engineering proceedings

Article

Optimization of Graded Arrays of Resonators for Energy Harvesting in Sensors as a Markov Decision Process Solved via Reinforcement Learning †

Luca Rosafalco *, Jacopo Maria De Ponti, Luca Iorio, Raffaele Ardito and Alberto Corigliano

Dipartimento di Ingegneria Civile ed Ambientale, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy
* Correspondence: luca.rosafalco@polimi.it; Tel.: +390223994273
† Presented at the 9th International Electronic Conference on Sensors and Applications, 1–15 November 2022; Available online: https://ecsa-9.sciforum.net/

Abstract: The design optimization of the grading of a resonator array for energy harvesting in sensors is described. Attention is paid to set the resonator heights, possibly removing resonators whenever convenient. Instead of employing time-consuming heuristic approaches that require to verify the physical understanding of the problem and to tune the design ruling parameters, the optimization task has been treated as a Markov decision process, in which states describe specific system configurations, and actions represent the modifications to the current design. The physics-based understanding of the problem is exploited to constrain the set of possible modifications to the mechanical system. Finite elements simulations have been exploited to evaluate the action effects and to inform the reinforcement learning agent. The proximal policy optimization algorithm has been employed to solve the Markov decision problem. The procedure is demonstrated able to automatically produce configurations enhancing the mechanical system performance. The proposed framework is generalizable to a large class of problems involving the design optimization of sensors.

Keywords: energy harvesting for sensors; metamaterials; reinforcement learning; Markov decision process

Citation: Rosafalco, L.; De Ponti, J. M.; Iorio, L.; Ardito, R.; Corigliano, A. Optimization of Graded Arrays of Resonators for Energy Harvesting in Sensors as a Markov Decision Process Solved via Reinforcement Learning . *Eng. Proc.* **2022**, *1*, 0. https://doi.org/

Academic Editor: Stefano Mariani

Published: 1 November 2022

1. Introduction

An elastic waveguide with a graded array of resonant bars was proposed for energy harvesting in [1,2] with possible applications in microsystems. This metamaterial structure features a spatial variation of mechanical properties allowing for manipulating propagating waves. Specifically, the grading enables both to enhance the wavefield amplitude in the resonator endowed with the harvester, typically realised through a piezoelectric material, and to enhance the interaction time between the waves and the resonators. Our aim is to improve energy harvesting capacities by tuning the lengths of the resonator bars. With a similar goal, [3,4] compared different grading laws.

The optimization of a mechanical system can be automatised by relying: on gradient based methods; on genetic algorithms [5]; on particle swarm optimisation [6]. However [7]: the first family of approaches is negatively affected by the non linear dependence between the optimisation object and the design parameters; the second suffers from a high computational cost; the third requires to constrain some parameters of the optimisation algorithm without any clear indications for doing so.

As done in [8], we propose to look at the optimisation task as a Markov Decision Process (MDP), in which states describe specific configurations, and actions represent the modification to the current design. The solution of the MDP has been based on the use of RL, and in particular of the Proximal Policy Optimisation (PPO) algorithm [9]. Finite Element (FE) simulations have been exploited to simulate wave propagation in order to provide information to the RL agent. In [10], experimental data were used with the same goal.

Another aspect of interest is the description adopted for the possible system configurations. Indeed, the physical understanding of the problem has suggested to set the resonator lengths, and possibly to modify the number of resonator, through few interpolation points and B–spline interpolation, similarly to what was done by [11] for structural shape optimisation.

The proposed procedure will be demonstrated able to lead to suboptimal configurations enhancing the mechanical system performance with respect to previously proposed configurations. The interest of the approach stays in the possible applications to a large class of optimisation problems involved by the design of sensors.

The reminder of the paper is arranged as follows. The proposed methodology is detailed in Section 2, while the results relevant to the optimisation of rainbow based metamaterial for energy harvesting are reported in 3. Final considerations are collected in Section 4.

