

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, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur 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

2 1 Introduction

3 2 A Multi-Agent Bidding Approach for Multi-Objective RL

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

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

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

15 **Definition 2** (MAB-MDP). Let $\mathcal{M} = (S, A, T, R, \mu_0)$ be an MO-MDP with m objectives and
 16 let $b \in \mathbb{Z}_{>0}$ be the bid upper bound. Also, define $M = \{\perp, 1, \dots, m\}$ be indices of the m agents
 17 corresponding to the m objectives along with \perp representing a null agent. Lastly, let $B = \{0, \dots, b\}$
 18 be the range of bids and $\rho > 0$ be the bid penalty factor. We define the multi-agent bidding Markov
 19 decision process (MAB-MDP) as a tuple $\mathcal{B}_{\mathcal{M}} = (\hat{S}, \hat{A}, \hat{T}, \hat{R}, \hat{\mu}_0)$ where

- 20 • $\hat{S} = S \times M$ is the original state space accompanied by an agent index,
- 21 • $\hat{A} = (A \times B)^m$ represents the action space of the m agents in which each agent selects an action
 22 from A and a bid from B ,
- 23 • $\hat{T} : \hat{S} \times \hat{A} \rightarrow \mathbb{D}(\hat{S})$ is the new transition function defined as,

$$\hat{T}((s, _), ((a_1, b_1), \dots, (a_m, b_m))) := \frac{1}{|B_{\max}|} \sum_{i \in B_{\max}} (T(s, a_i), i)$$

24 where $B_{\max} := \{i \mid b_i = \max\{b_1, \dots, b_m\}\}$ is the set of agents with maximal bids. The tuple
 25 $(T(s, a_i), i)$ represents the distribution over \hat{S} induced by the original transition function T such
 26 that the second component is fixed, and the weighted sum represents taking the weighted sums of
 27 the distributions over \hat{S} .

- 28 • $\hat{R} : \hat{S} \times \hat{A} \times \hat{S} \rightarrow \mathbb{R}^m$ is the reward function for the m agents with

$$\hat{R}_k((s_0, _), ((a_1, b_1), \dots, (a_m, b_m)), (s, i)) := \begin{cases} R_k(s_0, a_k, s) - \rho b_k & i = k \\ R_k(s_0, a_i, s) & \text{otherwise} \end{cases}$$

29 • $\hat{\mu}_0 := (\mu_0, \perp)$ is the initial state distribution over \hat{S} induced by μ_0 and the second component is
 30 fixed to be \perp .

31 Given an MAB-MDP $\mathcal{B}_{\mathcal{M}}$, a policy for each agent indexed by $i \in \{1, \dots, m\}$ takes a similar form
 32 $\pi_i : (\hat{S} \times \hat{A})^* \times \hat{S} \rightarrow \hat{A}$. Intuitively, a state $(s, i) \in \hat{S}$ encodes the agent that won the bidding and
 33 chose the action to reach s in the previous step. At each step, each of the agents choose an action and
 34 a bid, and an action amongst the set of highest bidders is chosen uniformly at random. The reward
 35 function is designed so that the agent whose action was chosen receives a bidding penalty.

36 **References**