# Beyond Markov Decision Process with Scalar Markovian Rewards

**Shuwa Miura** *

Manning College of Information and Computer Sciences, University of Massachusetts
140 Governors Drive Amherst, MA, USA
smiura@umass.edu.

**Abstract**

Real-world decision problems often involve multiple competing objectives or a complex reward structure that violate Markov assumption. However, the existing research on sequential decision making under uncertainty primarily focused on Markov Decision Processes (MDPs) with scalar Markovian reward signals. My thesis considers settings where scalar Markovian rewards are not sufficient to produce desired behaviors. The first part of my thesis develops algorithms to optimize lexicographically ordered objectives. The second part considers autonomous agents which incorporate the perspective of their observer. As the perspective of the observer can depend on how the agents behaved so far, rewards in this setting can depend on histories (non-Markovian). In the last part of my thesis, I hope to characterize when rewards beyond scalar Markovian signals are needed from the decision theoretic perspective.

## Introduction

The existing research on sequential decision making under uncertainty primarily focused on MDPs with scalar Markovian rewards. The implicit assumption is that scalar Markovian rewards are sufficient to characterize desired behaviors of autonomous agents. However, many real-world problems inherently involve multiple competing objectives and can involve rewards that depend on histories. Therefore, to deal with real-world problems, we need to go beyond traditional scalar Markovian reward settings.

The field of multi-objective decision making offers several fruitful approaches to formalize and solve decision problems with multiple objectives, such as approaches based on computing the *Pareto front* of the policy space (Roijers and Whiteson 2017) or Constrained MDPs (CMDPs) (Altman 1999). Among different approaches to handle problems with multiple objectives, my thesis focuses on MDPs with lexicographic preferences (LMDPs) (Mouaddib 2004). LMDPs offer intuitive formulations for problems when there is a clear ordering among objectives. The existing solution methods for LMDPs, however, lack scalability (Pineda, Wray, and Zilberstein 2015) and optimality guarantees (Wray, Zilberstein, and Mouaddib 2015).

*number: 4135129570. website: https://dosydon.github.io.



There are growing interests in MDPs with non-Markovian rewards (Bacchus, Boutilier, and Grove 1996; Brafman, Giacomo, and Patrizi 2018). Of particular interest to my thesis are agents cognizant of the perspective of their observers. Consider, for example, an autonomous vehicle (AV) and a pedestrian approaching a crosswalk. If the AV is aware of the pedestrian, it may slow down further away from the crosswalk to assure the pedestrian that it plans to stop. We call this kind of behavior an *observer-aware* behavior. Several existing frameworks offer different approaches to produce different kinds of observer-aware behaviors (Dragan, Lee, and Srinivasa 2013; Zhang et al. 2017), but there is no unifying framework that reveals the relationships among the approaches and the complexity of the problem.

The general goal of my thesis therefore is to develop autonomous agents that can handle complex objectives beyond traditional scalar Markovian rewards. I hope to make the following contributions in my thesis:

1. **Heuristic Search for Lexicographic MDPs**.
   (a) optimal heuristic search algorithm.
   (b) anytime search algorithm.
2. **Unifying Framework for Observer-Aware Planning**.
   (a) framework and its complexity.
   (b) initial empirical evaluations.
   (c) approximate solution methods for the problem.
   (d) empirical evaluations with autonomous vehicles.
   (e) balancing task completion and interpretability.
3. **Expressivity of Scalar Markovian Rewards**.

The first part of my thesis develops theories and algorithms for LMDPs. The second part of my thesis proposes an unifying framework to produce observer-aware behaviors called Observer-Aware MDP (OAMDP). OAMDP is a variant of MDP with non-Markovian rewards, where rewards depend on the belief of the observer. While the two problems above intuitively need complex reward signals beyond scalar Markovian rewards, one could still argue that carefully engineered scalar Markovian rewards are sufficient to characterize desired behaviors. The last part of the thesis explores exactly when that is the case from the decision theoretic perspective.

## Completed Work

I describe below the contributions that have been made so far toward the general goal of my thesis.

**Contribution 1 (a)** We proposed the first heuristic search algorithm for LMDPs (Miura, Wray, and Zilberstein 2022). Our algorithm is based on heuristic search algorithm for CMDPs (Trevizan et al. 2017). The algorithm starts with the initial state, and gradually expands states on the current best solution. We experimentally showed that using heuristic search can avoid visiting irrelevant states and can find an optimal policy faster than the naïve linear programming formulation proposed in (Pineda, Wray, and Zilberstein 2015).

