\bar{G}^c(\pi_{comm}, \pi_{act})$  do not have closed-form expressions based on occupancy measures. Instead, we employ stationary processes which induce the same occupancy measures, and derive expressions that are upper bounds for  $\bar{G}^i(\pi_{comm}, \pi_{act})$  and  $\bar{G}^c(\pi_{comm}, \pi_{act})$ .

As  $s^i = \langle o^i, l^i \rangle$  for some  $o^i \in \mathcal{O}^i$ ,  $l^i \in \mathcal{L}^i$ , we write  $\nu_{o^i, l^i, a^i}$  instead of  $\nu_{s^i, a^i}$ . For  $o \in \mathcal{O}$  and  $c \in \mathcal{A}_{comm}$ , we define  $\nu_{o,c} := \nu_o \cdot \pi_{comm}(o)(c) = \left( \sum_{l \in \mathcal{L}, a \in \mathcal{A}} \nu_{o,l,a} \right) \pi_{comm}(o)(c)$ .

For each agent  $i \in [N]$  and each set of agents  $c \in \mathcal{A}_{comm}$ , we consider the stationary process that induces the same occupancy measures  $\nu_{o, l^i, a^i}$  and  $\nu_{o, l^c, a^c}$ , respectively, as the joint policy. We establish the following proposition.

**Proposition 2.** Let

$$\begin{aligned} \bar{G}^i(\pi_{comm}, \pi_{act}) &= \\ &- \left( \sum_{o, l^i, a^i} \nu_{o, l^i, a^i} \cdot w'(o, i) \cdot \log \left( \frac{\nu_{o, l^i, a^i}}{\sum_{b^i} \nu_{o, l^i, b^i}} \right) \right) \\ &- \left( \sum_{o, l^i, a^i, o_1^i, l_1^i} \nu_{o, l^i, a^i} \cdot w'(o, i) \cdot h'(o^i, l^i, a^i, o_1^i, l_1^i) \right), \\ \bar{G}^c(\pi_{comm}, \pi_{act}) &= \\ &- \left( \sum_{o, l^c, a^c} \nu_{o, l^c, a^c} \cdot w''(o, c) \cdot \log \left( \frac{\nu_{o, l^c, a^c}}{\sum_{b^c} \nu_{o, l^c, b^c}} \right) \right) \\ &- \left( \sum_{o, l^c, a^c, o_1^c, l_1^c} \nu_{o, l^c, a^c} \cdot w''(o, c) \cdot h''(o^c, l^c, a^c, o_1^c, l_1^c) \right) \end{aligned}$$

$$w'(o, i) = \sum_{c \in \mathcal{A}_{comm}, i \notin c} \frac{\nu_{o,c}}{\sum_{c' \in \mathcal{A}_{comm}} \nu_{o,c'}},$$

$$w''(o, c) = \frac{\nu_{o,c}}{\sum_{c' \in \mathcal{A}_{comm}} \nu_{o,c'}},$$

$$\begin{aligned} h'(o^i, l^i, a^i, o_1^i, l_1^i) &:= \\ &P^i(o^i, l^i, a^i)(o_1^i, l_1^i) \cdot \log P^i(o^i, l^i, a^i)(o_1^i, l_1^i), \\ h''(o^c, l^c, a^c, o_1^c, l_1^c) &:= \\ &P^c(o^c, l^c, a^c)(o_1^c, l_1^c) \cdot \log P^c(o^c, l^c, a^c)(o_1^c, l_1^c), \\ P^c(\langle o^j \rangle_{j \in c}, \langle l^j \rangle_{j \in c}, \langle a^j \rangle_{j \in c}) &(\langle o_1^j \rangle_{j \in c}, \langle l_1^j \rangle_{j \in c}) = \\ &\prod_{j \in c} P^j(o^j, l^j, a^j)(o_1^j, l_1^j). \end{aligned}$$

Then, it holds that  $\bar{G}^i(\pi_{comm}, \pi_{act}) \leq \bar{G}^i(\pi_{comm}, \pi_{act})$  and  $\bar{G}^c(\pi_{comm}, \pi_{act}) \leq \bar{G}^c(\pi_{comm}, \pi_{act})$ .

