



# Harmonic-Based Optimal Motion Planning in Constrained Workspaces Using Reinforcement Learning

Panagiotis Rousseas, Charalampos Bechlioulis , *Member, IEEE*, and Kostas J. Kyriakopoulos , *Fellow, IEEE*

**Abstract**—In this work, we propose a novel reinforcement learning algorithm to solve the optimal motion planning problem. Particular emphasis is given on the rigorous mathematical proof of safety, convergence as well as optimality w.r.t. to an integral quadratic cost function, while reinforcement learning is adopted to enable the cost function's approximation. Both offline and online solutions are proposed, and an implementation of the offline method is compared to a state-of-the-art RRT\* approach. This novel approach inherits the strong traits from both artificial potential fields, i.e., reactivity, as well as sampling-based methods, i.e., optimality, and opens up new paths to the age-old problem of motion planning, by merging modern tools and philosophies from various corners of the field.

**Index Terms**—Motion and path planning, optimization and optimal control, reinforcement learning.

## I. INTRODUCTION

THE motion planning problem has been at the heart of robotics ever since the early days of the field. Nevertheless, it has proven to be a tedious problem, due to certain aspects, including non-holonomic constraints and safety specifications, which restrict the motion of a robot and render most classical tools unfit for tackling it. Although specific approaches to these types of problems have arisen, there exist no feedback-based methods for optimal motion planning. On the other hand, sampling-based methods, while successful in their own merit, suffer from lack of smoothness, continuity and deterministic guarantees. In this work, we provide a framework that optimizes the motion of a robot in planar, fully known, constrained spaces with fixed internal obstacles. Reinforcement learning is fundamental to this approach since it allows us bypassing the problem of solving a hard non-linear partial differential equation. In particular, the successive approximation theory is adopted to extract the underlying cost function, based on which we formulate both an offline and an online framework, with provably safe policies, almost globally convergent to the goal position and optimal w.r.t an integral quadratic cost function.

Manuscript received October 15, 2020; accepted February 6, 2021. Date of publication February 19, 2021; date of current version March 9, 2021. This letter was recommended for publication by Associate Editor A. Kuntz and N. Amato upon evaluation of the reviewers' comments. (*Corresponding author: Charalampos Bechlioulis.*)

The authors are with the School of Mechanical Engineering, Control Systems Laboratory, National Technical University of Athens, Athens 15780, Greece (e-mail: rousseas.p@gmail.com; chmpechl@mail.ntua.gr; kkyria@mail.ntua.gr). Digital Object Identifier 10.1109/LRA.2021.3060711

## A. Related Work

The motion planning problem has been thoroughly addressed by many researchers. The proposed methods are generally classified as discrete methodologies, e.g., Configuration Space Decomposition [1], [2], Probabilistic Sampling, e.g., Rapidly Exploring Random Trees [3], [4], Probabilistic Roadmaps [5], [6] or other categories like Manifold Samples [7], Receding Horizon control [8], [9] and Path Homotopy Invariants [10], [11]. Alternatively, the Artificial Potential Fields (APFs) [12] method tackles both safety and convergence concerns through the form of the gradient of a potential field. Nonetheless, APFs exhibit unwanted equilibria on account of their construction and the workspace topology [13]. A family of APFs was introduced by Rimon and Koditschek [14], namely Navigation Functions (NF) that are applied in the context of sphere worlds. In that work, Rimon and Koditschek alleviated some of the issues of the APFs, however, these functions are tedious in their implementation requiring fine-tuning in order to diminish local minima. Artificial Harmonic Potential Fields (AHPFs) [15], [20] were also introduced in order to overcome the limitations of APFs. Such artificial potential fields are free of local minima by design, alleviating many of the shortcomings of previous NFs. The aforementioned works inspired us to propose a novel optimization algorithm for the motion planning problem in [23]. In the present work, the concept of harmonic functions will also be adopted to provide globally convergent policies, free of local minima. This will be in the context of harmonic series, suitably chosen to enable safe policies w.r.t. the workspace boundary. Besides yielding optimal, safe and convergent solutions, our method consists of a closed-loop policy for optimal motion planning, in contrast to the majority of the related approaches that are essentially open-loop solutions.

Furthermore, this work expands upon [23] by resolving its main limitation i.e., the need for a diffeomorphic map of the workspace onto a disk. Moreover, we adopt a wider basis set for the cost function approximation structure to span a larger functional space. Additionally, the cost function chosen in [23] involves a quantity that lacks physical meaning (such as energy input minimization) and is only used to ensure safety and convergence. On the contrary, in this work a classical integral quadratic cost function for energy input minimization is introduced. Finally, it should be highlighted that our key contribution is combining the strong-points of two fundamentally different

methodologies, namely sampling-based and AHPFs methods. While AHPFs methods lack in optimality and sampling based methods lack in reactivity, the proposed method encompasses both attributes through an appropriately selected structure of the input control policy, and an optimal (w.r.t. an integral quadratic cost function) choice of the underlying AHPF parameters.

### B. Background

Our approach leans heavily on AHPFs [13], [14]. These methods essentially create an artificial potential over the workspace, with high values on the boundary of the free space. The disadvantage of such methods stems from the need to guarantee a single global minimum at the desired position. Nevertheless, they have been successfully and widely applied with numerous enhancements, such as AHPFs [15], [16]. Concerning optimality and how it is introduced in this work, Lewis *et al.* ([25], [19]), have provided a way for successively estimating a classical cost function for a general, unknown non-linear system with mild assumptions. Our work finds common ground between them and builds for the first time a solution to the reactive optimal motion planning problem.

