DOI: 10.1111/mice.12920

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Reinforcement learning-based control to suppress the transient vibration of semi-active structures subjected to unknown harmonic excitation

Dominik Pisarski | Łukasz Jankowski

Department of Intelligent Technologies, Institute of Fundamental Technological Research of the Polish Academy of Sciences, Warsaw, Poland

Correspondence
Dominik Pisarski, Department of Intelligent Technologies, Institute of Fundamental Technological Research of the Polish Academy of Sciences, Warsaw, Poland.
Email: dpisar@ippt.pan.pl

Abstract

The problem of adaptive semi-active control of transient structural vibration induced by unknown harmonic excitation is studied. The controller adaptation is attained by using a specially designed reinforcement learning algorithm that adjusts the parameters of a switching control policy to guarantee efficient dissipation of the structural energy. This algorithm relies on an efficient gradient-based sequence that accelerates the learning protocol and results in suboptimal control. The performance of this method is examined through numerical experiments for a span structure that is equipped with a semi-active device of controlled stiffness and damping parameters. The experiments cover a selection of control learning scenarios and comparisons to optimal open-loop and heuristic state-feedback control strategies. This study has confirmed that the developed method has high stabilizing performance, and the relatively low computational burden of the incorporated iterative learning algorithm facilitates its application to multi–degree-of-freedom structures.

1 | INTRODUCTION

1.1 | Motivation

The rapid growth of the scale and complexity of modern designs in civil and mechanical engineering (e.g., bridges, overpasses, skyscrapers, and also automotive, railway, aerospace, and robotic technologies) and the evidence that large-scale systems can be exceptionally sensitive to external perturbations have motivated intensive research into the design of reliable and high-performance structural controllers (Ghaedi et al., 2017; Gutierrez Soto & Adeli, 2017c; Li & Adeli, 2018). In line with the recent concept of smart cities (Li & Adeli, 2018) and smart structures (Adeli & Saleh, 1997), special attention has been devoted to adaptive controllers that can operate in dynamic and uncertain environmental conditions (Bitaraf et al., 2012; Li & Adeli, 2016; Naderpoor Shad & Taghikhany, 2022; Wang & Adeli, 2015b), adjusting their control decisions to changes in the internal system parameters or the external excitation forces. Given that these changes are usually unpredictable and rapid, the major challenge in designing an online adaptive controller is to find a compromise between the control performance and the computational complexity of the involved algorithms.

1.2 | Literature background

A typical approach to adapting control functions is based on model predictive control (MPC), which employs repetitive solutions to a finite horizon optimal control problem where the time horizon is constantly rolled back (this is often referred to as receding horizon control). Even though MPC controllers rely on the system model by definition, some level of uncertainty in the model parameters or inaccuracies in forecasting the external disturbances can be compensated by a state-feedback loop that accommodates the actual system response in the subsequent optimal control problems. Numerous MPC applications can be found in the optimization of industrial processes (Bordons & Camacho, 1998) and traffic flows (Ferrara et al., 2015), where the controllers are used to cope with time-varying parameters and evolving boundary conditions. MPC is of special importance for the coordination of wind farms (Vali et al., 2019), which are subject to permanent changes in wind direction. The MPC-based controllers have also confirmed their efficiency for autonomous driving, where vehicles confront dynamic obstacles (Babu et al., 2018). In structural control, the majority of MPC controllers rely on specifically designed dynamic models that predict the evolution of the external excitation forces. Oveisi et al. (2018) developed a recursive least squares algorithm to estimate the disturbance signal, which is constantly updated and used to determine the receding horizon control. The method was successfully validated for a piezo-laminated beam subjected to harmonic disturbances. In Wasilewski et al. (2019), earthquake excitation is recovered from an autoregressive model and fed-forward to the MPC controller, which stabilizes the vibration of a multistorey building with hydraulic actuators. In Zelleke and Matsagar (2019), an energy-based predictive control algorithm was developed to suppress the vibration of a multistorey building subjected to wind excitation. An alternative method to mitigate the vibration of slender buildings exposed to uncertain excitation, based on the probabilistic robust control approach, was proposed by Yuen et al. (2007). Five optimal and suboptimal MPC methods were tested in Takacs and Rohal'-Ilkiv (2014) to determine their computational complexity and capabilities for online implementation to mitigate the free, steady-state, and transient vibration of a cantilever beam equipped with piezoceramic control devices. The authors observed no significant diversity in the control performance between optimal and suboptimal strategies. They suggested that in practice the computationally efficient suboptimal methods (e.g., minimum-time explicit or Newton–Raphson's MPC) may be implemented for systems of larger dimensions without a considerable loss of performance.

The majority of the MPC-based adaptive methods have been confirmed to have a decent stabilization performance. Nevertheless, due to the high computational complexity of the search for the optimal solution, they are mostly restricted to applications in linear structures with active force control actuators. The recent trend in structural control promotes the use of semi-active devices (Cundumi & Suárez, 2008; Gutierrez Soto & Adeli, 2019; Naderpoor Shad & Taghikhany, 2021), in particular, those based on intelligent materials (Ostrowski et al., 2021; Szmidt et al., 2019) that offer robust, energy-efficient operation, and relatively easy deployment. However, semi-active devices introduce nonlinearities, which in the case of multi–degree-of-freedom systems result in highly complex optimal control problems. It is therefore essential to search for alternative approaches that allow for efficient online control adaptation. Appealing perspectives are offered by recent nonclassical computational approaches such as replicator dynamics (Gutierrez Soto & Adeli, 2017b, 2017a, 2018) or reinforcement learning (RL). The latter is a subfield of machine learning (Adeli & Hung, 1994; Amezquita-Sancheza et al., 2020) that is grounded on the idea of learning from interaction (Sutton & Barto, 2020). In view of the adaptive control design, an RL algorithm enables the control decisions to be adjusted based on the controller–system interaction. Therefore, the knowledge of the system model and its parameters may be much poorer than in the case of the MPC control. Furthermore, the computational complexity of successive updates of RL control is significantly lower than in the case of searching for the optimal control solutions.

1.3 | RL approaches

Approaches based on RL have recently achieved exceptionally successful results in a variety of hard real-world control-like problems, ranging from a superhuman level of proficiency in the games of chess and Go (Silver et al., 2018), through the thermal soaring of gliders (Reddy et al., 2016) and swimming by body undulation (Jiao et al., 2021), to bus traffic control (Shi et al., 2021) and autonomous car driving (Sallab et al., 2017; Shi et al., 2022). Despite these outstanding achievements and the conducive algorithmic structure, the possibilities offered by RL are only occasionally exploited in the field of structural control. There is only a handful of related publications. Although pioneering, they concern only active control or structures with a very limited number of degrees of freedom. Qiu et al. (2021) adopted a deep deterministic policy gradient RL algorithm to train the neural networks that are responsible for controlling a flexible hinged plate. The control was realized through piezoelectric actuators. The experiments confirmed that the developed method is superior to a PD (proportional derivative) controller. In Nagendra et al. (2017), three RL algorithms (i.e., temporal-difference, policy gradient actor–critic, and value function approximation) were studied in the context of stabilizing a benchmark cart-pole system with no prior knowledge of its parameters. The authors compared the algorithms for their convergence and control performance, and concluded that the value function approximation method was the most preferred option. In Khalatbarisoltani et al. (2019), the Q-learning RL algorithm was applied to tune a fuzzy-PD controller to stabilize the vibration of a high-rise building. The method was successfully verified for a selection of seismic scenarios, while taking into account time delays in the control loop. The potential of using RL to control the shape of an active tensegrity structure was studied in Adam and Smith (2008). The developed algorithm combines case-based reasoning and learning from errors. It has an increased control quality, while reducing the time for control computation. Dengler and Lohmann (2018) employed the RL actor–critic algorithm to stabilize a swinging chain at the desired position. The proposed active force control used incomplete state information. Although this method was outperformed by the control relying on the analytic solution, it can be viewed as a viable alternative for classical control designs if the model cannot be sufficiently accurate.

1.4 | Objectives and organization of this paper

This paper proposes a new RL-based control method to mitigate the transient vibration of semi-active structures subjected to an unknown repetitive harmonic excitation force. This method is dedicated for a class of bilinear systems that represent a wide range of structures of controlled internal parameters. The main contribution lies in the control design, where due to the nonlinear nature of the considered dynamical system, it is necessary to constitute a new unique control policy that guarantees the asymptotic stability of the homogeneous system and facilitates an efficient optimization to suppress the transient vibration induced by unknown excitation. The optimization is performed by using a specially developed RL actor-only algorithm that relies on the state measurements and structural model parameters, with no information from the external excitation force. It adapts the control policy using the derivatives of the assumed energy-related cost functional, which is defined directly over the parameter space of the assumed control policy. This technique exploits the convergence that is naturally inherited from the incorporated gradient descent method, accelerating the iterative learning protocol, and results in a suboptimal control. The proposed method is validated by numerical experiments for a span structure equipped with a semi-active device with controlled stiffness and damping parameters. The convergence and performance are examined for several learning scenarios, including random perturbations in the frequency of the excitation force. The designed RL controller is compared to the optimal open-loop solution and the heuristic strategy, which relies on an equivalent control function that is precomputed offline. The relatively low computational complexity of the iterative learning algorithm opens up new perspectives for its application in large-scale complex structures.

The remainder of this work is structured as follows. Section 2 provides the assumptions and definitions of the considered system. The problem of RL control under unknown excitation force is formulated and resolved in Section 3: The switching parameterized control policy is defined, and then an updating sequence for the policy parameter is defined and accommodated in an iterative learning algorithm. A definition of comparative control methods is given in Section 4. In Section 5, the proposed controller is investigated by means of numerical experiments. The concluding remarks are given in Section 5.

2 | THE INVESTIGATED SYSTEM

This paper studies a class of semi-active vibrating structures that are governed by the following dynamical equation:

$$\dot{x}(t) = Ax(t) + \sum_{j=1}^{m} u_j(t)B_j x(t) + f(t), x(0) = x_0. \qquad (1)$$

In Equation (1), $x = x(t) : [0, T_c] \to \mathbb{R}^n$ represents the state vector at time $t \in [0, T_c]$, where $T_c > 0$ is a considered control time. The initial state is denoted by x_0. Each of the control inputs $u_1, \dots, u_m$ is assumed to be bounded by the minimum and maximum admissible values; that is, $u_j(t) : [0, T_c] \to \mathcal{U} = [u_{min}, u_{max}]$, $j = 1, \dots, m$, $u_{min} < u_{max}$. These bounds correspond to the physical constraints of a semi-active device (e.g., extreme voltages). The $n \times n$ matrices A and B_j, $j = 1, \dots, m$ are assumed to be constant. The $n \times 1$ vector $f(t)$ represents a repetitive excitation defined by a harmonic function with unknown amplitude, frequency, and phase shift (see Figure 1), which repeats in identical time windows corresponding to the control time interval $[0, T_c]$. For each time window the excitation force $f(t) \neq 0$ is acting on a structure for $t \in [0, T_f]$, where $T_f < T_c$. The time T_f is assumed to be sufficiently small, so that the vibrations in the whole time interval $[0, T_f]$ are transient (the steady-state vibration is not considered here). For the remaining control time—that is, for $t \in (T_f, T_c]$—the problem of controlling free vibration is considered by setting $f(t) = 0$.

