

Measuring Policy Distance for Multi-Agent Reinforcement Learning

Tianyi Hu

Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
hutianyi2021@ia.ac.cn

Zhiqiang Pu

Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
zhiqiang.pu@ia.ac.cn

Xiaolin Ai

Institute of Automation, CAS
Beijing, China
xiaolin.ai@ia.ac.cn

Tenghai Qiu

Institute of Automation, CAS
Beijing, China
tenghai.qiu@ia.ac.cn

Jianqiang Yi

Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
jianqiang.yi@ia.ac.cn

ABSTRACT

Diversity plays a crucial role in improving the performance of multi-agent reinforcement learning (MARL). Currently, many diversity-based methods have been developed to overcome the drawbacks of excessive parameter sharing in traditional MARL. However, there remains a lack of a general metric to quantify policy differences among agents. Such a metric would not only facilitate the evaluation of the diversity evolution in multi-agent systems, but also provide guidance for the design of diversity-based MARL algorithms. In this paper, we propose the multi-agent policy distance (MAPD), a general tool for measuring policy differences in MARL. By learning the conditional representations of agents' decisions, MAPD can compute the policy distance between any pair of agents. Furthermore, we extend MAPD to a customizable version, which can quantify differences among agent policies on specified aspects. Based on the online deployment of MAPD, we design a multi-agent dynamic parameter sharing (MADPS) algorithm as an example of the MAPD's applications. Extensive experiments demonstrate that our method is effective in measuring differences in agent policies and specific behavioral tendencies. Moreover, in comparison to other methods of parameter sharing, MADPS exhibits superior performance.

KEYWORDS

Multi-Agent System; Reinforcement Learning; Diversity Measure

ACM Reference Format:

Tianyi Hu, Zhiqiang Pu, Xiaolin Ai, Tenghai Qiu, and Jianqiang Yi. 2024. Measuring Policy Distance for Multi-Agent Reinforcement Learning. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, Auckland, New Zealand, May 6 – 10, 2024, IFAAMAS, 9 pages.



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has achieved significant success in practical applications, such as multiplayer gaming [23], multi-robot system controlling [5], sensor networks [29] and autonomous driving [13]. Traditional MARL approaches commonly employ parameter sharing techniques to enhance sample efficiency [21, 25] and algorithm scalability [1]. Although parameter sharing has been shown to be useful for accelerating training [4], its excessive use can restrict the multi-agent system to fully shared policies, thus limiting their adaptability to complex tasks [12].

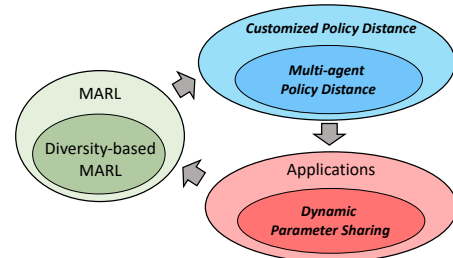


Figure 1: The relationship between our work and MARL. Our contributions are highlighted in bold and italicized.

The fundamental cause of this issue lies in the lack of policy diversity in multi-agent systems. This lack of diversity not only hinders the adaptability of multi-agent systems to complex tasks, but also weakens the overall exploration during learning. To address this problem, diversity-based MARL methods have emerged. One of the classic approaches to achieving policy diversity is to directly incorporate agent indices into the observations [6, 8]. This allows the agents to exhibit different behaviors while sharing network parameters. Other intuitive approaches involve grouping the agents based on prior knowledge or learned information, and only sharing parameters within each group. For instance, [2] leverages the heterogeneous characteristics of agents for grouping, [4] utilizes supervised learning and clustering to group the agents before training. Additionally, some methods introduce the concept of roles into MARL, controlling the parameters of the agents based on

their roles. [27] captures emergent roles from agents’ local observations, [28] learns roles by embedding decomposed action spaces in the initial training phase, and [15] learns roles during training and updates the role clusters at fixed time steps.

Although the above methods allow agents to emerge policy diversity, there is currently a lack of a general metric that can quantify the differences in agent policies and be applicable in various types of multi-agent systems. It is meaningful to measure such differences between agent policies, as it can help us analyze the diversity evolution in multi-agent systems during learning, and gain a deeper understanding of how MARL can benefit from diversity.

There have been previous attempts at measuring policy differences in MARL. [19] introduces a "policy distance" by sampling states and recording the ratio of different actions selected by the population. [9] employs KL divergence to measure the distances between action-based roles, [18] uses the maximum mean discrepancy to compare distributions over trajectories under different policies. [14] formulates the concept of behavioral diversity as the discrepancies of occupancy measures and utilizes f-divergence. Such methods based on divergence metrics fail to satisfy the triangle inequality. This results in increased challenges in analyzing multi-agent policies and complicates further applications. Some researches have delved into the representations of policies. In [7], episodes generated by agent policies are mapped to real-valued vectors, and the Euclidean distances between these vectors is utilized as the policy distances. [10] maps agent policies to a distribution, and employs KL divergence and one-dimensional Wasserstein distance to handle discrete and continuous action spaces, respectively. These approaches, which map the overall policies to low-dimensional vectors or distributions, inevitably result in significant information loss. The state-of-the-art method for measuring behavioral diversity at a system level is presented in [3]. This method calculates the behavioral distance between pairs of agents, and employs the average of these distances as an indicator of system diversity. However, this approach requires assumptions about the type of agent policies and presumes that the observation and action spaces of agents are identical, limiting its generality.