## II. PROBLEM FORMULATION

Consider a point robot<sup>1</sup> operating within a bounded and connected workspace  $\mathcal{G} \subset \mathbb{R}^2$  with inner distinct obstacles  $\mathcal{O}_i, i = 1, \dots, M$  and a desired position inside the free space,  $p_d \in \mathcal{W} - \partial\mathcal{W}$  with  $\mathcal{W} \triangleq \mathcal{G} - \cup_{i=1}^M \mathcal{O}_i$  denoting the free space with boundary  $\partial\mathcal{W}$ . Let also  $\partial\mathcal{G} = S_0(t) \in C^\infty : [0, 1] \rightarrow \mathbb{R}^2$  and  $\partial\mathcal{O}_i = S_i(t) \in C^\infty : [0, 1] \rightarrow \mathbb{R}^2, \forall i = 1, \dots, M$ , correspondingly denote a smooth parametrization of the outer boundary and the boundary of the internal obstacles, where  $C^\infty$  denotes the class of continuously differentiable functions. Let  $p = [x, y]^T \in \mathcal{W}$  be the state vector denoting the robot's position dictated by the single integrator model:

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

with  $u(t) : \mathbb{R} \rightarrow \mathbb{R}^2$  the control input (i.e., velocity command) and  $\bar{p} \in \mathcal{W}$  the initial position.

In this work, we consider the optimal motion planning problem, i.e., the problem of developing a control policy  $u$  that minimizes a cost function:

$$V(\bar{p}; p_d) = \int_0^\infty [Q(p(\tau; \bar{p}); p_d) + R(u(\tau))] d\tau \quad (2)$$

consisting of a state-related term  $Q(p(\tau; \bar{p}); p_d) = \alpha \|p(\tau; \bar{p}) - p_d\|^2$  and a control-input related term  $R(u(\tau)) = \frac{\beta}{2} \|u(\tau)\|^2$  with  $\|\cdot\|$  being Euclidean norm functions and  $p(t; \bar{p})$  denoting the solution of (1) from an initial state  $\bar{p}$  under the control policy  $u$ ,  $p_d$  being the goal position and  $\alpha, \beta > 0$  denoting weighting parameters.

<sup>1</sup>The results of this work can be readily employed for the navigation of disk robots with radius  $r > 0$ , by appropriately augmenting the workspace boundaries with the robot radius.

## III. HARMONIC SERIES APPROXIMATION

The reactive motion planning problem concerns essentially the design of a velocity vector field over the feasible workspace  $\mathcal{W}$ . In our work, the problem will be addressed as a Laplace problem, the solution of which will provide a suitable potential for navigation following its negated gradient. Additionally, von-Neumann conditions, i.e., conditions on the normal components of the gradient of the potential over the workspace boundary, will be imposed to ensure safety. Notice that the above formulation is not at all heuristic. The harmonic functions are solutions to the Laplace equation and exhibit desirable properties with respect to local minima, namely such points exist only on the boundary of the solution space. Consequently, we can ensure through the basis functions of the approximation structure both safety during navigation and convergence to the desired position, through the addition of an attractive term at the goal position  $p_d$  within the workspace. In that respect, the proposed solution comprises a series of harmonic functions with its parameters selected to satisfy both safety on the workspace boundary as well as optimality in view of (2).

### A. Smooth Boundaries

For workspaces with smooth boundaries, we will use the harmonic series approximation solution as provided in [21], where Trefethen demonstrates how harmonic series are employed to solve Laplace problems over smooth, multiply connected domains. In our work, we will adopt this type of functions to approximate the cost function (2). Studying the workspace as the complex plane, where  $z = x + iy \in \mathbb{C}$  with  $[x, y]^T = p \in \mathcal{W}$ , consider the following weighted series:

$$\begin{aligned} \hat{V}_i(z) = & -a_{0i} \ln(|z - c_i|) + \sum_{k=1}^{N_i} [a_{ki} \operatorname{Real}((z - c_i)^{-k})] \\ & + \sum_{k=1}^{N_i} [b_{ki} \operatorname{Imag}((z - c_i)^{-k})], \quad i = 1, \dots, M \end{aligned} \quad (3)$$

for each obstacle  $\mathcal{O}_i, i = 1, \dots, M$ , where  $c_i = c_{x,i} + ic_{y,i} \in \mathbb{C}$  with  $[c_{x,i}, c_{y,i}]^T \in \mathcal{O}_i - \partial\mathcal{O}_i$  and  $N_i$  denotes the number of adopted harmonic terms. For the goal position consider the term  $\ln(|z - c_0|)$  where  $c_0 = x_0 + iy_0 \in \mathbb{C}$  with  $[x_0, y_0]^T \triangleq p_d$  and for the outer boundary consider the weighted series:

$$\hat{V}_0(z) = \sum_{k=1}^{N_0} [a_{k0} \operatorname{Real}((z - c'_0)^k) + b_{k0} \operatorname{Imag}((z - c'_0)^k)]$$

where  $c'_0 = c_{x,0} + ic_{y,0} \in \mathbb{C}$  with  $[c_{x,0}, c_{y,0}]^T \in \mathcal{W} - \partial\mathcal{W}$ . In our case we select  $[c_{x,0}, c_{y,0}]^T = p_d$ . Concerning the above approximation structure, the constants  $c_i$  are essentially "centres" for the harmonic terms, placed appropriately inside the obstacles whereas the weights  $a_{ki}, b_{ki}$  are to be computed during the approximation process. Putting it all together, a harmonic potential  $\hat{V}_s(z)$  on a multiply connected space with

smooth boundary can be approximated through the series:

$$\hat{V}_s(z) \triangleq \sum_{i=0}^M \hat{V}_i(z) + a_{00} \ln(|z - c_0|), \quad a_{00} > 0 \quad (4)$$

with its gradient<sup>2</sup> calculated as:

$$\nabla \hat{V}_s(z) = \sum_{i=0}^M \nabla \hat{V}_i(z) + a_{00} \frac{z - c_0}{|z - c_0|^2} \quad (5)$$

