Opportunistic Qualitative Planning in Stochastic Systems with Incomplete Preferences over Reachability Objects

Abhishek N. Kulkarni*, and Jie Fu

Abstract—Preferences play a key role in determining what goals/constraints to satisfy when not all constraints can be satisfied simultaneously. In this paper, we study how to synthesize preference-satisfying plans in a stochastic system modeled as an MDP, given a (possibly incomplete) combinative preference model over temporally extended goals. We start by introducing new semantics to interpret preferences over infinite plays of the stochastic system. Then, we introduce a new notion of ‘improvement’ to enable comparison between two prefixes of an infinite play. Based on this, we define two solution concepts called Safe and Positively Improving (SPI) and Safe and Almost-Sure Improving (SASI) that enforce improvements with a positive probability and with probability one, respectively. We construct a model called an improvement MDP, in which the synthesis of SPI and SASI strategies that guarantee at least one improvement, reduces to computing positive and almost-sure winning strategies in an MDP. We present an algorithm to synthesize the SPI and SASI strategies that induce multiple sequential improvements. We demonstrate the proposed approach using a robot motion planning problem.

I. INTRODUCTION

With the rise of artificial intelligence, robotics, and autonomous systems are being designed to make complex decisions while reasoning about multiple goals at the same time. Preference-based planning (PBP) allows the systems to decide which goals to satisfy when not all of them can be achieved [1]. Even though PBP has been studied since the early 1950s, most works on preference-based temporal planning (c.f. [2]) fall into at least one of the following categories: (a) those which assume that all outcomes are pairwise comparable—that is, the preference relation is complete [3], [4], (b) those which study exclusionary preferences—that is, the set of outcomes is mutually exclusive (see [5] and the references within), (c) those which are interpreted over finite traces [6]. In this work, we study the PBP problem for the class of systems in which the preference model is incomplete, combinative (as opposed to exclusionary) and is interpreted over infinite plays of the stochastic system.

Incomplete, combinative preferences arise naturally in robotics [7], [8], economics [9], AI [10] etc. In practice, agents may need to make decisions with incomplete preferences because of (a) Incomplete information; for instance, when an agent who has lost communication with the server has to make a decision under a time constraint, and (b) Incomparability; for instance, in the trolley problem [11].

A. Kulkarni (*) corresponding author) and J. Fu are with the Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, Fl 32611 USA. {a.kulkarni2,fujie}@ufl.edu

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-21-1-0085.

an autonomous car may not decide whether sacrificing one person is preferable to sacrificing 5 people. On the other hand, combinative preferences allow the agent to express preferences over alternatives that may not be mutually exclusive. For example, consider a user preference for a robot that “visiting A is strictly preferred over visiting B.” The two alternatives, ‘visiting A’ and ‘visiting B,’ are not mutually exclusive because, for instance, a path that visits A may also visit B. To express combinative preferences as exclusionary, a user must consider 4 alternatives based on whether A and B are visited and rewrite the preference, often resulting in a much larger preference model [5]. As the number of alternatives increases, an exponential number of alternatives may be required to express a combinative preference as an exclusionary one. This, in addition to the fact that the behavior of several control and robotic systems is interpreted over infinite plays in a stochastic system, there is a need for a way to synthesize strategies that satisfy incomplete, combinative preferences in stochastic environments.

Preference-based planning problems over temporal goals have been well-studied for deterministic planning given both complete and incomplete preferences (see [2] for a survey). For preferences specified over temporal goals, several works [12], [13], [14] proposed minimum violation planning methods that decide which low-priority constraints should be violated in a deterministic system. Mehdipour et al. [15] associate weights with Boolean and temporal operators in signal temporal logic to specify the importance of satisfying the sub-formula and priority in the timing of satisfaction. This reduces the PBP problem to that of maximizing the weighted satisfaction in deterministic dynamical systems. However, the solutions to the PBP problem for deterministic systems cannot be applied to stochastic systems. This is because, in stochastic systems, even a deterministic strategy yields a distribution over outcomes satisfied by the resulting paths. To determine which strategy is better, we require a method to compare distributions over paths instead of comparing deterministic paths, which is what deterministic preference-based planners do.

Several works have studied the PBP problem for stochastic systems. Lahijanian and Kwiatkowska [16] considered the problem of revising a given specification to improve the probability of satisfaction of the specification. They formulated the problem as a multi-objective Markov Decision Process (MDP) problem that trades off minimizing the cost of revision and maximizing the probability of satisfying the revised formula. Li et al. [17] solve a preference-based probabilistic planning problem by reducing it to a multi-

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
objective model checking problem. However, all these works assume the preference relation to be complete. To the best of our knowledge, [8] is the only work that studies the problem of probabilistic planning with incomplete preferences. The authors introduce the notion of the value of preference satisfaction for planning within a pre-defined finite time duration and developed a mixed-integer linear program to maximize the satisfaction value for a subset of preference relations.

The work [8] studied preference-based quantitative planning. In comparison, this work focuses on qualitative planning of MDPs with preferences over a set of outcomes represented as reachability objectives. Qualitative analysis of an MDP with reachability objective involves computing the set of almost-sure and positive winning states from which the agent can achieve the objective with probability one or positive probability, respectively [18]. Qualitative analysis is widely used in control theory, and robotics applications [19]. However, the almost-sure and positive winning criteria cannot be used to synthesize preference satisfying strategies because they only reason about a single objective.

