Entropy-Based Local Fitnesses for Evolutionary Multiagent Systems

Ayhan Alp Aydeniz
Collaborative Robotics and Intelligent Systems Institute
Oregon State University
Corvallis, Oregon, USA
aydeniza@oregonstate.edu

Anna Nickelson
Collaborative Robotics and Intelligent Systems Institute
Oregon State University
Corvallis, Oregon, USA
nickelsa@oregonstate.edu

Kagan Tumer
Collaborative Robotics and Intelligent Systems Institute
Oregon State University
Corvallis, Oregon, USA
kagan.tumer@oregonstate.edu

ABSTRACT

Evolutionary multiagent systems have been successfully applied to many real world problems, including search and rescue and ocean exploration. However, as the number of agents increases in such problems, the evaluation function captures an individual agent's fitness less and less accurately. As a consequence, agents adopt a small set of acceptable behaviors that are neither optimal nor robust to environmental changes or teammate failures. Fitness shaping, intrinsic fitnesses, or multi-fitness learning alleviate some of these concerns but generally require domain knowledge or the functional form of the evaluation function. In this paper, we introduce Entropy-Based Local Fitnesses (EBLFs) that generate diverse behaviors for agents and produce robust team behaviors without requiring environmental knowledge. The key contribution of EBLFs is to inject a dense, entropy-based fitness into the agents' evolution without interfering with the sparse, high-level system evaluation function. Our results show that the agents using EBLFs learn new skills in difficult environments with sparse feedback without requiring domain knowledge. In addition, EBLFs generated new team-level behaviors that were not defined by a human operator, but beneficial to robust team performance.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**.

KEYWORDS

Evolutionary robotics, Multi-agent systems, Neuroevolution

ACM Reference Format:
Ayhan Alp Aydeniz, Anna Nickelson, and Kagan Tumer. 2022. Entropy-Based Local Fitnesses for Evolutionary Multiagent Systems. In *Genetic and Evolutionary Computation Conference Companion (GECCO '22 Companion), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3520304.3529035

GECCO '22 Companion, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9268-6/22/07.
https://doi.org/10.1145/3520304.3529035

1 INTRODUCTION

Multiagent robotic systems enable humans to conduct remote research in challenging environments, such as exoplanets and deep ocean. Research in these remote environments often consists of long-term tasks that are difficult to define in advance. Agent teams must have the capacity to explore and learn new team-level policies, as well as individual agent-level behaviors. Each agent must explore a diverse set of individual behaviors in order to distinguish those that contribute most to the team objectives.

Evolutionary methods learn through a team fitness score to encourage team-level learning. However, when the team fitness is sparse, agents struggle to generate effective local behaviors. Multiagent Evolutionary Reinforcement Learning (MERL) [6] offers a partial solution by combining the local gradient-based learning of Reinforcement Learning (RL) with the power of an Evolutionary Algorithm (EA). Agents learn local behaviors trained on agent-specific fitnesses using a gradient-based RL algorithm; team behaviors are learned through an EA whose objective is to maximize a sparse global team fitness and agent-specific behaviors are learned through the gradient-based learner. However, the local learner requires a pre-defined objective which is not capable of capturing dynamics of an environment with complex tasks.

Entropy maximization is a promising technique that encourages agents to learn diverse policies. Entropy-based approaches have gained traction in single agent systems; prior works maximize the entropy of policies through variance in the neural network weights or variability in the states (and/or actions) that policies visited [2]. Entropy-based methods are able to achieve unsupervised emergence of diverse skills; however, they do not directly extend to multiagent settings due to the added complexity of team dynamics.

In this paper we introduce Entropy-Based Local Fitnesses (EBLFs) for multiagent systems. Our method combines the diversity learned via an entropy-based learner with the power of MERL. We first semantically disambiguate states to distinguish "novel" states, then measure the entropy of the distribution of those states. As a result, an agent receives contribution of this state to the entropy as a local fitness. EBLFs enable agents to generate diverse agent-specific policies beneficial to the desired team behavior.