Notice that the gradients  $\nabla \hat{V}_i(z)$  may be easily calculated through the properties  $\nabla \text{Real}(f(z)) = f'(z)$ ,  $\nabla \text{Imag}(f(z)) = i f'(z)$ . The aforementioned weighted series essentially provides a set of basis functions for the approximation of a potential function over a two-dimensional, multiply connected space. Therefore, the parameters involved in (4), namely  $a_{00}, a_{ki}, b_{ki}, i = 1, \dots, M$  and  $k = 1, \dots, N_i$  should be selected to satisfy both safety on the workspace boundary as well as optimality in view of (2).

### B. Non-Smooth Boundaries

For workspaces with non-smooth boundaries, such as corners, we will expand the above set of basis functions with a set of particular solutions provided by [22]. Consider a  $\theta_v \in (0, 2\pi)$ -angle corner at vertex  $p_v \in \partial\mathcal{W}$ . Then, an approximate particular solution is given as:

$$\hat{V}_p(z) = \sum_{i=1}^{N_{p_v}} A_i r^i \cos(i\bar{\theta}) + B_0 \bar{\theta} + \sum_{i=1}^{N_{p_v}} B_i \Phi_i(r, \bar{\theta}) \quad (6)$$

where  $r = \|p - p_v\|$ ,  $\bar{\theta} = \angle(p - p_v)$ ,  $p \in \mathcal{W}$  and

$$\Phi_i(r, \theta) = \begin{cases} \frac{r^i \sin(i\theta)}{\sin(i\theta_v)}, & \text{if } i\theta_v \neq k\pi, k = 1, 2, \dots \\ \frac{(-1)^k}{\theta_v} \phi'_i(r, \theta), & \text{if } i\theta_v = k\pi, \text{ for some } k, \end{cases} \quad (7)$$

where  $\phi_k(r, \theta) = r^k (\ln r \cos(k\theta) - \theta \sin(k\theta))$ ,  $k = 1, \dots, N_{p_v}$ . The gradient of these functions can be calculated analytically first in polar coordinates and then converted to cartesian coordinates. We note that in [22] analytic solutions for specific angles are provided (e.g.,  $\frac{\pi}{2}, \pi, \frac{3\pi}{2}$ ), which can be used as a set of basis instead of (6). Finally, the potential for the whole workspace with non-smooth boundaries is approximated as  $\hat{V}_s(z) + \hat{V}_p(z)$ .

## IV. OFFLINE METHODOLOGY

### A. Optimal Control Policy

In this section, we study the optimal motion planning problem as a classical optimal control problem. Based on the cost function (2), we define the associated Hamiltonian as follows:

$$H(p, u, \nabla V) = \nabla V^T(p; p_d)u + \alpha \|p - p_d\|^2 + \frac{\beta}{2} \|u\|^2 \quad (8)$$

Hence, the Hamilton-Jacobi-Bellman (HJB) optimality condition is given by  $H(p, u^*, \nabla V^*) = 0$ , while the optimal control

<sup>2</sup>The gradient here is expressed in terms of complex variables where  $\nabla \hat{V}_s(z) = [\text{Real}(\nabla \hat{V}_s(z)), \text{Imag}(\nabla \hat{V}_s(z))]^T$ .

policy is given by the stationary condition  $\frac{\partial H(p, u, \nabla V^*)}{\partial u} \big|_{u=u^*} = 0$ , as:

$$u^* = -\frac{1}{\beta} \nabla V^*(\bar{p}; p_d). \quad (9)$$

Notice that an analytical expression for the value function  $V^*(p; p_d)$  is necessary to compute the optimal policy  $u^*$ . Such an expression could be attained if one were to substitute the optimal control policy in the HJB optimality condition, and subsequently solve a rather hard non-linear partial differential equation with extra difficulty stemming from the fact that safety conditions need to be satisfied on the workspace boundary. Instead, the successive approximation technique [18] will be applied to adjust appropriately the weights of the aforementioned harmonic series (see Section III) so as to satisfy both  $H(p, u^*, \nabla V^*) = 0$  and safety over the workspace boundary.

### B. Safety Over the Workspace Boundary

The safety specification during a navigation task may be expressed as follows:

$$v^T(p)u(p) \geq 0, \forall p \in \partial\mathcal{W} \quad (10)$$

where  $u(p)$  is the underlying vector field dictating the robot motion and  $v(p)$  denotes the normal vector at each point of the boundary pointing inwards. Although the aforementioned inequality involves the whole set of points over the boundary  $\partial\mathcal{W}$  of the workspace, we prove that such property can be relaxed if inequality (10) holds strictly for a finite set of boundary points.

