

Multi-Agent Collaboration with Ergodic Spatial Distributions

Derrik E. Asher^{*a}, Erin Zaroukian^a, Brandon Perelman^a, Jake Perret^b, Rolando Fernandez^a, Blaine Hoffman^c, Sebastian S. Rodriguez^d

^aUS CCDC Army Research Laboratory, ^bUniversity of Maryland; ^cUS CCDC Data Analytics Center, ^dUniversity of Illinois, Urbana-Champaign

Abstract

When considering collaboration among agents in multi-agent systems, individual and team measures of performance are used to describe the collaboration. Typically, the definition of collaboration is limited in that it is only indicative of coordination required for a small class of tasks wherein this coordination is necessary for task completion (e.g. two or more agents needed to lift a heavy object). In this work, we aim to present a method that may be used to classify individual and group behaviors, enabling the measurement of collaboration among agents. We demonstrate the capability to use performance and behavioral data from computational learning agents in a predator-prey pursuit task to produce ergodic spatial distributions. Ergodicity is shown quantitatively and used to benchmark performance. The ergodic distributions shown, reflect the learned policies developed through multi-agent reinforcement learning (MARL). We also demonstrate that independently trained models produce distinctly different behavior, as revealed through ergodic spatial distributions. The ergodicity of the agents' behavior provides both a potential path for classifying group behavior, predicting performance of group behavior with novel partners, and a quantifiable measure of collaboration built from explicitly aligned goals (i.e., cooperation) as a result of behavioral interdependencies.

Keywords— predator-prey pursuit, collaboration, coordination, simulation, spatial distributions, ergodicity, multi-agent, reinforcement learning

1. Introduction

The term collaboration is often ill-defined in multi-agent tasks and is typically used to imply that a benchmark level of performance has been achieved by a group of agents in cooperative domains. We reinterpret collaboration as the state achieved by a group of agents performing a task with aligned goals (i.e., cooperation) that exhibit measurable and observable characteristics (i.e., coordination) that result from convergent performance [1]. This definition for collaboration indicates that agents must execute measurable behaviors that account for the actions of their partners and the environment [2], [3]. Only recently has a repeatable method for measuring collaboration in multi-agent systems from behavior alone been shown [4], which is the inspiration for the current work.

In stochastic decision-making, if the probability of each event (i.e., decision selected) only depends on the state from the previous event, the process is considered Markovian. In other words, a Markov Decision Process (MDP) has the statistical properties of a Markov chain (i.e., transition density). This indicates that as time becomes large, an MDP's transition density will converge to its ergodic (limiting) distribution [5]. In the context of this paper, we observe different samples from fixed ergodic distributions of Markov chains to represent measurable behaviors of agents.

MDPs assume, among other things, stationary dynamics [6] and full observability [7]. Using RL in multi-agent domains typically violates the assumptions of MDPs. However, recent algorithmic approaches have shown that these multi-agent MDP violations can be overcome by using a centralized training scheme, as is the case with the multi-agent deep deterministic policy gradient (MADDPG) RL algorithm [8]. MADDPG utilizes an actor-critic with deep neural networks representing both the policy (actor) and Q-learners (critic). Multi-agent behaviors are achieved by passing information about each agent's state and actions to each critic network during training and removing the critic networks during testing. As such, the learning agents are joint action learners as opposed to independent learners, giving them an advantage when coordinated behaviors (i.e., behaviors through action interdependencies related to task performance) lead to better solutions. We hypothesize that during centralized training in multi-agent learning, such as the predator-

prey pursuit task, adversaries develop policies that coordinate with each other [9], [10]. However, although we emphasize that adversaries’ behavior can coordinate, we also stress that they may not cooperate due to the opposition of goals, and therefore do not collaborate.

The current work aims to demonstrate a novel collaboration measure by embedding the MADDPG RL algorithm in a simulated predator-prey pursuit task, which has an aligned goal for the collaborating predators. With data from this task, we conduct comparisons of agents’ learned behavior as represented by heat maps of their spatial locations throughout testing episodes, where noticeably different heat maps are interpreted as showing different learned behavior (agent’s policy) developed through training. We further speculate on how mixing agent policies (i.e., allowing the mixing of agents between independently trained models) can lead to better team performance. We hypothesize that understanding the link between these low-dimensional representations of agent behaviors and team performance will lead to a prediction of the types of human behavior that positively impact coordination in human-agent teams. The resulting improved coordination may optimize group performance in spatial tasks with shared goals. We first established that the learned behaviors of 10 independently trained models can be depicted as low-dimensional representations of trained policies, which we refer to as ergodic spatial distributions. We demonstrate, 1) that these distributions exhibit ergodicity and are quantitatively different from a non-ergodic model (random walk), 2) the ergodic nature of the distributions through detailing a single agent across 4 different data sets, and 3) that agents learned different ergodic spatial distributions while achieving the same level of performance across 10 independently trained models. From this work, we predict that group performance (and collaboration) with novel partners may be analytically evaluated with ergodic distributions.

