Reinforcement learning approach to nonequilibrium quantum thermodynamics

Sgroi, P., Palma, G. M., & Paternostro, M. (2021). Reinforcement learning approach to nonequilibrium quantum thermodynamics. *Physical Review Letters*, *126*(2), [020601]. https://doi.org/10.1103/PhysRevLett.126.020601

Published in:
Physical Review Letters

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2021, American Physical Society.
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:08. giu. 2023

Reinforcement learning approach to non-equilibrium quantum thermodynamics

Pierpaolo Sgroi,[1,2] G. Massimo Palma,[1,3] and Mauro Paternostro[2]
[1]*Dipartimento di Fisica e Chimica - Emilio Segré ,*
Università degli Studi di Palermo, via Archirafi 36, I-90123 Palermo, Italy
[2]*Centre for Theoretical Atomic, Molecular and Optical Physics,*
School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, United Kingdom
[3]*NEST, Istituto Nanoscienze-CNR, Piazza S. Silvestro 12, 56127 Pisa, Italy*
(Dated: December 10, 2020)

We use a reinforcement learning approach to reduce entropy production in a closed quantum system brought out of equilibrium. Our strategy makes use of an external control Hamiltonian and a policy gradient technique. Our approach bears no dependence on the quantitative tool chosen to characterize the degree of thermodynamic irreversibility induced by the dynamical process being considered, require little knowledge of the dynamics itself and does not need the tracking of the quantum state of the system during the evolution, thus embodying an experimentally non-demanding approach to the control of non-equilibrium quantum thermodynamics. We successfully apply our methods to the case of single- and two-particle systems subjected to time-dependent driving potentials.

The design, development, and optimization of quantum thermal cycles and engines is one of the most active and attention-catching research strand in the burgeoning field of quantum thermodynamics [1–4]. Besides being one of the most important applications of thermodynamics, thermal engines play also a fundamental role in the development of the theory of classical thermodynamics itself. It is thus not surprising that the community working in the field that explores the interface between thermodynamics and quantum dynamics is very interested in devising techniques for the exploitation of quantum advantages for the sake of of realizing quantum cycles and machines [3–5]. The overarching goal is to operate at much smaller scales than classical motors and engines and enhance the performance of such devices so as to reach classically unachievable efficiencies [1, 6–8].

However, the quasi-static approximation that allows us to describe thermodynamic transformations with an equilibrium theory does not hold for real finite-time thermal engines, which operate in non-equilibrium conditions. This is even more true for quantum engines: in order to exploit the potential benefits of quantum coherences, such devices should operate within the coherence time of the platforms used for their embodiment, which might be very short [5, 9, 10]. Any finite-time process gives rise to a certain degree of irreversibility, as quantified by entropy production, which enters directly into the thermodynamic efficiency of the process, limiting it [11]. Therefore, the control of non-equilibrium quantum processes is an important goal to achieve to enhance the efficiency of quantum engines [12].

For a closed system, a well-known quantum control approach consists of shortcuts-to-adiabaticity (STA) [13, 14]. This approach has been successfully applied to various platforms [15–21], and the possible application of STA to non-equilibrium thermodynamics has been explored [12, 22–27]. However, it bears considerable disadvantages as it requires extensive knowledge of the system dynamics. It is thus difficult to use STA as on-the-run experimental procedures. Moreover, they do not allow for the choice of the function characterizing the dissipative processes for the system and it is currently very difficult to incorporate in a working STA protocol any constraint on the energetic cost of its implementation [28, 29]. Therefore, alternative approaches are necessary to improve our control power over quantum systems subjected to non-equilibrium processes.

A possible approach to this problem is the use of machine learning techniques currently employed in growing number of problems. In particular, quantum physics is benefiting from machine learning in many ways in light of their capability to approximate high dimensional nonlinear functions that would be difficult to infer otherwise. Numerous applications have been developed, ranging from phase detection [30, 31] to the simulation of stationary states of open quantum many-body systems [32], from the research of novel quantum experiments [33] to quantum protocols design and state preparation [34–36, 41, 42], from the learning of states and operations [37–39] to the modelling and reconstruction of non-Markovian quantum processes [33, 40]. In particular, problems of planning or control can be successfully addressed through reinforcement learning (RL) [44].

