



Online identification and control of PDEs via reinforcement learning methods

Alessandro Alla¹ · Agnese Pacifico² · Michele Palladino³ · Andrea Pesare⁴

Received: 1 November 2023 / Accepted: 19 June 2024 / Published online: 1 August 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

We focus on the control of unknown partial differential equations (PDEs). The system dynamics is unknown, but we assume we are able to observe its evolution for a given control input, as typical in a reinforcement learning framework. We propose an algorithm based on the idea to control and identify on the fly the unknown system configuration. In this work, the control is based on the state-dependent Riccati approach, whereas the identification of the model on Bayesian linear regression. At each iteration, based on the observed data, we obtain an estimate of the *a-priori* unknown parameter configuration of the PDE and then we compute the control of the correspondent model. We show by numerical evidence the convergence of the method for infinite horizon control problems.

Keywords Reinforcement learning · System identification · Stabilization of PDEs · State-dependent Riccati equations · Bayesian linear regression · Numerical approximation

Mathematics Subject Classification (2010) 65Mxx · 93B30 · 49Mxx

1 Introduction

Reinforcement Learning (RL) is one of the main Machine Learning paradigms, together with supervised and unsupervised Learning. In RL, an agent interacts with

Communicated by: Stefan Volkwein

Andrea Pesare is an Independent Researcher by the time this manuscript is processed for publication.

A. Alla and A. Pacifico are members of the INdAM-GNCS activity group. A. Alla is part of INdAM - GNCS Project “Metodi numerici innovativi per equazioni di Hamilton-Jacobi” (CUP_E53C23001670001). The work of A.A. was carried out within the “Data-driven discovery and control of multi-scale interacting artificial agent systems,” and received funding from the European Union Next-GenerationEU - National Recovery and Resilience Plan (NRRP) - MISSION 4 COMPONENT 2, INVESTMENT 1.1 Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) - Project Code P2022JC95T, CUP H53D23008920001. The work of M. Palladino is partially funded by the University of L’Aquila Starting Project Grant “Optimal Control and Applications,” and by INdAM-GNAMPA project, n. CUP_E53C22001930001.

Extended author information available on the last page of the article

an unknown environment, aiming at an action-selection strategy to optimize a system's performance. Generally speaking, one can consider two main RL philosophies. The first one, called model-based, usually concerns the reconstruction of a model from the data trying to mimic the unknown environment. That model is then used to plan and compute a suboptimal policy. The second RL philosophy, called model-free, employs a direct approximation of the value function and/or a policy based on a dynamic programming like algorithm without using a model to simulate the unknown environment. Model-free methods include the famous Monte Carlo methods [39], Temporal-Difference Learning [36, 38] and Q-Learning [41] and more recent ones [18, 24, 26, 37]. An overview of the two RL approaches can be found in [39].

Since the philosophy of this work is to connect model-based RL problems and optimal control, we will first recall some classical approaches to optimal control problems. Specifically, we are interested in feedback control to obtain a state-dependent optimal control which is a valuable property as it makes the control system stable with respect to random disturbances. Dynamic Programming (DP, [5, 15]) considers a family of optimal control problems with different initial conditions and states and looks at the relationship between these problems. The main ingredient is the value function, defined as the minimum of the cost functional which is the solution of the Bellman equation: a nonlinear partial differential equation (PDE) of the Hamilton-Jacobi type. Once the value function has been obtained, it provides optimal feedback control. Although the DP approach is preferable from a theoretical point of view as it provides sufficient conditions and synthesis of optimal feedback control, it has always been challenging to apply it to real problems since it is very expensive from a computational point of view. It suffers from the so-called "curse of dimensionality" (an expression coined by Bellman himself [6]), which means that the computational cost necessary to solve the Bellman equation grows exponentially with respect to the system dimension. This led to the development of suboptimal solution methods to the Bellman equation, known as approximate dynamic programming [10], that could mitigate the curse of dimensionality. Later, we started calling these methods Reinforcement Learning [9, 40]. Optimal control and RL are strongly connected, as they deal with similar problems; in fact, both can be regarded as sequential decision problems, in which one has to make decisions in sequence, trying to optimize not only the immediate rewards but also the future, delayed ones. Recently, in [29, 30], it has been proposed a unified framework for all the families of sequential decision problems including OC and RL. More precisely, RL deals with control problems in which the system's dynamics is uncertain.

In this paper, we want to control an unknown nonlinear dynamics following a RL strategy. We will adopt an online strategy as explained below. We suppose that the system is described by a parametric PDE, whose parameters are unknown. We also assume to have a library which includes all possible terms of the PDE, so that a function of the library enters into the model if the corresponding parameter is not zero. Those parameters are the ones we need to discover to achieve our goal. The chosen library, in this work, will contain several models which are very well studied in the mathematical physics community. Furthermore, although the system dynamics is unknown, we assume it is always possible to observe the true evolution of the system for a given control input. The possibility to observe the unknown system is a

typical assumption in Reinforcement Learning where an agent takes action based on his observation. This will allow us to update the parameter estimate.

To achieve our goal, we propose the following workflow: “control–observe–estimate.” To set the method into perspectives, we begin with an initial parameter estimate that allows to compute the control for such configuration. Note that the control is computed for a problem that uses a parameter estimate and might be far from the true optimal control. Then, by applying that control, we observe the true system configuration by its trajectories. Thus, to update the parameter estimate, we set a linear system based on the observed trajectories which will be solved via Bayesian Linear Regression methods (see, e.g., [32, 33]). We iterate this procedure till the end of the chosen time horizon. We will also discuss a heuristic stopping criterion for the parameter estimation. As mentioned, this is an online approach since we update the parameter estimate every iteration. A first approach driven by the same workflow proposed here has been introduced in [28] for linear low-dimensional problems and quadratic cost functionals. Here, we extend to generic nonlinear control problems with a keen focus on the control of PDEs. The dimension of the discretized problem increases also the challenges of the problem. Our approach to the control of the PDE is based on the discretization by finite differences that reduces the problem to a large system of ordinary differential equations. In the paper, we also show numerically how the computed control stabilizes the PDE for smaller spatial discretization leading to the control of the continuous PDE.

Let us now comment on how, we solve the control problems. As already mentioned at the beginning of this section, control in feedback form is usually obtained by the solution of dynamic programming equations [5] or by Nonlinear Model Predictive Control (NMPC, [17]). An alternative, which combines elements from both dynamic programming and NMPC, is the state-dependent Riccati equation (SDRE) approach (see, e.g., [4, 14]). The SDRE method originates from the dynamic programming associated to infinite horizon optimal stabilization. It circumvents its solution by reformulating the feedback synthesis as the sequential solution of state-dependent Algebraic Riccati Equations (ARE), which are updated online along a trajectory. The SDRE feedback is implemented similarly as in NMPC, but the online solution of an optimization problem is replaced by a nonlinear matrix equation. Later, in [8], theoretical conditions for the stabilization of the problem have been studied. In [1], it has been shown an efficient method employing SDRE for large scale problems.

For the sake of completeness, we also recall that system identification of nonlinear dynamics is a very active and modern research area with a vast literature. Although our identification is strictly linked to a control problem, we briefly recall some literature. Clearly, Physics Informed Neural Networks (PINNs) deserve to be mentioned due to their innovative, accurate and efficient way to discover partial differential equations using neural networks and information from system in the definition of the loss function. We refer to, e.g., [20, 31] for a complete description of the method. It is worth to mention also methods based on variants of sparse optimization techniques such as Sparse Identification of Nonlinear Dynamics (SINDy) for ODEs [12] and for PDEs [34, 35]. SINDy was also applied for the identification of controlled problems (see, e.g., [19]). The authors used an external source as input to identify the system and then

apply NMPC to control the identified model. There, the authors used the workflow: identify first, control later which is different from the strategy presented in the current work. Similar ideas to our work were also presented later in [27]. Other strategies dedicated on control and system identification can be found in, e.g., [22] for PDEs and in, e.g., [25] for ODEs. Recently, a study on the control of an unknown problem with MPC has been introduced in [13]. There, the system was identified using the Extended Dynamic Mode Decomposition, i.e., a surrogate linear model in contrast to our work where we directly identify the nonlinear model.

The outline of the paper is the following. In Section 2, we recall the basics of Bayesian Linear Regression as a building block when adapting our parameter estimate. In Section 3, we briefly explain the state-dependent Riccati equation. In Section 4, we provide all the details of the method proposed in this paper. Later, numerical experiments to support our algorithm are presented in Section 5. Finally, conclusions are driven in Section 6.