2. Methods

Simulations consisted of a set of agents (depicted as circles) performing a predator-prey pursuit task in a continuous bounded 2D environment [8], [11]. Three predator agents scored points (i.e., received reward during training and tracked performance during testing) each time they came in contact with a prey agent (hit the prey agent). Predator agents had identical capabilities to one another (i.e., same size, velocity, and acceleration limitations), whereas the single prey agent was 50% smaller, had a 33% acceleration advantage, and a 30% max velocity advantage. All agents had the same mass, with minimal elasticity to provide a small bump force upon collisions. The simulation environment was adapted from OpenAI Gym [12] and was developed for the multi-agent deep learning algorithm discussed below [8]. Given the prey’s distinct movement advantage, it is inferred that the predator agents must work together (i.e., cooperate) to contact the prey, and a simple greedy agent policy (e.g., minimize distance to target) has been shown in previous work to be insufficient for predator success [1].

As stated previously, a multi-agent deep deterministic policy gradient (MADDPG) algorithm was used to train all agents simultaneously [8]. Predator agents received the same fixed reward when any one of them made contact with the prey agent (shared or joint reward), and the prey agent received the negative of the same fixed reward when contacted. The MADDPG algorithm utilizes a centralized critic network with access to all agents’ states and actions to train agents. This centralized training paradigm allows agents to develop policies based on other agent observations and actions. Agents’ positions and velocities were randomized at the start of each episode, and their accelerations were set to zero. Each agent’s state space contained its velocity, absolute position in the environment, relative distance to predators, and the prey’s velocity. The action output for an agent was accelerations in the x and y directions. During training, the number of episodes (10^5 episodes) and episode duration (25 timesteps) were based on previous work with convergent performance [1], [3], [4], [11] [13]. During testing (i.e., when learning was disabled and network weights were no longer changing), agents were run in a decentralized fashion where each agent’s policy depended only on their local observations (i.e., state space). Test data was collected from 10 independently trained models (each model had three predators and one prey guided by the MADDPG algorithm).

To demonstrate ergodicity, a single prey agent was selected from 1 of the 10 trained models and compared against a known non-ergodic random walk model across four different sets. Four different data sets were collected for both the trained and random models, where each data set differed in number of episodes and duration per episode, keeping the total number of timesteps constant at 1M (10^6): Data Set 1 (1 episode at 10^6 [1M] timesteps); Data Set 2 (100 episodes at 10^4 [10K] timesteps); Data Set 3 (10^3 [1K] episodes at 10^3 [1K] timesteps); Data Set 4 (10^4 [10K] episodes at 100

timesteps). In the case of the trained model, the primary differences between the four data sets are the randomized initial conditions and the stochasticity in multi-agent MDPs leading to different trajectories and resulting performance per episode. Trained model performance was measured as the collective number of contacts that predators had with the prey per episode, aggregated across all episodes and shown in the results section (see III. Results).

Given ergodicity in all trained models, 3 of the 10 models were selected to demonstrate observed differences in their respective spatial distributions per agent. These ergodic spatial distributions can be interpreted as low-dimensional representations of the agents’ learned policies. This indicates that the spatial distributions may provide insight into how RL agents with a centralized training scheme perform a pursuit task in a coordinated fashion. Further, the observed predator agent distributions may represent distinct approaches or group strategies for achieving collaboration, whereas the prey agents’ spatial distributions show some aspect of the resulting adversarial solution to the predators’ collaborative behavior.

3. Results

First, ergodicity is demonstrated by comparing a trained model to a random model, known to be non-ergodic, and verified through visual inspection. This similarity across data sets is quantified using RV coefficient [14], a correlation coefficient similar to the coefficient of determination that is frequently used to compare heat maps (i.e., matrices) in neuroimaging studies, and explored further through distributions of prey position (see Tables 1 and 2). Figure 1 shows the Model 1 prey agent’s movement, represented as discrete spatial locations (thus referred to here as spatial distributions) through all episodes and timesteps (1M total timesteps), along with the spatial locations of the prey upon contact with predators (referred to as contact distributions) across four data sets. Figure 1 demonstrates the remarkable similarity between both spatial locations and locations where the predators contacted the prey (hits) given the variability in number of episodes, episode length, randomized initial conditions (i.e., starting locations of predators and prey were uniformly randomized at the beginning of each episode), stochasticity of agent trajectories through episodes, and performance per episode. These results described herein provide the motivation for exploring the ergodic nature of agent behavior in multi-agent RL.

