

Reinforcement Learning for Solving Control Problems in Robotics [†]

Askhat Diveev ^{1,*}, Elena Sofronova ¹, Sergey Konstantinov ² and Viktoria Moiseenko ²

¹ Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44/2 Vavilova str., Moscow 119333, Russia; sofronova_ea@mail.ru

² RUDN University, 6 Miklukho-Maklaya str., Moscow 117198, Russia; svkonstantinov@mail.ru (S.K.); vika_moiseenko1999@mail.ru (V.M.)

* Correspondence: aidiveev@mail.ru; Tel.: +7-905-711-4427

[†] Presented at the 15th International Conference “Intelligent Systems” (INTELS’22), Moscow, Russia, 14–16 December 2022.

Abstract: The use of reinforcement learning technology for the optimal control problem solution is considered. To solve the optimal control problem an evolutionary algorithm is used that finds control to ensure the movements of a control object along different trajectories with approximately the same values of the quality criterion. Additional conditions for passing the trajectory in the neighbourhood of given areas of the state space are included in the quality criterion. To build a stabilization system for the movement of an object along a given trajectory, machine learning control by symbolic regression is used. An example of solving the optimal control problem for a quadcopter is given.

Keywords: optimal control; evolutionary algorithm; reinforcement learning; symbolic regression

1. Introduction

The optimal control problem with phase constraints often has a multi-modal functional. Therefore, with its numerical solution by direct approach, it is possible to obtain several control functions that ensure the movement of the object along different trajectories in the state space with approximately the same value of the control quality criterion which is close to the optimal.

A numerical solution to the optimal control problem leads to some difficulties. As a rule, in most optimal control problems, it is necessary to minimize not one but at least two criteria, reach the control goal or minimize the error of reaching the terminal state and still minimize the given quality criterion. Addition of weight coefficients into criteria does not significantly simplify the problem, since the problem of choosing weights arises.

Another search problem is defined as the loss of unimodality of the functional on the space of parameters of the approximating function. Even a piecewise linear approximation of the control function, when only one parameter needs to be found on each interval for each control component, does not guarantee the presence of a single minimum of the goal functional on the space of parameters.

The problem becomes more complicated in the presence of phase constraints that describe the areas of state space forbidden for the optimal trajectory. It is most likely that due to these reasons, and despite numerous attempts [1,2], a universal computational method for the optimal control problem has not been created.

Further studies have shown that if a strictly optimal solution is not needed and solutions close to the optimal are quite satisfactory, then evolutionary algorithms can be successfully applied to the optimal control problems [3].

Sometimes in practice the researcher knows how the object should move along the optimal trajectory, i.e., approximately knows the areas in the state space the optimal trajectory should pass through. If we introduce additional requirements in the form of passing



Citation: Diveev, A.; Sofronova, E.; Konstantinov, S.; Moiseenko, V. Reinforcement Learning for Solving Control Problems in Robotics. *Eng. Proc.* **2023**, *33*, 29. <https://doi.org/10.3390/engproc2023033029>

Academic Editors: Ivan Zelinka, Arutun Avetisyan and Alexander Ilin

Published: 15 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

through the given areas into the quality criterion, then the evolutionary algorithm should change the search area and look for a solution that satisfies the additional requirements. This approach is effective when using evolutionary algorithms. Due to the inheritance property, the improvement of the criterion value at each generation is performed on the basis of small evolutionary transformations of possible solutions to the previous generation. Therefore, if at some generation one of the possible solutions passes through the required areas specified by the researcher, then with a high probability the evolutionary algorithm will search for the optimal solution that preserves the obtained properties. A similar technique is used in machine learning with reinforcement [4,5], when the researcher awards the object by the change of the target functional value for the right actions. Currently, reinforcement learning is actively used in the practice of solving control problems [6]. The paper contains a formal description and practical application of reinforcement learning for solving the optimal control problem.

2. The Optimal Control Problem and Reinforcement Learning

Consider a formal statement of the optimal control problem.

