Proceedings of the Seventh International Conference on
Artificial Intelligence, Soft Computing, Machine Learning and Optimization,
in Civil, Structural and Environmental Engineering
Edited by: P. Iványi, J. Kruis and B.H.V. Topping
Civil-Comp Conferences, Volume 11, Paper 3.2
Civil-Comp Press, Edinburgh, United Kingdom, 2025
ISSN: 2753-3239, doi: 10.4203/ccc.11.3.2

Reinforcement Learning-Based Control Strategy for Semi-Active Energy Transfer in Beam Structures

D. Bogucki, M. Ostrowski and B. Blachowski

Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

Abstract

In this paper, a modern, reinforcement learning-based semi-active vibration control strategy is presented. Three different reinforcement learning algorithms are used to determine the control of the vibration process of a cantilever beam modeled as a system of two rigid links connected by a rotational spring. The control is achieved by blocking the connections between the links. This effect is achieved by introducing an equivalent rotational viscous damper. The obtained control signals are compared with the instantaneous optimal control, which greedily transfers the vibration energy from one mode to another. The quality of the control signal obtained using reinforcement learning confirms the ability of such algorithms to obtain results consistent with the analytical solution.

Keywords: semi-active control, vibration mitigation, reinforcement learning, lockable joints, modal analysis, structural dynamics

1 Introduction

Vibration attenuation in mechanical and civil engineering structures has been a topic of intense research over the past few decades. Among various possible options, i.e., passive, active, and semi-active, the latter has recently attracted significant attention due to its performance comparable to that of an active control system while removing the requirement for an efficient power supply. Within such approaches, semi-actively lockable joint constitutes an interesting solution [1]. The basic mechanism of operation of such joints allows them to operate in two opposite states, namely as pin joints or rigid connections. Dynamic switching between these two states enables the transfer of energy of vibration from low-frequency modes to high-frequency ones, at the same time reducing the amplitudes of oscillations. Another promising application of such a control strategy is enhancing of the energy harvesting process. Here, energy is to be precisely transferred from the currently excited vibration modes to the targeted one. It allows for the operation of the attached energy harvester under resonance conditions for wide-band or random external excitations.

Prior to the application of the lockable joints for structural vibration reduction, one has to select an appropriate control strategy. Traditionally, researchers used various strategies derived from classical control theory, such as Lyapunov switching control or linear quadratic regulator. These control strategies were successfully applied not only to the stabilization of linearly vibrating systems, but also nonlinear ones like underactuated robotic arms [2]. However, application of control strategies based on optimal control theory often requires solving complex differential equations. Therefore, nowadays, the machine learning approaches are more frequently chosen by researchers [3]. One such approach, particularly useful in solving control problems, is reinforcement learning (RL), described in the seminal book by Sutton and Barto [4]. RL allows to avoid finding analytical conditions of optimality and seeks the optimal solution in a trial and error manner. Moreover, as it was shown in a paper by Seyyedabbasi [5] on the problem of localizing mobile sensor nodes, RL-based algorithms explore the search space for optimal solution in effective way trying to find global optimum.

In this study, we investigate the possibility of applying reinforcement learning for semi-active control of a reconfigurable beam structure equipped with a semi-active joint. The goal of the RL agent is to find a sequence of switches for the state of the joint, allowing the transfer of vibration energy from one mode shape to the other. Three different algorithms of reinforcement learning have been used, namely Proximal Policy Optimization (PPO) [9], Advantage Actor-Critic (A2C) [6], and Deep Q-Network (DQN) [7, 8] . Additionally, the results obtained by the algorithms mentioned above have been compared with the solution obtained by Instantaneous Optimal Control.