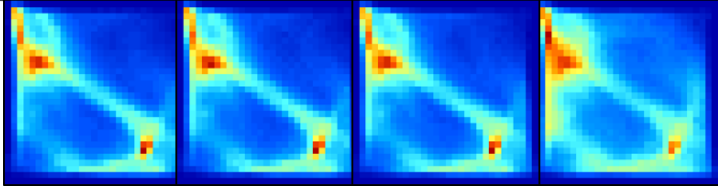
3.1 RV Coefficients and Ergodicity

In order to provide evidence that RL model-generated distributions exhibit ergodicity, we calculate and compare their RV coefficients to those generated by a random walk model. To achieve this, we implemented a simple random walk model in a 30x30 discretized environment that was permitted to move one cell per time step. This model was run using the same parameters as the RL model to produce four data sets with the same episode and time step variants described in the Methods section. Similarity between data sets is quantified by greater RV coefficients. In other words, we would expect RV coefficients calculated from ergodic distributions to be substantially greater than those calculated from non-ergodic distributions.

Table 1 shows a correlation matrix of the four data sets’ spatial distributions produced by the trained model (i.e., the RL-trained prey agent from Model 1). The RV coefficient values shown in Table 1 (see right column) suggest that shorter episodes (i.e., 100 timesteps) may be weighted by randomized initial positions of agents per episode. RV coefficients of at least .949 for spatial location were achieved for all pairwise comparisons (see Table 1), supporting the interpretation that these distributions have an extremely high similarity in shape and position and, more specifically, are ergodic. Further, it appears that spatial location distributions generated from episodes of at least 1K timesteps were sufficient to obtain RV coefficients = .999 with distributions containing longer episodes (see Table 1).

Table 1. RV Coefficient (similarity) between 2D heat maps (30x30 bins) of prey spatial distributions generated from 4 different data sets (pairwise comparisons of left column in Figure 1).

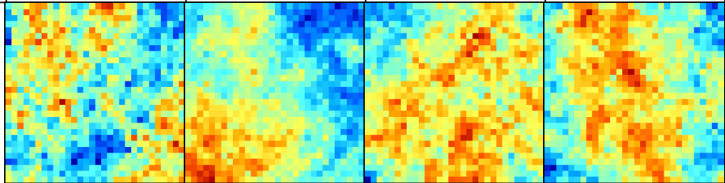
RV Coefficient Between Spatial Distributions Across 4 Data Sets (Model 1, Prey)				
	Data Set 1 (1 E, 1M ts)	Data Set 2 (100 E, 10K ts)	Data Set 3 (1K E, 1K ts)	Data Set 4 (10K E, 100 ts)
Data Set 1 (1 E, 1M ts)	-	.999	.999	.949
Data Set 2 (100 E, 10K ts)	-	-	.999	.953
Data Set 3 (1K E, 1K ts)	-	-	-	.961
Data Set 4 (10K E, 100 ts)	-	-	-	-



We demonstrate ergodicity in these distributions by comparing them to spatial distributions generated by a random walk model under similar conditions. Table 2 shows the same four data set parameters used to produce the distributions shown for the random walk model (i.e., non-ergodic distributions). In general, the random walk model's RV coefficient values in the correlation matrix are substantially lower than those of the ergodic spatial distributions. The random walk model results (Table 2) are provided for comparison to ergodic distributions as a baseline for evaluating RV coefficient values (Tables 1 and 2), congruence values (Tables 3 and 4), and visual comparisons (Figure 1) across the four data sets. Furthermore, the values do not support any trend related to episode length; for the random walk model, the RV coefficient values do not decrease with decreasing episode length (as would be expected due to distribution noise).

Table 2. RV Coefficient (similarity) between 2D heat maps (30x30 bins) of random walk model spatial distributions generated from 4 different data sets (pairwise comparisons for the heat maps shown below the respective columns).

RV Coefficient Between Spatial Distributions Across 4 Data Sets (Random Walk)				
	Data Set 1 (1 E, 1M ts)	Data Set 2 (100 E, 10K ts)	Data Set 3 (1K E, 1K ts)	Data Set 4 (10K E, 100 ts)
Data Set 1 (1 E, 1M ts)	-	.643	.346	.621
Data Set 2 (100 E, 10K ts)	-	-	.339	.412
Data Set 3 (1K E, 1K ts)	-	-	-	.509
Data Set 4 (10K E, 100 ts)	-	-	-	-



3.2 Ergodic Distributions

The RV coefficients show a high similarity between the models of agent behavior and difference from that generate by a random walk model, and the data can further be explored through comparison of the prey contact and prey spatial distributions across the data sets. Ergodicity in prey contact distributions (Figure 1, right column) is more visually remarkable than the observed similarity between prey spatial distributions (Figure 1, left column) because the contact distributions show the spatial locations throughout episodes where a predator achieved the task goal, effectively demonstrating a 2D representation of the solution space for the predators, likely dependent on the learned behavior (policies) of all agents (coordination between predators and prey). In contrast, the spatial distributions only show the learned behavior (policy) of a single agent (although the distributions have inherent dependencies due to centralized training and adversarial learning). Therefore, it is expected that similarity between contact distributions of the respective data sets shown in Figure 1 (right column) should be less similar than a single agent (we explore this more in section 3.3). In addition, the number of data points in the spatial distributions is 10^6 , whereas, the number of data points in the contact distributions is under 10^5 (Data Set 1: 90752, Data Set 2: 90704, Data Set 3: 90357, and Data Set 4: 84245) per data set. Not surprising, but interesting to note the dependence that performance has on the length of an episode (about 1% lower performance for shorter episodes - 100 timesteps).

