REINFORCEMENT LEARNING-BASED ESTIMATION FOR PARTIAL DIFFERENTIAL EQUATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In systems governed by nonlinear partial differential equations such as fluid flows, the design of state estimators such as Kalman filters relies on a reduced-order model (ROM) that projects the original high-dimensional dynamics onto a computationally tractable low-dimensional space. However, ROMs are prone to large errors, which negatively affects the performance of the estimator. Here, we introduce the reinforcement learning reduced-order estimator (RL-ROE), a ROM-based estimator in which the correction term that takes in the measurements is given by a nonlinear policy trained through reinforcement learning. The nonlinearity of the policy enables the RL-ROE to compensate efficiently for errors of the ROM, while still taking advantage of the imperfect knowledge of the dynamics. Using examples involving the Burgers and Navier-Stokes equations, we show that in the limit of very few sensors, the trained RL-ROE outperforms a Kalman filter designed using the same ROM. Moreover, it yields accurate high-dimensional state estimates for reference trajectories corresponding to various physical parameter values, without direct knowledge of the latter.

1 INTRODUCTION

Active control of turbulent flows has the potential to cut down emissions across a range of industries through drag reduction in aircrafts and ships or improved efficiency of heating and air-conditioning systems, among many other examples (Brunton & Noack, 2015). But real-time feedback control requires inferring the state of the system from sparse measurements using an algorithm called a state estimator, which typically relies on a model for the underlying dynamics (Simon, 2006). Among state estimators, the Kalman filter is by far the most well-known thanks to its optimality for linear systems, which has led to its widespread use in numerous applications (Kalman, 1960; Zarchan, 2005). However, continuous systems such as fluid flows are governed by partial differential equations (PDEs) which, when discretized, yield high-dimensional and oftentimes nonlinear dynamical models with hundreds or thousands of state variables. These high-dimensional models are too expensive to integrate with common state estimation techniques, including the Kalman filter or its numerous extensions. Thus, state estimators are instead designed based on a reduced-order model (ROM) of the system, in which the underlying dynamics are projected to a low-dimensional subspace that is computationally tractable (Barbagallo et al., 2009; Rowley & Dawson, 2017).

A big challenge is that ROMs provide a simplified and imperfect description of the dynamics, which negatively affects the performance of the state estimator. One potential solution is to improve the accuracy of the ROM through the inclusion of additional closure terms (Ahmed et al., 2021). In this paper, we leave the ROM untouched and instead propose a new design paradigm for the estimator itself, which we call a reinforcement-learning reduced-order estimator (RL-ROE). The RL-ROE is constructed from the ROM in an analogous way to a Kalman filter, with the crucial difference that the linear filter gain function, which takes in the current measurement data, is replaced by a nonlinear policy trained through reinforcement learning (RL). The flexibility of the nonlinear policy, parameterized by a neural network, enables the RL-ROE to compensate for errors of the ROM while still taking advantage of the imperfect knowledge of the dynamics. We describe how we frame the problem as a stationary Markov decision process in order to enable RL training, a non-trivial task since the RL-ROE must be able to estimate time-varying states. Finally, we show that in the limit of sparse measurements, the trained RL-ROE outperforms a Kalman filter designed using the

same ROM and displays robust estimation performance across different dynamical regimes. To our knowledge, the RL-ROE is the first application of RL to state estimation for systems governed by parametric PDEs.

2 General methodology

2.1 Problem formulation

Consider the parametric discrete-time nonlinear system given by

$$\boldsymbol{z}_{k+1} = \boldsymbol{f}(\boldsymbol{z}_k; \mu), \tag{1a}$$

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{z}_k, \tag{1b}$$

where $\boldsymbol{z}_k \in \mathbb{R}^n$ and $\boldsymbol{y}_k \in \mathbb{R}^p$ are respectively the state and measurement at time k, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a time-invariant nonlinear map from current to next state, $\mu \in \mathbb{R}$ is a physical parameter, and $\boldsymbol{C} \in \mathbb{R}^{p\times n}$ is a linear map from state to measurement. In this study, we assume that the dynamics given in (1) are obtained from the numerical discretization of a nonlinear partial differential equation (PDE), which typically requires a large number n of continuous state variables (on the order of at least a few hundreds). Note that we do not account for exogenous control inputs to the system, which will be studied in future extensions of the present work.

Here, we will focus on the post-transient dynamics of (1); these are the observed dynamics once the transients associated with the initial condition have died down. In particular, we consider systems whose post-transient dynamics are described by an attractor that is either a steady state, a periodic limit cycle or a quasi-periodic limit cycle, which encompasses the behavior of a large class of physical systems. The nature of the attractor is independent of the initial condition but depends on the value of μ, which we will consider to be in a range $[\mu_1, \mu_2]$.

The purpose of the present work is to combine reduced-order modeling (ROM) and reinforcement learning (RL) to construct a state estimator that solves the following problem: given a sequence of measurements $\{\boldsymbol{y}_0, \cdots, \boldsymbol{y}_k\}$ from a post-transient reference trajectory of (1), estimate the high-dimensional state $\boldsymbol{z}_k$ at current time k without knowledge of μ itself. The ROM procedure, which follows standard practices, is described in Section 2.2. The integration of the ROM with RL to solve the estimation problem, which constitutes the main novelty of the paper, is described in Section 2.3.

2.2 Reduced-order model

Since the high dimensionality of (1) renders online estimation impractical, it is customary to formulate a reduced-order model (ROM) of the dynamics (Rowley & Dawson, 2017). First, one chooses a suitable linearly independent set of modes $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$, where $\boldsymbol{u}_i \in \mathbb{R}^n$, defining an r-dimensional subspace of $\mathbb{R}^n$ in which most of the dynamics is assumed to take place. Stacking these modes as columns of a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, one can then express $\boldsymbol{z}_k \simeq \boldsymbol{U}\boldsymbol{x}_k$, where the reduced-order state $\boldsymbol{x}_k \in \mathbb{R}^r$ represents the coordinates of $\boldsymbol{z}_k$ in the subspace. Finally, one finds a ROM for the dynamics of $\boldsymbol{x}_k$, which is vastly cheaper to evolve than (1) when $r \ll n$.

There exist various ways to find an appropriate set of modes $\boldsymbol{U}$ and corresponding ROM for the dynamics of $\boldsymbol{x}_k$ (Taira et al., 2017). In this work, we employ the Dynamic Mode Decomposition (DMD), a purely data-driven algorithm that has found numerous applications in fields ranging from fluid dynamics to neuroscience (Schmid, 2010; Kutz et al., 2016). Importantly, we seek a single ROM to describe dynamics corresponding to various $\mu \in [\mu_1, \mu_2]$ since the state estimator that we will later construct based on this ROM does not have knowledge of μ.

In order to apply DMD, we first construct a training dataset by solving (1) for values of μ belonging to a finite set $S \subset [\mu_1, \mu_2]$, resulting in a concatenated collection of snapshots $\boldsymbol{Z}_{\text{train}} = \{\boldsymbol{Z}^\mu\}_{\mu\in S}$, where each $\boldsymbol{Z}^\mu = \{\boldsymbol{z}_0^\mu, \ldots, \boldsymbol{z}_m^\mu\}$ is a post-transient trajectory of (1a) for a specific value $\mu \in S$. The DMD then seeks a best-fit linear model of the dynamics in the form of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{z}_{k+1}^\mu \simeq \boldsymbol{A}\boldsymbol{z}_k^\mu$ for all k and μ, and computes the modes $\boldsymbol{U}$ as the r leading principal component analysis (PCA) modes of $\boldsymbol{Z}$. The transformation $\boldsymbol{z}_k \simeq \boldsymbol{U}\boldsymbol{x}_k$ and the orthogonality of $\boldsymbol{U}$ then yield a linear discrete-time ROM of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}_r\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{2a}$$

$$\boldsymbol{y}_k = \boldsymbol{C}_r \boldsymbol{x}_k + \boldsymbol{v}_k, \tag{2b}$$

where $\boldsymbol{A}_r = \boldsymbol{U}^\mathsf{T} \boldsymbol{A} \boldsymbol{U} \in \mathbb{R}^{r \times r}$ and $\boldsymbol{C}_r = \boldsymbol{C}\boldsymbol{U} \in \mathbb{R}^{p \times r}$ are the reduced-order state-transition and observation models, respectively. The (unknown) non-Gaussian process noise $\boldsymbol{w}_k$ and observation noise $\boldsymbol{v}_k$ account for the neglected PCA modes of $\boldsymbol{Z}$, as well as the error incurred by the linear approximation and effective averaging of the dynamics over a range of μ. Additional details regarding the calculation of $\boldsymbol{A}_r$ and $\boldsymbol{U}$ are provided in Appendix A of the supplementary materials.

