Generalizing Objective-Specification in Markov Decision Processes

Doctoral Consortium

Pedro P. Santos
INESC-ID, Instituto Superior Técnico
Lisbon, Portugal
pedro.pinto.santos@tecnico.ulisboa.pt

ABSTRACT

In this thesis, we address general utility Markov decision processes (GUMDPs), which generalize the standard Markov decision processes (MDPs) framework for decision-making by considering a broader range of objective functions that depend on the occupancy induced by a given policy. We aim to study GUMDPs from a theoretical perspective and develop new algorithms to solve GUMDPs by leveraging optimization techniques. We also aim to better understand how objective specification in GUMDPs compares to that of MDPs, further studying the connections between the two frameworks for sequential decision-making. We hope that, by achieving the proposed goals, the contributions of this thesis can lay down the foundations supporting the future development and deployment of agents that take advantage of the diverse set of objectives that can be encoded with GUMDPs.

KEYWORDS

Sequential Decision Making; Markov Decision Processes; Reinforcement Learning; Machine Learning.

ACM Reference Format:
Pedro P. Santos. 2024. Generalizing Objective-Specification in Markov Decision Processes: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Markov decision processes (MDPs) [23] provide a mathematical framework to study stochastic sequential decision-making. At a given point in the interaction, the MDP is at a particular state and the decision-maker chooses an action; given the chosen action, the process evolves to a new random state and the decision-maker receives a scalar reward value. The goal of the agent is to find a policy, i.e., a mapping from states to actions, such that some function of the stream of rewards yielded when interacting with the MDP is maximized. MDPs have found a wide range of applications in different domains [26], such as inventory management [7], optimal stopping [6] or queueing control [25].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

MDPs are also of key importance in the field of reinforcement learning (RL) [1] since the agent-environment interaction is typically formalized under the framework of MDPs. Recent years witnessed significant progress in solving challenging problems across various domains using RL [18, 20, 24]. Such results attest to the flexibility of MDPs as a general framework to study sequential decision-making under uncertainty, as well as to the power and convenience of RL methods that allow the learning of approximate optimal behavior under partial MDP specification via direct interaction with the environment.

As discussed, previous works attest to the flexibility of the MDPs framework in encoding different objectives via the specification of a scalar reward signal. However, there exist relevant objectives that cannot be easily specified under the MDP framework [2]. These include, for example, imitation learning [16, 22], pure exploration problems [14], risk-averse RL [11], diverse skills discovery [3, 9] and constrained MDPs [4, 8]. Such objectives, including the scalar reward objective of standard MDPs, can be formalized under the framework of general utility Markov decision processes (GUMDPs) [21, 28]. In GUMDPs, the objective is, instead, encoded as a function of the occupancy induced by a given policy, i.e., a function of the frequency of visitation of states (or state-action pairs) induced when running the policy on the MDP. Recent works have unified such objectives under the same framework and proposed general algorithms to solve GUMDPs under convex objective functions [12, 27, 28]. Extensions to the case of unknown dynamics are also provided by the aforementioned works.

In this thesis, we first aim to study GUMDPs from a theoretical perspective. In particular, we hope to contribute to a better understanding of the implicit assumptions of the GUMDPs framework and the types of objectives that can be encoded with GUMDPs, as well as the relation between GUMDPs and other decision-making-related frameworks. We further elaborate on this matter in Sec. 2. Second, we aim to develop new planning algorithms to solve GUMDPs by leveraging optimization techniques. We also aim to study the intersection between RL and GUMDPs, i.e., address the development of methods to learn approximately optimal behavior in an online, iterative fashion. We further elaborate on this matter in Sec. 3.

2 GUMDPS

The framework of GUMDPs allows for a variety of objectives, including many supervised and unsupervised RL problems. We aim to better understand, from a theoretical perspective, the implicit assumptions posed by GUMDPs, as well as the expressiveness of

objective-specification in GUMDPs in comparison to that of standard MDPs. We now further elaborate on this matter below, highlighting two research directions that we believe are of interest.

2.1 The infinite trials implicit assumption

As recently shown in [21], GUMDPs implicitly make an infinite trials/episodes assumption, i.e., GUMDPs implicitly assume the performance of a given policy is evaluated under an infinite number of episodes of interaction with the environment. Since this assumption may be violated under many interesting application domains, the authors introduce a modification of GUMDPs where the objective function depends on the empirical state (or state-action) occupancy induced over a finite number of episodes. The authors also motivate the use of non-Markovian policies under the proposed finite-trials GUMDPs formulation.

We aim to further extend the analysis in [21] by considering a more general setting where an agent interacts with an environment over multiple episodes, each with a random length [19]. In particular, we aim to study the impact of the number of trials/episodes used to evaluate the performance of a given policy when there is uncertainty in the episodes' length. Depending on the distribution over the episodes' length, we expect the mismatch between the infinite and finite trials formulation to be non-negligible. Considering randomness over the episodes' length is of key importance as it is related to discounting in sequential decision-making [10].

2.2 Connections with objective-specification in MDPs and beyond

The framework of GUMDPs differs from that of standard MDPs since, as previously discussed, the objective function depends on the occupancy induced by the policy. Such difference allows for a more general decision framework since it is known that certain objectives can be encoded using GUMDPs but cannot be encoded using the framework of MDPs [27]. Albeit being a more flexible framework for decision-making in terms of objective-specification, the temporal/sequential aspect of the decision-making process in GUMDPs becomes rather abstract in comparison to that of standard MDPs. This is because we can, equivalently, interpret the objective function in GUMDPs as encoding an ordering over the stationary distributions of the Markov chains that arise when we condition the transition probability function on different policies.