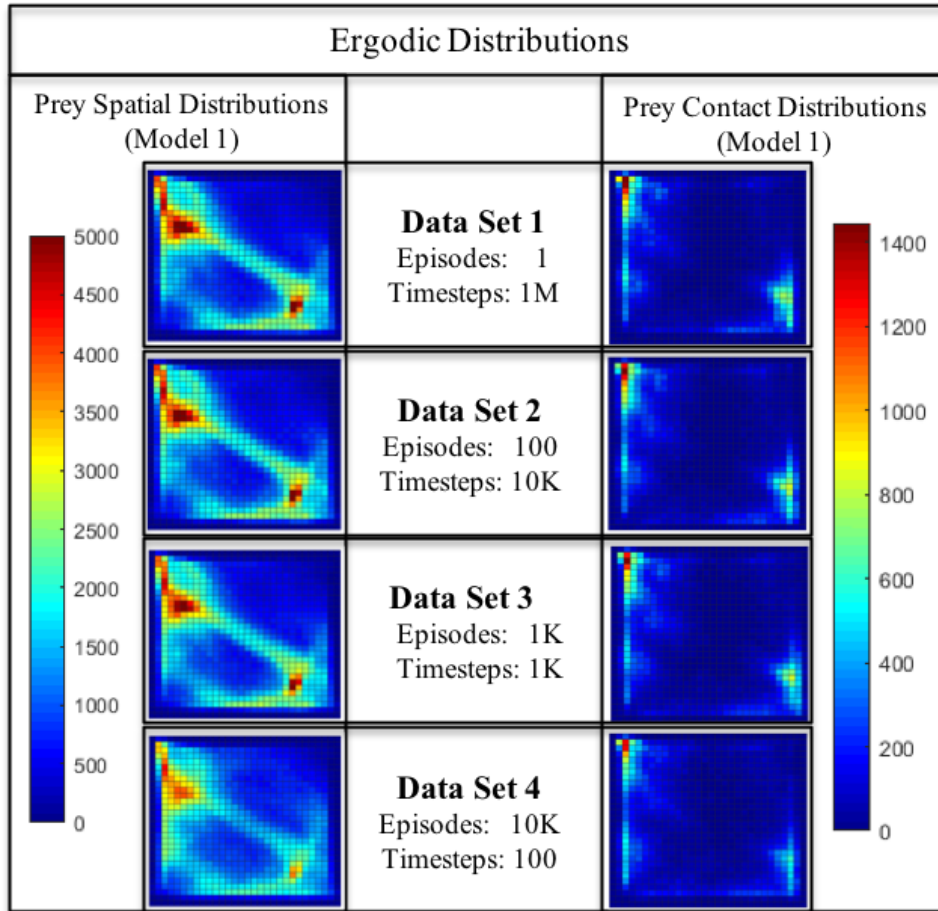


Figure 1. Ergodic distributions for the prey agent from Model 1 across 4 data sets. Left column shows 2D histograms (30x30 bins) of the prey's spatial locations across 1M timesteps with heat maps overlaid and a corresponding colorbar to the left ranging between [0, 5000]. The center column indicates the data set parameters per row defined by the number of episodes and timesteps per episode. The right column similar to the left column shows 2D histograms (30x30 bins) of the prey's locations upon coming in contact with any predator (hits) across all episodes and timesteps, with corresponding colorbar ranging between [0, 1450].

3.3 Simple Congruence Metrics

For these data sets, similarity is realized as the congruence of position data of the prey and instances of contact. In other words, higher similarity means that the relative position of agents at a timestep in one data set can be found in another. As the RV coefficients show (Table 1), the distributions are very close, and we expect that they are also congruent. Upon inspection of the spatial distributions collected from the 4 data sets, it can be seen that greater episode length leads to stronger similarities in the ergodic distributions (see Figure 1, Data Sets 1-3 vs. Data Set 4), which is confirmed by the similarity values shown in Tables 3 and 4 (compare right columns in Tables 3 and 4). These results suggest that shorter episodes (i.e., 100 timesteps) may be more heavily weighted by the randomized initial positions of the agents per episode, effectively adding noise to the ergodic distributions. Although the magnitude (i.e., height of the distribution peaks or histogram bins) of the distributions from Data Set 4 do not match those of the other data sets with longer episodes (i.e., 1M, 10K, and 1K), the respective shapes of the distributions (for spatial-left and contact-right locations of the prey across all episodes) are strikingly similar, with at least 80% congruence for spatial locations to the others (see Table 3) and at least 70% congruence for contact locations (see Table 4). Further, it appears that an episode length of at least 1K timesteps was sufficient to ascertain a congruence of greater than 96% for spatial locations (see Table 3). Note that the congruence similarity in Table 3 aligns well to the RV coefficients in Table 1.