Equation (1) can represent a wide range of semi-actively controlled structures, such as cantilever beams with elastomer-based blocks (Szmidt et al., 2017), span structures supported with magneto-rheological dampers

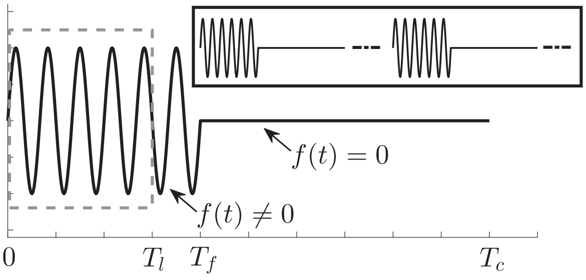

FIGURE 1 The assumed repetitive harmonic excitation with unknown amplitude, frequency, and phase shift. For the adaptation of the control decision, some learning time window denoted by T_l will be used.

(Pisarski & Myśliński, 2017; Wasilewski & Pisarski, 2020), frames with dry friction-based joints (Popławski et al., 2019), vehicular suspensions (Pepe & Carcaterra, 2016), or buildings with semi-active tuned mass dampers (Runlin et al., 2002). The assumed excitation is typical for repetitive industrial operations, such as drilling or grinding. Similar characteristics can model the influence of vehicle formations on the neighboring infrastructure. Temporary harmonic excitation is also found in the repetitive starting-up of rotor machines (e.g., mills, blowers, pumps, compressors), which is a result of subsynchronous resonances that are induced by electro–mechanical interaction between the motor's electric circuit and shaft structure (Szolc et al., 2019).

3 | RL-BASED CONTROL DESIGN

The aim is to design state feedback control functions $u_1, \dots, u_m$ (referred to as *policy*) for the system Equation (1) (*environment* in the RL terminology) that guarantee efficient suppression of the vibration induced by the excitation f. Regarding the excitation structure (Figure 1), two control phases are distinguished:

Phase I. The first phase is concerned with the transient vibration that is observed for $t \in [0, T_f]$. Here, an *actor-only RL* algorithm will be developed that allows the policy to be adapted to unknown characteristics of the excitation $f(t) \neq 0$. The algorithm will employ state measurements $x(t)$ for some learning time window $t \in [0, T_l]$ (where $T_l \leq T_f$) and it will provide succeeding reductions of the value of the cost functional:

$$J(T_l) = \int_0^{T_l} E(t)\mathrm{d}t. \tag{2}$$

In Equation (2), $E(t)$ stands for the structural energy:

$$E(t) = \frac{1}{2}x^T(t)Qx(t), \tag{3}$$

where $Q \succ 0$ is a positive definite $n \times n$ matrix.

Phase II. For the free vibration at $t \in (T_f, T_c]$, a policy will be used that assures the asymptotic stability of Equation (1) for $f(t) = 0$. It will be derived using the Lyapunov functions method and structural energy matrix Q.

3.1 | Parameterized policy

To construct a policy that is uniform for the control phases I and II, a parameterized state-dependent control law is employed. For each control time $t \in [0, T_c]$, the switching policy $u_1(t), \dots, u_m(t)$ is defined for Equation (1) as follows:

$$u_j(t) = \begin{cases} u_{\min}, x^T(t)K_j(t)x(t) \geq 0 \\ u_{\max}, x^T(t)K_j(t)x(t) < 0, j = 1, \cdots, m. \end{cases} \tag{4}$$

Here, $K_j(t)$ is an $n \times n$ matrix referred to as the policy parameter. Each of the policy parameters $K_j(t)$, $j = 1, \dots, m$ is structured by two subparameters. The first is denoted by K_j^* and iterated online by the learning algorithm in phase I; that is, for the time $t \in [0, T_f]$, where nonzero excitation force is acting on a structure. The second subparameter is K_j^0. It is precomputed offline based on the system structuring and remains constant for the free vibration in phase II, that is, for $t \in (T_f, T_c]$ when $f = 0$. Formally, the policy parameters can be written as follows:

$$K_j(t) = \begin{cases} K_j^*, 0 \leq t \leq T_f \\ K_j^0, T_f < t \leq T_c. \end{cases} \tag{5}$$

The method for iterative learning of K_j^*, $j = 1, \dots, m$ will be discussed in detail in the next section. First, computing the constant parameters K_j^0, $j = 1, \dots, m$ is considered. For each K_j^0, $j = 1, \dots, m$, it is assumed that:

$$K_j^0 = PB_j. \tag{6}$$

where P is an $n \times n$ symmetric matrix, which is computed as the solution to the following Lyapunov equation:

$$\left(A^T + \sum_{j=1}^{m} u_{\max}B_j^T\right)P + P\left(A + \sum_{j=1}^{m} u_{\max}B_j\right) = Q. \tag{7}$$

In Equation (7), the matrix $Q \succ 0$ is the same as in the definition of the energy function Equation (3). The assumed structuring of the policy parameters $K_j^0 = PB_j$ guarantees that the system Equation (1) for $f(t) = 0$ is asymptotically stable. To inspect this, let $V = V(x)$ be the Lyapunov function, which is defined by; see, for example, Sastry (1999):

$$V(x) = x^T P x. \tag{8}$$

The time derivative of V is:

$$\dot{V} = \dot{x}^T P x + x^T P \dot{x}. \tag{9}$$

The insertion of Equation (1) with $f(t) = 0$ into Equation (9) yields:

$$\dot{V} = x^T A^T P x + x^T P A x + \sum_{j=1}^{m} u_j x^T B_j^T P x + \sum_{j=1}^{m} u_j x^T P B_j x, \tag{10}$$

which can be written in the following form:

$$\dot{V} = x^T \left(A^T + \sum_{j=1}^{m} u_{\max} B_j^T \right) P x + x^T P \left(A + \sum_{j=1}^{m} u_{\max} B_j \right) x$$
$$+ \sum_{j=1}^{m} (u_j - u_{\max}) x^T B_j^T P x + \sum_{j=1}^{m} (u_j - u_{\max}) x^T P B_j x. \tag{11}$$

From the Lyapunov equation (7) and the symmetry of the matrix $P = P^T$, it can be eventually concluded that:

$$\dot{V} = -x^T Q x + 2 \sum_{j=1}^{m} (u_j - u_{\max}) x^T P B_j x. \tag{12}$$

The application of the switching policy Equation (4) ensures that:

$$\sum_{j=1}^{m} (u_j - u_{\max}) x^T P B_j x \le 0. \tag{13}$$

From $Q \succ 0$ and Equation (13), it follows that $\dot{V} < 0$ for every x, which guarantees the asymptotic stability of the closed-loop system Equation (1) with Equations (4) and (5) for $t \in (T_f, T_c]$.

3.2 | Policy parameter update

The policy parameter K_j^*, $j = 1, \ldots, m$ in Equation (5) will be updated using an approach that is in line with the method of actor-only RL. The actor-only method relies on the optimization of a cost functional that is defined directly over the parameter space of the policy (Grondman, 2015). Here, the energy-related objective functional J that is defined in Equation (2) will be optimized for the parameter space K_j^*, $j = 1, \ldots, m$. For each matrix K_j^*, the admissible set is defined as:

$$\mathcal{K}_j^* = \left\{ K_j^* = \left\{ K_{qr}^{*j} \right\}_{q,r=1}^{n} : K_{\min}^* \le K_{qr}^{*j} \le K_{\max}^* \right\}, \tag{14}$$

where $K_{min}^* < K_{max}^*$ are given real constants. The solution $x(t)$ to the system Equation (1) depends continuously on the policy parameters K_j^*, $j = 1, \ldots, m$ (see Chicone, 2006, Theorem 1.3). This result implies the continuity of the cost functional J in Equation (2) with respect to K_j^*, $j = 1, \ldots, m$ on the admissible set $\mathcal{K}_1^* \times \cdots \times \mathcal{K}_m^* \subset R^{m \cdot n^2}$. This set is finite-dimensional and compact in $R^{m \cdot n^2}$. Therefore, from the Weierstrass theorem (Liberzon, 2012), it follows that there is a set of policy parameters K_j^*, $j = 1, \ldots, m$ that minimizes J. The method of steepest descent is used to search for the optimal policy parameters, which relies on the following updating sequence:

$$K_j^{*(z)} = K_j^{*(z-1)} - \alpha_j \left(\frac{\mathrm{d}J}{\mathrm{d}K_j^*} \right)_{|K_j^* = K_j^{*(z-1)}}, z = 1, \cdots, z_{\max}, \tag{15}$$

where $\alpha_j > 0$, and z_{max} is the maximal number of learning iterations. The sequence Equation (15) will be initialized by setting $K_j^{*(0)} = K_j^0$, $j = 1, \ldots, m$, where K_j^0 is computed as in Equation (6).