The mathematical model of a control object is given in the Cauchy form of an ordinary differential equation system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \quad (1)$$

where $\mathbf{x}$ is a state space vector, $\mathbf{u}$ is a control vector, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in U \subseteq \mathbb{R}^m$, and U is a compact set.

The initial state is given by

$$\mathbf{x}(0) = \mathbf{x}^0 \in \mathbb{R}^n. \quad (2)$$

The terminal state is given by

$$\mathbf{x}(t_f) = \mathbf{x}^f \in \mathbb{R}^n, \quad (3)$$

where t_f is the time to reach the terminal state (3). Time t_f is not given, but it is limited to $t_f \leq t^+$, where t^+ is a given positive value.

The quality criterion is given by

$$J = \int_0^{t_f} f_0(\mathbf{x}, \mathbf{u}) dt \rightarrow \min_{\mathbf{u} \in U}. \quad (4)$$

Assume that the researcher knows the areas in the state space of where the optimal trajectory should be. Then, additional conditions are included in the quality criterion

$$J_1 = \int_0^{t_f} f_0(\mathbf{x}, \mathbf{u}) dt + p \sum_{i=1}^r \psi_i(\mathbf{x}(t)) \rightarrow \min_{\mathbf{u} \in U}, \quad (5)$$

where

$$\psi_i(\mathbf{x}) = \vartheta(\min\{\|\mathbf{z}^i - \mathbf{x}\|\} - \varepsilon_i)(\min\{\|\mathbf{z}^i - \mathbf{x}\|\} - \varepsilon_i), \quad (6)$$

p is a penalty coefficient, and $\vartheta(\alpha)$ is a Heaviside step function

$$\vartheta(\alpha) = \begin{cases} 1, & \text{if } \alpha > 0 \\ 0, & \text{otherwise} \end{cases}, \quad (7)$$

$\varepsilon_i, i = 1, \dots, r$ are given small positive values, and $\mathbf{z}^i, i = 1, \dots, r$ are the centres of known areas.

According to the introduced additional conditions, if an optimal trajectory does not pass near some given point $\mathbf{z}^i$, then value of criterion (5) will grow.

3. Computation Experiment

Consider the optimal control problem for the spatial movement of a quadcopter. The mathematical model of the control object is

$$\begin{aligned}\dot{x}_1 &= x_4, \\ \dot{x}_2 &= x_5, \\ \dot{x}_3 &= x_6, \\ \dot{x}_4 &= u_4(\sin(u_3) \cos(u_2) \cos(u_1) + \sin(u_1) \sin(u_2)), \\ \dot{x}_5 &= u_4 \cos(u_3) \cos(u_1) - g_c, \\ \dot{x}_6 &= u_4(\cos(u_2) \sin(u_1) - \cos(u_1) \sin(u_2) \sin(u_3)),\end{aligned}\quad (8)$$

where $g_c = 9.80665$.

Control is constrained

$$u_i^- \leq u_i \leq u_i^+, \quad i = 1, \dots, 4, \quad (9)$$

where $u_1^- = -\pi/12$, $u_1^+ = \pi/12$, $u_2^- = -\pi$, $u_2^+ = \pi$, $u_3^- = -\pi/12$, $u_3^+ = \pi/12$, $u_4^- = 0$ and $u_4^+ = 12$.

The initial state is given by

$$\mathbf{x}^0 = [0 \ 5 \ 0 \ 0 \ 0 \ 0]^T. \quad (10)$$

The terminal state is given by

$$\mathbf{x}^f = [10 \ 5 \ 10 \ 0 \ 0 \ 0]^T. \quad (11)$$

The phase constraints are given by

$$\varphi_k(\mathbf{x}) = r_k - \sqrt{(x_1 - x_{k,1})^2 + (x_3 - x_{k,3})^2} \leq 0, \quad (12)$$

where $k = 1, 2$, $r_1 = r_2 = 2$, $x_{1,1} = 2.5$, $x_{1,3} = 2.5$, $x_{2,1} = 7.5$, $x_{2,3} = 7.5$.