Table 3. Percent similarity between 2D histograms (30x30 bins) of prey spatial distributions generated from 4 different data sets (pairwise comparisons of left column in Figure 1).

Similarity Between Prey Spatial Distributions Across 4 Data Sets (Model 1)				
	Data Set 1 (1 E, 1M ts)	Data Set 2 (100 E, 10K ts)	Data Set 3 (1K E, 1K ts)	Data Set 4 (10K E, 100 ts)
Data Set 1 (1 E, 1M ts)	--	96.92%	96.24%	81.58%
Data Set 2 (100 E, 10K ts)	--	--	96.49%	82.07%
Data Set 3 (1K E, 1K ts)	--	--	--	84.00%
Data Set 4 (10K E, 100 ts)	--	--	--	--

As noted previously, the similarity values shown in Table 4 are substantially less than those in Table 3 due to at least 2 reasons, 1) there is an order of magnitude difference in number of data points between spatial distributions (1M) and contact distributions ($< 100K$), and 2) the contact distributions can be thought of as taking into account the spatial locations and time (i.e., when a predator contacted the prey) with a low-dimensional representation of the solution space for the predators, which is dependent on 2 or more agents occupying the same point in space at the same time. Therefore, it is not surprising that the ergodic contact distributions are less similar to each other than the spatial distributions (compare Table 3 to Table 4).

Table 4. Percent similarity between 2D histograms (30x30 bins) of prey contact distributions generated from 4 different data sets (pairwise comparisons of right column in Figure 1).

Similarity Between Prey Contact Distributions Across 4 Data Sets (Model 1)				
	Data Set 1 (1 E, 1M ts)	Data Set 2 (100 E, 10K ts)	Data Set 3 (1K E, 1K ts)	Data Set 4 (10K E, 100 ts)
Data Set 1 (1 E, 1M ts)	--	78.03%	72.75%	75.86%
Data Set 2 (100 E, 10K ts)	--	--	86.17%	72.91%
Data Set 3 (1K E, 1K ts)	--	--	--	72.83%
Data Set 4 (10K E, 100 ts)	--	--	--	--

3.4 Ergodicity Between Models

In order to determine whether the agents' behaviors were consistent or different across trained models, we compared the ergodic distributions, representing learned behaviors, of a few select different models. For this comparison, we selected Models 1 through 3, Data Set 3 (which corresponded to 1K episodes at 1K timestamps per episode). Figure 2 shows the ergodic spatial distributions for these models, each of which were independently trained. It is interesting to note that at least one predator for every model shown in Figure 2 learned a spatial distribution that mimics their respective prey (see Figure 2, Model 1 - Predator 1; Model 2 - Predator 1; Model 3 - Predators 2 and 3). In contrast, it also appears that at least one predator per model learned to mimic some rotation of their respective prey's spatial distribution (see Figure 2, Model 1 - Predator 2; Model 2 - Predator 3; Model 3 - Predator 2). It is important to note that the predator agent numbers do not have any significant meaning (e.g., Predator 1 has a specific role), indicating that the predator agents should be treated as homogeneous agents and any predator from a model might be compared to a predator from a different model. With homogeneity between predators in mind, Figure 2 shows similar ergodic spatial distribution patterns for all predators of Models 1 and 3. This is likely due to the symmetrically similar (roughly 90 degree counterclockwise rotation from Model 1 to Model 3) ergodic spatial distributions learned by their respective prey (compare Figure 2, Model 1 - Prey to Model 3 - Prey).

