2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

A Tube-Based Reinforcement Learning Approach for Optimal Motion Planning in Unknown Workspaces

†Panagiotis Rousseas, ‡Charalampos P. Bechlioulis, and †Kostas J. Kyriakopoulos

***Abstract*— In this work, a tube-based nearly optimal solution to motion planning in unknown workspaces is presented. The advantages of reactive motion planning are combined with a Policy Iteration Reinforcement Learning scheme to yield a novel solution for unknown workspaces that inherits provable safety, convergence and optimality. Moreover, in simply-connected workspaces, our method is proven to asymptotically provide the globally optimal path. Our method is compared against a provably asymptotically optimal RRT* method, as well as a relevant reactive method and provides satisfactory performance, closely matching or outperforming the former.**

I. Introduction

Over the past few years, robots are operating in increasingly challenging and complex environments, due to related workspace uncertainties and un-modeled features. Towards accomplishing any related task, planning a robot's motion within such environments is in most cases a pre-requisite, thus, the Motion Planning (MP) problem has been one of the most fundamental problems with both theoretical and practical challenges. In light of the above, this work aims at providing a provably safe, nearly optimal solution to motion planning in unknown, planar workspaces. While planning in unknown workspaces has been extensively treated with open-loop methods, the aim is to inherit the advantages of Reactive Approaches (RAs) (e.g. extension to drift dynamics, robustness). Concurrently, a Policy Iteration (PI), Reinforcement Learning (RL) scheme is formulated to provide a nearly optimal control law, without sacrificing the provable guarantees of RAs. However, this necessitates omitting internal obstacles, in order to avoid sacrificing optimality guarantees, or imposing discontinuity over the solution, as strict navigation is impossible through vector fields in manifolds with multiple connectivity (i.e., obstacles) [1]. Furthermore, in treating unknown workspaces, the challenges posed by limited sensing, (e.g. through a LiDAR sensor) necessitate that extra care is taken to explore the workspace prior to finding the optimal path around obstacles. Therefore, the herein proposed method consists a first step in tackling Optimal Motion Planning (OMP) in unknown workspaces through RAs.

Although employing RAs presents key advantages for Robotics applications, PI schemes are notorious for their computational complexity. While, in light of introducing provable guarantees in the context of RL, this might remain an attractive compromise, treating simply-connected (obstacle-free) unknown workspaces enables formulating a tube-based approach that limits the computational load of the resulting scheme, without sacrificing optimality, while safety guarantees during both the exploration, and training processes are critically preserved.

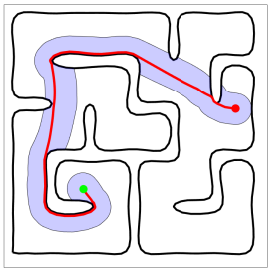

Fig. 1: A Synthetic Workspace (black lines), and an example of the optimal trajectory (red line) within a tube-like portion of the workspace (blue-shaded area).

II. Related Work

Some of the most prevalent methods for tackling the MP problem belong to the family of open-loop approaches, including Sampling-Based Methods (SBMs). These include the Rapidly-exploring Random Trees (RRT) algorithm [2], Probabilistic Roadmaps (PRM) [3], etc. and usually perform probabilistic sampling over a (continuous of discretized) workspace. On the other hand, grid-based approaches, such as A* [4] or Dijkstra's algorithm [5] (which are path length minimizers belonging to graph-based methods) solve the problem over a discretized version of the original workspace. In tackling optimality, RRT* [6] significantly improves upon its predecessor by introducing asymptotically optimal path length minimization, complex cost functions with energy input minimization, kinematic and dynamic constraints, etc. [7]. Consequent efforts, such as pruning [8], enhance further the computational behaviour of the method through biasing the sampling on the workspace. Notable extensions even extend the method to dynamic environments [9]. Finally, approaches such as interpolating curve algorithms [10] tackle the problem by employing parametrized solutions to path-planning, and optimizing the former.

Concerning RAs, these concentrate around the single integrator model, i.e., modelling the robot as a particle able to follow velocity commands exactly. The latter usually are in the form of the gradient of a field defined over the workspace,

†The authors are with the School of Mechanical Engineering, Control Systems Laboratory, National Technical University of Athens, Greece. ‡The author is with the Department of Electrical and Computer Engineering, University of Patras. E-mails: {prousseas,kkyria}@mail.ntua.gr, chmpechl@upatras.gr

which is designed to exhibit a single global minimum at the desired goal position, and maximal values over the workspace boundary, guaranteeing safety and convergence by construction. Such methods include Navigation Functions (NFs) [11] and Artificial Harmonic Potential Fields (AHPFs) [12]. However, NFs are hard to tune in order to alleviate any local minima and in most cases require a transformation from the original workspace to a star-shaped one. In order to amend these drawbacks, AHPFs were developed, requiring no extensive tuning. Nevertheless, the need for a workspace transformation persists, however this problem has seen some development in the form of harmonic maps [13], which transform any workspace to a punctured-disk one. In the context of optimality, however, RAs have seen limited development. Concerning unknown workspaces, there has been some reactive methods [14], [15]. The authors have previously concentrated on parametrized [16], [17] and non-parametrized [18] PI-RL methods, with an extension to unknown workspaces [19]. The herein proposed method, aims at extending previous efforts to the non-parametric regime, in order to negate the disadvantages of restricting the robot's policy to AHPFs, which hinders the optimality of the final output of the relevant method.

