

Multi-Objective Planning with Contextual Lexicographic Reward Preferences

Pulkit Rustagi
Oregon State University
Corvallis, OR, USA
rustagip@oregonstate.edu

Yashwanthi Anand
Oregon State University
Corvallis, OR, USA
anandy@oregonstate.edu

Sandhya Saisubramanian
Oregon State University
Corvallis, OR, USA
sandhya.sai@oregonstate.edu

ABSTRACT

Autonomous agents are often required to plan under multiple objectives whose preference ordering varies based on *context*. The agent may encounter multiple contexts during its course of operation, each imposing a distinct lexicographic ordering over the objectives, with potentially different reward functions associated with each context. Existing approaches to multi-objective planning typically consider a single preference ordering over the objectives, across the state space, and do not support planning under *multiple objective orderings within an environment*. We present Contextual Lexicographic Markov Decision Process (CLMDP), a framework that enables planning under varying lexicographic objective orderings, depending on the context. In a CLMDP, both the *objective ordering* at a state and the associated *reward* functions are determined by the context. We employ a Bayesian approach to infer a state-context mapping from expert trajectories. Our algorithm to solve a CLMDP first computes a policy for each objective ordering and then combines them into a single context-aware policy that is valid and cycle-free. The effectiveness of the proposed approach is evaluated in simulation and using a mobile robot.

KEYWORDS

Context-aware planning; multi-objective planning.

ACM Reference Format:

Pulkit Rustagi, Yashwanthi Anand, and Sandhya Saisubramanian. 2025. Multi-Objective Planning with Contextual Lexicographic Reward Preferences. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Detroit, Michigan, USA, May 19 – 23, 2025, IFAAMAS, 9 pages.

1 INTRODUCTION

In many real-world applications, such as navigation [12, 26], and warehouse management [13, 22], autonomous agents must optimize multiple, potentially competing objectives. The priority over these objectives can be conveniently expressed via a lexicographic ordering, and may vary based on the *context*. Consider a semi-autonomous car navigating in a city (Figure 1), where the possible set of objectives includes minimizing travel time, ensuring pedestrian safety, and minimizing going over uneven road surfaces. When the car navigates through a construction zone, it must prioritize minimizing going over uneven surfaces, followed by pedestrian

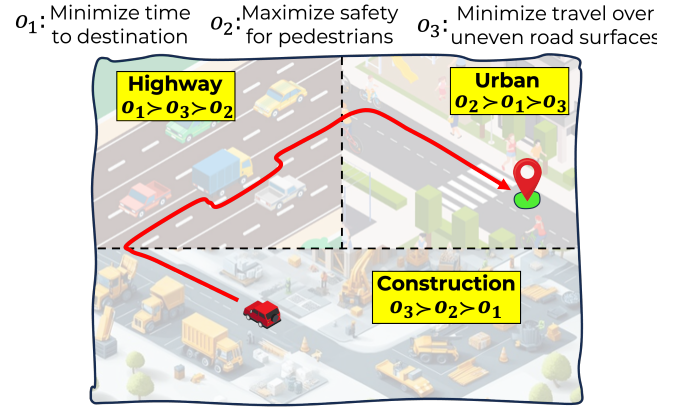


Figure 1: Example of a car navigating in a city environment with three different contexts (urban, highway, and construction), each imposing a unique ordering over the objectives.

safety, and finally the speed. When the car is on a highway, it is desirable to prioritize travel time over other objectives, since a highway is designed for high-speed navigation and is typically free of pedestrian traffic and other obstacles. When in urban areas with high foot traffic, ensuring pedestrian safety becomes the highest priority. Thus, during the course of its navigation, the car must optimize different *objective orderings*, corresponding to the context associated with the region. The reward functions associated with the objectives may also vary based on context. Using contextual information for decision-making has often proven to be valuable, such as in contextual bandits [17], context-aware planning [16], contextual information retrieval and text generation [14, 32]. The existing literature often defines context as a parameter that influences the environment dynamics and rewards [6, 16, 18], the position of obstacles and other agents [15, 29], or the area of operation [4]. We define context as a set of *exogenous features* in the state representation that determine the objective ordering and corresponding reward functions for each state.

We consider a lexicographic ordering over objectives, which is a convenient approach for representing relative preferences over multiple objectives. Traditional approaches for multi-objective decision-making typically support only a single preference ordering over objectives throughout the entire state space. A prior work [34] that supports planning with multiple state partitions could potentially be adapted to handle different priority orderings. However, it lacks a systematic method for defining partitions and does not provide an explicit mapping from states to their respective objective orderings. Their framework also does not support varying reward



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

functions for each partition or ensure that the resulting policy is valid and cycle-free. A recent work [36] uses a context map to track the scalarization weights for each objective in every state. However, it does not scale to larger state spaces.

Our paper addresses these challenges by introducing *Contextual Lexicographic Markov Decision Process* (CLMDP), a framework that supports seamless switching between different objective orderings, determined by the state’s context, during the course of agent operation. We present an algorithm to solve CLMDP by first computing a policy for each objective ordering across the state space, and then combining them into a single context-aware policy, given a state-context mapping. Our algorithm also ensures that the action recommendations across contexts do not conflict with each other, during the merge step. When state-context mapping is unknown, it is derived using Bayesian inference and limited expert trajectories.

Our primary contributions are: (1) formalizing the problem of contextual multi-objective planning as a contextual lexicographic MDP; (2) presenting an algorithm to solve a CLMDP, producing a policy that is free of conflicts in a setting where multiple contexts coexist; (3) theoretical analysis of the algorithm’s performance; (4) a Bayesian approach to infer the state-context mapping; and (5) empirical evaluation in simulation and using a physical robot.

2 RELATED WORK

Multi-objective planning. Multi-objective planning is gaining increasing attention. Common approaches to solve multi-objective planning problems include scalarization [2, 31, 33], lexicographic planning [19, 24, 25, 34], constraint optimization [9, 10], and Pareto optimization [3, 20]. *Scalarization* combines multiple objectives into a single objective value using weighted sum. While it is a popular approach, it often requires non-trivial parameter tuning [23]. Our work is based on the *lexicographic* formulation that considers a lexicographic ordering over objectives and solves them sequentially. While the lexicographic Markov decision process (LMDP) [34] can support multiple state partitions, each corresponding to a different objective ordering, it suffers from the following limitations: (1) lack of a principled approach to define state partitions, especially when a state could be associated with multiple contexts such as weather, road conditions, and time of day; (2) hard-coded partitions and lack of explicit state-to-context mapping which makes it difficult to adapt to newer settings where the set of contexts may change; (3) does not support scenarios where the context influences the reward function; and (4) lack of tools to resolve conflicts in action recommendation that arise from solving partitions sequentially, with no principled approach to determine the order. The CLMDP framework addresses these shortcomings and facilitates smooth transitions between contexts, avoiding the risk of conflicting actions. The explicit state-to-context mapping in CLMDP offers flexibility and scalability in handling new scenarios, *without* requiring manual interventions to redefine partitions or objective priorities.