*Theorem 1:* Consider the boundary  $\partial\mathcal{W}$  of the workspace as well as a finite number of uniformly distributed points  $p_j \in \partial\mathcal{W}$ ,  $j = 1, \dots, N$  along with their respective normal vectors  $v_j = v(p_j)$ ,  $j = 1, \dots, N$  pointing inwards the workspace. There exists a number  $N_0 \in \mathbb{N}$  such that

$$v_j^T u_j > 0, \quad \forall j = 1, \dots, N \quad \text{with } N \geq N_0 \quad (11)$$

guarantees safety over the whole boundary  $\partial\mathcal{W}$  as described by (10).

*Proof:* Consider the parametrized curves  $S_i(t) : [0, 1] \rightarrow \partial\mathcal{W}$ ,  $i = 0, 1, \dots, M$  that define the boundary of the workspace. Then, for any pair of consecutive points on the workspace boundary  $p_j, p_{j+1}$  there exists  $t_j, t_{j+1} \in [0, 1]$  such that  $S_i(t_j) = p_j$ ,  $S_i(t_{j+1}) = p_{j+1}$ , for some  $i \in \{0, 1, \dots, M\}$ . Moreover, we may express the velocity field on this curve as  $u(S_i(t)), t \in [0, 1]$ . Consider now the Taylor series expansion of the velocity field and the parametrized curve around  $t_i$ , as:

$$u(t) = u(t_i) + (t - t_i) \partial_t u|_{t_i} + \mathcal{O}_2$$

$$S(t) = S(t_i) + (t - t_i) \partial_t S|_{t_i} + \mathcal{O}_2$$

$$\frac{\partial S(t)}{\partial t} \big|_{t_i} = \partial_t S|_{t_i} + (t - t_i) \partial_t^2 S|_{t_i} + \mathcal{O}_2 \quad (12)$$

where  $\mathcal{O}_2$  include all higher order terms. Note that these functions are analytic and thus the sum of the Taylor Series converges. Now consider that the normal vector to any point of the parametrized curve is expressed as  $v(t) = R \partial_t S|_t$  where  $R$  is a  $\pm \frac{\pi}{2}$  rotation matrix, depending on the parametrization of each

curve so that  $v$  points inwards the workspace. Applying (12) to (10) and disregarding the higher order terms, we get:

$$\begin{aligned} & [u(t_i) + (t - t_i) \partial_t u|_{t_i}]^T R [\partial_t S|_{t_i} + (t - t_i) \partial_t^2 S|_{t_i}] = \\ & = [u_i + (t - t_i) \partial_t u_i]^T R [\partial_t S_i + (t - t_i) \partial_t^2 S_i] = \\ & = z^2 (\partial_t u_i^T \partial_t^2 S_i) + z [\partial_t (u_i^T R \partial_t S_i)|_{t_i} + 2\Delta t \partial_t u_i \partial_t^2 S_i] + \\ & + [u_i^T R \partial_t S_i + \Delta t^2 \partial_t u_i \partial_t^2 S_i + \Delta t \partial_t (u_i^T R \partial_t S_i)|_{t_i}] = \\ & = z^2 a + zb + c \end{aligned}$$

where  $\Delta t = t_{i+1} - t_i$  and  $z \triangleq t - t_{i+1} \in [-\Delta t, 0]$ . Since

$$\begin{aligned} c &= u_i^T R \partial_t S_i + \Delta t^2 \partial_t u_i \partial_t^2 S_i + \Delta t \partial_t (u_i^T R \partial_t S_i)|_{t_i} \\ &= u_{i+1}^T v_{i+1} + \Delta t^2 \partial_t u_i \partial_t^2 S_i \end{aligned}$$

and  $u_{i+1}^T v_{i+1} > 0$  while  $\Delta t^2 \partial_t u_i \partial_t^2 S_i$  decreases as  $\Delta t$  tends to zero, there exists a small  $\Delta t > 0$  such that  $c > 0$ . Notice that this holds true since  $\lim_{\Delta t \rightarrow 0^+} (c) = u_i^T v_i > 0$ , irrespectively of the choice of  $\Delta t$ . Therefore, we conclude that there exists a non zero  $\Delta t$  and consequently a fine discretization of  $N_0$  points over the workspace boundary, such that  $z^2 a + zb + c \geq 0, \forall z \in [-\Delta t, 0]$ , which concludes the proof. ■

### C. Successive Approximation of the Value Function

In this subsection, we first define the notion of admissible control policy and then propose a method for successively approximating the cost function (2).

**Definition 1 (Admissible Control):** Let  $V(p)$  be a solution to (8). The control vector  $u(p)$  is admissible with respect to (2) on  $\mathcal{W}$ , if  $u(p)$  is continuous on  $\mathcal{W}$ ,  $u(p)$  stabilizes the system at  $p_d \in \mathcal{W}$ ,  $V(p)$  is finite for all  $p \in \mathcal{W}$  and  $u(p)$  is safe for all  $p \in \partial\mathcal{W}$ .

Following the successive approximation theory [17], the equation for approximating the control policy at each step- $i$  is:

$$u^{(i)}(p) = -\frac{1}{\beta} \nabla V^{(i-1)}(p), \quad (13)$$

where  $V^{(j-1)}$  is the cost function approximation at the previous step. Notice that the admissibility of the control policy is ensured through the specific form of the adopted harmonic series approximation structure (see Section III) as well as Theorem 1, which enables us to prove that after each iteration, the calculated policy  $u^{(i+1)}$  is safe if (11) holds true for a finite set of points over the workspace boundary.