A. Outline and Contribution

The rest of the manuscript is organized as follows: In Section III, the reactive OMP problem is formulated, and the proposed framework is introduced in Section IV. A theorem concerning the optimality of the method is proven in Section IV-A, while details over the implementation of the proposed scheme are outlined in Section IV-B. Comparative simulations with both the authors' previous work on unknown workspaces [19] and an asymptotically optimal RRT* method are presented in Section V. The paper concludes with a discussion over the presented results and possible future directions in Section VI. The main contribution of this work lies in extending non-parametrized OMP [18] to unknown workspaces and providing a tube-based approach that does not hinder optimality, while restricting the computational complexity of the method.

III. Problem Formulation

In this work, we treat a point-robot[1], operating within a planar, bounded and connected workspace $\mathcal{W} \subset \mathbb{R}^2$[2] and adopt $\partial\mathcal{W}$ to denote the boundary of $\mathcal{W}$. Additionally, let $p \in \mathcal{W} - \partial\mathcal{W}$ denote the robot's position. We assume that the robot's position obeys the single integrator model (as usual in reactive methods):

$$\dot{p} = u,\ p(0) = \bar{p} \in \mathcal{W} - \partial\mathcal{W}, \tag{1}$$

where $u(t) : \mathbb{R}_+ \mapsto \mathbb{R}^2$ denotes the velocity control input and $\bar{p}$ denotes the robot's position at time $t = 0$. Furthermore, assume that the robot obtains workspace information through a LiDAR sensor with range r, where the sensed subset of the workspace is defined as follows:

$$\mathcal{S}(p) = \{q \in \mathcal{W} : (\mathcal{B}(p, r) \wedge \mathcal{L}(p, q)) \subseteq \mathcal{W}\}, \tag{2}$$

where $\mathcal{B}(p, r)$ is a disk with radius r, and $\mathcal{L}(p, q)$ is the line segment that connects p and q inclusively. In addition, the robot's path under the policy u, parametrized from time $t = 0$ to $t > 0$ is denoted as $p_u(t; \bar{p})$: $p_u(t; \bar{p}) = \int_0^t u(\tau)d\tau + \bar{p}$. Therefore, the explored region along the whole path $\mathcal{P}_u(t; \bar{p}) = \bigcup_{\tau\in[0,t]} p_u(\tau; \bar{p})$ is:

$$\mathcal{E}(\mathcal{P}_u) = \bigcup_{\tau\in[0,t]} \mathcal{S}(p_u(\tau; \bar{p})). \tag{3}$$

Furthermore consider the following cost function form [20], which is subject to minimization:

$$V_u(\bar{p}; p_d) = \int_0^\infty r(p_u(\tau; \bar{p}), u; p_d)\, d\tau, \tag{4}$$

where $r(p, u; p_d) \triangleq Q(p; p_d) + R(u)$ involves a state-related term $Q(p; p_d) = \alpha \|p - p_d\|^2$ and an input-related term $R(u) = \beta \|u\|^2$, with $\|\cdot\|$ denoting the Euclidean 2-norm, and where $\alpha, \beta > 0$ are weighting/design parameters. Thus, the R term minimizes input energy, while the Q term minimizes the settling time of the system, as it penalizes the robot for staying away from the goal position as time evolves.

IV. Methodology

1) Safety during Navigation: In order to ensure safety, Zeroing Barrier Function (ZBF) theory [21] is employed. We define the following ZBF $L(p) : \mathcal{X} \mapsto [0, 1]$ w.r.t. the safety-set $\mathcal{X} \subset \mathcal{W}$:

$$L(p; \mathcal{X}) = \begin{cases} 1 - \exp\left(-\left(\frac{d(p)}{d(p)-a}\right)^2\right), & d(p; \mathcal{X}) \leq a \\ 1, & d(p; \mathcal{X}) > a \end{cases}, \tag{5}$$

with $a \in \mathbb{R}_+$[3]. The function $d : \mathcal{X} \mapsto \mathbb{R}_+$ computes the current distance of the robot to the boundary of the set $\mathcal{X}$:

$$d(p; \mathcal{X}) = \min_{z\in\partial\mathcal{X}} \{\|p - z\|\} \tag{6}$$

The above ZBF $L(p; \mathcal{X})$ equals 1 in the interior of $\mathcal{X}$ at a distance-to-the-boundary larger than, or equal to a, while varying smoothly (but not analytically) from 1 to 0 for points with a distance less than a. System (1) is safe if:

$$\dot{L} + h(L) \geq 0 \Leftrightarrow (\nabla L)^{\mathrm{T}} u + h(L(p)) \geq 0, \tag{7}$$

where $h(\cdot)$ is a class $\mathcal{K}$ function [21] and the dependence on the set $\mathcal{X}$ is dropped for readability.

