

Motion Planning as Online Learning: A Multi-Armed Bandit Approach to Kinodynamic Sampling-Based Planning

Marco Faroni and Dmitry Berenson

Abstract—Kinodynamic motion planners allow robots to perform complex manipulation tasks under dynamics constraints or with black-box models. However, they struggle to find high-quality solutions, especially when a steering function is unavailable. This paper presents a novel approach that adaptively biases the sampling distribution to improve the planner’s performance. The key contribution is to formulate the sampling bias problem as a non-stationary multi-armed bandit problem, where the arms of the bandit correspond to sets of possible transitions. High-reward regions are identified by clustering transitions from sequential runs of kinodynamic RRT and a bandit algorithm decides what region to sample at each timestep. The paper demonstrates the approach on several simulated examples as well as a 7-degree-of-freedom manipulation task with dynamics uncertainty, suggesting that the approach finds better solutions faster and leads to a higher success rate in execution.

Index Terms—Motion and Path Planning, Integrated Planning and Learning, Planning under Uncertainty.

I. INTRODUCTION

PHYSICS simulators and deep-learning models allow robots to reason about complex manipulation tasks such as manipulation of deformable objects [1]–[3], liquid handling [4], [5], and contact-rich manipulation [6], [7]. Kinodynamic motion planning can find a sequence of controls that brings such systems to a desired state. For example, consider the tabletop scenario in Fig. 1: A compliant manipulator moves a heavy object across the table; because of the payload and the compliant control, the trajectory execution will deviate from the planned path, possibly causing unexpected collisions. Suppose we have a function that maps the robot state to an estimate of the end-effector Cartesian error; we can avoid unexpected collisions in execution by finding a trajectory that minimizes such a function.

Sampling-based planners, such as rapidly exploring random trees (RRT) [8] are widely used in robotics because of their effectiveness in high-dimensional problems. Despite the fact that asymptotically optimal algorithms [9], [10] ensure convergence to the optimal solution for an infinite number of iterations, their convergence rate is often slow for practical



Fig. 1: A 7-degree-of-freedom manipulator carrying a weight in a cluttered environment with uncertain tracking control.

applications. This issue holds especially if a steering function (*i.e.*, a function that connects two given states) is not available or computationally expensive, which is often the case for learned or simulated dynamics models [11].

Planning performance can be improved by biasing the sampling distribution, *e.g.*, to find a solution faster [12] or to reduce the cost of the solution [13]. RRT-like planners with biased sampling extend the tree by sampling a target state from a non-uniform distribution. The biased distribution can be learned offline [14]–[16] or adapted online based on previous iterations [17], [18]. We approach the problem of biased sampling from an online-learning perspective. That is, we consider biased sampling as a sequential decision-making process where each transition added to the tree is associated with a reward (dependent on the cost function). Then, we decide what transition to sample at the next iteration based on the rewards estimated from previous timesteps. In particular, this paper proposes an online learning approach to biasing samples in a kinodynamic RRT.

We use Multi-Armed Bandit (MAB) algorithms to shape the sampling probability distribution iteratively. The proposed method is illustrated in Fig. 2. Our approach builds on the asymptotically optimal framework AO-RRT [19], which runs kinodynamic RRT multiple times. Every time a new solution is found, transitions are clustered based on their reward and spatial position, and a non-stationary bandit algorithm biases samples based on the expected reward of each region.

The contributions of this paper are:

- An online learning approach to biasing samples in a motion planner that formulates the bias problem as a non-stationary Multi-Armed Bandit problem and trades off the

Manuscript received: March, 24, 2023; Revised May, 22, 2023; Accepted July, 29, 2023. This paper was recommended for publication by Editor Hanna Kurniawati upon evaluation of the Associate Editor and Reviewers’ comments.

This work was supported in part by Toyota Research Institute, the Office of Naval Research Grant N00014-21-1-2118, and NSF grants IIS-1750489, IIS-2113401, and IIS-2220876.

The authors are with the Robotics Department, University of Michigan, Ann Arbor, MI 48109, United States. {mfaroni; dmitryb}@umich.edu
Digital Object Identifier (DOI): see top of this page.

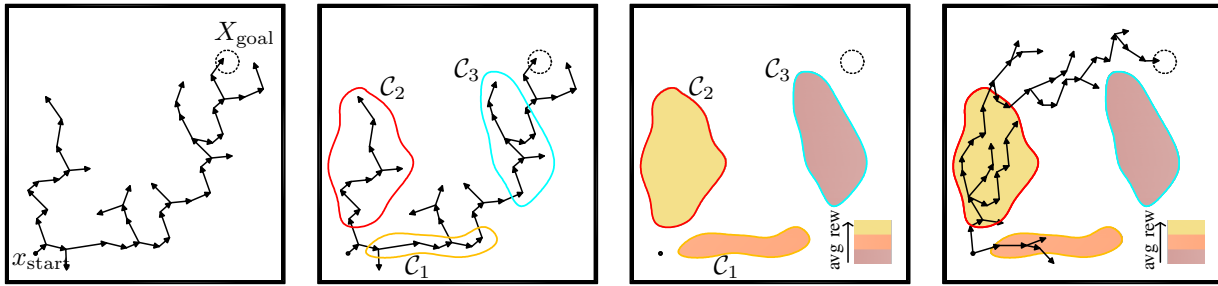


Fig. 2: Sketch of the proposed method. From left to right: (i) a randomized planner searches for a solution without bias; (ii) each time it finds a new solution, transitions are clustered based on their reward and position; (iii) each cluster is associated with an estimated reward; (iv) a non-stationary bandit algorithm biases sampling based on the regions' expected reward.

exploration and exploitation of high-reward regions based on the reward observed at previous timesteps.

- A kinodynamic planner that does not rely on a steering function and uses the proposed MAB approach to find better solutions faster.
- A demonstration of the proposed method on a 7-degree-of-freedom manipulation problem (Fig. 1), showing that the proposed approach improves the solution cost and, in the scenario at hand, leads to a higher success rate in execution. An empirical regret analysis of different sampling strategies also suggests that a better solution cost coincides with a lower cumulative regret.

II. RELATED WORK

In the last decade, sampling-based motion planning has seen a focus shift from finding feasible solutions to finding high-quality ones, especially after [9] provided conditions for asymptotic optimality for planners such as RRT* and PRM*. These conditions include the availability of a steering function, making these planners unsuitable for black-box dynamics models. To overcome this issue, variants of RRT* have been proposed by approximating the steering function [10], [20], [21], but these approaches are only suitable for a limited class of systems. Other works researched how to guarantee asymptotic optimality without a steering function [11], [19]. In particular, [19] proposed an asymptotically optimal meta-planning algorithm based on multiple runs of RRT in an augmented state-cost space, whereas [11] combined biased node expansion with pruning to refine an initial solution. Attempts to improve the convergence rate of these methods include using a heuristic to bias node expansion [22], [23], pruning and re-usage of previous edges [24], and building a PRM-like roadmap of edges offline [25].

