Multi-Objective POMDPs for Robust Autonomy

Kyle Hollins Wray
Alliance Innovation Lab Silicon Valley

Stefan J. Witwicki
Alliance Innovation Lab Silicon Valley

***Abstract*—Decision-making in real-world robots requires a robustness to uncertainty in dynamic environments with a balancing across multiple objectives. This paper proposes a general model for robust multi-objective reasoning called a topological partially observable Markov decision process (TPOMDP) and its fully observable subclass (TMDP). TPOMDPs and TMDPs allow for additional objective measures, such as maximizing safety, smoothness, and/or other human preferences, to be incorporated into a typical POMDP or MDP objective, such as minimizing time or distance traveled. To enable use on a real robot, we also present a scalable solver for TPOMDPs. The model is discussed through comparisons of behaviors produced by POMDP policies on a fully operational autonomous vehicle prototype acting in the real world.**

I. Introduction

Reasoning about multiple objectives is prevalent in many real-world domains that require robust and safe control solutions, such as water reservoir control [3], industrial scheduling [1], energy-conserving smart environments [4], and anthrax outbreak detection [9]. Recently, multi-objective reasoning techniques have also been starting to be applied to autonomous robots, such as through notions of safety in semi-autonomous vehicles [16, 13, 17, 11]. Models for multi-objective reasoning offer unique capabilities when designing robots with long-term autonomy, as they allow for explicitly modeled safety, risk, and any other robustness constraints in conjunction with environmental uncertainties.

Multi-objective Markov decision processes (MOMDPs) represent a model of multiple objectives with two main methodologies to structure their typically conflicting nature: scalarization and preference orderings. Scalarization approaches attempt to weigh each objective properly in a complex function, creating a single-objective MDP which can be solved with standard techniques [7]. However, finding this scalarziation function is non-trivial, and suffers from both computational complexity issues and the conflation of the reward function, losing any semantic meaning the objectives might have once had. We instead leverage the latter, using a preference ordering over objectives [5, 8], such as in *constrained (PO)MDPs (C(PO)MDPs)* [2] and *lexicographic (PO)MDPs (L(PO)MDPs)* [16, 13]. We assign a preference structure and only considers other objectives in the case of tie-breaking, combined with the notion of slack or constraints to liberate successive objectives' choice. Our proposed model defines this ordering via the topological order of a directed acyclic graph (DAG) over the constraints, generalizing both C(PO)MDPs and L(PO)MDPs, and is called a *topological (PO)MDP (T(PO)MDP)*.

II. Robustness With Multi-Objective Reasoning

A **topological partially observable Markov decision process (TPOMDP)** is a sequential decision-making model defined by the tuple $\langle S, A, \Omega, T, O, \mathbf{R}, E, \delta \rangle$:

- S is a finite set of states;
- A is a finite set of actions;
- Ω is a finite set of observations;
- $T : S \times A \times S \to [0, 1]$ is a state transition function such that $T(s, a, s') = Pr(s'|s, a)$ is the probability of successor s' given action a was performed in state s;
- $O : A \times S \times \Omega \to [0, 1]$ is an observation function such that $O(a, s', \omega) = Pr(\omega|a, s')$ is the probability of observing ω given action a was performed resulting in successor s';
- $\mathbf{R} = [R_1, \ldots, R_k]^T$ is a vector of reward functions for $K = \{1, \ldots, k\}$ such that $R_i : S \times A \to \mathbb{R}$ denotes an reward $R_i(s, a)$ for performing action a in state s;
- $E \subseteq K \times K$ is a set of edges over k rewards forming a directed acyclic graph, with one leaf/sink reward vertex which, without loss of generality, is reward vertex k; and
- $\delta : E \to \mathbb{R}^+$ is a function mapping edges $e = \langle i, j \rangle \in E$ to a non-negative slack constraint $\delta(e) \geq 0$, or also overloading notation by the equivalent $\delta(i, j) \geq 0$.