The key feature of EBLFs is to shift the paradigm of finding good domain-specific fitness functions to evolve agent skills to generating *domain-independent* fitness functions that rate the *exploration skills* of the agents. Our main contribution is to introduce a new fitness structure, Entropy Based Local Fitnesses (EBLFs), that enables agents to learn diverse skills in order to collectively solve complex tasks.

2 BACKGROUND AND RELATED WORK

In this section, we provide background information on methods used and concepts promoting diversity via entropy maximization as related works.

2.1 Evolutionary Reinforcement Learning

Both reinforcement learning (RL) algorithms and evolutionary algorithms (EAs) have strong advantages that has been shown through their successful applications [4]. RL based algorithms are able to learn fast, but they do not perform well when the fitnesses are sparse [10]. The population-based approach of EAs is advantageous for generating diverse experiences [1], especially incorporating the idea of *searching for novelty* [5]. Evolutionary Reinforcement Learning (ERL) [4] combines these two approaches to leverage the advantages of each. The EA generates a diverse set of experiences to train the RL agent (any RL agent utilizing an actor critic) and the gradient information of the RL agent is reincorporated in the EA. The transmission between these two is done through a replay buffer where every experience after each time step is stored. Because EA optimizes the global reward given at the end of each episode, it biases the exploration to the states that contribute most to long-term objectives. A version of this algorithm used in multiagent systems is Multiagent Reinforcement Learning (MERL) [6]. MERL utilizes a gradient-based algorithm to learn agent-specific skills and the EA learns the team skills by optimizing the global reward. In this paper, we utilize MERL as a learning framework, due to its two-tier structure and biasing towards the states having long-term returns. In our experiments, we use TD3 as the gradient-based learner of MERL.

2.2 Entropy Maximization in Reinforcement Learning

Designing a fitness function requires domain-specific objectives and pre-definition of certain behaviors [3]. However, capturing every need of a domain is not always possible, due to large state-spaces and the effect of complexity on the definition of a task. Entropy maximization in reinforcement learning solves this issue by providing learning frameworks utilizing discriminators or entropy functions that are used to learn without rewards [2]. In reinforcement learning, entropy is typically defined as the randomness of a skill, a skill set, or a policy (including neural network networks).

In multiagent systems, learning through entropy maximization presents challenges, as the learning agents must also account for team coordination to achieve a global task. In this work, we leverage entropy to generate dense local rewards.

3 ENTROPY-BASED LOCAL FITNESSES

The core idea which we rely on is that maintaining diversity through the states visited by agents will provide more diverse policies to the evolutionary algorithm. Throughout the paper, we define an *agent-specific behavior* as a policy that visits different states until an episode ends. We encourage agents to learn to visit as many different states as they can and distinguish the states that contribute most to the global team objective. The goal is to provide more diverse and discriminable policies to the evolutionary algorithm. This, in turn, will have significant impacts in achieving the global objective.

In continuous domains, it is uninformative to use raw sensor values to differentiate states, as each new sensor reading will look like a new state. Therefore, we first apply *quantization* methods to distinguish significantly different states. Second, we measure the entropy over a memory which is a history of states visited throughout an episode. We compute entropy-based local fitnesses (EBLFs) with an objective to maximize the entropy over this memory; therefore, we provide the contribution of an action to the entropy as fitnesses.

In this paper, we can define our path to diversity in policies of multiagent systems as:

- Explore *novel* states
- Learn policies that have more diversity at agent level
- Expand diversity of agents' experiences through evolution

3.1 Entropy Maximization

Entropy is a measure of information. We encourage agents to have more distinguishable policies and to search more information in the environments with high uncertainty and unknowns. Agents' own positions, the other agents' positions, the positions of the tasks can be seen as uncertainties and unknowns in an environment.

