

# Diverse, Top- $k$ , and Top-Quality Planning Over Simulators

Lyndon Benke<sup>a,b</sup>, Tim Miller<sup>b</sup>, Michael Papisimeon<sup>a</sup> and Nir Lipovetzky<sup>b</sup>

<sup>a</sup>Defence Science and Technology Group, Australia

<sup>b</sup>The University of Melbourne, Australia

**Abstract.** Diverse, top- $k$ , and top-quality planning are concerned with the generation of sets of solutions to sequential decision problems. Previously this area has been the domain of classical planners that require a symbolic model of the problem instance. This paper proposes a novel alternative approach that uses Monte Carlo Tree Search (MCTS), enabling application to problems for which only a black-box simulation model is available. We present a procedure for extracting bounded sets of plans from pre-generated search trees in best-first order, and a metric for evaluating the relative quality of paths through a search tree. We demonstrate this approach on a path-planning problem with hidden information, and suggest adaptations to the MCTS algorithm to increase the diversity of generated plans. Our results show that our method can generate diverse and high-quality plan sets in domains where classical planners are not applicable.

## 1 Introduction

Automated planning is a model-based approach to the discovery of solutions to sequential decision problems. In typical applications, a search algorithm generates a single strategy or action sequence to solve a specific planning task. This paper explores the generation of bounded sets of plans, described as *diverse*, *top- $k$* , or *top-quality* planning when the set is bounded by diversity, cardinality, or quality, respectively. Planners that produce plan sets have applications in a wide range of fields, including planning for problems with incomplete or unknown user preferences [18], scenario prediction for risk management [27], plan recognition [28], plan repair [10], and explanation generation [5]. Our research is motivated by the discovery of novel agent behaviours in the context of operations research [16, 19, 23], recently proposed as an input to deceptive mission planning [3].

Previous approaches to generating plan sets have focused on classical planning, typically employing repeated calls to off-the-shelf solvers [14, 24, 29, 30]. Such approaches take advantage of structural features of the planning problem for efficient solution discovery, and hence require a symbolic model of the domain and problem instance in a declarative planning language such as STRIPS or PDDL. We consider problems for which a declarative model is *not* available, and classical planners cannot be used off-the-shelf as action effects are not known before execution. Instead, a black-box simulator provides a sample of a successor state and reward given an input state and action. Such problems are particularly widespread in operations research, where Monte Carlo simulation is used to study complex sociotechnical systems [20]. These simulations include high-fidelity models of cyber-physical systems, such as fighter aircraft, that are impractical or impossible to model in symbolic planning languages.

Monte Carlo Tree Search (MCTS) is an alternative approach to

planning that has achieved considerable success in complex domains with large branching factors [6]. Unlike classic planners, MCTS does not require a symbolic model of the planning problem, and can be used when only a generative model is available. While it is commonly used for online, adversarial planning, building a new search tree at each time-step to estimate the best single action from a given state, it has also proven to be effective in offline and single-player contexts, such as those studied in previous research on diverse, top- $k$  and top-quality planning problems.

In this paper, we present a novel method for extracting bounded sets of plans from pre-generated search trees. While we focus on MCTS, our approach may be applied to any tree search algorithm that backpropagates node values. We additionally propose a metric for the relative quality of paths in a search tree, based on expected return, that guarantees best-first ordering. We present proofs for soundness and completeness, and discuss modifications to the MCTS algorithm to increase the diversity of generated plans. Finally, we provide an empirical evaluation in a path-planning scenario with hidden information, and discuss the practicality of our approach for diverse, top- $k$  and top-quality planning. The experimental results suggest that diverse planning is most effective for problems with moderate to high levels of hidden information, where any decrease in efficiency is offset by an increased probability of success.

In line with related work, this paper considers the generation of plan sets in the context of deterministic single-player planning problems. The extension of our approach to multi-agent problems is left for future work.

## 2 Related Work

Previous works have explored top- $k$ , top-quality and diverse planning using either generalisations of A\* search known as K\* [24], or replanning approaches that repeatedly reformulate the planning task to forbid a growing set of solutions [14, 29]. Approaches based on K\* require a heuristic function that may be difficult to obtain, and are outperformed by replanning approaches for small values of  $k$ , while replanning approaches incur significant additional cost as the size of the plan set increases. Additionally, replanning approaches require a symbolic model of the domain and problem instance in a declarative planning language such as STRIPS or PDDL, and as a result cannot be applied to problems for which only a black-box simulation model is available. This limitation also applies when deriving domain-independent heuristics for K\*-based approaches.

Recent works have proposed addressing diversity separately to top- $k$  and top-quality planning, by post-processing an existing set of plans. One approach proposes a greedy method, iteratively selecting plans

that maximise the diversity of the growing set [11], while another finds an optimally diverse subset by solving a mixed-integer linear programming problem [13]. However, since these approaches operate on pre-generated sets of plans, they cannot be applied over simulators without first using a method such as the one presented in this paper to generate a top- $k$  or top-quality set of plans.

MCTS has been applied to single-player problems in a variety of domains, in some cases outperforming classical methods and achieving a level of performance comparable to the best hand-crafted approaches [4, 26]. Relevant applications have included constructing an MCTS tree offline and querying the subsequent tree during gameplay [2], and employing offline MCTS to generate a Pareto set of possible winning positions [22]. While the extraction of a single optimal plan from an MCTS tree has been considered in the context of nested Monte Carlo search [8] and hierarchical task decomposition [21], to our knowledge no previous work has explored the generation of bounded *sets* of plans using MCTS; our synthesis of these concepts appears to be unique.

### 3 Background and Preliminaries

#### 3.1 Markov Decision Processes

A planning problem that provides rewards rather than explicit goals is typically formulated as a Markov Decision Process (MDP). In an MDP, an agent interacts with an environment by taking actions and receiving rewards. The objective of the agent is to maximise the expected cumulative reward it receives over time. An MDP is defined by a tuple  $\langle S, A, T, R \rangle$ , where  $S$  is a set of possible states,  $A$  is a set of actions available to the decision-maker,  $T(s, a, s')$  is a transition model, and  $R(s)$  is a reward function. The transition model determines the probability of reaching state  $s'$  after taking action  $a$  in state  $s$ , while the reward function defines the reward received after a transition. For deterministic problems, such as those considered here, taking action  $a \in A$  in state  $s \in S$  always results in a unique next state  $s' \in S$ . A solution to an MDP is a *plan* or *policy* that maximises the expected reward for the agent. A plan is a sequence of actions  $\pi = \{a_0, \dots, a_n\}$  that generates one or more sequences of states  $\{s_0, \dots, s_n\}$ , where  $s_0$  is the initial problem state. A policy is a mapping from states to actions that specifies the action to select from any given state  $s \in S$ .