Despite such differences, [27] connects GUMDPS with MDPs with non-stationary rewards. Other works provide connections between GUMDPs and other decision-making-related frameworks. For example, as shown in [27], the problem of solving GUMDPs with convex objective functions can be recast as a game between two players; the solution to the GUMDP is equivalent to finding a solution to a min-max game (saddle-point) involving a cost and a policy player. In [12], it is shown that GUMDPs with convex objectives are related to the concept of mean field games [15, 17], a continuous approximation of many-agent RL.

We aim to continue these lines of research by further investigating connections between GUMDPs and other frameworks for (sequential) decision-making. In particular, we aim to further investigate connections between GUMDPs and: (i) MDPs with non-stationary rewards; (ii) multi-objective MDPs [5]; and (iii) robust MDPs [13].

3 SOLVING GUMDPS

We now highlight below two research directions related to the development of planning and learning algorithms to solve GUMDPs.

3.1 Plannning for GUMDPs

Different works provide planning algorithms for GUMDPs when the objective function is convex such as [12, 14, 27]. As an example, [27] provides a meta-algorithm that, given an algorithm for a cost player and an algorithm for a policy player, returns a solution to the GUMDP to any desired tolerance.

In this thesis, we will focus our attention on the development of new algorithms to solve GUMDPs by leveraging techniques from the field of optimization. In particular, we also want to provide algorithms to approximately solve GUMDPs when the objective function is non-convex in the space of occupancies.

3.2 Learning in GUMDPs

Finally, we want to provide learning algorithms for GUMPS, i.e., algorithms that allow learning approximately optimal policies without requiring a complete specification of the GUMDP. First, we want to study the case where only the dynamics of the GUMDP are unknown. In a later stage, we also want to investigate the case of unknown of underspecified objective functions.

ACKNOWLEDGMENTS

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia (FCT), under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020), PTDC/CCI-COM/7203/2020, and PTDC/CCI-COM/5060/2021 (RELEvaNT). This research was also supported by the Air Force Office of Scientific Research under award number FA9550-22-1-047. The author acknowledges the FCT PhD grant 2021.04684.BD.

REFERENCES

[1] 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press.
[2] David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael L. Littman, Doina Precup, and Satinder Singh. 2022. On the Expressivity of Markov Reward. arXiv:2111.00876 [cs.LG]
[3] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. 2018. Variational Option Discovery Algorithms. arXiv:1807.10299 [cs.AI]
[4] E. Altman. 1999. *Constrained Markov Decision Processes.* Chapman and Hall.
[5] Krishnendu Chatterjee, Rupak Majumdar, and Thomas A. Henzinger. 2006. Markov Decision Processes with Multiple Objectives. 325–336.
[6] Y.S. Chow, H. Robbins, and D. Siegmund. 1971. *Great Expectations: The Theory of Optimal Stopping.*
[7] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. 1952. The Inventory Problem: II. Case of Unknown Distributions of Demand. *Econometrica* 20, 3 (1952), 450–466.
[8] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. 2020. Exploration-Exploitation in Constrained MDPs. *CoRR* abs/2003.02189 (2020).
[9] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Skills without a Reward Function. arXiv:1802.06070 [cs.AI]
[10] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. 2019. Hyperbolic Discounting and Learning over Multiple Horizons. arXiv:1902.06865 [stat.ML]

[11] Javier García, Fern, and o Fernández. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* 16, 42 (2015), 1437–1480.
[12] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. 2022. Concave Utility Reinforcement Learning: the Mean-Field Game Viewpoint. arXiv:2106.03787 [cs.LG]
[13] Vineet Goyal and Julien Grand-Clément. 2021. Robust Markov Decision Process: Beyond Rectangularity. arXiv:1811.00215 [math.OC]
[14] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. 2019. Provably Efficient Maximum Entropy Exploration. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 2681–2691.
[15] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems* 6, 3 (2006), 221 – 252.
[16] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.* 50, 2, Article 21 (apr 2017), 35 pages.
[17] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean Field Games. *Japanese Journal of Mathematics* 2 (03 2007), 229–260.
[18] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *CoRR* abs/1509.02971 (2016).
[19] Debmalya Mandal, Goran Radanovic, Jiarui Gan, Adish Singla, and Rupak Majumdar. 2023. Online Reinforcement Learning with Uncertain Episode Lengths. arXiv:2302.03608 [cs.LG]
[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2015. Playing Atari with Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533.
[21] Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. 2023. Convex Reinforcement Learning in Finite Trials. *Journal of Machine Learning Research* 24, 250 (2023), 1–42. http://jmlr.org/papers/v24/22-1514.html
[22] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. 2018. An Algorithmic Perspective on Imitation Learning. *Foundations and Trends in Robotics* 7, 1–2 (2018), 1–179.
[23] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
[24] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354–359.
[25] Shaler Stidham. 1978. Socially and Individually Optimal Control of Arrivals to a GI/M/1 Queue. *Management Science* 24, 15 (1978), 1598–1610.
[26] D. J. White. 1988. Further Real Applications of Markov Decision Processes. *Interfaces* 18, 5 (1988), 55–61.
[27] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. 2023. Reward is enough for convex MDPs. arXiv:2106.00661 [cs.AI]
[28] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. 2020. Variational Policy Gradient Method for Reinforcement Learning with General Utilities. arXiv:2007.02151 [cs.LG]