Acta Numerica (2022), pp. 265–345 Printed in the United Kingdom
doi:10.1017/S0962492922000058

Reduced basis methods for time-dependent problems

Jan S. Hesthaven
Ecole Polytechnique Federale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
E-mail: jan.hesthaven@epfl.ch

Cecilia Pagliantini
Eindhoven University of Technology,
5600 MB Eindhoven, Netherlands
E-mail: c.pagliantini@tue.nl

Gianluigi Rozza
SISSA – International School for Advanced Studies,
34136 Trieste, Italy
E-mail: gianluigi.rozza@sissa.it

Numerical simulation of parametrized differential equations is of crucial importance in the study of real-world phenomena in applied science and engineering. Computational methods for real-time and many-query simulation of such problems often require prohibitively high computational costs to achieve sufficiently accurate numerical solutions. During the last few decades, model order reduction has proved successful in providing low-complexity high-fidelity surrogate models that allow rapid and accurate simulations under parameter variation, thus enabling the numerical simulation of increasingly complex problems. However, many challenges remain to secure the robustness and efficiency needed for the numerical simulation of nonlinear time-dependent problems. The purpose of this article is to survey the state of the art of reduced basis methods for time-dependent problems and draw together recent advances in three main directions. First, we discuss structure-preserving reduced order models designed to retain key physical properties of the continuous problem. Second, we survey localized and adaptive methods based on nonlinear approximations of the solution space. Finally, we consider data-driven techniques based on non-intrusive reduced order models in which an approximation of the map between parameter space and coefficients of the reduced basis is learned. Within each class of methods, we describe different approaches and provide a comparative discussion that lends insights to advantages, disadvantages and potential open questions.

CONTENTS

1. Introduction

In numerous applications we are faced with the need to sample a large number of instances of a parametrized problem, often described by a parameter-dependent differential equation; this occurs in the context of optimization, control or uncertainty quantification, for example. In such models, found across all areas of science and engineering, the parameters can be used to characterize variations in geometric configurations, physical properties, and initial or boundary conditions.

The direct numerical simulation of the parametrized differential equations can be achieved by standard discretization methods such as the finite element method, the spectral method or finite volume methods. However, such models typically require the use of thousands of degrees of freedom to obtain sufficiently accurate solutions, thus leading to unmanageable demands on computational resources. This is not only exacerbated in the multi-query context discussed above but also in cases where real-time simulations are needed. In such cases there is a need to develop surrogate models that allow efficient evaluation of an output of interest when the input parameters are being varied, albeit without sacrificing the accuracy and fidelity of the model. In other words we are not looking for models of reduced accuracy in the physical description but for models that allow for rapid evaluation of the implicit connection between the parameter(s) and the output of interest.

Model order reduction (MOR) provides an effective way to reduce the computational cost of large-scale simulations by replacing the original high-dimensional problem with models that are faster to simulate yet accurately represent the original solution behaviour. The fundamental trade-off is one of obtaining accelerated models at the cost of a reduced generality of the surrogate model.

In this article we focus on a particular class of reduced order models, known as reduced basis methods (RBMs), which seek to construct low-dimensional approximations of parametrized solutions to partial differential equations by identifying a suitable problem-dependent basis from a collection of full order solutions at sampled values of time and parameters. A problem of lower dimension, called the reduced model, results from the projection of the original equation into the subspace spanned by this basis.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Early works based on this idea date back to the 1970s, in the context of many-query design evaluation (Fox and Miura 1971) and parameter continuation methods for nonlinear problems (Almroth, Stern and Brogan 1978, Noor and Peters 1980, Noor 1981, 1982). Extensions to parametrized finite-dimensional systems and certain classes of differential equations were proposed by Fink and Rheinboldt (1983, 1984), Noor, Balch and Shibut (1984), Porsching and Lee (1987), Peterson (1989) and Gunzburger (1989).

Subsequently the reduced basis method was put on a more sound mathematical footing and considerable improvements in the computational performances were achieved. In particular, substantial effort was devoted to the development of *a posteriori* error estimation and rigorous error bounds that allow the certification of the output of the reduced model for any parameter value (Prud'homme *et al.* 2002, Veroy, Prud'homme, Rovas and Patera 2003, Grepl and Patera 2005). See also Hesthaven, Rozza and Stamm (2015) and Quarteroni, Manzoni and Negri (2016) for an overview of the development of such techniques for a variety of different applications in science and engineering.

The main advantage of an *a posteriori* error estimator is the ability to determine the reliability of the output and to allow efficient sampling strategies across the parameter space in a greedy approach to the construction of the reduced basis. This has been investigated in several works (LeGresley and Alonso 2000, Ravindran 2000, Christensen, Brøns and Sørensen 1999, Willcox and Peraire 2002, Kunisch and Volkwein 2002, Bui-Thanh, Damodaran and Willcox 2003, Cuong, Veroy and Patera 2005, Rozza 2005, Gunzburger, Peterson and Shadid 2007, Patera and Rozza 2007, Rozza, Huynh and Patera 2008) and has led to efficient criteria for the selection of the basis functions.

A computational breakthrough was the development of a full decoupling of the reduced basis computational procedure into a parameter-independent offline phase and a parameter-dependent online phase. In this splitting, the complexity of the offline phase depends on the complexity of the numerical discretization of the parametrized differential equation, which is potentially very high, while the complexity of the online stage depends solely on the complexity of the reduced order model. While this offline–online decomposition is natural in the case of affine parameter dependence, the development of decoupling strategies in the non-affine case has been addressed via so-called hyper-reduction strategies; see Section 3.3 for a non-exhaustive literature overview.

During the last two decades, the development of reduced order models for elliptic, stationary and linear problems has reached a considerable level of maturity, as reflected by an extensive literature; see the reviews by Prud'homme *et al.* (2002), Rozza *et al.* (2008), Quarteroni, Rozza and Manzoni (2011), Rozza (2014) and Rozza, Huynh, Nguyen and Patera (2009), recent books by Hesthaven *et al.* (2015) and Quarteroni *et al.* (2016), and references therein.

Although these developments have paved the way for the use of reduced basis methods suitable for real-time and many-query simulations of complex problems,

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

the treatment of time-dependent non-coercive nonlinear equations still poses essential challenges. These are related to the lack of guarantees on the stability of the reduced order solutions, on the certification and reliability of the output, and on the efficiency of the reduced models. Indeed, for time-dependent problems, even in the parabolic case, the exponential growth of *a priori* and *a posteriori* error bounds limits their use to modest temporal intervals.

Moreover, independently of the accuracy of the reduced model, a classical projection-based RBM is not guaranteed to yield stable approximate solutions, even if the original high-dimensional scheme is stable. This lack of stability often results in spurious oscillations of the solution, blow-up of the system energy and violation of the conservation laws and invariants of the problem. The reason is that there is generally no relationship between the physics of the full order problem and that of the reduced model. This is problematic, especially for hyperbolic systems and conservative processes.

Finally, the success of an RBM relies on the assumption that the set of solutions, obtained as time and parameters vary, is intrinsically of low dimension. However, several important phenomena, such as convection-dominated problems, wave-type equations and conservation laws, do not generally exhibit such global low-dimensional structure. This implies that traditional reduced models derived via linear approximations are generally not effective. Nonlinear and local model order reduction techniques are then required to achieve sufficiently accurate solutions at manageable computational costs. Additionally, the upsurge of data-driven techniques has suggested the possibility of highly efficient reduced basis methods for general nonlinear problems.

In recent years considerable effort has been made to address these challenges by developing accurate, efficient and stable reduced basis methods for time-dependent nonlinear problems. Although some of the developed strategies are still in their early stages of development, promising first steps have been taken and various directions have been pursued. The purpose of this article is to draw together and highlight these recent advances.

Before discussing these developments in more detail, let us also briefly discuss a couple of alternative ideas that have recently been successful and offer some additional insight.

One approach to recovering efficient reduced order models for the time-dependent problem is to apply a Laplace transform in time and rely on the development of reduced basis methods for harmonic problems (Chen, Hesthaven, Maday and Rodríguez 2009, 2010, Hesthaven *et al.* 2015). This has been pursued with success by Huynh, Knezevic and Patera (2011) and Bigoni and Hesthaven (2020). However, the computational robustness and accuracy of the inverse Laplace transform remains a substantial challenge for this approach, in particular for long time intervals.

The temporal instability of the reduced models is often associated with the truncation of the reduced model, resulting in an accumulation of errors and an insufficient small-scale dissipation. One way to look at this problem is as a closure

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

problem, akin to that found in turbulence modelling. This point of view has led to recent work focusing on the formulation of closure models for time-dependent reduced order models, based on the Mori–Zwanzig formulation (Gouasmi, Parish, and Duraisamy 2017, Wang, Ripamonti and Hesthaven 2020), delay-differential equations (Gupta and Lermusiaux 2021), and a substantial effort for closure models inspired by developments in fluid dynamics, reviewed by Ahmed *et al.* (2021). This line of work is still at an early stage of development but these initial results, even if largely of a heuristic and problem-dependent nature, offer some interesting ideas for future work to improve the stability of reduced models and, as demonstrated in Wang *et al.* (2020), enable the development of reduced order models with predictive accuracy beyond the training interval.

The remainder of this work is organized as follows. Sections 2 and 3 present a brief overview of the state of the art of reduced basis methods for parametrized stationary elliptic problems. Section 2 introduces parametrized elliptic stationary partial differential equations. Their model order reduction via reduced basis methods is the topic of Section 3, where we discuss reduced basis generation via proper orthogonal decomposition and greedy algorithms, affine decomposition and hyper-reduction strategies.

This sets the stage for Section 4 as the central part of the article, focusing on reduced basis methods for time-dependent problems. Basis generation via POD-greedy methods is presented in Section 4.1, while Section 4.2 is devoted to structure-preserving techniques to ensure efficient approximations that retain the physical properties of the original model. To deal with the local reducibility properties of many important phenomena, recent works on localized and nonlinear reduced order approximations are discussed in Section 4.3. Section 4.4 focuses on data-driven methods which seek to develop reduced basis methods that are both stable and, more importantly, highly efficient, even for general nonlinear problems. This is achieved by introducing non-intrusive reduced order models and a data-driven map between parameter space and coefficients of the reduced basis to reconstructed. A few concluding remarks and thoughts on open questions are offered in Section 5.

2. Parametrized partial differential equations

We begin by considering parametrized stationary partial differential equations. Let $\Omega \in \mathbb{R}^d$ be a (suitably regular) physical domain, where $d = 1$, 2 or 3 denotes the spatial dimension. Let us also introduce a closed parameter domain $\mathbb{P} \in \mathbb{R}^P$ for a parameter point, or vector, or P-tuple, $\mu = (\mu_1, \mu_2, \ldots, \mu_P)$. This parameter set could in principle include time, but we will in general consider time explicitly as a special parameter.

We consider real-valued field variables $u \colon \Omega \to \mathbb{R}^{d_v}$, where $d_v = 1$ in the scalar case and $d_v = d$ for vector-valued fields, and assume that the field variables belong to a Hilbert space $\mathbb{V}$ equipped with the inner product $(w, v)_{\mathbb{V}}$, for all $w, v, \in \mathbb{V}$, and

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

the induced norm $\|w\|_{\mathbb{V}} = \sqrt{(w,w)_{\mathbb{V}}}$, for all $w \in \mathbb{V}$. The parametric field variable is then defined as $u \equiv (u_1, \ldots, u_{d_v})$: $\mathbb{P} \to \mathbb{V}$, and $u(\mu)$ denotes the field for parameter value $\mu \in \mathbb{P}$.

2.1. *Parametrized stationary problems in weak form*

Let us introduce a parametrized stationary model problem in a variational form. The abstract formulation reads as follows: given $\mu \in \mathbb{P}$, find $u(\mu) \in \mathbb{V}$ such that

$$a(u(\mu), v; \mu) = f(v; \mu) \quad \text{for all } v \in \mathbb{V}, \tag{2.1}$$

where $a: \mathbb{V} \times \mathbb{V} \times \mathbb{P} \to \mathbb{R}$ is a parametric form, assumed to be bilinear with respect to the first two variables, and $f: \mathbb{V} \times \mathbb{P} \to \mathbb{R}$ is taken to be linear with respect to the first variable. Given a parametric linear form $\ell: \mathbb{V} \times \mathbb{P} \to \mathbb{R}$, one might be interested in a function of the solution of (2.1), namely,

$$s(\mu) = \ell(u(\mu); \mu).$$

Here $s: \mathbb{P} \to \mathbb{R}$ denotes an output of interest, and ℓ plays the role of a linear 'output' functional which links the input to the output through the field variable $u(\mu)$.

The well-posedness of the weak formulation (2.1), for all parameter values $\mu \in \mathbb{P}$, can be established by the Lax–Milgram theorem, under the following further assumptions:

- $a(\cdot, \cdot; \mu)$ is coercive and continuous for all $\mu \in \mathbb{P}$ with respect to the norm $\|\cdot\|_{\mathbb{V}}$, *i.e.* for every $\mu \in \mathbb{P}$ there exists a positive constant $\alpha(\mu) \geq \alpha > 0$ and a finite constant $\gamma(\mu) \leq \gamma < \infty$ such that

$$a(v, v; \mu) \geq \alpha(\mu) \|v\|_{\mathbb{V}}^2 \quad \text{and} \quad a(w, v; \mu) \leq \gamma(\mu) \|w\|_{\mathbb{V}} \|v\|_{\mathbb{V}} \tag{2.2}$$

  for all $w, v \in \mathbb{V}$;
- $f(\cdot; \mu)$ is continuous for all $\mu \in \mathbb{P}$ with respect to the norm $\|\cdot\|_{\mathbb{V}}$, *i.e.* for every $\mu \in \mathbb{P}$ there exists a constant $\delta(\mu) \leq \delta < \infty$ such that

$$f(v; \mu) \leq \delta(\mu) \|v\|_{\mathbb{V}} \quad \text{for all } v \in \mathbb{V}.$$

The coercivity and continuity constants of $a(\cdot, \cdot; \mu)$ over $\mathbb{V}$ are, respectively, defined as

$$\alpha(\mu) = \inf_{v \in \mathbb{V}} \frac{a(v, v; \mu)}{\|v\|_{\mathbb{V}}^2} \quad \text{and} \quad \gamma(\mu) = \sup_{w \in \mathbb{V}} \sup_{v \in \mathbb{V}} \frac{a(w, v; \mu)}{\|w\|_{\mathbb{V}} \|v\|_{\mathbb{V}}} \tag{2.3}$$

for every $\mu \in \mathbb{P}$.

Model order reduction may target a continuous partial differential equation in infinite-dimensional function spaces as in (2.1) or, more frequently, consider the numerical discretizations of a PDE.

2.2. *High-fidelity model*

Let us formulate the numerical approximation of the parametric weak formulation (2.1). We focus on conforming approximations, namely where the finite-

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

dimensional approximation space $\mathbb{V}_h$ is a subset of $\mathbb{V}$. For each $\mu \in \mathbb{P}$, the discrete problem consists in finding $u_h(\mu) \in \mathbb{V}_h$ such that

$$a(u_h(\mu), v_h; \mu) = f(v_h; \mu) \quad \text{for all } v_h \in \mathbb{V}_h, \tag{2.4}$$

and evaluating $s_h(\mu) = \ell(u_h(\mu); \mu)$. This problem is described as the *truth problem* or high-fidelity problem. Solving the truth problem (2.4) can be potentially very expensive since a large number of degrees of freedom $N = \dim(\mathbb{V}_h)$ might be needed to achieve a sufficiently accurate approximation of the continuous solution of (2.1), to ensure that the error $\|u(\mu) - u_h(\mu)\|_{\mathbb{V}}$ is acceptably small. Depending on the properties of the PDE and on the solver of choice, the operation count of the map $\mu \mapsto s_h(\mu)$ is of the order of $O(N^\alpha)$, for $\alpha \geq 1$, resulting in a large computational cost in the case of the need for evaluations at many parameter values. It is exactly this problem that reduced basis methods seek to address, albeit without sacrificing the accuracy of the truth solution.

3. Reduced basis methods for elliptic problems

The numerical models emerging from discretization techniques for PDEs, such as finite element, discontinuous Galerkin and spectral methods, are usually very high-dimensional. The complexity and computational costs associated with solving the full parametric problem for each new parameter value rule out a direct approach.

The reduced basis method aims at constructing surrogate models that can be evaluated at a considerably reduced computational cost, yet without sacrificing the predictive accuracy of the complex model. The method is based on a two-stage procedure, comprising an offline and an online phase. The potentially very costly offline phase consists in the construction of a low number, n, of basis functions able to approximate any solution of the parametrized problem to within a prescribed accuracy. This process requires an empirical exploration of the set of all solutions under variation of the parameter and can therefore be very costly as it involves the solution of at least n truth problems, each with N degrees of freedom. The online stage consists of a Galerkin projection of the full model onto the space spanned by the selected basis functions. During this phase the parameter space can be explored at a substantially reduced cost, ideally at a cost independent of N.

3.1. The reduced basis approximation

The goal of reduced basis methods is to approximate the solution of the parametrized problem for any arbitrary value of the parameter in $\mathbb{P}$ using a low number of basis functions. Let us introduce the notion of *solution manifold*[1] as the set of all solutions of the parametrized problem under variation of the parameters. The

[1] It is common practice to use the term 'manifold', although strictly speaking $\mathcal{M}$ may not be a manifold in the differential geometry sense.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

solution manifold of the full model (2.1) is defined as

$$\mathcal{M} = \{u(\mu) \mid \mu \in \mathbb{P}\} \subset \mathbb{V},$$

where each $u(\mu) \in \mathbb{V}$ satisfies the problem (2.1): find $u(\mu) \in \mathbb{V}$ such that

$$a(u(\mu), v; \mu) = f(v; \mu) \quad \text{for all } v \in \mathbb{V}.$$

For the lack of an analytic expression of the exact solution of the parametrized problem in many cases of interest, we consider an approximate solution $u_h(\mu) \in \mathbb{V}_h$ of the discretized problem (2.4), referred to as the *truth* or *high-fidelity solution*. We assume that a computational model is available to solve the truth problem, and approximate the exact solution at any required accuracy. This accuracy requirement also implies that the computational cost of evaluating the truth model may be very high. Throughout the discussion we assume that $\|u(\mu) - u_h(\mu)\|_{\mathbb{V}}$ can be made arbitrarily small for any given parameter value, $\mu \in \mathbb{P}$. This simply states that we assume that a computational model is available to solve the truth problem at sufficient accuracy.

Analogously to the continuous problem, we also introduce the discrete version of the solution manifold,

$$\mathcal{M}_h = \{u_h(\mu) \mid \mu \in \mathbb{P}\} \subset \mathbb{V}_h, \tag{3.1}$$

where each $u_h(\mu) \in \mathbb{V}_h$ is the solution of (2.4).

A key aspect in model order reduction and in the construction of any reduced model is the reducibility property of the solution manifold. To understand this concept, it is instructive to introduce the notion of the Kolmogorov n-width. The Kolmogorov n-width of $\mathcal{M}_h$ is defined as

$$d_n(\mathcal{M}_h) = \inf_{\substack{\mathbb{V}_r \subset \mathbb{V} \\ \dim(\mathbb{V}_r)=n}} \sup_{u_h \in \mathcal{M}_h} \inf_{v_r \in \mathbb{V}_r} \|u_h - v_r\|_{\mathbb{V}}. \tag{3.2}$$

Hence the n-width measures how well $\mathcal{M}_h$ can be approximated by some n-dimensional linear subspace $\mathbb{V}_r$. Rapid decay of the n-width with increasing n suggests a compact and efficient approximation across the entire parameter space. In this case the solution manifold can be well approximated by the span of a low number of appropriately chosen basis functions, which form the so-called *reduced basis*.

In some cases the Kolmogorov n-width may even decay exponentially, that is, $d_n(\mathcal{M}_h, \mathbb{V}_r) \leq C\mathrm{e}^{-cn}$, *e.g.* for elliptic problems of high regularity (Benner, Gugercin and Willcox 2015). However, for other cases, such as a moving front solution to a linear transport equation, $d_n(\mathcal{M}_h, \mathbb{V}_r) \leq Cn^{-1/2}$ (Ehrlacher, Lombardi, Mula and Vialard 2020), suggesting that a straightforward construction of the basis is less efficient as a large number of basis functions will be needed to ensure high accuracy.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Let us initially assume that an n-dimensional reduced basis, denoted by $\{\xi_\ell\}_{\ell=1}^n \subset \mathbb{V}_h$, is available. The associated reduced basis space is given by

$$\mathbb{V}_r = \operatorname{span}\{\xi_1, \ldots, \xi_n\} \subset \mathbb{V}_h,$$

and any function $u_r(\mu)$ in $\mathbb{V}_r$ can be represented as

$$u_r(\mu) = \sum_{\ell=1}^{n} \mathrm{u}_{r,\ell}(\mu)\, \xi_\ell, \quad \mu \in \mathbb{P},$$

where $\boldsymbol{u}_r(\mu) \coloneqq (\mathrm{u}_{r,1}(\mu), \ldots, \mathrm{u}_{r,n}(\mu)) \in \mathbb{R}^n$ denotes the vector of the expansion coefficients in the reduced basis. Given the n-dimensional space $\mathbb{V}_r$, the reduced basis approximation is sought as follows: for any given $\mu \in \mathbb{P}$, find $u_r(\mu) \in \mathbb{V}_r$ such that

$$a(u_r(\mu), v_r; \mu) = f(v_r; \mu) \quad \text{for all } v_r \in \mathbb{V}_r. \tag{3.3}$$

In the following, let us discuss the problem of generating accurate reduced basis spaces during the offline stage, and how to recover the reduced basis solution efficiently during the online stage.

3.2. *Reduced basis generation*

While there are several strategies for generating reduced basis spaces, we focus on proper orthogonal decomposition (POD) and the greedy construction in the following. Let $\mathbb{P}_h \subset \mathbb{P}$ denote a discrete and finite-dimensional point-set of M points in $\mathbb{P}$; for example, it can consist of a regular lattice or a randomly generated point-set intersecting with $\mathbb{P}$. We can then introduce the following set:

$$\mathcal{M}_h(\mathbb{P}_h) = \{u_h(\mu) \mid \mu \in \mathbb{P}_h\}.$$

It holds that $\mathcal{M}_h(\mathbb{P}_h) \subset \mathcal{M}_h$ as $\mathbb{P}_h \subset \mathbb{P}$ and, if $\mathbb{P}_h$ is fine enough, $\mathcal{M}_h(\mathbb{P}_h)$ is also a good representation of $\mathcal{M}_h$.

Proper orthogonal decomposition (POD)

Proper orthogonal decomposition is an explore-and-compress strategy to reduce the dimensionality of a given dataset by retaining the components capturing the most important information. In the context of reduced basis methods, the idea is to compute truth solutions at sample points in the parameter space, transform them into a set of uncorrelated variables, called POD modes, and retain only a few of them. It is worth recalling that POD corresponds to principal component analysis (Jolliffe 1986) in multivariate statistics, to the Karhunen–Loève decomposition (Karhunen 1947, Loève 1955) in the theory of stochastic processes, and to the Hotelling transformation (Hotelling 1933).

Assume we have an ordering $\mu_1, \ldots, \mu_M$ of the values in $\mathbb{P}_h$, hence inducing an ordering $u_h(\mu_1), \ldots, u_h(\mu_M)$ of the elements of $\mathcal{M}_h(\mathbb{P}_h)$. To construct the

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

POD-space, let us introduce the space $\mathbb{V}_{\mathcal{M}} = \text{span}\{u_h(\mu) \mid \mu \in \mathbb{P}_h\}$ that spans the elements of $\mathcal{M}_h(\mathbb{P}_h)$, and let us define the operator $\mathsf{C}\colon \mathbb{V}_{\mathcal{M}} \to \mathbb{V}_{\mathcal{M}}$,

$$\mathsf{C}(v_h) = \frac{1}{M}\sum_{m=1}^{M}(v_h, u_h(\mu_m))_{\mathbb{V}}\, u_h(\mu_m), \quad v_h \in \mathbb{V}_{\mathcal{M}}.$$

The operator C is compact, self-adjoint and linear, and there exists a $\mathbb{V}$-orthonormal set $\{\xi_\ell\}_{\ell=1}^{M'}$ of $M' \leq M$ eigenvectors with real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{M'} > 0$, so that

$$(\mathsf{C}(\xi_\ell), u_h(\mu_m))_{\mathbb{V}} = \lambda_\ell(\xi_\ell, u_h(\mu_m))_{\mathbb{V}}, \quad 1 \leq m \leq M. \tag{3.4}$$

The orthogonal POD basis functions are given by the eigenfunctions $\xi_1, \ldots, \xi_M$. The span of the truncated basis $\xi_1, \ldots, \xi_n$ gives the n-dimensional space $\mathbb{V}_{\text{POD}}$ that minimizes the quantity

$$\sqrt{\frac{1}{M}\sum_{\mu\in\mathbb{P}_h}\inf_{v_r\in\mathbb{V}_r}\|u_h(\mu) - v_r\|_{\mathbb{V}}^2} \tag{3.5}$$

over all n-dimensional linear subspaces $\mathbb{V}_r$ of $\mathbb{V}_{\mathcal{M}}$. Further, the projection $P_n\colon \mathbb{V} \to \mathbb{V}_{\text{POD}}$ defined as $(P_n f, \xi_\ell)_{\mathbb{V}} = (f, \xi_\ell)_{\mathbb{V}}$, for $1 \leq \ell \leq n$, is given by

$$P_n f = \sum_{\ell=1}^{n}(f, \xi_\ell)_{\mathbb{V}}\, \xi_\ell.$$

In particular, if the projection is applied to all elements in $\mathcal{M}_h(\mathbb{P}_h)$, it satisfies

$$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\|u_h(\mu_m) - P_n u_h(\mu_m)\|_{\mathbb{V}}^2} = \sqrt{\sum_{m=n+1}^{M}\lambda_m}.$$

Remark 3.1. The computational cost of solving the eigenvalue problem for the correlation operator is potentially very high. A more efficient computational procedure consists in reformulating it as an M-dimensional eigenvalue problem for the Gram matrix. Let $\mathbf{C} \in \mathbb{R}^{M\times M}$ be the correlation matrix defined as

$$\mathbf{C}_{m,q} = \frac{1}{M}(u_h(\mu_m), u_h(\mu_q))_{\mathbb{V}}, \quad 1 \leq m, q \leq M.$$

Then the problem of finding the n largest eigenvalue–eigenvector pairs $(\lambda_\ell, \mathrm{v}_\ell)$ with $\|\mathrm{v}_\ell\|_{\ell^2(\mathbb{R}^M)} = 1$ such that $\mathbf{C}\,\mathrm{v}_\ell = \lambda_\ell \mathrm{v}_\ell$, with $1 \leq \ell \leq n$, is equivalent to (3.4). With the eigenvalues sorted in descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, the orthogonal POD basis functions $\{\xi_1, \ldots, \xi_n\}$ span the POD-space $\mathbb{V}_{\text{POD}}$ and are given by the linear combinations

$$\xi_\ell = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}(\mathrm{v}_\ell)_m\, u_h(\mu_m), \quad 1 \leq \ell \leq n, \tag{3.6}$$

where $(\mathrm{v}_\ell)_m$ denotes the mth coefficient of the eigenvector $\mathrm{v}_\ell \in \mathbb{R}^M$.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Proper orthogonal decomposition allows the construction of a reduced basis that is optimal in an ℓ^2-sense over the parameter space. However, its main limitation lies in the large computational cost: computing the reduced basis requires the solution of a large dense eigenvalue problem that scales as $O(nN^2)$. Moreover, a potentially large number $M \gg n$ of high-fidelity solutions may be required to ensure a reduced basis of sufficient accuracy.

A proper choice of M is not known or predictable for a general problem. This could result in a substantial computational overhead associated with the fact that a large number of full order solves is required but the majority of the resulting solutions do not contribute to the reduced basis.

Greedy algorithm

An efficient alternative to proper orthogonal decomposition is to attempt to build the reduced basis using an iterative approach. Such a greedy generation of the reduced basis space consists in adding a new basis function at each iteration in such a way that the approximation properties of the updated reduced basis are improved. This strategy allows us to minimize the number of snapshots to be evaluated in the construction of the reduced space. Indeed, the greedy algorithm requires one truth solution to be computed per iteration and a total of n truth solutions to generate the n-dimensional reduced basis space.

To describe this in a bit more detail, at the ℓth step of the algorithm, assuming that an ℓ-dimensional reduced basis space $\mathbb{V}_r$ is given, the next basis function is the element of the solution set that is worst approximated by the current reduced space $\mathbb{V}_r$ over $\mathbb{P}$. That is, we select

$$\mu_{\ell+1} = \arg\max_{\mu \in \mathbb{P}} \epsilon_\ell(\mu), \tag{3.7}$$

where $\epsilon_\ell(\mu) \coloneqq \|u_h(\mu) - u_r(\mu)\|_{\mathbb{V}}$ and $u_r(\mu)$ is the solution of (3.3) for the current reduced basis space $\mathbb{V}_r$. Note that a different norm for ϵ_ℓ can be chosen, or even a different error quantity, such as the measure of the output functional.

The truth solution $u_h(\mu_{\ell+1})$ is then selected to enrich the reduced basis space as $\mathbb{V}_r = \operatorname{span}\{u_h(\mu_1), \ldots, u_h(\mu_{\ell+1})\}$. This is repeated until the maximal error is below a fixed tolerance, resulting in a hierarchical sequence of approximation spaces. Since the greedy algorithm selects the parameter for which the error is maximum, the resulting reduced basis seeks to be optimal in the maximum norm over $\mathbb{P}$ rather than in L^2 as in proper orthogonal decomposition.

Similarly to POD, computing the maximum in (3.7) over the entire parameter space $\mathbb{P}$ is unfeasible and we introduce a finite point-set $\mathbb{P}_h$. Moreover, to be efficient, a greedy algorithm must be supported by an error $\eta(\mu)$ that provides an estimate of the error induced by replacing $\mathbb{V}_h$ with the reduced basis space $\mathbb{V}_r$, *i.e.* $\|u_h(\mu) - u_r(\mu)\|_\mu \leq \eta(\mu)$, for all $\mu \in \mathbb{P}$. If the error estimator $\eta(\mu)$ can be evaluated efficiently, the computation of $\mu_{\ell+1}$ can be significantly accelerated since no truth solution is required and the evaluation of the error estimator is embarrassingly

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

parallel. This implies that the training set $\mathbb{P}_h$ can be considerably denser than the one used in the construction of the POD basis. Other ways to accelerate this process are discussed in Hesthaven, Stamm and Zhang (2014). Furthermore, a considerable reduction in the computational cost of the reduced basis construction is given by the fact that only n truth solutions are computed in the greedy approach, in contrast to the M solutions needed for the POD basis generation, where $M \gg n$ in almost all cases.

Convergence results for the greedy approximation have been established by Binev *et al.* (2011) and Buffa *et al.* (2012), and generalized to Banach spaces by DeVore, Petrova and Wojtaszczyk (2013). They show that if the underlying problem allows an efficient and compact reduced basis, the greedy reduced basis approximation converges exponentially fast to the truth.

Theorem 3.2. If $\mathcal{M}$ has an exponentially small Kolmogorov n-width $d_n(\mathcal{M}) \leq c\mathrm{e}^{-an}$, with $a > \log a_0$, then the reduced basis method built using the greedy algorithm converges exponentially fast with respect to n, that is, there exists $\beta > 0$ such that

$$\|u_h(\mu) - u_r(\mu)\|_{\mathbb{V}} \leq C\mathrm{e}^{-\beta n} \quad \text{for all } \mu \in \mathbb{P}.$$

The value of a_0 depends on the error $\epsilon_\ell(\mu)$ used in (3.7), and we find that

$$a_0 = 2 \qquad \text{if } \epsilon_\ell(\mu) = \|u_h(\mu) - P_\ell u_h(\mu)\|_{\mathbb{V}},$$

$$a_0 = 1 + \sqrt{\frac{\gamma}{\alpha}} \quad \text{if } \epsilon_\ell(\mu) = \|u_h(\mu) - u_r(\mu)\|_{\mathbb{V}},$$

where α and γ are the parameter-independent coercivity and continuity constants of the bilinear form a introduced in (2.2).

3.3. Affine decomposition and hyper-reduction

The computational efficiency of traditional reduced basis methods is tied to the feasibility of a complete decoupling of the offline and online phases. Indeed, projection-based reduced basis methods allow significant computational savings as long as the computational complexity of any online query is independent of the size of the full order problem. Ideally, the cost of computing $u_r(\mu)$ by solving the reduced problem (3.3) should be independent of the complexity of the truth problem and should depend only on the size $n \ll N$ of the reduced basis approximation.