2.3 Reinforcement learning-based reduced-order estimator

Using the ROM (2) defined by $\boldsymbol{A}_r$, $\boldsymbol{C}_r$ and $\boldsymbol{U}$, we now want to solve the estimation problem defined in Section 2.1. To this effect, we propose to design a reduced-order estimator (ROE) of the form

$$\hat{\boldsymbol{x}}_k = \boldsymbol{A}_r \hat{\boldsymbol{x}}_{k-1} + \boldsymbol{a}_k, \tag{3a}$$

$$\boldsymbol{a}_k \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}(\,\cdot\,|\boldsymbol{y}_k, \hat{\boldsymbol{x}}_{k-1}), \tag{3b}$$

where $\hat{\boldsymbol{x}}_k$ is an estimate of the reduced-order state $\boldsymbol{x}_k$ and $\boldsymbol{a}_k \in \mathbb{R}^r$ is an action sampled from a nonlinear and stochastic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, which takes as input the current measurement $\boldsymbol{y}_k$ from the reference trajectory of (1a) and the previous state estimate $\hat{\boldsymbol{x}}_{k-1}$. The subscript $\boldsymbol{\theta}$ denotes the set of parameters that defines the policy, whose goal is to use the sparse measurements $\boldsymbol{y}_k$ to act on the dynamics of $\hat{\boldsymbol{x}}_k$ in (3a) so that the reconstructed high-dimensional state estimate $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$ converges towards the reference $\boldsymbol{z}_k$.

A Kalman filter is a special case of such an estimator, for which the action in (3b) is given by

$$\boldsymbol{a}_k = \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{C}_r \boldsymbol{A}_r \hat{\boldsymbol{x}}_{k-1}), \tag{4}$$

with $\boldsymbol{K}_k \in \mathbb{R}^{r \times p}$ the optimal Kalman gain. Although the Kalman filter is optimal when the state-transition and observation models are known exactly, its performance suffers in the presence of unmodeled dynamics.

In our case, such model errors are unavoidable due to the ROM (2) being an inherent approximation of the high-dimensional parametric dynamics (1), which motivates our adoption of the more general form (3b). This form retains the dependence of $\boldsymbol{a}_k$ on $\boldsymbol{y}_k$ and $\hat{\boldsymbol{x}}_{k-1}$ but is more flexible thanks to the nonlinearity of the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, which we train with deep RL in an offline stage. The stochasticity of $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ lets the RL algorithm explore different actions during the training process, in order to find eventually an optimal $\boldsymbol{\theta}^*$ such that $\mathbb{E}[\boldsymbol{z}_k - \hat{\boldsymbol{z}}_k]$ is minimized for a range of initial estimates $\hat{\boldsymbol{x}}_0$ and across parameter values $\mu \in [\mu_1, \mu_2]$. We call the estimator constructed and trained through this process an RL-trained ROE, or RL-ROE for short.

2.4 Summary of the proposed methodology

In summary, the methodology we propose consists of the following three steps. The first two are carried out offline using a training dataset $\boldsymbol{Z}_{\text{train}} = \{\boldsymbol{Z}^\mu\}_{\mu \in S}$ containing high-dimensional state snapshots from multiple trajectories of (1) for various μ. The third takes place online using sole knowledge of measurements $\{\boldsymbol{y}_0, \cdots, \boldsymbol{y}_k\}$ from a trajectory of (1) corresponding to unknown μ not necessarily belonging to S.

1. **Construction of the ROM (offline)**. A ROM of the form (2) is obtained by applying the DMD to the training dataset $\boldsymbol{Z}_{\text{train}}$.
2. **Training of the RL-ROE (offline)**. An RL-ROE of the form (3) is designed based upon the ROM constructed in Step 1, and its policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is trained using the reference trajectories contained in $\boldsymbol{Z}_{\text{train}}$.
3. **Deployment of the RL-ROE (online)**. Using measurements $\boldsymbol{y}_k$ from a reference trajectory of (1) corresponding to unknown μ, the trained RL-ROE returns an estimate $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$ for the corresponding (unobserved) high-dimensional state $\boldsymbol{z}_k$.

The RL-ROE, which combines a ROM with a nonlinear policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ trained in an RL framework to solve the estimation problem, constitutes the main contribution of the present paper. The training procedure for $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, which involves a non-trivial reformulation of the time-varying estimation problem into a stationary Markov decision process, is described in the next section.

3 TRAINING METHODOLOGY

In order to train $\boldsymbol{\pi_\theta}$ with reinforcement learning, we need to formulate the problem as a stationary Markov decision process (MDP). However, this is no trivial task given that the aim of the policy is to minimize the error between the state estimate $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$ and a time-dependent reference state $\boldsymbol{z}_k$. At first sight, such trajectory tracking problem requires a time-dependent reward function and, therefore, a time-varying MDP.

To be able to use off-the-shelf RL algorithms, we introduce a trick to translate this time-varying MDP to an equivalent stationary MDP based on an extended state. Indeed, we show hereafter that the problem can be framed as a stationary MDP by including $\boldsymbol{z}_k$ and μ into our definition of the MDP's state. Letting $\boldsymbol{s}_k = (\hat{\boldsymbol{x}}_{k-1}, \boldsymbol{z}_k, \mu) \in \mathbb{R}^{r+n+1}$ denote an augmented state at time k, we can define an MDP consisting of the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S} = \mathbb{R}^{r+n+1}$ is the augmented state space, $\mathcal{A} \subset \mathbb{R}^r$ is the action space, $\mathcal{P}(\cdot|\boldsymbol{s}_k, \boldsymbol{a}_k)$ is a transition probability, and $\mathcal{R}(\boldsymbol{s}_k, \boldsymbol{a}_k, \boldsymbol{s}_{k+1})$ is a reward function. At each time step k, the agent selects an action $\boldsymbol{a}_k \in \mathcal{A}$ according to the policy $\boldsymbol{\pi_\theta}$ defined in (3b), which can be expressed as

$$\boldsymbol{a}_k \sim \boldsymbol{\pi_\theta}(\,\cdot\,|\boldsymbol{o}_k), \tag{5}$$

where $\boldsymbol{o}_k = (\boldsymbol{y}_k, \hat{\boldsymbol{x}}_{k-1}) = (\boldsymbol{C}\boldsymbol{z}_k, \hat{\boldsymbol{x}}_{k-1})$ is a partial observation of the current state $\boldsymbol{s}_k$. The state $\boldsymbol{s}_{k+1} = (\hat{\boldsymbol{x}}_k, \boldsymbol{z}_{k+1}, \mu)$ at the next time step is then obtained from equations (1a) and (3a) as

$$\boldsymbol{s}_{k+1} = (\boldsymbol{A}_r\hat{\boldsymbol{x}}_{k-1} + \boldsymbol{a}_k, \boldsymbol{f}(\boldsymbol{z}_k; \mu), \mu), \tag{6}$$

which defines the transition model $\boldsymbol{s}_{k+1} \sim \mathcal{P}(\cdot|\boldsymbol{s}_k, \boldsymbol{a}_k)$. Finally, the agent receives the reward

$$r_k = \mathcal{R}(\boldsymbol{s}_k, \boldsymbol{a}_k, \boldsymbol{s}_{k+1}) = -\lambda_1||\boldsymbol{z}_k - \boldsymbol{U}\hat{\boldsymbol{x}}_k||^2 - \lambda_2||\boldsymbol{a}_k||^2, \tag{7}$$

where λ_1 and λ_2 are scalar weights. The first term in the reward function $\mathcal{R}$ penalizes the difference between the high-dimensional state estimate $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$ and the reference $\boldsymbol{z}_k$, which is only partially observed by the agent. The second term favors smaller values for the action $\boldsymbol{a}_k$, which acts as a regularization mechanism. Unless indicated otherwise, we will consider $\lambda_1 = 1$ and $\lambda_2 = 0$. Thanks to the incorporation of $\boldsymbol{z}_k$ into $\boldsymbol{s}_k$, the reward function (7) has no explicit time dependence and the MDP is therefore stationary.