Constraint optimization. approach optimizes a single objective and converts other objectives into constraints [9, 10]. This approach cannot efficiently balance the trade-offs between different objectives [11]. *Pareto* optimization finds non-dominated solutions across objectives [3, 20], making it unsuitable for scenarios where specific preference orderings must be satisfied.

Context-aware planning. Prior works use contextual information to determine environment dynamics and rewards [6, 16, 18], or represent specific configurations like obstacle layouts and operational areas [4, 15, 29]. While many definitions of context exist for different problem settings, we focus our discussion on those most pertinent to multi-objective scenarios. In multi-objective settings, context has been used to assign scalarization weights [36] but this approach struggles to scale in larger state spaces. We integrate contextual information in a lexicographic setting to enable planning with different objective ordering in different regions of the state space, with associated reward functions determined by the context.

Bayesian Inference. It is commonly used to infer values of unknown parameters by incorporating prior knowledge and updating beliefs as new information becomes available [1, 5, 37]. Bayesian inference is often used in inverse reinforcement learning (IRL) to estimate reward functions [7, 21], and goal inference in multi-agent settings [30, 37]. In multi-objective optimization problems, Bayesian methods are used to approximate the set of Pareto optimal solutions for competing objectives [5, 8], and identify solutions that satisfy user-defined preference ordering over competing objectives [1]. In this work, we apply Bayesian inference to determine the most likely context of a state, using limited number of expert trajectories.

3 CONTEXTUAL LEXICOGRAPHIC MDP

Consider an agent operating in an environment with multiple objectives modeled as a multi-objective Markov decision process (MOMDP). The agent must optimize a lexicographic ordering over the n primitive objectives $\mathbf{o} = \{o_1, \dots, o_n\}$. We focus on goal-oriented MDPs where one of the objectives o_i is to maximize the reward associated with reaching the goal. The lexicographic ordering over objectives at a *particular state* is determined by the *context* associated with it. The context, inferred from a set of *exogenous* features in the state representation, determines the objective ordering at a state and the reward functions associated with the objectives. In the example illustrated in Figure 1, the set of contexts in the environment are: highway, urban area, and construction zone. Each context imposes a unique ordering over the objectives, which determines an acceptable behavior for that context. Since an agent may encounter multiple contexts during the course of its operation, its reasoning module must be able to seamlessly switch between optimizing different objective orderings.

We introduce *Contextual Lexicographic Markov Decision Process* (CLMDP), a framework that facilitates optimizing different objective orderings in different regions of the state space.

DEFINITION 1. A contextual lexicographic Markov decision process (CLMDP) is denoted by $M = \langle C, \Omega, \mathbf{o}, \mathbf{w}, f_w, S, A, T, \mathbf{R}, f_R, \mathcal{Z} \rangle$, where:

- $C = \{c_1, \dots, c_m\}$ denotes a finite set of contexts;
- Ω is a lexicographic ordering over contexts in C ;
- $\mathbf{o} = \{o_1, \dots, o_n\}$ denotes the set of primitive objectives;
- $\mathbf{w} = \{\omega_1, \dots, \omega_m\}$ is a set of unique lexicographic orderings over all primitive objectives, with ω_i denoting a lexicographic ordering over all objectives such as $o_1 \succ \dots \succ o_n$, $o_i \in \mathbf{o}$;
- $f_w : C \rightarrow \mathbf{w}$ is a function that maps a context $c_j \in C$ to a lexicographic ordering over the primitive objectives $\omega_i \in \mathbf{w}$;
- S is a finite set of states, with initial state $s_0 \in S$;

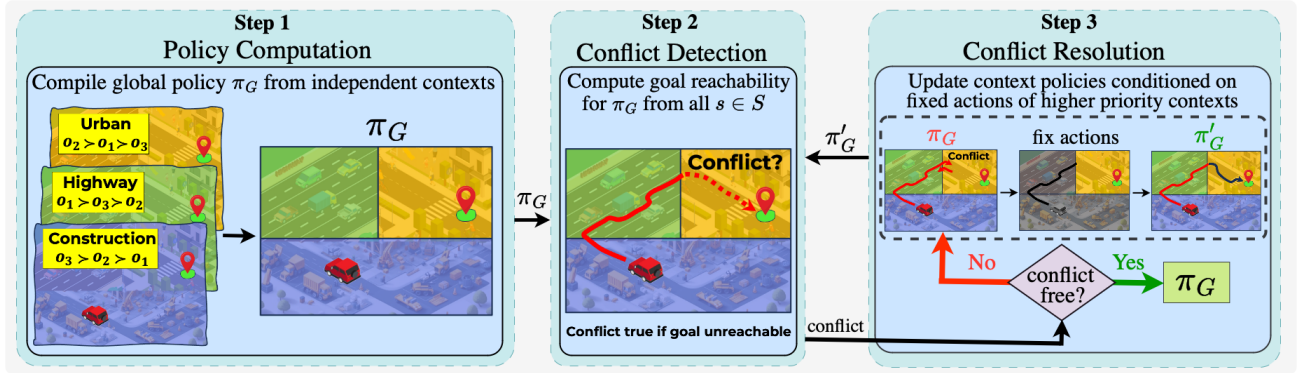


Figure 2: Overview of our solution approach for contextual planning. First, policies are calculated for each context in isolation, across the entire state space, and then compiled into a global policy π_G by mapping actions to states based on each state’s associated context. Second, π_G is analyzed for cycles by estimating goal reachability from each state. Finally, the detected conflicts are resolved by updating lower priority context policies conditioned on fixed actions of higher priority contexts.

- A is a finite set of actions;
- $T : S \times A \times S \rightarrow [0, 1]$ is the transition function that determines the probability of reaching state s' from state s by executing action a , independent of the context;
- $R = [R_1, \dots, R_n] \in \mathbb{R}^n$ is the space of possible reward vectors, where for each $o_i \in \mathbf{o}$, $R_i : S \times A \rightarrow \mathbb{R}$;
- $f_R : C \rightarrow \mathbb{R}$ maps context $c_j \in C$ to a vector of context-specific reward functions, $f_R(c_j) = [R_{1,j}, \dots, R_{n,j}] \in \mathbb{R}$; and
- $\mathcal{Z} : S \rightarrow C$ is a deterministic mapping of each $s \in S$ to $c \in C$

In a CLMDP, a state can belong to multiple contexts simultaneously, when some state features are shared across contexts. For example in Fig. 1, a road segment with uneven road surface may be associated with both urban and construction zone contexts. In such cases, \mathcal{Z} maps the state to the context with higher priority in the meta-ordering Ω . This ensures that the agent’s decision-making process remains consistent and aligned with the most critical objective at that state. When \mathcal{Z} maps every state to the same context $c \in C$, the CLMDP becomes an MDP with a single lexicographic ordering over the objectives as in [34]. In this paper, we consider \mathcal{Z} to be stationary and deterministic but the framework can be extended to support non-stationary \mathcal{Z} .