**Theorem 2 (Control Improvement):** Under the current formulation for the cost function, if  $V^{(j+1)}$  is the unique positive-definite function that satisfies the equation  $H(p, u^{(i+1)}, V^{(i+1)}) = 0$ , then

$$V^*(p) \leq V^{(i+1)}(p) \leq V^{(i)}(p), \forall p \in \mathcal{W}.$$

**Proof:** Along the trajectories of  $\dot{p} = u^{(i+1)}(p)$ ,  $\forall p \in \mathcal{W}$  we have:

$$\begin{aligned} & V^{(i+1)}(p) - V^{(i)}(p) = \\ & - \int_0^\infty \left[ \nabla_p^T V^{(i+1)}(p(\tau)) - \nabla_p^T V^{(i)}(p(\tau)) \right] u^{(i+1)}(p(\tau)) d\tau \end{aligned} \quad (14)$$

Now expressing the HJB equations for two successive approximations and adding them yields:

$$\begin{aligned} & - \left( \nabla V^{(i+1)}(p) \right)^T u^{(i+1)} \\ & = \frac{\beta}{2} \|u^{(i+1)}\|^2 - \frac{\beta}{2} \|u^{(i)}\|^2 - \left( \nabla V^{(j)}(p) \right)^T u^{(i)} \end{aligned}$$

Substituting the above in (14) and invoking (13) yields:

$$\begin{aligned} & V^{(i+1)}(p) - V^{(i)}(p) \\ & = \int_0^\infty \left[ \frac{\beta}{2} \|u^{(i+1)}\|^2 - \frac{\beta}{2} \|u^{(i)}\|^2 + \beta u^{(i+1)T} (u^{(i)} - u^{(i+1)}) \right] d\tau \end{aligned}$$

Finally, invoking the mean value theorem, we arrive at:  $V^{(i+1)}(p) - V^{(i)}(p) \leq 0$  and by contradiction we can easily conclude that this approximation procedure **has** to be bounded from below by  $V^*(p)$ , which completes the proof. ■

### D. Offline Successive Approximation

In this work, similarly to [25],  $V^{(j)}(p)$  is approximated by:

$$V^{(i)}(p) = w_{(i)}^T \phi(p) \quad (15)$$

with  $\phi(p) \in \mathcal{C}^1(\mathcal{W}) : \mathbb{R}^2 \rightarrow \mathbb{R}^L$  and weights  $w_{(i)} \in \mathbb{R}^L$ . This formulation relates to Section III, as the elements of the harmonic series discussed therein are chosen as the basis functions in  $\phi(p)$ . Consequently, the parameter value vector  $w_{(i)}$  contains all the respective weights, namely  $a_{lm}, b_{lm}$  with  $l = 1, \dots, N_m$   $m = 0, 1, \dots, M$ , and therefore,  $L = \sum_{m=0}^M N_m$ . However, owing to the fact that the basis set tends to  $+\infty$  due to the logarithmic term located at the desired point  $p_d$ , we scale the control vector as  $u_{(i+1)} = -\frac{\|p - p_d\|^2}{\beta} \nabla^T \phi(p) w_{(i)}$ , which is in fact a well-defined stabilizing policy. Moreover, we express the safety condition (11) as a set of linear inequalities w.r.t the weights  $w_{(i)}$  of the approximation structure:

$$A^T w_{(i)} < 0, \quad (16)$$

where  $A = [A_1, \dots, A_N]$  with  $A_j = \nabla^T \phi(p_j) v(p_j)$ ,  $j = 1, \dots, N$ , which based on Theorem 1 results in safe policies for the whole boundary of the workspace. Therefore, the resulting policy remains always **admissible**, according to Definition 1, as long as  $A^T w_{(i)} < 0$ ,  $i = 0, 1, 2, \dots$

Finally, a successive approximation algorithm is designed to minimize the HJB error

$$w_{(i)}^T \nabla \phi(p) u^{(j)}(p) + \alpha \|p - p_d\|^2 + \frac{\beta}{2} \|u^{(j)}(p)\|^2 = e(w_{(i)}) \quad (17)$$

over a number of samples within the workspace, subject to (16). Fortunately, the constrained minimization problem is formulated as a constrained quadratic problem, which can be efficiently solved by various numerical approaches. The algorithmic process implementing all the above steps is described in Algorithm 1.



**Algorithm 1:** Offline Successive Approximation.

- **Boundary Sampling;** Select  $N$  points  $p_j \in \partial\mathcal{W}, j \in 1 \cdots N$  associated with the normal vectors  $v_j, j \in 1 \cdots N$ . Create the linear constraint matrix  $A = [A_1, \dots, A_N]$  with  $A_j = \nabla^T \phi(p_j) v(p_j), j = 1, \dots, N$ .
- **Workspace Sampling;** Select  $P$  points within the workspace  $\mathcal{W}$ .
- **Initialize;** Set  $i = 0$  and select an initial admissible control policy  $u_{(0)} = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) w_{(0)}$  by solving the quadratic constrained problem:

$$w_{(0)} = \underset{w}{\operatorname{argmin}} \{w^T w\}$$

$$\text{s.t. } A^T w \leq -\epsilon$$

for a small positive constant epsilon  $\epsilon$ .