In this work, we focus on improving the solution quality of kinodynamic planners via adaptive sampling. Most sampling-based planners use a uniform sampling distribution; however, biasing the sampling has been a common strategy to improve the solution cost [14]–[16], [26]–[29]. The sampling bias is often tailored to the specific problem manually [30] or using machine learning techniques [14]–[16]. Other works leverage the knowledge gathered at the previous iterations to bias the sampling at the current one. For example, they use local bias to overcome narrow passages [31], switch between global and

local sampling to find a solution faster [32], or quickly refine the current solution [17], [33]. The works above are designed for holonomic motion planners. Informed sampling is another biased sampling technique that uses cost heuristics to discard regions with a null probability of improving previous solutions [13]. How to derive or approximate such heuristics for non-trivial cost functions is an open research question [33], [34].

Our approach leverages MAB algorithms to choose the sampling bias online. MAB is an online learning technique used for repeated decision-making under uncertainty. An MAB problem is defined by a set of actions (*arms*) associated with a belief of their reward function. At each iteration, an agent chooses an arm and updates the reward estimates according to its realized reward. MAB algorithms are typically characterized by their regret, *i.e.*, how much worse they perform compared to a strategy that picks the best arm at each iteration. Different approaches (and regret bounds) have been derived based on different assumptions on the reward distribution. Common algorithms are UCB-1 [35] and Thompson Sampling [36] for constant reward distributions and their variants for non-stationary rewards [37]. A comprehensive overview of the topic can be found in [38], [39].

Recent works applied the MAB framework to motion planning, aiming to automatically balance the trade-off between exploration and exploitation [40]–[42]. To the best of our knowledge, MAB was applied to sampling-based planning only to overcome narrow passages in bi-directional search [31], [43]. They consider trees as bandits' arms and decide which tree to expand depending on the estimated probability of a successful expansion. However, this technique is specific to narrow passages and works only with multiple trees and the availability of a steering function.

III. PROBLEM STATEMENT

Consider a dynamics model $\dot{x} = f(x, u)$, $x \in X$, $u \in U$, where X and U are the state and the control spaces and $X_{\text{free}} \subseteq X$ is the set of valid states. Solving a kinodynamic motion planning problem means finding a control function $\gamma : [0, T] \in U$ that induces a trajectory $\sigma : [0, T] \rightarrow X$ such that $\sigma(0) = x_{\text{start}}$ and $\sigma(T) \in X_{\text{goal}}$, and $\sigma(t) \in X_{\text{free}} \forall t \in [0, T]$. In optimal motion planning, we also aim to minimize a Lipschitz continuous cost function $c(\sigma)$. In this work, we restrict γ to be a staircase function defined by a sequence of

controls and control durations, $\{(u_i, d_i)\}$, so that $\sum_i d_i = T$ and $\gamma(t) = u_j$ with $j \in \mathbb{Z}^+ | \sum_{i=0}^j d_i \leq t_j < \sum_{i=0}^{j+1} d_i$.

Sampling-based planners such as kinodynamic RRT solve this problem by randomly sampling a target state, x_{trg} , retrieving the closest node on the tree, x_p , and expanding this node by forward propagation. We denote by $\tau = (x_p, u, d, x_c, x_{\text{trg}})$ the transition from x_p to x_c induced by u and d ; note that τ also stores the target x_{trg} from which τ originated.

The search strategy above is a sequential decision-making process, where the planner has to choose which node to expand next and in what direction. We consider this process from an online-learning perspective, where each new transition is associated with a reward $r = \rho(\tau)$, where $0 \leq \rho \leq 1^1$. The planner observes the reward of a transition after each iteration and chooses the next transition to maximize the total reward over K , possibly infinite, timesteps. MAB is a framework to address this kind of problems. In the MAB settings, an agent can choose among M actions (the *arms*) for K , possibly infinite, rounds. The goal is to maximize the cumulative reward, assuming each action yields a reward from an unknown distribution and the agent can only observe the reward of the selected action.

We frame the problem of choosing the next transition in a kinodynamic RRT as an MAB problem where the arms are sets of transitions. Selecting an arm then corresponds to sampling a transition τ from a certain set and using it to extend the tree. Our goal is to use MAB to improve the path cost over iterative runs of kinodynamic RRT by trading off the exploitation of high-reward regions (according to the reward obtained at previous runs) and the exploration of transitions with a highly uncertain reward (*i.e.*, less-explored regions).

IV. METHOD

Our approach can be summarized as follows: (a) we iteratively re-plan with kinodynamic RRT, using MAB to select regions for sampling transitions; (b) every time we find a new solution trajectory, we identify high-reward regions by clustering previous transitions; (c) our MAB method estimates the non-stationary reward distribution as we plan during a run of kinodynamic RRT. This section describes the planning framework and methods for clustering and biased sampling.

A. Planning framework

We propose a kinodynamic planner based on the AO-RRT meta-planning paradigm [19], which runs instances of RRT sequentially and keeps the best solution so far. Alg. 1 summarizes the proposed algorithm (the differences with respect to AO-RRT are in red). At each iteration, `sampleAndPropagate` (line 5) samples a transition. The new transition is added to the tree (line 7). If the goal condition is satisfied, the solution is retrieved by `retracePath` (lines 8–9) and RRT is reset (line 15). After K iterations, Alg. 1 returns the best solution so far, σ_{best} , and its cost, c_{best} . We embed our adaptive sampling strategy in AO-RRT in two steps: (a) clustering and (b) bandit-based sampling.

¹Note that $\rho(\tau)$ should be inversely proportional to the cost of τ ; however, its definition may be problem-specific. We propose examples of the reward function in Sec. V and VI.