To derive the formula to compute the derivative of objective functional J with respect to the policy parameters K_j^*, $j = 1, \ldots, m$ for Equation (15), the policy Equation (4) is first rewritten for $t \in [0, T_l]$ using the unit step function $\mathcal{U}(\cdot)$, as follows:

$$u_j(t) = c_1 + c_2 \mathcal{U}\left(x^T(t) K_j^* x(t) \right), j = 1, \cdots, m. \tag{16}$$

Here, it is assumed that

$$c_1 = u_{\max}, c_2 = u_{\min} - u_{\max}. \tag{17}$$

Next, the Hamiltonian for the cost functional Equation (2) is defined:

$$H(x, p, \{K_j^*\}_{j=1,\ldots,m})$$
$$= p^T \left(A\, x + \sum_{j=1}^{m} \left(c_1 + c_2\, \mathcal{U}\left(x^T K_j^*\, x \right) \right) B_j\, x + f \right)$$
$$- \frac{1}{2} x^T\, Q\, x \tag{18}$$

with the adjoint state $p = p(t) : [0, T_l] \rightarrow R^n$ satisfying the following differential equation:

$$\begin{aligned}\dot{p} &= -\frac{\partial H\left(x, p, \{K_j^*\}_{j=1,\cdots,m}\right)}{\partial x} \\ &= -A^T p - \sum_{j=1}^{m}\left(c_1 + c_2 \mathcal{U}\left(x^T K_j^* x\right)\right) B_j^T p \\ &\quad - \sum_{j=1}^{m} c_2 p^T B_j x \left(K_j^* + K_j^{*T}\right) x \delta\left(x^T K_j^* x\right) + Qx \\ p(T_l) &= 0, \end{aligned} \tag{19}$$

where $\delta(\cdot)$ stands for the Dirac delta function. From Equation (3) and Equation (18), the cost functional Equation (2) can be represented by:

$$J = \int_0^{T_l} \left(p^T \dot{x} - H(x, p, \{K_j^*\}_{j=1,\cdots,m})\right) \mathrm{d}t. \tag{20}$$

Let the functions $\delta x : [0, T_l] \rightarrow R^n$ and $\delta p : [0, T_l] \rightarrow R^n$ denote perturbations of the functions x and p with respect to the infinitesimal changes $\mathrm{d}K_j^* : R^n \times R^n \rightarrow R^n \times R^n$, $j = 1, \ldots, m$ of K_j^*, respectively. From the differentiability of the state vector x with respect to t, it follows that

$$\delta \dot{x} = \frac{d}{dt}(\delta x). \tag{21}$$

Now let K_j^{*r}, $r = 1, \ldots, n$ denote the vector corresponding to the rth column of the matrix K_j^*. Consistently, $\mathrm{d}K_j^{*r}$, $r = 1, \ldots, n$ will stand for the perturbation of the vector corresponding to rth column of the matrix K_j^*. From Equation (20), it follows that the differential $\mathrm{d}J$ of the cost functional Equation (2) with respect to perturbations $\mathrm{d}K_j^*$, $j = 1, \ldots, m$ is given by

$$\begin{aligned}\mathrm{d}J &= \int_0^{T_l} \left| - \sum_{j=1}^{m} \sum_{r=1}^{n} \left(\frac{\partial H}{\partial K_j^{*r}}\right)^T \mathrm{d}K_j^{*r} - \left(\frac{\partial H}{\partial x}\right)^T \delta x \right| \mathrm{d}t \\ &\quad + \int_0^{T_l} \left(p^T \delta \dot{x} + \left(\dot{x} - \frac{\partial H}{\partial p}\right)^T \delta p\right) \mathrm{d}t. \end{aligned} \tag{22}$$

Recall that (Chicone, 2006):

$$\dot{x}(t) = \frac{\partial H(x, p, \{K_j^*\}_{j=1,\cdots,m})}{\partial p}, \tag{23}$$

and thus the last term in Equation (22) can be canceled. Furthermore, integration by parts yields:

$$\int_0^{T_l} p^T \delta \dot{x} \mathrm{d}t = -\int_0^{T_l} \dot{p}^T \delta x \mathrm{d}t + [p^T \delta x]_0^{T_l}. \tag{24}$$

From $p(T_l) = 0$ and the initial condition in Equation (1) that implies $\delta x(0) = 0$, it follows that the last term in Equation (24) vanishes:

$$[p^T \delta x]_0^{T_l} = p(T_l)\delta x(T_l) - p(0)\delta x(0) = 0. \tag{25}$$

Taking into account Equations (23), (24), and (25) and inserting this into Equation (22), one obtains:

$$\begin{aligned}\mathrm{d}J &= \int_0^{T_l} \left| - \sum_{j=1}^{m} \sum_{r=1}^{n} \left(\frac{\partial H}{\partial K_j^{*r}}\right)^T \mathrm{d}K_j^{*r} - \left(\frac{\partial H}{\partial x}\right)^T \delta x \right| \mathrm{d}t \\ &\quad - \int_0^{T_l} \dot{p}^T \delta x \mathrm{d}t = -\int_0^{T_l} \sum_{j=1}^{m} \sum_{r=1}^{n} \left(\frac{\partial H}{\partial K_j^{*r}}\right)^T \mathrm{d}K_j^{*r} dt \\ &\quad - \int_0^{T_l} \left(\dot{p} + \frac{\partial H}{\partial x}\right)^T \delta x \mathrm{d}t. \end{aligned} \tag{26}$$

From the definition of the adjoint state in Equation (19), it can be observed that:

$$\mathrm{d}J = -\int_0^{T_l} \sum_{j=1}^{m} \sum_{r=1}^{n} \left(\frac{\partial H}{\partial K_j^{*r}}\right)^T \mathrm{d}K_j^{*r} \,\mathrm{d}t \tag{27}$$

From Equation (27), it follows that the derivative of the functional J with respect to the matrix K_j^* is given by:

$$\frac{\mathrm{d}J}{\mathrm{d}K_j^*} = -\int_0^{T_l} \frac{\partial H}{\partial K_j^*} \mathrm{d}t. \tag{28}$$

Using the definition of the Hamiltonian (18), it follows that:

$$\frac{\mathrm{d}J}{\mathrm{d}K_j^*} = -c_2 \int_0^{T_l} p^T B_j x x x^T \delta(x^T K_j^* x) \mathrm{d}t. \tag{29}$$

Applying the Dirac delta function's sifting property, the explicit formula is obtained:

$$\frac{\mathrm{d}J}{\mathrm{d}K_j^*} = -c_2 \sum_{i=1}^{s} p^T(\tau_i) B_j x(\tau_i) x(\tau_i) x^T(\tau_i), \tag{30}$$

where $\tau_1, \ldots, \tau_s$ denotes the sequence of time instants when the argument of the Dirac delta function in

Equation (29) equals zero:

$$\{\tau_1, \cdots, \tau_s\} = \{t \in [0, T_l] \, : \, x^T(t) K_j^* x(t) = 0, j = 1, \cdots, m\}. \tag{31}$$

To evaluate the cost derivative Equation (30) and perform the updating of the policy parameters Equation (15), the method has to rely on the information of the state $x(t)$ and adjoint state $p(t)$ for $t \in [0, T_l]$. While the state is assumed to be measurable or accessible through a state observer, the collection of the adjoint state requires an integration of the differential equation (19). The right-hand side of Equation (19) includes the Dirac delta function term, so that the solution is piecewise continuous and composed of a set of limiting functions; see Nedeljkov and Oberguggenberger (2012, Proposition 2.1). A convenient way to generate the trajectory of $p(t)$ is to start by detecting the sequence of time instants $\tau_1, \dots, \tau_s$ as in Equation (31). Next, the sequence of time steps $0 = t_0, t_1, \dots, t_\upsilon = T_l$ is assumed for the purpose of the backward integration of the equation:

$$\dot{p}(t) = -A^T p(t) - \sum_{j=1}^{m} \left(c_1 + c_2 \mathcal{U}\left(x^T(t) K_j^* x(t) \right) \right) B_j^T p(t) + Qx(t), \;\; p(T_l) = 0 \tag{32}$$

for $t \in [\tau_s^*, T_l]$. Here, τ_s^* denotes the time step that is the closest to the time instant τ_s, that is,

$$\tau_s^* = \underset{i=0,1,\cdots,\upsilon}{\operatorname{argmin}} |t_i - \tau_s|. \tag{33}$$

Then, for $t = \tau_s^*$ the following jump is performed:

$$p(\tau_s^*) = p(\tau_s^*) + \Delta p(\tau_s^*), \tag{34}$$

where the increment of the adjoint state $\Delta p(\tau_s^*)$ is computed applying the Dirac delta function's sifting property and the specific structuring of the right-hand side of adjoint state dynamical equation (19):

$$\Delta p(\tau_s^*) = c_2 p^T(\tau_s^*) B_j x(\tau_s^*) \left(K_j^* + K_j^{*T} \right) x(\tau_s^*). \tag{35}$$

Next, taking into account the value of $p(\tau_s^*)$ updated by the jump Equation (34), the backward integration of Equation (32) is continued until $t = \tau_{s-1}^*$; that is, until the time step that is the closest to the time instant τ_{s-1} and found in analogy to Equation (33). Again, a jump for $t = \tau_{s-1}^*$ is made in the same manner as in Equation (34). The operation is repeated unless $t = t_0$. A complete procedure for updating the policy parameters is demonstrated in Algorithm 1. In Figure 2, the RL process is visualized in the context of an actor–environment interaction.

Algorithm 1 Iterative learning algorithm to update the policy parameters $K_1^*, \dots, K_m^*$.

```
Step 1. Set $z = 0$ and the initial matrices $K_j^{*(z)} = K_j^0$, $j = 1, \dots, m$
          using Eq. (6).Select small positive numbers
          $\alpha_1, \dots, \alpha_m, \epsilon \in (0, 1)$.
          Set the maximal number of iterations $z_{max}$.
          Select the learning time window $[0, T_l]$.
Step 2. Apply the policy $u_1, \dots, u_m$ as in Eq. (4).
          Collect the state information $x(t)$ for $t \in [0, T_l]$.
Step 3. Using the state trajectory $x(t)$ detect the time instants as
          in Eq. (31) and compute the adjoint state trajectory
          $p(t)$ for $t \in [0, T_l]$ employing backward integration
          and Eqs. (32)–(35).
Step 4. Evaluate the derivatives $\frac{d\mathcal{J}}{dK_j^*}$, $j = 1, \dots, m$ as in Eq. (30)
          using the trajectories $x(t)$, $p(t)$ and matrices $K_j^{*(z)}$,
          $j = 1, \dots, m$.
Step 5. Set $z = z + 1$ and compute the updated
          policy parameters $K_j^{*(z)}$, $j = 1, \dots, m$ using Eq. (15) and
          $K_j^{*(z-1)}$, $\frac{d\mathcal{J}}{dK_j^*}$, $j = 1, \dots, m$.
          For every $j = 1, \dots, m$ project matrix $K_j^{*(z)}$ by Eq. (14)
          onto admissible set $\mathcal{K}_j^*$, $K_j^{*(z)} = \mathrm{Proj}_{\mathcal{K}_j^*}(K_j^{*(z)})$.
Step 6. Check if any of the terminal conditions is
          fulfilled:$\| \frac{d\mathcal{J}}{dK^*} \|_{|K^*=K^{*(z-1)}} < \epsilon$  or  $z = z_{max}$.
        If yes, then STOP. Otherwise go to Step 2.
```

Remarks 1.

R1. The value of the cost derivative in Equation (30) is related to the number of time instants $\tau_1, \dots, \tau_s$ defined in Equation (31). To guarantee a substantial decrease of the cost functional value when executing the sequence Equation (15), the selection of the step sizes $\alpha_1, \dots, \alpha_m \in (0, 1)$ in Step 1 should depend on the number s, and may vary for subsequent iterations z. Here, it will be assumed that $\alpha_j = \bar{\alpha}_j / s$, $j = 1, \dots, m$ for some small positive numbers $\bar{\alpha}_1, \dots, \bar{\alpha}_m$.