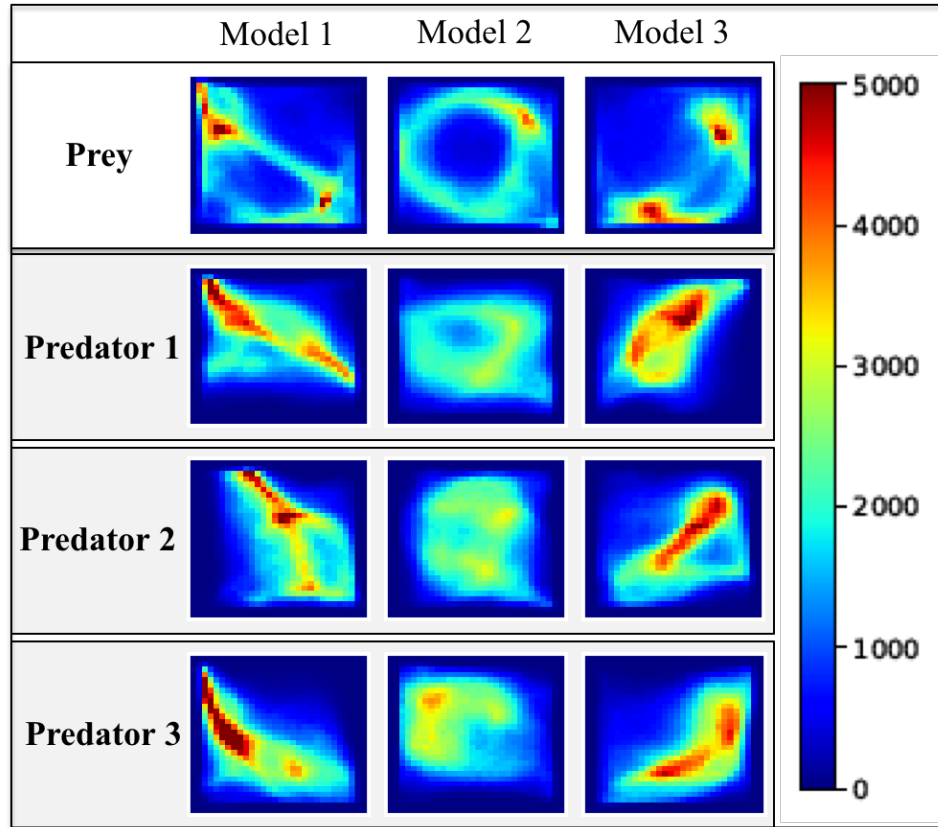


Figure 2. Ergodic spatial distributions for all agents (1 prey agent and 3 predator agents) from 3 representative models of the 10 independently trained models. Each column shows 2D histograms (30x30 bins) representing spatial distributions for the respective agents (rows) with heat maps overlaid and a corresponding colorbar ranging between [0, 5000] (right). The respective 2D histograms were generated from 3 data sets consisting of 1K episodes at 1K timestamps per episode.

Performance plots showing the distributions for hits per episode over 1K episodes at 1K time steps per episode are shown across the 10 independently trained models in Figure 3. The colored plots with legend associations correspond to the three representative models shown in Figure 2. No significant differences were found between the performance

distributions with 2-sample Kolmogorov-Smirnov tests (KS tests) across the 10 models ($p > 0.9$ for all pairwise comparisons). The light gray lines are shown to visualize the 7 models not shown in Figure 2. It is interesting to note that the peaks of the performance distributions align for Models 1 and 3 (see Figure 3, aligned peaks at 90 on the x-axis for Models 1 and 3), given that the spatial distributions for their respective prey agents shown in Figure 2 are symmetrically similar (appear to have the same shape with a 90 degree rotational difference). Although not significantly different, it is also interesting to note that the performance for Model 2 resulted in a peak shifted to the left (see Figure 3 peak for Model 2 at 80). This perhaps provides some insight into how the different observed ergodic spatial distributions correspond to different peaks in model performance, suggesting that the peak of model performance might indicate the group strategy employed by the collaborative predators and their corresponding (coordinated) adversarial prey.