In this Letter, we extend the range of quantum problems that can be tackled with machine learning approaches by demonstrating their successful use in the study of non-equilibrium thermodynamics of quantum processes. In particular, we propose an approach to reduce energy dissipation and irreversibility arising from a unitary work protocol using RL. Specifically, we employ a policy gradient technique to tackle out-of-equilibrium work-extraction protocols whose thermodynamic irre-

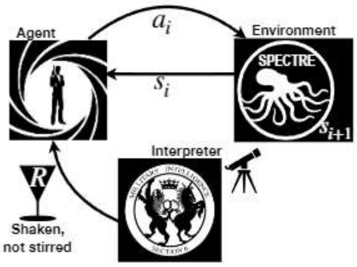

FIG. 1. Principles of RL: at the i^{th} step of the protocol, an agent observes an environment, acquiring its state s_i, upon which he decides to implement action a_i. As a result, the state of the environment is updated to s_{i+1}. Based on the actions of the agent and the states of the environment, an interpreter decides to grant the agent a reward R, which the agent aims to maximize.

versibility we aim at reducing. Our methodology allows us to address this problem with only little knowledge of the system dynamics and to choose how to quantify dissipations. Our study provides a significant contribution to the development of control strategies tailored for physically relevant non-equilibrium quantum processes, thus complementing the scenario drawn so far and based on optimal control and STA.

Background on reinforcement learning.– In the RL setting, an agent dynamically interacts with an environment and learns from such interaction how to behave in order to maximize a given reward functional [44, 45]. The process is typically divided in discrete interaction steps: at each step i, the agent makes an observation of the environment state s_i and – based on the outcomes of their observations – takes an action a_i. Based on this action, the environment state is updated to s_{i+1} and we repeat the procedure for the new step. This is iterated for a given number of steps or until we reach a certain state, when a third party (an interpreter) provides the agent with a reward $R(s_0, a_0, s_1, a_1, ...)$. Based on their past behaviour and the states of the environment, the agent change the way further actions are chosen so as to maximize the future reward (cf. Fig. 1). This procedure is repeated for many epochs until, if possible, the agent learn how to reach the maximum reward.

If the environment is completely observable, at each step the agent action and the reward depend only on the observation at the current step and the process is said to be a Markov decision process (MDP). In this case, we can describe the behaviour of the agent using a policy function $\pi(a_i|s_i)$. This represents the probability for the agent to choose the action a_i, given the state s_i of the environment. In a policy gradient approach, we parametrize the policy function $\pi_\theta(a_i|s_i)$ with a set of parameters θ, and change them accordingly to the reward. This can be done using a gradient ascent algorithm. If the reward is given to the agent at the end of each epoch, as in our case, the gradient ascent reads [46]

$$\Delta\theta = \eta R \sum_{a_i} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_i, s_i), \tag{1}$$

where η is the learning rate and the sum is calculated over the actions taken in any given *trajectory* $\{a_i\}_i$.

For a continuous action space, we assume a certain shape for the policy function and use a function approximator for one or more parameters of the probability distribution [44]. Here we assume the policy function to be a Gaussian and take

$$\pi_{\boldsymbol{\theta}}(a_i = a|s_i = s) = \exp\left[-\frac{(a - \mu_{\boldsymbol{\theta}}(s))^2}{2\sigma^2}\right] / (\sigma\sqrt{2\pi}), \tag{2}$$

where we treat σ as an external parameter and use a neural network for the parametrization of μ. Based on our choice for $\pi_{\boldsymbol{\theta}}(a_i|s_i)$, the condition in Eq. (1) is satisfied if the neural network is trained with a stochastic gradient descent method over the batch using the cost function $C = \frac{1}{2\sigma^2}\sum_{a_i} R|a_i - \mu_{\boldsymbol{\theta}}(s_i)|^2$.

Physical system and methodology.– Let us consider a closed quantum system evolving under a time dependent Hamiltonian $H_S(t)$ within the time interval $[0, \tau]$. We want to control the system evolution using an additional Hamiltonian $H_{\text{opt}}(t)$ such that $H_{\text{opt}}(0) = H_{\text{opt}}(\tau) = 0$.