Algorithm 1: MAB-RRT

Input: $x_{\text{start}}, X_{\text{goal}}, c(\cdot), \rho(\cdot), K > 0$
Output: $\sigma_{\text{best}}, c_{\text{best}}$

- 1 $\sigma_{\text{best}} \leftarrow \emptyset, c_{\text{best}} \leftarrow +\infty, \mathcal{T} \leftarrow \emptyset$
- 2 $\mathcal{G} \leftarrow \emptyset, \mathcal{G}.add(x_{\text{start}})$
- 3 **initializeBandits**(0,0)
- 4 **for** k **in** $1, \dots, K$ **do**
- 5 $\tau \leftarrow \text{sampleAndPropagate}(\mathcal{C}, \mathcal{G})$
- 6 **updateBanditArms**($\rho(\tau)$)
- 7 $\mathcal{G}.add(\tau)$
- 8 **if** $x_{\text{next}} \in X_{\text{goal}}$ **then**
- 9 $\sigma \leftarrow \text{retracePath}(\tau)$
- 10 **if** $c(\sigma) < c_{\text{best}}$ **then**
- 11 $\sigma_{\text{best}} \leftarrow \sigma, c_{\text{best}} \leftarrow c(\sigma)$
- 12 $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{G}$
- 13 $\mathcal{C}, \mathcal{R} \leftarrow \text{clustering}(\mathcal{T})$
- 14 **initializeBandits**($\mathcal{R}_1, \dots, \mathcal{R}_M$)
- 15 $\mathcal{G} \leftarrow \emptyset, \mathcal{G}.add(x_{\text{start}})$
- 16 **return** $\sigma_{\text{best}}, c_{\text{best}}$

1) *Clustering*: At the end of each RRT run, we add all the transitions to the set of all previous transitions, \mathcal{T} , and cluster them into a set of clusters \mathcal{C} (lines 12 and 13) by using HDBSCAN [44]. Then, we associate each cluster \mathcal{C}_i with a bandit's arm and use each cluster's average reward, \mathcal{R}_i , to initialize the bandit's expected rewards (line 14).

2) *Bandit-based sampling*: An MAB algorithm in `sampleAndPropagate` decides whether to sample the next transition τ from a cluster, the uniform distribution, or the goal set. Then, we extend the tree from $\tau.x_p$ by forward propagation. After the extension, the MAB updates the arms' rewards according to the reward realized by the new transition (line 6). Note that the reward function is non-stationary with respect to the tree, as detailed in Sec.IV-C and IV-D, because the transition reward depends on the current state of the tree.

The next sections detail the clustering and sampling phases. Tuning guidelines are given in Appendix A.

B. Online learning of high-reward regions

We aim to find groups of transitions that constitute high-reward regions. Because full state-space coverage is often intractable, we do not try to create a partition of the entire space of possible transitions, instead focusing on the transitions obtained from previous iterations. Given a set of transitions \mathcal{T} , we cluster them according to their reward and spatial distribution through the distance function:

$$d(\tau_1, \tau_2) = \|\tau_1.x_p - \tau_2.x_p\| + \|\tau_1.x_c - \tau_2.x_c\| + \lambda \|\rho(\tau_1) - \rho(\tau_2)\| \quad (1)$$

where $\lambda > 0$, and $\rho(\tau_j)$ is the reward of τ_j . Although any clustering techniques could be used, we use HDBSCAN because of its effectiveness at identifying irregular clusters and ease of tuning (see also Appendix A). Hence, we obtain a set of clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_N\}$ and each cluster is associated with an average reward $\mathcal{R}_i = 1/|c_i| \sum_{\tau_j \in \mathcal{C}_i} \rho(\tau_j)$. These clusters are subsets of transitions, which will be used to bias sampling.

Algorithm 2: sampleAndPropagate

Input: \mathcal{C}, \mathcal{G}
Output: $(x_p, u, d, x_c, x_{\text{trg}})$

```

1  $x_p \leftarrow \text{NULL}, x_{\text{trg}} \leftarrow \text{NULL}$ 
2 while  $x_p = \text{NULL}$  or  $x_{\text{trg}} = \text{NULL}$  do
3    $i \leftarrow \text{selectNextBanditArm}()$ 
4   if  $i = 1$  then
5      $x_{\text{trg}} \leftarrow \text{unif}(X)$ 
6      $x_p \leftarrow \mathcal{G}.\text{nearest}(x_{\text{trg}})$ 
7   else if  $i = 2$  then
8      $x_{\text{trg}} \leftarrow \text{unif}(X_{\text{goal}})$ 
9      $x_p \leftarrow \mathcal{G}.\text{nearest}(x_{\text{trg}})$ 
10  else
11     $x_p, x_{\text{trg}} \leftarrow \text{sampleCluster}(\mathcal{C}_i, \mathcal{G})$ 
12    if  $x_p = \text{NULL}$  or  $x_{\text{trg}} = \text{NULL}$  then
13       $\text{updateBanditArms}(0)$ 
14  $x_c, u, d \leftarrow \text{sampleTo}(x_p, x_{\text{trg}}, i)$ 
15 return  $(x_p, u, d, x_c, x_{\text{trg}})$ 

```

Algorithm 3: sampleCluster

Input: $\mathcal{C}_i, \mathcal{G}$
Output: x_p, x_{trg}
Parameters: $K > 0, \delta_1, \delta_2 > 0, \delta_3 > \delta_2, w > 0$

```

1  $x_p \leftarrow \text{NULL}, x_{\text{trg}} \leftarrow \text{NULL}$ 
2 for  $k$  in  $1, \dots, K$  do
3    $\tau_c \leftarrow \text{unif}(\mathcal{C}_i)$ 
4    $\tau_c.x_p = \tau_c.x_p + \text{unif}([-w, w])$ 
5    $\tau_c.x_{\text{trg}} = \tau_c.x_{\text{trg}} + \text{unif}([-w, w])$ 
6   if  $\|\tau_c.x_{\text{trg}} - \mathcal{G}.\text{nearest}(\tau_c.x_{\text{trg}})\| < \delta_1$  then
7     continue
8   if  $\|\tau_c.x_p - \mathcal{G}.\text{nearest}(\tau_c.x_p)\| < \delta_2$  then
9      $x_p \leftarrow \mathcal{G}.\text{nearest}(\tau_c.x_p)$ 
10     $x_{\text{trg}} \leftarrow \tau_c.x_{\text{trg}}$ 
11    break
12  if  $\|\tau_c.x_p - \mathcal{G}.\text{nearest}(\tau_c.x_p)\| < \delta_3$  then
13     $x_p \leftarrow \mathcal{G}.\text{nearest}(\tau_c.x_p)$ 
14     $x_{\text{trg}} \leftarrow \tau_c.x_p$ 
15 return  $x_p, x_{\text{trg}}$ 