However, since the parameter may enter the bilinear form $a(\cdot, \cdot; \mu)$, the linear terms $\ell(\cdot; \mu)$ and $f(\cdot; \mu)$, we generally need to first assemble the truth operators and then construct their approximation in the reduced basis space. This is a computation that depends on N and would severely limit the potential for rapid online evaluation. This limitation can be overcome under suitable assumptions on the problem operators, such as affine decomposability or parameter separability. The bilinear form $a(\cdot, \cdot; \mu)$ allows an affine decomposition if there exist coefficient functions $\theta_{\mathrm{a}}^q : \mathbb{P} \to \mathbb{R}$ for $q = 1, \ldots, Q_{\mathrm{a}}$, with $Q_{\mathrm{a}} \in \mathbb{N}$, and parameter-independent

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

continuous bilinear forms $a_q : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ such that

$$a(w, v; \mu) = \sum_{q=1}^{Q_{\mathrm{a}}} \theta_{\mathrm{a}}^q(\mu)\, a_q(w, v) \quad \text{for all } \mu \in \mathbb{P},\ w, v \in \mathbb{V}. \tag{3.8}$$

In this case the term associated to the parameter-independent form $a_q(\cdot, \cdot)$ can be precomputed, for every q, during the offline stage once the reduced basis space is known. Then, during the online stage, for any new instance of the parameter μ, only the factors $\{\theta_{\mathrm{a}}^q(\mu)\}_{q=1}^{Q_{\mathrm{a}}}$ need to be computed. This operation is independent of N and scales proportionally to $Q_{\mathrm{a}} n^2$. The treatment of the linear forms f and ℓ can be done similarly.

In the case of nonlinear operators, the affine parametric dependence is only a necessary condition to deliver computational efficiency. In the presence of low-order polynomial nonlinearities, tensorial techniques (Ştefănescu, Sandu and Navon 2014) can be used to restore the affine separability in the parameter by manipulating the order of computation of the various factors.

Dealing with general nonlinear terms remains one of the major computational bottlenecks in dimension reduction algorithms. The need for a further level of dimension reduction to deal with nonlinear operators has led to so-called *hyper-reduction* methods (Ryckelynck 2009). Most of these techniques consist in approximating the high-dimensional nonlinear terms using sparse sampling via interpolation among samples of the nonlinear operators. This is the rationale behind missing point estimation (Astrid, Weiland, Willcox and Backx 2008), the empirical interpolation method (EIM) (Barrault, Maday, Nguyen and Patera 2004, Grepl, Maday, Nguyen and Patera 2007), the discrete empirical interpolation method (DEIM) (Chaturantabut and Sorensen 2010), Gauss–Newton with approximated tensors (GNAT) (Carlberg, Farhat, Cortial and Amsallem 2013) and the trajectory piecewise linear (TPWL) method (Rewienski 2003). More recently, lifting transformations (Kramer and Willcox 2019) have been proposed, as well as data-driven approaches such as dynamic mode decomposition (DMD) (Schmid 2010, Kutz, Brunton, Brunton and Proctor 2016*a*) and operator inference (Peherstorfer and Willcox 2016*a*).

This problem is one we will return to in detail in Section 4.4, where ideas from machine learning will be introduced to address this bottleneck.

4. Reduced basis methods for time-dependent problems

In this section we consider reduced basis methods for parametrized evolution equations. We will focus on a model problem in an operator-based formulation rather than in a variational form, as this is more natural when considering numerical approximations such as finite difference and finite volume methods, usually not based on a variational formulation, and often advantageous for time-dependent

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

problems. However, for continuity with the discussion so far we shall also outline the equivalent variational form.

Let $\mathcal{T} = (t_0, T] \subset \mathbb{R}$ be a temporal interval. Consider the following model problem: given a parameter value $\mu \in \mathbb{P} \subset \mathbb{R}^P$, find $u(\mu) \in C^1(\mathcal{T}; \mathbb{V})$ such that

$$\partial_t u(\mu) + \mathcal{L}(u(\mu); \mu) = f(\mu) \quad \text{in } \Omega \times \mathcal{T}, \tag{4.1}$$

subject to initial condition $u(t_0; \mu) = u_0(\mu) \in \mathbb{V}$.

The corresponding variational statement is to find $u(\mu) \in C^0(\mathcal{T}; L^2(\Omega)) \cap L^2(\mathcal{T}; \mathbb{V})$ such that

$$(\partial_t u(\mu), v)_{L^2(\Omega)} + a(u(\mu), v; \mu) = g(t)h(v) \quad \text{for all } v \in \mathbb{V}, \tag{4.2}$$

subject to the initial condition $u(t_0; \mu) = u_0(\mu) \in L^2(\Omega)$. Here $g(t) \in L^2(\mathcal{T})$ is called the control function, and we assume that the right-hand side h is independent of the parameter, although this assumption can be relaxed. As for the elliptic case, we assume that the bilinear form $a(\cdot, \cdot; \mu)$ is coercive and continuous (2.2), satisfies the affine assumption (3.8) and is time-invariant. We let $(\cdot, \cdot)_{L^2(\Omega)}$ denote the $L^2(\Omega)$ scalar product.

As to the stationary case, we resort to an approximation of the problem (4.1) or (4.2). Let us consider a method of lines approach where the problem is first discretized in space and then in the temporal variable. Space–time discretization is also possible. The discrete solution manifold is given by the set of solutions at each instance of the parameter and of time, that is,

$$\mathcal{M}_h = \{u_h(t; \mu) \mid t \in \mathcal{T},\ \mu \in \mathbb{P}\} \subset \mathbb{V}_h, \tag{4.3}$$

where each $u_h(\mu) \in C^1(\mathcal{T}; \mathbb{V}_h)$ is the solution of the problem

$$\partial_t u_h(\mu) + \mathcal{L}_h(u_h(\mu); \mu) = f_h(\mu) \quad \text{in } \Omega \times \mathcal{T}, \tag{4.4}$$

and the initial condition $u_h(t_0, \mu)$ is given by a suitable projection of $u_0(\mu)$ onto $\mathbb{V}_h$. Here $\mathcal{L}_h$ and f_h are the discrete operators resulting from the spatial discretization of problem (4.1).

For the variational formulation the only difference is that $u_h(\mu) \in C^0(\mathcal{T}; L^2(\Omega)) \cap L^2(\mathcal{T}; \mathbb{V}_h)$ is the solution of the problem

$$(\partial_t u_h(\mu), v_h)_{L^2(\Omega)} + a(u_h(\mu), v_h; \mu) = g(t)h(v_h) \quad \text{for all } v_h \in \mathbb{V}_h, \tag{4.5}$$

subject to $(u_h(t_0, \mu), v_h)_{L^2(\Omega)} = (u_0(\mu), v_h)_{L^2(\Omega)}$, for all $v_h \in \mathbb{V}_h$. Reduced basis methods seek to approximate the solution of the parametrized time-dependent problem (4.4), or (4.5), for any arbitrary value of the parameter in $\mathbb{P}$ and at any time $t \in \mathcal{T}$ using a low number of basis functions.

Let us assume that an n-dimensional reduced basis, denoted by $\{\xi_\ell\}_{\ell=1}^n \subset \mathbb{V}_h$, is available. The associated reduced basis space is given by $\mathbb{V}_r = \text{span}\{\xi_1, \ldots, \xi_n\} \subset \mathbb{V}_h$, and any function $u_r(t; \mu)$ in $\mathbb{V}_r$ can be represented in terms of the reduced

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

basis as

$$u_r(t;\mu) = \sum_{\ell=1}^{n} \mathrm{u}_{r,\ell}(t;\mu)\,\xi_\ell, \quad t \in \mathcal{T},\ \mu \in \mathbb{P}, \tag{4.6}$$

where $\boldsymbol{u}_r(t;\mu) := (\mathrm{u}_{r,1}(t;\mu), \ldots, \mathrm{u}_{r,n}(t;\mu)) \in \mathbb{R}^n$ denotes the vector of the expansion coefficients in the reduced basis. Given the n-dimensional space $\mathbb{V}_r$, the reduced basis approximation is sought as follows: for any given $\mu \in \mathbb{P}$, find $u_r(\mu) \in C^1(\mathcal{T};\mathbb{V}_r)$ such that

$$\partial_t u_r(\mu) + \mathcal{L}_h(u_r(\mu);\mu) = f_h(\mu) \quad \text{in } \Omega \times \mathcal{T}, \tag{4.7}$$

where the initial condition $u_r(t_0,\mu)$ is given by the projection of $u_h(t_0,\mu)$ onto $\mathbb{V}_r$.

For the variational form, we seek $u_r(\mu) \in C^0(\mathcal{T};L^2(\Omega)) \cap L^2(\mathcal{T};\mathbb{V}_r)$ such that

$$(\partial_t u_r(\mu), v_r)_{L^2(\Omega)} + a(u_r(\mu), v_r;\mu) = g(t) f(v_r) \quad \text{for all } v_r \in \mathbb{V}_r,\ t \in \mathcal{T},$$

subject to $(u_r(t_0,\mu), v_r)_{L^2(\Omega)} = (u_h(t_0), v_r)_{L^2(\Omega)}$, for all $v_r \in \mathbb{V}_r$.

The offline–online procedure is now very similar to the steady case presented above. In the next subsection we discuss how to find the reduced basis space and how compact we can expect it to be, given available information about the Kolmogorov n-width.

4.1. POD-greedy approach

The simplest way to construct a reduced basis for a parametrized time-dependent problem is to treat time as an additional 'parameter'. Consider a partition of the time interval $\mathcal{T}$ into N_T subintervals of equal length $\Delta t = T/N_T$ and define $t_k = k\Delta t$, $0 \le k \le N_T$. Therefore, for each μ, the full order solution in the discretized time interval is given by the set $\{u_h(t_k,\mu)\}_{k=0}^{N_T} \subset \mathbb{V}_h$. We assume that Δt is sufficiently small and N is sufficiently large such that $u_h^k(\mu)$ is an arbitrarily accurate approximation of $u(t^k;\mu)$.

As in the steady case, we consider the discrete solution manifold

$$\mathcal{M}_h^{N_T} = \{u_h^k(\mu) \mid 0 \le k \le N_T,\ \mu \in \mathbb{P}\} \subset \mathbb{V}_h, \tag{4.8}$$

and seek a small representative basis thereof. We let $\mathbb{P}_h$ denote a finite sample of M points in $\mathbb{P}$, serving as training set for the construction of the reduced basis space. We introduce the set

$$\mathcal{M}_h^{N_T}(\mathbb{P}_h) = \{u_h^k(\mu) \mid 0 \le k \le N_T,\ \mu \in \mathbb{P}_h\}.$$

and generate an n-dimensional reduced basis space $\mathbb{V}_r$ by a sampling procedure which combines spatial snapshots in time and parameter space, as described below.

Proper orthogonal decomposition and greedy strategies (see Section 3.2) applied to the time-parameter snapshots from $\mathcal{M}_h^{N_T}(\mathbb{P}_h)$ may encounter difficulties. First, the time-parameter manifold might be very high-dimensional and characterized by extensive ranges of parameter variation. Moreover, to obtain the snapshots for fixed

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

μ_m, a complete time-trajectory needs to be computed, suitable time instants need to be selected, and unused information from intermediate time steps discarded.

To overcome these limitations, the greedy and POD approaches can be suitably combined in the so-called *POD-greedy* procedure (Grepl and Patera 2005, Haasdonk and Ohlberger 2008). The POD-greedy method combines the greedy algorithm in parameter space with proper orthogonal decomposition to select the time instants containing the maximal information of the trajectory associated to a given parameter.

As described in Hesthaven *et al.* (2015, Section 6.1.2), the POD-greedy sampling procedure requires the set of training parameters $\mathbb{P}_h$, an error tolerance and two suitable positive integers N_1 and N_2. The greedy algorithm provides the outer algorithm where we search the currently worst-resolved parameter μ_m in $\mathbb{P}_h$ using an error bound or indicator. Then the complete time-trajectory, $u_h^1(\mu_m), \ldots, u_h^{N_T}(\mu_m)$, associated with μ_m is computed and the first N_1 principal components are retained via POD. In a subsequent step, the existing n-dimensional reduced basis space is enriched with the components derived in the previous step and a new $(n + N_2)$-dimensional reduced basis is constructed via POD. In an alternative version proposed in Haasdonk and Ohlberger (2008), the second POD compression is avoided by considering the error trajectories rather than the trajectories. In this case POD with respect to time is performed to compress the error trajectory to its most important new information, and the POD mode is added to the current basis. Note that the POD-greedy algorithm generates hierarchical reduced spaces. Moreover, the successive greedy cycles allow us to retain new information and reject redundant information.

Concerning the computational cost of the POD-greedy procedure, we remark that the operation count is additive and not multiplicative in the number of training points in $\mathbb{P}_h$ and N. The fact that the high-fidelity trajectories are evaluated only for selected parameters in $\mathbb{P}_h$ allows us to consider relatively large training sets of parameters.

The convergence results of the greedy algorithm shown in Binev *et al.* (2011) and Buffa *et al.* (2012) (see Theorem 3.2) were extended by Haasdonk (2013) to the POD-greedy strategy of Haasdonk and Ohlberger (2008): exponential or algebraic convergence rates of the Kolmogorov n-width of the solution manifold are maintained by the POD-greedy algorithm with a change of the multiplicative factor and of the exponent.

4.2. *Structure-preserving methods*

The application of reduced basis methods to time-dependent problems might lead to unstable and qualitatively wrong solution behaviours: indeed, the reduced model generally does not inherit the physical properties of the original differential equation. The physical properties of a problem are mathematically encoded in so-called geometric structures, structural properties 'which can be defined independently of

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

particular coordinate representations of the differential equations' (Christiansen, Munthe-Kaas and Owren 2011). Examples of such structures are conservation laws, symmetries, symplecticity, reversibility and invariants of motion. The observation that inaccurate and unstable solutions resulting from standard numerical methods often originate from inconsistent discretizations of the fields and operators involved has led to the concepts of structure-preserving discretization (see *e.g.* Shashkov 1996, Hiptmair 2002, Arnold, Falk and Winther 2006) and geometric integration (Sanz-Serna and Calvo 1994, Leimkuhler and Reich 2004, Iserles, Munthe-Kaas, Nørsett and Zanna 2000, Hairer, Lubich and Wanner 2006, Christiansen *et al.* 2011). Formulating numerical approximations compatible with the geometric structure underlying the original model allows physically consistent simulations and leads to numerical schemes with superior accuracy and stability properties.

Structure-preserving model order reduction originates from discussion of the physical interpretability of reduced models. Indeed, failing to preserve intrinsic properties of the original model not only raises questions about the validity of the reduced models but is also associated with instabilities and exponential error growth, independent of the accuracy of the reduced space. Structure-preserving MOR aims at constructing reduced models that exactly retain (part of) the geometric structures of the full model.

In the following subsections we discuss structure-preserving reduced basis methods for finite-dimensional parametrized differential equations. These can be dynamical systems or they can result from the semidiscrete approximation of time-dependent PDEs. In the latter case, the geometric structures considered are strictly tied to the numerical techniques employed in the semidiscretization of the PDE of interest. First we discuss structure-preserving reduced basis methods of conservative dynamical systems in the framework of Lagrangian, Hamiltonian and Poisson dynamics. Then we deal with non-conservative systems with special types of dissipation and with port-Hamiltonian problems.

Remark 4.1 (notation). Throughout this subsection we will use the following notation: $\mathbb{V}_N$ denotes a general vector space of dimension N; N is the size of the truth problem; n is the size of the reduced order model; $2N$ and $2n$ are used to indicate that the full order and reduced order model, respectively, are even-dimensional; $\boldsymbol{u}(t;\mu)$ denotes the truth solution and it coincides with the vector $(\mathrm{u}_{h,1}(t;\mu),\dots,\mathrm{u}_{h,N}(t;\mu))$ of degrees of freedom of u_h whenever the ordinary differential equation comes from the semidiscretization of a PDE as in (4.4).

4.2.1. Conservative dynamics

Conservative processes are characterized by no dissipative effects, such as thermal conduction, electric resistivity or viscous dissipation. They are modelled via hyperbolic partial differential equations, conservation laws and Hamiltonian dynamical systems. Examples of finite-dimensional systems in this category are rigid body

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

motion and the n-body problem, the harmonic oscillator, the dynamical billiard, the Euler top equations, the Heisenberg equation, the May–Leonard model and certain Lotka–Volterra models. In the infinite-dimensional case, examples of conservative systems are Maxwell's equations, Schrödinger's equation, Korteweg–de Vries and the wave equation, compressible and incompressible Euler equations, Vlasov–Poisson and Vlasov–Maxwell equations.

Conservative systems are at the core of classical mechanics and they can be considered from two points of view: the Lagrangian formalism and the Hamiltonian formalism (Marsden and Ratiu 1999). Let the configuration manifold of a mechanical system be the manifold collecting all possible configurations of the system. Lagrangian mechanics is based on variational principles via a smooth function, the Lagrangian, defined on the velocity phase space, the tangent bundle of the configuration manifold. We will delve into details of the Lagrangian formalism at the end of the subsection. Hamiltonian mechanics instead relies on conjugate momenta and on the cotangent bundle, called momentum phase space. The cotangent bundle is endowed with a natural symplectic manifold structure, which underpins the physical properties of the system.

Definition 4.2 (symplectic vector space). Let $\mathbb{V}_{2N}$ be a $2N$-dimensional real vector space. A skew-symmetric bilinear form $\omega\colon \mathbb{V}_{2N} \times \mathbb{V}_{2N} \to \mathbb{R}$ is *symplectic* if it is non-degenerate, that is, if $\omega(\boldsymbol{u}, \boldsymbol{v}) = 0$, for any $\boldsymbol{v} \in \mathbb{V}_{2N}$, then $\boldsymbol{u} = 0$. The map ω is called a *linear symplectic structure* on $\mathbb{V}_{2N}$, and $(\mathbb{V}_{2N}, \omega)$ is called a *symplectic vector space*.

On a finite $2N$-dimensional smooth manifold $\mathbb{V}_{2N}$, let ω be a 2-form, that is, for any $\boldsymbol{u} \in \mathbb{V}_{2N}$, the map $\omega_{\boldsymbol{u}} : T_{\boldsymbol{u}}\mathbb{V}_{2N} \times T_{\boldsymbol{u}}\mathbb{V}_{2N} \to \mathbb{R}$ is skew-symmetric and bilinear on the tangent space $T_{\boldsymbol{u}}\mathbb{V}_{2N}$ to $\mathbb{V}_{2N}$ at $\boldsymbol{u}$, and it varies smoothly in $\boldsymbol{u}$. The 2-form ω is a *symplectic structure* if it is closed and $\omega_{\boldsymbol{u}}$ is symplectic for all $\boldsymbol{u} \in \mathbb{V}_{2N}$, in the sense of Definition 4.2. A manifold $\mathbb{V}_{2N}$ endowed with a symplectic structure ω is called a *symplectic manifold* and denoted by $(\mathbb{V}_{2N}, \omega)$. Every symplectic manifold $(\mathbb{V}_{2N}, \omega)$ is, in a suitable local coordinate system, a symplectic vector space. This means that in a neighbourhood of each $\boldsymbol{u} \in \mathbb{V}_{2N}$, there is a local coordinate chart in which ω is constant (Darboux 1882).

The algebraic structure of a symplectic manifold $(\mathbb{V}_{2N}, \omega)$ can be characterized by the definition of a bracket. Let $\mathsf{d}\mathcal{F}$ be the 1-form given by the exterior derivative d of a smooth function $\mathcal{F}$. Then, for all $\mathcal{F}, \mathcal{G} \in C^\infty(\mathbb{V}_{2N})$,

$$\{\mathcal{F}, \mathcal{G}\}_{2N}(\boldsymbol{u}) := {}_{T^*\mathbb{V}_{2N}}\langle \mathsf{d}\mathcal{F}, \mathcal{J}_{2N}(\boldsymbol{u})\, \mathsf{d}\mathcal{G}\rangle_{T\mathbb{V}_{2N}} = \omega(X_{\mathcal{F}}(\boldsymbol{u}), X_{\mathcal{G}}(\boldsymbol{u})), \tag{4.9}$$

where ${}_{T^*\mathbb{V}_{2N}}\langle \cdot, \cdot \rangle_{T\mathbb{V}_{2N}}$ denotes the duality pairing between the cotangent bundle $T^*\mathbb{V}_{2N}$ and the tangent bundle $T\mathbb{V}_{2N}$. The function $\mathcal{J}_{2N}(\boldsymbol{u})\colon T^*\mathbb{V}_{2N} \to T\mathbb{V}_{2N}$ is a contravariant 2-tensor on the manifold $\mathbb{V}_{2N}$, commonly referred to as a *Poisson tensor*, and $X_{\mathcal{F}} := \mathcal{J}_{2N}\mathsf{d}\mathcal{F}$ is called the *Hamiltonian vector field* associated with $\mathcal{F}$.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Hamiltonian systems on symplectic spaces

Any vector field on a manifold $\mathbb{V}_{2N}$ determines a phase flow. Hamilton's equations for the smooth function $\mathcal{H} \in C^\infty(\mathbb{V}_{2N})$ correspond to the following system of differential equations defined by the Hamiltonian vector field $X_\mathcal{H} \in T\mathbb{V}_{2N}$: find $\boldsymbol{u} \in C^1(\mathcal{T}; \mathbb{V}_{2N})$ such that

$$\begin{cases} \dot{\boldsymbol{u}}(t) = X_\mathcal{H}(\boldsymbol{u}(t)) & \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}(t_0) = \boldsymbol{u}_0 \in \mathbb{V}_{2N}. \end{cases} \tag{4.10}$$

In a local coordinate system on $\mathbb{V}_{2N}$, the Hamiltonian vector field $X_\mathcal{H}$ has the form $X_\mathcal{H}(\boldsymbol{u}) = \mathbf{D}_{2N}(\boldsymbol{u})\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u})$, where, for any $\boldsymbol{u} \in \mathbb{V}_{2N}$, $\nabla_{\boldsymbol{u}}$ denotes the gradient with respect to $\boldsymbol{u}$, and $\mathbf{D}_{2N}(\boldsymbol{u}) \in \mathbb{R}^{2N\times 2N}$ is a skew-symmetric matrix, representation of the Poisson tensor $\mathcal{J}_{2N}$ in the chosen local coordinate system. A set of local coordinates $(q^1(\boldsymbol{u}), \dots, q^N(\boldsymbol{u}), p_1(\boldsymbol{u}), \dots, p_N(\boldsymbol{u}))$, with $\boldsymbol{u} \in \mathbb{V}_{2N}$, is called canonical if

$$\{q^i, q^k\}_{2N} = \{p_i, p_k\}_{2N} = 0, \quad \{q^i, p_k\}_{2N} = \delta_{i,k} \quad \text{for all } i, k = 1, \dots, N.$$

In the local canonical coordinates, the Poisson tensor $\mathcal{J}_{2N}$ takes the so-called canonical symplectic form, defined as

$$\mathbf{J}_{2N} \coloneqq \begin{pmatrix} \mathbf{0}_N & \mathbf{I}_N \\ -\mathbf{I}_N & \mathbf{0}_N \end{pmatrix} \in \mathbb{R}^{2N\times 2N}, \tag{4.11}$$

where $\mathbf{I}_N, \mathbf{0}_N \in \mathbb{R}^{N\times N}$ are the identity and zero matrix, respectively. This implies that the Hamiltonian vector field associated with $\mathcal{H} \in C^\infty(\mathbb{V}_{2N})$, in the canonical coordinate system, is given by $X_\mathcal{H}(\boldsymbol{u}) = \mathbf{J}_{2N}\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u})$. Thus Hamilton's equations (4.10) in canonical coordinates read

$$\begin{cases} \dot{\boldsymbol{q}}(t) = \nabla_{\boldsymbol{p}}\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) & \text{for } t \in \mathcal{T}, \\ \dot{\boldsymbol{p}}(t) = -\nabla_{\boldsymbol{q}}\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) & \text{for } t \in \mathcal{T}, \\ \boldsymbol{q}(t_0) = \boldsymbol{q}_0, \ \boldsymbol{p}(t_0) = \boldsymbol{p}_0. \end{cases} \tag{4.12}$$

The phase flow of a vector field $X_\mathcal{H}$ on $\mathbb{V}_{2N}$ is the one-parameter group of diffeomorphisms $\Phi^t_{X_\mathcal{H}} : \mathbb{V}_{2N} \to \mathbb{V}_{2N}$ that satisfy $\Phi^0_{X_\mathcal{H}}(\boldsymbol{u}) = \boldsymbol{u}$ and

$$\dot{\Phi}^t_{X_\mathcal{H}}(\boldsymbol{u}) = X_\mathcal{H}(\Phi^t_{X_\mathcal{H}}(\boldsymbol{u})) \quad \text{for all } t \in \mathcal{T}, \ \boldsymbol{u} \in \mathbb{V}_{2N}. \tag{4.13}$$

The flow of a Hamiltonian vector field satisfies $(\Phi^t_{X_\mathcal{H}})^*\omega = \omega$, for each $t \in \mathcal{T}$, that is, it preserves the symplectic form ω. The converse also holds: see Marsden and Ratiu (1999, Proposition 10.2.3).

Definition 4.3 (symplectic map). Let $(\mathbb{V}_{2N}, \omega_{2N})$ and $(\mathbb{V}_{2n}, \omega_{2n})$ be symplectic manifolds of finite dimension $2N$ and $2n$ respectively, with $n \le N$. Let $\{\cdot,\cdot\}_{2N}$ and $\{\cdot,\cdot\}_{2n}$ denote the corresponding brackets. A smooth map $\Psi\colon (\mathbb{V}_{2N}, \omega_{2N}) \to (\mathbb{V}_{2n}, \omega_{2n})$ is called *symplectic* if it satisfies

$$\Psi^*\omega_{2n} = \omega_{2N},$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

or equivalently $\Psi^*\{\mathcal{F},\mathcal{G}\}_{2n} = \{\Psi^*\mathcal{F},\Psi^*\mathcal{G}\}_{2N}$, for any $\mathcal{F},\mathcal{G} \in C^\infty(\mathbb{V}_{2n})$.

In addition to possessing a symplectic phase flow, Hamiltonian dynamics is characterized by the existence of differential invariants, and symmetry-related conservation laws. This family of conserved quantities are related, by Noether's theorem (Noether 1918), to symmetries of the Hamiltonian. A function $\mathcal{I} \in C^\infty(\mathbb{V}_{2N})$ is an *invariant of motion* of the dynamical system (4.10), if $\{\mathcal{I},\mathcal{H}\}_{2N}(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in \mathbb{V}_{2N}$. Consequently $\mathcal{I}$ is constant along the orbits of $X_{\mathcal{H}}$. The Hamiltonian, if time-independent, is an invariant of motion. A particular subset of the invariants of motion of a dynamical system is given by the *Casimir invariants*, smooth functions $\mathcal{C}$ on $\mathbb{V}_{2N}$ that ω_{2N}-commute with every other functions, *i.e.* $\{\mathcal{C},\mathcal{F}\}_{2N} = 0$ for all $\mathcal{F} \in C^\infty(\mathbb{V}_{2N})$. Since Casimir invariants are associated with degeneracies of the symplectic form ω, symplectic manifolds only possess trivial Casimir invariants. In Section 4.2.4 we will see non-trivial Casimir invariants associated with Poisson manifold structures.

When dealing with symplectic vector spaces, the canonical coordinates allow us to identify a *Kähler structure*, namely a compatible combination of a scalar product and symplectic form, as follows. On a symplectic vector space $(\mathbb{V}_{2N},\omega)$, the operator $\mathbf{J}_{2N}^\top$ is an *almost complex structure*, *i.e.* a linear map on $\mathbb{V}_{2N}$ such that $\mathbf{J}_{2N}^\top \circ \mathbf{J}_{2N}^\top = -\mathbf{I}_{2N}$. Furthermore, $\mathbf{J}_{2N}^\top$ is *compatible* with the symplectic structure ω, namely, for any $\boldsymbol{u},\boldsymbol{v} \in \mathbb{V}_{2N}$, $\boldsymbol{u} \neq 0$, we find that

$$\omega\left(\mathbf{J}_{2N}^\top\boldsymbol{u},\mathbf{J}_{2N}^\top\boldsymbol{v}\right) = \omega(\boldsymbol{u},\boldsymbol{v}) \quad \text{and} \quad \omega\left(\boldsymbol{u},\mathbf{J}_{2N}^\top\boldsymbol{u}\right) > 0.$$

A symplectic form ω on a vector space $\mathbb{V}_{2N}$ together with a compatible positive almost complex structure $\mathbf{J}_{2N}^\top$ determines an inner product on $\mathbb{V}_{2N}$, given by

$$(\boldsymbol{u},\boldsymbol{v}) \coloneqq \omega\left(\boldsymbol{u},\mathbf{J}_{2N}^\top\boldsymbol{v}\right) \quad \text{for all } \boldsymbol{u},\boldsymbol{v} \in \mathbb{V}_{2N}. \tag{4.14}$$

A symplectic basis on $(\mathbb{V}_{2N},\omega)$ is an orthonormal basis for the compatible inner product (4.14), and we refer to it as orthosymplectic.

Definition 4.4 (orthosymplectic basis). The set of vectors $\{\boldsymbol{e}_i\}_{i=1}^{2N}$ is said to be *orthosymplectic* in the $2N$-dimensional vector space $\mathbb{V}_{2N}$ if

$$\boldsymbol{e}_i^\top\mathbf{J}_{2N}\boldsymbol{e}_k = (\mathbf{J}_{2N})_{i,k} \quad \text{and} \quad (\boldsymbol{e}_i,\boldsymbol{e}_k) = \delta_{i,k} \quad \text{for all } i,k = 1,\dots,2N,$$

where $(\cdot,\cdot)$ is the Euclidean inner product and $\mathbf{J}_{2N}$ is the canonical symplectic tensor (4.11) on $\mathbb{V}_{2N}$.

Orthosymplectic bases play a crucial role in model order reduction in the construction of low-dimensional spaces where surrogate dynamical systems are derived.

Lagrangian formalism

In the Lagrangian variational formalism, a system is defined by a Lagrangian function $L\colon T\mathbb{V}_N \to \mathbb{R}$, where $\mathbb{V}_N$ is the configuration manifold of dimension N,

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

and $T\mathbb{V}_N$ is the corresponding velocity phase space (the tangent bundle of $\mathbb{V}_N$). Let $\boldsymbol{q}\colon \mathcal{T} \to \mathbb{V}_N$ denote the time-dependent configuration variables, called *generalized coordinates*. The Lagrangian is typically given by the difference between the kinetic and potential energy of the system, namely,

$$L(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \frac{1}{2}\dot{\boldsymbol{q}}^\top \mathbf{M}\dot{\boldsymbol{q}} - V(\boldsymbol{q}), \tag{4.15}$$

where $\mathbf{M}$ is an $N \times N$ symmetric positive definite mass matrix and V is a nonlinear potential-energy function $V\colon \mathbb{V}_N \to \mathbb{R}$. The corresponding equations of motion are given by the Euler–Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\dot{\boldsymbol{q}}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}) - \nabla_{\boldsymbol{q}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}) = 0 \quad \text{for } t \in \mathcal{T}. \tag{4.16}$$

As discussed at the beginning of the subsection, autonomous Lagrangian systems (4.16) possess important physical properties. The energy is conserved along solution trajectories and, by Noether's theorem (Noether 1918), there is an invariant of motion associated with each symmetry that leaves the Lagrangian invariant. Moreover, the flow of the system, *i.e.* its time-evolution map, is a symplectic transformation (see Definition 4.3). To pass to the canonical Hamiltonian formalism (4.12), a change of variables based on the conjugate momenta $p_i := \partial L/\partial \dot{q}^i$ for $i = 1, \ldots, N$ is needed. With the change of variables $(\boldsymbol{q}, \dot{\boldsymbol{q}}) \mapsto (\boldsymbol{q}, \boldsymbol{p})$ and the Hamiltonian $\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{p}^\top \dot{\boldsymbol{q}} - L(\boldsymbol{q}, \dot{\boldsymbol{q}})$, the Euler–Lagrange equations are equivalent to Hamilton's equations (4.12).

In this review we focus on autonomous systems, namely with Hamiltonian and Lagrangian functions that do not explicitly depend on time. However, conservative systems with explicit time dependence in the Hamiltonian or Lagrangian can be studied by utilizing the extended Lagrangian mechanics framework or the symplectic extended phase space; see *e.g.* Lanczos (1949). This takes into account time variations in addition to variations in the configuration variables and consists in reformulating the problem as an autonomous system by defining an extended state with time as a variable. Structure-preserving model order reduction for non-autonomous Hamiltonian system can be derived from the autonomous case. We refer to Buchfink, Bhatt and Haasdonk (2019, Section 2.5) for a summary of the resulting algorithm.

4.2.2. Lagrangian model order reduction

Structure-preserving model order reduction of conservative dynamical systems was first approached from the Lagrangian perspective in Lall, Krysl and Marsden (2003) and extended to nonlinear parametric Lagrangian systems in Carlberg, Tuminaro and Boggs (2015). Let us consider parameters that describe, for example, variations in material or physical properties of the system. For any parameter $\mu \in \mathbb{P}$, consider the parametric Euler–Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\dot{\boldsymbol{q}}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mu) - \nabla_{\boldsymbol{q}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mu) = 0 \quad \text{for } t \in \mathcal{T}. \tag{4.17}$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

To apply model order reduction in the Lagrangian framework, the full model can be thought of as characterized by a configuration space and an associated Lagrangian function, from which the complete dynamics can then be specified by the variational principle. Lagrangian model order reduction consists in first approximating the quantities defining the problem's Lagrangian structure and subsequently deriving the equations of motion by applying the Euler–Lagrange equation to the reduced quantities. The rationale is that model reduction based on Galerkin projection of the Euler–Lagrange equations – and not on the first-order state-space form – preserves Lagrangian structure. As seen in Section 4.2.1, this mechanical structure is associated with key physical properties, such as energy conservation, symplecticity of the flow map, and conservation of quantities corresponding to symmetries of the system.