#### 3.2 Search Trees

A *game tree* is a hierarchical graph structure that is used to model and analyse sequential decision problems such as MDPs, where each node represents a game state, and each edge a possible transition from one state to another. A *complete* game tree includes every node that is reachable from the initial state, capturing the full search space for the decision problem. Due to computational limitations, in practical applications a partial game tree or *search tree* is usually used instead, representing a subgraph of the complete tree that is constructed through repeated sampling of the search space. For example, A\* search builds a partial tree by selecting nodes based on a heuristic, while Monte Carlo Tree Search employs simulated playouts. A plan in a search tree corresponds to a path from the tree root to a leaf node.

#### 3.3 Monte Carlo Tree Search (MCTS)

MCTS is a family of algorithms that apply stochastic game-tree search to sequential decision problems. A search tree is generated that contains nodes representing the visited subset of *state-action* pairs, with

each node maintaining an approximation of the value of a particular transition. During each iteration, the MCTS algorithm selects a leaf node by descending the tree from the root according to a *tree policy*, expands the selected node if possible by adding a new child node, executes a simulated playout according to a *default policy*, and back-propagates the result by updating the statistics for each of the visited nodes. After a predetermined time or number of iterations, the action corresponding to the best child of the root node is returned.

The tree policy balances exploitation of nodes with high value estimates against exploration of nodes with low visit counts by recursively applying a multi-armed bandit policy, such as UCB1 [1] which selects the node that maximises the value:

$$\text{UCB1}(\sigma) = Q(\sigma) + C \sqrt{\frac{2 \ln n}{n_\sigma}} \quad (1)$$

where  $Q(\sigma)$  is the value estimate for node  $\sigma$ ,  $n_\sigma$  and  $n$  are the number of times the node and its parent have been visited respectively, and  $C$  is a positive constant balancing exploration and exploitation. The node value considered by the tree policy,  $Q(\sigma)$ , is the average reward based on repeated Monte Carlo playouts, representing an estimate of the expected return for choosing the corresponding action in the previous state:

$$Q(\sigma) = Q(s, a) = \frac{z_\sigma}{n_\sigma} \quad (2)$$

where  $s$  and  $a$  are the state and action leading to the node, and  $z_\sigma$  is the total reward received at  $\sigma$ . As results are backpropagated, the  $q$ -value for each node also represents the weighted average of the values of its set of child nodes  $C(\sigma)$ :

$$Q(\sigma) = \frac{1}{\sum_{\sigma' \in C(\sigma)} n_{\sigma'}} \sum_{\sigma' \in C(\sigma)} n_{\sigma'} Q(\sigma') \quad (3)$$

Previous research has demonstrated that MCTS performance can be improved in single-player games by choosing the final move according to the *maximum* playout result rather than the average, since there is no adversary to prevent optimistic play [4]. In such cases, the  $q$ -value that is used for action selection is the maximum child value:

$$Q_{max}(\sigma) = \max_{\sigma' \in C(\sigma)} Q(\sigma') \quad (4)$$

For games that award a win or loss to the player,  $Q(\sigma)$  may be considered an estimate of the probability that the player will win after visiting the node, when following the tree and default policies.

#### 3.4 Diverse, Top-k and Top-Quality Planning

The generation of multiple solutions to a planning problem, rather than a single optimal plan, is known as *diverse*, *top-k*, or *top-quality* planning, depending on the attributes used to bound the plan set. Where top- $k$  planning restricts the set of plans to a fixed size, top-quality planning enforces a minimum solution quality, and diverse planning adds a constraint on plan similarity. We note here the distinction between a *plan* in the context of classical planning, being a sequence that successfully achieves an explicit goal, and a *plan* in the context of an MDP, being any sequence with an associated expected return.

Top- $k$  planning addresses the problem of finding a finite set of plans of size  $k$ , such that no plan with greater quality exists outside of the set according to some measure of plan quality provided by the user [9, 24, 29]. Top- $k$  planning extends the well-studied *k shortest paths* problem, and is typically applied to problems requiring sets

of high-quality but similar plans, such as plan repair [10]. Single planning may be considered a special case of top- $k$  planning, for which  $k = 1$ . Drawing on classical planning [29], a top- $k$  planning problem may be defined for an MDP as follows:

**Definition 1** (Top- $k$  planning problem).

Given a planning problem  $\Pi = \langle S, A, T, R \rangle$ , measure of plan quality  $Q_{plan}(\pi)$ , and natural number  $k$ , find a subset  $\mathcal{P}$  in the set of all plans  $\mathcal{P}_\Pi$  such that:

1. there exists no plan  $\psi \in \mathcal{P}_\Pi$  with  $\psi \notin \mathcal{P}$  that has quality  $Q_{plan}(\psi)$  greater than some plan  $\pi \in \mathcal{P}$ , and
2.  $|\mathcal{P}| = k$  if  $|\mathcal{P}_\Pi| \geq k$ , or  $|\mathcal{P}| = |\mathcal{P}_\Pi|$  otherwise.

Note that  $k$  is regarded as a constraint rather than a requirement [18], where some works consider the problem to be unsolvable if fewer than  $k$  plans are found.

Top-quality planning is a variation on top- $k$  planning wherein the plan set is subject to limits on quality, rather than cardinality [12, 25, 32]. This approach is useful when the user does not require a specific number of plans, but has well-defined constraints that bound acceptable solutions. Drawing on classical planning [12], a top-quality planning problem may be defined for an MDP as follows:

**Definition 2** (Top-quality planning problem).

Given a planning problem  $\Pi = \langle S, A, T, R \rangle$ , measure of plan quality  $Q_{plan}(\pi)$ , and quality constraint  $q$ , find the set of plans  $\mathcal{P} \subset \mathcal{P}_\Pi$ , where  $\mathcal{P} = \{\pi \in \mathcal{P}_\Pi \mid Q_{plan}(\pi) \geq q\}$ .