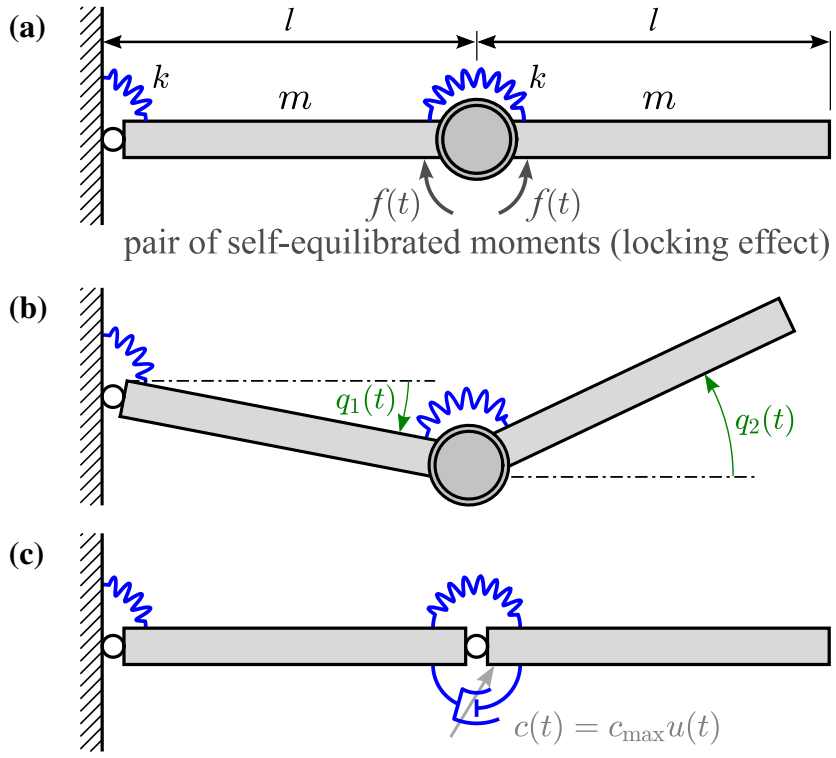


Figure 1: Considered structure equipped with lockable joint: (a) scheme of the structure including the locking effect represented by the pair of self-equilibrated moments, (b) rotational displacements, and (c) rotational viscous damper (with large damping factor) representing locking effect

2 Methods

2.1 Dynamics of the structure under investigation

2.1.1 Dynamic behaviour of a reconfigurable system

The controlled structure consists of two ideally rigid rods of mass m and length l, which are connected by rotational springs as shown in Figure-1a. The structure is equipped with a semi-actively controlled, lockable joint that changes its structural properties depending on the joint's state (locked or unlocked). The locking effect is represented by the pair of self-equilibrated bending moments keeping the connection rigid. When the joint is unlocked, then $f(t) = 0$, allowing free relative rotation between the adjacent rods. Motion of the structure is described by two rotational displacements $q_1(t)$ and $q_2(t)$ as shown in Figure 1b.

It is assumed that the transient time of the locking/unlocking of the joint is negligibly short, and the joint is fully locked or fully unlocked in its steady state. Taking into account these assumptions, the moment $f(t)$ transmitted by the lockable joint can be described using a switchable viscous damper as shown in Figure 1c. Then, this

moment is represented by the formula:

$$f(t) = -\underbrace{c_{\max} u(t)}_{c(t)} \left(\dot{q}_1(t) - \dot{q}_2(t)\right), \tag{1}$$

where: $c_{\max}$ is large constant damping factor, $u(t) \in \{0, 1\}$ is control signal taking value of 0 when the joint is unlocked (then also $f(t) = 0$) or value of 1 when the joint is locked. A sufficiently large $c_{\max}$ allows for representing the joint as a rigid connection when locked, without providing any significant damping to the system.

It is assumed that the system is undamped (except for negligible damping provided by the viscous joint model) and that vibration has small amplitudes. For these assumptions, the motion of the considered structure is described by the equation of motion below.

$$\begin{cases} \mathbf{M}\ddot{\mathbf{q}}(t) + u(t)\widetilde{\mathbf{C}}\dot{\mathbf{q}}(t) + \mathbf{K}\mathbf{q}(t) = \mathbf{d}(t) \\ \mathbf{q}(0) = \mathbf{q}_0, \dot{\mathbf{q}}(0) = \dot{\mathbf{q}}_0 \end{cases} \tag{2}$$

where:

$$\mathbf{M} = ml^2 \begin{bmatrix} 4/3 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}, \qquad \mathbf{K} = k \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix},$$

are mass and inertia matrices, respectively, $\mathbf{q}(t) = \begin{bmatrix} q_1(t) & q_2(t) \end{bmatrix}^{\mathrm{T}}$,

$$\widetilde{\mathbf{C}} = c_{\max} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

is a matrix employing a viscous model of the lockable joint, which couples rotational degrees of freedom (DOFs) when $u(t) = 1$, and $\mathbf{d}(t)$ is vector of external forces.

2.1.2 Vibration modes and modal energy transfer

Aiming at the investigation of the transfer of energy between vibration modes, the structural motion is to be expressed in modal coordinates $\boldsymbol{\eta}(t)$. They are related to rotational displacements by the following transformation:

$$\mathbf{q}(t) = \boldsymbol{\Phi}\boldsymbol{\eta}(t). \tag{3}$$

Matrix $\boldsymbol{\Phi}$ collects structure modeshapes: $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}^{(1)} & \boldsymbol{\phi}^{(2)} \end{bmatrix}$. They relate to the eigenvalue problem for the joint in the unlocked state:

$$\left(\mathbf{K} - \omega^2 \mathbf{M}\right) \boldsymbol{\phi} = \mathbf{0}, \tag{4}$$

where ω is the natural frequency.

The considered structure has 2 DOFs, hence it has two vibration modes which are depicted in Figure 2.