Combining Proposition 1 and Proposition 2, we obtain an upper bound  $\bar{D}_{(\pi_{comm}, \pi_{act})}$  on  $D_{(\pi_{comm}, \pi_{act})}$  based on occupancy measures. That is, we have  $D_{(\pi_{comm}, \pi_{act})} \leq \bar{D}_{(\pi_{comm}, \pi_{act})}$  for

$$\begin{aligned} \bar{D}_{(\pi_{comm}, \pi_{act})} &:= -H(X) + \sum_{i \in [N]} \bar{G}^i(\pi_{comm}, \pi_{act}) \\ &+ \sum_{c \in \mathcal{A}_{comm}} \bar{G}^c(\pi_{comm}, \pi_{act}). \end{aligned} \quad (1)$$

**Cost optimization problem.** Using the function  $\bar{D}$  defined in eq. (1), we formulate below, in (2), the optimization problem with decision variables  $x_{o,l,a}$  and  $x_{o,c}$ , representing the occupancy measures for each joint public state-local state-action triplet  $(o, l, a)$  and for each joint public state-communication action pair  $(o, c)$ , respectively. Through appropriate marginalization,  $x_{o, l^i, a^i}$  and  $x_{o, l^c, a^c}$  represent the public state-local state-action occupancy measures for individual agents and groups of agents, respectively. For synthesis, we assume that the occupancy measure is finite for all states  $s \in \mathcal{S} \setminus (\mathcal{S}_{avoid} \cup \mathcal{S}_{target})$ . We add absorbing sink-states and corresponding actions to  $\widehat{M}$ , denoted with  $(o_\alpha, l_\alpha) = ((o_\alpha^1, l_\alpha^1), \dots, (o_\alpha^n, l_\alpha^n))$  and  $a_\alpha = (a_\alpha^1, \dots, a_\alpha^n)$ , respectively. These states represent the end of the game concerning the reach-avoid objective, that is, for all  $(o, l) \in (\mathcal{S}_{target} \cup \mathcal{S}_{avoid})$  we have  $P((o, l), a_\alpha)((o_\alpha, l_\alpha)) = 1$ .

We consider the optimization problem (2) with the objective (2a) and constraints (2b) – (2h) given below. The value  $v^*$  computed in the first step is used to constrain from below the reach-avoid probability of the action policy. Additionally, the value of the function  $\bar{D}$  must be minimized.

$$\min_{(x_{o,l,a}, x_{o,c})} \bar{d} = -h + \sum_{i \in [N]} g^i + \sum_{c \in \mathcal{A}_{comm}} g^c \quad \text{s.t.} \quad (2a)$$

$$g^i = \dots \quad \forall i \in [N] \quad /* \text{ encodes } \bar{G}^i */ \quad (2b)$$

$$g^c = \dots \quad \forall c \in \mathcal{A}_{comm} \quad /* \text{ encodes } \bar{G}^c */ \quad (2c)$$

$$h = \dots \quad /* \text{ encodes } H(X) */ \quad (2d)$$

$$w'(o, i) = \dots, w''(o, c) = \dots \quad \forall o, i, c \quad (2e)$$

$$v^* \leq \dots \quad /* \text{ reach-avoid probability } v^* / \quad (2f)$$

$$\sum_{a \in \mathcal{A} \cup \{a_\alpha\}} x_{o,l,a} = \dots \quad \forall (o, l) \in \mathcal{S} \quad (2g)$$

$$x_{o,l,a} \geq 0, \quad x_{o_\alpha, l_\alpha, a} = 0 \quad \forall (o, l) \in \mathcal{S}, a \in \mathcal{A} \cup \{a_\alpha\}$$

$$x_{o,c} \geq 0, \quad x_{o_\alpha, c} = 0 \quad \forall o \in \mathcal{O}, c \in \mathcal{A}_{comm}$$

$$\sum_{l \in \mathcal{L}, a \in \mathcal{A}} x_{o,l,a} = \sum_{c \in \mathcal{A}_{comm}} x_{o,c} \quad \forall o \in \mathcal{O} \quad (2h)$$

The constraints (2b) – (2g) are presented below.

Constraints (2b), (2c) and (2d) capture the definitions of the expressions  $\bar{G}^i(\pi_{comm}, \pi_{act})$ ,  $\bar{G}^c(\pi_{comm}, \pi_{act})$  and  $H(X)$  from Proposition 2 and Proposition 1 respectively. Formally,