2 Bayesian linear regression

Bayesian Linear Regression (BLR, [32, 33]), is a probabilistic method for solving the classical linear regression (LR, [16]) problem. In LR, we consider *data* in the form of input–output pairs

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1, \dots, d}$$

and we suppose that the output variable $y_i \in \mathbb{R}$ can be expressed approximately as a linear function of the input variable $x_i \in \mathbb{R}^n$, i.e.,

$$y_i \approx x_i^T \theta, \quad \text{for } i = 1, \dots, d. \quad (1)$$

We look for a parameter $\theta \in \mathbb{R}^n$ such that (1) is satisfied. The (*ordinary*) *least squares* (LS) approach chooses θ by minimizing the sum of squared residuals

$$E(\theta) = \sum_{i=1}^d |y_i - x_i^T \theta|^2. \quad (2)$$

The *LS solution* can be computed analytically and is given by

$$\theta_{LS} = (X^T X)^{-1} X^T Y, \quad (3)$$

where we collected all the observed inputs in a matrix $X \in \mathbb{R}^{d \times n}$ and all the observed outputs in a vector $Y \in \mathbb{R}^d$:

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_d^T \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}. \quad (4)$$

In BLR, instead, we assume that the deviation of the data from the linear model can be described by a Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$:

$$y_i = x_i^T \theta + \varepsilon_i, \tag{5}$$

where $\theta \in \mathbb{R}^n$ is an unknown parameter to be determined and $\sigma > 0$. We will assume that the value of σ is known, though more general formulations apply Bayesian inference on σ as well. Equation (5) corresponds to fix a conditional distribution of the random variable y given the value of x and θ ,

$$p(y|x, \theta) \sim \mathcal{N}(x^T \theta, \sigma^2). \tag{6}$$

This is what in Bayesian inference is called the *likelihood function*. If we assume that the d observations are independent, the global likelihood function can be written as

$$p(Y|X, \theta) = \prod_{i=1}^d p(y_i|x_i, \theta) \sim \mathcal{N}(X\theta, \sigma^2 I_d), \tag{7}$$

where X, Y have been defined in (4), and I_d denotes the d -dimensional identity matrix.

The available information on the parameter θ is included in the model through the definition of a *prior distribution*, which we assume to be Gaussian with initial mean $m_0 \in \mathbb{R}^n$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{n \times n}$:

$$\theta \sim \mathcal{N}(m_0, \Sigma_0). \tag{8}$$

Bayesian formulas allow to compute the *posterior distribution* of the parameter θ , which is again a Gaussian distribution [11, 33]

$$p(\theta|X, Y) = \frac{p(\theta)p(Y|X, \theta)}{\int_{\mathbb{R}^n} p(\theta')p(Y|X, \theta')d\theta'} \sim \mathcal{N}(m, \Sigma), \tag{9}$$

where

$$\Sigma^{-1} = \frac{1}{\sigma^2} X^T X + \Sigma_0^{-1} \text{ and } m = \Sigma \left(\frac{1}{\sigma^2} X^T Y + \Sigma_0^{-1} m_0 \right). \tag{10}$$

From the posterior distribution one can extract a point estimate of the parameter θ , that is the posterior mean

$$\begin{aligned} \bar{\theta}_{BLR} &= \Sigma \left(\frac{1}{\sigma^2} X^T Y + \Sigma_0^{-1} m_0 \right) \\ &= \left(\frac{1}{\sigma^2} X^T X + \Sigma_0^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} X^T Y + \Sigma_0^{-1} m_0 \right). \end{aligned} \tag{11}$$

However, the advantage of BLR is that it provides a quantification of the uncertainty of this estimate. Finally, we remark that the estimate $\bar{\theta}_{BLR}$ converges to the LS solution (3), when the noise variance σ goes to 0.

3 Control of nonlinear problem via state-dependent riccati equation

In this section, we recall one possible approach to control nonlinear differential equations. We consider the following infinite horizon optimal control problem:

$$\min_{u(\cdot) \in \mathcal{U}} J(u(\cdot)) := \int_0^{\infty} \left(\|x(t)\|_Q^2 + \|u(t)\|_R^2 \right) dt \quad (12)$$

subject to the nonlinear dynamical constraint

$$\begin{aligned} \dot{x}(t) &= A(x(t))x(t) + B(x(t))u(t), \quad t \in (0, \infty), \\ x(0) &= x_0, \end{aligned} \quad (13)$$

where $x(t) : [0, \infty] \rightarrow \mathbb{R}^d$ denotes the state of the system, $A(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, the control signal $u(\cdot)$ belongs to $\mathcal{U} := L^\infty(\mathbb{R}^+; \mathbb{R}^m)$ and $B(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$. The running cost is given by $\|x\|_Q^2 := x^\top Qx$ with $Q \in \mathbb{R}^{d \times d}$, $Q \succ 0$, and $\|u\|_R^2 = u^\top Ru$ with $R \in \mathbb{R}^{m \times m}$, $R \succ 0$. This formulation corresponds to the asymptotic stabilization of nonlinear dynamics toward the origin.

We can synthesize a suboptimal feedback law by following, e.g., an approach known as the state-dependent Riccati equation (SDRE). We refer to, e.g., [4, 14] for more details on the topic.

The SDRE approach is based on the idea that infinite horizon optimal feedback control for systems of the form (13) is linked to a state-dependent algebraic Riccati equation (ARE):

$$A^\top(x)\Pi(x) + \Pi(x)A(x) - \Pi(x)B(x)R^{-1}B^\top(x)\Pi(x) + Q = 0. \quad (14)$$

We note that equation (14) is an ARE that changes every iteration, in fact it depends on the current state x . This makes the difference with respect to the standard LQR¹ problem where the ARE is constant and it is solved just once. SDRE might be thought as an MPC algorithm (see, e.g., [17]) where the inner optimization problem is solved by (14).

When equation (14) admits a solution, it leads to a state-dependent Riccati operator $\Pi(x)$, from where we obtain a nonlinear feedback law given by

$$u(x) := -R^{-1}B^\top(x)\Pi(x)x. \quad (15)$$

We will refer to the feedback gain matrix as $K(x) := R^{-1}B^\top(x)\Pi(x)$. It is important to observe that the ARE (14) admits an analytical solution in a limited number of cases and the obtained control is only suboptimal. More general approaches following, e.g., the dynamic programming approach [7] might be used. This goes beyond the scope of this work; however, one can easily replace, throughout the paper, the SDRE approach with a different (feedback) control method.

¹ LQR deals with the constant matrices $A(x(t)) = A$, $B(x(t)) = B$ in (13).

In [4], it is shown that the SDRE method provides asymptotic stability if $A(\cdot)$ is C^1 for $\|x\| \leq \delta$ and some $\delta > 0$, $B(\cdot)$ is continuous and the pair $(A(x), B(x))$ is stabilizable for every x in a non-empty neighborhood of the origin. Thus, the closed-loop dynamics generated by the feedback law (15) are locally asymptotically stable. The SDRE algorithm proposed in [4] is summarized below.

Algorithm 1 SDRE method

Require: $\{t_0, t_1, \dots\}$, model (13), R and Q ,
 1: **for** $i = 0, 1, \dots$ **do**
 2: Compute $\Pi(x(t_i))$ from (14)
 3: Set $K(x(t_i)) := R^{-1}B^T(x(t_i))\Pi(x(t_i))$
 4: Set $u(t) := -K(x(t_i))x(t)$, for $t \in [t_i, t_{i+1}]$
 5: Integrate the system dynamics with $u(t) := -K(x(t_i))x(t)$ to obtain $x(t_{i+1})$
 6: **end for**

Assuming the stabilization hypothesis above, the main bottleneck in the implementation of Algorithm 1 is the high rate of calls to an ARE solver for (14). We refer to [1] for efficient methods related to large scale problems. In this work, we will deal with small scale problems and thus solve the AREs with the Matlab function `icare`.

4 Identification and control of unknown nonlinear dynamics

The system we want to identify and control is taken from (13) and reads

$$\begin{aligned} \dot{x}(t) &= \sum_{j=1}^n \mu_j A_j(x(t))x(t) + B(x(t))u(t), \quad t \in (0, \infty), \\ x(0) &= x_0, \end{aligned} \tag{16}$$

with the matrix function $A(x)$ in (13) given by $A(x) = \sum_{j=1}^n \mu_j A_j(x)$ and $A_j(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ for $j = 1, \dots, n$. The functions $A_j(x)$ may be thought as a library with terms that have to be selected by the coefficients μ_j 's. Note that this sum is not unique and that we can include extra basis functions by simply setting the corresponding μ_j 's to be zero. Throughout this work, we will assume that the terms $A_j(x)$'s and $B(x)$ are known. Thus, the system (16) is fully identified by the knowledge of the coefficient $\mu = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$ which is considered unknown in the present work.

We assume that there exists a true system configuration $\mu^* \in \mathbb{R}^n$ which is not known but observable through the dynamics (16) setting $\mu = \mu^*$. In other words, we assume that the dynamics generated by this true model configuration μ^* is always observable as a black box, i.e., if we choose a control we can compute the solution of (16) with the true parameter without knowing μ^* explicitly. This is a typical assumption in the Reinforcement Learning setting, where an agent can take actions and observe how the environment responds to them.

The cost functional we want to minimize is adapted from (12) and reads

$$\min_{u(\cdot) \in \mathcal{U}} J(u(\cdot), \mu^*) := \int_0^\infty \left(\|x(t; \mu^*)\|_Q^2 + \|u(t)\|_R^2 \right) dt, \tag{17}$$

where the dependence on μ^* stresses that trajectories x are observed from the true system configuration for a given input.