Contributions. We first introduce new semantics to interpret an incomplete, combative preference model over temporally extended objectives in a stochastic system. We observe that uncertainties in the planning environment combined with infinite plays might give rise to opportunities to improve the outcomes achieved by the agent. Thus, analogous to the idea of an improving flip [20], we define the notion of improvement that compares two prefixes of an infinite play to determine which one is more preferred, based on their different prospects regarding the set of possible, achievable objectives. Based on whether a strategy exists to enforce an improvement with a positive probability or with probability one, we introduce two solution concepts called safe and positively improving (SPI), and safe and almost-surely improving (SASI) which ensure an improvement can be made with positive probability and with probability one, respectively. The synthesis of Safe and Positive Improving (SPI) and Safe and Almost-Sure Improving (SASI) strategies is through a construction called improvement MDP and a reduction to that of computing positive and almost-sure winning strategies for some reachability objectives of the improvement MDP. In the case of almost-surely improvement, we also provide an algorithm to determine the maximum number of improvements achievable given any given state. The correctness of the proposed algorithms is demonstrated through a robot motion planning example.

II. PRELIMINARIES

Notation. Given a finite set $X$, the powerset of $X$ is denoted as $\wp(X)$. The set of all finite (resp., infinite) ordered sequences of elements from $X$ is denoted by $X^*$ (resp., $X^\omega$). The set of all finite ordered sequences of length $> 0$ is denoted by $X^+$. We write $D(X)$ to denote the set of probability distributions over $X$. The support of a distribution $D \in D(X)$ is denoted by $\text{Supp}(D) = \{x \in X \mid D(x) > 0\}$.

In this paper, we consider a class of decision-making problems in stochastic systems modeled as a MDP without the reward function [21]. We then introduce a preference model over the set of infinite plays in the MDP.

Definition 1 (MDP). An MDP is a tuple $M = (S, A, T, \iota)$, where $S$ and $A$ are finite state and action sets, $\iota \in S$ is an initial state (or an initial distribution over states), and $T : S \times A \rightarrow D(S)$ is the transition probability function such that $T(s, a, s')$ is the probability of reaching the state $s' \in S$ when action $a \in A$ is chosen at the state $s \in S$.

A play in an MDP $M$ is an infinite sequence of states $\rho = s_0 s_1 \ldots \in S^\omega$ such that, $\iota = s_0$ (or $\iota(s_0) > 0$) and for every integer $i \geq 0$, there exists an action $a \in A$ such that $T(s_i, a, s_{i+1}) > 0$. We denote the set of all plays starting from $s \in S$ in the MDP by $\text{Plays}(M, s)$, and the set of all plays in $M$ is denoted by $\text{Plays}(M) = \bigcup_{s \in S} \text{Plays}(M, s)$. The set of states occurring in a play is given by $\text{Occ}(\rho) = \{s \in S \mid \exists i \geq 0, s_i = s\}$. A prefix of a play $\rho$ is a finite sub-sequence of states $\nu = s_0 s_1 \ldots s_k \geq 0$, whose length is $|\nu| = k + 1$. The set of all prefixes of a play $\rho$ is denoted by $\text{Pref}(\rho)$. The set of all prefixes in $M$ is denoted by $\text{PrefPlays}(M) = \bigcup_{\rho \in \text{Plays}(M)} \text{Pref}(\rho)$. Given a prefix $\nu = s_0 s_1 \ldots s_k \in \text{PrefPlays}(M)$, the sequence of states $s_{k+1} s_{k+2} \ldots \in S^\omega$ is called a suffix of $\nu$ if the play $\nu' = s_0 s_1 \ldots s_k s_{k+1} s_{k+2} \ldots$ is an element of $\text{Plays}(M)$.

In this MDP, we consider reachability objectives for the agent. Given a set $F \subseteq S$, a reachability objective is characterized by the set $\text{Reach}(F) = \{s \in \text{Plays}(M) \mid \text{Occ}(\rho) \cap F \neq \emptyset\}$, which contains all the plays in $M$ starting at the state $s \in S$ that visit $F$. Any play $\rho \in \text{Plays}(M)$ that satisfies a reachability objective Reach$(F)$ has a good prefix $\nu \in \text{Pref}(\rho)$ such that the last state of $\nu$ is in $F$ [22].

A finite-memory (resp., memoryless), non-deterministic strategy in the MDP is a mapping $\pi : S^+ \rightarrow \wp(A)$ (resp., $\pi : S \rightarrow \wp(A)$) from a prefix to a subset of actions that can be taken from that prefix. The set of all finite-memory, non-deterministic strategies is denoted $\Pi$. Given a prefix $\nu = s_0 \ldots s_k \in \text{PrefPlays}(M)$, a suffix $\rho = s_{k+1} s_{k+2} \ldots \in S^\omega$ is consistent with $\pi$, if for all $i \geq 0$, there exists an action $a \in \pi(s_0 \ldots s_k s_{k+1})$ such that $T(s_i, a, s_{i+1}) > 0$. Given an MDP $M$, a prefix $\nu \in \text{PrefPlays}(M)$ and a strategy $\pi$, the cone is defined as the set of consistent suffixes of $\nu$ w.r.t. $\pi$, that is

$$\text{Cone}(M, \nu, \pi) = \{\rho \in S^\omega \mid \nu\rho \text{ is consistent with } \pi\}.$$

Given a prefix $\nu \in \text{PrefPlays}(M)$ and a reachability objective Reach$(F)$, a (finite-memory/memoryless) strategy $\pi_{\text{Reach}(F)}$ is said to be positive winning if $\text{Cone}(M, \nu, \pi) \cap \text{Reach}(F) \neq \emptyset$. Similarly, a (finite-memory/memoryless) strategy $\pi_{\text{ASWin}(F)}$ is said to be almost-sure winning if $\text{Cone}(M, \nu, \pi) \subseteq \text{Reach}(F)$.

The set of states in the MDP $M$, starting from which the agent has an almost-sure (resp. positive) winning strategy to satisfy a reachability objective $F \in \mathbb{F}$ is called the almost-sure (resp., positive) winning region and is denoted by $\text{ASWin}(F)$ (resp., $\text{PWin}(F)$). The almost-sure winning

3542

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
region and strategies can be synthesized in polynomial time and linear time, respectively [21].