```

C. Adaptive sampling of high-reward regions

We bias the probability of sampling cluster \mathcal{C}_i according to its expected reward. We model the problem of finding the optimal sampling bias as an MAB problem, where each \mathcal{C}_i is an arm and \mathcal{R}_i is its initial reward. We also consider uniform and goal sampling as arms of the MAB problem. Therefore, we define a set of M arms, $\{a_1, \dots, a_M\}$, with $M = N + 2$:

$$a_i = \begin{cases} \text{uniform sampling over } X & \text{if } i = 1 \\ \text{uniform sampling over } X_{\text{goal}} & \text{if } i = 2 \\ \text{sampling } \mathcal{C}_{i-2} & \text{if } i \geq 3 \end{cases} \quad (2)$$

MAB selects where to sample the next transition from (line 3 of Alg. 2). If the MAB selects the first or second arm, $\tau.x_{\text{trg}}$ is a random state or a goal state, respectively, while $\tau.x_p$ is the node of the tree closest to $\tau.x_{\text{trg}}$ (lines 4–9 of Alg. 2). For all other arms, the algorithm tries to sample a transition from the selected cluster (Alg. 3). Cluster sampling depends on the search tree because:

- i We want to avoid over-sampling regions that the current tree has already explored; thus, we discard a candidate τ_c if $\tau_c.x_{\text{trg}}$ is too close to the current tree;
- ii We select τ_c only if its pre-conditions are met; *i.e.*, if $\tau_c.x_p$ is close enough to the current search tree;
- iii If it is impossible to sample a transition whose pre-conditions are met, we try to extend the tree in such a way as to meet the pre-conditions in future timesteps; *i.e.*, we set $\tau_c.x_p$ as the target of the new sample τ .

Alg. 3 implements these considerations. It randomly draws candidate transitions, τ_c , from a cluster and perturbs them until $\tau_c.x_{\text{trg}}$ is further than δ_1 from \mathcal{G} and $\tau_c.x_p$ is closer than δ_2 to \mathcal{G} (lines 6–11). If it is impossible to satisfy the second condition (*i.e.*, \mathcal{G} is too far from the selected cluster), we try to expand the tree toward that cluster by looking for a transition whose parent is closer than $\delta_3 > \delta_2$ to \mathcal{G} (line 12) by selecting its parent as the new target. If we could not draw

a valid transition from the selected cluster, we discourage its future selection by assigning a reward equal to zero (line 13). Finally, `sampleTo` (line 14 of Alg. 2) generates a transition by propagating a random action for a random duration if $i \leq 2$. If $i > 2$, `sampleTo` tries to extend toward x_{trg} by sampling N_p random controls, $u \in U$, and durations, $d \in [0, T_p]$, with $T_p > 0$, and selecting the closest transition to x_{trg} .

D. Non-stationary rewards

We update the estimated reward distributions when we add a transition to the tree (line 6 of Alg. 1) and if we could not sample a valid candidate transition from the selected cluster \mathcal{C}_i (line 13 of Alg. 2). In the first case, we use the reward $\rho(\tau)$ realized by the transition; in the second case, we use a reward equal to zero to discourage sampling \mathcal{C}_i if it does not satisfy conditions i, ii, or iii from Sec. IV-C. In both cases, the reward realized by a transition depends on the tree state. We model the variability of the reward with respect to the tree state as a non-stationary MAB problem, where the reward distribution can vary over iterations. Standard MAB algorithms such as UCB-1 [35] perform poorly under these conditions because they adapt too slowly to the reward changes [37]. We therefore use a non-stationary bandit algorithm, which accounts for shifts in the reward distribution. Specifically, we use the Kalman Filter-Based solution for Non-stationary Multi-Arm Bandit (KF-MANB) algorithm [37]. KF-MANB models each arm's reward as a normal distribution. At each iteration, it selects the next arm via Thompson Sampling [36]. When it observes the reward, KF-MANB updates the estimated distributions using a Kalman Filter update rule, which allows for tracking non-stationary rewards over time.

E. Completeness and optimality

Note that not all variants of RRT are probabilistically complete [45]. Assumed that (i) the dynamics system is

Lipschitz continuous, and (ii) there exists a robust solution with clearance $\delta > 0$, kinodynamic RRT is probabilistically complete if it extends the tree by forward propagating random controls $u \in U$ for a random duration $d \in [0, T_p]$, with $T_p > 0$ [24]. This is true for our method because `sampleTo` in Alg. 2 chooses random controls and durations when the first two arms (*i.e.*, uniform and goal sampling) are chosen by the MAB algorithm. Note that these arms have a non-zero probability of being selected at each iteration (like any other arm of the MAB). Therefore, under the assumptions that $f(x, u)$ from Sec. III is Lipschitz continuous and there exists a solution with clearance $\delta > 0$, each run of RRT in MAB-RRT is probabilistically complete.

As for asymptotic optimality, [19] proved that AO-RRT is asymptotically optimal if the underlying RRT is *well-behaved* in the augmented state-cost space. In a refined version of the proof, [46] argues that well-behavedness holds if the dynamics system and the cost function derivative are Lipschitz continuous, and the optimal trajectory is robust with clearance $\delta > 0$, proving the asymptotic optimality of a single-tree version of AO-RRT. Because Alg. 1 uses a multi-tree implementation and runs RRT in the state space, we cannot derive asymptotic optimality directly from [46]. Nonetheless, results in Sec. V and VI suggest that the solution cost decreases consistently with iterations. We therefore leave the formal analysis of MAB-RRT asymptotic optimality as future work.

V. SIMULATION RESULTS

This section shows that our approach improves the solution cost faster than AO-RRT and yields smaller cumulative regret with 2D problems. We consider a single integrator $\dot{x} = u$, $X = [0, 1]^2$, $U \in [0.5, 0.5]^2$, and the five scenarios in Fig. 3. We consider the reward function $\rho(\tau) = 0.5(\rho_x(\tau.x_p) + \rho_x(\tau.x_c))$ where $\rho_x \in [0, 1]$ as in Fig. 3 and the cost function $c(\sigma) = \sum_{\tau \in \sigma} (1 - \rho(\tau)) \|\tau.x_p - \tau.x_c\|$.

The scenarios serve as illustrative examples of problems with different features. For example, in Scenario A, the optimal solution should take a long path through a narrow passage, while Scenarios C and D are examples of “trap” problems, where the high-reward region leads to a dead end. Scenario E combines these issues into a more complex problem.