**Contribution 2 (a)** We proposed a unifying framework to produce observer-aware behaviors called Observer-Aware MDP (OAMDP) (Miura and Zilberstein 2021). OAMDP is a variant of MDP with non-Markovian rewards, where rewards depend on the belief of the observer. We discussed its relationships with other models (POMDP and I-POMDP) and proved that computing an optimal policy for OAMDP is PSPACE-hard in the worst case. We presented initial results on solving OAMDPs using Monte-Carlo Tree Search.

**Contribution 2 (b)** We conducted initial empirical evaluations of policies computed by OAMDP (Miura, Cohen, and Zilberstein 2021). The participants in our experiment observed an agent moving in GridWorld-like environment. There were three possible goals for the agent. We asked the participants which of the three goals is most likely given trajectories. The results indicated that observer-aware behaviors can convey the agent's goal better to observers.

## Directions for Future Work

**Contribution 1 (b)** Based on our previous work on optimal heuristic search for L-MDPs, I plan to work on developing anytime search algorithm for L-MDPs. As autonomous agents in the real world have limited time before acting, they need to come up with a reasonable action to take within given time while taking multiple objectives into account.

**Contribution 2 (c), (d), and (e)** As we showed computing an optimal policy for OAMDP is intractable in the worst-case, I plan to develop approximate algorithms to solve OAMDP. Moreover, as our initial evaluations are limited to simple virtual environments, I hope to conduct evaluations in more realistic settings such as autonomous vehicles. Another interesting direction is to explore how to balance task completion and interpretability as optimizing interpretability alone can lead to degenerate behaviors.

**Contribution 3** Intuitively, we need rich reward signals beyond scalar Markovian rewards to characterize complex behaviors. But one might argue that we can tell agents to do anything by carefully engineering scalar Markovian rewards. Recent work pointed out that a set of good policies cannot necessarily be characterized by scalar Markovian rewards (Abel et al. 2021). I started working on the extension of the work (Miura 2022). Our initial result provides both necessary and sufficient conditions for the existence of such Markov reward function.

## References


Abel, D.; Dabney, W.; Harutyunyan, A.; Ho, M. K.; Littman, M.; Precup, D.; and Singh, S. 2021. On the Expressivity of Markov Reward. *Advances in Neural Information Processing Systems*, 34.

Altman, E. 1999. *Constrained Markov Decision Processes*, volume 7. CRC Press.

Bacchus, F.; Boutilier, C.; and Grove, A. 1996. Rewarding Behaviors. In *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, volume 2, 1160–1167.

Brafman, R.; Giacomo, G. D.; and Patrizi, F. 2018. LTLf/LDLf Non-Markovian Rewards. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 1771–1778.

Dragan, A. D.; Lee, K. C. T.; and Srinivasa, S. S. 2013. Legibility and predictability of robot motion. In *Proceedings of Eighth ACM/IEEE International Conference on Human-Robot Interaction*, 301–308.

Miura, S. 2022. On the Expressivity of Multidimensional Markov Reward. In *RLDM Workshop on RL as a Model of Agency*.

Miura, S.; Cohen, A.; and Zilberstein, S. 2021. Maximizing Legibility in Stochastic Environments. In *Proceedings of the Thirtieth IEEE International Conference on Robot and Human Interactive Communication*.

Miura, S.; Wray, K.; and Zilberstein, S. 2022. Heuristic Search for SSPs with Lexicographic Preferences over Multiple Costs. In *Proceedings of the Fifteenth International Symposium on Combinatorial Search*.

Miura, S.; and Zilberstein, S. 2021. A Unifying Framework for Observer-Aware Planning and its Complexity. In *Proceedings of the Thirty-Seventh Uncertainty in Artificial Intelligence*, 610–620.

Mouaddib, A.-I. 2004. Multi-objective decision-theoretic path planning. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*, volume 3, 2814–2819.

Pineda, L. E.; Wray, K. H.; and Zilberstein, S. 2015. Revisiting multi-objective MDPs with relaxed lexicographic preferences. In *2015 AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*.

Roijers, D. M.; and Whiteson, S. 2017. Multi-objective decision making. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 11(1): 1–129.

Trevizan, F.; Thiébaux, S.; Santana, P.; and Williams, B. 2017. I-dual: Solving constrained SSPs via heuristic search in the dual space. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4954–4958.

Wray, K. H.; Zilberstein, S.; and Mouaddib, A.-I. 2015. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3418–3424.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan Explicability and Predictability for Robot Task Planning. In *International Conference on Robotics and Automation*, 1313–1320.