Towards Sample-Efficient Multi-Objective Reinforcement Learning

Doctoral Consortium

Lucas N. Alegre
Institute of Informatics - Federal University of Rio Grande do Sul
Porto Alegre - RS, Brazil
Artificial Intelligence Lab - Vrije Universiteit Brussel
Brussels, Belgium
lnalegre@inf.ufrgs.br

ABSTRACT

In sequential decision-making problems, the objective that a reinforcement learning agent seeks to optimize is often modeled via a reward function. However, in real-world problems, agents often have to optimize multiple (possibly conflicting) objectives. This setting is known as multi-objective reinforcement learning (MORL). In MORL, the goal of the agent is not to learn a single policy, but a *set* of policies, each of which specialized in optimizing a single objective or a combination of objectives. In my Ph.D., I investigate methods that allow the agent to learn a carefully-constructed set of policies that can be combined to solve challenging MORL problems in a sample-efficient manner. In this paper, I present a brief overview of my work on this topic and focus on two main contributions: (i) a novel algorithm for optimal policy transfer based on theoretical equivalences between successor features and MORL; and (ii) a novel MORL algorithm based on generalized policy improvement that learns a set of policies that is guaranteed to contain an optimal policy for *any* possible agent's preferences over objectives.

KEYWORDS

Multi-Objective Reinforcement Learning; Generalized Policy Improvement; Successor Features; Model-Based RL

ACM Reference Format:
Lucas N. Alegre. 2023. Towards Sample-Efficient Multi-Objective Reinforcement Learning: Doctoral Consortium. 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

Reinforcement learning (RL) [16] has been successfully applied to solve challenging decision-making problems [6, 8, 12, 17]. In RL, a task is often represented by a single scalar reward function that encodes the agent's goal. However, in many real-world problems, the agent must optimize behaviors that balance multiple (possibly) conflicting objectives, each encoded by a different reward function. For example, a robot may need to balance speed, battery usage, and accuracy in reaching a goal location. Multi-objective RL (MORL) [10] algorithms tackle such a challenge.

Standard RL algorithms can be sample-inefficient as they may require the agent to interact with the environment a large number of times [16]. This can make them impractical in situations where interactions are costly or dangerous. The problem of sample efficiency is further exacerbated in MORL algorithms because they require the agent to learn a *set* of policies instead of a single one, each aimed at optimizing a different trade-off between the agent's objectives [19]. This further increases the number of interactions required, making it even more challenging to implement these algorithms in real-world scenarios.

Until now, during my Ph.D., I have explored different strategies to design sample-efficient algorithms capable of learning multiple policies specialized to different preferences over objectives/features. First, in [1], we introduced a novel method based on theoretical equivalences between the optimal policy transfer problem tackled by *successor features* (SFs) [4] and the MORL problem, that learns a set of policies that is guaranteed to contain an optimal policy for *any* possible agent's preferences over objectives. Next, in [3], we further explored these connections and introduced a novel MORL algorithm that employs *generalized policy improvement* (GPI) [5] to (i) identify promising preferences to train on and (ii) identify which previous experiences are most relevant when learning a policy for a particular preference. I discuss both contributions in Section 2 and Section 3, respectively.

2 OPTIMAL POLICY TRANSFER

When reward functions are expressed as linear combinations of features, and the agent has previously learned a set of policies for different tasks, the framework of *successor features* (SFs) and *generalized policy improvement* (GPI) [4] can be exploited to identify reasonable policies for new tasks in a zero-shot manner. Intuitively, GPI generalizes the policy improvement step by improving a given policy, tasked with solving a particular task, over a *set* of policies, instead of a single one. However, the resulting policy is not guaranteed to be optimal. In [1], we introduce a novel algorithm that addresses this limitation and solves the following *optimal policy transfer* problem: *how to construct a set of policies such that combining them directly leads to the optimal policy for any novel linearly-expressible tasks?*

We first show (under mild assumptions) that the transfer learning problem tackled by SFs is equivalent to the problem of learning to optimize multiple objectives in RL. We then introduce SFOLS, an SF-based extension of the *Optimistic Linear Support* [14] algorithm to learn a set of policies whose SFs form a convex coverage set (CCS). We prove that policies in this set can be combined via GPI to

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.