It is necessary to find a control function, taking into account the constraints in (9), that minimizes the following criterion

$$J_1 = t_f + p_1 \sum_{k=1}^2 \int_0^{t_f} \vartheta(\varphi_k(\mathbf{x})) dt \rightarrow \min_{\mathbf{u} \in \mathbf{U}}, \quad (13)$$

where $t_f \leq t^+ = 5.6$, $p_1 = 3$.

To solve the control problem numerically, let us use a piecewise linear approximation. The time axis is divided into equal intervals Δt , and the search for constant parameters is performed at the interval boundaries for each control component. Control is a piecewise linear function that consists of segments connecting points at the bounds of intervals. Given the control constraints, the desired control function is as follows

$$u_i = \begin{cases} u_i^+, & \text{if } \hat{u}_i \geq u_i^+ \\ u_i^-, & \text{if } \hat{u}_i \leq u_i^- \\ \hat{u}_i, & \text{otherwise} \end{cases}, \quad i = 1, \dots, 4, \quad (14)$$

where

$$\hat{u}_i = d_{i+(j-1)m} + (d_{i+jm} - d_{i+(j-1)m}) \frac{t - (j-1)\Delta t}{\Delta t}, \quad i = 1, \dots, 4, j = 1, \dots, K, \quad (15)$$

and K is a number of time interval boundaries

$$K = \left\lfloor \frac{t^+}{\Delta t} \right\rfloor + 1 = \left\lfloor \frac{5.6}{0.4} \right\rfloor + 1 = 15. \quad (16)$$

When solving the problem by a direct approach, the condition of reaching the terminal state is included in the quality criterion

$$J_2 = t_f + p_1 \sum_{k=1}^2 \int_0^{t_f} \vartheta(\varphi_k(\mathbf{x})) dt + p_2 \|\mathbf{x}^f - \mathbf{x}(t_f)\| \rightarrow \min_{\mathbf{u} \in U}, \quad (17)$$

where $p_2 = 1$,

$$t_f = \begin{cases} t, & \text{if } t \leq t^+ \text{ and } \|\mathbf{x}^f - \mathbf{x}(t)\| \leq \varepsilon = 0.01 \\ t^+, & \text{otherwise} \end{cases}. \quad (18)$$

To solve the problem, a hybrid evolutionary algorithm [7] is used.

Figures 1 and 2 show projections on the horizontal plane $\{x_1; x_3\}$ of the two found optimal trajectories. The big circles present the phase constraints in (12).

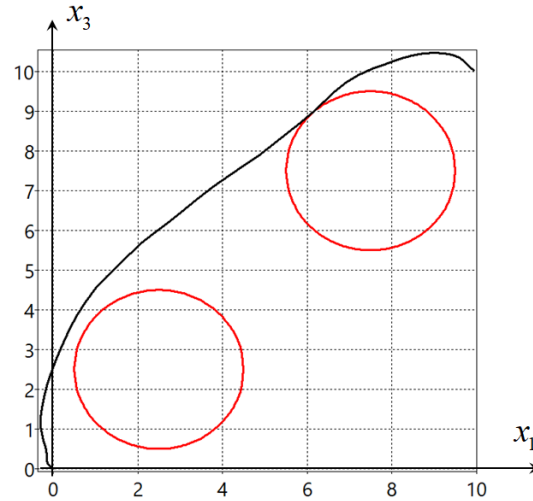


Figure 1. Projection of optimal trajectory 1 on the horizontal plane.

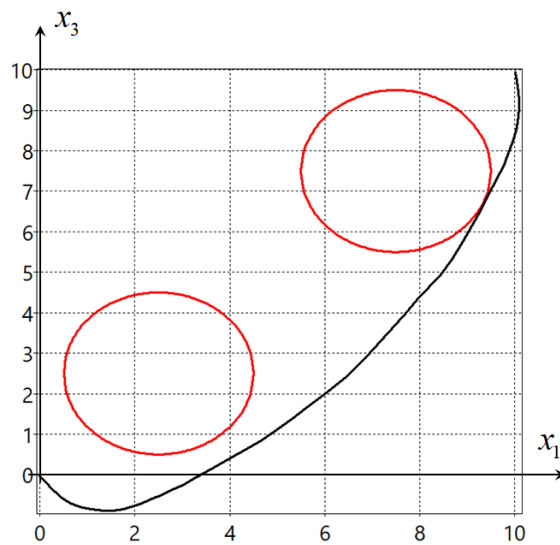


Figure 2. Projection of optimal trajectory 2 on the horizontal plane.