The goal of the RL training process is then to find the optimal policy parameters

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{\tau\sim\boldsymbol{\pi_\theta}}[R(\tau)], \tag{8}$$

where the expectation is over trajectories $\tau = (\boldsymbol{s}_1, \boldsymbol{a}_1, \boldsymbol{s}_2, \boldsymbol{a}_2, \dots)$, and $R(\tau) = \sum_{k=1}^{K} r_k$ is the finite-horizon undiscounted return, with the integer K denoting the length of each training episode. At the beginning of every episode, the environment is reset according to the distributions

$$\hat{\boldsymbol{x}}_0 \sim \boldsymbol{p}_{\hat{\boldsymbol{x}}_0}(\cdot), \quad \boldsymbol{z}_0 \sim \boldsymbol{p}_{\boldsymbol{z}_0}(\cdot), \quad \mu \sim p_\mu(\cdot), \tag{9}$$

from which the augmented state $\boldsymbol{s}_1 = (\hat{\boldsymbol{x}}_0, \boldsymbol{z}_1, \mu) = (\hat{\boldsymbol{x}}_0, \boldsymbol{f}(\boldsymbol{z}_0; \mu), \mu)$ follows immediately; $\boldsymbol{s}_1$ thus constitutes the start of the agent-environment interactions. The distribution $\boldsymbol{p}_{\hat{\boldsymbol{x}}_0}(\cdot)$ bestows robustness of the learned policy with respect to the initial state estimate $\hat{\boldsymbol{x}}_0$, while the distributions $\boldsymbol{p}_{\boldsymbol{z}_0}(\cdot)$ and $p_\mu(\cdot)$ enable the same policy $\boldsymbol{\pi_\theta}$ to be trained on several reference trajectories of (1) corresponding to various $\mu \in [\mu_1, \mu_2]$. In practice, we reuse the same reference trajectories corresponding to $\mu \in S$ from the training dataset $\boldsymbol{Z}$ utilized to construct the DMD in Section 2.2, so that we do not have to keep solving (1a) during the training process. Thus, at the beginning of every episode, we draw a random μ from S and we initialize $\boldsymbol{z}_0$ as the starting state $\boldsymbol{z}_0^\mu$ of the corresponding trajectory $\boldsymbol{Z}^\mu = \{\boldsymbol{z}_0^\mu, \dots, \boldsymbol{z}_m^\mu\}$.

To find the optimal policy parameters $\boldsymbol{\theta}^*$, we employ the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017), which belongs to the class of policy gradient methods (Sutton et al., 2000). PPO alternates between sampling data by computing a set of trajectories $\{\tau_1, \tau_2, \tau_3, \dots\}$ using the most recent version of the policy, and updating the policy parameters $\boldsymbol{\theta}$ in a way that increases the probability of actions that led to higher rewards during the sampling phase. The policy $\boldsymbol{\pi_\theta}$ encodes a diagonal Gaussian distribution described by a neural network that maps from observation to mean action, $\boldsymbol{\mu_{\theta'}}(\boldsymbol{o}_k)$, together with a vector of standard deviations $\boldsymbol{\sigma}$, so that $\boldsymbol{\theta} = \{\boldsymbol{\theta}', \boldsymbol{\sigma}\}$. We utilize the Stable Baselines3 (SB3) implementation of PPO (Raffin et al., 2019) and define our MDP as a custom environment in OpenAI Gym (Brockman et al., 2016). Besides the discount factor γ,

all results to follow are obtained with the default PPO hyperparameters in SB3, which demonstrates the robustness of our approach with respect to the RL hyperparameters.

Remark. Since the policy (5) is conditioned on a partial observation $\boldsymbol{o}_k$ of the state $\boldsymbol{s}_k$, the MDP we have defined in this section is, in fact, a partially observable MDP (POMDP). In this case, it is known that the globally optimal policy depends on a summary of the history of past observations and actions, $\boldsymbol{h}_k = \{\boldsymbol{o}_1, \boldsymbol{a}_1, \ldots, \boldsymbol{o}_k\}$, rather than just the current observation $\boldsymbol{o}_k$ (Kaelbling et al., 1998). However, policies formulated based on an incomplete summary of $\boldsymbol{h}_k$ are common in practice and still achieve good results (Sutton & Barto, 2018). We therefore pursue this approach in the present paper, and leave for future work testing the generalization of our policy input to a more complete summary of $\boldsymbol{h}_k$. We also note that policy gradient methods, which PPO belongs to, do not require the Markov property of the state (that is, conditional independence of future states on past states given the present state) and can therefore be readily applied to the POMDP setting. For our problem, this guarantees that the PPO algorithm will converge to a locally optimum policy.

4 RESULTS

We evaluate the state estimation performance of the RL-ROE for systems governed by the Burgers equation and Navier-Stokes equations. For each system, we first compute various solution trajectories corresponding to different physical parameter values, which we use to construct the ROM and train the RL-ROE following the procedure outlined in Section 2.4. The trained RL-ROE is finally deployed online and compared against a time-dependent Kalman filter constructed from the same ROM, which we refer to as KF-ROE. The KF-ROE is given by equations (3a) and (4), with the calculation of the time-varying Kalman gain detailed in Appendix B of the supplementary materials.

4.1 BURGERS EQUATION

The forced Burgers equation is a prototypical nonlinear hyperbolic PDE that takes the form

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = f(x,t), \tag{10}$$

where $u(x,t)$ is the velocity at position $x \in [0, L]$ and time t, $f(x,t)$ is a distributed time-dependent forcing, and the scalar ν acts like a viscosity. Here, we choose a forcing of the form

$$f(x,t) = 2\sin(\omega t - kx) + 2\sin(3\omega t - kx) + 2\sin(5\omega t - kx), \tag{11}$$

where $k = 2\pi/L$, and we let ν and ω be related through a scalar parameter $\mu \in [0, 1]$ as follows:

$$\nu = \nu_1 + (\nu_2 - \nu_1)\mu, \qquad \omega = \omega_1 + (\omega_2 - \omega_1)\mu. \tag{12a}$$

Thus, μ can be regarded as a physical parameter that affects the dynamics of the forced Burgers equation through both ν and ω. We consider periodic boundary conditions and choose $L = 1$, $\nu_1 = 0.01$, $\nu_2 = 0.1$, $\omega_1 = 0.2\pi$, $\omega_1 = 0.4\pi$.

We solve the forced Burgers equation using a spectral method with $n = 256$ Fourier modes and a fifth-order Runge-Kutta time integration scheme. We define the discrete-time state vector $\boldsymbol{z}_k \in \mathbb{R}^n$ that contains the values of u at n equally-spaced collocation points and at discrete time steps $t = k\Delta t$, where $\Delta t = 0.05$. To generate the training dataset $\boldsymbol{Z}_{\text{train}} = \{\boldsymbol{Z}^\mu\}_{\mu \in S}$ used for constructing the ROM and training the RL-ROE, we compute solutions of the Burgers equation corresponding to $\mu \in S = \{0, 0.1, 0.2, \ldots, 0.1\}$. For each μ, we discard the transient portion of the dynamics and save 201 snapshots $\boldsymbol{Z}^\mu = \{\boldsymbol{z}_0^\mu, \ldots, \boldsymbol{z}_{200}^\mu\}$ in the post-transient regime. We retain $r = 10$ modes when constructing the ROM, corresponding to an-order-of-magnitude reduction in the dimensionality of the system. We train the RL-ROE using episodes of length $K = 200$ steps to make full use of the trajectories stored in $\boldsymbol{Z}_{\text{train}}$, and we end the training process when the return no longer increases on average. The RL hyperparameters and learning curve displaying the performance improvement of the RL-ROE during the training process are reported in Appendix D.