From a Euclidean configuration space $\mathbb{V}_N$, Lagrangian MOR derives an approximate configuration space of lower dimension, namely a submanifold $\mathbb{V}_n \subset \mathbb{V}_N$, using standard dimension reduction techniques, *e.g.* POD or modal decomposition. Then the reduced mechanical system is constrained to the configuration space given by $\mathbb{V}_n$. This approach separates the configuration space selection from the dynamic reconstruction. The original Lagrangian

$$L(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mu) = \frac{1}{2}\dot{\boldsymbol{q}}^\top \mathbf{M}(\mu)\dot{\boldsymbol{q}} - V(\boldsymbol{q}; \mu),$$

restricted to the constraint submanifold $\mathbb{V}_n$, gives a new Lagrangian L_r on $\mathbb{V}_n$:

$$L_r(\boldsymbol{q}_r, \dot{\boldsymbol{q}}_r; \mu) = \frac{1}{2}\dot{\boldsymbol{q}}_r^\top \mathbf{U}^\top M(\mu)\mathbf{U}\dot{\boldsymbol{q}}_r - V(\mathbf{U}\boldsymbol{q}_r; \mu). \tag{4.18}$$

Here $\mathbf{U} \in \mathbb{R}^{N\times n}$ is a parameter-independent matrix whose orthonormal columns represent a reduced basis spanning an n-dimensional subspace of $\mathbb{V}_N$, and $\boldsymbol{q}_r \in \mathbb{V}_n$ is such that $\mathbf{U}\boldsymbol{q}_r$ approximates the full model solution $\boldsymbol{q} \in \mathbb{V}_N$. The reduced Lagrangian (4.18) defines the Euler–Lagrange equations for the reduced dynamics on $\mathbb{V}_n$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\dot{\boldsymbol{q}}_r} L_r(\boldsymbol{q}_r, \dot{\boldsymbol{q}}_r; \mu) - \nabla_{\boldsymbol{q}_r} L(\boldsymbol{q}_r, \dot{\boldsymbol{q}}_r; \mu) = 0 \quad \text{for } t \in \mathcal{T}. \tag{4.19}$$

Since $\mathbf{U}^\top\mathbf{M}(\mu)\mathbf{U}$ is a low-dimensional symmetric positive definite matrix, the first term of (4.18) represents kinetic energy. Moreover, the reduced dynamics inherits, by construction, the geometric structure of the full model.

Although the Lagrangian model order reduction method proposed in Lall *et al.* (2003) yields equations of motion (4.19) of lower dimension, the computational cost of solving the reduced dynamics scales with the dimension of the full model whenever the operators cannot be assembled in the offline phase, namely when they exhibit arbitrary parameter dependence and the potential is nonlinear. In order to reduce the complexity of reduced order models of nonlinear Lagrangian systems, Carlberg *et al.* (2015) proposed further approximating the reduced Lagrangian quantities before deriving the Euler–Lagrange equations. To approximate the

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

reduced symmetric positive definite matrix $\mathbf{U}^\top\mathbf{M}(\mu)\mathbf{U}$, two alternatives are proposed in Carlberg *et al.* (2015): a *reduced basis sparsification* method that consists in projecting the full matrix $\mathbf{M}$ onto a sparse basis, and a *gappy POD* procedure applied to $\mathbf{M}$ that approximates the reduced matrix as a linear combination of precomputed reduced matrices.

To deal with nonlinearities in the potential V, first the reduced basis $\mathbf{U}$ is replaced by a sparse parameter-dependent matrix $\mathbf{U}_V(\mu) \in \mathbb{R}^{N\times n}$, and then the gradient of the corresponding reduced potential $\nabla_{\boldsymbol{q}_r} V(\mathbf{U}_V(\mu)\boldsymbol{q}_r;\mu)$ is approximated with its Taylor polynomial expansion about the reference configuration, truncated at the second term. This technique, although not effective for general systems, provides an efficient and good approximation when the expansion is performed around an equilibrium point and the dynamical system is assumed to be asymptotically stable.

4.2.3. Symplectic model order reduction

A second family of structure-preserving model order reduction techniques for conservative systems relies on the Hamiltonian formalism. Let us first consider parametric Hamiltonian systems, in the so-called canonical symplectic form (4.12): for any $\mu \in \mathbb{P}$ and for $\boldsymbol{u}_0(\mu) \in \mathbb{V}_{2N}$, find $\boldsymbol{u}(\mu) \in C^1(\mathcal{T};\mathbb{V}_{2N})$ such that

$$\begin{cases} \dot{\boldsymbol{u}}(t;\mu) = \mathbf{J}_{2N}\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u}(t;\mu);\mu) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}(t_0,\mu) = \boldsymbol{u}_0(\mu). \end{cases} \tag{4.20}$$

A Petrov–Galerkin projection-based approach consists in constructing a reduced basis $\mathbf{U} \in \mathbb{R}^{2N\times 2n}$ spanning the reduced space $\mathbb{V}_{2n}$ and a matrix $\mathbf{W} \in \mathbb{R}^{2n\times 2N}$ such that $\mathbf{W}^\top\mathbf{U} = \mathbf{I}_{2n}$. The reduced order model is then derived by requiring that the equation residual vanishes in the space spanned by the columns of the projection matrix $\mathbf{U}\mathbf{W}^\top$; thereby

$$\begin{cases} \dot{\boldsymbol{u}}_r(t;\mu) = \mathbf{W}^\top\mathbf{J}_{2N}\nabla_{\boldsymbol{u}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t;\mu);\mu) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}_r(t_0;\mu) = \mathbf{W}^\top\boldsymbol{u}_0(\mu), \end{cases} \tag{4.21}$$

where $\boldsymbol{u}_r \in \mathbb{V}_{2n}$ is such that $\mathbf{U}\boldsymbol{u}_r$ is an approximation of the full model solution $\boldsymbol{u} \in \mathbb{V}_{2N}$.

Reduced basis methods aimed at preserving the physical and geometric properties of Hamiltonian dynamics (4.20) hinge on two main aspects: (i) an approximate phase space of reduced dimension endowed with the same symplectic structure of the full phase space, and (ii) a projection operator that preserves the geometric properties of the flow field (4.13).

Proposition 4.5. For any fixed parameter $\mu \in \mathbb{P}$, the reduced model (4.21) is Hamiltonian provided that the reduced basis and the projection operator satisfy the constraint

$$\mathbf{W}^\top\mathbf{J}_{2N} = \mathbf{J}_{2n}\mathbf{U}^\top. \tag{4.22}$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Moreover, under the constraint (4.22), the resulting Hamiltonian system reads

$$\begin{cases} \dot{\boldsymbol{u}}_r(t;\mu) = \mathbf{J}_{2n}\nabla_{\boldsymbol{u}_r}\mathcal{H}_r(\boldsymbol{u}_r(t;\mu);\mu) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}_r(t_0;\mu) = \mathbf{W}^\top \boldsymbol{u}_0(\mu), \end{cases} \tag{4.23}$$

where the Hamiltonian is given by $\mathcal{H}_r(\cdot\,;\mu) \coloneqq \mathcal{H}(\cdot\,;\mu)\circ\mathbf{U}\colon \mathbb{V}_{2n} \to \mathbb{R}$.

Proof. Condition (4.22) applied to (4.21) gives

$$\begin{aligned} \dot{\boldsymbol{u}}_r(t;\mu) &= \mathbf{W}^\top\mathbf{J}_{2N}\nabla_{\boldsymbol{u}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t;\mu);\mu) \\ &= \mathbf{J}_{2n}\mathbf{U}^\top\nabla_{\boldsymbol{u}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t;\mu);\mu) \\ &= \mathbf{J}_{2n}\nabla_{\boldsymbol{u}_r}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t;\mu);\mu) \\ &= \mathbf{J}_{2n}\nabla_{\boldsymbol{u}_r}\mathcal{H}_r(\boldsymbol{u}_r(t;\mu);\mu), \end{aligned}$$

with $\mathcal{H}_r(\boldsymbol{x};\mu) = \mathcal{H}(\mathbf{U}\boldsymbol{x};\mu)$ for any $\boldsymbol{x} \in \mathbb{V}_{2n}$ and $\mu \in \mathbb{P}$. □

The first method following this construction was proposed in Peng and Mohseni (2016*b*), where the reduced space is spanned by an *orthosymplectic* basis and a *symplectic* Galerkin projection yields a reduced Hamiltonian system on this subspace.

To construct a Galerkin projection, symplectic MOR relies on the fact that a linear map $\mathbf{A}\colon (\mathbb{V}_{2N},\omega) \to (\mathbb{V}_{2n},\omega)$, $N \geq n$, is symplectic if and only if the corresponding matrix representation $\mathbf{A} \in \mathbb{R}^{2n\times 2N}$ satisfies $\mathbf{A}\mathbf{J}_{2N}\mathbf{A}^\top = \mathbf{J}_{2n}$. The *symplectic right inverse* of $\mathbf{A}$ is defined as the matrix $\mathbf{B} \coloneqq \mathbf{J}_{2N}\mathbf{A}^\top\mathbf{J}_{2n}^\top \in \mathbb{R}^{2N\times 2n}$. The symplectic condition $\mathbf{A}\mathbf{J}_{2N}\mathbf{A}^\top = \mathbf{J}_{2n}$ is equivalent to $\mathbf{B}^\top\mathbf{J}_{2N}\mathbf{B} = \mathbf{J}_{2n}$. Owing to this equivalence, with a small abuse of notation, the space

$$\mathrm{Sp}(2n,\mathbb{R}^{2N}) \coloneqq \{\mathbf{M} \in \mathbb{R}^{2N\times 2n}\colon \mathbf{M}^\top\mathbf{J}_{2N}\mathbf{M} = \mathbf{J}_{2n}\}$$

is often used to denote the space of symplectic matrices and $\mathbf{M}^+ \coloneqq \mathbf{J}_{2n}^\top\mathbf{M}^\top\mathbf{J}_{2N} \in \mathbb{R}^{2n\times 2N}$ defines the symplectic inverse of $\mathbf{M} \in \mathrm{Sp}(2n,\mathbb{R}^{2N})$; see Peng and Mohseni (2016*b*, Definition 3.2).

Symplectic rectangular matrices that are also orthogonal are called *orthosymplectic*; thereby

$$\mathcal{U}(2n,\mathbb{R}^{2N}) \coloneqq \mathrm{St}(2n,\mathbb{R}^{2N}) \cap \mathrm{Sp}(2n,\mathbb{R}^{2N}), \tag{4.24}$$

where $\mathrm{St}(2n,\mathbb{R}^{2N}) \coloneqq \{\mathbf{M} \in \mathbb{R}^{2N\times 2n}\colon \mathbf{M}^\top\mathbf{M} = \mathbf{I}_{2n}\}$ is the Stiefel manifold. Orthosymplectic rectangular matrices can be characterized as follows; see *e.g.* Buchfink *et al.* (2019, Proposition 4).

Lemma 4.6. Let $\mathbf{M} \in \mathrm{Sp}(2n,\mathbb{R}^{2N})$ and let $\mathbf{M}^+ = \mathbf{J}_{2n}^\top\mathbf{M}^\top\mathbf{J}_{2N} \in \mathbb{R}^{2n\times 2N}$ be its symplectic inverse. Then $\mathbf{M}^\top\mathbf{M} = \mathbf{I}_{2n}$ if and only if $\mathbf{M}^+ = \mathbf{M}^\top$.

The method proposed in Peng and Mohseni (2016*b*) consists in setting $\mathbf{W}^\top = \mathbf{U}^+$ in (4.21) so that condition (4.22) is satisfied. This choice corresponds to a symplectic projection in the sense that $\omega(\boldsymbol{u} - \mathbf{U}\mathbf{U}^+\boldsymbol{u}, \boldsymbol{w}) = 0$ for any $\boldsymbol{w}$ in the range of $\mathbf{U}$.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Stability and conservation properties of the reduced order model

The fact that the reduced model constructed by the symplectic projection is Hamiltonian has the important consequence that energy and stability are preserved during the time evolution. As shown in Peng and Mohseni (2016*b*, Theorem 3.10) and Maboudi Afkham and Hesthaven (2017, Theorem 18), if $\boldsymbol{u}^* \in \mathrm{Im}(\mathbf{U})$ is a stable equilibrium of the full model (4.20), then it is a stable equilibrium of the reduced system (4.23) constructed by the symplectic projection.

Concerning the invariants of motion of the dynamics, the error of the Hamiltonian evaluated in the full solution $\boldsymbol{u}$ and in the reduced solution $\boldsymbol{u}_r$, for any parameter $\mu \in \mathbb{P}$, is given by

$$\begin{aligned} |\mathcal{H}(\boldsymbol{u};\mu) - \mathcal{H}(\mathbf{U}\boldsymbol{u}_r;\mu)| &= |\mathcal{H}(\boldsymbol{u};\mu) - \mathcal{H}_r(\boldsymbol{u}_r;\mu)| \\ &= |\mathcal{H}(\boldsymbol{u}_0;\mu) - \mathcal{H}_r(\boldsymbol{u}_r(t_0;\mu);\mu)| \\ &= |\mathcal{H}(\boldsymbol{u}_0;\mu) - \mathcal{H}(\mathbf{U}\mathbf{W}^\top \boldsymbol{u}_0;\mu)|, \end{aligned} \tag{4.25}$$

where, in the second line, we have used the fact that the Hamiltonian is a conserved quantity. This implies that the error in the Hamiltonian is constant in time and vanishes if the initial condition $\boldsymbol{u}_0$ belongs to the range of $\mathbf{U}$. The latter condition can be enforced by either including the vector $\boldsymbol{u}_0 \in \mathbb{R}^{2N}$ in the reduced basis $\mathbf{U}$ or by reformulating the Hamiltonian system (4.20) for the shifted variable $\boldsymbol{u}(t;\mu) - \boldsymbol{u}_0(\mu)$ as in Gong, Wang and Wang (2017, Section 4.2) and Hesthaven and Pagliantini (2021, Section 3.2.1).

With the exception of the Hamiltonian, if $\mathcal{I} \in C^\infty(\mathbb{V}_{2N})$ is an invariant of motion of the Hamiltonian system (4.20), $\mathbf{U} \circ \mathcal{I} \in C^\infty(\mathbb{V}_{2n})$ is not necessarily an invariant of the reduced order model (4.23), since $\mathbf{U}$ is not a symplectic map. This is possible under the stronger condition that the Hamiltonian $\mathcal{H}(\cdot;\mu)$ belongs to the range of the pullback of $\mathbf{U}^+$ for any $\mu \in \mathbb{P}$; see Hesthaven and Pagliantini (2021, Lemma 3.8).

Construction of the reduced basis

Let $\mathcal{M}_h$ denote the solution set which collects, as in (4.3), solutions of (4.20) under variation of time and parameter. Different data-driven techniques have been proposed in the literature to construct structure-preserving reduced bases from partial information on the solution space given by full order snapshots $\mathbf{S} := [\boldsymbol{u}(\eta_1), \ldots, \boldsymbol{u}(\eta_M)] \in \mathbb{R}^{2N\times M}$. All these methods consist in selecting the linear subspace of $\mathbb{V}_{2N}$ in which the projection error of the snapshots is minimal, namely, $\mathbb{V}_{2N}$ is spanned by the column of the matrix $\mathbf{U} \in \mathbb{R}^{2N\times 2n}$ that satisfies

$$\mathbf{U} := \underset{\mathbf{W}\in\mathcal{W}}{\arg\min} \|\mathbf{S} - \mathcal{P}_{\mathbf{W}}\mathbf{S}\|, \tag{4.26}$$

where $\mathcal{P}_{\mathbf{W}}$ denotes a suitable Petrov–Galerkin projection operator, $\mathcal{W}$ is a matrix space and $\|\cdot\|$ is a given matrix norm. The geometric structure of the reduced space is enforced by constraining $\mathbf{W}$ to $\mathcal{W}$ in the minimization problem (4.26): the choice $\mathcal{W} = \mathrm{St}(2n, \mathbb{R}^{2N})$ yields an orthogonal reduced basis, while $\mathcal{W} = \mathrm{Sp}(2n, \mathbb{R}^{2N})$ yields a symplectic basis. The minimization problem associated with the latter

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

choice is significantly more challenging than the orthogonal case, and rigorous results are not available. However, two families of numerical algorithms have been proposed: SVD-based methods and greedy algorithms. We discuss them in the following.

Proper symplectic decomposition (PSD), introduced by Peng and Mohseni (2016*b*), is the extension to the orthosymplectic case of the proper orthogonal decomposition (POD) (see Section 3) archetype of SVD-based algorithms to find the best low-rank approximation of $\mathbf{S}$. Three different PSD algorithms are derived in Peng and Mohseni (2016*b*) in a non-parametric setting.

(i) The *cotangent lift* (Peng and Mohseni 2016*b*, Section 4.1) consists in constructing a symplectic reduced basis in the subset of block diagonal matrices with equal diagonal blocks $\Phi \in \mathbb{R}^{N\times n}$ that are orthogonal. An optimal basis in this subset can be found by deriving Φ via SVD of the snapshot matrix

$$\mathbf{S} \coloneqq [\boldsymbol{q}(t_1), \dots, \boldsymbol{q}(t_M), \boldsymbol{p}(t_1), \dots, \boldsymbol{p}(t_M)] \in \mathbb{R}^{N\times 2M}.$$

(ii) *Complex SVD* (Peng and Mohseni 2016*b*, Section 4.2) relies on the isomorphism between the set $\mathcal{U}$ of orthosymplectic matrices (4.24) and the complex Stiefel manifold $\mathrm{St}(2n, \mathbb{C}^{2N}) \coloneqq \{\mathbf{M} \in \mathbb{C}^{2N\times 2n} : \mathbf{M}^*\mathbf{M} = \mathbf{I}_{2n}\}$. An optimal orthosymplectic reduced basis results from the unitary matrix of left singular vectors obtained via SVD of the snapshot matrix

$$\mathbf{S} \coloneqq [\boldsymbol{q}(t_1) + \imath\boldsymbol{p}(t_1), \dots, \boldsymbol{q}(t_M) + \imath\boldsymbol{p}(t_M)] \in \mathbb{C}^{N\times M}.$$

(iii) The *nonlinear programming* (NLP) algorithm (Peng and Mohseni 2016*b*, Section 4.3) provides a (suboptimal) solution to the minimization problem (4.26) without imposing the orthogonality constraint in the feasibility set. It consists of an optimization strategy where a POD symplectic basis is constructed from a linear transformation of an orthosymplectic matrix, as follows. First an orthosymplectic basis $\mathbf{U}^* \in \mathcal{U}(2r, \mathbb{R}^{2N})$ is generated using the complex SVD method, then the algorithm looks for a basis $\mathbf{U} \in \mathrm{Sp}(2n, \mathbb{R}^{2N})$, with $n < r \ll N$, such that $\mathbf{U} = \mathbf{U}^*\mathbf{C}$, for some $\mathbf{C} \in \mathbb{R}^{2r\times 2n}$. Using this expression for $\mathbf{U}$ in (4.26) results in a minimization problem for the coefficient matrix $\mathbf{C}$. Although the latter is of significantly smaller dimension ($4nr$ unknowns) compared to the original minimization problem for $\mathbf{U}$ ($4Nn$ unknowns), no optimality results are available in this setting.

A slightly different SVD-based algorithm is the Galerkin POD method proposed in Gong *et al.* (2017). The method considers Hamiltonian systems (4.20) where the Poisson tensor $\mathbf{J}_{2N}$ (4.11) is replaced by any skew-symmetric matrix $\mathbf{D}_{2N} \in \mathbb{R}^{2N\times 2N}$. The reduced basis $\mathbf{U} \in \mathbb{R}^{2N\times 2n}$ is then assumed to be block diagonal with blocks $\mathbf{U}_{\boldsymbol{q}} \in \mathrm{St}(n, \mathbb{R}^N)$ and $\mathbf{U}_{\boldsymbol{p}} \in \mathrm{St}(n, \mathbb{R}^N)$. These blocks are obtained via POD from the snapshot matrices $\mathbf{S}_{\boldsymbol{q}} \coloneqq [\boldsymbol{q}(t_1), \dots, \boldsymbol{q}(t_M)]$ and $\mathbf{S}_{\boldsymbol{p}} \coloneqq [\boldsymbol{p}(t_1), \dots, \boldsymbol{p}(t_M)]$,

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

respectively. Although the resulting reduced basis is not guaranteed to be optimal, it satisfies $\mathbf{U}^\top \mathbf{D}_{2N} = \mathbf{D}_{2n}\mathbf{U}^\top$, with $\mathbf{D}_{2n} := \mathbf{U}^\top \mathbf{D}_{2N}\mathbf{U}$, and thus choosing $\mathbf{W} = \mathbf{U}$ in (4.21) yields a Hamiltonian reduced model.

A different approach to solving the minimization problem (4.26) is via a greedy strategy. The greedy method is an iterative algorithm that consists in enlarging, at each iteration, the reduced basis in the direction that gives the worst projection error of the full model snapshots. In the symplectic greedy algorithm of Maboudi Afkham and Hesthaven (2017), the error in the Hamiltonian (4.25) is used as an error indicator to iteratively select a direction in the parameter space, namely, at the kth iteration of the algorithm

$$\mu^* = \underset{\mu \in \mathbb{P}_h \subset \mathbb{P}}{\arg\max} \, |\mathcal{H}(\boldsymbol{u}_0(\mu)) - \mathcal{H}(P_{\mathbf{W}_{k-1}} \boldsymbol{u}_0(\mu))|, \tag{4.27}$$

where $\mathbb{P}_h$ is a finite subset of the parameter space and $\mathbf{W}_{k-1}$ is the reduced basis constructed at the $(k-1)$th step. Since the error in the Hamiltonian depends only on the initial condition, it does not require the integration of the full model (4.20) over the whole temporal interval and for all parameter values, thus making the procedure faster compared to the SVD-type algorithms described in (i)–(iii). On the other hand, if the initial condition $\boldsymbol{u}_0(\mu)$ belongs to the range of $\mathbf{W}_{k-1}$ the procedure will fail and the error indicator needs to be replaced with the more expensive projection error. After a parameter has been selected according to (4.27), the projection error in the snapshots $\mathbf{S}_* := [\boldsymbol{u}(t_1, \mu^*), \ldots, \boldsymbol{u}(t_{N_T}, \mu^*)]$ is used to select as the new basis vector $\boldsymbol{x} := \boldsymbol{u}(t^*, \mu^*)$, where

$$t^* := \underset{1 \le i \le N_T}{\arg\max} \, \|\boldsymbol{u}(t_i, \mu^*) - P_{\mathbf{W}_{k-1}} \boldsymbol{u}(t_i, \mu^*)\|.$$

To enforce the orthosymplecticity of the resulting reduced basis, the algorithm adds, at each iteration, the new vector $\boldsymbol{x}$ and its symplectic dual $\mathbf{J}_{2N}^\top \boldsymbol{x}$. Moreover, the orthonormalization step of traditional greedy algorithms is replaced by a variation of the QR method, known as the SR method (Salam 2005), which is based on the symplectic Gram–Schmidt method leading to a symplectic basis. If the projection error is used as the criterion to select the new basis elements, the greedy algorithm is shown in Maboudi Afkham and Hesthaven (2017, Theorem 20) to converge at an exponential rate, under the assumption of good approximability properties of the full model (*i.e.* exponential decay of the Kolmogorov n-width of the solution space).

Reduced bases that are symplectic but do not satisfy the orthogonality constraints were proposed in Buchfink *et al.* (2019). The gist of the proposed algorithm is to perform a symplectic SVD-like decomposition of the snapshot matrix and to retain the $2n$ modes associated with the largest so-called *weighted symplectic singular values*. In more detail, a result by Xu (2003) ensures that any matrix $\mathbf{S} \in \mathbb{R}^{2N \times M}$ admits a factorization $\mathbf{S} = \mathbf{V}\mathbf{D}\mathbf{Q}$, where $\mathbf{V} \in \mathrm{Sp}(2N, \mathbb{R}^{2N})$, $\mathbf{Q} \in \mathrm{St}(M, \mathbb{R}^M)$ and

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

$\mathbf{D} \in \mathbb{R}^{2N\times M}$ is of the form

$$\mathbf{D} = \begin{array}{c} \begin{matrix} p & q & p & n-2p-q \end{matrix} \\ \begin{bmatrix} \Sigma & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array} \begin{matrix} p \\ q \\ m-p-q \\ p \\ q \\ m-p-q \end{matrix},$$

with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, and $\sigma_i > 0$, for all $i = 1, \ldots, p$, known as *symplectic singular values*. The weighted symplectic singular values of $\mathbf{S}$ are defined as

$$w_i = \begin{cases} \sigma_i \sqrt{\|V_i\|_2^2 + \|V_{n+i}\|_2^2}, & 1 \le i \le p, \\ \|V_i\|_2, & p+1 \le i \le p+q, \end{cases}$$

where $V_i \in \mathbb{R}^{2N}$ denotes the ith column of $\mathbf{V}$ and $\|\cdot\|_2$ is the Euclidean norm. Starting from the factorization of the snapshot matrix $\mathbf{S} \coloneqq [\boldsymbol{u}(\eta_1), \ldots, \boldsymbol{u}(\eta_M)] \in \mathbb{R}^{2N\times M}$, a symplectic reduced basis $\mathbf{U} \in \mathrm{Sp}(2n, \mathbb{R}^{2N})$ is obtained by selecting the pairs of columns from the symplectic matrix $\mathbf{V}$ corresponding to the n largest energy contributors, namely $\mathbf{U} = [V_{i_1}, \ldots, V_{i_n}, V_{N+i_1}, \ldots, V_{N+i_n}]$, where

$$\{i_k\}_{k=1}^n \coloneqq \underset{\mathcal{I}\subset\{1,\ldots,p+q\}}{\arg\max} \left(\sum_{i\in\mathcal{I}} w_i^2 \right).$$

Similarly to POD (see Section 3), the projection error of the snapshots in the derived reduced basis is bounded by the neglected weighted symplectic singular values (Buchfink *et al.* 2019, Proposition 11).

4.2.4. Conservative systems on Poisson manifolds

In this subsection we consider the generalization of Hamiltonian systems on symplectic manifolds to more general phase space manifold structures. Many dynamical systems describing conservative processes are indeed characterized by a phase space whose manifold structure is degenerate, meaning that the Poisson tensor is rank-deficient. Moreover, the Poisson tensor $\mathcal{J}_{2N}(\boldsymbol{u})$ introduced in (4.9) might depend on the state variable $\boldsymbol{u}$ so that the geometric structure of the phase space has a purely local characterization. Despite the complexity of such phase space manifold structure, many problems of interest in scientific applications can be considered within this framework, for example rigid body motion, the Toda lattice problem (Kostant 1979), certain Lotka–Volterra models and, in infinite dimensions, the Korteweg–de Vries equation (Miura, Gardner and Kruskal 1968), compressible (Marsden *et al.* 1983) and incompressible (Arnol'd 1966) Euler equations, and the Vlasov–Poisson and Vlasov–Maxwell equations (Morrison 1980).

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

The phase space of a general Hamiltonian system is a Poisson manifold. An N-dimensional Poisson manifold is a smooth differential manifold $\mathbb{V}_N$ endowed with a bracket $\{\cdot,\cdot\}_N\colon C^\infty(\mathbb{V}_N)\times C^\infty(\mathbb{V}_N)\to C^\infty(\mathbb{V}_N)$, namely a bilinear form satisfying the following properties, for all $\mathcal{F},\mathcal{G},\mathcal{I}\in C^\infty(\mathbb{V}_N)$:

$$\begin{aligned}&\text{skew-symmetry}\quad \{\mathcal{F},\mathcal{G}\}_N=-\{\mathcal{G},\mathcal{F}\}_N,\\ &\text{Leibniz rule}\quad \{\mathcal{F}\mathcal{G},\mathcal{I}\}_N=\{\mathcal{F},\mathcal{I}\}_N\,\mathcal{G}+\mathcal{F}\{\mathcal{G},\mathcal{I}\}_N,\\ &\text{Jacobi identity}\quad \{\mathcal{F},\{\mathcal{G},\mathcal{I}\}_N\}_N+\{\mathcal{G},\{\mathcal{I},\mathcal{F}\}_N\}_N+\{\mathcal{I},\{\mathcal{F},\mathcal{G}\}_N\}_N=0.\end{aligned}\tag{4.28}$$

The bracket can be equivalently defined via the Poisson tensor as

$$\{\mathcal{F},\mathcal{G}\}_N(\boldsymbol{u})={}_{T^*\mathbb{V}_N}\langle \mathrm{d}\mathcal{F},\mathcal{J}_N(\boldsymbol{u})\,\mathrm{d}\mathcal{G}\rangle_{T\mathbb{V}_N}\quad\text{for all }\mathcal{F},\mathcal{G}\in C^\infty(\mathbb{V}_N).$$

Allowing degeneracies of the Poisson structure entails that there exists a set of functions $\mathcal{C}\in C^\infty(\mathbb{V}_N)$ such that $\{\mathcal{C},\mathcal{F}\}_N=0$ for all $\mathcal{F}\in C^\infty(\mathbb{V}_N)$. These are called *Casimir invariants*. Contrary to the invariant of motions, Casimir invariants are independent of the dynamics and only depend on the Poisson structure of the manifold, in particular its degeneracy. Most Hamiltonian systems on Poisson manifolds that appear in applications are characterized by globally conserved quantities, such as energy, angular momentum, helicity and vorticity. Hence we assume that the phase space $\mathbb{V}_N$ is a *regular* Poisson manifold, that is, the rank of the Poisson tensor is constant, $\operatorname{rank}(\mathcal{J}_N(\boldsymbol{u}))=2R$, for all $\boldsymbol{u}\in\mathbb{V}_N$, with $R\in\mathbb{N}$, $2R\le N$.

A fundamental property of Poisson manifolds is that they can be expressed as a union of symplectic manifolds. Two points $\boldsymbol{u}$ and $\boldsymbol{v}$ on a Poisson manifold $\mathbb{V}_N$ are said to be on the same *symplectic leaf* of $\mathbb{V}_N$ if there is a piecewise smooth curve joining $\boldsymbol{u}$ and $\boldsymbol{v}$, each segment of which is a trajectory of a locally defined Hamiltonian vector field. Any Poisson manifold $\mathbb{V}_N$ is the disjoint union of its symplectic leaves, and the dimension of the leaf through a point $\boldsymbol{u}$ equals the rank of the Poisson tensor at that point. This result allows us to introduce local coordinate charts as the generalization of the canonical coordinates on symplectic manifolds: for any ball $\mathbb{B}\subset\mathbb{V}_N$ and $\boldsymbol{u}\in\mathbb{B}$, there exist

$$\big(q^1(\boldsymbol{u}),\dots,q^R(\boldsymbol{u}),p_1(\boldsymbol{u}),\dots,p_R(\boldsymbol{u}),c^1(\boldsymbol{u}),\dots,c^{N-2R}(\boldsymbol{u})\big),$$

such that $\{q^i,q^k\}_N=\{p_i,p_k\}_N=\{q^i,c^k\}_N=\{p_i,c^k\}_N=0$ and $\{q^i,p_k\}_N=\delta_{i,k}$ for all $i,k=1,\dots,R$ and $k=1,\dots,N-2R$. In the neighbourhood $\mathbb{B}$ of $\boldsymbol{u}$, the coordinates $\{c^k\}_{k=1}^{N-2R}$ correspond to the Casimir invariants, whereas $\{(q^i,p_k)\}_{i,k=1}^{R}$ are the symplectic canonical coordinates. For any $\boldsymbol{u}\in\mathbb{V}_N$, the Poisson tensor $\mathcal{J}_N(\boldsymbol{u})$ in the local canonical coordinates takes the form

$$\mathbf{J}_N\coloneqq\begin{pmatrix}\mathbf{0}_R & \mathbf{I}_R & \mathbf{0}_{R,q}\\ -\mathbf{I}_R & \mathbf{0}_R & \mathbf{0}_{R,q}\\ \mathbf{0}_{q,R} & \mathbf{0}_{q,R} & \mathbf{0}_q\end{pmatrix}\in\mathbb{R}^{N\times N},\tag{4.29}$$

where $\mathbf{0}_{R,q}$ is the matrix of size R times q with all zero entries. There are many advantages in using canonical coordinates and having the Poisson tensor in the

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

constant-valued form (4.29). However, canonical coordinates for general Poisson tensors are often not known and not trivial to construct.

In this subsection we consider the following general parametric Hamiltonian systems: for any $\mu \in \mathbb{P}$ and for $\boldsymbol{u}_0(\mu) \in \mathbb{V}_N$, find $\boldsymbol{u}(\mu) \in C^1(\mathcal{T}; \mathbb{V}_N)$ such that

$$\begin{cases} \dot{\boldsymbol{u}}(t;\mu) = \mathbf{D}_N(\boldsymbol{u}) \nabla_{\boldsymbol{u}} \mathcal{H}(\boldsymbol{u}(t;\mu);\mu) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}(t_0,\mu) = \boldsymbol{u}_0(\mu), \end{cases} \tag{4.30}$$

where $\mathbf{D}_N(\boldsymbol{u}) \in \mathbb{R}^{N\times N}$ is the matrix representation of the Poisson tensor $\mathcal{J}_N(\boldsymbol{u})$ in a local coordinate system on $\mathbb{V}_N$, not necessarily the canonical one. In the following, we consider degenerate Poisson structures: first the constant-valued case and then the state-dependent case.