Our definition of an *agent-specific behavior* leads us to keep a memory of state observations for each agent and we measure the entropy within these memories. Because each policy will result in a different distribution within these memories, maximizing diversity within these memories also enhances diversity within these policies.

$$SMemory_{t_j} = \{s_1, \ldots, s_i, \ldots, s_{t_j}\} \tag{1}$$

Our aim is to minimize the recurrence of a state, s_i in $SMemory_j$. Shannon's entropy of $SMemory_j$ is mathematically described as,

$$H(SMemory_{t_j}) = -\sum_{i=1}^{n_{t_j}} p_{s_i} \ln p_{s_i} \tag{2}$$

where $H(SMemory_{t_j})$ is the entropy of the memory that has the observations generated by a policy until the time step, t_j, n_{t_j} is the amount of all observations generated until time, t_j, and p_{s_i} is the probability of the state observation [9]. It is important to note that EBLFs are not the entropy of $SMemory_j$, but they are the probabilistic contribution of an action to this entropy. Our objective minimizes p_{s_i}, thus $H(SMemory_{t_j})$ of equation 2. Maximum entropy is reached by a skill that is able to achieve a uniform distribution over the observation vector, $SMemory_j$. Hence, the probability of a randomly chosen s_i needs to be low to maximize the diversity generated by a policy.

$$p_{s_i} = \frac{n_{s_i}}{n_{t_j}} \tag{3}$$

In the Equation 3, n_{s_i} is the frequency (or the recurrence) of states. State quantization helps us to calculate the frequency of states, because we aim to use EBLFs in continuous domains.

$$f_{local_s} = \frac{1}{n_{s_i}} \tag{4}$$

In order to provide fitness minimizing the occurrence of a state, s_i, we design our fitnesses as given in the equations 4. Later, the entropy can be measured as shown in Equation 5.

$$H(SMemory_{t_j}) = -\sum_{i=1}^{n_{t_j}} \frac{n_{s_i}}{n_{t_j}} \ln \frac{n_{s_i}}{n_{t_j}} \tag{5}$$

To give an example, the policy generated $SMemory_{t_3}$ already visited the state, s_y; therefore, we do not want the agent to visit a similar or the same state again. Because it adds one more s_y to its state vector below at t_4, it receives a discounted reward, $1/2$, given as below.

$$SMemory_{t_3} = \{s_x, s_y, s_z\} \tag{6}$$

$$SMemory_{t_4} = \{s_x, s_y, s_z, s_y\} \tag{7}$$

$$f_{local_{t_4}} = \frac{1}{2} \tag{8}$$

However, when agents enter states that have useful information like a target, they receive the given reward. In this paper, domain used in the experiments has different points of interest (POIs) [6]. We incorporated the value of the POIs visited to the reward via multiplying with the value of sub-tasks, so that the agents are encouraged to visit the states having benefit to the overall team behavior.

4 EXPERIMENTS

In our experiments, we adopt the environment used in the works [6, 8]. This allows us to test agents learning through EBLFs in a continuous multiagent domain.

4.1 Multi-Rover Exploration Domain

We use Multi-Rover Exploration Domain in the traditional settings where there are multiple rovers and multiple points of interest (POI)s in a continuous environment. The rovers need to cooperate with each other to achieve a team goal. Each POI can be seen as a sub-task and must be observed simultaneously by a number of rovers, which is determined by the coupling requirement. Rovers are not given any information about the other teammates or the POIs and there is no explicit communication among the agents.

4.2 Experimental Design

We design experiments to compare our approach, S-EBLFs, against the domain-specific objective function used in the paper proposed MERL, denoted D-MERL [6]. Through our experiments, we show that achieving the global goal without any domain-specific knowledge helps agents build a robust team behavior.

$$f_{local_s} = \frac{1}{n_{s_i}} V_{POI} \tag{9}$$

Equation 9 shows how POI values incorporated into the equation 4 to compute EBLFs by a simple multiplication. This simple modification to the fitness functions are of significant impact on the robustness of team behavior and we will discuss it further in our results.

4.2.1 Parameters: We outline three sets of parameters used for our experiments: **pre-defined fitness**, **state quantization**, and **environmental settings**.