The following section presents an algorithm to solve a CLMDP, given \mathcal{Z} . In Section 5, we present a Bayesian approach to infer \mathcal{Z} .

4 SOLUTION APPROACH

We begin with an overview of our solution approach, illustrated in Figure 2. Given \mathcal{Z} , our approach solves CLMDPs by computing a policy π_i for each context $c_i \in C$ independently, and then combines them into a global policy π_G , based on the context associated with each state. Combining different π_i into π_G may result in cycles as each policy is computed independent of other policies and contexts. The cycles are detected and resolved by updating the policies associated with lower priority contexts, based on Ω .

4.1 Computing global policy π_G

A policy π_i for each context $c_i \in C$ is first computed independent of other contexts. We do this by considering each $c_i \in C$ to be the

only context across the entire state space and use the corresponding lexicographic ordering over objectives ($f_w(c_i)$) and the associated reward functions ($f_R(c_i)$) over all states. This multi-objective problem, with a single ordering over the objectives, is solved using lexicographic value iteration algorithm [34]. Each π_i specifies the optimal actions for its respective context. These individual policies are then compiled into a global policy π_G , where actions are mapped to each state, based on the actual context of each state, following state-to-context mapping \mathcal{Z} .

4.2 Conflict Detection

Combining multiple policies, computed under different contexts and associated objective orderings, can lead to conflicting action recommendations that result in cycles and affect goal reachability.

DEFINITION 2. A policy π_G , in a goal-oriented MDP with goal state s_g , is said to have a **conflict** if there exists at least one state s from which the probability of reaching the goal, $\Pr(s_g | s, \pi_G) = 0$, $\exists s \in S$.

DEFINITION 3. A **conflict-free policy** π_G has a non-zero probability of reaching the goal from every state s , $\Pr(s_g | s, \pi_G) > 0$, $\forall s \in S$.

ASSUMPTION 1. There exists a conflict-free policy under Ω .

We assume that there exists an underlying conflict-free policy for CLMDP and any cycles detected are due to π_G computation based on policies calculated for individual contexts. To detect conflicts that introduce dead-ends, Algorithm 1 estimates the goal reachability for each state, under π_G . The goal reachability is calculated using a Bellman backup-based method (lines 11–18) with a reward function R_e defined as follows:

$$R_e(s) = \begin{cases} +1, & \text{for } s = s_{goal} \\ 0, & \text{otherwise.} \end{cases}$$

States with $V_r(s) = 0$ do not have a path to the goal, indicating a conflict (line 25). Thus, if $V_r(s) = 0$ for any $s \in S$, it indicates a conflict. Conversely, if all states have $V_r > 0$, the policy is conflict-free. To handle large state spaces, a modified value iteration is used, marking states as solved once they individually converge

Algorithm 1 ConflictChecker

```

1: Input Global policy  $\pi_G$ , state space  $S$ , transition function  $T$ ,
   discount factor  $\gamma$ , goal state  $s_g$ 
2: Initialize conflict  $\leftarrow$  False,  $\delta \leftarrow 0$ ,  $\epsilon \leftarrow 10^{-6}$ 
3: Initialize  $V_r \leftarrow \{s : 0 \mid s \in S\}$ ,  $\Delta \leftarrow \{s : 0 \mid s \in S\}$ 
4: Initialize  $R_e \leftarrow \{s : 0 \mid s \in S\}$ ,  $R_e(s_g) \leftarrow 1.0$ 
5: Initialize  $S_{solved} \leftarrow []$ ,  $S_{unsolved} \leftarrow S$ ,  $S_{conflict} \leftarrow []$ 
6:  $S_{solved} \cdot \text{APPEND}(s_g)$   $\triangleright$  only  $s_g$  is solved initially
7:  $S_{unsolved} \cdot \text{REMOVE}(s_g)$ 
8: while  $S_{unsolved} \neq \emptyset$  do  $\triangleright$  continue if unsolved states remain
9:    $V'_r \leftarrow V_r$ 
10:   $S_{check} \leftarrow []$ 
11:  for  $s \in S_{unsolved}$  do  $\triangleright$  update  $V_r$  for all unsolved states
12:     $V_r(s) \leftarrow R_e(s)$ 
13:    for  $s' \in S$  &  $T(s, \pi_G(s), s') > 0$  do
14:       $V_r(s) \leftarrow V_r(s) + \gamma \cdot T(s, \pi_G(s), s') \cdot V_r(s')$ 
15:      if  $s' \in S_{solved}$  &  $s \notin S_{check}$  then
16:         $S_{check} \cdot \text{APPEND}(s)$   $\triangleright$  if  $s$  leads to a solved state
17:       $\delta \leftarrow \max(\delta, |V'_r(s) - V_r(s)|)$ 
18:       $\Delta(s) \leftarrow |V'_r(s) - V_r(s)|$ 
19:  for  $s \in S_{check}$  do
20:    if  $\Delta(s) < \epsilon$  then  $\triangleright$  label  $s$  as solved if  $V_r(s)$  converged
21:       $S_{solved} \cdot \text{APPEND}(s)$ 
22:       $S_{unsolved} \cdot \text{REMOVE}(s)$ 
23:  if  $\delta < \epsilon$  or  $S_{unsolved} = \emptyset$  then
24:    break
25:   $S_{conflict} \leftarrow [s \mid s \in S \text{ if } V_r(s) = 0]$   $\triangleright$  list of conflict states
26:  if  $S_{conflict} \neq \emptyset$  then
27:    conflict  $\leftarrow$  True
28: return conflict,  $S_{conflict}$ 

```

(lines 19–22). Since a converged state's value no longer changes, skipping further updates ensures computational efficiency without compromising the correctness of the solution, allowing for early termination. For problems with very long trajectories, Bellman backups can be executed in logarithmic space to avoid false positives caused by product over low values [35]. Algorithm 1 outputs a Boolean variable indicating the presence of conflicts, along with all the states from which the goal is unreachable.

4.3 Conflict Resolver

To resolve conflicts identified by Algorithm 1, conflict resolver in Algorithm 2 updates policies starting from the lowest-priority context involved in the conflict and moves upwards to the highest priority, while keeping the policies of higher-priority contexts fixed. The algorithm begins by identifying a set of contexts requiring updates, denoted as C_{update} (lines 4–10), starting with the lowest priority context involved in the conflict, based on Ω . This set expands *iteratively*, adding higher-priority contexts to C_{update} if conflicts persist. At each step, actions for states in higher-priority contexts outside of C_{update} are fixed according to π_G , creating an updated action space A_{new} (lines 11–12). This ensures that these contexts remain unaffected during updates to other contexts. The policies for contexts in C_{update} are then updated sequentially from highest to lowest