For simplicity, we consider $H_{\text{opt}}(t) = f_{\text{opt}}(t)M_{\text{opt}}$ where the operator M_{opt} is kept fixed and we control the function $f_{\text{opt}}(t)$ (enforcing the boundary conditions $f_{\text{opt}}(0) = f_{\text{opt}}(\tau) = 0$ so as to fulfil the requests made on the Hamiltonian) to optimize the process. The total Hamiltonian of the system during its evolution is thus

$$H(t) = H_S(t) + f_{\text{opt}}(t)M_{\text{opt}}. \tag{3}$$

We divide the system evolution in a certain number of discrete time steps. At each step t_i, the agent makes an observation s_i and chooses an action a_i. This is done by extracting a random number according to Eq. (2), based on the prediction of the neural network for $\mu_{\boldsymbol{\theta}}(s_i)$. We then take $f_{\text{opt}}(t) = a_i$ in the interval $[t_i, t_{i+1}[$. We limit the maximum and the minimum output of the network $|\mu_{\boldsymbol{\theta}}(s_i)| < \mu^*$ so that we can control the maximum amount of energy spent for the optimization. This is important when dealing with thermal engines, as we want to spend les energy for the control than what we extract from the process. This is done in parallel for a batch of systems and, at the end of the evolution, the neural network is trained on this batch and the corrisponding rewards. The procedure is repeated for many epochs, each time resetting the system and the Hamiltonian to the original state and value. The process is run again and the value of $f_{\text{opt}}(t)$ maximizing the reward over the batch is chosen.

We now comment on the quantifier of irreversibility addressed in our study and the different approaches that we will consider to reduce the system dissipations. The first approach aims to reduce the mean entropy production of the system [1, 47–51], calculated as the relative entropy between the final state of the system $\rho(\tau)$ and the corresponding instantaneous thermal equilibrium state $\rho^{eq}(t) = e^{-\beta H_S(t)}/Z_S(t)$ with $Z_S(t) = \mathrm{tr}[e^{-\beta H_S(t)}]$ the partition function of the system. We thus consider

$$\Sigma = S(\rho(\tau)\|\rho^{eq}(\tau)) \tag{4}$$

where $S(\sigma\|\chi) = \mathrm{tr}[\sigma(\log\sigma - \log\chi)]$ is the quantum relative entropy [52]. For this purpose, we use a Dense-layers Neural Network [53] taking as inputs the time step and the density matrix. In this case, the agent reward is $R = -\Sigma$, which suits our goal well: the agent is rewarded by reducing the degree of irreversibility of the process.

In the second approach, we assume to measure the energy of the system before the evolution [54, 56]. We consider the case of non-degenerate energy levels and use as reward the square-root of the fidelity between the final state of the system and the corresponding adiabatic final state, thus having $R = |\langle\phi(t)|\phi_{ad}(t)\rangle|$. This approach too benefits of the use of a Dense-layers Neural Network with inputs embodied by the time step and the (pure) quantum state of the system.

The third approach uses the same ideas laid out above. However, this time we want the model to be useful as a control technique even when we are not able to simulate or track the dynamics of the system. We thus use a different input, while still considering MDP. We use a Long-Short-Term-Memory (LSTM) Neural Network [46] taking as inputs the energy measured at the beginning of the evolution, and the time steps.

If the observation of our agent at a given time step contains all the informations about the initial state of the system and the control term of the Hamiltonian at any previous time, the knowledge of the current quantum state is no longer required in order to have a MDP. However, we can avoid to use such a large imput at each step if we use a LSTM network instead. The output of a LSTM network does not only depend on the input at a given time step, but also on all the previous imputs and outputs. Such neural networks can retain long-term dependencies and are widely used for tasks that involve sequential data, such as speech recognition.

For these reasons, we just need to take measurements at the beginning and at the end of the evolution. Needless to say, this embodies a significant reduction on the practical complexity of the control protocol, as the scheme only requires two measurements, and thus leaves room for a non-demanding experimental implementations that does not need to track the evolution of the system.

