Multiple Plans are Better than One: Diverse Stochastic Planning

Mahsa Ghasemi, [1*] **Evan Scope Crafts,** [2*] **Bo Zhao,** [2, 3] **Ufuk Topcu** [4]

[1]Electrical and Computer Engineering
[2]Oden Institute for Computational Engineering and Sciences
[3]Biomedical Engineering
[4]Aerospace Engineering and Engineering Mechanics
University of Texas at Austin, Austin, Texas 78712, USA
{mahsa.ghasemi, escopec, bozhao, utopcu}@utexas.edu

Abstract

In planning problems, it is often challenging to fully model the desired specifications. In particular, in human-robot interaction, such difficulty may arise due to human's preferences that are either private or complex to model. Consequently, the resulting objective function can only partially capture the specifications and optimizing that may lead to poor performance with respect to the true specifications. Motivated by this challenge, we formulate a problem, called *diverse stochastic planning*, that aims to generate a set of representative — small and diverse — behaviors that are near-optimal with respect to the known objective. In particular, the problem aims to compute a set of diverse and near-optimal policies for systems modeled by a Markov decision process. We cast the problem as a constrained nonlinear optimization for which we propose a solution relying on the Frank-Wolfe method. We then prove that the proposed solution converges to a local optimum and demonstrate its efficacy in several planning problems.

Introduction

Solution diversity has value in numerous planning applications, including collaborative systems, reinforcement learning, and preference-based planning. In human groups and, more generally, animal groups, the so-called notion of behavioral diversity leads to the group members' heterogeneous behavior. This heterogeneity ensures that the members learn complementary skills, thus improving the group's overall performance. An agent learning a task in an unknown environment may benefit from inducing diversity in its decisions to explore the environment more efficiently. In planning with unknown preferences, one can use diversity to construct a set of behaviors that are suitable for different preferences.

Algorithms that use notions of diversity to address one or more of these applications are known as *quality diversity (QD)* algorithms. A key component of QD algorithms is a way to summarize the important properties of different solutions. This description, known as a *behavior characterization*, is used to define diversity-based metrics. Without proper behavior characterization, solutions with trivial differences can have high values of diversity as measured by the resulting metric.

Our work is motivated by planning in settings where, in addition to a known objective, there exist some unknown objectives. The unknown objectives may represent a human user or designer's preference, which is either private or complex to model. In these settings, we propose a QD-based approach to construct a "representative" — small and diverse — set of near-optimal policies with respect to the known objective and then present that to the human to select from according to their unknown objectives. This approach allows the human to have the ultimate control over the behavior, without requiring prior knowledge of the human's preferences.

Formally, we consider the multi-objective optimization problem of returning a set of feasible policies for an infinite horizon Markov decision process (MDP) that is both near-optimal and diverse. We define the optimality of a set of policies as the sum of each policy's expected average reward in the set. Diversity captures the representativeness of a set of policies. We characterize the behavior of policies using their state-action occupancy measures and quantify diversity by the sum of pairwise divergences between the state-action occupancy measures of the policies in the set.

Our main contribution is the behavior characterization of policies using their state-action occupancy measures. This approach is domain-independent and fully encapsulates the dynamics of a given policy. We use this characterization to define the diversity of a set of policies using the pairwise Jensen-Shannon divergences between the occupancy measures. We then formulate the objective as a linear combination of the sum of the policies' rewards and their diversity and show this can be viewed as a constrained optimization problem. Due to the constraints' linearity, we can efficiently solve the problem using the Frank-Wolfe algorithm. We also prove that the algorithm is guaranteed to converge to a local optimum. Furthermore, in a series of simulations, we evaluate the proposed algorithm's performance and show its efficacy.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Section 3 provides the required background and formalizes the main problem. In Sec-

*These authors contributed equally to the manuscript.

tion 4, we introduce the proposed algorithm and detail its implementation. Section 5 presents a series of results demonstrating the algorithm's performance in multiple applications. Lastly, in Section 6, we state the concluding remarks and point to future research directions.

Related Work

Research on the development of QD algorithms has occurred within two different communities. In the field of optimization, perspectives on evolution as a process that finds distinct niches for different species have motivated the use of diversity. Simultaneously, there has been significant interest in the use of diversity to provide high-quality solutions for unknown objectives within the planning community.