Algorithm 2 ConflictResolver

```

1: Input policy  $\pi_G$ , set of contexts  $C$ , state to context mapping
   function  $\mathcal{Z}$ , context to objective ordering map  $f_w$ , meta-
   ordering over contexts  $\Omega \equiv c_1 \succ \dots \succ c_m$ ,  $S$ ,  $A$ ,  $T$ ,  $R$ ,  $\gamma$ , goal
    $s_g$ , list conflict states from ConflictChecker  $S_{conflict}$ 
2: Initialize  $A_{new} \leftarrow \{\}$ ,  $\Pi \leftarrow \{\}$ ,  $c^* \leftarrow c_1$ 
3: if conflict = False then return  $\pi_G$ 
4: for  $s \in S_{conflict}$  do
5:   if  $c^* \succ_{\Omega} \mathcal{Z}(s)$  then
6:      $c^* \leftarrow \mathcal{Z}(s)$   $\triangleright$  stores lowest priority conflict context
7: for  $c_i$  varying from  $c^*$  to  $c_1$  do  $\triangleright$  low to high priority
8:    $C_{update} \leftarrow \{\}$ ,  $A_{new} \leftarrow A$ 
9:   for  $c_j$  varying from  $c_i$  to  $c_m$  do
10:     $C_{update} \leftarrow C_{update} \cup c_j$   $\triangleright$  contexts to be updated
11:   for  $s \in S$  and  $\mathcal{Z}(s) \in C \setminus C_{update}$  do
12:      $A_{new}[s] \leftarrow \pi_G[s]$   $\triangleright$  fix actions for  $s \notin C_{update}$ 
13:   for  $c_j$  varying highest to lowest priority in  $C_{update}$  do
14:      $\Pi[c_j] \leftarrow LVI(S, A_{new}, T, R, f_w(c_j))$   $\triangleright$  new policy for  $c_j$ 
15:     for  $s \in S$  and  $\mathcal{Z}(s) = c_j$  do
16:        $A_{new}[s] \leftarrow \Pi[c_j][s]$   $\triangleright$  fix actions for  $s \in c_j$ 
17:        $\pi_G[s] \leftarrow \Pi[c_j][s]$ 
18:   conflict,  $\_ \leftarrow$  ConflictChecker( $\pi_G$ ,  $S$ ,  $T$ ,  $\gamma$ ,  $s_g$ )
19:   if no conflict then return  $\pi_G$ 
20: return  $\pi_G$ 

```

priority, following Ω and using A_{new} to condition policy updates on the fixed actions (line 14). After updating the policy for a context in C_{update} , the actions for states associated with that context are fixed, and the action space is updated (lines 16–17).

Once all contexts in C_{update} are updated, Algorithm 1 is invoked to check for conflicts (line 18). If the conflict is resolved, the new conflict-free policy is returned (line 19). Otherwise, the process is repeated with an expanded C_{update} that includes the next higher-priority context, using the loop variable c_i (line 7). In the worst case, during the last iteration (when $c_i = c_1$), all contexts are updated starting with highest-priority context, sequentially followed by all lower-priority contexts, ensuring alignment with Ω . Algorithm 2 is guaranteed to find a conflict-free policy, when one exists.

4.4 Theoretical Analysis

In this section, we analyze the correctness and time complexities of Algorithms 1 and 2.

PROPOSITION 4.1. *Algorithm 1 correctly identifies conflicts.*

PROOF. We first describe V_r convergence and then show that states from which the goal cannot be reached have $V_r(s) = 0$ under reward R_e defined in Sec. 4.2. Since Algorithm 1 uses Bellman backup (lines 12–14), V_r is guaranteed to converge [28].

Let S_u denote the set of all states from which s_{goal} is unreachable, and let S_r represent the set of all states that can reach the goal, such that $S = S_u \cup S_r$. In the worst case, the unsolved set may be the entire state space, $S_{unsolved} = S$. V_r is first initialized to zero for all states (line 3): $V_r^{(k=0)}(s) = 0, \forall s \in S$. We now show using induction that $V_r(s_u) = 0, \forall s_u \in S_u$ always. For $k = 0$, $V_r^{(0)}(s_u) = 0$ (by

initialization). Using Bellman equation, the value of a state s in the k^{th} iteration is updated as:

$$V_r^{(k)}(s) = R_e(s) + \gamma \max_a \sum_{s' \in S} T(s, a, s') V_r^{(k-1)}(s') \quad (1)$$

where $T(s, a, s')$ is the transition probability from s to s' using action a , and γ is the discount factor. Note that, while all actions in a state s are considered here for generality, Algorithm 1 only requires $\pi_G(s)$ for conflict detection. By Definition 2, $T(s_u, a, s_r) = 0$ and so the values of any $s_u \in S_u$ never get updated. Hence for k^{th} iteration: $V_r^{(k)}(s_u) = 0$. For iteration $k + 1$, using Equation 1,

$$V_r^{(k+1)}(s_u) = R_e(s_u) + \gamma \max_a \sum_{s' \in S} T(s_u, a, s') V_r^{(k)}(s'), \forall s_u \in S_u.$$

For $V_r^{(k+1)}(s_u)$ to be non-zero, at least one of its successors must have a path to goal, $s' \in S_r$. Substituting $R_e(s_u) = 0$ and $T(s_u, a, s_r) = 0$ in the above equation, we get $V_r^{(k+1)}(s_u) = 0$. Thus, Algorithm 1 identifies a conflict, when there exists one. \square

PROPOSITION 4.2. *The time complexity of Algorithm 1 is $O(|S|^2)$.*

PROOF. The complexity of Algorithm 1 depends on the time taken for calculating V_r and the time taken to check for conflicts based on the computed V_r . Computing V_r (lines 8–24) is essentially policy evaluation with π_G under reward R_e . Hence the complexity of this step is $O(|S|^2)$. The worst case complexity of conflict detection step (lines 25–27) is $O(|S|)$ since it cycles through the entire state space. Thus, the worst case time complexity of Algorithm 1 is $O(|S|^2 + |S|) \implies O(|S|^2)$. \square

PROPOSITION 4.3. *Algorithm 2 guarantees a conflict-free policy.*

PROOF. We prove this using the property of lexicographic ordering, where the optimality of a policy in a higher-priority context does not depend on those of lower priority. Consider a meta-ordering of contexts $\Omega \equiv c_1 \succ \dots \succ c_m$. Algorithm 2 identifies the lowest-priority context involved in the conflict, c^* (lines 4–6). The algorithm then marks all contexts from the identified c^* (in line 6) to least priority context c_m for updates (lines 7–10). Higher-priority contexts are kept fixed (lines 11–12) while the policies for the identified contexts are updated using lexicographic value iteration (line 14). After the context's policy is updated, the actions are fixed (lines 15–17) before moving on to lower priority contexts following Ω .