We are desired to develop a general tool for measuring policy difference among agents. This tool not only allows us to gain a deeper understanding of the impact of diversity in MARL, but also provides guidance for the design of diversity-based algorithms. We model agent policies as action distributions conditioned on observations, and discuss the limitations of directly using action distributions for measuring. Our key insight, is to learn latent conditional representations of agents’ decisions under each observation, and measure the accumulation of differences between these representations. Additionally, we extend MAPD to a customizable version, to quantify policy differences on specific aspects which the users are interest in. Based on the online deployment of MAPD, we design multi-agent dynamic parameter sharing (MADPS) as a paradigm for the utilization of MAPD. MADPS is an algorithm that can automatically adjust the parameter sharing scheme of agents during training, helping multi-agent systems find the appropriate balance between policy diversity and parameter sharing. The relationship between our work and MARL is illustrated in Figure 1.

Our contributions can be summarized as follows:

- (1) Proposing MAPD, a generalized measure of policy differences for multi-agent reinforcement learning.
- (2) Developing a customizable version of MAPD for measuring policy differences on specific aspects.
- (3) Designing MADPS, a multi-agent dynamic parameter sharing algorithm based on the online deployment of MAPD.
- (4) Demonstrating the utilization of MAPD through cases in the multi-agent particle environment (MPE) [16], and applying MADPS on both MPE and StarCraft II micromanagement environment [22] to showcase its superior performance.

2 BACKGROUND

In this section, we introduce partially observable markov game (POMG) [2], and give a general mathematical definition of agent policies to establish the foundation of policy distance measurement.

A partially observable markov game can be represented by a 8-tuple:

$$\langle \mathcal{N}, \mathcal{S}, \{\Omega^i\}_{i \in \mathcal{N}}, \{O^i\}_{i \in \mathcal{N}}, \{A^i\}_{i \in \mathcal{N}}, \{R^i\}_{i \in \mathcal{N}}, \mathcal{T}, \gamma \rangle,$$

where $\mathcal{N} = \{1, \dots, n\}$ defines the set of all agents, \mathcal{S} is the global state space. For an agent $i \in \mathcal{N}$, Ω^i denotes the observation space, O^i denotes the corresponding observation function, A^i is the action space and $R^i : \mathcal{S} \times \{A_i\}_{i \in \mathcal{N}} \times \mathcal{S} \mapsto \mathbb{R}$ is the reward function. $\mathcal{T} : \mathcal{S} \times \{A_i\}_{i \in \mathcal{N}} \mapsto \mathcal{S}$ represents the global state transition function, and γ refers to the discount factor.

At each time t , an agent i receives a partial observation $o_t^i \sim O^i(s_t)$ from the environment, and takes an action $a_t^i \in A^i$ according to its policy, where $s_t \in \mathcal{S}$ is the global state of the environment. Each agent in a POMG aims to learn an optimal policy π_i^* that, given an observation at a specific time step, outputs the optimal action. The optimal policy can be learned by maximizing the total expected reward accumulated by the agent.

In order to establish a general agent policy difference metric, the first thing we need to do is to give a general definition of agent policy. An agent’s policy can always be represented as a **conditional probability distribution of actions under condition** o , that is, $\pi_i := \pi_i(a|o)$. This definition naturally holds when agents adopt stochastic policies and take actions based on policy-based MARL methods. In the case of value-based approaches, the action-values can be converted into a probability distribution across all actions, e.g., through softmax. Even when deterministic policies are utilized, actions can still be represented as a probability distribution, with a 100% probability assigned to the chosen action. According to this definition, the policies of agents with discrete actions are discrete conditional distributions, while those corresponding to continuous actions are continuous distributions.

3 MEASURING POLICY DISTANCE BETWEEN AGENTS

In this section, we present the desired input-output format and properties of MAPD. Following that, we discuss the reasons and methods for learning a conditional representation for the agent decision-making. Finally, we show how to utilize this representation to compute a policy distance.

3.1 Analysis

Our Objective. This work aims to develop a general tool, termed multi-agent policy distance (MAPD), to measure the policy difference between any two agents in a multi-agent system,

$$\text{MAPD} : \{(\pi_i, \pi_j)\}_{i,j \in \mathcal{N}} \mapsto \mathbb{R}_{\geq 0},$$

where π_i, π_j represent the policies of two agents in a multi-agent system. When the policies π_i, π_j are given, MAPD can output a non-negative scale value, indicating the degree of difference between these two policies. And we hope this computed scale value to possess some distance-like properties.

Properties of MAPD. For ease of exposition, we use π_i and θ_i to denote the policy and the network parameters of an agent i , respectively. The policy distance between agent i and j is represented by d_{ij} , and the properties of MAPD are as follows:

- Symmetry: $d_{ij} = d_{ji}$.
- Non-negativity: $d_{ij} \geq 0$.
- Identicals of indiscernibility (policies): $d_{ij} = 0 \iff \pi_i = \pi_j$.
- Identicals of indiscernibility (parameters): $d_{ij} = 0 \implies \theta_i = \theta_j$.
- Triangle inequality: $d_{ij} \leq d_{ik} + d_{kj}$.

Among the above properties, *symmetry* implies that the policy distance between two agents is not affected by their order. *Non-negativity* specifies that the minimum policy distance is 0, indicating that two policies are identical, which corresponds to *identicals of indiscernibility*. In order to emphasize the distinction between the policies and parameters of agents, we divide the concept of *identicals of indiscernibility* into two aspects: one relating to the policies and the other to the parameters. A policy distance of 0 implies that the policies of two agents are identical, and thus their parameters are also the same, the reverse is not true because the output of the agent’s network may not equal to that of the policy. Therefore, directly comparing parameters of the neural network won’t necessarily indicate policy similarity [11, 24].