A. Cost analysis

We compare our MAB-RRT-KFMANB with AO-RRT [19] and other variants of MAB-RRT using UCB-1 [35] and Thompson Sampling [36]. Fig. 4 shows the average cost trends for 30 repetitions. Except for Scenario C, MAB-RRT-KFMANB improves the solution significantly faster than AO-RRT, suggesting that the online bias learning drives the search to more promising regions. In Scenario C, MAB-RRT-KFMANB has a slightly worse convergence rate because the high-reward region (yellow in Fig. 3) leads to a dead end (notice that the optimal path mainly lies in the low-reward region). MAB-RRT-KFMANB outperforms AO-RRT even in Scenario D, where the high-reward region leads to a dead end: after an initial exploration of the high-reward, MAB-RRT spots the medium-reward region and quickly improve the solution

Algorithm 4: Regret computation

Input: tree \mathcal{G} , iteration k
Output: expected regret $E[R_{\text{strategy}}]$ for all strategies

- 1 $\text{strategies} = \{\text{kfmanb}, \text{ucb1}, \text{TS}, \text{random}, \text{astar}\}$
- 2 **for** strategy **in** strategies **do**
- 3 **if** strategy **in** $\{\text{kfmanb}, \text{ucb1}, \text{TS}\}$ **then**
- 4 draw a batch of points for each arm of the MAB and a batch of points using Alg. 2
- 5 **if** $\text{strategy} = \text{random}$ **then**
- 6 draw a batch of points using the random sampling strategy
- 7 **if** $\text{strategy} = \text{astar}$ **then**
- 8 draw a batch of points using the `astar` sampling strategy
- 9 For all batches of points, generate the corresponding transitions with respect to \mathcal{G} ;
- 10 For all batches of transitions, compute the expected reward $(\bar{r}_{\text{rnd}}, \bar{r}_{A^*}, \bar{r}_{\text{kfmanb},i}, \bar{r}_{\text{ucb1},i}, \bar{r}_{\text{TS},i} \forall i=1\dots M)$
- 11 Compute the best expected reward $E[r^*(k)]$ as the maximum of the average rewards of all batches;
- 12 **for** strategy **in** strategies **do**
- 13 regret $E[R_{\text{strategy}}] = E[r^*(k)] - \bar{r}_{\text{strategy}}$;

cost. As expected, MAB-RRT-KFMANB outperforms the stationary variants. Overall, MAB-RRT-UCB1 and MAB-RRT-TS perform comparably to AO-RRT, showing the importance of the non-stationary MAB to account for the changing reward.

B. Regret analysis

A standard metric to evaluate MAB is regret, *i.e.*, the difference between the reward one would have obtained by sampling the best action and the reward realized by the chosen action at iteration k . We can define the expected regret of a sampling strategy over a (possibly infinite) horizon K as

$$E[R(K)] = \sum_{k=1}^K E[r_k^*] - E[r_k] \quad (3)$$

where r_k^* and r_k are the rewards of the best sampling strategy and the chosen sampling strategy at iteration k , respectively. We evaluate the regret of the following sampling strategies:

- `kf-manb`, `ucb1`, `TS`: our method as described in Alg. 2 using KF-MANB, UCB-1, and Thompson Sampling;
- `random`: uniform sampling over X , as in standard RRT;
- `astar`: it mimics the A^* search strategy; the control space is discretized as $\bar{U} = \{-0.5, -0.25, 0, 0.25, 0.5\}^2$ and, at each iteration, the node with the lower estimated cost and with unexplored children is expanded. We use $h(x_1, x_2) = (1 - \max_x \rho_x(x)) \|x_1 - x_2\|$ as admissible heuristic, where $\max_x \rho_x(x) = 0.99$ according to Fig. 3.

The regret computation is described in Alg. 4, which runs at the beginning of each iteration k in Alg. 1. It samples a batch of points from each arm of each sampling strategy (lines 2–9) and computes the average reward of each one (line 10). The best arm reward is the maximum average reward across all

batches (line 11) and is used to compute the regret of each strategy (line 13). Note that the regret comparison requires the sampling strategies to be evaluated at each iteration given the same tree. For this reason, we grow the tree by using samples from `kfmanb` to obtain the tree at the next timestep.

Results are in Fig. 5. Interestingly, `astar` yields very small regret in four out of five scenarios; `kfmanb` yields significantly lower regret than `random`, `ucbl` and `TS`, suggesting a correlation between the cost and the regret for all scenarios.

C. Discussion

The results confirm that the proposed sampling technique reduces the cumulative regret compared to uniform sampling. The regret reduction coincides with smaller trajectory costs compared to AO-RRT. Interestingly, A^* accumulates the lowest regret because its heuristic search tends to expand nodes with low cost-to-come, focusing more on high-reward regions. Unfortunately, A^* is inefficient for high-dimensional problems, making the proposed method attractive from the perspective of high-dimensional kinodynamic planning.

Computationally, MAB-RRT differs from AO-RRT in the MAB-based sampling and the clustering. Given the same number of control propagations, N_p , in the `sampleTo` function, AO-RRT and MAB-RRT perform the same number of collision checks and forward dynamics propagations, leading to similar average iteration time (0.26 ms and 0.29 ms, respectively, with $N_p = 100$). Clustering time grows with the number of transitions and the state-space dimensionality. With an off-the-shelf Python implementation of HDBSCAN [47] we observed an almost-linear clustering time between 2 and 80 ms for 100 and 5000 transitions. Because of clustering and computational overheads, the average total planning time was 0.60 s for MAB-RRT and 0.43 s for AO-RRT. Note that such difference is expected to become thinner for more complex scenarios, where the iteration time is predominant compared to the clustering time.

VI. EXPERIMENTS

We demonstrate our approach on a manipulation task with uncertain dynamics and show that it finds trajectories with a lower cost and a higher success rate in execution. We consider a tabletop application where a Kuka IIWA 7 arm moves a dumbbell (6 kg) across the table, as in Fig. 1. Note that the payload (gripper + dumbbell) is around 9 kg, exceeding the maximum payload of the robot. We implemented all the planners with OMPL [48] in a ROS/Gazebo [49] simulation environment. The robot is controlled in joint-space impedance mode, which makes it compliant with the environment, yet causes large tracking errors with large payloads. The state and control spaces are joint position and velocity, respectively. We devise a cost function proportional to the robot Cartesian-space tracking error and dependent on the robot joint states:

$$c(\sigma) = \sum_{\tau \in \sigma} \hat{e}(\tau) \|\tau.x_p - \tau.x_c\| \text{ and } \rho(\tau) = 1 - \hat{e}(\tau) \quad (4)$$

where $\hat{e} \in [0, 1]$ is proportional to the end-effector Cartesian-space error (see Appendix B for its derivation). By minimizing

c , the planner is expected to find trajectories that avoid large tracking errors, thus reducing the risk of unexpected collisions.