R2. Respecting the structuring of the admissible sets Equation (14), the projection $\mathrm{Proj}_{\mathcal{K}_j}(K_j^*) = \{K_{qr}^{*j}\}_{q,r=1}^n$ used in Step 5 is defined as follows:

$$K_{qr}^{*j} = \begin{cases} K_{min}^* & \text{if } K_{qr}^{*j} < K_{min}^*, \\ K_{qr}^{*j} & \text{if } K_{min}^* \leq K_{qr}^{*j} \leq K_{max}^* \\ K_{max}^* & \text{if } K_{qr}^{*j} > K_{max}^*. \end{cases} \tag{36}$$

R3. The norm of the cost derivative in the terminal condition in Step 6 is defined as the maximal value of the absolute entries of matrices Equation (30) for

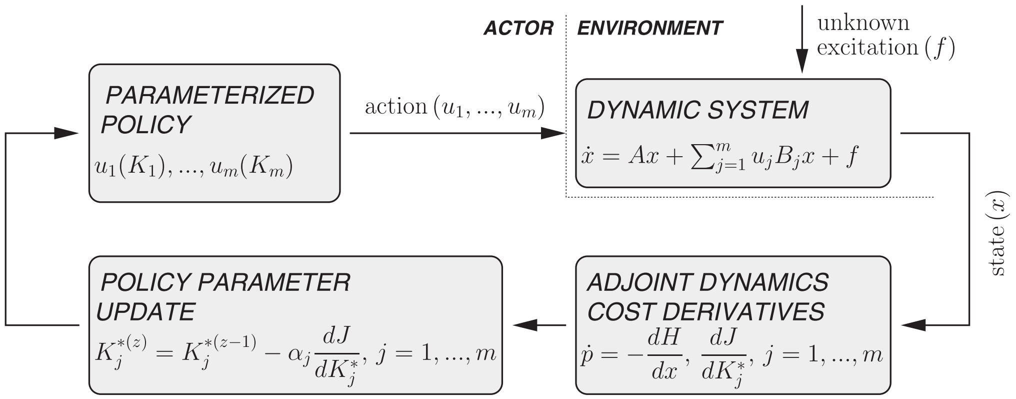
ACTOR
ENVIRONMENT
unknown excitation (f)
PARAMETERIZED POLICY
$u_1(K_1), ..., u_m(K_m)$
action $(u_1, ..., u_m)$
DYNAMIC SYSTEM
$\dot{x} = Ax + \sum_{j=1}^{m} u_j B_j x + f$
state (x)
POLICY PARAMETER UPDATE
$K_j^{*(z)} = K_j^{*(z-1)} - \alpha_j \frac{dJ}{dK_j^*}, j = 1, ..., m$
ADJOINT DYNAMICS COST DERIVATIVES
$\dot{p} = -\frac{dH}{dx}, \frac{dJ}{dK_j^*}, j = 1, ..., m$

FIGURE 2 Scheme of the reinforcement learning process. Updating the control policy parameters $K_1^*, \ldots, K_m^*$ to the unknown excitation force is based on the actor–environment (controller–dynamic system) interaction.

$j = 1, \ldots, m$, that is,

$$\left\| \frac{\mathrm{d}J}{\mathrm{d}K^*} \right\| = \max_{j=1,\ldots,m} \max_{q,r=1,\ldots,n} \left| \frac{\mathrm{d}J}{\mathrm{d}K_{qr}^{*j}} \right|. \tag{37}$$

4 | COMPARATIVE CONTROLS

To assess the efficiency of the developed method, it will be compared with the optimal open-loop solution, a heuristic control, and a passive strategy. The focus will be on optimality in suppressing the transient vibration in the control phase I. For that purpose, the open-loop optimal control for $t \in [0, T_f]$ will be computed assuming a complete information of the excitation $f(t)$. An examination of the overall stabilization capabilities, including the control phase I and the free vibration in the control phase II (i.e., for $t \in [0, T_c]$), will be performed by comparison to the heuristic and passive strategies.

4.1 | Open-loop optimal control

The open-loop optimal control $u_1^O(t), \ldots, u_m^O(t)$ for $t \in [0, T_f]$ will be established as the solution to the problem of minimizing the cost functional $J(T_f)$ as in Equation (2), that is,

$$\{u_1^O, \cdots, u_m^O\} = \underset{u_1,\cdots,u_m \in \mathcal{U}}{\operatorname{argmin}} \; J(T_f)$$

$$= \frac{1}{2} \int_0^{T_F} x^T(t) Q x(t) \mathrm{d}t$$

$$\text{subject to } \dot{x}(t) = Ax(t) + \sum_{j=1}^{m} u_j(t) B_j x(t) + f(t),$$

$$x(0) = x_0. \tag{38}$$

Assuming the set of admissible controls $\mathcal{U} = [u_{min}, u_{max}]$ and employing the Pontryagin Maximum Principle (Pontryagin et al., 1962) lead to the following solution to the problem Equation (38):

$$u_j^O(t) = \begin{cases} u_{\min}, p^T(t) B_j x(t) \le 0 \\ u_{\max}, p^T(t) B_j x(t) > 0, j = 1, \cdots, m, \end{cases} \tag{39}$$

where $p(t)$ stands for the adjoint state that is computed using the Hamiltonian associated to the problem Equation (38) (see Mohler, 1973). To determine the trajectories of $u_1^O(t), \ldots, u_m^O(t)$, the method based on the gradient descent will be used (see Pisarski, 2012).

4.2 | Heuristic control

Heuristic control is based on the concept of instantaneous optimization of the rates of change of the system energy (Pisarski, 2018). The control functions $u_1^H(t), \ldots, u_m^H(t)$ for $t \in [0, T_c]$ that provide the best instantaneous decrease of the energy $E(t)$ (see Equation (3) are given as the solution to the following problem:

$$u_j^H(t) = \underset{u_1,\cdots,u_m \in \mathcal{U}}{\operatorname{argmin}} \; \dot{E}(t), j = 1, \cdots, m. \tag{40}$$

Computing the time derivative of the energy function (see Equation 3):

$$\dot{E} = \dot{x}^T Q x + x^T Q \dot{x}, \tag{41}$$

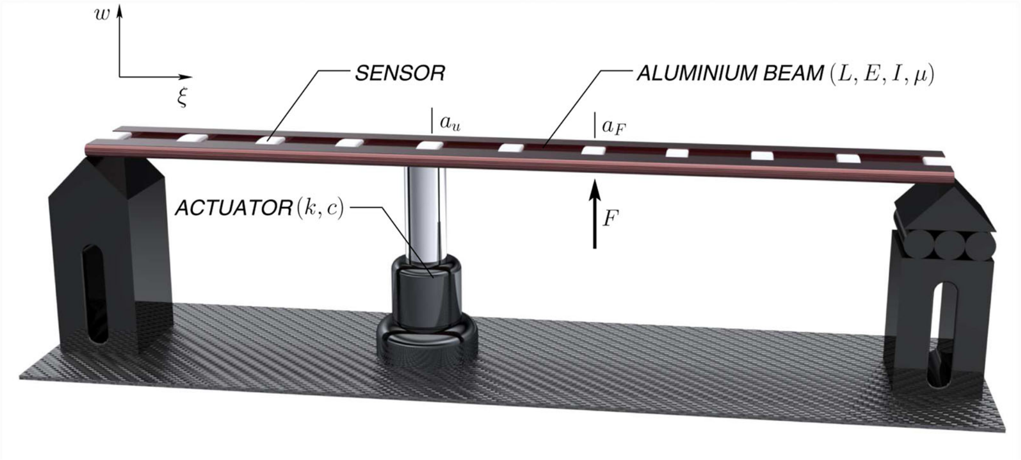

FIGURE 3 Scheme of the system used in the simulations: A simply supported aluminium beam subjected to unknown excitation and equipped with 11 state sensors and a semi-active device of controlled stiffness and damping parameters

using the symmetry of the matrix Q and inserting the state equation, Equation (1), into Equation (41) leads to

$$u_j^H(t) = \begin{cases} u_{\min}, x^T(t)Q_jB_jx(t) \geq 0 \\ u_{\max}, x^T(t)Q_jB_jx(t) < 0, j = 1, \cdots, m. \end{cases}$$

It can be observed that the control Equation (42) guarantees the asymptotic stability of Equation (1) if $f(t) = 0$ and A is a Hurwitz matrix (the last condition is fulfilled for a majority of structures as a consequence of material or viscous damping).

TABLE 1 Parameters of the investigated structure and actuator

Length of the beam (L)	1 (m)
Young's modulus (E)	70 (GPa)
Moment of inertia (I)	$0.7142 \cdot 10^{-10}$ (m^4)
Mass per unit length (μ)	0.2 (kg/m)
External damping coefficient (σ)	0.01 (Ns/m)
Actuator's stiffness coefficient (k)	200 (N/m)
Actuator's damping coefficient (c)	0.5 (Ns/m)

4.3 | Passive strategy

In this method, constant control functions $u_1^P(t), \ldots, u_m^P(t)$ for $t \in [0, T_c]$ will be assumed, where each actuator operates at the maximal admissible value; that is, $u_j^P(t) = u_{max}$, $j = 1, \ldots, m$. In the majority of the semi-active controlled structures, this operation is equivalent to the optimal passive strategy (see, e.g., Szmidt et al., 2017).

5 | CASE STUDY

5.1 | The analyzed structure

A span structure supported by a semi-active actuator will be investigated as depicted in Figure 3. For the span, a slender elastic body is assumed that is subjected to small deflections. The height and the depth of the span are small when compared to the length L. The span can be thus represented by the Euler–Bernoulli beam equation parameterized by the bending stiffness EI and length density μ. It is subjected to an external damping of air that is characterized by the coefficient σ. The semi-active actuator is attached at position $a_u = 0.4L$. Note that the ith mode shape of the assumed simply supported beam at the coordinate ξ is characterized by $\theta_i(\xi) = \sin(i\pi\xi/L)$, and for the first four modes $i = 1, \ldots, 4$ the assumed actuator's position guarantees $\theta_i(a_u) \neq 0$. Therefore, the actuator's position allows the first four modes to be controlled, but besides, it is selected arbitrarily. For the actuator, a controlled input u is assumed that influences the damping $c(u)$ and stiffness $k(u)$ parameters. Each of these parameters depends linearly on the control variable, that is,

$$k(u) = uk, c(u) = uc, \tag{43}$$

where k and c are assumed to be constant. The force generated by the actuator is assumed to be equal to the sum of the elastic and damping forces that are, respectively, proportional to the beam's traverse deflection and velocity at point a_u. The unknown external force F is of short duration and acts on the span at point $a_F = 0.6L$. The parameters assumed for the simulations are listed in Table 1.