III. PREFERENCE MODEL

In this section, we introduce a new combinative preference model over an infinite set of infinite plays.

Definition 2. A preference model is a tuple $(U, \succeq)$, where $U$ is a countable set of outcomes and $\succeq$ is a reflexive and transitive binary relation on $U$.

Given $u_1, u_2 \in U$, we write $u_1 \succeq u_2$ if $u_1$ is weakly preferred to (i.e., is at least as good as) $u_2$; and $u_1 \sim u_2$ if $u_1 \succeq u_2$ and $u_2 \succeq u_1$, that is, $u_1$ and $u_2$ are indifferent. We write $u_1 \succ u_2$ to mean that $u_1$ is strictly preferred to $u_2$, i.e., $u_1 \succeq u_2$ and $u_2 \not\succeq u_1$. We write $u_1 \nmid u_2$ if $u_1$ and $u_2$ are incomparable. Since Def. 3 allows outcomes to be incomparable, it models incomplete preferences [23].

We consider planning objectives specified as preferences over reachability objectives.

Definition 3. A preference model over reachability objectives in an MDP $M$ is a tuple $(\mathcal{F}, \succeq)$, where $\mathcal{F} = \{\text{Reach}(F_1), \text{Reach}(F_2), \ldots, \text{Reach}(F_n)\}$ is a set of reachability objectives such that $F_1, \ldots, F_n$ are subsets of $S$.

Intuitively, a preference $\text{Reach}(F_1) \succeq \text{Reach}(F_2)$ means that any play in $\text{Reach}(F_1)$ is weakly preferred to any play in $\text{Reach}(F_2)$. The strict preference ($\succ$), indiscernibility, and incomparability are understood similarly.

The model $(\mathcal{F}, \succeq)$ is a combinative preference model, as opposed to an exclusionary one. This is because we do not assert the exclusivity condition $\text{Reach}(F_1) \cap \text{Reach}(F_2) = \emptyset$. This allows us to represent a preference such as “Visiting A and B is preferred to visiting A,” where the less preferred outcome must be satisfied first in order to satisfy the more preferred outcome. In literature, it is common to study exclusionary preference models (see [2], [4] and the references within) because of their simplicity [5]. However, we focus on planning with combinative preferences since they are more expressive than the exclusionary ones [24]. In fact, every exclusionary preference model can be transformed into a combinative one, but the opposite is not true.

When a combinative preference model is interpreted over infinite plays, the agent needs a way to compare the sets of reachability objectives satisfied by two plays. For instance, consider the preference that “Visiting A and B is preferred to visiting A.” Let $\rho_1, \rho_2$ be two plays. Suppose that $\rho_1$ visits both A and B, and $\rho_2$ visits A only. Therefore, $\rho_1$ satisfies two outcomes $\text{Reach}(F_B)$ and $\text{Reach}(F_A)$, whereas $\rho_2$ satisfies only $\text{Reach}(F_A)$. To determine the preference between the two plays, the agent compares the set $\{\text{Reach}(F_B), \text{Reach}(F_A)\}$ with $\{\text{Reach}(F_A)\}$ to conclude that the $\rho_1$ is preferred over $\rho_2$. However, suppose the given preference is that “visiting A is preferred over visiting B,” then the two plays would be indiscernible since both satisfy the more preferred objective of visiting A. In this case, the less preferred objective of visiting B does not influence the comparison of the sets. To formalize this notion, we define the notion of most-preferred outcomes.

Given a non-empty subset $\mathcal{X} \subseteq \mathcal{F}$, let $\text{MP}(\mathcal{X}) \triangleq \{R \in \mathcal{X} \mid \nexists R' \in \mathcal{X} : R' \succ R\}$ denote the set of most-preferred outcomes in $\mathcal{X}$.

Definition 4. Given a preference model $(\mathcal{F}, \succeq)$ and a play $\rho \in \text{Plays}(M)$, the set of most-preferred outcomes satisfied by $\rho$ is given by $\text{MP}(\rho) \triangleq \text{MP}(\{\text{Reach}(F) \in \mathcal{F} \mid \forall \nu \in \text{Pref}(\rho) : \nu \text{ is a good prefix for } \text{Reach}(F)\})$.

By definition, there is no outcome included in $\text{MP}(\rho)$ that is preferred to any other outcome in $\text{MP}(\rho)$. Thus, we have the following result.

Lemma 1. For any play $\rho \in \text{Plays}(M)$, every pair of outcomes in $\text{MP}(\rho)$ is incomparable to each other.

Now, we formally define the interpretation of $(\mathcal{F}, \succeq)$ in terms of the preference relation it induces on $\text{Plays}(M)$.

Definition 5. Let $(\text{Plays}(M), \succeq)$ be the preference model induced by $(\mathcal{F}, \succeq)$. Then, for any $\rho_1, \rho_2 \in \text{Plays}(M)$, we have

- $\rho_1 \succ \rho_2$ if and only if there exist a pair of outcomes $R \in \text{MP}(\rho_1)$ and $R' \in \text{MP}(\rho_2)$ such that $R \succ R'$, and there does not exist a pair of outcomes $R \in \text{MP}(\rho_1)$ and $R' \in \text{MP}(\rho_2)$ such that $R' \succ R$.
- $\rho_1 \sim \rho_2$ if and only if $\text{MP}(\rho_1) = \text{MP}(\rho_2)$.
- $\rho_1 \not\succeq \rho_2$, otherwise.