Diverse planning considers the problem of generating sets of plans that are meaningfully different according to some measure of plan distance. Diversity is defined as the minimum or average pairwise distance within a set of plans, and has been approached using both qualitative measures [17], and domain-independent, quantitative measures [18]. Diverse planning is typically combined with a bound on plan set cardinality [7, 18], in some cases with an additional bound on plan quality [11, 25]. We consider the problem of maximising quality given a bound on diversity, in contrast to recent works that maximise diversity given a bound on quality [11, 13]. This approach is more applicable to our domains of interest, which typically require the highest-quality plans with sufficient diversity, rather than the most diverse set of plans. A quality-optimal, diversity-bounded top- $k$  planning problem may be defined for an MDP as follows:

**Definition 3** (Diverse top- $k$ -quality planning problem).

Given a planning problem  $\Pi = \langle S, A, T, R \rangle$ , natural number  $k$ , measure of plan quality  $Q_{plan}(\pi)$ , quality constraint  $q$ , measure of minimum pairwise distance  $D(\pi, \mathcal{P})$ , and distance constraint  $d$ , find a maximal subset  $\mathcal{P}$  in the set of all plans  $\mathcal{P}_\Pi$  such that:

1. there exists no plan  $\psi \in \mathcal{P}_\Pi$ ,  $\psi \notin \mathcal{P}$  that has diversity  $D(\psi, \mathcal{P}) \geq d$  and quality  $Q_{plan}(\psi)$  greater than some plan  $\pi \in \mathcal{P}$ , and
2. for all  $\pi \in \mathcal{P}$ ,  $D(\pi, \mathcal{P} - \pi) \geq d$ , and
3. for all  $\pi \in \mathcal{P}$ ,  $Q_{plan}(\pi) \geq q$ , and
4.  $|\mathcal{P}| \leq k$ .

The solution to a diverse planning problem is hence a Pareto set balancing diversity and quality. Diverse planning by this definition may be considered a generalisation of top- $k$  and top-quality planning, since either alternative may be achieved by setting the diversity constraint to zero.

## 4 Plan Extraction from Search Trees

MCTS is typically used for online decision-making, informing the selection of a *single action* by generating a new partial tree at each

time step that provides an estimate of the value of each transition from the current state. However, MCTS may also be used to generate *plans* by extracting action sequences from the search tree, where a plan in this context corresponds to any path from the tree root to a leaf node. In the simplest case, the optimal plan may be extracted by starting at the root node and recursively selecting the child node with the maximum  $q$ -value until a leaf node is reached, returning the corresponding action sequence (plan  $\alpha$  in Fig. 1). Generating a set of plans is more complex, requiring a method for extracting and comparing alternative plans in a search tree without the benefit of explicit plan costs. The following sections introduce a measure for the relative quality of paths through a search tree, and a process for extracting plans in best-first order subject to various constraints.

### 4.1 Plan Quality Metric

Unlike in classical planning, where the quality of a plan is naturally defined by the sum of action costs [12], in standard MCTS implementations rewards are assumed to be received only in terminal states. Nodes instead track the backpropagated results of simulated playouts, typically in the form of a win or loss. We propose a quality measure that is based on expected return relative to the optimal plan, rather than plan cost:

**Definition 4** (Relative Plan Quality).

Given a node sequence (plan) representing a path through a search tree,  $\pi = \{\sigma_0, \dots, \sigma_n\}$ , the relative quality of  $\pi$  is defined as the product of the regret due to suboptimal action choices, quantified as the reduction in expected return attributed to each node in the plan:

$$Q_{plan}(\pi) = \prod_{i=0}^{n-1} \frac{Q(\sigma_{i+1})}{\max_{\sigma' \in C(\sigma_i)} Q(\sigma')} \quad (5)$$

where  $Q(\sigma) \in \mathbb{R}^+$  is the backpropagated quality of node  $\sigma$ , and  $C(\sigma_i)$  is the set of child nodes of  $\sigma_i$ .

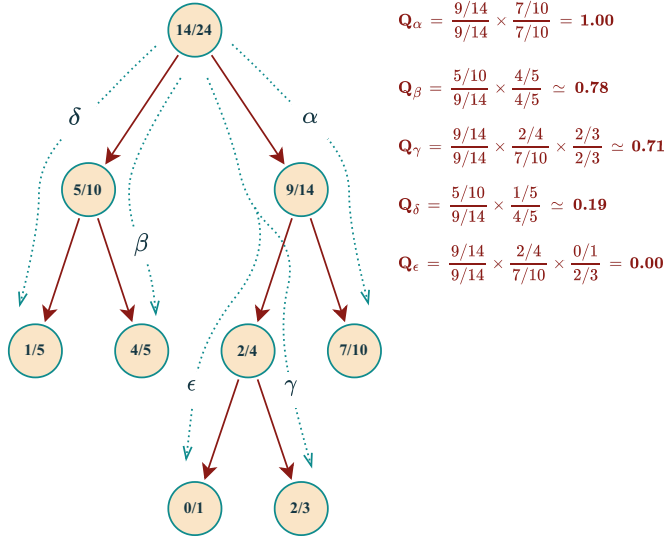
An example of plan quality calculation for a search tree is provided in Fig. 1. It follows that the top plan in any tree has a relative quality equal to one, since by definition the optimal plan selects the best available action in each state. This approach provides a natural basis for comparing plans with different lengths: given two plans of unequal length  $p$  and  $q$  with  $|p| < |q|$ , the quality of  $p$  is equal to the quality of a plan  $p'$  padded with optimal actions such that  $|p'| = |q|$ .

The *absolute* quality for any plan, corresponding to the expected return at the initial state when following the tree policy, may be calculated by multiplying the relative plan quality by the  $q$ -value of the root node. This absolute quality may vary greatly, since it depends on both the problem definition and the growth of the MCTS tree. By defining plan quality *relative* to the optimal plan, the metric defined above can be used to bound plan sets without knowledge of the actual  $q$ -value estimates in the tree.

Since the relative quality of a plan is defined as the product of the quality of its action choices, it follows that the quality for any sub-plan must be greater than or equal to the quality of the complete plan:

**Lemma 1.** *If  $\pi$  is a sequence of nodes  $\{\sigma_0, \dots, \sigma_n\}$  representing a path through a search tree with backpropagated node values, then any subsequence  $\psi \subset \pi$  has quality  $Q_{plan}(\psi) \geq Q_{plan}(\pi)$ .*

*Proof.* Let  $Q_{plan}(\pi)$  be the quality of  $\pi$  as defined by Eq. 5. Since  $Q_{plan}(\pi)$  is the product of the  $q$ -value ratio corresponding to each node  $\sigma \in \pi$ , and each ratio lies in the range  $[0, 1]$  as the denominator cannot be greater than the divisor, then  $\forall (\psi \subset \pi) : Q_{plan}(\psi) \geq Q_{plan}(\pi)$ .  $\square$



**Figure 1:** Illustration of relative plan quality calculation for a search tree with backpropagated rewards. Plan quality is defined by the reduction in expected return attributed to each node on the path (Eq. 5).