We now show that the final iteration, where all context policies are updated in sequence according to Ω , yields a conflict-free policy. Specifically, considering policy update in line 14:

$$\pi_{c_i}(s) = \arg \max_{a \in A_{c_i}^*} \left[R(s) + \gamma \sum_{s' \in S} T(s, a, s') V_{c_i}(s') \right], \quad \forall 2 \leq i \leq m$$

Here, $A_{c_i}^*$ represents the action space, with actions fixed for states in the higher-priority contexts c_1 through c_{i-1} . Since these updates do not affect the higher-priority contexts, no new conflicts can be introduced in c_1, \dots, c_{i-1} . Thus, the sequential updates across all contexts, respecting the lexicographic ordering Ω , ensure that no conflicts arise in higher-priority contexts. The resulting policy is also verified for conflicts using Algorithm 1 (line 18), ensuring that Algorithm 2 terminates with a conflict-free policy. \square

PROPOSITION 4.4. *The time complexity of Alg. 2 is $O(|C|^2|S|^2|A|)$.*

PROOF. The time complexity of Algorithm 1 can be broken down into three primary components: identifying the minimum priority context involved in a conflict (lines 4–6), updating the policy based on the lexicographic ordering over conflicts (lines 8–17), and verifying that the updated policy is conflict-free (line 19). The time complexity of identifying the minimum priority context (lines 4–6) is $O(|S|)$ time in the worst case. The time taken by the main outer loop (lines 7–19) is $O(|C|)$ since it iterates over all contexts in C . Within this loop, compiling the list of contexts to be updated (lines 9–10) requires $O(|C|)$ time as it cycles through all contexts. Fixing actions for the remaining contexts (lines 11–12) takes $O(|S|)$ time in the worst case if actions need to be fixed for all states. Updating policies for contexts in C_{update} (lines 13–17) involves iterating through all contexts in C_{update} , which takes $O(|C|)$ time. For each context, the policy computation using lexicographic value iteration takes $O(|S|^2|A|)$, and the subsequent policy update and fixing loop requires $O(|S|)$ time. Therefore, this step takes $O(|C|(|S|^2|A| + |S|))$. Finally, conflict checking (line 18) has a time complexity of $O(|S|^2)$ (from Proposition 4.2). Thus, the overall time complexity of Algorithm 2 is $O(|C|^2|S|^2|A|)$. \square

5 LEARNING STATE-CONTEXT MAPPING

The mapping between the states and contexts, \mathcal{Z} is critical for effective planning but this information may sometimes be unavailable to the agent a priori. We address this challenge by using a Bayesian approach to infer the likely context of a state, using a limited number of expert demonstrations. Note that we assume access to all other parameters of CLMDP, except \mathcal{Z} .

Consider a set of expert trajectories $\tau = \{\tau_1, \dots, \tau_N\}$, where each τ_i is a trajectory that originates at a random start state and terminates at the goal state s_g , $\tau_i(s) = \{(s, a_0), \dots, (s_g, a_n)\}$. Given τ , the posterior probability of a context c is computed as,

$$Pr(c|s, \vec{r}, \tau) \propto \sum_{a \in A} Pr(\vec{r}|s, a, c) \cdot Pr(a|s, c, \tau) \cdot Pr(c|\tau) \quad (2)$$

where $Pr(\vec{r}|s, a, c)$ is the probability of observing the reward vector \vec{r} when executing action a in state s under context c , $Pr(a|s, c, \tau)$ is the probability of the expert executing action a in state s if c is the underlying context associated with s , and $Pr(c|\tau)$ denotes the prior probability, given τ . We marginalize over actions to ensure that the computed posterior reflects the overall probability of context c across all potential actions that the expert could have taken, given the state and reward (which varies based on the context). We first describe how these terms are calculated for state-action pairs in the dataset and then discuss how to estimate the posterior for other states. Algorithm 3 outlines the steps involved in inferring \mathcal{Z} .

Likelihood estimation: The probability of observing a particular reward vector \vec{r} in the data is either one or zero, based on f_R in the CLMDP (line 10). That is, $Pr(\vec{r}|s, a, c) = 1$ when \vec{r} matches the reward vector under context c ($f_R(c)$), and $Pr(\vec{r}|s, a, c) = 0$ otherwise.

Action probability: For a state s in the expert data, the probability of taking an action a under context c is either one or zero depending on whether the expert followed that action during demonstration. To estimate this probability, we calculate a set of *possible contexts*, $\mathcal{P}(s_k, a_k)$, that the expert might have followed for each state-action pair in the data, as follows. First, we compute a policy π_c for each

Algorithm 3 Infer State-Context Mapping, \mathcal{Z}

```

1: Input  $C, o, f_w, \Omega, S, A, T, R, s_g$ , Set of expert trajectories  $\tau$ 
2:  $\mathcal{P}(s, a) \leftarrow [], \forall (s, a) \in \tau$ 
3: for  $c \in C$  do
4:    $\pi_c \leftarrow LVI(S, A, T, R, f_w(c))$   $\triangleright$  compute policy for  $c$ 
5:    $\mathcal{P}(s, \pi_c(s)).\text{APPEND}(c), \forall (s, \pi_c(s)) \in \tau$   $\triangleright$  possible contexts
6: for  $s \in S$  do
7:    $\mathcal{Z}(s) \leftarrow \{\}$ 
8:   for  $c \in C$  do
9:     for  $a \in A$  do
10:      Calculate  $Pr(\vec{r}|s, a, c)$  based on  $f_R(c)$ 
11:      if  $s \in \tau$  then  $\triangleright$  state is in expert data
12:         $a_E \leftarrow$  expert action in  $s$ 
13:        Calculate  $Pr(a = a_E|s, c \in \mathcal{P}(s, a))$ 
14:         $Pr(c) \leftarrow Pr(c|\mathcal{P}(s, a))$ 
15:      if  $s \notin \tau$  then  $\triangleright$  state is not in expert data
16:        Calculate  $Pr(a = \pi_c(s)|s, c)$ 
17:         $Pr(c) \leftarrow$  uniform distribution over  $C$ 
18:      Compute posterior  $Pr(c|s, \vec{r})$  using Eqn. 2
19:       $\hat{c} \leftarrow (c \in C : Pr(c|s, r) = \max_{c' \in C} (Pr(c'|s, r)))$ 
20:       $\mathcal{Z}(s) \leftarrow \text{argmax}_{\hat{c}} \Omega$ 
21: return  $\mathcal{Z}$ 

```

$c \in C$, assuming c is the only context and *all* states are mapped to it (line 4). The objective ordering and the reward functions associated with c are determined using $f_w(c)$ and $f_R(c)$ respectively. Note that this problem has a single context (and hence a single ordering over primitive objectives) and is solved using lexicographic value iteration (LVI) [34]. Second, the expert's action for a state s in the data is compared with $\pi_c(s)$. A context c is considered as a potential context if $\pi_c(s)$ matches expert action at state s (line 5). The action probability $Pr(a|s, c, \tau) = 1$, if the expert followed action a and $c \in \mathcal{P}(s, a)$. Otherwise, $Pr(a|s, c, \tau) = 0$ (line 13).

Calculating informed prior: The prior probability $Pr(c|\tau)$ is often a uniform distribution. However, when expert data is available, an *informed prior* can be calculated (line 14). The prior $Pr(c|\tau)$ for a state-action pair (s, a) in the dataset is calculated as the fraction of occurrences of c in the possible contexts set $\mathcal{P}(s, a)$.