Hamiltonian systems with degenerate constant-valued structures

On the N-dimensional vector space $\mathbb{V}_N$ we consider, for each parameter $\mu \in \mathbb{P}$, the initial value problem (4.30) with constant-valued Poisson tensor, namely,

$$\dot{\boldsymbol{u}}(t;\mu) = \mathbf{D}_N \nabla_{\boldsymbol{u}} \mathcal{H}(\boldsymbol{u}(t;\mu);\mu) \quad \text{for } t \in \mathcal{T}, \tag{4.31}$$

where $\mathbf{D}_N \in \mathbb{R}^{N\times N}$ is a skew-symmetric linear operator. Every skew-symmetric matrix of odd dimension can be seen as the representation of the Poisson tensor (4.11) in a given coordinate system. In particular, even-dimensional skew-symmetric operators are associated with symplectic structures. Hesthaven and Pagliantini (2021, Proposition 2.11) showed that when N is odd, any skew-symmetric matrix $\mathbf{D}_N \in \mathbb{R}^{N\times N}$ of rank $2R < N$ admits a decomposition of the form

$$\Psi \mathbf{D}_N \Psi^\top = \mathbf{J}_N, \tag{4.32}$$

where $\Psi \in \mathbb{R}^{N\times N}$ is invertible – but not orthogonal in general – and $\mathbf{J}_N \in \mathbb{R}^{N\times N}$ is the matrix representation of the Poisson tensor in canonical form (4.29). The factorization (4.32) is unique up to transformations in the symplectic group $\mathrm{Sp}(2R, \mathbb{R}^{2R})$. Although the factorization (4.32) entails that the phase space of any system of the form (4.31) has a degenerate symplectic structure, orthogonal and symplectic projection-based model order reduction fails to preserve this geometric structure and its degeneracy.

A structure-preserving reduced basis method was proposed in Hesthaven and Pagliantini (2021). The gist of the method is to recast the full model using a suitable coordinate system where the null space of the structure is isolated, and then to apply symplectic model order reduction only on the non-degenerate component of the dynamics. In more detail, the factorization (4.32) ensures that there exists a linear bijective map Ψ such that

$$\Psi \colon \mathbb{V}_N \longrightarrow \mathbb{V}_{2R} \times \mathcal{N}, \quad \Psi \mathbf{D}_N \Psi^\top = \mathbf{J}_N, \tag{4.33}$$

where $\mathbb{V}_{2R}$ is a symplectic manifold of dimension $2R$ and $\mathcal{N}$ is a submanifold whose dimension equals q, the number of independent Casimir invariants associated with

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

the operator $\mathbf{D}_N$. This entails that the full model dynamics can be decoupled into the dynamics on the symplectic leaf and the trivial dynamics of the Casimir invariants, that is, (4.31) can be recast in canonical form as follows: find $z(\cdot;\mu) \in C^1(\mathcal{T};\mathbb{V}_N)$ such that

$$\begin{cases} \dot{z}(t;\mu) = \mathbf{J}_N \nabla_z \mathcal{H}^c(z(t;\mu);\mu) & \text{for } t \in \mathcal{T}, \\ z(t_0,\mu) = \Psi \boldsymbol{u}_0(\mu), \end{cases} \tag{4.34}$$

where $\mathcal{H}^c := (\Psi^{-1})^* \mathcal{H}$ for every $\mu \in \mathbb{P}$.

Since Ψ is a Poisson map, the invariants of motion of the system (4.31) are in one-to-one correspondence with the invariants of motion of (4.34) (Hesthaven and Pagliantini 2021, Corollary 3.2).

The idea of the method proposed in Hesthaven and Pagliantini (2021) is to exploit the splitting of the dynamics (4.33), to perform symplectic model order reduction on the symplectic manifold $\mathbb{V}_{2R}$, while leaving unchanged the submanifold $\mathcal{N}$ associated with the centre of the Lie algebra $C^\infty(\mathbb{V}_N)$. Formally this means that a reduced basis $\mathbf{U} \in \mathbb{R}^{n\times N}$ is constructed to span an n-dimensional vector space $\mathbb{V}_n$, for $n \ll N$, such that the rank of the canonical Poisson tensor $\mathbf{J}_n$ on $\mathbb{V}_n$, $\operatorname{rank}(\mathbf{J}_n) =: 2r$, satisfies $n - 2r = q$, namely,

$$\mathbf{U}\colon \mathcal{V}_{2r} \times \mathcal{N} \longrightarrow \mathcal{V}_{2R} \times \mathcal{N}, \quad \mathbf{U} = \mathbf{U}_s \times \mathbf{I},$$

where $\mathbf{U}_s$ is taken to be an injective ℓ^2-orthogonal symplectic map, *i.e.* $\mathbf{U}_s \in \mathrm{Sp}(2r, \mathbb{R}^{2R})$. This is constructed via one of the techniques presented in Section 4.2.3 using the snapshot matrix $\mathbf{S} = [\Psi_s \boldsymbol{u}(t_1), \ldots, \Psi_s \boldsymbol{u}(t_{N_T})]$, where the map Ψ_s is defined from Ψ by $\mathbb{V}_{2R} = \Psi_s(\mathbb{V}_N)$. If $\mathbf{U}^+ := \mathbf{U}_s^+ \times \mathbf{I}$, and $\mathbf{U}_s^+ \in \mathbb{R}^{2R\times 2r}$ is the symplectic inverse of $\mathbf{U}$, then a reduced model is obtained via the Poisson projection $\mathbf{U}\mathbf{U}^+$ onto $\mathrm{Im}(\mathbf{U}) \subset \mathbb{V}_N$ of the dynamical system (4.34) in canonical Poisson form: for $t \in \mathcal{T}$ and $\mu \in \mathbb{P}$, the approximation $\mathbf{U}z_r(t;\mu)$ of $z(t;\mu)$ is such that

$$\begin{cases} \dot{z}_r(t;\mu) = \mathbf{J}_n \nabla_{z_r} \mathcal{H}_r(z_r(t;\mu);\mu) & \text{for } t \in \mathcal{T}, \\ z_r(t_0,\mu) = \mathbf{U}^+ \Psi \boldsymbol{u}_0(\mu), \end{cases} \tag{4.35}$$

where $\mathcal{H}_r := \mathbf{U}^* \mathcal{H}^c$ for all $\mu \in \mathbb{P}$.

The structure-preserving properties of this reduced basis method have important effects: the reduced dynamics preserves the Lyapunov stable equilibria contained in $\mathrm{Im}(\mathbf{U})$, the Poisson map $\mathbf{U}^+$ provides a Hamiltonian-preserving model reduction and, since the null space $\mathcal{N}$ is not affected by the reduction, the Casimir invariants of the full model are exactly conserved in the reduced problem. Furthermore, *a priori* convergence estimates of the L^2-error – both in time and parameter space – between the full model solution and the reduced solution were established in Hesthaven and Pagliantini (2021, Proposition 3.13), including the case when the symplectic DEIM of Peng and Mohseni (2016*b*) and Maboudi Afkham and Hesthaven (2017) is used as hyper-reduction of general nonlinear terms.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Hamiltonian systems with state-dependent and degenerate structures

The numerical discretization and model order reduction of problems with state-dependent and degenerate structures, $\mathbf{D}_N = \mathbf{D}_N(\boldsymbol{u})$ in (4.30), is challenged by the time dependence and nonlinearity intrinsic to the manifold structure. Approximations of the Poisson tensor may indeed destroy this structure since the Jacobi identity (4.28) generally fails to hold for the approximate tensor.

Hesthaven and Pagliantini (2021) introduced a reduced basis method that preserves general Poisson structures for non-parametric systems. The method uses the temporal discretization of the dynamics to 'freeze' the phase space manifold structure in each discrete temporal interval, then recasts the local problem in canonical form, and subsequently constructs a local reduced model in canonical Hamiltonian form.

Let $\mathbb{V}_N$ be a Poisson manifold and let $\mathbf{D}_N(\boldsymbol{u})$ denote the Poisson tensor at each state $\boldsymbol{u} \in \mathbb{V}_N$. On the temporal mesh $\mathcal{T}_h = \bigcup_{k\in\Upsilon_h} \mathcal{T}_k$, with $\Upsilon_h \subset \mathbb{N}$ a given set of indices, the approach of Hesthaven and Pagliantini (2021) considers the following discretization of (4.30): for $\boldsymbol{u}_0 \in \mathbb{V}_N$, find $\{\boldsymbol{u}^{k+1}\}_{k\in\Upsilon_h} \subset \mathbb{V}_N$ such that

$$\begin{cases} \boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t\, \mathbf{D}_N(\widehat{\boldsymbol{u}}^k)\nabla\mathcal{H}(\widehat{\boldsymbol{u}}^k) \quad \text{for } k \in \Upsilon_h, \\ \boldsymbol{u}^0 = \boldsymbol{u}_0, \end{cases} \tag{4.36}$$

where $\widehat{\boldsymbol{u}}^k \in \mathbb{V}_N$ is determined by the temporal discretization. Alternative discretizations of the Poisson tensor and of the Hamiltonian are possible. To recast the local dynamical system (4.36) in canonical Poisson form, each local Poisson tensor $\mathbf{D}_N(\widehat{\boldsymbol{u}}^k)$, with $k \in \Upsilon_h$, is factorized as in (4.32). In more detail, if $\mathbb{V}_{N,k}$ denotes an open subset of $\mathbb{V}_N$ containing the states $\boldsymbol{u}^{k+1}$, $\boldsymbol{u}^k$ and $\widehat{\boldsymbol{u}}^k$, (4.32) identifies a bijective linear function $\psi_{k+1/2}\colon \mathbb{V}_{N,k} \to \mathbb{V}_{N,k}$ that satisfies $\psi_{k+1/2}\mathbf{D}_N(\widehat{\boldsymbol{u}}^k)\psi_{k+1/2}^\top = \mathbf{J}_N$ at the state(s) $\widehat{\boldsymbol{u}}^k \in \mathcal{T}_k$ dictated by the temporal discretization (4.36). Each map $\psi_{k+1/2}$ provides the local splitting $\psi_{k+1/2}\colon \mathbb{V}_{N,k} \to \mathbb{V}_{2R} \times \mathcal{N}_k$, where $\mathcal{N}_k$ is the approximation of the subspace associated with the kernel of the Poisson tensor at $\widehat{\boldsymbol{u}}^k$. Similarly to the degenerate symplectic case described earlier in this subsection, the function $\psi_{k+1/2}$ introduces a change of coordinates on $\mathbb{V}_{N,k}$ so that, if $\boldsymbol{u}^{k+1} \in \mathbb{V}_N$ is the solution of (4.36) in $\mathcal{T}_k$, then the function $\boldsymbol{z}^{k+1} = \psi_{k+1/2}\boldsymbol{u}^{k+1} \in \mathbb{V}_{2R} \times \mathcal{N}_k$ satisfies

$$\begin{cases} \boldsymbol{z}^{k+1} = T_k\boldsymbol{z}^k + \Delta t\, \mathbf{J}_N\nabla\mathcal{H}^k(\widehat{\boldsymbol{z}}^k) \quad \text{for } k \in \Upsilon_h, \\ \boldsymbol{z}^0 = \psi_{1/2}\boldsymbol{u}_0, \end{cases} \tag{4.37}$$

where $\widehat{\boldsymbol{z}}^k := \psi_{k+1/2}\widehat{\boldsymbol{u}}^k$, $\mathbf{J}_N$ is as in (4.29) and the Hamiltonian is $\mathcal{H}^k(\boldsymbol{z}) := \mathcal{H}(\psi_{k+1/2}^{-1}\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathbb{V}_{N,k}$. The functions $\{T_k\}_{k\in\Upsilon_h}$ are defined as *transition maps* between neighbouring subsets,

$$T_k\colon \psi_{k-1/2}(\mathbb{V}_{N,k-1} \cap \mathbb{V}_{N,k}) \longrightarrow \psi_{k+1/2}(\mathbb{V}_{N,k-1} \cap \mathbb{V}_{N,k}),$$

with $T_k := \psi_{k+1/2} \circ \psi_{k-1/2}^{-1}$, for $k \in \Upsilon_h \setminus \{0\}$ and $T_0 := I$.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

A reduced model for the discrete dynamical system (4.36) is derived by an approach analogous to the degenerate symplectic case, with the difference that here the projection is performed *locally*, in each temporal interval. The reduced phase space is an n-dimensional Poisson manifold, $n \ll N$, endowed with the canonical Poisson tensor $\mathbf{J}_n$ such that $n - \operatorname{rank}(\mathbf{J}_n) = q$, with q the dimension of the null space of $\mathbf{D}_N(\boldsymbol{u})$. This is constructed by means of a *global* linear surjective map $\mathbf{U}^+$ such that, for every $k \in \Upsilon_h$,

$$\mathbf{U}^+ \colon \mathbb{V}_{2R} \times \mathcal{N}_k \longrightarrow \mathbb{V}_{2r} \times \mathcal{N}_k, \quad \mathbf{U}^+ = \mathbf{U}_s^+ \times \mathbf{I},$$

where $\mathbf{U}_s^+ \in \mathcal{U}(2r, \mathbb{R}^{2R})$. In the offline phase, snapshots of system (4.36) in each temporal interval $\mathcal{T}_k$ are collected, and the corresponding maps $\{\psi_{k+1/2}\}_k$ are computed via (4.32). This information is used to construct a global symplectic reduced basis via one of the algorithms presented in the previous subsection. The local projection of (4.37) onto $\operatorname{Im}(\mathbf{U}) \cap \mathbb{V}_{N,k} \subset \mathbb{V}_N$ yields the following reduced model: for $\boldsymbol{u}_0 \in \mathbb{V}_N$, find $\{z_r^{k+1}\}_{k \in \Upsilon_h} \subset \mathbb{V}_n$ such that

$$\begin{cases} z_r^{k+1} = \tau_k z_r^k + \Delta t\, \mathbf{J}_n \nabla \mathcal{H}_r^k(\widehat{z}_r^k) & \text{for } k \in \Upsilon_h, \\ z_r^0 = \mathbf{U}^+ \psi_{1/2} \boldsymbol{u}_0, \end{cases} \tag{4.38}$$

where the quantity $\mathbf{U}\widehat{z}_r^k$ approximates the pseudo-state $\widehat{z}^k$, the reduced Hamiltonian is $\mathcal{H}_r^k(z_r) \coloneqq \mathcal{H}(\psi_{k+1/2}^{-1}\mathbf{U}z_r)$ for all $z_r \in \mathbb{V}_n$, and the reduced transition maps τ_k are defined as $\tau_k \coloneqq \mathbf{U}^+ \circ T_k \circ \mathbf{U}$ for all $k \in \Upsilon_h \setminus \{0\}$, with $\tau_0 \coloneqq I$, and can be precomputed in the offline phase.

The global Poisson structure, the stability and the conservation properties of the phase flow are retained by the reduced model up to local errors in the Poisson tensor approximation; see Hesthaven and Pagliantini (2021, Section 4.2). If the approximate maps $\{\psi_{k+1/2}\}_{k \in \Upsilon_h}$ are constructed to be continuous at the interface between temporal intervals, *i.e.* $\psi_{k-1/2}u^k = \psi_{k+1/2}\boldsymbol{u}^k$, where $\boldsymbol{u}^k$ is the solution of (4.36) in $\mathcal{T}_k$, then the reduced model exactly preserves the Hamiltonian and the global manifold structure.

Dynamical systems of the form (4.30), with state-dependent operator $\mathbf{D}_{2N}(\boldsymbol{u})$ and no parameter dependence, have been considered in Miyatake (2019). However, there $\mathbf{D}_{2N}(\boldsymbol{u})$ is not a Poisson tensor, namely, it is assumed to be skew-symmetric but it does not satisfy the Jacobi identity (4.28). Any dynamical system with a conservation law can be represented in such a form. Miyatake (2019) derived a reduced model with the method proposed in Gong *et al.* (2017); thereby

$$\begin{cases} \dot{\boldsymbol{u}}_r(t) = \mathbf{U}^\top \mathbf{D}_{2N}(\mathbf{U}\boldsymbol{u}_r)\mathbf{U}\mathbf{U}^\top \nabla_{\boldsymbol{u}} \mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t)) & \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}_r(t_0) = \mathbf{U}^\top \boldsymbol{u}_0, \end{cases}$$

where $\mathbf{U} \in \mathbb{R}^{2n \times 2N}$ is orthogonal and the gradient $\nabla_{\boldsymbol{u}}\mathcal{H}$ has been approximated by its projection into the reduced space. The skew-symmetry of the reduced operator $\mathbf{D}_{2n}(\boldsymbol{u}_r) \coloneqq \mathbf{U}^\top \mathbf{D}_{2N}(\mathbf{U}\boldsymbol{u}_r)\mathbf{U}$, for any $\boldsymbol{u}_r \in \mathbb{R}^{2n}$, is guaranteed by the orthogonality transformation and ensures preservation of the conservation law of the full model.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Moreover, its efficient computation is carried out in Miyatake (2019) via a matrix DEIM strategy (Wirtz, Sorensen and Haasdonk 2014, Carlberg *et al.* 2015).

The structure-preserving reduced basis methods described in the previous subsections has proved successful in the efficient and stable solution of several problems such as the linear wave equation (Peng and Mohseni 2016*b*, Maboudi Afkham and Hesthaven 2017, Gong *et al.* 2017), sine-Gordon equation (Peng and Mohseni 2016*b*), nonlinear Schrödinger equation (Maboudi Afkham and Hesthaven 2017, Karasözen and Uzunca 2018), shallow water equations (Karasözen, Yıldız and Uzunca 2021), Korteweg–de Vries equation (Gong *et al.* 2017, Hesthaven and Pagliantini 2021, Uzunca, Karasözen and Yıldız 2021) and population dynamics models (Hesthaven and Pagliantini 2021).

4.2.5. Methods based on constraints

A different class of projection-based reduced basis methods for conservative systems relies on the preservation of certain geometric structures in the reduced models by enforcing suitable constraints.

Carlberg, Choi and Sargsyan (2018) constructed a model order reduction method for finite-volume models to guarantee conservation over each control volume of the computational mesh. This projection-based method endows the optimization problems of Galerkin and least-squares Petrov–Galerkin projection with a minimum-residual objective function and nonlinear equality constraints that explicitly enforce conservation over subdomains. These constrained optimization problems are solved online, at each time step. Moreover, to efficiently handle nonlinearities in the flux and source terms while respecting the finite-volume discretization, hyper-reduction techniques are applied to the nonlinear residuals that appear in the constraints.

General constrained-optimization formulations for projection-based model reduction were proposed in Schein, Carlberg and Zahr (2021) to enforce the resulting reduced models to satisfy specific physical properties such as conservation of invariants.

An energy-conserving POD-Galerkin reduced basis method for the incompressible Navier–Stokes equations is proposed in Sanderse (2020). The method relies on a finite-volume discretization of the full model that preserves mass, momentum and kinetic energy. Moreover, it possesses a skew-symmetric convective operator, a symmetric diffusive term, and satisfies the compatibility relation between divergence and gradient. Based on this semidiscrete problem, the method of Sanderse (2020) constructs a reduced velocity space from snapshots of divergence-free velocities via POD with a weighted inner product. To enforce exact conservation of momentum and energy in the reduced model – in the inviscid case and under periodic boundary conditions – the minimization problem that underlies the SVD is replaced by a constrained minimization problem. The formulation of dynamical systems and fluid flow equations in a skew-symmetric form, and its numerical discretization, has received considerable attention thanks to the conservation prop-

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

erties associated with the preservation of the skew-symmetric form at the discrete level, *e.g.* quadratic invariants such as energy. In Maboudi Afkham, Ripamonti, Wang and Hesthaven (2020), nonlinear conservation laws with quadratic nonlinearities are reduced using a POD-Galerkin reduced basis method applied to the skew-symmetric form of the semidiscrete full model. By preserving the skew-symmetry of the differential operators, conservation of the quadratic invariants of the reduced model is guaranteed. The loss in energy is associated with the approximation of the full model invariants due to the model order reduction, which, however, remains constant in time.

4.2.6. Non-conservative dynamics

The favourable properties of the reduced models obtained via structure-preserving MOR have also stimulated the development of analogous techniques in the context of non-conservative systems, whose dynamics can be described by special geometric structures. In this setting, model order reduction methods have been developed to address forced Lagrangian systems (Carlberg *et al.* 2015), forced Hamiltonian systems (Peng and Mohseni 2016*a*), time dispersive and dissipative (TDD) problems (Maboudi Afkham and Hesthaven 2019) and port-Hamiltonian systems (Polyuga and van der Schaft 2010, 2011, Beattie and Gugercin 2011, Chaturantabut, Beattie and Gugercin 2016, Beattie, Gugercin and Mehrmann 2019). As discussed below, several of these techniques are based on modifications of the Lagrangian and symplectic model order reduction techniques introduced in the previous subsections.

Forced Hamiltonian systems

Carlberg *et al.* (2015) dealt with parametric mechanical systems in Lagrangian form subject to the external and dissipative forces arising from Rayleigh viscous damping. The equations of motion of such systems are given by the Euler–Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\dot{\boldsymbol{q}}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mu) - \nabla_{\boldsymbol{q}} L(\boldsymbol{q}, \dot{\boldsymbol{q}}; \mu) = f(\boldsymbol{q}, \dot{\boldsymbol{q}}, t; \mu)) - \nabla_{\dot{\boldsymbol{q}}} F(\dot{\boldsymbol{q}}; \mu), \tag{4.39}$$

where L is the Lagrangian as in (4.16), f denotes the external force that is derived from the Lagrange–D'Alembert variational principle, and dissipation is modelled by the Rayleigh dissipation function

$$F(\dot{\boldsymbol{q}}; \mu) = \frac{1}{2}\dot{\boldsymbol{q}}^{\top}\mathbf{C}(\mu)\dot{\boldsymbol{q}}, \tag{4.40}$$

where $\mathbf{C}(\mu)$ is a parametric symmetric positive semidefinite matrix.

The model order reduction strategy described in Section 4.2.2 is based on the Galerkin projection of the Euler–Lagrange equations into the reduced space spanned by the columns of an orthonormal matrix $\mathbf{U} \in \mathbb{R}^{N\times n}$. The resulting reduced order model preserves the Lagrangian structure even in the non-conservative case. Indeed, it leads to a positive semidefinite reduced dissipation function

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

$F_r(\dot{\boldsymbol{q}}_r;\mu) = \frac{1}{2}\dot{\boldsymbol{q}}_r^\top \mathbf{U}^\top \mathbf{C}(\mu)\mathbf{U}\dot{\boldsymbol{q}}_r$, and a reduced external force f_r derived by applying the Lagrange–D'Alembert principle to the full model force f restricted to variations in the configuration space Q_r. To deal with non-affine dependence and nonlinear terms, the complexity reduction strategies used in the conservative case (see Section 4.2.2) can be applied to the non-conservative forces *mutatis mutandis*, as shown in Carlberg *et al.* (2015, Sections 4.3–4.4). In particular, a reduced basis sparsification method and a gappy POD procedure can be used to approximate the reduced damping matrix $\mathbf{U}^\top \mathbf{C}(\mu)\mathbf{U}$ to maintain the Rayleigh-damping structure. Concerning the external force f_r, traditional DEIM or gappy POD can be employed since this function has no particular geometric structure.

Analogously to the Lagrangian case, non-conservative forces often consist of applied external forces and dissipative effects that can be modelled as perturbation of a Hamiltonian system; thereby

$$\begin{cases} \dot{\boldsymbol{u}} = \mathbf{J}_{2N}\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u}(t;\mu);\mu) + f(\boldsymbol{u}(t;\mu),t) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}(t_0,\mu) = \boldsymbol{u}_0(\mu). \end{cases} \tag{4.41}$$

Peng and Mohseni (2016*a*) considered non-parametric dynamical systems with forced Hamiltonian structure (4.41) in the special case where $f : \mathbb{V}_{2N} \times \mathcal{T} \to \mathbb{V}_{2N}$ is a *vertical* vector field, namely, it has zero component in the generalized coordinate $f(\boldsymbol{u},t) = (0, f_H(\boldsymbol{u},t))$. To obtain a reduced order model that retains a forced Hamiltonian structure, Peng and Mohseni (2016*a*) proposed performing a structure-preserving projection of the full model as described in Proposition 4.5, using an orthosymplectic reduced basis $\mathbf{U} \in \mathcal{U}(2n, \mathbb{R}^{2N})$. The resulting reduced dynamics possesses a forced Hamiltonian structure for any Hamiltonian functions $\mathcal{H}$ and force fields f_H, under the further constraint that the upper right block of $\mathbf{U}$ of size $N \times n$ vanishes; see Peng and Mohseni (2016*a*, Lemma 3.3). This condition ensures that the reduced force field is a vertical vector.

Moreover, the variation of the energy of the system is preserved by the reduced model if the force field f belongs to the reduced space at each instance of the reduced state $\mathbf{U}\boldsymbol{u}_r$ (or an equivalent condition is satisfied: see Peng and Mohseni 2016*a*, Theorem 3.13). Although this assumption can be hardly satisfied in practice, Peng and Mohseni (2016*a*) proposed several algorithms to construct the reduced space in such a way that the above property on the time derivative of the energy holds approximately.

In the framework of Peng and Mohseni (2016*a*), dissipative Hamiltonian systems can be seen as special forced Hamiltonian systems, where the system energy is decreasing with time. When the dissipation is Rayleigh dissipation as in (4.39), *i.e.* $f_H(\boldsymbol{u}) = f_H(\boldsymbol{q},\boldsymbol{p}) = -\mathbf{C}\boldsymbol{p}$, with $\mathbf{C}$ that might depend on $\boldsymbol{q}$ or a parameter μ, Peng and Mohseni (2016*a*) showed that the aforementioned structure-preserving approach exactly preserves the dissipativity independently of the construction of the reduced basis.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

A different approach is proposed in Maboudi Afkham and Hesthaven (2019) to deal with non-parametric linear dissipative problems. The idea is to include dissipation by recasting (4.41) as time dispersive and dissipative (TDD) (Figotin and Schenker 2007); thereby

$$\begin{cases} \dot{\boldsymbol{u}} = \mathbf{J}_{2N}\mathbf{K}^\top \boldsymbol{g}(t) \quad \text{for } t \in \mathcal{T}, \\ \boldsymbol{u}(t_0) = \boldsymbol{u}_0, \end{cases} \tag{4.42}$$

where $\mathbf{K} \in \mathbb{R}^{2N\times 2N}$ is such that the Hamiltonian reads $\mathcal{H}(\boldsymbol{u}) = (\boldsymbol{u}^\top \mathbf{K}^\top \mathbf{K}\boldsymbol{u})/2$, and $\boldsymbol{g}$ is the solution of the Volterra integral equation

$$\boldsymbol{g}(t) + \int_{t_0}^{t} \chi(t-s)\boldsymbol{g}(s)\,\mathrm{d}s = \mathbf{K}\boldsymbol{u},$$

where $\chi\colon \mathbb{R}^+ \to \mathbb{R}^{2N\times 2N}$ is a matrix-valued function (bounded in the Frobenius norm) called *general susceptibility*. Under suitable assumptions on χ (see Figotin and Schenker 2007 and Theorem 2 of Maboudi Afkham and Hesthaven 2019), problem (4.42) can be recast as an extended closed conservative system by coupling the full model with a canonical heat bath that absorbs the dissipated energy of the original system. This extension is obtained by defining a suitable Hilbert space $(\mathcal{H}^{2N}, \|\cdot\|_{\mathcal{H}^{2N}})$ and an isometric injection $I\colon \mathbb{R}^{2N} \to \mathbb{R}^{2N} \times \mathcal{H}^{2N}$, such that

$$\begin{cases} \dot{\boldsymbol{u}} = \mathbf{J}_{2N}\mathbf{K}^\top \boldsymbol{g}(t), \\ \partial_t \phi = \theta(t,x), \\ \partial_t \theta = \partial_x^2 \phi(t,x) + \sqrt{2}\delta_0(x)\cdot \chi^{1/2}\boldsymbol{g}(t), \\ \boldsymbol{u}(t_0) = \boldsymbol{u}_0,\ \theta(t_0,\cdot) = 0,\ \phi(t_0,\cdot) = 0, \end{cases} \tag{4.43}$$

where ϕ and θ are vector-valued functions in $\mathcal{H}^{2N}$, δ_0 is the Dirac delta and $\boldsymbol{g}$ solves $\boldsymbol{g}(t) + \sqrt{2}\chi^{1/2}\phi(t,0) = \mathbf{K}\boldsymbol{u}(t)$. Problem (4.43) is a conserved Hamiltonian system with Hamiltonian given by

$$\mathcal{H}_{\mathrm{ex}}(\boldsymbol{u},\phi,\theta) = \frac{1}{2}\left(\|\mathbf{K}\boldsymbol{u} - \phi(t,0)\|_2^2 + \|\theta(t)\|_{\mathcal{H}^{2N}}^2 + \|\partial_x\phi(t)\|_{\mathcal{H}^{2N}}^2\right).$$

By choosing a reduced basis that is orthosymplectic, the actions of model reduction and Hamiltonian extension commute and the reduced model inherits the closed Hamiltonian structure of the full model (Maboudi Afkham and Hesthaven 2019, Section 3.3). The reformulation of the model as an extended Hamiltonian system and its structure-preserving projection allows us to retain stability properties of the full model and to integrate the reduced system with a symplectic integrator, which guarantees conservation of the system energy and the correct dissipation of energy.

Port-Hamiltonian systems

Port-Hamiltonian systems rely on a system-theoretic network modelling paradigm that formalizes the interconnection of subsystems (van der Schaft 2006). They can

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

be formulated as dynamical systems of the form

$$\begin{cases} \dot{\boldsymbol{u}}(t) = (\mathbf{D}_{2N} - \mathbf{R}_{2N})\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u}(t)) + \mathbf{B}\boldsymbol{w}(t) & \text{for } t \in \mathcal{T}, \\ \boldsymbol{y}(t) = \mathbf{B}^{\top}\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u}(t)) & \text{for } t \in \mathcal{T}, \end{cases} \tag{4.44}$$

under suitable initial conditions. Here $\boldsymbol{w}\colon \mathcal{T} \to \mathbb{R}^m$ represents the system input, $\mathbf{D}_{2N} \in \mathbb{R}^{2N\times 2N}$ is a skew-symmetric matrix describing the interconnection of energy storage elements in the system, $\mathbf{R}_{2N} \in \mathbb{R}^{2N\times 2N}$ is a symmetric positive definite matrix, the *dissipation matrix*, describing energy loss in the system, $\mathbf{B} \in \mathbb{R}^{2N\times m}$ is the *port matrix* describing how energy enters and exits the system, and the Hamiltonian $\mathcal{H}$ is assumed to be a positive function. Port-Hamiltonian systems are characterized by stability and passivity via

$$\mathcal{H}(\boldsymbol{u}(t_2)) - \mathcal{H}(\boldsymbol{u}(t_1)) \leq \int_{t_1}^{t_2} \boldsymbol{y}(t)^{\top}\boldsymbol{w}(t)\,\mathrm{d}t, \tag{4.45}$$

and it is desirable to retain such properties in reduced models.

With the intention of closely mimicking the input–output response of systems of the form (4.44), Beattie and Gugercin (2011) developed Petrov–Galerkin reduced models that preserve the port-Hamiltonian structure, as follows. As explained at the beginning of Section 4.2.3, a projection-based reduced basis method consists in first constructing reduced spaces spanned by the bases $\mathbf{U}, \mathbf{W} \in \mathbb{R}^{2N\times 2n}$, $n \ll N$, respectively. Then the system trajectory $\boldsymbol{u} \in \mathbb{V}_{2N}$ is approximated by $\mathbf{U}\boldsymbol{u}_r$ so that the reduced system reads

$$\begin{cases} \dot{\boldsymbol{u}}_r(t) = \mathbf{W}^{\top}(\mathbf{D}_{2N} - \mathbf{R}_{2N})\nabla_{\boldsymbol{u}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t)) + \mathbf{W}^{\top}\mathbf{B}\boldsymbol{w}(t) & \text{for } t \in \mathcal{T}, \\ \dot{\boldsymbol{y}}_r(t) = \mathbf{B}^{\top}\nabla_{\boldsymbol{u}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_r(t)) & \text{for } t \in \mathcal{T}. \end{cases} \tag{4.46}$$

The reduced system (4.46) does not inherit the port-Hamiltonian structure of (4.44) and does not even satisfy the energy dissipation (4.45). The method proposed in Beattie and Gugercin (2011) ensures preservation of the port-Hamiltonian structure by assuming $\mathbf{W}^{\top}\mathbf{U} = \mathbf{I}$ and by approximating the gradient of the Hamiltonian in the space spanned by $\mathbf{W}$, *i.e.* $\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u}) \approx \mathbf{W}\mathbf{U}^{\top}\nabla_{\boldsymbol{u}}\mathcal{H}(\boldsymbol{u})$ for all $\boldsymbol{u}$ in the range of $\mathbf{U}$, or equivalently, the Hamiltonian $\mathcal{H}(\boldsymbol{u})$ is approximated by $\mathcal{H}(\mathbf{W}\mathbf{U}^{\top}\boldsymbol{u})$. The resulting approximate reduced system reads

$$\begin{cases} \dot{\boldsymbol{u}}_{a,r}(t) = (\mathbf{D}_{2n} - \mathbf{R}_{2n})\nabla_{\boldsymbol{u}_{a,r}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_{a,r}(t)) + \mathbf{W}^{\top}\mathbf{B}\boldsymbol{w}(t) & \text{for } t \in \mathcal{T}, \\ \dot{\boldsymbol{y}}_{a,r}(t) = \mathbf{B}^{\top}\mathbf{W}\nabla_{\boldsymbol{u}_{a,r}}\mathcal{H}(\mathbf{U}\boldsymbol{u}_{a,r}(t)) & \text{for } t \in \mathcal{T}, \end{cases} \tag{4.47}$$

where $\mathbf{D}_{2n} \coloneqq \mathbf{W}^{\top}\mathbf{D}_{2N}\mathbf{W}$, $\mathbf{R}_{2n} \coloneqq \mathbf{W}^{\top}\mathbf{R}_{2N}\mathbf{W}$, and $\mathbf{U}\boldsymbol{u}_{a,r}$ is an approximation of the full model solution $\boldsymbol{u} \in \mathbb{V}_{2N}$. The reduced basis $\mathbf{U}$ is constructed via POD to minimize the projection error of the matrix of the snapshots, while $\mathbf{W}$ minimizes the projection error in the Hamiltonian gradient evaluated at the snapshots; see Beattie and Gugercin (2011, Algorithm 1) for further details. Although such an approach preserves the port-Hamiltonian structure and ensures energy dissipation,

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

it does not guarantee the same energy distribution between dissipative and null energy contributors.

