# Deep Reinforcement Learning for Concentric Tube Robot Control with Goal Based Curriculum Reward

K. Iyengar[1], D. Stoyanov[1]

[1] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, UK,*
*keshav.iyengar@ucl.ac.uk*

## INTRODUCTION

Concentric tube robots (CTRs) are composed of pre-curved, super-elastic tubes that behave like continuum robots with a rotation and translation degree of freedom per tube [1]. The tubes interact with each other when rotating and translating to bend and twist in a manner resulting in curvilinear paths as seen in Figure 1. CTRs are clinically employed in minimally invasive surgery (MIS) in applications of actuated steerable needles or teleoperated manipulators. Ophthalmological, endonasal and fetal surgery have been explored interventions that may benefit highly from the dexterity, compliance and flexibility of CTRs [2]. In such applications, the main benefit is having a flexible articulated instrument while maintaining a small incision point to minimize trauma at the entry point.

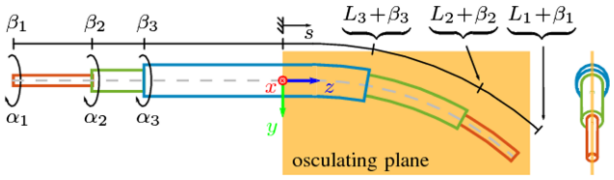


**Figure. 1.** Concentric tube robot kinematic inputs for 3 tubes adapted from [6].

Compared to rigid link robots, the kinematic modelling of CTRs is more complex due to the non-linear interactions of the tubes. Previous model-based work has throughly investigated various solutions with respect to computation complexity and accuracy. A geometrically exact model [3] is one such solution but has its limitations in terms of transverse shear, elongation and friction. A generally applicable modelling technique has not yet emerged.

In this work, we present a goal based curriculum reinforcement learning approach to learn the inverse kinematics of CTRs. Reinforcement learning is an iterative paradigm where an agent aims to learn the optimal sequence of actions to achieve a goal summarized by a reward signal. We focus on simulation results in this work because transfer learning is an active research area with various strategies such as domain randomization being explored. Although previous work has investigated model-free kinematics for CTRs [4,5,6,7], our contributions in this work are a novel training strategy with goal based curriculum and confirming previous reinforcement learning approach results [7] with a more accurate kinematics model [3] used for simulation.

## MATERIALS AND METHODS

As described in [7], we formulate a Markov Decision Process (MDP) with states, actions and reward function. The state is defined as rotation ($\alpha_i$) and extension ($\beta_i$) position for each tube i in trigonometric form [6], desired goal, g, achieved goal or robot tip position, $\hat{g}$ and goal tolerance $\delta(t)$. The trigonometric form, $\gamma_i$, is defined as

$$\gamma_i = \{\gamma_{1,i}, \gamma_{2,i}, \gamma_{3,i}\} = \{\cos\alpha_i, \sin\alpha_i, \beta_i\}$$

Thus, we can define the state at timestep t as

$$s_t = \{\gamma_1, \gamma_2, \gamma_3, g - \hat{g}, \delta(t)\}$$

The extension joint, $\beta_i$ can directly be obtained from $\gamma_{3,i}$ and the rotation joint $\alpha_i$ can be obtained by

$$\alpha_i = \text{atan2}(\gamma_{2,i}, \gamma_{1,i})$$

The kinematic input variables (α, β) are shown in Figure 1 where $L_i$ is the overall length of tube i. Actions are defined as changes in rotation and extension positions in a single timestep. The desired goal, g, is defined as a Cartesian point in the achievable workspace. The desired goal, $\hat{g}$, is the tip position of the robot determined with forward kinematics in simulation. The reward at timestep t, is defined with a reward function as follows,

$$r_t = \begin{cases} 0, & e_t \leq \delta(t) \\ -1, & otherwise \end{cases}$$

where $\delta(t)$ is the goal tolerance as a function of timesteps and $e_t$ is the Cartesian error from the robot tip to desired goal.

To investigate a goal-based curriculum, we propose three goal tolerance functions. Using a starting goal tolerance $\delta_{\text{initial}}$, and final goal tolerance $\delta_{\text{final}}$, with $N_{ts}$ being the total number of timesteps to apply the function, we can fully define these functions. The first function is a constant tolerance,

$$\delta(t) = \begin{cases} \delta_{\text{final}}, & t \leq N_{ts} \\ \delta_{\text{final}}, & otherwise \end{cases}$$

The second function is a linear,

$$\delta(t) = \begin{cases} at + b, & t \leq N_{ts} \\ \delta_{\text{final}}, & otherwise \end{cases}$$

$$a = (\delta_{\text{final}} - \delta_{\text{initial}})/N_{ts}$$

$$b = \delta_{\text{initial}}$$

The final function is exponentially decaying with $a$ as an initial tolerance and $r$ as the rate of decay.