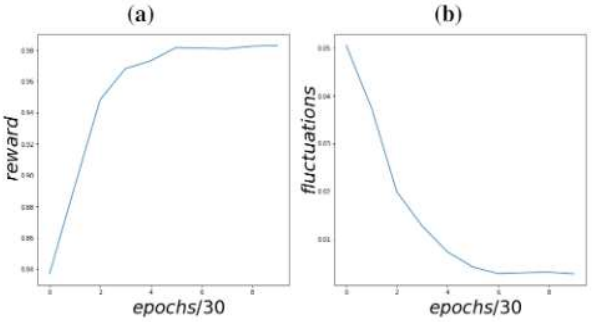

FIG. 2. Panel (**a**): average reward over the batch and 30 epochs as a function of the number of epochs of training. Panel (**b**): fluctuations around the average reward. The asymptotic behavior of both curves demonstrates success of the training.

For inaccessible initial states of the system, or should one want to avoid performing a measurement at the start of the dynamics, if we assume the initial density matrix of the system to be always the same, we can still use a LSTM network in a way similar to the first approach, with just the time steps as inputs and a reward $R = -\Sigma$. The advantage of our approach with respect to other techniques lies on the number of runs needed to achieve good performances, which is significantly smaller than what is needed from standard numerical optimization (see Ref. [55] for details).

Case studies.– We now apply our methods to simple yet physically relevant models. We address the cases of a single spin-like system exposed to a time-dependent field and a two-spin model with a time-dependent coupling.

Single spin-$1/2$ *particle in a time-dependent field*: Let us consider a qubit evolving in the interval $t \in [0, \tau]$ under an Hamiltonian (we take units such that $\hbar = 1$)

$$H_S(t) = [\sigma_x B_x(t) + \sigma_z B_z(t)]/2 \tag{5}$$

with $B_x^2(t) + B_z^2(t) = B_0^2 \ \forall t \in [0, \tau]$, thus modelling a spin subjected to a rotating magnetic field. We assume that the system is initialized in a thermal state at inverse temperature β. This is relevant only when we do not take an energy measurement at the beginning of the evolution. Our optimization Hamiltonian is $H_{\mathrm{opt}}(t) = -f_{opt}(t)\sigma_y$ so that $H(t) = H_S(t) - f_{opt}(t)\sigma_y$.

We start with the first approach, introduced in the previous section, that aims to reduce the relative entropy. We introduce the entropy production reduction $\Delta\Sigma = 1 - \Sigma_{opt}/\Sigma_{free}$, where Σ_{opt} is given by Eq. (4) and Σ_{free} is associated to the case without optimization term in the Hamiltonian. Likewise, the reduction of the work done on the systems is $\Delta W = 1 - (\Delta U_{opt} + E_{in})/\Delta U_{free}$, where E_{in} is an estimation of the energy spent for the optimization, defined as [22]

$$E_{in} = \left| \int_0^1 \mathrm{tr}(\rho(t) f_{opt}(t)\sigma_y) dt \right|. \tag{6}$$

and ΔU is the change of the internal energy $U(t) = \mathrm{tr}(\rho(t)H(t))$ of the system between initial and final state.

As our control process starts only after a measurement, the second approach to the quantification of irreversibility gives additional information about the system. Our $f_{opt}(t)$ then depends on the initial state. Based on our knowledge of the initial pure state of the system, we want the final state as close as possible to the adiabatic one (that is, the corresponding eigenvector of $H_S(\tau)$). Therefore, our performance measure for this approach will be the fidelity of the final state with the adiabatic target $|\langle\phi(\tau)|\phi_{ad}(\tau)\rangle|^2$. For the third approach, we solve the previous problem, this time with a LSTM Neural Network, as discussed in **Physical system and methodology**.

We divided the dynamics of our system in 10 steps and set $\mu^* = 3$. We considered $B_x(t) = B_0 \sin(\frac{\pi t}{2\tau})$ in Eq. (5). For each of the RL methods considered here, we ran 20 simulations of a training consisting in 300 epochs with a batch of 30 systems. In Fig. 2 we show a typical example of a learning curve with successful training. Using the first approach with an initial thermal state with $\beta = 1$, we successfully reduced both the relative entropy $\Delta\Sigma = (99 \pm 1)\%$ and the work done on the system $\Delta W = (91 \pm 9)\%$. Examples of $f_{opt}(t)$ are given in Fig. 3. When the second approach to irreversibility was used, we obtained the fidelity with the adiabatic target $F_{ad}(\tau) = |\langle\phi(\tau)|\phi_{ad}(\tau)\rangle|^2 = 0.997 \pm 0.002$. In Fig. 4 we show an example of $f_{opt}(t)$ for this case. Finally, for the third approach we obtained $F_{ad}(\tau) = 0.998 \pm 0.001$.