## 4.2 Plan Extraction Algorithm

We present a process for extracting top plans from a search tree in Algorithm 1. The algorithm is general and can be applied to diverse, top- $k$ , or top-quality planning problems by adjusting the bounding constraints and metrics accordingly. The search tree is traversed by incrementally expanding plans in order of quality, representing a best-first search of the plan space defined by the pre-generated tree. During execution of the algorithm, the open list  $\mathcal{O}$  contains a queue of node sequences ordered by quality using the metric defined in Section 4.1. These partial-depth sequences extend from the root node and are described as *plan stems* following a similar concept [33]. On each iteration of the outer loop, the highest-priority plan stem  $\pi$  is removed from the open list (line 6), and a set of extended plans is generated, each with length  $|\pi| + 1$ , by successively appending each child of the final node of  $\pi$  (lines 7–10). The quality of these new plan stems is evaluated, and any that meet the required minimum quality  $q$  are added to the open list (lines 11–15). If  $\pi$  was not expanded, then the plan has reached a leaf node, and if it meets the specified diversity requirement  $d$ , the complete plan is added to the final set of top plans  $\mathcal{P}$  (lines 17–19). Ties in plan quality are broken using diversity by continuing to evaluate plans until a lower-quality plan is found (lines 20–28). The process is repeated until the open list is empty, or the desired number of plans,  $k$ , has been generated.

**Lemma 2.** *At each step of the algorithm, the open list  $\mathcal{O}$  contains a set of plan stems guaranteed to produce one or more complete plans with quality  $Q_{plan}(\psi \in \mathcal{O}) \geq q$ .*

*Proof.* Let  $\sigma$  be the final node of a plan stem  $\pi$  in  $\mathcal{O}$  with plan quality  $Q_{plan}(\pi) \geq q$ . Since the  $q$ -value of node  $\sigma$ ,  $Q(\sigma)$ , is equal to the weighted average (Eq. 3) or maximum (Eq. 4) of the  $q$ -values of its child nodes, there necessarily exists at least one child  $\sigma'$  such that  $Q(\sigma') \geq Q(\sigma)$ . By Eq. 5, the quality of the extended plan corresponding to the best child is equal to  $Q_{plan}(\pi)$ . Applied recursively, a sequence  $\psi$  must exist extending to a leaf node with quality  $Q_{plan}(\psi) = Q_{plan}(\pi)$ , and hence each plan stem in  $\mathcal{O}$  is guaranteed to produce at least one complete plan with quality  $Q_{plan}(\psi) \geq q$ .  $\square$

### Algorithm 1: Plan Extraction

---

**Input:**  $\mathcal{T}$  ▷ root node of tree  
 $Q_{plan}(\pi)$  ▷ quality metric  
 $D(\pi, \mathcal{P})$  ▷ diversity metric  
 $k \in \mathbb{Z}^+$  ▷ cardinality constraint  
 $q \in [0, 1]$  ▷ quality constraint  
 $d \in [0, 1]$  ▷ diversity constraint

**Output:**  $\mathcal{P}$  ▷ set of top plans

```

1  $\mathcal{P} \leftarrow \emptyset$ 
2  $\pi_{\mathcal{T}} \leftarrow \langle \mathcal{T} \rangle$  ▷ seed plan containing tree root
3  $\mathcal{O} \leftarrow \text{PriorityQueue}()$  ▷ open list
4  $\mathcal{O}.\text{push}(\pi_{\mathcal{T}}, \text{priority} = 1)$ 
5 while  $\mathcal{O} \neq \emptyset$  do
6    $\pi \leftarrow \mathcal{O}.\text{pop-max}()$  ▷ highest-priority plan stem
7    $\sigma \leftarrow \pi.\text{get-last}()$  ▷ last node of plan stem
8    $\text{expanded} \leftarrow \text{False}$ 
9   foreach  $\sigma' \in \sigma.\text{child-nodes}()$  do
10     $\pi' \leftarrow \pi + \sigma'$  ▷ extended plan stem
11     $q_{\pi'} \leftarrow Q_{plan}(\pi')$  ▷ quality of new plan
12    if  $q_{\pi'} \geq q$  then
13       $\mathcal{O}.\text{push}(\pi', \text{priority} = q_{\pi'})$ 
14       $\text{expanded} \leftarrow \text{True}$ 
15    end
16  end
17  if  $\neg \text{expanded} \wedge (D(\pi, \mathcal{P}) \geq d)$  then
18    if  $|\mathcal{P}| < k$  then
19       $\mathcal{P}.\text{push}(\pi)$  ▷ add to set of top plans
20    else
21       $q_{min} \leftarrow \min_{x \in \mathcal{P}} Q_{plan}(x)$ 
22      if  $(d = 0) \vee (Q_{plan}(\pi) < q_{min})$  then
23        break
24       $\mathcal{P}_{q_{min}} \leftarrow \{x \mid x \in \mathcal{P}, Q_{plan}(x) = q_{min}\}$ 
25       $\pi_{dmin} \leftarrow \text{argmin}_{x \in \mathcal{P}_{q_{min}}} D(x, \mathcal{P} - x)$ 
26      if  $D(\pi, \mathcal{P}) > D(\pi_{dmin}, \mathcal{P} - \pi_{dmin})$  then
27         $\mathcal{P}.\text{replace}(\pi_{dmin}, \pi)$  ▷ break quality tie
28      end
29    end
30  end
31 end
32 return  $\mathcal{P}$ 

```

---

## 4.3 Theoretical Analysis

We describe Algorithm 1 as sound and complete for a given search tree if the set of returned plans satisfies the conditions for top- $k$ , top-quality or diverse planning as defined in Section 3.4. That is, the algorithm is sound if each plan in the returned set is guaranteed to meet the requirements for quality and diversity, and complete if every such plan in the tree is guaranteed to be found.

