Multi-Objective RL With Multi-Agent Bidding

Anonymous authors
Paper under double-blind review

Keywords: Multi-Objective RL, Multi-Agent RL

Summary

The summary appears on the cover page. Although it can be identical to the abstract, it does not have to be. One might choose to omit the stated contributions in the Summary, given that they will be stated in the box below. The original abstract may also be extended to two paragraphs. The authors should ensure that the contents of the cover page fit entirely on a single page. The cover page does **not** count towards the 8–12 page limit.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contribution(s)

1. Provide a succinct but precise list of the contribution(s) of the paper. Use contextual notes to avoid implications of contributions more significant than intended and to clarify and situate the contribution relative to prior work (see the examples below). If there is no additional context, enter “None”. Try to keep each contribution to a single sentence, although multiple sentences are allowed when necessary. If using complete sentences, include punctuation. If using a single sentence fragment, you may omit the concluding period. A single contribution can be sufficient, and there is no limit on the number of contributions. Submissions will be judged mostly on the contributions claimed on their cover pages and the evidence provided to support them. Major contributions should not be claimed in the main text if they do not appear on the cover page. Overclaiming can lead to a submission being rejected, so it is important to have well-scoped contribution statements on the cover page.
   Context: None
2. The submission template for submissions to RLJ/RLC 2025
   Context: Built from previous RLC/RLJ, ICLR, and TMLR submission templates
3. *[Example of one contribution and corresponding contextual note for the paper “Policy gradient methods for reinforcement learning with function approximation” (**?**).]*
   This paper presents an expression for the policy gradient when using function approximation to represent the action-value function.
   Context: Prior work established expressions for the policy gradient without function approximation (**?**).

Multi-Objective RL With Multi-Agent Bidding

Anonymous authors
Paper under double-blind review

Abstract

1 Introduction

2 Compositional RL on Multi-Objective MDPs

Mainstream RL algorithms consider Markov decision processes (MDP) equipped with a *single* reward function, pertaining to a single task or *objective* for the system. In reality, a majority of real-world applications of RL requires satisfying multiple, partly contradictory objectives. We model such multi-objective decision-making problems using multi-objective MDPs (MO-MDP), as formally defined below. Intuitively, an MO-MDP has the exact same syntax as a regular MDP, except that it now has multiple reward functions pertaining to the different objectives.

Definition 1 (MO-MDP). *A multi-objective Markov decision process (MO-MDP) with $m \in \mathbb{Z}_{>0}$ objectives is specified by a tuple $\mathcal{M} = (S, A, T, R, \mu_0)$, where*

- *S is the set of states,*
- *A is the set of actions,*
- *$T : S \times A \to \mathbb{D}(S)$ is the transition function mapping a state-action pair to a distribution over the successor states,*
- *$\{R_i : S \times A \times S \to \mathbb{R}\}_{i\in[1;m]}$ is the set of reward functions pertaining to different objectives, and*
- *$\mu_0 \in \mathbb{D}(S)$ is the initial state distribution.*

The notions of policies and paths induced by them are exactly the same as in classical MDPs, which we briefly recall below. A *policy* in an MO-MDP $\mathcal{M}$ is a function $\pi : (S \times A)^* \times S \to \mathbb{D}(A)$ that maps a history of state-action pairs and the current state to a distribution over actions. A *path* on $\mathcal{M}$ induced by π is a sequence $(s_0, a_0)(s_1, a_1), \ldots \in (S \times A)^\infty$ such that for every $i \geq 0$, $\pi((s_0, a_0) \ldots (s_i, a_i), s_{i+1})(a_{i+1}) > 0$ and $T(s_i, a_i)(s_{i+1}) > 0$. A path can be either finite or infinite, and we will write $\mathit{Paths}(\mathcal{M}, \pi)$ to denote the set of all infinite paths fo $\mathcal{M}$ induced by π. Given a finite path $\rho = (s_0, a_0) \ldots (s_i, a_i)$, the probability that ρ occurs is given by: $\mu_0(s_0) \cdot \prod_{j=0}^{i-1} T(s_j, a_j)(s_{j+1}) \cdot \pi((s_0, a_0) \ldots (s_j, a_j), s_{j+1})(a_{j+1})$. This can be extended to a probability measure over the set of all infinite paths in $\mathcal{M}$ using standard constructions, which can be found in the literature (Baier & Katoen, 2008). Given a measurable set of paths Ω and a function $f\colon \mathit{Paths}(\mathcal{M}, \pi) \to \mathbb{R}$, we will write $\mathbb{P}^{\mathcal{M},\pi}[\Omega]$ and $\mathbb{E}^{\mathcal{M},\pi}[f]$ to denote, respectively, the probability measure of Ω and the expected value of f evaluated over random infinite paths.