We have rounded our analysis by running a single simulation for a different choice of time-dependent field, namely $B_x(t) = B_0 \sin[\frac{\pi}{2}\sin^2(\frac{\pi t}{2\tau})]$, obtaining a value of the adiabatic fidelity as large as $F_{ad}(\tau) \approx 0.998$. The corresponding functions $f_{opt}(t)$ are shown in Fig. 5.

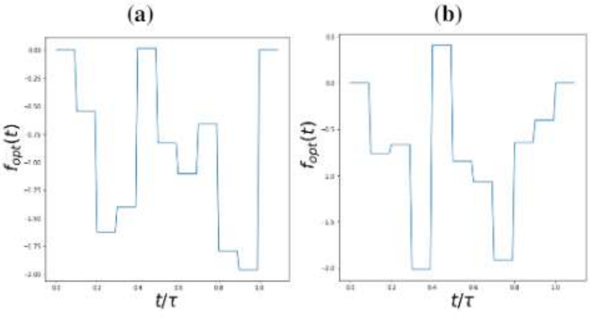

FIG. 3. We show the form taken by $f_{opt}(t)$ for two different runs of the optimization process. Although they both reduce the entropy production of approximately the same amount (99.86% in panel **(a)** and 99.80% in panel **(b)** respectively), the trends followed by the control function are visibly different.

$J(t)$	Σ_{free}/β	Σ_{opt}/β	ΔU_{free}	ΔU_{opt}	E_{in}
$J_1(t)$	0.600644	0.367022	0	−0.233621	0.005392
$J_2(t)$	0.575289	0.366835	−0.025354	−0.233808	0.004581

TABLE I. Results for a simulation of the free and optimized evolution for the different choices of $J(t)$ in the text. The optimized quantities are very close for both the different interaction term functions. In both cases, we achieved an error of 10^{-6}.

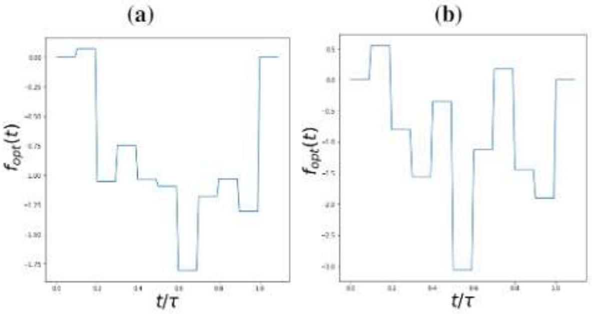

FIG. 4. Panel **(a)** [**(b)**]: example of $f_{opt}(t)$ for an initial $|\uparrow\rangle$ [$|\downarrow\rangle$] state of $H_S(0)$. Here, $\sigma_z|\uparrow\rangle = |\uparrow\rangle$ while $\sigma_z|\downarrow\rangle = -|\downarrow\rangle$. The corresponding fidelity with the targets is 0.997.

Time-dependent coupling of spin-$1/2$ *particles*: We now consider a slightly more complicated system composed of two two-level systems with Hamiltonian

$$H_S(t) = \sigma_z^1 + \frac{1}{2}\sigma_z^2 + J(t)(\sigma_x^1\sigma_x^2 + \sigma_y^1\sigma_y^2), \quad (7)$$

where the coupling strength $J(t)$ evolves in the time interval $t \in [0, \tau]$. We start with both spins in a termal state with an inverse temperature β. Our control term is

$$H_{opt}(t) = f_{opt}(t)(\sigma_x^1\sigma_y^2 - \sigma_y^2\sigma_x^1)/2. \quad (8)$$

We aim at minimizing Eq. (4), this time using a LSTM Neural Network, as described in **Physical system and methodology**. As the variation in the free energy between the initial and final state [47, 49, 50] $\Delta F = \Delta U - \Sigma/\beta$ for both the free and the optimized process must be the same, we set the error in our energy measurements to be the difference in this quantity for the two processes.