**while** *Weights have not converged* **do**

- **Weight Improvement Step;**

$$w_{(i+1)} = \underset{w}{\operatorname{argmin}} \{w^T \bar{H}_{(i)}(p_j) w + \bar{B}_{(i)}(p_j; p_d) \cdot w\}$$

$$\text{s.t. } A^T w \leq -\epsilon$$

where  $\bar{H}_{(i)} \triangleq \sum_{j=1}^P \left( H_{(i)}(p_j) H_{(i)}^T(p_j) \right)$  and  $\bar{B}_{(i)}(p_j; p_d) \triangleq \sum_{j=1}^P \left( B_{(i)}(p_j; p_d) H_{(i)}^T(p_j) \right)$ , with  $H_{(i)}(p) = \nabla \phi(p) u_{(i)}(p)$  and  $B_{(i)}(p; p_d) = \alpha \|p - p_d\|^2 + \frac{\beta}{2} \|u_{(i)}(p)\|^2$ .

- **Policy Improvement Step;**

$$u^{(i+1)} = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) w_{(i)}$$

- $i \leftarrow i + 1$

**end**

Upon convergence :  $u^*(p) = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) w_{(i)}$

**Algorithm 2:** Online Value Iteration.

- **Initialize;** Select an initial admissible control policy  $u_{(0)} = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) w_{(0)}$  by solving the quadratic constrained problem:

$$w_{(0)} = \underset{w}{\operatorname{argmin}} \{w^T w\}$$

$$\text{s.t. } A^T w \leq -\epsilon$$

for a small positive constant epsilon  $\epsilon$ .

**while** *Weights have not converged* **do**

- **Policy Evaluation Step;** Solve the linear regression problem w.r.t.  $w^{(i+1)}$ :

$$w_{(i+1)}^T \phi(p(t)) =$$

$$= \int_t^{t+T} \left[ \alpha \|p(\tau; \bar{p}) - p_d\|^2 + \frac{\beta}{2} \|u(\tau)\|^2 \right] d\tau +$$

$$+ w_{(i)}^T \phi(p(t+T))$$

subject to the safety constraints:

$$A^T w_{(i+1)} \leq -\epsilon$$

- **Policy Improvement Step;** Update the policy as:

$$u^{(i+1)} = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) w^{(i+1)}$$

**end**

with  $T > 0$ , which has been proven equivalent to (8) (see [19]), we can formulate a valid online successive approximation algorithm that inherits the same properties proven in Section IV-C. In particular, we propose a Value Iteration (VI) scheme, utilizing the cost function approximation (19) as described in Algorithm 2. Note that the samples for the regression vectors are written in terms of points within the interval  $(t, t+T)$  but are essentially successive “measurements” in the robot’s trajectory. Furthermore, the integral of the cost function for the two points that need to be calculated in (19) requires the same samples. Hence, such measurements are combined in sufficiently large vectors for the regression to be numerically stable.

## VI. RESULTS

In this section, we present the results of the aforementioned methods. The proposed method is also compared against an RRT\* approach [24], which is based on sampling both the position as well as the control space, for the same cost function (2) (the weighting parameters of the cost function were set as  $\alpha = \beta = 0.5$ ) with the input velocity varying linearly w.r.t. time during any point to point transition in the workspace. For both spaces we used a  $100 \times 100$  grid with essentially no limit on maximum step size. All simulations were implemented in Matlab, on a PC running Windows 10, on an intel-i7 quad-core processor.

### A. Offline Results

*Smooth Boundary:* A smooth workspace was created, consisting of an elliptical outer boundary and star-shaped internal obstacles. In Figure 1(a) and Table I the offline method

## V. ONLINE METHODOLOGY

In this section, we present an online approach to address the optimal motion planning problem based on Integral Reinforcement Learning (IRL) [25]. The cost function to be approximated has the same form as in (2). Thus, invoking the approximation capabilities of (15), the cost function is approximated as follows:

$$V(p) = \hat{w}^T \phi(p) + \epsilon_V(p)$$

where  $\epsilon_V(p)$  denotes the modelling error. The admissible policy is also selected as follows:

$$u = u(p, \hat{w}) = -\frac{\|p-p_d\|^2}{\beta} \nabla \phi^T(p) \hat{w}. \quad (18)$$

Following the IRL formulation of the cost function:

$$V(p(t)) = V(p(t+T))$$

$$+ \int_t^{t+T} \left[ \alpha \|p(\tau; \bar{p}) - p_d\|^2 + \frac{\beta}{2} \|u(\tau)\|^2 \right] d\tau \quad (19)$$

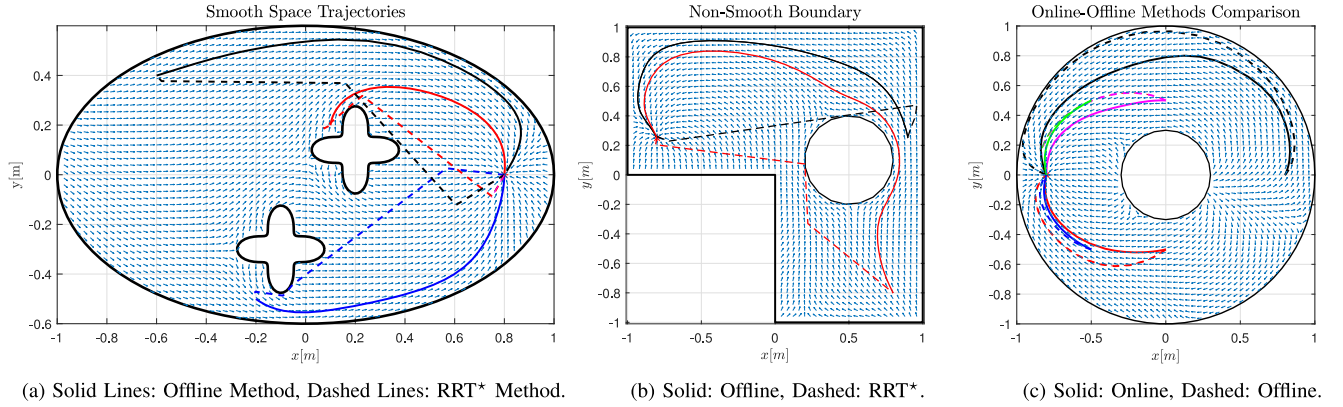


Fig. 1. Simulation Results: Different colors correspond to different starting positions. The normalized vector field is also depicted.