IV. SOLUTION CONCEPT

In preference-based planning, the agent is to choose its next action given a finite prefix $\nu \in \text{Pref}(\text{Plays}(M))$ in order to satisfy the given preference relation on a set of outcomes. A naive approach to this problem is to follow the strategy to satisfy a most-preferred outcome from the set of almost-surely achievable outcomes given $\nu$. However, this is not sufficient, as illustrated by the following example.

Example 1. Consider the toy MDP shown in Fig. 1. The exact probabilities are omitted because we analyze the MDP qualitatively. The transitions are understood as follows: Given action $a$ at state $s_0$, it is possible to reach both $s_3$ and $s_1$ with positive probabilities.

Let $F_1 = \{s_1, s_3\}$, $F_2 = \{s_2, s_4\}$ and $F_3 = \{s_3\}$ be three sets of final states. Let $(\mathcal{F}, \succeq)$ be the preference model such that $\mathcal{F} = \{\text{Reach}(F_1), \text{Reach}(F_2), \text{Reach}(F_3)\}$ and $\text{Reach}(F_2) \succ \text{Reach}(F_1) \succ \text{Reach}(F_3)$. Clearly, $\text{Reach}(F_2)$ and $\text{Reach}(F_3)$ are incomparable. Therefore, the play $\rho_1 = s_0s_3s_4$, which satisfies $\text{Reach}(F_3)$, is strictly preferred to the play $\rho_2 = s_0s_3s_4s_1$, which satisfies $\text{Reach}(F_1)$. Whereas, $\rho_1$ is incomparable to the play $\rho_3 = s_0s_3s_4$ because it satisfies $\text{Reach}(F_2)$.

Consider the state $s_0$ at which the agent is to choose its next action. From $s_0$, the agent can visit $F_1$ almost surely by choosing $a$. It, however, does not have an almost sure winning strategy to visit either $F_2$ or $F_3$, individually. But, by choosing $b$ at $s_0$, the agent almost surely visits either $F_2$ or $F_3$ and achieves a strictly better outcome than $F_1$. 3543
The example highlights that the almost sure winning solution concept is not suitable for preference-based planning because it reasons about exactly one outcome at a time. As a result, the agent cannot reason about opportunities to achieve a better outcome that may become available due to stochasticity in the environment.

In the sequel, we introduce two new solution concepts for probabilistic planning under incomplete preferences interpreted over infinite plays. Our solution concepts are based upon the notion of an improvement that generalizes the idea of improving flip [20] which is defined for propositional preferences. An improving flip compares two outcomes representable as propositional logic formulas to determine which is more preferred. Analogously, an improvement compares two prefixes of a play to determine which one can yield a more preferred outcome with probability one.

Definition 6. Given a play $\rho \in \text{Plays}(M)$ and two of its prefixes $\nu, \nu' \in \text{Pref}(\rho)$ such that $|\nu'| > |\nu|$, $\nu'$ is said to be an improvement of $\nu$ if there exists a pair of outcomes $R \in \text{MP}(\text{Outcomes}(\nu))$ and $R' \in \text{MP}(\text{Outcomes}(\nu'))$ such that $R' \triangleright R$. And, $\nu'$ is said to be a weakening of $\nu$ if there exists a pair of outcomes $R \in \text{MP}(\text{Outcomes}(\nu))$ and $R' \in \text{MP}(\text{Outcomes}(\nu'))$ such that $R' \triangleright R$.

Given a prefix $s_0 s_1 \ldots s_k \in \text{PrefPlays}(M)$, the transition from $s_{k-1}$ to $s_k$ is said to be an improving transition if the prefix $s_0 s_1 \ldots s_{k-1} s_k$ is an improvement over $s_0 s_1 \ldots s_{k-1}$. A play that contains an improving transition is called an improving play. It is noted that a prefix $\nu'$ can simultaneously be an improvement and a weakening of a prefix $\nu$.

Next, we define the two solution concepts that, while avoiding any weakening, induce improvements either with positive probability or with probability one.

Definition 7 (SPI/SASI Strategy). Given a prefix $\nu = s_0 s_1 \ldots s_k \in \text{PrefPlays}(M)$, a strategy $\pi : S^+ \rightarrow 2^{A}$ is said to be safe and positively (resp., safe and almost-surely) improving for $\nu$ if the following conditions hold:

1) (Safety) For all $\rho \in \text{Cone}(M, \nu, \pi)$, the play $\nu \rho$ satisfies that $s_0 s_1 \ldots s_j$ is not a weakening of $s_0 s_1 \ldots s_k$ for any integer $j > k$.
2) (Improvement) There exists (resp., for any) $\rho \in \text{Cone}(M, \nu, \pi)$, the play $\nu \rho$ satisfies the condition that there exists an integer $j > k$ such that $s_0 s_1 \ldots s_j$ is an improvement over $s_0 s_1 \ldots s_k$.

We now state our problem statement.

Problem 1. Given an MDP $M$ and a preference model $\langle F, \succeq \rangle$, design an algorithm to synthesize an SPI and a SASI strategy.

V. OPPORTUNISTIC QUALITATIVE PLANNING WITH INCOMPLETE PREFERENCES

Our approach to synthesize SPI and SASI strategies distinguishes between opportunistic states, i.e., the states from which an improvement could be made, and non-opportunistic states. We now introduce a new model called an improvement MDP to synthesize the SPI and SASI strategies.

To facilitate the definition, we slightly abuse the notation and let $\text{MP}(s) \triangleq \text{MP}(\{\text{Reach}(F) \in F \mid s \in \text{ASWin}(F)\})$ be the set of outcomes almost surely achievable from state $s$ in $M$.