Substituting Equation (3) into (2) and left-multiplying by transpose of $\boldsymbol{\Phi}$ the following modal equation of motion is obtained:

$$\ddot{\boldsymbol{\eta}}(t) + u(t)\widetilde{\boldsymbol{\Gamma}}\dot{\boldsymbol{\eta}}(t) + \boldsymbol{\Omega}^2\boldsymbol{\eta}(t) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{d}(t), \tag{5}$$

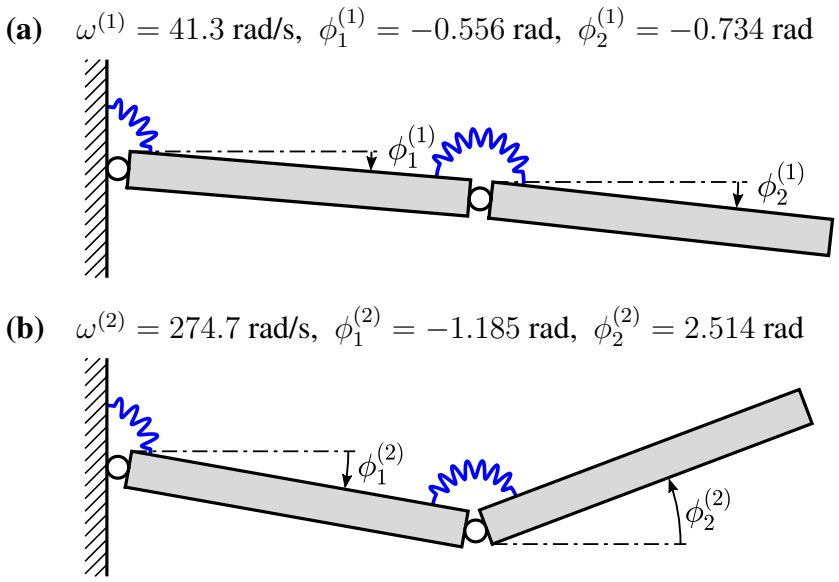


Figure 2: Vibration modes of the considered structure

where: $\boldsymbol{\Omega}^2 = \mathrm{diag}(\omega^{(1)2}, \omega^{(2)2})$ and $\widetilde{\boldsymbol{\Gamma}} = \boldsymbol{\Phi}^{\mathrm{T}}\widetilde{\mathbf{C}}\boldsymbol{\Phi}$ is matrix representing the locking effect in modal coordinates. This matrix is not diagonal, hence it is responsible for modal coupling between vibration modes obtained for the joint in the unlocked state. It results in the energy exchange between vibration modes.

The modal basis corresponding to the unlocked joint is selected because locking of the joint imposes a kinematic constraint (softened for the viscous model, but for large $c_{\max}$, the constraint is approximated well) and effectively removes one rotational DOF. Thus, the modal basis obtained for the locked joint is insufficient to describe the whole state of the system (in this case, we would have a single-degree-of-freedom system).

Mechanical energy of the structure is the sum of the energies related to particular vibration modes (namely: modal energies), according to the equation below.

$$\begin{aligned} E(t) &= \frac{1}{2}\dot{\boldsymbol{\eta}}^{\mathrm{T}}(t)\underbrace{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}\boldsymbol{\Phi}}_{\mathbf{I}}\dot{\boldsymbol{\eta}}(t) + \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}(t)\underbrace{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{K}\boldsymbol{\Phi}}_{\boldsymbol{\Omega}^2}\boldsymbol{\eta}(t) = \\ &= \sum_{i=1}^{2}\frac{1}{2}\left(\dot{\eta}_i^2(t) + \omega^{(i)2}\eta_i^2(t)\right) = \sum_{i=1}^{2}E_i(t). \end{aligned} \tag{6}$$

Due to the off-diagonal elements in matrix $\widetilde{\boldsymbol{\Gamma}}$, one particular energy can increase at the expense of the other one. Due to the fact that a locked joint does not perform any work over the structure, the total balance of increments and decrements of the modal energies is zero. Hence, the remainder of this study is devoted to an RL-based algorithm that locks or unlocks the joint at suitable time instants, thereby providing an efficient transfer of energy to the preselected vibration mode. The exception when the joint can dissipate the energy is an inelastic collision occurring if the joint is being

locked when rotational velocities are different: $\dot{q}_1(t) \neq \dot{q}_2(t)$. The algorithm should avoid such a jerking of the structure.

2.2 Reinforcement learning-based semi-active control

2.2.1 Semi-active control in the form of a Markov Decision Process

Prior application of reinforcement learning for semi-active control of reconfigurable structure described in previous sections, it is convenient to represent our problem in the form of the sequential process known as the Markov Decision Process (MDP) [4]. A finite MDP consists of:

- Finite set of states S, with subset of terminal states $\bar{S} \subset S$
- Finite set of actions A
- Reward function $\mathcal{R} : S \times A \times S \to \mathbb{R}$
- State transition probability function $\mathcal{T} : S \times A \times S \to [0, 1]$ such that

$$\forall s \in S, a \in A : \sum_{s' \in S} \mathcal{T}(s, a, s') = 1$$

- Initial state distribution $\mu : S \to [0, 1]$ such that

$$\sum_{s \in S} \mu(s) = 1 \text{ and } \forall s \in \bar{S} : \mu(s) = 0$$

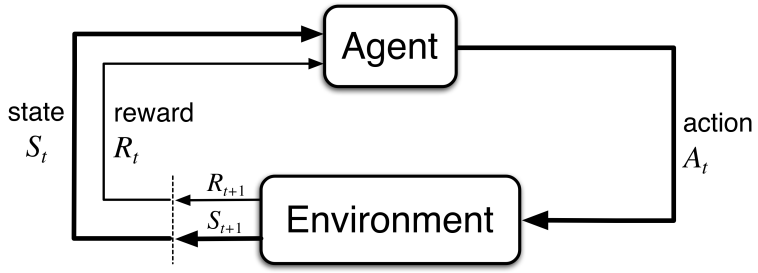


Figure 3: Visualization of MDP framework

The above formulas represent a general probabilistic MDP, which starts in an initial state $s_0 \in S$, which is sampled from μ. Then, at time t, the agent observes the current state $s_t \in S$ of the system and chooses an action $a_t \in A$ with probability given by its policy, $\pi(a_t|s_t)$, which is conditioned on the state. Next, given the state s_t and action a_t, the MDP transitions into a next state $s_{t+1} \in S$ with probability given by $\mathcal{T}(s_t, a_t, s_{t+1})$, and the agent receives a reward $r_t = \mathcal{R}(s_t, a_t, s_{t+1})$. This probability

can be written as $\mathcal{T}(s_{t+1}|s_t, a_t)$ to emphasize that it is conditioned on the state-action pair s_t, a_t. Finally, these steps are repeated until the process reaches a terminal state $s_t \in \bar{S}$ or after completing a maximum number of N time steps. Each repetition of this process is called an episode.

In this study, however, we apply deterministic equations of motion under specified initial conditions. As a result, the state transition probability function is simply a set of first-order differential equations that describe the dynamics of the system in a state space.

2.2.2 Environment definition and experiments setup

For the 2-rod semi-active control problem investigated in this study, we defined our MDP environment as follows.

Action space - we consider a single-valued action u_t. At every time step, it takes one of two discrete values $\{0, 1\}$, where 1 represents a locked joint between the rods and 0 implies the elements moving freely.

Observation space - the observation vector consists of 7 components, which represent the full state of the environment. Variables include:

- q_1, q_2 are continuous variables for rotational displacement of both rods (2),
- $\dot{q}_1$, $\dot{q}$ are continuous variables for rotational velocities of both rods (2),
- $\ddot{q}_1$, $\ddot{q}$ are continuous variables for rotational accelerations of both rods (2),
- a_{t-1} is an additional discrete variable of the agent's previous action (1).

Observation variables are standardized to improve their numerical properties and to minimize the impact of outliers. Examples of such extreme values are large acceleration peaks present during locking of the joint when the rotational velocities are different $\dot{q_1}(t) \neq \dot{q_1}(t)$. It results in the structure jerking. The standardization formula looks as follows:

$$x_{stand} = \min\left(\max\left(\frac{x_{original}}{\max\limits_{1,\dots,T} x_{a=0,t}}, -2\right), 2\right) \tag{7}$$

where $x_{original}$ is the value to be standardized, T is the number of simulation steps, and $\max\limits_{1,\dots,T} x_{a=0,t}$ is the maximal value registered during the freely moving rods episode. Agent's previous action, as a discrete action, is not standardized.

Initial and terminal states - we consider a fixed initial state, that is, the system displacement in the second vibration mode with no initial velocity. The terminal state is driven by overall simulation steps and is limited to 400 of 0.01s steps.

Reward formula - we use an energy-based formula as a reward granted in every

simulation step of the environment. The reward function is equal to the ratio of the first modal energy to the total initial mechanical energy of the system:

$$r_t = E_1(t)/(E_1(0) + E_2(0)) \tag{8}$$

Under the above assumptions, the goal of the RL agent is to find a switching sequence u_t that takes the system from oscillation in one modal form to another. The reward definition encourages the agent to transfer energy from the initial vibration mode to the targeted one as quickly as possible. The reward defined above is expected to avoid jerking of the structure and limit excess damping, but does not directly address the potential chattering problem.

Theoretically, such a goal could be achieved using a full enumeration process of the search space, as shown in Figure 4. However, even moderate discretization of the process involving N time steps would result in an enormous number of 2^N possible combinations, which makes this approach infeasible for any practical engineering problem.