The *triangle inequality* is applicable in comparing the policies of multiple agents. We consider this property to be of great importance, since it determines that MAPD is not isolated between two agents, but also considers the influence of other agents. This property allows us not only to estimate other policy distances based on a given set of distances, but also to guide the design of dynamic parameter sharing in Section 5.1.

According to Prop. 2 and Prop. 3, a policy distance equals to 0 when the agent policies are identical, whereas a policy distance bigger than 0 when there are differences in policies. Thus, we need to think about what would cause an increase in the differences of policies, or in other words, an increase in policy distance between two agents. A straightforward approach is to compare the action distributions under the same observations, and accumulate the differences between these distributions. To capture even the slightest differences related to the overall policy, these calculations need to be conducted across the entire observation space. Therefore, a vanilla method [3] of measuring policy distance can be obtained as follows:

$$d_{ij}^{\text{vanilla}} = \int_{\mathcal{O}} f[\pi_i(a|o), \pi_j(a|o)] do, \quad (1)$$

where $\pi_i(a|o)$ denotes the action distribution of an agent i , f is a certain metric for measuring the distance between two distributions, \mathcal{O} is the observation space.

However, **directly measuring the action distributions may not be an ideal approach** due to certain requirements related to the type of distribution. Experiments in [3] assume Gaussian-distributed actions in all continuous settings. Contrarily, agents might take actions under different distributions, such as a bimodal one. When the action distributions between two agents exhibit diverse forms or correspond to different action spaces, measuring the distance between these distributions becomes unfeasible.

Things may get worse in scenarios where agents take discrete actions. Consider a scenario with five discrete actions: given the same observation, an agent A outputs an action distribution of $[1.0, 0, 0, 0, 0]$, indicating a 100% preference for action No.1. Agent B and C have distributions $[0, 0, 0, 1.0, 0]$ and $[0, 0, 0, 0, 1.0]$, respectively. According to the definition of Wasserstein distance (WD) [26] and Hellinger distance (HD) [20], the WD between A - B is three times as big as that between B - C (3 and 1), while both A - B and B - C have identical HD of 1. The crux of this matter lies in the non-sequential nature of action indices, where index-close actions can denote either closely related or vastly different operations.

Hence, our idea is to identify a distribution $p_i(z|o)$ in a latent space to represent the agent’s action distribution $\pi_i(a|o)$. Within this designated latent space, the inherent non-sequential nature can be eradicated, and the types of distributions can also be standardized. This can be achieved by assuming the latent variables following to a certain multi-dimensional prior distribution. We can measure distances between these distributions as an alternative to measuring action distributions, while the semantics of each dimension within this distribution is not of concern to us.

3.2 Learning the conditional representations of agents’ decisions

We hope to find a mapping function, which can output a latent distribution when the observation o and an agent i ’s decision $a_i \sim \pi_i(a|o)$ are given. Thus, this latent distribution can be presented as a posterior $q(z|a_i, o)$ under o and a_i , where q denotes the desired mapping function.

In order to ensure the representational capability of $q(z|a_i, o)$, we employ an auto-encoder architecture to learn q . The encoder, which approximates the posterior distribution, takes a_i and o as input and outputs a distribution of the latent variable z . On the other hand, the decoder tries to reconstruct a_i by drawing samples z from $q(z|a_i, o)$. Solely using z as input is insufficient since the agent’s actions inherently depend on o , so the decoder also incorporates o to reconstruct the action.

However, the auto-encoder framework is widely used as a generative model [17], which predicts the underlying distribution of sample data (in our case $a \sim \pi_i(a|o)$). Since the input action a is sampled from the agent’s action distribution, it requires an ample amount of data to fit the action distribution $\pi_i(a|o_x)$ for agent i under all possible observations o_x , where x denotes the index of observation.

Fortunately, the agent’s policy is fixed during the training of the auto-encoder, meaning the action distributions for each observation

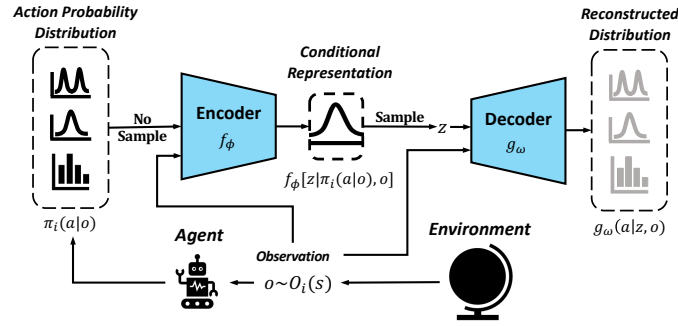


Figure 2: Learning the conditional representation of an agent's decision.

is known. Thus, we directly use $\pi_i(a|o)$ as the encoder's input and explicitly incorporate agent's observation o . The decoder no longer reconstructs the action codes but the action distribution. In this case, the conditional latent variable distribution changes from $q(z|a_i, o)$ to a posterior $q[z|\pi_i(a|o), o]$.

The architecture for learning this conditional representation is illustrated in Figure 2. An agent i receives observation $o \sim O_i(s)$ through its observation function and outputs the corresponding action distribution $\pi_i(a|o)$, where $O_i(s)$ represents the observation function, s is the global state. The encoder f , parameterized by ϕ , accepts inputs $\pi_i(a|o)$ and o , and produces the parameters of the latent variable's posterior probability distribution. The decoder g uses the sampled z and o to yield the reconstructed action distribution, with its parameters denoted by ω . This auto-encoder model is trained by minimizing the following loss function:

$$\begin{aligned} \mathcal{L}(\phi, \omega) = & \mathbb{E}_{z \sim f_\phi[z|\pi_i(a|o), o]} [D_\Pi(g_\omega(a|z, o) || \pi_i(a|o))] \\ & + D_{KL}(f_\phi[z|\pi_i(a|o), o] || p(z|o)), \end{aligned} \quad (2)$$

where the first term is the reconstruction loss, and the second term is the KL loss for bring generated latent distribution $f_\phi[z|\pi_i(a|o), o]$ closer to the prior $p(z|o)$. In our experiments, $p(z|o)$ is a multi-dimensional Gaussian distribution.