Posterior estimation for states not in dataset: For states not in τ , the posterior is estimated using $Pr(c|s, \vec{r}) \propto \sum_{a \in A} Pr(\vec{r}|s, a, c) \cdot Pr(a|s, c) \cdot Pr(c)$, where each term is calculated based on the policy computed for each context, instead of using expert data. The action probability is determined by the policy π_c that is computed by considering all states mapped to context c . $Pr(a|s, c) = 1$ if $\pi_c(s) = a$, and $Pr(a|s, c) = 0$ otherwise (line 16). The prior $Pr(c)$ is a uniform distribution over all contexts (line 17). $Pr(\vec{r}|s, a, c)$ is determined based on f_R , similar to estimating it under expert data (line 10).

Finally, the context with the highest posterior probability is mapped to the state s . If multiple contexts are equally probable, then the context with a higher priority in the meta-ordering Ω is assigned to the state (lines 19–20), ensuring that the most critical objective in that state is prioritized. A natural question then is why we need expert data if \mathcal{Z} can be inferred without it. When

\mathcal{Z} is inferred without using expert data, it does not provide any information as to which context should be followed in the states. Expert data offers insights into potential contexts for a subset of states, enabling the agent to compute a conflict-reduced policy.

6 EXPERIMENTS

We evaluate our approach against six baselines using three domains in simulation and validate our results through hardware experiments. Unless specified otherwise, all algorithms were implemented by us in Python and the simulations were run on an iOS machine with 18 GB RAM.¹

Baselines. We evaluate the performance of two variants of our approach, *contextual planning with conflict resolver*: one with a provided state-to-context mapping \mathcal{Z} (**O1**), and the other where \mathcal{Z} is learned by Bayesian approach (**O2**) using Algorithm 3. The performances are compared with six baselines:

- **B1: Task only**—a single objective planning that focuses solely on reaching the goal state and ignores the context;
- **B2: LMDP using Ω** —applies the lexicographic MDP formulation [34] assuming the entire state space falls under the highest priority context in Ω ;
- **B3: Scalarization using Ω** —policy computation using scalarization, with all states mapped to the highest priority context in Ω ;
- **B4: LMDP for contexts**—modifies LMDP [34] to plan for multiple contexts in descending order of priority following Ω ;
- **B5: Adaptive scalarization from Yang et al. (2019)** [36]—uses a deep neural network (DNN) to learn a policy based on state and scalarization weights as input, producing actions as output;
- **B6: Contextual planning w/o resolver**—our approach with \mathcal{Z} given but without using the conflict resolver;

B2 and B3 are multi-objective planning but with a single ordering over objectives. The following domains are used for evaluation.

Sample Collection using Salp. A salp-inspired underwater robot [27] optimizes the collection and deposition of chemical sample (o_1) while minimizing coral damage (o_2) and optimizing battery usage by avoiding eddy currents (o_3). The robot operates under three contexts: $\{c_1: \text{task completion}, c_2: \text{coral}, c_3: \text{eddy}\}$. Context c_1 prioritizes depositing the sample over minimizing coral damage and battery usage: $o_1 \succ o_2 \succ o_3$. Context c_2 prioritizes minimizing coral damage, followed by depositing the sample and battery usage: $o_2 \succ o_1 \succ o_3$. Context c_3 prioritizes battery usage, then depositing the sample, and lastly minimizing coral damage: $o_3 \succ o_1 \succ o_2$.

States where the agent carries the samples and is around corals are mapped to the coral context (c_2). Locations with eddy currents are assigned to context c_3 . All remaining states are mapped to c_1 . The meta-ordering of these contexts is $\Omega \triangleq c_2 \succ c_1 \succ c_3$.

Semi-Autonomous Taxi. We modify the semi-autonomous driving domain from [34] to consider three objectives and multiple contexts. The agent operates under three objectives: quickly dropping off passengers (o_1), maximizing travel on autonomy-enabled roads (o_2), and minimizing passenger discomfort by avoiding potholes (o_3). The contexts in this environment are: $\{c_1: \text{urban transit}, c_2: \text{self-driving}, c_3: \text{rough terrain}\}$. Context c_1 prioritizes passenger drop-off

¹Code: <https://tinyurl.com/Contextual-LMDP>.

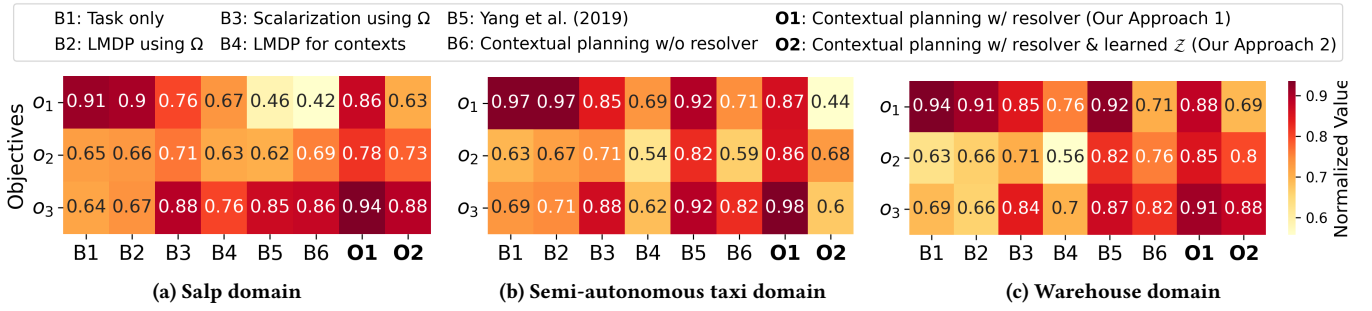


Figure 3: Performance in all objectives (o_1 , o_2 , o_3) normalized against maximum reward, averaged over 100 trials across 5 instances of each domain, represented as a heatmap with darker shade denoting higher and lighter denoting lower reward.

time over comfort and traveling on autonomy-enabled roads, $o_1 \succ o_3 \succ o_2$. Context c_2 prioritizes travel on autonomy-enabled roads followed by drop-off and comfort: $o_2 \succ o_1 \succ o_3$. Context c_3 prioritizes minimizing discomfort followed by drop-off and autonomy-enabled travel: $o_3 \succ o_1 \succ o_2$. States where the taxi is on autonomy-enabled roads are mapped to context c_2 , while those going over potholes when a passenger is onboard, are mapped to c_3 . All other states are assigned to the urban transit context (c_1) that prioritizes passenger drop-off. The context ordering for this domain is $\Omega \triangleq c_2 \succ c_1 \succ c_3$.

Package Delivery in Warehouse. The robotic agent in a warehouse [13] operates under three objectives: quick package delivery (o_1), minimizing damage from slippery tiles (o_2), and reducing inconvenience to human workers by avoiding narrow corridors (o_3). The contexts in this environment are: $\{c_1$: normal operation, c_2 : caution zone, c_3 : worker zone}. Context c_1 prioritizes quick package delivery over minimizing slip damage and reducing inconvenience to human workers: $o_1 \succ o_2 \succ o_3$. Context c_2 prioritizes minimizing slip damage, followed by quick package delivery and minimizing inconvenience: $o_2 \succ o_1 \succ o_3$. Context c_3 prioritizes reducing inconvenience to human workers, followed by optimizing package delivery and minimizing slip damage: $o_3 \succ o_1 \succ o_2$. States located areas with slippery tiles are mapped to the caution zone (c_2), while those involving navigating through narrow corridors with human workers are mapped to c_3 . All remaining states are designated to c_1 . Ordering over contexts for this domain is $\Omega \triangleq c_2 \succ c_1 \succ c_3$.