This addresses the problem of system identification together with the control of (16). Indeed, we consider two unknowns: (i) the parameter configuration μ which is required to converge to μ^* and (ii) the control $u(t)$. The computation of the control will be done using Algorithm 1 which clearly depends on the parameter configuration. For an estimated parameter $\tilde{\mu} \in \mathbb{R}^n$ such that $\tilde{\mu} \neq \mu^*$, the control will be denoted by $u(t; \tilde{\mu})$ to stress the dependence on the particular parameter configuration $\tilde{\mu}$ in (16). Instead, the observed trajectory will be denoted by $x(t; u(t; \tilde{\mu}), \mu^*)$. This notation considers a trajectory computed with the control $u(t; \tilde{\mu})$ plugged into the true system configuration. Furthermore, if we want to represent the solution at the discrete time t_i , we will identify $x^i(u(t_i; \tilde{\mu}), \mu^*) = x(t_i; u(t_i; \tilde{\mu}), \mu^*)$.

Remark 1 (Notation) Let us briefly summarize the notations valid throughout the whole paper for the parameter configuration: μ is a generic parameter, μ^* is the true system configuration, and $\tilde{\mu}$ is an estimated system configuration.

4.1 The method

Let us now explain how we identify the system. We remind that our goal is to steer to the origin of a partially unknown nonlinear system. We also aim at discovering the system on the fly through real-time observation of the trajectories. The workflow of our proposed method goes under the paradigm “control first and identify later” as follows:

1. Pick a parameter configuration,
2. Compute the corresponding control,
3. Observe the trajectories,
4. Update the parameter configuration based on the observations,
5. Go to the second item.

We use a Bayesian Linear Regression algorithm to estimate the system configuration μ^* from the observed trajectory data, as described in Section 2. We will now provide all the details of the proposed method which is summarized at the end of this subsection by Algorithm 2.

Initial configuration To begin with, we provide an initial estimate² $\tilde{\mu}^0 \in \mathbb{R}^n$ for the true system configuration μ^* . To give an example, in the numerical tests, we will set $(\tilde{\mu}^0)_k = 1$ for $k = 1, \dots, n$, but if prior information about μ^* is available, it can be used to choose a proper $\tilde{\mu}^0$. Note that $\tilde{\mu}^0$ will act as m_0 in the prior distribution (8) of the BLR algorithm. We also need to choose an initial covariance matrix $\Sigma_0 \in \mathbb{R}^{n \times n}$. We observed heuristically that $\Sigma_0 = cI_n$, where I_n is the $n \times n$ identity matrix and $c > 0$ is large enough, works well in general.

Computation of the control At time t_i , we compute an approximate solution for the Algebraic Riccati Equation (14) corresponding to the current parameter estimate

² The notation $\tilde{\mu}^0$ refers to the parameter estimate at time t_0 . This will become clearer later in this subsection.

$\tilde{\mu}^i$. Then, we can set the feedback gain matrix $K(x(t_i); \tilde{\mu}^i)$ and the feedback control $u(t; \tilde{\mu}^i)$.

Observation of the trajectories At each iteration, we apply a constant control $u(t; \tilde{\mu}^i)$ with $t \in [t_i, t_{i+1}]$ and *observe* the trajectory at time t_{i+1} , which will be either the actual trajectory, if we are dealing with a real physical system, or a simulated one, if we are simulating the physical system with some numerical methods. Thus, for a given configuration estimate $\tilde{\mu}^i$ and its control input $u(t_i; \tilde{\mu}^i)$ computed following Algorithm 1, we will observe the trajectory $x^{i+1}(u(t_i; \tilde{\mu}^i); \mu^*)$.

The observation of the true trajectory has to be thought as a black box that provides the solution, or approximate solution, of the original controlled problem for a given control. The black box takes the control and the initial state as input and provides the trajectories as output. The observation of only one trajectory is due to the fact that we aim at discovering and controlling on the fly while updating the parameter estimate at each time instance. In RL, such methods are referred to as *online*, in contrast with *offline* methods where an agent can use multiple offline observations of the system to build a control.

Update of the parameter estimate We now provide the crucial part of the method, that is how we update the parameter estimate using a Bayesian Linear Regression. To apply BLR and obtain a problem in the form (1), we have to discretize the system (16). We provide, without loss of generality, an example through an implicit Euler scheme to discretize (16). Thus, the discretization of (16), using, e.g., an implicit Euler method and the correspondent feedback gain matrix $K^i := K(x(t_i))$ (see Section 3), reads

$$\frac{x^{i+1} - x^i}{\Delta t} \approx \sum_{j=1}^n \tilde{\mu}_j^i A_j(x^{i+1})x^{i+1} - B^i K^i x^{i+1}, \quad i = 0, 1, \dots \tag{18}$$

where we have dropped the dependence on the control for x and we recall that x^i is the short notation for $x^i(u(t_i, \tilde{\mu}^{i-1}), \mu^*)$, whereas $B^i = B(x(t_i))$.³ In our numerical simulations, we deal with the control explicitly, that is why we use K^i in (18) and we consider the control constant in each interval $[t_i, t_{i+1}]$. We employ an implicit scheme due to numerical stability and our application to PDEs later in Section 5.

Once we plug the true, observed trajectory in equation (18), i.e., x^i and x^{i+1} , we obtain a linear system of equations that the true system configuration μ_j^* solves, at least up to a certain approximation error. We use it to update our estimate $\tilde{\mu}^i$ of the system configuration. Starting from equation (18), we can write

$$\frac{x^{i+1} - x^i}{\Delta t} + B^i K^i x^{i+1} \approx \sum_{j=1}^n \tilde{\mu}_j^i A_j(x^{i+1})x^{i+1}.$$

Thus, we obtain d equations for the n coefficients $\tilde{\mu}_j^i$ as in (1), which we can write in a more compact form

$$Y^i \approx X^i \tilde{\mu}^i, \tag{19}$$

³ We remark that in all our simulations the matrix B is constant.

where $\tilde{\mu}^i \in \mathbb{R}^n$, $X^i := [A_1(x^{i+1})x^{i+1}, \dots, A_n(x^{i+1})x^{i+1}] \in \mathbb{R}^{d \times n}$ and $Y^i := \frac{x^{i+1} - x^i}{\Delta t} + B^i K^i x^{i+1} \in \mathbb{R}^d$. The notation $\tilde{\mu}^i$ stresses the fact that, at each time iteration, we look for a parameter configuration that may differ on time. This problem fits into the structure presented in Section 2, and the solution for $\tilde{\mu}^i$ is given by (11).

Algorithm Our proposed idea is finally summarized in Algorithm 2 below.

Algorithm 2 Online identification and control

Require: $\{t_0, t_1, \dots\}$, model $\{A_j(x)\}_{j=1}^n, B, R, Q, \tilde{\mu}_0, \Sigma_0$

- 1: **for** $i = 0, 1, \dots$ **do**
 - 2: Solve (14) and obtain $\Pi(x(t_i); \tilde{\mu}^i)$
 - 3: Set $K(x(t_i); \tilde{\mu}^i) := R^{-1} B^T(x(t_i)) \Pi(x(t_i); \tilde{\mu}^i)$
 - 4: Set $u(t_i; \tilde{\mu}^i) := -K(x(t_i); \tilde{\mu}^i)x(t_i)$
 - 5: Apply the control $u(t_i; \tilde{\mu}^i)$ and observe the trajectories $x^{i+1}(u(t_i; \tilde{\mu}^i), \mu^*)$
 - 6: Compute $\tilde{\mu}^{i+1}$ as in (11) from (19)
 - 7: **end for**
-

Remark 2 There might be cases in Algorithm 2 where the ARE does not provide a solution and this will depend on the parameter estimate. In those cases, we fix the feedback gain equal to the zero vector and we go to the next step. This is equivalent to addressing the uncontrolled problem within that specific time window. However, in our simulations, we did not observe this behavior after the first iteration (where we decided to set $u \equiv 0$).

Remark 3 Theoretical convergence of the parameter $\tilde{\mu}^i$ to the true parameter μ^* for $i \rightarrow +\infty$ is not guaranteed and goes beyond the scope of this paper. We decided to keep this study for follow-up work. However, in the numerical tests in Section 5, we observed numerical convergence of the method. The identification of the system configuration can be stopped if, for a certain $\bar{i} > 0$, we obtain $\|\tilde{\mu}^{\bar{i}} - \tilde{\mu}^{\bar{i}-1}\|_\infty < tol_\mu$ with $tol_\mu > 0$ being the desired threshold. Note this criteria is only heuristic and that Algorithm 2, as it is, does not need the parameter convergence. The primary goal is to stabilize an unknown control system at 0.

4.2 Application to PDEs

Our ultimate goal is the application of Algorithm 2 to identify and control nonlinear PDEs given by

$$\left\{ \begin{array}{l} y_t(t, \xi) = \sum_{j=1}^n \mu_j F_j(y(t, \xi), y_\xi(t, \xi), y_{\xi\xi}(t, \xi), y_{\xi\xi\xi}(t, \xi), \dots) + B^T(\xi)u(t), \\ y(0, \xi) = y_0(\xi), \\ y(t, a) = 0, \quad y(t, b) = 0, \end{array} \right. \begin{array}{l} t \in [0, +\infty), \xi \in (a, b), \\ \xi \in [a, b], \\ t \in [0, +\infty). \end{array} \tag{20}$$

where $y : [0, \infty] \times \mathbb{R} \rightarrow \mathbb{R}$, $\mu_j \in \mathbb{R}$, $u(t) : [0, \infty) \rightarrow \mathbb{R}^m$ and $B(\xi) : [a, b] \rightarrow \mathbb{R}^m$. Without loss of generality, we set zero Dirichlet boundary conditions. We assume that

the model is given by the sum of simple monomial basis functions F_j of y and its derivatives. Similarly to (16), the functions F_j 's may be thought as a library with terms that has to be selected by the coefficients μ_j 's. Note that this sum is not unique and that we can include extra basis functions by simply setting the corresponding μ_j 's to be zero.