The trained RL-ROE and the KF-ROE are now compared based on their ability to track reference trajectories corresponding to various μ using sparse measurements of u from a limited number p of equally-spaced sensors. The evaluation is carried out using 20 initial state estimates sampled from a Gaussian distribution with unit standard deviation. Beginning with $p = 4$ sensors, Figure 1 reports the mean (lines) and standard deviation (shaded areas) of the normalized L_2 error for 3 values of μ not present in the training dataset $\boldsymbol{Z}_{\text{train}}$ used to construct the ROM and train the RL-ROE.

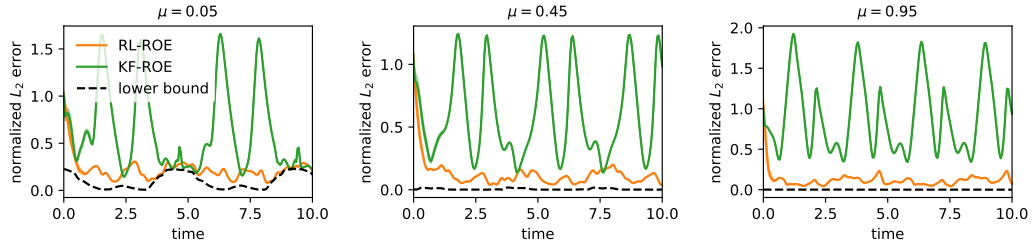

Figure 1: Burgers equation with $p = 4$ sensors. Normalized L_2 error of the RL-ROE and KF-ROE for the estimation of trajectories corresponding to values of μ not seen during training.

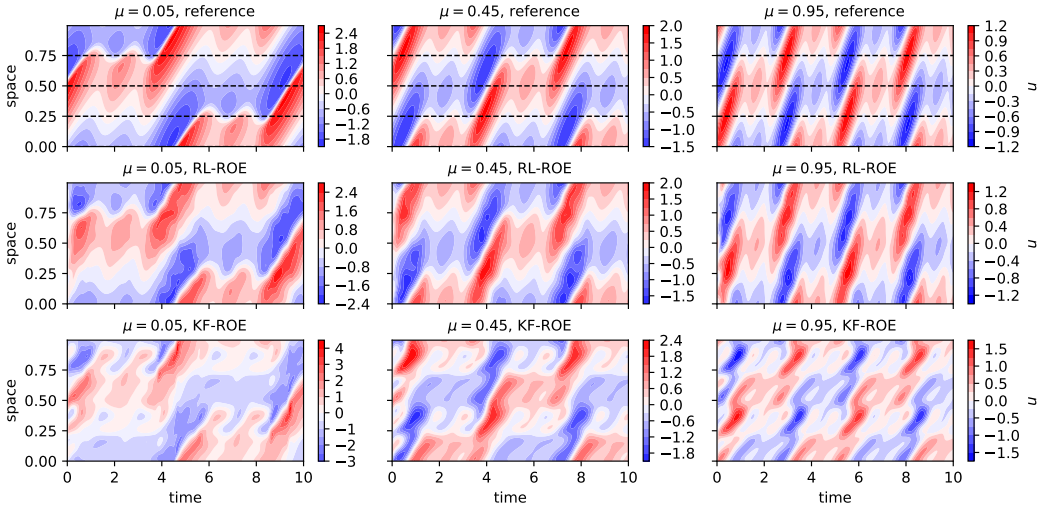

Figure 2: Burgers equation with $p = 4$ sensors. Reference trajectories for values of μ not seen during training and corresponding RL-ROE and KF-ROE estimates. The dashed lines on the reference trajectory plots indicate the sensor data seen by the RL-ROE and KF-ROE.

The normalized L_2 error is defined as $||\hat{\boldsymbol{z}}_k - \boldsymbol{z}_k||/||\boldsymbol{z}_k||$, where $\boldsymbol{z}_k$ is the high-dimensional reference solution and $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$ is the high-dimensional reconstruction of the RL-ROE or KF-ROE estimate $\hat{\boldsymbol{x}}_k$. The error of the RL-ROE is very close to the lower bound, which is the error incurred by projecting the reference solution $\boldsymbol{z}_k$ to the modes $\boldsymbol{U}$, i.e. the lowest possible error achievable by an ROE based on $\boldsymbol{U}$. Spatio-temporal contours of the reference solutions for the same 3 values of μ and the corresponding RL-ROE and KF-ROE high-dimensional estimates are shown in Figure 2. The RL-ROE vastly outperforms the KF-ROE, which demonstrates the superiority of a nonlinear correction to the estimator dynamics (3a). We emphasize that the Kalman filter was tuned to obtain the best possible results for the KF-ROE.

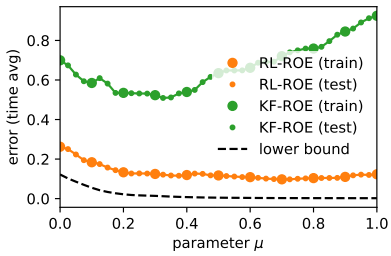

Figure 3: Burgers equation with $p = 4$ sensors. Time average of the normalized L_2 error versus μ. Values of μ present in $\boldsymbol{Z}_{\text{train}}$ shown by large circles.

Figure 3 reports the time average of the normalized L_2 error as a function of μ for $p = 4$, with the values present in $\boldsymbol{Z}_{\text{train}}$ indicated by large circles. The RL-ROE exhibits robust performance across the entire parameter range $\mu \in [0, 1]$, including when estimating trajectories unseen during the training.

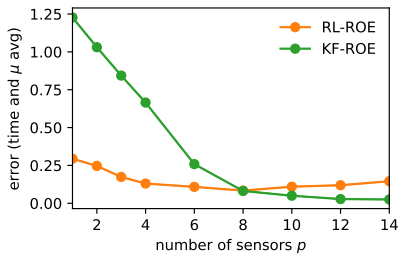

Figure 4: Burgers equation. Average over time and over μ of the normalized L_2 error versus number p of sensors.

Finally, Figure 4 displays the average over time and over μ of the normalized L_2 error for varying number p of sensors. Note that each value of p corresponds to a separately trained RL-ROE. As the number of sensors increases, the KF-ROE performs better and better until its accuracy overtakes that of the RL-ROE. We hypothesize that the accuracy of the RL-ROE is limited by the inability of the RL training process to find an optimal policy, due to both the non-convexity of the optimization landscape as well as shortcomings inherent to current deep RL algorithms. This being said, the strength of the nonlinear policy of the RL-ROE becomes very clear in the very sparse sensing regime; its performance remains remarkably robust as the number of sensors reduces to 2 or even 1. In Appendix E, spatio-temporal contours (similar as in Figure 2) of the reference solution and corresponding estimates for $p = 2$ and 12 illustrate that the slight advantage held by the KF-ROE for $p = 12$ is reversed into clear superiority of the RL-ROE for $p = 4$.

4.2 Navier-Stokes equations

The Navier-Stokes equations are a set of nonlinear PDEs that describe the motion of fluids flows. For incompressible fluids, the Navier-Stokes equations take the form

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = -\nabla p + \frac{1}{Re}\Delta \boldsymbol{u}, \tag{13a}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{13b}$$

where $\boldsymbol{u}(\boldsymbol{x}, t)$ and $p(\boldsymbol{x}, t)$ are the velocity vector and pressure at position $\boldsymbol{x}$ and time t, and the scalar Re is the Reynolds number. We consider the classical problem of a flow past a cylinder in a 2D domain, which is well known to exhibit a Hopf bifurcation from a steady wake to periodic vortex shedding at a critical Reynolds number $Re_c \sim 40$ (Jackson, 1987). For our study, we focus on the range $Re \in [10, 110]$, which makes the estimation problem very challenging since this range includes the bifurcation and therefore comprises solution trajectories with very different dynamics – steady for $Re < Re_c$, periodic limit cycle for $Re > Re_c$. Furthermore, the shedding frequency and spacing between consecutive vortices in the limit cycle regime both vary with Re.