**Theorem 1 (Soundness).** *The plan extraction process presented in Algorithm 1 is sound with respect to a pre-generated search tree, such that no plan of higher quality can exist in the tree outside of the returned plan set.*

*Proof.* By contradiction: Assume that there exists in the tree some plan  $\psi$ , with quality  $Q_{plan}(\psi)$  that is greater than  $Q_{plan}(\pi)$ , the quality of a plan  $\pi$  in the top plan set. Recall that a plan corresponds



to a sequence of nodes starting at the tree root, and hence that  $\psi$  and  $\pi$  must share a common branching point. From [Algorithm 1](#), a new plan is added to the open list for each child of the node at the branching point with sufficient corresponding plan quality (lines 9–15). Following [Lemma 1](#), any subsequence of  $\psi$ , including all possible prefixes, must have quality greater than or equal to  $Q_{plan}(\psi)$ . Since  $Q_{plan}(\psi) > Q_{plan}(\pi)$ , a prefix of  $\psi$  must exist in the open list with quality greater than the quality of the plan  $\pi$ . The outer loop of [Algorithm 1](#) selects plans from the open list in order of plan quality, and since a prefix of  $\psi$  with quality  $Q_{plan}(\psi) > Q_{plan}(\pi)$  exists in the list, it is necessarily selected before  $\pi$ . Thus  $\psi$  must be added to the top plan set before  $\pi$ , so the assumption is not true, and the theorem holds.  $\square$

**Theorem 2** (Completeness). *The plan extraction process presented in [Algorithm 1](#) is complete with respect to a pre-generated search tree, such that the set of returned plans contains every top plan in the tree that meets the requirements for quality and diversity.*

*Proof.* From [Lemma 1](#), for each top plan in the search tree, every subsequence must have greater or equal quality. Since the open list is populated by expanding nodes starting at the root node, a prefix of each top plan is necessarily added to the open list on the first iteration, and by [Theorem 1](#) the complete plan must eventually be added to the set of top plans before any plan with lower quality. The outer loop continues until either  $k$  plans have been generated or the open list is empty, and hence all top plans in the tree are extracted by the algorithm.  $\square$

It has been demonstrated that given an infinite number of episodes, the probability of UCB1 selecting a suboptimal action at any given node converges to zero [\[15\]](#), and MCTS with UCB1 is hence optimal. It follows that if [Algorithm 1](#) is sound and complete with respect to a given search tree, and MCTS converges to the optimal policy given sufficient time and memory, then [Algorithm 1](#) is also sound and complete with respect to the complete decision problem.

#### 4.4 Complexity Analysis

The worst-case complexity for top- $k$  plan extraction given a pre-generated search tree is  $O(k \cdot d)$ , where  $d$  is the depth of the tree. Since the algorithm extracts plans in order of quality, the first  $k$  iterations of the loop produce the top- $k$  plans, and each iteration expands at most  $d$  nodes. The complexity for top-quality is  $O(p \cdot d)$ , where  $p$  is the number of leaf nodes (paths) in the tree and  $d$  is the depth. However, this worst case is unlikely in practice, and will occur only if all plans in the tree are above the quality limit. The algorithm assesses the plans in order of quality, so if there are  $k$  plans with a quality above the threshold, the algorithm assesses  $k \cdot d$  nodes; however, we do not know  $k$  a priori. For diverse planning, the worst-case complexity is  $O(p \cdot d)$  – at worst, we are required to assess all paths in the tree. For all of these, in practice, since the tree is fully traversed at least once during its creation, even the worst case represents a fraction of the planning process. We provide empirical results in [Section 6](#).

## 5 Diverse Planning with MCTS

When applied to top- $k$  or top-quality planning problems, the algorithm presented in the previous section bounds the set of generated solutions according to a cardinality constraint  $k$ , a quality constraint  $q$ , or both. This approach guarantees the generation of a set of plans such that no plan of greater quality exists outside of the returned set. However,

these plans are not guaranteed to differ significantly from one another, and may represent minor variations of the top plan, particularly given MCTS biases exploration toward the most promising branches of the game tree. In some applications this is desirable, such as plan recognition [\[28\]](#) or plan repair [\[10\]](#). In other situations, such as course of action generation or travel planning [\[17\]](#), the user desires a set of plans that are *meaningfully* distinct according to some qualitative or quantitative measure, to provide greater coverage of the solution space.

### 5.1 Diverse Plan Extraction

To enforce diversity in the plan set generated by [Algorithm 1](#), each complete plan is evaluated against the growing set of top plans, and any that do not meet the required diversity are rejected (line 17). The diversity of a candidate plan,  $D(\pi, \mathcal{P})$ , is a measure of its dissimilarity to the set of plans  $\mathcal{P}$ . Plan diversity is calculated from the *minimum pairwise distance* relative to  $\mathcal{P}$ , which may be based on qualitative measures that are specific to the domain, or quantitative measures that are domain-independent.

A number of domain-independent plan distance measures have been presented in previous works, typically comparing states, actions, causal links, or landmarks [\[7, 30\]](#), requiring varying degrees of problem and domain theory. Given a basis for comparison, each plan is aggregated as either an unordered set or a position-sensitive sequence, and a difference is calculated between each pair of plans. We assume a distance measure is provided by the user, and for the remainder of this paper employ a simple diversity metric based on state set distance:

$$\delta(\pi, \psi) = \frac{|\pi - \psi|}{|\pi|} \quad (6)$$

where  $\pi$  and  $\psi$  are plans, and  $\pi - \psi$  is a one-way set difference. That is,  $\delta(\pi, \psi)$  is the fraction of unique states in  $\pi$  that are not in  $\psi$  (for simplicity, a plan here represents a set of states). The diversity metric used when considering a new plan  $\pi$  in [Algorithm 1](#) is then its minimum distance from any plan in the existing plan set  $\mathcal{P}$ :

$$D(\pi, \mathcal{P}) = \min_{\psi \in \mathcal{P}} \delta(\pi, \psi) \quad (7)$$

**Theorem 3.** *The diverse plan set generated by [Algorithm 1](#) is Pareto-optimal: there exists no alternative set that dominates the returned set with respect to quality and diversity.*

*Proof.* From [Algorithm 1](#), plans are added to the final plan set in order of quality (lines 6 and 19), breaking ties using diversity (lines 20–28), and bounded by the diversity constraint  $d$  (line 17). It follows that every plan  $\pi \in \mathcal{P}$  has quality no less than any plan  $\psi \notin \mathcal{P}$  with diversity  $D(\psi, \mathcal{P}) \geq d$ , and hence there is no set outside of  $\mathcal{P}$  that can increase quality without decreasing diversity, or vice versa. Therefore, the set is Pareto-optimal with respect to quality and diversity.  $\square$

### 5.2 Diverse Multi-Armed Bandit Policies

The diverse-planning approach described in the previous section collects diverse plans from a pre-generated search tree, and as a result relies on sufficient diversity in the planning process used to build the tree. MCTS naturally biases exploration toward the most promising areas of the search space. As a result, the highest-quality branches in the tree are likely to be clustered with the optimal plan. While this side effect may be desirable for top- $k$  and top-quality planning, it may lead to starvation of branches that are distant from the optimal plan, and hence reduce the diversity of the final plan set.