2. Methodology

The metamaterial optimisation is organised in a sequence of T actions A_t, with $t = 1, \dots, T$, taken by an agent, producing a modification of the system state S_t. The performance of the obtained configuration is measured by the reward R_t, here defined as the average value in time of the elastic energy of the bar endowed with the harvester. This quantity is strictly related to the energy obtained by exploiting a piezoelectric material to convert mechanical into electrical energy. States and rewards define the environment in which the agent plays. Given that the probability to get into a state S_t depends only on S_{t-1} and on A_{t-1}, a MDP has been used to formalise the sequential decision process. Considering a certain state S_t, the optimisation problem coincides with the maximisation of the expected return G_t defined as

$$G_t = R_{t+1} + R_{t+2} + \ldots + R_T. \tag{1}$$

The agents actions are guided by a policy π, here treated as a stochastic entity associating a Probability Density Function (PDF) over the set of possible actions to a given state of the system. Stochasticity is required to allow the exploration of the state space. To understand if a policy π is preferable than a second policy π', value functions $v_\pi(s)$ are used, where s is treated as a random variable whose possible realisations at time t are indicated by S_t. Value functions are defined as

$$v_\pi = \mathbb{E}_\pi[G_t | S_t = s], \tag{2}$$

where $\mathbb{E}_\pi$ is the expected value of G_t starting from S_t and using π to guide the following actions. Other two quantities, namely the action–value function $q_\pi(s, a)$ and the advantage function $d_\pi(s, a)$, are similarly defined as

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a], \tag{3a}$$

$$d_\pi(s, a) = q_\pi(s, a) - v_\pi(s). \tag{3b}$$

The notion of $d_\pi(s, a)$ is exploited by PPO, a policy gradient algorithm. This family of RL approaches explicitly looks for the best policy π^* by exploiting a (large) number of agent–environment interactions. Outcome of the procedure is typically a suboptimal policy. However, approximating π^* does not preclude to enhance the system performance with respect to already known configurations.

Before presenting PPO, we discuss the description adopted for the states. The possibility of representing the state through a vector collecting the resonator lengths has been discarded because modifying one by one the resonator length produces reward alterations too limited to inform the RL agent. A more convenient option is to employ a limited number N_s of continuous variables by constraining the state and action spaces through the enforcement of smooth graded patterns of the resonator lengths. This strategy is motivated by the problem insight gained in previous works [1,2]. Specifically, the coordinates of few points have been employed as state variables, while the envelope of the resonator array has been obtained by interpolating these points through cubic B–splines. Figure 1 has been reported to exemplify the adopted state representation. Actions coincide with modifying the coordinates of the light blue starts, as it will be further specified in Section 3.

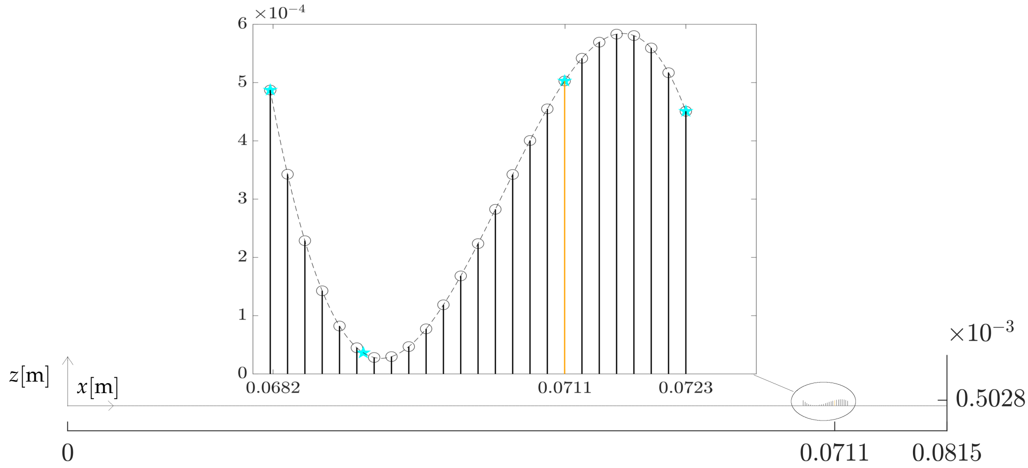

Figure 1. Use of interpolation points (light blue markers) to define the envelope curve (dotted line) setting the resonator lengths. The resonator endowed with the harvester is plotted with an orange line. The circles recall the lumped mass-spring description adopted for the resonators.