We solve the Navier-Stokes equations with the open source finite volume code OpenFOAM using a mesh consisting of 18840 nodes and a second-order implicit scheme with time step 0.05. The discrete-time state vector $\boldsymbol{z}_k \in \mathbb{R}^{37680}$ contains the two velocity components of $\boldsymbol{u}$ at discrete time steps $t = k\Delta t$, where we choose $\Delta t = 0.25$. To generate the training dataset $\boldsymbol{Z}_{\text{train}} = \{\boldsymbol{Z}^{Re}\}_{Re \in S}$ for constructing the ROM and training the RL-ROE, we run simulations of the Navier-Stokes equations for $Re \in S = \{10, 20, 30, \ldots, 110\}$. We discard the transient portion of the dynamics (for the cases $Re > Re_c$) and save 201 snapshots $\boldsymbol{Z}^{Re} = \{\boldsymbol{z}_0^{Re}, \ldots, \boldsymbol{z}_{200}^{Re}\}$ in the post-transient regime. We retain $r = 20$ modes when constructing the ROM, corresponding to a three-orders-of-magnitude reduction in the dimensionality of the system. These modes are shown in Appendix D. We train the RL-ROE using episodes of length $K = 200$ steps to make full use of the trajectories stored in $\boldsymbol{Z}_{\text{train}}$ and end the training process when the return no longer increases on average. The RL hyperparameters and learning curve displaying the performance improvement of the RL-ROE during the training process are reported in Appendix D.

The trained RL-ROE and the KF-ROE are now compared based on their ability to track reference trajectories corresponding to various Re using sparse measurements of $\boldsymbol{u}$ from a limited number p of sensors randomly distributed in the wake region behind the cylinder. The evaluation is carried out using 5 initial state estimates sampled from a Gaussian distribution with unit standard deviation. Beginning with $p = 3$ sensors, Figure 1 reports the mean (lines) and standard deviation (shaded areas) of the normalized L_2 error for 3 values of Re not present in the training dataset $\boldsymbol{Z}_{\text{train}}$ used to construct the ROM and train the RL-ROE. The RL-ROE is almost as accurate as the lower bound, while the KF-ROE has normalized error around 1, which indicates complete failure. The velocity magnitude of the reference solutions for the same 3 values of Re and the corresponding RL-ROE and KF-ROE reconstructed high-dimensional estimates are shown at $t = 50$ in Figure 6.

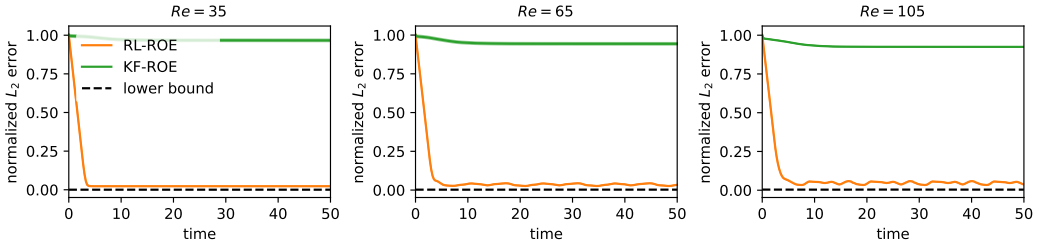

Figure 5: Navier-Stokes equation with $p = 3$ sensors. Normalized L_2 error of the RL-ROE and KF-ROE for the estimation of trajectories corresponding to values of Re not seen during training.

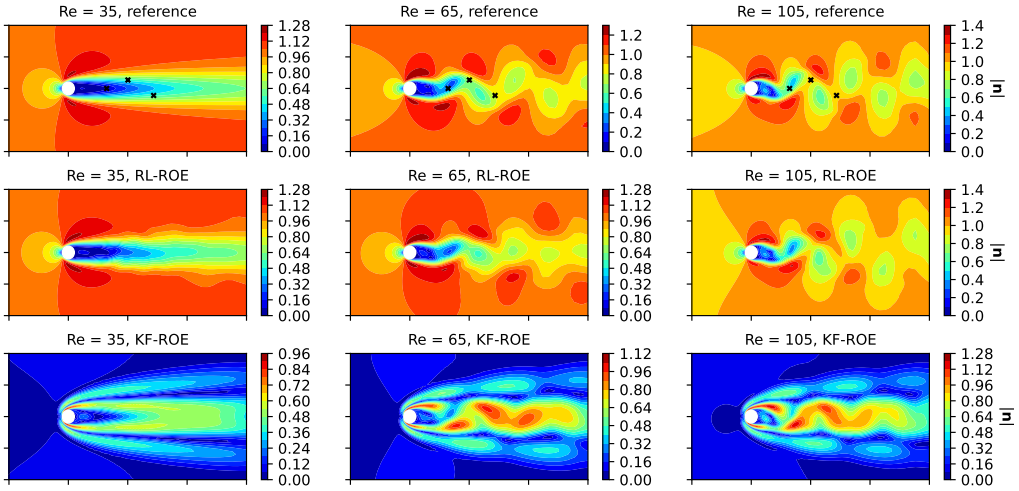

Figure 6: Flow past a cylinder with $p = 3$ sensors. Velocity magnitude at $t = 50$ of the reference trajectories for values of Re not seen during training and corresponding RL-ROE and KF-ROE estimates. The black crosses in the contours of the reference solutions indicate the sensor locations.

Remarkably, the RL-ROE is manages to estimate very precisely the entire flow field across different dynamical regimes, with the steady wake at $Re = 35$ being reproduced equally well as the wake of vortices at $Re = 65$ and 105. The KF-ROE, on the other hand, struggles to estimate the flow fields for all 3 Reynolds numbers, and instead predicts a velocity field that is almost everywhere zero except in the wake. Again, the superiority of the RL-ROE is granted by the nonlinearity of its policy – in fact, note that bifurcations such as the one exhibited by this flow are inherently nonlinear phenomena.

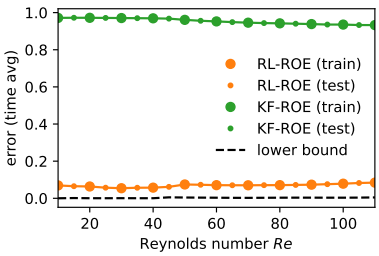

Figure 7: Flow past a cylinder with $p = 3$ sensors. Time average of the normalized L_2 error versus μ. Values of μ present in $\boldsymbol{Z}_{\text{train}}$ shown by large circles.

Figure 7 reports the time average of the normalized L_2 error as a function of μ for $p = 3$, with the values present in $\boldsymbol{Z}_{\text{train}}$ indicated by large circles. The RL-ROE exhibits robust performance across the entire range of Reynolds numbers, including in the vicinity of the bifurcation at $Re_c \sim 40$ as well as when estimating trajectories unseen during the training.

Finally, Figure 8 displays the average over time and over Re of the normalized L_2 error for varying number p of sensors. We observe the same trends as in the case of the Burgers equation.

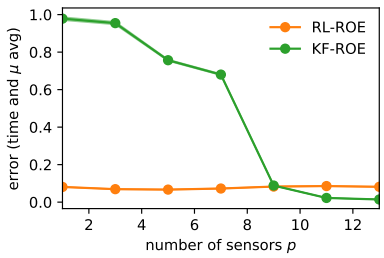

Figure 8: Flow past a cylinder. Average over time and over μ of the normalized L_2 error versus number p of sensors.