[1] A disk robot can also be treated through inflating the workspace boundary inwards by a distance equal to the robots radius. An example of such a transformation is: $T : \partial\mathcal{W} \to \partial\mathcal{W}'$, $p' = T(p) = p + R \times n(p)$, $\forall p \in \partial\mathcal{W}$, where R denotes the robot's radius, and $n(p)$ denotes the unitary, inwards-pointing normal vector to the boundary at each point $p \in \partial\mathcal{W}$, and $\partial\mathcal{W}'$ denotes the transformed boundary.

[2] In case of a simply connected, i.e., obstacle-free workspace, the method will be proven to provide the globally optimal solution, whereas otherwise, a local solution, dependent on the topological properties of the initial path will be extracted.

[3] The use of the English letter a in (5) is not to be mistaken for the Greek letter α employed in the definition of the metric Q.

2) Optimality: In order to solve the OMP problem in unknown workspaces, we develop a tube-based Policy Iteration (PI) method. Extracting the optimal policy can be accomplished through forming the following Hamilton-Jacobi-Bellman (HJB) equation from the differential form of (4):

$$H\left(p,u;\nabla V_u;p_d\right)=\left(\nabla V_u\right)^{\mathrm{T}}u+r\left(p,u;p_d\right). \tag{8}$$

The optimal input/cost function pair satisfies the following equation:

$$\min_u\{H\left(p,u;\nabla V^\star;p_d\right)\}=0. \tag{9}$$

In order to account for the safety specifications as defined in (7), the following constrained optimization problem is instead proposed:

$$\begin{aligned}&\min_u\{H\left(p,u;\nabla V^\star;p_d\right)\}=0,\\ &\text{s.t.: } C(p;u)\triangleq\left(\nabla L\right)^{\mathrm{T}}u+h\left(L(p)\right)\geq 0.\end{aligned} \tag{10}$$

The above minimization problem cam be tackled through the addition of a Lagrange multiplier $\lambda\in\mathbb{R}_+$. The corresponding Karush–Kuhn–Tucker (KKT) stationary condition yields:

$$\nabla V^\star+\frac{\partial r\left(p,u\right)}{\partial u}\bigg|_{u^\star}-\lambda\frac{\partial C(p;u)}{\partial u}\bigg|_{u^\star}=0, \tag{11}$$

which yields the optimal constrained control as follows:

$$u^\star=-\frac{1}{2\beta}\left(\nabla V^\star-\lambda^\star\nabla L\right). \tag{12}$$

The Lagrange multiplier can be extracted by considering the optimal condition $\left(\lambda^\star\right)C\left(p;u^\star\right)=0$ (for $\lambda^\star\neq 0$). Thus, the optimal Lagrange multiplier is:

$$\lambda^\star(p)=\begin{cases}0 & \text{if } C^\star\geq 0\\ \frac{-2\beta h(L(p))+(\nabla L)^{\mathrm{T}}(\nabla V^\star)}{\|\nabla L\|^2} & \text{if } C^\star<0\end{cases}, \tag{13}$$

where $C^\star=C\left(p;-1/2\beta\nabla V^\star\right)$. Substituting the above into (8) yields a hard, non-linear Partial Differential Equation (PDE). To avoid solving this PDE, we employ a PI-based, RL method, based upon successive approximation [22], which entails successively approximating the optimal cost function.

3) Tube-Based Policy Iteration: In order to implement PI in unknown workspaces we propose a tube-based optimization framework, which presumes a method for navigating towards a desired final position in unknown workspaces. A plethora of such methods have been proposed in the literature, from SBMs [4], [23] to reactive ones [15]. Hence, given a path $\mathcal{P}_u$ from an initial $\bar{p}$, to a final position p_d, a corresponding tube is defined as (see Fig. 1):

$$\mathcal{B}_R\left(\mathcal{P}_u\right)=\{x\in\mathcal{E}\left(\mathcal{P}_u\right)|\min_{z\in\mathcal{P}_u}\{\|x-z\|\}\leq R\}, \tag{14}$$

where $R>0$ defines a buffer distance around the path and $\mathcal{P}_u=\mathcal{P}_u\left(t\to\infty;\bar{p}\right)$ is employed to denote the entire path till convergence for brevity. In presenting the PI scheme, we begin by defining an *Admissible Policy* on a tube:

Definition 1: (**Admissible Policy**) A policy $u\left(p\right):\mathcal{B}\mapsto\mathcal{A}\left(\mathcal{B}\right)$, where $\mathcal{A}\left(\mathcal{B}\right)$ denotes the set of admissible policies w.r.t. the tube $\mathcal{B}$, is defined as admissible with respect to the cost function (4) over the tube $\mathcal{B}$, if: **1)** u is continuous on $\mathcal{B}$, **2)** $u\left(p_d\right)=\tilde{0}$, **3)** $u\left(p\right)$ stabilizes (1) on $\mathcal{B}$, **4)** $V_u\left(p\right)$ is finite $\forall p\in\mathcal{B}$ and **5)** the resulting trajectories of (1) under the control law $u=u\left(p\right)$ are safe, i.e., for any $\bar{p}\in\left(\mathcal{B}-\partial\mathcal{B}\right)$ it holds that $\mathcal{P}_u\left(t\to\infty,\bar{p}\right)\bigcap\partial\mathcal{B}=\emptyset$[4].

In contrast to previous works, where the PI scheme is implemented over the entire sensed subset of the workspace [19] herein the optimization is restricted on the smaller tube subset to alleviate the computational load in the context of unknown workspaces. Therefore, we propose the following PI scheme: Given an initial policy $u^{(i=0)}\in\mathcal{A}\left(\mathcal{B}_R\right)$ that guides the robot to the goal position (and an appropriate tube $\mathcal{B}_R$), a sequence of policies is extracted:

$$u^{(i+1)}=-\frac{1}{2\beta}\left(\nabla V^{(i)}-\lambda^{(i)}(p)\nabla L\right), \tag{15}$$

where

$$\lambda^{(i)}=\begin{cases}0 & \text{if } C^{(i)}\geq 0\\ \frac{-2\beta h(L(p))+(\nabla L)^{\mathrm{T}}(\nabla V^{(i)})}{\|\nabla L\|^2} & \text{if } C^{(i)}<0\end{cases}. \tag{16}$$

where $C^{(i)}=C\left(p;-\frac{\nabla V^{(i)}}{2\beta}\right)$. In the tube case, the ZBF (5) is computed w.r.t. the boundary of the **tube** $\mathcal{B}_R$, i.e. the ZBF becomes: $L(p)\triangleq L\left(p;\mathcal{B}_R\right)$, which, since the tube is a subset of $\mathcal{W}$, ensures safety w.r.t. the workspace boundary. In a previous work [18], we have demonstrated how the resulting sequence of policies yields admissible policies (per Def. 1), as well as decreasing cost values for the respective trajectories. Furthermore, in the limit where $i\to\infty$ the sequence yields a nearly optimal policy w.r.t. to the domain of solution of the optimization problem[5]. However, the limit of the above sequence is optimal w.r.t. the tube $\mathcal{B}_R$, which is not necessarily (and most likely never) identical to the optimal policy for the initial/final position pair $\{\bar{p},p_d\}$ in $\mathcal{W}$. Therefore, a possible solution is to compute the nearly optimal policy $u'\in\mathcal{A}\left(\mathcal{B}_R\right)$ for the tube $\mathcal{B}_R$, which results in a new path $\mathcal{P}'=\mathcal{P}_{u'}$. Then, a *new tube* $\mathcal{B}'_R$ can be extracted. Extracting the optimal policy for $\mathcal{B}'_R$ and repeating the above process yields the optimal policy for the initial/final position pair $\{\bar{p},p_d\}$ in $\mathcal{W}$ upon convergence. This claim is proven in the following section. In order to formalize the preceding elements, consider a policy and respective tube, indexed by (i,j) and j respectively, the first index denoting the PI step and the second index denoting the iterative process of updating the tube:

$$\bar{u}^{(i,j)}\in\mathcal{A}\left(\mathcal{B}_R^{(j)}\left(\mathcal{P}_{u^{(i)}}\right)\right),\ i,j\in\{1,2,\cdots,\infty\}, \tag{17}$$

where $\bar{u}^{(i,j)}\triangleq u^{(i)}$, given by Eqs. (15), (16) for the corresponding ZBF given by $L^{(j)}\triangleq L\left(p;\mathcal{B}_R^{(j)}\right)$ (see Eqs. (5), (6)). Hence the proposed scheme entails repeating the following steps until convergence:

[4]This definition ensures that under the control law u the robot does not collide with the tube boundary at any point along all trajectories (for any initial position $\bar{p}$).

[5]This can be proven as an extension of Theorem 5 in [24], for the constrained optimization problem that is treated herein.