Handling continuous state and action spaces forces to approximate $v_\pi(s)$ and $q_\pi(s,a)$ by parametric functions

$$v_\pi(s) \approx v(s, \boldsymbol{\theta}_v), \tag{4a}$$
$$q_\pi(s,a) \approx q(s,a,\boldsymbol{\theta}_q). \tag{4b}$$

whose tunable weights are collected in $\boldsymbol{\theta}_v \in \mathbb{R}^{N_{\theta v}}$ and in $\boldsymbol{\theta}_q \in \mathbb{R}^{N_{\theta q}}$, respectively. A similar treatment has been done for the advantage function $d_\pi(s,a) \approx d(s,a,\boldsymbol{\theta}_v)$.

By associating the PDF characterising a Gaussian distribution to the policy, a tunable parametric function has been exploited to establish a mapping between the state and the statistical moments of the PDF, namely the mean $\mu(s,\boldsymbol{\theta}_p)$ and the standard deviation $\sigma(s,\boldsymbol{\theta}_p)$. The weight tuning both the advantage function and the function having as output the policy moments is done through PPO. In particular, two fully connected Neural Networks (NN) featuring 32 neurons in each layer have been employed for modelling $d(s,a,\boldsymbol{\theta}_v)$ and the function with output $[\mu(s,\boldsymbol{\theta}_p), \sigma(s,\boldsymbol{\theta}_p)]$. Thanks to NN differentiability, $\boldsymbol{\theta}_p$ is updated to maximise the objective function of PPO

$$\mathcal{L}(\boldsymbol{\theta}_p) = \hat{\mathbb{E}}_e\left[\min\left(\frac{\pi(a|s,\boldsymbol{\theta}_p)}{\pi_{\text{old}}(a|s,\boldsymbol{\theta}_{p_{\text{old}}})}\hat{d}_e(s,a,\boldsymbol{\theta}_v), \text{clip}\left(\frac{\pi(a|s,\boldsymbol{\theta}_p)}{\pi_{\text{old}}(a|s,\boldsymbol{\theta}_{p_{\text{old}}})}, 1-\epsilon, 1+\epsilon\right)\hat{d}_e(s,a,\boldsymbol{\theta}_v)\right)\right], \tag{5}$$

via Adam [12], where: $\epsilon = 0.2$; $\hat{\mathbb{E}}_e$ and $\hat{d}_e$ are computed over N_e episodes; an episode is a complete sequence of agent–environment interaction $t = (1, \ldots, T)$.

Specifically, indicating by $y(\boldsymbol{\theta}_p)$ the ratio between $\pi(a|s,\boldsymbol{\theta}_p)$ and $\pi_{\text{old}}(a|s,\boldsymbol{\theta}_{p_{\text{old}}})$, the "min" and "clip" operations allow to define the following probability distribution

$$\begin{cases} y(\theta_p)\hat{d}_e(s,a,\boldsymbol{\theta}_v) & \text{for } \hat{d}_e(s,a,\boldsymbol{\theta}_v) > 0 \text{ and } y(\theta_p) < 1+\epsilon, \\ & \text{or } \hat{d}_e(s,a,\boldsymbol{\theta}_v) < 0 \text{ and } y(\theta_p) > 1-\epsilon, \\ (1+\epsilon)\hat{d}_e(s,a,\boldsymbol{\theta}_v) & \text{for } \hat{d}_e(s,a,\boldsymbol{\theta}_v) > 0 \text{ and } y(\theta_p) > 1+\epsilon, \\ (1-\epsilon)\hat{d}_e(s,a,\boldsymbol{\theta}_v) & \text{for } \hat{d}_e(s,a,\boldsymbol{\theta}_v) < 0 \text{ and } y(\theta_p) < 1-\epsilon, \end{cases} \tag{6}$$

whose expected mean is the objective of PPO. The update of $d(s,a,\boldsymbol{\theta}_v)$ is separately done every N_e episodes according to the actor–critic scheme of the PPO algorithm [13]. Further details on PPO can be found in [9].