To address this, we propose two approaches to increasing the diversity of the pre-generated MCTS tree. The first and most obvious solution is to increase the bias constant that balances exploration and exploitation in the selection policy used during construction of the tree (e.g.  $C$  in Eq. 1), to encourage exploration of less-promising areas of the solution space. This approach incurs a performance penalty, sacrificing tree depth for breadth and hence requiring a higher number of iterations for the MCTS tree to produce reliable value estimates. Furthermore, the additional exploration is undirected and does not necessarily bias the search toward high-quality diverse plans.

To direct tree exploration toward diverse solutions, we propose a second approach that extends the bandit policy by explicitly considering diversity. During the selection phase of the MCTS algorithm, the value of each node  $\sigma$  considered by the multi-armed bandit policy is augmented with the diversity of the corresponding plan stem  $\pi_\sigma = \{\sigma_0, \dots, \sigma\}$ . For example, the UCB1 formula (Eq. 1) may be modified to include an additional term:

$$\text{DiverseUCB1}(\sigma) = Q(\sigma) + C\sqrt{\frac{2\ln n}{n_\sigma}} + D(\pi_\sigma, \mathcal{P}) \quad (8)$$

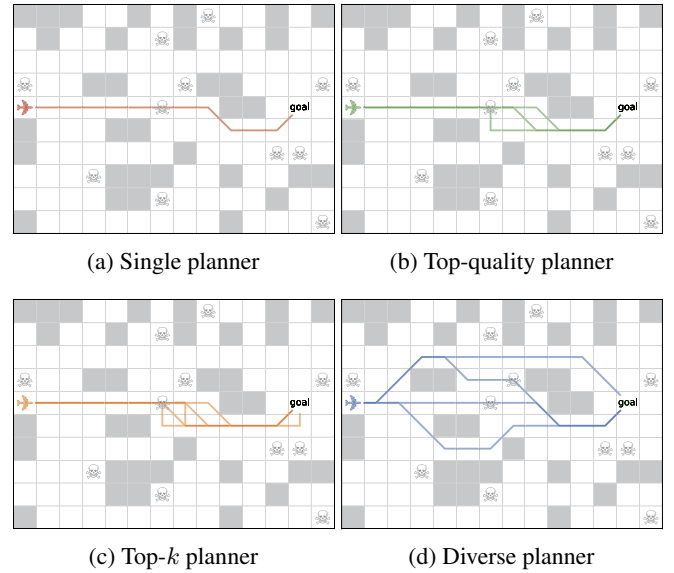
where  $D(\pi_\sigma, \mathcal{P})$  is the diversity of the plan stem  $\pi$  relative to a set of plans  $\mathcal{P}$ . Similar to Algorithm 1, plan diversity is defined relative to a bounded set of high-quality plans extracted from the partially-constructed tree, in this case updated dynamically as the tree is constructed. Exploration of this approach is left for future work.

## 6 Experimental Validation

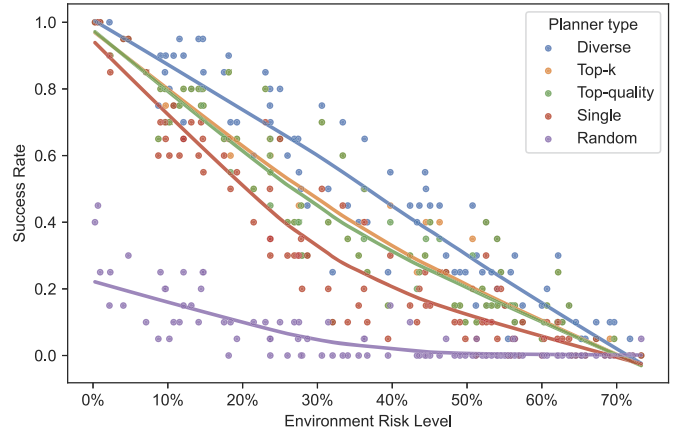
To evaluate our approach, we test the plan extraction algorithm over a path-planning problem with hidden information. This experiment is provided as a proof of concept for generating plan sets over black-box simulators, in the absence of a benchmark for such problems. In the problem scenario, a mission planner must deliver medical supplies on a battlefield using a set of delivery drones, while avoiding hostile forces. The planner is provided with a map and goal location, but the locations of enemy combatants are unknown. Enemy combatants will shoot down the drones if they pass nearby. The scenario objective is to generate a set of plans that maximises the probability of at least one drone successfully reaching the target. The source code for the simulator and experiments is available at <http://bit.ly/3WnQxme>.

Three configurations of the algorithm are considered: a top- $k$  planner, a top-quality planner, and a diverse planner. Each of the planners generates a set of plans for up to five drones (i.e.  $k \leq 5$ ). The top-quality planner restricts the plans to a minimum 80% of the quality of the optimal plan ( $q = 0.8$ ), while the diverse planner enforces a minimum 50% plan distance ( $d = 0.5$ ). For the purposes of these experiments, a relatively simple diversity metric is used, based on state set difference (Eq. 7). Two baselines are implemented for comparison: a planner that extracts the single optimum plan for each problem instance, and a random planner that extracts a set of paths selected uniformly from the search tree.

For the experiment, 2000 randomised problem instances are generated across 100 risk levels. The risk level for each instance is controlled by varying the level of hidden information, represented by the percentage of map locations occupied by enemy combatants. For each unique problem instance, an MCTS tree is constructed over 20,000 iterations ( $\approx 1.3$  seconds on an Intel Core i7-8650U processor), using a simulation model with hostile forces removed to reflect the limited information available to the planners. The plan set generated by each planner type is then executed in the true environment (Fig. 2) where



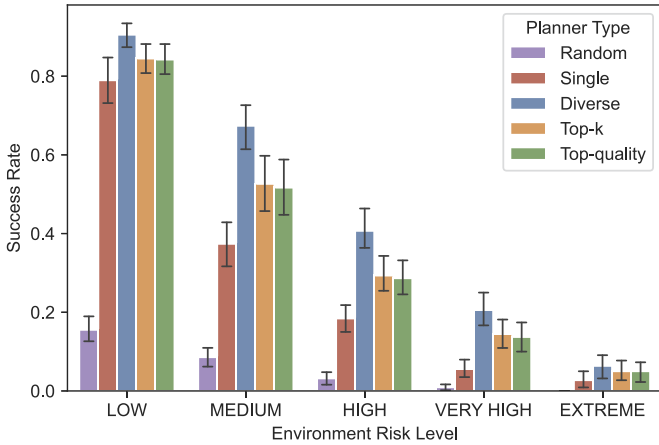
**Figure 2:** Example of a solution set generated by each planner type. In this example, with risk level 10%, only the diverse planner is successful.