Definition 8 (Improvement MDP). Given an MDP $M = \langle S, A, T, \nu \rangle$ and a preference model $\langle F, \succeq \rangle$, an improvement MDP is the tuple,

$$M = \langle V, A, \Delta, \nu_0, F \rangle,$$

where $V = S \times \{0, 1\}$ is the set of states, $A$ is the same set of actions as $M$, $\nu_0 = (\nu, 0)$ is the initial state, and $F = \{(s, 1) \mid s \in S\}$ is a set of final states that can only be reached by making an improvement. The transition function $\Delta : V \times A \rightarrow D(V)$ is defined as follows: For any states $\nu = (s, m), \nu' = (s', m') \in V$ and for any action $a \in A$, $\Delta(\nu, a, \nu') > 0$ holds if and only if the following conditions hold:

1) $T(s, a, s') > 0$.
2) (Safety) For all pairs of outcomes $R \in \text{MP}(s)$ and $R' \in \text{MP}(s')$, we have $R \not\succ R'$.
3) (Improvement) If there exists a pair $R \in \text{MP}(s)$ and $R' \in \text{MP}(s')$ such that $R' \succ R$, then $m' = 1 \text{ else } m' = 0$.

Every play $\rho = s_0 s_1 \ldots \in \text{Plays}(M)$ induces a play $\varrho = \nu_0 \nu_1 \nu_2 \ldots$ in $M$ such that for all $i = 0, 1, \ldots, \nu_i = (s_i, m_i)$ where $m_i \in \{0, 1\}$ represents a memory element such that $m_i = 1$ if and only if the transition from $s_{i-1}$ to $s_i$ is improving. The following proposition highlights important features of the improvement MDP. Before that, we note the following fact to prove Proposition 1.

Lemma 2. For every prefix $\nu = s_0 s_1 \ldots s_k \in \text{PrefPlays}(M)$, it holds that $\text{Outcomes}(\nu) = \text{Outcomes}(s_k)$ and thus $\text{MP}(\text{Outcomes}(\nu)) = \text{MP}(\text{Outcomes}(s_k))$.

The proof follows from the fact that memoryless strategies are sufficient to ensure the satisfaction of reachability objectives in MDPs [25]. In other words, if an outcome is almost surely achievable given a prefix $\nu = s_0 s_1 \ldots s_k$, then it is almost surely achievable given $s_k$.

For convenience, we write $\text{MP}(s) = \text{MP}(\text{Outcomes}(s))$ to denote the set of most preferred outcomes satisfactory/achievable with some strategy from a state $s \in S$.
Proposition 1. For any play $\rho = v_0v_1 \ldots \in \text{Plays}(\mathcal{M})$ such that $v_i = (s_i, m_i)$ for all $i = 0, 1, \ldots$, the following statements hold.

1) (Safety). For every prefix $v_0v_1 \ldots v_j \in \text{PrefPlays}(\rho)$, $s_0s_1 \ldots s_j$ is not a weakening of $s_0s_1 \ldots s_i$ for any $0 \leq i < j$.

2) (Improvement). For every integer $k > 0$ such that $s_k \in \mathcal{F}$, the prefix $s_0s_1 \ldots s_k$ is an improvement of $s_0s_1 \ldots s_{k-1}$.

Proof (Sketch). For statement (1) to hold, it must be the case that $R \neq R'$ holds for all pairs of outcomes $R, R' \in \text{MP}(s_i)$ and $R' \in \text{MP}(s_j)$. This is true because of Lma. 2 and the fact that every transition from $v_i$ to $v_{i+1}$, $j < i \leq k$, that violates the condition is disabled by Def. 8.

To see why statement (2) holds, consider an integer $k > 0$ such that $v_k \in \mathcal{F}$. Then, by construction, there exists a pair $R \in \text{MP}(s_{k-1})$ and $R' \in \text{MP}(s_k)$ such that $R' \uparrow R$. □

In words, the improvement GDPRs guarantees by construction that no play in $\text{Plays}(\mathcal{M})$ violates the safety condition of Def. 7. Moreover, it helps identify the opportunistic states as the ones that have an outgoing transition into $\mathcal{F}$.

Corollary 1. A play $\rho \in \text{Plays}(\mathcal{M})$ is improving if and only if $\text{Occ}(\rho) \cap \mathcal{F} \neq \emptyset$.

As a result, the problem of determining whether an improvement is possible from a state $v \in \mathcal{V}$ reduces to checking whether a state in $\mathcal{F}$ can be reached from $v$ with a positive probability (in case of SPI strategy) or with probability one (in case of SASI strategy).

Theorem 1. The following statements hold:

1) Any positive winning strategy $\pi^{\text{PosWin}}(\mathcal{F})$ in $\mathcal{M}$ is an SPI strategy.

2) Any almost-sure winning strategy $\pi^{\text{ASWin}}(\mathcal{F})$ in $\mathcal{M}$ is an SASI strategy.

The proof follows from the fact that there exists a (resp., every) play $\rho \in \text{Cone}(\mathcal{M}, v_0, \pi)$ induced by any positive (resp., almost-sure) winning strategy $\pi$ visits $\mathcal{F}$ with positive probability (resp., probability one) [26]. Therefore, Thm. 1 establishes that by following $\pi^{\text{PosWin}}(\mathcal{F})$ (resp., $\pi^{\text{ASWin}}(\mathcal{F})$), the agent is ensured to make an improvement with a positive probability (resp., with probability one). It is noted that an SPI (resp., SASI) strategy exists if and only if the corresponding positive (resp., almost-sure) winning strategy exists in $\mathcal{M}$.

The SPI and SASI strategies from Thm. 1 guarantee that at least one improvement will occur with positive probability or with probability one. Next, we present Alg. 1, which uses which we can determine the maximum number of improvements that can almost surely be made from a given state in $\mathcal{M}$. The algorithm to determine the maximum number of improvements possible from a given state in $\mathcal{M}$ with a positive probability and its properties are similar to Alg. 1.

