Diversity Through Exclusion (DTE): Niche Identification for Reinforcement Learning through Value-Decomposition

Extended Abstract

Peter Sunehag
DeepMind

Alexander Sasha Vezhnevets
DeepMind

Edgar A. Duéñez-Guzmán
DeepMind

Igor Mordatch
Google Brain

Joel Z. Leibo
DeepMind

ABSTRACT

Many environments contain numerous available niches of variable value, each associated with a different local optimum in the space of behaviors (policy space). In this work we propose a generic reinforcement learning (RL) algorithm where multiple sub-policies are learnt in a manner inspired by fitness sharing in evolutionary computation and applied in reinforcement learning using Value-Decomposition-Networks in a novel manner for a single-agent's internal population. Further, we introduce an artificial chemistry inspired platform where it is easy to create tasks with multiple rewarding strategies utilizing different resources (i.e. multiple niches). We show that agents trained this way can escape poor-but-attractive local optima to instead converge to harder-to-discover higher value strategies in both the artificial chemistry environments and in simpler illustrative environments.

KEYWORDS

Reinforcement Learning, Artificial Chemistry, Ecology

ACM Reference Format:
Peter Sunehag, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Igor Mordatch, and Joel Z. Leibo. 2023. Diversity Through Exclusion (DTE): Niche Identification for Reinforcement Learning through Value-Decomposition: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

This is a short version of [11]. The principle of competitive exclusion in ecology says that two species cannot inhabit the same niche, or put differently, cannot rely on the same set of limited resources [13]. The repulsive effect that an occupied niche has on nearby species who might otherwise have occupied it is an important part of the reason why evolution on Earth produced a diversity of species [6], not just a single generically optimal species like you might expect from a black box optimization algorithm.

Here we study how ecological concepts of competitive exclusion, niche discovery, and diversity can be productively applied to a reinforcement learning (RL) agent operating in a complex environment. While evolution may be understood as a "space-filling" process, which over time generates a species for every as-yet-unfilled point in niche space [8], the standard image of an RL agent's learning process is the trajectory of a single point traveling through policy space by following a gradient and, after enough time, arriving at a globally optimal rest point [12]. In most of the RL literature, the central obstacle acknowledged to getting RL agents to discover radically new behaviors that a random policy would never emit by chance, is that of sparse reward and vanishing gradients when far from a local optima. This is why the field has been mainly interested in models where an exploration bonus is added to the environment's default reward [1, 9]. The aim is to create new gradients where none would otherwise have existed (e.g. [3]). In this paper, we explore how the ecological perspective may provide an alternative organizing metaphor that can be used productively to motivate a new reinforcement learning algorithm.

2 DIVERSITY THROUGH EXCLUSION (DTE)

We consider a single agent as though it contains multiple cooperating agents. Thus it makes sense to view it through the Dec-POMDP framework of cooperative MARL and rely on the approach of Value-Decomposition-Networks (VDN) [10], though here the implications of the VDN approach are different from its original setting. Since all the "sub-agents" are inside one agent, and only one can act at a time, credit should flow only to the agent who actually acted. Please note, we will below be assuming for simplicity that all rewards are non-negative. Our approach is based on sampling a policy π_i from a population for each episode and using it to generate an experience trajectory. These policies will, in our experiments, be defined by N different policy heads on top of a common core encoder network similar to Bootstrapped DQN [7].

Applying the value decomposition networks approach [10] we let the global state-action value function Q be decomposed as

$$Q(o(s), a) = \sum_{i=1}^{N} Q_i(o_i(s), a_i)$$

where Q_i is the state-action value function of policy i but here $o_i(s) = o(s)$ and $a_i = a$ since all sub-agents experience the same observations and actions.

In our case, we have more information than in cooperative MARL. For each state s we know which sub-agent j was responsible for selecting the joint action $\vec{a} = a_j(o(s))$ and we can directly use this knowledge to assign credit. This amounts to updating the responsible sub-agent toward a target using the true sequence of rewards while simultaneously updating all other sub-agents with the same observation and action sequence but all rewards replaced by 0.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

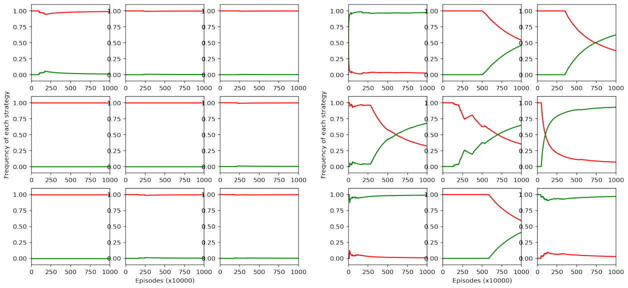

Figure 1: 9-headed DQN (left) vs 9-headed DTE (right) on Simple Artificial Chemistry

Thus for every trajectory $(o(s), a)$ and policy i, a learning update is performed where a gradient step is taken for the loss

$$L(o(s), a, i, \{Q_j\}_j, R) = (Q_i(o(s), a) - R)^2 + (\sum_{j \neq i} Q_j(o(s), a))^2,$$

which by the triangle inequality is not smaller than the standard VDN loss $(\sum_j Q_j(o(s), a) - R)^2$.