A deflection of the span at the coordinate ξ and time t is denoted by $w(\xi, t)$. The Dirac delta function $\delta(\xi)$ is used to describe the contact point between the span and actuator or external force. Based on these assumptions, the structure can be represented by the following partial differential equation:

$$EI\frac{\partial^4 w(\xi,t)}{\partial \xi^4} + \sigma\frac{\partial w(\xi,t)}{\partial t} + \mu\frac{\partial^2 w(\xi,t)}{\partial t^2}$$
$$= -\left(u(t)kw(a_u,t) + u(t)c\frac{\partial w(a_u,t)}{\partial t}\right)\delta(\xi - a_u)$$
$$+F(t)\delta(\xi - a_F). \tag{44}$$

The left-hand side of Equation (44) consists of the common elements of the Euler–Bernoulli beam equation that characterize the potential, air damping, and inertial forces of the span. On the right-hand side there are the terms that stand for the viscoelastic forces generated by the semi-active actuators and unknown force acting on the structure. The assumed endpoint supports (see Figure 3) enforce the boundary conditions:

$$\begin{aligned} w(0,t) &= 0, (L,t) = 0,\\ \tfrac{\partial^2 w(0,t)}{\partial \xi^2} &= 0, \tfrac{\partial^2 w(L,t)}{\partial \xi^2} = 0. \end{aligned}$$

For each simulation, a zero initial condition is assumed:

$$w(\xi,0) = 0, \dot{w}(\xi,0) = 0\text{for}\xi \in [0,L]. \tag{46}$$

The finite element method is employed to represent Equation (44) in the form of an ordinary differential equation, as in Equation (1). For the span structure, 10 identical elements and 11 uniformly distributed nodes are used (nodes 1 and 11 are located at positions $\xi = 0$ and $\xi = L$, respectively). Introducing the vector of nodal displacements $(Y_1, \ldots, Y_{11}) = (w_1, \ldots, w_{11})$ $(w_i, i = 1, \ldots, 11$ represents the span's displacement at the ith node's position) and angles of rotation $(Y_{12}, \ldots, Y_{22}) = (\phi_1, \ldots, \phi_{11})$ $(\phi_i,\ i = 1, \ldots, 11$ represents the span's angle of rotation at the ith node's position), the system Equation (44) can be approximated by the second-order differential equation:

$$M\ddot{Y}(t) + D\dot{Y}(t) + SY(t)$$
$$= -u(t)H_2\dot{Y}(t) - u(t)H_1Y(t) + \bar{F}(t) \tag{47}$$

In Equation (47), M, D, and S are, respectively, the 22×22 mass, damping, and stiffness matrices, H_1, H_2 are the 22×22 matrices that accommodate the elastic and damping forces generated by the actuators, and $\bar{F}$ represents the 22×1 vector that incorporates an unknown external force. The composition of these terms results from a standard finite element approach that involves the shape functions based on a third-degree polynomial (Bathe, 1996). Define the state vector:

$$x = [x_1, \cdots, x_{44}]^T = [Y_1, \cdots, Y_{22}, \dot{Y}_1, \cdots, \dot{Y}_{22}]^T, \tag{48}$$

the system matrices as:

$$A = \begin{bmatrix} 0 & I \\ -M^{-1}S & -M^{-1}D \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ -M^{-1}H_1 & -M^{-1}H_2 \end{bmatrix}, \tag{49}$$

and the external force vector:

$$f(t) = \begin{bmatrix} 0 \\ M^{-1}\bar{F}(t) \end{bmatrix}. \tag{50}$$

In Equation (49), 0 and I denote 22×22 zero and identity matrix, respectively. In Equation (50), 0 stands for 22×1 zero vector. Using Equations (48)–(50), the system Equations (44)–(46) can be eventually represented in the form of a first-order ordinary differential equation, as in Equation (1):

$$\dot{x}(t) = Ax(t) + u(t)Bx(t) + f(t), x(0) = 0. \tag{51}$$

From Equations (47) and (48), it follows that the energy matrix Q in Equation (3) is computed as:

$$Q = \begin{bmatrix} S & 0 \\ 0 & M \end{bmatrix}. \tag{52}$$

To reconstruct the state vector Equation (48), it is assumed that 11 state sensors are located at the structure's node positions (see Figure 3). The sensors permanently collect the local information of the transverse displacements $w_1, \ldots, w_{11}$, the transverse velocities $\dot{w}_1, \ldots, \dot{w}_{11}$, the angles of rotation $\phi_1, \ldots, \phi_{11}$, and the angular velocities $\dot{\phi}_1, \ldots, \dot{\phi}_{11}$.

5.2 | Controller settings

For each control function u (Section 3.1), u^0 (Section 4.1), u^H (Section 4.2), and u^P (Section 4.3), it is assumed that $u_{min} = 0.02$ and $u_{max} = 1$. To simulate the transient response of the semi-active device, any change between the extreme control values is realized at a constant rate, and this change takes 0.002 (s); for an alternative filtering approach, see Wang and Adeli (2015a). For the control time, $T_c = 2$ (s) is selected (see Figure 1). An unknown excitation force $F(t) \neq 0$ is acting on a structure for $T_f = 0.2$ (s).

For the designed RL algorithm, the maximal number of iterations $z_{max} = 500$ (Step 1) was assumed and the terminal condition parameter $\epsilon = 0.001$ (Step 6). The learning process was repeated for three lengths of the learning time window: $T_l = 0.4\,T_f$, $T_l = 0.75\,T_f$, and $T_l = T_f$. Based on several test runs, the step size for the updating sequence Equation (15) was selected as $\alpha_j = 0.0015/s$ for $T_l = 0.4\,T_f$ and $T_l = 0.75\,T_f$, and $\alpha_j = 0.0005/s$ for $T_l = T_f$, where s was computed at each iteration using Equation (31) (see Step 1 and remark R1). To solve the adjoint state equation (Step 3), the Runge–Kutta fourth-order scheme was employed with a time step of 0.0001 (s). The controller was implemented in the MATLAB programming language and run using a workstation with an Intel Xeon, 3.00 GHz, 16 GB, that operated on the Linux platform.

5.3 | Simulation results

The proposed method will be examined for three scenarios of the external excitation force $F(t)$. In the first scenario (Case A), the excitation will repeat at each learning iteration with an identical frequency. The convergence of the policy parameters will be then analyzed, as well as the stabilization performance, in comparison to the optimal, heuristic, and passive strategies. The two subsequent scenarios (Cases B and C) will validate the robustness of the learning process. Here, the policy parameter will be updated for the excitation force with either a randomly perturbed frequency (Case B) or an additional high-frequency harmonic perturbation of a random amplitude (Case C). For the comparisons, the cost functional as in Equation (2) will be used that is computed either for the transient vibration in the control phase I—that is, applying $J(T_f)$—or for the overall process, including the transient and free vibration in control phases I and II, assuming $J(T_c)$. Finally, the computational effort will be assessed by varying the number of finite elements that compose the structure from 10 to 100.

5.3.1 | Case A

The external excitation is assumed to constantly repeat for $T_f = 0.2$ (s) with an identical characteristic given by $F(t) = \mathcal{A}\sin(2\pi\omega t)$, where amplitude $\mathcal{A} = 100$ (N) and frequency $\omega = 25$ (Hz) are constant. The duration between subsequent repetitions is assumed to be sufficiently large. Therefore, for each repetition, zero initial conditions are assumed for Equation (51). It can be emphasized that the specified excitation parameters are unknown to the designed controller. The reader can also observe that the proposed controller does not require any information on which point of a structure the force is acting on.

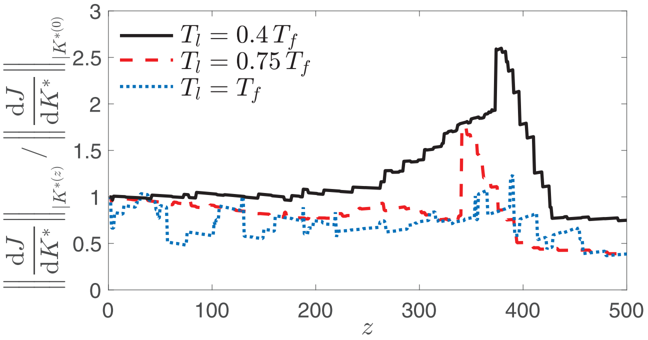

FIGURE 4 Evolution of the derivative norm with respect to the updating iteration for the assumed learning time windows. For each case, the norm is normalized to its initial value.

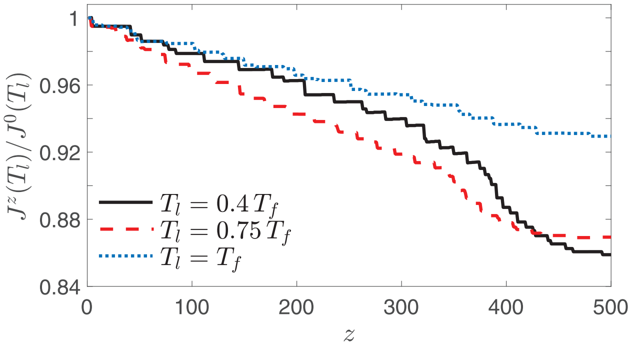

FIGURE 5 Evolution of the cost functional value computed for the assumed learning time windows. For each case, the cost is normalized to its initial value.

To investigate the proposed control learning process realised through Algorithm 1, the evolution of the derivative norm (see Equation 37) with respect to the updating iterations can be analyzed for the assumed learning time windows, as depicted in Figure 4. For each case, the optimization procedure was terminated by the condition $z = z_{max}$. For each of the curves, smooth sections can be observed that are separated by instant jumps. The latter indicates the iterations where the derivative changes the number s of added up components (see Equation 31). Although there are the sections where the value of the derivative norm is increasing, the general downward trend (for each case the final value is lower than the initial one) validates the convergence of the sequence for updating the policy parameter Equation (15). The convergence of the steepest gradient procedure can also be inspected by analyzing the evolution of the cost functional value Equation (2), as presented in Figure 5. Here, a substantial decrease in the value of the cost functional can be observed for each of the assumed learning time windows $T_l = 0.4\,T_f$, $T_l = 0.75\,T_f$, and $T_l = T_f$. The notably slower

TABLE 2 Comparison of the cost functional $J(T_f)$ obtained in the case of the designed method and comparative controls. For each of the control cases, the values are normalized to the passive strategy.