The numerical discretization of (20), by, e.g., finite differences method [23], provides a system in the form (16), where each component of $x \in \mathbb{R}^d$ corresponds to the grid points, say $x_i(t) \approx y(t, \xi_i)$ for $i = 1, \dots, d$ and $A_j(x(t))x(t) \approx F_j$, where $F_j \in \mathbb{R}^d$ denotes the basis function evaluated at all the grid points such that $(F_j)_i = F_j(y(t, \xi_i), y_{\xi}(t, \xi_i), y_{\xi\xi}(t, \xi_i), y_{\xi\xi\xi}(t, \xi_i), \dots)$. In Section 5, we will explain in detail how to obtain each term $A_j(x)$. We note that the matrices $A_j(x)$ take into account the boundary conditions.

The continuous cost functional we want to minimize is

$$J(u; \mu^*) = \int_0^\infty \left(\|y(t, \cdot; \mu^*)\|_{L^2(a,b)}^2 + \|u(t)\|_R^2 \right) dt \tag{21}$$

with R defined after equation (13) and we stress the dependence of the trajectory y on the true system configuration. The discretization of (21) corresponds to the choice $Q = \Delta\xi I_d$ in (12) with $\Delta\xi > 0$ being the spatial step size and I_d the $d \times d$ identity matrix.

5 Numerical experiments

In this section, we will show our numerical examples to validate the proposed method. To set the section into perspective, we provide the continuous PDE model studied which reads:

$$\begin{cases} y_t(t, \xi) = \mu_1 y_{\xi\xi}(t, \xi) + \mu_2 y_{\xi}(t, \xi) + \mu_3 y(t, \xi) \\ \quad + \mu_4 y^2(t, \xi) + \mu_5 y^3(t, \xi) + \mu_6 y(t, \xi) y_{\xi}(t, \xi) \\ \quad + \mu_7 y_{\xi\xi\xi}(t, \xi) + B^T u(t) & t \in [0, t_{end}], \xi \in (a, b), \\ y(0, \xi) = y_0(\xi) & \xi \in [a, b], \\ y(t, a) = 0 = y(t, b) & t \in [0, t_{end}]. \end{cases} \tag{22}$$

where for numerical reasons, we have to choose a finite horizon with $t_{end} > 0$ large enough to simulate the infinite horizon problem and such that for $t > t_{end}$ the controlled solution will not change significantly. We remark that, based on the choice of the parameters, the model (22) includes the control of, e.g., heat equation, advection equation, diffusion–reaction–convection equation, burgers equation, viscous burgers, etc. Many of these models have different physical interpretations between them. This is the reason behind our choice of (22).

In this model, we fix $n = 7$ libraries and in order to fit into the desired canonical form (16) we use the finite difference (FD) method (see, e.g., [23]) where the discrete state $x(t)$ corresponds to the approximation of $y(t, \xi)$ at the grid points.

The term $A(x)$ will be given by

$$A(x) = \mu_1 \Delta_d + \mu_2 T + \mu_3 I_d + \mu_4 \text{diag}(x) + \mu_5 \text{diag}(x \circ x) + \mu_6 \tilde{D}(x) + \mu_7 M$$

where the symbol \circ denotes the Hadamard or component-wise product and

- $\Delta_d \in \mathbb{R}^{d \times d}$ is the FD approximation of the Dirichlet Laplacian with $\Delta_d := \Delta \xi^{-2} \text{tridiag}([1, -2, 1], d)$,⁴
- $T \in \mathbb{R}^{d \times d}$ is the FD upwind or downwind approximation of the advection term such that
 - if $\mu_2 > 0$, $T = T_{neg} = \Delta \xi^{-1} \text{tridiag}([-1, 1, 0], d)$,
 - if $\mu_2 < 0$, $T = T_{pos} = \Delta \xi^{-1} \text{tridiag}([0, -1, 1], d)$,
- $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix,
- $\text{diag}(x) \in \mathbb{R}^{d \times d}$ is a diagonal matrix with the components of the vector x
- $\tilde{D}(x) \in \mathbb{R}^{d \times d}$ is a matrix such that its i -th row $\tilde{D}(x)_i$ is
 - if $(\mu_6 x)_i > 0$, $\tilde{D}(x)_i = (\text{diag}(x) T_{neg})_i$
 - if $(\mu_6 x)_i < 0$, $\tilde{D}(x)_i = (\text{diag}(x) T_{pos})_i$

where $(\mu_6 x)_i$ indicates the i -th element of the vector $\mu_6 x$ and $(\tilde{D}(x))_i$ indicates the i -th row of the matrix between parentheses

- $M \in \mathbb{R}^{d \times d}$ is a FD approximation of the third order derivative: $M = -\frac{1}{2\Delta \xi^3} \text{pentadiag}([1, -2, 0, 2, -1], d)$ ⁵.

Controlled trajectories are integrated in time using an implicit Euler method (see (18)), which is accelerated using a Jacobian-Free Newton Krylov method (see, e.g., [21]) using 10^{-5} as the threshold for the stopping criterion of the Newton method and less than 500 iterations. As mentioned in Remark 3, in our numerical simulations, we have observed convergence of our estimated configuration to the true one. Therefore, we have added to Algorithm 2 the stopping criterion with $tol_\mu = 10^{-5}$.

We present three numerical test cases with nonlinear PDEs. Those are the PDEs we can observe. The first test is a nonlinear diffusion–reaction equation, known as the Allen-Cahn equation ($\mu^* = [1, 0, 11, 0, -11, 0, 0]$). The second test studies the viscous Burgers’ equation ($\mu^* = [0.01, 0, 0, 0, 0, 1, 0]$), and the third one the so-called Korteweg-de Vries (KdV) model ($\mu^* = [0.5, 0, 0, 0, 0, 6, -1]$). The goal of all our tests is the stabilization of the (unknown) dynamics to the origin by means of the minimization of the cost functional (21), that can be approximated $\|y(t, \cdot; \mu)\|_{L^2(a,b)}^2 \simeq \sum_{i=1}^d \Delta \xi y(t, \xi_i; \mu)^2 = x^T(t) Q x(t) = \|x(t)\|_Q^2$, where $\xi_i = a + \Delta \xi i$ and $Q = \Delta \xi I_d$ and with $R = 0.01$, thus obtaining (17).

In all our tests, we will plot (i) the *uncontrolled solution* of the true dynamical system, i.e., the solution of (22) obtained choosing $u(t) \equiv 0$ and $\mu = \mu^*$, (ii) the

⁴ The notation $\text{tridiag}([a, b, c], d)$ stands for a tridiagonal $d \times d$ matrix having the constant values $b \in \mathbb{R}$ on the main diagonal, $a \in \mathbb{R}$ on the lower diagonal and $c \in \mathbb{R}$ on the upper diagonal.

⁵ The notation $\text{pentadiag}([a, b, c, e, f], d)$ stands for a pentadiagonal $d \times d$ matrix having the constant values $c \in \mathbb{R}$ on the main diagonal, $b \in \mathbb{R}$ on the lower and $e \in \mathbb{R}$ on the upper diagonal and $a \in \mathbb{R}$ on the second diagonal below and $f \in \mathbb{R}$ on the second diagonal above the main diagonal.

controlled solution based on the SDRE method where μ^* is known, and (iii) our *RL solution* identified by Algorithm 2 where μ^* has to be discovered. We will then compare the optimal control computed by Algorithm 1 and our method Algorithm 2 and the evaluation of the cost functionals. Furthermore, the history of the estimated coefficients $\tilde{\mu}^i$ over time will be presented. Finally, we will also discuss a numerical convergence toward the control of the continuous PDE problem. To do that, we will compute the control for a given spatial discretization $\Delta\xi$ and show that the obtained control is robust enough to stabilize the same problem with decreasing values of the spatial discretization. This will show the numerical mesh independence and the robustness of the proposed method.

The RL assumption relies on the observability of the dynamics with the true system configuration μ^* . This should be thought of as a black box where the true model can be computed (or approximated). In this work, since we do not know the exact solution, we will use two different numerical approaches to obtain the observed trajectories: (i) we use the same scheme, e.g., backward Euler method, used in (18) but with the true parameter μ^* and (ii) an explicit Runge Kutta scheme for stiff problems.

For the sake of completeness, we provide some more numerical details on the two schemes. Again, we stress that those details are not critical to the algorithm, but they are only needed for the numerical simulations. Indeed, one could use any method or even a “real” black box.⁶ These two methods will have different way to approach the feedback control. Indeed, the first approach is “implicit” in the control term Kx , so will be called “implicit approach” or “implicit algorithm” in the following; in this case, we will have the feedback control in the form $K^i x^{i+1}$, mainly for stability reasons. The second scheme is explicit and the feedback control will be $K^i x^i$. In the paper, the latter approach has been implemented using the Matlab function `ode15s`. We remark that the second approach could be used in a real application, replacing the result of the `ode15s` function with an observation of the system evolution, and will be called the “black box algorithm.” Note that in this case, we have⁷, e.g., $Y^i = \frac{x^{i+1}-x^i}{\Delta t} + BK^i x^i$ in (19).