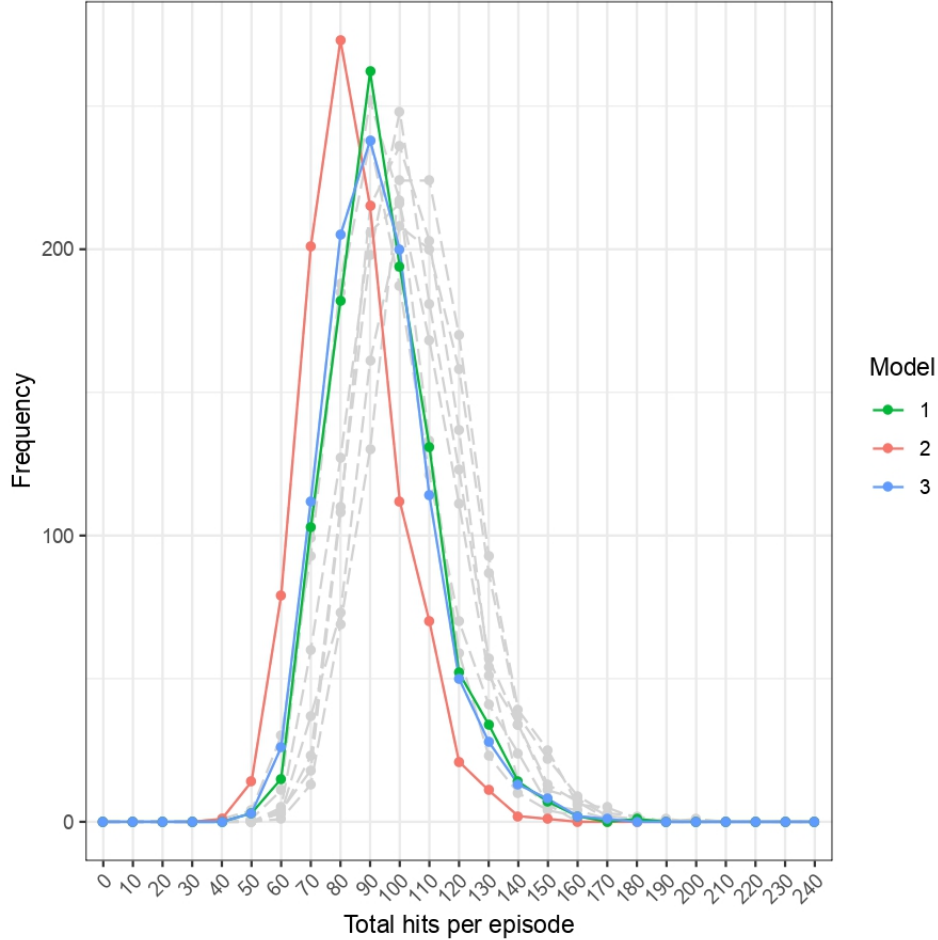


Figure 3. Performance line plots for the 10 independently trained models. These plots show the distribution per model for predator agent hits (i.e., contacts with the prey agent) per episode over 1K episodes at 1K timesteps per episode. Each line plot is a histogram showing the performance of a select model in terms of contacts per episode. The x-axis (Total hits per episode) shows histogram bins (bin width of 10) for total number of predator hits on the prey per episode, while the y-axis (Frequency) shows the count of episodes that reached the corresponding x-axis performance for each data point (filled circles) in the histogram line plot. The performance line plots in color correspond to the 3 representative models shown in Figure 2. The gray dashed line plots represent the other 7 trained models.

Together these results are intended to show our motivation for exploring ergodicity (see Figure 1 and Tables 1 – 4). The similarities and differences observed in ergodic (see Figure 2 and Tables 1, 3, and 4) and non-ergodic (see Table 2) distributions generated from independently trained models, along with the performance distributions associated with the learning agents' ergodic distributions (see Figure 3), are showing a consistent metric dependent on multi-agent training that may be useful for understanding and controlling the behavior of agents in multi-agent systems. In summary, these

results illuminate the possibility of using ergodic distributions for, 1) exploring coordination in multi-agent systems (cooperative and/or adversarial), 2) evaluating or identifying different learned group behaviors or strategies emerging from the use of MARL, and 3) describing collaboration with a combination of performance and ergodic distributions.

4. Discussion

In this article, we explored the stochastic decision-making property of ergodicity associated with MDPs in a 2D continuous multi-agent predator-prey pursuit task. 10 models (each consisting of 3 homogeneous predator agents and 1 faster prey agent) were trained independently to convergent performance with a centralized MARL algorithm (MADDPG). As expected, ergodicity was encountered in distributions across every measurable agent dimension (i.e., position, velocity, acceleration, and prey contacts/hits) in each model (all data was not reported). Ergodic distributions were demonstrated by selecting the prey agent from a model (Model 1), collecting 4 different data sets with a varying number of episodes and time steps per episode, and quantitatively comparing the distributions across the data sets (see Figure 1, Tables –1, 3, and 4). To further demonstrate ergodicity, a random walk model was run with the same data set parameters to contrast ergodicity with non-ergodicity (see Table 2). To explore what agents learned with convergent performance, ergodic spatial distributions were compared across 3 of the 10 models to show observable similarities and differences (see Figure 2). Performance distributions showed that non-significantly different model performances could result in very different agent behaviors (see Figures 2 and 3). This work provides evidence that ergodic distributions may be used to understand, classify coordinated or adversarial behaviors of agents trained with centralized MARL algorithms, or possibly even control or modify the behavior of agents for general collaboration with other agents (human or independently trained agents).

The ergodic distributions shown so far in this article have allowed us to observe some characteristics of collaboration (i.e., coordination) among the predators working towards a shared goal within the predator-prey pursuit task (see Figure 2). Because we have observed this characteristic of collaboration it should be measurable but we have yet to explore various techniques to quantify collaboration with respect to ergodic distributions. As an example of the observed coordination, each of the predators appears to select a region of the play area (left/top, right/bottom, or center) to attempt to control or corral the prey into an advantageous position for one of them to achieve their aligned goal (i.e., cooperation) for a hit. Leveraging ergodic distributions, we may be able to provide either the predators or the prey with information about their partners and/or opponents in novel environments, potentially allowing for a reduction in training time and/or for new behaviors to emerge.