Note that the reduced dynamical system (4.46) inherits the port-Hamiltonian structure of the full model (4.44) without further approximation if

$$\mathbf{W}^\top \mathbf{D}_{2N} = \mathbf{D}_{2n}\mathbf{U}^\top \quad \text{and} \quad \mathbf{W}^\top \mathbf{R}_{2N} = \mathbf{R}_{2n}\mathbf{U}^\top$$

for a suitable $\mathbf{D}_{2n} \in \mathbb{R}^{2n\times 2n}$ skew-symmetric matrix and $\mathbf{R}_{2n} \in \mathbb{R}^{2n\times 2n}$ symmetric positive definite matrix.

Besides projection-based reduced basis techniques, dimension reduction of linear control systems and linear or quadratic bilinear port-Hamiltonian problems has also been developed via balanced truncation (Reis and Stykel 2008, Ionescu, Fujimoto and Scherpen 2010), interpolatory techniques and moment-matching (Freund 2003, Beattie and Gugercin 2009, Wolf, Lohmann, Eid and Kotyczka 2010, Polyuga and van der Schaft 2010, 2011, Benner and Breiten 2012, Gugercin, Polyuga, Beattie and van der Schaft 2012, Ionescu and Astolfi 2013, Beattie *et al.* 2019).

4.2.7. Hyper-reduction of nonlinear gradient fields

As described at the beginning of Section 4, hyper-reduction methods provide a further level of dimension reduction via convenient approximations of nonlinear and non-affine operators. Although these methods have led to the successful construction of inexpensive low-dimensional models, little attention has been paid to retaining, during the approximation, specific structures of the nonlinear operators, which are generally destroyed in the hyper-reduction step. This is, in particular, the case of gradient fields: if $X = \nabla\mathcal{H}$, for some $\mathcal{H}\colon \mathbb{R}^{2N}\times\mathbb{P}\to\mathbb{R}$, as in *e.g.* (4.23), the approximate vector field ensuing from the hyper-reduction of X is no longer a gradient, which implies that hyper-reduced models of Hamiltonian and Lagrangian systems do not result in Hamiltonian and Lagrangian flows, and hence the underlying physics is compromised. This drawback becomes particularly troublesome when aiming at stable and efficient model order reduction of conservative systems.

A symplectic discrete empirical interpolation method (SDEIM) has been proposed by Peng and Mohseni (2016*b*) and Maboudi Afkham and Hesthaven (2017), and consists in applying a DEIM approximation to the nonlinear Hamiltonian gradient. Although the proposed method does not exactly preserve the gradient structure of the Hamiltonian vector field, the energy of the system is shown to be bounded; see Peng and Mohseni (2016*b*, Theorem 5.1). Moreover, by choosing $(\mathbf{U}^+)^\top$ as the DEIM basis and constructing a sufficiently large reduced space via snapshots of the solution and of the nonlinear terms, the reduced dynamics possesses asymptotically (in n) a Hamiltonian structure.

Chaturantabut *et al.* (2016) have proposed a variation of the DEIM that preserves the Hamiltonian structure and consists in approximating the nonlinear Hamiltonian velocity field in the space where the DEIM projection is orthogonal. Since orthogonal projections preserve gradient structures, the Hamiltonian dynamics is retained, but there is no guarantee on the accuracy of the approximation.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

A different strategy consists in considering the reduced functions rather than the high-dimensional operators: first the full order nonlinear terms are mapped into the reduced space via structure-preserving projections, and then the resulting reduced functions are approximated. This is the idea behind the cubature approach called the energy-conserving sampling and weighting (ECSW) scheme, proposed in Farhat, Chapman and Avery (2015), that guarantees exact preservation of the gradient structure. The nonlinear vector field obtained from the semidiscretization of a Hamiltonian PDE is approximated with a weighted average of the field components on a coarser mesh. The application of this method is, however, limited to dynamical systems ensuing from the finite element discretization of PDEs, and requires a very expensive offline phase, especially for parametric problems.

4.3. Localized and adaptive methods

Traditional dimension reduction techniques are based on *global* and *linear* approximations of the solution space. However, some major challenges need to be faced when dealing with complex problems: (i) slowly decaying Kolmogorov n-widths, namely there might not exist a low-dimensional manifold on which lie *all* solutions obtained under variations of time and parameters (DeVore 2017); (ii) knowledge of the dynamics from simulation data needs to span a sufficiently large and informative portion of the parameter space and time horizon, but simulation data are expensive to collect; (iii) the reduced models derived need to be robust enough to deal with solution behaviour that was not encountered in the offline training phase. Nonlinear and adaptive model order reduction has been proposed in recent years to overcome the limitations of global reduced models based on approximations on linear subspaces.

In the following subsections we let $\eta \in \Gamma \subset \mathbb{R}^s$, $s \geq 1$, denote a generic quantity that represents the parameter of the problem of interest. Depending on the application and on the reference, η can be a time instant $t \in \mathcal{T}$, a parameter $\mu \in \mathbb{P}$ or a pair $(t; \mu) \in \mathcal{T} \times \mathbb{P}$. We will refer to it as a *generalized parameter*.

Localized reduced basis methods

A major issue concerns the accuracy of the reduced order model (ROM) solution, as, in many cases, the number of basis functions required to guarantee a certain error might be too large to benefit computationally from the dimension reduction. Assume we have an error estimate for the reduced solution (see Section 4.1). In the case in which the ROM solution is identified as not sufficiently accurate, a procedure that is easy to implement is to enrich the reduced basis by resorting to full model solutions from multiple parameter configurations, called *calibration points*, and then proceed with the enriched reduced order model (Weickum, Eldred and Maute 2008). These points are chosen either randomly or via adaptive search strategies, such as trust-region optimization. Since enrichment procedures require us to revert to the high-fidelity model and to solve the associated high-dimensional equations online, they incur large computational costs.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

A more effective approach is *localized model order reduction*, which hinges on an adaptive partition of the (generalized) parameter domain and a collection of low-dimensional approximate models associated with these clusters of parameters (Eftang, Patera and Rønquist 2010, Haasdonk, Dihlmann and Ohlberger 2011, Maday and Stamm 2013, Hess *et al.* 2019, Bonito *et al.* 2021). In the context of parametrized time-dependent PDEs, the idea is to consider a set of reduced order spaces $\{\mathbb{V}_r^{(k)}\}_{k=1}^K$, each of them spanned by a different set of reduced basis functions, namely,

$$\mathbb{V}_r^{(k)} = \operatorname{span}\{\xi_1^{(k)}, \ldots, \xi_n^{(k)}\}, \quad 1 \le k \le K.$$

To each reduced space is associated a local reduced order model that provides accurate approximations to the full order solution for parameters belonging to a certain subdomain of the parameter space. It is additionally required to construct an assignment function $\mathcal{F}\colon \Gamma \to \{1, \ldots, K\}$ that assigns the current time-parameter instance to one of the reduced spaces under consideration. This function will then be employed to select the set of reduced basis functions to be used for the expansion, namely (4.6) becomes

$$u_r(\eta) = \sum_{\ell=1}^{n} u_{r,\ell}(\eta)\, \xi_\ell^{(\mathcal{F}(\eta))}.$$

The main ingredients of this approach can be summarized as follows (see Hess *et al.* 2019):

- select sample points in the generalized parameter space Γ;
- cluster the corresponding snapshots so that each cluster is associated with parameters in a different part of the parameter domain;
- construct the local bases corresponding to each cluster;
- assign a cluster to any new parameter and use the corresponding local basis to construct a reduced order model.

Note that although the snapshots are used to determine the clusters, the identification and assignment of the local basis to a new parameter is performed in the online phase and thus cannot rely on the expensive snapshot but needs to be done in parameter space.

There are several approaches to creating dictionary models of this type, which differ mainly in the adaptive partitioning step of the space Γ. Many approaches are based on a notion of distance. Eftang *et al.* (2010) have considered stationary problems and proposed a hierarchical splitting of the parameter domain based on the proximity to chosen parameter anchor points in each subdomain. A reduced basis is then constructed on these subdomains by a greedy sampling procedure. Subdomains can be added in certain regions of the parameter space by adapting (offline) the training sample with the inclusion of new points at each level of the

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

h-refinement. A similar adaptive parameter domain partition, based on structured meshes, was proposed in Haasdonk *et al.* (2011).

In the localized MOR approach of Maday and Stamm (2013), the reduced solution for a certain given parameter $\eta = \mu \in \mathbb{P}$ is computed in the approximation space spanned by the (precomputed) basis functions whose corresponding parameter values lie in a ball around μ. The distance function used to define the ball takes into account the geometry of the solution manifold and the local anisotropies in the parameter space. It is constructed empirically by approximating the geodesic between two elements in the parameter space.

The splitting of the parameter space in localized MOR allows reduced spaces of lower dimension compared to global strategies, and it is particularly effective for problems in which the solution presents very different behaviours in different regions of the parameter domain. This is typical of bifurcation phenomena. Hess *et al.* (2019) have proposed a model for PDEs with bifurcating solutions in which the assignment function is based on the parameter belonging to one of the clusters computed via the k-means algorithm (MacQueen 1967) on the offline snapshots. Peherstorfer, Butnaru, Willcox and Bungartz (2014) have used k-means clustering and nearest neighbour classifiers with respect to parameters or a low-dimensional representation of the current state in combination with hyper-reduction strategies.

We remark that it is possible to operate the localization procedure with respect to the system state, and thus choose a local reduced space according to the subregion of the solution space where the current high-fidelity solution lies (Amsallem, Zahr and Farhat 2012). Similarly, Amsallem and Haasdonk (2016) use the projection error, rather than the Euclidean distance, as a partitioning criterion for a given set of snapshots.

Transformation-based model order reduction

A second family of model order reduction methods has been developed to deal with transport phenomena efficiently. Indeed, the application of reduced basis methods to convection-dominated problems and wave-type solutions might lead to poor approximations. Sufficiently large spaces are needed to achieve even moderate accuracy in the approximation of sets of solutions – characterized by travelling waves, moving shocks, sharp gradients and discontinuities – which do not possess a global low-rank structure and are therefore characterized by slowly decaying Kolmogorov n-widths. As a simple example, a shift of the solution cannot be represented by a linear combination of global modes.

To overcome the difficulties of approximating advection phenomena, a common strategy consists in recasting the problem in a coordinate frame where it is more amenable to reduction via a global approach. Given the solution manifold (4.3), the idea is to construct a mapped solution manifold

$$\widehat{\mathcal{M}}_h = \{\mathcal{G}_\eta(u_h(\eta)) \in \mathbb{V}_h \mid \eta \in \Gamma\},$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

that has a smaller Kolmogorov n-width compared to $\mathcal{M}_h$. The transformation $\mathcal{G}_\eta : \mathbb{V}_h \to \mathbb{V}_h$ is generally nonlinear; it depends on the generalized parameter $\eta \in \Gamma$ and possibly on hyperparameters. A particular case is $\mathcal{G}_\eta(u_h(\cdot, \eta)) = u_h(\cdot, \eta) \circ \Phi_\eta$, where $\Phi_\eta : \Omega \to \Omega$ represents a suitable change of coordinates of the spatial domain Ω, so that the reduced solution (4.6) has the form

$$u_r(\eta) = \sum_{\ell=1}^{n} \mathrm{u}_{r,\ell}(\eta)\, \xi_\ell \circ \Phi_\eta^{-1}.$$

In this framework, albeit for problems with no dependence on parameters, Rowley and Marsden (2000) have focused on removing the discrete translational symmetries by applying POD in a shifted frame of reference, with the travelling speed determined using template fitting and reconstruction. A generalization to self-similar solutions was proposed later by Rowley, Kevrekidis, Marsden and Lust (2003) with the implementation of both translation and scaling in space and time. Sesterhenn and Shahirpour (2019) also proposed a space–time transformation of the solution snapshots to construct a reference frame along the characteristics defined by the group velocity of travelling waves. The resulting approach, called *characteristic DMD*, hinges on a modal decomposition in space and time along the characteristics, which allows a POD or DMD with only a few modes. With a similar goal, *shifted POD* (Reiss, Schulze, Sesterhenn and Mehrmann 2018) and *transport reversal* (Rim, Moe and LeVeque 2018) introduce time-dependent shifts of the snapshot matrix based on the dominant transport velocities of the problem. Reiss *et al.* (2018) determined the velocities by front tracking or by considering the dependence of the singular values of the shifted snapshot matrix on the time-dependent shift. The corresponding low-dimensional subspace is then used to build a reduced model using standard techniques such as Galerkin projection. Rim *et al.* (2018) developed a greedy generalization of template-fitting to approximate the snapshot matrix by the superposition of multiple transport dynamics.

Several other techniques have looked for a spatial change of coordinates for better representation of the solution manifold. The *transformed snapshot interpolation* (Welper 2017) considers the problem of interpolation of parameter-dependent jumps and kinks. The idea is to perform a transformation Φ_η, for any parameter η, that allows us to align the discontinuities of a set of snapshots to the discontinuity of the target function $\boldsymbol{u}(\eta)$. The transformation Φ_η is computed by solving a non-smooth optimization problem that minimizes the sup-norm over the parameter domain of the approximation error on a training sample of snapshots. With a similar goal of constructing the frame of reference of the solution, a Lagrangian projection-based model order reduction has been proposed in Mojgani and Balajewicz (2017). This approach was motivated by the observation that certain wave-type and moving shock solutions exhibit low-rank structures in the Lagrangian frame of reference. A projection-based approach using global basis functions is to approximate both the state variable and the Lagrangian computational grid that is evolving in time.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Cagniart, Maday and Stamm (2019) have constructed mapped solution snapshots offline based on snapshots of the high-fidelity solution. In the online phase, the map Φ_η is computed, for any new instance of η, via a minimization of the fully discrete equation residual, in the spirit of shock-fitting methods. The algorithm simultaneously looks for an appropriate change of coordinates Φ_η, and thus a reduced basis, and for the expansion coefficients $\{u_{r,i}(\eta)\}_i$ in this reduced basis. A related approach is the *method of freezing* (Ohlberger and Rave 2013, Beyn and Thümmler 2004), where the transformation of the solution manifold is interpreted in the frame of Lie groups as a decomposition of the full model solution into a phase/group component Φ_η and a shape component $v := u(\cdot, \eta) \circ \Phi_\eta$. Such a decomposition can be performed for arbitrary Lie group actions, provided they satisfy an equivariance condition (Ohlberger and Rave 2013, equation (4)). Rather than solving an optimization problem as in Cagniart *et al.* (2019), the full model is reformulated as a partial differential equation for the shape component, which is much simpler to solve than the original model. The phase component is determined via a set of algebraic constraints, called *phase conditions*. A further method in this direction has been proposed by Iollo and Lombardi (2014) in a non-parametric setting, *i.e.* $\eta = t$. The method approximates the full model solution with a modal decomposition that takes into account advection via a ranked sequence of mappings and diffusion via a ranked global mode expansion. The advection modes are obtained by fitting the Wasserstein distance between snapshots and modes derived from Monge–Kantorovich optimal transport problems relative to a reference mode $\boldsymbol{u}_0$. However, the method is limited to problems where the solution exhibits such a main mode.

The optimization procedure for the construction of Φ_η in Iollo and Lombardi (2014) is related with the *registration method* of Taddei (2020). In this approach for stationary parametrized PDEs, a bijective map Φ_η is constructed from a set of snapshots with the aim of low-dimensional representations of the mapped solution manifold. The method is based on a nonlinear non-convex minimization of the difference between a reference state $u_h(\overline{\eta})$ and the mapped snapshots $u_h(\eta) \circ \Phi_\eta$ at the training parameters. To determine the transformation Φ_η for all parameters in $\mathbb{P}$ outside the training set, a kernel-based multitarget regression procedure is proposed.

Ehrlacher *et al.* (2020) have proposed a data-driven approach that targets one-dimensional transport problems characterized by slowly decaying Kolmogorov n-width. The idea is to perform a nonlinear model reduction in metric spaces via tangent principal component analysis or a barycentric greedy algorithm.

Online adaptive model order reduction

The main limitation of most localized and transformation-based approaches lies in their *a priori adaptivity*, that is, the construction of the different reduced order models is done during the offline phase, following which it is therefore not possible to incorporate new information into the model. Online adaptive model order

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

reduction techniques have been developed to overcome this limitation by updating the reduced space online according to various criteria associated with changes of the system dynamics in parameters and time.

One of the first examples of online adaptive MOR relies on subspace enrichment by refining the reduced order space based on *a posteriori* estimates of the reconstruction error. The method proposed in Carlberg (2015) draws inspiration from mesh-adaptive h-refinement to generate a hierarchy of subspaces online. Similarly to other enrichment techniques (*e.g.* Weickum *et al.* 2008), the refinement is built following a tree structure constructed offline but the enrichment step does not require high-fidelity solves. Starting from an initial reduced basis and ROM solution, the method identifies a set of basis vectors using a dual-weighted residual approach that aims to reduce error in an output quantity of interest. The reduced basis is refined by splitting the selected reduced basis vectors into multiple vectors with disjoint support. This procedure is justified by the heuristic according to which two strongly correlated state variables can be described by the same reduced variables and must belong to the support of the same reduced basis function.

Another class of methods is based on the construction of reduced basis spaces $\mathbb{V}_r$ that directly depend on time,

$$\mathbb{V}_r^{(t)} = \text{span}\{\xi_1(t), \ldots, \xi_n(t)\},$$

so that the reduced solution is given as an expansion with time-dependent basis and coefficient, namely,

$$u_r(t;\mu) = \sum_{\ell=1}^{n} \mathrm{u}_{r,\ell}(t;\mu)\,\xi_\ell(t), \quad t \in \mathcal{T},\ \mu \in \mathbb{P}. \tag{4.48}$$

Specifically, this type of procedure entails two sets of operations: (i) compute or infer the expansion coefficients with respect to the adaptive basis functions, and (ii) update the basis.

Low-rank updating of the reduced operators is at the basis of dynamic data-driven model reduction (Peherstorfer and Willcox 2015*a*), in which the SVD updating method (Brand 2006) is exploited to compute the reduced operators of the current iteration as a function of the previous ones. Although the resulting reduced models adapt directly from sensor data without recourse to the full model, the sensor samples need to measure the high-dimensional state of the system. This limitation does not preclude the use of the methods in applications where dynamic sensor data are available, such as structural assessment. Moreover, a generalization to the case of incomplete sampling has been proposed in Peherstorfer and Willcox (2016*b*). The aforementioned dynamic data-driven methods consider problems with affine parameter dependence in both the linear operator and the source term. Although the affine decomposability of more general operators can be obtained via hyper-reduction strategies (see Section 4), a low-rank online adaptation of

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

nonlinear operators has been developed in Peherstorfer and Willcox (2015*b*). The *adaptive discrete empirical interpolation method* (ADEIM) of Peherstorfer and Willcox (2015*b*) and Zimmermann, Peherstorfer and Willcox (2018) adapts the DEIM reduced space at each time step via a low-rank correction of the reduced basis from a previous time step. The update is obtained by querying the full order system, in a certain temporal window, at a few selected components in order to guarantee a computationally efficient adaptation.

An alternative to a discrete-in-time update of the (hyper-) reduced space consists in prescribing a differential evolution equation for the reduced basis. This is done in the *approximated Lax pair* (ALP) algorithm of Gerbeau and Lombardi (2014), which deals with non-parametric evolution PDEs characterized by the propagation of fronts. In this work, the time-dependent reduced basis functions are the eigenfunctions of a linear Schrödinger operator $\mathcal{L}(\boldsymbol{u}(t))\cdot = -\Delta\cdot - \chi\boldsymbol{u}(t)$, where χ is a positive constant and $\boldsymbol{u}(t)$ is the solution of the PDE of interest at time t. These eigenfunctions are shown to satisfy an evolution PDE. The original PDE is projected into the reduced space associated with the time-dependent basis and, in a final post-processing step, the reduced solution is reconstructed on the full order space by propagating the reduced order basis in time with an approximation of a Lax operator. Although the ALP algorithm provides an interesting criterion to select the reduced basis, it does not appear to cure the large increase in the dimension of the reduced space when the accuracy requirement is tightened.

The idea of approximating solutions of evolution problems via a modal decomposition with time-dependent modes, as in (4.48), has a surprisingly long history and is common to various fields. Low-rank approximations based on such decompositions have already been widely studied in quantum mechanics by Dirac (1930), and later in the *multiconfiguration time-dependent Hartree* (MCTDH) method; see *e.g.* Beck, Jäckle, Worth and Meyer (2000) and Lubich (2008). A related approach, known as *dynamical low-rank approximation*, has been developed by Koch and Lubich (2007) for the approximation of time-dependent data matrices. For the discretization of time-dependent stochastic PDEs, Sapsis and Lermusiaux (2009) proposed the so-called *dynamically orthogonal* (DO) scheme, where the deterministic approximation space adapts over time by evolving according to the differential operator describing the stochastic problem. All these methods share a common paradigm: an approximation of the solution, possibly a matrix-valued function, is obtained by projecting its time derivative onto the tangent space of a reduced manifold at the current approximation. In the context of model order reduction of evolution equations, this means that the full model solution is approximated by the trajectory in the reduced space associated with the best approximation of the velocity field in the tangent space to the reduced manifold. A connection between dynamical low-rank approximations and DO methods was established in Musharbash, Nobile and Zhou (2015) and a geometric interpretation of this connection was investigated in Feppon and Lermusiaux (2018). In recent years, dynamical model order reduction and DO discretizations have also addressed the challenge of

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

preserving geometric structures (see Section 4.2). A DO discretization of stochastic PDEs that possesses a symplectic Hamiltonian structure (4.10) at the semidiscrete level has been developed in Musharbash and Nobile (2017). In Pagliantini (2021), reduced models for parametrized Hamiltonian systems are derived using a dynamical approach that yields a low-dimensional Hamiltonian system coupled with an evolution equation for the reduced basis dictated by the original dynamics. Pagliantini (2021) also proposed a structure-preserving temporal discretization of these reduced models. Hesthaven, Pagliantini and Ripamonti (2022) showed that a dynamical MOR approach, with adaptive size of the reduced basis, allows a small reduced basis, improving accuracy and efficiency of the reduced models.

While all the aforementioned adaptive procedures achieve speed-ups compared to traditional reduced models, their computational cost scales with the complexity of the full model. The pay-off is the absence of a – generally expensive – offline phase.

4.4. Data-driven methods

The data-driven revolution represents one of the most important consequences of the advent of the new digital era, and is destined to change the way of working and managing data and information in many sectors. In the last twenty years, so-called 'big data' has been the driving force behind the digital transformation: both the data collection and the underlying management process changed and was enhanced by technological developments in the sensor field and the growing availability of low-cost media for their storage. However, despite the former collection occurring in all sectors and with low costs, it is not sufficient for the generation of insights. In particular, a greater amount of information generally translates into a more difficult analysis. So the real revolution is the enormous progress in *data science* (Baesens 2014). This discipline aims to extract insights of various kinds from collected statistics, to open up new strategies for decision support and predictions of scenarios at different scales. This is done through the intersection of a wide range of multidisciplinary fields, including statistics, data mining, machine learning, complex systems, network science and applied mathematics.

However, despite the aforementioned impact of the data-driven revolution, the consequences it will have on the scientific sector, which is a major producer of big data, are unknown. For example, it is estimated that the Square Kilometre Array project (Dewdney, Hall, Schilizzi and Lazio 2009) (an intergovernmental radio telescope) will generate up to 1 *exabyte* (10^{18} *bytes*) of raw data per day, which is more than the daily global mobile data registered in 2017 (CISCO 2019). This amount of data cannot currently be analysed or understood via traditional methods. In fact, in the past, the collection of data in science and engineering was exclusively motivated by the validation of hypotheses on the models and theories proposed to describe certain phenomena. As such, data collection was based on the creation of *ad hoc* experiments, which made the process sporadic and difficult to implement.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Data science is promoting a rapid paradigm shift in the scientific sector too: the wide availability of large datasets can lead to the discovery of patterns that can be incorporated into the modelling process and used to better understand the complexity of the phenomena under investigation. A prime example is the concept of the digital twin, an industry 4.0 paradigm that aims to create a virtual replica of physical, potential and actual resources connected to objects, processes, people, places, infrastructure, systems and devices. Since a static digital is incapable of delivering the performance needed in the modern digital era, one must provide the model with dynamic integration of real-time information coming from sources such as sensors, external services and simulations. In particular, this change in approach was initially pioneered in those sectors that did not have adequate modelling tools, leading to very successful applications, such as for recommender systems. However, recently data-driven approaches have also affected traditional sectors, which had so far been able to provide excellent results both in terms of modelling and design. For example, in the field of computational fluid dynamics (CFD), several data-driven models (Brunton, Noack and Koumoutsakos 2020) have been proposed to address problems such as the development of turbulence closure models (Hijazi, Stabile, Mola and Rozza 2020), shock wave propagation (Dupuis, Jouhaud and Sagaut 2018), optimal flow control (Novati, Mahadevan and Koumoutsakos 2019), velocity reconstruction for PIV images (Semeraro, Bellani and Lundell 2012, Lee, Yang and Yin 2017) and hydrodynamic instabilities (Stegeman, Ooi and Soria 2015, Gao *et al.* 2021). For a more complete review on the topic we refer to Rozza *et al.* (2018) and Yu, Yan and Guo (2019) for the CFD field and to Montáns, Chinesta, Gómez-Bombarelli and Kutz (2019) for a broader discussion of applications to other science and engineering fields.

Non-intrusive reduced basis method

In the field of reduced order modelling, the most popular data-driven approaches are surrogate *data-fit* (Eldred and Dunlavy 2006) reduced models. The term surrogate refers to replacing the original full order system with a reduced counterpart that allows for real-time evaluation, while the data-fit nature concerns the use of *interpolation* or *regression* to map the input parameters to quantities of interest. These methods represent an alternative to the classical projection-based approach, in which the full order operators of the original problem are projected onto a reduced order space of a much smaller dimension.

Both data-driven and projection-based methods start with a set of full order trajectories, which are compressed to unveil a lower-dimensional representation of the solution manifold. This can be accomplished in various ways, using for example linear methodologies such as POD, but also non-linear methods such as *autoencoders*. The substantial difference lies in how the evolution of the general coordinates is recovered in the compressed solution space. As anticipated, in the case of projection-based approaches this result is obtained by employing a Galerkin or least-squares projection of the original system onto the reduced space.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

This means having access to the high-dimensional algebraic system, but it has the big advantage of preserving the underlying structure of the original problem, and providing a framework for deriving rigorous error estimates for the reconstructed quantities.

Data-driven methods try to infer the map between the parameters and the expansion coefficients of the solution by using only the data. This approach allows us to operate in a *non-intrusive* manner, that is, the simulation software can be considered as a *black box* that produces the datasets upon which the model is built, following the prescription of an appropriate set of parameters. Non-intrusive approaches are easier to implement, and since there is no need to access the underlying system of equations, one can make use of commercial codes for the data generation. However, a purely data-driven technique usually lacks conservation of physical principles and a rigorous error certification.

Recently some methods have been proposed to deal with this limitation, by integrating the knowledge of the model within the data learning framework; this is what is done in physics-informed neural networks (PINNs) (Raissi, Perdikaris and Karniadakis 2019), *i.e.* neural networks that are trained to respect the imposition of a generic nonlinear partial differential equation. An alternative approach results from applications in the context of *system identification* (Brunton, Proctor and Kutz 2016), which attempt to reconstruct the system dynamics from noisy samples of the trajectories.

Another major limitation concerns the need for a large amount of data, and in fact, within the data-scarce regime, the projection-based methods generally perform better. This last scenario is of crucial importance, since for high-dimensional complex systems, obtaining solutions to use for training the model can be prohibitively expensive. Examples of this situation can be found in fields such as geophysics, neuroscience and fluid-dynamic turbulence, which differ greatly from traditional machine learning tasks where training data are massively available.

In the following, we will review the most used techniques in the field of data-driven non-intrusive model order reduction. In particular, in Section 4.4.1 we will review some of the best-known techniques for dimension reduction. In Section 4.4.2 we will discuss *proper orthogonal decomposition with interpolation* (PODI), an abstract framework that builds an interpolation phase on top of POD-based ROMs. Finally, Section 4.4.4 presents the algorithm known as *dynamic mode decomposition* (DMD), which uses regression to learn adaptively the best-fit linear dynamical model.

4.4.1. Dimension reduction

Let $\boldsymbol{u}_h^k(\mu) \in \mathbb{R}^N$ denote the vector of degrees of freedom associated with the discrete truth solution $u_h^k(\mu)$ at time t^k belonging to the solution manifold $\mathcal{M}_h^{N_T}$ in (4.8). We consider the snapshot matrix

$$\mathbf{S} = \left[\boldsymbol{u}_h^0(\mu_1) \mid \ldots \mid \boldsymbol{u}_h^{N_T}(\mu_1) \mid \ldots \mid \boldsymbol{u}_h^0(\mu_M) \mid \ldots \mid \boldsymbol{u}_h^{N_T}(\mu_M)\right] \in \mathbb{R}^{N \times N_s}, \quad (4.49)$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

where $N_s := M(N_T + 1)$ is the total number of samples. To ease the notation, we combine the indices associated with time and parameter into a single index, so that $\boldsymbol{u}_h^{(j)} := \boldsymbol{u}_h^k(\mu_m)$, where

$$j = (m-1)(N_T+1) + (k+1), \quad 0 \le k \le N_T, \quad 1 \le m \le M.$$

The snapshot matrix can then be written as

$$\mathbf{S} = \left[\boldsymbol{u}_h^{(1)} \mid \ldots \mid \boldsymbol{u}_h^{(N_s)}\right] \in \mathbb{R}^{N \times N_s}.$$

Proper orthogonal decomposition, discussed in Section 4.1, can also be implemented on the Gram matrix $\mathbf{C} = \mathbf{S}^\top \mathbf{S}$ of inner products of the snapshots.

Some of the disadvantages to using POD concern the loss of information related to the phase of modes and small field fluctuations in CFD, as well as the difficulty in representing moving discontinuities, due to the linearity of the method. In general, nonlinear reduction techniques are found to better span the nonlinear manifold associated with a parametrized PDE, which is why they have also become popular in the ROM community (see Section 4.3).

A nonlinear generalization of proper orthogonal decomposition is kernel POD (KPOD) (Schölkopf, Smola and Müller 1997). The idea of KPOD is to project the full order snapshots into a high-dimensional space via a nonlinear map and then to apply classical POD in the mapped space. KPOD starts by computing the kernel (similarity) matrix $\mathbf{K} \in \mathbb{R}^{N_s \times N_s}$ as

$$(\mathbf{K})_{ij} = \kappa\left(\boldsymbol{u}_h^{(i)}, \boldsymbol{u}_h^{(j)}\right),$$

where $\kappa \colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a bivariate symmetric form, referred to as a *kernel function*. A popular choice for the kernel function is given by radial basis functions, such as the squared exponential

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right) \quad \text{with } \gamma \in \mathbb{R}_+.$$

The result of the projection of the full order model (FOM) solutions onto the principal components is then given by

$$\boldsymbol{u}_{r,i} = \left[\sum_{j=1}^{N_s} \frac{(\boldsymbol{v}_1)_j}{\sqrt{\lambda_1}} \kappa\left(\boldsymbol{u}_h^{(j)}, \boldsymbol{u}_h^{(i)}\right), \ldots, \sum_{j=1}^{N_s} \frac{(\boldsymbol{v}_d)_j}{\sqrt{\lambda_d}} \kappa\left(\boldsymbol{u}_h^{(j)}, \boldsymbol{u}_h^{(i)}\right)\right],$$

where $\{(\boldsymbol{v}_i, \lambda_i)\}_{i=1}^d$ are the first d dominant eigenvalue–eigenvector pairs of $\mathbf{K}$, and the notation $(\boldsymbol{v}_i)_j$ refers to the jth component of vector $\boldsymbol{v}_i$. Note that unlike the POD procedure, here the projection is performed without having an explicit expression for the principal components, that is, the high-dimensional feature space is used only implicitly via the kernel function.