First, note the following properties of the improvement MDP which follow from the construction of MDP.

Algorithm 1 Level set for constructing SASI strategy

Inputs: Improvement MDP, $\mathcal{M} = (\mathcal{V}, A, \Delta, v_0, \mathcal{F})$.

Outputs: Level set, $\mathcal{W}$.

1: $i \leftarrow 0$
2: $R_i \leftarrow \mathcal{F}$
3: while $R_i$ is not empty do
4:   $W_{i+1} \leftarrow \text{ASWin}(R_i)$
5:   $R_{i+1} \leftarrow \{(s, 1) \in \mathcal{F} \mid (s, 0) \in W_{i+1}\}$
6: if $i = 0$ then
7:   Add $V \setminus W_{i+1}$ to level 0 in $\mathcal{W}$.
8: Add $W_{i+1}$ to level $i + 1$ in $\mathcal{W}$.
9: $i \leftarrow i + 1$
10: return $\mathcal{W}$

Proposition 2. Consider two states $(s, 0), (s, 1) \in \mathcal{V}$, it holds that for any action $a \in A$, we have $\text{Supp}(\Delta((s, 0), a)) = \text{Supp}(\Delta((s, 1), a))$.

The proof is straightforward because given $(s, 0), (s, 1)$, for any action $a \in A$, if a transition from $s$ to $s'$ given $a$ is improving, then $\Delta((s, 0), a, (s', 1)) > 0$ and $\Delta((s, 1), a, (s', 1)) > 0$. Else, $\Delta((s, s), a, (s', 0)) > 0$ and $\Delta((s, 1), a, (s', 0)) > 0$.

Corollary 2. The final states $\mathcal{F}$ can be visited again from a state $(s, 1) \in \mathcal{V}$ with a positive probability (resp., with probability one) if and only if $\mathcal{F}$ can be visited from $(s, 0)$ with a positive probability (resp., with probability one).

Proof. Let $\pi$ be a positive winning strategy to visit $\mathcal{F}$ from $(s, 0)$. Let $Y = \text{Supp}(\Delta((s, 0), a))$ for some $a \in \pi((s, 0))$. By the property of a positive winning strategy, a state in $\mathcal{F}$ is reached with positive probability by following $\pi$ from any state in $Y$. By Proposition 2, $Y = \text{Supp}(\Delta((s, 1), a))$. Therefore, by choosing $a$ at $(s, 1)$ and then following $\pi$, a state in $\mathcal{F}$ is visited with positive probability from $(s, 1)$. The proof for almost-sure winning is similar. □

Intuitively, Alg. 1 constructs a set $\mathcal{W}$ of level sets such that from any state that appears at $k$-th level in $\mathcal{W}$, at least $k$ visits to $\mathcal{F}$ are guaranteed and, thereby, at least $k$ improvements can be made.

For this purpose, it iteratively computes the almost-sure winning region to visit the states in $R_i \subseteq \mathcal{F}$, from which $\mathcal{F}$ can be visited at least $i$ times. We denote by $W_i$ the $i$-th level set. The level-0 of $\mathcal{W}$ contains the states $V \setminus \text{ASWin}(\mathcal{F})$ from which $\mathcal{F}$ cannot be visited again with probability one. That is, 0-visits to $\mathcal{F}$ are guaranteed from any state in level-0 of $\mathcal{W}$. Every state in level-1 of $\mathcal{W}$ is almost surely winning to visit $\mathcal{F}$. Hence, at least one visit to $\mathcal{F}$ is guaranteed. Now, consider the subset $R_1 = \{(s, 1) \in \mathcal{F} \mid (s, 0) \in W_1\}$ of final states $\mathcal{F}$. By Corollary 2, because $(s, 0) \in W_1 = \text{ASWin}(\mathcal{F})$, there exists a strategy from every state in $R_1$ to visit $\mathcal{F}$ with probability one. Therefore, from any state $(s, 0) \in W_2 = \text{ASWin}(R_1)$ at least two improvements are guaranteed—first, when visiting $(s', 1) \in R_1$ and, second, when visiting $R_0 = \mathcal{F}$ by following the almost-sure winning strategy at $(s', 1)$. 
Repeating a similar argument, it follows that at least $k$-visits are guaranteed almost surely from states at $k$-th level in $\mathcal{W}$.

The largest integer $k \geq 0$ such that the state $(s, 0) \in V$ appears at $k$-th level of $\mathcal{W}$ is called the rank of the states $(s, 0)$ and $(s, 1)$, denoted as $\text{rank}(s, 0) = \text{rank}(s, 1) = k$.

**Proposition 3.** From any state $v = (s, m) \in V$, $m \in \{0, 1\}$, there exists a strategy to visit $\mathcal{F}$ at least $\text{rank}(v)$-many times.

**Proof.** We construct the strategy that achieves $\text{rank}(v)$ improvements: First, if $\text{rank}(v) = k$, then by construction it is in $\text{ASWin}(R_{k-1})$. Following the almost-sure winning strategy a state in $R_{k-1}$ can be reached with probability one and thus the first improvement is made. Upon reaching a state, say $(s', 1)$, in $R_{k-1}$, we have $(s', 0) \in W_{k-1}$. Because $W_{k-1} = \text{ASWin}(R_{k-2})$, an almost-sure winning strategy exists to reach $R_{k-2}$ and hence the second improvement. Repeating similar steps, eventually, $R_0$ will be reached after the $k$-th improvement.

**Corollary 3.** From any state $v = (s, m) \in V$ at most $\text{rank}(v)$-many visits to $\mathcal{F}$ are almost surely guaranteed.