3. Results

To compute the reward related to a certain state, wave propagation is simulated through FEs for $T = 1.25 \times 10^{-5}$ s with a time step of 3×10^{-9}. The waveguide has been discretised using 376 Euler Bernoulli beams, while a mass–spring schematisation has been employed for the resonating bars. The lengths of the FE have been set to 0.0344×10^{-3} m in between the resonators and to 0.344×10^{-3} m elsewhere. The mesh refinement has been required to catch the effects of the resonator interactions. Two absorbing layers, one at the beginning of the waveguide and the other at the end, have been placed to avoid reflections as suggested in [14]. The employed material is aluminium with density $\rho = 2710$ g/m^3 and Young's modulus $E = 70$ GPa. Concerning the cross sectional area and moment of inertia, the ones of the waveguide are $B_w = 3 \times 10^{-6}$ m^2 and $I_w = 2.5 \times 10^{-13}$ m^4, while the relevant moment of inertia I_r of the resonating bars is equal to 0.4909×10^{-13} m^4. An initial number of 25 resonators with spacing close to $\lambda_w/11$ have been considered, where $\lambda_w = 1.8 \times 10^{-3}$ m is the length of the flexural wave travelling on the elastic beam without resonators.

The excitation generating the propagating wave is reported in Figure 2. It mimics the one experimentally adopted by [2]. The frequency content of the excitation matches the first bending frequency $\omega_h = 17.67$ MHz of the resonator endowed with the harvester. The four points depicted as light blue markers in Figure 1 have been employed to define the arrangement of the resonating bars. Specifically, the number N_s of continuous variables has been set to 4. They coincide with the z coordinate of the 1st and 4th points, and with the (x,z) coordinates of the 2nd point. The 3rd point, placed at the tip of the bar equipped with the harvester, has been fixed. The order of the agent action has been set too, see Table 1.

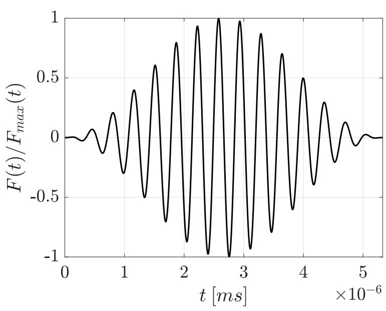

Figure 2. Load applied to the rainbow based metamaterial.

Except for the way in which the state space has been constrained, no other physical knowledge of the system has been exploited. As starting state, the z coordinates of all the points have been set equal to the length of the harvester bar $l_h = 5.028 \times 10^{-4}$ m. The range of variation of the coordinate points is also reported in Table 1. The value $l^{\text{max}} = 9.156 \times 10^{-4}$ m allows to have a 10% attenuation of the forced response of a bar with length l_r^{max} and moment of inertia I_r excited by an oscillating force with frequency equal to ω_h. If bars with length smaller than $l_h/20$ results by the interpolation, they are removed from the system, in this way enabling to modify the number of resonators.

Table 1. Description and ordering of the agent actions.

Action Ordering	What is Modified	Variable Value at the Starting State	Range of Possible Values
1	1st point z coordinate	5.028×10^{-4} m	$[0, 9.156 \times 10^{-4}$ m$]$
2	4th point z coordinate	5.028×10^{-4} m	$[0, 9.156 \times 10^{-4}$ m$]$
3	2th point x coordinate	0.0697 m	$[0.0682, 0.0711]$ m
4	2th point z coordinate	5.028×10^{-4} m	$[0, 9.156 \times 10^{-4}$ m$]$

The outcomes of the optimisation process are evaluated in terms of the reward R_T of the last episode configuration. This value has been divided by the reward R_T^H featuring the waveguide with just one resonator. The interest is to judge the performance improvement with respect to the configuration featuring a linear grading reported in Figure 3b, originally proposed by [1] on the basis of physical considerations.