Other reduction techniques try to optimize a non-convex objective function, which therefore contains local optima. An example in this context is the Sammon mapping (Sammon 1969), in which the contribution of each sample pair in the cost function is weighted by their mutual distance. The minimization, in this case, can

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

be conducted through the Newton method. A more important example, because of its recent wide diffusion, is given by the use of neural networks (NNs) (Goodfellow, Bengio and Courville 2016). In Section 4.4.2 we will introduce the main concepts for this type of regression, and discuss their application in retrieving the map from the parameter space to the space of expansion coefficients with respect to the basis obtained by POD. However, a special type of NN, known as an *autoencoder* (Wang, Yao and Zhao 2016), also plays a key role in the preliminary compression phase of the solution manifold. This particular architecture of unsupervised neural network aims at reconstructing an identity map, *i.e.* $\widetilde{\mathcal{I}}\colon \boldsymbol{u}_h \mapsto \widetilde{\boldsymbol{u}}_h$. The lower-dimensional representation is obtained via an *encoder* architecture $\mathcal{F}_{\text{en}}(\boldsymbol{w}_{\text{en}})\colon \boldsymbol{u}_h \mapsto \boldsymbol{u}_r$, whose output is subsequently mapped back to the high-dimensional space by a *decoder* network $\mathcal{F}_{\text{de}}(\boldsymbol{w}_{\text{de}})\colon \boldsymbol{u}_r \to \widetilde{\boldsymbol{u}}_h$ (see Figure 4.1). The network parameters $(\boldsymbol{w}_{\text{en}}, \boldsymbol{w}_{\text{de}})$ are obtained via a training process, which minimizes a given loss function,

$$\left(\boldsymbol{w}^*_{\text{en}}, \boldsymbol{w}^*_{\text{de}}\right) = \underset{\boldsymbol{w}_{\text{en}}, \boldsymbol{w}_{\text{de}}}{\arg\min}\, \mathcal{L}(\widetilde{\boldsymbol{u}}_h, \boldsymbol{u}_h),$$

where $\mathcal{L}$ is typically the mean square error between the input and the output of the network.

The main limitation in the use of autoencoders is the high number of parameters to train, in the case of systems with high-dimensional state. In fact, the number of parameters of fully connected autoencoders far exceeds the size of the state itself. For this reason it is common to use *convolutional autoencoders* (Masci, Meier, Cireşan and Schmidhuber 2011), which have layers characterized by the sharing of parameters and local connectivity, that is, the nodes of a feature map are linked only to a local number of the input vector, and therefore extract local features. The above properties allow considerable reduction of the number of parameters with respect to feedforward autoencoders, making it possible to deal with slowly decaying Kolmogorov n-width problems, such as advection-dominated problems (Lee and Carlberg 2020). Following the same idea, a deep-learning-based ROM (POD-DL-ROM) was introduced in Fresca and Manzoni (2022).

Other dimension reduction techniques which result in solving a generalized eigenvalue problem for a full matrix are Isomap (Tenenbaum, de Silva and Langford 2000), maximum variance unfolding (Weinberger, Sha and Saul 2004) and diffusion maps (Lafon and Lee 2006). These are all global techniques, in that like POD they attempt to maintain the globalized properties of the full order data in the reduced order representation. There are also local methods that operate this conservation of solution properties only in the neighbourhood of the solution samples of the training dataset. To this category belong the Laplacian eigenmaps (Belkin and Niyogi 2001), local tangent space analysis (Zhang and Zha 2004) and t-distributed stochastic neighbour embedding (van der Maaten and Hinton 2008).

The above techniques can act on any set of raw data, although they may prove more efficient for dealing with a particular type of problem; for instance, convolutional NNs are the standard network architecture for computer vision tasks.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

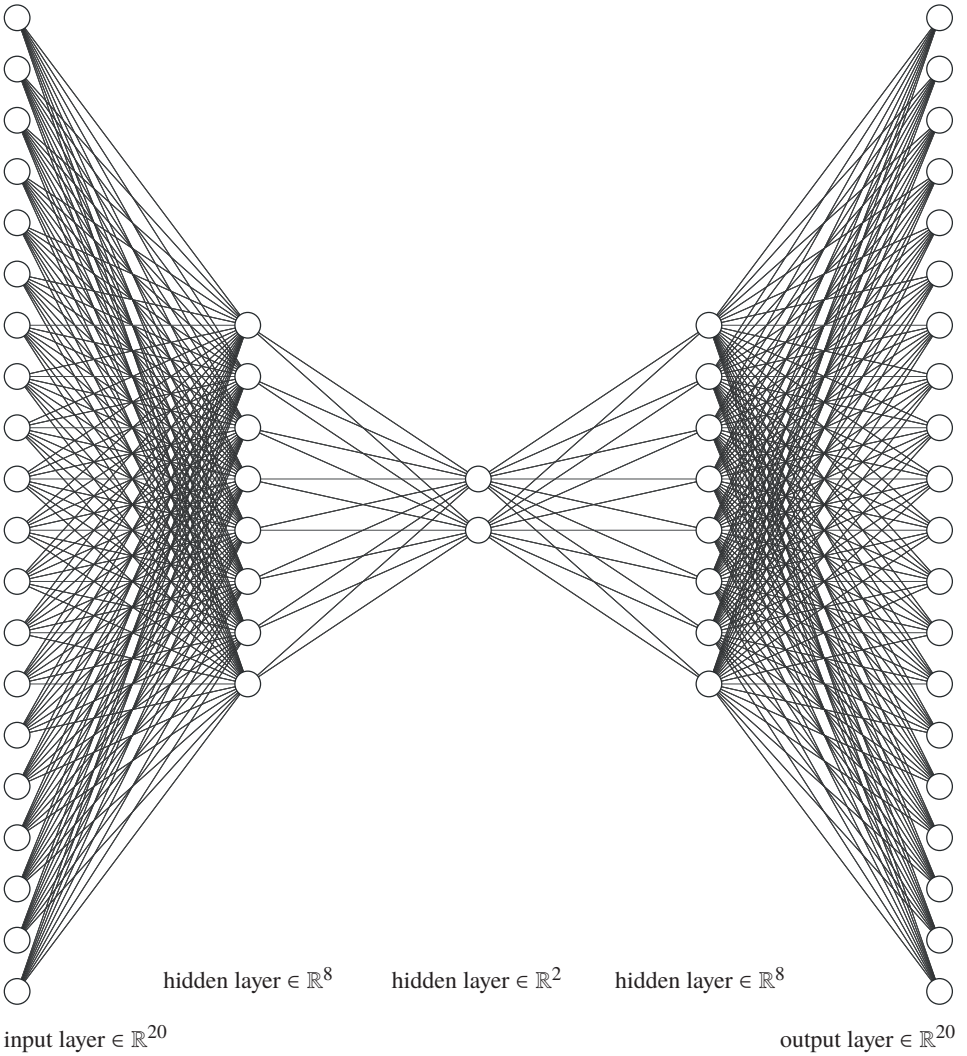

Figure 4.1. The architecture of a fully connected autoencoder: the latent *reduced* representation of the input is given by the values in the middle hidden layer.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

One technique that has not yet been mentioned but will be discussed separately in Section 4.4.4 is dynamic mode decomposition (DMD), which provides an eigen-decomposition for the best-fit linear operator that describes the temporal dynamics of snapshots in a given dataset.

Parameter space reduction

Data-driven reduced order models were developed to address problems involving the simulation of complex systems and requiring high computational resources. However, all data-driven ROMs demand the preliminary calculation of a set of high-fidelity solutions, whose number depends on the number of parameters taken into consideration. In this context it is typical to speak of the *curse of dimensionality*, referring to the fact that the amount of data required to obtain a reliable model grows exponentially with the dimension of the parameter space. This is problematic for methods that require a significant amount of data to be valid, making them inefficient or inoperative otherwise. Typical examples in this regard are optimization, optimal control, uncertainty quantification and inverse problems. In these applications the number of parameters can be $O(N^s)$, where N is the state dimension and s is a positive integer, an exponent dictated by the particular problem (Himpe and Ohlberger 2015). Since the industrial field widely exploits the former types of analysis, this plays a crucial role from an application perspective.

The above considerations make it necessary to operate with a reduced number of parameters, as the possibility to explore the solution manifold effectively during the offline stage is strictly connected to the cardinality of the set of independent parameters considered. For this reason, various strategies have been proposed to operate a compression of the parameter space, as is done for the state space.

A simple example involves the use of Galerkin or Petrov–Galerkin projections for the low-rank approximation of the parameter vector $\mu \in \mathbb{P} \subset \mathbb{R}^P$. In such a case one will be operating with a reduced parameter vector $\mu_r \in \mathbb{R}^{P_r}$ given by

$$\mu_r = \mathbf{P}\mu \in \mathbb{R}^{P_r},$$

where $\mathbf{P} \in \mathbb{R}^{P_r \times P}$ is the projection matrix into the low-dimensional parameter space, and satisfies $\mathbf{P}^\top \mathbf{P} = \mathbf{I}$. The matrix $\mathbf{P}$ can be obtained, for example, via an iterative procedure based on a greedy approach, which tries to minimize the reconstruction error. For this purpose, the reduced parameter set is obtained by iteratively maximizing an appropriate objective function, such as the one proposed by Bashir *et al.* (2008), that is,

$$J(\theta) = \alpha\, \epsilon(\boldsymbol{u}_r(\mu_r)) - \beta \|\mu_r\|^2_{\mathbf{K}^{-1}},$$

where $\alpha, \beta \in [0, 1]$ are appropriate hyperparameters, while

$$\|\mu_r\|^2_{\mathbf{K}^{-1}} := (\mathbf{P}\mu_r)^T \mathbf{K}^{-1} (\mathbf{P}\mu_r)$$

is the weighted 2-norm induced by the inverse of the prior covariance matrix $\mathbf{K} \in \mathbb{R}^{P \times P}$. The quantity $\epsilon(\boldsymbol{u}_r(\mu_r))$ is an estimate of the reconstruction error, which can

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

be obtained for example by using a p-norm, *i.e.* $\epsilon(\boldsymbol{u}_r(\mu_r)) = \|\boldsymbol{u}_r(\mu_r) - \boldsymbol{u}_r(\mu)\|_p$, with $p \in [0, +\infty)$. Clearly, since the former reconstruction error may be excessively expensive to evaluate, it is convenient to use a suitable *a posteriori* error estimation.

Another conceivable approach exploits a possible low-dimensional structure directly at the level of the model function. In this context a very popular data-driven technique is the *active subspace* (AS) method (Constantine, Dow and Wang 2014, Zahm, Constantine, Prieur and Marzouk 2020), which aims at constructing a ridge approximation of the function by using the gradient of the function itself. A ridge function is a real-valued multivariate function $F\colon \mathbb{R}^P \to \mathbb{R}$ such that

$$F(\mu_1, \mu_2, \dots, \mu_P) = F(a_1\mu_1 + a_2\mu_2 + \cdots + a_P\mu_P) = f(\boldsymbol{a} \cdot \mu),$$

where $\boldsymbol{a} \in \mathbb{R}^P$ is called *direction* and is such that F is constant on the hyperplanes $\boldsymbol{a} \cdot \mu = c \in \mathbb{R}$. A generalization of the above concept concerns the existence of multiple directions, which gives rise to generalized ridge functions:

$$F(\mu_1, \mu_2, \dots, \mu_P) = f(\mathbf{A}^\top \mu),$$

where $\mathbf{A} \in \mathbb{R}^{P\times s}$ is a matrix with $1 \le s < P$, and $f\colon \mathbb{R}^s \to \mathbb{R}$.

The AS method aims at approximating a prescribed target function with a generalized ridge function. This is done by finding the leading eigenspaces of the second moment matrix of the model function's gradient. Through this procedure it is possible to identify and exploit the directions of maximum variability of the model function, to build its low-dimensional representation, chosen through the resolution of the minimization problem

$$\underset{P_r \in \mathbb{R}^{P\times P}}{\arg\min}\, \mathbb{E}_\rho\left[\|F(\mu) - \widetilde{f}(P_r\mu)\|_V^2\right], \tag{4.50}$$

with the following quantities:

- $\rho\colon \mathbb{R}^P \to \mathbb{R}$ is the probability density of μ with support $\mathbb{P} \subset \mathbb{R}^P$;
- $P_r : \mathbb{R}^P \to \mathbb{R}^P$ is the r-rank projection operator within the parameter space;
- $V = (\mathbb{R}^s, \mathbf{R}_V)$ is the Euclidean space for the output, whose distance is induced by the norm $\|\boldsymbol{y}\|_V := \boldsymbol{y}^T\mathbf{R}_V\boldsymbol{y}$, with $\mathbf{R}_V \in \mathbb{R}^{s\times s}$ a symmetric and positive definite matrix;
- $\widetilde{f}\colon (\mathbb{R}^P, \mathcal{B}(\mathbb{R}^P), \rho) \to V$ is the ridge-regression profile associated to the target $F\colon \mathbb{P} \subset \mathbb{R}^P \to V$, with $\mathcal{B}(\mathbb{R}^P)$ the Borel σ-algebra of $\mathbb{R}^P$.

The linear space $\mathrm{Im}(P_r)$ will be the reduced parameter space and the profile $\widetilde{f}$ is such that

$$\widetilde{f} \circ P_r = \mathbb{E}_\rho[F \mid \sigma(P_r)], \tag{4.51}$$

for a given σ-algebra $\sigma(P_r)$. For a scalar model function, the covariance matrix of ∇F can be replaced by

$$\mathbf{H} = \int_{\mathbb{P}} (D_\mu F(\mu))^T \mathbf{R}_V(\rho)(D_\mu F(\mu))\, \mathrm{d}\rho \in \mathbb{R}^{P\times P},$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

where $D_\mu F(\mu)$ is the Jacobian of F. The solution to the minimization problem (4.50) is then given by

$$P_r = \sum_{j=1}^{r} \boldsymbol{v}_j \otimes \boldsymbol{v}_j,$$

where $\{\boldsymbol{v}_j\}_{j=1}^{r}$ are the r leading eigenvectors of $\mathbf{H}$. The dimension r is chosen in order to satisfy an upper bound for $\mathbb{E}_\rho[\|F(\mu) - f(P_r\mu)\|_V^2]$ obtained by a Poincaré-like inequality (Zahm *et al.* 2020). Once P_r is computed, the response surface in equation (4.51) can be evaluated using the *Monte Carlo* method or exploiting regression (Constantine *et al.* 2014).

Other techniques that aim at identifying low-dimensional structures within the parameter space are the *sliced inverse regression* (SIR) method (Li 1991), and more generally any supervised dimension reduction technique in the field of manifold learning algorithms (see Section 4.4.1), such as supervised *kernel principal component analysis* (Barshan, Ghodsi, Azimifar and Zolghadri Jahromi 2011).

Although the above approaches concern a parameter space reduction only, they can be combined with a model order reduction technique, to deal with very complex numerical studies. For instance, the iterative procedure of correcting the projection operator on the reduced parameter space has been combined with goal-oriented model constrained optimization (Bui-Thanh, Willcox, Ghattas and van Bloemen Waanders 2007) to solve large-scale inverse problems (Himpe and Ohlberger 2015, Lieberman, Willcox and Ghattas 2010). Instead, within the AS context, various analyses have been conducted combining this technique with POD-Galerkin (Tezzele, Ballarin and Rozza 2018*a*), proper orthogonal decomposition with interpolation (Demo, Tezzele and Rozza 2019) (see Section 4.4.2) and dynamic mode decomposition (Tezzele *et al.* 2018*b*) (see Section 4.4.4).

4.4.2. POD with interpolation (PODI)

Proper orthogonal decomposition with interpolation (PODI) is a technique widely used in the ROM community, which builds data-driven ROMs by replacing the projection phase typical of the classical approach with an interpolation phase, used to approximate the solution manifold.

As described in Section 4.1, POD allows extraction of an optimal basis from a set of high-dimensional snapshots. We additionally assume that we do not know or do not want to employ the underlying mathematical model used to obtain the high-fidelity solutions. In fact the PODI methodology can be applied even if the numerical solver that computed the solutions is not available; this also allows its application to data coming from experimental measurements, such as *particle image velocimetry* (PIV) images (Semeraro *et al.* 2012). The non-intrusiveness of this approach resides exactly in the lack of modification to the simulation software, as often happens with data-driven methodologies. However, we recall that there are

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

also non-data-driven techniques that allow operation in a non-intrusive way (Zou, Conti, Díez and Auricchio 2018).

The basic steps of PODI are:

- collection of the solution database;
- extraction of a reduced basis via POD;
- use of a regression model to recover the reduced basis coefficients.

Assume that we have a training database $\mathcal{D} = \{(\eta_j, \boldsymbol{u}_h^{(j)})\}_{j=1}^{N_{\text{tr}}}$ consisting of $N_{\text{tr}} \leq N_s$ pairs of time-parameter values $\eta_j = (t_k, \mu_m)$, with $j = (m-1)(N_T+1) + (k+1)$, and degrees of freedom $\boldsymbol{u}_h^{(j)} \in \mathbb{R}^N$ of the associated full order solution. Typically this database coincides with that used to perform POD. Since we are interested in recovering the map from the time-parameter space to the reduced order coefficients, we need to project the high-fidelity solutions onto the reduced space $\mathbb{V}_r$. This can be done in terms of the coefficient vectors via

$$\boldsymbol{u}_r^{(j)} := \mathbf{U}^\top \boldsymbol{u}_h^{(j)}, \quad 1 \leq j \leq N_{\text{tr}},$$

where $\mathbf{U} \in \mathbb{R}^{N \times n}$ collects the left-singular vectors of the snapshot matrix. At this point we have a database $\mathcal{D}_r = \{(\eta_j, \boldsymbol{u}_r^{(j)})\}_{j=1}^{N_{\text{tr}}}$, which can be used to fit the regression model.

The final task will in fact consist in constructing a regression $\pi : \mathcal{T} \times \mathbb{P} \to \mathbb{R}^n$ which approximates the function $\mathscr{F} : (t; \mu) \in \mathcal{T} \times \mathbb{P} \mapsto \boldsymbol{u}_r \in \mathbb{R}^n$. Through the database $\mathcal{D}_r$ we have N_{tr} known points of $\mathscr{F}$, and any interpolation technique can be used to evaluate the reduced coefficients at other parameter values. The full order solution will then be approximated as

$$\boldsymbol{u}_h(t; \mu) = \mathbf{U}\pi(t; \mu) \quad \text{for all } t \in \mathcal{T},\ \mu \in \mathbb{P}. \tag{4.52}$$

Note that the reconstruction of the solution is conducted through the matrix–vector multiplication in equation (4.52) and consequently requires $O(Nn)$ operations. As can be seen from the above discussion, it is theoretically possible to treat time as an additional parameter. Therefore, in the following, the various regression methods will be presented to construct a map $\pi : \mathbb{P} \to \mathbb{R}^n$, thus assuming that the parameter also incorporates the variable t, *i.e.* $\mu = [t, \mu_1, \mu_2, \ldots, \mu_{P-1}]$. This is not the only possible approach, and, for this reason, in the next subsections we will first discuss different strategies to deal with the time dependence, and then introduce the main regression methods used in the ROM community, namely radial basis function (RBF) regression, Gaussian process regression (GPR) and artificial neural networks (ANNs).

Time-dependent problems

Regarding time dependence, we have seen that the above framework allows us to treat the time variable as an additional parameter, for which the model must learn the dynamics. As pointed out in Wang, Hesthaven and Ray (2019), the problem

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

with the above approach is that the regression task is more complex and the cost of POD increases considerably. For these reasons many works exploit a two-level POD (Chen *et al.* 2018, Xiao, Fang, Pain and Hu 2015), which we will describe using the approach presented by Audouze, De Vuyst and Nair (2009). We consider the following two discretizations for the time domain and the parameter space:

$$\mathcal{T}_h = \{t_1, t_2, \ldots, t_{N_T}\} \subset \mathcal{T}, \quad \mathbb{P}_h = \{\mu_1, \mu_2, \ldots, \mu_M\} \subset \mathbb{P}.$$

We then consider the set obtained via numerical resolution of the PDE under consideration for a given value of the parameter $\mu_m \in \mathbb{P}_h$:

$$\mathcal{M}_h^{N_T}(\mu_m) = \{u_h^k(\mu_m) \mid 1 \leq k \leq N_T\}.$$

We note that the above set simply represents the discrete trajectory of the dynamical system, by fixing the value of the parameter. The idea is to compress this trajectory by extracting spatial and temporal modes separately, with the aim of speeding up the offline computations. A set of K_m reduced spatial basis functions is obtained via POD of the following snapshot matrix:

$$\mathbf{S}_m^N = \left[\boldsymbol{u}_h^1(\mu_m) \mid \ldots \mid \boldsymbol{u}_h^{N_T}(\mu_m)\right] \in \mathbb{R}^{N \times N_T}.$$

Let $\mathbf{U}_m \in \mathbb{R}^{N \times K_m}$ denote the matrix whose columns are the resulting POD basis functions. After obtaining a set of bases for each value of the parameter $\mu_m \in \mathbb{P}_h$, we proceed to compute a global spatial basis $\{\xi_i\}_{i=1}^{N_{r,x}}$ via a high-level POD, which will be performed on the $\mathbf{U}$ matrix obtained by concatenating the matrices $\mathbf{U}_m$ from the above step, namely,

$$\mathbf{U} = [\mathbf{U}_1 \mid \ldots \mid \mathbf{U}_M] \in \mathbb{R}^{N \times K},$$

where $K = \sum_{m=1}^{M} K_m$ and K_m is the number of POD modes extracted at the mth iteration. Typically this procedure is conducted by setting a tolerance threshold on the energy retained by the spatial modes for each value of the parameter $\mu_m \in \mathbb{P}_h$. This causes the number of bases K_m extracted for each value of the parameter to be different.

To obtain temporal modes, the procedure is analogous: the snapshots are obtained by fixing the spatial degree of freedom, that is,

$$\mathbf{S}_m^{N_T} = [\boldsymbol{u}_{h,1}(\mu_m) \mid \ldots \mid \boldsymbol{u}_{h,N}(\mu_m)] \in \mathbb{R}^{N_T \times N},$$

where $\boldsymbol{u}_{h,j}$ denotes the N_T-dimensional vector that collects all time instances of the jth degree of freedom of $\boldsymbol{u}_h(\mu_m)$. Reproducing the two-level POD for time mode extraction results in a temporal reduced basis $\{\varphi_j\}_{j=1}^{N_{r,t}}$. The expansion expressed by (4.6) is replaced by

$$u_r(\mu) = \sum_{i=1}^{N_{r,x}} \sum_{j=1}^{N_{r,t}} \alpha_{(i,j)}(\mu)\, \varphi_j\, \xi_i.$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Therefore, within the data-driven context, the approximation will involve the functions $\alpha_{(i,j)}(\mu)\colon \mathbb{P} \to \mathbb{R}$.

Another effective approach to addressing the problem involves using tensor decomposition for the reduced coefficients (Guo and Hesthaven 2019). This technique consists in organizing the reduced coefficients into matrices that exploit the time-parameter grid structure $\mathcal{T}_h \times \mathbb{P}_h$. We decompose the matrix

$$[\mathbf{P}_k]_{(i,j)} = \boldsymbol{u}_{r,k}(t_i, \mu_j), \quad 1 \le i \le N_T,\ 1 \le j \le M,$$

by relying again on POD for the extraction of time modes $\{\boldsymbol{\psi}_\ell^{(k)}\}_{\ell=1}^n$ and parameter modes $\{\boldsymbol{\phi}_\ell^{(k)}\}_{\ell=1}^n$, namely,

$$\mathbf{P}_k \approx \widetilde{\mathbf{P}}_k = \sum_{\ell=1}^{n} \lambda_\ell^{(k)} \boldsymbol{\psi}_\ell^{(k)} (\boldsymbol{\phi}_\ell^{(k)})^\top,$$

where $\lambda_\ell^{(k)}$ is the ℓth singular value. At this point the elements of the former discrete modes are used as a dataset to approximate the functions

$$\begin{aligned} t &\mapsto \widehat{\psi}_\ell^{(k)}(t) \quad \text{such that} \quad \widehat{\psi}_\ell^{(k)}(t_i) \approx (\boldsymbol{\psi}_\ell^{(k)})_i, \\ \mu &\mapsto \widehat{\phi}_\ell^{(k)}(\mu) \quad \text{such that} \quad \widehat{\phi}_\ell^{(k)}(\mu_j) \approx (\boldsymbol{\phi}_\ell^{(k)})_j. \end{aligned}$$

A continuous regression function $\widehat{\boldsymbol{u}}_{r,k}$ for the kth projection coefficient with respect to time-parameter values can therefore be recovered as

$$\boldsymbol{u}_{r,k}(t;\mu) \approx \widehat{\boldsymbol{u}}_{r,k}(t;\mu) = \sum_{\ell=1}^{n} \lambda_\ell^{(k)} \widehat{\psi}_\ell^{(k)}(t)\, \widehat{\phi}_\ell^{(k)}(\mu) \quad \text{for all } (t;\mu) \in \mathcal{T} \times \mathbb{P}.$$

A final approach can be implemented when we are not interested in treating time as a continuous variable (Berzins, Helmig, Key and Elgeti 2020). In such a case we can construct a snapshot matrix whose columns are represented by *complete* trajectories of the solution, *i.e.* $\mathbf{S} \in \mathbb{R}^{N N_T \times M}$. The basis extracted from POD in this case is spatio-temporal and thus regression should be performed only to recover the dynamics with respect to the parameter.

Radial basis function interpolation

Radial basis function (RBF) interpolation (Lazzaro and Montefusco 2002) is one of the primary tools used for multivariate scattered function interpolation. Its coupling with proper orthogonal decomposition (POD-RBF) has found wide acceptance in the ROM community due to its remarkable flexibility. In particular, it was applied by Audouze *et al.* (2009) to solve the parametrized Burgers' equation and a parametrized convection–reaction–diffusion problem. In the CFD context it was effective for solving the Navier–Stokes equations in both two (Xiao *et al.* 2015) and three dimensions (Walton, Hassan and Morgan 2013). The various investigations have shown that this method can effectively capture the nonlinearities of PDEs and can provide a spectral convergence rate (Buhmann and Dyn 1993).

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Table 4.1. Some examples of radial basis functions in common use. The parameter σ typically scales the width of the function.

Radial basis function	Expression
Gaussian	$\psi(\mu) = \mathrm{e}^{-(\|\mu-\mu_c\|/\sigma)^2}$
Linear spline	$\psi(\mu) = \|\mu - \mu_c\|$
Multiquadric	$\psi(\mu) = \sqrt{\|\mu - \mu_c\|^2 + \sigma^2}$
Inverse multiquadric	$\psi(\mu) = 1/\sqrt{\|\mu - \mu_c\|^2 + \sigma^2}$
Cubic spline	$\psi(\mu) = \|\mu - \mu_c\|^3$

A generic radial function $\psi : \mathbb{R}^P \to \mathbb{R}$ is a multivariate function, which however depends only on an appropriate norm $\|\cdot\|$ of its argument: $\psi = \psi(\|\mu\|)$. A typical example consist in taking the Euclidean distance of the argument from a point, with respect to which the function will be radially symmetric $\psi = \psi(\|\mu - \mu_c\|)$. Some popular choices concerning radial basis functions are given in Table 4.1.

Within the regression task, multiple radial functions are used, namely,

$$\psi_m = \psi(\|\mu - \mu_m\|), \quad 1 \le m \le M,$$

which means that we are taking M different basis function by shifting the centre within the discretized parameter space $\mathbb{P}_h$.

The idea is to approximate component-wise the vector of reduced coefficients using the former set of functions as a basis, that is,

$$\boldsymbol{u}_{r,k}(\mu) \approx \widetilde{\pi}_k(\mu; \boldsymbol{w}_k) = \sum_{m=1}^{M} (\boldsymbol{w}_k)_m \psi(\|\mu - \mu_m\|). \tag{4.53}$$

The coefficients of the expansion $\boldsymbol{w}_k \in \mathbb{R}^M$ can be derived by requiring exact interpolation of the training data:

$$\widetilde{\pi}_k(\mu_m; \boldsymbol{w}_k) = \boldsymbol{u}_{r,k}(\mu_m), \quad 1 \le m \le M.$$

Substituting the above M conditions in equation (4.53), we arrive at the following linear system:

$$\mathbf{A}\boldsymbol{w}_k = \boldsymbol{u}_{r,k}, \quad \text{where} \quad \begin{cases} \mathbf{A}_{(i,j)} = \psi(\|\mu_i - \mu_j\|), & 1 \le i, j \le M, \\ (\boldsymbol{u}_{r,k})_j = \boldsymbol{u}_{r,k}(\mu_j), & 1 \le j \le M. \end{cases} \tag{4.54}$$

Note that we must solve a total of n systems (one for each component of the reduced vector $\boldsymbol{u}_r$) but that, as pointed out in system (4.54), the matrix $\mathbf{A} \in \mathbb{R}^{M \times M}$ (and its subsequent decomposition) is the same for all the systems under consideration. This makes the resolution computationally efficient. All the weights can then be

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

collected in a single matrix given by

$$\mathbf{W} = [\boldsymbol{w}_1 \mid \ldots \mid \boldsymbol{w}_n] \in \mathbb{R}^{M \times n}.$$

The map for the reduced coefficients will then be given by

$$\pi(\mu) = \mathbf{W}^T \boldsymbol{\Psi}(\mu), \quad \text{with } (\boldsymbol{\Psi}(\mu))_m = \psi(|\|\mu - \mu_m\|), \ 1 \leq m \leq M.$$

Gaussian process regression

Regression with a Gaussian process (GP) (Rasmussen and Williams 2005) is a supervised Bayesian approach to regression. This technique is closely related to Bayesian linear regression; indeed, in the case of using a kernel, these two approaches coincide. Moreover, it has been shown that GPs are also equivalent to NNs, with a single hidden layer of infinite width (Neal 1996). In the context of the reduced basis methods, this approach was first tested to evaluate particular outputs of interest of the dynamical system (Nguyen and Peraire 2016). Guo and Hesthaven (2018) proposed recovering the solution field of a nonlinear structural analysis problem. Later this technique was also applied to time-dependent problems (Guo and Hesthaven 2019) and industrial applications, such as shape optimization for naval engineering (Ortali, Demo and Rozza 2022).

This approach is applied initially to a literature case, the simulation of the Stokes problem, and in the following to a real-world industrial problem, within a shape optimization pipeline for a naval engineering problem.

The main idea of GP regression lies in learning a distribution over functions. This provides a significant advantage over classical Bayesian linear regression, which requires specifying the set of basis functions in advance and restricts the approximation to reside in the space generated by the bases. In contrast, GPs do not assume a parametric form for the approximate function, but learn a distribution over what is the output of the unknown underlying function for any point in its domain. By doing so, we obtain a flexible model that allows the capture of any function that interpolates the available dataset.