In the reconstruction loss term, $D_\Pi(x||y)$ is a generalized distribution distance which measures the similarity between two distributions x and y . In Equation 2, x refers to the reconstructed action distribution $g_\omega(a|z, o)$ and y is the original action distribution $\pi_i(a|o)$. The specific form of D_Π depends on the form of $\pi_i(a|o)$. A question may arise: since learning this representation is to avoid directly measuring the distance between agents' action distributions, why the (action distribution) distance reoccurs here? In fact, this measurement only ensures a close alignment of the reconstructed action distribution with the original one, without requiring the various properties necessary for MAPD. Therefore, D_Π can be a WD (standard distance metrics), a KL divergence (f-divergences), or even the MSE Loss of the action distribution vectors.

3.3 Multi-Agent Policy Distance

After obtaining the conditional representations of an agent's decisions, MAPD uses these learned distributions to replace the action distributions under each observation. Since the action distributions are mapped into a latent distribution in a same feature space, we are able to measure the policy distance between a pair of agents

which have different types of actions. Considering that the observation spaces may also be different, MAPD measures the policy differences between two agents on a common space O_p , which is a set formed by padding the observations of these two agents. The policy distance between two agents following MAPD can be computed as follows:

$$d_{ij} = \int_{O' \subseteq O_p} W[p_i(z|o), p_j(z|o)] do, \quad (3)$$

where $p_i(z|o) = f_\phi[z|\pi_i(a|o), o]$, denotes the latent distribution of agent i under observation o , W denotes a distribution distance, O' is a subset of the observation space O_p . More details about Equation 3 can be found in Appendix 1.1.

In practical applications, it is intractable to compute the integral by traversing the observation space. In fact, the set of observations involved in multi-agent tasks (determined by the initial state and the state transition function) may only constitute a small portion of the entire observation space. Hence, we focus only on the relevant observations $o \in O'$, specifically those that occur in the actual tasks. In experiments, we use Monte Carlo sampling to estimate the integral and employ multi-rollout to reduce estimation variance.

3.4 Case Study of MAPD

In this section, we use a multi-agent spread task [16] as a case study to demonstrate the effectiveness of MAPD. In this scenario, there are 15 agents with 3 colors: the first 5 agents are assigned color No.1, the next 5 and last 5 are assigned color No.2 and No.3, respectively. Each agent must move exclusively towards the landmark matching its assigned color. The positions of the agents and landmarks are randomly initialized, and the agents need to avoid collisions with other agents, while moving towards the corresponding landmarks. The neural networks of all agents are set to be independent, and the parameters are initialized as the same.

Figure 3 shows parts of the policy distance matrices in two scenarios. In scenario (a), the observation spaces of agents are the same, while in scenario (b), the observations are randomly shuffled, resulting in different observation spaces. Both matrices are obtained at 10000 steps during training over 4 random seeds, with only 100 steps of data used for auto-encoder training and policy distance computation. We can observe that each agent has a policy distance of 0 with itself, corresponding to Prop. 3 of MAPD. Agents of the

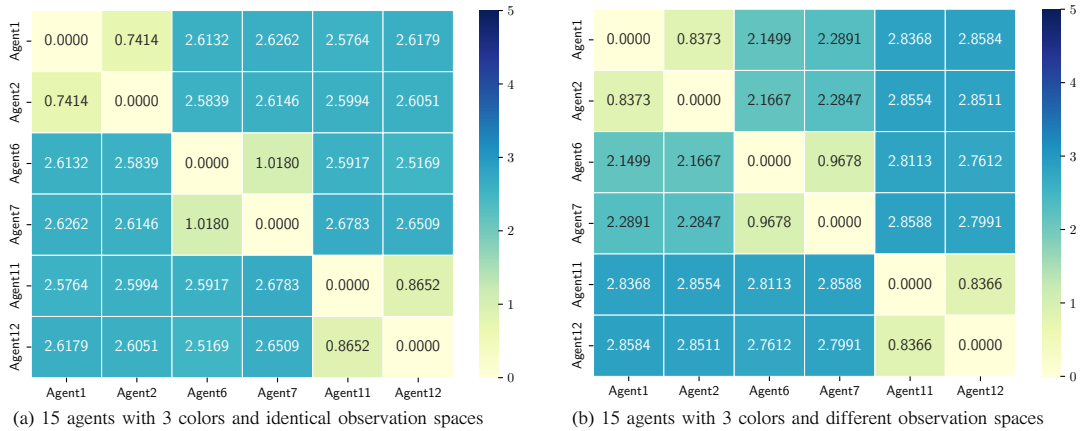


Figure 3: Policy distance matrices in multi-agent spread tasks. In this scenario, there are 15 agents with 3 colors: agents numbered 1-5, 6-10, and 11-15 are each given the colors No.1, No.2, and No.3. The agents must move to the specific landmarks that matches their color. These matrices show the policy distances just for the first two agents in each of colored groups.

same color should move to the same landmark (similar policies), as indicated by their policy distances being less than 1.1.

On the other hand, agents of different colors, due to their different tasks, need to move to different landmarks even under the same observation, resulting in their policy distances all bigger than 2.1. Furthermore, this phenomenon exists regardless of whether the agents have the same observation space. This indicates that the proposed MAPD is able to effectively capture and quantify the differences in agent policies.