Pre-Defined Fitness for Comparisons: In multi-rover exploration domain, the objective is to observe POIs as a team of rovers; therefore, Khadka *et al.* [6] defines the objective function as to minimize the distance to the closest POI as defined in the work [6].

$$f_{merl} = \frac{r_{act}}{d_{POI_c}} \tag{10}$$

where f_{merl} is the fitness function used in the paper [6], r_{act} is the activation radius of a POI, d_{POI_c} is the distance to the closest POI, POI_c. Here, minimizing distance is what we define as domain-specific objective, because the global objective is to observe POIs within a certain radius. To compute this fitness function, an operator needs to have more information than an agent can provide through its sensor. For example, f_{merl} requires finding the closest POI which is not captured directly in rover domain where sensors capturing the ratio of the value of a POI to the square of its distance.

State quantization is used to make states more significantly distinguishable in our work (to distinguish pragmatically and semantically different states). In our experiments, we set the level of quantization as 1. In other words, if a sensor captures a POI or another teammate, convert the sensory value to 1, otherwise convert to 0. Increasing this level increases resolution of an observation; therefore, the observation space used to compute EBLFs expands significantly. The probabilistic values used in the equations 2 decrease as the observation space expands, so observations generated through a policy will converge to a uniform distribution, as we increase the quantization level. Though this sounds promising, this will result in premature increase in the entropy and fitnesses will become less informative; as a result, learning will decrease.

Environmental settings in multi-rover domain is crucial as it allows us to test the number of tasks and varying features of tasks. In the first setting, we test 6 agents and 4 POIs (randomly distributed and all with the same value) in a 15x15 environment and we conduct experiments to test how agents handle **coupling** (complexity). As our second setting of experiments, we use 6 agents and 6 POIs (randomly distributed POIs with the value of 2 on a inner circle, and randomly distributed POIs with the value of 5 on a outer circle) in a 20x20 environment where we test agents within more **diverse tasks** and **tight-coupling** to see how EBLFs help agents explore more.

4.2.2 Experiments: our experiments demonstrate the impact of EBLFs on three factors: **tight-coupling**, **task diversity**, and **team behavior**.

Tight-coupling in multiagent systems requires agents to coordinate to contribute to the overall team behavior and we use it as a measure of task complexity throughout this paper. In the experiments, we test EBLFs to show how agents learning through entropy-based fitnesses generate more robust team behaviors. Increasing the coupling factor does not expand the state-space, thus we only consider its effect on the task complexity. Team performance will certainly decrease, when we increase the coupling factor because we do not provide more mission time to achieve tasks.

Task diversity is a inevitable aspect of most of the real world problems. Multiple tasks may have varying outcomes. The experiments testing EBLFs with a diverse set of tasks includes POIs having different values. We designed a task configuration to test the effect of EBLFs on exploration within the second environmental settings. Additionally, we demonstrate how exploration is affected by the complexity of tasks.

Team behavior is a crucial as we design EBLFs for complex multiagent systems. In our experiments, we compare the learned team behavior learned by teams using D-MERL and EBLFs within second environmental settings.

5 RESULTS AND DISCUSSION

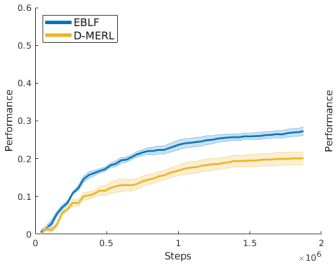

Figure 1: 6 Rovers and 4 POIs (Randomly Dist.) Coupling = 4

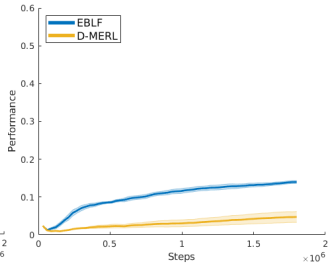

Figure 2: 6 Rovers and 6 POIs (3 POIs on inner and 3 POIs on outer circles) Coupling = 4

The performance metric is calculated as: The team with the highest fitness, *champion*, is selected in every generation, then it was tested on 10 rollouts in the environment. Then, we recorded the average score achieved by the *champion*. The x-axis shows the number of steps taken in the environment (a local reward is given). These metrics and the parameters of MERL are determined according to the work [6]. Each plot is generated with the average of 5 statistical runs.