As in a POMDP, the TPOMDP operates over a *belief* $b \in B \subseteq \triangle^{|S|}$ of the world. Given belief b, after performing a and observing ω, the next belief $b_{ba\omega}$ over state s' is:

$$b_{ba\omega}(s') \propto O(a, s', \omega) \sum_{s \in S} T(s, a, s') b(s) \tag{1}$$

A **topological Markov decision process (TMDP)** is a fully observable a TPOMDP with $\Omega = S$ and $O(a, s', s') = 1$, such that the reachable beliefs $b \in B$ are $b(s) = 1$ for all $s \in S$.

The agent chooses which action to perform via a *policy* $\pi : B \to A$. The objective in an *infinite horizon* TPOMDP seeks to maximize the expected discounted reward from an initial belief b^0 with discount factor $\gamma \in [0, 1)$. Formally, for a policy π, it follows: $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathbf{R}(b^t, \pi(b^t))|\pi, b^0]$ with b^t denoting the random variable for the belief at time t generated following T and O. The *value* $V^\pi : B \to \mathbb{R}$ is the expected reward at belief b following:

$$\mathbf{V}^\pi(b) = \mathbf{R}(b, \pi(b)) + \gamma \sum_{\omega \in \Omega} Pr(\omega|b, \pi(b)) \mathbf{V}^\pi(b'_{b\pi(b)\omega}) \tag{2}$$

and $\mathbf{R}(b, a) = \sum_s b(s) \mathbf{R}(s, a)$ and $b'_{\pi(b)\omega}$ following the belief update in Equation 1. Also, $\mathbf{Q}^\pi(b, a)$ refers to the one step deviation of π following action a in belief b instead of $\pi(b)$.

Algorithm 1 Local action restriction (LAR) approximation.

Require: $\langle S, A, \Omega, T, O, \mathbf{R}, E, \delta \rangle$: The TPOMDP.
Require: B: The set of beliefs $B \subseteq \triangle^{|S|}$.
Require: η: The local slack specific for each $e \in E$.

procedure LOCALACTIONRESTRICTION(K, E, $\mathbf{x}$, k, η)
  $\langle \pi^*, \hat{\mathbf{V}}^*, \mathbf{A} \rangle \leftarrow \langle \pi^0, \{\}, \{\} \rangle$
  for $i \leftarrow x_1, \ldots, x_k$ **do**
    $\mathcal{P}_i \leftarrow \{j \in K | \exists \langle j, i \rangle \in E\}$
    for $b \in B$ **do**
      $A_i(b) \leftarrow A \cap (\bigcap_{j \in \mathcal{P}_i} \{a \in A_j(b) | \hat{\epsilon}_j(b, a) \leq \eta(j, i)\})$
    $\langle \pi_i^*, \hat{V}_i^* \rangle \leftarrow$ PBVI(S, A_i, Ω, T, O, R_i, B)
    $\langle \pi^*, \hat{\mathbf{V}}^*, \mathbf{A} \rangle \leftarrow \langle \pi_i^*, \hat{\mathbf{V}}^* \cup \{\hat{V}_i^*\}, \mathbf{A} \cup \{A_i\} \rangle$
  return π^*
$\mathbf{x} \leftarrow$ REVERSEPOSTORDERDFS(K, E, k, η)
return LOCALACTIONRESTRICTION(K, E, $\mathbf{x}$, k, η)

We leverage the piecewise linear convex property of a similar **finite horizon** TPOMDP objective to approximate the infinite horizon TPOMDP. Following the same logic as POMDPs, we use a set of α-vectors Γ_i for each objective $i \in K$, with their collection $\mathbf{\Gamma} = \{[\alpha_1, \ldots, \alpha_k]^T \in \mathbb{R}^k | \forall i \in K, \alpha_i \in \Gamma_i\}$, to represent the value function. The equation for π at b is:

$$\mathbf{V}^{\pi}(b) = \mathbf{R}(b, \pi(b)) + \gamma \sum_{\omega \in \Omega} \max_{\alpha' \in \mathbf{\Gamma}} \sum_{s \in S} b(s) \sum_{s' \in S} T(s, \pi(b), s') O(\pi(b), s', \omega) \alpha'(s').$$

Point-based value iteration (PBVI) methods [6, 10, 14] apply this at a fixed set of beliefs B, as used here. Controller-based methods [15] can also be used with slack constraints [19].