The criterion for the solutions found had the following values: for the solution in Figure 1 $J_2 = 5.6434$, for the solution in Figure 2 $J_2 = 5.6330$.

As can be seen from the experiment, the values of the criteria practically coincide, the solutions found ensure the movement of the object from the given initial state (10) to the given terminal state (11) without violation of the phase constraints. In the series of experiments, the hybrid evolutionary algorithm found solutions that bypass the phase constraints either from above, as in Figure 1, or from below, as in Figure 2.

Suppose that we need the control object to move between obstacles. For this purpose the desired areas on the horizontal plane are defined. It is known that in the presence of interfering phase constraints, the optimal trajectory should be close to the boundary of these constraints. For the given problem four desired areas are defined as

$$\begin{aligned} \mathbf{z}^1 &= [2.5 \ 0.4]^T, \ \varepsilon_1 = 0.6, \\ \mathbf{z}^2 &= [4.5 \ 2.5]^T, \ \varepsilon_2 = 0.6, \\ \mathbf{z}^3 &= [5.5 \ 7.5]^T, \ \varepsilon_3 = 0.6, \\ \mathbf{z}^4 &= [7.5 \ 9.6]^T, \ \varepsilon_4 = 0.6. \end{aligned} \quad (19)$$

The conditions for passing through the desired areas (19) are included in the quality criterion

$$J_3 = t_f + p_1 \sum_{k=1}^2 \int_0^{t_f} \vartheta(\varphi_k(\mathbf{x})) dt + p_2 \|\mathbf{x}^f - \mathbf{x}(t_f)\| + p_3 \sum_{i=1}^4 \psi_i(\mathbf{x}) \rightarrow \min_{\mathbf{u} \in U}, \quad (20)$$

where $p_3 = 3$.

The hybrid evolutionary algorithm found the following optimal solution $\mathbf{d} = [d_1 \dots d_{60}]^T = [-11.1092, 5.9957, -0.0532, 7.0045, 17.5091, 18.1172, -1.6764, 18.7012, -16.0121, 10.5543, -19.9307, 12.1721, -6.0892, 0.5339, -0.8616, 19.2556, -13.7218, 15.1266, 0.3982, 14.2650, -1.1768, 2.9832, 4.3286, 15.1508, -8.9240, -19.6814, 4.5363, 15.9879, -0.0026, 1.1203, 13.2592, -6.6358, -6.2012, -0.5328, -0.0354, 4.2548, 11.6764, -4.3345, -6.7336, 19.8643, 0.3360, -8.9741, -2.6648, 12.5608, 19.6577, -19.9308, -1.6252, 19.3797, -1.1954, 2.2625, 5.9582, 16.0807, -0.8272, 2.3167, 0.9842, 14.2695, -6.3767, 2.3895, 0.3742, 16.2710]^T$.

Figure 3 shows the projection of the found optimal trajectory for the solution with quality criterion $J_3 = 5.5730$. Small dashed circles are the desired areas, while big circles are the constraints.

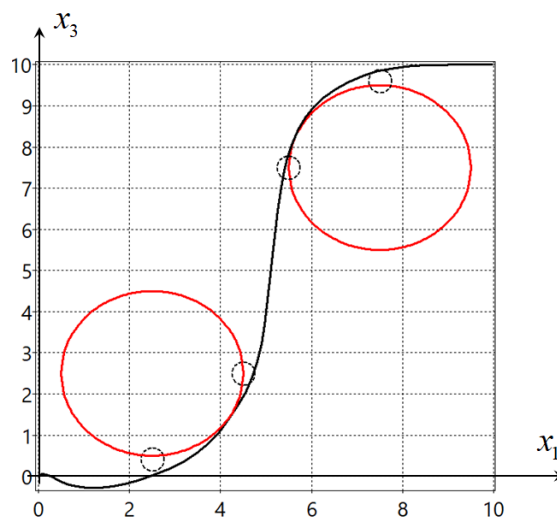


Figure 3. Projection on the horizontal plane of the optimal trajectory found by reinforcement learning.