1) Starting from a valid path $\mathcal{P}^{(0)} \triangleq \mathcal{P}_{u^{(0)}}$ and policy $\bar{u}^{(0,1)}$ compute the tube $\mathcal{B}^{(j=1)} \triangleq \mathcal{B}_R\left(\mathcal{P}_{u^{(i)}}\right) \subset \mathcal{W}$.
2) Implement the PI scheme in [18] to obtain the optimal policy in $\mathcal{B}^{(1)}$,
3) Compute the new path $\mathcal{P}^{(\infty)}$ (with an abuse of notation where $i \to \infty$) and the new tube $\mathcal{B}_R^{(2)} \triangleq \mathcal{B}_R\left(\mathcal{P}^{(\infty)}\right)$ for the above path,
4) Compute a new addmissible policy $\bar{u}^{(0,2)} \in \mathcal{A}\left(\mathcal{B}_R^{(2)}\right)$ for the new tube.

A. Technical Results

In this section we prove how the proposed scheme results in the optimal policy for an initial/final position configuration in the case of simply connected workspaces, irrespective of the initial path.

Theorem 1: Consider a simply-connected, bounded workspace $\mathcal{W} \subset \mathbb{R}^2$, along with a valid path $\mathcal{P}_u$ that connects an initial/final position pair $\{\bar{p}, p_d\}$. Then the sequence of reactive policies described in Section IV approaches the optimal policy for the tube $\mathcal{B}_R^{(j\to\infty)}$. Additionally, this policy is optimal w.r.t. the workspace $\mathcal{W}$ for the initial position $\bar{p} \in \mathcal{B}_R^{(j\to\infty)}$.

Proof: Consider the PI scheme for a tube $C_R^{(j)},\ j \in \mathbb{N}^+$, starting from an initial policy $u^{(0,j)} \in \mathcal{A}\left(\mathcal{B}_R^{(j)}\right)$. According to [18], this yields a decreasing sequence of cost functions, i.e.

$$V^{(i+1,j)}\left(p; p_d\right) \leq V^{(i,j)}\left(p; p_d\right),\ i \in \{1, 2, \cdots, \infty\}, \quad (18)$$

for any $p \in \mathcal{B}_R^{(j)}$. Consider now the tube $\mathcal{B}_R^{(j+1)}$. Since by definition -see (14)-, the initial point $\bar{p}$ lies within both $\mathcal{B}_R^{(j)}$ and $\mathcal{B}_R^{(j+1)}$, it holds that, implementing the same PI scheme for $\mathcal{B}_R^{(j+1)}$:

$$V^{(\infty,j+1)}(p; p_d) \leq V^{(\infty,j)}(p; p_d), \quad (19)$$

which can be seen through the following: According to [18], $V^{(\infty,j)}(p; p_d)$ is the optimal cost for $p \in \mathcal{B}_R^{(j)}$. The optimal trajectory from $\bar{p}$ inside the set $\mathcal{B}_{(j,j+1)} \triangleq \mathcal{B}_R^{(j)} \cup \mathcal{B}_R^{(j+1)}$ belongs either entirely, partially, or not at all inside $\mathcal{B}_R^{(j)}$. In the first case, since $\mathcal{B}_R^{(j)} \cap \mathcal{B}_R^{(j+1)} \neq \emptyset$ (and the tube $\mathcal{B}_R^{(j+1)}$ is extracted from the trajectory of the policy $u^{(\infty,j)}$), the PI scheme employed inside $\mathcal{B}_R^{(j+1)}$ guarantees that the equality in (19) holds. In the last case, by definition the optimal policy can be extracted through the PI scheme in $\mathcal{B}_R^{(j+1)}$ and hence the inequality in (19) holds strictly. In case where the optimal trajectory in $\mathcal{B}_{(j,j+1)}$ belongs partially in both constituent sets, then by Bellman's optimality principle it can be shown that the trajectory belongs to the set $\left(\mathcal{B}_R^{(j)} \cap \mathcal{B}_R^{(j+1)}\right) \cup \left(\mathcal{B}_R^{(j+1)} \setminus \mathcal{B}_R^{(j)}\right) \neq \emptyset$. Assuming the opposite would imply that at least part of the trajectory resulting from $u^{(\infty,j)}$ is not optimal in $\mathcal{B}_R^{(j)}$.

Additionally, the decreasing sequence of cost function values $V^{(\infty,j)},\ j = 1, 2, \cdots, \infty$, is bounded below by the optimal cost function. However, this does not necessarily imply that the sequence converges to the latter. However, the preceding arguments (along with the near global optimality proof in [24]) imply that the optimal cost function is indeed approximated as $i \to \infty$ for each subset $\mathcal{B}_R^{(j)}$. Hence, at the limit where $j \to \infty$ (that is, in the worst case where the tubes are very small), the tubes converge to the tube around the globally optimal policy $u^{(\infty)}$. ■

Notice how, due to the nature of the PI scheme, the initial policy $\bar{u}^{(0,j)}$ for each tube $\mathcal{B}_R^{(j)}$ is irrelevant for extracting the respective optimal policy within the tube, as long as it is admissible w.r.t. the latter, per Def. 1. The necessity for the simply-connected workspace case also now becomes apparent, as the aforementioned iterative scheme provides a sequence of *homeomorphic two-dimensional curves* in $\mathcal{W}$. Any obstacles in $\mathcal{W}$ would impact the optimality of the final result, as the initial policy $\bar{u}^{(0,j)}$ defines an equivalence class of homeomorphic curves which does not necessarily include the truly optimal path. This is left untreated in this work and left as part of our intended future research efforts.