$$\begin{aligned} g^i &= - \left( \sum_{o,l^i,a^i} x_{o,l^i,a^i} \cdot w(o, i) \cdot \log \left( \frac{x_{o,l^i,a^i}}{\sum_{b^i} x_{o,l^i,b^i}} \right) \right) \\ &\quad - \left( \sum_{o,l^i,a^i} x_{o,l^i,a^i} \cdot w(o, i) \cdot P^i(o^i, l^i, a^i)(o_1^i, l_1^i) \right. \\ &\quad \cdot \log P^i(o^i, l^i, a^i)(o_1^i, l_1^i) \left. \right) \\ g^c &= - \left( \sum_{o,l^c,a^c} x_{o,l^c,a^c} \cdot w(o, c) \cdot \log \left( \frac{x_{o,l^c,a^c}}{\sum_{b^c} x_{o,l^c,b^c}} \right) \right) \\ &\quad - \left( \sum_{o,l^c,a^c} x_{o,l^c,a^c} \cdot w(o, c) \cdot P^c(o^c, l^c, a^c)(o_1^c, l_1^c) \right. \\ &\quad \cdot \log P^c(o^c, l^c, a^c)(o_1^c, l_1^c) \left. \right) \\ h &= - \left( \sum_{s,a'} x_{s,a'} \cdot \log \left( \frac{x_{s,a'}}{\sum_b x_{s,b}} \right) \right) \\ &\quad - \left( \sum_{s,a',s'} x_{s,a'} \cdot P(s, a')(s') \cdot \log P(s, a')(s') \right). \end{aligned}$$

Similarly, constraints (2e) encode the respective definitions from Proposition 2. Formally, we have

$$\begin{aligned} w(o, i) &= \frac{\sum_{c \in \mathcal{A}_{comm}, i \notin c} x_{o,c}}{\sum_{c' \in \mathcal{A}_{comm}} x_{o,c'}} \\ w(o, c) &= \frac{x_{o,c}}{\sum_{c' \in \mathcal{A}_{comm}} x_{o,c'}}. \end{aligned}$$

Constraint (2f) lower-bounds the reach-avoid probability for the sought action policy, and constraint (2g) enforces the usual flow constraint for occupancy measures:

$$\begin{aligned} v^* &\leq \sum_{\substack{(o,l) \in \mathcal{S} \setminus (\mathcal{S}_{avoid} \cup \mathcal{S}_{target}) \\ a \in \mathcal{A}, (o', l') \in \mathcal{S}_{target}}} x_{o,l,a} P(o, l, a)(o', l'), \\ \sum_{a \in \mathcal{A} \cup \{a_\alpha\}} x_{o,l,a} &= \sum_{\substack{(o', l') \in \mathcal{S} \\ b \in \mathcal{A} \cup \{a_\alpha\}}} x_{o', l', b} P(o', l', b)(o, l) + \mathbb{1}_{\{s_{init}=s\}}. \end{aligned}$$

Finally, constraints (2h) express the relationship between the occupancy measures  $x_{o,l,a}$  and  $x_{o,c}$  via the occupancy measure of  $o \in \mathcal{O}$ .

**Policies from an optimal solution.** Let  $(x_{o,l,a}^*, x_{o,c}^*)$  be an optimal solution to the optimization problem (2). We define

the pair  $(\pi_{comm}^*, \pi_{act}^*)$  of policies by

$$\begin{aligned} \pi_{comm}^*(o)(c) &= \frac{x_{o,c}^*}{\sum_{d \in \mathcal{A}_{comm}} x_{o,d}^*}, \\ \pi_{act}^*(o, l)(a) &= \frac{x_{o,l,a}^*}{\sum_{b \in \mathcal{A}} x_{o,l,b}^*}. \end{aligned} \quad (3)$$

The next theorem states that  $(\pi_{comm}^*, \pi_{act}^*)$  has the desired properties, namely  $\pi_{act}^*$  ensures  $v^*$  under unrestricted communication, and  $(\pi_{comm}^*, \pi_{act}^*)$  minimizes the value of  $\bar{D}$ .

**Theorem 2.** Let  $\widehat{M}$  be a cooperative Markov game with MMDP  $M$ ,  $(\mathcal{S}_{target}, \mathcal{S}_{avoid})$  be a reach-avoid objective, and  $(\pi_{comm}^*, \pi_{act}^*)$  be the pair of policies defined by (3) for an optimal solution to (2a). Then, we have  $\mathbb{P}_{M_{\pi_{act}^*}}((\neg \mathcal{S}_{avoid}) \mathcal{U} \mathcal{S}_{target}) = \sup_{\pi} \mathbb{P}_{M_{\pi}}((\neg \mathcal{S}_{avoid}) \mathcal{U} \mathcal{S}_{target})$ . Furthermore, for every pair  $(\pi'_{comm}, \pi'_{act}) \in \Pi_{comm}^{pos}(\mathcal{O}, K) \times \Pi_{act}^{pos}(M)$  of positional joint communication and action policies, if the policy  $\pi'_{act}$  is optimal, that is, if  $\mathbb{P}_{M_{\pi'_{act}}}((\neg \mathcal{S}_{avoid}) \mathcal{U} \mathcal{S}_{target}) = \sup_{\pi} \mathbb{P}_{M_{\pi}}((\neg \mathcal{S}_{avoid}) \mathcal{U} \mathcal{S}_{target})$ , then for function  $\bar{D}$  defined in eq. (1) we have  $\bar{D}_{(\pi_{comm}^*, \pi_{act}^*)} \leq \bar{D}_{(\pi'_{comm}, \pi'_{act})}$ .