**Proof (Sketch).** By contradiction. Suppose that $\text{rank}(v) = k$ but $k+1$ visits to $\mathcal{F}$ are possible from $v$. Since $k+1$ visits are possible from $v$, by definition of $\mathcal{W}$, it must be the case that $v \in W_{k+1}$. If $v$ is at $(k+1)$-th level in $\mathcal{W}$ then its rank must be at least $k+1$—a contradiction.

The proof of Proposition 3 defines the strategy that allows the agent to make $\text{rank}(v)$ improvements from any state $v$.

**Complexity.** Alg. 1 runs in polynomial time with respect to the size of $\mathcal{M}$ since the while loop can run no more than $|V|$ times and the complexity of $\text{ASWin}$ is quadratic in the size of $\mathcal{M}$ [21].

**VI. EXAMPLE: ROBOT MOTION PLANNING**

We illustrate our approach using a motion planning problem for a robot in a $5 \times 5$ gridworld as shown in Figure 2. The gridworld environment consists of seven regions: \{A : (0, 0), B : (2, 0), C : (4, 0), D : (2, 4), E : (4, 4), F : (1, 2)\} from which the robot must pick up an item. There is a charging station at cell (4, 2). Each cell is denoted using the convention (row, col). The robot can choose among four actions $n, s, e, w$ to deterministically move north, east, south, and west by one cell. The actions $e, w$ are disabled in the cells (4, 2) and (2, 2). The cells (1, 1), (3, 1), (1, 3), (3, 3) are slippery; that is, whenever the robot moves into any of these cells, say (1, 1), it may non-deterministically end up in either the same cell (1, 1), or the cell north to it (2, 1), or south to it (0, 1). In any cell, if applying an action results in a cell that is outside the gridworld or contains an obstacle, the robot returns to the same cell. The robot has a limited battery of 8 units, which it may recharge by visiting the charging station. The robot spends 1 unit to execute each action.

Initially, only the items at $A, B,$ and $C$ are available for pickup. That is, if the robot visits the charging station or regions $D, E, F$, then neither its battery will be recharged nor will it be able to pick up items $D, E, F$. When the robot picks up an item at $A$ or $B$, the charging station and the items at $D, E$ become available. When the robot picks up an item at $C$, the charging station and the items at $E, F$ become available. The following preference about picking up the items is given to the robot: $D \triangleright A, E \triangleright A, D \triangleright B, E \triangleright B, E \triangleright C, F \triangleright C$. By default, picking up any item is preferred to not picking up any item.

Note that the preference model given to the robot is incomplete as well as combinative. It is incomplete because picking up items $A, B, C$ are mutually incomparable outcomes. Similarly, picking up items $D, E, F$ are mutually incomparable. It is combinative because, for instance, any play in which robot picks up an item from $D$ or $E$ is considered preferred to a play in which robot only picks an item from $A$ or $B$, even though to pick an item from $D$ or $E$ an item from $A$ or $B$ must be picked first.

We implemented the example in Python 3.9 on a Windows 10 machine with a core i7, 2.80GHz CPU, and 32GB memory. The SPI and SASI strategies are computed using set-based positive and almost-sure winning algorithms implemented in https://github.com/abhibp1993/ggsolver/. We discuss a few noteworthy observations next. The improvement MDP for this case has 3600 states and 18496 transitions, whereas the improvement MDP has 7200 states and 35524 transitions. The time required for constructing the improvement MDP is 9.47 seconds which includes time required to solve for almost-sure winning regions to visit $A\,\sim\,F$ independently. Whereas, the construction of SASI and SPI strategies took 6.54 seconds and 7.23 seconds, respectively.

Consider the initial state $s_0 = (2, 2, 8, (1, 1, 0, 0, 0, 0, 0))$ in which the robot is at cell (2, 2) with 8 units of battery. The fourth component of the state denotes which items are available for pickup, with the last element of the tuple reserved for the availability of the charging station. In this state, the robot has no almost-sure winning strategy to visit any of $A, B,$ or $C$. This is because to visit, say, $A$; the robot must visit the slippery cell (1, 1). But whenever (1, 1) is visited, the robot may reach (2, 1) with a positive probability. Hence, $MP(\text{Outcomes}(s_0)) = \emptyset$.  

![Figure 2](https://github.com/abhibp1993/ggsolver/)
TABLE I

| Rank-1 | SASI | SPI |
|--------|------|-----|
| 768    |      |     |
| 98     |      | 167 |

NUMBER OF STATES FROM WHICH THE ROBOT HAS SPI AND SASI STRATEGIES TO MAKE AT LEAST 1 OR AT LEAST 2 IMPROVEMENTS.

When we use the SASI concept, the rank of the state \((s_0, 1)\) is 2, indicating that two improvements are almost surely guaranteed. This is understood by observing the SASI strategy which chooses action \(N\) at \((s_0, 0)\) to reach \(s_1 = (3, 2, 7, (1, 1, 1, 0, 0, 0, 0))\). At \((s_1, 0)\) the strategy selects \(W\) and visits either \(B\) or \(C\) with probability one. Since a pickup from \(B\) and \(C\) are incomparable, both actions \(N\) and \(S\) are deemed valid under SASI strategy at \((3, 1, 6, (1, 1, 1, 0, 0, 0, 0))\). On visiting either \(B\) or \(C\), the SASI strategy follows the almost-sure winning strategy to visit either \(D\) or \(E\) to make a second improvement. Since visiting cell \((3, 3)\) may result in returning back to cell \((3, 2)\) with a positive probability, the robot can recharge itself until a successful visit to \(E\) or \(D\) is made.

The SASI strategy at \((s_0, 0)\) does not select \(S\) because a second improvement cannot be guaranteed with probability one after visiting \(A\) since the robot may remain at the cell \((0, 1)\) until its battery runs out. However, we observe that the SPI strategy at \((s_0, 0)\) allows the selection of both actions \(N\), \(S\) at \((s_0, 0)\) since in both cases, two improvements are possible with positive probability.