TABLE I  
COST OF RRT\* AND OUR METHOD - SMOOTH BOUNDARY

Initial Pos.[m]	Our Method	RRT* Mean	RRT* S. Dev	RRT* Trials
$(-0.2, -0.5)^T$	0.835	1.042	0.05659	10
$(0.1, 0.2)^T$	0.429	0.604	0.06690	10
$(-0.6, 0.4)^T$	1.525	1.621	0.03790	13

TABLE II  
COST OF RRT\* AND OUR METHOD - NON-SMOOTH BOUNDARY

Initial Pos.[m]	Our Method	RRT* Mean	RRT* S. Dev	RRT* Trials
$(0.8, -0.8)^T$	4.162	3.929	0.6418	10
$(0.9, 0.25)^T$	2.748	3.641	0.6170	13

is contrasted to the RRT\* method. The goal position is set at the coordinates  $[0.8, 0]^T$ , starting from the points  $[-0.2, 0.5]^T$ ,  $[0.1, 0.2]^T$  and  $[-0.6, 0.4]^T$ .

As seen in Table I, the offline approach outperforms the RRT\* method. From the respective figures it is evident that, while the behaviour is similar in both cases, with similar distance travelled, our method prevails with respect to the required control effort. To elaborate more on this, notice that while the RRT\* method can easily find straight paths in the workspace, it is very unlikely that these paths also exhibit the optimal control policy for each point.

*Non-Smooth Boundary:* An L-shaped workspace with corners and one internal disk obstacle was created. In Figure 1(b) and Table II the offline method is contrasted to the RRT\* method. The goal position is set at the coordinates  $[-0.8, 0.25]^T$ , starting from the points  $[0.8, -0.8]^T$  and  $[0.9, 0.25]^T$ .

Our method operates better, or similarly to the RRT\* method in this case. As previously, while our method results in longer trajectories, a more “efficient” control input selection results in diminished cost.

*Remarks:* The proposed offline methodology outperforms an RRT\* method, which converges asymptotically in probability to the optimal solution as the sampling becomes denser. Our method however exhibits the following advantages. The solutions are smooth, but more importantly, our method computes an optimal control policy with respect to the selected parametrization for **every** initial configuration, in contrast to the RRT\*

TABLE III  
RESULTED COST

Run #	Initial Pos.[m]	Our Method	HAPF	NF
1	$(-0.2, -0.5)^T$	0.835135	1.0302	2.2581
2	$(0.1, 0.2)^T$	0.428804	1.2091	8.1184
3	$(-0.6, 0.4)^T$	1.524973	2.1763	2.89

TABLE IV  
COMPARATIVE SIMULATION RESULTS

Traj. #	Start Pos.[m]	Cost Online	Cost Offline
1	$(0, 0.5)$	0.5781	0.6608
2	$(0, -0.5)$	0.6576	0.7611
3	$(0.8, 0)$	2.9577	3.5718
4	$(-0.5, 0.5)$	0.2058	0.2110
5	$(-0.5, -0.5)$	0.2322	0.2300

method in which a single run computes a single trajectory for a specific starting point.

### B. Comparison With Potential-Based Methods

Finally, we compare our method against two existing potential-based methods, namely NFs [26] and HAPFs [16]. The results on the workspace illustrated in Fig. 1(a) are given in Table III. Our method outperforms both methods, which is to be expected as no optimality is considered in the latter.

### C. Online Results

The classical annulus problem was considered. The online training was implemented starting from a 10 % deviation from the optimal parameter values. The trajectories were implemented by applying an excitation signal to cover the workspace. Convergence was achieved within 500 iterations and the cost function was successfully approximated. The results of the training process are summarized in Fig. 1(c). In this case, the online method was able to improve upon the cost of the offline methodology, finding locally optimal solutions w.r.t. the cost function (see Table IV). The local improvement is owing to the cost improvement from Theorem 2, which is expected as the online method did not cover the entire workspace. Hence, the offline method is

expected to outperform the online one for initial configurations close to, or inside, the subsets of the workspace that were not sufficiently covered.

## VII. LIMITATIONS AND FUTURE WORK

While our method presents clear advantages over other existing ones, there are some caveats to it. Firstly, the method has been developed for single integrator dynamics. Thus, any other model, noise, or stochastic elements will render it inoperable. Furthermore, it can only be applied in the context of two-dimensional navigation and only for Euclidian components. Additionally, as the number of obstacles increase, the problem of finding safe and optimal vector fields becomes increasingly difficult. Finally, moving obstacles are also not addressed by our approach. Many of these constraints are addressed by existing algorithms -e.g. RRT\*- , nevertheless we believe that our method presents a step towards a right direction for the future of motion planning. Therefore, and given the aforementioned promising results, we intend to expand the present ideas for unknown workspaces, with the robot implementing policies that exhibit safety, convergence and optimality. The advantages of our approach in tackling such a significant problem are evident, providing robust yet applicable solutions via reinforcement learning. We further intend to expand our method to higher dimensions (beyond two), and research stochastic and non-linear dynamics. Finally, we intend to render our method more robust as the number of obstacles increases, thus addressing more effectively navigation problems in obstacle populated problems such as logistics etc.