A. Optimality Criterion

The topologically ordered constraints can subject predecessor objectives to satisfying slack at the initial belief or across all beliefs, called **initial slack** and **universal slack**, respectively. An initial slack TPOMDP objective for initial belief b^0 is the recursively defined objective to find a policy π that maximizes the expected value for reward $i \in K$ following:

$$\begin{array}{ll} \text{maximize} & V_i^{\pi}(b^0) \\ \text{subject to} & V_w^*(b^0) - V_w^{\pi}(b^0) \leq \delta(w, v) \\ & \forall v \in \mathcal{A}_i \cup \{i\}, \forall w \in \mathcal{P}_v \end{array} \quad (3)$$

with $V_w^*(b^0)$ denoting the optimal value of ancestor w recursively following this same constrained objective. The difference is also denoted $\epsilon_w(b, a) = V_w^*(s) - Q_w^*(s, a)$. Universal slack ensures the constraints are satisfied at all beliefs $b \in B$.

B. Scalable Approximate Algorithm

From Equation 3 we can recognize that applying universal slack at each belief $b \in B$, with a local slack of $\eta(j, i) \leq (1 - \gamma)\delta(j, i)$ ensures the global slack $\delta(j, i)$ is satisfied [12]. We call this **local action restriction (LAR)**. Algorithm 1 implements this approach. Crucially, this is for real-world robots and must be tractable. LAR is highly scalable, with a complexity of PBVI times the number of objectives k.

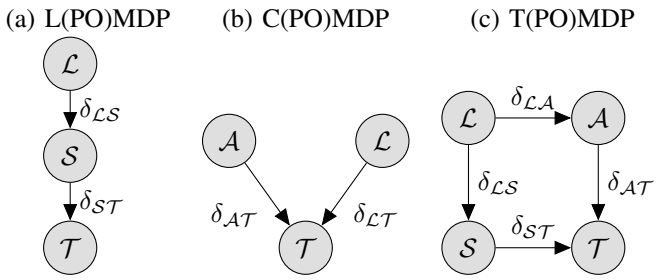

Fig. 1. Graphical notation representing robust autonomous vehicle T(PO)MDP topological constraints E. The vertices $K = \{\mathcal{L}, \mathcal{A}, \mathcal{S}, \mathcal{T}\}$ denote following the law, assertiveness, smoothness, and time objectives, respectively. For each vertex $i \in K$: (a) a three-reward L(PO)MDP; (b) a two-constraint C(PO)MDP; and (c) a general T(PO)MDP able to capture a richer landscape of robustness constraints.

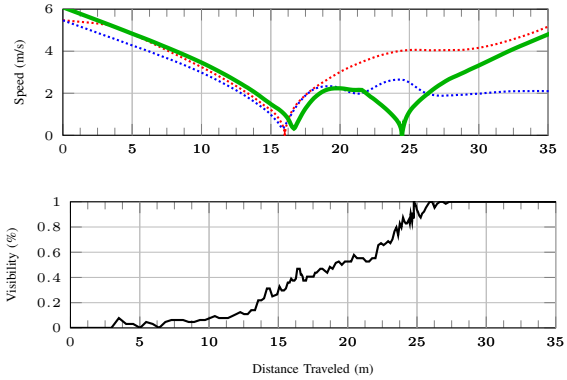

Fig. 2. Experiments of three POMDP policies on a fully operational autonomous vehicle prototype at an occluded T-intersection. For a TPOMDP implementation (e.g., following Fig 1.(c)), in all cases the law to stop at the stop line must be followed $\delta_{\mathcal{LA}} = \delta_{\mathcal{LS}} = 0$. The red baseline shows possible behavior of an assertive policy with $\delta_{\mathcal{AT}} = 0$. The blue baseline shows possible behavior of a smooth comfortable policy with $\delta_{\mathcal{ST}} = 0$. The green line shows desired behavior, which has careful motion for visibility, provided enough slack ($\delta_{\mathcal{AT}} > 0$ and $\delta_{\mathcal{ST}} > 0$) to maximize the time objective $\mathcal{T}$.