**Figure 3:** Experimental results for each planner type. The environment risk level is controlled by varying the percentage of map locations occupied by enemy combatants.

the location of enemy combatants is known. The set of plans is considered successful if at least one plan in the set reaches the objective without encountering hostile forces. The average success rate for each planner is recorded over 20 replications at each risk level.

Figs. 3 and 4 capture the results. For this domain, our algorithm shows a significant improvement in success rate over each of the baseline approaches. As expected, the single planner baseline performs well when the risk level is low and the optimal path is unlikely to encounter hidden enemy combatants. As the risk level increases, performance for the single planner rapidly decreases relative to the top- $k$ , top-quality and diverse planners. At medium to high risk levels, where up to 80% of map locations are occupied by hostile forces, the diverse planner has a success rate 1.8–3.7 $\times$  that of the single planner (95% CI). While the difference in success rate between the diverse planner and all other planners is statistically significant, the difference between the top- $k$  and top-quality planners is not. This is to be expected, as the plan sets generated by these planners differ only



**Figure 4:** Comparison of mean success rate for each planner (error bars denote 95% confidence interval).

when there are fewer than  $k$  plans that meet the quality requirement  $q$ . The poor performance of the random planner baseline suggests that the benefit provided by the top- $k$ , top-quality and diverse planners is due to effective plan discovery, rather than simply due to greater coverage of the state space compared to the single planner.

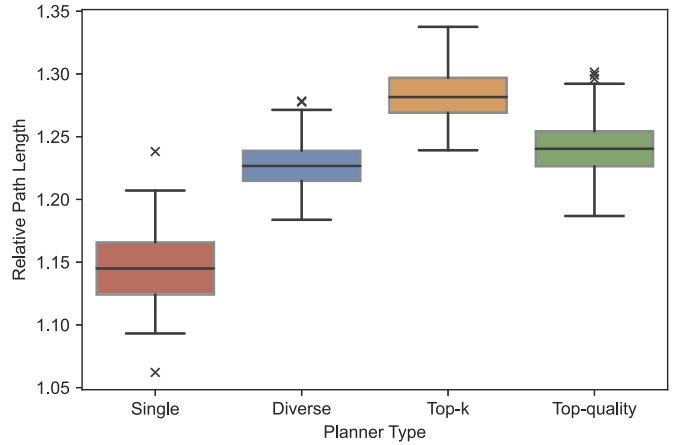
Fig. 5 demonstrates the average path cost for each planner relative to the shortest unobstructed path between the initial location and the objective. Our algorithm generates plans in order of expected return, and does not explicitly minimise path cost. However, it can be seen that the plans generated by the diverse, top-quality and top- $k$  planners add minimal additional path length over the optimal plan in this application.

Although not proposed as an optimised algorithm, we additionally consider the computational overhead of plan set extraction using our approach. Given an MCTS tree constructed over  $1.5s \pm 820ms$  (95% CI), the time required to generate a plan set during an experiment run ranges from  $312\mu s \pm 28.6\mu s$  (95% CI) for the single planner, to  $1.76ms \pm 152.4\mu s$  (95% CI) for the diverse planner. Compared to the cost of generating the MCTS tree, these costs are considered to be negligible.

## 7 Discussion

Our experimental results demonstrate that plan extraction from pre-generated search trees is a viable approach to the generation of bounded plan sets. However, empirical observation suggests that further work is needed to improve the diversity of search trees generated using MCTS algorithms. Additionally, while a state-set diversity metric was effective for the discrete path-planning domain used for our experiments, previous work has shown that more sophisticated measures are needed for general planning problems [7]. In particular, it has been observed that plans that are distant with respect to state and action sequences may be fundamentally the same when considered semantically. A possible solution may lie in sequence comparison methods used in genomics, repurposed to identify causal patterns in agent trajectories [31].

As discussed in Section 4.1, our definition of plan quality is based on expected return rather than plan cost. As a result, while the plans generated by our approach are guaranteed to have the highest expected return, they are not guaranteed to be the most efficient with respect to other measures such as path length, unless these are integrated into the reward function. Previous work has explored the trade-off between



**Figure 5:** Average path length generated by each planner, relative to the shortest unobstructed path to the objective.

efficiency and diversity [25]. Our experiments suggest that where efficiency is an important consideration, the optimal level of plan diversity may be driven by the expected risk level of the environment. If the risk is low, such as in fully-observable and deterministic domains, then single planning may be optimal. If the risk is intermediate, and the successful solution is expected to be close to the optimal plan, then top- $k$  or top-quality planning may be sufficient. If the risk level is high or unknown, then the additional cost of diverse planning may be justified by an increased probability of success.

As a post-processing method, our algorithm provides a novel capability: regeneration of alternative plan sets without replanning. Previous approaches have generated multiple plans by repeatedly solving modified versions of the planning problem, often incurring significant additional cost over single planning [14, 18, 24, 29]. In contrast, our approach operates over a pre-generated search tree, and does not require replanning to produce multiple plans. Once the tree has been generated, new sets of plans can be extracted with different quality and diversity constraints for negligible additional cost. This feature is particularly useful for planning problems requiring long processing times, as it enables the user to experiment with constraints and diversity measures without incurring the cost of solving the planning problem anew.

## 8 Conclusion

In this paper, we have introduced a novel approach to top- $k$ , top-quality and diverse planning over black-box simulators that removes the requirement for a symbolic model of the problem. We have presented a metric for the relative qualities of paths in a search tree, and an algorithm for extracting sets of plans in best-first order, along with a theoretical analysis showing that our approach finds exactly the set of top plans. Further, we have discussed strategies to improve the diversity of generated plans by modifying the multi-armed bandit selection policy used during search. Finally, we have provided an empirical evaluation to demonstrate the practicality of our approach.

There are a number of areas of continuing research, including extending our algorithm to multi-agent problems, and further developing our proposed diverse bandit policy. In future work we will explore more sophisticated diversity measures, such as causal pattern extraction, to improve the generation of meaningfully diverse sets of plans.