We compare MAB-RRT (with KF-MANB) and AO-RRT over three queries repeated 30 times. Fig. 6a and 6b show the path cost and the execution success rate on the real robot (*i.e.*, the percentage of runs that reached the goal without collisions). To compare different queries, costs were normalized with respect to the best cost found for the corresponding query. Fig. 6a shows that MAB-RRT's cost after 1500 iterations is significantly smaller than that of AO-RRT (-25%). Intuitively, a good solution avoids configurations where the payload causes a large deviation from the path. This translates into a higher success rate when the trajectory is executed in the real-world (+65%). As shown in Fig. 7 and in the attached video, the paths found by MAB-RRT are more likely to avoid obstacles by retracting the arm (top images). On the contrary, the path computed by AO-RRT (bottom images) passes above the obstacle while stretching the arm. In this configuration, the high payload causes a large path deviation, resulting in an unexpected collision. Fig. 6c also shows the cumulative regret of `kfmanb` and `random` for a single experiment. The trend qualitatively confirms the regret results discussed in Sec. V-B.

VII. CONCLUSIONS

We presented an online learning approach to biased sampling for kinodynamic motion planning. The approach runs RRT multiple times and uses MAB to choose between uniform sampling and sampling regions identified during the previous runs. We showed that the proposed approach finds better solutions faster than an unbiased planner. The experiments also suggested a correlation between low regret and cost in different scenarios. Future works will investigate how to improve the performance of the approach, *e.g.*, via pruning and a single-tree implementation as in [46] and [11].

APPENDIX

A. Parameter tuning

We tune the parameters of MAB-RRT planner almost independently for each main module of the method.

1) *Clustering*: HDBSCAN requires the minimum number N_{\min} of points in a cluster. Small values of N_{\min} favor the identification of small clusters with sparse data (which is likely the case for high-dimensional planning problems). Because small values of N_{\min} allow for spotting small high-reward clusters, we empirically set N_{\min} at random between 2 and 5 at each clustering. Moreover, the reward weight λ in (1) is needed to define the clustering distance function. Large values of λ tend to favor clusters with similar rewards; small values favor state-space proximity. We observed a low sensitivity to λ in all our experiments; $\lambda \in [1, 10]$ yielded satisfactory performance.

2) *Bandit algorithm*: KF-MANB requires initializing the expected rewards $\bar{\mu}_i(0)$, their variance $\sigma_i^2(0)$, and the Kalman filter's noise factors σ_{obs}^2 , σ_{tr}^2 and η . σ_{obs}^2 controls how much we believe the new observed reward (the larger σ_{obs}^2 , the faster the Kalman Filter adapts the arm's distribution mean). σ_{tr}^2 increases the variance of non-selected arms to favor their exploration. η is a tuning parameter to scale the

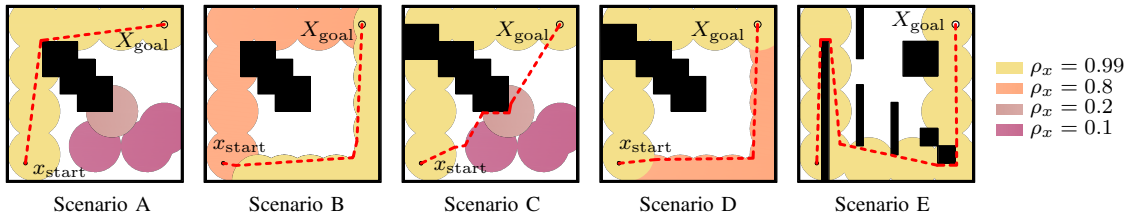


Fig. 3: Scenarios used for the numerical analysis in Sec. V. Dashed red lines are the optimal paths.

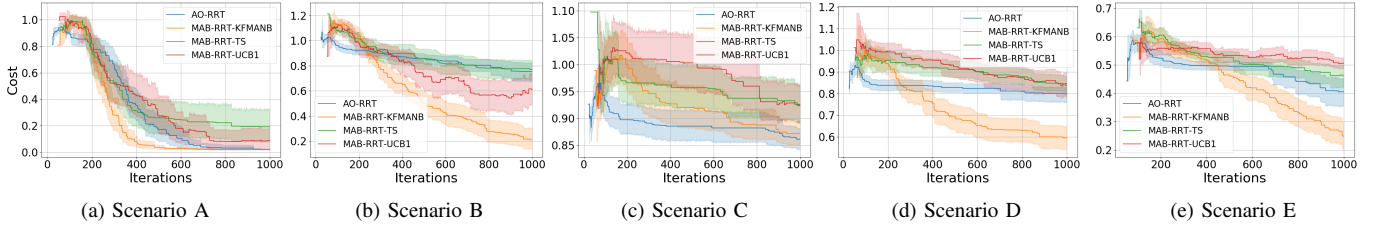


Fig. 4: Cost trends for the scenarios of Fig. 3 (30 repetitions; solid lines: mean values; shadow: 95% confidence interval).

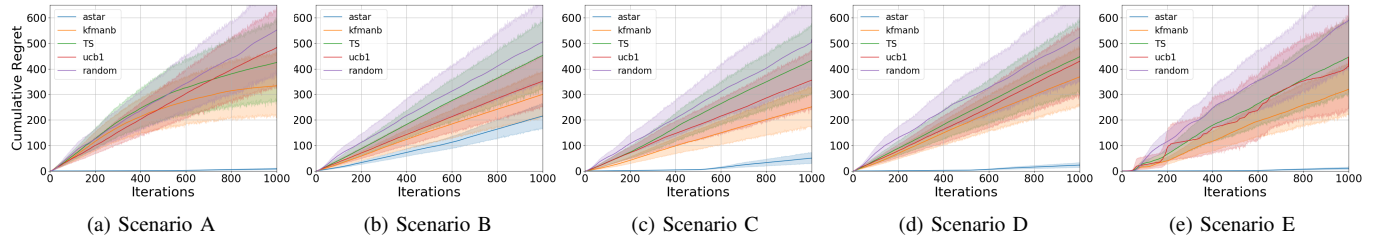


Fig. 5: Cumulative regret for the scenarios of Fig. 3 (mean values of 30 repetitions).

covariance to match the reward scale. While performance on individual experiments could be marginally improved by using different values, we found that $\sigma_i(0) = 0.2 \forall i \in 1, \dots, M$, $\sigma_{\text{obs}}^2 = 10^{-4}$ and $\sigma_{\text{tr}}^2 = 10^{-4}$ gave satisfactory results for all of our planning tasks. Concerning the initial rewards, we set $\bar{\mu}_i(0) = \mathcal{R}_i \forall i \in 1, \dots, M$. Finally, we set η dynamically to $\eta(k+1) = \max(10^{-10}, 0.9\eta(k) + 0.1|r(k+1)|)$ as in [50].