4 MEASURING CUSTOMIZED POLICY DISTANCE

In some cases, beyond identifying the overall difference between two agent policies, we may also be curious about their differences in certain aspects. Agents might have distinct patterns in some aspects of policy, while exhibiting similar patterns in others. Measuring the *policy distance in certain aspects* can help us determine in which areas the agents truly differ, and in which they show similar inclinations.

We refer to this special policy distance as the **customized policy distance**. In this section, we will introduce our method of measuring customized policy distance, and give a case study to illustrate its application.

4.1 Learning customized representations

In fact, it is challenging to define and quantify certain aspects of policies. This is because this abstract concept is manually defined, rather than just being a part of the action or the policy. However, these manually defined aspects can be linked to feedback from environments or agents in MARL. For example, if we want to measure the differences in *aggressive* and *defensive* aspects of units' policies in StarcraftII [22], we can find the underlying *bonds* between their actions and the "killing reward" and "defending reward" in the environment, and then use these bonds to measure the differences.

We refer to such information from environments or agents as **customized features**, and the bond as **customized representations**. Customized features possess the following properties:

- Customized features can be modeled as stochastic variables.
- Customized features follow a conditional probability distribution based on observations.

The process of measuring the customized policy distance is similar to the methodology of measuring policy distance in Section 3. Both need to learn a conditional representation of agent's decisions, but the customized one no longer reconstruct the policy itself, but to predict the conditional distribution of the customized features. For a clearer exposition of customized feature and customized policy distance, we denote the customized feature as c , which follows some probability distribution $p(c|o)$ conditioned on observation. In order to learn a meaningful latents representing the customized feature, our objective is to maximize the likelihood $\log p(c|o)$ of all observed c under condition o [17].

To optimize the $\log p(c|o)$, an Evidence Lower Bound (ELBO) of the likelihood can be derived as follows:

$$\log p(c|o) \geq \mathbb{E}_{f_{\phi^c}[z|\pi_i(a|o), o]} \left[\log \frac{p(c, z|o)}{f_{\phi^c}[z|\pi_i(a|o), o]} \right], \quad (4)$$

where $f_{\phi^c}[z|\pi_i(a|o), o]$ represents the posterior probability distribution of the latent variable generated by the encoder, and $p(c, z|o)$ denotes a joint probability distribution concerning the customized feature and latent variable, conditioned on o .

Considering that the ELBO includes an unknown joint probability distribution, we can further decompose it by using the posterior probability distributions from the encoder and decoder:

$$\begin{aligned} & \mathbb{E}_{f_{\phi^c}[z|\pi_i(a|o), o]} \left[\log \frac{p(c, z|o)}{f_{\phi^c}[z|\pi_i(a|o), o]} \right] \\ &= \underbrace{\mathbb{E}_{f_{\phi^c}[z|\pi_i(a|o), o]} [\log g_{\omega}(c|z, o)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(f_{\phi^c}[z|\pi_i(a|o), o] \| p(z|o))}_{\text{prior matching term}}. \end{aligned} \quad (5)$$

In Equation 5, $f_{\phi^c}[z|\pi_i(a|o), o]$ and $g_{\omega}(c|z, o)$ are the posteriors from the encoder and decoder, respectively. The detailed derivation can be found in Appendix 1.2. Thus, the ELBO can be decomposed into a reconstruction term of the customized feature, and a prior matching term of the posterior and the prior. For the training of the auto-encoder, we utilize both the reconstruction and matching components as loss functions.

It is not difficult to notice that the form of Equation 5 is very similar to that of Equation 2. As a matter of fact, the normal policy distance is a special case of the customized distance when c is the action, as the actions follow a conditional distribution based on observations. Furthermore, the theory of customized distance demonstrates the rationality of the loss design in Section 3.2. Similar to MAPD, the customized distance between agents can be computed as follows:

$$d_{ij}^c = \int_{O' \subseteq O_P} W \left[p_i^c(z|o), p_j^c(z|o) \right] do, \quad (6)$$

where d_{ij}^c denotes the customized policy distance between agent i and agent j , $p_i^c(z|o) = f_{\phi^c}[z|\pi_i(a|o), o]$ is the customized representations of agent i 's decisions.

4.2 Case Study of Customized MAPD

To give an example of measuring the customized policy distance, we refer to the multi-agent spread scenarios in Section 3.4. In this section, we measure the policy distances of the multi-agent system based on two aspects: 1) *The tendency of moving towards a same landmark* and 2) *The tendency of moving towards the matching landmark*.

In this case, we choose the distance dis_i between an agent i and the landmark as the customized feature, to link agent's decisions and the specific aspects. This distance is sampled at step $t + 1$ corresponding to the observation o_i^t . Thus, the likelihood term in Equation 4 corresponds to $p(c|o) = p(dis_i^{t+1}|o_i^t)$.

As for the experiments, we use the same parameter settings in Section 3.4, and record customized policy distances of the multi-agent system trained for 10000 steps (only 100 steps of data for auto-encoder training and policy distance computation), the results are shown in Figure 4.

It can be observed that, when considering the tendency of moving towards a same landmark (in our experiments the landmark No.1), the customized policy distances varies significantly depending on the colors of agents. The customized policy distances among agents with different colors are three times as big as that among agents with the same colors. This ratio is similar to the ratio of policy distances shown in Figure 3, indicating that the tendency of spreading towards landmark No.1 is one of the factors contributing to the differences in agent policies.

As for the tendency of moving towards the matching landmarks, all the customized policy distances become extremely small (all less than 0.01). This result is reasonable because all agents have the same goal of moving towards landmarks of their own color, thus they are consistent in this aspect. These results demonstrate the effectiveness of our proposed method, in quantifying the differences in certain aspects of agent policies.