We remark that in both cases, say the use of the implicit algorithm or the use of the “black box”, we added noise to the data used for regression. After computing the control $u(t_i)$ and the trajectory $x(t_{i+1})$, we obtained the matrix $X = [A_1(x(t_{i+1}))x(t_{i+1}), \dots, A_n(x(t_{i+1}))x(t_{i+1})] \in \mathbb{R}^{d \times n}$. To each column of X , we added a vector of independent Gaussian random variables, each with mean 0 and standard deviation given by 0.01 times the mean of the absolute values of the components in the column itself as follows:

$$X_{i,j} \leftarrow X_{i,j} + \mathcal{N}\left(0, \left(\frac{0.01}{d} \sum_{k=1}^d |X_{k,j}|\right)^2\right).$$

This will be referred to as 1% relative noise in the following and has been used in every numerical test, in order to simulate noise on data from real applications. The

⁶ By the term “real” black box, we intend something that takes an input and provides the trajectories without knowing how they are computed.

⁷ Note that B is constant in this section.

noise can be also interpreted as a variation on the observed system that adds negligible terms not in the library.

Finally, we note that in all the tests the *prior distribution* on the parameter μ was initialized as described in Section 4.1, i.e., we started with a normal distribution with mean $\tilde{\mu}^0 = [1, 1, 1, 1, 1, 1]^T$ and covariance matrix $\Sigma_0 = cI_7$ with $c = 1000$ for test 2 and 3 and $c = 200000$ for test 1. In general, $c > 0$ must be chosen large enough to guarantee flexibility to the model. Indeed, the smaller it is, the closer the final approximation of μ will tend to be to the chosen initial *prior* distribution.

The tests presented in this paper were performed on a DELL Latitude 7200, Intel(R) Core(TM) i5-8265U CPU 1.60GHz, using MATLAB.

5.1 Test 1: Allen-Cahn

Our first test is inspired by the example in [3] where it is shown that an MPC approach, for a short prediction horizon, does not stabilize the following equation:

$$\begin{cases} y_t(t, \xi) = y_{\xi\xi}(t, \xi) + 11(y(t, \xi) - y^3(t, \xi)) + u(t), & t \in (0, 0.5], \xi \in (0, 1) \\ y(0, \xi) = 0.2 \sin(\pi\xi) & \xi \in (0, 1), \\ y(t, 0) = 0, y(t, 1) = 0, & t \in [0, 0.5]. \end{cases}$$

This model is known as the Allen-Cahn equation, or Chaffee-Infante equation. Note that for this example a small horizon with $t_{end} = 0.5$ is enough to simulate the infinite horizon problem since the control problem will be stabilized before as it is shown in Fig. 1. Here, we use the same settings of [3], i.e., $\Delta\xi = 0.01 = \Delta t$ and we obtain its discrete version as described in (16) where the B vector is given by a vector of ones. The dimension of the discrete problem is $d = 101$. The only difference with respect to [3] is that we introduce the control as a time-dependent function instead of dealing with a control as a function of time and space. The parameter used in the observable trajectories are $\mu_1^* = 1, \mu_3^* = 11, \mu_5^* = -11$ in (22) subject to the cost functional recalled in Section 5. In the left panel of Fig. 1, we show the solution to the uncontrolled problem whereas in the middle panel, the trajectory is computed using Algorithm 1. Both simulations have been computed knowing the true system configuration. It is clear that the solution is stabilized. We remark that the SDRE method is able to stabilize the problem with an infinite prediction horizon for the linearized problem, whereas the method in [3] uses a finite prediction horizon, that is required to be of size at least $11\Delta t$, for the nonlinear equation. Clearly, our inner

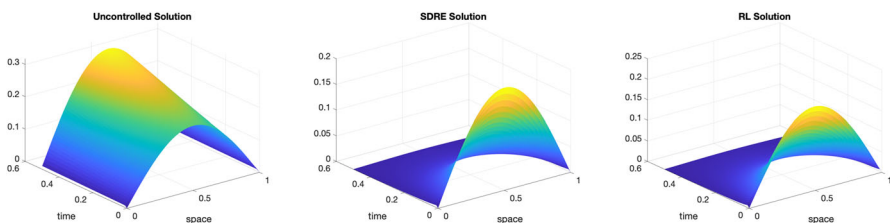


Fig. 1 Test 1: Allen-Cahn, $\Delta\xi = 0.01, \Delta t = 0.01, 1\%$ relative noise

minimization problem is different from the approach proposed in [3] but nevertheless, it is interesting to see its stabilization through SDRE.

Finally, in the right panel of Fig. 1, we show the solution of Algorithm 2. It is clear that with our method we can also stabilize the problem and identify the correct model as shown in Table 1.

In Table 1, we show the results of Algorithm 2 concerning the parameter configuration $\tilde{\mu}$ estimated. We can see that the reconstructed values (second row of the table) are very close to the desired configuration considering the discretization $\Delta\xi = 0.01 = \Delta t$ and the noise added at each iteration.

In Fig. 2, we compare the control of the SDRE algorithm and the RL-based one. One can see that at the beginning the RL control starts from 0 because we decided not to act at the first iteration since we do not have information at that stage. Then, we can see that, slowly, the RL control tends to the SDRE one which is our reference control. In the middle plot of Fig. 2, we show the evaluation of the cost functional. One can see that the RL algorithm is very close to the SDRE method and, as expected, the SDRE cost functional provides lower values. This is clear since our method starts without any knowledge of the model which is learnt on the fly. Finally, for completeness, in the right panel of Fig. 2, we show the convergence history of the parameter configurations. In this example, until the end of the chosen time interval, the algorithm never stops updating the distribution. It would stop at $t = 0.58$ (after 58 iterations) if a longer time interval was considered.

5.2 Test 2: viscous burgers

The equation we study in this test is the viscous Burgers problem which reads:

$$\left\{ \begin{array}{ll} y_t(t, \xi) = 0.01 y_{\xi\xi}(t, \xi) + y(t, \xi) y_{\xi}(t, \xi) \\ \quad \quad \quad + B(\xi)^T u(t), & t \in [0, 2], \xi \in (-1.5, 1.5), \\ y(0, \xi) = \sin(\pi\xi) \chi_{[0,1]}(\xi) & \xi \in (-1.5, 1.5), \\ y(t, -1.5) = 0 = y(t, 1.5), & t \in [0, 2]. \end{array} \right. \tag{23}$$

with $B(\xi)^T = (\chi_{[0.25,0.5]}(\xi), \chi_{[0.75,1]}(\xi))$. The true system configuration is given by $\mu_1^* = 0.01$ and $\mu_6^* = 1$ in (22). We note that in this example $u(t) \in \mathbb{R}^2$, and we set in (17) $Q = \Delta\xi I_d, R = 0.01$. The discretization of (23) is done with $\Delta\xi = 0.025 = \Delta t$. This discretization leads to a problem of dimension $d = 121$. In Table 2, one can find the true coefficients versus the reconstructed ones at the last iteration using Algorithm 2. We can see that our algorithm matches the desired configuration considering the order of the finite discretization used.

The three trajectories are compared in Fig. 3. One can see a clear difference between the uncontrolled solution (left panel) and the other two plots. Moreover, the controlled trajectories show a similar behavior.

A more detailed comparison between Algorithm 1 and Algorithm 2 is shown in Fig. 4. Indeed, in the left plot, we show the two controls obtained from each algorithm that are very close to each other for both components. The evaluation of cost functional is shown in the middle plot of Fig. 4 and one can see that, again as expected, the value

Table 1 Test 1: Reconstructed parameter configuration for Allen-Cahn with $\Delta\xi = 0.01$, $\Delta t = 0.01$, 1% relative noise

True μ^*	1	0	11	0	-11	0	0
Estimated $\tilde{\mu}$	0.9992	-0.0017	11.0008	-0.0653	-10.8232	0.0431	0

The values reported in this table are taken at time $t = 0.5$

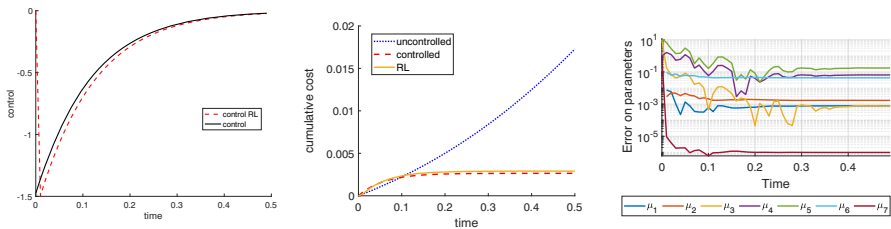


Fig. 2 Test 1: Allen-Cahn, $\Delta\xi = 0.01$, $\Delta t = 0.01$, 1% relative noise. On the left, the comparison between the control found using knowledge of the true μ and the control found by the RL algorithm is shown. In the middle, the cumulative cost. On the right, the error on the parameter estimation at each time