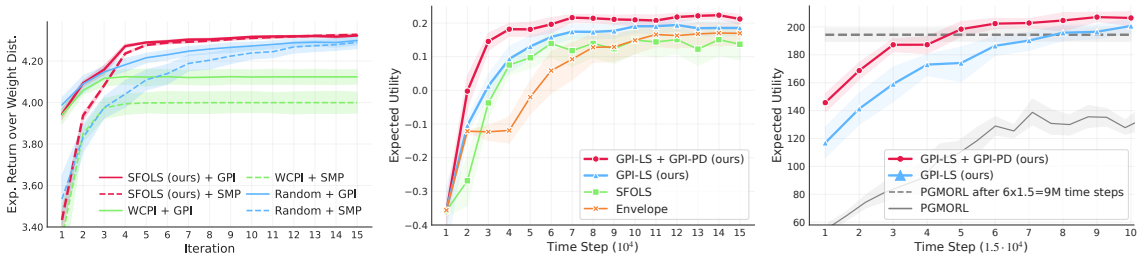

Figure 1: (Left) Expected performance over the preference distribution of SFOLS on the Reacher domain; (Middle) Expected utility of GPI-LS on the Minecart domain; (Right) Expected utility of GPI-LS in the MO-Hopper domain.

construct optimal behaviors for any new linearly-expressible tasks, without requiring additional training samples.

We empirically show that our method outperforms state-of-the-art competing algorithms in discrete and continuous domains under value function approximation. For instance, we compare SFOLS with competing algorithms (WCPI [20] and a random baseline) on the MO-Reacher domain. In this domain, the agent controls a two-joint arm that must reach different target locations. A weight vector encodes the linear preferences over each feature (the distance to each target location) of the reward function. In the left panel of Fig. 1, we show that SFOLS learns a set of policies that outperform the competing algorithms in terms of expected return over the weight distribution in few iterations. It does so by following the GPI policy (solid lines) or the best learned policy (dashed lines).

3 SAMPLE-EFFICIENT MORL VIA GPI

In [3], we introduce a novel MORL algorithm, with important theoretical guarantees, that improves sample efficiency via two novel prioritization techniques which are based on GPI.

If the utility function of a MORL problem is a linear combination of the agent's objectives, optimal solutions are sets of policies known as *convex coverage sets* (CCS) [15]. Given a CCS, agents can *directly* identify the optimal solution to any novel linear preferences. MORL algorithms that learn a CCS (e.g., [13]) may be sample inefficient ***(i)*** due to the heuristics they use to determine which preferences to train on, at any given moment during the construction of a CCS; and ***(ii)*** because they can only improve a CCS after optimal (or near-optimal) policies are identified—which may require a large number of samples acquired via environment interactions.

We address the first issue via a novel algorithm, GPI-LS, which employs a GPI-based prioritization technique for selecting which preferences to train on. GPI-LS prioritizes preferences based on a lower bound on performance improvements guaranteed to be achievable via GPI, which accurately and reliably identifies the most relevant preferences to train on when learning a CCS. To address the second issue, we show that our method is an anytime algorithm that monotonically improves the quality of its CCS, even if given *intermediate* (possibly sub-optimal) policies for different preferences. This improves sample efficiency: our method identifies intermediate CCSs with formally bounded maximum utility loss even if there are constraints on the number of times the agent can interact with its environment. GPI-LS is guaranteed to always converge to an optimal solution in a finite number of steps, or an ϵ-optimal solution (for a bounded ϵ) if the agent is limited and can only identify possibly sub-optimal policies.

A complementary approach for increasing sample efficiency (which has been increasingly studied in the context of deep RL [11]) is to use a model-based approach to accelerate learning. In MORL, once a model is learned, it can be used to identify policies for *any* preferences, thus minimizing the required number of interactions with the environment. Dyna algorithms are based on generating simulated experiences to more rapidly update a value function or policy. An important question, however, is *which* artificial experiences should be generated to accelerate learning. We introduce GPI-Prioritized Dyna (GPI-PD), a novel Dyna-style MORL algorithm—the first model-based MORL technique capable of dealing with continuous state spaces. It introduces a new, principled GPI-based prioritization technique for identifying which experiences are most relevant to rapidly learn the optimal policy for novel preferences.

In the middle and right panels of Fig. 1, we compare GPI-LS (and its model-based extension with GPI-PD) with competing algorithms in the Minecart and MO-Hopper domains, respectively. These domains are available in the MO-Gym library, which we introduced in [2]. Our methods consistently identify optimal solutions, reach near-zero maximum utility loss, and achieve performance metrics that strictly dominate that of competitors. This is the case even when we allow the PGMORL algorithm [18] to collect 9 million experiences: they could interact with their environment ten times more often than our method/agent. Even then, GPI-PS (with or without using a learned model) consistently achieved higher expected utility during learning, and converged to a final solution with superior performance.