3) *Cluster sampling*: Alg. 3 requires thresholds δ_1 , δ_2 , and δ_3 . We relate their values to the dispersion of each cluster so that, for all clusters, $\delta_1 = \delta_2 = \text{median}(\{d_1, \dots, d_{|\mathcal{T}_i|}\})$, where $d_j = \min_z \|z - \tau\| \forall \tau \in \mathcal{T}_i$, and $\delta_3 = 2\delta_2$.

B. Cost function proportional to Cartesian error

Assuming we do not know the actual controller parameters, we consider a simplified proportional joint-space controller $\xi_{\text{mot}} = \hat{H}(q)K_p e_p + \hat{C}(q, \dot{q})\dot{q} + \hat{g}(q)$, where $\xi_{\text{mot}} \in \mathbb{R}^7$ is the torque required to the joint motors, $K_p > 0$, and \hat{H} , \hat{C} , \hat{g} are the estimated inertia, Coriolis, and gravity matrices. Assuming quasi-static conditions and perfect knowledge of robot inverse dynamics, we can write $\hat{H}(q)K_p e_p - J(q)^T f_{\text{ext}} = 0$, where $f_{\text{ext}} \in \mathbb{R}^6$ is the external wrench (owed to the payload), and J is the robot Jacobian. By approximating $\Delta x \approx J\Delta q$ for small Δq , the estimated maximum Cartesian-space position error is $e_{xyz} = K_p^{-1} [I_3 \ 0_{3 \times 3}] J(q)H(q)J(q)^T f_{\text{ext}}$. Because K_p is a constant scalar, we can set $K_p = 1$ and scale e_{xyz} between 0 and 1 to obtain $\hat{e}_q = \min(\max(|e_{xyz}|), e_{\text{max}}, 1)$ and $\hat{e}(\tau) = 0.5(\hat{e}_q(\tau.x_p) + \hat{e}_q(\tau.x_c))$, where e_{max} is an empirical estimate of the maximum value of $\max(|e_{xyz}|)$. In our experiments, we set $e_{\text{max}} = 70$ by computing the

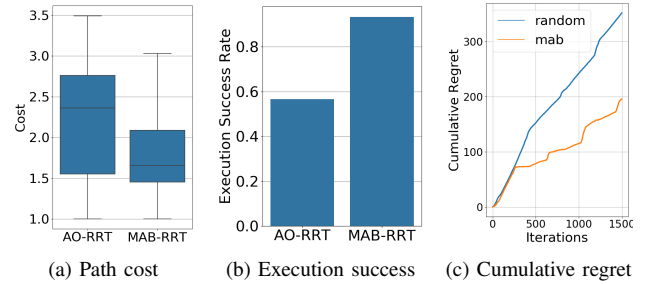


Fig. 6: Experimental results of the manipulation scenario.

maximum value of $\max(|e_{xyz}|)$ from 10^5 random q . \hat{H} and J are from the URDF model provided by the robot manufacturer.

REFERENCES

- [1] P. Mitrano, D. McConachie, and D. Berenson, “Learning where to trust unreliable models in an unstructured world for deformable object manipulation,” *Science Robotics*, vol. 6, no. 54, p. eabd8170, 2021.
- [2] X. Lin, C. Qi, Y. Zhang, Z. Huang, K. Fragkiadaki, Y. Li, C. Gan, and D. Held, “Planning with spatial-temporal abstraction from point clouds for deformable object manipulation,” in *CoRL*, 2022.
- [3] M. Lippi, P. Poklukar, M. C. Welle, A. Varava, H. Yin, A. Marino, and D. Kragic, “Enabling visual action planning for object manipulation through latent space roadmap,” *IEEE T-RO*, vol. 39, pp. 57–75, 2023.
- [4] P. Mitrano, A. LaGrassa, O. Kroemer, and D. Berenson, “Focused adaptation of dynamics models for deformable object manipulation,” *Robotics: Science and Systems*, 2022.
- [5] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, and F. Chen, “Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects,” *IEEE RAL*, vol. 7, no. 2, pp. 5159–5166, 2022.
- [6] J. Liang, X. Cheng, and O. Kroemer, “Learning preconditions of hybrid force-velocity controllers for contact-rich manipulation,” 2022.
- [7] M.-T. Khoury, A. Orthey, and M. Toussaint, “Efficient sampling of transition constraints for motion planning under sliding contacts,” in *IEEE CASE*, 2021, pp. 1547–1553.

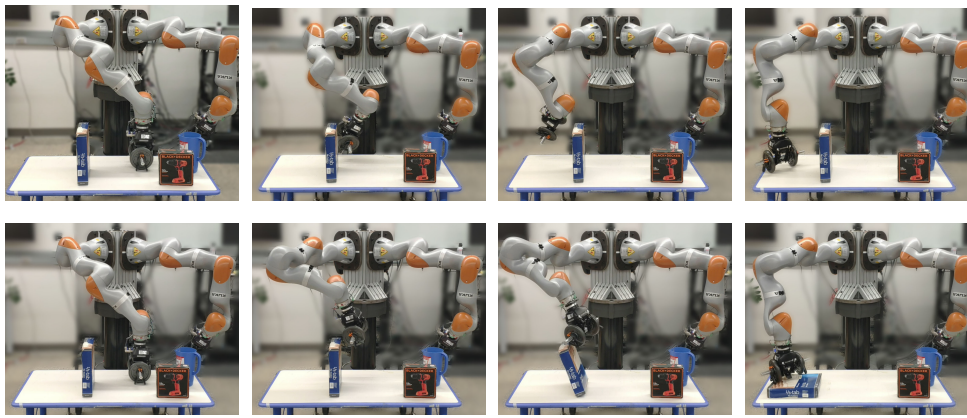


Fig. 7: Examples of executions of two trajectories planned with MAB-RRT (top) and AO-RRT (bottom).