According to this perspective, let us consider a fixed point $\mu^* \in \mathbb{P}$. We are trying to approximate the kth component $\boldsymbol{u}_{r,k}(\mu)$ of the vector $\boldsymbol{u}_r(\mu)$, so in the continuation we will denote such scalar regression by $f : \mathbb{P} \to \mathbb{R}$. The evaluation of the regression function must then be considered as a Gaussian distribution, *i.e.* $f(\mu^*) \sim \mathcal{N}(m(\mu^*), \sigma^2)$. We now consider a finite set of inputs that we collect into the vector $\boldsymbol{x} = [\mu_1, \mu_2, \ldots, \mu_M]$. The values of the output then result in N different Gaussian distributions. However, since all the outputs in question are to be considered the values of the function when evaluated at different points, then these distributions must necessarily be correlated to each other. For example, if we start with a value in parameter space for which the function assumes a certain value with relative uncertainty, we expect that a small perturbation of the input will correspond to a limited change in both the output and the uncertainty value because of correlation. The former requirement means that the input $\boldsymbol{x}$ is associated to a

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

multivariate distribution for the output $\boldsymbol{y} = [f(\mu_1), f(\mu_2), \ldots, f(\mu_M)]$, that is,

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}, \mathbf{K}), \quad \text{where} \quad \begin{cases} (\boldsymbol{m})_i = \mathbb{E}[y_i], \\ \mathbf{K}_{(i,j)} = \mathbb{E}[(y_i - m_i)(y_j - m_j)], \end{cases} \tag{4.55}$$

where $\boldsymbol{m}$ is the mean vector and $\mathbf{K}$ is the covariance matrix. The above considerations suggest that the regression function must be an infinite (continuous) multivariate Gaussian distribution, which is what we refer to as a *Gaussian process*:

$$f(\mu) \sim \mathcal{GP}(m(\mu), \kappa(\mu, \mu')) \quad \text{for all } \mu, \mu' \in \mathbb{P}. \tag{4.56}$$

More formally, a GP is a stochastic process $f(\mu)$ such that any finite number of random variables taken from the collection that forms the random process itself has a joint Gaussian probability distribution. Equation (4.56) represents a distribution *over functions*. In fact, as for a regular Gaussian distribution, we specify a mean and a covariance. However, since the dimensions are infinite, the vector $\boldsymbol{m}$ is replaced by the mean function $m(\cdot)$, and the matrix $\mathbf{K}$ is replaced by the two-dimensional kernel covariance function $\kappa(\cdot, \cdot)$, such that

$$\begin{aligned} m(\mu) &= \mathbb{E}[f(\mu)], \\ k(\mu, \mu') &= \mathbb{E}[(f(\mu) - m(\mu))(f(\mu') - m(\mu')]. \end{aligned}$$

In particular, the quantities that appear in equation (4.55) turn out to be particular realizations of a GP, over a finite subset of the input. That is, if we consider a vector $\boldsymbol{x}_{\text{tr}} = [\mu_1, \ldots, \mu_M]$ that contains all elements of the training set $\mathbb{P}_h$, we will obtain

$$\begin{aligned} \boldsymbol{m} &= m(\boldsymbol{x}_{\text{tr}}) \coloneqq [m(\mu_i)]_{1 \le i \le M} \in \mathbb{R}^M, \\ \mathbf{K} &= k(\boldsymbol{x}_{\text{tr}}, \boldsymbol{x}_{\text{tr}}) \coloneqq [k(\mu_i, \mu_j)]_{1 \le i,j \le N} \in \mathbb{R}^{M \times M}. \end{aligned}$$

As for the regression task, it begins by choosing prior distributions for $m(\cdot)$ and $\kappa(\cdot, \cdot)$ that must be representative of our assumptions about the profile of these functions before looking at the observed values. A widely used assumption for $m(\cdot)$ is that of zero mean, *i.e.* $m(\mu) = 0$. This assumption is particularly suitable in the case in which the set $\mathbb{P}_h$ is processed by subtracting the mean. Regarding the covariance function, different choices greatly influence the regression profile. For example, a common choice is the *squared exponential*,

$$\kappa(\mu, \mu') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2} \|\mu - \mu'\|_{\mathbb{R}^P}^2 \right),$$

where σ_f is the standard deviation parameter that scales the uncertainty value outside the training set, and ℓ is the correlation length scale. If we have more information about the correlation between pairs of parameter components, we can use the anisotropic *squared exponential kernel*:

$$\kappa(\mu, \mu') = \sigma_f^2 \exp\left(-\sum_{i=1}^{P} \frac{(\mu_i - \mu_i')^2}{2\ell^2} \right).$$

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

The *posterior* distribution is obtained by conditioning the *prior* based on the observed data, namely,

$$\mathrm{P}(f(\cdot) \mid \boldsymbol{x}_{\mathrm{tr}}, \boldsymbol{y}_{\mathrm{tr}}) = \mathcal{GP}(\bar{f}(\cdot), k'(\cdot,\cdot)),$$

where the revised mean $\bar{f}(\cdot)$ and covariance $\kappa'(\cdot,\cdot)$ are given by the following expressions:

$$\begin{cases} \bar{f}(\cdot) = k(\cdot, \boldsymbol{x}_{\mathrm{tr}})\mathbf{K}^{-1}\boldsymbol{y}_{\mathrm{tr}}, \\ k'(\cdot,\cdot) = k(\cdot,\cdot) - k(\cdot, \boldsymbol{x}_{\mathrm{tr}})\mathbf{K}^{-1}k(\boldsymbol{x}_{\mathrm{tr}}, \cdot). \end{cases} \tag{4.57}$$

We want to use the former expressions for prediction purposes. In particular, we imagine that we have a set of new instances of the parameters that we collect in the vector $\boldsymbol{x}_{\mathrm{pr}}$. We want to compute the expected value of the output $\boldsymbol{y}_{\mathrm{pr}}$, namely $m(\boldsymbol{x}_{\mathrm{pr}})$. By using the expressions in equation (4.57), together with the properties of GPs, we obtain

$$\boldsymbol{y}_{\mathrm{pr}} \mid \boldsymbol{x}_{\mathrm{pr}}, \boldsymbol{x}_{\mathrm{tr}}, \boldsymbol{y}_{\mathrm{tr}} \sim \mathcal{N}(\mathbf{m}_{\mathrm{pr}}, \mathbf{K}'_{\mathrm{pr}}),$$

with mean vector and covariance matrix given by

$$\begin{aligned} \mathbf{m}_{\mathrm{pr}} &= \overline{f}(x_{\mathrm{pr}}) = k(\boldsymbol{x}_{\mathrm{pr}}, \boldsymbol{x}_{\mathrm{tr}})\mathbf{K}^{-1}\boldsymbol{y}_{\mathrm{tr}}, \\ \mathbf{K}'_{\mathrm{pr}} &= k(\boldsymbol{x}_{\mathrm{pr}}, \boldsymbol{x}_{\mathrm{pr}}) - k(\boldsymbol{x}_{\mathrm{pr}}, \boldsymbol{x}_{\mathrm{tr}})\mathbf{K}^{-1}k(\boldsymbol{x}_{\mathrm{tr}}, \boldsymbol{x}_{\mathrm{pr}}). \end{aligned}$$

The values of the hyper-parameters that appear in the kernel function have an important effect on the predictive performances of the model. For instance, in the case of the squared exponential we need to find a suitable value for $\theta = [\sigma_f, \ell]$. This is achieved by maximizing the log of the conditional density of $\boldsymbol{y}_{\mathrm{tr}}$ given $\boldsymbol{x}_{\mathrm{tr}}$:

$$\begin{aligned} \theta &= \arg\max_{\widehat{\theta}} \ \log p(\boldsymbol{y}_{\mathrm{tr}} \mid \boldsymbol{x}_{\mathrm{tr}}, \widetilde{\theta}) \\ &= \arg\max_{\widehat{\theta}} \left\{ -\frac{1}{2}\boldsymbol{y}_{\mathrm{tr}}^{\top}\mathbf{K}^{-1}(\widehat{\theta})\boldsymbol{y}_{\mathrm{tr}}^{s} - \frac{1}{2}\log|\mathbf{K}^{-1}(\widehat{\theta})| - \frac{M}{2}\log(2\pi) \right\}. \end{aligned}$$

As mentioned in Section 4.4.2, the GP regression can be integrated with proper orthogonal decomposition, using the PODI framework. The regression map $\pi_{GP} : \mathbb{P} \to \mathbb{R}^n$ is constructed component-wise by using the former procedure. In particular, the kth entry of this map will be obtained by considering $\boldsymbol{x}_{\mathrm{tr}} = [\mu]_{\mu\in\mathbb{P}_h}$ and $\boldsymbol{y}_{\mathrm{tr}} = [\boldsymbol{u}_{r,k}(\mu)]_{\mu\in\mathbb{P}_h}$ as training data. The result will consist of n independent Gaussian distributions, whose mean functions can be used to predict the reduced coefficients at new parameter points.

4.4.3. Artificial neural networks

In the field of machine learning, an artificial neural network (ANN) is a computational model composed of artificial 'neurons', loosely inspired by the simplification of a biological neural network. Like any other machine learning algorithm, this technique involves a learning process, which improves the performance of the

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

model by using available data. In particular, these models rely on learning not only the map from input to output but also a representation of the input, which can better justify the observed data. In the following, we will describe a particular ANN architecture, known as *multilayer perceptrons* (MLPs) (Goodfellow *et al.* 2016) or a *feedforward neural network*. This type of network constructs an approximation of a function by composing multiple simple functions. The enormous diffusion of this type of model has also reached the ROM community, and in fact Hesthaven and Ubbiali (2018) have proposed the so-called POD-NN method, which consists in using a neural network in the regression phase of PODI. This method has since been used to conduct investigations in a variety of physical contexts, including instability in the convection-dominated regime (Gao *et al.* 2021), aerostructural design optimization (Park *et al.* 2013), combustion (Wang *et al.* 2019), bifurcations in CFD (Pichi, Ballarin, Rozza and Hesthaven 2021), complex flow problems (Berzins *et al.* 2020) and turbulence (Zancanaro *et al.* 2021). The potential of this type of architecture can be understood via the universal approximation theorem (Cybenko 1989), which states that for any continuous function f, on a compact set $K \subset \mathbb{R}^n$, there exists a standard MLP with a single hidden layer that uniformly approximates f in K with arbitrary tolerance $\epsilon > 0$. Moreover, in the case of two hidden layers, this approximation property is extended to *any* function. As appealing as the above result is, we recall that it gives no indication of how to achieve the approximation, and whether a single (or double) layer is the most efficient choice.

The operational unit of the network is represented by the so-called *perceptron* (Rosenblatt 1958), which receives as input a vector whose components are the activation states of the sending neurons, *i.e.* $\boldsymbol{y}_s = [y_{s_1}, \dots, y_{s_m}]^\top \in \mathbb{R}^m$. This input is converted into the scalar c_j by means of a weighted sum,

$$c_j(\boldsymbol{y}_s, \boldsymbol{w}_{s,j}) = \boldsymbol{w}_{s,j}^\top \boldsymbol{y}_s,$$

which is afterwards transformed into the *excitation state* of the neuron through a scalar activation function:

$$y_j = f_{\text{act}}(c_j; b_{0,j}) = f_{\text{act}}(\boldsymbol{w}_{s,j}^\top \boldsymbol{y}_j + b_{0,j}).$$

A typical example of activation function is the *hyperbolic tangent*:

$$f_{\text{act}}(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}}.$$

The idea for MLPs is to connect various perceptrons together, by organizing them into consecutive layers. In particular, the first layer represents the input of the network and thus $\boldsymbol{y}_0 = \mu$, while the last layer $\boldsymbol{y}_N \in \mathbb{R}^P$ is the output. In the interlude there are $N_\ell - 1$ *hidden* layers $\{\boldsymbol{y}_i\}_{1 \le i \le N_\ell - 1}$, composed of perceptrons that work according to the model just described. An example of an MLP with a single hidden layer is proposed in Figure 4.2. The model is also called *feedfoward* because the information flow unidirectionally from the input to the output, *i.e.* $\boldsymbol{y}_0 \to \boldsymbol{y}_1 \to \dots \to \boldsymbol{y}_{N_\ell}$.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

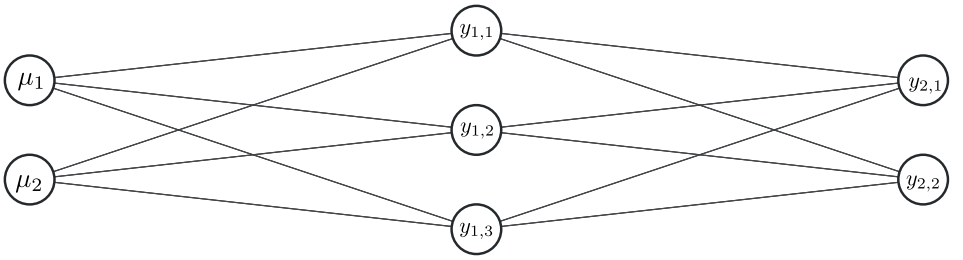

Figure 4.2. An example of multilayer perceptron with a single hidden layer: $\boldsymbol{y}_0 = \mu \in \mathbb{R}^2$ and $\boldsymbol{y}_{\text{out}} = \boldsymbol{y}_2 \in \mathbb{R}^2$.

In practice, each layer of the network is obtained by first applying a linear transformation to the vector of the previous layer, and then applying a nonlinear transformation represented by the activation function. We can then model the overall behaviour of the network as follows:

$$\pi_{\text{NN}}(\mu;\Theta)\begin{cases}\boldsymbol{y}_0 = \mu,\\ \boldsymbol{y}_j = \phi_j(\mathbf{W}_j\boldsymbol{y}_{j-1} + \boldsymbol{b}_j) \quad \text{for } j = 1,\dots,N_\ell,\\ \boldsymbol{y}_{\text{out}} = \boldsymbol{y}_{N_\ell},\end{cases}$$

where $\mathbf{W}_j \in \mathbb{R}^{N_j \times N_{j-1}}$ and $b_j \in \mathbb{R}^{N_j}$ are the weights and the biases for the jth hidden layer, and N_j is the number of nodes for the same layer. In particular, if we introduce the intermediate transformation maps $\pi_j(\boldsymbol{y}_{j-1}) = \phi_j(\mathbf{W}_j\boldsymbol{y}_{j-1} + \boldsymbol{b}_j)$, we can compactly rewrite

$$\pi_{\text{NN}} : (\mu;\Theta) \mapsto \pi_N(\mu;\mathbf{W}_N;\boldsymbol{b}_N) \circ \dots \circ \pi_1(\mu;\mathbf{W}_1;\boldsymbol{b}_1), \tag{4.58}$$

where we have collected all the network parameters (weights and biases) within Θ. Equation (4.58) corresponds to the intuition related to the construction of a regression map via the composition of simpler functions.

So far we have only seen how to use the neural network for predictions. However, these predictions will only be accurate if the value of Θ is correctly prescribed. For this reason, a supervised learning paradigm is used, that is, given a set of N_s input–output pairs $\{\mu_i, \boldsymbol{y}^i\}_{i=1}^{N_s}$, find the optimal parameter vector Θ by optimizing a loss (or cost) function on the training data:

$$\Theta = \arg\min_{\widehat{\Theta}} \mathcal{J}(\widehat{\Theta}) = \arg\min_{\widehat{\Theta}} \frac{1}{N_s}\sum_{i=1}^{N_s} \mathcal{L}(\boldsymbol{y}^i, \boldsymbol{y}^i_{N_\ell}; \widehat{\Theta}), \tag{4.59}$$

where $\mathcal{J}$ is the loss function and $\mathcal{L}$ the per-sample loss function. A common choice is given by measuring the per-sample error using the squared Euclidean distance, to which corresponds the cost function known as *cumulative mean square error*

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

(MSE):

$$\mathcal{J}(\Theta) = \frac{1}{N_s} \sum_{i=1}^{N_s} \|\boldsymbol{y}^i - \boldsymbol{y}^i_{N_\ell}\|^2_{\mathbb{R}^{N_\ell}}.$$

A factor that has promoted the recent huge success of NNs is the possibility of solving problem (4.59) efficiently. In particular, the most popular method is *stochastic gradient descent* (SGD), whose efficient implementation requires us to use *back-propagation* (Rumelhart, Hinton and Williams 1986) to compute the partial derivatives of the loss function with respect to the parameters. In particular, the gradient of $\mathcal{J}$ is approximated using a minibatch of m samples belonging to the training set:

$$\nabla_\Theta \mathcal{J}(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\Theta \mathcal{L}(\boldsymbol{y}^i, \boldsymbol{y}^i_N; \Theta).$$

For an extended description of the alternative techniques in use for solving problem (4.59), we refer to Goodfellow *et al.* (2016).

4.4.4. Dynamic mode decomposition

Dynamic mode decomposition (DMD) is a recent data-driven technique used to obtain linear reduced order models for high-dimensional complex systems. This method extracts from the given dataset a few spatio-temporal coherent structures, which dominate the dynamics of the measured data. First introduced by Schmid (2010) in the computational fluid dynamics field, it was later applied successfully to a diverse range of problems (Kutz *et al.* 2016*a*). Its wide range of applications depends on its purely data-driven nature, which makes it useful whenever there is a lack of knowledge of the underlying equation of motion.

The DMD standard procedure (Schmid 2010) starts by considering a dynamic process encoded by a generic ordinary differential equation:

$$\frac{\mathrm{d}\boldsymbol{u}_h(t)}{\mathrm{d}t} = f(\boldsymbol{u}_h(t), t) \quad \text{for } t \in \mathcal{T}, \tag{4.60}$$

with $\boldsymbol{u}_h \in \mathbb{R}^N$, $N \gg 1$. Typically the function f is nonlinear with respect to the solution $\boldsymbol{u}_h$, which makes the resolution non-trivial. However, in the DMD setting we assume that the function f is unknown, which happens for example when we do not have governing first principles for the problem at hand. However, assume that we can measure the state of the system, *i.e.* the solution $\boldsymbol{u}_h$ at different times. Even when it is not clear if we have measurements of $\boldsymbol{u}_h$, we can consider taking as their proxy certain outputs of the dynamical system, *i.e.* $\boldsymbol{y} = g(\boldsymbol{u}_h)$. The main idea of DMD is to approximate equation (4.60) by means of a linear dynamical system, that is,

$$\frac{\mathrm{d}\boldsymbol{u}_h(t)}{\mathrm{d}t} = \mathbf{A}\boldsymbol{u}_h(t) \quad \text{for } t \in \mathcal{T}, \tag{4.61}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$. Clearly the linear system in equation (4.61) becomes easy to

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

treat since there is a closed form for its solutions:

$$\boldsymbol{u}_h(t) = \mathrm{e}^{\mathbf{A}t}\boldsymbol{u}_h(t_0),$$

where $\mathrm{e}^{\mathbf{A}t} \in \mathbb{R}^{N\times N}$ is the exponential of the matrix $\mathbf{A}$. Once the matrix $\mathbf{A}$ is obtained, we proceed to compute the set $\{(\lambda_j, \boldsymbol{\Phi}_j)\}_{1\le j\le N}$ of its eigenvalues and eigenvectors. In particular, the eigenvectors $\boldsymbol{\Phi}_j$ are called DMD modes, while the eigenvalues λ_j provide information regarding the frequency ω_j associated to the corresponding mode, via

$$\omega_j = \frac{\mathrm{Im}(\log \lambda_j)}{2\pi\Delta t}.$$

We must remember that the function f in equation (4.60) is unknown and consequently we cannot solve the problem by a simple linearization of the operator. For the moment we assume that we have access to the complete state vector and that we have some measurements $\boldsymbol{u}_h^k = \boldsymbol{u}_h(t_k)$. These snapshots are used to create the following snapshot matrices:

$$\mathbf{S} = \begin{bmatrix} | & | & & | \\ \boldsymbol{u}_h^1 & \boldsymbol{u}_h^2 & \dots & \boldsymbol{u}_h^{m-1} \\ | & | & & | \end{bmatrix}, \quad \mathbf{S}' = \begin{bmatrix} | & | & & | \\ \boldsymbol{u}_h^2 & \boldsymbol{u}_h^3 & \dots & \boldsymbol{u}_h^{m} \\ | & | & & | \end{bmatrix}.$$

The idea is that the matrix $\mathbf{A}$ will perform a linear transformation of the current snapshot database in the shifted one, that is,

$$\mathbf{S}' = \mathbf{A}\mathbf{S}. \tag{4.62}$$

One can therefore proceed by multiplying equation (4.62) by the pseudo-inverse of $\mathbf{S}$, in order to obtain

$$\mathbf{A} = \mathbf{S}'\mathbf{S}^{\dagger}. \tag{4.63}$$

The former procedure is called *exact DMD* and involves the calculation of the pseudo-inverse of the snapshot matrix, which we recall is a least-squares regression algorithm. An important consideration is that the matrix $\mathbf{A}$ is the one that minimizes the error $\|\mathbf{S}' - \mathbf{A}\mathbf{S}\|_F$ in the Frobenius matrix norm. This means that the matrix $\mathbf{A}$ minimizes $\|\boldsymbol{u}_h^{k+1} - \mathbf{A}\boldsymbol{u}_h^k\|$ across all snapshots of the solution for $1 \le k \le m-1$.

However, this way of proceeding ignores the considerations already made in Section 4.4.1, namely that many of the systems that are measured show low-rank structures. For this reason, the DMD algorithm involves studying the dynamics of the system over the subspace provided by POD. This is done by first substituting the r-rank POD approximation $\mathbf{S} \approx \mathbf{U}\Sigma\mathbf{V}^*$ in equation (4.63):

$$\mathbf{A} = \mathbf{S}'\mathbf{V}\Sigma^{-1}\mathbf{U}^*.$$

We then project $\mathbf{A}$ onto the dominant single vectors, in order to obtain the operator

$$\widetilde{\mathbf{A}} = \mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{U}^*\mathbf{S}'\mathbf{V}\Sigma^{-1},$$

which provides information about the time evolution of the POD modes. One

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

can prove that the eigenvalues of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ coincide, whereas the eigenvectors are related by

$$\mathbf{\Phi} = \mathbf{X}'\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{W},$$

where $\mathbf{\Phi}$ (respectively $\mathbf{W}$) collects the eigenvectors of $\mathbf{A}$ (respectively $\widetilde{\mathbf{A}}$) column-wise.

Finally, we can compute future state predictions as

$$\mathbf{u}_h(t) \approx \sum_{j=1}^{r} \mathbf{\Phi}_j \lambda_j^{t/\Delta t} b_j = \mathbf{\Phi}\mathbf{\Lambda}^{t/\Delta t}\boldsymbol{b}. \tag{4.64}$$

The amplitude vector $\boldsymbol{b}$ can be computed by considering (4.64) at initial time $t = 0$, which leads to

$$\boldsymbol{b} = \mathbf{\Phi}^{\dagger}\boldsymbol{u}_h^0. \tag{4.65}$$

In particular, (4.65) can be used by substituting another snapshot $\boldsymbol{u}_h^k$ for $\boldsymbol{u}_h^0$ to improve the approximation accuracy in the neighbourhood of that specific time instant.

Following the remarkable application success of DMD, several variants have been proposed that account for the limitations of this method, such as randomized DMD (Bistrian and Navon 2016), optimized DMD (Chen, Tu and Rowley 2012) and multiresolution DMD (Kutz, Fu and Brunton 2016*b*). For a broader discussion of extensions and applications we refer the interested reader to Kutz *et al.* (2016*a*).

5. Concluding remarks and outlook

The development of reduced basis methods to recover accurate and efficient surrogate models for solutions to parametrized solutions has resulted in the very successful applications of such methods to a variety of increasingly complex applications, including Maxwell equations, acoustic and elastic problems, heat-conduction, and the Navier–Stokes equations of incompressible fluid flows. We have discussed some of these applications in this article, and Hesthaven *et al.* (2015) and Quarteroni *et al.* (2016) have provided many further examples. With a primary focus on stationary problems, these developments have resulted in powerful methods that are mathematically rigorous and are often complemented by an error theory to certify the accuracy of the reduced models. This latter element has been achieved for most linear problems and even for certain types of nonlinear problems with quadratic nonlinearities.

However, in contrast to most past work we focus on the special case of time-dependent problems. This focus adds additional challenges to the development of reduced basis methods – challenges that originate in the stability of the reduced models and in questions of efficiency of the reduced models, related both to reducibility of the parametrized problem and to the efficient evaluation of nonlinear terms.

https://doi.org/10.1017/S0962492922000058 Published online by Cambridge University Press

Recent advances in structure-preserving reduced order models have demonstrated that by insisting on preserving invariants, inherited from the continuous model, stability of the reduced models can be preserved, and we have discussed many of these developments in detail. Such techniques also typically result in methods of increased accuracy. While the extension of such methods to complex applications in fluid or plasma dynamics is non-trivial, the general approach has successfully demonstrated a path forward.

The use of data-driven techniques has likewise shown itself to be a key tool in the development of reduced basis methods that are both stable and, most importantly, highly efficient, even for general nonlinear problems. This is achieved by introducing non-intrusive reduced order models in which a data-driven map is learned as the map between parameter space and coefficients of the reduced basis to reconstruct the solutions. In other words, the complex nonlinear relationship between parameters and coefficients is embedded into the data-driven map, *e.g.* a neural network. This approach has the additional advantage that only access to solutions is needed to build reduced order models.

While these are important steps towards the goal of making reduced basis methods an accurate and efficient tool to build surrogate model for general problems, possibly of industrial complexity, a number of key questions remain open to fully achieve this goal.

As discussed earlier, the reducibility of problem is highly problem-dependent, as expressed in the Kolmogorov n-width discussed in Section 3.1. A particular implication of this is that a nonlinear basis is required to ensure a compact basic for transport-dominated problems to reach the promise of an substantial acceleration of the surrogate. This challenge was discussed in Section 4.3 by the explicit update of the basis and in Section 4.4 with the generation of a nonlinear basis via a data-driven approach. However, much work is need to achieve this in a robust and general fashion and enable substantial acceleration of the reduced models for transport-dominated problems.

An attractive feature of the original development of reduced basis methods for linear parametrized PDEs is the ability to develop a rigorous error theory to qualify the output of the reduced basis method; that is, not only is the output a function of interest but we are also interested in explicit estimates of the errors associated with this output as a function of the size of the reduced basis (Hesthaven *et al.* 2015, Quarteroni *et al.* 2016). This last step implies that one can have full confidence in the output of the reduced model and choose the size of the reduced basis based solely on accuracy requirements.

For time-dependent problems or general nonlinear problems, the analysis leading to such estimates often either results in overly conservative error estimates or is simply not possible with the tools available. There is some limited work in that direction (Nguyen, Rozza and Patera 2009, Yano, Patera and Urban 2014), but there are substantial open questions that need to be understood in order to make this a practical tool.

This quest for a rigorous understanding of the accuracy of reduced basis methods is even more challenging when turning towards data-driven methods, in particular for problems based on the use of neural networks. Some initial results have been obtained by Kutyniok, Petersen, Raslan and Schneider (2022), but a more comprehensive understanding of the errors associated with such an approach remains elusive.

The majority of problems for which reduced basis methods have been developed so far are physics problems for which a global truth solver is available. However, for many complex problems, *e.g.* in an industrial setting, a global solver is not available. Instead the system is understood via a collection of local models, each of which can be a surrogate model of some kind. An open question then arises as to whether the availability of a collection of reduced models, coupled in a predefined way, is sufficient to characterize the behaviour of the global system. There are some early results in this direction for the harmonic Maxwell's equations in Ganesh, Hesthaven and Stamm (2012) and recently some also for time-dependent problems (Carlberg, Guzzetti, Khalil and Sargsyan 2019, Discacciati and Hesthaven 2021). Nonetheless, how to achieve this for more general time-dependent multi-physics coupled systems remains open – in particular the question of how to guarantee that global emergent behaviour, arising as a consequence of the model coupling, can be established. The ability to rapidly sample the reduced models and parametrize couplings suggests a path forward through iterations, but substantial advances are needed to establish such an approach to modelling large-scale coupled systems of industrial complexity.

Acknowledgements

GR would like to thank Mr Moaad Khamlich (SISSA) and Dr Federico Pichi (EPFL) for their valuable insights and support. We acknowledge the support provided by the European Research Council Executive Agency by the Consolidator Grant project AROMA-CFD 'Advanced Reduced Order Methods with Applications in Computational Fluid Dynamics' – GA 681447, H2020-ERC CoG 2015 AROMA-CFD, PI G. Rozza, INdAM-GNCS 2019-2020 projects, MUR PRIN 2019 project NA-FROM-PDEs. The work was partially supported by AFOSR under grant FA9550-17-1-9241.

References

S. E. Ahmed, S. Pawar, O. San, A. Rasheed, T. Iliescu and B. R. Noack (2021), On closures for reduced order models: A spectrum of first-principle to machine-learned avenues, *Phys. Fluids* **33**, 091301.

B. O. Almroth, P. Stern and F. A. Brogan (1978), Automatic choice of global shape functions in structural analysis, *AIAA J.* **16**, 525–528.

D. Amsallem and B. Haasdonk (2016), PEBL-ROM: Projection-error based local reduced-order models, *Adv. Model. Simul. Engrg Sci.* **3**, doi:10.1186/s40323-016-0059-7.

D. Amsallem, M. Zahr and C. Farhat (2012), Nonlinear model order reduction based on local reduced-order bases, *Internat. J. Numer. Methods. Engrg* **92**, 891–916.
D. N. Arnold, R. S. Falk and R. Winther (2006), Finite element exterior calculus, homological techniques, and applications, in *Acta Numerica*, Vol. 15, Cambridge University Press, pp. 1–155.
V. I. Arnol'd (1966), On the topology of three-dimensional steady flows of an ideal fluid, *J. Appl. Math. Mech.* **30**, 223–226.
P. Astrid, S. Weiland, K. Willcox and T. Backx (2008), Missing point estimation in models described by proper orthogonal decomposition, *IEEE Trans. Automat. Control* **53**, 2237–2251.
C. Audouze, F. De Vuyst and P. B. Nair (2009), Reduced-order modeling of parameterized PDEs using time–space-parameter principal component analysis, *Internat. J. Numer. Methods. Engrg* **80**, 1025–1057.
B. Baesens (2014), *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, Wiley and SAS Business Series, Wiley.
M. Barrault, Y. Maday, N. C. Nguyen and A. T. Patera (2004), An 'empirical interpolation' method: Application to efficient reduced-basis discretization of partial differential equations, *C. R. Math. Acad. Sci. Paris* **339**, 667–672.
E. Barshan, A. Ghodsi, Z. Azimifar and M. Zolghadri Jahromi (2011), Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, *Pattern Recognition* **44**, 1357–1371.
O. Bashir, K. Willcox, O. Ghattas, B. Van, B. Waanders and J. Hill (2008), Hessian-based model reduction for large-scale systems with initial-condition inputs, *Internat. J. Numer. Methods Engrg* **73**, 844–868.
C. Beattie and S. Gugercin (2009), Interpolatory projection methods for structure-preserving model reduction, *Systems Control Lett.* **58**, 225–232.
C. Beattie and S. Gugercin (2011), Structure-preserving model reduction for nonlinear port-Hamiltonian systems, in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 6564–6569.
C. Beattie, S. Gugercin and V. Mehrmann (2019), Structure-preserving interpolatory model reduction for port-Hamiltonian differential-algebraic systems. Available at arXiv:1910.05674. To appear in a Festschrift to honour the 70th birthday of A. Antoulas.
M. H. Beck, A. Jäckle, G. A. Worth and H.-D. Meyer (2000), The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propagating wavepackets, *Phys. Rep.* **324**, 1–105.
M. Belkin and P. Niyogi (2001), Laplacian eigenmaps and spectral techniques for embedding and clustering, in *Advances in Neural Information Processing Systems 14* (T. Dietterich *et al.*, eds), MIT Press.
P. Benner and T. Breiten (2012), Interpolation-based H_2-model reduction of bilinear control systems, *SIAM J. Matrix Anal. Appl.* **33**, 859–885.
P. Benner, S. Gugercin and K. Willcox (2015), A survey of projection-based model reduction methods for parametric dynamical systems, *SIAM Rev.* **57**, 483–531.
A. Berzins, J. Helmig, F. Key and S. Elgeti (2020), Standardized non-intrusive reduced order modeling using different regression models with application to complex flow problems. Available at arXiv:2006.13706.
W.-J. Beyn and V. Thümmler (2004), Freezing solutions of equivariant evolution equations, *SIAM J. Appl. Dyn. Syst.* **3**, 85–116.

C. Bigoni and J. S. Hesthaven (2020), Simulation-based anomaly detection and damage localization: An application to structural health monitoring, *Comput. Methods Appl. Mech. Engrg* **363**, 112896.
P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova and P. Wojtaszczyk (2011), Convergence rates for greedy algorithms in reduced basis methods, *SIAM J. Math. Anal.* **43**, 1457–1472.
D. A. Bistrian and I. M. Navon (2016), Randomized dynamic mode decomposition for nonintrusive reduced order modelling, *Internat. J. Numer. Methods. Engrg* **112**, 3–25.
A. Bonito, A. Cohen, R. DeVore, D. Guignard, P. Jantsch and G. Petrova (2021), Nonlinear methods for model reduction, *ESAIM Math. Model. Numer. Anal.* **55**, 507–531.
M. Brand (2006), Fast low-rank modifications of the thin singular value decomposition, *Linear Algebra Appl.* **415**, 20–30.
S. L. Brunton, B. R. Noack and P. Koumoutsakos (2020), Machine learning for fluid mechanics, *Annu. Rev. Fluid Mech.* **52**, 477–508.
S. L. Brunton, J. L. Proctor and J. N. Kutz (2016), Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proc. Natl. Acad. Sci. USA* **113**, 3932–3937.
P. Buchfink, A. Bhatt and B. Haasdonk (2019), Symplectic model order reduction with non-orthonormal bases, *Math. Comput. Appl.* **24**, 43.
A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme and G. Turinici (2012), *A priori* convergence of the greedy algorithm for the parametrized reduced basis method, *ESAIM Math. Model. Numer. Anal.* **46**, 595–603.
M. Buhmann and N. Dyn (1993), Spectral convergence of multiquadric interpolation, *Proc. Edinb. Math. Soc.* **36**, 319–333.
T. Bui-Thanh, M. Damodaran and K. Willcox (2003), Proper orthogonal decomposition extensions for parametric applications in compressible aerodynamics, in *21st Applied Aerodynamics Conference*. AIAA paper 2003-4213.
T. Bui-Thanh, K. Willcox, O. Ghattas and B. van Bloemen Waanders (2007), Goal-oriented, model-constrained optimization for reduction of large-scale systems, *J. Comput. Phys.* **224**, 880–896.
N. Cagniart, Y. Maday and B. Stamm (2019), Model order reduction for problems with large convection effects, in *Contributions to Partial Differential Equations and Applications* (B. Chetverushkin *et al.*, eds), Vol. 47 of Computational Methods in Applied Sciences, Springer, pp. 131–150.
K. Carlberg (2015), Adaptive *h*-refinement for reduced-order models, *Internat. J. Numer. Methods Engrg* **102**, 1192–1210.
K. Carlberg, Y. Choi and S. Sargsyan (2018), Conservative model reduction for finite-volume models, *J. Comput. Phys.* **371**, 280–314.
K. Carlberg, C. Farhat, J. Cortial and D. Amsallem (2013), The GNAT method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows, *J. Comput. Phys.* **242**, 623–647.
K. Carlberg, S. Guzzetti, M. Khalil and K. Sargsyan (2019), The network uncertainty quantification method for propagating uncertainties in component-based systems. Available at arXiv:1908.11476.
K. Carlberg, R. Tuminaro and P. Boggs (2015), Preserving Lagrangian structure in nonlinear model reduction with application to structural dynamics, *SIAM J. Sci. Comput.* **37**, B153–B184.

S. Chaturantabut and D. C. Sorensen (2010), Nonlinear model reduction via discrete empirical interpolation, *SIAM J. Sci. Comput.* **32**, 2737–2764.

S. Chaturantabut, C. Beattie and S. Gugercin (2016), Structure-preserving model reduction for nonlinear port-Hamiltonian systems, *SIAM J. Sci. Comput.* **38**, B837–B865.