III. ROBUST AUTONOMOUS ROBOTS

In order to incorporate robustness optimization into an autonomous robot, we now can employ a T(PO)MDP. Real world robotic domains require more than just safety to be successfully deployed; the robots need to be robust across multiple considerations. The objectives can explicitly model robustness considerations (e.g., following the law, assertiveness, smoothness, safety, etc.) in addition to the main objective (e.g., minimizing time). Figure 1 shows an example of three T(PO)MDP graphs for an autonomous vehicle domain. Figure 2 illustrates an example of the kind of difference the graph E and slacks $\delta(e)$ will introduce in policy execution behavior, as shown by actual robot experiments using multiple distinct POMDPs. The POMDPs are those used in *MODIA* [18], a framework enabling scalable decision-making in robots.

TPOMDPs and TMDPs allow for a rich landscape of robustness constraints to be described in the theoretically grounded model [12]. This model facilitates the design of scalable algorithms, such as LAR in Algorithm 1, that enable TPOMDPs to be deployed on robots operating in the real world which actualize aspects of robust autonomy.

References

[1] Nassima Aissani, Bouziane Beldjilali, and Damien Trentesaux. Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach. *Engineering Applications of Artificial Intelligence*, 22(7): 1089–1103, 2009.

[2] Eitan Altman. *Constrained Markov decision processes*. Chapman & Hall/CRC Press, England, 1999.

[3] Andrea Castelletti, Francesca Pianosi, and Rodolfo Soncini-Sessa. Water reservoir control under economic, social and environmental constraints. *Automatica*, 44(6): 1595–1607, 2008.

[4] Jun-Young Kwak, Pradeep Varakantham, Rajiv Maheswaran, Milind Tambe, Farrokh Jazizadeh, Geoffrey Kavulya, Laura Klein, Burcin Becerik-Gerber, Timothy Hayes, and Wendy Wood. SAVES: A sustainable multiagent application to conserve building energy considering occupants. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 21–28, 2012.

[5] L. G. Mitten. Preference order dynamic programming. *Management Science*, 21(1):43–46, 1974.

[6] Joelle Pineau, Geoffrey Gordon, and Sebastian Thrun. Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.

[7] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[8] Matthew J. Sobel. Ordinal dynamic programming. *Management Science*, 21(9):967–975, 1975.

[9] Harold Soh and Yiannis Demiris. Multi-reward policies for medical applications: Anthrax attacks and smart wheelchairs. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, pages 471–478, 2011.

[10] Matthijs Spaan and Nikos Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, 24:195–220, 2005.

[11] Justin Svegliato, Kyle Hollins Wray, Stefan J. Witwicki, Joydeep Biswas, and Shlomo Zilberstein. Belief space metareasoning for exception recovery. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1224–1229, 2019.

[12] Kyle Hollins Wray. *Abstractions in Reasoning for Long-Term Autonomy*. PhD thesis, University of Massachusetts, Amherst, MA, 2019.

[13] Kyle Hollins Wray and Shlomo Zilberstein. Multi-objective POMDPs with lexicographic reward preferences. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1719–1725, 2015.

[14] Kyle Hollins Wray and Shlomo Zilberstein. A parallel point-based POMDP algorithm leveraging GPUs. In *Proceedings of the 2015 AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, pages 95–96, 2015.

[15] Kyle Hollins Wray and Shlomo Zilberstein. Generalized controllers in POMDP decision-making. In *2019 IEEE International Conference on Robotics and Automation*, pages 7166–7172, 2019.

[16] Kyle Hollins Wray, Shlomo Zilberstein, and Abdel-Illah Mouaddib. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3418–3424, 2015.

[17] Kyle Hollins Wray, Luis Pineda, and Shlomo Zilberstein. Hierarchical approach to transfer of control in semi-autonomous systems. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 517–523, 2016.

[18] Kyle Hollins Wray, Stefan J. Witwicki, and Shlomo Zilberstein. Online decision-making for scalable autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4768–4774, 2017.

[19] Kyle Hollins Wray, Akshat Kumar, and Shlomo Zilberstein. Integrated cooperation and competition in multi-agent decision-making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4751–4758, 2018.