We will use the standard discounted reward objectives. Suppose $\gamma \in [0, 1]$ is a fixed discounting factor. Let $\theta = r_0 r_1 \ldots \in \mathbb{R}^\omega$ be an infinite sequence of real numbers. Define the discounted sum function: $f_{\mathrm{ds}}(\theta) \coloneqq \sum_{i=0}^{\infty} \gamma^i \cdot r_i$. Then, the *value* of the policy ρ for $\mathcal{M}$ is the expected value of the discounted sum of rewards we can secure by executing ρ on $\mathcal{M}$, i.e, $\mathit{val}^{\mathcal{M}}(\pi) = \mathbb{E}^{\mathcal{M},\pi}[f_{\mathrm{ds}}]$.

3 A Multi-Agent Bidding Approach for Multi-Objective RL

Definition 2 (MO-MDP)**.** *A multi-objective Markov decision process (MO-MDP) with* $m \in \mathbb{Z}_{>0}$ *objectives is specified by a tuple* $\mathcal{M} = (S, A, T, R, \mu_0)$*, where*

- S *is the set of states,*
- A *is the set of actions,*
- $T : S \times A \to \mathbb{D}(S)$ *is the transition function mapping a state-action pair to a distribution over states,*
- $R : S \times A \times S \to \mathbb{R}^m$ *is the reward function with each output component corresponding to the different objectives, and*
- $\mu_0 \in \mathbb{D}(S)$ *is the initial state distribution.*

A *policy* in an MO-MDP $\mathcal{M}$ is a function $\pi : (S \times A)^* \times S \to \mathbb{D}(A)$ that maps a history of state-action pairs and the current state to a distribution over actions.

Definition 3 (MAB-MDP)**.** *Let* $\mathcal{M} = (S, A, T, R, \mu_0)$ *be an MO-MDP with* m *objectives and let* $b \in \mathbb{Z}_{>0}$ *be the bid upper bound. Also, define* $M = \{1, \ldots, m\}$ *be indices of the* m *agents corresponding to the* m *objectives along with* $\perp$ *representing a null agent. Lastly, let* $B = \{0, \ldots, b\}$ *be the range of bids and* $\rho > 0$ *be the bid penalty factor. We define the multi-agent bidding Markov decision process (MAB-MDP) as a tuple* $\mathcal{B}_{\mathcal{M}} = (\hat{S}, \hat{A}, \hat{T}, P, \hat{R}, \hat{\mu}_0)$ *where*

- $\hat{S} = M \times S$ *is the new state space augmented with the index of the agent that won the previous round of bidding,*
- $\hat{A} = A^m \times B^m$ *represents the action space of the* m *agents in which each agent selects an action from* A *and a bid from* B*,*
- $\hat{T} : \hat{S} \times \hat{A} \to \mathbb{D}(\hat{S})$ *is the new transition function defined as,*

$$\hat{T}((\_, s), (\mathbf{a}, \mathbf{b})) \coloneqq \frac{1}{|B_{\max}|} \sum_{i \in B_{\max}} (T(s, a_i), i)$$

  where $B_{\max} \coloneqq \{i \mid b_i = \max\{b_1, \ldots, b_m\}\}$ *is the set of agents with maximal bids. The tuple* $(T(s, a_i), i)$ *represents the distribution over* $\hat{S}$ *induced by the original transition function* T *such that the second component is fixed, and the weighted sum represents taking the weighted sums of the distributions over* $\hat{S}$*.*
- $P : \hat{A} \times M \to \mathbb{R}^m$ *is the bidding penalty for the* m *agents and the second component is the index of the agent that won the bidding.*
- $\hat{R} : \hat{S} \times \hat{A} \times \hat{S} \to \mathbb{R}^m$ *is the reward function for the* m *agents with*

$$\hat{R}_k((\_, s_0), (\mathbf{a}, \mathbf{b}), (i, s)) \coloneqq R_k(s_0, a_i, s) - P_k((\mathbf{a}, \mathbf{b}), i)$$

  where $i \in M$ *is the index of the agent that won the bid and chose the action.*
- $\hat{\mu}_0 \coloneqq (\mu_0, 1)$ *is the initial state distribution over* $\hat{S}$ *induced by* μ_0 *and the second component is fixed to be* 1 *without loss of generality.*

Given an MAB-MDP $\mathcal{B}_{\mathcal{M}}$, a *policy* for each agent indexed by $i \in \{1, \ldots, m\}$ takes a similar form: $\pi_i : (\hat{S} \times \hat{A})^* \times \hat{S} \to \hat{A}$. Intuitively, a state $(i, s) \in \hat{S}$ encodes the agent that won the bidding and chose the action to reach s in the previous step. At each step, each of the agents choose an action and a bid, and an action amongst the set of highest bidders is chosen uniformly at random. The reward function includes a penalty term that captures the desired bidding mechanism.

References

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT press, 2008.