Table 2 Test 2. Reconstructed parameter configuration for viscous Burger with $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise

True μ^*	0.01	0	0	0	0	1	0
Estimated $\tilde{\mu}$	0.0096	0	-0.0008	0.002	-0.001	0.9999	0

The values reported in this table are taken at time $t = 0.625$

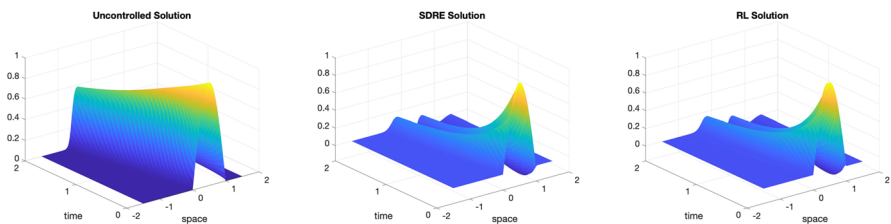


Fig. 3 Test 2: Viscous Burgers, $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise

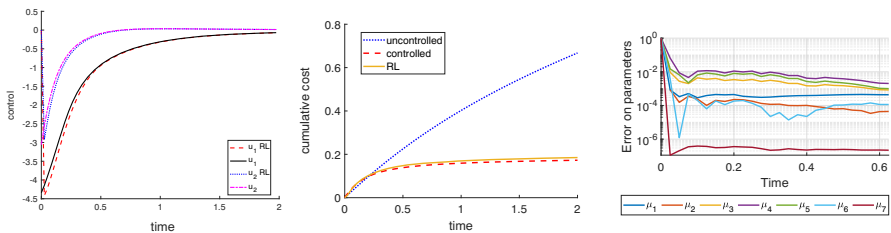


Fig. 4 Test 2: Viscous Burgers, $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise. The first plot shows the comparison between each of the two components of the control found using knowledge of the true μ and the control found by the RL algorithm. The other plots show the cumulative cost and the error on the parameter estimation at each time until the update stop

of the RL-based method is slightly larger to the SDRE method. Finally, in the right plot, we show the error in the convergence of the parameter configuration. The method stops updating the configuration estimate at time $t = 0.625$ (i.e., after 25 iterations out of 80).

Results with a black box. For this test case, we also show the results obtained using a real black box. At each iteration t_i we first solved the Riccati equation, thus obtaining K^i , then we found the control $u = K^i x^i$ and finally, we let the system evolve, thus obtaining x^{i+1} . For the evolution, we used the MATLAB function `ode15s` at each iteration. Table 3 shows the final approximations of the parameters. We can see that the term $\tilde{\mu}_5$, which was close to the correct value 0 using an implicit scheme in Table 2, appears in the reconstruction with a value of 0.1762 in this case. Nevertheless, the algorithm is still able to provide a stabilizing control which is very close to the one obtained using the implicit scheme, as shown in Fig. 5.

We then tested our algorithm using different libraries, i.e., considering only some F_j 's. In Table 4, we report the results when not considering the 5th term $y^3(t, x)$, whose parameter μ_5 is the extra term appearing when working with the whole library. It is then clear that our method works accurately. Furthermore, we report in Table 5 the results when considering only the terms that belong to the problem; again, our algorithm was able to approximate them well.

The reason why the algorithm is not able to approximate μ_5 might be that the components of the solution vector x all tend to zero with the applied control, so the components of x^3 (each element of x raised to the power of 3), which is the term that must be multiplied by μ_5 , tend to 0 very rapidly. This is also justified by the error in the infinity norm we computed in Table 6. There, we have computed the difference between the controlled solution and our RL solutions using the full library in the first column, the library without the μ_5 -term in the second column, and the library with only the correct terms in the last column. One can see that there is no difference in using the libraries chosen. Indeed, the computed controls appear to be the same and our method is always able to stabilize the problem even in the case of the full library (see Table 3). This further validates our method, which is able to stabilize the problem even if the discovered model does not match perfectly with the true system configuration. The reason is that even if a configuration $\tilde{\mu}$ doesn't match exactly the true configuration μ^* , by construction it solves the linear system (18) and so it well approximates the system dynamics, at least along the controlled trajectory.

Finally, for the sake of completeness, we show more details in Fig. 5 on the results, obtained with a black box, where the whole library was used. The top left panel shows the solution and the top right panel shows the error on parameters. In the bottom left panel, we show a comparison between the control found with the black box and the control computed by the algorithm that uses the implicit formula. The last plot shows

Table 3 Test 2: Viscous Burgers, $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise

True μ^*	0.01	0	0	0	0	1	0
Estimated $\tilde{\mu}$	0.0096	-0.0002	0.0008	0.0004	0.1762	1.021	0

Results with a black box, all parameters considered. The values reported in this table are taken at time $t = 2$. The stopping criterion was never matched

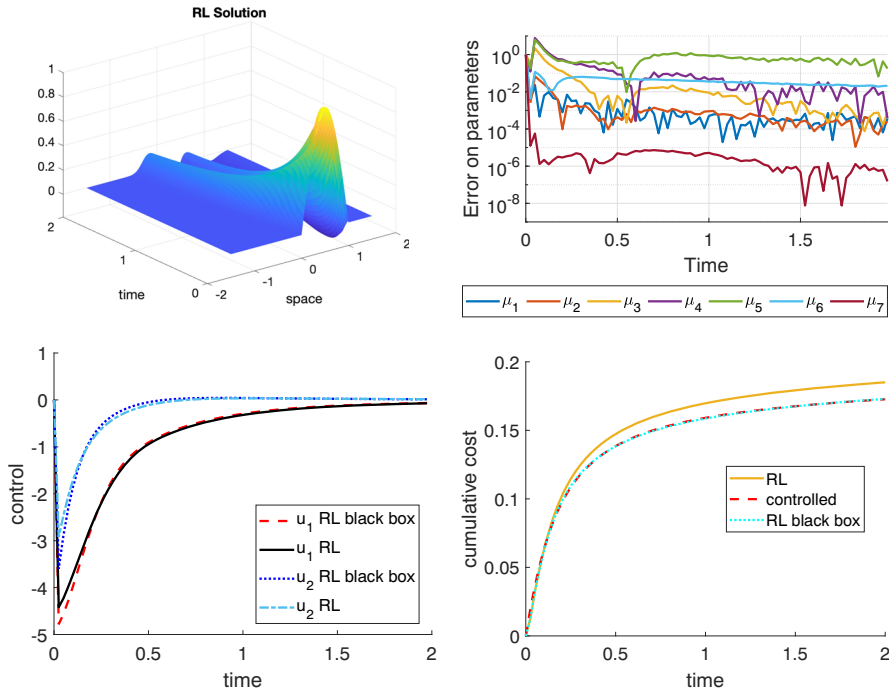


Fig. 5 Test 2: Viscous Burgers, $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise. Results with a black box, all parameters considered

Table 4 Test 2: $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise

True μ^*	0.01	0	0	0	–	1	0
Estimated $\tilde{\mu}$	0.0101	–0.0003	0.0007	0.0116	–	1.0199	0

Results with a black box, μ_5 not considered. The values reported in this table are taken at time $t = 2$

Table 5 Test 2, $\Delta\xi = 0.025$, $\Delta t = 0.025$, 1% relative noise

True μ^*	0.01	–	–	–	–	–	1
Estimated $\tilde{\mu}$	0.0099	–	–	–	–	–	1.0564

Results with a black box, only μ_1 and μ_6 considered. The values reported in this table are taken at time $t = 0.35$

Table 6 Difference between the controlled approximation y_c with Algorithm 1 for the known problem, the RL approximation y_{RL} and the RL with a black box y_{RL-bb}

	full library	No μ_5	Only μ_1, μ_6
$\ y_{RL} - y_{RL-bb}\ _\infty$	0.054254	0.054254	0.054254
$\ y_c - y_{RL-bb}\ _\infty$	0.113177	0.113177	0.113177
$\ y_c - y_{RL}\ _\infty$	0.111801	0.111801	0.111801

In the first column the error is computed using the results with a full library, in the second excluding the term μ_5 and in the third using the library only contains the terms μ_1 and μ_6 . The error shown in this table has to be understood in space-time domain

Table 7 Test 3: Reconstructed parameter configuration for Korteweg-de Vries, $\Delta\xi = 0.1$, $\Delta t = 0.025$, 1% relative noise

True μ^*	0.5	0	0	0	0	6	-1
Estimated $\tilde{\mu}$	0.4931	0.0012	0.0004	0.001	-0.0016	5.9943	-0.9999

The values reported in this table are taken at time $t = 1.275$

a comparison between the costs of the controlled solution and the two RL solutions (implicit and black box). Note that, even if the model parameters found with the black box algorithm are less accurate than the ones found with the implicit one, the cost of the applied control is very similar.