5 DYNAMIC PARAMETER SHARING: AN APPLICATION OF MAPD FOR MARL

Thanks to the low data requirement of MAPD, it becomes possible to deploy MAPD online during the training of multi-agent systems. Having a precise online measurement will enable a real-time evaluation of the policy diversity evolution within the multi-agent system, and such evolution can be intervened if needed. This section provides an application of online deployment of MAPD, multi-agent dynamic parameter sharing (MADPS).

5.1 Multi Agent Dynamic Parameter Sharing

The motivation of MADPS comes from the need for a balance between policy diversity and parameter sharing in multi-agent systems. As parameter sharing can improve the efficiency of MARL, it can also constrain the policy diversity of multi-agent systems, and consequently affect the performance of MARL. If a dynamic parameter sharing can be accomplished during training and agents' policies are appropriately shared when necessary, the multi-agent system can not only adapt to tasks with emergent diversity, but also benefit from the sample efficiency of parameter sharing.

The basic idea of MADPS is to *fuse* the policies of closely related agents while *dividing* the policies of agents with significant differences. We employ the normal/customized policy distance to measure the similarity between agent policies, and conduct a dynamic *fusion* and *division* of agent policies. During each fusion operation, only a portion of the network parameters is shared, to prevent multi-agent policies from getting stuck in local optima. Similarly, during each division operation, we only split a portion of the network parameters. MADPS uses a fusion threshold ϵ_1 and a division threshold ϵ_2 to determine whether fusion or separation should be conducted, as shown in Figure 5.

However, when addressing parameter sharing among multiple agents, the interrelationship between ϵ_1 and ϵ_2 and the order of fusion and division should be considered carefully. Based on the triangular inequality property of MAPD, we indicate that the division threshold should satisfy $\epsilon_2 \geq 2\epsilon_1$. In the experiments, we directly set $\epsilon_2 = 2\epsilon_1$ to reduce the number of hyperparameters. Additionally, an algorithm for the fusion and division of multi-agent policies is developed, in order to avoid various corresponding contradictions (refer to Appendix 2 for algorithm details and relevant explanations).

Thus, we introduce this algorithm of multi-agent policy fusion and division in MARL, resulting in MADPS. MADPS allows a multi-agent system to begin training with flexible forms of parameter sharing initialization (e.g., fully shared policies, independent policies, or selective parameter sharing [4].) In every T time steps, MADPS computes a policy distance matrix of the multi-agent system. By comparing the policy distances with predefined thresholds, MADPS conducts appropriate fusion and division of policies, until a stable parameter sharing structure is achieved, striking a balance between parameter sharing and policy diversity.

6 EXPERIMENTS

In Section 3.4 and Section 4.2, we employ a toy environment to demonstrate the effectiveness of MAPD. In this section, the online deployment of MAPD and MADPS are applied to more challenging

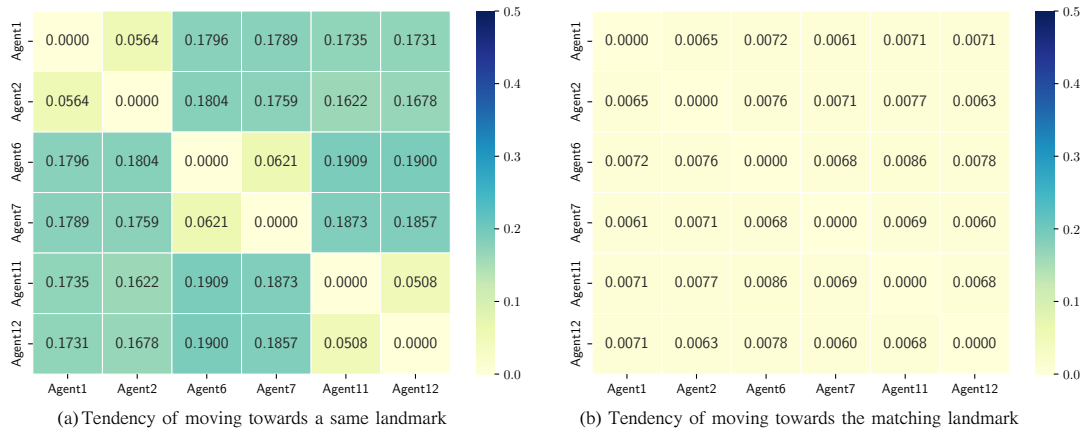


Figure 4: Customized policy distance matrices in multi-agent spread tasks. Figure (a) demonstrates the policy distances between agents on the tendency of moving towards a same colored landmark, figure (b) demonstrates the policy distances on the tendency of moving towards the matching landmark.

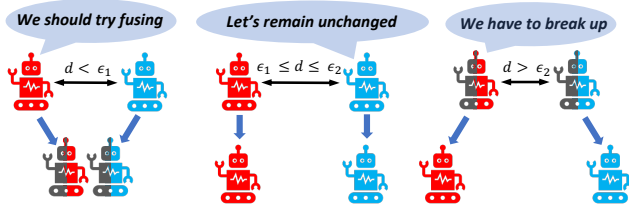


Figure 5: The basic idea of dynamic parameter sharing.

multi-agent spread tasks and StarCraft II micromanagement environment (SMAC). By comparing MADPS with other methods of parameter sharing, we analyze the performance improvement of MARL algorithms through appropriate policy diversity and parameter sharing, and demonstrate the superior performance of MADPS based on MAPD. The code is available here¹.