We conclude with Table I that shows the number of states from which the robot has an SPI and SASI strategies to make at least 1 or 2 improvements, since the maximum number of improvements possible under given preference model is 2. We note that the states from which a SASI strategy exists are a subset of states from which an SPI strategy exists.

VII. CONCLUSION

In this paper, we introduced two solution concepts, namely SPI and SASI to solve a preference-based planning problem given a combinative, incomplete preference model over infinite plays of a stochastic system. In the improvement MDP, we showed that the synthesis of SPI and SASI strategies reduces to that of computing positive and almost-sure winning strategies. Finally, we designed an algorithm using which we can synthesize a strategy that induces the maximum number of improvements under the SASI concept. Building on this work, there are a number of future directions: 1) it is possible to consider a preference over temporal objectives that encompass more general properties such as safety, recurrence, and liveness; 2) it remains open as how to connect qualitative reasoning with quantitative planning with such preference specifications.

REFERENCES

[1] R. Hastie and R. M. Dawes, Rational choice in an uncertain world: The psychology of judgment and decision making. Sage, 2010.
[2] J. A. Baier and S. A. McIraith, “Planning with Preferences,” *AI Magazine*, vol. 29, no. 4, p. 25, 2008.
[3] T. C. Son and E. Pontelli, “Planning with preferences using logic programming,” *Theory and Practice of Logic Programming*, vol. 6, no. 5, pp. 559–607, 2006.
[4] M. Bienvenu, C. Fritz, and S. A. McIraith, “Specifying and computing preferred plans,” *Artificial Intelligence*, vol. 175, no. 7-8, pp. 1308–1345, 2011.
[5] S. O. Hansson and T. Gråne-Yanoff, “Preferences”, the stanford encyclopedia of philosophy (spring 2022 edition). [Online]. Available: https://plato.stanford.edu/archives/spr2022/entries/preferences/
[6] H. Rahmani, A. N. Kulikarni, and J. Fu, “Probabilistic planning with partially ordered preferences over temporal goals,” in *2021 American Control Conference (ACC)*. IEEE, 2023 (accepted).
[7] M. Lahijanian, S. Almagor, D. Fried, L. Kavraki, and M. Vardi, “This time the robot settles for a cost: A quantitative approach to temporal logic planning with partial satisfaction,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.
[8] J. Fu, “Probabilistic planning with preferences over temporal goals,” in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 4854–4859.
[9] A. Sen, “Maximization and the act of choice,” *Econometrica: Journal of the Econometric Society*, pp. 745–779, 1997.
[10] C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade, “Preferences in ai: An overview,” *Artificial Intelligence*, vol. 175, no. 7, pp. 1037–1052, 2011, representing, Processing, and Learning Preferences: Theoretical and Practical Challenges.
[11] J. J. Thomson, “Killing, letting die, and the trolley problem,” *The monist*, vol. 59, no. 2, pp. 204–217, 1976.
[12] J. Tumova, G. R. Santhanam, S. Basu, and V. Honavar, “Representing and Reasoning about Preferences in AI: An overview,” *Artificial Intelligence and Machine Learning*, vol. 25, no. 6, pp. 1370–1381, 2009.
[13] T. Wongpiromsarn, K. Slutzky, E. Frazzoli, and U. Topcu, “Minimum-violation planning for autonomous systems: Theoretical and practical considerations,” in *2021 American Control Conference*, 2021, submitted.
[14] H. Rahmani and J. M. O’Kane, “What to do when you can’t do it all: Temporal logic planning with soft temporal logic constraints,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 6619–6626.
[15] N. Mehdipour, C.-I. Vasile, and C. Bela, “Specifying User Preferences Using Weighted Signal Temporal Logic,” *IEEE Control Systems Letters*, vol. 5, no. 6, pp. 2006–2011, Dec. 2021.
[16] M. Lahijanian and M. Kwiatkowska, “Specification revision for Markov decision processes with optimal trade-off,” in *Proc. 55th Conference on Decision and Control (CDC’16)*, 2016, pp. 7411–7418.
[17] M. Li, A. Turrini, E. M. Hahn, Z. She, and L. Zhang, “Probabilistic preference planning problem for markov decision processes,” *IEEE transactions on software engineering*, 2020.
[18] K. Chatterjee, L. Doyen, and T. A. Henzinger, “Qualitative analysis of partially-observable markov decision processes,” in *Mathematical Foundations of Computer Science 2010*, P. Hliněný and A. Kučera, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 258–269.
[19] K. Hress-Gazit, G. E. Fainekos, and G. J. Pappas, “Temporal-logics-based reactive mission and motion planning,” *IEEE transactions on robotics*, vol. 25, no. 6, pp. 1370–1381, 2009.
[20] G. R. Santhanam, S. Basu, and V. Honavar, “Representing and Reasoning with Qualitative Preferences: Tools and Applications,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 10, pp. 1–154, Jan. 2016.
[21] C. Baier and J.-P. Katoen, *Principles of model checking*. MIT press, 2008.
[22] O. Kupferman and M. Y. Vardi, “Model checking of safety properties,” *Formal methods in system design*, vol. 19, no. 3, pp. 291–314, 2001.
[23] D. Bouyssou, D. Dubois, and M. Pirol, *Concepts & Methods of Decision-Making*. John Wiley & Sons Inc., 2009.
[24] S. O. Hansson, *The structure of values and norms*. Cambridge University Press, 2001.
[25] L. de Alfaro and R. Majumdar, “Quantitative solution of omega-regular games,” in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 675–683.
[26] K. Chatterjee and T. A. Henzinger, “A survey of stochastic ω-regular games,” *Journal of Computer and System Sciences*, vol. 78, no. 2, pp. 394–413, 2012.