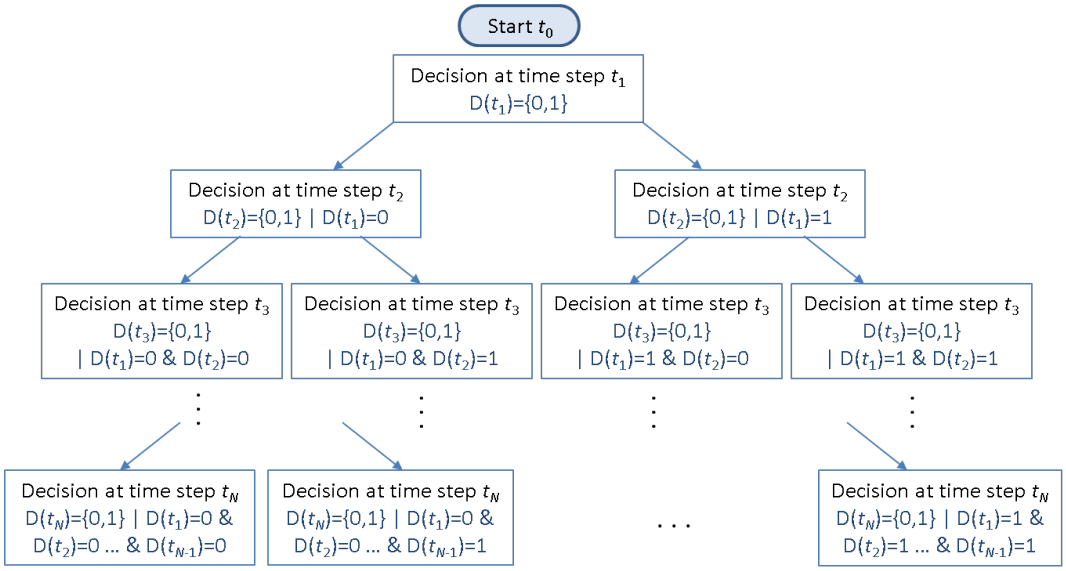


Figure 4: Decision tree of the sequence of switching process

2.2.3 Algorithms and experiment setup

We used popular discrete action space algorithms implementations from the Stable-Baselines3 benchmark algorithms library. DQN [7, 8] is representative of off-policy learning. A2C [6], PPO [9] are on-policy algorithms. We compared the performance of these algorithms in our environment, considering different discount factor.

Discount factor (γ) is one of the most essential parameters in RL. It represents how prone the agent is to postpone immediate rewards to gain higher future rewards. Dis-

count factor close to zero represents a myopic agent that considers only the nearest rewards. On the other hand, high discount factor considers the long-term consequences of the actions, but sometimes struggles to converge as current expectations of future rewards might be mistaken [4]. The current expected return of future rewards G_t is represented by the formula:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1}) \tag{9}$$

Where γ is the discount factor, t is the current time step, and k is the future rewards time horizon. In infinite k time horizon problems, discount factor has to be below 1 to avoid the summation of infinitely many future rewards. Depending on the environment, the value of γ coefficient is typically around 0.9 to 0.99. Our simulation has a fixed time horizon of 400 decision steps; therefore, we can consider unit discount factor, which tries to find a global solution to the problem. In this energy transfer problem, discount factor of 0 represents an agent that focuses only on the instant reward, which is similar to the solution proposed by [1]. We keep learning rates, exploration formulas, and other parameters as suggested in the Stable-Baselines3 library.

For each algorithm setup we run 10 experiments with different initial random generator seed. Each run was trained for 1,000 episodes, consisting of 400 decision steps each. Every episode was evaluated. We gathered and compared the results of the best episode from each experiment.

3 Results

Table 1 summarizes the best control found for each of 10 runs for A2C, DQN, and PPO algorithms and different discount factors. We analyze the stability of the results by examining the minimum, maximum, average, and standard deviation of the values. The best results in columns are bolded. The analytical solution proposed by [1] scored 304.36. The reward represents the share of the first modal energy of the total initial mechanical energy for every 400 simulation steps; the total achievable reward could vary from 0 to 400.

Algorithms using a discount factor of 0 achieved a better average score compared to those with a far-sighted approach. Such an unintuitive results prove that simple immediate control is, by default, better than heuristic search for the global optimum. Usage of RL in this case limited the RL algorithms to simple next-state approximators under a given discrete action.

Simple DQN appeared to be the most robust algorithm in this problem. It achieved the highest average best result and the lowest standard deviation of the best results among all 10 runs. The best solution found with the DQN algorithm and default discount factor can be seen in Figure 5.

PPO was less stable in performance, but managed to discover the best strategy

Experiments results					
Algorithm	γ coefficient	Min	Max	Std	Mean
A2C	0.0	73.09	210.62	58.51	137.42
A2C	0.99 (default)	73.09	183.33	35.39	87.62
A2C	1.0	25.57	75.3	15.12	68.56
DQN	0.0	**262.11**	298.2	**11.66**	**281.45**
DQN	0.99 (default)	227.78	294.85	23.69	269.48
DQN	1.0	235.59	289.94	20.14	264.46
PPO	0.0	198.65	**304.18**	37.67	267.25
PPO	0.99 (default)	73.09	286.14	68.0	209.61
PPO	1.0	73.09	241.85	64.2	156.79

Table 1: Comparison of sample algorithms with different discount factors

(a)