B. Implementation Details

1) Cost Function Approximation: In this section we present the implementation of the method outlined in Section IV. First of all, in order to employ the PI scheme, the cost function of each policy should be acquired at each iteration. Thus, an approximation structure is employed, in the form of a Radial Basis Function (RBF) Approximation Structure (AS): $V^{(i,j)} = \phi^{\mathrm{T}}(p) w^{(i,j)} + \epsilon,\ i, j \in \mathbb{N}$, where $\phi^{\mathrm{T}}(p) w^{(i,j)}$ denotes the cost function approximation, ϵ denotes the approximation error, $\phi : \mathcal{W} \mapsto \mathbb{R}^n$ denotes a function basis, and $w^{(i,j)} \in \mathbb{R}^n$ denotes the weights of the AS[6]. The cost function is approximated through the HJB equation (8) as in [22], [25]:

$$\begin{gathered}\left(w^{(i,j)}\right)^{\mathrm{T}} \phi(p) u^{(i,j)}(p; p_d) + r\left(p, u^{(i,j)}; p_d\right) = 0 \Rightarrow \\ \left(w^{(i,j)}\right)^{\mathrm{T}} A^{(i,j)}(p; p_d) = B^{(i,j)}(p; p_d),\end{gathered} \quad (20)$$

which forms a linear regression problem, as the above expression is evaluated over a sufficiently rich (to render the respective matrix pseudo-invertible) number of collocation points $\{p_1^j, p_2^j, \cdots, p_K^j\},\ p_k^j \in \mathcal{B}_R^{(j)}, \forall k, j \in \mathbb{N}^+$.

2) Initial Policy Extraction: The proposed scheme requires an initial admissible policy for each tube $\mathcal{B}_R^{(j)}$ of the iterative process. In order to acquire such a policy, methods such as [17] or [13] can be employed, in order to provide a policy that restricts the motion of the robot within the tube $\mathcal{B}_R^{(j)}$. This can be accomplished with minor alterations of the aforementioned works, by defining the safety constraints over the boundary of the tube, instead of the workspace boundary.

3) Proposed Algorithm: The preceding elements are summarized in this section in Algorithm 1, where there are several modifications w.r.t. the method discussed in Section IV. Specifically, at each step of the policy iteration, the tube

[6]The cost function approximation should satisfy $V(p_d) = 0$, which can be enforced as a linear constrained. Nevertheless, in practice, the solutions to the approximation problem tend to satisfy this constraint without explicitly requiring it.

is altered based on a threshold, in order to avoid unecessary computations. Since a new tube essentially necessitates re-starting the PI scheme, recomputing the tube every time upon convergence would slow down the method. Therefore, for each tube, the PI is not implemented until convergence of the latter. Instead, in order to balance the effect of the above heuristic, a number of in-tube iterations is chosen in order to enable the method converging prior to updating the tube based on the most recent path.

Briefly, the method consists of running trajectories, computing the respective tube and optimizing the policy over the latter for a specific number of iterations. If the final trajectory results in a significantly altered tube, then the tube is recomputed and the above process is repeated. The whole scheme is iterated until convergence of the weights of the approximation structure, which is guaranteed owing to the universal approximation of RBFs, as well as the convergence of the cost function proven in Theorem 1.

V. Results

In this section we present simulations of the proposed method in synthetic workspaces. All simulations were carried out on a PC with 50Gb RAM and an Intel-i7 processor running Ubuntu version 18.04LTS in the environment of MATLAB 2022a. We evaluate our method's performance against the method in [19] and an RRT* one. We underline that the main advantage of our method compared to RRT* rests on its application to unknown workspaces; nevertheless the comparison with the latter is made in order to asses the overall optimality of the final result. Since RRT* provides a set of quasi-linear segments for known workspaces, the optimality of the final output of our algorithm is benchmarked against the latter. In order to compute the equivalent cost (4), the paths are imbued with the optimal velocity norm [18]. In Figs. 2, 3, results for a single initial/final position pair are depicted, in the form of the iterative trajectory generation, along with the results from [19], enhanced however with the optimal velocity norm, as defined in [18]. It is evident that our method produces superior results. In Fig. 4, a travelling salesman problem is solved, as described in [19], in order to demonstrate a use-case for the proposed method. The initial plan ($1 \rightarrow 2 \rightarrow 5 \rightarrow 6 \rightarrow 4 \rightarrow 3$ - which turns out to be the final one as well) along with the final plan trajectories are depicted. In Table I, the cost of the initial plan (with and without artificially imbuing the solution with the optimal velocity norm) are compared to the herein proposed scheme, as well as an RRT* method. For statistical significance, 50 trials were carried out for the letter. Our method is placed at a disadvantage, owing to the optimal velocity norm for the rest of the methods, however, its performance is nearly globally optimal, as can be evidenced through the RRT* results. Taking into account the fact that the results from the RRT* are non-smooth and that implementing the optimal velocity norm would be challenging in the RRT* case due to non-smoothness, the proposed scheme is shown to be very effective in providing the almost globally optimal policy for unknown workspaces.