			RL control				
		Passive strategy	$T_l = 0.4T_f$	$T_l = 0.75T_f$	$T_l = T_f$	**Heuristic control**	**Optimal control**
Case A	$J(T_f)$	1.0000	0.6416	0.6412	0.6956	0.8186	0.6340

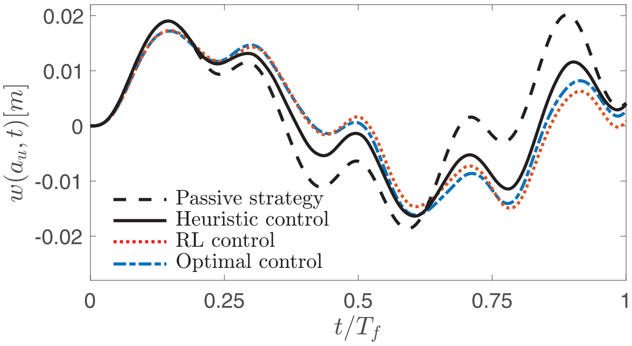

FIGURE 6 Comparison of the beam's deflection for the considered controllers measured for the control phase I at the actuator's position

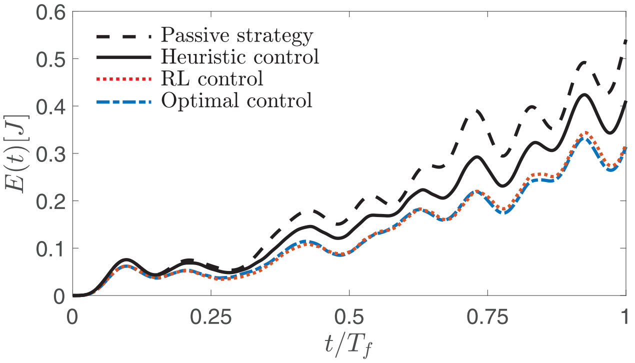

FIGURE 7 Comparison of the system's energy for the considered controllers measured for the control phase I

convergence rate in the case of $T_l = T_f$ follows from the selection of a lower step size compared to the remaining cases (see Section 5.2). This selection was motivated by the fact that a longer learning time window has a larger variation in number s (see Equation 31), which in turn implies a larger variation in the derivative value Equation (30). The assumed lower step size in the case of the largest time window allowed overshooting to be avoided and guaranteed a stable reduction of the cost functional value. It should be noted that a too short learning time window may significantly degrade the control efficiency. The piecewise sections in the cost trajectories are concerned with the discrete-time solution of Equation (51), which for some subsequent algorithm iterations results in identical switching times of the control policy function Equation (4). The effect of some policy parameter updates on the state response cannot be then detected.

After completing Algorithm 1 for $T_l = 0.4T_f$, $T_l = 0.75T_f$, and $T_l = T_f$, the comparative study can be performed. Regarding the assumed cost functional Equation (2), which is computed for the transient state vibration at the control phase I—that is, for $t \in [0, T_f]$ (see Section 3)—the RL-based control (in short, RL control) exhibits a comparable performance for all of the assumed learning time windows (see Table 2). Moderately poorer efficiency in the case of $T_l = T_f$ is concerned with the previously investigated convergence in the learning protocol. Assuming $T_l = 0.75T_f$, the RL control is marginally outperformed by the optimal one by 1.13%. Compared to the heuristic and passive strategies, this RL control obtains a cost reduction of 27.6% and 55.9%, respectively. Figure 6 depicts the beam's deflection simulated for the control phase I at the location of the semi-active device a_u for the RL controller of the learning time window $T_l = 0.75T_f$ with the other controllers. It can be observed that the RL and optimal control result in almost identical responses for $t < 0.5T_f$. For the remaining time, a gradual divergence of the trajectories generated by the RL and optimal control can be detected with 30.6% of the relative difference in their last peak amplitudes at $t \approx 0.9T_f$. A significantly larger divergence in the deflection trajectories can be found in the case of heuristic and passive strategies, where (compared to the RL control) the final peak amplitude has increased by 84.8% and 221%, respectively. The similarity of the dynamic response of the system to the RL and optimal control is also confirmed by the characteristics of the energy function Equation (3), as demonstrated in Figure 7. For the ending time of control phase I, the RL control results in a negligible increase of 1.91% of the energy when compared to the optimal solution. For the heuristic and passive methods, this increase is 31.0% and 72.1%, respectively.

Figure 8 compares the switching patterns of the RL control of $T_l = 0.75T_f$ and the optimal open-loop control. They do not match, although both functions are generated through the optimization of the same cost functional. This mismatch is essentially concerned with the state-feedback structuring that is imposed on the RL control. To a lesser, but not negligible, extent, the observed mismatch is caused by different selections of the time horizon assumed for the optimization (respectively, $0.75T_f$ and T_f in the case of the RL and optimal control). Even though the RL control is unable to reproduce the optimal control pattern, the critical switches of the optimal control are here replicated much more accurately than in case of the heuristic control. In particular, the first two switching actions of

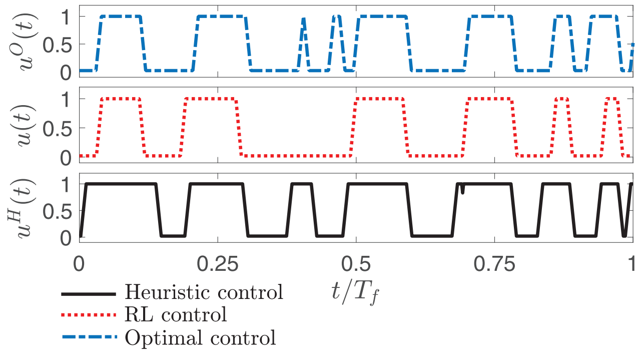

FIGURE 8 Comparison of the control signals for the control phase I

the RL and optimal controls at $t \approx 0.03\,T_f$ and $t \approx 0.11\,T_f$ almost coincide, while these switching times are evidently advanced and retarded in the case of the heuristic control.

The overall stabilization performance of the proposed method can be justified by investigating the deflection (Figure 9a), energy (Figure 9b), and frequency (Figure 9c) characteristics that are obtained for the transient and free vibration—that is, for $t \in [0, T_c]$ (control phases I and II)—where the RL control of $T_l = 0.75\,T_f$ and the heuristic strategy were employed. For deflection and energy trajectories, the RL control resulted in a reduction of the peak amplitudes for phase I and faster convergence to the equilibrium for phase II. The reduction of the deflection amplitudes is also confirmed by the frequency characteristics (Figure 9c), where we observe a decrease of 19.4% of the peak value at 6 Hz, which corresponds to the first natural frequency of the beam structure. The implementation of the RL control also resulted in a decrease of the cost functional value $J(T_c)$ by 15.2% and 47.8% when compared to the heuristic and passive methods, respectively.

5.3.2 | Case B

The following simulation was carried out to investigate how the changes in the frequency of the excitation force during the realization of Algorithm 1 influence the efficiency of the resulting RL control. For this purpose, the learning time window $T_l = 0.75\,T_f$ was selected and $z = 500$ iterations were performed following Steps 2–6. At each iteration, for the excitation force, $F(t) = \mathcal{A}\,\sin(2\pi(\omega + \delta\omega(z))\,t)$ was assumed. Here, similarly to Case A, the constant amplitude $\mathcal{A} = 100$ (N) was used, but the frequency was given by $\omega + \delta\omega(z)$, where $\omega = 25$ (Hz) and $\delta\omega(z)$ is a perturbation that was selected randomly at each iteration z and fulfilled the condition $\max_{z=1,\ldots,500}|\delta\omega(z)|/\omega \leq \delta_\omega$. The learning protocol was performed while assuming different frequency perturbation magnitudes δ_ω, from 0.05 to 0.20. For each perturbation magnitude, the algorithm was repeated three times. Next, each of the obtained RL controls was applied to the unperturbed case—that is, assuming $F(t) = \mathcal{A}\,\sin(2\pi\omega t)$. Eventually, for each perturbation magnitude, the cost functional value $J(T_f)$ was computed and averaged for the assumed three repetitions. The attained controls were compared to the corresponding RL control computed in the previous section; that is, for $\delta_\omega = 0$ (see Table 3). The increase of the cost functional value remains below 0.5% for all of the considered cases, which confirms that moderate perturbations have no significant impact on the control performance. The learning protocol was also carried out for $\delta_\omega > 0.2$, where difficulties gradually appeared in selecting a relevant step size for a stable cost descending (concerned with an increased variation in the cost derivative values). As a result, there was a loss in the control performance (8.2% increase in the cost functional value for $\delta_\omega = 0.3$).

5.3.3 | Case C

In order to examine the developed method for a more complex excitation force, Algorithm 1 was executed assuming $F(t) = \mathcal{A}\,\sin(2\pi\omega t) + \mathcal{A}^d(z)\,\sin(2\pi\omega^d t)$. Here, the first term stands for the dominant harmonic excitation where—as in Case A—the constant amplitude $\mathcal{A} = 100$ (N) and frequency $\omega = 25$ (Hz) were assumed. The second term characterizes an additional harmonic disturbance of a constant frequency $\omega^d = 100$ (Hz) and an amplitude $\mathcal{A}^d(z)$ that was randomly selected for each learning iteration $z = 1, \ldots, 500$ and fulfilled the condition $0 \leq \max_{z=1,\ldots,500}\mathcal{A}^d(z) \leq \mathcal{A}^d$. Assuming the learning time window $T_l = 0.75\,T_f$, the procedure was carried out for different values of $\mathcal{A}^d$, ranging from from 20 to 100 (N); that is, 10%–50% of the amplitude $\mathcal{A}$ of the dominant excitation. For each limiting value $\mathcal{A}^d$ the procedure was repeated three times. The obtained RL controls were then applied in two scenarios. In the first scenario (Case C1), it was assumed that the excitation force is unperturbed; that is, inserting $F(t) = \mathcal{A}\,\sin(2\pi\omega t)$. In the second scenario (Case C2), the applied excitation force included the additional harmonic disturbance with a constant amplitude equal to half of the limiting value $\mathcal{A}^d$, namely, $F(t) = \mathcal{A}\,\sin(2\pi\omega t) + 0.5\mathcal{A}^d\,\sin(2\pi\omega^d t)$. For each scenario and limiting value $\mathcal{A}^d$, the cost functional value $J(T_f)$ was computed and averaged for the assumed three repetitions (see Table 4). Analyzing the cost values obtained for Case C1, where the drop of the control performance for each of the perturbed cases remains below 1.9%, it can be concluded that the learning algorithm is robust to disturbances imposed on the dominant characteristic of the excitation. Furthermore, the results obtained for Case C2 confirm that the method can be successfully applied for more complex polyharmonic forces, also in the case of random perturbations of the amplitude.