We ran a simulation where the dynamics of our system is divided in 10 steps and took $\mu^* = 3$. We used a batch

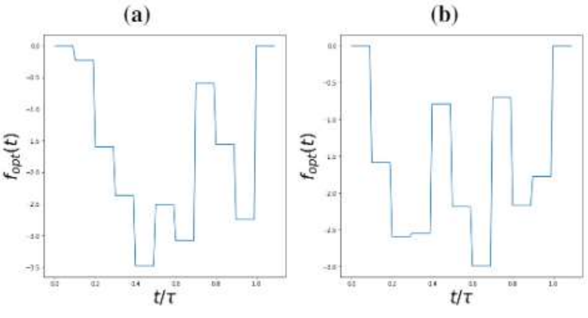

FIG. 5. Panel **(a)** [Panel **(b)**] example of $f_{opt}(t)$ for an initial $|\uparrow\rangle$ [$|\downarrow\rangle$] state of $H_S(0)$. Here $B_x(t) = B_0 \sin[(\pi/2)\sin(\pi t/2\tau)^2]$.

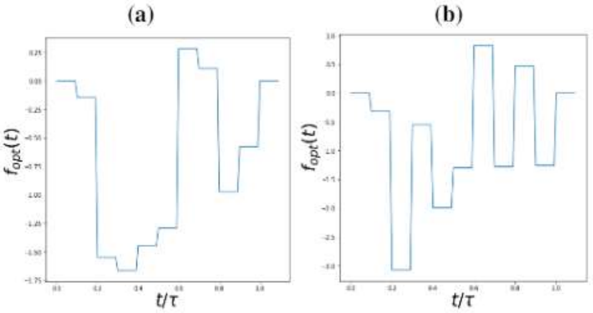

FIG. 6. We plot $f_{opt}(t)$ for the time-dependent coupling rates $J_1(t)$[panel (**a**)] and $J_2(t)$ [panel (**b**)].

of 30 systems and considered 100 epochs, choosing the time-dependent coupling rate $J_1(t) = \chi(t/\tau - 0.5)$ with $\chi(t - t_0) = 1$ at $t = t_0$ and being null otherwise. We have also considered $J_2(t) = \sin[\pi/2 - (\pi/2)\cos(\frac{\pi t}{2\tau})]$, both for an initial thermal state with $\beta = 1$. Our results are shown in Fig. 6 and Table I. A successful reduction of entropy production is achieved in both cases. Moreover, the entropy production Σ_{opt} for both optimized processes takes very similar values. Similar considerations hold for ΔU_{opt}. This is encouraging, although not surprising, as for both processes we have $J(0) = 0$ and $J(\tau) = 1$ so that the corresponding adiabatic process is the same, and we have in fact the same ΔF.

Next, using $J_1(t)$, we changed the temperature of the initial state of the system in the range $\beta \in [0.1, 2.1]$, dividing this interval in 20 steps. Running a single simulation for each value of β, we obtained a mean entropy production reduction $\Delta\Sigma \approx 36\%$ in this interval.

Conclusions.–We have proposed and benchmarked a technique a deep RL-based approach to reduce the degree of irreversibility resulting from a non-equilibrium thermodynamic transformation of a closed quantum system. Our method can be used with an arbitrary choice of the function characterizing the dissipative process undergone by the system and requires little knowledge of the system dynamics. Moreover, it can be applied without tracking the state of the system during the evolution, thus potentially easing any experimental implementations.

We applied our technique to two simple yet relevant problems: we successfully reduced the entropy production and the distance of the final state from the adiabatic target for a spin-1/2 particle subjected to a time-dependent magnetic field and the entropy production resulting from the time-dependent coupling between two spin-1/2 particles. While we focused on simple models, it would be interesting to apply the proposed approach to many-body quantum systems. This could help significantly in the development of an efficient mesoscopic thermal engine operating under realistic conditions. In particular, our third approach could be advantageous when tackling high-dimensional systems. However, such systems will still embody a challenge, as they could require a large number of control terms, and the optimization would still suffer from the increase of dimensionality of the agent actions space. This could lead to a decrease in the performances or at least to an increase in the number of experiments required for a successful optimization, which are issues that we will address in future. A natural further development of our work would be the extension to open quantum systems dynamics.