## 5 Experimental Evaluation

We evaluated our approach on four multi-robot navigation scenarios. In each scenario, we consider a MAS with three agents and  $K = 2$ . We used the formalization in Section 4 to synthesize for each problem a pair of communication and action policies. All experiments were performed on a Macbook Pro with an Apple M2 chip and 32GB memory. The two optimization problems are solved using GLPK [Makhorin, 2000] and SNOPT [Gill *et al.*, 2018], respectively. The detailed setup and results of all scenarios are included in [Soudijani and Dimitrova, 2025].

With this evaluation we aim to demonstrate the following.

1. Our method synthesizes policies with zero communication cost (which means they fully conform to the communication restrictions) when such policies exist.
2. Our method is capable of synthesizing policies that adhere to communication restriction while incurring zero communication cost—an outcome that cannot be achieved through approaches based solely on minimizing total correlation as an objective function.
3. Our method can synthesize communication policies that adapt dynamically to the current state of public information so that the set of communicating agents changes.
4. There is a trade-off between performance (the reach-avoid probability) and the value of our cost function.

Next, we describe the four scenarios and the respective results. Further details can be found in the extended version of this paper [Soudijani and Dimitrova, 2025].

### Scenario #1 with Navigation Tasks

We consider the environment in Figure 1. The task of the robots, initialized as marked in the figure, is to navigate to one of their target cells, labeled  $T1$ ,  $T2$ , and  $T3$ , respectively,