6.1 Experiment Settings

Multi-agent environments. This part introduces the testing environment for our method. We increase the difficulty of the original multi-agent spread task, by requiring the agents not only to spread towards corresponding landmarks, but also to gather as closely as possible once they reach the landmarks. Besides the original penalties for collisions and guidance, a sparse reward for encourage agents moving closely is added in the overall rewards. This sparse reward is only given to agents when the distance between the agents and the landmark is smaller than a threshold, increases as the distance between the agents decreases, and decreases with the number of agents (which meeting the threshold condition) decreases. Another testing environment used in this study is SMAC. All the testing methods are evaluated on four super-hard maps in SMAC, namely, *corridor*, *6h_vs_8z*, *3s5z_vs_3s6z* and *MMM2*. The first two maps involve homogeneous agent scenarios, while the last two represent scenarios with heterogeneous agents.

¹<https://github.com/Harry67Hu/MADPS>

Baselines. We compare the proposed dynamic parameter sharing with other methods of parameter sharing. To ensure fairness, all methods employ the same actor-critic architecture for training, and the network structures and hyperparameter settings are kept identical. (More details can be found in Appendix 3.2). The baselines can be summarized as follows:

- No parameter sharing (NPS): each agent has an independent network with no parameter sharing.
- Fully parameter sharing (FPS): all agents fully share the same network parameters.
- Fully parameter sharing with index (FPS-id): all agents share the network parameters, but a unique index is attached to each agent’s observation.
- Selective parameter sharing (SePS): before the training of MARL, agents are clustered according to learned latent representations, and parameter sharing is restricted to each agent cluster [4].

6.2 Superior Performance of Our Method

Our method and the baselines are tested on four super-hard maps in SMAC and a series of harder multi-agent spread tasks. The variations in multi-agent spread tasks include the number of agents, the number of agent colors, the allocation of agent colors, and the structure of agent observations.

We represent a scene with x agents and y colors as xa_{yc} , and use *super* to denote scenes where agent colors are not evenly allocated, and *shuffle* to indicate scenes where agent observations are randomly shuffled. For example, *15a_3c_shuffle* represents a scene with 15 agents, where the agents are evenly allocated among 3 colors based on their indices (the agents numbered 1-5, 6-10, and 11-15 are given the colors No.1, No.2, and No.3, respectively), and the agent observations are shuffled.

Figure 6 illustrates the performance of all methods on a selection of the above environments. More experimental results can be found in Appendix 3.2.

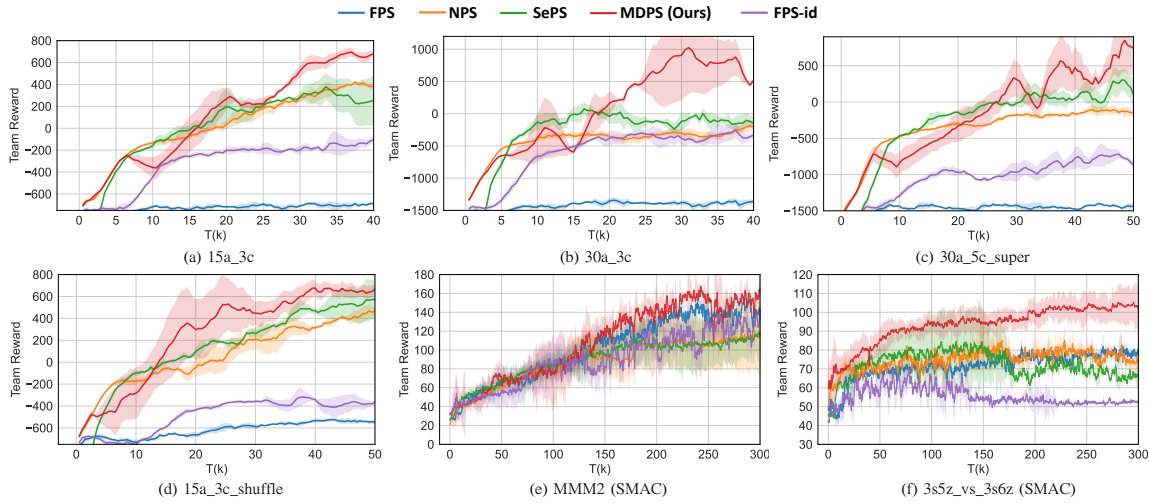


Figure 6: Performance comparison with baselines on a series of multi-agent spread tasks and super-hard tasks in SMAC.

It can be observed that FPS cannot learn effective policies in all multi-agent spread tasks. Because in these scenarios, agents of different colors move in different directions due to guided penalties, but these experiences are inappropriately used in the training of the shared network, causing the policy to fluctuate between different goals and leading to training failure. The curves of FPS-id can converge to a value close to 0 in scenarios (a) and (b), but they still fail to trigger the designed sparse rewards. Additionally, in the *shuffle* scenario (d), the shared network cannot effectively recognize the types of agents due to the shuffled observations, thus preventing the learning of effective policies.

On the other hand, NPS can achieve a mediocre performance on all tasks. In multi-agent spread tasks, this approach actually allows agents to learn unique policies, thus achieving a overall behavior diversity in multi-agent systems. But in other tasks such as *MMM2*, the training of NPS is slow due to its independent network setting, and it cannot use the experience of other agents to train networks.

By analyzing the curves of NPS, FPS, and FPS-id on different tasks, we can gain a rough understanding of how important policy diversity is in each task. For instance, in all multi-agent spread tasks, policy diversity is crucial for improving algorithm performance, but excessive policy diversity can hinder agents from achieving higher rewards. Additionally, the performance difference between FPS and FPS-id is much smaller in *MMM2* compared to *3s5z_vs_3s6z*. This indicates that although the both are heterogeneous scenarios, the task of *3s5z_vs_3s6z* requires more policy diversity for agents than *MMM2*.