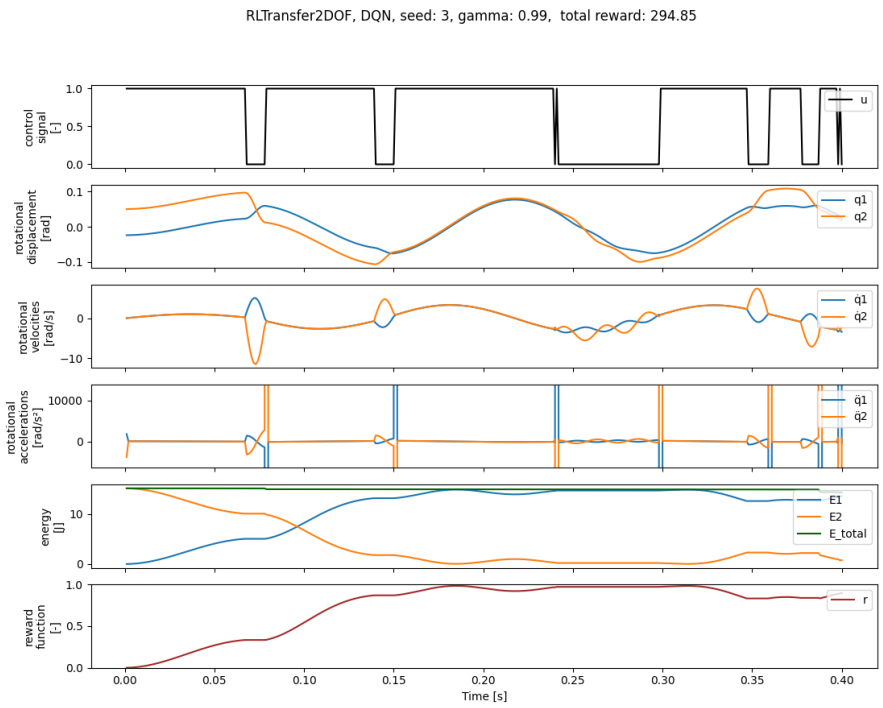
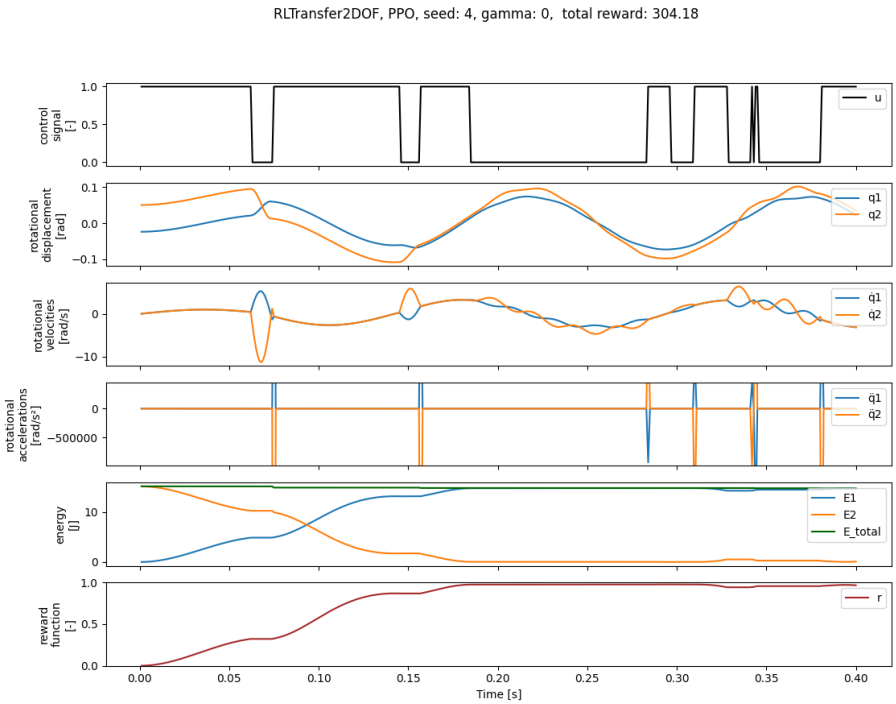


Figure 5: Solution found with DQN algorithm with default value of the discount factor

overall. It was very similar to one described in [1]. A comparison of both approaches is shown in Figure 6.

Experiments based on the A2C algorithm yielded consistently worse solutions than those of other algorithms in this problem. The runs typically got stuck in local maxima, which were related to blocking or unblocking the joint for the whole episode without any changes in control. Typical behaviour can be seen in Figure 7.