Expert trajectories for inference. We simulate an expert with complete knowledge of the CLMDP. For each problem instance, 10 trajectories are sampled, with random start states.

7 RESULTS

We evaluate the effectiveness of various approaches for contextual multi-objective decision making in terms of: (1) how well a technique can balance the trade-off between different objective values, measured by minimum value across objectives, and (2) validity of the resulting policy, measured in terms of number of conflicts detected and resolved. Additionally, we validate our approach using a mobile robot in an indoor setup of warehouse domain (Figure 5).

Balancing trade-offs between objectives. A heat map showing the performance of each technique, across various objectives and domains, is presented in Figure 3. Darker shades indicate higher

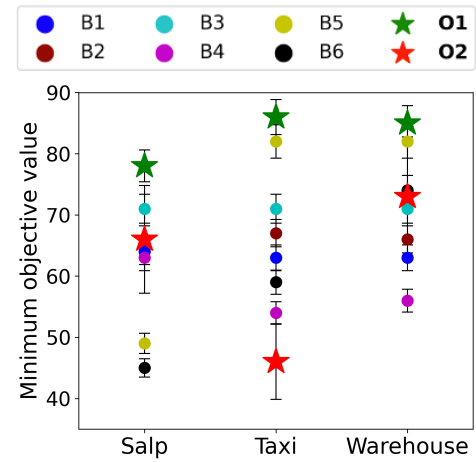


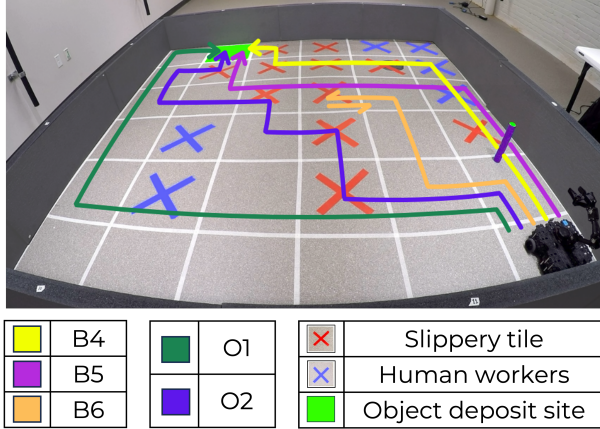
Figure 4: Minimum objective value from each technique averaged over 100 trials in five instances of each domain.

(better) objective values. The approach from Yang et. al, (2019) [36] occasionally surpasses our approach in individual objectives, but it does not perform consistently well across all domains. Our approach performs consistently well, across objectives and domains, when the state-context mapping \mathcal{Z} is known (O1).

While some baseline techniques perform well in certain objectives, even outperforming our approach, they tend to score lower on others. In a multi-objective setting, the performance of a technique must be evaluated based on how well it can perform *across all objectives*. We evaluate this performance trade-off across objectives via a percentile comparison showing the minimum objective value. We compare the lowest-performing objective value, across all objectives, normalized against the maximum achievable reward in that objective. This value is averaged over 100 trials, across five instances in each domain. Figure 4 shows the percentile comparison showing the minimum objective value for each technique across all domains. Note that our approach consistently achieves the highest minimum objective value. Additionally, we also evaluate our second approach, contextual planning with resolver with a learned \mathcal{Z} (O2), where the state-context mapping is inferred from expert trajectories (as described in Section 5). While this approach shows

Table 1: Percentage conflicts and task completion for multi-context techniques, averaged over 100 trials in five instances.

Technique	Salp Domain		Warehouse Domain		Taxi Domain	
	%Conflicts	%Goal Reached	%Conflicts	%Goal Reached	%Conflicts	%Goal Reached
B4: LMDP for contexts	40.4 ± 12.1	39.2 ± 3.5	34.2 ± 4.5	28.3 ± 5.2	13.1 ± 3.2	88.7 ± 9.6
B5: Yang et. al, 2019 [36]	28.7 ± 4.6	62.4 ± 7.1	11.2 ± 3.3	94.1 ± 5.1	4.2 ± 3.3	98.6 ± 13.4
B6: Contextual planning w/o resolver	19.8 ± 5.3	88.4 ± 8.1	34.2 ± 7.4	83 ± 6.2	6.7 ± 1.5	95.2 ± 14.1
O1: Contextual planning w/ resolver	0 ± 0	100 ± 0	0 ± 0	100 ± 0	0 ± 0	100 ± 0
O2: Contextual planning w/ resolver & learned \mathcal{Z}	0 ± 0	97.2 ± 1.72	0 ± 0	96.4 ± 2.24	0 ± 0	62.8 ± 35.65

**Figure 5: Comparison of the paths taken by our warehouse agent during delivery task in our indoor setup.**

weaker performance, it should be noted that this is due to inaccuracies in predicting the context with limited expert trajectories, affecting both the objective ordering and the associated reward function. The performance with learned \mathcal{Z} will likely improve with the availability of additional data.

Resolving conflicts. Table 1 provides a comparative analysis of the percentage of conflicts (cycles) and task completions for techniques that plan using *context* information. Our approach (O1) successfully avoids all conflicts and achieves 100% task completion in every trial across all domains. This consistent performance highlights the effectiveness of our conflict detection and resolution mechanisms (Alg. 1 and 2). Our approach with learned \mathcal{Z} (O2) does not always reach the goal, even in the absence of conflicts in the policy. This is due to *incorrect attribution of contexts*, which leads to suboptimal policies that hinder the agent’s ability to reach the goal state. When context partitions are incorrectly assigned, Assumption 1 does not hold. By Defn. 2, a policy is said to have a conflict only if the goal reachability is zero. Some policies may have a non-zero (but less than one) probability of reaching the goal. Thus, some conflict-free policies may not always reach the goal during execution.

Evaluation using a mobile robot. We conduct a series of experiments using a TurtleBot in an indoor warehouse domain setup. The robot autonomously collects and delivers an object using a LiDAR and a map of the area for active localization so as to determine its state and execute actions based on its computed policy. Figure 5 shows the trajectories corresponding to different techniques.

Table 2: Normalized objective values from hardware experiments in the warehouse domain, averaged over five trials.

Technique	o_1	o_2	o_3
B4: LMDP for contexts	0.84	0.72	0.94
B5: Yang et. al, 2019 [37]	0.48	0.43	0.83
B6: Contextual planning w/o resolver	0.0	0.0	0.0
O1: Contextual planning w/ resolver	0.98	0.90	0.93
O2: Contextual planning w/ resolver & learned \mathcal{Z}	0.73	0.88	0.99

Table 2 shows the normalized objective values. Our contextual planning approach performs well across all objectives. Without conflict resolver, the robot does not reach the goal, emphasizing the importance of conflict resolver module. Overall, the results demonstrate the method’s applicability in real-world robotic tasks with context-specific preferences over multiple objectives.