5.2.1 Test 3: Korteweg-de Vries

In the third model, we study the well-known Korteweg-de Vries (KdV) equation, with an additional diffusion term, which reads:

$$\begin{cases} y_t(t, \xi) = \frac{1}{2}y_{\xi\xi\xi}(t, \xi) + 6y(t, \xi)y_{\xi}(t, \xi) - y_{\xi\xi\xi}(t, \xi) \\ \quad + \chi_{[1,4]}(\xi)u(t), & t \in [0, 2], \xi \in (-10, 7), \\ y(0, \xi) = \chi_{[0,6]}(\xi)\left(\cos\left(\frac{\pi}{3}(\xi - 3)\right) + 1\right) & \xi \in (-10, 7), \\ y(t, -10) = 0, y(t, 7) = 0, & t \in [0, 2]. \end{cases}$$

Thus, it is a special case of (22) when $\mu_1^* = 0.5$, $\mu_6^* = 6$, $\mu_7^* = -1$. Note that in this test there is a third derivative in the equation. The boundary conditions are of Dirichlet type and the relative noise added was 1%. The finite difference discretization is performed choosing $\Delta\xi = 0.1$ which leads a problem (13) of dimension $d = 171$, and integrated in time with $\Delta t = 0.025$. The configuration found from our Algorithm 2 can be seen in Table 7. The update of the estimated configuration stopped after $t = 1.275$ (i.e., after 51 iterations out of 80). The reconstructed parameter configuration has a difference of order less than Δt with respect to the true configuration. Note that, from Table 7, we obtained $\|\mu^* - \tilde{\mu}\|_{\infty} = 0.0069 < \Delta t = 0.025$.

Again, as seen in the previous examples, this confirms the accurateness of our method.

The trajectories are presented in Fig. 6. One can see that the middle and right panels have a similar behavior whereas the uncontrolled simulation has a completely different evolution.

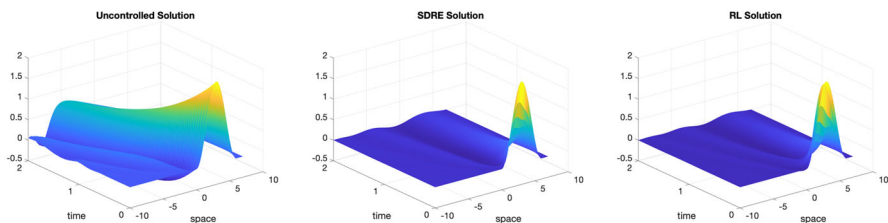


Fig. 6 Test 3: Korteweg-de Vries, $\Delta\xi = 0.1$, $\Delta t = 0.025$, 1% relative noise

Finally, we show the computed controls in the left panel of Fig. 7 which, after the first iterations, follow the same behavior. A more qualitative result is given in the middle panel of Fig. 7 where we can see the evaluation of cost functional (21). Again (and as expected) Algorithm 1 performs slightly better than Algorithm 2 but is still very close. To finalize, the history of the parameter configuration is shown in the right panel.

5.3 CPU times

In this subsection, we report in Table 8 the CPU times of the tests presented above. We compare the time needed to compute the uncontrolled, controlled, and RL solutions for each of the three presented test cases for the implicit scheme.

Since we have random components, the table has been obtained by executing the algorithm 50 times and then considering the arithmetic mean of the execution times. Table 8 shows that the time needed to obtain the solution with Algorithm 1 is similar to the time needed with our proposed method. This is because the number of PDEs solved is the same, the computation of the Bayesian linear regression is neglectable since we do not deal with large-scale problems, and in our problem, we have to solve one ARE less than SDRE since at the first iteration we decide to start with 0 control. In the third test, our method is slightly faster than SDRE, this also depends on the number of iterations needed in the Newton method which may be different since we opt for different control strategies.

To make the comparison fair, we consider the time needed to approximate the PDE in each method. Theoretically, one could think the black box in our method as an offline strategy with no cost.

Figure 8 shows the execution time needed to conclude each iteration of the solution (and control) computation for the uncontrolled, controlled, and RL cases. The final iteration times correspond to the times in Table 8. The uncontrolled case only requires the solution computation. We can observe that, in the first two tests, at the beginning the RL algorithm is slightly faster than the controlled one, and this is due to the choice of using a fixed control at the first iteration. Then, RL algorithm iterations are slightly slower, since more operations are carried out (e.g., Bayesian regression). This behavior is different for Test 3 as already commented. We also note that the uncontrolled KdV problem takes more time than the other two examples.

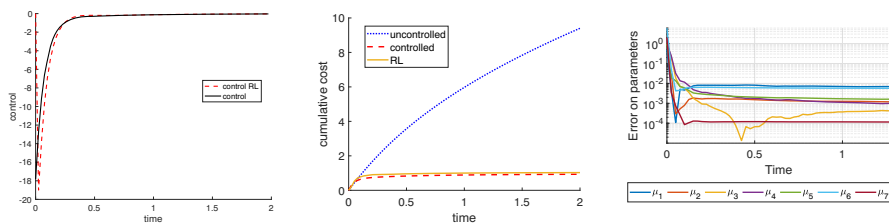


Fig. 7 Test 3: Korteweg-de Vries, $\Delta\xi = 0.1$, $\Delta t = 0.025$, 1% relative noise. On the left, the comparison between the control found using knowledge of the true μ and the control found by the RL algorithm is shown. In the middle the cumulative cost. On the right, the error on the parameter estimation at each time until the update stop

Table 8 CPU times in seconds of the three presented tests

	uncontrolled	Algorithm 1	Algorithm 2
Test 1	0.69s	8.7s	10.1s
Test 2	0.87s	19.9s	21.3s
Test 3	12.1s	67.7s	64.2s

The times have been computed as the arithmetic mean of the time required to complete 50 algorithm's executions. The runtimes for Algorithm 2 are rather constant with a small standard deviation

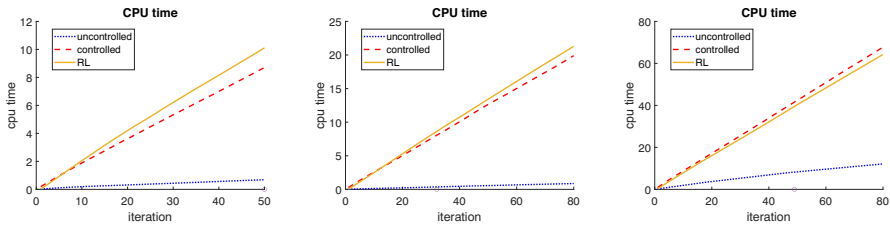


Fig. 8 Cumulative execution times at each iteration for Test 1 (left), Test 2 (middle), and Test 3 (Right). Mean times over 50 executions are considered

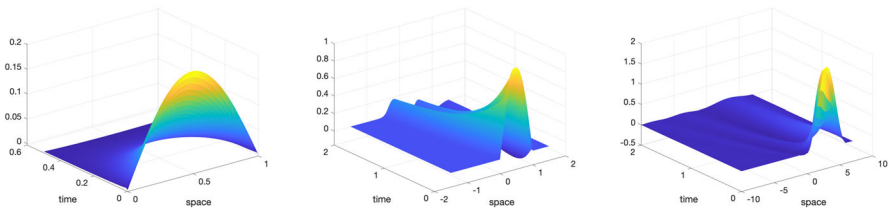


Fig. 9 Trajectories for Test 1 (left), Test 2(middle), Test 3(right) with spatial discretization $\frac{\Delta\xi}{2}$ using the RL control computed with $\Delta\xi$

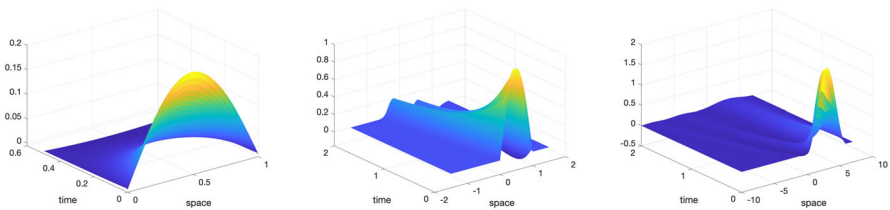


Fig. 10 Trajectories for Test 1 (left), Test 2(middle), Test 3(right) with spatial discretization $\frac{\Delta\xi}{4}$ using the RL control computed with $\Delta\xi$

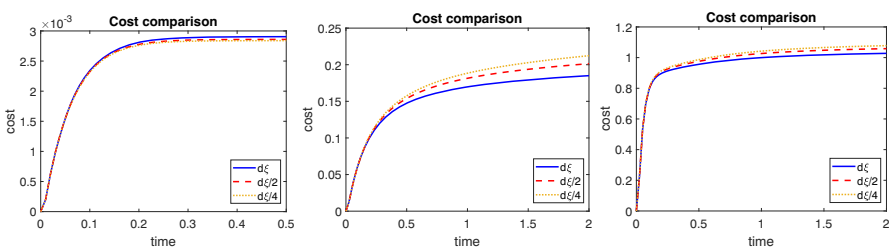


Fig. 11 Comparison of the cost functionals for Test 1 (left), Test 2(middle), Test 3(right)

5.4 Convergence to the PDE

To conclude, we provide a numerical assessment of the convergence of our method in a PDE control framework. We consider the examples of the previous sections and study the convergence of the control for increasing the dimension of the problem using the same time discretization grid used for each example to study the role of the mesh toward the control of the PDE. Thus, we have tested the control obtained for a discretized problem of dimension d (step $\Delta\xi$) using our Algorithm 2 and plugged into finer discretizations of the reference PDE of dimension, say $2d$ (step $\Delta\xi/2$) and $4d$ (step $\Delta\xi/4$). This has been done because, even if we use the true parameter μ^* for the evolution, the obtained dynamics is still an approximation of the PDE evolution, due to the use of numerical schemes. Finer grids allow us to better investigate the behavior of the system after the application of the computed control. For all three numerical examples, we plot the 3D solution generated with the finer grids (Figs. 9 and 10) and the cost computed accordingly (Fig. 11). We can see that the control found stabilizes the system also in these cases.