(a)



(b)

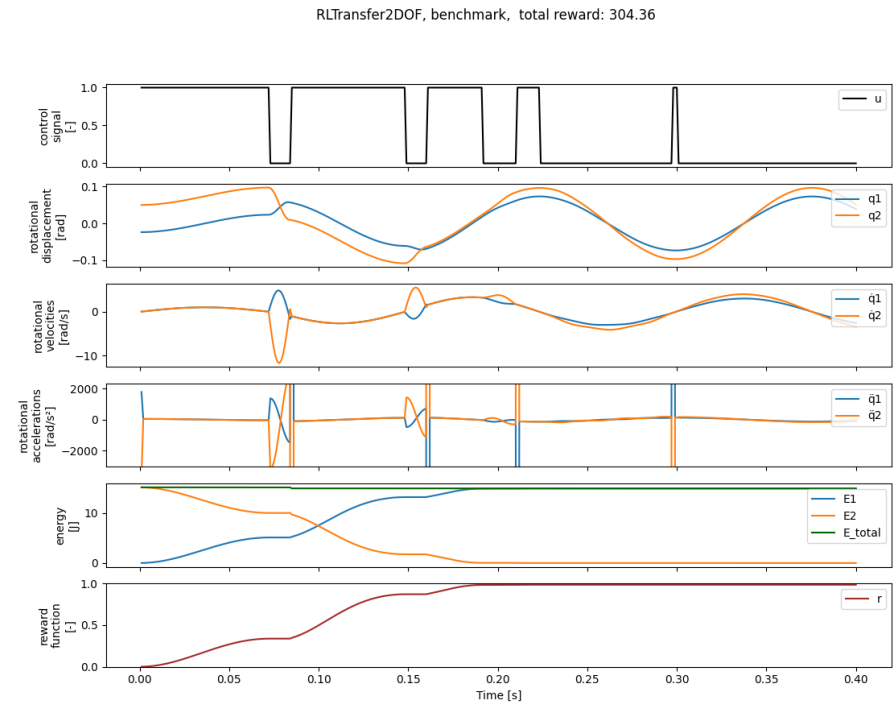


Figure 6: Comparison of the PPO best solution and analytical benchmark

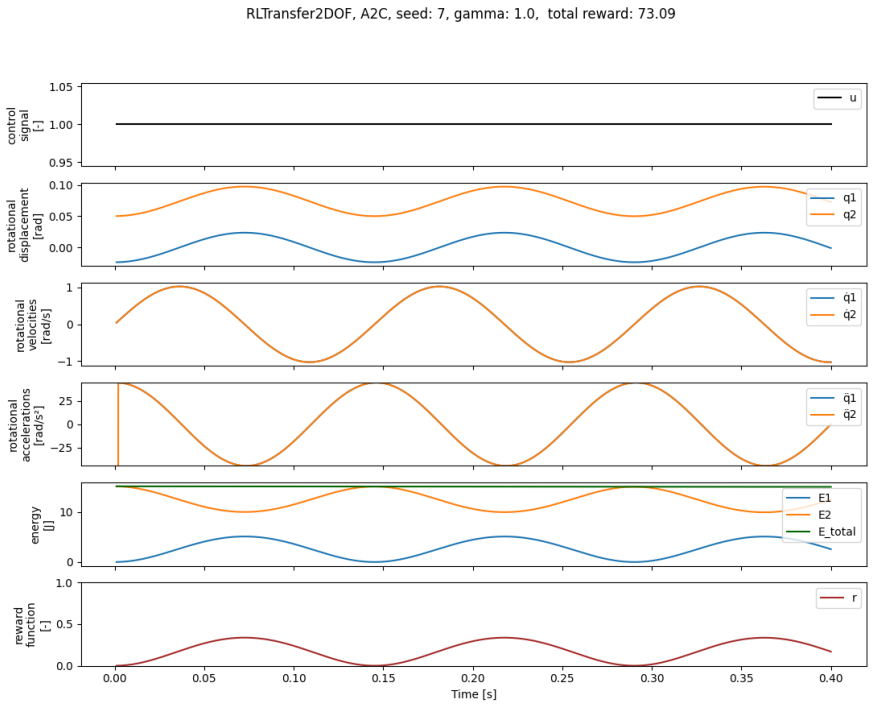


Figure 7: Typical solution found by A2C algorithms

4 Conclusions

This paper describes a possible application of the reinforcement learning algorithm in the task of semi-active vibration control of a beam structure. A beam with two degrees of freedom, connected by rotational springs, is used as an illustrative example. Three different reinforcement learning algorithms are analyzed, achieving results similar to the analytical solution obtained by minimizing the instantaneous quality factor describing the energy flow between the controlled modes. Visualization of the most efficient RL-based control with the analytical one is presented in Figure 8. At the top of the image is the beginning of the simulation. The black dot represents a locked joint.

Although achieving comparable performance to the benchmark solution, RL usage in this problem was limited to a simple next-state approximator. Experiments using gamma other than zero typically had significantly worse performance, implying that a simple greedy control might be sufficient for this problem. Heuristic search for a global optimal solution didn't bring any increase in the results. Possibly, RL-based approaches might be more beneficial for the efficient control of complex systems.