PS thanks the Centre for Theoretical Atomic, Molecular, and Optical Physics, School of Mathematics and Physics, Queen's University Belfast for hospitality during the initial phases of this work. We acknowledge support from the H2020-FETOPEN-2018-2020 TEQ (grant nr. 766900), the DfE-SFI Investigator Programme (grant 15/IA/2864), COST Action CA15220, the Royal Society Wolfson Research Fellowship (RSWF\R3\183013), the Leverhulme Trust Research Project Grant (grant nr. RGP-2018-266), the UK EPSRC (grant nr. EP/T028106/1), and the PRIN project 2017SRN-BRK QUSHIP funded by MIUR.

---

[1] S. Vinjanampathy and J. Anders, Contemp. Phys. **57**, 545 (2016).

[2] S. Deffner and S. Campbell, *Quantum Thermodynamics* (Morgan and Claypool Publishers, 2019).

[3] M. T. Mitchison, Contemp. Phys. **60**, 164 (2019).

[4] R. Kosloff and A. Levy, Annual Review of Physical Chemistry **65**, 365 (2014).

[5] G. Barontini and M. Paternostro, New Journal of Physics **21**, 063019 (2019).

[6] W. Niedenzu, V. Mukherjee, A. Ghosh, A. G. Kofman, and G. Kurizki, Nat. Comm. **9** (2018).

[7] B. K. Agarwalla, J.-H. Jiang, and D. Segal, Phys. Rev. B **96**, 104304 (2017).

[8] J. Klaers, S. Faelt, A. Imamoglu, and E. Togan, Phys. Rev. X **7**, 031044 (2017).

[9] D. von Lindenfels, O. Gräb, C. T. Schmiegelow, V. Kaushal, J. Schulz, M. T. Mitchison, J. Goold, F. Schmidt-Kaler, and U. G. Poschinger, Phys. Rev. Lett. **123**, 080602 (2019).

[10] J. P. S. Peterson, T. B. Batalhão, M. Herrera, A. M. Souza, R. S. Sarthour, I. S. Oliveira, and R. M. Serra, Phys. Rev. Lett. **123**, 240601 (2019).

[11] G. T. Landi and M. Paternostro, arXiv:2009.07668 (2020).

[12] A. del Campo, J. Goold, and M. Paternostro, Sci. Rep. **4**, 6208 (2014).

[13] E. Torrontegui, S. Ibáñez, S. Martínez-Garaot, M. Modugno, A. del Campo, D. Guéry-Odelin, A. Ruschhaupt, X. Chen, and J. G. Muga, *Adv. At. Mol. Opt. Phys.*, edited by E. Arimondo, P. R. Berman, and C. C. Lin, Vol. 62 (Academic Press, 2013) p. 117.

[14] M. V. Berry, J. Phys. A: Math. Theor. **42**, 365303 (2009).

[15] D. Guéry-Odelin, A. Ruschhaupt, A. Kiely, E. Torrontegui, S. Martínez-Garaot, and J. G. Muga, Rev. Mod. Phys. **91**, 045001 (2019).

[16] M. Palmero, S. Wang, D. Guéry-Odelin, J.-S. Li, and J. G. Muga, New Journal of Physics **18**, 043014 (2016).

[17] H. L. Mortensen, J. J. W. H. Sørensen, K. Mølmer, and J. F. Sherson, New J. Phys. **20**, 025009 (2018).

[18] S. Deng, P. Diao, Q. Yu, A. del Campo, and H. Wu, Phys. Rev. A **97**, 013628 (2018).

[19] F.-H. Ren, Z.-M. Wang, and Y.-J. Gu, Physics Letters A **381**, 70 (2017).

[20] H. Saberi, T. Opatrny, K. Mølmer, and A. del Campo, Physical Review A **90**, 060301(R) (2014).

[21] S. Campbell, G. De Chiara, M. Paternostro, G. M. Palma, and R. Fazio, Phys. Rev. Lett. **114**, 177206 (2015).