Two resonator arrangements have been found out by the RL agent. The best discovered configuration depicted in Figure 3c has been generated after roughly 5000 agent–environment interactions, much before that the total number N_I of interactions, here set to $N_I = 100,000$, has been run out. Instead, the converged RL policy configuration shown in Figure 3d has been produced by the quasi–deterministic policy obtained at the end of the agent training. This policy is a suboptimal solution of the MDP. They both outperform the linear grading rule by $\approx 4.7\%$ and by $\approx 1.0\%$, respectively.

The suboptimality of the converged RL policy and the better performance of the other discovered configuration should not appear to undermine the value of the method. Indeed: the obtained configurations are close in terms of R_T/R_T^H; they confirm the physical intuition of the problem; discovering the reported best configuration has been allowed by the first policy updates; a closest approximation of the optimal policy could have been obtained, but only at the cost of a huge increase of the computational time [13]. On the contrary, the small number of agent–environment interactions needed to discover the configuration in Figure 3c promises a successful application of this RL and MDP based optimisation approach to other sensor design problems possibly involving more complex and time demanding simulations, even in the realm of multiphysics.

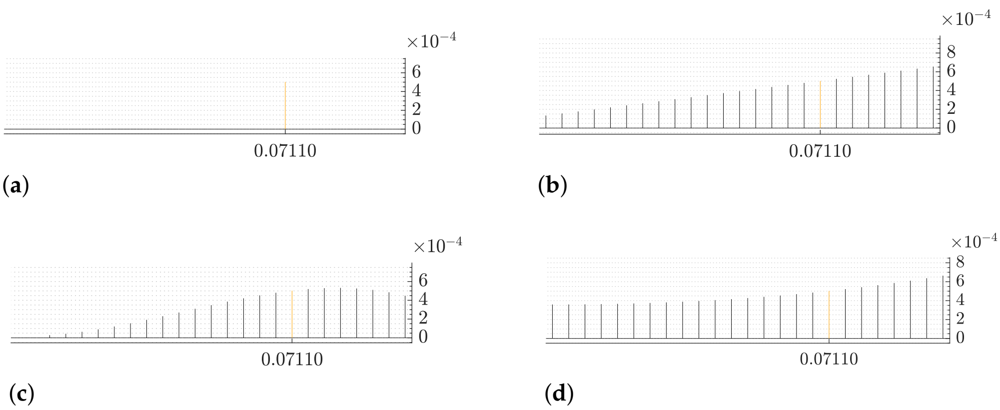

Figure 3. Optimised and reference configurations of the bar arrangement together with the relevant reward R_T/R_T^H. (**a**) Harvester only configuration, $R_T/R_T^H = 1.000$; (**b**) Reference optimised configuration, $R_T/R_T^H = 3.504$; (**c**) Best RL discovered configuration, $R_T/R_T^H = 3.669$; (**d**) Converged RL policy configuration, $R_T/R_T^H = 3.537$.

Moreover, it is worth to remember that these configurations have been obtained without exploiting the physical understanding of the problem, such as the notion that an initial linear ascending grading both enhances the interaction time between the waves and the resonators and increases the wavefield amplitude in the resonator endowed with the harvester. On the contrary, a greater insight into the surface wave propagation in rainbow based structure should be obtained by explaining the reason behind the improved performances of the discovered configurations. For example, the concave curvature reported at the beginning of the grading deserves a deepen comprehension. Moreover, it is shown that the best performance has been obtained when the number of resonators has been reduced to 23. These and other aspects are currently under investigation.

4. Conclusions

In this work, the grading optimisation of a resonator array for energy harvesting with possible applications in sensor design has been performed exploiting an innovative reinforcement learning approach. Using few points and interpolation functions to describe the space of the possible system states, the proximal policy optimisation algorithm has led to two resonator configurations both improving the performance with respect to a reference linear grading rule. The optimisation outcome has confirmed the physical comprehension of the problem already in possession, promising to open the understanding of more subtle mechanical aspects. The procedure is suitable to be generalised to other optimisations of sensor systems.