- [8] S. M. LaValle and J. James J. Kuffner, "Randomized kinodynamic planning," *Int. J. Robot. Res.*, vol. 20, no. 5, pp. 378–400, 2001.
- [9] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 846–894, 2011.
- [10] D. J. Webb and J. Van Den Berg, "Kinodynamic RRT*: Asymptotically optimal motion planning for robots with linear dynamics," in *ICRA*, 2013, pp. 5054–5061.
- [11] Y. Li, Z. Littlefield, and K. E. Bekris, "Asymptotically optimal sampling-based kinodynamic planning," *IJRR*, vol. 35, no. 5, pp. 528–564, 2016.
- [12] C. Urmson and R. Simmons, "Approaches for heuristically biasing RRT growth," in *IROS*, vol. 2, 2003, pp. 1178–1183.
- [13] J. D. Gammell, T. D. Barfoot, and S. S. Srinivasa, "Informed sampling for asymptotically optimal path planning," *IEEE T-RO*, vol. 34, 2018.
- [14] B. Ichter, J. Harrison, and M. Pavone, "Learning sampling distributions for robot motion planning," in *ICRA*, 2018, pp. 7087–7094.
- [15] C. Chamzas, Z. Kingston, C. Quintero-Peña, A. Shrivastava, and L. E. Kavraki, "Learning sampling distributions using local 3d workspace decompositions for motion planning in high dimensions," in *ICRA*, 2021.
- [16] R. Cheng, K. Shankar, and J. W. Burdick, "Learning an optimal sampling distribution for efficient motion planning," in *IROS*, 2020.
- [17] S. Choudhury, J. D. Gammell, T. D. Barfoot, S. S. Srinivasa, and S. Scherer, "Regionally accelerated batch informed trees (rabit): A framework to integrate local information into optimal path planning," in *ICRA*, 2016.
- [18] J. D. Gammell, S. S. Srinivasa, and T. D. Barfoot, "Batch Informed Trees (BIT*): Sampling-based optimal planning via the heuristically guided search of implicit random geometric graphs," in *ICRA*, 2015.
- [19] K. Hauser and Y. Zhou, "Asymptotically optimal planning by feasible kinodynamic planning in a state–cost space," *IEEE T-RO*, vol. 32, 2016.
- [20] D. S. Yershov and E. Frazzoli, "Asymptotically optimal feedback planning using a numerical hamilton-jacobi-bellman solver and an adaptive mesh refinement," *Int. J. Robot. Res.*, vol. 35, no. 5, pp. 565–584, 2016.
- [21] J.-S. Ha, J.-J. Lee, and H.-L. Choi, "A successive approximation-based approach for optimal kinodynamic motion planning with nonlinear differential constraints," in *IEEE CDC*, 2013, pp. 3623–3628.
- [22] Z. Littlefield and K. E. Bekris, "Efficient and asymptotically optimal kinodynamic motion planning via dominance-informed regions," in *IROS*, 2018, pp. 1–9.
- [23] M. G. Westbrook and W. Ruml, "Anytime kinodynamic motion planning using region-guided search," in *IROS*, 2020, pp. 6789–6796.
- [24] M. Kleinbort, K. Solovey, Z. Littlefield, K. E. Bekris, and D. Halperin, "Probabilistic completeness of prr for geometric and kinodynamic planning with forward propagation," *IEEE RAL*, vol. 4, pp. 277–283, 2019.
- [25] R. Shome and L. E. Kavraki, "Asymptotically optimal kinodynamic planning using bundles of edges," in *ICRA*, 2021, pp. 9988–9994.
- [26] J. D. Gammell and M. P. Strub, "Asymptotically optimal sampling-based motion planning methods," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 295–318, 2021.
- [27] N. M. Amato, O. B. Bayazit, L. K. Dale, C. Jones, and D. Vallejo, "Obprm: An obstacle-based prm for 3d workspaces," in *Int. Workshop on Algorithmic Foundations of Robotics*, 1998, pp. 155–168.
- [28] A. Attali, S. Ashur, I. B. Love, C. McBeth, J. Motes, D. Uwacu, M. Morales, and N. M. Amato, "Evaluating guiding spaces for motion planning," 2022.
- [29] S. Dalibard and J.-P. Laumond, "Linear dimensionality reduction in random motion planning," *IJRR*, vol. 30, pp. 1461–1476, 2011.
- [30] F. U. González, J. Rosell, and R. Suárez, "Task space vector field guiding for motion planning," in *IEEE ETFA*, 2022, pp. 1–7.
- [31] T. Lai and F. Ramos, "Adaptively exploits local structure with generalised multi-trees motion planning," *IEEE RAL*, vol. 7, no. 2, pp. 1111–1117, 2021.
- [32] T. Lai, P. Morere, F. Ramos, and G. Francis, "Bayesian local sampling-based planning," *IEEE RAL*, vol. 5, no. 2, pp. 1954–1961, 2020.
- [33] M. P. Strub and J. D. Gammell, "Advanced BIT* (ABIT*)- sampling-based planning with advanced graph-search techniques," in *ICRA*, 2020.
- [34] —, "Adaptively informed trees (AIT*): Fast asymptotically optimal path planning through adaptive heuristics," in *ICRA*, 2020.
- [35] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.
- [36] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conf. on learning theory*, 2012, pp. 39–1.
- [37] O.-C. Granmo and S. Berg, "Solving non-stationary bandit problems by random sampling from sibling kalman filters," in *Trends in Applied Intelligent Systems*, 2010, pp. 199–208.
- [38] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.
- [39] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
- [40] M. Phillips, V. Narayanan, S. Aine, and M. Likhachev, "Efficient search with an ensemble of heuristics," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [41] J. Lee, D. Yi, and S. S. Srinivasa, "Sampling of pareto-optimal trajectories using progressive objective evaluation in multi-objective motion planning," in *IROS*, 2018, pp. 1–9.
- [42] M. C. Koval, J. E. King, N. S. Pollard, and S. Srinivasa, "Robust trajectory selection for rearrangement planning as a multi-armed bandit problem," in *IROS*, 2015, pp. 2678–2685.
- [43] W. Wang, L. Zuo, and X. Xu, "A learning-based multi-rrt approach for robot path planning in narrow passages," *Journal of Intelligent & Robotic Systems*, vol. 90, pp. 81–100, 2018.
- [44] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*, 2013, pp. 160–172.
- [45] T. Kunz and M. Stilman, "Kinodynamic rrt with fixed time step and best-input extension are not probabilistically complete," in *WAFR*, 2015, pp. 233–244.
- [46] M. Kleinbort, E. Granados, K. Solovey, R. Bonalli, K. E. Bekris, and D. Halperin, "Refined analysis of asymptotically-optimal kinodynamic planning in the state–cost space," in *IEEE ICRA*, 2020, pp. 6344–6350.
- [47] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, 2017.
- [48] I. A. Şucan, M. Moll, and L. E. Kavraki, "The Open Motion Planning Library," *IEEE Robot. Autom. Mag.*, vol. 19, no. 4, pp. 72–82, 2012.
- [49] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IROS*, 2004.
- [50] D. Mcconachie and D. Berenson, "Estimating model utility for deformable object manipulation using multiarmed bandit methods," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 967–979, 2018.