6 Conclusions

We proposed a new algorithm designed to control/stabilize unknown PDEs under certain assumptions. The strength of the method is the identification of the system on the fly, where at each iteration we provide a parameter estimate of the unknown system by Bayesian Linear regression. The update of the parameter configuration is based on the RL assumption where the user is always able to observe the true system configuration without its explicit knowledge. Numerical experiments have shown convergence results that validate our proposed approach. Since, to the best of the authors' knowledge, this is the first approach of this kind for nonlinear problems, we leave several open problems, such as efficient algorithms for higher dimensional problems and a theoretical study of the convergence for the proposed method. A preliminary work for two dimensional PDEs combined with model order reduction has been studied in [2]. Then, it will be interesting to add further unknowns in the problem, such as, e.g., the $B(x)$ term in the model and the quantity Q in the cost.

Acknowledgements The authors want to express their deep gratitude to Maurizio Falcone. Thanks to him the authors met up and started to collaborate on this project.

Data availability No data has been used in this paper.

Code availability The MATLAB source code for the implementations used to compute the presented results can be downloaded from <https://github.com/alessandroalla/SDRE-RL> upon request to the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

References


1. Alla, A., Kalise, D., Simoncini, V.: State-dependent Riccati equation feedback stabilization for nonlinear PDEs. *Adv. Comput. Math.* **49** (2023). <https://doi.org/10.1007/s10444-022-09998-4>
2. Alla, A., Pacifico, A.: A pod approach to identify and control PDEs online through state dependent Riccati equations. *Tech. Rep. arXiv:2402.08186* (2024)
3. Altmüller, N., Grüne, L., Worthmann, K.: Receding horizon optimal control for the wave equation. In: 49th IEEE Conference on Decision and Control (CDC), pp. 3427–3432 (2010). <https://doi.org/10.1109/CDC.2010.5717272>
4. Banks, H.T., Lewis, B.M., Tran, H.T.: Nonlinear feedback controllers and compensators: a state-dependent Riccati equation approach. *Comput. Optim. Appl.* **37**(2), 177–218 (2007)
5. Bardi, M., Capuzzo-Dolcetta, I.: Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Birkhauser (1997)
6. Bellman, R.: The theory of dynamic programming. *Bullet. American Math. Soc.* **60**(6), 503–515 (1954)
7. Bellman, R.: Adaptive control processes: a guided tour. Princeton University Press, Princeton, N.J. (1961)
8. Benner, P., Heiland, J.: Exponential stability and stabilization of extended linearizations via continuous updates of Riccati-based feedback. *Int. J. Robust Nonlinear Control* **28**(4), 1218–1232 (2018). <https://doi.org/10.1002/rnc.3949>
9. Bertsekas, D.: Reinforcement and optimal control. Athena Scientific (2019)
10. Bertsekas, D.P.: Approximate dynamic programming (2008)
11. Box, G.E., Tiao, G.C.: Bayesian inference in statistical analysis, vol. 40. John Wiley & Sons (2011)
12. Brunton, S., Proctor, J., Kutz, J.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 3932–3937 (2016)
13. Casper, S., Fuertinger, D.H., Kotanko, P., Mechelli, L., Rohleff, J., Volkwein, S.: Data-driven modeling and control of complex dynamical systems arising in renal anemia therapy. In: Ehrhardt, M., Günther, M. (eds.) *Progress in Industrial Mathematics at ECMI 2021*, pp. 155–161. Springer International Publishing, Cham (2022)
14. Cloutier, J.R.: State-dependent Riccati equation techniques: an overview. In: *Proceedings of the 1997 American Control Conference (Cat. No.97CH36041)*, vol. 2, pp. 932–936 vol.2 (1997)
15. Falcone, M., Ferretti, R.: Semi-Lagrangian approximation schemes for linear and Hamilton—Jacobi equations. SIAM (2013)
16. Freedman, D.A.: Statistical models: theory and practice. Cambridge University Press (2009)
17. Grüne, L., Pannek, J.: Nonlinear model predictive control. *Communications and Control Engineering Series*. Springer, London (2011). Theory and algorithms
18. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *International Conference on Machine Learning*, pp. 1861–1870. PMLR (2018)
19. Kaiser, E., Kutz, J.N., Brunton, S.L.: Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **474**(2219), 20180335 (2018). <https://doi.org/10.1098/rspa.2018.0335>
20. Karniadakis, G., Kevrekidis, I., Lu, L., Perdikaris, P., Wang, S., Yang, L.: Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 686–707 (2021). <https://doi.org/10.1038/s42254-021-00314-5>
21. Knoll, D.A., Keyes, D.E.: Jacobian-free Newton-Krylov methods: a survey of approaches and applications. *J. Comput. Phys.* **193**(2), 357–397 (2004)
22. Krstic, M., Smyshlyaev, A.: Adaptive control of PDEs. *IFAC Proceedings Volumes, 9th IFAC Workshop on Adaptation and Learning in Control and Signal Processing* **40**(13), 20–31 (2007). <https://doi.org/10.3182/20070829-3-RU-4911.00004>
23. LeVeque, R.J.: Finite difference methods for ordinary and partial differential equations. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2007). Steady-state and time-dependent problems. <https://doi.org/10.1137/1.9780898717839>
24. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. In: 4th International Conference on Learning Representations (ICLR) (2016)

25. Martinsen, A.B., Lekkas, A.M., Gros, S.: Combining system identification with reinforcement learning-based MPC. *IFAC-PapersOnLine* **53**(2), 8130–8135 (2020). <https://doi.org/10.1016/j.ifacol.2020.12.2294>. 21st IFAC World Congress
26. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. *Nature* **518**(7540), 529–533 (2015)
27. Zolman, N., Fasel, U., Kutz, J.N. and Brunton, S.L.: SINDy-RL: interpretable and efficient model-based reinforcement learning. *Tech. Rep.* (2024). [arXiv:2403.09110](https://arxiv.org/abs/2403.09110)
28. Pacifico, A., Pesare, A., Falcone, M.: A new algorithm for the LQR problem with partially unknown dynamics. In: Lirkov, I., Margenov, S. (eds.) *Large-Scale Scientific Computing*, pp. 322–330. Springer International Publishing, Cham (2022)
29. Powell, W.B.: *Approximate dynamic programming: solving the curses of dimensionality*, vol. 703. John Wiley & Sons (2007)
30. Powell, W.B.: From reinforcement learning to optimal control: a unified framework for sequential decisions. In: *Handbook of Reinforcement Learning and Control*, pp. 29–74. Springer (2021)
31. Raissi, M., Perdikaris, P., Karniadakis, G.: Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Computat. Phys.* **378**, 686–707 (2019). <https://doi.org/10.1016/j.jcp.2018.10.045>
32. Rasmussen, C., Williams, C.: *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA (2006)
33. Rossi, P.E., Allenby, G.M., McCulloch, R.: *Bayesian statistics and marketing*. John Wiley & Sons (2012)
34. Rudy, S., Alla, A., Brunton, S.L., Kutz, J.N.: Data-driven identification of parametric partial differential equations. *SIAM J. Appl. Dynamical Syst.* **18**(2), 643–660 (2019). <https://doi.org/10.1137/18M1191944>
35. Rudy, S., Brunton, S., Proctor, J., Kutz, J.: Data-driven discovery of partial differential equations. *Sci. Adv.* **3** (2017)
36. Rummery, G.A., Niranjan, M.: *On-line Q-learning using connectionist systems*, vol. 37. Citeseer (1994)
37. Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. In: *International conference on machine learning*, pp. 1889–1897. PMLR (2015)
38. Sutton, R.S.: Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**(1), 9–44 (1988)
39. Sutton, R.S., Barto, A.G.: *Reinforcement learning: an introduction*, vol. 1, first edn. MIT Press, Cambridge, MA (1998)
40. Sutton, R.S., Barto, A.G.: *Reinforcement learning: an introduction*, 2nd edn. MIT Press, Cambridge, MA (2018)
41. Watkins, C., Hellaby, J.C.: *Learning from delayed rewards* (1989)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Alessandro Alla¹  · **Agnese Pacifico**² · **Michele Palladino**³ · **Andrea Pesare**⁴

✉ Alessandro Alla
alessandro.alla@unive.it

Agnese Pacifico
agnese.pacifico@uniroma1.it

Michele Palladino
michele.palladino@univaq.it

Andrea Pesare
andreapesare1@gmail.com

- ¹ Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari, Venezia, Italy
- ² Department of Mathematics, Sapienza University of Rome, Rome, Italy
- ³ Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy
- ⁴ Viale Molire 51, Rome 00142, Italy