4 FUTURE WORK

In future work, I would like to explore how other model-based techniques can be employed to increase the sample efficiency of MORL algorithms. For instance, an interesting direction would be to use predecessor/backward models [7] or value equivalent models [9] to learn policies for different agent preferences. Another important challenge is dealing with uncertainties in the learned model and/or action-value function predictions. Because GPI relies on the value function of all the available policies, a single error in the value estimate of a policy's action may ruin the performance of the GPI policy. Thus, knowing when to trust the value estimates of each policy is crucial in policy transfer settings.

ACKNOWLEDGMENTS

This study was financed in part by the following Brazilian agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, and CNPq (grant 140500/2021-9). This research was partially supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program and the Research Foundation Flanders (FWO) [G062819N].

REFERENCES

[1] Lucas N. Alegre, Ana L. C. Bazzan, and Bruno C. da Silva. 2022. Optimistic Linear Support and Successor Features as a Basis for Optimal Policy Transfer. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 394–413. https://proceedings.mlr.press/v162/alegre22a.html

[2] Lucas N. Alegre, Florian Felten, El-Ghazali Talbi, Grégoire Danoy, Ann Nowé, Ana L. C. Bazzan, and Bruno C. da Silva. 2022. MO-Gym: A Library of Multi-Objective Reinforcement Learning Environments. In *Proceedings of the 34th Benelux Conference on Artificial Intelligence BNAIC/Benelearn 2022*.

[3] Lucas N. Alegre, Diederik M. Roijers, Ann Nowé, Ana L. C. Bazzan, and Bruno C. da Silva. 2023. Sample-Efficient Multi-Objective Learning via Generalized Policy Improvement Prioritization. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. To appear.

[4] Andre Barreto, Will Dabney, Remi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. 2017. Successor Features for Transfer in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[5] André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. 2020. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30079–30087. https://doi.org/10.1073/pnas.1907370117

[6] Marc G. Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, and Ziyu Wang. 2020. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* 588, 7836 (01 Dec 2020), 77–82. https://doi.org/10.1038/s41586-020-2939-8

[7] Veronica Chelu, Doina Precup, and Hado P van Hasselt. 2020. Forethought and Hindsight in Credit Assignment. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 2270–2281.

[8] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* 602, 7897 (01 Feb 2022), 414–419. https://doi.org/10.1038/s41586-021-04301-9

[9] Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. 2021. Proper Value Equivalence. In *Proceedings of the 35th Conference on Neural Information Processing Systems*. Sydney, Australia.

[10] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (13 Apr 2022), 26. https://doi.org/10.1007/s10458-022-09552-y

[11] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. 2019. When to Trust Your Model: Model-Based Policy Optimization. In *Advances in Neural Information Processing Systems (NIPS) 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 12519–12530. http://papers.nips.cc/paper/9416-when-to-trust-your-model-model-based-policy-optimization.pdf

[12] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nazi, Jiwoo Pak, Andy Tong, Kavya Srinivasa, William Hang, Emre Tuncer, Quoc V. Le, James Laudon, Richard Ho, Roger Carpenter, and Jeff Dean. 2021. A graph placement methodology for fast chip design. *Nature* 594, 7862 (01 Jun 2021), 207–212. https://doi.org/10.1038/s41586-021-03544-w

[13] Hossam Mossalam, Yannis M. Assael, Diederik M. Roijers, and Shimon Whiteson. 2016. Multi-Objective Deep Reinforcement Learning. *CoRR* abs/1610.02707 (2016). arXiv:1610.02707 http://arxiv.org/abs/1610.02707

[14] Diederik Roijers. 2016. *Multi-Objective Decision-Theoretic Planning*. Ph.D. Dissertation. University of Amsterdam.

[15] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A Survey of Multi-Objective Sequential Decision-Making. *J. Artificial Intelligence Research* 48, 1 (Oct. 2013), 67–113.

[16] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: An introduction* (second ed.). The MIT Press.

[17] Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. 2022. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* 602, 7896 (01 Feb 2022), 223–228. https://doi.org/10.1038/s41586-021-04357-7

[18] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. 2020. Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

[19] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Eds.). 14610–14621.

[20] Tom Zahavy, Andre Barreto, Daniel J Mankowitz, Shaobo Hou, Brendan O'Donoghue, Iurii Kemaev, and Satinder Singh. 2021. Discovering a set of policies for the worst case reward. In *Proceedings of the 9th International Conference on Learning Representations*.