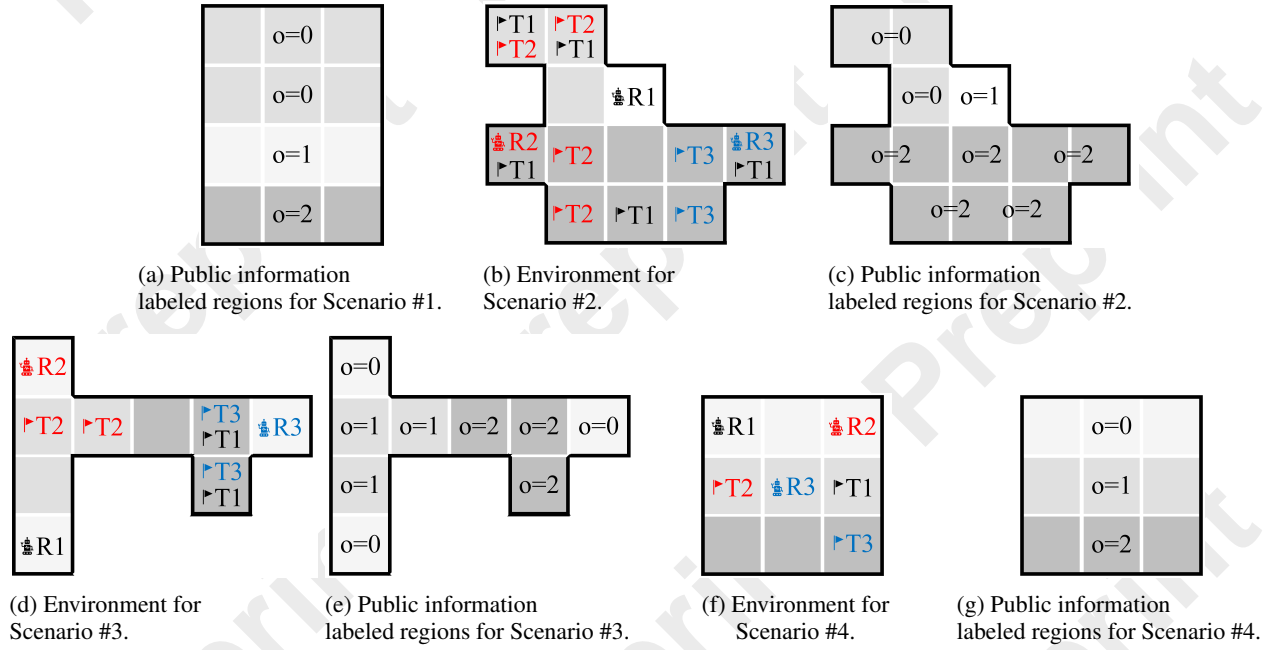


Figure 2: Grid environments and regions labeled with public information for the scenarios in Section 5. Robots’ initial positions are indicated by  $R1$ ,  $R2$ , and  $R3$ , and their target positions by  $T1$ ,  $T2$ , and  $T3$ .

avoiding collision. At any given time, only two robots can communicate and share precise locations. Which ones communicate is decided based on public information, which is the current regions of the robots. The regions, labeled  $o = 0$ ,  $o = 1$ , and  $o = 2$  are shown in Figure 2a. The possible actions are moving in one of the four cardinal directions or remaining in place. Moves succeed with probability 0.9 and fail, resulting in remaining in the current cell, with probability 0.1. The remain action and impossible actions stay in place.

**Result.** The method proposed in Section 4 generates a pair of policies where the action policy is optimal under unrestricted communication, with reach-avoid probability 0.99. The synthesized communication policy results in *cost zero*. It assigns probability 1 to robots  $R1$  and  $R2$  communicating. The synthesized action policy matches that, unlike other possible action policies with reach-avoid probability 0.99. Thus, our method identifies a suitable action policy that can be equipped with a communication policy achieving cost zero.

### Scenario #2 with a Swarm Intersection

The goal of this scenario is to compare our method with the approach based on minimizing the total correlation. In the scenario depicted in Figure 2b, three robots are required to navigate to their respective target locations while avoiding collisions. The publicly available information corresponding to each region is illustrated in Figure 2c. Again, only two out of three agents can communicate their precise location without an extra communication cost.

The work closest to ours is [Karabag *et al.*, 2022], which employs total correlation to measure dependencies between agents and to synthesize an (action) policy that is robust to communication loss. However, [Karabag *et al.*, 2022] does not consider communication policies since robustness

is w.r.t. complete loss of communication, not to partial limitation of the communication. Thus, there is a baseline for comparison only for the action policies synthesized by our approach (which are optimal), namely the policies computed using total correlation in the objective function. Minimizing total correlation cannot yield action policies that satisfy communication restrictions, whereas our method can do so, and also generates a matching communication policy.

**Result.** In scenario #2, our approach finds a pair of action and communication policies that satisfy the reach-avoid property with probability 1 and ensure communication cost 0. Thus, it finds policies that satisfy the communication restriction. In contrast, the action policy generated using total correlation violates the communication restriction at time  $t = 2$  when it requires coordination among three agents. Here, the minimum total correlation is 0.591, while the total correlation of the policy computed by our method is 0.693, and thus it will not be computed by minimizing total correlation.

### Scenario #3 with a Hallway

In this scenario, we consider the environment in Figure 2d. The public information regions, labeled  $o = 0$ ,  $o = 1$ , and  $o = 2$ , are depicted in Figure 2e. In this scenario, coordination among certain robots is critical at certain time steps to avoid collisions. As before, only two of the robots are allowed to communicate at any given time.

**Result.** The synthesized action policy achieves a reach-avoid probability of 1, and the communication policy ensures zero communication cost by selecting the appropriate set of robots to communicate in different public information states.



#### Scenario #4 with High Uncertainty

The robots in the environment shown in Figure 2f must navigate to their respective target cells while avoiding collisions. The public information regions here correspond to the rows, as shown in Figure 2g. The actions are the same as in Scenario #1. Here, however, the move actions lead to the desired cell with probability 0.9, and the remaining 0.1 probability is redistributed across the current cell and all other neighboring cells. For impossible move actions, the full transition probability 1.0 is redistributed among the current cell and all valid neighboring cells. As in previous scenarios, when communication is restricted, only two of three agents can communicate. In this scenario, a crowded intersection necessitates communication among all agents to avoid collisions.

**Result.** Here, the optimal reach-avoid probability under full communication is 0.958. No pair of action and communication policies exists that achieves this probability with zero communication cost. If, however, we lower the threshold and only ask for a policy that guarantees 0.92 reach-avoid probability under unrestricted communication, then our method synthesizes one with zero cost.