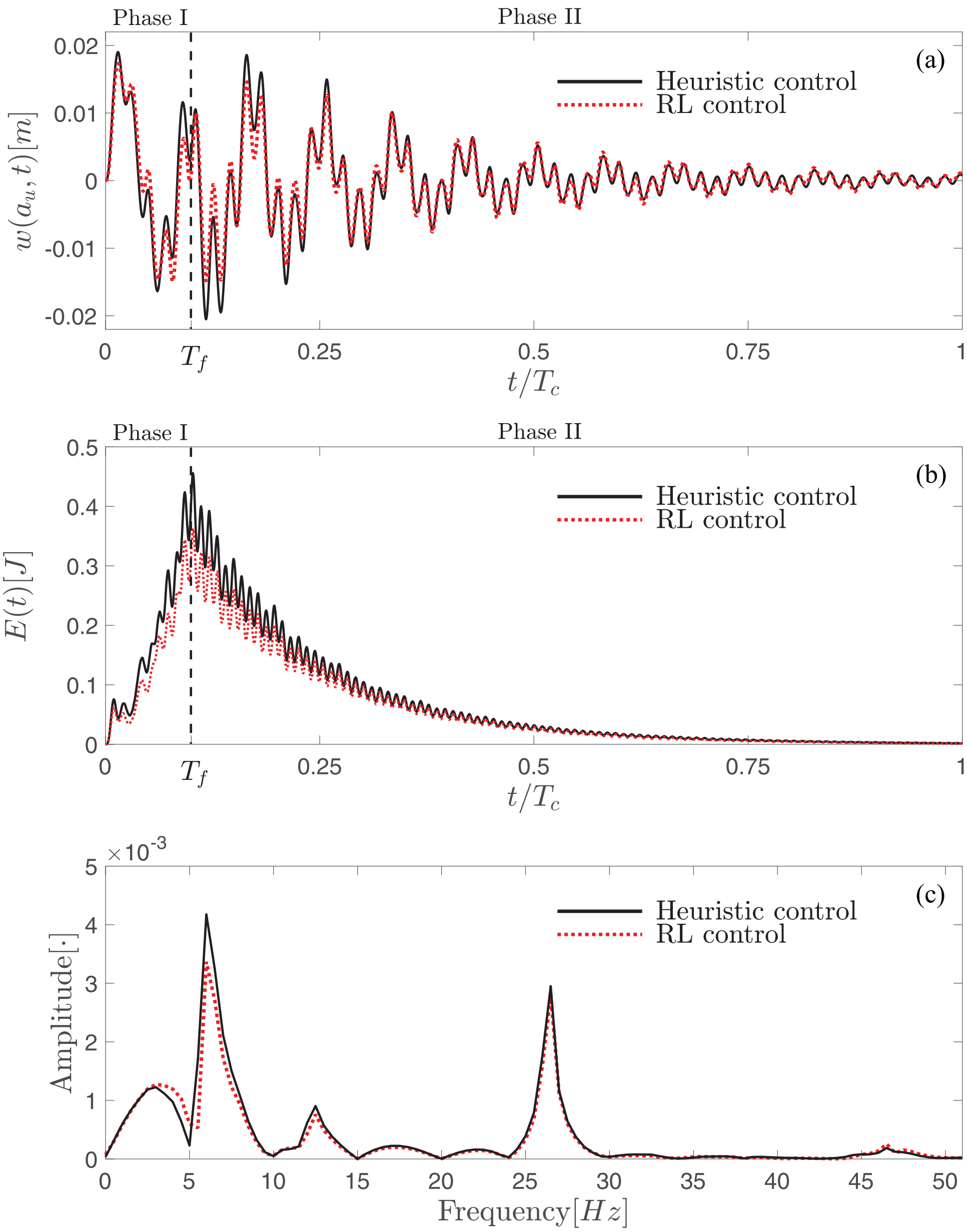

FIGURE 9 Comparison of the beam's deflection for the RL and heuristic controllers measured for control phases I and II at the actuator's position (a). Comparison of the system's energy for the RL and heuristic controllers measured for control phases I and II (b). Frequency characteristics of the deflection signals presented in (c)

TABLE 3 Comparison of the cost functional $J(T_f)$ in the case of the RL control of $T_l = 0.75\,T_f$ where the updating algorithm is performed with different magnitudes of perturbations in the frequency of the excitation force. For each perturbation magnitude the cost functional value was averaged for the assumed three repetitions and normalized to the case with no perturbation.

Frequency perturbation	**$\delta_\omega = 0$**	**$\delta_\omega = 0.05$**	**$\delta_\omega = 0.10$**	**$\delta_\omega = 0.15$**	**$\delta_\omega = 0.20$**
Case B $J(T_f)$	1.0000	1.0002	1.0011	1.0026	1.0044

5.3.4 | The computational effort

The final set of simulations was performed to investigate the capabilities of the designed algorithm with respect to systems with a larger number of state variables. The aim was to analyze the computational time required for updating the policy parameter (Steps 3–6 in Algorithm 1), when assuming different sizes for the dynamic Equation (51). In Cases A–C, the investigated structure was represented by 10 finite elements that resulted in the state vector of 44 components. The learning procedure was repeated assuming 20, 40, 60, 80, and 100 elements in the finite element

TABLE 4 Comparison of the cost functional $J(T_f)$ in the case of the RL control of $T_l = 0.75\,T_f$ where the updating algorithm is performed with an additional harmonic disturbance of the constant frequency and a randomly selected amplitude limited by different boundaries $\mathcal{A}^d$. In Case C1, each value is averaged taking into account the assumed three repetitions and normalized to the case with no disturbance; that is, when $\mathcal{A}^d = 0$. In Case C2, each value is averaged taking into account the assumed three repetitions and normalized to the case where the learning protocol was carried out for constant amplitude of the additional disturbance that was equal to the half of the limiting value $\mathcal{A}^d$.

Disturbance amplitude	$\mathcal{A}^d = 0$	$\mathcal{A}^d = 20$	$\mathcal{A}^d = 40$	$\mathcal{A}^d = 60$	$\mathcal{A}^d = 80$	$\mathcal{A}^d = 100$
Case C1 $J(T_f)$	1.0000	1.0001	1.0006	1.0033	1.0079	1.0186
Case C2 $J(T_f)$	1.0000	1.0002	1.0011	1.0016	1.0028	1.0074

TABLE 5 Comparison of the computational times required for updating the policy parameter when assuming different numbers of elements in the finite element model Equation (47)

Number of finite elements	10	20	40	60	80	100
Computational time (s)	0.077	0.124	0.552	0.818	1.209	8.163

model, which resulted in the state vector in Equation (51) of the size 84, 164, 244, 324, and 404, respectively. The greatest computational cost was concerned with the integration of the adjoint state equation in Step 3 (performed using the Runge–Kutta fourth-order scheme), which for the assumed time step of 0.0001 (s) and learning time window $T_l = 0.75\,T_f$ required 1500 time samples. The obtained computational times are summarized in Table 5. A single iteration in the case of 60 finite elements (244 components in the state vector) remained below 1 (s). Furthermore, the use of the steepest descent approach to update the policy parameter guaranteed that an increase in the size of the system did not significantly influence the rate of convergence in the overall learning process (in the case of the 60 finite elements, the algorithm required 612 iterations to reach the same cost functional value as in Case A). It can be concluded that the method can be effectively used in multidimensional systems.

6 | CONCLUSIONS

An RL-based semi-active control method for suppressing structural vibration that is induced by unknown harmonic excitation has been proposed. This method relies on a state-feedback switching control law that includes a parameter matrix to be updated by means of the developed actor-only iterative learning algorithm. In view of the stabilization performance, this method can be perceived as suboptimal. Its efficiency has been validated via numerical experiments for a span structure that is equipped with an actuator of controlled stiffness and damping parameters. In terms of the assumed energy-related cost functional, the method resulted in a marginal degradation (of 1.13%) when compared to the optimal open-loop control, while it significantly outperformed a heuristic control (by 27.6%) that employed an analogous control law and an identical amount of state information. The relatively low computational burden of the proposed iterative learning algorithm allows this method to be applied to multidimensional systems (a single iteration required less than 0.08 (s) for the system represented by the state vector of 44 components). The method has been designed and validated for repetitive transient vibration. Nevertheless, assuming an appropriately selected moving learning time window, the proposed algorithm can also be adapted to a steady-state vibration. The ongoing works include the development of an adaptive scheme to select the moving learning time window that guarantees the best convergence of the updating sequence and the design of a test stand platform that simulates a real environment for experimental validation.

ACKNOWLEDGMENT

This research has been supported by the National Science Centre, Poland, under grant agreements 2020/39/B/ST8/02615 and DEC-2017/26/D/ST8/00883.

REFERENCES

Li, Z., & Adeli, H. (2016). New discrete-time robust h_2/h_∞ algorithm for vibration control of smart structures using linear matrix inequalities. *Engineering Applications of Artificial Intelligence*, *55*, 47–57.

Adam, B., & Smith, I. (2008). Reinforcement learning for structural control. *Journal of Computing in Civil Engineering*, *22*(1), 133–139.

Adeli, H., & Hung, S. (1994). *Machine learning : Neural networks, genetic algorithms and fuzzy systems*. John Wiley and Sons Ltd.

Adeli, H., & Saleh, A. (1997). Optimal control of adaptive/smart bridge structures. *Journal of Structural Engineering*, *123*(2), 218–226.

Amezquita-Sancheza, J., Valtierra-Rodriguez, M., & Adeli, H. (2020). Machine learning in structural engineering. *Scientia Iranica*, *27*(6), 2645–2656.

Babu, M., Oza, Y., Singh, A. K., Krishna, K. M., & Medasani, S. (2018). Model predictive control for autonomous driving based on time scaled collision cone. In *2018 European control conference* (pp. 641–648). IEEE.

Bathe, K. J. (1996). *Finite element procedures*. Prentice-Hall.

Bitaraf, M., Hurlebaus, S., & Barroso, L. R. (2012). Active and semi-active adaptive control for undamaged and damaged building structures under seismic load. *Computer-Aided Civil and Infrastructure Engineering*, *27*(1), 48–64.

Bordons, C., & Camacho, E. (1998). A generalized predictive controller for a wide class of industrial processes. *IEEE Transactions on Control Systems Technology*, *6*, 372–387.

Chicone, C. (2006). *Ordinary differential equations with applications*. Springer Science and Business Media.