## REFERENCES

- [1] J. T. Schwartz and M. Sharir, "On the 'piano movers' problem I: The case of a two-dimensional rigid polygonal body moving amidst polygonal barriers," *Commun. Pure Appl. Math.*, vol. 36, pp. 345–398, 1983.
- [2] J. Canny, *The Complexity of Robot Motion Planning*. Cambridge, MA, USA: MIT Press, 1988.
- [3] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 846–894, 2011.
- [4] Z. Kingston, M. Moll, and L. E. Kavraki, "Sampling-based methods for motion planning with constraints," *Annu. Rev. Control, Robot., Auton. Syst.*, vol. 1, no. 1, pp. 159–185, 2018.
- [5] L. E. Kavraki, P. Svestka, J. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Trans. Robot. Automat.*, vol. 12, no. 4, pp. 566–580, Aug. 1996.
- [6] R. Bohlin and L. E. Kavraki, "Path planning using lazy prm," in *IEEE Proc. Millennium Conf. Int. Conf. Robot. Automat. Symp. Proc.*, vol. 1, Apr. 2000, pp. 521–528.
- [7] O. Salzman, M. Hemmer, and D. Halperin, "On the power of manifold samples in exploring configuration spaces and the dimensionality of narrow passages," *IEEE Trans. Automat. Sci. Eng.*, vol. 12, no. 2, pp. 529–538, Apr. 2015.
- [8] P. Ogren and N. E. Leonard, "A convergent dynamic window approach to obstacle avoidance," *IEEE Trans. Robot.*, vol. 21, no. 2, pp. 188–195, Apr. 2005.
- [9] S. Kousik, S. Vaskov, F. Bu, M. Johnson-Roberson, and R. Vasudevan, "Bridging the gap between safety and real-time performance in receding-horizon trajectory design for mobile robots," *CoRR*, 2018, [arXiv:1809.06746](https://arxiv.org/abs/1809.06746).
- [10] S. Bhattacharya and R. Ghrist, "Path homotopy invariants and their application to optimal trajectory planning," *Ann. Math. Artif. Intell.*, Aug. 2018, [arXiv:1710.02871](https://arxiv.org/abs/1710.02871).
- [11] J. Gregoire, M. Čáp, and E. Frazzoli, "Locally-optimal multi-robot navigation under delaying disturbances using homotopy constraints," *Auton. Robots*, vol. 42, no. 4, pp. 895–907, Apr. 2018.
- [12] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Proc. IEEE Int. Conf. Robot. Automat.*, vol. 2, Mar. 1985, pp. 500–505.
- [13] D. Koditschek, "Exact robot navigation by means of potential functions: some topological considerations," in *Proc. IEEE Int. Conf. Robot. Automat.*, vol. 4, Mar. 1987, pp. 1–6.
- [14] E. Rimon and D. Koditschek, "Exact robot navigation using artificial potential fields," *IEEE Trans. Robot. Automat.*, vol. 8, no. 5, pp. 501–518, Oct. 1992.
- [15] S. G. Loizou, "Closed form navigation functions based on harmonic potentials," in *Proc. 50th IEEE Conf. Dec. Control Eur. Control Conf.*, 2011, pp. 6361–6366.
- [16] C. Vrohidis, P. Vlantis, C. P. Bechlioulis and K. J. Kyriakopoulos, "Robot navigation in complex workspaces using harmonic maps," *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 1726–1731.
- [17] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, pp. 779–791, 2005.
- [18] C. S. Lee and G. Saridis, "An approximation theory of optimal control for trainable manipulators," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 3, pp. 152–159, 1979.
- [19] F. L. Lewis, D. L. Vrabie, and V. L. Syrmos, *Optimal Control*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc, 2012.
- [20] J. O. Kim and P. K. Khosla, "Real-time obstacle avoidance using harmonic potential functions," *IEEE Trans. Robot. Automat.*, vol. 8, no. 3, pp. 338–349, Jun. 1992.
- [21] N. Trefethen and D. Lloyd, "Series solution of laplace problems," *ANZIAM J.* vol. 60, 2018, pp. 1–26.
- [22] Zi-Cai Li, Tzon-Tzer Lu, Hsin-Yun Hu, Alexander H. D. Cheng, "Particular solutions of laplace's equations on polygons and new models involving mild singularities," *Eng. Anal. Boundary Elements*, vol. 29, 2005, pp. 59–75.
- [23] P. Rousseas, C. Bechlioulis, K. Kyriakopoulos, "Optimal motion planning in constrained workspaces using reinforcement learning," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* Oct. 2020, pp. 25–29.
- [24] I. Noreen, A. Khan, Z. Habi, "Optimal Path Planning using RRT\* based Approaches: A Survey and Future Directions," *Int J. Advanced Comput. Sci. Appl.*, vol. 7, doi: [10.14569/IJACSA.2016.071114](https://doi.org/10.14569/IJACSA.2016.071114).
- [25] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Jul.–Sep. 2009.
- [26] E. Rimon, and D. E. Koditschek, "Exact robot navigation using artificial potential functions," *IEEE Trans. Robot. Automat.*, vol. 8, no. 5, pp. 501–518, Oct. 1992.