Algorithm 1: Tube-Based PI Algorithm

Parameters: Weight convergence threshold $\epsilon_w > 0$, tube recomputing threshold $0 < \epsilon_{\mathcal{B}} < 1$, in-tube number of iterations $N_c \in \mathbb{N}$, tube radius $R > 0$.

• Given a Workspace $\mathcal{W}$, and a starting-ending configuration pair $\{\bar{p}, p_d\} \in \mathcal{W}$,

• Navigate from $\bar{p}$ towards p_d using [15],

• Set $j \leftarrow 1$, $i \leftarrow 0$ and weights_converged $\leftarrow$ FALSE, tube_has_changed $\leftarrow$ TRUE,

• Compute the tube $\mathcal{B}_R^{(j)}$, and the resulting initial policy $u^{(i,j)} \in \mathcal{A}(\mathcal{B}^{(j)})$,

while NOT(weights_converged) **do**

• $i \leftarrow 0$

if tube_has_changed **then**

• Compute the initial policy $u^{(i,j)} \in \mathcal{A}(\mathcal{B}^{(j)})$.

end if

for $i = 1 : N_c$ **do**

• Compute the linear regression matrices from (20),

• Acquire the cost function approximation for $V^{(i,j)}$ through the vector $w^{(i,j)}$ through solving (20),

• Acquire the next policy through (15) and (16), where $\nabla V^{(i,j)} \approx \nabla \Phi^{\mathrm{T}}(x) w^{(i,j)}$,

end for

• Compute the trajectory from the resulting input, i.e.: $\mathcal{P}_{u^{(N_c,j)}}$,

• $j \leftarrow j + 1$

• Compute the new tube $\mathcal{B}_R^{(j)} = \mathcal{B}_R\left(\mathcal{P}_{u^{(N_c,j)}}\right)$

if $\text{Area}\left(\mathcal{B}_R^{(j)} \setminus \mathcal{B}_R^{(j-1)}\right) / \text{Area}\left(\mathcal{B}_R^{(j-1)}\right) < \epsilon_{\mathcal{B}}$ **then**

• tube_has_changed $\leftarrow$ FALSE,

• $\mathcal{B}_R^{(j)} \leftarrow \mathcal{B}_R^{(j-1)}$,

else

• tube_has_changed $\leftarrow$ TRUE,

end if

if $\|w^{(N_c,j)} - w^{(N_c,j-1)}\| \leq \epsilon_w$ **then**

• weights_converged $\leftarrow$ TRUE,

end if

end while

• The optimal policy is $u^\star \approx u^{(N_c,j-1)}$

VI. Conclusion and Future Work

In this work, a novel tube-based PI-RL method for OMP in unknown, simply connected workspaces was introduced. Safety and optimality are inherited from previous works and the method is demonstrated to outperform the authors' previous work, as well as match an RRT* asymptotically optimal method. While the results of the proposed scheme are promising, a main limitation stems from the local optimality of the scheme in the presence of obstacles. Furthermore, while the optimal trajectories lie close to the boundary, this might be underisable in case of unknown workspaces due to the need to take more conservative paths to ensure safety in realistic applications. Future directions include, first and foremost, tackling internal obstacles, as well as including non-linear drift dynamics and dynamic environments.

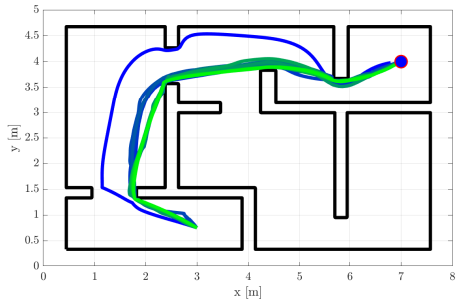

Fig. 2: Results for a Synthetic Workspace. The workspace boundary (black line), along with trajectories during the iterative PI scheme are presented. The initial trajectory (blue line) and the final one (green line), along with intermediate trajectories (colours between blue and green).

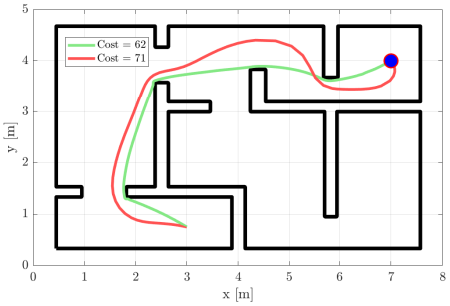

Fig. 3: Comparative Results between our method (green line) and [19] (red line). The respective costs are depicted in the legend.