Cundumi, O., & Suárez, L. E. (2008). Numerical investigation of a variable damping semiactive device for the mitigation of the seismic response of adjacent structures. *Computer-Aided Civil and Infrastructure Engineering*, *23*(4), 291–308.

Dengler, C., & Lohmann, B. (2018). Actor-critic reinforcement learning for the feedback control of a swinging chain. *IFAC Papers*, *51*, 378–383.

Ferrara, A., Sacone, S., & Siri, S. (2015). Event-triggered model predictive schemes for freeway traffic control. *Transportation Research Part C: Emerging Technologies*, *58*, 554–567.

Ghaedi, K., Ibrahim, Z., Adeli, H., & Javanmardi, A. (2017). Invited review: Recent developments in vibration control of building and bridge structures. *Journal of Vibroengineering*, *19*(5), 3564–3580.

Grondman, I. (2015). *Online model learning algorithms for actor-critic control* (doctoral thesis). Delft Center for Systems and Control, TU Delft, The Netherlands.

Gutierrez Soto, M., & Adeli, H. (2017a). Many-objective control optimization of high-rise building structures using replicator dynamics and neural dynamics model. *Structural and Multidisciplinary Optimization*, *56*(6), 1521–1537.

Gutierrez Soto, M., & Adeli, H. (2017b). Multi-agent replicator controller for sustainable vibration control of smart structures. *Journal of Vibroengineering*, *19*(6), 4300–4322.

Gutierrez Soto, M., & Adeli, H. (2017c). Recent advances in control algorithms for smart structures and machines. *Expert Systems*, *34*(2), e12205.

Gutierrez Soto, M., & Adeli, H. (2018). Vibration control of smart base-isolated irregular buildings using neural dynamic optimization model and replicator dynamics. *Engineering Structures*, *156*, 322–336.

Gutierrez Soto, M., & Adeli, H. (2019). Semi-active vibration control of smart isolated highway bridge structures using replicator dynamics. *Engineering Structures*, *186*, 536–552.

Jiao, Y., Ling, F., Heydari, S., Heess, N., Merel, J., & Kanso, E. (2021). Learning to swim in potential flow. *Physical Review Fluids*, *6*, 050505.

Khalatbarisoltani, A., Soleymani, M., & Khodadadi, M. (2019). Online control of an active seismic system via reinforcement learning. *Structural Control and Health Monitoring*, *26*(3), e2298.

Li, Z., & Adeli, H. (2018). Control methodologies for vibration control of smart civil and mechanical structures. *Expert systems*, *35*(6), e12354.

Liberzon, D. (2012). *Calculus of variations and optimal control theory: A concise introduction*. Princeton University Press.

Mohler, R. R. (1973). *Bilinear control processes*. Academic Press.

Naderpoor Shad, P., & Taghikhany, T. (2022). Seismic adaptive control of building structures with simultaneous sensor and damper faults based on dynamic neural network. *Computer-Aided Civil and Infrastructure Engineering*, *37*(11), 1402–1416.

Naderpoor Shad, P., & Taghikhany, T. (2022). Seismic adaptive control of building structures with simultaneous sensor and damper faults based on dynamic neural network. *Computer-Aided Civil and Infrastructure Engineering*, *37*(11), 1402–1416.

Nagendra, S., Podila, N., Ugarakhod, R., & George, K. (2017). Comparison of reinforcement learning algorithms applied to the cart-pole problem. In *2017 international conference on advances in computing, communications and informatics* (pp. 26–32). IEEE.

Nedeljkov, M., & Oberguggenberger, M. (2012). Ordinary differential equations with delta function terms. *Publications de l'Institut Mathématique*, *91*(105), 125–135.

Ostrowski, M., Blachowski, B., Poplawski, B., Pisarski, D., Mikulowski, G., & Jankowski, L. (2021). Semi-active modal control of structures with lockable joints: General methodology and applications. *Structural Control and Health Monitoring*, *28*(5), e2710.

Oveisi, A., Hosseini-Pishrobat, M., Nestorović, T., & Keighobadi, J. (2018). Observer-based repetitive model predictive control in active vibration suppression. *Structural Control and Health Monitoring*, *25*, e2149.

Pepe, G., & Carcaterra, A. (2016). VFC—Variational feedback controller and its application to semi-active suspensions. *Mechanical Systems and Signal Processing*, *76*, 72–92.

Pisarski, D. (2012). *Semi-active control system for trajectory optimization of a moving load on an elastic continuum* (doctoral thesis). Institute of Fundamental Technological Research of Polish Academy of Sciences, Warsaw, Poland.

Pisarski, D. (2018). Decentralized stabilization of semi-active vibrating structures. *Mechanical Systems and Signal Processing*, *100*, 694–705.

Pisarski, D., & Myśliński, A. (2017). Online adaptive algorithm for optimal control of structures subjected to travelling loads. *Optimal Control Applications and Methods*, *38*(6), 1168–1186.

Pontryagin, L., Boltyanskii, V. G., & Gamkrelidze, R. V. (1962). *The mathematical theory of optimal processes*. Wiley.

Popławski, B., Mikułowski, G., Pisarski, D., Wiszowaty, R., & Jankowski, Ł. (2019). Optimum actuator placement for damping of vibrations using the prestress-accumulation release control approach. *Smart Structures and Systems*, *24*(1), 27–35.

Qiu, Z., Chen, G., & Zhang, X. (2021). Reinforcement learning vibration control for a flexible hinged plate. *Aerospace Science and Technology*, *118*, 107056.

Reddy, G., Celani, A., Sejnowski, T., & Vergassola, M. (2016). Learning to soar in turbulent environments. *Proceedings of the National Academy of Sciences*, *113*(33), E4877–E4884.

Runlin, Y., Xiyuan, Z., & Xihui, L. (2002). Seismic structural control using semi-active tuned mass dampers. *Earthquake Engineering and Engineering Vibration*, *1*(1), 111–118.

Sallab, A., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, *19*, 70–76.

Sastry, S. (1999). *Nonlinear systems: Analysis, stability, and control*. Springer.

Shi, H., Nie, Q., Fu, S., Wang, X., Zhou, Y., & Ran, B. (2021). A distributed deep reinforcement learning-based integrated dynamic bus control system in a connected environment. *Computer-Aided Civil and Infrastructure Engineering*, *Early view*, 1–17.

Shi, H., Zhou, Y., Wang, X., Fu, S., Gong, S., & Ran, B. (2022). A deep reinforcement learning-based distributed connected automated vehicle control under communication failure. *Computer-Aided Civil and Infrastructure Engineering*, *Early view*, 1–19.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Sutton, R., & Barto, A. (2020). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Szmidt, T., Pisarski, D., Bajer, C., & Dyniewicz, B. (2017). Double-beam cantilever structure with embedded intelligent damping block: Dynamics and control. *Journal of Sound and Vibration*, *401*, 127–138.

Szmidt, T., Pisarski, D., Konowrocki, R., Awietjan, S., & Boczkowska, A. (2019). Adaptive damping of a double-beam structure based on magnetorheological elastomer. *Shock and Vibration*, *2019*, 8526179–1–16.

Szolc, T., Konowrocki, R., Pisarski, D., & Pochanke, A. (2019). Influence of various control strategies on transient torsional vibrations of rotor-machines driven by asynchronous motors. In Cavalca, K. L., Weber, H. I. (Eds.), *Proceedings of the 10th international conference on rotor dynamics—IFToMM* (pp. 205–220). Springer International Publishing.

Takacs, G., & Rohal'-Ilkiv, B. (2014). Model predictive control algorithms for active vibration control: A study on timing, performance and implementation properties. *Journal of Vibration and Control*, *20*(13), 2061–2080.

Vali, M., Petrović, V., Boersma, S., van Wingerden, J.-W., Pao, L., & Kühn, M. (2019). Adjoint-based model predictive control for optimal energy extraction in waked wind farms. *Control Engineering Practice*, *84*, 48–62.

Wang, N., & Adeli, H. (2015a). Robust vibration control of wind-excited highrise building structures. *Journal of Civil Engineering and Management*, *21*(8), 967–976.

Wang, N., & Adeli, H. (2015b). Self-constructing wavelet neural network algorithm for nonlinear control of large structures. *Engineering Applications of Artificial Intelligence*, *41*, 249–258.

Wasilewski, M., & Pisarski, D. (2020). Adaptive semi-active control of a beam structure subjected to a moving load traversing with time-varying velocity. *Journal of Sound and Vibration*, *481*, 115404–1–20.

Wasilewski, M., Pisarski, D., & Bajer, C. (2019). Adaptive optimal control for seismically excited structures. *Automation in Construction*, *106*, 102885.

Yuen, K., Shi, Y., & Beck, J. (2007). Structural protection using MR dampers with clipped robust reliability-based control. *Structural and Multidisciplinary Optimization*, *34*, 431–443.

Zelleke, D. H., & Matsagar, V. A. (2019). Semi-active algorithm for energy-based predictive structural control using tuned mass dampers. *Computer-Aided Civil and Infrastructure Engineering*, *34*(11), 1010–1025.

How to cite this article: Pisarski, D., & Jankowski, Ł. (2022). Reinforcement learning-based control to suppress the transient vibration of semi-active structures subjected to unknown harmonic excitation. *Computer-Aided Civil and Infrastructure Engineering*, 1–17. https://doi.org/10.1111/mice.12920

APPENDIX: LIST OF SYMBOLS

x, p	State and adjoint state vectors
$u_1, \dots, u_m$	Control policies
u_{min}, u_{max}	Min and max admissible control value
$K_1^*, \dots, K_m^*$	Iterated policy parameters
$K_1^0, \dots, K_m^0$	Initial policy parameters
$A, B_1, \dots, B_m$	System matrices
f	Excitation vector
J	Cost functional
E	Structural energy
Q	Energy weighting matrix
V	Lyapunov function
P	Solution to the Lyapunov equation
H	Hamiltonian function
t	Simulation time
T_l	Learning time window
T_f, T_c	Force acting time, total control time
$\tau_1, \dots, \tau_s$	Switching times
M, D, S	Mass, damping, and stiffness matrices
H_1, H_2	Matrices accommodating control forces
F	Excitation force
ω, $\mathcal{A}$	Frequency and amplitude of F
ω^d, $\mathcal{A}^d$	Frequency and amplitude of the disturbance
w	Deflection of the beam
a_u, a_F	Positions of the actuator and excitation force
$\alpha_1, \dots, \alpha_1$	Step sizes for policy parameter updating
z_{max}	Maximal number of iterations
ϵ	Terminal condition parameter

This research was funded in whole or in part by National Science Centre, Poland (grant numbers 2020/39/B/ST8/02615 and 2017/26/D/ST8/00883). For the purpose of Open Access, the authors have applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.