Author Contributions: Conceptualization, L.R., J.M.D.P., R.A. and A.C.; methodology, formal analysis, investigation, L.R. and J.M.D.P.; software, validation, resources, visualisation, L.R., J.M.D.P. and L.I; writing—original draft preparation, L.R.; writing—review and editing, J.M.D.P., L.I., R.A. and A.C.; supervision, project administration, funding acquisition, R.A. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partially funded by the support of the H2020 FET–proactive project Metamaterial Enabled Vibration Energy Harvesting (MetaVEH) project under Grant Agreement No. 952039.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

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

References

1. De Ponti, J.M.; Colombi, A.; Ardito, R.; Braghin, F.; Corigliano, A.; Craster, R.V. Graded elastic metasurface for enhanced energy harvesting. *New J. Phys.* **2020**, *22*, 013013. https://doi.org/10.1088/1367-2630/ab6062.
2. De Ponti, J.M.; Colombi, A.; Riva, E.; Ardito, R.; Braghin, F.; Corigliano, A.; Craster, R.V. Experimental investigation of amplification, via a mechanical delay–line, in a rainbow–based metamaterial for energy harvesting. *Appl. Phys. Lett.* **2020**, *117*, 143902. https://doi.org/10.1063/5.0023544.
3. Alshaqaq, M.; Erturk, A. Graded multifunctional piezoelectric metastructures for wideband vibration attenuation and energy harvesting. *Smart Mater. Struct.* **2020**, *30*, 1–11.
4. Zhao, B.; Thomsen, H.R.; De Ponti, J.M.; Riva, E.; Van Damme, B.; Bergamini, A.; Chatzi, E.; Colombi, A. A graded metamaterial for broadband and high-capability piezoelectric energy harvesting. *Energy Convers. Manag.* **2022**, *269*, 116056. https://doi.org/10.1016/j.enconman.2022.116056.
5. Jenkins, W. Towards structural optimization via the genetic algorithm. *Comput. Struct.* **1991**, *40*, 1321–1327. Special Issue: Computational Structures Technology. https://doi.org/10.1016/0045-7949(91)90402-8.
6. Perez, R.; Behdinan, K. Particle swarm approach for structural design optimization. *Comput. Struct.* **2007**, *85*, 1579–1588. https://doi.org/10.1016/j.compstruc.2006.10.013.
7. Viquerat, J.; Rabault, J.; Kuhnle, A.; Ghraieb, H.; Larcher, A.; Hachem, E. Direct shape optimization through deep reinforcement learning. *J. Comput. Phys.* **2021**, *428*, 110080. https://doi.org/10.1016/j.jcp.2020.110080.
8. Ororbia, M.E.; Warn, G.P. Design Synthesis Through a Markov Decision Process and Reinforcement Learning Framework. *J. Comput. Inf. Sci. Eng.* **2021**, *22*. 021002, https://doi.org/10.1115/1.4051598.
9. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347. https://doi.org/10.48550/ARXIV.1707.06347.
10. Fan, D.; Yang, L.; Wang, Z.; Triantafyllou, M.S.; Karniadakis, G.E. Reinforcement learning for bluff body active flow control in experiments and simulations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26091–26098. https://doi.org/10.1073/pnas.2004939117.

11. Papadrakakis, M.; Lagaros, N.D.; Tsompanakis, Y. Structural optimization using evolution strategies and neural networks. *Comput. Methods Appl. Mech. Eng.* **1998**, *156*, 309–333. https://doi.org/10.1016/S0045-7825(97)00215-6.
12. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980; pp. 1–13.
13. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
14. Rajagopal, P.; Drozdz, M.; Skelton, E.A.; Lowe, M.J.; Craster, R.V. On the use of absorbing layers to simulate the propagation of elastic waves in unbounded isotropic media using commercially available Finite Element packages. *NDT E Int.* **2012**, *51*, 30–40. https://doi.org/10.1016/j.ndteint.2012.04.001.