Although the KF-ROE eventually becomes very accurate in the presence of a large number p of sensors, the accuracy of the RL-ROE remains remarkably stable as p decreases, allowing it to vastly outperforms the KF-ROE as soon as $p < 8$. Once again, this showcases the benefits of using a nonlinear policy to correct the estimator dynamics (3a).

5 Related work

Previous studies have already proposed designing state estimators using policies trained through reinforcement learning. Morimoto & Doya (2007) introduced an estimator of the form $\hat{\boldsymbol{x}}_k = \boldsymbol{f}(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{L}(\hat{\boldsymbol{x}}_{k-1})(\boldsymbol{y}_{k-1} - \boldsymbol{C}\hat{\boldsymbol{x}}_{k-1})$, where $\boldsymbol{f}(\cdot)$ is the state-transition model of the system, and the state-dependent filter gain matrix $\boldsymbol{L}(\hat{\boldsymbol{x}}_{k-1})$ is defined using Gaussian basis functions whose parameters are learned through a variant of vanilla policy gradient. Hu et al. (2020) proposed an estimator of the form $\hat{\boldsymbol{x}}_k = \boldsymbol{f}(\hat{\boldsymbol{x}}_{k-1}) + \boldsymbol{L}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{f}(\hat{\boldsymbol{x}}_{k-1}))$, where $\boldsymbol{L}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)$ is approximated by neural networks trained with a modified Soft-Actor Critic algorithm (Haarnoja et al., 2018). Although they derived convergence properties for the estimate error, the dependence of the filter gain $\boldsymbol{L}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)$ on the reference state $\boldsymbol{x}_k$ limits its practical application. Furthermore, a major difference between these past studies and our work is that they only consider low-dimensional systems with four state variables at most, and scaling up their methodology to systems governed by PDEs would be computationally challenging. Our RL-ROE, on the other hand, handles parametric PDEs described by tens of thousands of state variables as shown in the previous section.

In another line of works, reinforcement learning has been applied to learn control policies for the torques at the joints of simulated characters that enable the characters to imitate given reference motions consisting of a sequence of target poses; see e.g. Peng et al. (2018) or Lee et al. (2019). At first sight, these trajectory tracking problems are reminiscent of ours since the goal of the policy is to drive the pose of the character towards one defined by the reference motion. Similar to us, these studies also propose to transform the problem into a stationary MDP by augmenting the state; but they do so by appending a scalar phase variable $\phi \in [0, 1]$ that represents the normalized time elapsed in the reference motion. The reward can then be formulated in terms of $q(\phi)$, where $q(\cdot)$ describes the reference motion, and no explicit time dependence appears. However, this approach would not work for our purpose since the RL-ROE cannot be restricted to estimating a single or a small finite number of reference trajectories. Instead, by augmenting the state with entire snapshots from the reference trajectory instead of just a phase variable, we ensure that there is no theoretical limit to the number of reference trajectories that the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ can learn.

6 Conclusions

In this paper, we have introduced the reinforcement learning reduced-order estimator (RL-ROE), a new methodology for estimating the state dynamical systems governed by parametric PDEs. Our approach relies on the construction of a computationally inexpensive reduced-order model (ROM) to approximate the dynamics of the system. The novelty of our contribution lies in the design, based on this ROM, of a reduced-order estimator (ROE) in which the feedback correction term is given by a nonlinear stochastic policy trained through reinforcement learning. To be able to use off-the-shelf RL algorithms, we introduce a trick to translate this time-dependent trajectory tracking problem to an equivalent stationary MDP based on an augmented state. We demonstrate using simulations of the Burgers and Navier-Stokes equations that in the limit of very few sensors, the trained RL-ROE vastly outperforms a Kalman filter designed using the same ROM, which is attributable to the nonlinearity of its policy. Finally, the RL-ROE also yields accurate high-dimensional state estimates for reference trajectories corresponding to various parameter values without direct knowledge of the latter.

REFERENCES

Shady E Ahmed, Suraj Pawar, Omer San, Adil Rasheed, Traian Iliescu, and Bernd R Noack. On closures for reduced order models—a spectrum of first-principle to machine-learned avenues. *Physics of Fluids*, 33(9):091301, 2021.

Alexandre Barbagallo, Denis Sipp, and Peter J Schmid. Closed-loop control of an open cavity flow using reduced-order models. *Journal of Fluid Mechanics*, 641:1–50, 2009.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Steven L Brunton and Bernd R Noack. Closed-loop turbulence control: Progress and challenges. *Applied Mechanics Reviews*, 67(5), 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 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.

Liang Hu, Chengwei Wu, and Wei Pan. Lyapunov-based reinforcement learning state estimator. *arXiv preprint arXiv:2010.13529*, 2020.

CP Jackson. A finite-element study of the onset of vortex shedding in flow past variously shaped bodies. *Journal of fluid Mechanics*, 182:23–45, 1987.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

J Nathan Kutz, Steven L Brunton, Bingni W Brunton, and Joshua L Proctor. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

Seunghwan Lee, Moonseok Park, Kyoungmin Lee, and Jehee Lee. Scalable muscle-actuated human simulation and control. *ACM Transactions On Graphics (TOG)*, 38(4):1–13, 2019.

Jun Morimoto and Kenji Doya. Reinforcement learning state estimator. *Neural computation*, 19(3): 730–756, 2007.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

Clarence W Rowley and Scott TM Dawson. Model reduction for flow analysis and control. *Annual Review of Fluid Mechanics*, 49:387–417, 2017.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Kunihiko Taira, Steven L Brunton, Scott TM Dawson, Clarence W Rowley, Tim Colonius, Beverley J McKeon, Oliver T Schmidt, Stanislav Gordeyev, Vassilios Theofilis, and Lawrence S Ukeiley. Modal analysis of fluid flows: An overview. *Aiaa Journal*, 55(12):4013–4041, 2017.

Jonathan H Tu, Clarence W Rowley, Dirk M Luchtenburg, Steven L Brunton, and J Nathan Kutz. On dynamic mode decomposition: theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014.

Paul Zarchan. *Progress in astronautics and aeronautics: fundamentals of Kalman filtering: a practical approach*, volume 208. Aiaa, 2005.

A Dynamic Mode Decomposition

In this appendix, we describe the DMD algorithm (Schmid, 2010; Tu et al., 2014), which is a popular data-driven method to extract spatial modes and low-dimensional dynamics from a dataset of high-dimensional snapshots. Here, we use the DMD to construct a ROM of the form (2) given an observation model $\boldsymbol{C}$ and a concatenated collection of snapshots $\boldsymbol{Z}_{\text{train}} = \{\boldsymbol{Z}^{\mu}\}_{\mu \in S}$, where each $\boldsymbol{Z}^{\mu} = \{\boldsymbol{z}_0^{\mu}, \dots, \boldsymbol{z}_m^{\mu}\}$ contains snapshots from a trajectory of (1a) for a specific value μ.

Fundamentally, the DMD seeks a best-fit linear model of the dynamics in the form of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{z}_{k+1} \simeq \boldsymbol{A}\boldsymbol{z}_k$. First, arrange the snapshots into two time-shifted matrices

$$\boldsymbol{X} = \{\boldsymbol{z}_0^{\mu_1}, \dots, \boldsymbol{z}_{m-1}^{\mu_1}, \dots, \boldsymbol{z}_0^{\mu_q}, \dots, \boldsymbol{z}_{m-1}^{\mu_q}\}, \tag{14a}$$

$$\boldsymbol{Y} = \{\boldsymbol{z}_1^{\mu_1}, \dots, \boldsymbol{z}_m^{\mu_1}, \dots, \boldsymbol{z}_1^{\mu_q}, \dots, \boldsymbol{z}_m^{\mu_q}\}, \tag{14b}$$

where q denotes the number of elements in S. The best-fit linear model is then given by $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{X}^{\dagger}$, where $\boldsymbol{X}^{\dagger}$ is the pseudoinverse of $\boldsymbol{X}$. The ROM is then obtained by projecting the matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ onto a basis $\boldsymbol{U}$ consisting of the r leading left singular vectors of $\boldsymbol{X}$, which approximate the r leading PCA modes of $\boldsymbol{Z}$. Using the truncated singular value decomposition

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}} \tag{15}$$

where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$, the resulting reduced-order state-transition and observation models are given by

$$\boldsymbol{A}_r = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{Y}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}, \tag{16a}$$

$$\boldsymbol{C}_r = \boldsymbol{C}\boldsymbol{U}. \tag{16b}$$

Conveniently, the ROM matrices $\boldsymbol{A}_r$ and $\boldsymbol{C}_r$ can be calculated directly from the truncated SVD of $\boldsymbol{X}$, which avoids forming the large $n \times n$ matrix $\boldsymbol{A}$.