[22] O. Abah and E. Lutz, EPL (Europhysics Letters) **118**, 40005 (2017).

[23] S. Deng, A. Chenu, P. Diao, F. Li, S. Yu, I. Coulamy, A. del Campo, and H. Wu, Sci. Adv. **4** (2018).

[24] O. Abah and M. Paternostro, Phys. Rev. E **99**, 022110 (2019).

[25] B. Çakmak and Ö. E. Müstecaplıoğlu, Phys. Rev. E **99**, 032108 (2019).

[26] O. Abah and E. Lutz, Phys. Rev. E **98**, 032121 (2018).

[27] O. Abah, M. Paternostro, and E. Lutz, arXiv:1911.00373 (2019).

[28] Y. Zheng, S. Campbell, G. De Chiara, and D. Poletti, Phys. Rev. A **94**, 042132 (2016).

[29] A. C. Santos and M. S. Sarandy, Sci. Rep. **5**, 15775 (2015).

[30] P. Broecker, F. Assaad, and S. Trebst, arXiv:1707.00663 (2017).

[31] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 1745 (2017).

[32] N. Yoshioka and R. Hamazaki, Phys. Rev. B **99**, 214306 (2019).

[33] A. A. Melnikov, H. Poulsen Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, Proc. Natl. Acad. Sci. **115**, 1221 (2018).

[34] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Commun. Phys. **2**, 2399 (2019).

[35] I. Paparelle, L. Moro, and E. Prati, Phys. Lett. A **384**, 126266 (2020).

[36] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Phys. Rev. X **8**, 031086 (2018).

[37] L. Innocenti, L. Banchi, A. Ferraro, S. Bose, and M. Paternostro, New J. Phys. **22**, 065001 (2020).

[38] C. Harney, S. Pirandola, A. Ferraro, and M. Paternostro, New J. Phys. **22**, 045001 (2018).

[39] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Phys. Rev. X **8**, 031084 (2018).

[40] L. Banchi, E. Grant, A. Rocchetto, and S. Severini, New J. Phys. **20**, 123030 (2018).

[41] T. Giordani, E. Polino, S. Emiliani, A. Suprano, L. Innocenti, H. Majury, L. Marrucci, M. Paternostro, A. Ferraro, N. Spagnolo, and F. Sciarrino, Phys. Rev. Lett. **122**, 020503 (2019).

[42] T. Giordani, A. Suprano, E. Polino, F. Acanfora, L. Innocenti, A. Ferraro, M. Paternostro, N. Spagnolo, and F. Sciarrino, Phys. Rev. Lett. **124**, 160401 (2020).

[43] J. V. der Plas, *Python Data Science Handbook* (O'Reilly Media, 2016).

[44] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press Cambridge, Massachusetts, London, England, 2015).

[45] V. Dunjko and H. J. Briegel, Rep. Prog. Phys. **81**, 074001 (2018).

[46] F. Marquardt, "Machine learning for physicists," `"https://machine-learning-for-physicists.org"`, [Online; accessed April-2019].

[47] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[48] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[49] S. Deffner and E. Lutz, Phys. Rev. Lett. **107**, 140404 (2011).

[50] S. Deffner and E. Lutz, Phys. Rev. Lett. **105**, 170402 (2010).

[51] S. Deffner and S. Campbell, *Quantum Thermodynamics*, 2053-2571 (Morgan and Claypool Publishers, 2019).

[52] V. Vedral, Rev. Mod. Phys. **74**, 197 (2002).

[53] M. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).

[54] P. Talkner, E. Lutz, and P. Hänggi, Phys. Rev. E **75**, 050102(R) (2007).

[55] Supplemental Material reporting technical details and a comparison with other optimization methods is available from XXX.

[56] L. Mazzola, G. De Chiara, and M. Paternostro, Phys. Rev. Lett. **110**, 230602 (2013).

[57] F. Chollet *et al.*, "Keras," `https://keras.io` (2015).

[58] D. Kingma and J. Ba, International Conference on Learning Representations (2014).

[59] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, npj Quant. Inf. **5**, 85 (2019).

[60] P. Virtanen, *et al.*, Nature Methods **17**, 261 (2020).