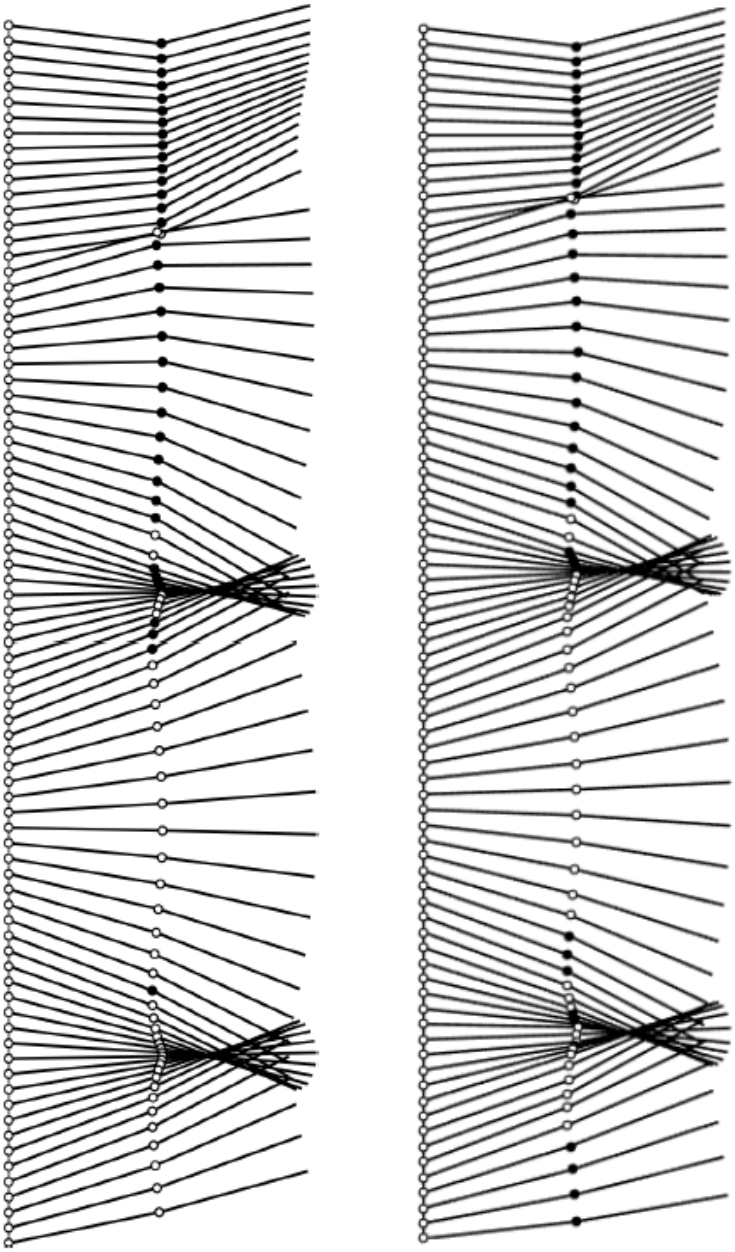


Figure 8: Time-lapse comparison of switching sequence obtained for analytical (LHS) and the best one found by RL-based (RHS) approach.

References

[1] M. Ostrowski, B. Błachowski, B. Popławski, D. Pisarski, G. Mikułowski, Ł. Jankowski, "Semi-active modal control of structures with lockable joints: general methodology and applications", Structural Control and Health Monitoring, Vol.28, No.5, pp.e2710-1-24, DOI: 10.1002/stc.2710, 2021.

[2] M.W. Spong, “The swing up control problem for the Acrobot”, IEEE Control Systems Magazine, 15(1), 49-55, DOI: 10.1109/37.341864, 1995.
[3] D. Pisarski, Ł. Jankowski, “Reinforcement learning-based control to suppress the transient vibration of semi-active structures subjected to unknown harmonic excitation”, Computer-Aided Civil and Infrastructure Engineering, Vol.38, No.12, 1605-1621, DOI: 10.1111/mice.12920, 2023.
[4] R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 2018.
[5] A. Seyyedabbasi, “A reinforcement learning-based metaheuristic algorithm for solving global optimization problems”, Advances in Engineering Software, Vol. 178, 103411, DOI: 10.1016/j.advengsoft.2023.103411, 2023.
[6] V. Mnih, A.P. Badia, M. Mirza, A. Graves, T.P. Lillicrap, T. Harley, D. Silver and K. Kavukcuoglu. “Asynchronous Methods for Deep Reinforcement Learning”, arXiv:1602.01783, DOI: 10.48550/arXiv.1602.01783, 2016.
[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning”, arXiv:1312.5602, DOI: 10.48550/arXiv.1312.5602, 2013.
[8] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, “Human-level control through deep reinforcement learning”, Nature, Vol. 518, pp. 529-533, DOI: 10.1038/nature14236, 2015.
[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Proximal Policy Optimization Algorithms”, arXiv:1707.06347, DOI: 10.48550/arXiv.1707.06347, 2017.