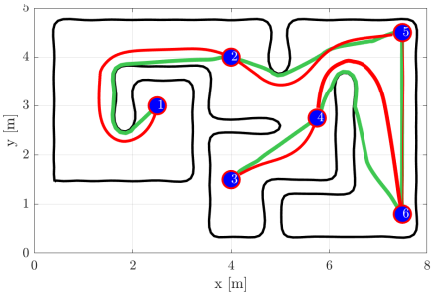

Fig. 4: Results for a Travelling Salesman problem in a Synthetic Workspace. The workspace boundary (black continuous line), initial trajectories (red lines) and final trajectories (green lines) for positions depicted through blue disks.

TABLE I: Comparative Results for Workspace of Fig. 4.

Path Index	Costs Ours	Costs Initial	Costs Initial Plan Optimal Vel	RRT*		
				min	mean	max
$1 \to 2$	17.12	301.06	19.72	**16.82**	17.52	18.27
$2 \to 5$	13.71	179.02	15.14	**13.54**	13.74	14.12
$5 \to 6$	13.68	150.67	**13.66**	13.68	13.76	13.81
$6 \to 4$	**10.44**	260.05	12.74	10.64	11.67	11.81
$4 \to 3$	**4.62**	50.12	4.8	**4.62**	4.65	4.73

REFERENCES

[1] D. E. Koditschek and E. Rimon, "Robot navigation functions on manifolds with boundary," *Advances in Applied Mathematics*, vol. 11, no. 4, pp. 412–442, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0196885890090017S

[2] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 846–894, 2011.

[3] L. Kavraki, P. Svestka, J.-C. Latombe, and M. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[4] X. Liu and D. Gong, "A comparative study of a-star algorithms for search and rescue in perfect maze," in *International Conference on Electric Information and Control Engineering*, 2011, pp. 24–27.

[5] N. Anastopoulos, K. Nikas, G. Goumas, and N. Koziris, "Early experiences on accelerating dijkstra's algorithm using transactional memory," in *Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium*, 2009.

[6] I. Noreen, A. Khan, and Z. Habib, "Optimal path planning using rrt* based approaches: A survey and future directions," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp. 1–16, 2016.

[7] 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.

[8] Z. Wang, Y. Li, H. Zhang, C. Liu, and Q. Chen, "Sampling-based optimal motion planning with smart exploration and exploitation," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2376–2386, 2020.

[9] J. Wang, M. Q.-H. Meng, and O. Khatib, "Eb-rrt: Optimal motion planning for mobile robots," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 4, pp. 2063–2073, 2020.

[10] C. Zhou, B. Huang, and P. Fränti, "A review of motion planning algorithms for intelligent robots," *Journal of Intelligent Manufacturing*, vol. 33, p. 387–424, 02 2022.

[11] E. Rimon and D. E. Koditschek, "Exact robot navigation using artificial potential functions," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 5, pp. 501–518, 1992.

[12] S. G. Loizou, "Closed form navigation functions based on harmonic potentials," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 6361–6366.

[13] P. Vlantis, C. Vrohidis, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Robot navigation in complex workspaces using harmonic maps," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1726–1731.

[14] P. D. Grontas, P. Vlantis, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Computationally efficient harmonic-based reactive exploration," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2280–2285, 2020.

[15] P. Rousseas, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Trajectory planning in unknown 2d workspaces: A smooth, reactive, harmonics-based approach," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1992–1999, 2022.

[16] ——, "Optimal robot motion planning in constrained workspaces using reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 6917–6922.

[17] P. Rousseas, C. Bechlioulis, and K. Kyriakopoulos, "Harmonic-based optimal motion planning in constrained workspaces using reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2005–2011, 2021.

[18] P. Rousseas, C. P. Bechlioulis, and K. J. Kyriakopoulos, "A continuous off-policy reinforcement learning scheme for optimal motion planning in simply-connected workspaces," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 247–10 253.

[19] ——, "Optimal motion planning in unknown workspaces using integral reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6926–6933, 2022.

[20] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Boletín de la Sociedad Matemática Mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[21] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2017.

[22] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[23] S. Koenig and M. Likhachev, "D*lite." in *Proceedings of the National Conference on Artificial Intelligence*, 2002, pp. 476–483.

[24] F.-Y. Wang and G. Saridis, "Suboptimal control for nonlinear stochastic systems," in *[1992] Proceedings of the 31st IEEE Conference on Decision and Control*, 1992, pp. 1856–1861 vol.2.

[25] M. Abu-Khalaf and F. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," *Automatica*, vol. 41, pp. 779–791, 05 2005.