In the Figures 1, and 2, EBLFs outperform the pre-defined fitness, D-MERL. Particularly, in the circular POI settings, agents using EBLFs learn a unexpected specific behavior that allows them to explore POIs with higher values, whereas agents using D-MERL are not capable of exploring under more complex task. The Figure 3 shows how agents using EBLFs are more scalable than D-MERL agents as the task complexity increases.

6 CONCLUSION AND FUTURE WORK

This work considers a simple and novel fitness structure for multiagent systems that enables agents to generate diverse agent-specific policies beneficial to the desired team behavior. Our method is based on entropy maximization. Unlike previous entropy-based methods in RL, EBLFs do not require any modification on a learning framework. We propose EBLFs within MERL framework and show that agents can generate diverse behaviors without relying on any domain-specific objective.

Our results show that pre-defined objectives that heavily rely on domain-specific information do well, but only when the environments are relatively simple. As we move agents learning through

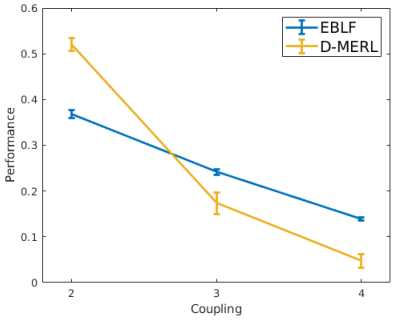

Figure 3: Slopes represent scalability to more complex tasks for EBLFs and D-MERL - 6 Rovers, 6 POIs (3 POIs on inner and 3 POIs on outer circles) - Coupling 2, 3, 4

these objectives to more complex tasks, they struggle or fail to solve these tasks as a team. EBLFs are able to achieve similar performance in simple domains and maintain higher performance as the environmental complexity increases. The results are proof of the concept that learning with no pre-defined fitnesses results in unsupervised emergence of new behaviors that contribute to team behavior in evolutionary multiagent systems.

As future work, we propose investigating the effect of EBLFs on neural networks to show how EBLFs can contribute adaptive team behavior. Some evolutionary concepts like phenotypic plasticity of neural networks [7] or agents' policies can be applied throughout EBLFs and show how agents trained through EBLFs can adapt to different environments as well (like Transfer Learning).

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under an AI Institute grant No. 2112633 and by the Air Force Office of Scientific Research under grant No. FA9550-19-1-0195.

REFERENCES

[1] Dave Cliff, Phil Husbands, and Inman Harvey. 1993. Explorations in evolutionary robotics. *Adaptive behavior* 2, 1 (1993), 73–110.
[2] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070* (2018).
[3] Atil Iscen, Ken Caluwaerts, Jonathan Bruce, Adrian Agogino, Vytas SunSpiral, and Kagan Tumer. 2015. Learning tensegrity locomotion using open-loop control signals and coevolutionary algorithms. *Artificial life* 21, 2 (2015), 119–140.
[4] Shauharda Khadka and Kagan Tumer. 2018. Evolution-guided policy gradient in reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1196–1208.
[5] Joel Lehman, Kenneth O Stanley, et al. 2008. Exploiting open-endedness to solve problems through the search for novelty.. In *ALIFE*. Citeseer, 329–336.
[6] Somdeb Majumdar, Shauharda Khadka, Santiago Miret, Stephen Mcaleer, and Kagan Tumer. 2020. Evolutionary Reinforcement Learning for Sample-Efficient Multiagent Coordination. In *International Conference on Machine Learning*. PMLR, 6651–6660.
[7] Stefano Nolfi, Orazio Miglino, and Domenico Parisi. 1994. Phenotypic plasticity in evolving neural networks. In *Proceedings of PerAc'94. From Perception to Action*. IEEE, 146–157.
[8] Aida Rahmattalabi, Jen Jen Chung, Mitchell Colby, and Kagan Tumer. 2016. D++: Structural credit assignment in tightly coupled multiagent domains. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4424–4429.
[9] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
[10] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.