Both MADPS and SePS encourage policy diversity among agents, but our MADPS uses a smarter approach. SePS utilizes each agent’s trajectory to learn a latent representation of its index, and employs this representation for clustering, thus dividing agent network parameters based on groups. However, such grouping is fixed during training, and may be overly deliberate, resulting in less favorable experimental performance in some scenarios. To ensure a good performance, SePS requires prior knowledge of each scenario, necessitating a setting of unique hyperparameters for each scenario.

On the other hand, MADPS only employs one hyperparameter across a series of scenarios, and can automatically adjust the parameter sharing scheme during training.

These experimental results demonstrate the effectiveness of the idea, i.e., dynamically adjusting parameter sharing in MARL. The essence behind this idea is the indispensable need for appropriate policy diversity as well as parameter sharing techniques in MARL. Thanks to our proposed MAPD, we are able to accurately capture the differences in policies between agents. This enabling a further understand of policy diversity for specific tasks, and provides valuable guidance for the design of MARL algorithms in practical applications.

7 CONCLUSION

This paper proposes MAPD, a method of quantifying differences in policies of MARL. Instead of directly measuring the policies, MAPD learns conditional representations of agents’ decisions, and computes the policy distance by integrating the distances of the latent distributions. This approach is applicable to various policy forms of multi-agent systems. The variant of MAPD, customized policy distance, allows for measuring the behavioral differences between agents on specific aspects of policy. Based on MAPD, we design a dynamic parameter sharing algorithm, which can achieve a dynamic fusion and division of agents’ policies during the MARL training. Experimental results demonstrate that our method can accurately quantify the diversity in agent policies, and MADPS outperforms other baseline methods of parameter sharing. We believe that MAPD can serve as an important tool for studying and leveraging behavioral diversity in multi-agent systems, ultimately promoting the advancement of MARL.

ACKNOWLEDGMENTS

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27030204, the National Natural Science Foundation of China under Grant 62073323, and the Beijing Nova Program under Grant 20220484077.

REFERENCES

- [1] Akshat Agarwal, Sumit Kumar, and Katia Sycara. 2019. Learning transferable cooperative behavior in multi-agent teams. *arXiv preprint arXiv:1906.01202* (2019).
- [2] Matteo Bettini, Ajay Shankar, and Amanda Prorok. 2023. Heterogeneous Multi-Robot Reinforcement Learning. *arXiv preprint arXiv:2301.07137* (2023).
- [3] Matteo Bettini, Ajay Shankar, and Amanda Prorok. 2023. System Neural Diversity: Measuring Behavioral Heterogeneity in Multi-Agent Learning. *arXiv preprint arXiv:2305.02128* (2023).
- [4] Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. 2021. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*. PMLR, 1989–1998.
- [5] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. 2016. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*. PMLR, 1329–1338.
- [6] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems* 29 (2016).
- [7] Aditya Grover, Maruan Al-Shedivat, Jayesh Gupta, Yuri Burda, and Harrison Edwards. 2018. Learning policy representations in multiagent systems. In *International conference on machine learning*. PMLR, 1802–1811.
- [8] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers 16*. Springer, 66–83.
- [9] Siyi Hu, Chuanlong Xie, Xiaodan Liang, and Xiaojun Chang. 2022. Policy diagnosis via measuring role diversity in cooperative multi-agent rl. In *International Conference on Machine Learning*. PMLR, 9041–9071.
- [10] Haobin Jiang, Yifan Yu, and Zongqing Lu. 2021. Metric Policy Representations for Opponent Modeling. *arXiv preprint arXiv:2106.05802* (2021).
- [11] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. 2019. Policy consolidation for continual reinforcement learning. *arXiv preprint arXiv:1902.00255* (2019).
- [12] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 3991–4002.
- [13] Dong Li, Dongbin Zhao, Qichao Zhang, and Yaran Chen. 2019. Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. *IEEE Computational Intelligence Magazine* 14, 2 (2019), 83–98.
- [14] Xiangyu Liu, Hangtian Jia, Ying Wen, Yujing Hu, Yingfeng Chen, Changjie Fan, Zhipeng Hu, and Yaodong Yang. 2021. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. *Advances in Neural Information Processing Systems* 34 (2021), 941–952.
- [15] Yuntao Liu, Yuan Li, Xinhai Xu, Donghong Liu, and Yong Dou. 2022. ROGC: Role-Oriented Graph Convolution Based Multi-Agent Reinforcement Learning. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [16] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).
- [17] Calvin Luo. 2022. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970* (2022).
- [18] Muhammad A Masood and Finale Doshi-Velez. 2019. Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies. *arXiv preprint arXiv:1906.00088* (2019).
- [19] Kevin R McKee, Joel Z Leibo, Charlie Beattie, and Richard Everett. 2022. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 21.
- [20] Frank Nielsen and Sylvain Boltz. 2011. The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory* 57, 8 (2011), 5455–5466.
- [21] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research* 21, 1 (2020), 7234–7284.
- [22] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).
- [23] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
- [24] Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. 2019. Designing neural networks through neuroevolution. *Nature Machine Intelligence* 1, 1 (2019), 24–35.
- [25] Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems* 29 (2016).
- [26] Cédric Villani et al. 2009. *Optimal transport: old and new*. Vol. 338. Springer.
- [27] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. 2020. Roma: Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039* (2020).
- [28] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. 2020. Rode: Learning roles to decompose multi-agent tasks. *arXiv preprint arXiv:2010.01523* (2020).
- [29] Chongjie Zhang and Victor Lesser. 2013. Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 1101–1108.