To implement the obtained solution according to the extended statement of the optimal control problem, it is necessary to build a system to stabilize the movement of the object along the optimal trajectory [8]. For this purpose, machine learning control is used [9]. The control function structure search is carried out by symbolic regression [10].

The obtained solution is

$$u_i = \begin{cases} u_i^+, & \text{if } \tilde{u}_i \geq u_i^+ \\ u_i^-, & \text{if } \tilde{u}_i \leq u_i^- \\ \tilde{u}_i, & \text{otherwise} \end{cases}, i = 1, \dots, 4, \quad (21)$$

where

$$\tilde{u}_1 = \mu(G), \quad (22)$$

$$\tilde{u}_2 = (\tilde{u}_1 - \tilde{u}_1^3) \rho_{17}(A + \mu(G)) \vartheta(F) \rho_{17}(x_4^* - x_4), \quad (23)$$

$$\tilde{u}_3 = \tilde{u}_2 + \tanh(\tilde{u}_1) + \rho_{19}(A + \mu(G)) + \rho_{17}(W), \quad (24)$$

$$\tilde{u}_4 = \tilde{u}_3 + \ln|\tilde{u}_2| + \operatorname{sgn}(A + \mu(G)) \sqrt{|A + \mu(G)|} + \rho_{19}(A) + \arctan(C) + \operatorname{sgn}(E) + \arctan(F) + \exp(q_2(x_2^* - x_2)) + \sqrt{q_1}, \quad (25)$$

$$A = B \operatorname{sgn}(D) \sqrt{|D|} \tanh(D) \exp(H), \quad B = \exp(C) + \rho_{17}(F) + \cos(q_6(x_6^* - x_6)),$$

$$C = D + \tanh(E) + \rho_{18}(V), \quad D = E + \sqrt[3]{F} + \sin(W),$$

$$E = F + G + \exp(H) - V, \quad F = H + \operatorname{sgn}(x_5^* - x_5) + (x_2^* - x_2)^3,$$

$$G = q_6(x_6^* - x_6) + q_3(x_3^* - x_3) + \operatorname{sgn}(x_2^* - x_2) \sqrt{|x_2^* - x_2|},$$

$$H = \rho_{17}(q_6(x_6^* - x_6) + q_3(x_3^* - x_3)) + V^3 + W + q_6 q_5^2 (x_5^* - x_5)^2 + (x_5^* - x_5)^2,$$

$$V = \sin(q_6(x_6^* - x_6)) + q_5(x_5^* - x_5) + q_2(x_2^* - x_2) + \cos(q_1) + \exp(x_5^* - x_5) + \vartheta(x_2^* - x_2),$$

$$W = q_4(x_4^* - x_4) + q_1(x_1^* - x_1) + \sin(q_6),$$

$$\mu(\alpha) = \begin{cases} \alpha, & \text{if } |\alpha| \leq 1 \\ \operatorname{sgn}(\alpha), & \text{otherwise} \end{cases}, \quad \rho_{17}(\alpha) = \operatorname{sgn}(\alpha) \ln(|\alpha| + 1),$$

$$\rho_{18}(\alpha) = \operatorname{sgn}(\alpha) (\exp(|\alpha|) - 1), \quad \rho_{19}(\alpha) = \operatorname{sgn}(\alpha) \exp(-|\alpha|),$$

where $q_1 = 13.02930$, $q_2 = 11.21509$, $q_3 = 15.91016$, $q_4 = 14.33447$, $q_5 = 14.67798$, $q_6 = 9.91431$ and $\mathbf{x}^* = [x_1^* \dots x_6^*]^T$ is a state vector of the reference model.

Figure 4 shows the trajectories from eight initial states on the horizontal plane.