3 EXPERIMENTAL EVALUATION USING AN ARTIFICIAL CHEMISTRY PLATFORM

To enable the creation of rich environments with many completely different rewarding strategies we create an artificial chemistry platform where environments are defined by reaction graphs [2, 5] (see Fig. 2) and reactions occur stochastically when reactants are brought near one another. Our main environment is inspired by work on autocatalytic sets (RAF sets) [5]. The most interesting environment we consider here used here involved metabolic cycles aimed at being what Hordijk et al. [5] calls Reflexively Autocatalytic and Food generating (RAF). The notion of [5], that autocatalytic systems have such RAF subsystems, has critically guided our task design.

This first task only features two kinds of molecules, red and green. While all reactions in this task has the same rate, the red identity reaction gives reward 0.1 and green only 0.75. However, both are also involved in a pair identity reaction where two red (reward 0.15) are both reactant and product or two green (reward 0.25) react but remains two green. An agent might first learn to consistently go to red and then later discover that moving that red over to the other gives more, while never discovering that bringing the two green molecules together is even better.

Fig. 1 indeed shows that for DQN, all 9 heads only utilize the red molecules while DTE early is having two heads going for the more rewarding green strategy while as learning goes on, eventually all heads learns to utilize the green pair reaction.

In the more complex Metabolic cycles with distractors task, individuals benefit from different food generating cycles of reactions that rely on energy which dissipates over time if unused (red molecules). There are two possible autocatalytic reaction cycles, one cycle consists of molecules in three different shades of blue and the other cycle of molecules in three different shades of green. Both cycles require energy to continue. When they progress, they generates side products that comes in two types. If the two are brought together then that reaction generates new energy such that the cycles can continue, as well as a high reward for the agent. The population needs to keep both cycles running in order to sustain the system over time[1].

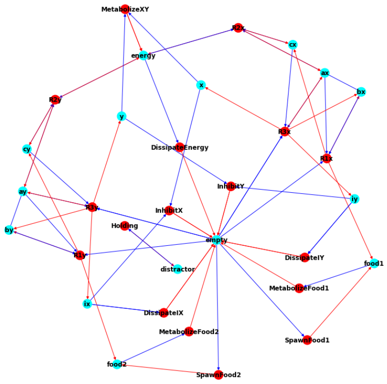

Figure 2: Reaction Graph for molecules in Chemistry Metabolic Cycles with Distractors. The artificial chemistry we use represents the possible reactions with a directed multigraph [4]. The blue nodes are compounds and the red nodes are reactions. Arrows point from reactant compounds in to reaction nodes and arrows point out of reaction nodes toward product compounds.

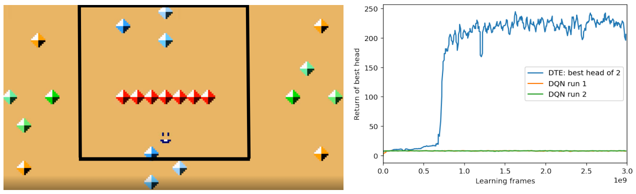

Figure 3: Left: Initial arrangement of molecules in Chemistry Metabolic Cycles with Distractors. The agent's partial viewing window is highlighted with a black rectangle. It is 11×11 sprites, each sprite being $8{\times}8$ pixels (making an $88{\times}88{\times}3$ RGB image observation). Right: Results for the Chemistry Metabolic Cycles with Distractors task, DTE's best head learns the sustainable cycles while both individual DQN runs gets permanently stuck with the distractor.

This task also contains one easier way to earn a small amount of rewarding and it is achieved by simply holding a molecule we call the distractor (orange in Fig. 3). where we can see 4 of them, 1 towards each corner) in its inventory[2]. This is the simplest rewarding strategy and represents a shallow local optima that agents consistently learns very early. The plot in Fig. 3 shows two DQN runs both stuck on this shallow local optima while the best of two DTE head (we let one head get 5 times the data of the other) learns to run all the cycles sustainably and efficiently.

[1] see https://youtu.be/FSStB7RpYis for an example video.
[2] see https://youtu.be/87NSFEPL7fk for an example video

REFERENCES

[1] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. 2020. Agent57: Outperforming the Atari Human Benchmark. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 507–517.

[2] W. Banzhaf and L. Yamamoto. 2015. *Artificial Chemistries.* MIT Press, Cambridge, MA.

[3] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations.*

[4] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[5] Wim Hordijk, Mike Steel, and Stuart Kauffman. 2012. The structure of autocatalytic sets: Evolvability, enablement, and emergence. *Acta biotheoretica* 60, 4 (2012), 379–392.

[6] Jonathan M Levine and Janneke HilleRisLambers. 2009. The importance of niches for the maintenance of species diversity. *Nature* 461, 7261 (2009), 254–257.

[7] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16).* Curran Associates Inc., Red Hook, NY, USA, 4033–4041.

[8] Dolph Schluter. 2000. *The ecology of adaptive radiation.* Oxford University Press.

[9] Satinder Singh, Andrew G. Barto, and Nuttapong Chentanez. 2004. Intrinsically Motivated Reinforcement Learning. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada].* 1281–1288.

[10] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* 2085–2087.

[11] Peter Sunehag, Alexander Sasha Vezhnevets, Edgar Duéñez-Guzmán, Igor Mordatch, and Joel Z. Leibo. 2023. Diversity Through Exclusion (DTE): Niche Identification for Reinforcement Learning through Value-Decomposition. https://arxiv.org/abs/2302.01180

[12] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[13] Vito Volterra. 1928. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science* 3, 1 (1928), 3–51.