B Kalman filter

The time-dependent Kalman filter that we use as a benchmark in this paper, KF-ROE, is based on the same ROM (2) as the RL-ROE, with identical matrices $\boldsymbol{A}_r$, $\boldsymbol{C}_r$ and $\boldsymbol{U}$. Similarly to the RL-ROE, the reduced-order estimate $\hat{\boldsymbol{x}}_k$ is given by equation (3a), from which the high-dimensional estimate is reconstructed as $\hat{\boldsymbol{z}}_k = \boldsymbol{U}\hat{\boldsymbol{x}}_k$. However, the KF-ROE differs from the RL-ROE in its definition of the action $\boldsymbol{a}_k$ in (3a), which is instead given by the linear feedback term (4). The calculation of the optimal Kalman gain $\boldsymbol{K}_k$ in (4) requires the following operations at each time step:

$$\boldsymbol{P}_k^{-} = \boldsymbol{A}_r\boldsymbol{P}_{k-1}\boldsymbol{A}_r^{\mathsf{T}} + \boldsymbol{Q}_k, \tag{17}$$

$$\boldsymbol{S}_k = \boldsymbol{C}_r\boldsymbol{P}_k^{-}\boldsymbol{C}_r^{\mathsf{T}} + \boldsymbol{R}_k, \tag{18}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_k^{-}\boldsymbol{C}_r^{\mathsf{T}}\boldsymbol{S}_k^{-1}, \tag{19}$$

$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C}_r)\boldsymbol{P}_k^{-}, \tag{20}$$

where $\boldsymbol{P}_k^{-}$ and $\boldsymbol{P}_k$ are respectively the a priori and a posteriori estimate covariance matrices, $\boldsymbol{S}_k$ is the innovation covariance, and $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ are respectively the covariance matrices of the process noise $\boldsymbol{w}_k$ and observation noise $\boldsymbol{v}_k$ in the ROM (2). We tune these noise covariance matrices to yield the best possible results; this resulted in $\boldsymbol{Q}_k = 10^3\boldsymbol{I}$ and $\boldsymbol{R}_k = \boldsymbol{I}$ for Burgers, and $\boldsymbol{Q}_k = 10^9\boldsymbol{I}$ and $\boldsymbol{R}_k = \boldsymbol{I}$ for Navier-Stokes. At time step $k = 0$, the a posteriori estimate covariance is initialized as $\boldsymbol{P}_0 = \text{cov}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{z}_0 - \hat{\boldsymbol{x}}_0)$, which can be calculated from the distributions (9).

C TRAINING HYPERPARAMETERS AND LEARNING CURVES

For both the Burgers and Navier-Stokes examples, the stochastic policy $\boldsymbol{\pi_\theta}$ is trained with PPO using the default hyperparameters from Stable Baselines3, except for the discount factor γ which we choose as 0.75. The mean output of the stochastic policy and the value function are approximated by two neural networks, each containing two hidden layers with 64 neurons and tanh activation functions. The input to the policy is normalized using a running average and standard deviation during the training process, which alternates between sampling data for 10 trajectories (of length 200 timesteps each) and updating the policy. Each policy update consists of multiple gradients steps through the most recent data using 10 epochs, a minibatch size of 64 and a learning rate of 0.0003. The policy is trained for a total of one to three million timesteps, corresponding to 5000 to 15000 trajectories, which depending on the dimensionality of the ROM takes between 15 min and an hour on a Core i7-12700K CPU. Figure 9 reports the learning curves corresponding to the unforced and forced cases. During training, the policy is tested (with stochasticity switched off) after each update using 20 separate test trajectories, and is saved if it outperforms the previous best policy. Finally, the RL-ROE is assigned the latest saved policy upon ending of the training process, and the stochasticity of the policy is switched off during subsequent evaluation of the RL-ROE.

D MODES FOR FLOW PAST A CYLINDER

The first 18 modes (i.e. columns of $\boldsymbol{U}$) of the 20-dimensional ROM for the flow past a cylinder are displayed in Figure 10.

E ADDITIONAL RESULTS ON THE BURGERS EQUATION

Figure 11 shows spatio-temporal contours of the reference solution and corresponding estimates, for $p = 2$ and 12.

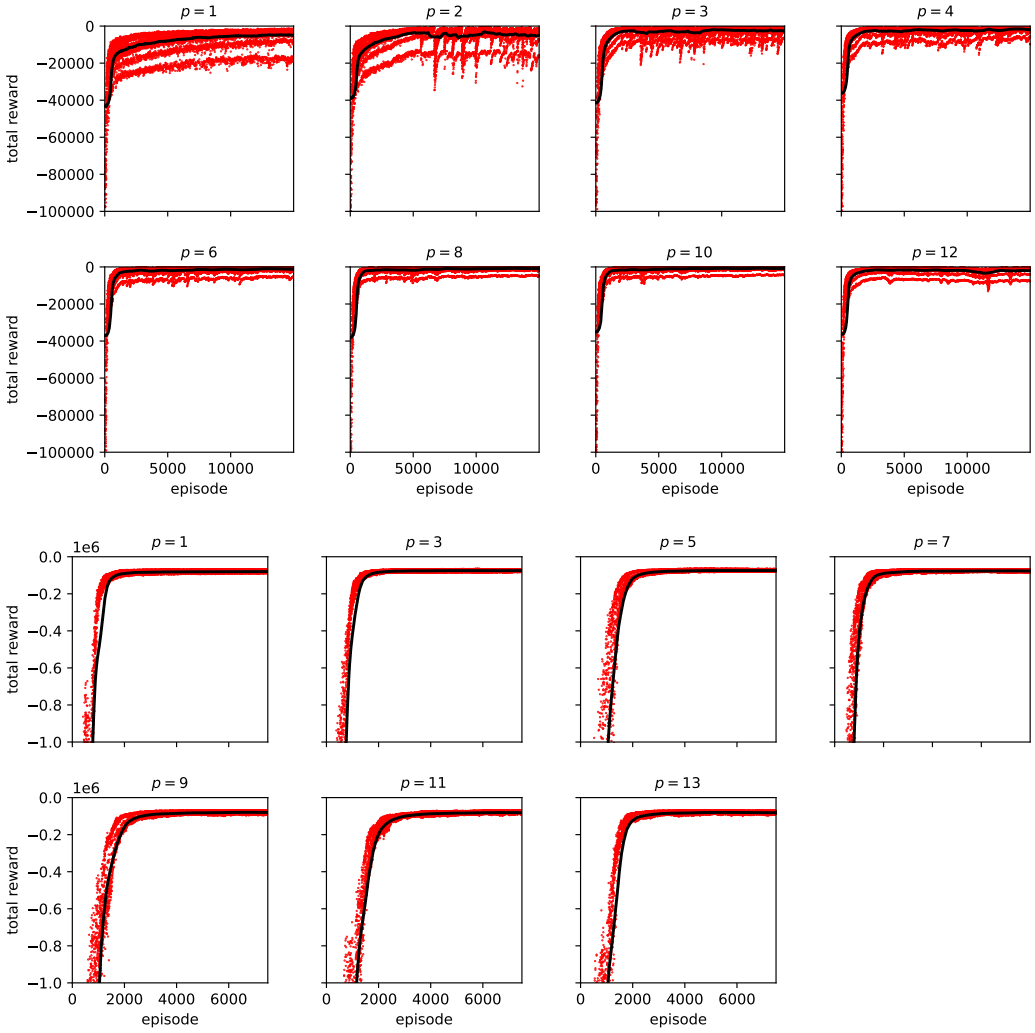

Figure 9: Learning curves for the stochastic policy of the RL-ROE for the Burgers (top 2 rows) and Navier-Stokes (bottom 2 rows) cases. Each plot corresponds to a different number p of sensor measurements. The line and shaded area show the mean and standard deviation of the results over 10 runs, each one smoothed with a moving average of size 100 episodes.