8 CONCLUSION

This paper presents contextual lexicographic MDP, a framework for multi-objective planning in scenarios where the relative preferences between objectives and the associated reward functions are determined by the context. We also present an algorithm to solve the CLMDP, by combining independently computed policies from different contexts into a global policy that is context-aware and cycle-free. We also analyze the algorithm’s correctness and complexity. Empirical evaluations in simulation and using a mobile robot demonstrate the effectiveness of our approach in balancing the performance trade-offs across objectives, under multiple contexts, and the feasibility of applying our approach in the real world.

In the future, we aim to extend our algorithm to support non-stationary state-context mapping. This work addresses the technical challenges in planning with context-specific objective orderings, assuming this information is known or can be inferred from expert data. How to define a context for a problem and how to define a mapping between objective orderings and contexts extend beyond technical considerations and into ethical challenges. Prioritizing one objective over another are often based on subjective input from domain experts and we acknowledge that these decisions carry significant ethical implications that cannot be overlooked.

ACKNOWLEDGMENTS

This work was supported in part by ONR grant N00014-23-1-2171. We thank Akshaya Agarwal for their help with robot experiments.

REFERENCES

- [1] Majid Abdolshah, Alistair Shilton, Santu Rana, Sunil Gupta, and Svetha Venkatesh. 2019. Multi-objective Bayesian optimisation with preferences over objectives. *Advances in neural information processing systems* 32 (2019).
- [2] Mridul Agarwal, Vaneet Aggarwal, and Tian Lan. 2022. Multi-objective reinforcement learning with non-linear scalarization. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 9–17.
- [3] Ayhan Alp Aydeniz, Enrico Marchesini, Christopher Amato, and Kagan Tumer. 2024. Entropy Seeking Constrained Multiagent Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2141–2143.
- [4] Ali Bahrani, Jun Yuan, Chukwuemeka David Emele, Daniele Masato, Timothy J Norman, and David Mott. 2008. Collaborative and context-aware planning. In *MILCOM 2008-2008 IEEE Military Communications Conference*. IEEE, 1–7.
- [5] Syrine Belakaria, Aryan Deshwal, Nitthilan Kannappan Jayakodi, and Janardhan Rao Doppa. 2020. Uncertainty-aware search framework for multi-objective Bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10044–10052.
- [6] Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. 2022. Contextualize Me—The Case for Context in Reinforcement Learning. *arXiv preprint arXiv:2202.04500* (2022).
- [7] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. 2020. Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences. In *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119.
- [8] Samuel Daulton, Sait Cakmak, Maximilian Balandat, Michael A. Osborne, Enlu Zhou, and Eytan Bakshy. 2022. Robust Multi-Objective Bayesian Optimization Under Input Noise. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 4831–4866.
- [9] Kalyanmoy Deb. 2001. Nonlinear goal programming using multi-objective genetic algorithms. *Journal of the Operational Research Society* 52, 3 (2001), 291–302.
- [10] Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. 2016. Multi-objective optimization. In *Decision sciences*. CRC Press, 161–200.
- [11] Kalyanmoy Deb and Ankur Sinha. 2010. An efficient and accurate solution methodology for bilevel multi-objective programming problems using a hybrid evolutionary-local-search algorithm. *Evolutionary computation* 18, 3 (2010), 403–449.
- [12] Kikuo Fujimura. 1996. Path planning with multiple objectives. *IEEE Robotics & Automation Magazine* 3, 1 (1996), 33–38.
- [13] Jianqi Gao, Yanjie Li, Yunhong Xu, and Shaohua Lv. 2022. A two-objective ILP model of OP-MATSP for the multi-robot task assignment in an intelligent warehouse. *Applied Sciences* 12, 10 (2022), 4843.
- [14] Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: recent advances and beyond. *IEEE Access* (2023).
- [15] Jakub Hvězda, Tomáš Rybecký, Miroslav Kulich, and Libor Přeučil. 2018. Context-aware route planning for automated warehouses. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2955–2960.
- [16] Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, and Jonghyun Choi. 2023. Context-aware planning and environment-aware memory for instruction following embodied agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10936–10946.
- [17] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.
- [18] Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. 2018. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*. PMLR, 597–618.
- [19] A-I Mouaddib. 2004. Multi-objective decision-theoretic path planning. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, Vol. 3. IEEE, 2814–2819.
- [20] Anna Nickelson and Kagan Tumer. 2024. Redefining the Behavior Space for Multi-Objective MAP-Elites. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 355–358.
- [21] Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning. In *IJCAI*.
- [22] Edgar Reehuis and Thomas Bäck. 2010. Mixed-integer evolution strategy using multiobjective selection applied to warehouse design optimization. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*. 1187–1194.
- [23] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.
- [24] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2021. A multi-objective approach to mitigate negative side effects. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 354–361.
- [25] Joar Skalse, Lewis Hammond, Charlie Griffin, and Alessandro Abate. 2022. Lexicographic multi-objective reinforcement learning. In *Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*.
- [26] Earl B Smith and Reza Langari. 2003. Fuzzy multiobjective decision making for navigation of mobile robots in dynamic, unstructured environments. *Journal of Intelligent & Fuzzy Systems* 14, 2 (2003), 95–108.
- [27] Kelly R. Sutherland and Laurence P. Madin. 2010. Comparative jet wake structure and swimming performance of salps. *Journal of Experimental Biology* 213, 17 (2010), 2967–2975.
- [28] Richard S Sutton, Andrew G Barto, et al. 1998. Introduction to reinforcement learning. Vol. 135.
- [29] Adriaan W Ter Mors, Cees Witteveen, Jonne Zutt, and Fernando A Kuipers. 2010. Context-aware route planning. In *Multiagent System Technologies: 8th German Conference, MATES 2010, Leipzig, Germany, September 27-29, 2010. Proceedings 8*. Springer, 138–149.
- [30] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. 2009. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems* 22 (2009).
- [31] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. 2013. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*. IEEE, 191–199.
- [32] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *Comput. Surveys* 56, 7 (2024), 1–33.
- [33] Nils Wilde, Stephen L Smith, and Javier Alonso-Mora. 2024. Scalarizing Multi-Objective Robot Planning Problems using Weighted Maximization. *IEEE Robotics and Automation Letters* (2024).
- [34] Kyle Wray, Shlomo Zilberstein, and Abdel-Ilah Mouaddib. 2015. Multi-objective MDPs with conditional lexicographic reward preferences, Vol. 29.
- [35] Kyle Hollins Wray, Dirk Ruiken, Roderic A Grupen, and Shlomo Zilberstein. 2016. Log-space harmonic function path planning. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1511–1516.
- [36] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems* 32 (2019).
- [37] Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. 2020. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems* 33 (2020), 19238–19250.