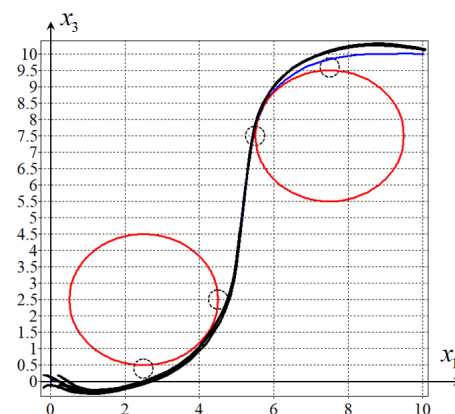


Figure 4. Projection on the horizontal plane of the optimal trajectories from eight initial states.

4. Results

The paper presents the use of machine learning technology with reinforcement to solve the optimal control problem with phase constraints using an evolutionary algorithm. To implement reinforcement learning, additional conditions defining the form of the optimal trajectory are introduced into the quality criterion. The optimal trajectory should pass through specified areas whose positions depend on the phase constraints. An example of solving the optimal control problem for a quadcopter by machine learning with reinforcement was given.

5. Discussion

The use of reinforcement learning technology to solve the optimal control problems of robotic devices is advisable, since in most cases the developer approximately knows the form of the optimal trajectory for the problem being solved.

Author Contributions: Conceptualization, A.D. and E.S.; methodology, A.D. and S.K.; software, A.D., S.K. and E.S.; validation, A.D. and V.M.; formal analysis, E.S.; investigation, S.K.; data curation, A.D., S.K. and E.S.; writing—original draft preparation, A.D., E.S. and V.M.; writing—review and editing, E.S.; visualization, V.M.; supervision, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work was performed with partial support from the Russian Science Foundation, Project No 23-29-00339.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in <https://cloud.mail.ru/public/pKW4/fmrfqjkv3> (for optimal trajectory 1), <https://cloud.mail.ru/public/mVZ5/Kr2yb9jJn> (for optimal trajectory 2).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rao, A.V. A Survey of Numerical Methods for Optimal Control. *Adv. Astronaut. Sci.* **2010**, *135*, 497–528.
2. Grachev, N.I.; Evtushenko, Y.G. A Library of Programs for Solving Optimal Control Problems. *USSR Comput. Math. Math. Phys.* **1980**, *2*, 99–119. [\[CrossRef\]](#)
3. Diveev, A.I.; Konstantinov, S.V. Study of the Practical Convergence of Evolutionary Algorithms for the Optimal Program Control of a Wheeled Robot. *J. Comput. Syst. Sci. Int.* **2018**, *4*, 561–580. [\[CrossRef\]](#)
4. Brown, B.; Zai, A. *Deep Reinforcement Learning in Action*; Manning, Publications Co.: Shelter Island, NY, USA, 2019; 475p.
5. Morales, M. *Grokking Deep Reinforcement Learning*; Manning Publications Co.: Shelter Island, NY, USA, 2020; 472p.
6. Duriez, T.; Brunton, S.; Noack, B.R. *Machine Learning Control—Taming Nonlinear Dynamics and Turbulence*; Springer: Cham, Switzerland, 2017; 229p.
7. Diveev, A. Hybrid Evolutionary Algorithm for Optimal Control Problem. In *Lecture Notes in Networks and Systems*; Springer: Berlin/Heidelberg, Germany, 2023; Volume 543, pp. 726–738.
8. Diveev, A.; Sofronova, E. Synthesized Control for Optimal Control Problem of Motion Along the Program Trajectory. In Proceedings of the 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul, Turkey, 17–20 May 2022; pp. 475–480.
9. Diveev, A.I.; Shmalko, E.Y. *Machine Learning Control by Symbolic Regression*; Springer: Cham, Switzerland, 2021; 155p.
10. Koza, J.R.; Keane, M.A.; Streeter, M.J.; Mydlowec, W.; Yu, J.; Lanza, G. *Genetic Programming IV. Routine Human-Competitive Machine Intelligence*; Springer: Boston, MA, USA, 2003; 590p.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.