Exploring the use and exploitation of ergodicity from MDPs in MARL provides a potential method for controlling agents, predicting the impact of agent behavior in novel situations, and integrating the behaviors of agents with novel partners (including humans). Ergodic distributions are static and therefore provide both a level of understanding and an opportunity for exploitation. On the understanding side, we can use MARL to train agents to work together but do not have sufficient methods for assessing the solutions encountered. Classification of ergodic distributions can provide insight or possibly even conclusive evidence for the emergent individual or group strategies learned to complete tasks. In addition, we may classify visually or otherwise superficially different, yet equivalent (performance non-significantly different), solutions encountered through the learning process (cooperative or adversarial). On the exploitation side, we can evaluate ergodic distributions to predict if a novel partner can increase, decrease, or have no effect on group performance. We may be able to use ergodic distributions of an adversary as a training signal for maximally exploitative supervised training. Further, we could use ergodic distributions to help human partners understand something about the expected behavior of computational partners. Finally, we might be able to identify ergodic distributions in humans for optimal integration with computational partners.

More explicitly we propose future research to investigate, 1) whether ergodic distributions may be used to predict the behavior of learning agents (collaborative or adversarial) when paired with new partners (i.e., agents that were trained separately), 2) that the property of ergodicity may be used to predict how trained agents may pair with human partners in collaborative paradigms, and 3) that an adversary may be exploited if its ergodic distributions are known, leading to improved performance without additional training of learning agents. Furthermore, we will evaluate the feasibility of visualization techniques to quantify and facilitate human understanding of similarity among resultant agent ergodic distributions. These potential avenues of ergodic distribution exploration were not evaluated in this paper. Instead, we have provided evidence to support the potential exploitation of ergodicity in learning agents.

In another extension of this paradigm, we intend to investigate the impact of novel partners on ergodic distributions. Given that an ergodic distribution is a low-dimensional representation of an agent’s learned policy, it is not clear that the ergodic distributions described and evaluated in the current paper have some flexibility to them. In other words, it is possible that manipulations to the environment or behaviors of agents may change the ergodic distributions (at least to some degree). If this is not the case, then we can truly exploit these distributions, otherwise, we need to understand how flexible ergodicity is in these trained policies. Therefore, it is imperative that we continue to explore the flexibility of ergodic distributions resulting from MDPs to understand how best to use them in future work.

5. References

- [1] S. L. Barton, N. R. Waytowich, E. Zaroukian, and D. E. Asher, “Measuring Collaborative Emergent Behavior in Multi-agent Reinforcement Learning,” in *Human Systems Engineering and Design*, Cham, 2018, pp. 422–427, doi: 10.1007/978-3-030-02053-8_64.
- [2] G. Klien, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, “Ten challenges for making automation a ‘team player’ in joint human-agent activity,” *IEEE Intell. Syst.*, vol. 19, no. 6, pp. 91–95, Nov. 2004, doi: 10.1109/MIS.2004.74.
- [3] E. Zaroukian *et al.*, “Algorithmically identifying strategies in multi-agent game-theoretic environments,” in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, May 2019, vol. 11006, p. 1100614, doi: 10.1117/12.2518609.
- [4] D. E. Asher, M. Garber-Barron, S. S. Rodriguez, E. Zaroukian, and N. R. Waytowich, “Multi-Agent Coordination Profiles Through State Space Perturbations,” *Springer*, 2020.
- [5] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- [6] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, “Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems,” *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 1–31, 2012.
- [7] C. C. White, “A survey of solution techniques for the partially observed Markov decision process,” *Ann. Oper. Res.*, vol. 32, no. 1, pp. 215–230, 1991.
- [8] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” presented at the Advances in neural information processing systems, 2017, pp. 6379–6390.
- [9] S. L. Barton, N. R. Waytowich, and D. E. Asher, “Coordination-driven learning in multi-agent problem spaces,” *ArXiv Prepr. ArXiv180904918*, 2018.
- [10] V. Behzadan and A. Munir, “Models and Framework for Adversarial Attacks on Complex Adaptive Systems,” *ArXiv Prepr. ArXiv170904137*, 2017.
- [11] S. L. Barton, E. Zaroukian, D. E. Asher, and N. R. Waytowich, “Evaluating the Coordination of Agents in Multi-agent Reinforcement Learning,” in *Intelligent Human Systems Integration 2019*, Cham, 2019, pp. 765–770, doi: 10.1007/978-3-030-11051-2_116.
- [12] G. Brockman *et al.*, “Openai gym,” *ArXiv Prepr. ArXiv160601540*, 2016.

- [13] D. Asher, S. Barton, E. Zaroukian, and N. Waytowich, "Effect of cooperative team size on coordination in adaptive multi-agent systems," presented at the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, 2019, vol. 11006, p. 110060Z.
- [14] P. Robert and Y. Escoufier, "A unifying tool for linear multivariate statistical methods: the RV-coefficient," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 25, no. 3, pp. 257–265, 1976.