K. K. Chen, J. H. Tu and C. W. Rowley (2012), Variants of dynamic mode decomposition: Boundary condition, Koopman, and Fourier analyses, *J. Nonlin. Sci.* **22**, 887–915.

W. Chen, J. S. Hesthaven, B. Junqiang, Y. Qiu, Z. Yang and Y. Tihao (2018), Greedy nonintrusive reduced order model for fluid dynamics, *AIAA J.* **56**, 4927–4943.

Y. Chen, J. S. Hesthaven, Y. Maday and J. Rodríguez (2009), Improved successive constraint method based *a posteriori* error estimate for reduced basis approximation of 2D Maxwell's problem, *ESAIM Math. Model. Numer. Anal.* **43**, 1099–1116.

Y. Chen, J. S. Hesthaven, Y. Maday and J. Rodríguez (2010), Certified reduced basis methods and output bounds for the harmonic Maxwell's equations, *SIAM J. Sci. Comput.* **32**, 970–996.

E. A. Christensen, M. Brøns and J. N. Sørensen (1999), Evaluation of proper orthogonal decomposition–based decomposition techniques applied to parameter-dependent non-turbulent flows, *SIAM J. Sci. Comput.* **21**, 1419–1434.

S. H. Christiansen, H. Z. Munthe-Kaas and B. Owren (2011), Topics in structure-preserving discretization, in *Acta Numerica*, Vol. 20, Cambridge University Press, pp. 1–119.

CISCO (2019), Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022.

P. G. Constantine, E. Dow and Q. Wang (2014), Active subspace methods in theory and practice: Applications to kriging surfaces, *SIAM J. Sci. Comput.*

R. Ştefănescu, A. Sandu and I. M. Navon (2014), Comparison of POD reduced order strategies for the nonlinear 2D shallow water equations, *Internat. J. Numer. Methods Fluids* **76**, 497–521.

N. N. Cuong, K. Veroy and A. T. Patera (2005), Certified real-time solution of parametrized partial differential equations, in *Handbook of Materials Modeling* (S. Yip, ed.), Springer, pp. 1529–1564.

G. Cybenko (1989), Approximation by superpositions of a sigmoidal function, *Math. Control Signals Systems* **2**, 303–314.

G. Darboux (1882), Sur le problème de Pfaff, *Bulletin des Sciences Mathématiques et Astronomiques* **6**, 14–36.

N. Demo, M. Tezzele and G. Rozza (2019), A non-intrusive approach for the reconstruction of POD modal coefficients through active subspaces, *Comptes Rendus Mécanique* **347**, 873–881.

R. A. DeVore (2017), The theoretical foundation of reduced basis methods, in *Model Reduction and Approximation: Theory and Algorithms*, SIAM, chapter 3, pp. 137–168.

R. A. DeVore, G. Petrova and P. Wojtaszczyk (2013), Greedy algorithms for reduced bases in Banach spaces, *Constr. Approx.* **37**, 455–466.

P. E. Dewdney, P. J. Hall, R. T. Schilizzi and T. J. L. W. Lazio (2009), The square kilometre array, *Proc. IEEE* **97**, 1482–1496.

P. A. M. Dirac (1930), Note on exchange phenomena in the Thomas atom, *Math. Proc. Cambridge Philos. Soc.* **26**, 376–385.

N. Discacciati and J. S. Hesthaven (2021), Modeling synchronization in globally coupled oscillatory systems using model order reduction, *Chaos* **31**, 053127.

R. Dupuis, J.-C. Jouhaud and P. Sagaut (2018), Surrogate modeling of aerodynamic simulations for multiple operating conditions using machine learning, *AIAA J.* **56**, 3622–3635.
J. Eftang, A. Patera and E. Rønquist (2010), An 'hp' certified reduced basis method for parametrized elliptic partial differential equations, *SIAM J. Sci. Comput.* **32**, 3170–3200.
V. Ehrlacher, D. Lombardi, O. Mula and F.-X. Vialard (2020), Nonlinear model reduction on metric spaces: Application to one-dimensional conservative PDEs in Wasserstein spaces, *ESAIM Math. Model. Numer. Anal.* **54**, 2159–2197.
M. Eldred and D. Dunlavy (2006), Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models, in *11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, pp. 1–20. AIAA paper 2006-7117.
C. Farhat, T. Chapman and P. Avery (2015), Structure-preserving, stability, and accuracy properties of the energy-conserving sampling and weighting method for the hyper reduction of nonlinear finite element dynamic models, *Internat. J. Numer. Methods Engrg* **102**, 1077–1110.
F. Feppon and P. F. J. Lermusiaux (2018), A geometric approach to dynamical model order reduction, *SIAM J. Matrix Anal. Appl.* **39**, 510–538.
A. Figotin and J. H. Schenker (2007), Hamiltonian structure for dispersive and dissipative dynamical systems, *J. Statist. Phys.* **128**, 969–1056.
J. P. Fink and W. C. Rheinboldt (1983), On the error behavior of the reduced basis technique for nonlinear finite element approximations, *Z. Angew. Math. Mech.* **63**, 21–28.
J. P. Fink and W. C. Rheinboldt (1984), Solution manifolds and submanifolds of parametrized equations and their discretization errors, *Numer. Math.* **45**, 323–343.
R. L. Fox and H. Miura (1971), An approximate analysis technique for design calculations, *AIAA J.* **9**, 177–179.
S. Fresca and A. Manzoni (2022), POD-DL-ROM: Enhancing deep learning-based reduced order models for nonlinear parametrized PDEs by proper orthogonal decomposition, *Comput. Methods Appl. Mech. Engrg* **388**, 114181.
R. W. Freund (2003), Model reduction methods based on Krylov subspaces, in *Acta Numerica*, Vol. 12, Cambridge University Press, pp. 267–319.
M. Ganesh, J. S. Hesthaven and B. Stamm (2012), A reduced basis method for electromagnetic scattering by multiple particles in three dimensions, *J. Comput. Phys.* **231**, 7756–7779.
Z. Gao, Q. Liu, J. S. Hesthaven, B.-S. Wang, W. S. Don and X. Wen (2021), Non-intrusive reduced order modeling of convection dominated flows using artificial neural networks with application to Rayleigh–Taylor instability, *Commun. Comput. Phys.* **30**, 97–123.
J.-F. Gerbeau and D. Lombardi (2014), Approximated Lax pairs for the reduced order integration of nonlinear evolution equations, *J. Comput. Phys.* **265**, 246–269.
Y. Gong, Q. Wang and Z. Wang (2017), Structure-preserving Galerkin POD reduced-order modeling of Hamiltonian systems, *Comput. Methods Appl. Mech. Engrg* **315**, 780–798.
I. Goodfellow, Y. Bengio and A. Courville (2016), *Deep Learning*, MIT Press.
A. Gouasmi, E. Parish, and K. Duraisamy (2017), *A priori* estimation of memory effects in reduced order modeling of nonlinear systems using the Mori–Zwanzig formalism, *Proc. Royal Soc. Ser A* **473**, 20170385.
M. A. Grepl and A. T. Patera (2005), *A posteriori* error bounds for reduced-basis approximations of parametrized parabolic partial differential equations, *ESAIM Math. Model. Numer. Anal.* **39**, 157–181.

M. A. Grepl, Y. Maday, N. C. Nguyen and A. T. Patera (2007), Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations, *M2AN Math. Model. Numer. Anal.* **41**, 575–605.
S. Gugercin, R. V. Polyuga, C. Beattie and A. van der Schaft (2012), Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems, *Automatica J. IFAC* **48**, 1963–1974.
M. D. Gunzburger (1989), *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*, Computer Science and Scientific Computing, Academic Press.
M. D. Gunzburger, J. S. Peterson and J. N. Shadid (2007), Reduced-order modeling of time-dependent PDEs with multiple parameters in the boundary data, *Comput. Methods Appl. Mech. Engrg* **196**, 1030–1047.
M. Guo and J. S. Hesthaven (2018), Reduced order modeling for nonlinear structural analysis using Gaussian process regression, *Comput. Methods Appl. Mech. Engrg* **341**, 807–826.
M. Guo and J. S. Hesthaven (2019), Data-driven reduced order modeling for time-dependent problems, *Comput. Methods Appl. Mech. Engrg* **345**, 75–99.
A. Gupta and P. F. J. Lermusiaux (2021), Neural closure models for dynamical systems, *Proc. R. Soc. A.* **477**, 20201004.
B. Haasdonk (2013), Convergence rates of the POD-greedy method, *ESAIM Math. Model. Numer. Anal.* **47**, 859–873.
B. Haasdonk and M. Ohlberger (2008), Reduced basis method for finite volume approximations of parametrized linear evolution equations, *M2AN Math. Model. Numer. Anal.* **42**, 277–302.
B. Haasdonk, M. Dihlmann and M. Ohlberger (2011), A training set and multiple bases generation approach for parameterized model reduction based on adaptive grids in parameter space, *Math. Comput. Model. Dyn. Syst.* **17**, 423–442.
E. Hairer, C. Lubich and G. Wanner (2006), *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Vol. 31 of Springer Series in Computational Mathematics, second edition, Springer.
M. Hess, A. Alla, A. Quaini, G. Rozza and M. Gunzburger (2019), A localized reduced-order modeling approach for PDEs with bifurcating solutions, *Comput. Methods Appl. Mech. Engrg* **351**, 379–403.
J. S. Hesthaven and C. Pagliantini (2021), Structure-preserving reduced basis methods for Poisson systems, *Math. Comp.* **90**, 1701–1740.
J. S. Hesthaven and S. Ubbiali (2018), Non-intrusive reduced order modeling of nonlinear problems using neural networks, *J. Comput. Phys.* **363**, 55–78.
J. S. Hesthaven, C. Pagliantini and N. Ripamonti (2022), Rank-adaptive structure-preserving model order reduction of Hamiltonian systems, *ESAIM Math. Model. Numer. Anal.* **56**, 617–650.
J. S. Hesthaven, G. Rozza and B. Stamm (2015), *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, Springer Briefs in Mathematics, Springer.
J. S. Hesthaven, B. Stamm and S. Zhang (2014), Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods, *ESAIM Math. Model. Numer. Anal.* **48**, 259–283.
S. Hijazi, G. Stabile, A. Mola and G. Rozza (2020), Data-driven POD–Galerkin reduced order model for turbulent flows, *J. Comput. Phys.* **416**, 109513.

C. Himpe and M. Ohlberger (2015), Data-driven combined state and parameter reduction for inverse problems, *Adv. Comput. Math.* **41**, 1343–1364.
R. Hiptmair (2002), Finite elements in computational electromagnetism, in *Acta Numerica*, Vol. 11, Cambridge University Press, pp. 237–339.
H. Hotelling (1933), Analysis of a complex of statistical variables into principal components, **24**, 417–441.
D. B. P. Huynh, D. J. Knezevic and A. T. Patera (2011), A Laplace transform certified reduced basis method: Application to the heat equation and wave equation, *C. R. Acad. Sci. Paris, Ser. I* **349**, 401–405.
A. Iollo and D. Lombardi (2014), Advection modes by optimal mass transfer, *Phys. Rev. E* **89**, 022923.
T. C. Ionescu and A. Astolfi (2013), Moment matching for nonlinear port Hamiltonian and gradient systems, *IFAC Proceedings Volumes* **46**, 395–399.
T. C. Ionescu, K. Fujimoto and J. M. A. Scherpen (2010), Dissipativity preserving balancing for nonlinear systems: A Hankel operator approach, *Systems Control Lett.* **59**, 180–194.
A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett and A. Zanna (2000), Lie-group methods, in *Acta Numerica*, Vol. 9, Cambridge University Press, pp. 215–365.
I. T. Jolliffe (1986), *Principal Component Analysis*, Springer Series in Statistics, Springer.
B. Karasözen and M. Uzunca (2018), Energy preserving model order reduction of the nonlinear Schrödinger equation, *Adv. Comput. Math.* **44**, 1769–1796.
B. Karasözen, S. Yıldız and M. Uzunca (2021), Structure preserving model order reduction of shallow water equations, *Math. Methods Appl. Sci.* **44**, 476–492.
K. Karhunen (1947), Über lineare Methoden in der Wahrscheinlichkeitsrechnung, *Ann. Acad. Sci. Fenn. Ser. A. I. Math.-Phys.* **1947**, 79.
O. Koch and C. Lubich (2007), Dynamical low-rank approximation, *SIAM J. Matrix Anal. Appl.* **29**, 434–454.
B. Kostant (1979), The solution to a generalized Toda lattice and representation theory, *Adv. Math.* **34**, 195–338.
B. Kramer and K. E. Willcox (2019), Nonlinear model order reduction via lifting transformations and proper orthogonal decomposition, *AIAA J.* **57**, 2297–2307.
K. Kunisch and S. Volkwein (2002), Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics, *SIAM J. Numer. Anal.* **40**, 492–515.
G. Kutyniok, P. Petersen, M. Raslan and R. Schneider (2022), A theoretical analysis of deep neural networks and parametric PDEs, *Constr. Approx.* **55**, 73–125.
J. N. Kutz, S. L. Brunton, B. W. Brunton and J. L. Proctor (2016*a*), *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*, SIAM.
J. N. Kutz, X. Fu and S. L. Brunton (2016*b*), Multiresolution dynamic mode decomposition, *SIAM J. Appl. Dyn. Syst.* **15**, 713–735.
S. Lafon and A. B. Lee (2006), Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization, *IEEE Trans. Pattern Anal. Mach. Intel.* **28**, 1393–1403.
S. Lall, P. Krysl and J. E. Marsden (2003), Structure-preserving model reduction for mechanical systems, *Phys. D* **184**, 304–318.
C. Lanczos (1949), *The Variational Principles of Mechanics*, Vol. 4 of Mathematical Expositions, University of Toronto Press.
D. Lazzaro and L. B. Montefusco (2002), Radial basis functions for the multivariate interpolation of large scattered data sets, *J. Comput. Appl. Math.* **140**, 521–536.

K. Lee and K. T. Carlberg (2020), Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders, *J. Comput. Phys.* **404**, 108973.

Y. Lee, H. Yang and Z. Yin (2017), PIV-DCNN: Cascaded deep convolutional neural networks for particle image velocimetry, *Exp. Fluids* **58**, 171.

P. A. LeGresley and J. J. Alonso (2000), Airfoil design optimization using reduced order models based on proper orthogonal decomposition, in *Fluids 2000 Conference and Exhibit*. AIAA paper 2000-2545.

B. Leimkuhler and S. Reich (2004), *Simulating Hamiltonian Dynamics*, Vol. 14 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.

K.-C. Li (1991), Sliced inverse regression for dimension reduction, *J. Amer. Statist. Assoc.* **86**, 316–327.

C. Lieberman, K. Willcox and O. Ghattas (2010), Parameter and state model reduction for large-scale statistical inverse problems, *SIAM J. Sci. Comput.* **32**, 2523–2542.

M. Loève (1955), *Probability Theory: Foundations, Random Sequences*, Van Nostrand.

C. Lubich (2008), *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*, Zurich Lectures in Advanced Mathematics, European Mathematical Society.

B. Maboudi Afkham and J. S. Hesthaven (2017), Structure preserving model reduction of parametric Hamiltonian systems, *SIAM J. Sci. Comput.* **39**, A2616–A2644.

B. Maboudi Afkham and J. S. Hesthaven (2019), Structure-preserving model-reduction of dissipative Hamiltonian systems, *J. Sci. Comput.* **81**, 3–21.

B. Maboudi Afkham, N. Ripamonti, Q. Wang and J. S. Hesthaven (2020), Conservative model order reduction for fluid flow, in *Quantification of Uncertainty: Improving Efficiency and Technology* (M. D'Elia *et al.*, eds), Vol. 137 of Lecture Notes in Computational Science and Engineering, Springer, pp. 67–99.

J. MacQueen (1967), Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, *Theory of Statistics* (L. Lecam and J. Neyman, eds), University of California Press, pp. 281–297.

Y. Maday and B. Stamm (2013), Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces, *SIAM J. Sci. Comput.* **35**, A2417–A2441.

J. E. Marsden and T. S. Ratiu (1999), *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, Vol. 17 of Texts in Applied Mathematics, second edition, Springer.

J. E. Marsden, A. Weinstein, T. Ratiu, R. Schmid and R. G. Spencer (1983), Hamiltonian systems with symmetry, coadjoint orbits and plasma physics, *Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur.* **117**, 289–340.

J. Masci, U. Meier, D. Cireşan and J. Schmidhuber (2011), Stacked convolutional auto-encoders for hierarchical feature extraction, in *Artificial Neural Networks and Machine Learning (ICANN 2011)* (T. Honkela *et al.*, eds), Springer, pp. 52–59.

R. M. Miura, C. S. Gardner and M. D. Kruskal (1968), Korteweg–de Vries equation and generalizations, II: Existence of conservation laws and constants of motion, *J. Math. Phys.* **9**, 1204–1209.

Y. Miyatake (2019), Structure-preserving model reduction for dynamical systems with a first integral, *Japan. J. Ind. Appl. Math.* **36**, 1021–1037.

R. Mojgani and M. Balajewicz (2017), Lagrangian basis method for dimensionality reduction of convection dominated nonlinear flows. Available at arXiv:1701.04343.
F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli and J. N. Kutz (2019), Data-driven modeling and learning in science and engineering, *Comptes Rendus Mécanique* **347**, 845–855.
P. J. Morrison (1980), The Maxwell–Vlasov equations as a continuous Hamiltonian system, *Phys. Lett. A* **80**, 383–386.
E. Musharbash and F. Nobile (2017), Symplectic dynamical low rank approximation of wave equations with random parameters. Technical report 18.2017, EPFL-SB-Institute of Mathematics-Mathicse.
E. Musharbash, F. Nobile and T. Zhou (2015), Error analysis of the dynamically orthogonal approximation of time dependent random PDEs, *SIAM J. Sci. Comput.* **37**, A776–A810.
R. M. Neal (1996), Priors for infinite networks, in *Bayesian Learning for Neural Networks*, Vol. 118 of Lecture Notes in Statistics, Springer, pp. 29–53.
N. C. Nguyen and J. Peraire (2016), Gaussian functional regression for output prediction: Model assimilation and experimental design, *J. Comput. Phys.* **309**, 52–68.
N.-C. Nguyen, G. Rozza and A. T. Patera (2009), Reduced basis approximation and *a posteriori* error estimation for the time-dependent viscous Burgers' equation, *Calcolo* **46**, 157–185.
E. Noether (1918), Invariante variationsprobleme, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1918**, 235–257.
A. K. Noor (1981), Recent advances in reduction methods for nonlinear problems, *Computers & Structures* **13**, 31–44.
A. K. Noor (1982), On making large nonlinear problems small, *Comput. Methods Appl. Mech. Engrg* **34**, 955–985.
A. K. Noor and J. M. Peters (1980), Reduced basis technique for nonlinear analysis of structures, *AIAA J.* **18**, 455–462.
A. K. Noor, C. D. Balch and M. A. Shibut (1984), Reduction methods for nonlinear steady-state thermal analysis, *Internat. J. Numer. Methods Engrg* **20**, 1323–1348.
G. Novati, L. Mahadevan and P. Koumoutsakos (2019), Controlled gliding and perching through deep-reinforcement-learning, *Phys. Rev. Fluids* **4**, 093902.
M. Ohlberger and S. Rave (2013), Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing, *C. R. Math. Acad. Sci. Paris* **351**, 901–906.
G. Ortali, N. Demo and G. Rozza (2022), A Gaussian process regression approach within a data-driven POD framework for engineering problems in fluid dynamics, *Math. Engrg* **4**, 1.
C. Pagliantini (2021), Dynamical reduced basis methods for Hamiltonian systems, *Numer. Math.* **148**, 409–448.
K. H. Park, S. O. Jun, S. M. Baek, M. H. Cho, K. J. Yee and D. H. Lee (2013), Reduced-order model with an artificial neural network for aerostructural design optimization, *J. Aircraft* **50**, 1106–1116.
A. T. Patera and G. Rozza (2007), *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*, MIT-Pappalardo Graduate Monographs in Mechanical Engineering, Massachusetts Institute of Technology.
B. Peherstorfer and K. Willcox (2015*a*), Dynamic data-driven reduced-order models, *Comput. Methods Appl. Mech. Engrg* **291**, 21–41.

B. Peherstorfer and K. Willcox (2015*b*), Online adaptive model reduction for nonlinear systems via low-rank updates, *SIAM J. Sci. Comput.* **37**, A2123–A2150.
B. Peherstorfer and K. Willcox (2016*a*), Data-driven operator inference for nonintrusive projection-based model reduction, *Comput. Methods Appl. Mech. Engrg* **306**, 196–215.
B. Peherstorfer and K. Willcox (2016*b*), Dynamic data-driven model reduction: Adapting reduced models from incomplete data, *Adv. Model. Simul. Engrg Sci.* **3**, 11.
B. Peherstorfer, D. Butnaru, K. Willcox and H.-J. Bungartz (2014), Localized discrete empirical interpolation method, *SIAM J. Sci. Comput.* **36**, A168–A192.
L. Peng and K. Mohseni (2016*a*), Geometric model reduction of forced and dissipative Hamiltonian systems, in *2016 IEEE 55th Conference on Decision and Control (CDC)*, IEEE, pp. 7465–7470.
L. Peng and K. Mohseni (2016*b*), Symplectic model reduction of Hamiltonian systems, *SIAM J. Sci. Comput.* **38**, A1–A27.
J. S. Peterson (1989), The reduced basis method for incompressible viscous flow calculations, *SIAM J. Sci. Statist. Comput.* **10**, 777–786.
F. Pichi, F. Ballarin, G. Rozza and J. S. Hesthaven (2021), An artificial neural network approach to bifurcating phenomena in computational fluid dynamics. Available at arXiv:2109.10765.
R. V. Polyuga and A. van der Schaft (2010), Structure preserving model reduction of port-Hamiltonian systems by moment matching at infinity, *Automatica J. IFAC* **46**, 665–672.
R. V. Polyuga and A. van der Schaft (2011), Structure preserving moment matching for port-Hamiltonian systems: Arnoldi and Lanczos, *IEEE Trans. Automat. Control* **56**, 1458–1462.
T. A. Porsching and M. L. Lee (1987), The reduced basis method for initial value problems, *SIAM J. Numer. Anal.* **24**, 1277–1287.
C. Prud'homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, A. T. Patera and G. Turinici (2002), Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods, *J. Fluids Engrg* **124**, 70–80.
A. Quarteroni, A. Manzoni and F. Negri (2016), *Reduced Basis Methods for Partial Differential Equations*, Springer.
A. Quarteroni, G. Rozza and A. Manzoni (2011), Certified reduced basis approximation for parametrized PDE and applications, *J. Math Industry* **1**, 3.
M. Raissi, P. Perdikaris and G. E. Karniadakis (2019), Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.* **378**, 686–707.
C. E. Rasmussen and C. K. I. Williams (2005), *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, MIT Press.
S. S. Ravindran (2000), A reduced-order approach for optimal control of fluids using proper orthogonal decomposition, *Internat. J. Numer. Methods Fluids* **34**, 425–448.
T. Reis and T. Stykel (2008), A survey on model reduction of coupled systems, in *Model Order Reduction: Theory, Research Aspects and Applications*, Vol. 13 of Mathematics in Industry, Springer, pp. 133–155.
J. Reiss, P. Schulze, J. Sesterhenn and V. Mehrmann (2018), The shifted proper orthogonal decomposition: A mode decomposition for multiple transport phenomena, *SIAM J. Sci. Comput.* **40**, A1322–A1344.
M. Rewienski (2003), A trajectory piecewise-linear approach to model order reduction of nonlinear dynamical systems. PhD thesis, Massachusetts Institute of Technology.

D. Rim, S. Moe and R. J. LeVeque (2018), Transport reversal for model reduction of hyperbolic partial differential equations, *SIAM/ASA J. Uncertain. Quantif.* **6**, 118–150.
F. Rosenblatt (1958), The perceptron: A probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65 6**, 386–408.
C. W. Rowley and J. E. Marsden (2000), Reconstruction equations and the Karhunen–Loève expansion for systems with symmetry, *Phys. D* **142**, 1–19.
C. W. Rowley, I. G. Kevrekidis, J. E. Marsden and K. Lust (2003), Reduction and reconstruction for self-similar dynamical systems, *Nonlinearity* **16**, 1257–1275.
G. Rozza (2005), Reduced-basis methods for elliptic equations in sub-domains with *a posteriori* error bounds and adaptivity, *Appl. Numer. Math.* **55**, 403–424.
G. Rozza (2014), Fundamentals of reduced basis method for problems governed by parametrized PDEs and applications, in *Separated Representation and PGD Based Model Reduction: Fundamentals and Applications* (F. Chinesta and P. Ladevèze, eds), Vol. 554 of CISM International Centre for Mechanical Sciences, Springer, pp. 153–227.
G. Rozza, D. B. P. Huynh and A. T. Patera (2008), Reduced basis approximation and *a posteriori* error estimation for affinely parametrized elliptic coercive partial differential equations: Application to transport and continuum mechanics, *Arch. Comput. Methods Engrg* **15**, 229–275.
G. Rozza, D. B. P. Huynh, N. C. Nguyen and A. T. Patera (2009), Real-time reliable simulation of heat transfer phenomena, in *ASME 2009 Heat Transfer Summer Conference collocated with the InterPACK09 and 3rd Energy Sustainability Conferences*, American Society of Mechanical Engineers, pp. 851–860.
G. Rozza, H. Malik, N. Demo, M. Tezzele, M. Girfoglio, G. Stabile and A. Mola (2018), Advances in reduced order methods for parametric industrial problems in computational fluid dynamics, in *ECCOMAS ECFD 7: Proceedings of 6th European Conference on Computational Mechanics (ECCM 6) and 7th European Conference on Computational Fluid Dynamics (ECFD 7)* (R. Owen *et al.*, eds), pp. 59–76.
D. E. Rumelhart, G. E. Hinton and R. J. Williams (1986), Learning representations by back-propagating errors, *Nature* **323**, 533–536.
D. Ryckelynck (2009), Hyper-reduction of mechanical models involving internal variables, *Internat. J. Numer. Methods Engrg* **77**, 75–89.
A. Salam (2005), On theoretical and numerical aspects of symplectic Gram–Schmidt-like algorithms, *Numer. Algorithms* **39**, 437–462.
J. W. Sammon (1969), A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.* **C-18**, 401–409.
B. Sanderse (2020), Non-linearly stable reduced-order models for incompressible flow with energy-conserving finite volume methods, *J. Comput. Phys.* **421**, 109736.
J. M. Sanz-Serna and M. P. Calvo (1994), *Numerical Hamiltonian Problems*, Vol. 7 of Applied Mathematics and Mathematical Computation, Chapman & Hall.
T. P. Sapsis and P. F. J. Lermusiaux (2009), Dynamically orthogonal field equations for continuous stochastic dynamical systems, *Phys. D* **238**, 2347–2360.
A. Schein, K. T. Carlberg and M. J. Zahr (2021), Preserving general physical properties in model reduction of dynamical systems via constrained-optimization projection, *Internat. J. Numer. Methods. Engrg* **122**, 3368–3399.
P. J. Schmid (2010), Dynamic mode decomposition of numerical and experimental data, *J. Fluid Mech.* **656**, 5–28.

B. Schölkopf, A. Smola and K.-R. Müller (1997), Kernel principal component analysis, in *Artificial Neural Networks (ICANN'97)* (W. Gerstner *et al.*, eds), Springer, pp. 583–588.

O. Semeraro, G. Bellani and F. Lundell (2012), Analysis of time-resolved PIV measurements of a confined turbulent jet using POD and Koopman modes, *Exp. Fluids* **53**, 1203–1220.

J. Sesterhenn and A. Shahirpour (2019), A characteristic dynamic mode decomposition, *Theoret. Comput. Fluid Dynamics* **33**, 281–305.

M. Shashkov (1996), *Conservative Finite-Difference Methods on General Grids*, Symbolic and Numeric Computation series, CRC Press.

P. Stegeman, A. Ooi and J. Soria (2015), Proper orthogonal decomposition and dynamic mode decomposition of under-expanded free-jets with varying nozzle pressure ratios, in *Instability and Control of Massively Separated Flows* (V. Theofilis and J. Soria, eds), Vol. 107 of Fluid Mechanics and its Applications, Springer, pp. 85–90.

T. Taddei (2020), A registration method for model order reduction: Data compression and geometry reduction, *SIAM J. Sci. Comput.* **42**, A997–A1027.

J. B. Tenenbaum, V. de Silva and J. C. Langford (2000), A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319–2323.

M. Tezzele, F. Ballarin and G. Rozza (2018*a*), Combined parameter and model reduction of cardiovascular problems by means of active subspaces and POD-Galerkin methods, in *Mathematical and Numerical Modeling of the Cardiovascular System and Applications*, SEMA SIMAI Springer Series, Springer, pp. 185–207.

M. Tezzele, N. Demo, M. Gadalla, A. Mola and G. Rozza (2018*b*), Model order reduction by means of active subspaces and dynamic mode decomposition for parametric hull shape design hydrodynamics, in *Technology and Science for the Ships of the Future: Proceedings of NAV 2018: 19th International Conference on Ship & Maritime Research*, IOS Press, pp. 569–576.

M. Uzunca, B. Karasözen and S. Yıldız (2021), Structure-preserving reduced-order modeling of Korteweg–de Vries equation, *Math. Comput. Simul.* **188**, 193–211.

L. van der Maaten and G. Hinton (2008), Visualizing data using t-SNE, *J. Mach. Learning Res.* **9**, 2579–2605.

A. van der Schaft (2006), Port-Hamiltonian systems: An introductory survey, in *International Congress of Mathematicians 2006*, European Mathematical Society, pp. 1339–1365.

K. Veroy, C. Prud'homme, D. V. Rovas and A. T. Patera (2003), *A posteriori* error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations, in *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*. AIAA paper 2003-3847.

S. Walton, O. Hassan and K. Morgan (2013), Reduced order modelling for unsteady fluid flow using proper orthogonal decomposition and radial basis functions, *Appl. Math. Model.* **37**, 8930–8945.

Q. Wang, J. S. Hesthaven and D. Ray (2019), Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem, *J. Comput. Phys.* **384**, 289–307.

Q. Wang, N. Ripamonti and J. S. Hesthaven (2020), Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori–Zwanzig formalism, *J. Comput. Phys.* **410**, 109402.

Y. Wang, H. Yao and S. Zhao (2016), Auto-encoder based dimensionality reduction, *Neurocomput.* **184**, 232–242.
G. Weickum, M. S. Eldred and K. Maute (2008), A multi-point reduced-order modeling approach of transient structural dynamics with application to robust design optimization, *Struct. Multidiscip. Optim.* **38**, 599.
K. Q. Weinberger, F. Sha and L. K. Saul (2004), Learning a kernel matrix for nonlinear dimensionality reduction, in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML '04)*, Association for Computing Machinery, p. 106.
G. Welper (2017), Interpolation of functions with parameter dependent jumps by transformed snapshots, *SIAM J. Sci. Comput.* **39**, A1225–A1250.
K. Willcox and J. Peraire (2002), Balanced model reduction via the proper orthogonal decomposition, *AIAA J.* **40**, 2323–2330.
D. Wirtz, D. C. Sorensen and B. Haasdonk (2014), *A posteriori* error estimation for DEIM reduced nonlinear dynamical systems, *SIAM J. Sci. Comput.* **36**, A311–A338.
T. Wolf, B. Lohmann, R. Eid and P. Kotyczka (2010), Passivity and structure preserving order reduction of linear port-Hamiltonian systems using Krylov subspaces, *Eur. J. Control* **16**, 401–406.
D. Xiao, F. Fang, C. Pain and G. Hu (2015), Non-intrusive reduced order modelling of the Navier–Stokes equations based on RBF interpolation, *Internat. J. Numer. Methods Fluids* **79**, 580–595.
H. Xu (2003), An SVD-like matrix decomposition and its applications, *Linear Algebra Appl.* **368**, 1–24.
M. Yano, A. T. Patera and K. Urban (2014), A space-time hp-interpolation-based certified reduced basis method for Burgers' equation, *Math. Models Methods Appl. Sci.* **24**, 1903–1935.
J. Yu, C. Yan and M. Guo (2019), Non-intrusive reduced-order modeling for fluid problems: A brief review, *Proc. Inst. Mech. Engrs G. J. Aerosp. Engrg* **233**, 5896–5912.
O. Zahm, P. Constantine, C. Prieur and Y. Marzouk (2020), Gradient-based dimension reduction of multivariate vector-valued functions, *SIAM J. Sci. Comput.* **42**, A534–A558.
M. Zancanaro, M. Mrosek, G. Stabile, C. Othmer and G. Rozza (2021), Hybrid neural network reduced order modelling for turbulent flows with geometric parameters, *Fluids* **6**, 296.
Z. Zhang and H. Zha (2004), Principal manifolds and nonlinear dimension reduction via local tangent space alignment, *SIAM J. Sci. Comput.* **26**, 313–338.
R. Zimmermann, B. Peherstorfer and K. Willcox (2018), Geometric subspace updates with applications to online adaptive nonlinear model reduction, *SIAM J. Matrix Anal. Appl.* **39**, 234–261.
X. Zou, M. Conti, P. Díez and F. Auricchio (2018), A non-intrusive proper generalized decomposition scheme with application in biomechanics, *Internat. J. Numer. Methods. Engrg* **113**, 230–251.