## References

- [1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer, 'Finite-time analysis of the multiarmed bandit problem', *Machine learning*, **47**(2), 235–256, (2002).
- [2] David Auger, 'Multiple tree for partially observable monte-carlo tree search', in *European Conference on the Applications of Evolutionary Computation*, pp. 53–62. Springer, (2011).
- [3] Lyndon Benke, Michael Papisimeon, and Tim Miller, 'Modelling Strategic Deceptive Planning in Adversarial Multi-Agent Systems', in *Deceptive AI*. Springer Nature Switzerland AG, (2021).
- [4] Yngvi Björnsson and Hilmar Finnsson, 'Cadiaplayer: A simulation-based general game player', *IEEE Transactions on Computational Intelligence and AI in Games*, **1**(1), 4–15, (2009).
- [5] Mark S. Boddy, Johnathan Gohde, Thomas Haigh, and Steven A. Harp, 'Course of Action Generation for Cyber Security Using Classical Planning', in *ICAPS*, pp. 12–21, (2005).
- [6] Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton, 'A Survey of Monte Carlo Tree Search Methods', *IEEE Transactions on Computational Intelligence and AI in Games*, **4**(1), 1–43, (March 2012).
- [7] Daniel Bryce, 'Landmark-based plan distance measures for diverse planning', in *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 24, pp. 56–64, (2014).
- [8] Tristan Cazenave and Stefan Edelkamp, 'High-Diversity Monte-Carlo Tree Search', in *German Conference on Artificial Intelligence (KI)*, Klagenfurt, (2016).
- [9] Shuiguang Deng, Bin Wu, Jianwei Yin, and Zhaohui Wu, 'Efficient planning for top-K Web service composition', *Knowledge and Information Systems*, **36**(3), 579–605, (September 2013).
- [10] Maria Fox, Alfonso Gerevini, Derek Long, and Ivan Serina, 'Plan Stability: Replanning versus Plan Repair', in *ICAPS*, volume 6, pp. 212–221, (2006).
- [11] Michael Katz and Shirin Sohrabi, 'Reshaping diverse planning', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9892–9899, (2020).
- [12] Michael Katz, Shirin Sohrabi, and Octavian Udrea, 'Top-quality planning: Finding practically useful sets of best plans', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9900–9907, (2020).
- [13] Michael Katz, Shirin Sohrabi, and Octavian Udrea, 'Bounding Quality in Diverse Planning', *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(9), 9805–9812, (June 2022).
- [14] Michael Katz, Shirin Sohrabi, Octavian Udrea, and Dominik Winterer, 'A Novel Iterative Approach to Top-k Planning', in *ICAPS18*, p. 9, (2018).
- [15] Levente Kocsis and Csaba Szepesvári, 'Bandit Based Monte-Carlo Planning', in *Proceedings of the 17th European Conference on Machine Learning*, pp. 282–293, Berlin, Heidelberg, (2006). Springer-Verlag.
- [16] Martin Masek, Chiou Peng Lam, Lyndon Benke, Luke Kelly, and Michael Papisimeon, 'Discovering Emergent Agent Behaviour with Evolutionary Finite State Machines', in *PRIMA 2018: Principles and Practice of Multi-Agent Systems*, pp. 19–34, (2018).
- [17] Karen L. Myers and Thomas J. Lee, 'Generating qualitatively different plans through metatheoretic biases', in *AAAI/IAAI*, pp. 570–576, (1999).
- [18] Tuan Anh Nguyen, Minh Do, Alfonso Emilio Gerevini, Ivan Serina, Biplav Srivastava, and Subbarao Kambhampati, 'Generating diverse plans to handle unknown and partially known user preferences', *Artificial Intelligence*, **190**, 1–31, (2012).
- [19] Michael Papisimeon and Lyndon Benke, 'Multi-Agent Simulation for AI Behaviour Discovery in Operations Research', in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pp. 72–85. Springer, (2021).
- [20] Michael Papisimeon, Lyndon Benke, Richard Brain, and Lilia Finkelshtein, 'Multiagent Simulation of Adversarial Socio-Technical Systems', in *International Conference on Autonomous Agents and Multiagent Systems*, p. 9, Stockholm, (2018).
- [21] Giambattista Parascandolo, Lars Buesing, Josh Merel, Leonard Hasenclever, John Aslanides, Jessica B. Hamrick, Nicolas Heess, Alexander Neitz, and Theophane Weber, 'Divide-and-conquer monte carlo tree search for goal-directed planning', *arXiv preprint arXiv:2004.11410*, (2020).
- [22] Diego Perez, Spyridon Samothrakis, and Simon Lucas, 'Online and offline learning in multi-objective Monte Carlo Tree Search', in *2013 IEEE Conference on Computational Intelligence in Games (CIG)*, pp. 1–8. IEEE, (2013).
- [23] Miquel Ramirez, Michael Papisimeon, Nir Lipovetzky, Lyndon Benke, Tim Miller, Adrian R Pearce, Enrico Scala, and Mohammad Zamani, 'Integrated Hybrid Planning and Programmed Control for Real-Time UAV Maneuvering', in *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, (2018).
- [24] Anton Riabov, Shirin Sohrabi, and Octavian Udrea, 'New algorithms for the top-k planning problem', in *Proceedings of the Scheduling and Planning Applications Workshop (SPARK) at the 24th International Conference on Automated Planning and Scheduling (ICAPS)*, pp. 10–16, (2014).
- [25] Mark Roberts, Adele E. Howe, and Indrajit Ray, 'Evaluating diversity in classical planning', in *Twenty-Fourth International Conference on Automated Planning and Scheduling*, (2014).
- [26] Maarten PD Schadd, Mark HM Winands, HJVD Herik, Guillaume MJ-B. Chaslot, and Jos WHM Uiterwijk, 'Single-player monte-carlo tree search', in *International Conference on Computers and Games*, pp. 1–12. Springer, (2008).
- [27] Shirin Sohrabi, Anton Riabov, Michael Katz, and Octavian Udrea, 'An AI planning solution to scenario generation for enterprise risk management', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, (2018).
- [28] Shirin Sohrabi, Anton V. Riabov, and Octavian Udrea, 'State projection via AI planning', in *Thirty-First AAAI Conference on Artificial Intelligence*, (2017).
- [29] David Speck, Robert Mattmüller, and Bernhard Nebel, 'Symbolic Top-k Planning', *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(06), 9967–9974, (April 2020).
- [30] Biplav Srivastava, Tuan Anh Nguyen, Alfonso Gerevini, Subbarao Kambhampati, Minh Binh Do, and Ivan Serina, 'Domain Independent Approaches for Finding Diverse Plans', in *IJCAI*, pp. 2016–2022, (2007).
- [31] Archana Vadakattu, Michelle Blom, and Adrian R. Pearce, 'Strategy Extraction in Single-Agent Games', in *Proceedings of the Adaptive and Learning Agents Workshop (ALA 2023)*, London, (2023).
- [32] Satya Gautam Vadlamudi and Subbarao Kambhampati, 'A combinatorial search perspective on diverse solution generation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, (2016).
- [33] Terry Zimmerman and Subbarao Kambhampati, 'Generating parallel plans satisfying multiple criteria in anytime fashion', in *Proc. Workshop on Planning and Scheduling with Multiple Criteria (AIPS-2002)*, pp. 56–66, (2002).