In the optimization community, recent interest in QD algorithms has been driven by the success of the Novelty Search algorithm (Lehman and Stanley 2008). The original Novelty Search algorithm eschews the use of notions of solution quality entirely; its sole goal is to find a set of solutions that are diverse with respect to some distance measure. Surprisingly, this approach is able to find solutions with better performance on difficult tasks, such as maze navigation, than algorithms relying on an objective function. This result has led to considerable interest in the development of new QD algorithms to address tasks that were previously considered to be too difficult (Lehman and Stanley 2010, 2011a,b; Kistemaker and Whiteson 2011; Mouret 2011; Risi, Hughes, and Stanley 2010; Mouret and Doncieux 2012; Cully and Mouret 2013; Gomes and Christensen 2013; Gomes, Urbano, and Christensen 2013; Liapis et al. 2013; Martínez et al. 2013; Naredo and Trujillo 2013). For a review, see (Pugh, Soros, and Stanley 2016).

The type of behavior characterization used in these works varies and can be domain-dependent. For example, in navigation problems, diversity can be defined using Euclidean distances between points visited. Another approach, used by the popular MAP elites algorithm, is to assume that a domain-dependent behavior characterization is given (Mouret and Clune 2015). A promising area of research is the development of new approaches to behavior characterization (Gaier, Asteroth, and Mouret 2020).

The success of the Novelty Search and MAP elites algorithms has inspired the use of diversity in reinforcement learning, with the hope that diversity can help avoid poor local minima. Different methods of behavior characterization for policies have been used, including methods based on sequences of actions (Jackson and Daley 2019), state trajectories (Eysenbach et al. 2018), or diversity through determinants of actions in states (Parker-Holder et al. 2020). Similarly to our work, (Parker-Holder et al. 2020) considers an explicit trade-off between the quality and diversity of the policies. However, our approach differs in that we leverage knowledge of the system dynamics to characterize policies in a way that includes information about both the states visited and the policy actions, and to develop a solution algorithm with guaranteed convergence to a local minimum.

Behavior characterization has also been a key focus of QD-based work in the planning community. For example, in an approach similar to MAP elites, Myers and Lee (1999) and Myers (2006) assume that there is a meta-description of the planning domain. They then define an approach that obtains solutions that are diverse with respect to the meta-description. Another approach to behavior characterization is through the use of domain landmarks, which are disjunctive sets of propositions that plans must satisfy, such as a set of states that a plan must transition through before reaching a goal state (Hoffmann and Nebel 2001). If the set of landmarks can be computed, a greedy algorithm can be used to iteratively select landmarks from the set and find a plan that satisfies the landmark (e.g., reaches a certain state) (Bryce 2014). Behavior characterization based on the plan actions, as in the RL community, is also a common technique (Coman and Munoz-Avila 2011; Nguyen et al. 2012; Katz and Sohrabi 2020).

The way behavioral characterization and diversity metrics are incorporated into planning algorithms varies. In some cases, the problem is formulated as maximizing the diversity of the set of solutions (Coman and Munoz-Avila 2011), or as finding a set of solutions that satisfy a diversity threshold (Nguyen et al. 2012; Srivastava et al. 2007). In other cases, like our work, there exists both an unknown objective and a known objective, and the problem is formulated in terms of a trade-off between the diversity of the solution set and the optimality of each of the candidate solutions (Coman and Munoz-Avila 2011; Katz and Sohrabi 2020; Petit and Trapp 2015). Our work is distinct from these approaches because we develop a new method for behavior characterization and consider a stochastic setting modeled as an MDP. In addition, unlike many QD-based planning algorithms, our approach does not rely on greedy strategies. While greedy algorithms have near-optimality guarantees in some settings, such as when the problem is submodular (Bach 2013), our problem is supermodular and in general no such guarantee exists.

References

Bach, F. 2013. *Learning with Submodular Functions: A Convex Optimization Perspective*, volume 6 of *Foundations and Trends in Machine Learning*. now publishers inc.