$$\delta(\text{t}) = \begin{cases} a(1-r)^t, & t \leq N_{ts} \\ \delta_{\text{final}}, & otherwise \end{cases}$$

$$a = \delta_{\text{initial}}$$

$$r = 1 - (\delta_{\text{final}}/\delta_{\text{initial}})^{N_{ts}^{-1}}$$

For the second function, the tolerance remains at $\delta_{\text{final}}$ for $t > N_{ts}$ till the end of training. We train a three tube robot system with parameters listed in Table 1, where stiffness is 5 GPa and torsional stiffness is 2.3 GPa with deep deterministic policy gradient [8] and with hindsight experience replay [9]. We use parameters found in [7] with multi-variate Gaussian noise. Each experiment was trained for a total of 300,000 and $N_{ts} = 150{,}000$, $\delta_{\text{final}} = 20$ mm and $\delta_{\text{initial}} = 1$ mm and 19 parallel workers.

| $L$ (mm) | $L_{curved}$ (mm) | $d_{inner}$ (mm) | $d_{outer}$ (mm) | Pre-curvature |
|---|---|---|---|---|
| 215.0 | 14.9 | 1.0 | 2.4 | 15.82 |
| 120.2 | 21.6 | 3.0 | 3.8 | 11.8 |
| 48.5 | 8.8 | 4.4 | 5.4 | 20.04 |

**Table. 1.** Simulation robot parameters. From inner to outer.

## RESULTS

To compare convergence of training of the experiments we plot error through training with the shaded standard deviation area in Figure 2 and the cumulative episode rewards through training in Figure 3.

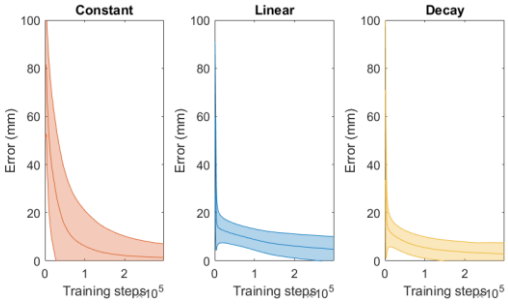


**Figure. 2.** Comparing error during training for constant, linear and decay functions.

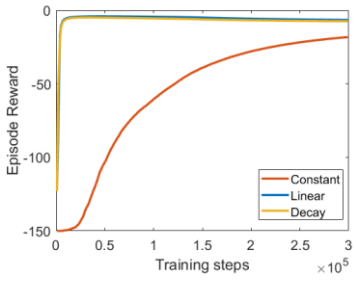


**Figure. 3.** Comparing error during training for constant, linear and decay functions

The final error found at the end of training was 1.4mm for the constant function, 4.8mm for linear and 2.9mm for the decay function in Cartesian space. With 2 million training steps, these methods are in-line with [7] which was on par with other simulation results [4].

## CONCLUSION AND DISCUSSION

Conducting this study we reached two conclusions. First, Although the final error for the constant function is lower than linear or decay, varying the goal tolerance had a large effect in the beginning of training by reducing errors very quickly with high rewards as compared to the constant function. Second, there is less deviation in the linear and decay function during the initial stages of training.

In this study, a novel training methodology for concentric tube robots with reinforcement learning based on goal tolerance has been suggested to be useful at the initial stages of training. Expanding on work conducted in [7] that used simplified kinematics, this study trains with accurate kinematics of CTRs. In future work we will experiment with combining functions to improve convergence speed and lower errors.

## REFERENCES


[1] P. E. Dupont, J. Lock, B. Itkowitz, and E. Butler, "Design and control of concentric-tube robots," *IEEE Trans. Robot.*, vol. 26, no. 2, pp. 209–225, 2010, doi: 10.1109/TRO.2009.2035740.

[2] J. Burgner-Kahrs, D. C. Rucker, and H. Choset, "Continuum Robots for Medical Applications: A Survey," *IEEE Transactions on Robotics*, vol. 31, no. 6. Institute of Electrical and Electronics Engineers Inc., pp. 1261–1280, 01-Dec-2015, doi: 10.1109/TRO.2015.2489500.

[3] D. C. Rucker, B. A. Jones, and R. J. Webster, "A geometrically exact model for externally loaded concentric-tube continuum robots," *IEEE Trans. Robot.*, vol. 26, no. 5, pp. 769–780, 2010, doi: 10.1109/TRO.2010.2062570.

[4] C. Bergeles, F.-Y. Lin, and G. Z. Yang, "Concentric Tube Robot Kinematics Using Neural Networks," in *The Hamlyn Symposium on Medical Robotics*, 2015, no. June, pp. 13–14, doi: 10.1109/MPUL.2012.2182857.

[5] A. Kuntz, A. Sethi, R. J. Webster, and R. Alterovitz, "Learning the Complete Shape of Concentric Tube Robots," *IEEE Trans. Med. Robot. Bionics*, vol. 3202, no. c, pp. 1–1, 2020, doi: 10.1109/tmrb.2020.2974523.

[6] R. Grassmann, V. Modes, and J. Burgner-Kahrs, "Learning the Forward and Inverse Kinematics of a 6-DOF Concentric Tube Continuum Robot in SE(3)," in *IEEE International Conference on Intelligent Robots and Systems*, 2018, pp. 5125–5132, doi: 10.1109/IROS.2018.8594451.

[7] K. Iyengar, G. Dwyer, and D. Stoyanov, "Investigating exploration for deep reinforcement learning of concentric tube robot control," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 7, pp. 1157–1165, 2020, doi: 10.1007/s11548-020-02194-z.

[8] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015.

[9] M. Andrychowicz *et al.*, "Hindsight experience replay," in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-Decem, pp. 5049–5059.