#### Performance

Our evaluation is focused on evaluating the quality of our approach on a range of relatively small but interesting problem instances, as well as its principle feasibility. Although our method requires solving large nonlinear optimization problems to synthesize a pair of policies, the running time remains reasonable for the considered benchmarks. The run-times range from a few minutes for Environment #3 (with 528 constraints and 62,581 variables) to one hour for Environment #4 (with 531 constraints and 62,956 variables) and Environment #2 (with 1,332 constraints and 163,081 variables), and up to three hours for Environment #1 (with 1,344 constraints and 164,581 variables). Clearly, the performance depends on the number of agents and the sparsity of the transition probability matrices. In the future, we plan to conduct larger-scale experiments and develop techniques for further improving performance.

## 6 Discussions and Future Work

**Limitations and extensions.** One limitation of the model we study in this paper is that the number  $K$  of agents always allowed to communicate is fixed. We can incorporate a dynamically changing  $K$  as part of the state, making the model and cost function more complex. Another limitation is the restriction to positional policies, which allows us to formulate the problem via occupancy measures. In the future we will consider extensions with bounded memory communication policies. Another challenge is the scalability, in particular with growing number of agents, which we plan to address by developing methods that iteratively improve the policies for subsets of agents. Additionally, we assume that the model, including public information is given. While in many cases identifying public information is natural (regions, fixed capabilities), exploring relations to agents' observations and approaches to deriving such information is an interesting direction.

**Richer communication models.** The focus of our work is to establish a rigorous theoretical foundation and provide novel insights and methodology. Our main aim is to provide the necessary basis for further theoretical exploration and practical applications. Moving forward, we plan to study extensions with richer communication restrictions (such as dynamic changes in communication availability) and explore ways to make the approach robust to implementation aspects such as delayed or noisy communication.

**Identifying public information.** The granularity of the public information affects the number of decision variables (for the communication policy) and hence the performance of policy synthesis. More (i.e., finer) public information leads to higher computation times. The choice of public information depends on the application and available communication bandwidth. In many cases, this is natural, such as coarser geographic regions or agents' capabilities. Developing techniques for identifying public information is one interesting direction for future work.

## References

- [Biondi *et al.*, 2014] Fabrizio Biondi, Axel Legay, Bo Friis Nielsen, and Andrzej Wasowski. Maximizing entropy over markov processes. *J. Log. Algebraic Methods Program.*, 83(5-6):384–399, 2014.
- [Cover and Thomas, 2006] Thomas M. Cover and Joy A. Thomas. *Elements of information theory* (2. ed.). Wiley, 2006.
- [Gill *et al.*, 2018] Philip E Gill, Walter Murray, Michael A Saunders, and Elizabeth Wong. User's guide for snopt 7.7: Software for large-scale nonlinear programming. *Center for Computational Mathematics Report CCoM*, 15(3), 2018.
- [Goldman and Zilberstein, 2004] Claudia V. Goldman and Shlomo Zilberstein. Decentralized control of cooperative systems: Categorization and complexity analysis. *J. Artif. Intell. Res.*, 22:143–174, 2004.
- [Guestrin *et al.*, 2001] Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored mdps. *Advances in neural information processing systems*, 14, 2001.
- [Karabag *et al.*, 2022] Mustafa O. Karabag, Cyrus Neary, and Ufuk Topcu. Planning not to talk: Multiagent systems that are robust to communication loss. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor, editors, *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, pages 705–713. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.
- [Makhorin, 2000] Andrew Makhorin. Modeling language gnu mathprog. *Relatório Técnico, Moscow Aviation Institute*, 63, 2000.
- [Melo and Veloso, 2011] Francisco S Melo and Manuela Veloso. Decentralized mdps with sparse interactions. *Artificial Intelligence*, 175(11):1757–1789, 2011.

- [Rizk *et al.*, 2018] Yara Rizk, Mariette Awad, and Edward W Tunstel. Decision making in multiagent systems: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):514–529, 2018.
- [Shannon and Weaver, 1949] Claude E Shannon and Warren Weaver. The mathematical theory of communication. university of illinois. *Urbana*, 117:10, 1949.
- [Soudijani and Dimitrova, 2025] Saleh Soudijani and Rayna Dimitrova. Synthesis of communication policies for multi-agent systems robust to communication restrictions. <https://arxiv.org/abs/2505.13311>, 2025.
- [Wu *et al.*, 2011] Feng Wu, Shlomo Zilberstein, and Xiaoping Chen. Online planning for multi-agent systems with bounded communication. *Artif. Intell.*, 175(2):487–511, 2011.