Bryce, D. 2014. Landmark-based plan distance measures for diverse planning. In *Proceedings of the Twenty-Fourth International Conferenc on International Conference on Automated Planning and Scheduling*, 56–64.

Coman, A.; and Munoz-Avila, H. 2011. Generating Diverse Plans Using Quantitative and Qualitative Plan Distance Metrics. In *AAAI*, 946–951. Citeseer.

Cully, A.; and Mouret, J.-B. 2013. Behavioral repertoire learning in robotics. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, 175–182.

Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070* .

Gaier, A.; Asteroth, A.; and Mouret, J.-B. 2020. Automating Representation Discovery with MAP-Elites. *arXiv preprint arXiv:2003.04389* .

Gomes, J.; and Christensen, A. L. 2013. Generic behaviour similarity measures for evolutionary swarm robotics. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, 199–206.

Gomes, J.; Urbano, P.; and Christensen, A. L. 2013. Evolution of swarm robotics systems with novelty search. *Swarm Intelligence* 7(2-3): 115–144.

Hoffmann, J.; and Nebel, B. 2001. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research* 14: 253–302.

Jackson, E. C.; and Daley, M. 2019. Novelty search for deep reinforcement learning policy network weights by action sequence edit metric distance. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 173–174.

Katz, M.; and Sohrabi, S. 2020. Reshaping Diverse Planning. In *AAAI*, 9892–9899.

Kistemaker, S.; and Whiteson, S. 2011. Critical factors in the performance of novelty search. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 965–972.

Lehman, J.; and Stanley, K. O. 2008. Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, 329–336.

Lehman, J.; and Stanley, K. O. 2010. Revising the evolutionary computation abstraction: minimal criteria novelty search. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, 103–110.

Lehman, J.; and Stanley, K. O. 2011a. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation* 19(2): 189–223.

Lehman, J.; and Stanley, K. O. 2011b. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 211–218.

Liapis, A.; Martínez, H. P.; Togelius, J.; and Yannakakis, G. N. 2013. Transforming exploratory creativity with DeLeNoX .

Martínez, Y.; Naredo, E.; Trujillo, L.; and Galván-López, E. 2013. Searching for novel regression functions. In *2013 IEEE congress on evolutionary computation*, 16–23. IEEE.

Mouret, J.-B. 2011. Novelty-based multiobjectivization. In *New horizons in evolutionary robotics*, 139–154. Springer.

Mouret, J.-B.; and Clune, J. 2015. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909* .

Mouret, J.-B.; and Doncieux, S. 2012. Encouraging behavioral diversity in evolutionary robotics: An empirical study. *Evolutionary computation* 20(1): 91–133.

Myers, K. L. 2006. Metatheoretic Plan Summarization and Comparison. In *ICAPS*, 182–192.

Myers, K. L.; and Lee, T. J. 1999. Generating qualitatively different plans through metatheoretic biases. In *AAAI/IAAI*, 570–576.

Naredo, E.; and Trujillo, L. 2013. Searching for novel clustering programs. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, 1093–1100.

Nguyen, T. A.; Do, M.; Gerevini, A. E.; Serina, I.; Srivastava, B.; and Kambhampati, S. 2012. Generating diverse plans to handle unknown and partially known user preferences. *Artificial Intelligence* 190: 1–31.

Parker-Holder, J.; Pacchiano, A.; Choromanski, K.; and Roberts, S. 2020. Effective Diversity in Population-Based Reinforcement Learning. *arXiv preprint arXiv:2002.00632* .

Petit, T.; and Trapp, A. C. 2015. Finding diverse solutions of high quality to constraint optimization problems. In *IJCAI. International Joint Conference on Artificial Intelligence*.

Pugh, J. K.; Soros, L. B.; and Stanley, K. O. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI* 3: 40.

Risi, S.; Hughes, C. E.; and Stanley, K. O. 2010. Evolving plastic neural networks with novelty search. *Adaptive Behavior* 18(6): 470–491.

Srivastava, B.; Nguyen, T. A.; Gerevini, A.; Kambhampati, S.; Do, M. B.; and Serina, I. 2007. Domain Independent Approaches for Finding Diverse Plans. In *IJCAI*, 2016–2022.