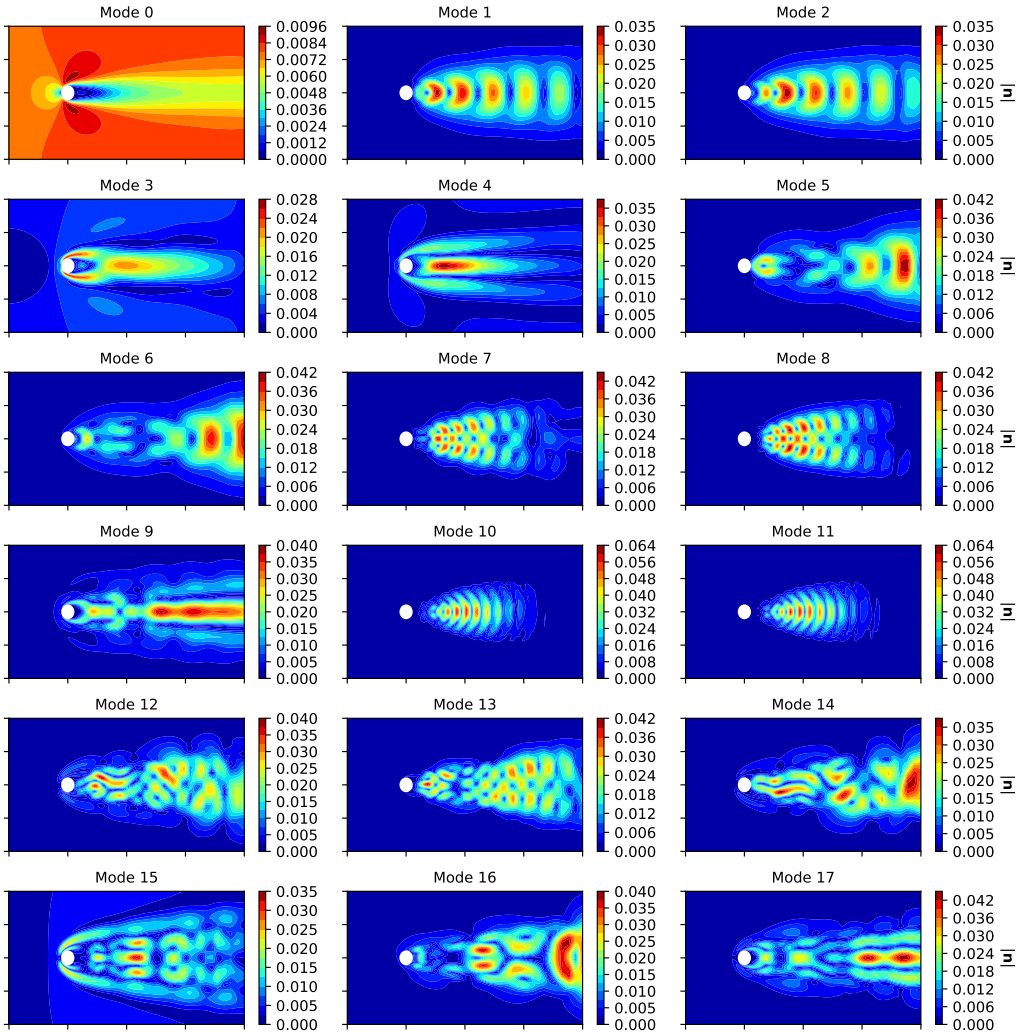

Figure 10: First 18 modes of the 20-dimensional ROM for the flow past a cylinder.

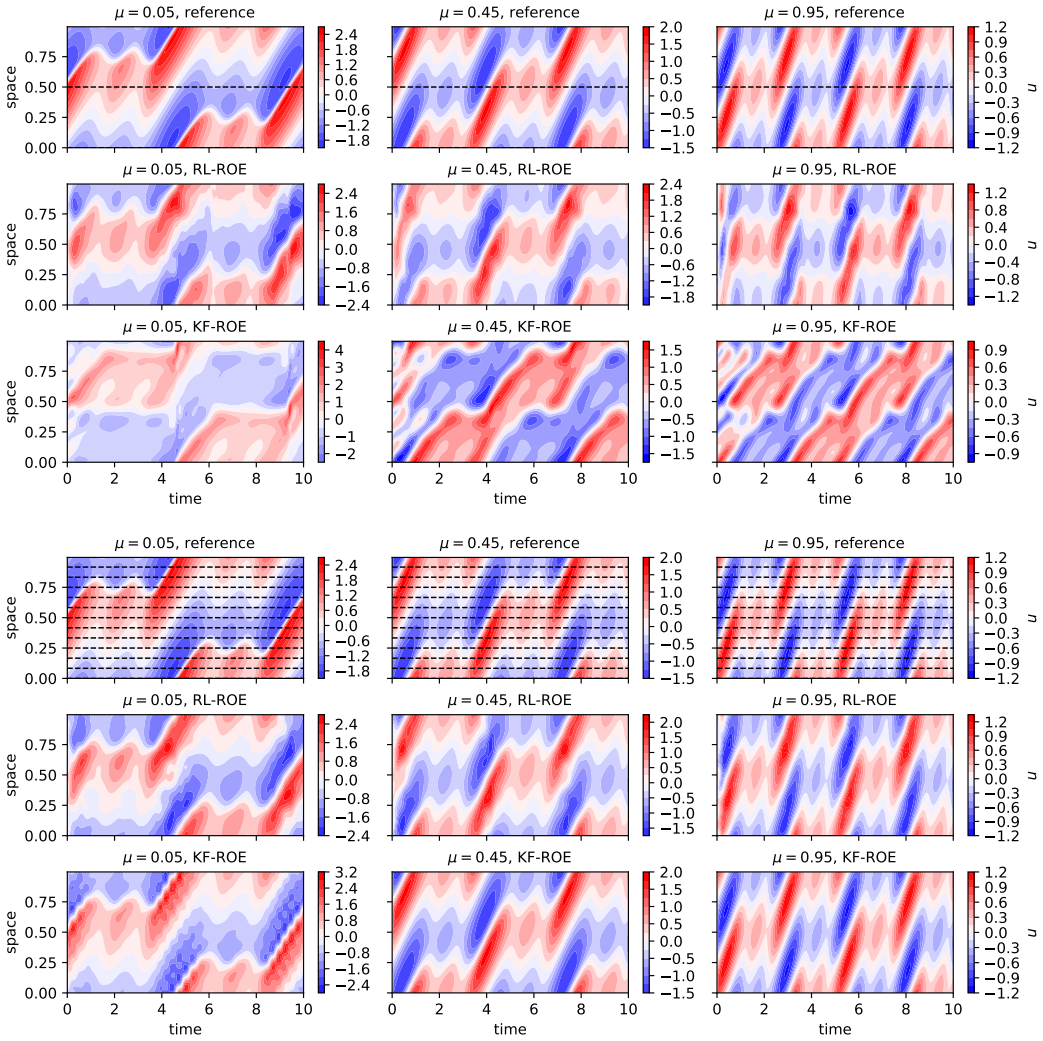

Figure 11: Burgers equation with $p = 2$ (top 3 rows) and $p = 12$ (bottom 3 rows) sensors. Reference trajectories for values of μ not seen during training and corresponding RL-ROE and KF-ROE estimates. The dashed lines on the reference trajectory plots indicate the sensor data seen by the RL-ROE and KF-ROE.